AI端侧落地元年，先进半导体大有可为

2023/11/23

行业评级：增持

姓名：王聪（分析师）
邮箱：wangcong@gtjas.com
电话：13918566336
证书编号：S0880517010002

姓名：舒迪（分析师）
邮箱：shudi@gtjas.com
电话：13381980096
证书编号：S0880521070002

姓名：文紫妍（分析师）
邮箱：wenziyan@gtjas.com
电话：18818210951
证书编号：S0880523070001

姓名：陈豪杰（研究助理）
邮箱：chenhaojie026733@gtjas.com
电话：18811361286
证书编号：S0880122080153

姓名：刘晓（研究助理）
邮箱：liuxiao026731@gtjas.com
电话：18225512797
证书编号：S0880122070050

投资要点（行业评级：增持）

01

半导体行业周期反转，制造环节国内先进制程急需突破

2023年以来全球半导体销售额月度销售连续7个月环比增长，库存压力下降，行业呈现恢复态势。制造环节中，我国在成熟制程领域市占率较高，先进制程急需突破。

02

Vision Pro引领空间运算新时代，AI端侧进入落地元年

Vision Pro强调与现实世界的交流和融合，而非隔绝；将在方方面面改变人们的生活与生产，将引领空间运算时代。多家大厂积极布局AI终端，引领消费电子新浪潮。

03

AI带来算力需求提升，chiplet规模化落地可期

AI应用需要大量算力作为支撑，算力芯片等产业链公司迎来发展机遇。Chiplet生态有望逐步落地，价值量的增长点主要集中在封测端和材料端。

04

推荐标的

立讯精密、中芯国际、韦尔股份。

05

风险提示

下游需求恢复不及预期；AI产业化进度不及预期；国际贸易风险

目录 | CONTENTS

01

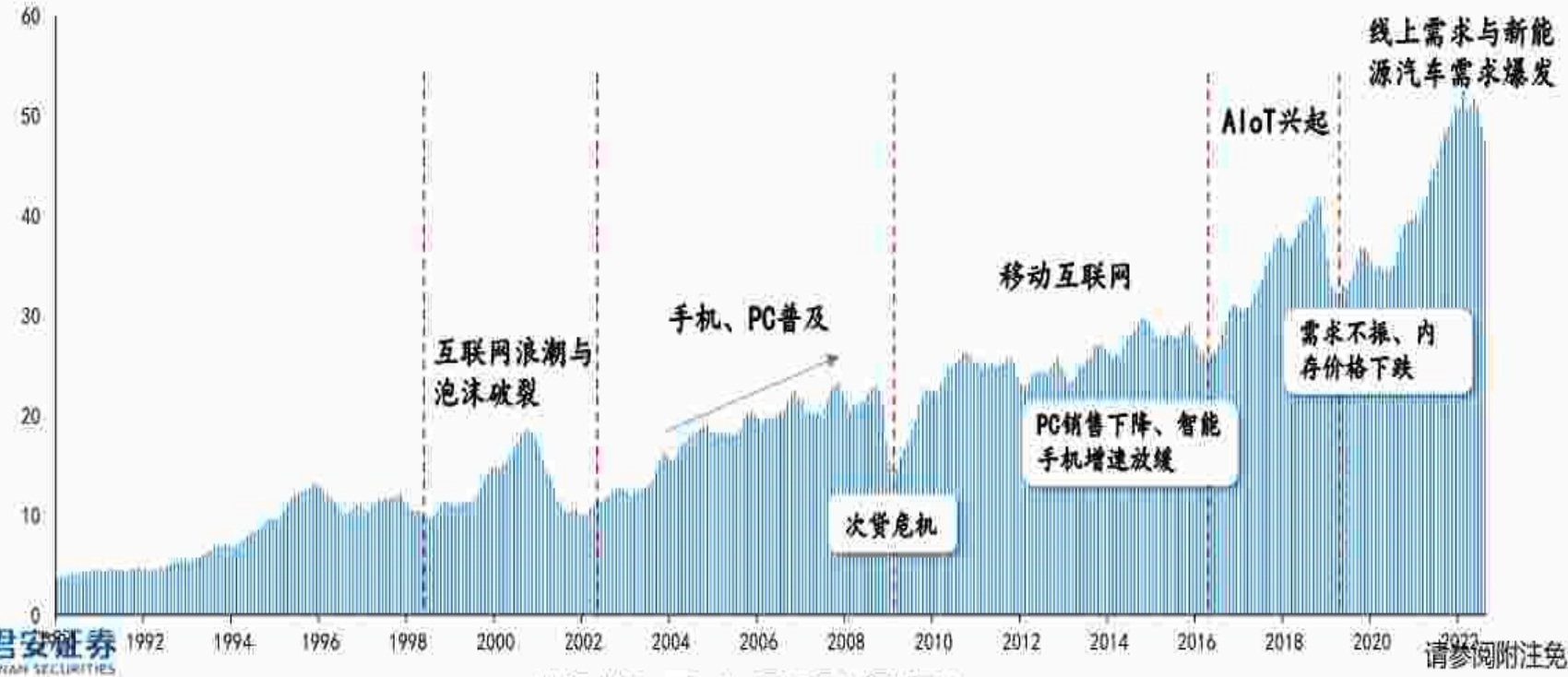
半导体行业周期反转，景气度向上

半导体是具备明显周期成长属性的行业，本质上是长短2个周期相互嵌套：即“科技创新周期+行业供需周期”。

（1）1994~2004，市场规模达2104.3亿美金：由PC/laptop市场驱动
（2）2004~2013，市场规模达3034.8亿美金：CAGR 5%，由智能手机市场驱动
（3）2013~2017，市场规模达4050.8亿美金：CAGR 7%，由数据市场驱动
（4）2017~2021，市场规模达5559亿美金：CAGR 6.2%，由5G+AIOT+国产化驱动

半导体是周期波动中的成长行业

全球半导体销售额月度数据（十亿美元）



数据来源：Wind，国泰君安证券研究

数据来源：Wind，国泰君安证券研究

全球半导体销售：全球半导体销售额月度销售连续7个月环比增长，库存压力下降，行业呈现恢复态势

2023三季度全球半导体销售额环比增长6.3%，连续第七次环比增长。根据美国半导体工业协会（SIA），2023年9月全球半导体行业销售情况。2023年9月的全球半导体销售额与2023年8月相比增长1.9%，2023年第三季度全球半导体销售总额为1347.亿美元，较2023年第二季度增长6.3%，较2022年第三季度下降4.5%，虽然同比有所下降，但随着全球半导体产业泡沫消散，库存渐趋减少，行业正呈现出逐渐恢复态势，随着人工智能、个人电脑、智能手机等相关需求增加，明年复苏态势料将延续下去。据半导体产业咨询公司nternational Business Strategies预计，今年全球半导体销售额同比将减少12%，但明年料反弹11%以上，达到5500亿美元。

全球半导体销售额同比回暖

数据来源：Wind，国泰君安证券研究

中国半导体销售额同比回暖

数据来源：Wind，国泰君安证券研究

01 销售回暖终端销售出货量增加

核心终端回暖情况：PC出货量连续两个季度增加，渠道库存健康，受益于生成式人工智能和设备换机周期，PC市场将迎来复苏

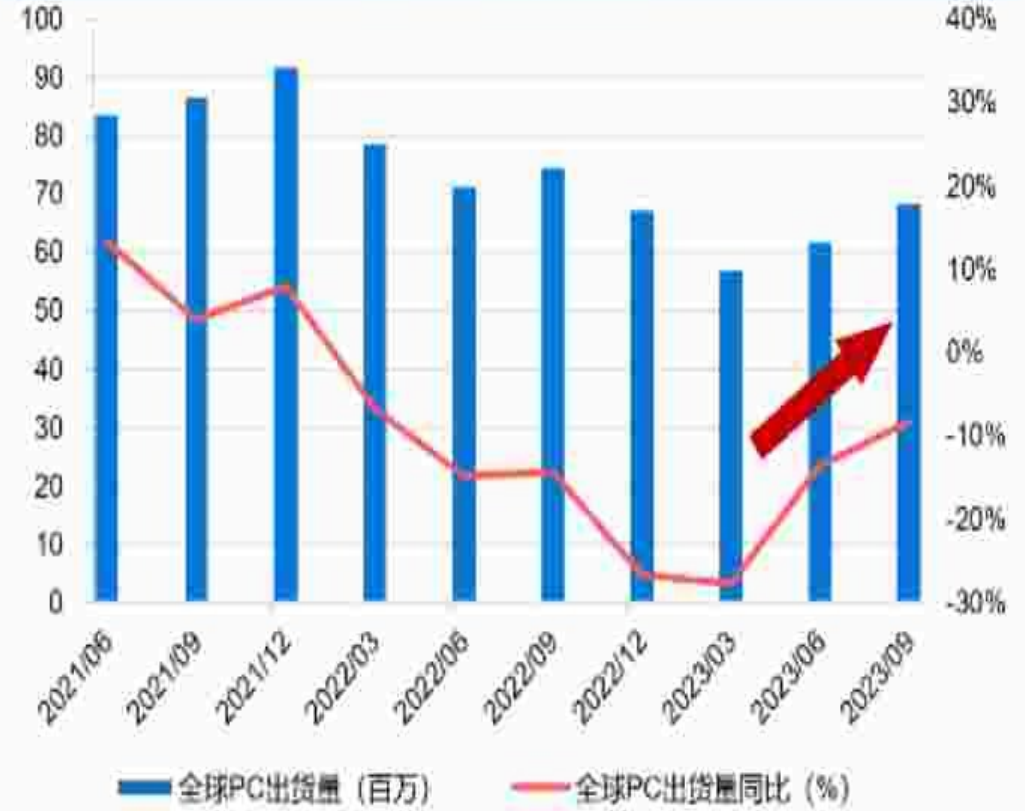
PC

手机

面板

全球PC出货量连续两个季度增加，库存呈下降趋势，市场逐渐走出低谷。根据IDC数据，2023年第三季度PC出货量继续螺旋式下降，全球出货量6,820万台，同比下降7.6%，同比下降速度趋缓，这表明市场已经走出低谷。库存方面，近几个月PC库存也呈下降趋势，在大多数渠道中都接近健康水平。PC行业正缓慢复苏，因为设备换机周期到来且Windows 10系统停用将有助于推动2024年下半年及以后的销售，同时生成式人工智能也将驱动PC单价和销量的提升。

全球PC出货量回暖



数据来源：IDC，国泰君安证券研究

中国PC出货量回暖

数据来源：IDC，国泰君安证券研究

核心终端回暖情况：手机智能手机降幅收窄至1%，库存水平相对健康，未来预计将实现温和增长

PC

手机

全球智能手机降幅收窄，库存状况得到改善。根据Canalys数据，2023年第三季度手机厂商库存状况得到改善，并在三季度推出新品，出货量达到2.95亿部，降幅收窄至1%。Canalys预计2024年全球智能手机市场将在谨慎态势下实现温和增长。各厂商在2023年末预计会有相对健康的库存水平，并有足够的空间为迎接潜在的需求复苏而重建库存。

国内手机市场出货量开始同比转正，市场触底新换机周期到来。据中国信通院官网数据，2023年1-8月，国内市场手机总体出货量累计1.67亿部，同比下降4.5%，但是，8月份的出货量已经开始同比转涨0.03%。

全球手机出货量回暖

数据来源：Canalys，国泰君安证券研究

中国手机出货量回暖

数据来源：信通院，国泰君安证券研究

核心终端回暖情况：服务器开启触底反弹，传统云服务需求复迹象低于AI服务器

PC

手机

服务器

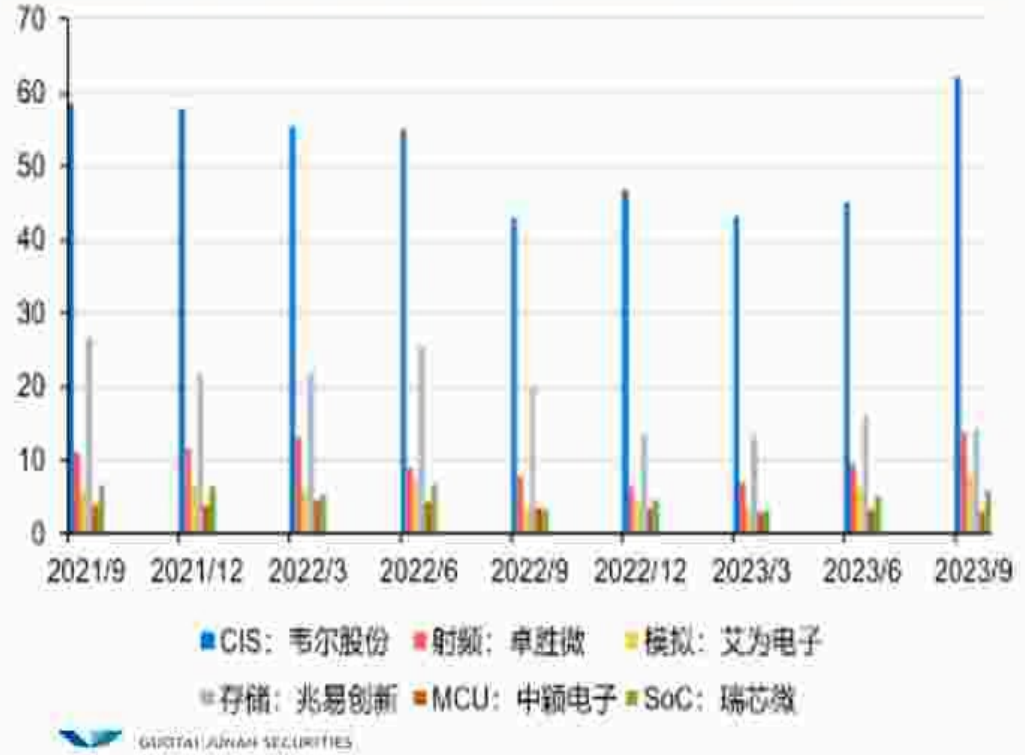
传统服务器市场修复强度弱，AI服务器出货量增长但无法反转整体疲态。根据TrendForce，2023年三季度全球服务器出货量温和反弹1.5%。传统云服务需求复苏迹象降低，北美CSP加大对高端AI服务器的投入，其价格比普通服务器贵数十倍，导致整体服务器出货量环比小幅下降。AI服务器出货量同比增长将逾10%。，但由于AI服务器目前占整体服务器出货比例仍不及1成，故尚无法反转整体服务器疲弱态势。随着第四季度服务器品牌新产品出货量持续增加，美国和中国的数据中心运营商将增加通用服务器和人工智能服务器的采购。

数据来源：Trendforce，国泰君安证券研究

01 板块回暖情况：收入端单季度同比修复力度较大

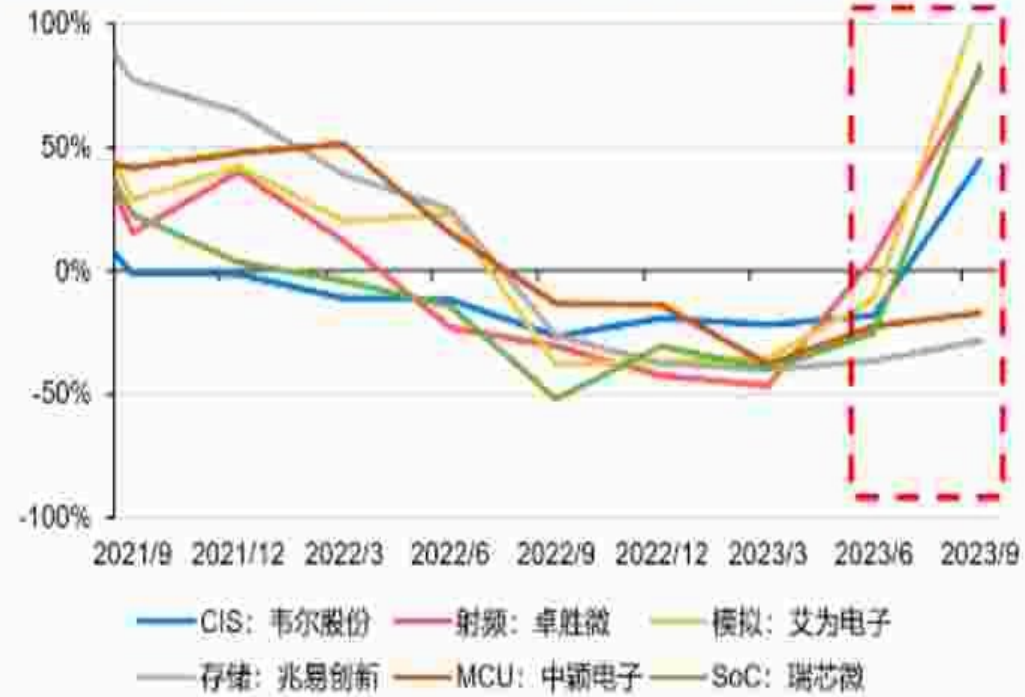
设计板块 **设计板块代表公司单季度营收修复，环比数据增长较大，同比降幅收窄。**收入端看，随着AI大模型从文字、语音到视频图像等应用不断落地和智能手机/智能穿戴/PC等终端需求的温和回暖，三季度芯片设计板块营收实现较高增长，环比数据实现较高增长，同比数据部分公司实现正增长，未实现正增长的公司同比降幅也有所收窄。展望未来，生成式AI也将继续催化端侧人工智能及AI服务器需求，进一步存进公司营收增长。

总量角度：设计板块代表公司单季度营收（亿元）



数据来源：Wind，国泰君安证券研究

边际角度：模拟芯片公司单季度营收同比回暖（%）



数据来源：Wind，国泰君安证券研究

设计板块

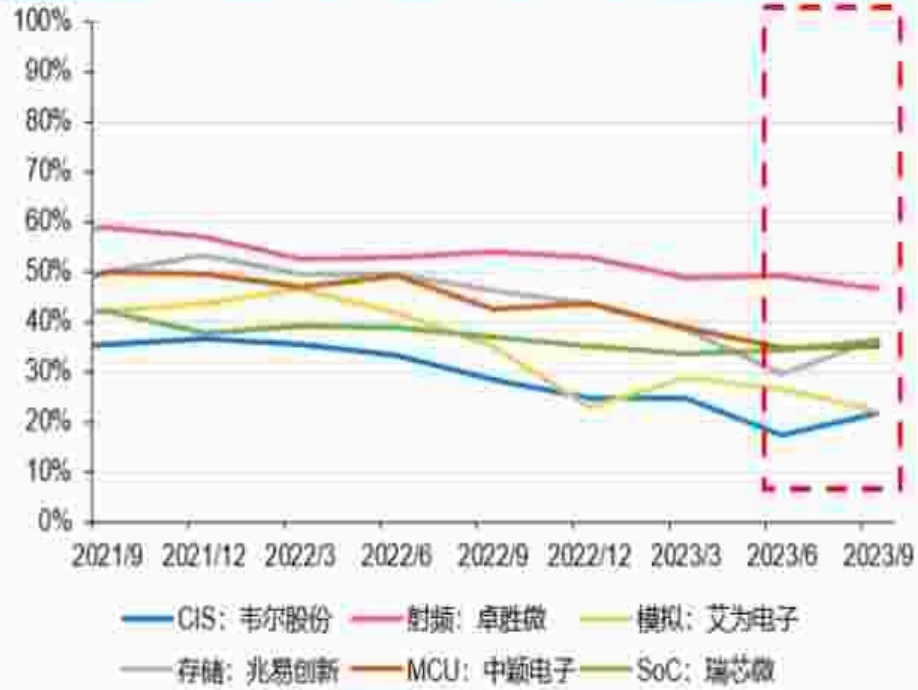
存货端，设计板块代表公司存货端去库显著，存货周转天数显著下降。2023Q3，设计板块行业库存环比状况较2022年同期有明显改善，继续2023Q2的库存去化趋势，库存压力进一步缓解预计，据「法国外贸银行亚洲制造业库存压力指标」分析显示，亚洲电子、半导体库存压力自2023年7月以来有所减轻。

利润端，设计板块代工公司毛利率修复尚不显著。部分细分领域由于竞争加剧，降价清库存等原因毛利率环比继续下滑。

存货角度：设计板块代表公司存货周转天数（天）

数据来源：Wind，国泰君安证券研究

利润角度：设计板块代表公司销售毛利率（%）



数据来源：Wind，国泰君安证券研究

设计板块回暖情况：MCU芯片-分销商数据

		23Q1			23Q2			23Q3		
厂商	产品类别	货期(周)	货期趋势	价格趋势	货期(周)	货期趋势	价格趋势	货期（周）	货期趋势	价格趋势
Infineon + Cypress	8位MCU	45-52	维稳	维稳	26-52	维稳	维稳	26-52	维稳	维稳
	32位MCU	45	延长	维稳	26-52	缩短	维稳	10-52	缩短	维稳
	汽车MCU	32-45	维稳	维稳	32-45	缩短	维稳	32-45	维稳	维稳
Infineon	汽车MCU	紧缺	维稳	维稳	紧缺	维稳	维稳	紧缺	维稳	维稳
Microchip	8位MCU	36-52+	缩短	上涨	36-52+	缩短	维稳	26-52+	缩短	维稳
	32位MCU	36-52+	缩短	上涨	36-52+	缩短	维稳	26-52+	缩短	维稳
NXP	8位MCU	35-52	缩短	维稳	35-52	缩短	维稳	26-52	缩短	维稳
	32位MCU	26-52	缩短	维稳	26-52	缩短	维稳	13-52	缩短	维稳
	汽车MCU	35-52	维稳	维稳	35-52	维稳	维稳	35-52	维稳	维稳
Renesas	8位MCU	40	缩短	上涨	18-24	缩短	维稳	18-24	缩短	维稳
	32位MCU	40	缩短	上涨	18-24	缩短	维稳	18	缩短	维稳
	汽车MCU	45	维稳	维稳	45	维稳	维稳	45	维稳	维稳
ST	8位MCU	48	缩短	维稳	35-52	缩短	维稳	35-52	缩短	维稳
	32位MCU	20-26	缩短	维稳	16-20	缩短	维稳	10-28	缩短	维稳
	汽车MCU	40-52	维稳	维稳	40-52	维稳	维稳	40-52	维稳	维稳

数据来源：富昌电子官网，国泰君安证券研究

代工板块回暖情况：晶圆代工厂

营收角度

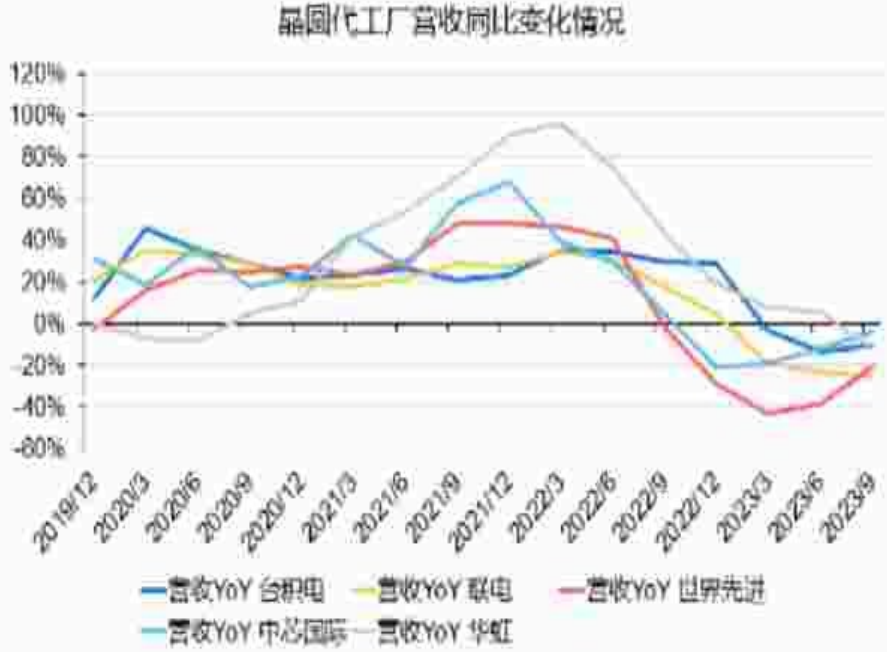
整体而言，第三季全球前十大晶圆代工产值将有望自谷底反弹，后续缓步成长。3Q23台积电营收172.86亿美元，环比+11.34%，同比呈边际改善趋势。其他晶圆厂如联电、世界先进、中芯国际、华虹营收同比变化均呈边际改善趋势。

产能利用率角度

总量：晶圆代工厂营收情况

数据来源：Wind，国泰君安证券研究

边际：晶圆代工厂营收同比变化情况



数据来源：Wind，国泰君安证券研究

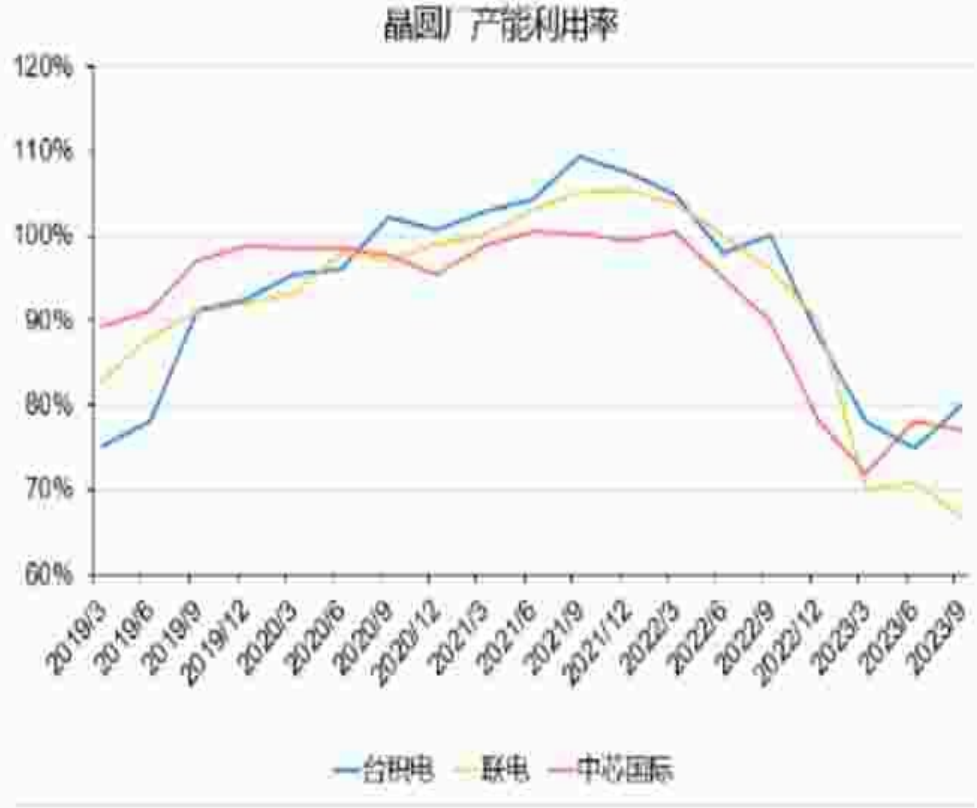
01 板块回暖情况：代工板块

代工板块回暖情况：

- 营收角度
- **产能利用率角度**

晶圆代工产线正逐步恢复产线订单量，产能利用率正逐步回升。从2Q23以来部分晶圆厂产能利用率环比复苏，如SMIC 2Q23产能利用率78%，环比+10pcts，主要系国内手机终端消费电子芯片库存开始下降，客户逐步恢复下单需求。伴随下游手机、PC等终端需求的缓慢复苏，晶圆厂产能利用率有望逐步回升。

总量：晶圆厂产能利用率情况



数据来源：公司公告，国泰君安证券研究

边际：晶圆厂产能利用率同比变化情况

数据来源：公司公告，国泰君安证券研究

封测板块回暖情况：

营收角度

2023下半年以来封测板块营收和同比变化数据均出现大面积增长。3Q23长电科技、通富微电、华天科技、甬矽电子营收分别82.57亿元、59.99亿元、29.80亿元、6.48亿元，同比变化-10.09%、+4.29%、+2.55%、+11.92%，营收同比变化呈边际改善趋势。

总量：封测公司营收情况

数据来源：Wind，国泰君安证券研究

边际：封测公司营收同比变化情况

数据来源：Wind，国泰君安证券研究

封测板块回暖情况：

利润角度

2023下半年以来封测板块毛利率数据呈现边际改善。3Q23长电科技、通富微电、华天科技、甬矽电子毛利率分别14.36%、12.72%、9.56%、16.82%，环比呈逐季改善趋势。

总量：封测公司毛利率情况

数据来源：Wind，国泰君安证券研究

边际：封测公司毛利率同比变化情况

数据来源：Wind，国泰君安证券研究

价格回暖情况：NAND晶圆涨势强劲

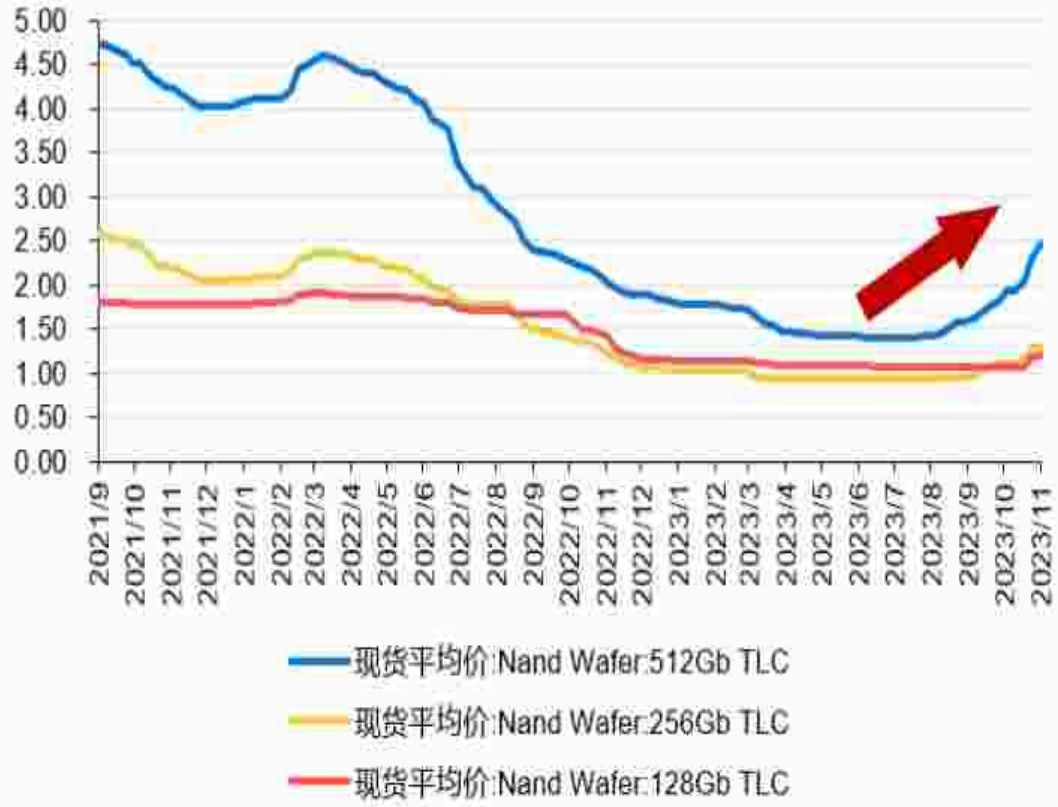
上游晶圆

中游模块

下游终端

以NAND晶圆为例，从8月起涨势强劲。根据TrendForce数据，NAND芯片价格自8月起涨，4Q23企业级固态硬盘合约价格可望上涨约5%-10%；用户端固态硬盘方面，随着供应商议价能力提高，高低阶用户端固态硬盘产品可望同步上涨，预计四季度合约价将扬升8%至13%。

NAND晶圆价格触底回升（单位：美元）



数据来源：Wind，国泰君安证券研究

数据来源：Wind，国泰君安证券研究

价格回暖情况：DRAM价格全面上涨

自2023下半年起，各型号DRAM价格开始全面上涨。以DDR4(16GB,3200Mbps)为例，现货均价从2.9美元左右涨至3.2美元左右。展望2024年第一季，预估整体存储器的涨势将延续，据TrendForce数据，4Q23Mobile DRAM合约价季涨幅预估将扩大至13~18%。

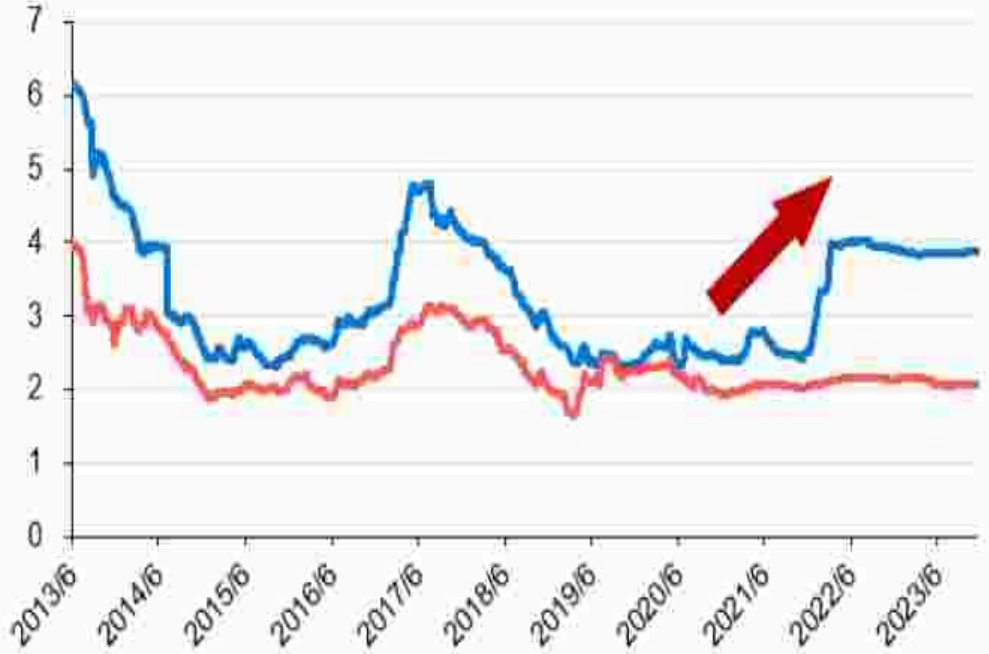
DDR4价格触底回升（单位：美元）

数据来源：Wind，国泰君安证券研究

价格回暖情况：NAND价格上涨

伴随原厂积极减产和控制供应，存储供需状况得到明显改善。由于NAND wafer和成品端价格全面反弹，NAND指数快速收复失地，以据CFM闪存市场数据显示，9月初至10月底，NAND指数涨幅达26.4%。展望后市，根据TrendForce数据，预估NAND Flash方面，eMMC、UFS第四季合约价涨幅约10~15%，1Q24NAND Flash合约价仍会续涨。

NAND价格触底回升（单位：美元）



数据来源：Wind，国泰君安证券研究

价格回暖情况：国内电视终端市场掀起涨价潮

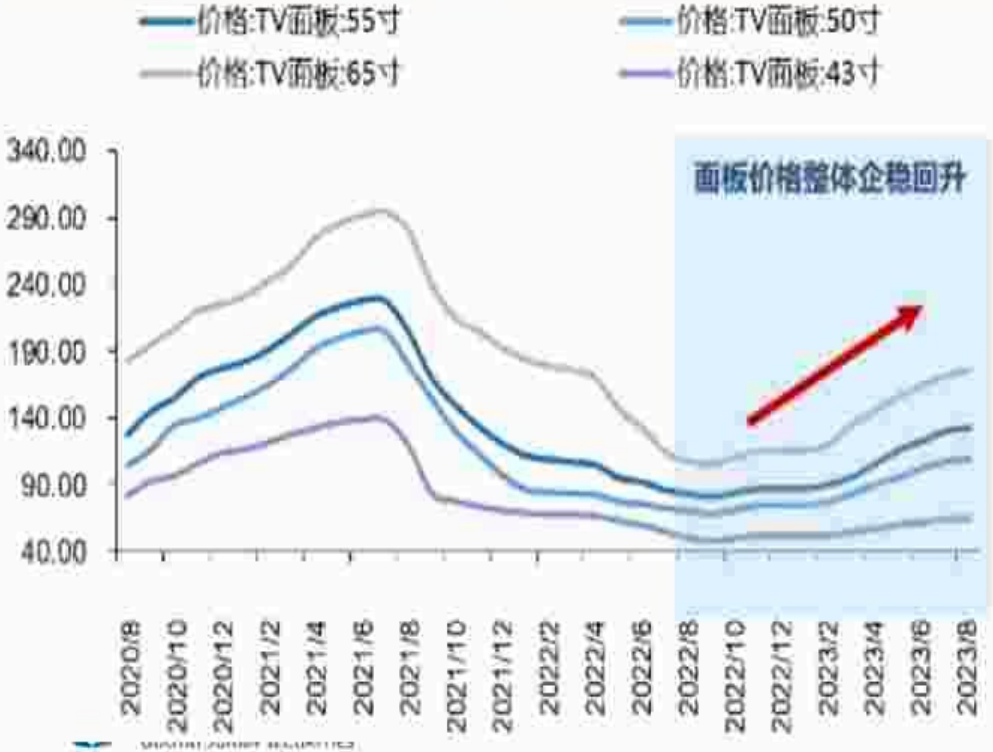
上游面板

中游模组

下游终端

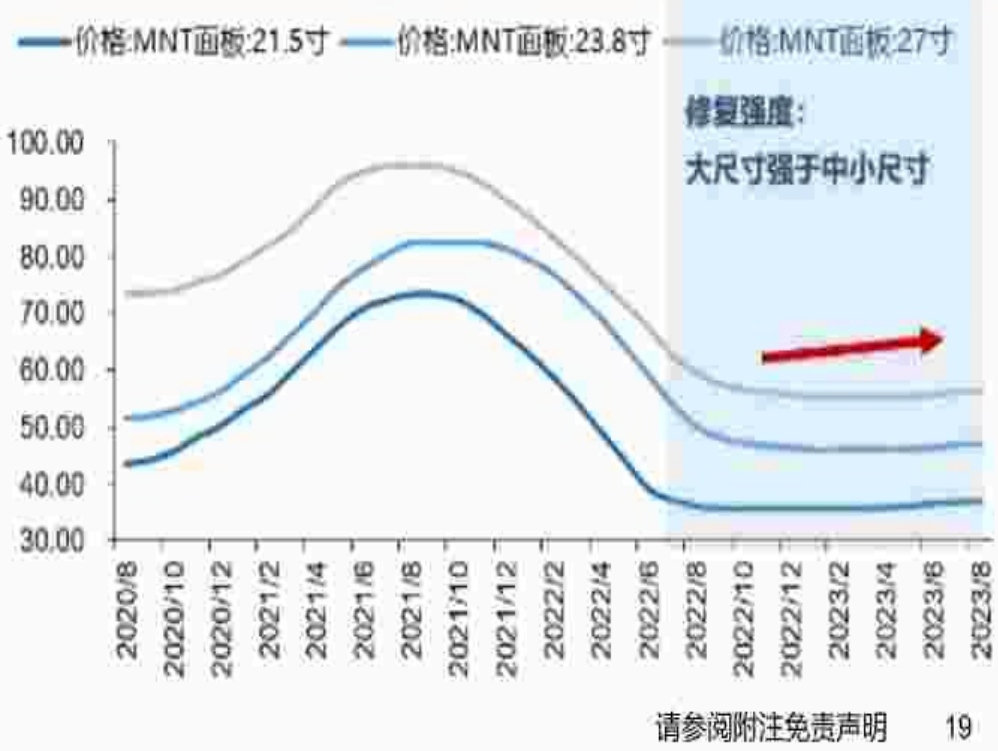
下游终端方面，面板价格连续7个月上涨，带动液晶电视价格上涨。根据洛图科技，上游液晶电视面板从2月起，单边上涨长达7个月，至第三季度末达到今年以来的最高值，以55寸为例，积累涨幅高达52%。受核心部件面板成本持续上涨的影响，电视终端市场在第三季度被掀起涨价潮。

TV面板价格回升（单位：元）



数据来源：Wind，国泰君安证券研究

MNT面板价格回升（单位：元）



数据来源：Wind，国泰君安证券研究

02

Vision Pro，开启空间运算新时代

- **Vision Pro通过VST（video see through）实现MR（数字与现实的混合）。**Vision Pro 通过传感器实时感知现实环境，将数字化的现实运算后发给显示屏幕，通过光机系统真实的发送给佩戴者。

通过调节Vision Pro的旋钮，用户可以获得不同的沉浸度体验

完全沉浸时，EyeSight技术让走近的同伴出现在用户视野中

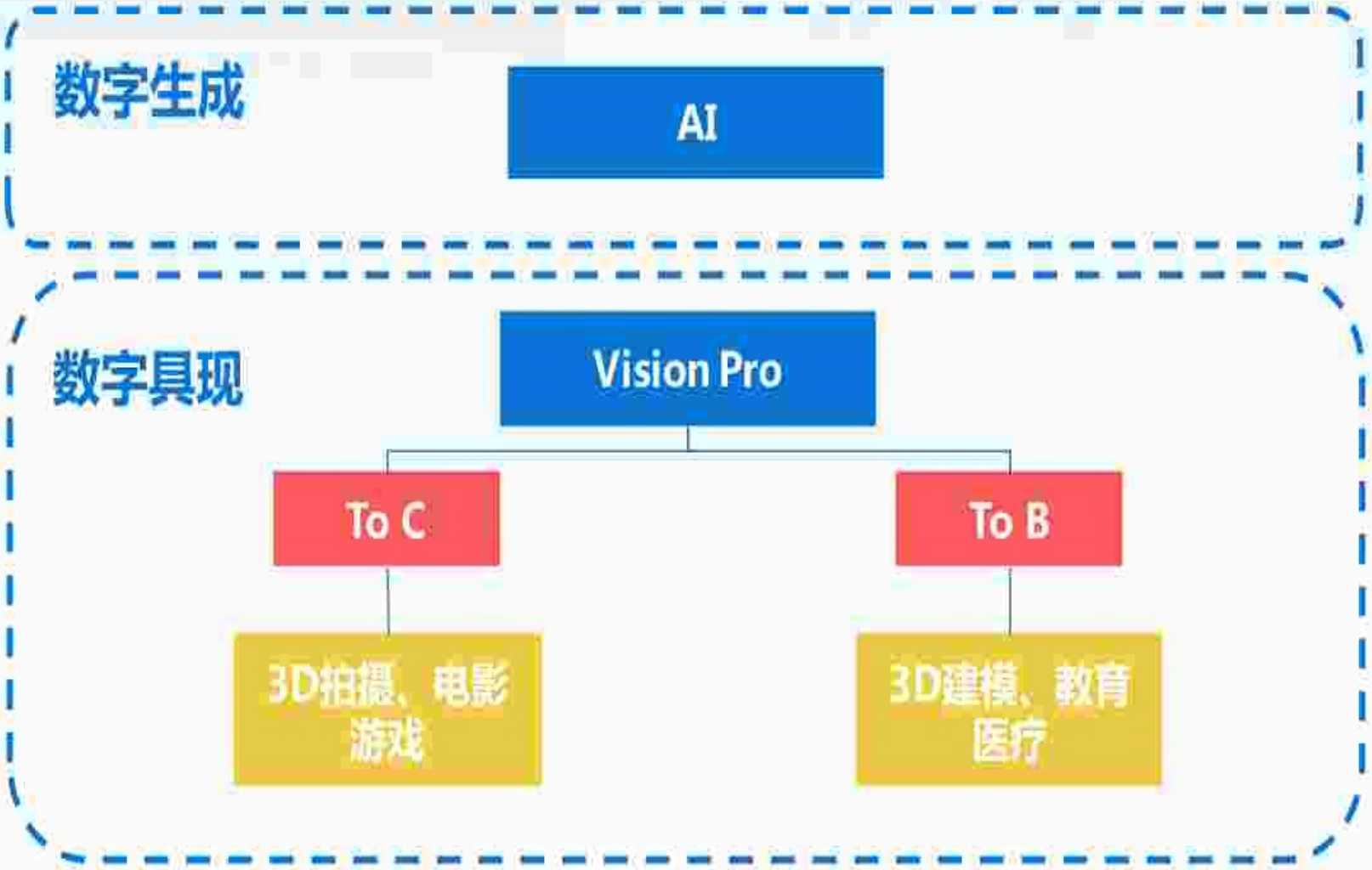
数据来源：WWDC23

苹果Vision Pro可通过VST方案实现MR（混合现实）

概念	VR（虚拟现实）	MR（混合现实）	AR（增强现实）
定义	利用电脑模拟产生三维的虚拟世界，提供用户感官模拟，让用户身临其境，无限制观察三维空间内的事物	结合真实和虚拟世界创造了新的环境和可视化，物理实体和数字对象共存并能实时相互作用	透过摄影机影像的位置及角度精算，加上图像分析技术，让虚拟世界与现实场景结合和交互
关键要素	沉浸感、交互性、假想性	现场感、混合性、逼真性	现场感、增强性、相关性
基本原理	利用计算机生成一种多源信息融合的、交互式的三维动态视景和实体行为的系统仿真	将感官增强功能实时添加到真实环境中，强调虚拟图像的真实性。通常为VST透视	将感官增强功能实时添加到真实环境中，强调与现实交互。通常为OST透视

