浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

ChatGPT芯片算力：研究框架

——【AIGC算力时代系列报告】

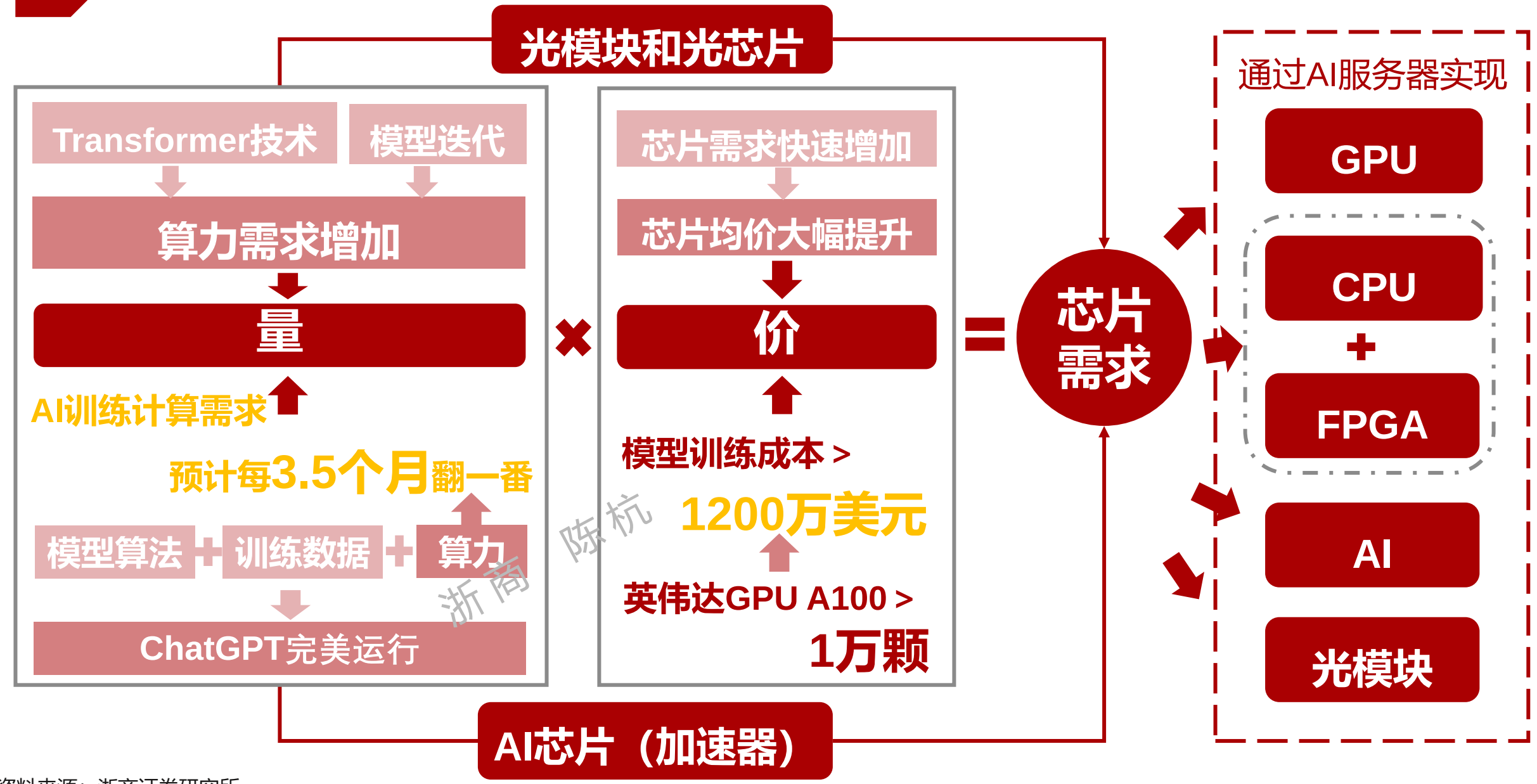
行业评级：看好

2023年2月10日

分析师 陈杭
邮箱 chenhang@stocke.com.cn
证书编号 S1230522110004

研究助理 安子超
邮箱 anzichao@stocke.com.cn

算力需求爆发拉动芯片量价齐升



资料来源：浙商证券研究所

摘要

ChatGPT热潮席卷全球。ChatGPT（Chat Generative Pre-trained Transformer）是由OpenAI于2022年12月推出的对话AI模型，一经面世便受到广泛关注，其2023年1月月活跃用户达到1亿，是史上月活用户增长最快的消费者应用。在问答模式的基础上ChatGPT可以进行推理、编写代码、文本创作等等，这样的特殊优势和用户体验使得应用场景流量大幅增加。

1、▲芯片需求=量↑ x 价↑， AIGC拉动芯片产业量价齐升

1）量：AIGC带来的全新场景+原场景流量大幅提高。① 技术原理角度：ChatGPT基于Transformer技术，随着模型不断迭代，层数也越来越多，对算力的需求也就越来越大；② 运行条件角度：ChatGPT完美运行的三个条件：训练数据+模型算法+算力，需要在基础模型上进行大规模预训练，存储知识的能力来源于1750亿参数，需要大量算力。

2）价：对高端芯片的需求将拉动芯片均价。采购一片英伟达顶级GPU成本为8万元，GPU服务器成本通常超过40万元。支撑ChatGPT的算力基础设施至少需要上万颗英伟达GPU A100，高端芯片需求的快速增加会进一步拉高芯片均价。