数据来源：亿欧智库，国泰君安证券研究

- Vision Pro强调与现实世界的交流和融合，而非隔绝；将在方方面面改变人们的生活与生产。



消除数字与现实的鸿沟，开启空间运算新时代

- **Vision Pro真正意义上实现了数字与现实的连接转化，从“将数字世界融入现实世界”将过渡至“运算现实世界”。**

真实的具现

- **Vision Pro可实现“真实的具现”，以满足人眼的视觉体验需求，**在屏幕、光路、芯片等方面全新培养供应链，以追求“真实的具现”。
- **“真实的具现”是拓宽用户群体的第一基础：**“真实的具现”也是带给用户三维视觉震撼体验的第一基础。

图：Vision Pro实现“真实的具现”

全新的交互

- **Vision Pro开启全新的交互方式，眼睛、手势、声音开启空间交互新模式，**放置十余颗摄像头传感器，以捕捉用户眼球、手势与语音信息。
- **三维的视觉体验与全新的空间交互方式，将进一步打破人与数字交互的界限。**

图：用户通过注视选择应用，通过捏合手指即可打开应用

图：用户手势的多种指令

可运算的现实

- **Vision Pro对现实的运算，是真正意义上的数字与现实的融合，也是与AI的完美结合。**
- **算力芯片是对现实进行运算的基础，AI将无处不在。**Vision Pro对现实的运算由M2和R1两颗强大的芯片保障算力，开启了空间运算新时代。

图：基于机器学习运算后创造的现实形象

数据来源：WWDC23

- **庞大二维存量生态的三维全新升级：Apple Vision pro 开创了一类崭新的计算设备，能将数字化内容融入真实世界实现增强现实。**如同Mac将我们带入个人计算时代，iPhone将我们带入移动计算时代，**Apple Vision Pro将带我们进入空间计算时代。**这是一个伟大历程的起点，一个强大个人化科技的全新维度。
- **Apple Vision Pro带来自由的空间和全三维互动界面。**Vision Pro的体验，不受显示框限制，它会自有地填满周围的环境，可以轻松随意缩小放大app界面。

用户能够使用眼镜双手和语音，与三维互动应用操作界面互动

数据来源： WWDC23

用户可以自由移动并调整app界面尺寸大小

数据来源： WWDC23

Vision Pro强调与现实的互动，而非隔绝

数据来源： WWDC23

工作场景

- **Vision Pro会融合数码和现实世界，非常适合在办公场景中使用。**整个世界都变成了各种app的画布，用户可以随处摆放他们，调整大小来适合手头的任务。

Vision Pro可以处理3D信息

Vision Pro工作时并不与现实脱节

设备能够与 iPhone 蓝牙配件互联

家庭场景

- **Vision Pro非常适合在家居场景中使用。**它将空间计算带入日常家居生活的方方面面，改写了重视珍贵瞬间的方式——为照片和视频增加一个新维度。

Vision Pro苹果首台3D照相机设备，提供全新的3D瞬间回忆

娱乐场景

- **Vision Pro提供全新的娱乐体验。**用户能使用设备沉浸式观看电影、玩游戏，并享受空间环绕音频。

Vision Pro用户可使用设备沉浸式娱乐

数据来源：WWDC23

- **Apple Vision Pro将庞大的存量Apple生态进行了3D移植。**将二维的存量内容进行三维显示，并与现实融合，结合全新的交互方式，将带来开创性的空间体验，拓宽无限应用场景。
- **Apple Vision Pro将带来沉浸式赛事直播体验。**NBA确认将在赛事直播中引入苹果Vision Pro头显，为球迷提供场边第一视角的线上观赛体验。NBA与苹果的合作将创新体育比赛观赛方式，重新构建用户的赛事直播体验。

Vision Pro目前已经拥有超过100款Arcade游戏可供游玩

数据来源：WWDC23

Immersive Videos提供180度高分辨率视频

数据来源：feefreetickets

NBA确认赛事直播引入苹果MR头显

数据来源：feefreetickets

VisionOS 将创造全新的交互体验，全面融入用户生活和生产

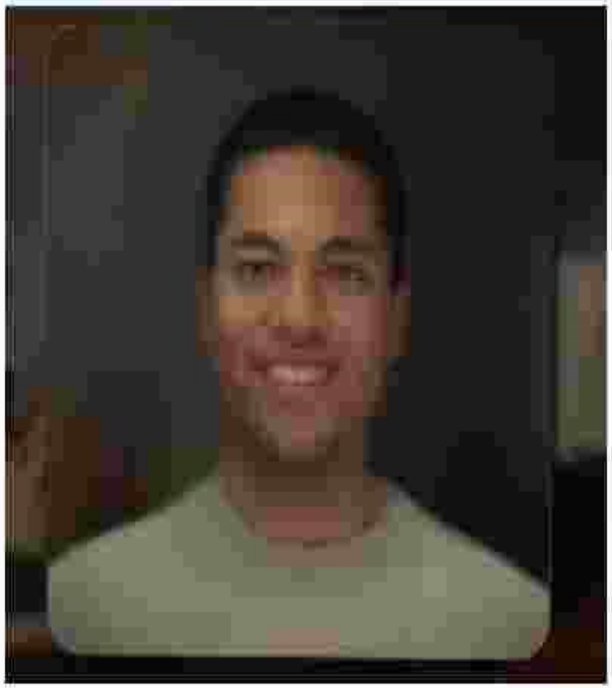
- **VisionOS将创造全新的交互体验，全面融入用户生活和生产**：苹果的visionOS基于macOS、iOS和iPadOS多年的技术积累，又做了诸多改进以适应空间计算的低延迟需求。
- 现有应用系统框架也为原生MR空间体验做了改进，**VisionOS是苹果首款专门为空间计算打造的操作系统，是整个新平台的起点。**
- **VisionOS使用了与iPadOS和iOS相同的架构，意味着基于后者开发的成千上万应用也可以在Vision Pro上使用。**Vision Pro也会拥有全新的App Store，其中包含专门为Vision Pro设备开发的应用，以及其他可以在iPhone和iPad上使用的应用。

VisionOS基于macOS、iOS和iPadOS多年的技术积累	基于VisionOS开发的应用	开发可以继续使用熟悉的开发工具	基于iPadOS和iOS开发的应用也可以在Vision Pro上使用

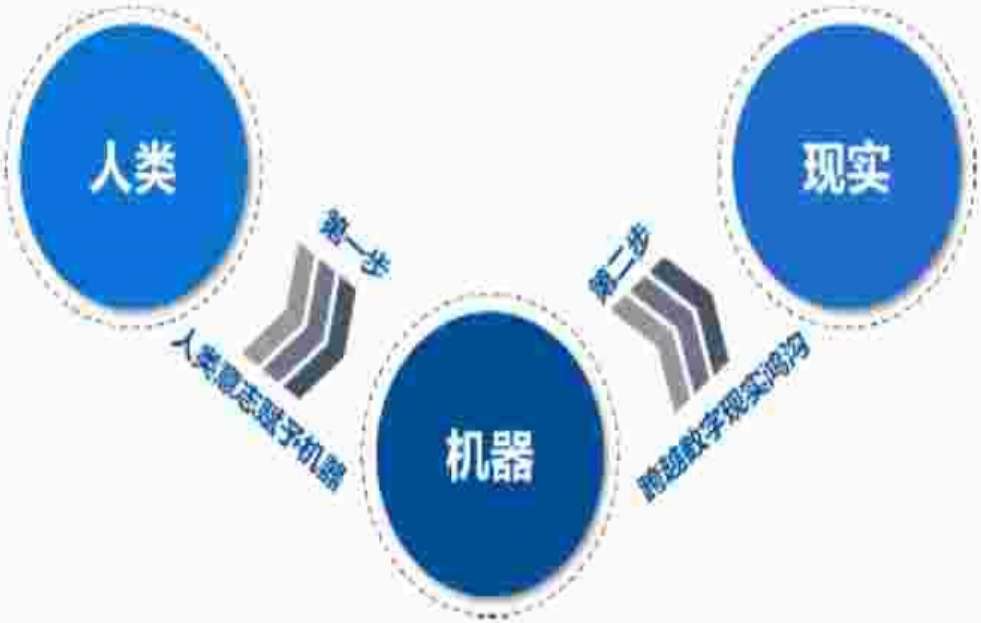
数据来源：WWDC23

- **Vision Pro与AI结合生成Persona虚拟现实形象。**Vision Pro采用机器学习技术，通过前置传感器和神经网络生成用户专属形象，并动态模仿用户的手部和面部动作，创建具有立体感的人物形象Persona

基于机器学习创建的用户形象



- **Vision Pro与AI天然绝配，生成的数字人类将彻底改变生活与生产，进而改变整个世界。**AI打造数字世界的内核，对“可运算的数字化现实”进行再定义，Vision展现数字世界的形貌。

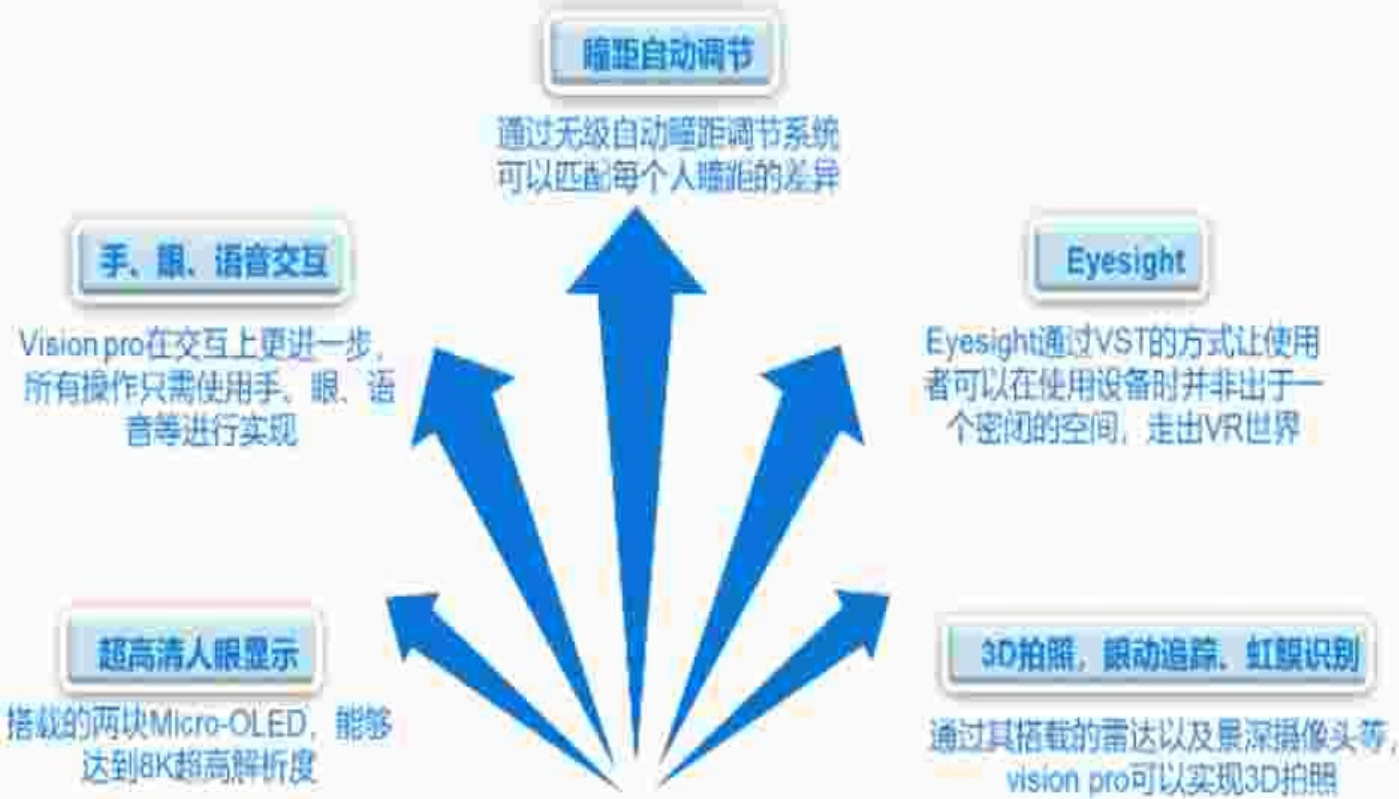


数据来源：WWDC23，国泰君安证券研究

- **硬件配置远超当前市场主流产品，赋能其“空间计算”、“可运算的现实”能力。**芯片方面，Vision pro搭载了苹果自研的M2和R1双芯片；光学方面，整台设备配备了两块索尼4K Micro OLED屏幕、3P Pancake光学镜片。
- **整台Vision Pro还拥有12颗摄像头、5个传感器、6个麦克风，可以实现包括但不局限于以下各类功能：**

vision pro正面配置了多种类型的传感器

vision pro内侧传感器用于眼动追踪、瞳距调节以及虹膜识别



数据来源：WWDC23，国泰君安证券研究

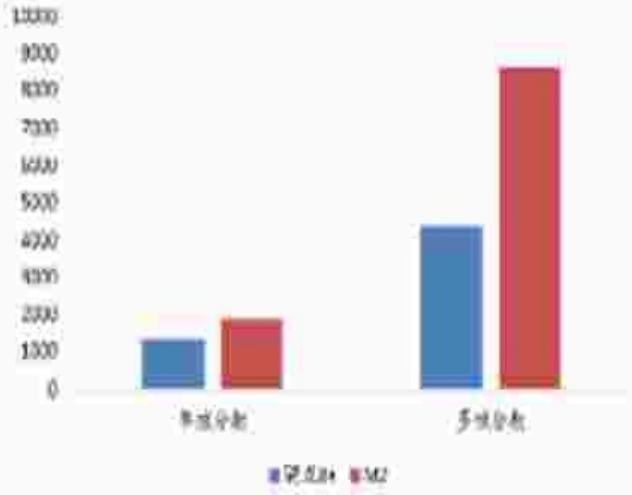
芯片：配置自研M2+R1芯片，提供强大运算基础和性能

- **M2+R1双芯片方案，为空间计算打下坚实底座。**苹果vision pro采用了双芯片设计，即M2+R1。M2芯片主要用于提供强大的计算性能，R1为苹果专门为处理实时传感数据而设计的芯片。
- **M2提供超强的运算能力，R1降低设备的延时。**R1芯片能够在12毫秒内将新图像传输到显示屏，速度比眨眼还要快8倍。而R1和M2的强大之处还在于，将外界的图形数字化，经过处理和渲染后，能够在12毫秒将处理完的虚拟与现实融合的图像传输到人眼前。

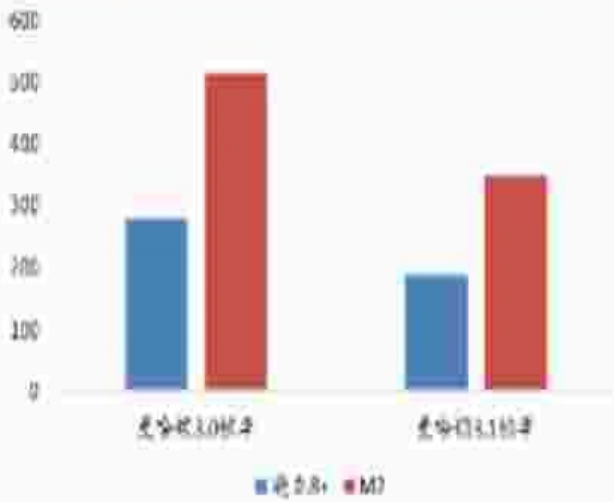
vision pro搭载了苹果自研的M2和R1芯片

数据来源：WWDC23

M2的CPU性能远超骁龙8+



M2的GPU性能更加流畅



数据来源：Geekbench5，国泰君安证券研究

芯片：配置自研M2+R1芯片，提供强大运算基础和性能

M2

- **更强的M2芯片已经推出，芯片制程也有望迎来升级。**台积电的3nm制程也已经实现量产，若下一代苹果的头显设备搭载更高规格或者制程更为先进的M系列芯片，其计算和传输性能有望再上一个台阶。

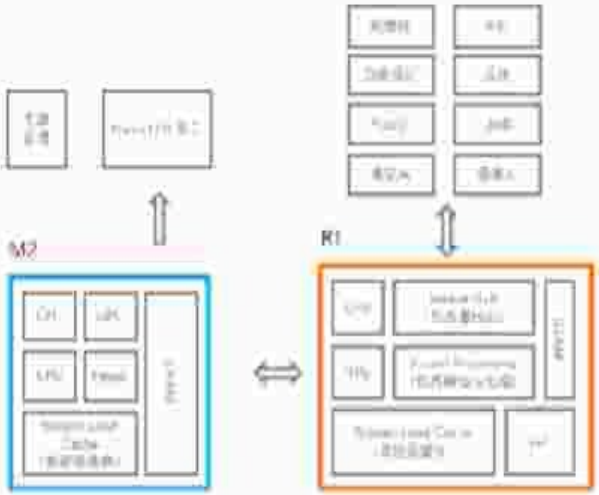
R1

- **R1芯片为苹果全新设计，专门负责处理来自相机、传感器和麦克风的输入。**R1芯片将图像实时传输到显示器，实现几乎无延迟的实时视图，响应时间仅为12毫秒，以展现世界的真实感。
- **R1芯片是 Vision Pro 的关键部件，提供流畅和身临其境的体验。**它具有许多对于流畅和身临其境的MR体验必不可少的性能：高速传感器处理、低延迟、高电源效率等。

M2 Ultra 为苹果在 WWDC2023 上推出的芯片

数据来源：WWDC23

R1架构与周边连接估计

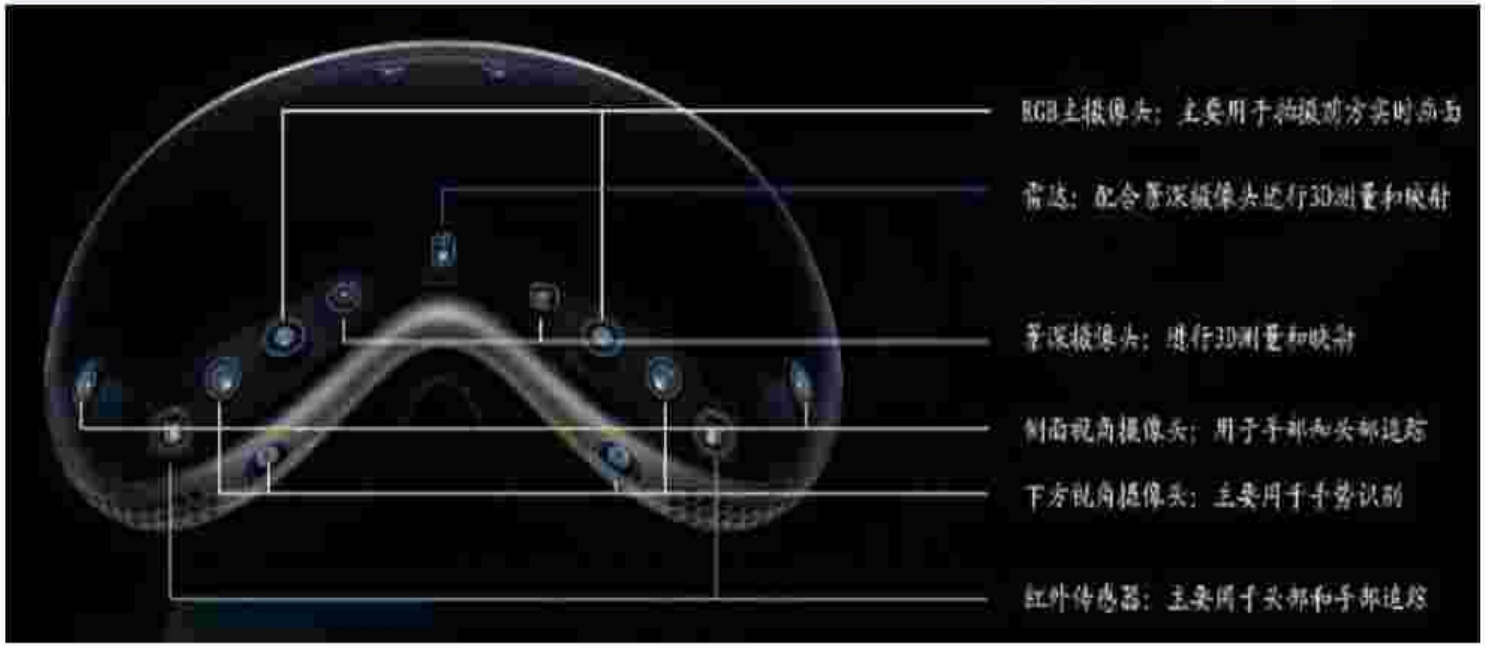


数据来源：陈巍谈芯

传感器：种类及数量均为顶级配置，赋能其空间运算能力

- **传感器数量及种类配置拉满，为应用和交互端提供无限可能。**Vision pro整台设备共搭载了12颗摄像头、6个麦克风、5个传感器。5个传感器部分包括1个激光雷达、2颗景深摄像头、2个红外传感器。同时设备内侧还配置了一圈LED。

vision pro 外侧配置了大量传感器



数据来源：WWDC23，国泰君安证券研究

传感器：种类及数量均为顶级配置，赋能其空间运算能力

Vision pro正面配置的超高清主摄不间断的收集外界图像

Vision pro真正实现了裸手交互

雷达及景深摄像头进行3D映射

雷达可以更好的对周边环境进行3D建模

雷达及景深摄像头进行3D映射

Vision pro使用Optic ID虹膜解锁

- **多颗摄像头用于各个角度实时画面及动作捕捉**：超高清摄像头+强大的运算能力，VST 实现无边界的虚拟与现实融合。
- **侧面视角、下侧视角摄像头及红外发射器实现精准的手部和头部追踪。**裸手的追踪相对较为困难，vision pro 将这一功能变为现实，通过多颗传感器让手部摆脱了手柄的限制，通过视觉、红外传感器以及软件算法能力真正实现了裸手交互。
- **雷达及景深摄像头实现3D拍照：Vision pro是苹果首个搭载3D相机的设备。**雷达与景深摄像头能更精准的测量深度信息，将2D图像转换为带有深度信息的3D图像。
- 雷达+ 景深摄像头的组合，对外界环境有着精准的感知，结合苹果自身的软件 算法优势，让 3D 成像更加真实。
- **内侧红外摄像头和LED矩阵可以实现眼动追踪及Optic ID功能**：红外摄像头+LED矩阵实时追踪眼球位置，实现眼球即鼠标。
- **Vision pro使用Optic ID虹膜解锁，可用于Apple Pay 等其他功能。**此外，Optic ID也可以用于Apple Pay的支付功能等其他场景。

数据来源： WWDC23

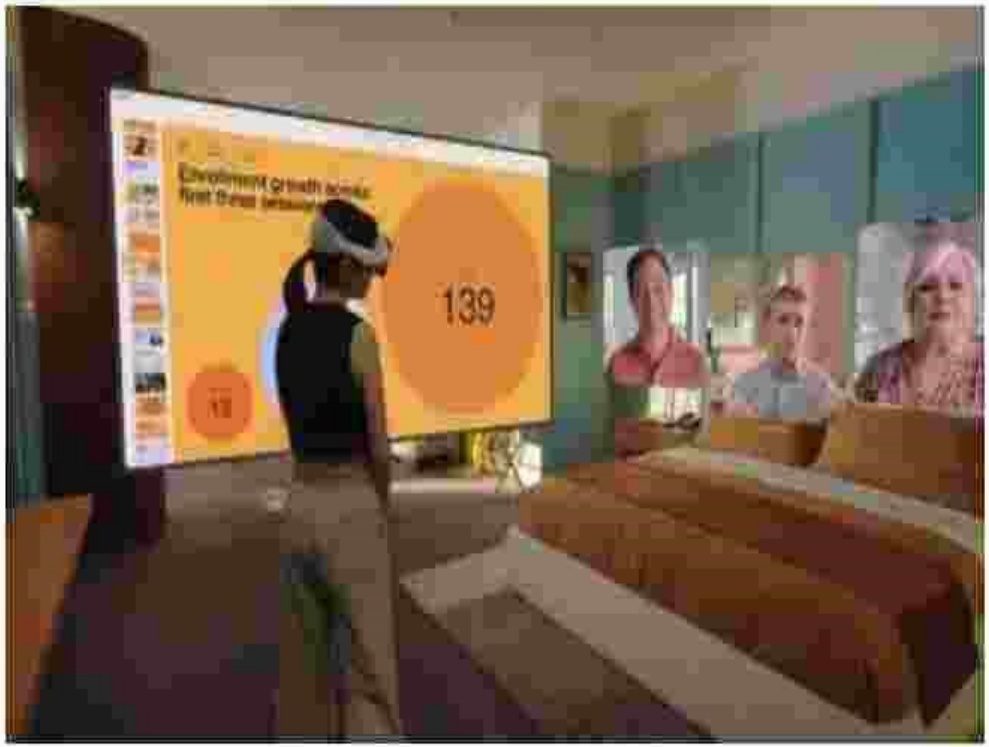
传感器：种类及数量均为顶级配置，赋能其空间运算能力

- **真实沉浸的空间音频系统：Vision pro搭载了目前苹果最先进的空间音频系统，让用户更加身临其境。**Vision pro会分析周围环境的声学特性，匹配环境所需的音频效果，尤其是在使用Facetime通话时，声音仿佛就是从头像窗口处发出。
- **声学测算+动态追踪，带来绝佳的音频体验。**Vision pro通过精准定位，实现对头部动态的追踪，带来更具沉浸感的音频体验。

Vision pro使用了目前苹果最先进的空间音频系统

数据来源：WWDC23

Vision pro让Facetime通话更具沉浸感



数据来源：WWDC23

光学：Pancake+Micro OLED的顶配方案

- **3层曲面pancake提供极高的清晰度和通透度：苹果新一代MR产品使用了3P Pancake光学方案，为当前VR设备的顶尖配置。**苹果Vision pro达到了120°的视场角，远大于目前其它主流厂家最新旗舰产品能够提供的视场角范围，且减少了传统菲涅尔透镜边缘的畸变、暗角问题，成像质量更高。
- **Pancake折叠光路方案利用偏振光原理，包含两组透镜。**显示屏上的画面进入半反半透功能的分束镜后，光线在镜片、相位延迟片以及反射式偏振膜经由多次折返，最终通过偏振反射镜进入人眼。
- **Pancake折叠光路具有佩戴舒适、成像质量高、屈光可调节的优势，是当前改善VR用户体验的最佳成熟方案。**Pancake通过折叠光路减小光路系统空间的物理距离，镜头总长比常见菲涅尔透镜方案减小超过一半。且该方案的镜片组无需和显示屏保持距离，进一步降低VR头显的厚度。随着镜头总长减小，产品重量也大幅减轻。以使用Pancake光学方案的Huawei VR为例，其重量仅为166克。

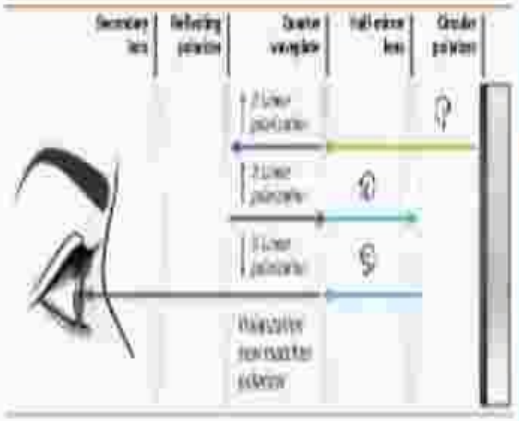
Vision pro配置了3P pancake光学方案

数据来源：WWDC23

pancake比菲涅尔透镜更加轻薄

数据来源：Oculus

pancake光学设计图



数据来源：智东西

主流厂家方案对比

	Huawei VR	Quest Pro	Quest2
重量	166克	722克	503克
光学方案	Pancake	菲涅尔	菲涅尔

数据来源：VRcoast、IT之家、公司官网

光学：Pancake+Micro OLED的顶配方案

- **Pancake 是当前 VR 光学主流方案**：Pancake可支持的理论FOV上限为200°，面板分辨率无上限，目前技术仍有较大发展和降本空间，具有充足想象力。

	非球面透镜	菲涅尔透镜	折叠光路Pancake	多重折返式自由曲面
视场角FOV	90°-180°	90°-120°	70°-120°	80°-100°
镜头总长TTL	40-50mm	40-50mm	15-20mm	40-45mm
成像质量	边缘成像好	容易产生伪影和畸变	较好，易伪影	容易产生畸变
优点	成本便宜	较轻薄便宜	轻薄，成像质量好	有利于眼动元器件布置
量产价格	5-10元	15-20元	120-180元	50-100元
发展阶段	逐渐淡出市场	主流选择	高端产品应用	小众市场
代表产品	PS VR	Meta Quest2	苹果MR	Lynx

数据来源：wellsennxr，国泰君安证券研究

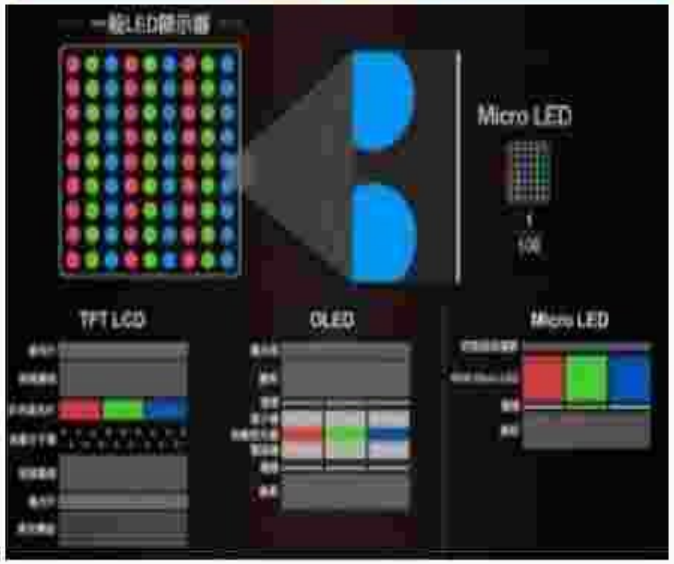
- **Micro OLED单眼4K分辨率，PPD遥遥领先：苹果Vision pro搭载Micro OLED屏幕，任何角度都足够清晰，文字清晰锐利。**据Latepost，苹果Vision pro PPD达到了~40的水平，作为对比，Meta Quest Pro/Pico 4的PPD分别约为17/20.6，远落后于vision pro。
- **硅基OLED是综合考虑性能、技术成熟度和价格后，可用于MR显示的最佳技术方案。**Micro OLED微显示器件采用单晶硅晶圆（Wafer）为背板，在CMOS驱动电路顶层制作发光层，是兼顾性能、价格、量产能力的最佳方案。
- **硅基OLED性能突出，自发光、产品轻薄、响应速度快等一系列优势，将改善用户体验。**硅基OLED的自发光特性使得产品无需背光源，且功耗仅为LCD的30-40%，提升续航能力，为电池减重提供了空间。硅基OLED的体积仅为传统现实器材1/10，有效提升像素密度且重量减少50%以上，且拥有纳秒级响应速度，远远快于毫秒级的LCD和微秒级的OLED。

雷达及景深摄像头进行3D映射

Vision pro的PPD远超市场其他产品

数据来源：WWDC23，Latepost

Micro OLED、普通LED、TFT-LCD、OLED技术对比



数据来源：poppur

Micro OLED、LCD、OLED 性能对比

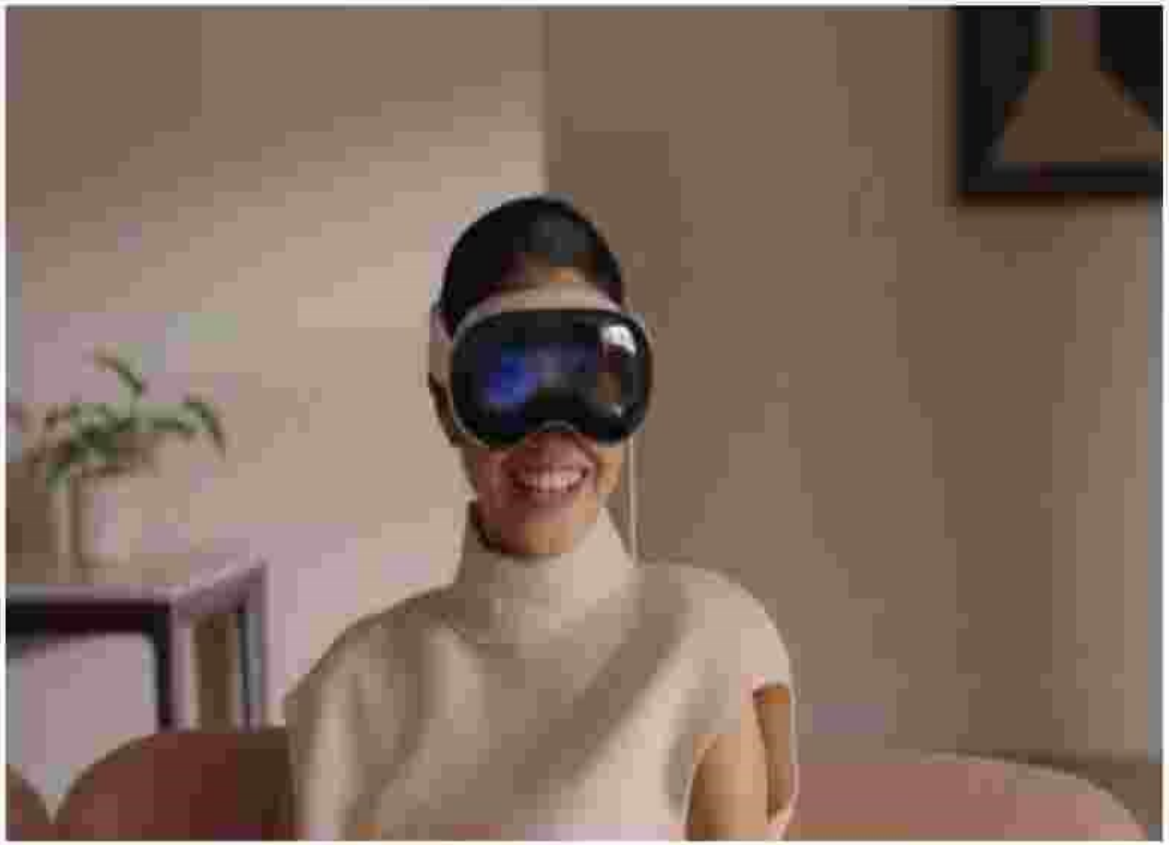
	LCD	OLED	Micro OLED
技术类型	背光板/LED	自发光	自发光
对比度	5000:1	无上限	无上限
寿命	中等	中等	长
反应时间	毫秒	微秒	纳秒
运作温度	摄氏-40~100度	摄氏-30~85度	摄氏-100~120度
成本	低	中等	高
能源消耗量	高	中等	低
可视角度	低	中等	高
PPI（穿戴式）	最高250 PPI	最高300 PPI	1500 PPI以上
PPI（VR）	最高500 PPI	最高600 PPI	1500 PPI以上

数据来源：poppur，国泰君安证券研究

光学：Pancake+Micro OLED的顶配方案

- **外屏：实现EyeSight功能，让用户与周围的人保持连接，**Vision pro在设备外部加了一块弧形OLED屏幕，做出双向透视的效果。

Vision pro可以让用户与身边人保持自然交流



数据来源：WWDC2023

光学：Pancake+Micro OLED的顶配方案

- **自动瞳距调节功能可以大幅降低甚至消除眩晕感**：Vision pro配置了两组独立的瞳距调节装置，可以大幅减甚至消除晕眩的情况。
- **电动无级瞳距调节已成为高端MR/VR设备基础配置。**vision pro的自动瞳距调节会比Meta quest pro更快、更无缝，无需手动拨盘或者通过滑块来进行设置，提高了用户的体验感，让用户能够更方便的上手使用。

Vision pro配置了两组独立的瞳距调节系统

数据来源：WWDC2023，国泰君安证券研究

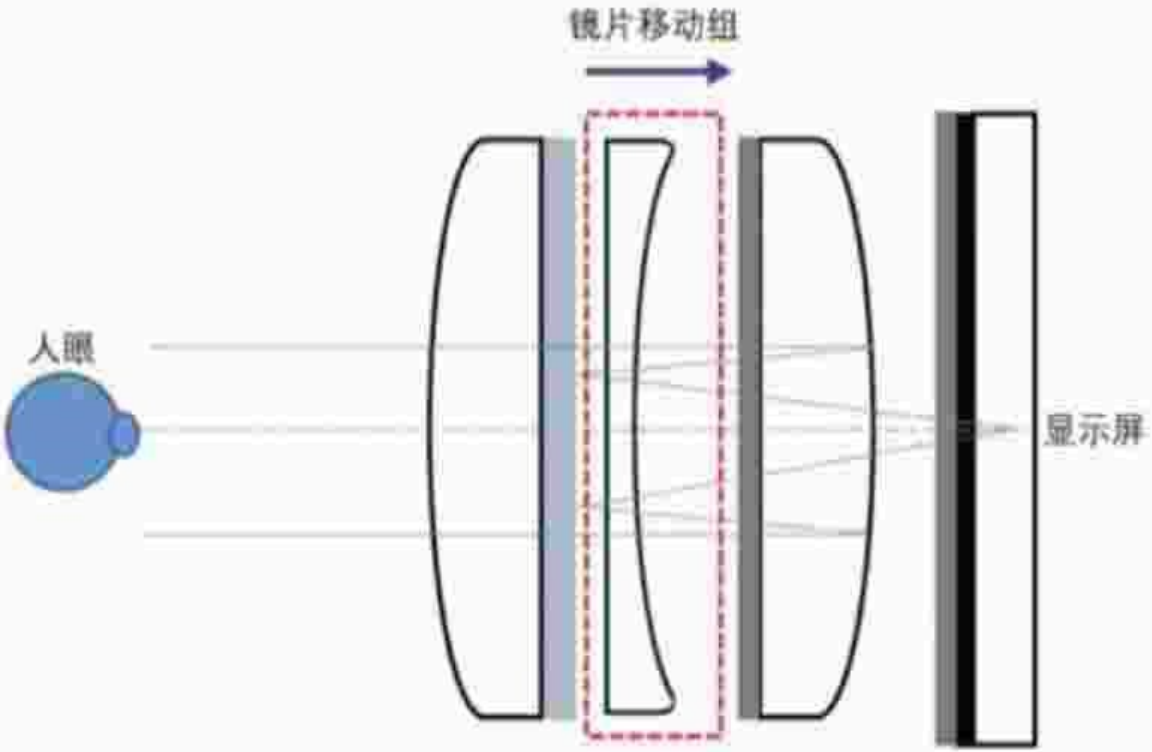
Pico4、Dream Pro、Meta quest pro对比

Pico 4	奇遇Dream Pro	Meta quest pro
62-72mm 无级电动调节	57-69mm 自适应调节	55-75mm 连续IPD

数据来源：中关村在线，国泰君安证券研究

- **Pancake光学方案给屈光度调节留有空间**：当前苹果针对近视用户在设备上设置了一个磁吸镜片接口，未来屈光度调节有望导入。

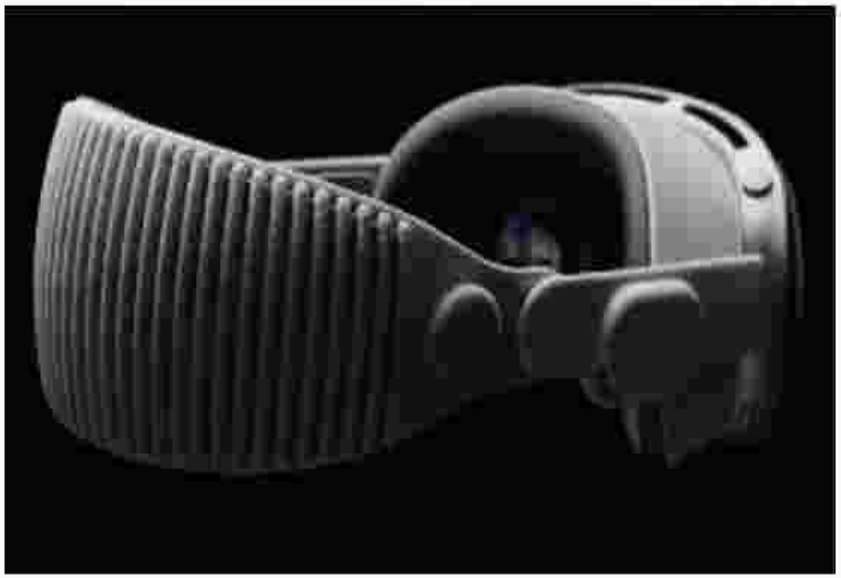
pancake光学方案可以通过调整镜片位置完成屈光度调节



数据来源：智东西

- **模块化设计，让整机看起来更具美感。Apple vision pro还配置了以下结构及功能件：**
- （1）三维构造的层压玻璃（2）外框上的两个物理按钮（3）特制铝合金外框（4）轻质眼罩（5）3D编织头带透气性（6）外置高性能电池
- **MR产业链基本定型，长期投入的厂商或享受到产业爆发红利，供应链上游将受益。**

Vision pro的头带舒适性、透气性俱佳



数据来源： WWDC23

Vision pro的模块化设计兼具美感与实用性

数据来源： WWDC23

替代传统存量市场

VR/AR设备
Meta Quest 2
Pico Neo 3

体感游戏设备

智能穿戴设备
iWatch
AirPods

便携智能设备
iPhone
Mac book
ipad

Vision Pro

重新”iPhone时刻”

重塑场景颠覆体验

重塑to B场景
对标便携智能设备
有望实现千万市场规模。

重塑to C场景
对标VR/AR、家用游戏设备
有望实现千万市场规模。

3D摄影颠覆性创新
取代传统相机手机摄影设备
成为市场的新引爆点

数据来源：国泰君安证券研究

- **Vision Pro的应用场景所对应的传统市场主要包含四个方面：VR/AR设备、体感游戏设备、智能穿戴设备、便携智能设备。**
- 本部分我们将分析传统存量市场中这四类产品的销量表现，以作为Vision Pro产品的市场判断参考。

数据来源：国泰君安证券研究

- VR/AR设备的销售量级为1000万左右，家用游戏设备为2000万左右，智能穿戴设备在2000万左右，便携智能设备在3000万左右。

未来，Vision设备将对这一存量市场的部分产品形成替代。

产品类型	产品名称	公司	价格	产品销量
VR/AR设备	Quest 2	Meta	299美元	2021年约2000万台
	Pico Neo 3	Pico	2699元	2022年约50万台
家用游戏设备	PS5	索尼	499美元	2022财年约1700万台
	Xbox Series X/S	微软	499美元	2021年约900万台
	Switch	任天堂	300美元	2021财年2306万台
智能穿戴设备	iWatch series 6	苹果	399美元	2021年约4160万只
	AirPods一代	苹果	159美元	2019年约3500万副
便携智能设备	iPad系列	苹果	599美元起	2022年约6180万台
	iPhone 13	苹果	799美元起	2022年约6500万台
	iMac 2020系列	苹果	999美元起	2021年2570万台

数据来源：Businessofapps，公司公告，国泰君安证券研究 备注：鉴于数据的缺失，部分产品销量采用整个系列，部分采用单个型号，仅供参考

Vision Pro 技术卓越，有望重现 iPhone 时刻——重塑to B场景

- **Vision Pro可为办公场景提供多样化的应用。**有望重塑沉浸式办公、更便捷的团队协作等to B场景。对标便携智能设备市场规模，**Vision有望实现千万市场规模。**
- **Vision Pro不仅适用于个人办公，也可以支持团队协作和跨部门沟通，**Vision Pro 支持用户在虚拟空间中创建、编辑和共享文档、图表等办公文件，也可以与远程同事进行实时的视频会议和协作。

Vision Pro可实现沉浸式便捷办公

数据来源：WWDC2023

Vision Pro让协作更加便捷

数据来源：WWDC2023

- **Vision Pro的虚拟现实功能，为个人生活娱乐提供了更多、更为便捷的选择。**Vision Pro 是一款能够将数字内容与物理世界无缝融合的空间计算机，让用户在保持与他人联系的同时，享受沉浸式的娱乐体验。

Vision Pro可实现沉浸式VR/AR游戏体验

数据来源：WWDC2023

Vision Pro 技术卓越，有望重现 iPhone 时刻——颠覆性创新

- **得益于3D传感器的加入**，苹果Vision Pro支持拍摄3D照片、录制3D视频，并且融合空间音频，让用户在拍摄和回忆照片和视频时，享受全新的视觉体验。**这一颠覆性创新，有望重新“iPhone时刻”。**

Vision Pro 3D摄影有望取代传统相机

数据来源：WWDC2023

Vision Pro 点燃 MR 赛道，有望重现“iPhone时刻”

数据来源：Captain Gizmo，国泰君安证券研究

Vision Pro 点燃 MR 赛道，有望重现“iPhone时刻”

- **我们预计，Vision Pro产品发展的第一阶段，将类似iPhone初期的创新与突破过程。**第一阶段的3年时间，Vision Pro将实现由百万级销量向千万级销量的增长过程。

年份	产品名称	产品销量
2007年	iPhone一代	约610万台
2008年	iPhone 3G	约2507万台
2009年	iPhone 3GS	约3559万台
2010年	iPhone 4	约5071万台
2011年	iPhone 4s	约6010万台
2012年	iPhone 5	约7040万台
2013年	iPhone 5s, iPhone 5c	约7669万台
2014-2015年	iPhone 6系列	约3.47亿台
2016年	iPhone 7系列与iPhone SE第一代	约1.13亿台
2017-2018年	iPhone 8系列与iPhone X	约1.61亿台
2018年	iPhone Xs系列	约1.23亿台
2019年	iPhone 11系列	9个月销量破1亿

数据来源：Captain Gizmo，国泰君安证券研究

对Vision系列产品的价格分析——当下价格不是关键

- **价格不是划时代产品成败的关键**：iPhone 3GS、AirPods 1、iWatch 1与当前主流产品相比，价格都超出数倍，然而由于苹果产品的划时代竞争优势，未来均实现了迅猛的增长。
- **Vision Pro的划时代价值，是其成为未来利润增长点的关键。**Vision Pro实现的对to B、to C场景的重塑，颠覆性的3D摄影创新，是现有设备无可比拟的。这些划时代的进步，将在不久的将来得到市场的认可，成为苹果新的利润增长点

数据来源：国泰君安证券研究

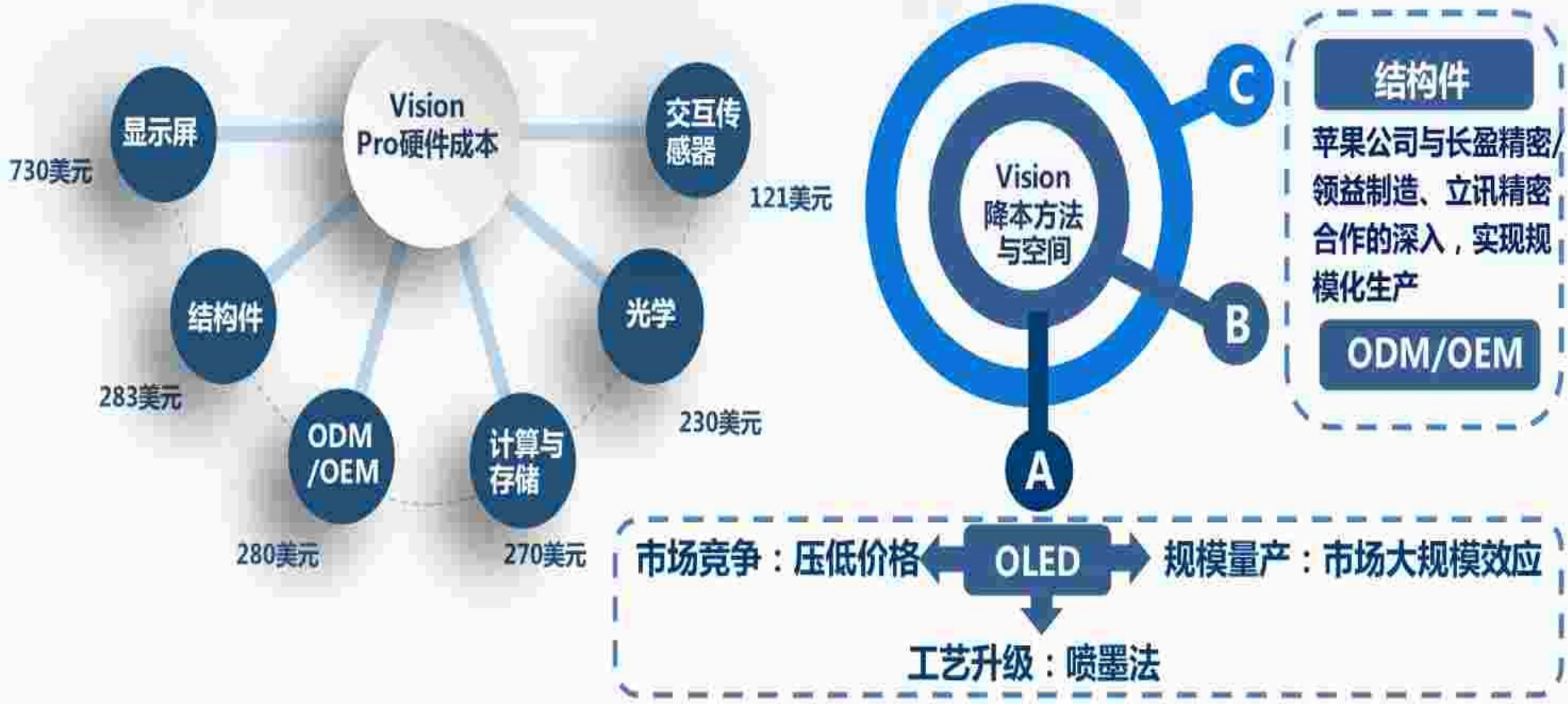
苹果产品与同时代竞品价格对比

厂商	手机价格	耳机价格	VR价格
苹果	iPhone 3GS约4500元	AirPods 1约1600元	iWatch 1约2500元
主流竞品1	安卓塞班机约1000元	有线耳机约 20元	电子手表约 50元
主流竞品2	小灵通约 100元	挂脖无线耳机约 200-300元	功能复杂的电子手表/环约300-500元

数据来源：根据历史商城信息整理，国泰君安证券研究

对Vision系列产品的价格分析——未来价格有望下降

- **主要成本：显示内屏、结构件、ODM/OEM，**随着未来良率上升，产能提升的规模化效应，有望实现价格下降。



数据来源：正观新闻，国泰君安证券研究

03

AI端侧落地元年，引领消费电子新浪潮

AI 手机

- **大厂开启 AI 探索之路，打造智能随行助理。**谷歌、华为、小米、OPPO、vivo等纷纷推出接入大模型的AI手机。核心芯片更新迭代，高通推出高性能AI引擎“骁龙 8 Gen 3”支持多项AI功能，相关开发平台构建丰富AI生态。
- **“智能便携”属性增强，智能加工应用升级。**AI助力智能搜索性能提升，整机管理效率增强，同时在增图像视频增强、语言感知增强两方面为用户提供个性化加工窗口。
- **计算、感知、连接硬件全线升级。**AI功能产生更多任务、计算需要，推动CPU、专用芯片、ISP全面升级并将促进手机在内存、无线通信、光感知领域模块的升级。

AI PC

- **巨头产品落地加速，构建全场景生态。**联想发布全球首款AI PC产品，提出“混合人工智能”计划，解决用户信息安全问题。英特尔、AMD、高通发布PC AI芯片，助力终端生成AI应用。操作软件方面，微软接入copilot免费更新，默认集成于任务栏，培养用户人工智能使用习惯；英特尔宣布AI PC加速计划，构造AI PC生态系统解决方案。
- **解放生产力，赋能创造力。**AI PC通过各类助手工具提高用户工作效率，解放生产力；通过整合人际语言，人机语言、机机语言,输入自然语言即可生成图片、视频、代码降低创造门槛，赋能创造力。
- **以消费级PC AI处理器为核心的一系列硬件升级。**AI将促进以消费级PC AI处理器为核心的硬件升级，驱动DDR5、PCIe5.0加速替代，催生更高效的电脑散热解决方案。

AI PIN

- **智能终端新品涌现，AI引领消费电子换机新浪潮。**AI终端在保护用户隐私的前提下增强用户个性化体验，从而促进用户对AI智能化的感知，坚定用户对AI端侧应用的信赖，刺激用户换机需求，推动新一轮的手机PC换机浪潮。

谷歌：推出Pixel 8系列新机，内置谷歌AI基础模型

终端厂商

谷歌

华为

小米

VIVO

OPPO

- **Pixel 8** 搭载 Google AI 定制的 Tensor G3 芯片，其中包括最新一代的 ARM CPU，升级的 GPU，新的 ISP 和成像 DSP 以及谷歌的下一代 TPU。与配置上一代 Tensor 的 Pixel 6 相比，Pixel 8 上运行的ML模型数量提升至 2 倍，模型结构也更加复杂（比去年 Pixel 7 中最复杂的模型复杂 150 倍）。Tensor G3 通过将机器算法直接构建到芯片的方式，以更少功耗实现更强大的功能。此外 Pixel 8 还预装最新 Android 14，并提供7年的安卓大版本软件更新，同时支持 WiFi-7 连接。

谷歌Pixel 8支持丰富AI功能

AI应用领域	功能
AI助手	网页页面内容翻译、总结、关键词搜索、阅读
	骚扰电话过滤
	电话接听助手
	即时的录音转录文字
图片功能	人物或物体的擦除、添加、移动，背景换色
	多人合照用换脸优化表情
视频功能	视频降噪
其他	面部解锁增强

数据来源：谷歌官网，国泰君安证券研究

终端厂商

- 荣耀
- 华为
- 小米
- VIVO
- OPPO

- **新一代智能操作系统HarmonyOS 4发布，接入AI大模型。**小艺成为首个具有AI大模型能力的终端语音助手，具备适用于生活、办公场景的助理功能，并在智慧交互、生产效率提升、创建专属场景和个性化服务方面获得提升。

（1）智慧交互：在原有的语音交互的基础上，新版小艺能够接受文字、图片、文档等多种形式的输入。华为小艺也能够基于这些信息调用相应APP插件，完成添加手机通讯录、美团搜索等操作；

（2）生产效率：能够对图片完成文字提取、摘要总结、表格提取、文档扫描等办公功能，让用户更高效的办公学习；

（3）个性化服务：在与用户的交流中，能够持续进行信息收集、理解，在保护隐私的前提下提供个性化服务。

华为鸿蒙4.0接入AI模型，集成AI语音助手小艺

AI应用	功能
“小艺建议”聊天助手	具备对话功能，文本创作能力
	能够辅助调用系统原装APP完成定时、查询日程等生活功能
小艺通话	能够对图片完成文字提取、表格提取、文档扫描等办公功能
	自动接电话、回复的语音助手

数据来源：华为官网，国泰君安证券研究

华为小艺能力矩阵

数据来源：华为官网

终端厂商

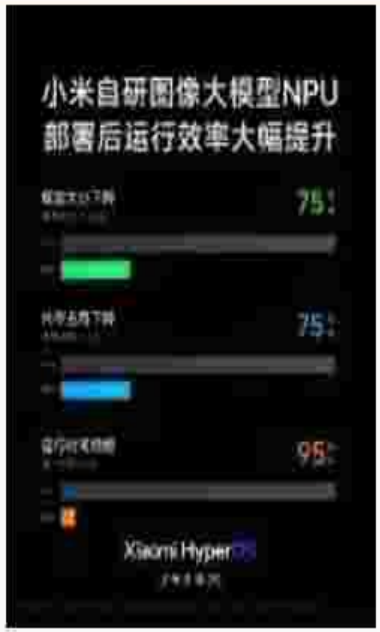
- 苹果
- 小米
- VIVO
- OPPO

- **小米14系列搭载高通“AI引擎”骁龙8 Gen3，CPU多核性能超越苹果A17 Pro，NPU赋予强大AI功能。**小米自研图像大模型在NPU部署后运行效率大幅提升，内存占用减少75%、运行时间缩短95%。同时，小米宣布与金山办公WPS进行合作。将WPS AI移动APP的能力嵌入到小米澎湃OS当中。用户通过小米手机拍摄文档后，可以进行内容总结；也可以上传文档来实现快速翻译和内容概括。

小米14支持多款AI应用

数据来源：IT之家

小米14内置强大NPU性能，助力AI在端运行



数据来源：IT之家

终端厂商

- 荣耀
- 华为
- 小米
- VIVO
- OPPO

- **VIVO X100系列手机发布，并全球首发天玑9300。**X100内部搭载基于百亿蓝心大模型BlueLM的AI助手“蓝心小V”，具备自然会话、信息处理和洞察能力，能够实现自然语言对话、AI路人隐身、文案写作、思维导图生成等功能。

蓝心小V应用矩阵

数据来源：VIVO X100系列新品发布会

终端厂商

- 荣耀
- 华为
- 小米
- VIVO
- OPPO

- **全新ColorOS 14操作系统发布，Pantanal与AndesGPT双模型实现智慧互融，具备闪速抠图、智能摘要、图片智能消除、内容创作、聊天助手等AI功能。**系统在端支持70亿参数的AndesGPT·Tiny大模型，云端支持AndesGPT·Turbo/Titan大模型，并以端云协同的方式为小布助手提供自然语言交互能力。

OPPO双大模型实现智慧互融功能

数据来源：OPPO开发者大会

高通：推出高性能 AI 引擎“骁龙8 Gen3”，支持在端运行大模型

核心硬件

高通

- **AI引擎“骁龙8 Gen3”支持在端运行100亿参数的多模态生成式AI模型（Stable diffusion可在1秒内生成图像或社交微博；70亿参数 Llama 2运行速率高达20 tokens每秒），并配套实现AI面部识别、智能拍摄对焦、视频物体擦除、视频降噪等基础AI功能。**为实现相应功能，骁龙8 Gen3底层核心全面升级：采用台积电4nm工艺，配置高通Kryo CPU（比前代产品骁龙8 Gen2性能提升30%、效能提升20%）、高通Hexagon NPU（性能提升98%、效能提升40%）。骁龙8 Gen3还配备Snapdragon X75 5G调制解调器、支持LPDDR5X RAM、UFS 4.0、Wi-Fi 7、蓝牙5.4等。高通骁龙8 Gen3将在全球OEM厂商和智能手机品牌的终端上得到广泛采用，有望开启生成式AI手机的新时代。

高通推出高性能AI引擎“骁龙8 Gen3”

数据来源：高通官网

高通推出高性能AI高通骁龙8 Gen3支持在端运行100亿大模型，并配套实现多种AI功能“骁龙8 Gen3”

AI应用领域	功能
AI大模型	支持大型语言模型LLM，语言视觉模型LVM，以及基于变压器网络的自动语音识别ASR
	支持INT4、INT8、INT16、FP16
	支持混合精度（INT8+INT16）
相机摄影	基于A的人脸检测
	3A功能：自动对焦、自动曝光和自动白平衡
	针对4K60FPS夜视视频抓取的RAW AI降噪
5G射频系统	AI增强信道状态反馈
	AI增强型天线调谐
	AI增强型全球导航卫星定位系统Gen 2
	基于AI的5G张量加速器Gen 2
虚拟助手	利用Sensing Hub安全合规地收集用户数据（如活动爱好、健身水平、位置信息），使AI虚拟助手提供个性化回复

数据来源：高通官网，国泰君安证券研究

联发科：推出天玑9300，最高支持在端运行330亿参数AI大语言模型

核心硬件

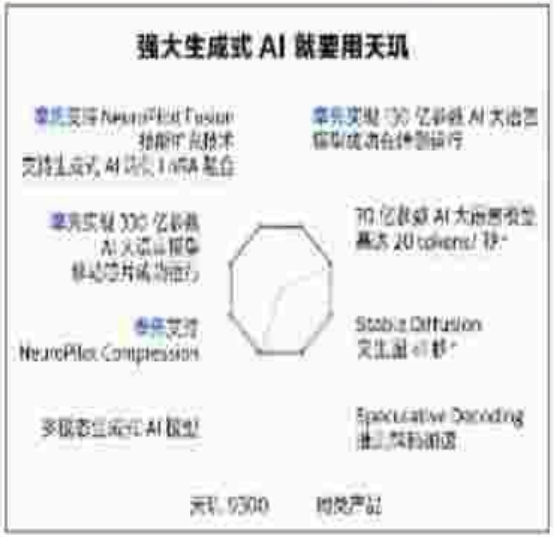
联发科

- **硬件方面**：天玑9300CPU采用“4超大核+4大核”设计，性能较上一代提升40%；集成全新第七代APU处理器APU 790，内置了硬件级生成式AI引擎，深度适配Transformer模型；集成新一代GPU，峰值性能提升46%，功耗节省40%；集成Imagiq 990 ISP，可进行16层图像语义分割引擎并逐帧优化，提升视频录制细节；内存率先支持LPDDR5T 9600Mbps。同时，联发科开发了混合精度INT4 技术，结合公司的内存硬件压缩技术，能够高效利用内存带宽，大幅减少内存占用。
- **AI进展方面**：天玑9300支持终端运行10亿、70亿、130亿和330亿参数的AI大语言模型，其中70亿参数LLM已在VIVO旗舰手机终端首发落地，并与VIVO成功在端运行130亿参数LLM。天玑9300在移动显示、音频降噪、5G通信等领域也融合了AI技术。预计首部搭载天玑9300芯片的手机将在2023年末上市。

APU 790最高支持330亿参数AI大模型

数据来源：联发科技公众号

天玑3900支持多项AIGC功能



数据来源：联发科技公众号

联发科天玑9300支持在端运行330亿参数LLM，并配套多领域AI功能

AI应用领域	功能
AI大模型	最高可支持330亿参数的AI大语言模型 混合精度INT4，2倍整数和浮点运算速度 高度适配生成式AI transformer，运算速度快8倍 首款生成式 AI 端侧技能扩充（LoRA Fusion）技术
AI摄影	AI 语义分割视频引擎，支持16类场景分割调整，提供视觉效果更佳的专业级电影视频捕捉体验 借助 ISP 和APU直联系统，可实现录制视频低延迟预览 先进的AI智能景深引擎
AI显示	支持 Google Ultra HDR 显示技术——从拍摄到显示的端到端HDR 率先支持环境光自适应 HDR 技术
Sub-6GHz 全频段 5G 网络	内置具有情境感知功能的AI

数据来源：联发科官网，国泰君安证券研究

系统应用

- **高通：发布高通AI Stack，端到端的AI软件解决方案。**高通AI软件栈是面向OEM厂商和开发者的一套完整的AI解决方案，通过丰富的AI软件权限和兼容性，能够支持各种智能终端，包括智能手机、汽车、XR、计算、物联网和云平台。高通AI软件栈支持包括TensorFlow、PyTorch和ONNX在内的不同AI框架与主流runtimes，以及开发者库与服务、系统软件、工具和编译器，使得任何面向单一终端开发的AI特性都可在其他终端上轻松部署。以Stable Diffusion为例，高通从Hugging Face的FP32版本1-5开源模型开始，通过量化、编译和硬件加速进行优化，最后实现在搭载骁龙8 Gen 2移动平台的手机上运行。
- **联发科：发布AI开发平台NeuroPilot，构建丰富AI生态。**NeuroPilot支持Android、Meta Llama 2、百度文心一言大模型、百川智能百川大模型等前沿主流AI大模型，为用户带来包含文字、图像、音乐等领域在内的终端侧生成式AI的创新体验。

高通发布高通AI Stack

数据来源：高通官网

联发科发布NeuroPilot

数据来源：联发科技公众号

智能搜索升级

- 智能手机是日常快捷信息的来源，伴随着大量搜索需求。基于生成式AI的查询能够提供精准、多轮次的交互回答，逐步改变用户的搜索方式。通过自然语言理解、图像理解、视频理解、语言转文字、文本生成等模型，在端AI能够理解用户的各种输入，并提供建议与操作。智能手机将成为真正的数字助手。

整机性能管理

- AI通过学习用户的使用习惯，能够实现个性化整机性能管理。整机性能管理主要体现两方面：能耗管理和系统优化管理。
  - ✓ **能耗管理：**例如谷歌在Android P系统中加入的“自适应电池”功能，随时监控用户电量消耗情况，让手机智能判断用户对APP的使用情况，并自动关闭无用的后台应用，大幅降低耗电量；
  - ✓ **系统内部资源（内存）智能感知分类：**由于无用权限获取、碎片化文件、垃圾数据日积月累等因素，导致手机内存不堪重负，卡顿增加。AI通过训练感知用户的使用习惯，进而预测用户未来行为，并进行有效资源的再分配。

- **现象级AI应用爆火，推理功能向边端转移。**ChatGPT iOS版APP上架首周下载量超过50万次；聊天机器人应用Chatacter.AI上市不到一周便吸引170万用户安装。AI为手机带来全新功能，深度参与日常生活当中。目前，性能强大的生成式 AI 模型正在逐渐变小；同时终端侧处理能力持续提升。如StableDiffusion 等参数超过10亿的模型已经能够在手机上运行，且性能和精确度达到与云端处理类似的水平。AI在端运行不受网络影响并能更快反应，极具应用前景。在不久将来，或将看到拥100亿或更多参数的生成式AI模型能够在终端上运行。**除搜索、聊天功能外，AI功能将主要在图片视频、语音两部分实现，能够增强设备的环境感知能力，并为用户提供个性化加工窗口。**

数据来源：高通《混合AI是AI的未来》

AI SoC：“CPU+GPU+专用芯片+ISP”全面升级

- **AI功能产生更多任务、计算需要，推动CPU、专用芯片、ISP全面升级：1）CPU进入全大核时代。**随着智能手机处理器高性能需求与日俱增，推动CPU向全大核设计发展。联发科新推出的天玑9300上采用4个最高频率可达3.25GHz的Cortex-X4超大核，以及4个主频为2.0GHz的Cortex-A720大核设计；**2）专用芯片（NPU、APU）实现机器学习以及AI大模型在端运行。**专用芯片为AI运行加速，目前高通NPU支持10亿参数、联发科APU支持百亿参数；谷歌采用多个机器学习小模型的软硬件配合设计。随着AI在端运行逐步普及，需求将向着更高性能、更高精度的更大模型发展，有望推动专用芯片迭代升级；**3）ISP集成AI引擎，并与专用芯片协作，实现图片视频拍摄优化**

骁龙8 Gen3和天玑9300 SoC全线升级

硬件	高通-骁龙8 Gen3	联发科-天玑9300
CPU	性能提升30% 效能提升20%	性能提升40% 多核功耗较节省33%
GPU	性能提升25% 效能提升20% 支持硬件光线追踪和240FPS游戏	峰值性能提升46%，功耗降低40% 在复杂游戏场景中可节省40%的内存带宽 移动端硬件光线追踪性能提升46%
专用芯片（NPU、APU）	性能提升98% 效能提升40%	生成式AI transformer运算速度快8倍 2倍整数和浮点运算速度 功耗较前一代降低45%
ISP	首个认知ISP，利用A神经网络让摄像头在情境中感知人脸、面部特征、头发、衣服和天空等，分别进行独立优化，	AI语义分割视频引擎 4K深度与光斑双引擎 全像素自动对焦+倍增无损变焦引擎 率先采用独立OIS光学防抖硬件，防抖运算速度提升3倍
内存	支持LPDDR5X RAM	率先支持LPDDR5T 9600Mbps

数据来源：高通官网，联发科官网，国泰君安证券研究

内存：LPDDR5渗透率提升

- **AI内存需求推动内存需求提升以及LPDDR5普及。**对于130亿参数AI大模型而言，如采用INT8精度，则除操作系统外至少需要12.1GB内存才能运行。而旗舰智能手机的内存容量大多也只有16GB，存在内存限制。目前高通与联发科分别采用了精度下调和内存硬件压缩的方法来缩小内存需求。随着模型增大、精度提升，更高速、更大容量的内存需求也将显现。高通骁龙8 Gen3、联发科天玑9300均已采用LPDDR5系列内存，或将成为趋势。

无线通信：WiFi 7迭代升级

- **AI推动AIoT需求，WiFi 7有望迭代升级。**在AIoT时代，手机成为智能生活的控制中心，通过WiFi、蓝牙、物联网等无线通信方式，能够与电脑实现跨屏操作、与电视实现高保真投屏、与智能穿戴设备实现信息交互、远程控制智能家电。随着AI进驻手机，这将大幅提升其感知用户需求的能力，并以指令分发的形式控制各个智能生活终端，实现更为流畅的生活体验。Wi-Fi 7在Wi-Fi 6的基础上引入了320MHz带宽、4096-QAM、多链路操作等技术，能够提供更高的数据传输速率和更低的时延，为AI数据交互赋能。

小米手机与各类终端互联

数据来源：小米官网

光感知：高性能潜望式镜头方案替代加速

- **AI优化潜望式镜头方案，有望推动落地进展。**在中国移动互联网应用中，短视频应用时长占比高达28.50%，是核心的应用场景，拍摄功能最受消费者关注。从物体到拍摄图像，需要经过光学系统、传感系统、计算系统的整套光感知系统。随着AI在手机端落地，多模态大模型将重塑光感知系统，创造更高性能、更高质量的拍照功能。潜望式镜头可以有效提高光学变焦倍数，大幅提升防抖能力和成像效果，但需要解决色散、聚焦、功耗等问题，AI算法有望弥补缺陷，推动潜望式镜头落地进展。

数据来源：艾瑞咨询

- **2023年第三季度全球智能手机出货量同比下降，AI有望驱动需求反弹。**据IDC数据，2023年第三季度全球智能手机出货量下跌1%，同比持续改善。在区域性复苏和新产品升级需求的带动下，全球智能手机市场在第三季度达到两位数的环比增长，市场逐步复苏。随着谷歌Pixel 8发布，高通骁龙8 Gen3、联发科天玑9300相继问世，各大厂均进入AI手机赛道。AI全新应用有望带动换机热潮。

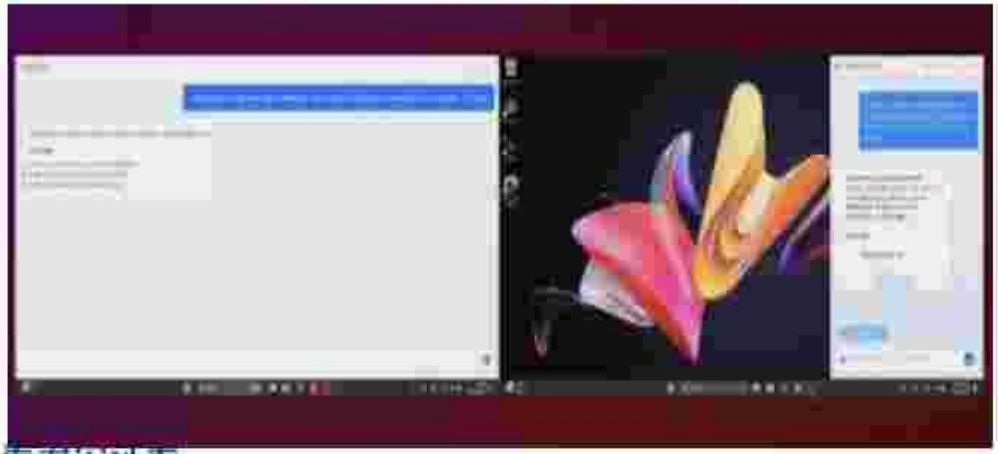
数据来源：IDC，国泰君安证券研究

联想：发布全球首款 AI PC，提出针对企业端的混合人工智能计划

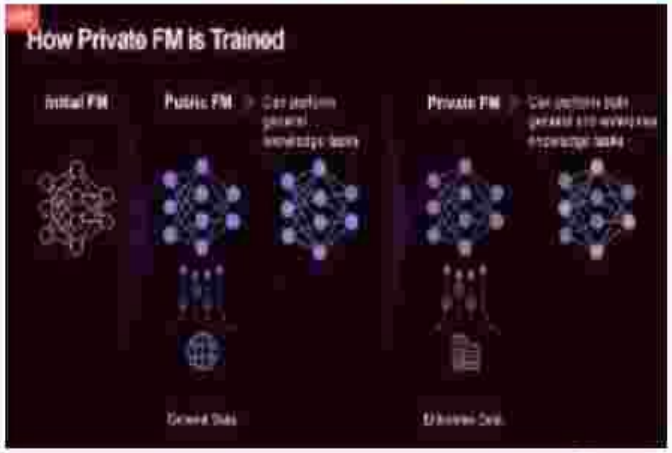
终端产品

- **联想：发布全球首款AI PC：可运行个人大模型。**联想AI PC能够创建个性化的本地知识库，通过模型压缩技术运行个人大模型，实现AI自然交互，为每个人量身定制的全新智能生产力工具，进一步提高生产力、简化工作流程，并保护个人隐私数据安全，将人工智能带给每一位用户。通过结合个人数据，可以做到更好的大模型效果，在联想AI PC模型演示中，针对同一个提问，PC级AI Lenovo AI Now相较于云端AI生成回答速度稍慢，但更具个性化。
- **针对企业端，联想提出混合人工智能计划。**联想认为，通过公共大模型与企业大模型相结合，可以解决企业的数据安全担忧。在最初大模型的基础上，企业根据特定数据进行额外的训练和微调并在端侧加入企业知识矢量数据库中的企业特定知识，最后，链接旧有的ERP 系统、CRM 系统、MES 系统等供应商数据库，即可得到一个混合的AI 系统，既能够既保证数据安全，也具有泛化知识，同时能够回答企业相关的特定问题，帮助企业规划相关活动。

联想PC大模型与云端大模型并列演示



企业训练私有大模型



数据来源：联想发布会演讲公告

03 英特尔：Meteor Lake处理器平台，首次将神经网络处理单元（NPU）集成到PC处理器

核心硬件

英特尔

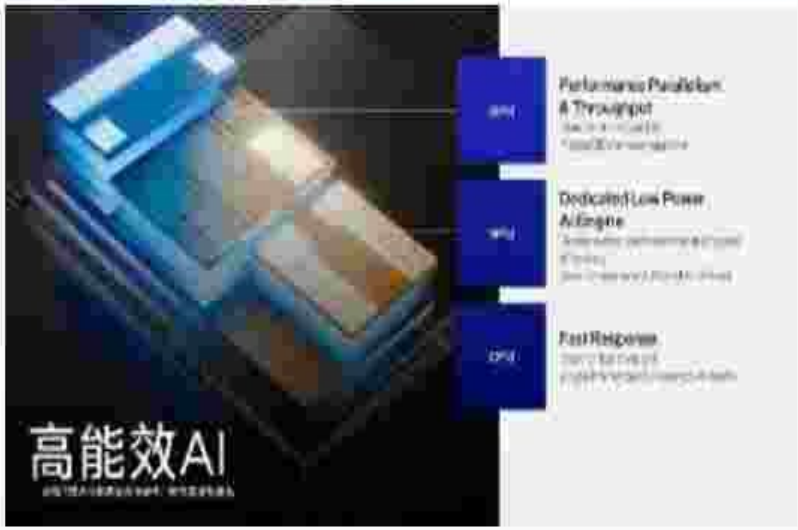
AMD

高通

- **在英特尔on技术创新大会上，英特尔推出了首款基于Intel 4制程工艺打造的Meteor Lake处理器平台。**Meteor Lake采用分离式模块架构，由计算模块、SoC模块、图形模块以及IO模块这4个独立模块组成，并通过业界出众的Foveros 3D封装技术连接。在SoC模块中，Meteor Lake采用了创新的低功耗岛设计，集成了NPU，为PC带来了高能效的AI功能表现，并兼容OpenVINO等标准化程序接口，便于AI的开发及应用普及。新的低功耗能效核，进一步优化节能与性能间的平衡。NPU除了专为持续的 AI带来高能低耗的表现，还可以通过AI卸载能力，通过NPU降低CPU和GPU的AI工作负载。实现PC上AI场景的长续航加速。

Meteor Lake中的SOC模块

Meteor Lake中的高效能AI功能



数据来源：英特尔发布会演讲公告

核心硬件

英特尔

AMD

英伟

- **微软公布未来AI 处理器蓝图，推动高能效AI规模化发展。**下一代的Arrow Lake将覆盖桌面和移动平台，主要侧重于提供更高的功率和性能，采用了与Meteor Lake相同的设计方法，不过会改用更新的Intel 20A工艺制造。LunarLake则为英特尔下一代低功耗架构，将进一步提升人工智能加速效率，并对Meteor Lake和Arrow Lake的多芯片设计做了改进，计划在面向移动平台的酷睿Ultra第2代处理器Arrow Lake之后发布。Lunar Lake将采用LionCove架构的P-Core和Skymont架构的E-Core，全新的微架构将提供突破性的每瓦性能优势，同时会采用Intel18A工艺制造，标志着该技术的首次商业应用。

微软未来AI PC处理器蓝图

微软未来AI PC处理器工艺节点

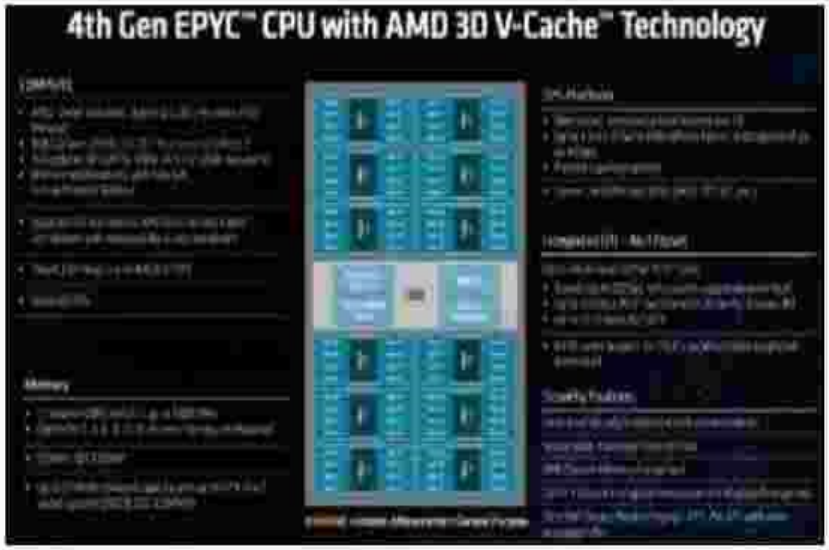
数据来源：英特尔发布会演讲公告

AMD：提供从数据中心到工作站、笔记本电脑的人工智能优化解决方案

核心硬件

- **在联想2023 Tech World大会上，ADM提出将为联想提供从数据中心ThinkSystem到ThinkStation工作站和ThinkPad笔记本电脑的人工智能优化解决方案。**AMD将为联想个人电脑和数据中心产品组合提供人工智能优化解决方案，产品组合将由AMD Ryzen、EPYC™和Instinct™处理器驱动。（1）第四代AMD EPYC处理器最多可配备96个核心，可用于加速一系列数据中心和边缘应用，包括客户支持、零售、汽车、金融服务、医疗和制造业。（2）Instinct为AMD的加速卡系列，Instinct MI300为全球首款同时集成CPU、GPU的数据中心APU。Instinct MI300A一共有多达13颗小芯片，其中计算部分9颗，都是5nm工艺制造。CPU部分为Zen4架构，三颗CCD芯片，24个核心，GPU为最新的CDNA3架构，六颗XCD芯片，核心单元数量仍未公布，还有128GB容量的HBM3高带宽内存，可以为CPU、GPU所共享。

AMD 第四代 EPYC



数据来源：AMD官网

AMD Instinct GPU

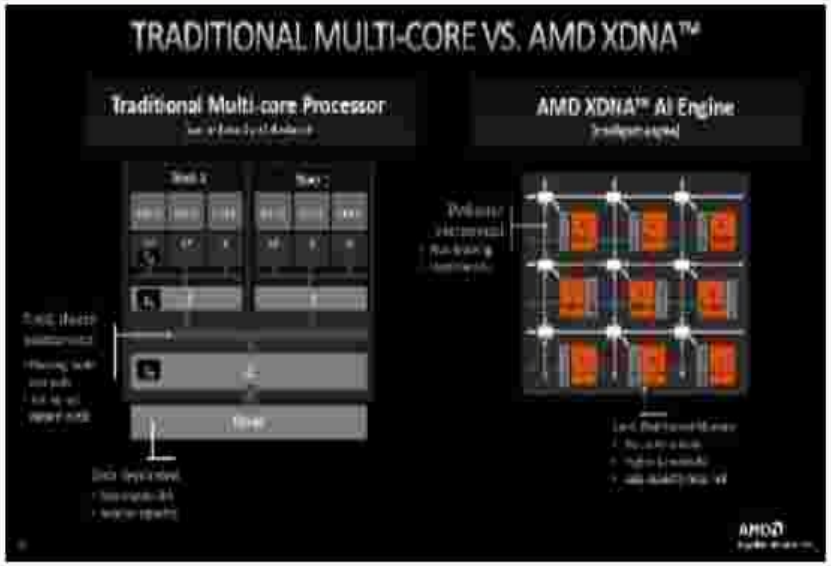
数据来源：AMD官网

03 AMD：锐龙7040系列处理器中加入Ryzen AI引擎，首款基于x86处理器的专用人工智能硬件

核心硬件

- AMD的Ryzen AI引擎基于专门设计的AMD XDNA AI架构，其核心是多个独立的AIE单元，而且与传统的CPU多核运算相比，XDNA AI架构有多个独立的内存电压单元以及对应的内存控制器并拥有独立的高速互连通道，在进行AI推理时更加灵活，效率更高，性能也更强。移动平台的XDNA AI架构有着超高的能效、强大的算力，它支持不同的AI神经网络，比如CNN（卷积神经网络）、RNN（循环神经网）、LSTM（长短时记忆）等。它还支持Int8/16/32、BFloat16等各种高级数据类型，同时XDNA AI架构还具备实时多任务能力，可处理最多4条并发空间流。

AMD XDNA AI架构不同于传统多核运算



数据来源：AMD官网

Ryzen AI引擎基于专门设计的AMD XDNA AI架构

数据来源： AMD官网

核心硬件

英特尔
AMD
高通

- 骁龙X Elite支持在终端侧运行超过130亿参数的生成式AI模型，引入了高通AI引擎核心Hexagon NPU，通过全新供电系统升级使NPU按照工作负载适配功率，兼顾高性能与低能耗，引入微切片推理，加速Transformer网络等复杂AI模型研发提供支持。同时，高通对张量加速器进行了升级，大矩阵处理速度提升2.5倍。共享内存规模增加了一倍，便于容纳更大的神经网络，使NPU实现了45TOPS的AI性能。骁龙X Elit 采用4nm制程工艺打造，拥有12个高性能核心，集成的Adreno GPU支持每秒4.6万亿次浮点运算，主频高达3.8GHz，支持双核增强，最大可达4.3GHz，也是首个4GHz以上的ARM架构CPU核心。此外，它拥有高达42MB的总缓存容量，内存带宽136GB/s，并支持八通道LPDDR5x。

骁龙X Elite性能

数据来源：高通官网

系统应用

微软

- Windows Copilot是微软推出的一款利用AI技术，帮助用户在Windows系统中更高效、个性化地完成工作、创作和娱乐任务的智能助手，能够预览并加速任务，减少摩擦，节省时间，并提供个性化的答案、灵感和任务帮助。Copilot系统权限高于Cortana（将于2023年底停止支持），可以实现自然语言交互级别的对话，与系统应用更深度地进行绑定，例如它直接读取浏览器页面内容，生成总结文字。Copilot还更新了画图、照片、Clipchamp等应用，其中画图功能可以智能移除图片背景并增加图层功能，照片新增了AI背景虚化功能，Clipchamp提供自动剪辑图片和视频与场景描述功能。微软 Copilot 作为 Windows 11 免费更新的一部分直接嵌入每一位用户的操作系统，优化了系统存在感，默认集成于任务栏，具备专门的Win+C热键，很大程度上培养了用户对人工智能的使用习惯。

Windows Copilot系统

微软365 Copilot为微软企业办公系列核心

数据来源：微软官网

03 英特尔：AI PC加速计划，构建AI PC生态系统解决方案，预计2025年达成逾1亿台AI应用

系统应用

英特尔

- AI PC加速计划旨在联结独立硬件供应商（IHV）和独立软件供应商（ISV），并充分利用英特尔在AI工具链、协作共创、硬件、设计资源、技术经验和共同推广的市场机会等资源。这些资源将帮助产业合作伙伴充分发挥英特处理器的技术和相关的硬件优势，以尽可能最大限度发挥AI和机器学习（ML）应用的性能，加速全新应用案例，并吸引更广泛的PC产业伙伴融合到AI PC生态系统的解决方案中。

英特尔的AI PC加速计划

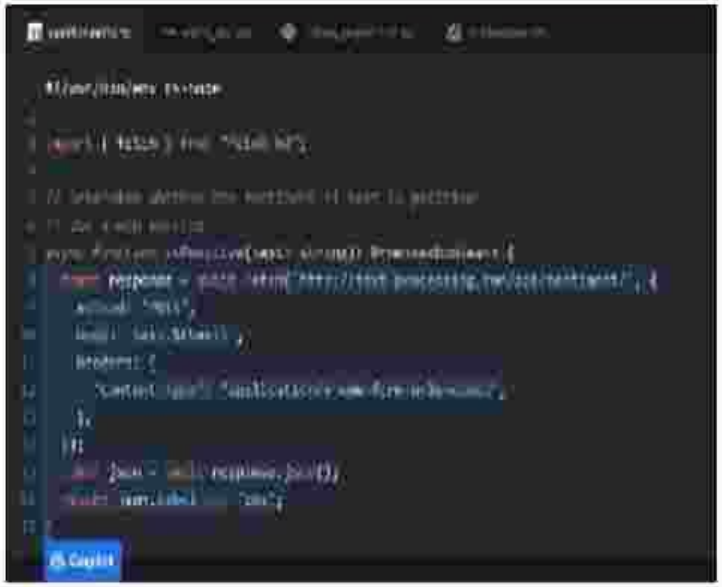
数据来源：英特尔官网

- **解放生产力**：作为设备、边缘计算和云技术的颠覆性混合体，AI PC将不仅具有强大的计算能力和先进的AI技术，还能全面满足新的生成式AI工作负载需求。以Microsoft Office 365 为例，全球有超过 4 亿 Microsoft Office 365 商业付费席位和个人订阅者，如果将生成式 AI 集成至用户日常工作流将带来重大影响。例如：人工智能编程助手GitHub Copilot能够提高开发效率、降低编程门槛和成本。
- **赋能创造力**：AI PC统一人际语言、人机语言和机机语言。例如，先进的AI系统（如OpenAI的Codex）能够将自然语言描述转换为代码，为非专业开发者提供了编程的可能性，使得从事编程工作的门槛大大降低，创造性的软件开发不再是少数专家的专利。借助AI PC，每个有兴趣的人都可以参将自己的想法转化为实际的应用程序，将真正的创造力交给了每一个人。