2、ChatGPT的“背后英雄”：芯片，看好国内GPU、CPU、FPGA、AI芯片及光模块产业链

1）GPU：支撑强大算力需求。由于具备并行计算能力，可兼容训练和推理，目前GPU被广泛应用于加速芯片。看好**海光信息、景嘉微**；

2）CPU：可用于推理/预测。AI服务器利用CPU与加速芯片的组合可以满足高吞吐量互联的需求。看好**龙芯中科、中国长城**；

3）FPGA：可通过深度学习+分布集群数据传输赋能大模型。FPGA具备灵活性高、开发周期短、低延时、并行计算等优势。看好**安路科技、复旦微电、紫光国微**；

4）ASIC：极致性能和功耗表现。AI ASIC芯片通常针对AI应用专门设计了特定架构，在功耗、可靠性和集成度上具有优势。看好**寒武纪、澜起科技**；

5）光模块：容易被忽略的算力瓶颈。伴随数据传输量的增长，光模块作为数据中心内设备互联的载体，需求量随之增长。看好**德科立、天孚通信、中际旭创。**

风险提示

1、AI技术发展不及预期

2、版权、伦理和监管风险

3、半导体下游需求不及预期

目录

CONTENTS

01

算力需求爆发拉动芯片量价齐升

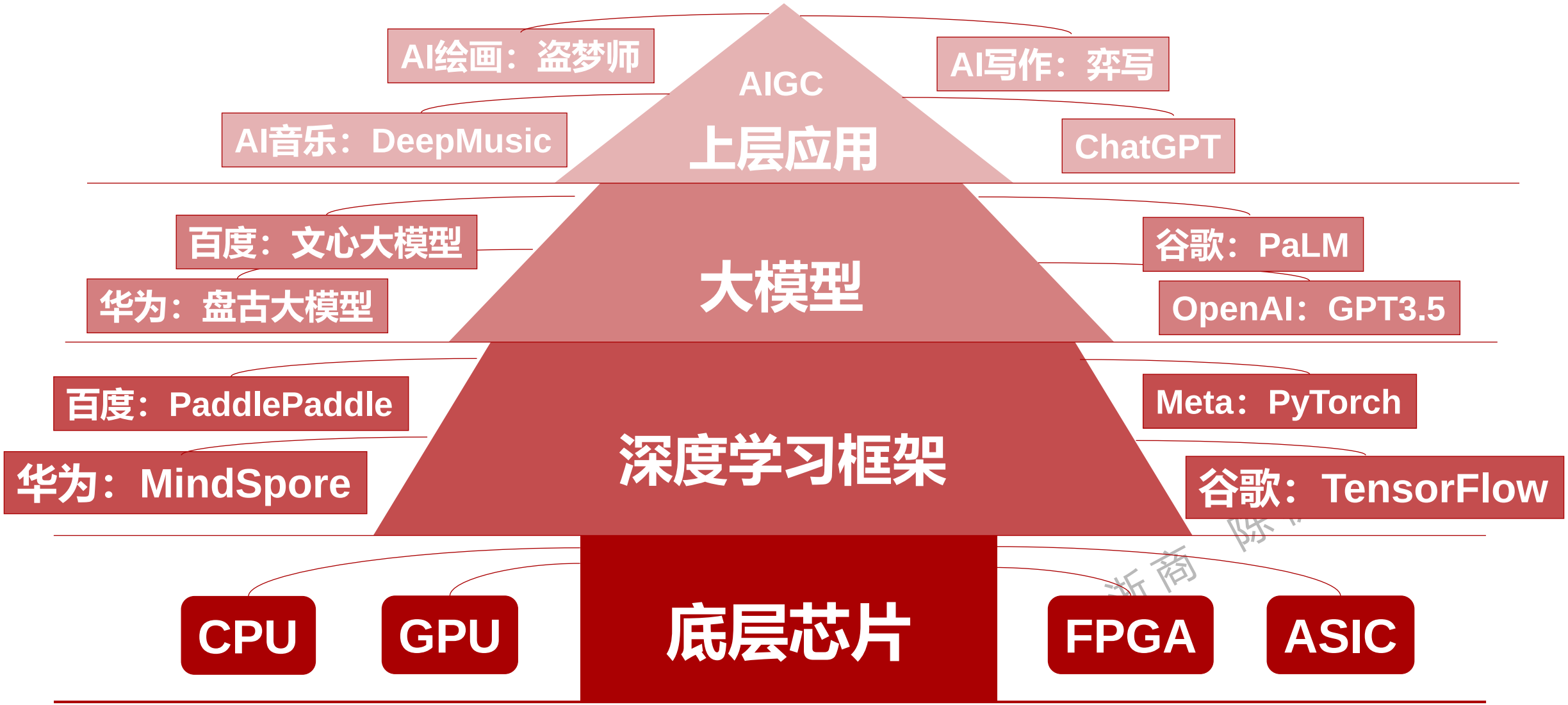
AI计算需要各类芯片支撑

算力需求爆发，芯片量价齐升

AI服务器为算力载体

CPU、GPU、FPGA、ASIC、光模块各司其职

人工智能四层架构，芯片为底层支撑



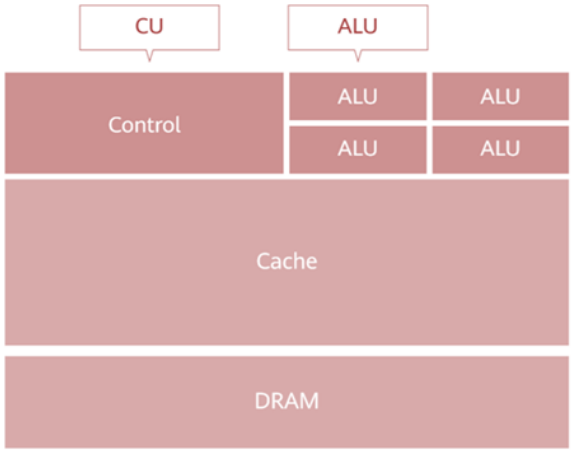
资料来源：澜舟科技，IDC，量子位，电子发烧友等，浙商证券研究所

人工智能不同计算任务需要各类芯片实现

通用性强，应用方便

- 强大的调度、管理、协调能力；
- 应用范围广
- 开发方便灵活

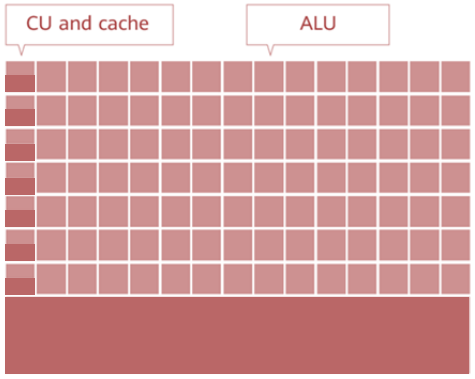
CPU



✓ 逻辑判断
✓ 任务调度与控制

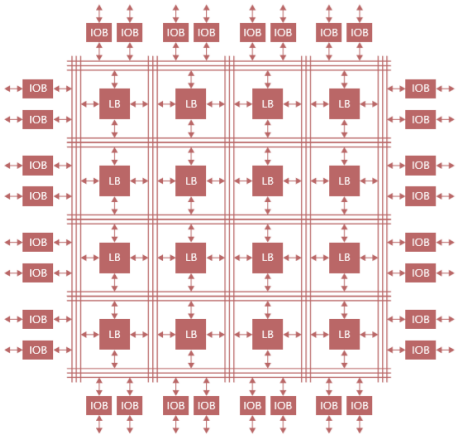
- 并行架构
- 计算单元多
- 适合大量逻辑确定的重复计算

GPU



✓ 模型训练

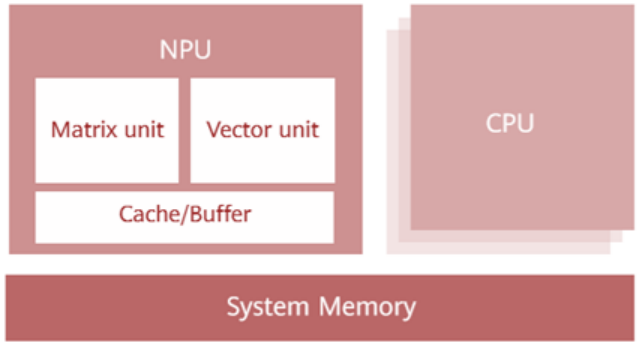
- 低延时
- 开发周期短
- 硬件可根据需求调整
- 成本和壁垒高



FPGA

✓ 研发阶段
✓ 数据中心
✓ AI推理

- 成本低
- 能耗低
- 性能强
- 针对AI设定特定架构



AI用ASIC

✓ 成熟量产阶段

性能更优，能效更高

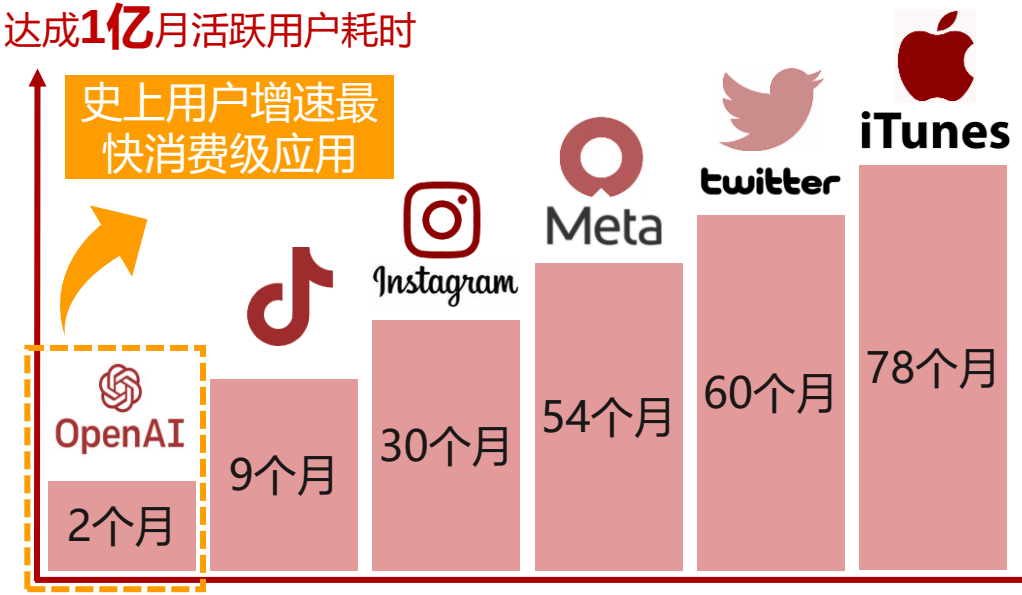
资料来源：Huawei forum，浙商证券研究所

01 ChatGPT流量激增，为AI服务器带来重要发展机遇

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

原场景流量提升+新应用场景 → **服务器算力要求提升** → **AI服务器需求增加**

达成1亿月活跃用户耗时



终端用户使用频率提高，数据流量暴涨，对服务器的数据处理能力、可靠性及安全性等要求相应提升

数据的质和量发生变化，非结构化数据占比激增

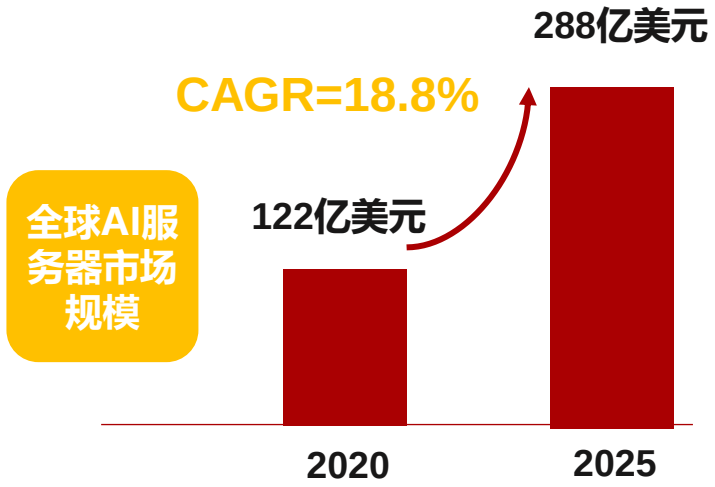
传统CPU服务器
通用性较强，专用性较弱

↓

算力无法满足

↓

AI服务器需求



原场景流量提升

ChatGPT在问答模式的基础上进行推理、编写代码、文本创作等，用户人数及使用次数均提升。

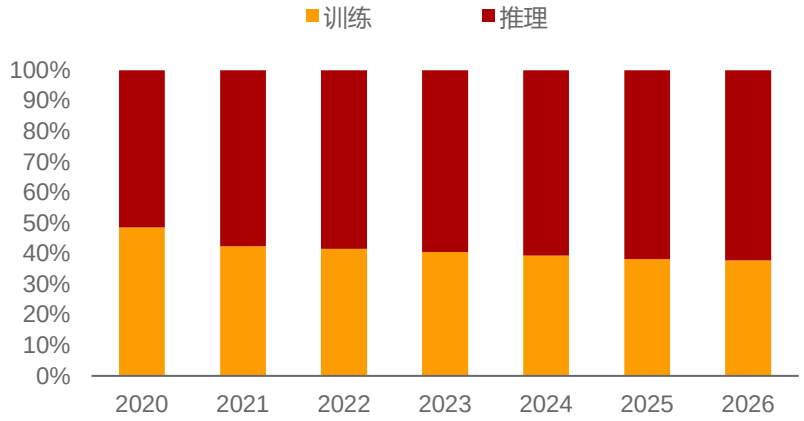
创造新应用场景

- 智能客服
- 智能音箱
- 内容生产
- 游戏NPC
- 陪伴型机器人
- ……

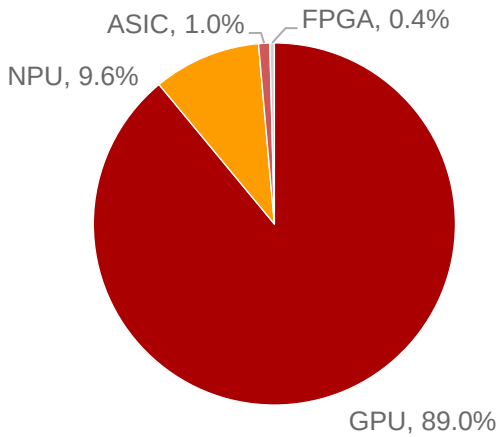
资料来源：Sensor Tower，World of Engineering，头豹研究院，华经产业研究院，浙商证券研究所

AI服务器快速增长，大力拉动芯片需求

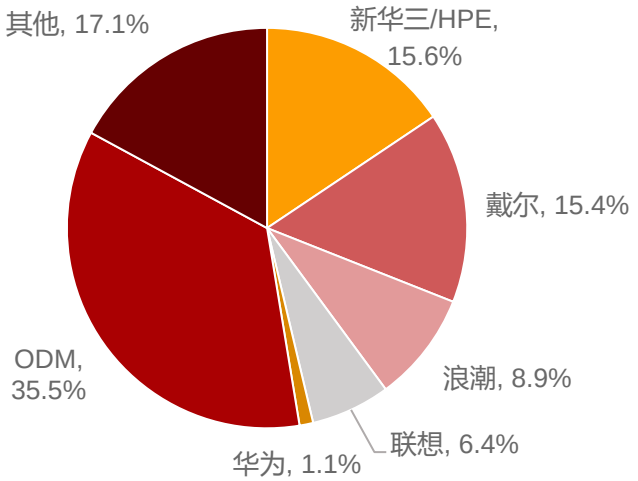
中国人工智能服务器工作负载预测



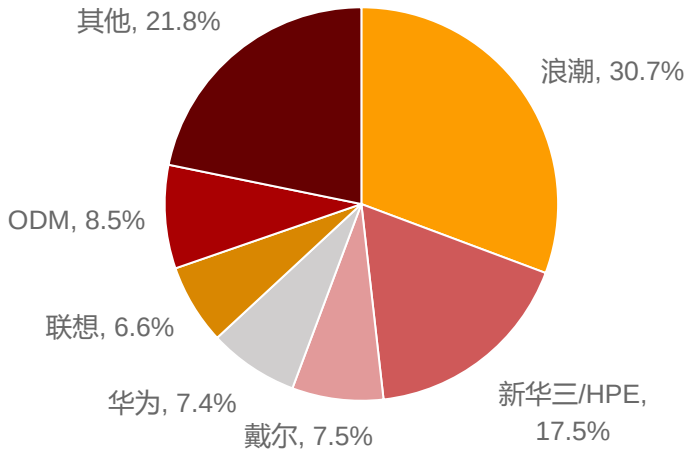
中国人工智能芯片市场规模占比



2021年全球服务器市场格局



2021年中国服务器市场格局



AI服务器=？

异构形式

CPU

+

GPU

或

FPGA

或

ASIC

资料来源：IDC，《2022-2023中国人工智能计算力发展评估报告》，芯八哥，wind，浙商证券研究所整理

AI服务器芯片构成——CPU+加速芯片

AI服务器 → 应用领域 → 应用场景

CPU+加速芯片：通常搭载GPU、FPGA、ASIC等加速芯片，利用CPU与加速芯片的组合可以满足高吞吐量互联的需求

计算机视觉　**机器学习**

自然语言处理

芯片种类	优点	缺点
GPU	提供了多核并行计算的基础结构，核心数多，可支撑大量数据的并行计算，拥有更高浮点运算能力	管理控制能力弱，功耗高
FPGA	可以无限次编程，延时性较低，拥有流水线并行（GPU只有数据并行），实时性最强，灵活性最高	开发难度大，只适合定点运算，价格比较昂贵
ASIC	与通用集成电路相比体积更小，重量更轻，功耗更低，可靠性提高，性能提高，保密性增强，成本降低	灵活性不够，价格高于FPGA