AI PC发展方向

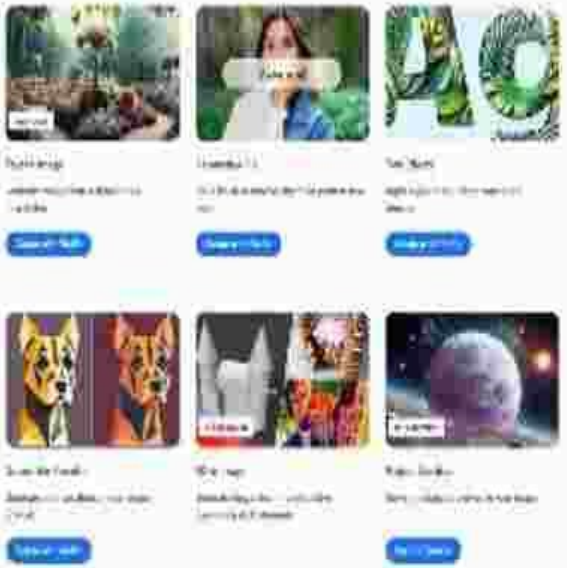
数据来源：群智咨询

人工智能编程助手GitHub Copilot使用界面



数据来源： GitHub Copilot

创意内容生成大模型Adobe Firefly的六大功能



数据来源： Adobe官网

- **AI 能为 XR 带来巨大前景。**它有潜力普及 3D 内容创作，并真正实现虚拟化身。下一代 AI渲染工具将赋能内容创作者使用如文本、语音、图像或视频等各种类型的提示，生成 3D 物体和场景，并最终创造出完整的虚拟世界。此外，内容创作者将能够利用文本生成文本的大语言模型，为能够发出声音并表达情绪的虚拟化身生成类人对话。高通认为，在未来几年里首批文本生成 3D 和图像生成 3D 类的模型将可能实现边缘侧部署，生成高质量的3D 物体点云。随后几年里模型将提升至能从零开始生成高质量 3D 纹理物体的水平。十年内模型将更进一步支持由文本或图像生成的高保真完整3D空间和场景。用户最终或能从零开始生成3D虚拟世界，例如自动构建满足用户任何想象的 3D 虚拟环境。

3D 内容创作为AI PC发展方向之一

数据来源：高通《混合AI是AI的未来》

硬件升级

处理器

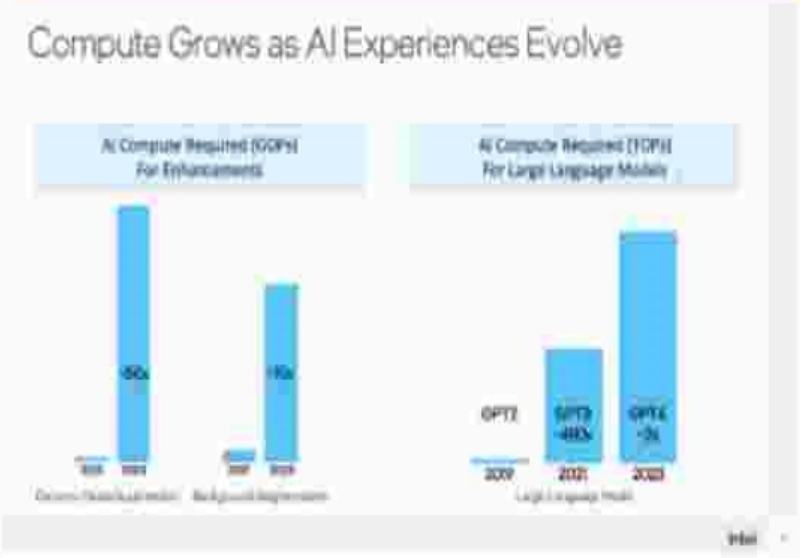
内存

接口

散热系统

- **AI PC应用的稳定运行需要硬件提供充足算力支持，消费级PC AI芯片将走向市场。**AI PC处理器将在提高能效和散热的同时，拥有更多的核心和处理单元，以提供更高的并行处理能力；集成专门的AI加速单元，如TPU、NPU等以快速处理矩阵乘法。多家芯片巨头已入局AI PC市场：高通发布采用ARM内核的首款PC芯片X Elite；英特尔计划于发布整合人工智能技术的Meteor Lake处理器，AMD在Ryzen 7040系列PC处理器中配备了基于Xilinx IP的专用AI引擎，可加速PyTorch和TensorFlow等机器学习框架的运行。

PC应用场景的AI算力需求提升



数据来源：微软官网

Meteor Lake集合 CPU、GPU、VPU单元

数据来源：英特尔官网

硬件升级

处理器

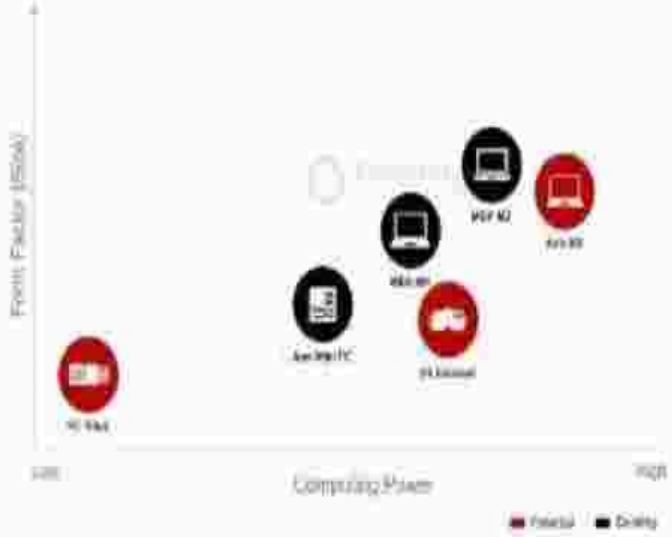
内存

接口模组

散热系统

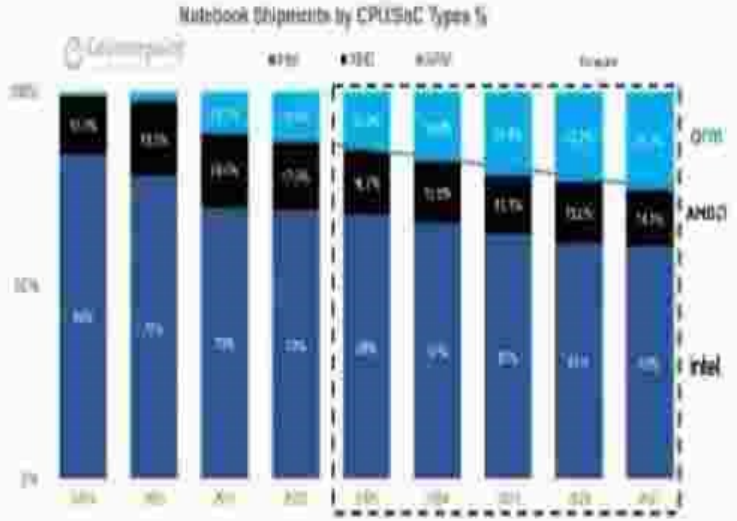
- **各大厂商陆续发布直接最新Arm PC规划，促进Arm PC生态发展。**Arm架构指令的精简使得平均指令运行周期缩短，增加了通用寄存器以减少读写操作，程序的执行效率与能耗比因此得到了大幅度提升。高通发布了新款PC处理器骁龙X Elite，英伟达与AMD也被曝正在采用ARM架构技术设计可运行微软Windows的CPU。随着各大厂商陆续发布最新的Arm PC，Counterpoint预计，随着更多芯片厂商推出Arm架构的PC芯片，Arm架构的市场份额有望上升。到2027年，Arm架构芯片在PC市场的份额预计为25.3%，较2022年增长近一倍；x86架构的总体份额将下滑至74.4%，其中英特尔的份额下滑至60%，但依然将保持大比例领先。

PC应用场景的AI算力需求提升



数据来源：Counterpoint

按CPU/SoC类型划分的PC出货量份额



数据来源：Counterpoint

硬件升级

- 处理器
- 内存
- 接口模组
- 散热系统

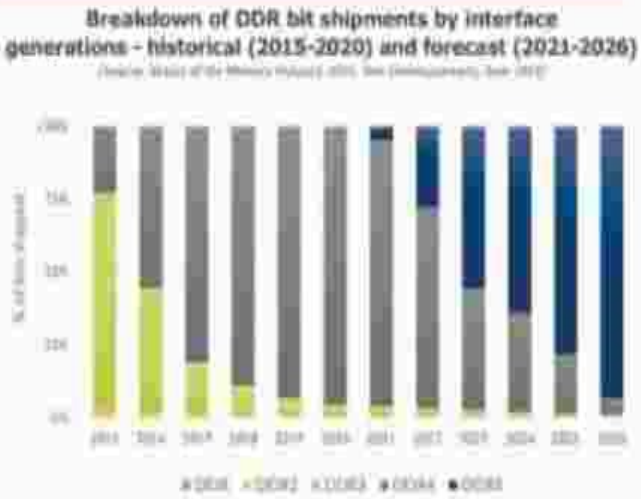
- **为满足AI应用对速度和效率的日益增长的需求，AI PC采用DDR5或成必然趋势。**较DDR4，DDR5可支持单一芯片容量最高容量上限可增加到128GB，传输速度提升两倍。电源管理IC方面，电源管理芯片内建至DDR5内存中，除可提高讯号完整性及噪声辨识能力，还能更具效率与扩展性，此外，双通道的设计也大幅降低延迟，提高内存效率并提升系统稳定度。根据研究机构 Yole Developments 的预测，到 2023 年，DDR5 内存出货量将超过 DDR4，到 2026 年 DDR5 销量将占市场的 90%。随着AI应用程序变得越来越复杂，对内存的需求也随之增加。DDR5通过提供更大的容量和更快的速度，使得单个系统能够处理以往需要分布式或专业硬件处理的任务。此外，DDR5的低功耗特性对于构建能源效率更高的系统至关重要，这对于遏制运行成本和环境影响尤其关键。

DDR5规格与效能整理

Max Die Density				
Max DIMM Size				
Max Data Rate				
Channels				
Width (per ch)				
Banks (Per Group)				
Bank Groups				
Burst Length				
Voltage (Vdd)				
VPP				

数据来源：JEDEC

DDR 5出货份额预测



数据来源：Yole

硬件升级

处理器
内存
总线接口
散热系统

- **传输速率成倍提升。**PCIe5.0的规范于2019年发布，相比PCIe4.0，PCIe5.0最大的提升莫过于速率的提升，相比PCIe4.0 16GT/s的传输速率，PCIe5.0提升到了32GT/s，速率整整提升了一倍。
- **优化信号完整性。**PCIe5.0向后兼容用于外接插件卡的CEM连接器。在AI应用中，尤其是需要实时反应的场合（如自动驾驶或者高频交易），PCIe 5.0能够提供更低的延迟，使得决策过程更加迅速和精确。
- **支持高速存储设备。**AI应用经常涉及到大型数据集，PCIe 5.0可以支持更快的NVMe存储设备，这些设备可以提供快速的数据读写速度，有利于大数据量的处理，赋能高性能计算。

PCIe 5.0 传输速度远快于上一代

	x1 Max Unidirectional Bandwidth	x16 Max Unidirectional Bandwidth	Maximum Bidirectional Bandwidth
PCIe 1.0	250MB/s	4GB/s	8GB/s
PCIe 2.0	500MB/s	8GB/s	16GB/s
PCIe 3.0	1GB/s	16GB/s	32GB/s
PCIe 4.0	2GB/s	32GB/s	64GB/s
PCIe 5.0	4GB/s	64GB/s	128GB/s

数据来源：英特尔官网，国泰君安证券研究

硬件升级

- 处理器
- 内存
- 接口模组
- **散热系统**

- **相较于普通PC，AI PC高性能的处理器进行大量并行计算时会产生显著的热量，需要更高效的散热方案。**未来的AI PC可能在以下几方面改善散热系统：（1）风冷系统：更大的散热片与设计更复杂的风道来提高空气流动效率，增加风扇转速，以及应用更为先进的热传导材料（如石墨烯等）更快地将热量从处理器传导到散热器。（2）液冷系统：采用定制的水冷套件或闭环液体冷却系统，使用具有高热传导率的冷却液，以及细化和优化散热片设计，使冷却系统可以覆盖多个发热组件，如CPU、GPU和AI专用加速器，将水冷系统的应用由台式机扩展到笔记本电脑。

新型散热材料石墨烯

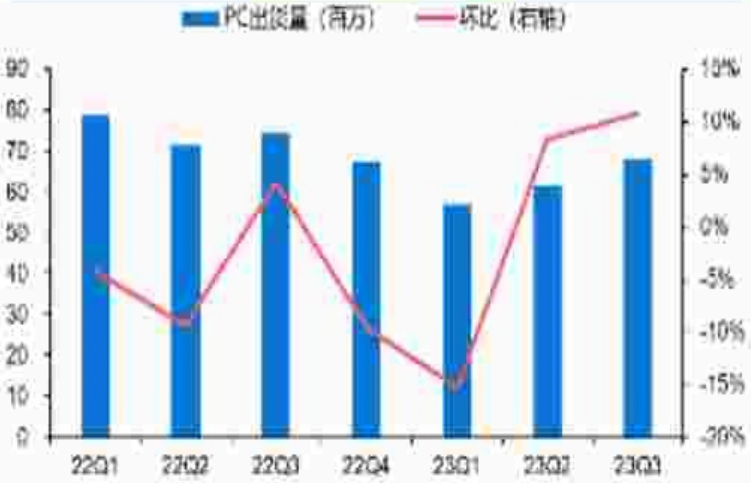
数据来源：涂料技术学会

联想Legion 9i 全球首款内置液冷的笔记本电脑

数据来源：联想官网

- **2023年第三季度全球PC出货量再次下降，生成式人工智能有望成为PC行业的分水岭。**根据IDC数据，2023年第三季度PC出货量继续螺旋式下降，全球出货量6,820万台，同比下降7.6%。尽管市场需求和全球经济仍然低迷，但个人电脑出货量在过去两个季度均有增加，同比下降速度趋缓，这表明市场已经走出低谷。生成式人工智能有望成为PC行业的分水岭。人工智能PC的开发者向使用机构承诺，他们有能力在保护数据隐私和主权的同时深层次个性化用户体验。随着明年更多相关设备问世，预计整体售价将有所上涨。
- **AI PC将会促进用户对AI智能化的强劲感知，从而坚定用户对于AI PC的信赖度，刺激用户换机需求。**根据群智咨询预测，2024年伴随着AI CPU与Windows 12的发布，将成为AI PC规模性出货的元年。预计2024年全球AI PC整机出货量将达到约1300万台。在2025年至2026年，AI PC整机出货量将继续保持两位数以上的年增长率，并在2027年成为主流化的PC产品类型，意味着未来五年内全球PC产业将稳步迈入AI时代。

PC出货量（百万）



数据来源：IDC，国泰君安证券研究

AI PC出货量预测

数据来源：群智咨询

- **Humane发布AI PIN，穿戴式AR人工助理。**Ai Pin的定位是人工智能个人助理，配备了AI聊天助手。AI PIN像个精美的小方盒，与领夹式麦克风外形接近，共有3种颜色，分别是Eclipse（日食）、Lunar（月光）、Equinox（破晓）。此外，AI PIN还有一个磁吸式背夹，通过具有磁力的夹子来固定在衣服，提供便捷的佩戴，整机仅重34.2g。

AI PIN产品形态

数据来源：humane官网

- **搭载高通骁龙芯片，集成摄像头、激光投影器、深度传感器。**配置方面，AI PIN搭载了2.1GHz八核高通骁龙芯片，RAM内存为4GB，ROM内存为8GB；支持蓝牙5.1连接与eSIM功能，提供多种连接方式。设备表面有两个指示灯，可作为来电提醒、消息通知。AI PIN正面集成了摄像头、激光投影仪、深度传感器、双麦克风阵列以及用于触摸交互的触摸板。AI PIN没有电子屏幕，通过内置的小型投影仪，可在用户的手掌或是任何固体表面上，投射一个分辨率为720p的画面，用以显示时间、日期、查看收到的信息等。

AI PIN计算端配置

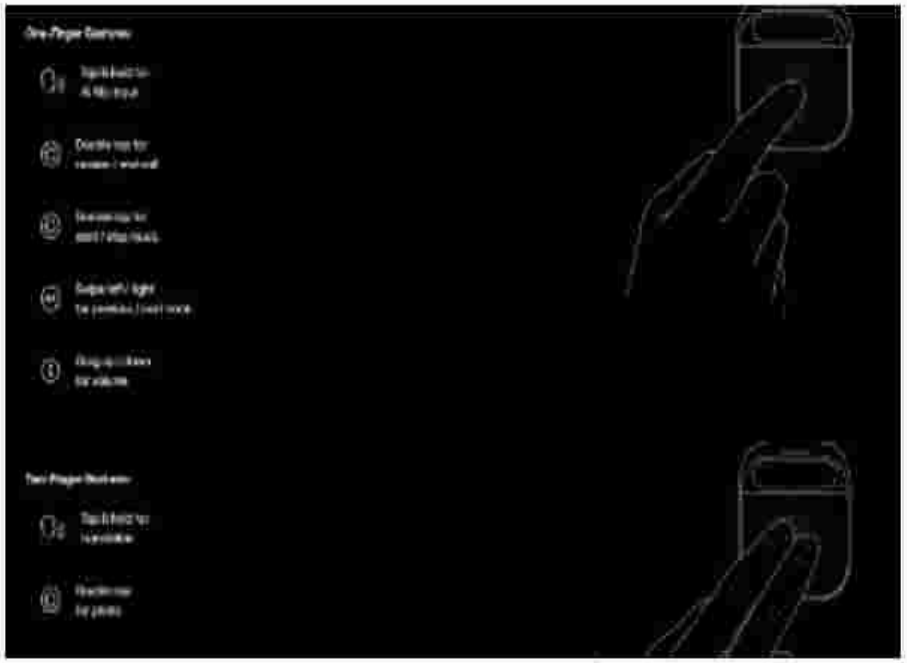
AI PIN链接端配置

数据来源：humane官网

- **脱离显示屏交互，采用声音+触摸+手势+投影四种核心交互方式。**AI PIN使用触摸板代替屏幕，对人类的手势做出反应。而且在不连接智能手机的情况下就可以使用，主要是通过T-Mobile 的无线网络服务来实现。使用者可以通过点击、手势和语音命令来控制AI PIN，两根手指在设备正面触摸板上轻按两次即可拍摄照片；同样双击然后按住该位置即可录制视频；轻按Pin，然后将手掌移入视野中，即可激活激光投影，激光将图像和文本投射到用户的手上。AI PIN为用户提供了一种无屏幕且直观的与人工智能互动的方式。凭借其激光投影系统和与强大的人工智能算法的集成，有潜力重新定义电子产品的交互方式，便利使用者的日常生活。

AI PIN通过上下摆动手掌来“选择”按钮，捏起拇指和食指来“确认”

AI PIN的交互触摸动作



数据来源：humane官网

- **AI PIN内置AI大模型和聊天助手Ai Mic。**Ai Mic类似ChatGPT，用户长按Pin即可与Ai Mic进行自然语言交流、网络搜索。相关AI功能还包括根据语气撰写短信、充当实时外语翻译、推荐餐厅、识别食物营养含量等。AI助手还能够梳理短信、邮件等不同渠道的信息并总结成回复，实现进阶搜索功能。
- **AI PIN彻底脱离了APP和显示屏交互，功能实现高度便携。**Humane采用AI操作系统Cosmos和AI软件框架Ai Bus。用户无需下载、管理或启动应用程序。系统能够快速了解用户的需求，并自动连接到正确的AI体验或服务。

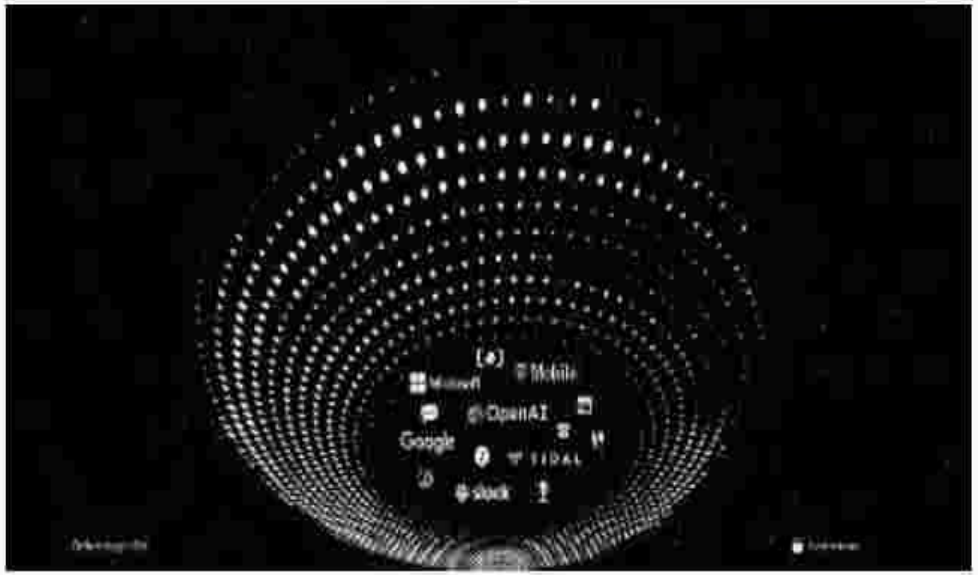
AI PIN会自动检测周围环境的常用语言，并在对话过程中把当地方言和用户语言互译，打破沟通的语言壁垒

AI PIN能够根据自然语言自动生成短信内容并发送

数据来源：humane官网

- **用户的照片、视频、笔记均上传至云端Humane.Center账户，以供查看、共享和搜索。**该平台将充当AI PIN的服务中心枢纽，来实现个人信息存储和个性化操作设置。
- **AI PIN定价699美元起，另需支付24美元/月订阅费，预计将于2024年初开始发货。**订阅费包括访问Ai Bus、无限制的无线数据服务计划（包括通话、短信、数据）和云存储、不断增长的AI服务套件的完全访问权限、以及无限制的查询次数。目前OpenAI、微软、谷歌、Slack等均为AI PIN提供服务。

OpenAI、微软、谷歌、Slack等均为AI PIN提供服务



AI PIN获取的内容均上传至云端账户以供查看

数据来源：humane官网

04

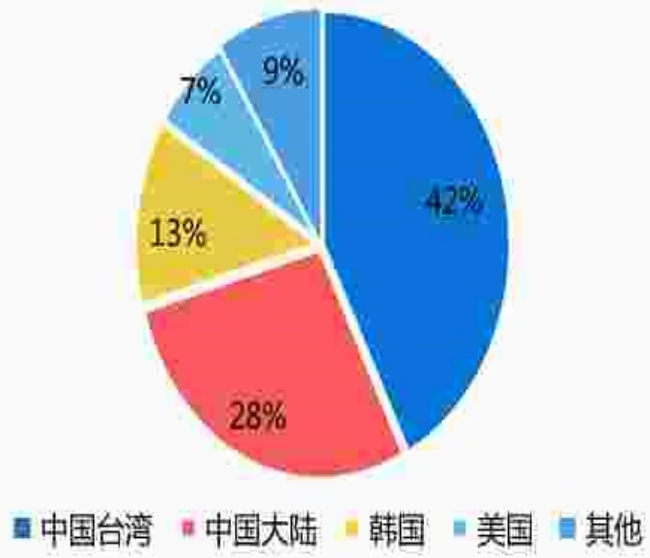
成熟制程国内产能市占率较高，先进制程急需突破

全球晶圆代工厂：整体2024年有望回暖，市占率呈现一超多强局面

- 自2022年下半年以来，受整体市况不佳，终端需求疲软，供应链库存持续去化影响，晶圆代工厂的整体产能利用率偏低，全球整体晶圆代工市场销售额出现下滑。**2023年全球晶圆代工产业的营收同比下滑12.5%。**
- 随着终端市场的逐步回暖，预计2024年整个市场将会出现6.4%的同比增长，达到1272.71亿美元。
- 市场份额方面，台积电以60%份额高居第一；产能方面，中国台湾企业处于主导位置，大陆晶圆厂积极追赶，产能占比有望从2022的24%上升至2027年的28%

2024年全球晶圆厂市场份额占比预测

2027年全球晶圆厂产能占比预测



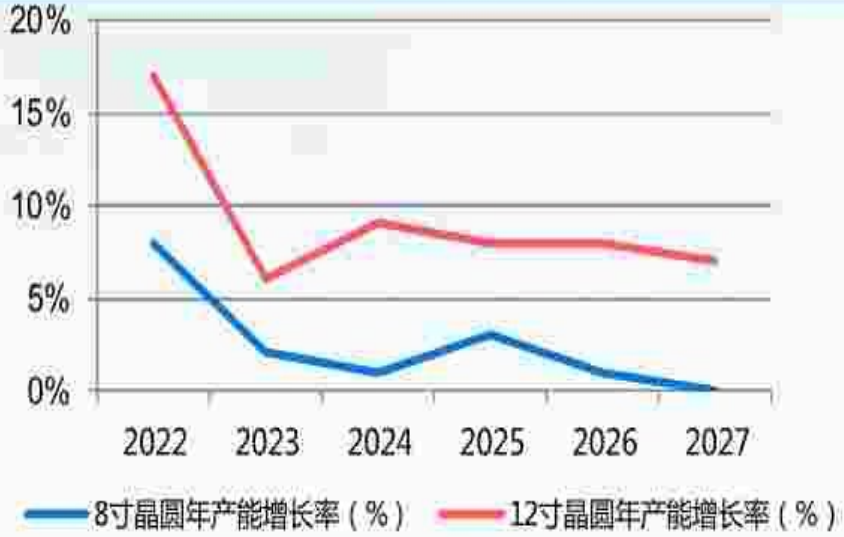
资料来源：TrendForce，国泰君安证券研究

全球晶圆产能：8寸及12寸产能利用率回升，未来晶圆扩产以12寸为主

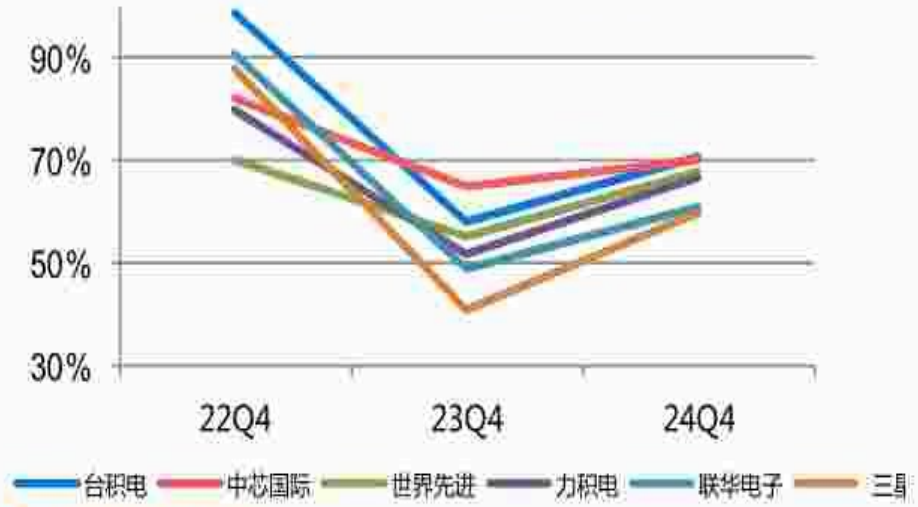
目前12寸晶圆已成主流。从未来晶圆产能的增长来看，到2027年，12寸晶圆年复合增长率将达7.4%；8吋晶圆产能年复合增长率将只有1.4%。

晶圆产能利用率逐步回暖。受2021年持续缺芯影响，晶圆厂产能利用率达到高点，自22年Q2开始下滑。预计进入2024年，随市场持续复苏，大部分晶圆代工厂产能利用率将持续攀升。

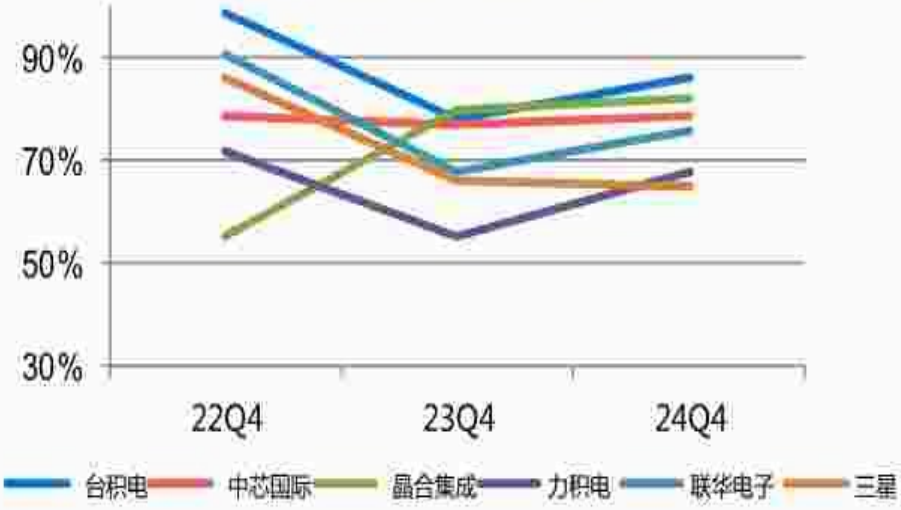
晶圆厂扩产以12寸为主



8寸晶圆产能利用率逐步回升



12寸产能利用率缓步上行



资料来源：TrendForce，国泰君安证券研究

大陆晶圆厂：目前建成44座晶圆厂，构建强大代工能力

公司	名称	地点	晶圆尺寸(英寸)	当前产能(万片/月)
中芯国际	S1（FAB1、2、3）	上海	8	11.5
	SN1	上海	12	1.5
	B1（FAB4、6）	北京	12	5.2
	B2A、B2B	北京	12	6.2
	FAB15	深圳	8	4.4
	FAB16A	深圳	12	0
	FAB7P2	天津	8	9.5
中芯集成	-	绍兴	8	4.25
中芯宁波	N1	宁波	8	4.25
	N2	宁波	8	1.5
华虹集团	Fab1-3	上海	8	17.8
	Fab5	上海	12	3.5
	Fab6	上海	12	3
	Fab7	无锡	12	2.5
华润微	-	重庆	12	2
	-	重庆	8	5.7
	晶圆二厂	无锡	8	7.8
	晶圆一厂	无锡	6	23
晶合集成	N1、N2	合肥	12	4
武汉新芯	Fab1	武汉	12	2.5
	Fab2	武汉	12	2.5
闻泰安世	-	上海	12	3

公司	名称	地点	晶圆尺寸(英寸)	当前产能(万片/月)
长江存储	Fab2	武汉	12	0
	Fab3	武汉	12	0
合肥长鑫	Fab1	合肥	12	4
士兰微	Fab1	杭州	8	3.5
	Fab1	厦门	12	4
杭州富芯	-	杭州	12	5
广义微电子	-	四川	6	15
上海新进芯	-	上海	8	1.5
英锐半导体	-	盐城	6	2.5
福建晋华	F1-F2	泉州	12	0
芯睿电子	-	新乡	6	2
三星	Fabx1	西安	12	12
	Fabx2	西安	12	8
英特尔	Fab68二期	大连	12	4
SK海力士	HC1	无锡	12	10
	HC2	无锡	12	10
德州仪器	-	成都	8、12	5
台积电	NJFab16	南京	12	2
	FAB10	上海	8	3.5
上海先进	-	上海	8	2.3
联电-厦门联芯	FAB12x	厦门	12	2
联电-和舰科技	-	苏州	8	10

资料来源：TrendForce，国泰君安证券研究

大陆晶圆厂：持续推动本土化生产，在建及规划晶圆厂达32座

22座在建晶圆厂

公司	名称	地点	晶圆尺寸（英寸）	规划产能（万片/月）
中芯国际	SN2	上海	12	3.5
中芯东方	-	上海	12	10
中芯国际	B3P1	北京	12	10
中芯国际	FAB16B	深圳	12	10
中芯西青	-	天津	12	10
中芯集成	-	绍兴	12	1
华虹集团	Fab9	无锡	8、12	8.3
海辰半导体	-	无锡	8	10.5
华润微	-	深圳	12	48
长江存储	Fab1	武汉	12	10
紫光集团	CD	成都	12	30
粤芯半导体	粤芯三期	广州	12	4
增芯科技	南沙项目	广州	12	6
芯恩集成	芯恩二期	青岛	12	8
芯恩集成	-	青岛	8	5
士兰微	Fab2	杭州	8	4
积塔半导体	临港二期	上海	12	5
积塔半导体	-	上海	8	6
燕东微电子	-	北京	8	5
赛莱克斯	-	北京	8	3
万国半导体	CQ	重庆	12	7
华微电子	-	吉林	8	2

10座晶圆厂规划中

公司	名称	地点	晶圆尺寸（英寸）	规划产能（万片/月）
中芯国际	B3P2	北京	12	5
中芯国际	B3P3	北京	12	5
中芯国际	B3P4	北京	12	5
华虹集团	Fab8	上海	12	4
晶合集成	N3	合肥	12	4
晶合集成	N4	合肥	12	4
合肥长鑫/兆易创新	Fab2/Fab3	合肥	12	12.5
士兰微	Fab2	厦门	12	8
中科晶芯	-	成都	8	0
矽力杰	-	青岛	12	4

资料来源：TrendForce，国泰君安证券研究

先进制程、成熟制程总体情况及下游应用

- 按照产品对工艺先进度的要求来分，半导体工艺制程可以分为特色工艺和逻辑工艺。其中特色工艺产品市场占比约为40%，逻辑工艺占比约为60%。
- 其中逻辑工艺又分为成熟工艺（28纳米及以上）和先进工艺（28nm以下的节点，目前主要为16/14nm及以下）。
- 2023~2027年全球晶圆代工成熟制程（28nm及以上）及先进制程（16nm及以下）产能比重大约维持在7:3。

晶圆尺寸	工艺制程	应用领域
12英寸（先进制程）	≤10nm	高端智能手机处理器（如苹果A12、骁龙855、麒麟970等）、高性能计算（如个人电脑、服务器CPU）等
	16/14nm	智能手机处理器（如骁龙660、骁龙821）、存储芯片、个人电脑CPU、高端显卡（NVIDIA Volta、AMD Vega20）、服务器处理器（FPGA芯片）等
	20-22nm	存储芯片、中低端智能手机处理器、数字电视、移动影像等
12英寸（成熟制程）	28-32nm	WiFi/蓝牙芯片、音效处理芯片、存储芯片、FPGA芯片、ASIC芯片等
	45-65nm	DSP处理器、传感器、射频、WiFi/蓝牙/GPS/NFC通信芯片、非易失性存储芯片等
	65-90nm	模拟芯片、功率器件、物联网MCU、射频芯片等
8英寸	90nm-0.13μm	汽车MCU、基站通信设备DSP、物联网MCU、射频芯片、模拟芯片、功率器件等
	0.13μm-0.15μm	指纹识别芯片、影像传感器、通信MCU、电源管理芯片、功率器件、LED驱动IC、传感器芯片等
	0.18μm-0.5μm	MOSFET、LGBT等功率器件、嵌入式非易失性存储器芯片等

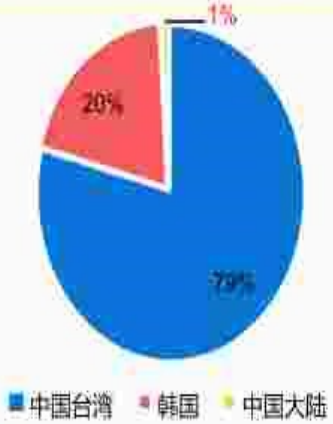
资料来源：TrendForce，国泰君安证券研究

大陆成熟制程占比上升，先进制程有待突破

- 大陆不断提高成熟制程本土化生产比例，放眼中长期，持续大幅扩产可能造成全球成熟制程产能过剩，未来可能引起价格战。
- 大陆先进制程产能占比较低，先进制程半导体设备与国外尚存差距，先进制程设备急需突破。

2022年成熟制程产能占比

2022年先进制程产能占比（大陆仅占1%）



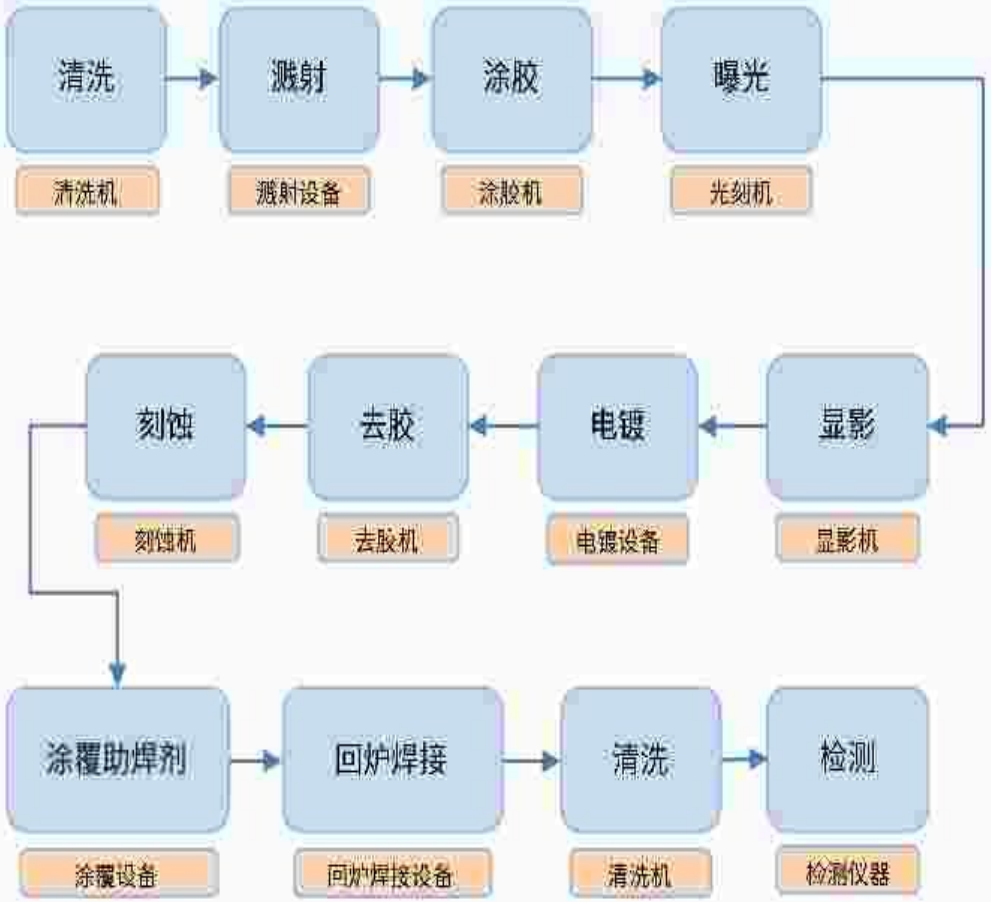
2027年成熟制程产能占比预测

资料来源：TrendForce，国泰君安证券研究

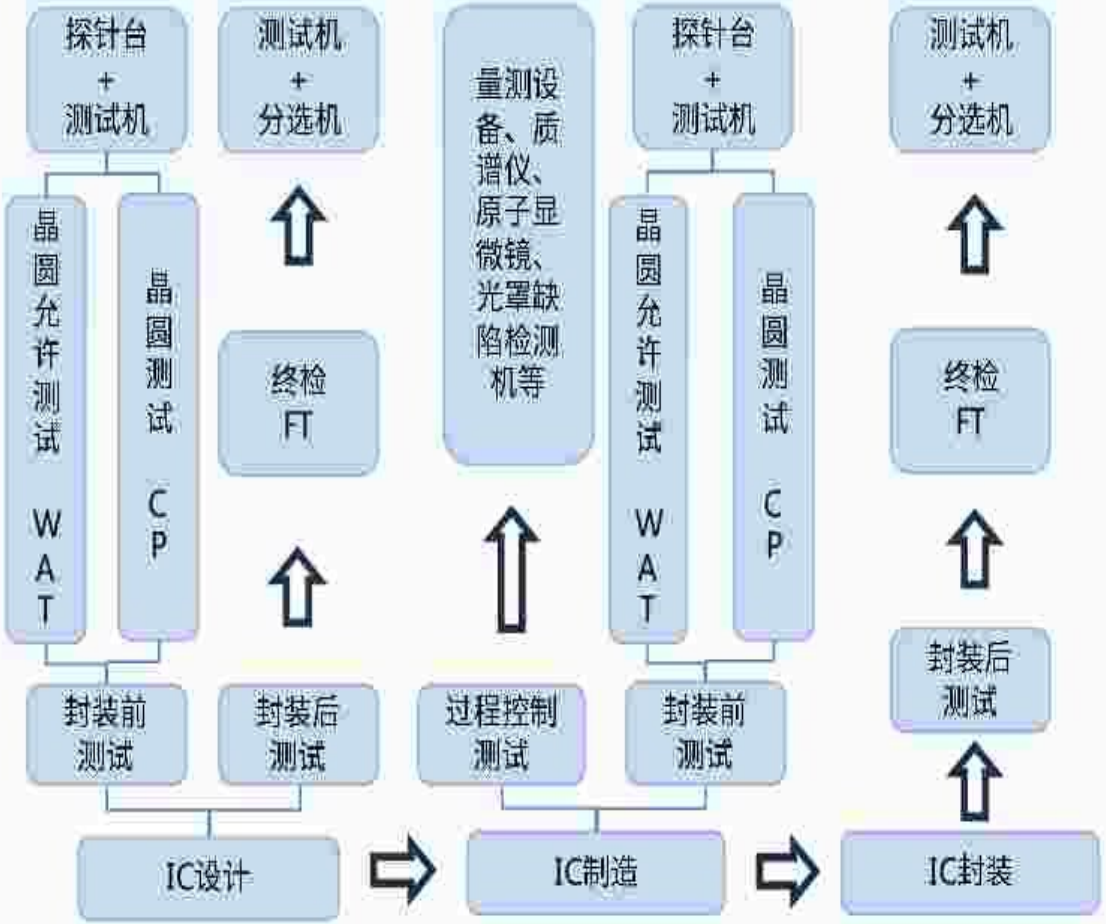
前道晶圆加工、芯片制造工艺流程及设备

资料来源：芯源微招股说明书，国泰君安证券研究

后道先进封装工艺流程及设备



测试工艺流程及设备



资料来源：中微公司招股说明书，国泰君安证券研究

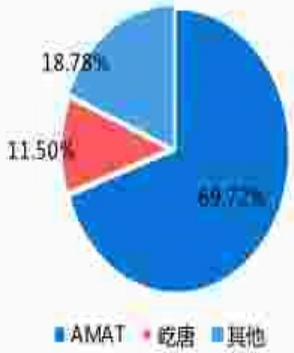
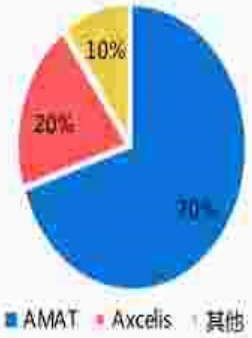
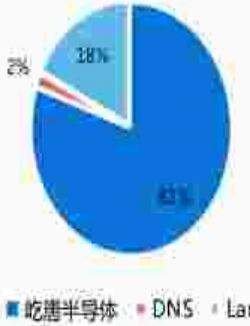
设备端：中国半导体设备和国外仍有差距

	薄膜沉积设备	刻蚀设备	光刻设备
2022年市场规模	212亿美元	135亿美元	180亿美元
市场竞争格局	PVD：AMAT 85%，其他 15% CVD：CVD 30%，AMAT 30%，Lam 19%，TEL 21% ALD：TEL 31%，ASM 29%，其他 40%	Lam 47%，TEL 27%，AMAT 17%，其他 9%	ASML 88%，Nikon 5%，Canon 32%
国内主要公司	北方华创、拓荆科技、中微公司	北方华创、中微公司、屹唐股份	上海微电子
国产化率	5.5%	20%	1.2%

	前道测量/检测	清洗设备	涂胶显影设备	CMP设备
2022年市场规模	118亿美元	47亿美元	24.76亿美元	27.78亿美元
市场竞争格局	KLA 51%，AMAT 28%，Hitachi 12%，其他 9%	DNS 45%，东京电子 27%，拉姆研究 12%，SEMES 5%，盛美上海 3%，北方华创 1%，其他 2%	TEL 89%，其他 11%	AMAT 63.10%，EBARA 29.13%，KC Tech 4.30%，东京精密 2.40%，其他 0.01%
国内主要公司	上海精测、中科飞测	盛美上海、芯源微、北方华创、至纯科技	芯源微	华海清科、烁科精微电子
国产化率	2%	31%	1.1%	18%

资料来源：SEMI，Gartner，拓荆科技招股说明书，中商情报网，芯源微招股说明书，智研咨询，公司公告，国泰君安证券研究

设备端：中国半导体设备和国外仍有差距

	热处理设备	离子注入设备	去胶设备	测试设备
2022年市场规模	20亿美元	16亿美元	5亿美元	124亿美元
市场竞争格局				
国内主要公司	屹唐股份、北方华创	屹唐股份、北方华创	屹唐股份、芯源微、中电科45	华峰测控、长川科技
国产化率	>20%	1.4%	81%	<10%