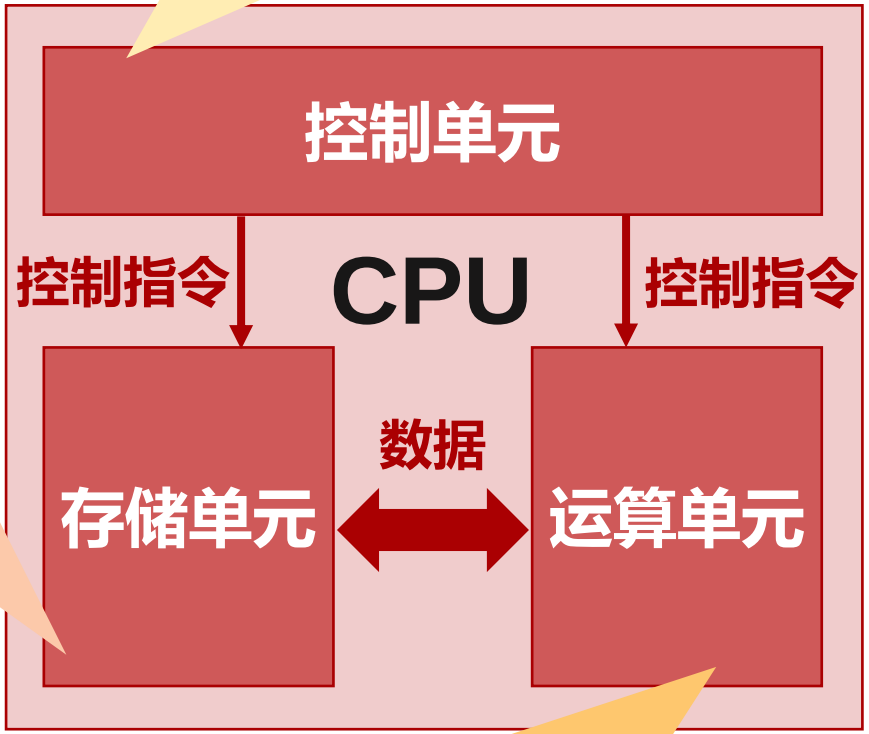
高度适配

- **训练**（算力 >30TOPS）
- **推理**（算力 >30TOPS）
- **计算**（5~30TOPS）

资料来源：华经产业研究院，亿欧智库，浙商证券研究所

01 CPU擅长逻辑控制，可用于推理/预测

整个CPU的指挥控制中心，由指令寄存器IR、指令译码器ID和操作控制器OC等组成。



暂时存放数据的区域，保存等待处理或已经处理过的数据。

执行部件，运算器的核心。可以执行算术运算和逻辑运算。运算单元所进行的全部操作都是由控制单元发出的控制信号来指挥。

CPU运行原理

取指令 → 指令译码 → 执行指令 → 修改指令计数器

作为计算机系统的**运算和控制核心**，是**信息处理、程序运行**的最终**执行单元**。

优势	有大量的缓存和复杂的逻辑控制单元，擅长逻辑控制、串行的运算。
劣势	计算量较小，且不擅长复杂算法运算和处理并行重复的操作。

在深度学习中可用于**推理/预测**

资料来源：CSDN，浙商证券研究所整理

01 服务器CPU向多核心发展，满足处理能力和速度提升需要

单核心CPU

串行单任务处理 → 分时多任务处理

“一心一用” → “一心多用”

处理的任务更多、处理速度更快

多核心CPU

分时多任务处理

“多心多用”

系统性能优劣不能只考虑CPU核心数量，还要考虑**操作系统、调度算法、应用和驱动程序**等。

英特尔

从单核到多核

2005	2010	2017	2020	2023
奔腾D系列	酷睿i7-980X	酷睿i9	Lakefield	Sapphire Rapids
史上第一个双核处理器	首款6核处理器	18核处理器	首款采用混合架构的x86 5核处理器	拥有56个核心

AMD

从双核到96核

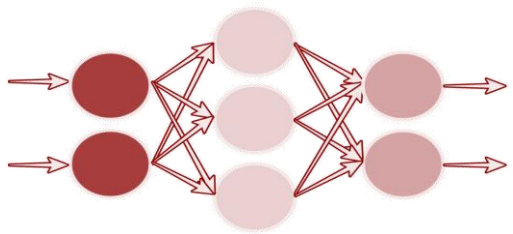
2005	2007	2018	2020	2023
Athlon 64 X2	Phenom9500	第二代锐龙Threadripper	锐龙Threadripper 3990X	霄龙9004
同一块芯片内整合两个K8核心	首款原生4核处理器	最大核心数量已达到32核	拥有64核	核心数量最多可达96个

资料来源：芯论语，半导体产业纵横，AMD官网，浙商证券研究所

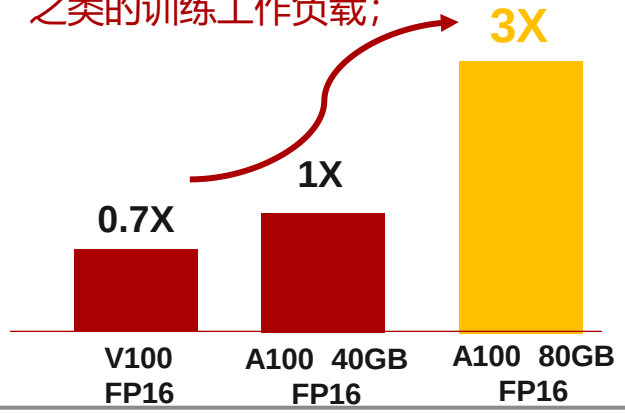
AI模型构建（以英伟达A100为例）

训练过程

GPU的并行计算高度适配神经网络

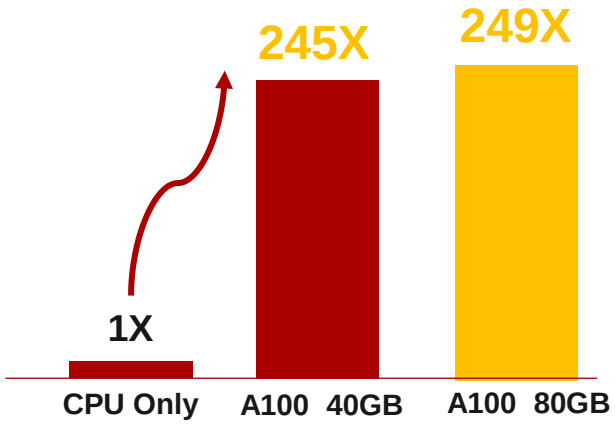


- **GPU帮助高速解决问题：**2048 个 **A100** GPU 可在一分钟内成规模地处理 BERT 之类的训练工作负载；



推理过程

- 多实例 GPU (MIG) 技术允许多个网络同时基于单个 A100 运行，从而优化计算资源的利用率。
- 在 A100 其他推理性能增益的基础之上，仅结构稀疏支持一项就能带来高达两倍的性能提升。
- 在 BERT 等先进的对话式 AI 模型上，**A100** 可将推理吞吐量提升到高达 CPU 的 249 倍；



ChatGPT引发GPU热潮

百度：即将推出文心一言（ERNIE Bot）

苹果：引入AI加速器设计的M2系列芯片（M2 pro和M2 max）将被搭载于新款电脑

OpenAI：随着ChatGPT的使用量激增，OpenAI需要更强的计算能力来响应百万级别的用户需求，因此增加了对英伟达GPU的需求

AMD：计划推出与苹果M2系列芯片竞争的台积电4nm工艺"Phoenix"系列芯片，以及使用Chiplet工艺设计的"Alveo V70"AI芯片。这两款芯片均计划在今年推向市场，分别面向消费电子市场以及AI推理领域

资料来源：英伟达官网，镁客网，腾讯网，浙商证券研究所

FPGA：可通过深度学习+分布集群数据传输赋能大模型

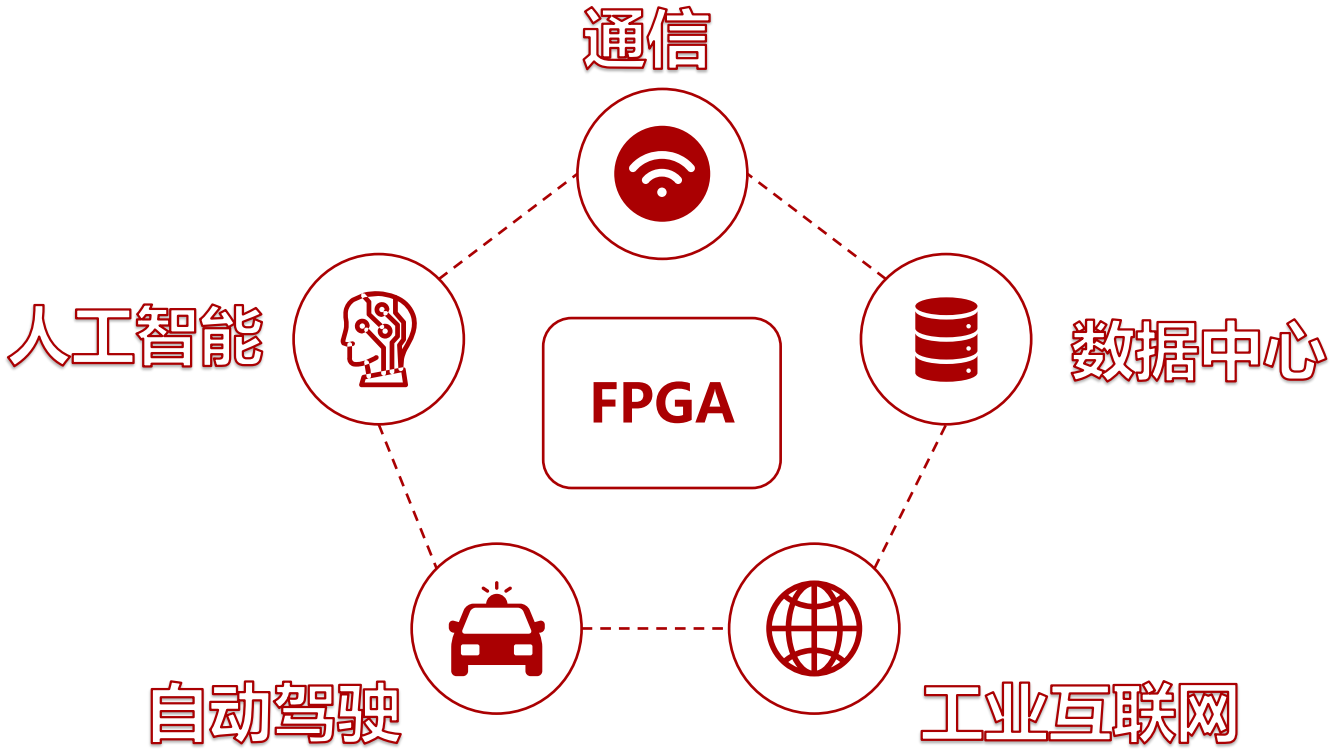
可编程灵活性高：半定制电路，理论上可以实现任意ASIC和DSP的逻辑功能

开发周期短：可通过设计软件处理布线、 布局及时序等问题。

现场可重编功能：可以远程通过软件实现自定义硬件功能。

低延时：逻辑门通过硬件线连接，不需要时钟信号

方便并行计算：集成了大量基本门电路，一次可执行多个指令算法



深度学习	异构计算、并行计算
通信接口	数据高速收发、交换

	数据中心	边缘端
推理	Intel，AMD (Xilinx)，亚马逊，微软，百度，阿里，腾讯	AMD (Xilinx)
训练	Intel，AMD (Xilinx)	/

资料来源：ofweek，摩尔星球，C114，知识产权课堂，亿欧智库，浙商证券研究所

ASIC可进一步优化性能与功耗，全球巨头纷纷布局

国内外ASIC芯片龙头布局

随着机器学习、边缘计算、自动驾驶的发展，大量数据处理任务的产生，对于芯片计算效率、计算能力和计能耗比的要求也越来越高，**ASIC通过与CPU结合的方式被广泛关注**，国内外龙头厂商纷纷布局迎战AI时代的到来。

国外

谷歌：张量处理器——TPU

- 最新的TPU v4集群被称为Pod，包含4096个v4芯片，可提供超过1 exaflops的浮点性能

英伟达：GPU+CUDA

- 主要面向大型数据密集型 HPC 和 AI 应用；
- 基于 Grace 的系统与 NVIDIA GPU 紧密结合，性能比NVIDIA DGX 系统高出 10 倍；