资料来源：SEMI，Gartner，智研瞻行业研究院，中国国际招标网，QY Research，思瀚研究院，行行查，公司公告，国泰君安证券研究

设备端：攻坚核心技术领域，加深空白环节覆盖度

国内厂商在半导体前道设备领域加速突破

设备类型	国产化率	国内公司	当前进展
热处理设备	20%	屹唐股份	主要可用于90纳米到5纳米逻辑芯片、1y到2x纳米系列DRAM芯片以及32层到128层3D闪存芯片制造中若干关键步骤大规模量产，市占率全球第二。
		北方华创	主要为真空热处理领域产品，包括真空热处理设备、气氛保护热处理设备、连续式热处理设备。真空热处理设备市场拓展顺利，应用范围涵盖真空电子、半导体材料、高端磁性材料等领域。立式氧化炉工艺达28nm水平，成为3D NAND客户的POR机台。
光刻机	1.2%	上海微电子	目前已可量产 90nm 分辨率的 ArF 光刻机，28nm 分辨率的光刻机也有望取得突破。
涂胶显影设备	1.1%	芯源微	在28nm及以上工艺节点的多项关键技术方面取得突破
清洗设备	31%	盛美半导体	18腔300mm UltraCVI单晶圆清洗设备已成功投入量产，支持先进逻辑、DRAM和3D NAND制造所需的大多数半导体清洗工艺，产能较12腔设备提升50%
		北方华创	单片清洗机覆盖 Al/Cu 制程全部工艺，是国内主流厂商后道制程的优选机台；槽式清洗机已覆盖 RCA、Gate、PR strip、磷酸、Recycle 等工艺制并在多家客户端实现量产，屡获重复订单。
		芯源微	前道物理清洗机可用于8/12英寸单晶圆处理，在自动刷压控制技术上已达到国际先进水平。
		至纯科技	12寸单片湿法清洗设备和槽式湿法设备工艺领先。单片湿法设备多工艺已通过验证并交付。将持续开发多反应腔-18腔、超临界清洗等高阶工艺设备。
刻蚀设备	20%	中微公司	CCP设备获客户批量订单，在5nm芯片生产线及下一代更先进的生产线上均实现了多次批量销售，针对逻辑器件的一体化大马士革刻蚀工艺和存储器件的极高深宽比刻蚀技术进行技术攻关，并取得良好进展；ICP设备方面，推出了用于高深宽比结构刻蚀的 Nanova VE 和用于高均匀性刻蚀的 Nanova UE 两种设备，在全面满足 55nm，40nm 和 28 纳米逻辑芯片制造中的 ICP 刻蚀工艺的基础上，拓展了在 DRAM、3D NAND 存储芯片和特色器件等芯片制造中的可刻蚀应用范围。
		北方华创	刻蚀机主要为 ICP，覆盖 8 英寸、12 英寸 55-28nm 制程，已进入中芯国际14nm产线验证阶段。
		屹唐股份	干法刻蚀设备可用于 65nm-5nm 逻辑芯片。
离子注入设备	1.4%	凯世通	低能大束流离子注入机2021年产线验证顺利， 2022年上半年取得在手订单超过11亿元。高能离子注入机设备已顺利通过验证并完成验收。新一代光伏离子注入机进入验证阶段，与某光伏公司签订试用订单。
		中科信	12英寸45-22nm低能大束流离子注入机研发及产业化项目的实施则进入一个全新的自主创新阶段。
去胶设备	81%	屹唐股份	干法去胶设备领域市占率全球第一，正研发应用于3nm及更先进逻辑芯片、先进10nm系列DRAM芯片、176层到256层3D闪存芯片制造的干法去胶设备和工艺。
		芯源微	已经推出了单片式湿法去胶产品，可用于8/12英寸单晶圆处理，先进封装工艺过程中的晶圆去胶制程和金属剥离制程。
		中电科45	研制的双8英寸全线自动化湿法整线设备进入国内主流FAB厂，满足8英寸90nm-130nm工艺节点，适用于8-12英寸BCD芯片工艺中的湿化学制程，可实现全自动湿法去胶。
薄膜沉积设备	5.5%	北方华创	PVD优势明显，制程制造覆盖90nm -14nm，公司在国内产线导入的国产 PVD 设备中占比较高。PVD、ALD、CVD设备新产品市场导入节奏加快。
		拓荆科技	国内唯一产业化应用的集成电路PECVD设备和SACVD设备厂商，ALD设备国内领先。PECVD设备应用于28nm及以上逻辑芯片、3D NAND FLASH、DRAM存储芯片制造等领域。SACVD产品在12英寸40/28nm以及8英寸90nm以上逻辑芯片广泛应用，取得现有及新客户订单。ALD 设备已完成产品开发并取得客户订单，PE-ALD设备在逻辑芯片领域实现产业化应用。
		中微公司	LPCVD的钨填充CVD设备已通过关键客户的工艺验证，能够满足先进逻辑器件接触孔填充应用，及64层、128层和200层以上3D NAND应用。高端逻辑器件和先进存储芯片应用的CVD和ALD设备研发中。EPI研发进入样机调试阶段。
CMP设备	18%	华海清科	国内唯一12英寸CMP设备制造商 ，28-14nm关键节点金属CMP已实现产线量产，14nm逻辑芯片用CMP研发中 ，128层以上制成的 3D NAND与1X/1Y制成的DRAM所用CMP设备实现产线量产。
		北京烁科精微电子	研发制造的8英寸CMP 设备已搬入中芯国际产线。

数据来源：SEMI，前瞻产业研究院，ASML公司公告，国泰君安证券研究

设备端：攻坚核心技术领域，加深空白环节覆盖度

国内厂商在半导体后道封装设备领域加速突破

封装设备	国际主要厂商	中国大陆厂商	当前进展
晶圆减薄机	DISCO、东京精密、冈本工机	中电科	全自动晶圆减薄机产业化机型实现了8-12英寸全自动系列减薄设备国产化替代。
		兰新高科	
		深圳方达	WG1211S自动晶圆减薄机可兼容4/5/6/8/12英寸晶圆，最薄可减薄到100μm以下，可对第三代半导体材料进行高速减薄和研磨,平面度可达1μm,厚度公差1μm
划片机	DISCO、东京精密	中电科	6-12英寸系列产品，全系列拥有手动、半自动及全自动型，适用于IC、LED晶圆、分立器件等晶圆制造行业。
		沈阳仪器	研制的全自动12英寸划片机加工尺寸从8英寸提高到12英寸，加工速度从600mm/s提高到1000mm/s，定位精度从5μm/210mm提高到5μm/310mm，技术指标全面达到国际先进水平
		兰新高科	
		汇盛电子	
		江苏京创	成功率先实现 12 英寸全自动精密划片机产业化的国产替代，系列产品已批量适用于各类半导体材料或泛半导体材料的复杂精密划切
		大族激光	离线式晶圆紫外激光切割系统，配备大族自主知识产权的355nm紫外激光器，该切割设备性能稳定，光斑好，适应长期稳定运行
固晶机	Besi、K&S、ASM Pacific	新益昌	在研项目mini 背光大基板新式固晶机，实现固晶机使用三联体结构，优化固晶工艺,使一台机达到常规三台机的效率
		艾科瑞思	
		东莞普莱信	独家采用刺晶模式的倒装COB固晶工艺以及Pick&Place固晶工艺，最小支持50μm的芯片尺寸，最快每小时产能可以做到180K，精度达到±15微米
		万福达	WFD8970B单机产能实测高达75K/H，RGB三台串联即WFD8916A产能实测200K/H，在业内处于领先水平
引线键合机	K&S、Shinkawa、ASM Pacific	中电科	
		成都宇芯	
		深圳嘉涛	公司焊线机在性能指标上接近或已达到国际先进水平，可提供“固晶机+焊线机”成套半导体封装核心设备。
		北京创世杰	F&S5830:楔焊工艺线径17.6~76μm适用金、铝丝，楔焊工艺角度支持45°,60°，超声具备60~140KHz可选,最大30w；F&S5832: 楔焊工艺线径17.6~76μm 金、铝丝以及30×12.5~250×25μm 金带，楔焊工艺角度支持90°深腔楔焊，超声具备60~140KHz可选,最大30w
		开玖自动化	主流机型K940全自动焊线机在光通讯领域（如2.5G、10G、25G、40G光模块元器件）和激光显示领域应用广泛
		凌波微步半导体科技	IC球焊机率先攻破技术壁垒，已经逐步开始量产，能够满足国内产能大概20-30%
倒装焊机	ASM Pacific、K&S、Shinkawa	中电科	
		大连佳峰	
塑封机	Towa、Besi、Yamada、ASM Pacific	富仕三佳	
		耐科科技	
切筋成型设备	Besi、ASM Pacific	三佳山田	
		耐科科技	
		富仕三佳	

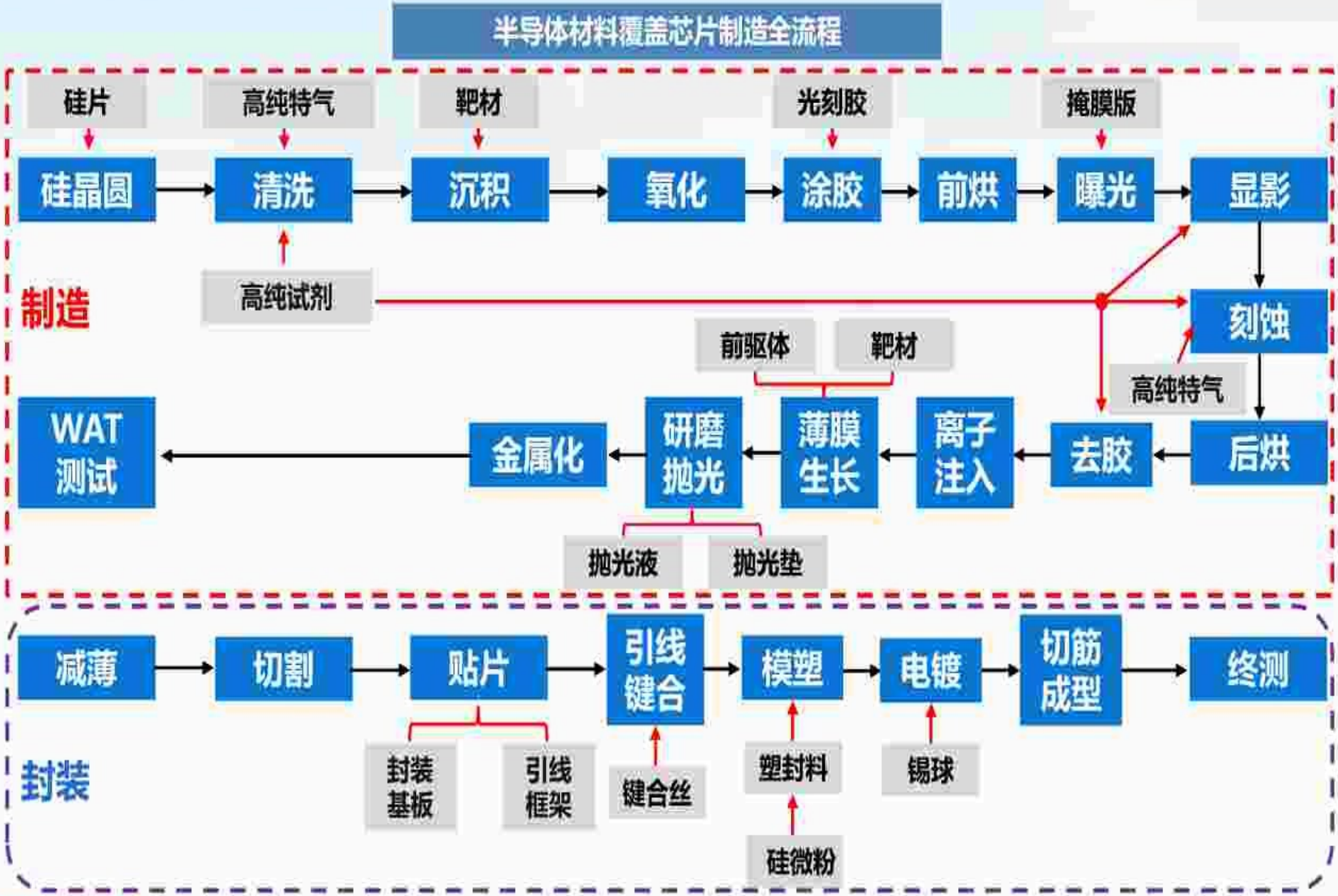
数据来源：SEMI，公司官网，国泰君安证券研究

国内厂商在半导体测试封装设备领域加速突破

测试设备		国际主要厂商	中国大陆厂商	当前进展
分选机		科休、爱德万	长川科技	分选机持续推出新功能，新增三温ATC测试、ART、RTC、2DID识别、5G测试等功能
探针台		东京精密、东京电子、SEMICS	长川科技	国内首台自主研发的CP12探针台，可兼容8/12英寸晶圆，被广泛应用于SoC、逻辑、存储等晶圆测试领域
			深圳矽电	研发出的PT-920 12英寸高精度全自动探针台可满足大规模集成电路对探针台多PIN及多芯的测试要求
			森美协尔	研发出的A12（12/8英寸）量产型全自动晶圆探针台，通过使探针卡与晶圆Pad点之间精准接触，实现完成晶圆WAT/CP测试
测试机	SoC测试机	爱德万、泰瑞达、科休	华峰测控	目前主要100M的8300实现量产，预计第二代400M以上的8300将在年内形成样机
			长川科技	数字测试机D9000，集合1024 个数字通道、200MHz数字测试速率实现快速放量
			华兴源创	新一代T7600 系列SoC 测试机最高速率可支持400MHz Pattern memory 512M、可达到0.5mV的电压精度、完整的混合信号板卡、64通道,每通道1.5A,最高 96A输出
	模拟/混合测试机	爱德万、泰瑞达	华峰测控	传统8200市占率高，向IPM、三代半过渡。
			长川科技	新品8290d获得一致好评，处于快速增长阶段
			佛山联动	QT-8200 系列产品是国内少数能满足 Wafer level CSP(晶圆 级封装)芯片量产测试要求的数模混合信号测试系统之一
	存储器测试机	泰瑞达、爱德万、东京电子、SEMICS	悦芯科技	正在开发验证的存储器测试设备TM 8000填补国产高端集成电路自动化测试设备领域的空白
			武汉精鸿	在BI测试、CP/FT测试已经基本实现小批量产，短期内可实现规模量产

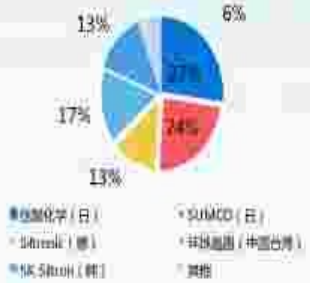
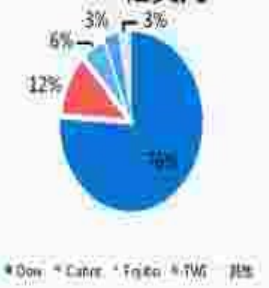
资料来源：SEMI，公司官网，国泰君安证券研究

材料端：半导体材料在制造流程中的应用环节



数据来源：芯源微、鼎龙股份、联瑞新材等公司官网，国泰君安证券研究

材料端：中国半导体材料和国外仍有差距

	硅片	光刻胶	CMP抛光液	CMP抛光垫
2022年市场规模	150.2亿美元	24.24亿美元	20.6亿美元	20.18亿美元
市场竞争格局				
国内主要公司	沪硅产业、立昂微电子	晶瑞电材、南大光电、彤程新材	安集科技	鼎龙股份
22年公司相关业务营收（亿元）	沪硅产业：36.00 立昂微：17.46	晶瑞电材：1.40 彤程新材：1.77	安集科技：9.51	鼎龙股份：5.22

	半导体靶材	电子特气	湿电子化学品
2022年市场规模	18.27亿美元	48.45亿美元	64.34亿美元
市场竞争格局			
国内主要公司	江丰电子、有研新材	华特气体、金宏气体、南大光电	江化微
22年公司相关业务营收（亿元）	江丰电子：19.69 有研新材：9.43	华特气体：13.22 金宏气体：7.44 南大光电：11.95	江化微：9.39

数据来源：Wind，TECHCET，中商情报网，华经情报网，智研咨询，新微超凡，沪硅/立昂微/晶瑞电材等上述公司公告，国泰君安证券研究

国内硅片厂商积极扩张产能，新增硅片产能聚焦12英寸

硅片厂商	子公司	8英寸硅片产能情况	12英寸硅片产能情况
沪硅产业	上海新昇	/	已建产能30万片/月，新增规划产能30万片/月转移
	新傲科技	8寸以下20万片/月，其中8寸SOI技改，达到4万片/月	建设40万片/年的12寸高端硅基材料研发中试线
	Okmetic	20万片/月8寸，2万片/月的SOI产能	
立昂微电子	金瑞泓	已建产能40万片/月	已建产能15万片/月，规划产能45万片/月
超硅半导体	上海超硅	/	/
	重庆超硅	已建产能50万片/月	/
	国晶半导体	/	已建产能15万片/月，二期规划产能30万片/月
	中环股份	已建产能70万片/月，规划产能100万片/月	已建产能17万片/月，规划产能60万片/月
	神工股份	已建产能5万片/月，规划产能15万片/月	/
	麦斯克电子	预计至2024年新增产能20万片/月	预计至2024年新增产能5万片/月
	有研半导体	已建产能276万片/年	扩产项目一期于2020年年底量产，已建成300吨12-18英寸硅单晶的生产能力，二期建设目标为360万片/年
	中欣晶圆	已建产能45万片/月，规划产能55万片/月	现有产能10万片/月，预计2022年年底二期建成新增产能10万片/月
	郑州合晶	已建产能20万片/月	已完成1万片/月试产，规划产能28万片/月
	晶睿电子	已建产能13万片/月，2022年年底规划产能38万片/月	预计2022年年底产能达10万片/月
	鑫晶半导体	新建产能30万片/月	已建成产能10万片/月，预计2022年扩充至30万片/月，2023年扩充至60万片/月

数据来源：公司公告，国泰君安证券研究

各大国产厂商在高端光刻胶领域逐步实现量产和规模出货

产品	细分二级	2020年国产化率	主要用途、技术制程节点	相关公司	量产化阶段	产能及时间节点
PCB光刻胶	干膜光刻胶	几乎全进口	微细图形加工	/		
	湿膜光刻胶及阻焊油墨	50%		容大感光		
				飞凯材料		
LCD光刻胶	彩色光刻胶	5%	制备彩色滤光片（占成本27%）	永太科技	产能建成	
				雅克科技	收购LG彩胶	
				彤程新材（北旭电子45%）		
	黑色光刻胶	5%		上海新阳	产能在建	
	触摸屏光刻胶	/	玻璃基板上沉积TO制作			
	TFT-LCD光刻胶	大部分进口	微细图形加工	晶瑞股份	量产	
				彤程新材（北京科华42.26%）	量产	年产1.1万吨半导体、平板显示用光刻胶及2万吨相关配套试剂项目，因为疫情有所推迟，将在2022下半年完成建设并投入生产。
				容大感光		
				飞凯材料	产能扩建	
				彤程新材（北旭电子45%）	产能扩建	北旭湖北工厂年产6000吨正型光阻项目已于2022年1月份正式投产
半导体光刻胶	g(436nm)/i(265nm)线光刻胶	10%	g:6英寸晶圆（0.5微米以上）；i:6、8英寸晶圆（0.35-0.5以上）	容大感光	产能扩建	
				晶瑞股份	量产	
				彤程新材（北京科华42.26%）	量产	500吨/年
				华懋科技（博康化学29.7%）	量产	
	KrF光刻胶（248nm）	1%	8英寸（0.15-0.25以上）	晶瑞股份	中试完成	
				彤程新材（北京科华42.26%）	量产	10吨/年
				华懋科技（博康化学29.7%）	量产	
	ArF光刻胶（193nm）	1%	12英寸（干法：65-130nm;湿法：45nm以下）	晶瑞股份	研发	
				彤程新材（北京科华42.26%）	研发	
				上海新阳	产能建设	ArF浸没21年少量销售，22年量产；干法22年少量销售，23年量产。23年预计合计销售额2亿
				南大光电	02建设	
				华懋科技（博康化学29.7%）	规模化量产	
	EUV光刻胶（13.5nm）	研发阶段	12英寸（32nm以下）	彤程新材（北京科华42.26%）	02专项研发通过验收	

数据来源：萝卜投研，前瞻产业研究院，国泰君安证券研究

材料端：CMP——鼎龙与安集合力，打破CMP材料垄断

安集科技抛光液部分产品技术已经接近国际最高水平

产品种类	技术节点	预计总投资额（万元）	进展
铜抛光液	28nm-14nm	15000	相关产品已在28nm产线和14nm产线实现量产，并持续在更多产品上验证
	10nm以下		10nm-7nm技术节点的产品平台研发完成，并在相关客户端测试优化
阻挡层抛光液	28-14nm	7000	相关产品已在28nm产线和14nm产线实现量产，并持续优化并验证
	高去除速率		性能满足要求，正在多家客户端验证
	10nm以下		10nm-7nm技术节点的产品平台研发完成，并在相关客户端测试优化
钨抛光液	高选择比	8000	相关产品在3D NAND和DRAM全面量产，在Logic上测试验证
	中低选择比		相关产品在3D NAND和DRAM全面量产，在Logic上测试验证
钴抛光液		1500	产品性能基本达到要求，在客户端持续测试验证中

鼎龙股份抛光材料产品进展顺利

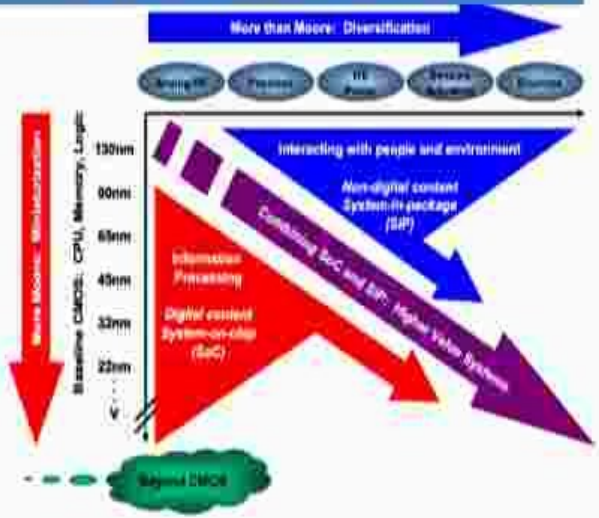
产品种类	技术节点	产品验证	产能
抛光垫	28nm及以上	已通过28nm产品全制程（ILD/STI/W/Cu/HKMG）的验证并获得订单	一期和二期将合计产能30万片/年，潜江三期建设已经投产，设计产能50万片/年
	14nm及以下	DH5XXX系列新产品在客户端验证进展顺利，将在2021年底获得新突破	
抛光液		公司进行Oxide，SiN，Poly，Cu，Al等CMP制程抛光液产品多线布局，目前在客户端的验证反馈情况良好，部分产品已通过各项技术指标测试，其中Oxide制程某抛光液产品已取得小量订单，Al制程某抛光液产品在28nm技术节点HKMG工艺中通过客户验证，进入吨级采购阶段。	公司已初步建成武汉本部一期年产2000吨清洗液产线，并计划在产业园启动年产10000吨产线建设

数据来源：公司公告，国泰君安证券研究

05 Chiplet：延续摩尔定律，规模化落地可期

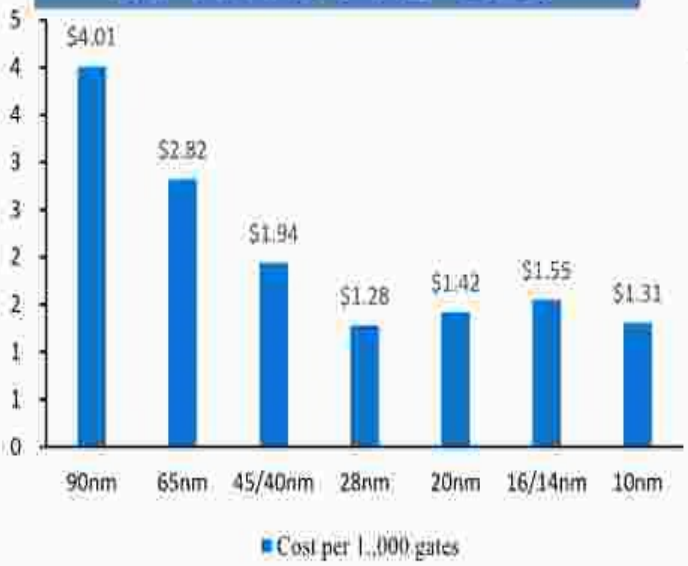
- **摩尔定律实现的维度主要分为制造、设计、封装三方面。**在制造方面，主要通过晶体管微缩工艺实现，从130nm逐步向5nm甚至是2nm迈进；在设计方面，主要通过各种架构演进、方案设计等方式实现；在封装方面，主要通过不同模块的异质集成来实现，通过SiP、WLP等方法不断提高系统化的集成密度。
- **摩尔定律在制造端的提升已经逼近极限，开始逐步将重心转向封装端和设计端。**随着AI、数字经济等应用场景的爆发，对算力的需求更加旺盛，芯片的性能要求也在不断提高，业界芯片的制造工艺从28nm向7nm以下发展，TSMC甚至已经有了2nm芯片的风险量产规划。但随着线宽逐步逼近原子级别，工艺制程升级带来的性能、功耗提升的性价比越来越低，封装端和设计端维度的提升开始逐步进入视野。

摩尔定律的不同的实现方式



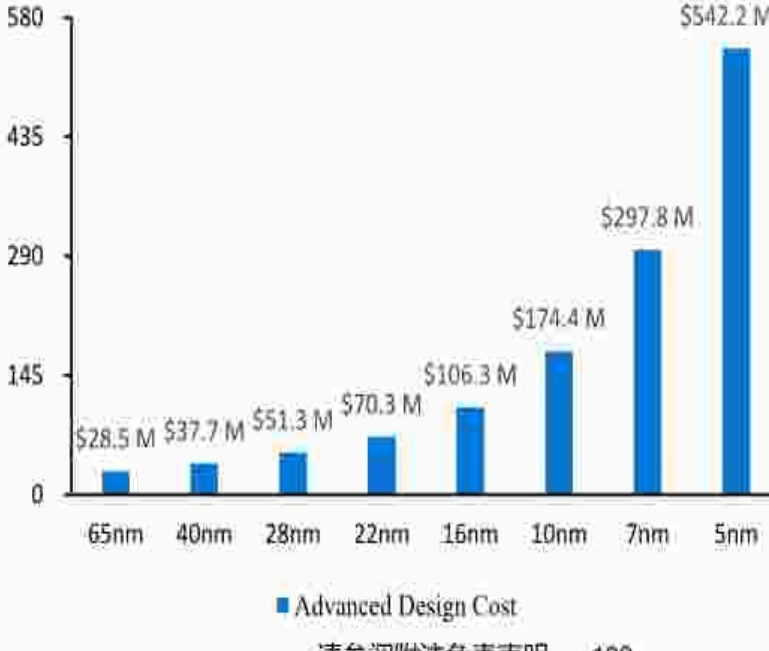
数据来源：ITRS

随着工艺制程的进步，单个晶体管的成本不再下降（单位：美元）



数据来源：芯启源，《Measuring Moore’s Law》，国泰君安证券研究

先进工艺的流片成本越来越高（单位：美元）



数据来源：芯启源，国泰君安证券研究

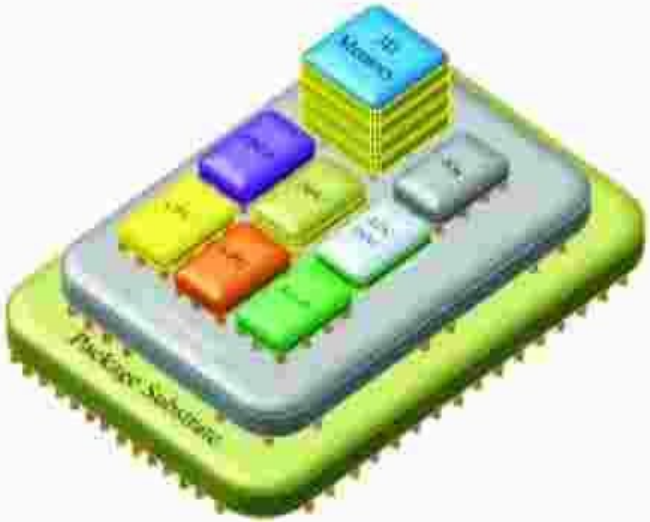
- **Chiplet方案正是一种通过在封装端和设计端的提升，来进一步提升芯片的集成化密度，从而延续摩尔定律的新型半导体技术方案。**其方案核心主要包含三个概念，分别是小芯粒、异构异质和系统级集成。

1）小芯粒：原有SoC芯片由各种IP内核设计组成，小芯粒即在设计端将各种IP单个拆分，进行芯片化。

2）异构异质：将类似CPU、GPU、DRAM等不同结构\工艺\材质的芯片合在一起，从而减少传输延迟、提高集成度。

3）系统级集成：在前两者的基础上，通过软件设计系统级高密度的方案，利用各种堆叠封装技术，将更多的异构异质的小芯片进行高密度封装集成，从而实现良率、成本、性能、商业风险等方面的综合提升。

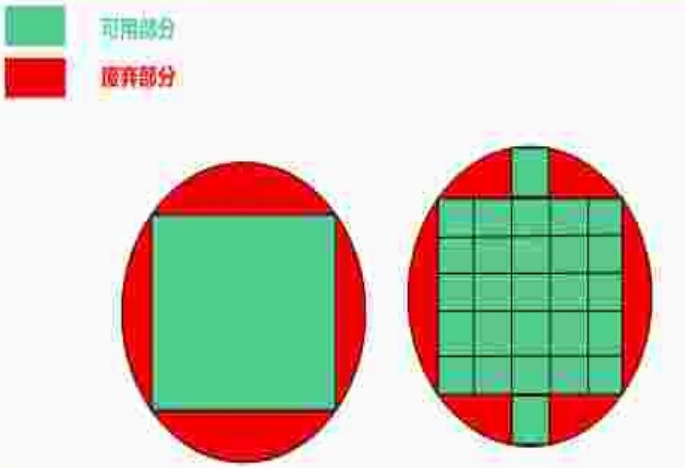
Chiplet方案概念图



数据来源：电子技术设计

- **Chiplet将芯片分解成特定的模块，这可以使单个芯片面积更小并可选择最合适的工艺，从而提高良率、降低制造成本和门槛。**
- **在降低成本方面**：当切割芯片的面积越小，如下左图所示，绿色芯片的数量就越多，整体晶圆中可用的芯片面积就越大，单位面积芯片的成本就越低。另外，硅片化IP的复用，也可以显著降低成本。
- **在提高良率方面**：晶圆中存在各种缺陷，当芯片的面积越大，它受影响的芯片数量比例就越大。例如，如下右图所示，一块晶圆中切割3片芯片，有一片受到缺陷影响，良率为2/3；当一块晶圆切割25片芯片，缺陷影响了3片芯片，良率为22/25，整体良率大于2/3。
- **在降低门槛方面**：小芯片化后，不同的芯片可以采用最合适的工艺和架构进行设计制造。例如I/O die因为更加先进的工艺对其性能的提升有限，可以采用12nm工艺进行设计制造，CPU die因为对先进工艺要求更高，可以采用7nm/5nm工艺进行设计制造。整体无需像SoC一样，I/O和CPU的IP都必须采用最先进的工艺设计制造。

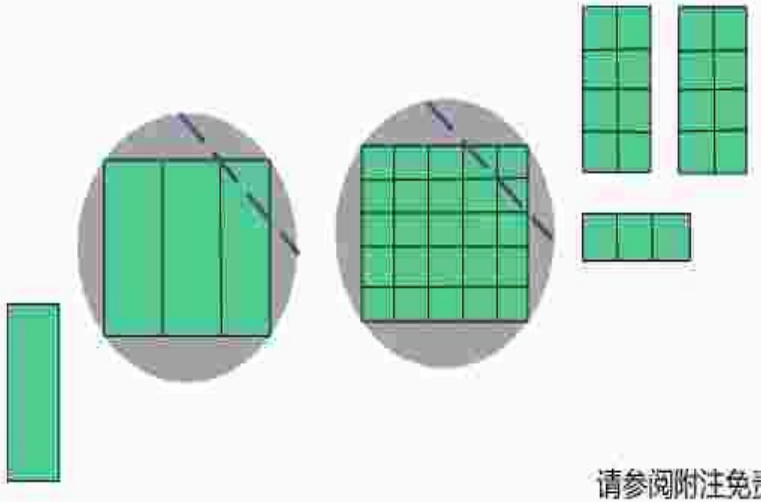
小芯片在单晶圆中的可用面积更大



由图可见，切割面积越小，绿色部分方块越多，可用的圆的面积越多

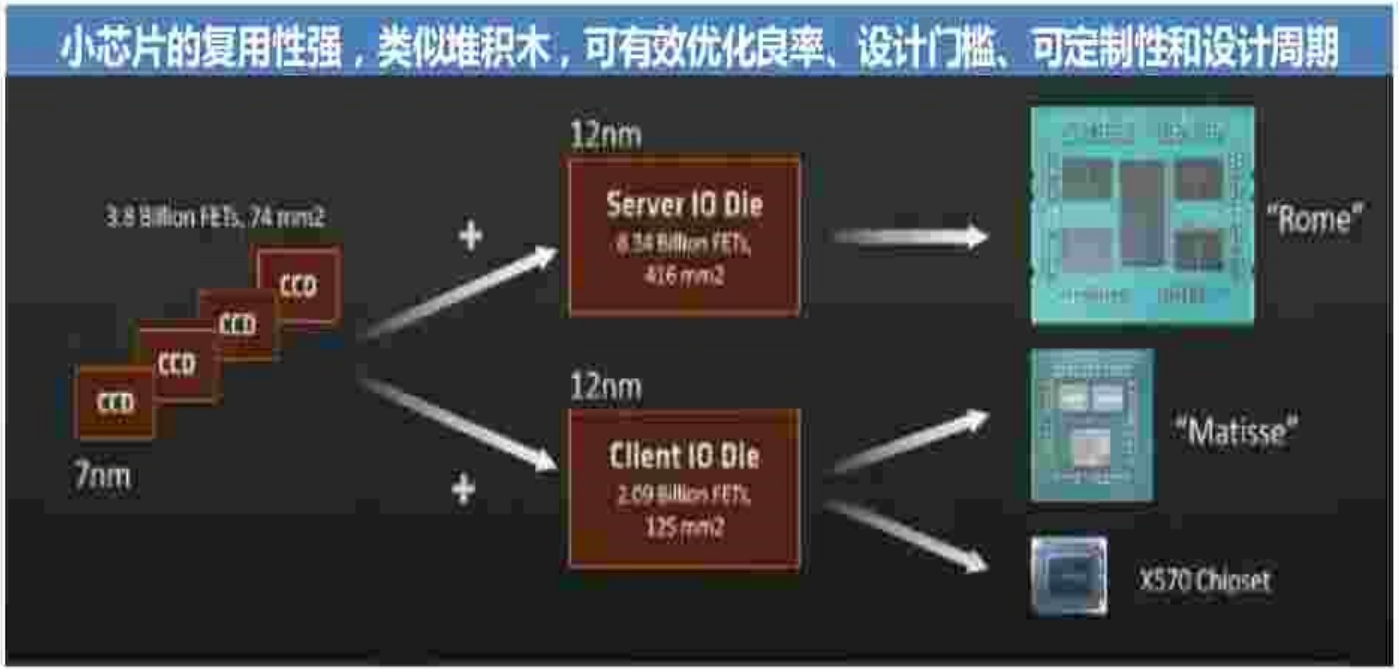
数据来源：国泰君安证券研究

小芯片在单晶圆中的良率更高



数据来源：国泰君安证券研究 备注：紫线为缺陷

- **以AMD的系列产品为例，将处理器芯片进行解耦合**，分成单个CCD（Core Chiplet Die）芯片和一个I/O die，CCD和I/O核之间采用第二代Infinity Fabric总线连接。其中CCD采用7nm工艺，I/O核采用12nm工艺。8个CCD和1个Server I/O die可组装成EPYC Rome（霄龙）服务器处理器；8个CCD和1个Client I/O die 可组装成Ryzen（锐龙）3000系列（代号Matisse）桌面服务器；AMD的X570 Chipset也可用现有的小芯片进行组装设计。
- **这种固定模块的小芯片方式，多个小芯片无需重复设计，具有复用价值，而且芯片可采用最合适的工艺制程，可有效提高良率以及降低设计门槛。在可定制性、设计周期方面、降低成本，进行极大优化。**



数据来源：Hotchips，AMD官网

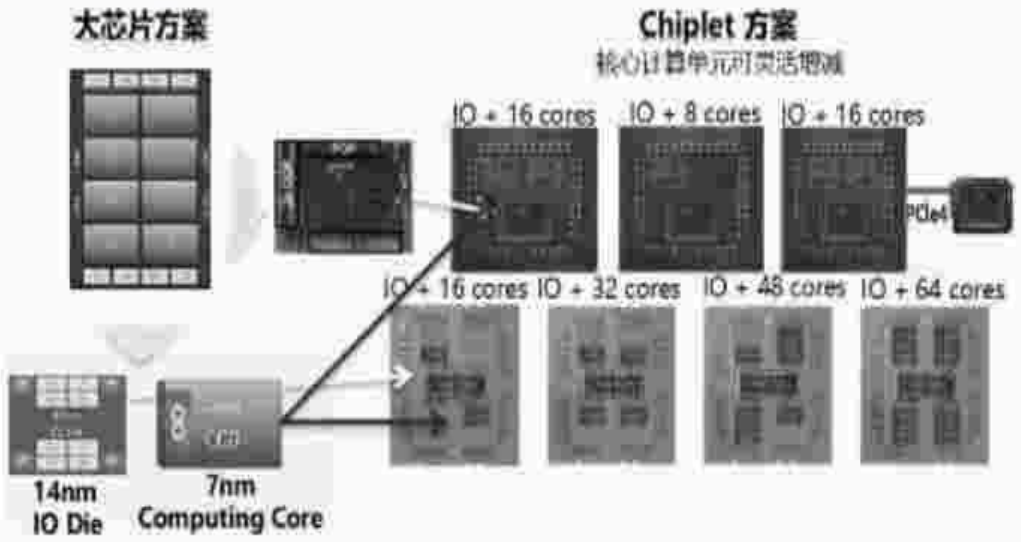
- **小芯片可高度集成化：小芯片利用芯片互连技术和高密度封装技术可轻易集成多核，满足高效能运算处理器的需求。**单片SoC的方案，在集成多核方案时，受制于可用的光罩尺寸、良率等问题，芯片面积最多只能达到800mm2。Chiplet核心计算单元可从16核堆积到64核，甚至96核以上。另外，对于内存和Cache方面，也能实现高密度集成，从而实现更低的延迟或者更高的并行运算速度。

Chiplet方案相较于大芯片方案，具有多方面的优势

类别	SoC	Chiplet技术	分立器件
设计费用	最高	较低	最低
设计费用	最长，一般超过18个月	较短，大概12个月	最短，大概6个月
设计风险	最高	较低	最低
性能	最高	较高	低
功耗	最低	较低，接近SoC	最高
可定制性	困难	容易	非常容易
上市时间	最慢	较快	最快
面积大小	最小	较小	最大

数据来源：《Chiplet接口IP 3DIC混合信号仿真验证》，国泰君安证券研究

Chiplet方案可轻易集成多核，满足高性能计算的需求



数据来源：《Chiplet接口IP 3DIC混合信号仿真验证》

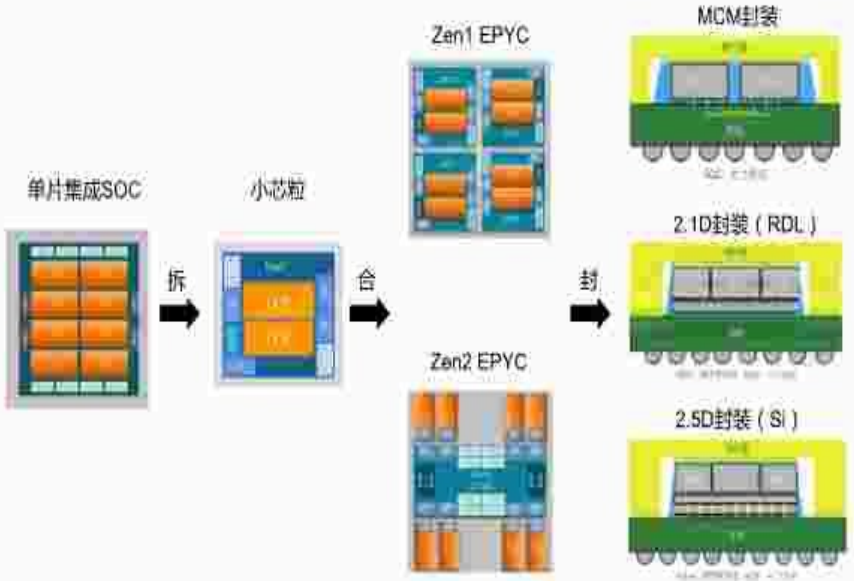
- **Chiplet方案主要由三大环节组成，分别是拆、合、封。**

1）在“拆”的环节：将原有多个IP组成的SoC大芯片进行拆分，形成多个不同的CPU、I/O等小芯片。拆解后的小芯片可以采用更加适配的工艺节点和材质。其中架构设计是关键，需要考虑访问频率、缓存一致性等各问题。

2）在“合”的环节：将不同的小芯片利用内部总线互连技术进行电路连接，各个电路互相组合，在功耗、通信延迟、带宽等方面达到最优的效果。与SoC不同的是，前者是芯片间的互连，而后者是IP内核间的互连。

3）在“封”的环节：将组合后的不同的芯片，利用RDL、TSV、硅转接板、晶圆等高密度集成的先进封装技术，进行组合。

Chiplet方案主要由拆、合、封组成



数据来源：电子技术设计，chiplet关键技术与挑战，AMD，国泰君安证券研究

- **Chiplet方案的实现包括Chiplet的设计制造和连接侧的互连制造。**依据主要的产业链制造顺序而言：

  1）在设计端：利用EDA和IP核进行分割后的Chiplet的设计、连接侧包括硅转接板或者RDL层的互连建模，之后两者协同仿真，得到完整的封装方案的模型。针对该模型依次进行时序分析、电源网络分析、可靠性分析以及PPA优化分析等，从而实现Chiplet和连接侧结合的系统性方案。

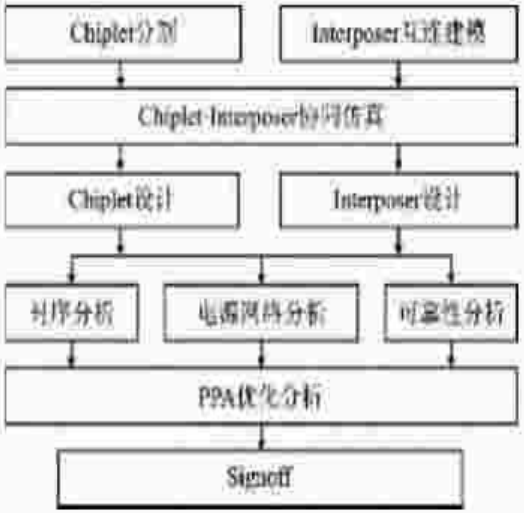
  2）在封装端：利用晶圆厂制造完成的Chiplet与连接侧方案进行连接，以2.5D的硅转接板为例，将Chiplet和进行TSV打孔的硅转接板相连，利用硅转接板内部的RDL层进行各个Chiplet之间的互连，最后将硅转接板与基板进行连接，即完成整体Chiplet系统性方案的制造。