Habana （Intel收购）

- 已推出云端 AI 训练芯片 Gaudi 和云端 AI 推理芯片 Goya；

国内

阿里巴巴：含光800AI芯片

- 硬件：自研芯片架构；
- 软件：集成达摩院先进算法，可实现大网络模型在一颗NPU上完成计算。

百度：昆仑2代AI芯片

- 采用全球领先的7nm 制程，搭载自研的第二代 XPU 架构，相比一代性能提升2-3倍；
- 昆仑芯3代将于2024年初量产。

华为：昇腾910

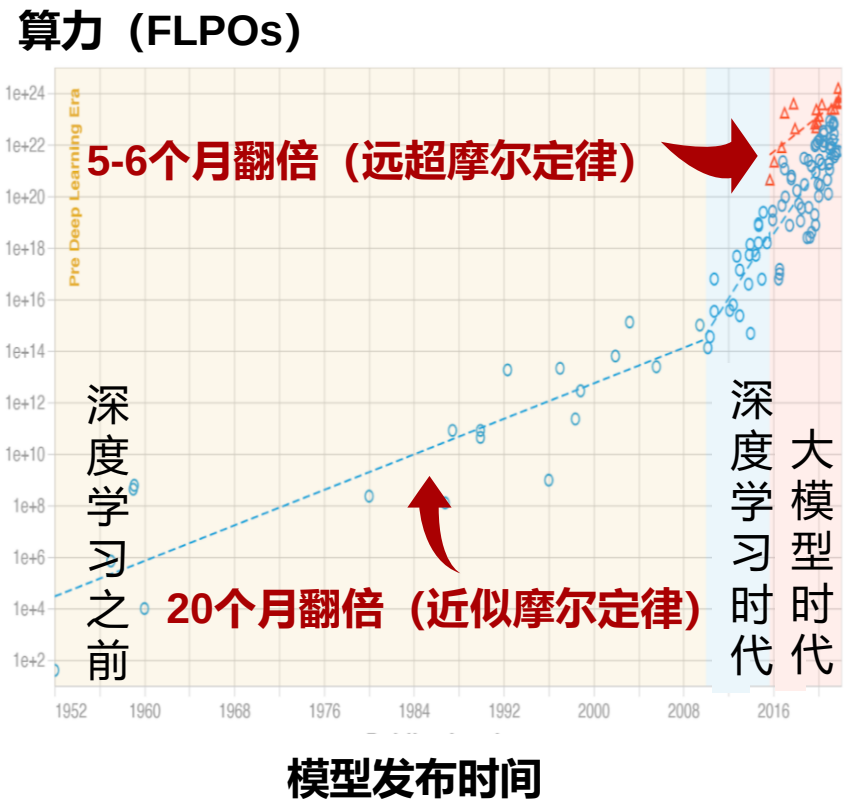
- 业界算力最强的AI处理器，基于自研华为达芬奇架构3D Cube技术；

资料来源： 机器之心，华为官网，半导体产业纵横，公开资料整理，浙商证券研究所

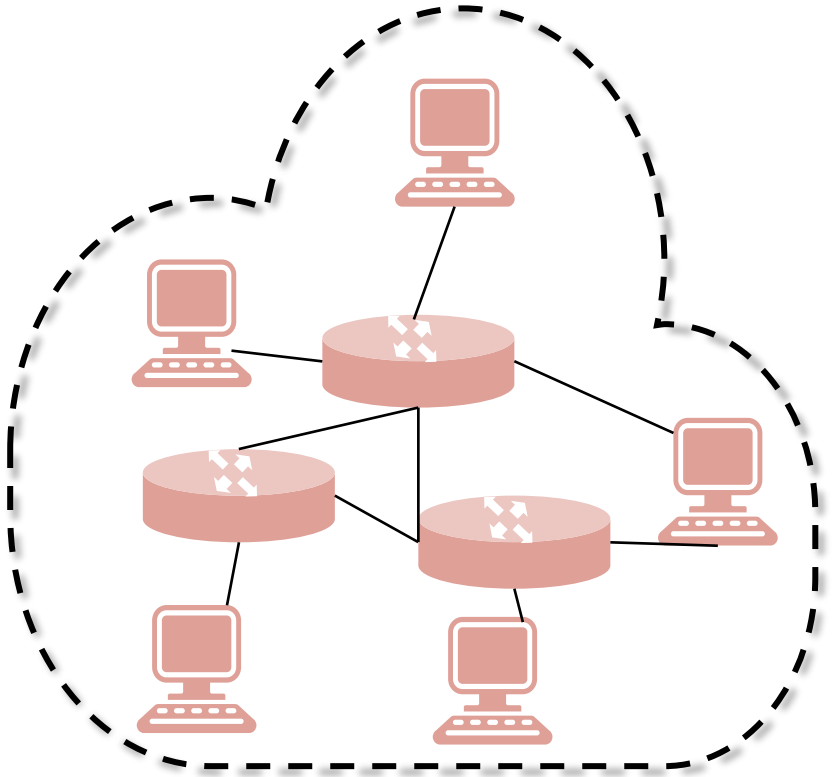
数据传输速率：容易被忽略的算力瓶颈

算力需求：超摩尔发展 → **算力供给：芯片提升+并行计算** → **并行瓶颈：数据传输速率**

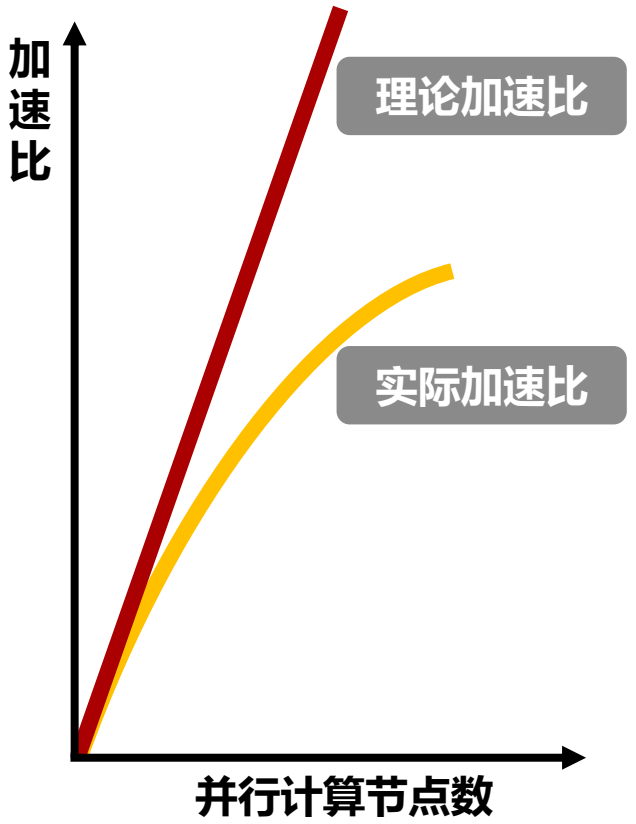
AI时代模型算力需求以超过摩尔定律增长



数据中心通过交换机网络实现设备互联



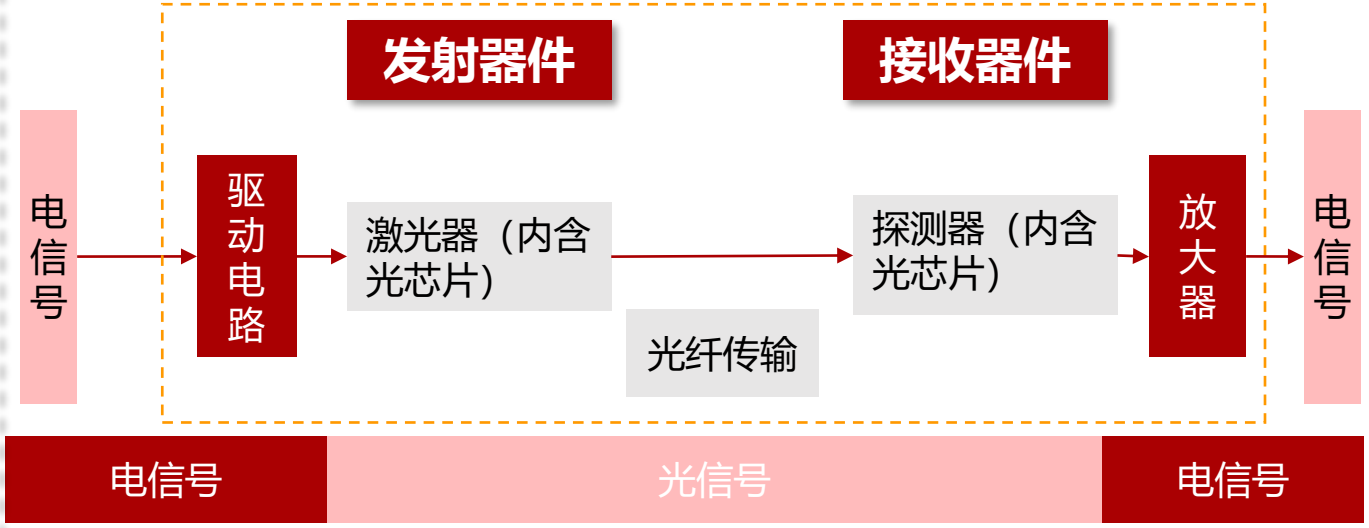
通信延时导致加速放缓



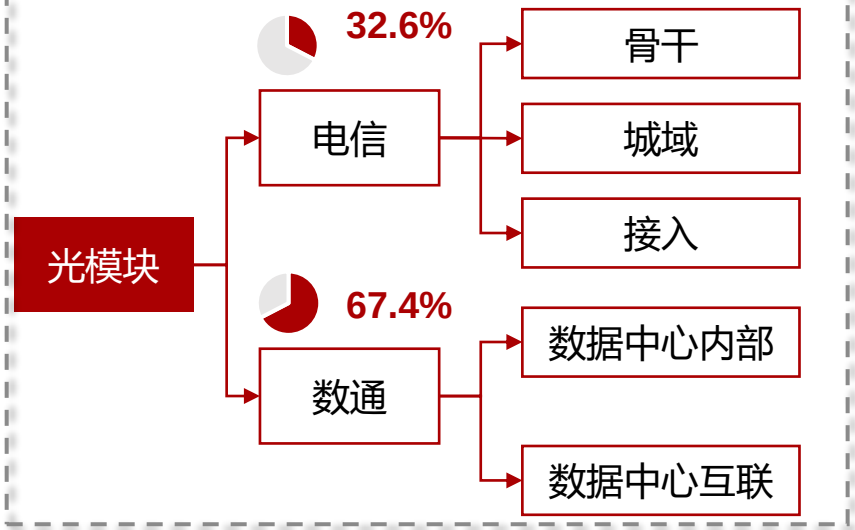
资料来源：Google Scholar，并行处理中节点间通信对加速比的影响，浙商证券研究所

数据传输核心器件：光模块

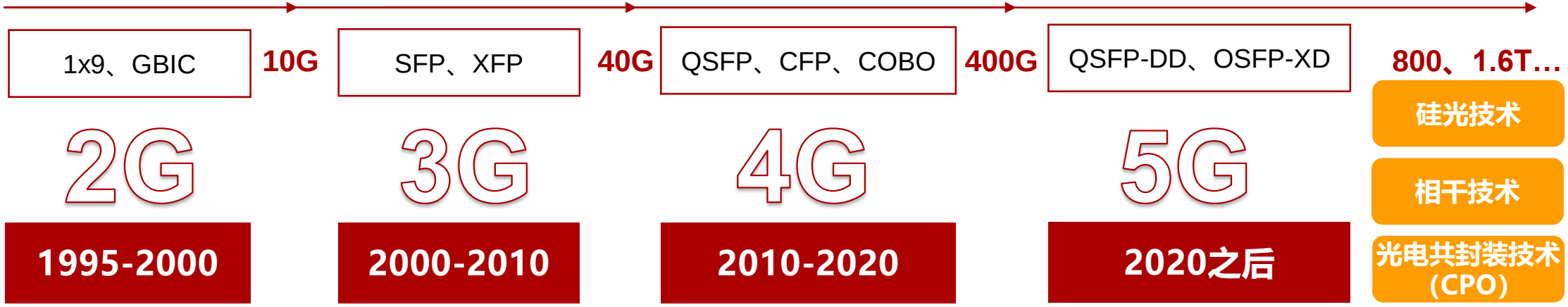
光模块包含发射器件和接收器件，光纤通信时的光电转换



数据中心占光模块一半以上市场 (2021Q4)