- **上述在设计端和封装端的步骤，刚好对应拆、合、封三大环节。**

产业链上下游结构

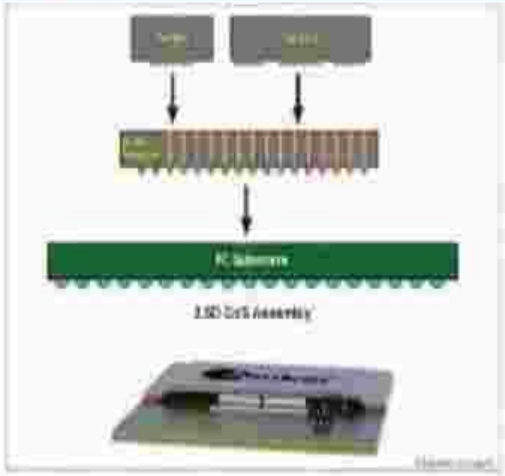
数据来源：国泰君安证券研究

Chiplet在芯片设计端的流程示意图



数据来源：Chiplet方案研究与展望

Chiplet在芯片设计端的流程示意图

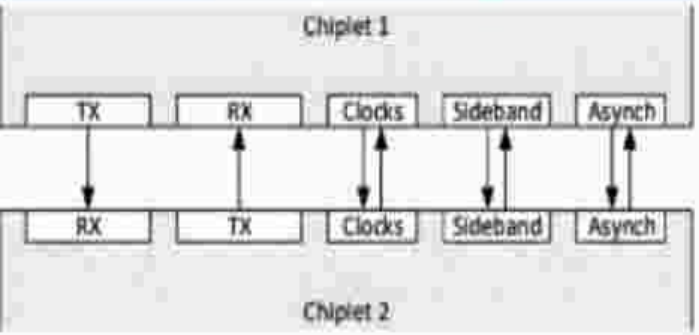


数据来源：Amkor Technology

Chiplet的统一接口和标准

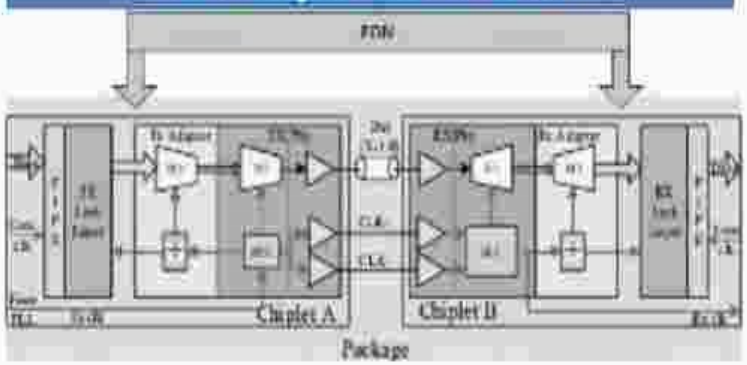
- **互连接口与协议的落地和推行是实现技术标准化和产品规模化的关键。**2022年3月，Intel、AMD、ARM、台积电、日月光等巨头成立Chiplet标准联盟，制定了通用Chiplet的高速互联标准UCIe（Universal Chiplet Interconnect Express）。2021年5月，CCITA（中国计算机互连技术联盟）针对Chiplet标准《小芯片接口总线技术要求》展开标准制定工作，集结了国内产业链60多家单位共同参与研究。
- **Chiplet总线互连接口与协议可以划分为物理层（PHY层）、数据链路层、网络层以及传输层。**数据链路层及以上的其他接口更多依赖沿用或扩展已有接口标准及协议。最重要的是物理层的接口研究，因为它与工艺、功耗和性能等息息相关。物理层主要分为串行和并行两种数据通信技术，串行主要分为串行器和解串器SerDes，并行则包括低电压封装互连LIPINCON技术（TSMC提出）、AIB高级接口总线（Intel提出）以及信号引线物理互连BoW技术（OCP提出）等。

物理层接口示意图



数据来源：《Chiplet 接口 IP 3DIC混合信号仿真验证》

两个Chiplet Die互连场景



数据来源：《Chiplet 接口 IP 3DIC混合信号仿真验证》

物理层并行互连的技术对比

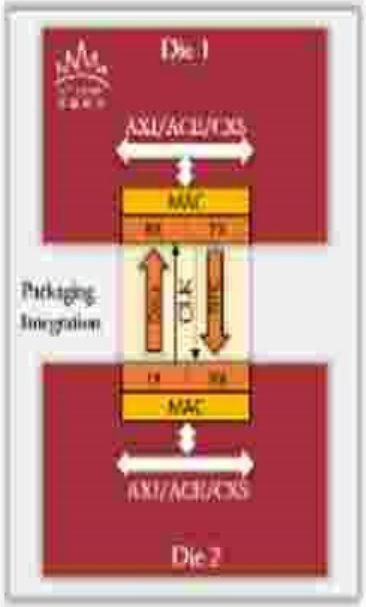
参数	AIB(第一代)	MDIO(第一代)	LIPINCON2	BoW
单lane数据率/Gb·s¹	2	5.4	8	16
shoreline带宽密度/(Gb·s¹/mm)	63	200	67	200
Areal带宽密度(Gb·s¹/mm²)	150	198	198	148
单位功耗/(pJ/bit)	0.85	0.5	0.56	0.5
封装技术	EMIB	EMIB,Foveros	CoWoS	MCP

数据来源：《异构集成芯片关键技术研究》，国泰君安证券研究

Chiplet的统一接口和标准

- **互连是技术标准化的重点之一，但芯片间互连协议的标准化方面仍处于发展演进阶段，相互竞争的标准较多。**包括CXL、CCIX、NVLink等标准，都已经在复杂的处理器芯片中得到应用。其中虽然CXL发布较晚，但因为Intel的业内影响力和产品效应，大多数厂商纷纷跟随并采纳，技术发展较快。国内以CCITA为主导的技术联盟正在进行相关技术和标准的研发中。相关国内公司例如超摩科技也已经宣布量产Chiplet互联IP整体解决方案CLCI，其协议标准主要采用自有方案，未来会考虑协议间的兼容性。

超摩科技宣布量产高性能Chiplet互联IP整体解决方案CLCI



EDA工具链和生态系统的完整性、可持续性

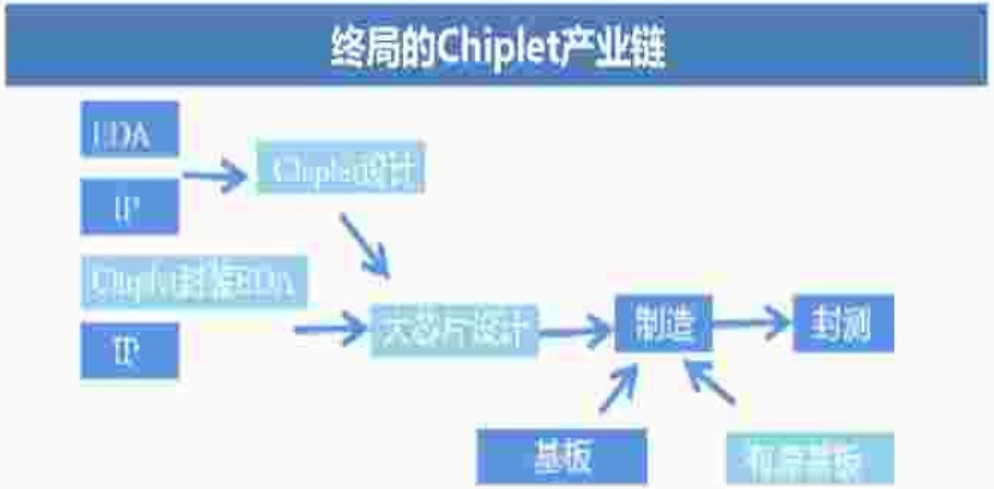
- **小芯片之间更密集的互连+Chiplet封装EDA的更高要求**

Chiplet方案将芯片进行精细化切割，并进行更为密集的互连，例如HBM的芯片间的互连位宽为1028bit，从而使其整体性能达到接近甚至超过SoC内部的传输效率。对于Chiplet的封装，也需要进行额外的EDA设计，这些都对EDA工具提出了更高的要求。

- **系统性方案带来的更严苛的可靠性挑战**

Chiplet方案作为一个整体的系统性方案，对热效应、电磁挑战、电容耦合、电感耦合、信号完整性等方面都提出了全新的要求，需要进行针对性的仿真建模，这是原有主要针对SoC芯片的EDA工具相对薄弱的点。当第三方Chiplet开始被采用时，对于完整系统的可靠性要求将会更高。第一种挑战可能可以采用Cadence等工具组合设计，但针对于第二种可靠性调整，则需要进行针对性优化升级。

- **考虑到无论是EDA工具链还是之前的协议标准抑或是制造封装技术都处于发展初期，为了实现有效的正反馈优化，将终端的测试纠错信息及时反馈到上游的EDA、设计端并进行改进，构建一个完整的、可持续的生态系统是极其重要的。**



数据来源：电子技术设计，《chiplet关键技术与挑战》，国泰君安证券研究

核心封装技术的选择

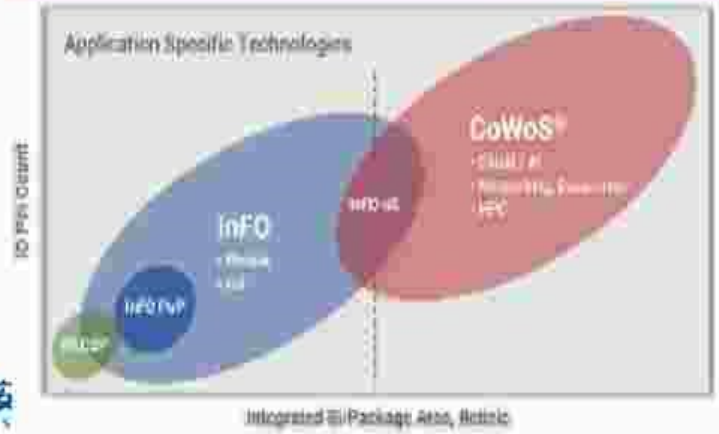
- **Chiplet方案对应的封装技术包括2D的MCM、2.1D的RDL方案、2.5D的CoWoS和3D的HBM等多种技术，需要根据功耗、性能、成本等多方面进行综合考虑。（基于PPA的芯片评价体系+实现系统效率最大化）**

1）2D的MCM/WLCSP技术属于典型的封装技术，将多个不同的芯片在基板上进行集成，属于成本低复杂度低，但能有效增加管脚数量，提高芯片集成密度的方案，在AMD、国内诸如超摩科技等多种产品中使用，是当前较为主流的方案。

2）InFO技术属于2.1D方案，介于MCM和2.5D的CoWoS之间，利用RDL层进行集成，线间距接近2微米，引脚数量约2500个，多用于手机和IoT中，苹果最新的M1等芯片就是采用该方案。

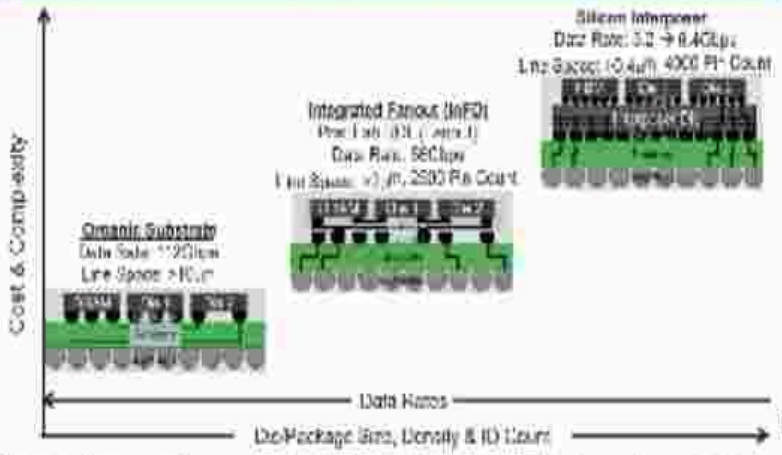
3）2.5D和3D技术可以在前两者的基础上，利用硅转接板等就技术极强地增大管脚数量和集成密度，例如2.5D的方案相较于InFO方案，线间距减小到0.4微米，引脚数量增加到4000个，是InFO方案的1.6倍，但由于成本过高，多用于云计算、HPC、数据中心中。

TSMC的多种多芯片封装集成方案



数据来源：TSMC

不同的成本和性能要求对应不同封装方案



数据来源：《The new frontier of die-to-die interface IP》

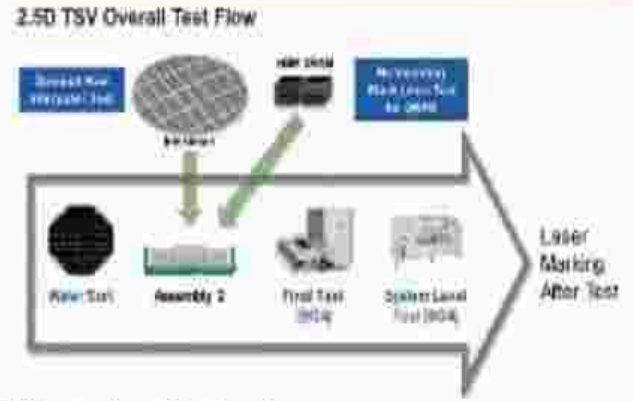
产品测试的复杂性

- **Chiplet方案由于互连封装方案的不同，其测试大多为定制化方案，且包含更多的测试流程。**除了常规的单片集成SoC芯片所需的CP测试（芯片针测）、FT测试（终测），还要包括介质层测试、MT（中段测试）、SLT（系统级测试等）。
- **测试流程中，KGSD（已知良好堆叠芯片）测试需要包含更多的可靠性测试，是主要的难点之一。**以DRAM和HBM为例进行对比：

1）在晶圆级测试环节，DRAM晶圆的测试基本相同，HBM额外增加针对逻辑晶圆的逻辑测试，包括测试IP、PHY电路中缺陷等。但是考虑到单颗小芯片的缺陷就会导致堆叠的KGSD芯片的性能失败，因此对单颗小芯片的测试性能要求会更高。

2）在KGSD测试环节，传统的DRAM封装级产品测试设备和解决方法将无法有效试用，其测试的挑战包括动态向量老化应力测试、大量内部TSV结构的可靠性测试、高速性能测试、2.5D SIP测试等。

先进封装的整体测试流程



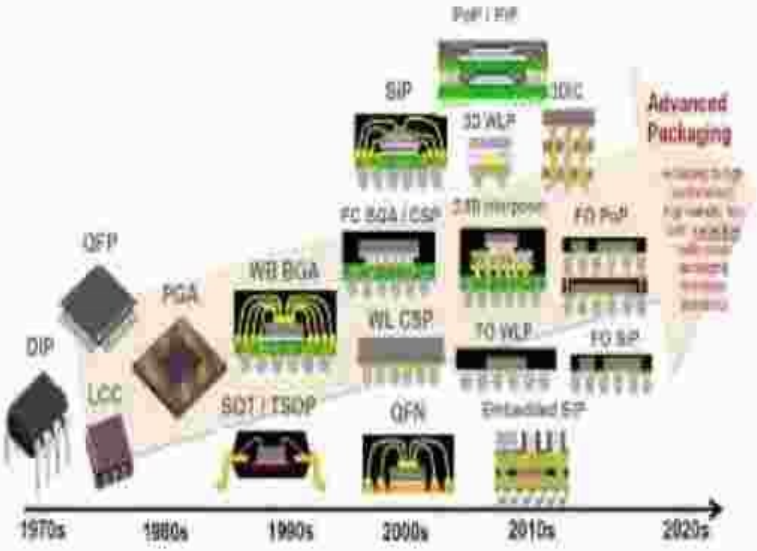
数据来源：Amkor Technology

HBM测试比DRAM测试要求更加高

数据来源：《高带宽存储器的技术演进和测试挑战》

- **先进封装技术不同于传统封装技术，其主要包含RDL、Bump、Wafer和TSV四个要素。**传统封装主要包括DIP、QFP等引脚封装和引线框架封装，而诸如FC-BGA、FO WLP和FI WLP等包含RDL、Bump、Wafer和TSV四个要素之一，均属于先进封装。
- **Chiplet封装方案是小芯粒的异构异质高密度集成方案，对应不同的封装类别，以先进封装技术为基础，可主要分为2D、2.1D、2.5D和3D四大类。**考虑到市场上各家公司对于封装方案的定义并不明确，本文粗浅根据在基板基础上是否有RDL层和硅桥、是否有无源硅转接板、是否有有源硅板之间的堆叠，进行分类，依次划分为2D、2.1D、2.5D和3D四大类，其中2D方案由于不使用任何额外高密度RDL/硅等转接板，性价比高，在Chiplet的发展初期，产品中应用广泛。

从传统封装向先进封装发展



国泰君安证券 GUOTAI JUNAN SECURITIES

数据来源：Yole

Chiplet的高密度封装技术主要分为2D、2.1D、2.5D和3D四大类，均有相关产品应用量产

类型	技术特点	焊点间距	说明	技术	代表产品应用
2D	MCM，厚度很薄	90微米	①直接通过封装基板走线实现互连；②无需基板，直接通过RDL层进行互连	InFO（普通），FC-MCM(ABF基板良率低，无法支撑多芯片应用）	AMD的Zen架构产品
2.1D	RDL转接板/硅桥（在基板基础上）	30-45微米	在基板基础上利用高密度的RDL层/内嵌硅桥的方式实现互连	EMIB，InFO-SoW，InFO-R/InFO-oS，InFO-LSI，FOCoS(B,CF,CL)，XDFOI，cow os-R，cow os-L，	苹果的M1 Ultra，Intel的CPU
2.5D	无源硅转接板	25微米	在基板基础上利用硅转接板实现互连，可实现更高密度的互连（成本高）	Cow os-S，I-Cube，VISionS	海思的鲲鹏920和昇腾910，AMD的Zen2/3/4架构产品
3D	有源硅之间的堆叠	10微米	多在2.5D基础上，利用混合键合实现芯片之间的垂直堆叠	Co-EMIB，Foveros，X-Cube，WIDE-IO，SoIC，HBM，HMC，3D V-Cache	英伟达的GPU，AMD的Zen2/3/4架构产品

数据来源：电子技术设计，《chiplet关键技术与挑战》，AMD，国泰君安证券研究

客户和产品应用

2D 方案

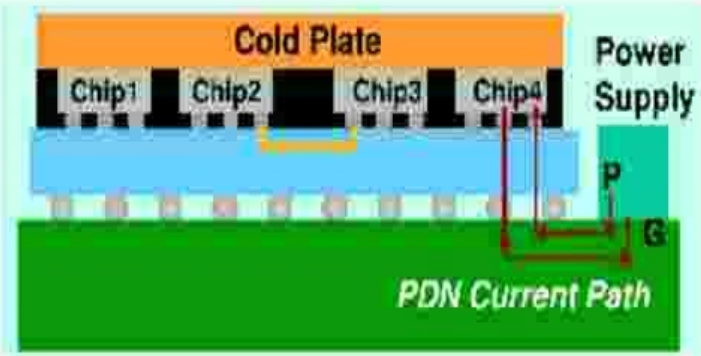
2.1D方案

2.5D方案

3D 方案

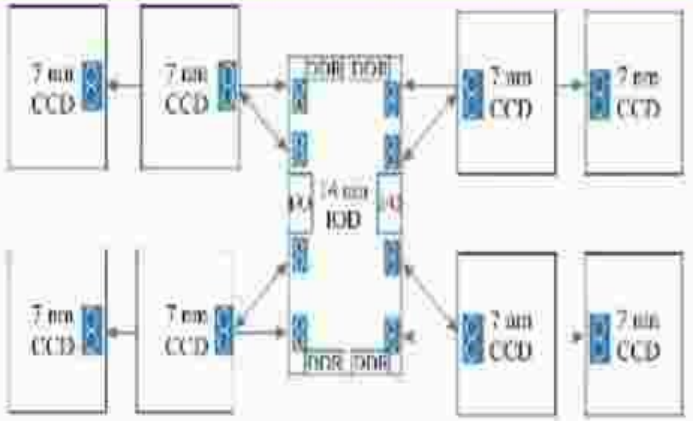
- **2D方案主要为简单的MCM方案，无需额外的转接板，成本低，性价比高，应用较为广泛，但无法支撑多芯片大面积应用，在性能提升上空间有限。**2D方案整体厚度较薄，主要分为FC-MCM类的直接通过封装基板走线实现互连和普通InFO类的无需基板直接通过RDL层进行互连。FC-MCM类受限于ABF基板良率低，无法支撑多芯片大面积的应用。普通InFO类由于没有基板，仅凭PI材料的RDL层，硬度不够，同样无法支撑大面积的多芯片集成。
- **2D方案受益于性价比，国内外客户多家产品有量产，在四种类别中应用最广，发展最快。**AMD的最初Zen架构的系列产品采用的就是MCM方案，如锐龙、霄龙等。另外，国内包括超摩科技（高性能CPU）、龙芯中科等都有相关方案研究。

Flip-Chip MCM方案概念图



数据来源：TSMC

AMD的第二代EPYC采用了MCM方案



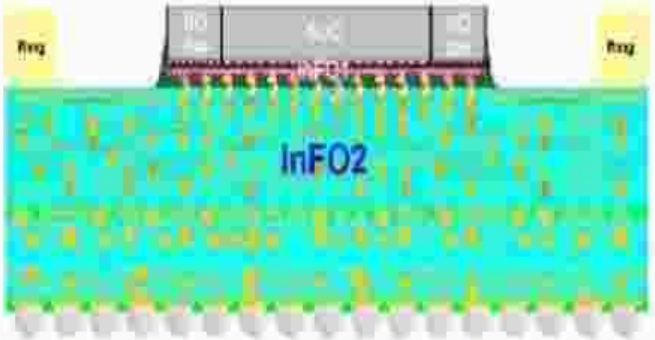
数据来源：AMD，《Chiplet封装结构与通信结构综述》

客户和产品应用

- 2.1D方案
- 2.5D方案
- 3D 方案

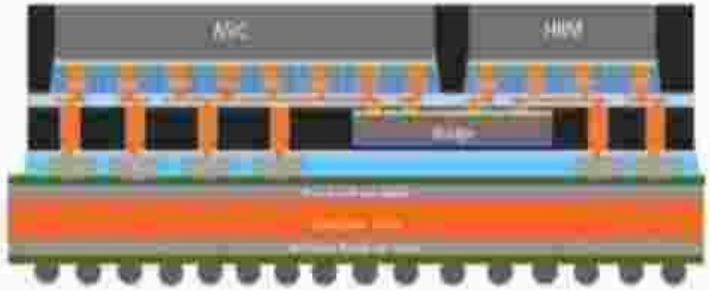
- **2.1D方案介于2D的MCM和2.5D硅转接板之间，成本相对适中，可集成度较高，可适用于大规模多芯片集成。**2.1D方案主要在基板上采用高密度的RDL层或者在RDL层/基板中内嵌硅桥来增大集成密度。高密度的RDL层方案包括特斯拉的InFO-SoW（六层RDL）、TSMC的InFO-R/InFO-oS/InFO-LSI系列、长电的XDFOI（五层RDL）等。内嵌硅桥的方案以Intel的EMIB、日月光的FOCoS-B为主。
- **2.1D方案的主要缺点在于技术难度相对较大，目前只在少数客户中使用。**例如高密度RDL层的InFO-R中，本身InFO工艺就较为复杂，还需要在PI树脂中进行多层RDL高密度布线，难度更加巨大。目前主要在苹果的M1 MAX芯片中使用该方案较多。例如内嵌硅桥的EMIB和FOCoS方案中，需要额外考虑硅桥和RDL层/基板的兼容性，目前主要在Intel的产品中使用较多。

Tesla的InFO-SoW方案布置6层RDL



数据来源：TESLA官网，TSMC

FOCoS-B在扇出RDL层中内嵌硅桥



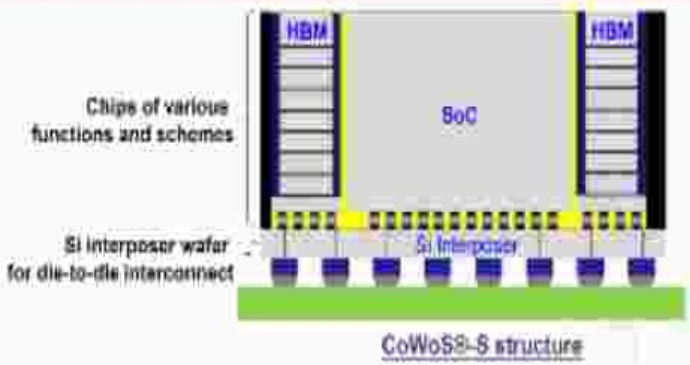
数据来源：日月光官网

客户和产品应用

2.5D方案

- **2.5D方案利用无源硅转接板方案，可实现更高密度、大面积多芯片的集成方案，传输速度高，性能优越，是潜在延续摩尔定律成长空间的主要方案。**无源硅转接板利用内部RDL和TSV可实现内部的高密度互连，加上硅技术较为成熟，成为替代先进工艺延续摩尔定律的中坚力量。另外3D方案的拓展也主要建立在2.5D方案的基础上。主要方案包括台积电的CoWoS系列方案，三星的I-Cube，通富的VISionS等。
- **2.5D方案整体性能更为优越，但由于增加硅转接板，成本较高，主要用在服务器、**数据中心等高端应用中，发展前景巨大。鲲鹏920、AMD的Zen2以上架构产品诸如Rome、Milan等服务器芯片都应用CoWoS方案。以AMD的Zen4架构的EPYC 7004服务器芯片为例，其内部可封装的CCD数量增加到12个，内核增加到96个，可支持12通道的DDR5内存，提供128条PCIE 5.0通道，性能十分突出。

TSMC的2.5D方案利用硅转接板集成芯片



数据来源：TSMC

AMD的Zen4架构EPYC 7004服务器处理器芯片架构

数据来源：AMD

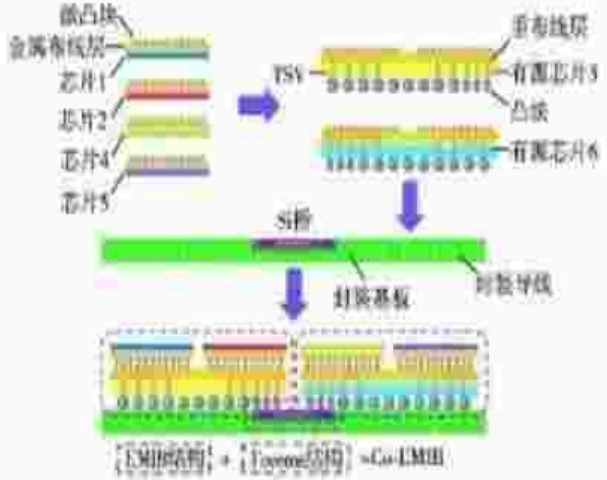
客户和产品应用

- 2.1D 方案
- 2.3D方案
- 2.5D方案
- **3D 方案**

• **3D方案主要在2.5D基础上，利用混合键合等方式实现芯片间的垂直互连，集成密度最大，性能提升也十分可观，但成本非常高。**3D方案为有源硅之间的互连，即芯片之间的互连，为满足足够的信息带宽，使用的互连线的数量和密度都远大于前三种，而且混合键合的难度也远大于bump键合，整体成本非常高。主要方案包括Intel的Co-EMIB/Foveros、三星的X-Cube、TSMC的SoIC、HBM、3D V-Cache等技术。

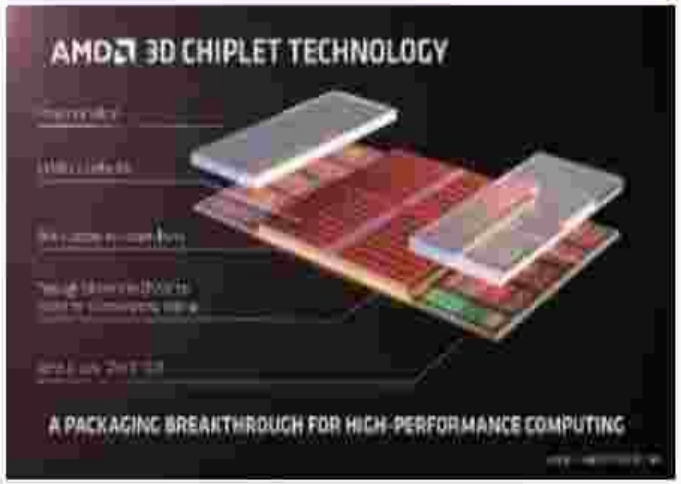
• **3D方案由于成本非常高，相关应用较少，主要在对性能要求非常苛刻的高端应用领域。**相关的HBM、3D V-Cache等产品主要用在对计算要求较高的AI芯片中或者对延迟要求非常高的游戏CPU芯片中。HBM主要将各种DRAM芯片进行堆叠，从而扩大内存容量，在高性能计算领域需求量较大。3D V-Cache主要将L3 cache堆叠在CPU上，以减小延迟，这在游戏领域需求量较大。

Intel的3D Co-EMIB方案集成度非常高



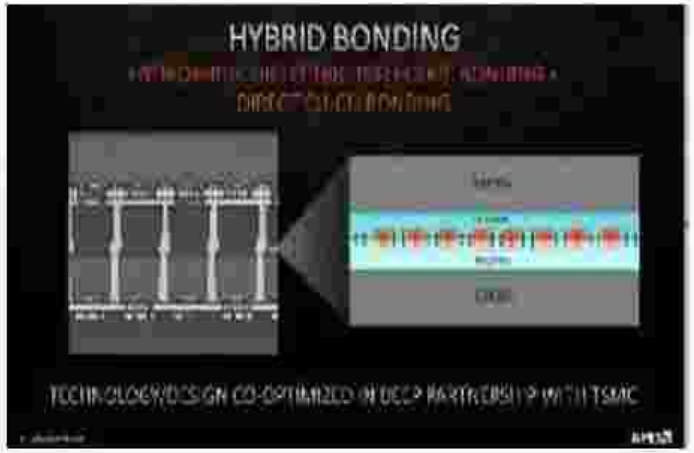
数据来源：TSMC

AMD的3D V-Cache架构



数据来源：AMD

混合键合可实现9微米的间距，实现更高集成度

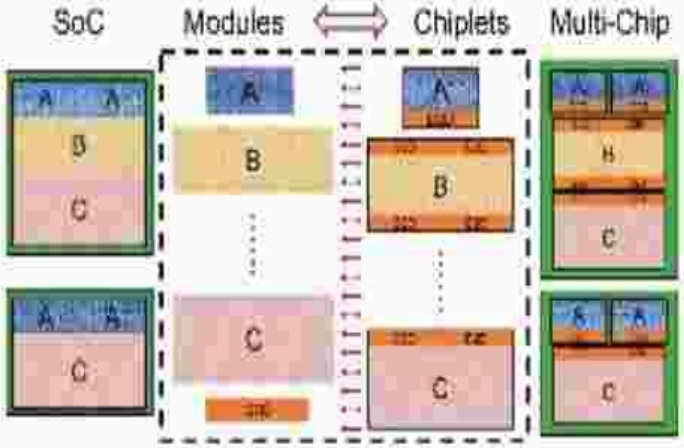


数据来源：AMD

高性能大芯片是实现成本性能最优化的应用

- **多芯片集成的Chiplet方案是在以先进工艺为基础的SoC方案遇到摩尔定律发展的门槛时，所延伸的提升性能、减小成本、优化性价比的方案。**SoC方案为将A、B、C等各种IP内核进行组合搭配，无需D2D（Die to Die）的IP；而Chiplet方案为将A、B、C等各种内核分别与D2D IP进行组合，依次封装，并在基板或者硅转接板上进行互连组合，并利用高密度集成封装方案进行封装。
- **Chiplet方案的成本随着集成密度的提高而不断提高，需要和小芯片的成本进行综合考量，实现最优综合性能。**例如2D方案的MCM封装集成密度最低，bump密度为90微米，成本也最低。而RDL Interposer和Si Interposer的集成密度逐步提高，bump密度分别达到45/30微米，成本也相对提升，其中硅转接板的成本最高。3D封装的bump密度达到9微米，成本是所有集成封装方案中最高的。

SoC和Chiplet方案的概念图



数据来源：《Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration》

Chiplet封装密度越高，成本也越高

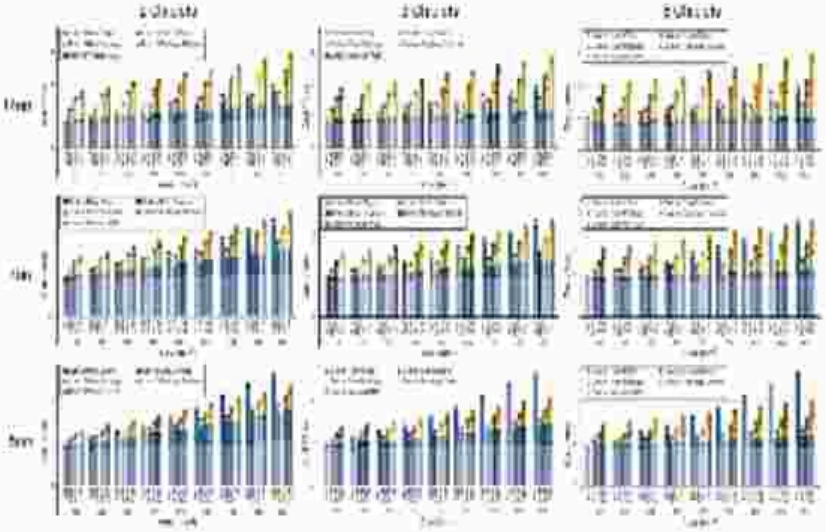
性能参数	MCM	RDL Interposer	Si Interposer	3D封装
集成密度	低	较高	较高	高
布线密度(μm/μm)	45272	44959	0.4/0.4	0.4/0.4
bump密度/μm	90	45	30	9
设计复杂度	低	中	较高	高
信号传输长度/mm	<10	<5	<5	<0.03
成本	低	中	较高	高
供应商	封测厂	晶圆厂/封测	晶圆厂	晶圆厂

数据来源：《Chiplet关键技术与挑战》，国泰君安证券研究

高性能大芯片是实现成本性能最优化的应用

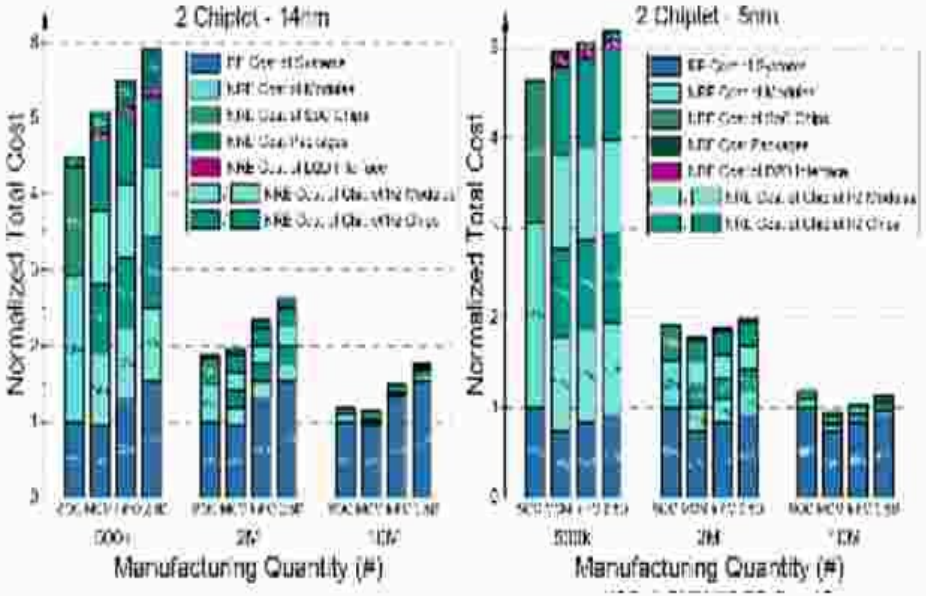
- **就成本角度进行考量，一块单片SoC芯片或者Chiplet芯片，主要成本可粗略划分为RE（recurring engineering）成本和NRE（non-recurring engineering）成本。**NRE成本为电路设计中的一次性成本，包括软件、IP授权、模块/芯片/封装设计、验证、掩模版等费用，针对于单颗芯片是摊销后的成本。RE成本为大规模量产中的制造成本，包括晶圆、封装、测试等。
- **RE成本方面：**①工艺节点越小，芯片面积越大，多芯片集成的Chiplet方案带来的好处越大。②小芯粒数量的提升，对成本的优化具有一定效果。
- **NRE成本方面：**多芯片Chiplet方案会造成非常高的额外NRE成本，只有当量产数量足够高，才有足够性价比。

不同工艺节点下，不同芯片集成的归一化RE成本比较



数据来源：《Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration》，备注：图片中的成本均为相对于100平方毫米的SoC芯片的成本进行归一化的结果

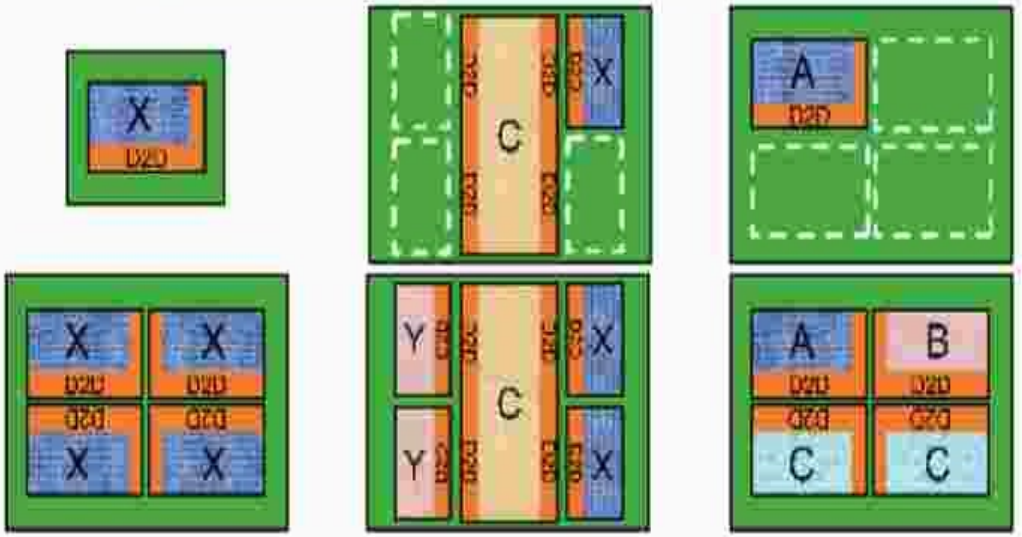
多芯片Chiplet方案会造成非常高的额外NRE成本，只有当量产数量足够高，才有足够性价比



数据来源：《Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration》

- **除了面积、工艺、小芯片的数量以外，Chiplet在多芯片架构复用和异构方面存也在着巨大的成本优势。**多芯片复用架构主要分为三类：①SCMS（单芯片多系统）；②OCME（一中心多拓展）；③FSMC（固定插座多组合）。
- **SCMS：**芯片的复用，使Chiplet相较于SoC而言节省一次性投入成本。该种方案只需要一个芯片即可，适用于同一产品线不同等级的产品。AMD和国内最初的产品架构就是采用该方案。
- **OCME：**实现了异构工艺，将不同的成熟工艺产品和先进工艺产品进行拼接。诸如AMD的ZEN3架构采用的就是该方案。
- **FSMC：**将复用的可能性最大化，即将可复用的芯片最小化，这样一次性投入成本摊销的收益就越大。多芯片集成的Chiplet方案的成本优势将会最大化。

多芯片复用架构主要分为三类，成本效益逐级提高

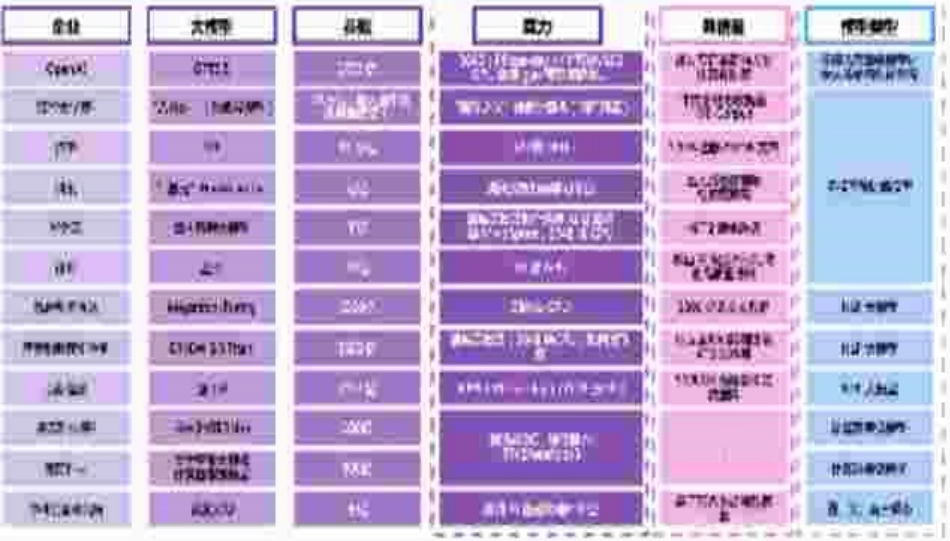


数据来源：《Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration》

AI+数字经济催生高算力需求，Chiplet深度受益

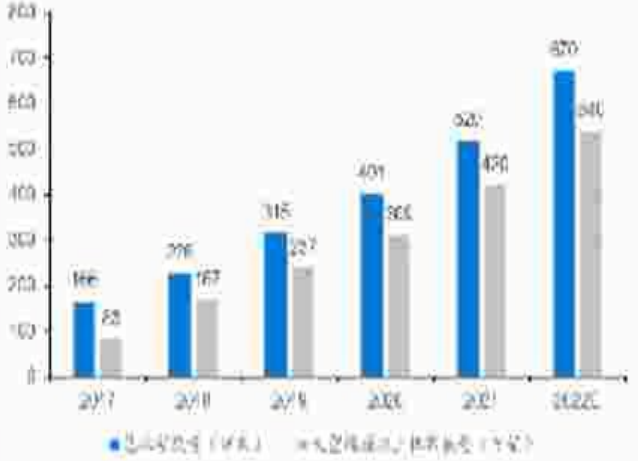
- **ChatGPT作为生成式AI的现象级产品，将催生庞大的产业链算力需求。**ChatGPT是OpenAI开发的聊天机器人，在2022年11月推出，一经推出，就成为迄今为止用户量增长最快的消费应用程序，仅用2月就积累1亿用户数量，即使是海外现象级应用TikTok也用了9个月的时间。未来国内外诸如百度等大模型公司、科大讯飞等应用端公司都在积极参与，带来庞大的算力需求。
- **数字经济推动数据中心建设快速发展，带动计算需求增长。受益于5G、人工智能、大数据、云计算等新兴产业发展，对海量数据处理的需求不断提升，数据中心成为数字化发展的重要基础设施。**截止2021年底，我国在用数据中心机架规模达到520万架，近五年CAGR超过30%，其中大型以上机架规模达420万架，占比达80%。进入数字经济时代，数据量呈指数级增长，对算力提出了巨大需求。据Cisco预计，2021年计算能力更强的超级数据中心将达到628座，占数据中心总量的53%。

ChatGPT大模型是AI基础设施，将进一步推动AIGC行业发展



数据来源：甲子光年

中国数据中心机架规模快速增长（万架）



数据来源：《数据中心白皮书（2022年）》，国泰君安证券研究

全球超级数据中心数量快速增长（座）

数据来源：Cisco Global Cloud Index，国泰君安证券研究

- **受益于AI和数字经济的需求，全球GPU、MPU、AI芯片等大算力芯片需求大幅提升。**根据华经产业研究院的数据，2027年全球GPU市场规模预计达到1853.1亿美元，21-27年CAGR为33%。2022年MPU的全球市场规模也已经突破1000亿美元。2024年AI的中国市场规模也预计突破785亿元，21-24年CAGR为46%。
- **Chiplet方案是继续提升大芯片算力的主要方案之一，将伴随高性能算力需求的爆发而强势增长。**根据Yole，2021年先进封装市场收入达374亿美元，预计2027年将达到650亿美元，CAGR为10%。其中2.5D/3D的市场规模预计27年将达到150亿美元，21-27年CAGR为14%。

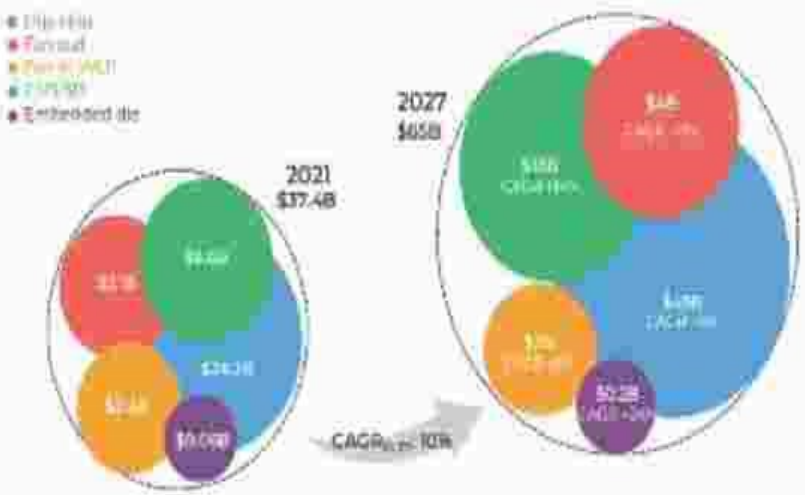
全球GPU市场规模快速增长（亿美元）

数据来源：VMR，华经产业研究院，国泰君安证券研究

全球MPU市场规模超过千亿美金

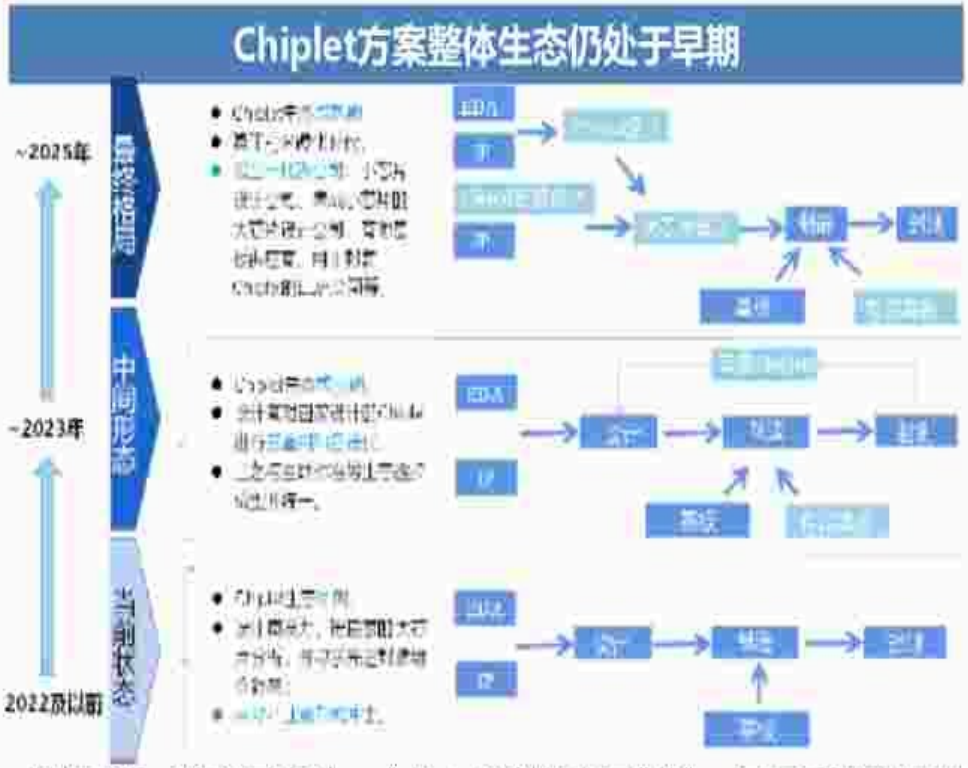
数据来源：IC Insights，国泰君安证券研究

27年先进封装市场空间将突破650亿美元



数据来源：Yole

- **Chiplet生态仍处于发展早期，就产业链而言，价值量的增长点主要集中在封测端和材料端。**目前产业仍处于Chiplet生态成长期，设计厂商主要采用已有的EDA和IP针对Chiplets进行自重用和自迭代，工艺和互连标准尚未统一。产业链中最大的价值量增长源于新的高密度集成的封装方案带来的封测端和材料端的应用，未来随着生态和技术的成熟，EDA等更上游的价值量也会逐步增加。
- **Chiplet业务链中，晶圆厂和封测厂都逐步向产业链下游垂直整合，以扩大自身的业务空间和利润增长点。**晶圆厂围绕硅互连技术进行发展，从带TSV的转接板向RDL层、微凸点等领域拓展，自上而下，拓展价值空间。封测厂在争取从原有的基板、C4凸点向上游Chiplet业务链中的RDL层、TSV转接板、微凸点等方向发展，因为该块业务精细度不高但有较大业务量。不过，封测厂话语权不如晶圆厂，大多封测厂更多向下游拓展，将更多的元器件、射频器件、PMIC等集成到基板中，以期获得更大的价值量增长。



数据来源：电子技术设计，《chiplet关键技术与挑战》，国泰君安证券研究

- **在封装端，对于封装厂而言，价值量额外增长主要集中在微凸点、转接板、线互连等领域。**

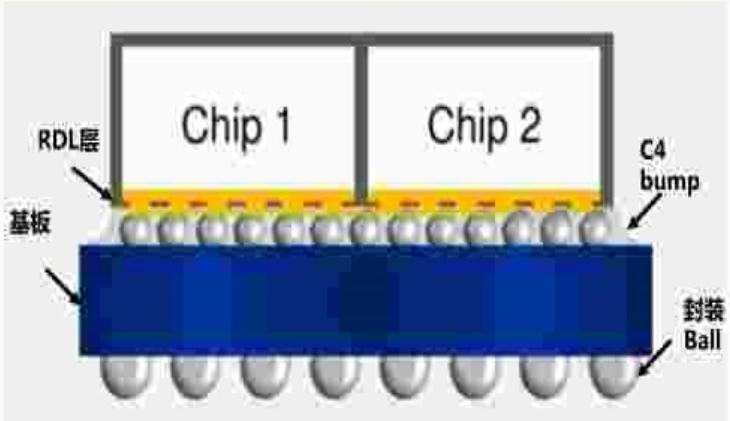
①在MCM的2D领域，只增加了额外的基板内互连，价值量增长最低。

②在RDL转接板的2.1D领域（RDL整体较薄，介于2.5D和2D之间，又可称为2.1D），主要为台积电的InFO和长电的XDFO等方案。InFO方案是Chipfirst技术，没有微凸点，由于该类方案主要由TSMC主导，下游封测厂话语权较小，价值量仍主要局限于原有封测领域，如C4 bump和基板等。XDFOI方案是chiplast方案，存在微凸点，该类方案包含多层RDL层、微凸点、互连线等，封测厂可做价值量更大。

③在硅转接板的2.5D领域，主要为台积电的CoWoS等方案，该方案价值量较多，包括微凸点、RDL、硅转接板、TSV等，但同样受限于TSMC等晶圆厂较为强势的话语权，大多硅转接板等价值量都被晶圆厂拿走。但是台积电等晶圆厂开价过高，终端厂等正尝试分散供应链，各环节找不同的厂商，以实现利益最大化。

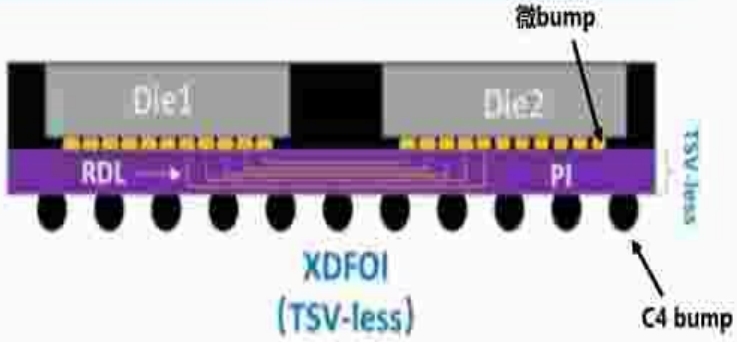
④在3D领域，如HBM方案，由于精细要求较高，这部分基本全部依赖晶圆厂，在晶圆制造领域直接堆叠完成。

InFO这类RDL Interposer封装的额外价值量主要在RDL层等（chip first）



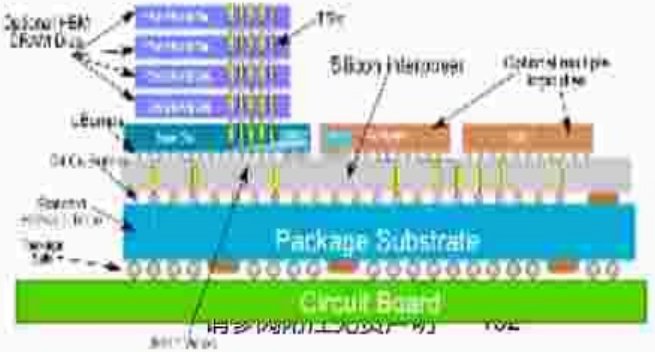
数据来源：TSMC官网

XDFOI这类RDL Interposer封装的额外价值量主要在RDL层、微凸点、互连线等（chip last）



数据来源：长电科技官网，国泰君安证券研究

硅转接板方案的额外价值量主要在微凸点、硅转接板、线互连等



数据来源：TSMC官网

产业生态发展早期，封测端是主要受益点

- **封装端受益公司包括通富微电、长电科技、甬矽电子等，相关公司均有属于自己的Chiplet方案，预计都将批量量产。**在全球封测企业中，不止考虑OSAT，长电科技2021年营收排名第四，有XDFOI平台；通富微电2021年营收排名第七，有VISionS平台；华天2021年营收排名第八，积极布局先进封装业务；甬矽电子营收排名相对靠后，但业务均是先进封装业务。
- **在测试端，受益于小芯粒带来更多的测试需求以及KGSD带来更复杂的测试要求，相关测试公司和测试设备公司将深度受益。**例如伟测科技、长川科技、和林微纳等都将较为受益。
- **在材料端，受益于Chiplet的突破和高算力的需求，ABF膜的需求在不断增长，相关基板产业链公司将深度受益。**例如生益科技、深南电路等都将较为受益。

06

算力是智能世界的基础，产业生态和投资图谱逐步清晰

算力是智能世界的基础，算力基础设施从云向算泛在演进

- **智能世界三要素**：数据、算力、算法是智能世界三要素，其中算力平台是核心基础。
- **算力两大类：通用算力、HPC（高性能计算，High-performance computing）算力。**其中通用算力计算量小，用于常规应用。HPC算力是一个计算机集群系统，通过各种互联技术将多个计算机系统连接在一起，利用所有被连接系统的综合计算能力来处理大型计算问题。
- 算力基础设施从云向算泛在演进，其位置的分布从中心向边缘和端侧泛在延伸，将出现云、边、端三级算力架构。

构建数据、算力、智能之间的互联网络

专用算力是算力中极为重要一环

科学计算类：物理化学，气象环保，生命科学，天文探测等
工程计算类：计算机辅助工程/制造，电子设计自动化，电磁仿真等
智能计算类：机器学习，深度学习，数据分析等

算力基础设施从云向算泛在演进

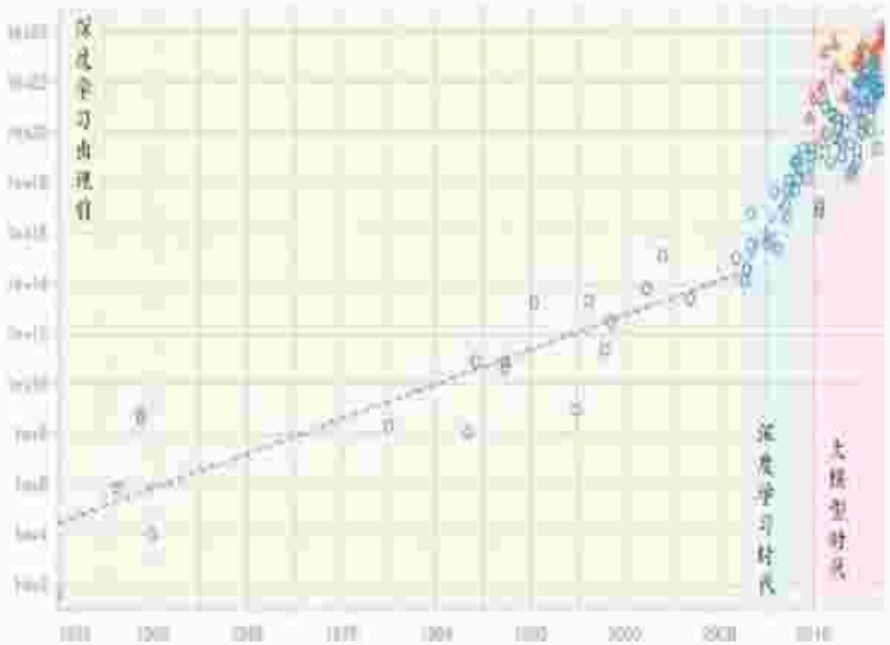
国泰君安证券 GUOTAI JUNAN SECURITIES

数据来源：国泰君安证券研究

AI大模型带动算力需求超越摩尔定律增长

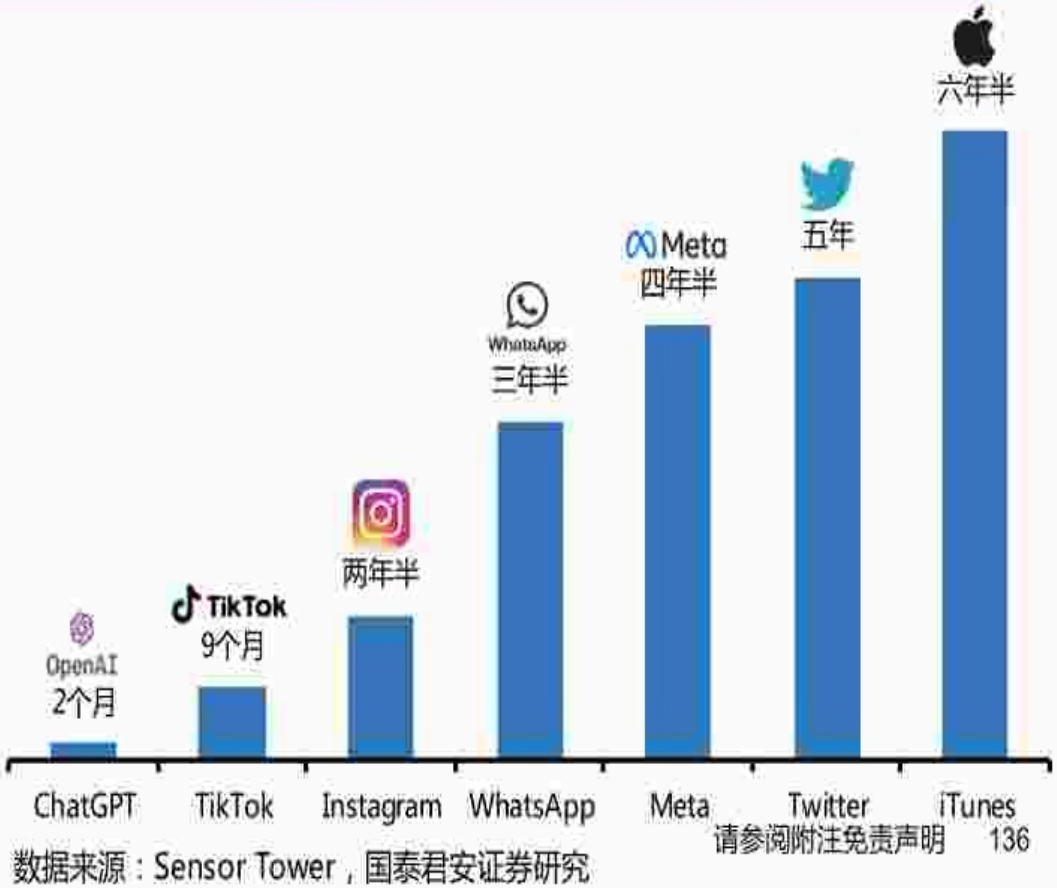
- **AI模型训练算力增长速度超越芯片摩尔定律。**AI训练任务中的算力增长（所需算力每3.5个月翻一倍）已经超越摩尔定律（晶体管数量每18月翻一倍）。
- **ChatGPT仅推出两个月，月活跃用户数预计已达1亿。**ChatGPT在2023年1月达到1亿月活跃用户，平均每天有1,300多万访客，用2个月时间达到1亿月活数，成为史上最快达到1亿月活跃用户的应用，TikTok、Instagram、Facebook、Twitter则分别用了9个月、2年半、4年半、5年的时间。

大模型时代算力需求增长超越摩尔定律



数据来源：Google Scholar，国泰君安证券研究

chatgpt仅用2个月月活用户突破1亿

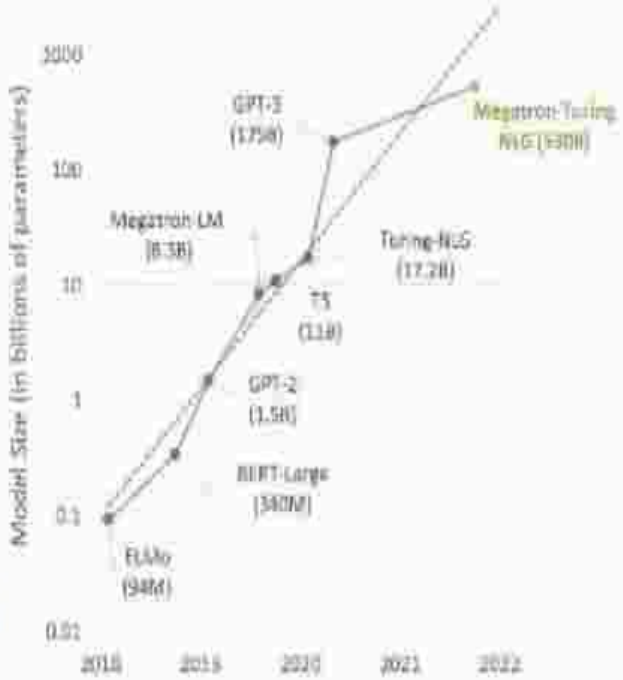


数据来源：Sensor Tower，国泰君安证券研究

AI大模型核心在于数据、算力与模型，需要消耗大量算力

- **预训练算力需求**：训练一次13亿参数的GPT-3 XL模型需要的全部算力约为27.5PFlop/s-day，而训练一次1,746亿参数的GPT-3模型需要的算力约为3,640 PFlop/s-day，对应的单次训练成本高达460万美元。
- **日常运营算力需求**：ChatGPT在日常与用户交互过程中需要大量的算力支持，结合访问量与内容量测算，单月运营算力约4,800PFlop/s-day；2023年1月ChatGPT官网总访问量已经6.16亿次，而ChatGPT每次交互产生的算力云服务成本约1~5美分，对应的单月运营成本高达千万美元。
- **调优迭代算力需求**：ChatGPT模型达到需要不断进行Finetune模型调优，以确保模型处于最佳应用状态；预计每月模型调优带来的算力需求为82.5~137.5 PFlop/s-day。

模型的尺寸在过去5年增长了5000倍



不同 NLP 模型参数量及训练算力对比

	模型	总计算量（PFlop/s-day）	总计算量（Flops）	参数量（百万个）	令牌数量（十亿）
T5模型	T5-Small	2.08E+00	1.80E+20	60	1000
	T5-Base	7.64E+00	6.60E+20	220	1000
	T5-Large	2.67E+01	2.31E+21	770	1000
	T5-3B	1.04E+02	9.00E+21	3000	1000
	T5-11B	3.82E+02	3.30E+22	11000	1000
BERT 模型	BERT-Base	1.89E+00	1.64E+20	109	250
	BERT-Large	6.16E+00	5.33E+20	355	250
	ROBERTa-Base	1.74E+00	1.50E+21	125	2000
	ROBERTa-Large	4.93E+01	4.26E+21	355	2000
GPT 模型	GPT-3 Small	2.60E+00	2.25E+20	125	300
	GPT-3 Medium	7.42E+00	6.41E+20	356	300
	GPT-3 Large	1.58E+01	1.37E+21	760	300
	GPT-3 XL	2.75E+01	2.38E+21	1320	300
	GPT-3 2.7B	5.52E+01	4.77E+21	2650	300
	GPT-3 6.7B	1.39E+02	1.20E+22	6660	300
	GPT-3 13B	2.68E+02	2.31E+22	12850	300
	GPT-3 175B	3.64E+03	3.14E+23	174600	300

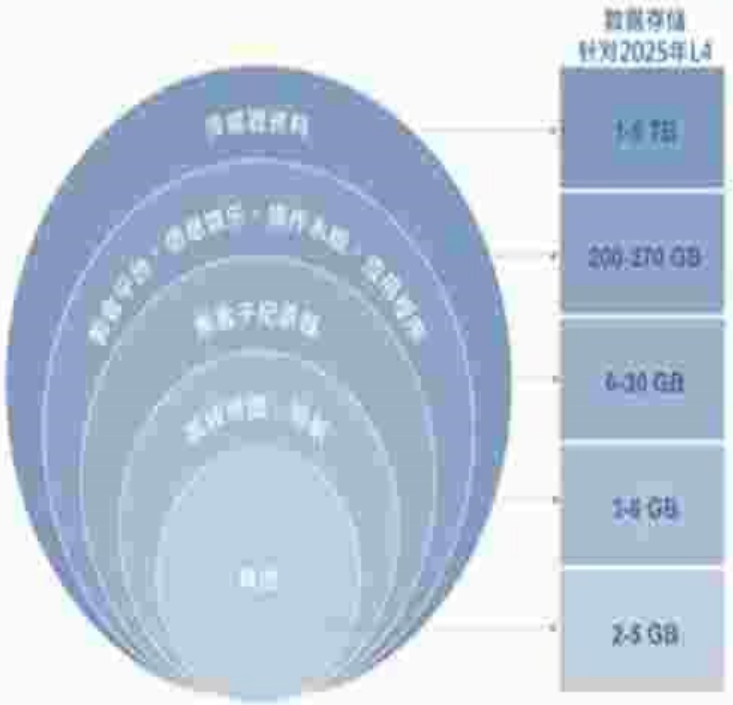
数据来源：Open AI，国泰君安证券研究

- **MR的推出更带来对低延时网络传输和底层算力技术升级的需求。**虚拟世界需要强大的图像实时渲染能力、计算和存储海量数据资源，头显交互设备的出现将进一步增加对云计算和边缘计算的应用需求。云计算能将终端渲染逐步迁移至云端，基于规模效应摊低运算成本，提升服务器使用效率，提升虚拟世界的可进入性。而边缘计算则更能满足实时数据分析需求、缓解中心云的计算负载。
- **汽车智能化需求持续升级带来数据流量的急剧飙升。**随着自动驾驶等级提升，车载信息娱乐系统、长续航电池及5G网络的引入，车辆要面对的计算量越来越大，网络架构升级、本地实时处理能力、“大容量缓存和存储”规格将成为硬需求。

算力升级是支撑虚拟世界内容创作与真实交互的保障

数据来源：中国信通院、华为

2025年L4等级无人自动驾驶数据存储需求



数据来源：Counterpoint Technology Market Research

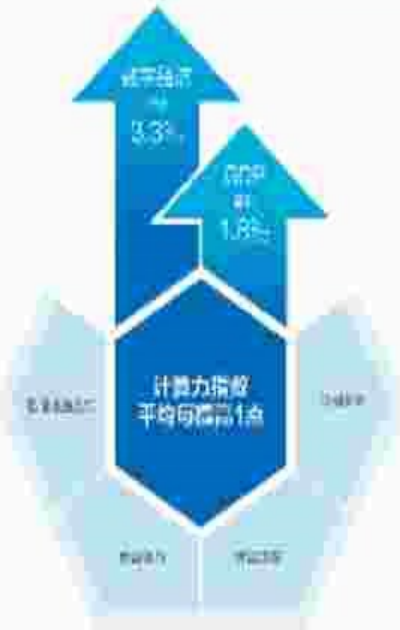
算力与经济水平直接挂钩，且具有倍增效应

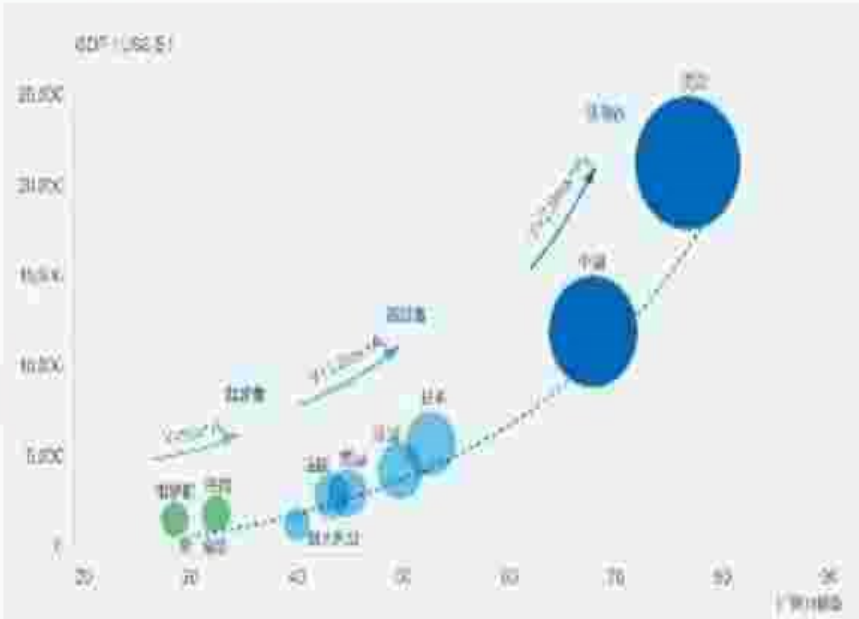
- **全球各国算力规模与经济发展水平呈现正相关。**2021年算力规模前20的国家中有17个是全球排名前20的经济体，并且前五名排名一致。
- **算力对经济有倍增效应。**数字经济作为GDP的组成部分，占比正在逐年增加，而算力是数字化技术持续发挥效益的根本性要素。根据IDC的报告，计算力指数平均每提高1个点，数字经济和GDP将分别增长3.3‰和1.8‰。当一个国家的计算力指数达到40分以上时，指数每提升1点，对于GDP增长的拉动将提高到1.5倍；当计算力指数达到60分以上时，对GDP的拉动将进一步提升至2.9倍。

算力排名与经济排名较为吻合

数据来源：中国信通院信通院2022年算力白皮书

算力对经济有倍增效应





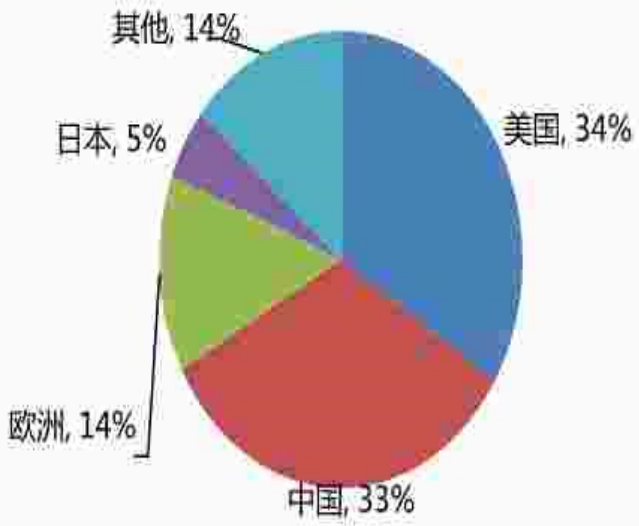
数据来源：IDC

中美算力布局较为领先，中国智能算力在追赶

- **智能算力规模和增速亮眼。**根据信通院算力白皮书，2021年全球算力增速超过40%，华为GIV预测2030年人类将迎来YB数据时代，全球算力平均年增速达到65%,其中基础算力平均年增速达27%;智能算力占大头,平均年增速超过80%;超算算力平均年增速超过34%。
- **中美算力在全球属于领先地位。**美国、中国、欧洲、日本在全球算力规模中的份额分别为34%、33%、14%和5%，其中全球基础算力美国份额达37%，中国以26%份额排名第二;智能算力方面，中国、美国分别占比为45%和28%；美国、日本、中国在超级计算综合性能指标方面份额分别为48%、22%、18%。

全球算力规模增长速度在40%以上

中美全球算力分布较为领先



中国占比较大为基础算力，智能算力当前快速赶上

数据来源：信通院算力白皮书，国泰君安证券研究

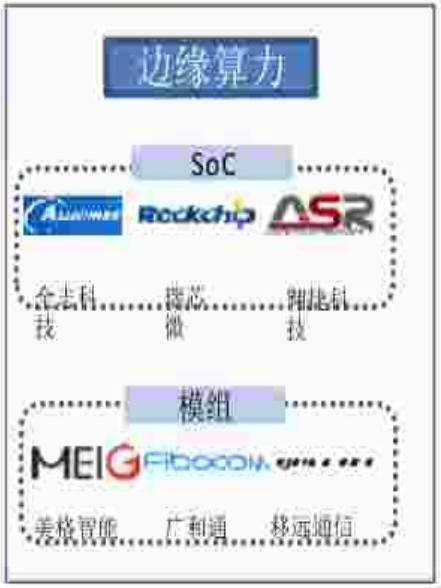
海外以互联网大厂为主导，资本开支向AI大幅倾斜

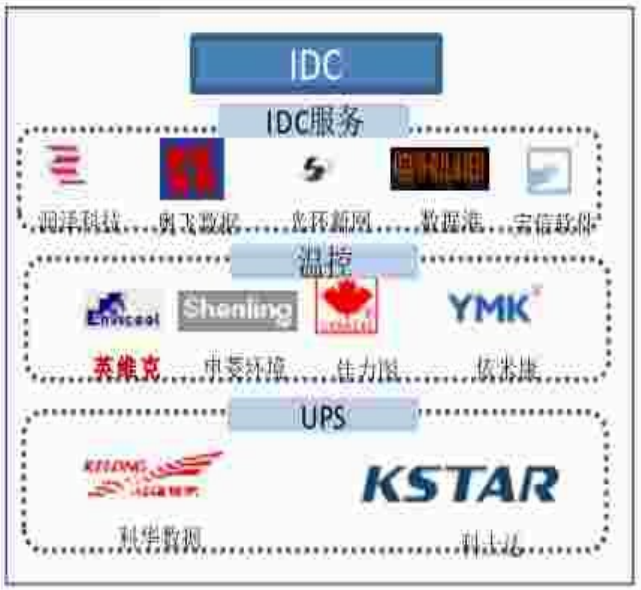
- 美国人工智能公司在过去的五年间获得的投资占到了全球的56%,数百亿美元，谷歌、微软、亚马逊、Meta四家美国科技巨头在经历了2022和2023年初史无前例的大裁员之后，这些公司一边降本增效，一边All in AI。

北美大厂积极自研和购买算力芯片

	自研芯片	算力布局	2023Q1	2023年展望	投入领域
AWS	Inferentia、trainium	AWS宣布即将推出的EC2超级集群（EC2 P5实例）可扩展至20000个互连的H100。	我们预计在数据中心建设和服务器的投入会在Q2以及后面持续的提升	我们预计总体资本开支会略微高于2022年，同时资本开支会在技术基础设施显著提升，而在办公室设施下降	数据中心，服务器，供应链
Google	TPUv4	已经建成包含26000个H100的A3；当前已经部署了数十台TPU v4超级计算机，每台拥有4096个TPU芯片；	发布了专门用于推理或训练的机器学习定制化芯片	资本开支包括金融租赁，2023年预计资本开支会低于2022年的590亿美金，去年主要因为完善物流网络的投资，未来这个数字会逐步减少。我们将持续投入基础设施来支持AWS客户需求，包括支持LLM和生成式AI	AWS
Meta	MTIA	已经包含有2000个DGX A100，配备16000个 A100；semianalysis认为其是2023年H100 GPU最大的买家，据专家交流可能采购了3万片H100以上。	投入的三个领域：1）非AI计算需求：计算和存储来支撑现有业务；2）核心的AI投资，支持Discovery engine、排序广告等的建设；当我们评估ROI感觉可以的话，这些都会提高我们对AI的投入；3）支持生成式AI，现在虽然难评估，但我们未来会提高资本开支，同时平衡好AI能力的建设	我们预期资本开支在300-330亿元，与前次预测无异，我们继续建设AI能力来支持ADS、Feed和Reels，以及将提升我们在生成式AI的容量投资（earning transcript）	数据中心
Microsoft	Athena	Azure云拥有10000个GPU和285000个CPU内核；据专家交流北美各家可能采购了3万片H100以上。	我们希望引领AI平台的浪潮并做相应投资来实现它	我们预计资本开支会有显著的后续环比增长，主要驱动为Azure AI基础设施。注意可能有正常的季度支出波动	Azure AI基础设施

数据来源：各公司官网、国泰君安证券研究





数据来源：中兴通讯等上述各公司官网，国泰君安证券研究

算力芯片：AI服务器带动高端GPU用量及ASP提升

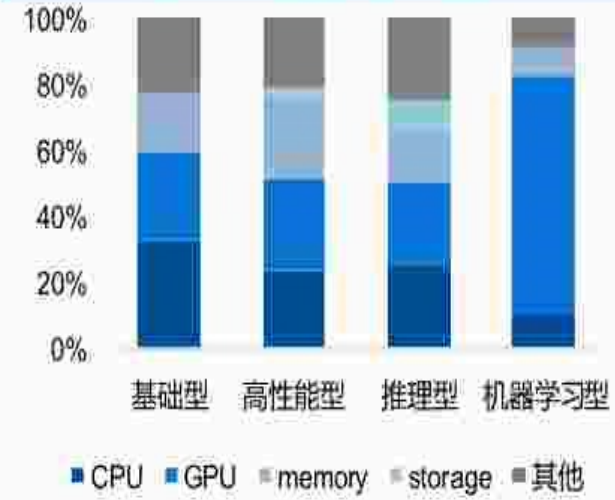
- **AI服务器采用异构式架构，GPU数量远高于普通服务器。**AI服务器和普通服务器的主要区别在于：1）架构不同，AI服务器采用CPU+GPU/FPGA/ASIC等异构式架构，而普通服务器一般是CPU架构；2）GPU数量差别巨大，AI服务器单服务器GPU用量通常在4 颗以上。例如：NVIDIA DGX A100包括8个A100 GPU + 2个AMD Rome CPU，而浪潮英信服务器NF5280M6仅配置1-2个英特尔第三代Xeon处理器。
- **GPU 架构为主流加速架构，是服务器核心成本构成。**GPU采用并行计算，适用于处理密集型运算，如图形渲染、机器学习等场景，AI算力需求的提升推动了GPU卡的运算速度和用量需求进一步增长。根据 IDC 数据，2022年GPU加速卡占据AI市场89%的份额，在机器学习型服务器中GPU成本占比达72.8%。

AI服务器和普通服务器的区别

类型	典型产品	芯片	价格	数量
AI服务器	NVIDIA DGX A100	A100 Tensor Core GPU	USD14,999	8
		64 core AMD Rome CPU	~USD7,000	2
普通服务器	浪潮英信服务器NF5280M6	Intel第三代Xeon处理器	RMB64,000	1~2

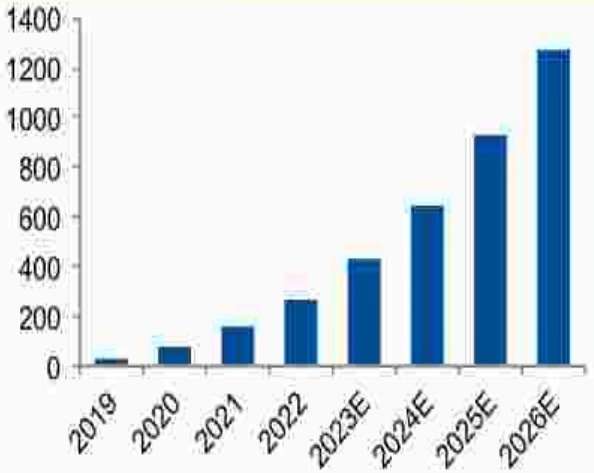
数据来源：英伟达官网，浪潮信息官网，国泰君安证券研究

服务器成本构成情况



数据来源：IDC，国泰君安证券研究

中国智能算力规模及预测（单位：EFLOPS）



数据来源：IDC，国泰君安证券研究

算力芯片：多卡互联&软件生态铸就英伟达强力护城河

- **高速互连技术开创者，多卡互联优势显著。**为实现超算模型的高速通信需求，英伟达开创式提出NVLink和NVSwitch技术：NVLink主要用于连接多个GPU，以加速高性能计算和深度学习等应用；NVSwitch用于连接多个GPU和CPU，形成高性能计算系统，适用于更复杂和大规模的场景，用户可根据具体应用需求和系统配置来决定使用NVLink或NVSwitch。GH200超级芯片所采用的NVLink-C2C技术，通过Chiplet工艺将CPU+GPU组合到同一封装，相比于PCIe5在能效方面提升25倍，面积效率提升90倍。
- **CUDA生态不断演进，满足各类行业需求。**英伟达依托于CUDA软件栈进行第三方应用及工具的扩展，形成了广义的CUDA生态，并在此基础上向上扩展出CUDA-X，以对接不同的行业应用需求，分为面向AI计算的CUDA-X AI和面向HPC计算的CUDA-X HPC。

NVLink和NVSwitch技术

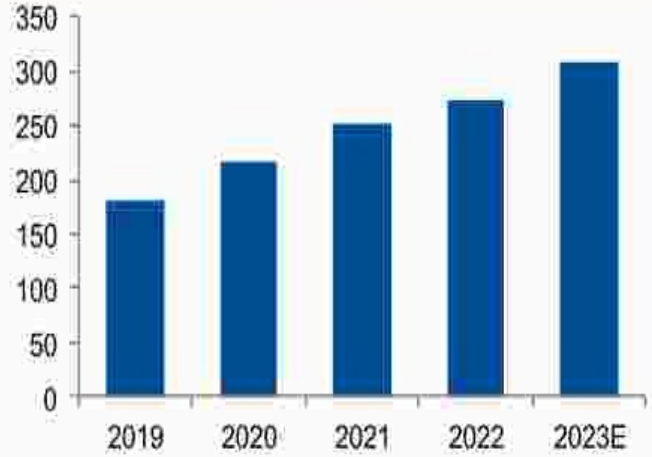
NVLink	第二代	第三代	第四代
NVLink 总带宽	300GB/s	600GB/s	900GB/s
每个GPU 最大链路数	6	12	18
架构支持	NVIDIA Volta	NVIDIA Ampere	NVIDIA Hopper
NVSwitch	**第一代**	**第二代**	**第三代**
直连或节点中GPU数量	最多8个	最多8个	最多8个
NVSwitch GPU之间带宽	300GB/s	600GB/s	900GB/s
聚合总带宽	2.4TB/s	4.8TB/s	7.2TB/s
架构支持	NVIDIA Volta	NVIDIA Ampere	NVIDIA Hopper

数据来源：英伟达官网，国泰君安证券研究

数据来源：英伟达官网

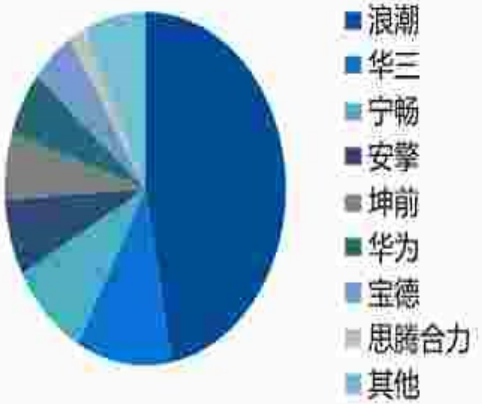
- **受益大模型热潮，国内AI服务器市场增量明显。** ChatGPT横空出世，引发科技企业大模型竞赛，全球算力需求呈指数级增加，带动国内AI服务器市场快速增长，以浪潮信息为主的国内厂家占据国内AI服务器主要市场。
- **头部厂商持续加单，国内AI芯片需求强劲。**全球头部互联网厂商相继入局大模型赛道，以英伟达GPU为代表的算力核心产品订单暴增，一批中国AI芯片企业立足于不同技术路径开展研发，面向云计算、汽车、智能家居等领域，国内AI芯片市场同样前景广阔

2022年中国AI服务器市场规模（亿美元）



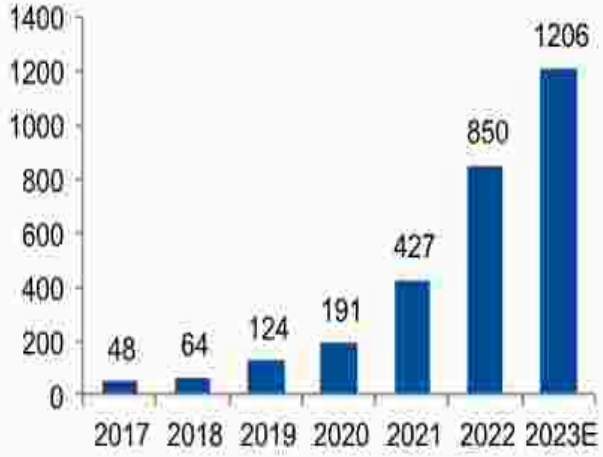
数据来源：中商产业研究院，国泰君安证券研究

2022年中国AI服务器市场份额



数据来源：华经产研，国泰君安证券研究

中国AI芯片市场规模（亿元）



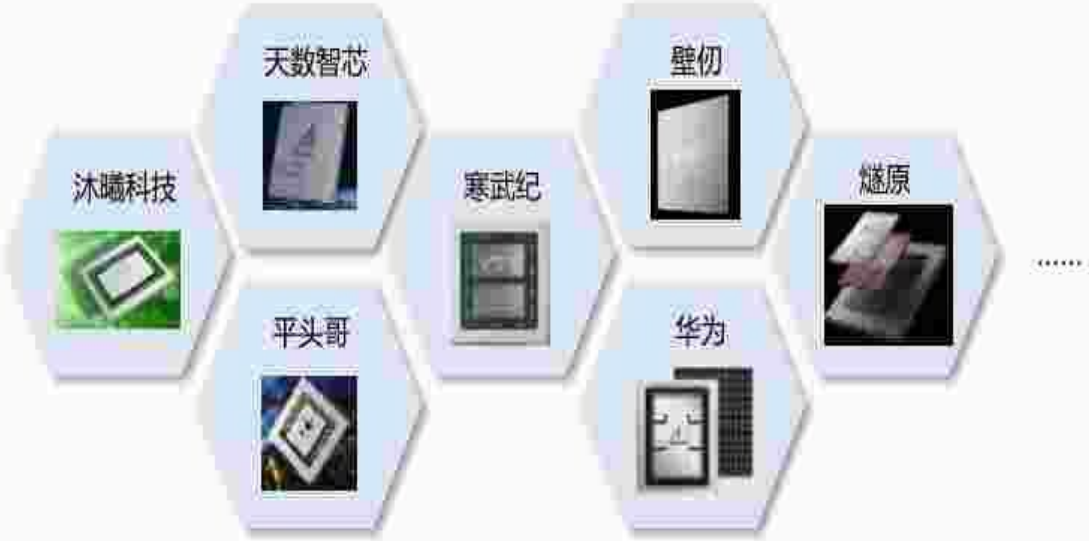
数据来源：IDC，国泰君安证券研究

自主可控：国内企业高算力AI芯片的布局和发展进度

- **国际巨头炙手可热，高性能GPU一芯难求。**科技企业大模型竞赛下，凭借A100、H100等绝对主流算力芯片，海外巨头英伟达订单火爆，芯片价格节节攀升；同时在地缘政治摩擦背景下，国产替代需求迫切。
- **在国家政策的指引下，国产公司遍地开花，各施所长不断缩短差距。**中国主要的AI芯片公司，寒武纪已量产四代芯片，其在研思元590性能预计能达到A100的70%，有望部分场景实现替代；华为昇腾采用独家达芬奇架构，昇腾910性能优越，处理速度达到同类产品180%；阿里平头哥另辟蹊径，其含光800推理性能和能效均达到世界前列水平；沐曦科技的通用芯片曦云MXC500对标A100；壁仞的BR100、燧原的邃思2.5以及天数智芯的智铠100等一系列高性能芯片即将面世。

2022年中国AI芯片市场规模占比

数据来源：IDC，国泰君安证券研究



数据来源：各公司官网

- **大力发展硬件的同时，软件也是及其重要的一环。**英伟达不仅在硬件方面具有统治力，在软件平台也具有很强的竞争力，CUDA生态已成为行业标的。对国内企业而言，兼顾软硬的发展路径至关重要，大力发展算力的同时，国产软件生态的建立刻不容缓。
- **共建生态开发平台，加速AI芯片落地。**寒武纪不仅实现了终端、云端、边缘端产品的完整布局，还为云边端全系列智能芯片与处理器产品提供统一的平台级基础系统软件Cambricon Neuware，使开发的应用可以在云边端互相兼容，大幅减少云边端不同平台的开发和应用迁移成本。华为同样致力于"一平台双驱动"为核心的昇腾AI生态，已有200多家合作伙伴经过认证，围绕昇腾的开发者超30万，其中核心开发者超2000，并在100多所高校开展了昇腾的人工智能课程。