光模块向高速传输发展，以顺应数据传输量增长趋势



资料来源：中际旭创年报，光纤在线，电子发烧友，华经情报网，浙商证券研究所

02

技术创新引领本土产业链弯道突围

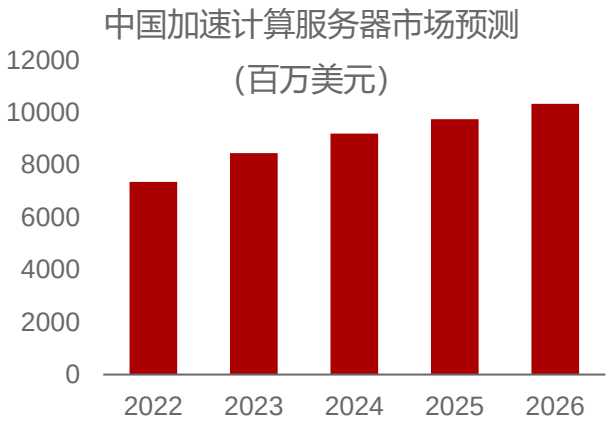
国产服务器CPU发展之路

通过CHIPLET布局先进制程，服务器芯片广泛应用

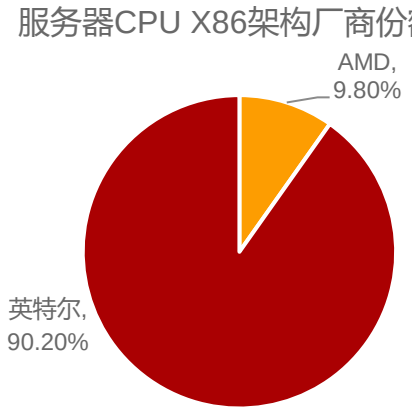
存算一体打破“存储墙”限制，实现降本增效

服务器CPU需求增长，国化产三条发展路线

中国服务器市场规模



服务器CPU市场格局



国产服务器CPU发展之路

IP内核授权	CISC	X86架构	上海兆芯	
			海光信息	
指令集架构授权	RISC	ARM架构	华为鲲鹏	HUAWEI
			天津飞腾	Phytium 飞腾
授权+自主研制指令集		MIPS架构	龙芯中科	龙芯中科 LOONGSON TECHNOLOGY
		MIPS架构	申威科技	

- **自主化程度**：低，未来扩充指令集难度较大，但生态迁移成本小、性能高
- **缺点**：安全基础不牢靠

- **自主化程度**：较高，安全基础相对牢靠，拥有自主发展权
- **缺点**：生态构建较为困难

- **自主化程度**：极高，申威科技已基本实现完全自主可控
- **缺点**：生态构建极其困难

资料来源：IDC，wind，浙商证券研究所整理

02 未来算力升级路径：CHIPLET、存算一体

近期CHATGPT的兴起推动着人工智能在应用端的蓬勃发展，这也对计算设备的运算能力提出了前所未有的需求。虽然AI芯片、GPU、CPU+FPGA等芯片已经对现有模型构成底层算力支撑，但面对未来潜在的算力指数增长，**短期使用CHIPLET异构技术加速各类应用算法落地，长期来看打造存算一体芯片（减少芯片内外的数据搬运），或将成为未来算力升级的潜在方式。**

2016 Google DeepMind AlphaGo
AlphaGo (DeepMind)

2019 Google DeepMind 游戏
AlphaStar (DeepMind)

2021 Google DeepMind 蛋白质结构
AlphaFold2 (DeepMind)

2023 OpenAI ChatGPT
ChatGPT (OpenAI)

2018 Google AI Google
BERT (Google AI)

2020 OpenAI GPT-3
GPT-3 (OpenAI)

2022 Google DeepMind Gato
Gato (DeepMind)

大模型　通用AI

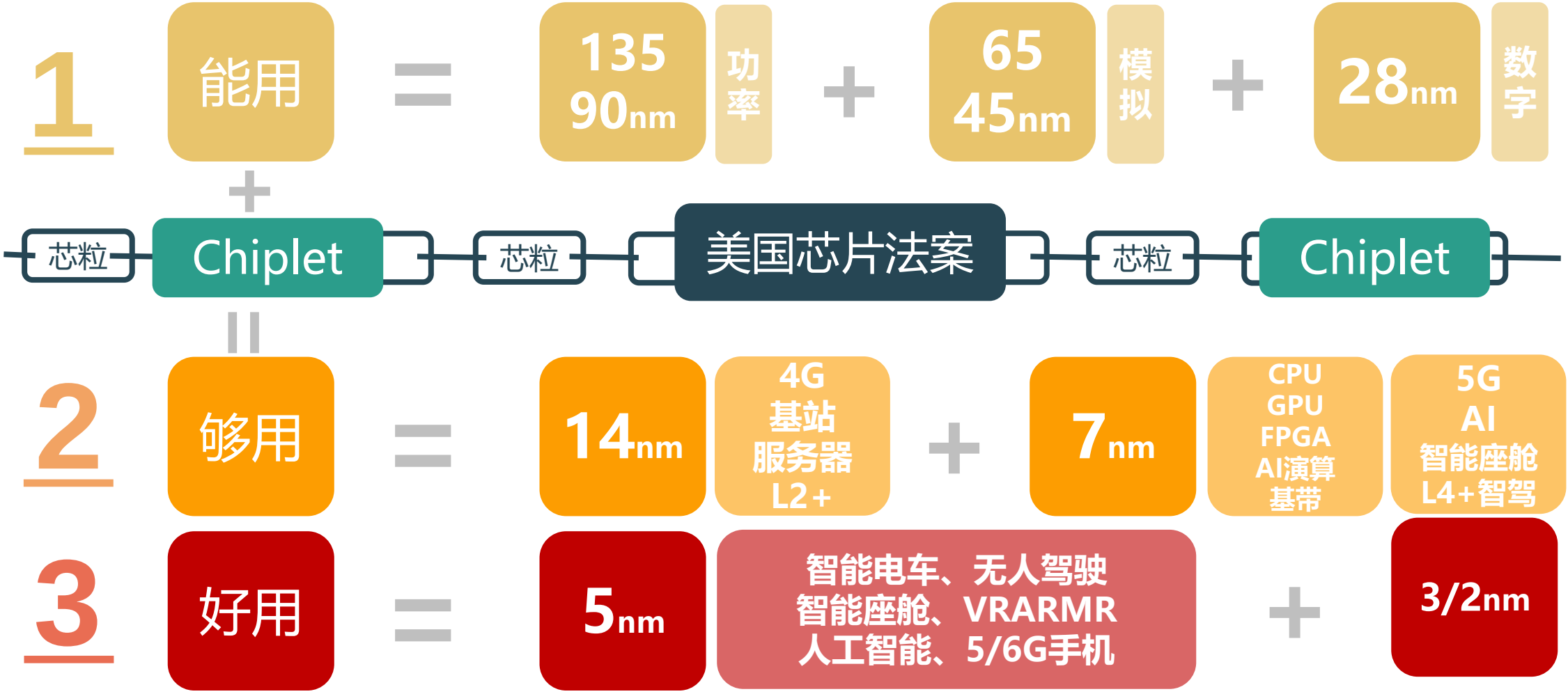
大算力

CPU → GPU → 未来：Chiplet? → 未来：存算一体?