寒武纪Cambricon NeuWare

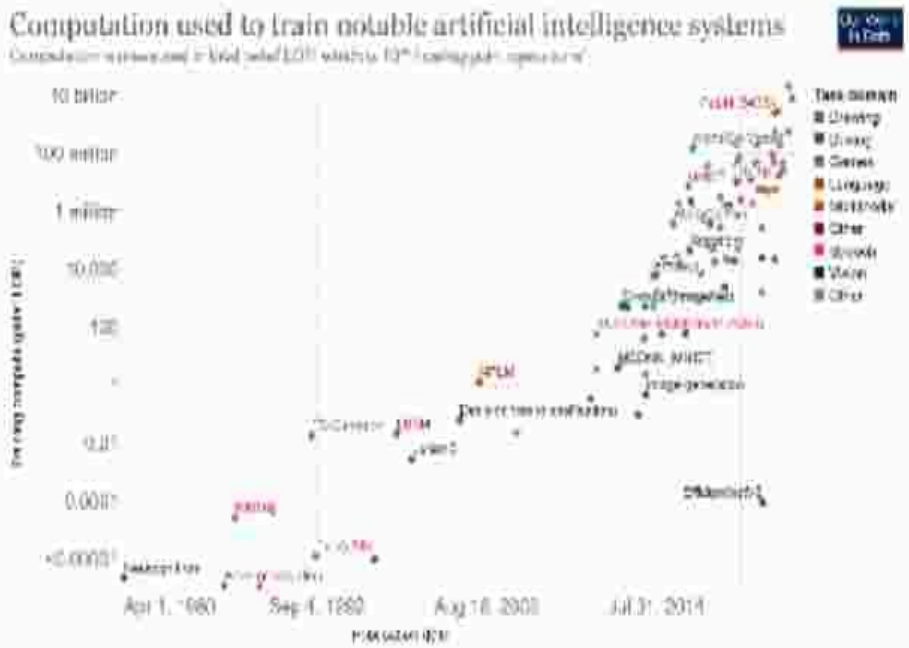
数据来源：寒武纪开发者论坛

华为昇腾产业生态

数据来源：华为官网

- **AI大模型的数据处理需求驱动服务器硬件升级与扩容。**目前AI大模型处理数据的吞吐量呈指数级增长。AI服务器对内存、硬盘存储容量要求更高。主要的提升是支持更大容量的内存满足当下实时负载增加的需求，提供更多外置硬盘插槽，并广泛支持NVME / PCIE等SSD，满足数据洪流需求。根据半导体观察，AI服务器容量是普通服务器的8倍，NAND容量是普通服务器的3倍。

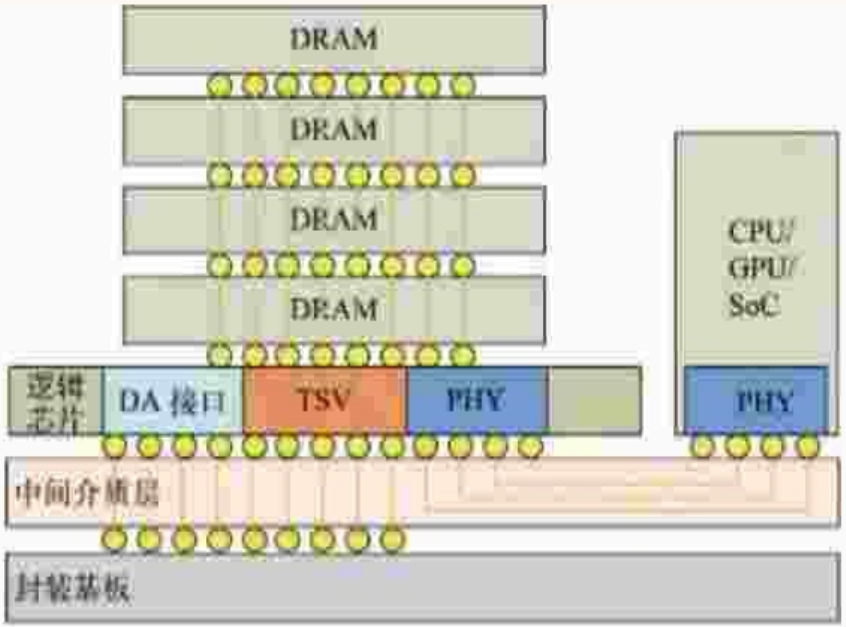
AI 大模型处理数据的吞吐量增长趋势



数据来源： Our WROLD in Data

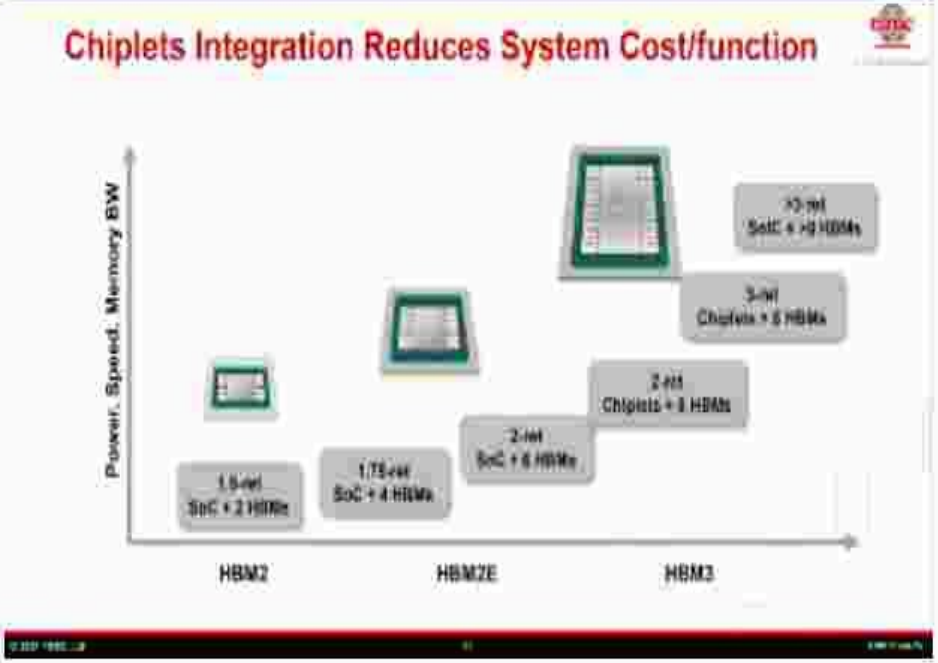
- **HBM内存技术：新型高性能存储产品的竞争与短缺。**HBM（High Bandwidth Memory，高带宽内存）是一款新型的 CPU/GPU 内存芯片，是将多个 DDR 芯片堆叠在一起后和 GPU 封装在一起，实现大容量和高位宽的 DDR 组合阵列。目前 HBM 占整个 DRAM 市场比重约 1.5%，为新型高性能存储产品，处于缺货低库存阶段。SK海力士、三星、美光等存储巨头都在HBM领域展开了升级竞赛。

HBM 的堆叠结构



数据来源：半导体产业纵横，《电子与封装》

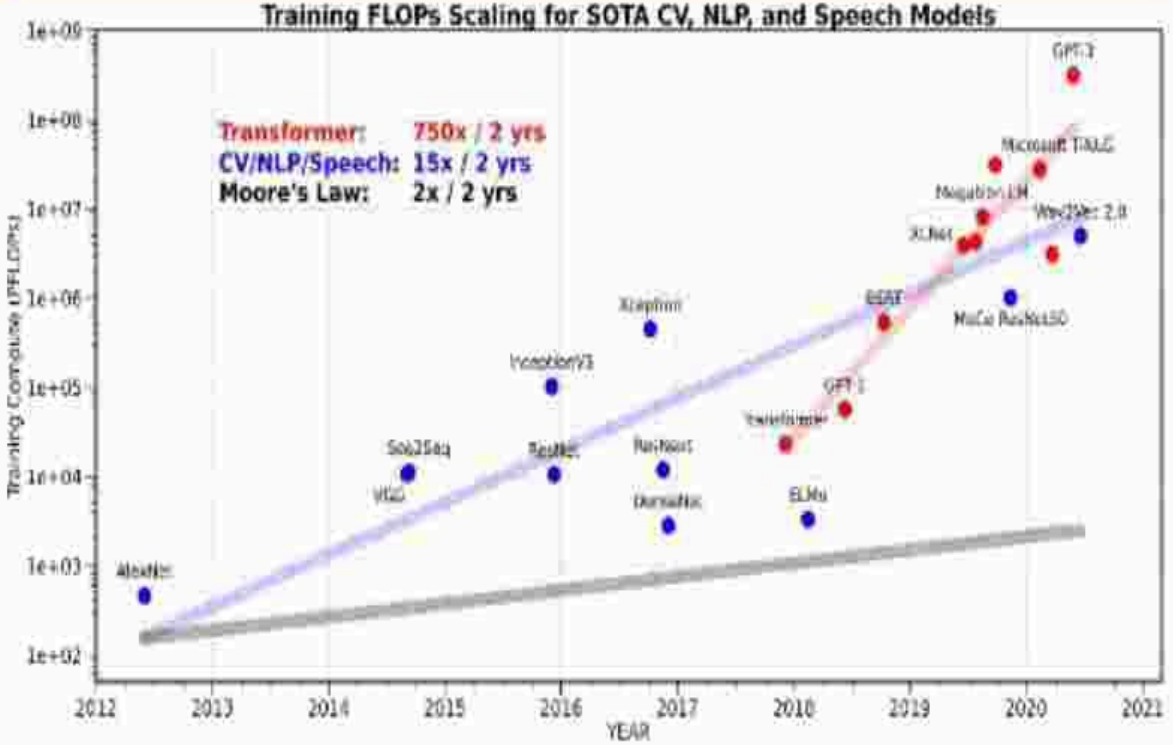
HBM 提供更快的数据处理速度



数据来源：TSMC

- **HBM：突破‘内存墙’的新一代3D DRAM解决方案。**“内存墙”是指处理器的运算能力超过了存储芯片的读取和写入能力，这导致了整体的计算能力被存储器所限制。3D化的DRAM是解决“内存墙”问题的主要途径。其中，HBM是3D DRAM的一种形式，相比其他DRAM的集成方式，它的数据传输速度最快，损耗最小，因此被认为是目前最理想的3D DRAM形式。HBM突破了内存容量和带宽的瓶颈，打破了"内存墙"对提升算力的束缚，被看作是新一代DRAM的解决方案。

存储带宽落后于算力成长速度形成“内存墙”



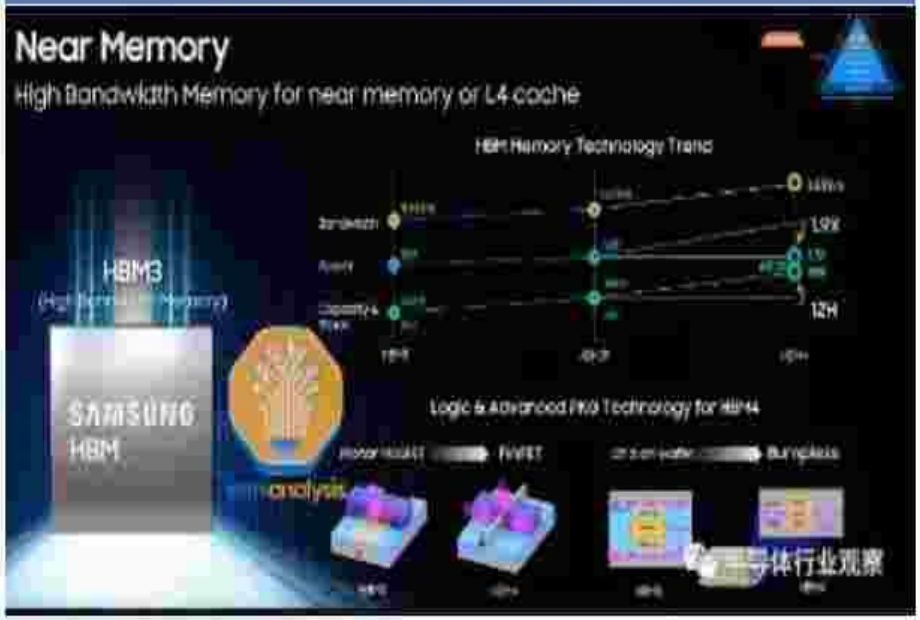
数据来源：RISElab

- **HBM市场：SK海力士占据主导地位。**据半导体行业观察，作为HBM的先驱，SK海力士是拥有最先进技术路线的领导者。SK海力士于2022年6月开始生目前唯一一家批量出货HBM3的供应商，拥有超过95%的市场份额，这是大多数H100 SKU所使用的。HBM现在的最大产HBM3，是配置为8层16GB HBM3模块。SK Hynix正在为AMD MI300X和Nvidia H100刷新生产数据速率为5.6 GT/s的12层24GB HBM3。三星紧随Hynix之后，预计将在2023年下半年发货HBM3，并正在大力投资以追赶市场份额。美光科技由于直到2018年，才开始从HMC转向HBM路线图，仍然停留在HBM2E，在HBM方面排名落后。

SK海力士在新一代技术方面保持最强

数据来源：半导体行业观察

三星预计2023下半年发布HBM3

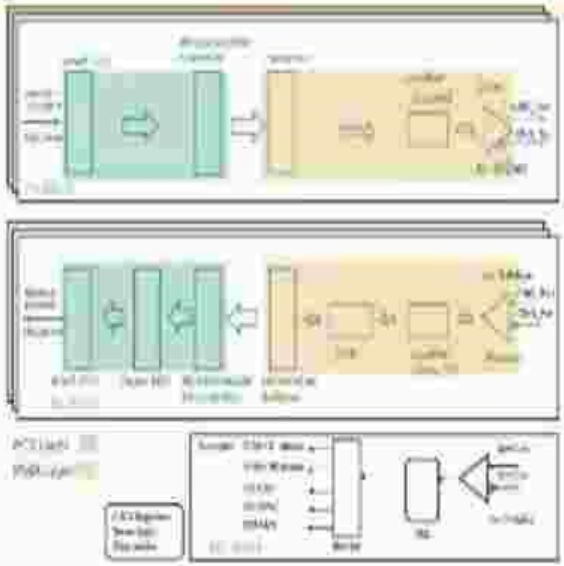


数据来源：半导体行业观察

交换芯片：AI服务器对SerDes提出更高性能要求

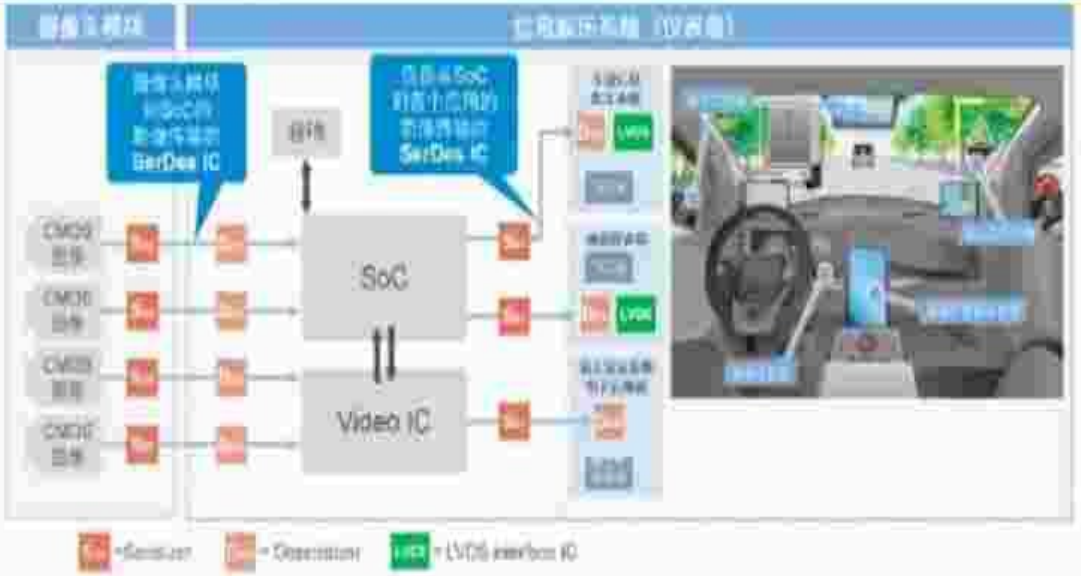
- **SerDes作为底层接口技术，是充分发挥AI硬件算力效能的关键。**SerDes是Serializer/Deserializer的缩写，即串行器和解串器，是目前主流的串行通信技术。通过数据在发送端并转串—串行传输—在接收端串转并，实现芯片间信号的有线传输。相比于传统并行接口传输，SerDes具有更高的速率（Gbps级）、更低的功耗，以及显著的成本优势，能够满足AI训练&推理等场景下高带宽、低延迟的数据传输要求，适用于电信、汽车、工业等领域。

SerDes结构



数据来源：CSDN

SerDes汽车电子应用场景



数据来源：罗姆半导体

交换芯片：AI服务器对SerDes提出更高性能要求

- **AI服务器算力提升显著，带动SerDes通信带宽需求激增。**AI服务器网络模块升级主要表现为带宽增加，主要涉及的芯片是SerDes。高性能计算机性能提升主要源于单个结点计算能力增强和系统中结点数增加。一般而言，结点对互连带宽的要求与其处理能力呈正比。随着结点计算能力迅速提高，系统对互连网络带宽的需求更加迫切。
- 以"天河二号"为例，单个计算结点性能3TFlops，结点通信带宽112Gbps，结点带宽性能比0.037。在未来 E级高性能计算机中，单结点计算能力约为10TFlps，若要将结点带宽性能比维持在0.04，则结点通信带宽需增加至400Gbps，对SerDes芯片性能提出了更高的要求。通过交替增加SerDes通道数量和每个通道传输速率，实现容量翻倍。

SerDes技术演进

年份	2010	2012	2014	2016	2018	2020	2020E	2022E
SerDes数量	64	128	128	256	256	512	256	512
SerDes速率（Gbps）	10	10	25	25	50	50	100	100
调制方式	NRZ	NRZ	NRZ	NRZ	PAM4	PAM4	PAM4	PAM4
容量（Tb/s）	0.64	1.28	3.2	6.4	12.8	25.6	25.6	51.2

数据来源：IET Optoelectronics ，国泰君安证券研究

交换芯片：AI服务器对SerDes提出更高性能要求

- **国内SerDes发展现状：**
- 上游：半导体IP龙头芯原股份于2021年2月获得加拿大高速接口领域全球领导者Alphawave公司在国内一系列多标准SerDes IP的独家经销权。
- 中游：芯片设计领域，专注于高速混合芯片设计公司龙讯股份基于单通道12.5Gbps SerDes技术研发的通用高速信号延长芯片可在5G通信领域实现国产化应用。国产以太网PHY芯片龙头裕太微研究形成高性能SerDes技术，可实现1.25~5G等不同数据率，应用于多款量产产品；10G SerDes现已通过实验室性能测试，在FR4电路板上传输距离长达40英寸，适配以太网、PCIE等多种上层协议。

高速 SerDes技术和各种接口的关系

数据来源：IP与Soc设计

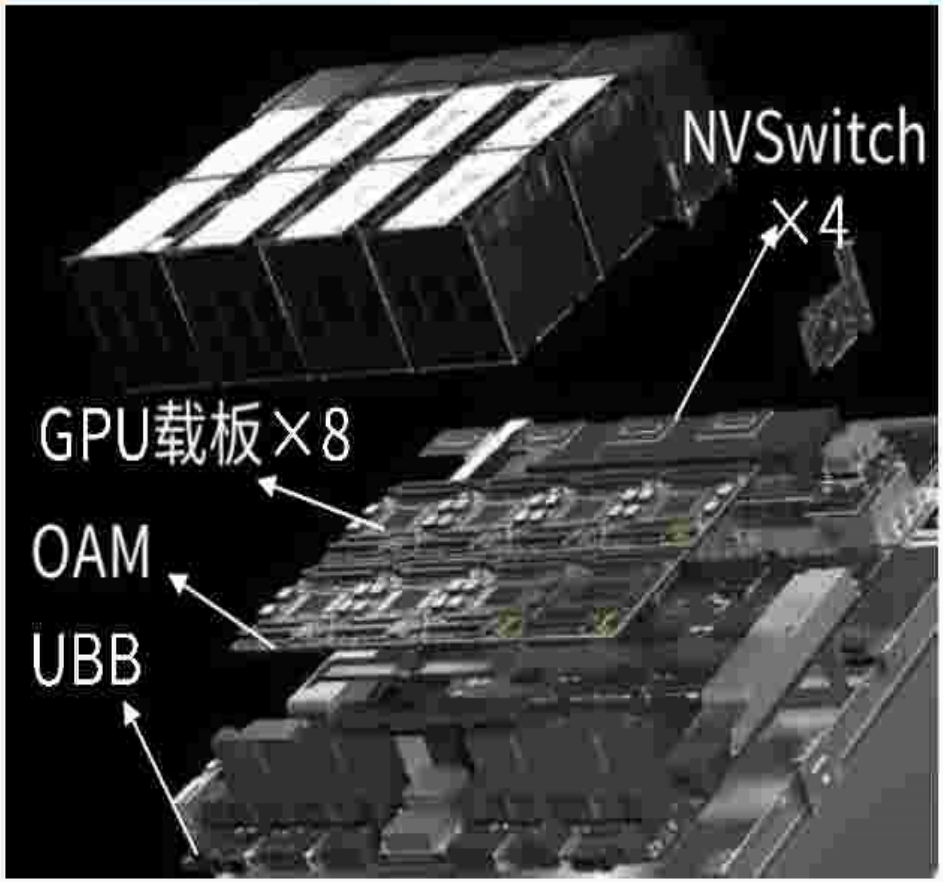
全球SerDes市场规模（单位：亿美元）

数据来源：iCroudNewswire，国泰君安证券研究

- **相比于普通服务器，AI服务器的PCB增量主要体现在GPU板组中。**由于AI服务器比普通服务器的GPU用量从1颗提升到8颗（参考英伟达DGX H100），相应GPU板组的PCB需求量增大且要求进一步提高。

NVIDIA DGX H100 爆破图

NVIDIA DGX H100 GPU板组爆破图



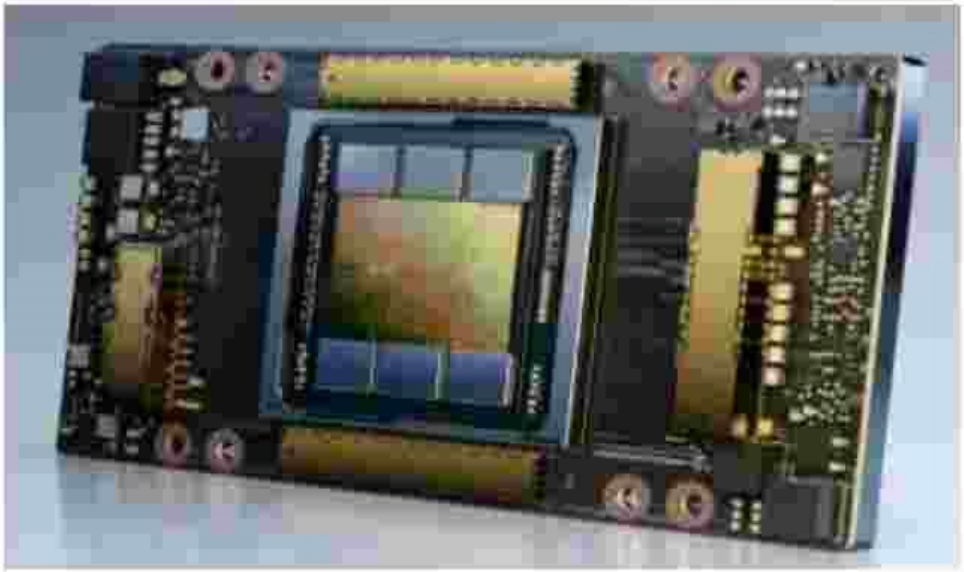
数据来源：NVIDIA官网

数据来源：NVIDIA官网

PCB及载板：规格加速升级，单机ASP持续上升

- OAM卡，是承载GPU加速卡片的PCB板，以NVIDIA DGX H100服务器为例，其可搭载8颗GPU，显著高于普通服务器，因此其PCB用量显著高于普通服务器。
- **AI服务器的OAM卡需要用更高层数的PCB板，价值量更大。**由于AI服务器电路更加复杂，需要更大带宽和更高传输速率，因此OAM需要更高层数PCB。NVIDIA OAM共两个版本，SXM约需要20层PCB，而Pcie版本层数相对较少；相比传统服务器，AI服务器的PCB层数更高，单台PCB价值量大幅提升。
- **AI服务器的OAM由于芯片性能的提升，对布线密度提出了更高要求。**其需要4阶及以上HDI加工工艺，根据靖邦电子，HDI板增加一阶，成本增加18%左右，因此带动OAM的ASP上升。
- 目前国内企业鹏鼎控股、沪电股份、奥士康、胜宏科技等均有领先布局。

NVIDIA DGX H100 OAM 实物图

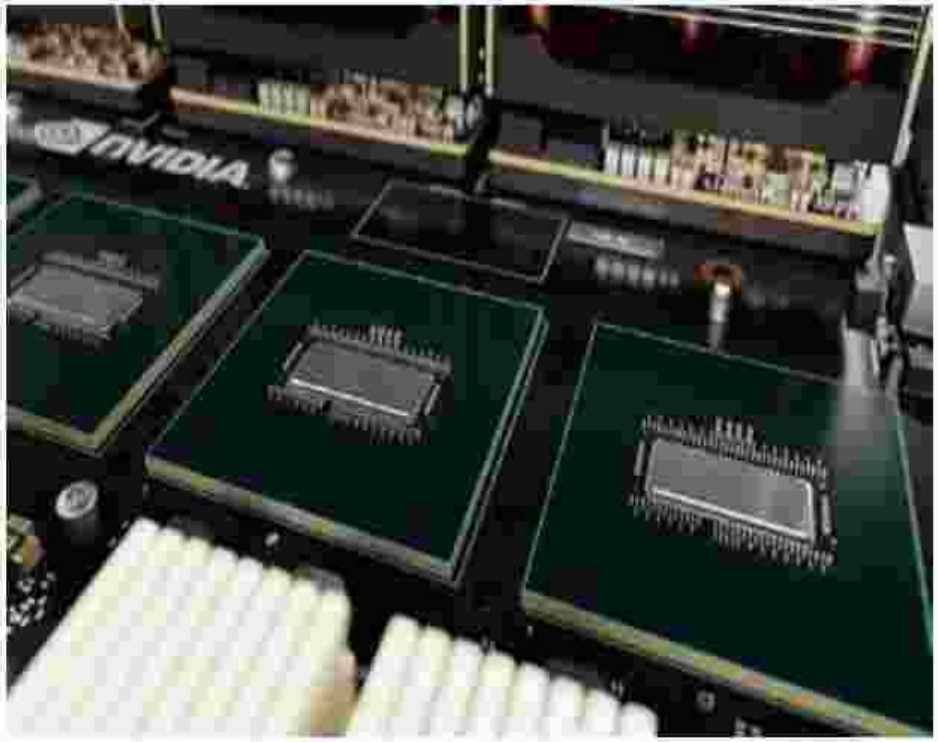


数据来源：NVIDIA官网技术文件

PCB及载板：规格加速升级，单机ASP持续上升

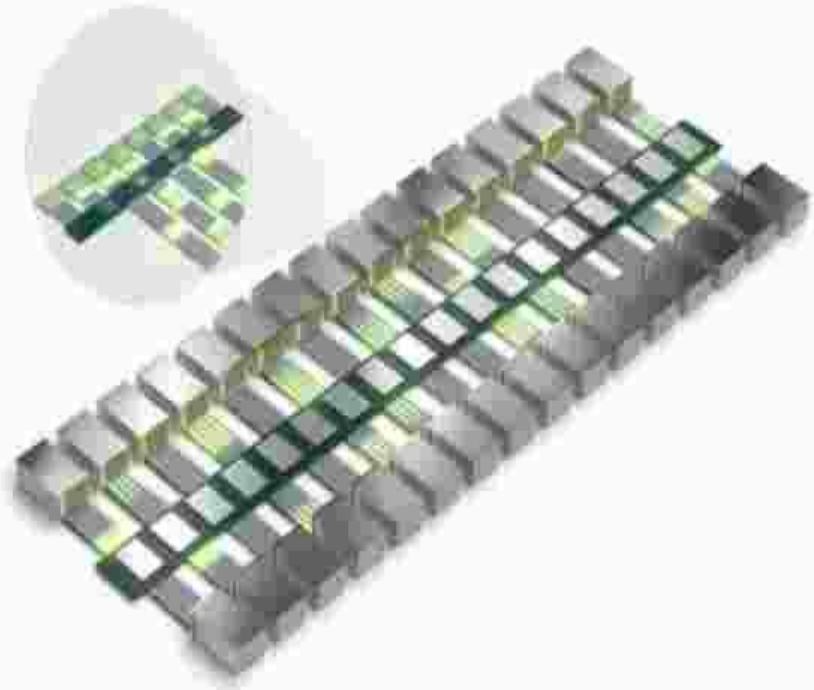
- **NVSwitch是GPU之间的通信模块**，在单节点内和节点间实现以 NVLink 能够达到的最高速度进行多对多 GPU 通信。NVIDIA DGX H100的GPU板组包含4个NVSwitch。
- **NVSwitch对PCB板的高速传输有更高要求。**根据立鼎产业研究院，NVIDIA DGX H100在4个NVSwitch加持下总带宽达到PCIe5.0的7倍，意味着其覆铜板材料至少需要使用Ultra Low Loss级别。

NVIDIA DGX H100 NVSwitch 实物图



数据来源：NVIDIA官网

NVIDIA DGX H100 NVSwitch 概念图

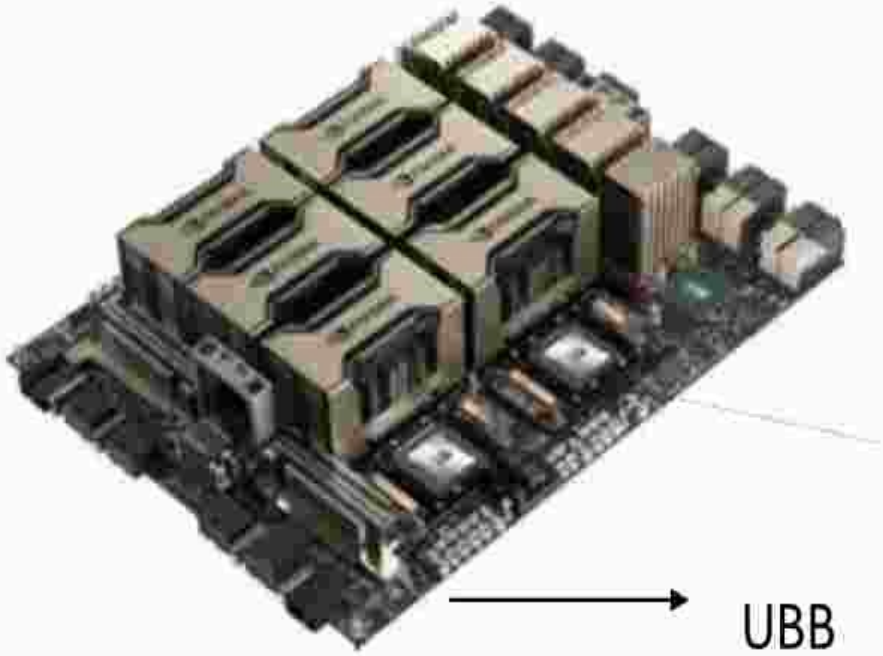


数据来源：NVIDIA官网

PCB及载板：规格加速升级，单机ASP持续上升

- **UBB是GPU模组板，用以承载OAM、NVSwitch等模块**，面积随GPU、NVSwitch数量增加而增大。
- 由于其上集成部件较多，布线较为复杂，**通常需要24~26层的超高层PCB板**，ASP提升较大。
- 作为搭载整个GPU板组的板块，**对于高频高速具有较高的要求，需要使用的覆铜板等级为Ultra Low Loss。**NVIDIA DGX H100的UBB设计更为紧凑，使用HDI技术会进一步提高ASP。
- 目前沪电股份在超高层和高密度PCB中优势领先，未来深度受益于AI服务器需求释放。

NVIDIA DGX H100 UBB 实物图



数据来源：NVIDIA官网

- **从覆铜板（CCL）技术升级角度，将目前最新的Intel Eagle Stream平台与前代平台对比，可明显看出服务器平台用覆铜板升级处于一个阶梯跨越至另一个阶梯的关键转型期。**最新的Eagle Stream平台要求CCL的介电损失因数Df达到0.002-0.004，介电常数Dk达到3.3-3.6。

Intel 不同平台CCL性能对比

服务器平台升级要求传输速率提高，Dk与Dr值下降

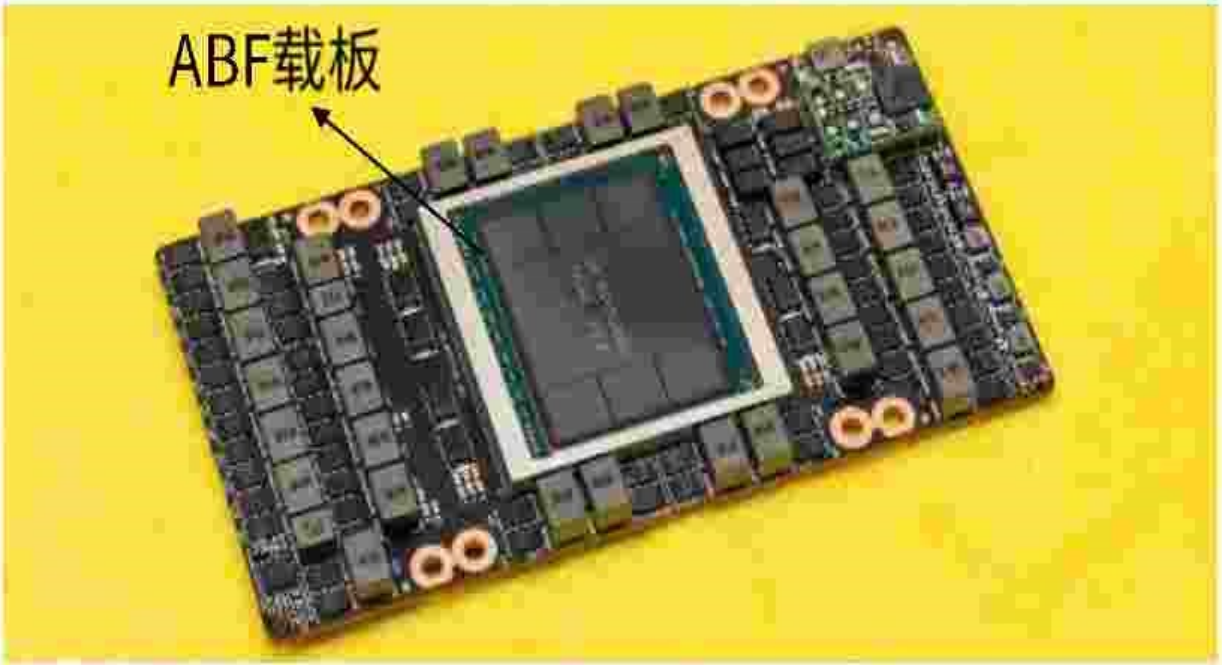
项目	Grantley平台	Purley平台	Whitley平台	Eagle Stream
传输速率(Gbps)	28及以下	28	56	112
高速覆铜板类型	Mid-loss	Mid-loss	Low-loss	Ultra-Low-loss
典型Dk值	4.1-4.3	4.1-4.3	3.7-3.9	3.3-3.6
典型Df值	0.008-0.010	0.008-0.010	0.005-0.008	0.002-0.004
对标松下电工产品型号(注)	M4以下	M4以下	M4及以上	M6及以上

数据来源：Prismark，国泰君安证券研究

PCB及载板：规格加速升级，单机ASP持续上升

- **H100等GPU的芯片封装通常使用2.5D/3D封装技术，而ABF载板是2.5D/3D封装的核心材料之一。**随着GPU需求持续走高，**ABF载板需求也相应增加。**并且由于AI大模型要求多张GPU之间相互通信形成超级CPU，对ABF载板的**高频高速需求更高，ABF载板量价齐升。**
- **从产业链看，**上游ABF薄膜基本由日本味之素垄断，中游制造有日本的Ibiden、 Shinko, 韩国的Semco，中国台湾的Kinsus、Unimicron等。**在ABF载板市场，中国台湾、日本、韩国总份额高达80%，大陆内资厂2020年全球占比仅为5.3%。**目前国内深南电路、兴森科技等产品逐步突破，未来有望深度受益于国产替代。

H100 TENSOR CORE GPU 实物图



数据来源：CNET

AI服务器：需求加速爆发，云服务商是采购主力

算力时代来临，高ASP的AI服务器需求加速爆发。传统服务器数据处理能力有限，AI服务器通过异构形式实现并行计算，可以更好满足算力要求。相较于普通服务器，AI服务器采用多芯片组合，算力硬件成本更高，品牌与白牌组装市场有望进一步扩容。

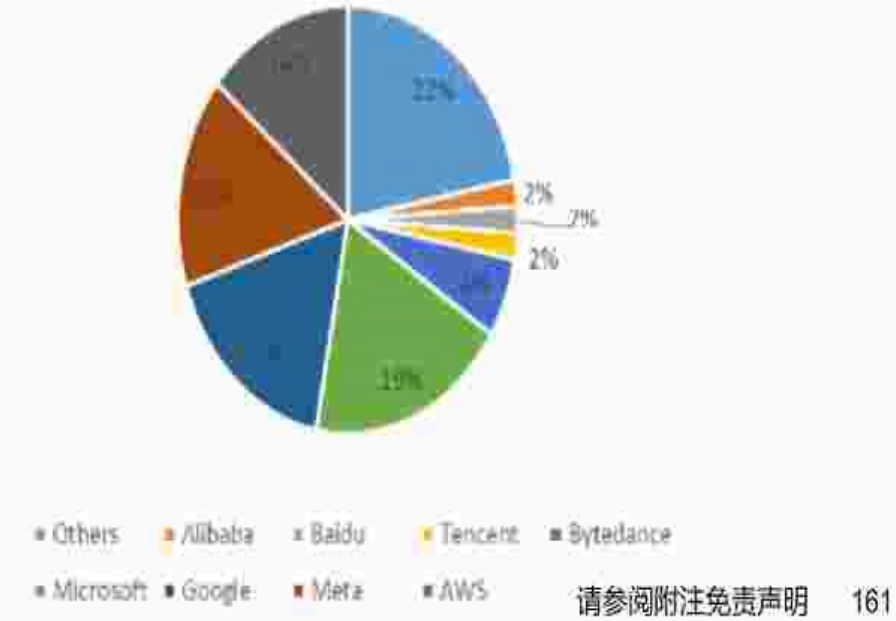
AI服务器需求量加速增长，2026年将超过200w台。根据TrendForce数据，2022年全球AI服务器出货量为85万台，预计2023年AI服务器出货量将增长到118.3万台，同比增长38.4%，预计到2026年出货量将达到236.9万台。

云服务厂商是AI服务器主要的采购主力。根据TrendForce，2022年北美四大云端厂商谷歌、亚马逊AWS、Meta、微软合计占据全球AI服务器采购量的66.2%，国内互联网/云计算厂商字节跳动、腾讯、百度、阿里巴巴分别占比6%、2%、2%、2%。

2022-2026全球AI服务器出货量预估（单位：千台）

数据来源：TrendForce，国泰君安证券研究

2022全球AI服务器市场客户采购结构



数据来源：TrendForce，国泰君安证券研究

AI服务器：量价齐升，组装市场进一步扩容

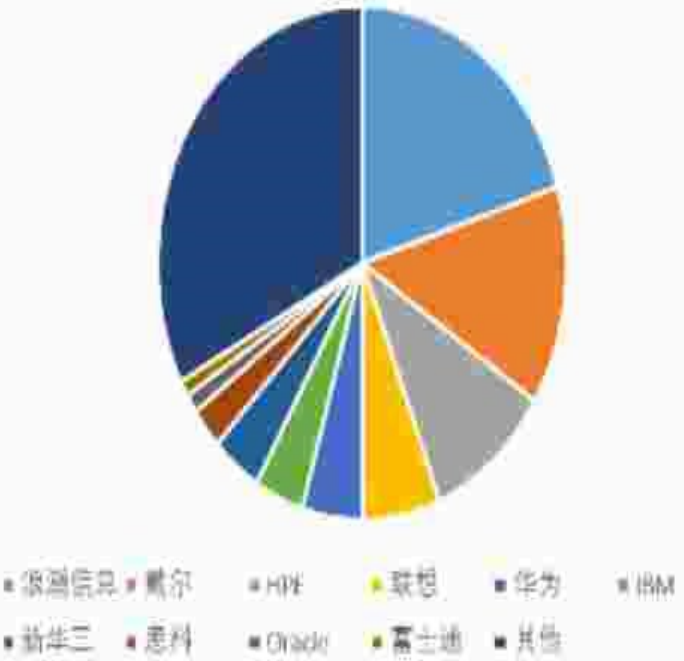
相较于普通服务器，AI服务器在规格以及价格方面均迎来升级，整机组装市场有望进一步扩容，且技术壁垒将同步提升。

- 总体服务器市场：根据Counterpoint，2021年全球服务器市场中排名前三的品牌分别为戴尔、HPE及浪潮，份额分别为14.5%、12.4%及10.1%。ODM厂商销售规模302.3亿美元，占比达31.1%，其中工业富联销售规模最大，在ODM市场占比42%。
- AI服务器市场：参照2022年上半年全球AI服务器市场中，浪潮、戴尔、惠普分别以20.2%、13.8%、9.8%位列前三。工业富联则在AI服务器ODM领域份额一家独大。

2021全球服务器供应商收入（单位：百万美元）

数据来源：Counterpoint

2022H1全球AI服务器市场份额占比统计



数据来源：IDC，国泰君安证券研究

07 重点公司盈利预测与估值

公司代码	公司名称	股价(元)	每股收益(元)			市盈率(X)			投资评级
			2023E	2024E	2025E	2023E	2024E	2025E	
000475	立讯精密	30.18	1.54	1.91	2.30	20	16	13	增持
688981	中芯国际	52.64	0.75	1.06	1.41	70	50	37	增持
603501	韦尔股份	103.83	6.09	7.59	9.31	17	14	11	增持

资料来源：wind，国泰君安证券研究（收盘价参考2023年11月22日，盈利预测来自国君电子外发报告）

08 风险提示

- **下游需求恢复不及预期：**虽新兴领域AI及智能终端等的需求较为旺盛，但消费电子领域的部分需求并未看到强劲复苏，若需求恢复不及预期，则会对产业链的公司业绩造成影响。

- **AI产业进度不及预期：**AI产业化目前正快速发展，带动产业链相关公司的需求上行，若需求阶段性下滑或者产业化进度不及预期，则可能会影响相关公司业绩。

- **国际贸易风险：**由于中国终端产品出口较多，若未来发生贸易纠纷或者政策变化等，有可能会对公司业绩产生不利影响。

THANKS FOR LISTENING

国泰君安证券研究所电子团队

免责声明

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰君安证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

投资建议的比较标准		评级	说明
投资评级分为股票评级和行业评级。	股票投资评级	增持	相对沪深300指数涨幅10%以上
以报告发布后的12个月内的市场表现为比较标准，报告发布日后的12个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深300指数涨跌幅为基准。		谨慎增持	相对沪深300指数涨幅介于5%～10%之间
		中性	相对沪深300指数涨幅介于-5%～5%
		减持	相对沪深300指数下跌5%以上
	行业投资评级	增持	行业指数相对于沪深300指数超配
		中性	行业指数相对于沪深300指数标配
		减持	行业指数相对于沪深300指数低配

国泰君安证券研究所 E-mail：gtjaresearch@gtjas.com		
上海	地址	上海市静安区新闸路669号博华广场20层
	邮编	200041
	电话	（021）38676666
深圳	地址	深圳市福田区益田路6003号荣超商务中心B栋27层
	邮编	518026
	电话	（0755）23976888
北京	地址	北京市西城区金融大街甲9号 金融街中心南楼18层
	邮编	100032
	电话	（010）83939888