资料来源：知乎《先进存算一体芯片设计》（陈巍等），浙商证券研究所

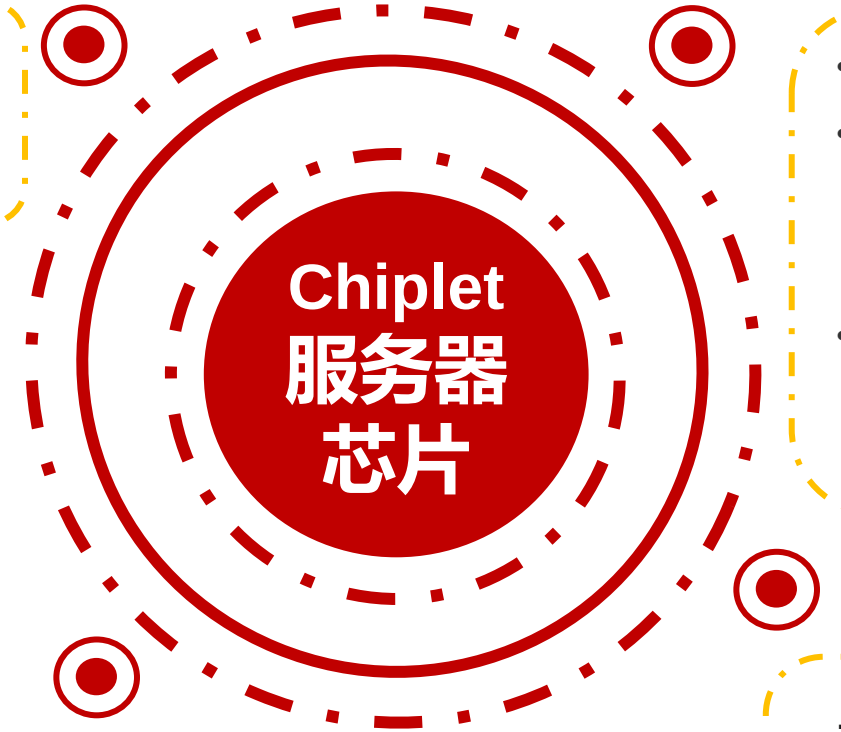
CHIPLET是布局先进制程、加速算力升级的关键技术

Chiplet异构技术不仅可以突破先进制程的封锁，并且可以大幅提升大型芯片的良率、降低设计的复杂程度和设计成本、降低芯片制造成本。Chiplet技术加速了算力升级，但需要牺牲一定的体积和功耗，因此将率先在基站、服务器、智能电车等领域广泛使用。



资料来源：芯智讯，浙商证券研究所

CHIPLET已广泛应用于服务器芯片



华为海思：鲲鹏920

- **采用7nm制造工艺，基于ARM架构授权**
- 由华为公司自主设计完成。典型主频下，SPECint Benchmark评分超过930。

寒武纪：云端AI芯片思元370

- **基于7nm制程工艺，是寒武纪首款采用chiplet（芯粒）技术的AI芯片**
- 集成了390亿个晶体管，最大算力高达256TOPS(INT8)，是寒武纪第二代产品思元270算力的2倍。
- 内存带宽是上一代产品的3倍，访存能效达GDDR6的1.5倍。

AMD：EPYC 第1代至第4代

- **Chiplet服务器芯片的引领者，4代产品采用5nm**
- 基于chiplet 的第一代AMD EPYC 处理器中，装载 8 个“Zen ”CPU 核，2 个DDR4 内存通道和32 个 PCIe 通道，以满足性能目标。
- 2022年AMD正式发布第四代EPYC处理器，拥有高达96颗5nm的Zen 4核心，并使用新一代的Chiplet工艺，结合5nm和6nm工艺来降低成本。

英特尔：第14代酷睿 Meteor Lake

- **首次采用 intel 4工艺，首次引入chiplet小芯片设计，预计将于23年下半年推出**
- 至少性能功耗比的目标要达到 13 代 Raptor Lake 的 1.5 倍水平。

龙芯中科：龙芯3D5000（试验）

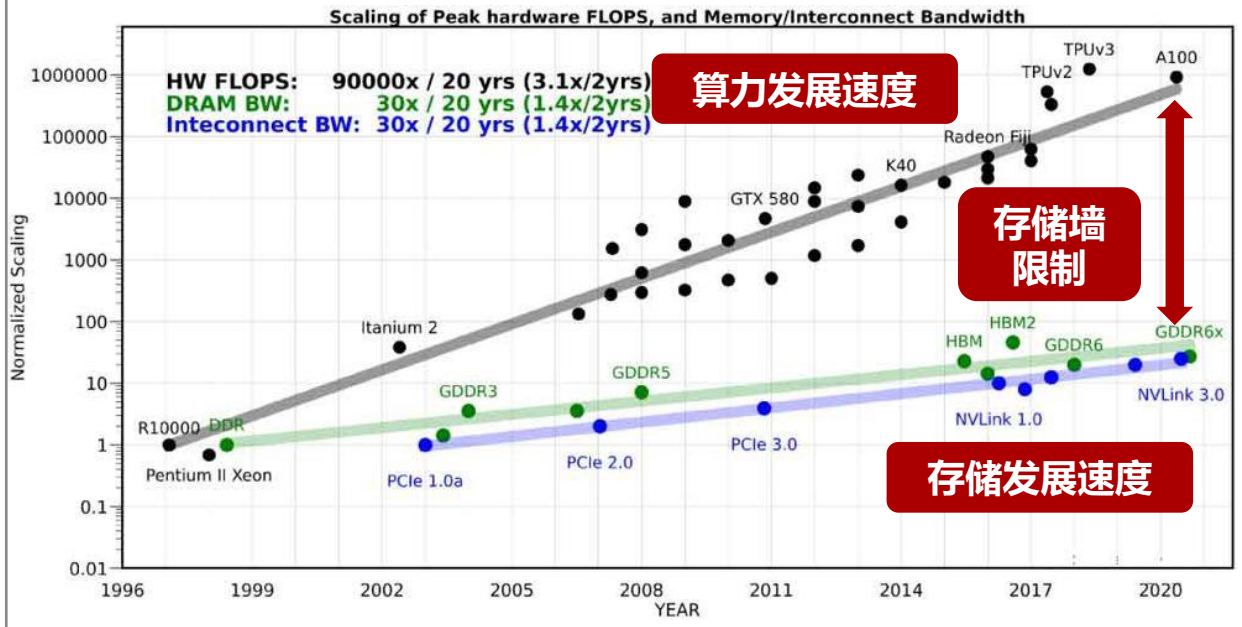
- **面向服务器市场的 32 核 CPU 产品**，通过Chiplet技术把两个 3C5000 硅片封装在一起， 集成了32 个 LA464 处理器核和 64MB 片上共享缓存，22年末初样试验成功

资料来源：各公司官网，浙商证券研究所

存算一体：打破“存储墙”限制，技术迭代演进

“存储墙”成为了数据计算应用的一大障碍

面对计算中心的数据洪流，数据搬运慢、搬运能耗大等问题成为了计算的关键瓶颈。从处理单元外的存储器提取数据，搬运时间往往是运算时间的成百上千倍，整个过程的无用能耗大概在60%-90%之间，能效非常低。



存算技术演进路线

- **查存计算（Processing With Memory）**
  - GPU对复杂函数的运算
  - 最早期技术
- **近存计算（Computing Near Memory）**
  - AMD的Zen系列CPU
  - 三星HBM-PIM
- **存内计算（Computing In Memory）**
  - Mythic
  - 千芯科技
  - 闪存
  - 知存
- **存内逻辑（Logic In Memory）**
  - TSMC
  - 千芯科技
  - 满足大模型计算精度要求

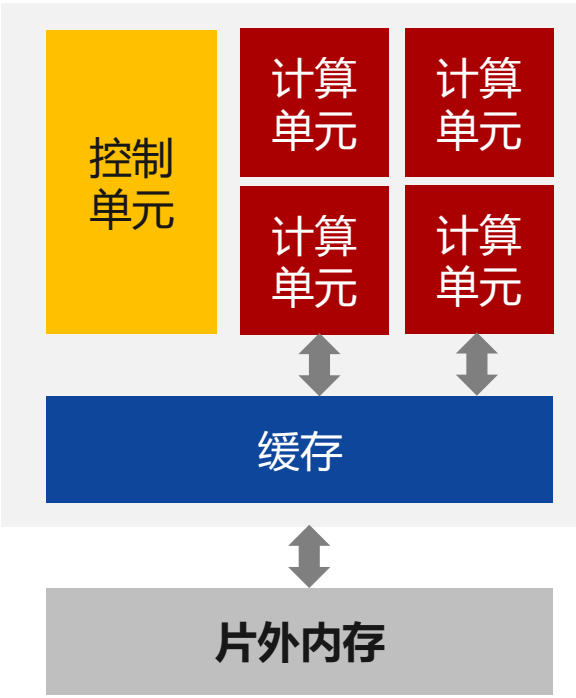
资料来源：知乎《先进存算一体芯片设计》（陈巍等），浙商证券研究所

存算一体：更大算力、更高能效、降本增效

存算一体就是存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵计算。**存算一体的优势**包括：（1）具有更大算力（1000TOPS以上）（2）具有更高能效（超过10-100TOPS/W），超越传统ASIC算力芯片（3）降本增效（可超过一个数量级）

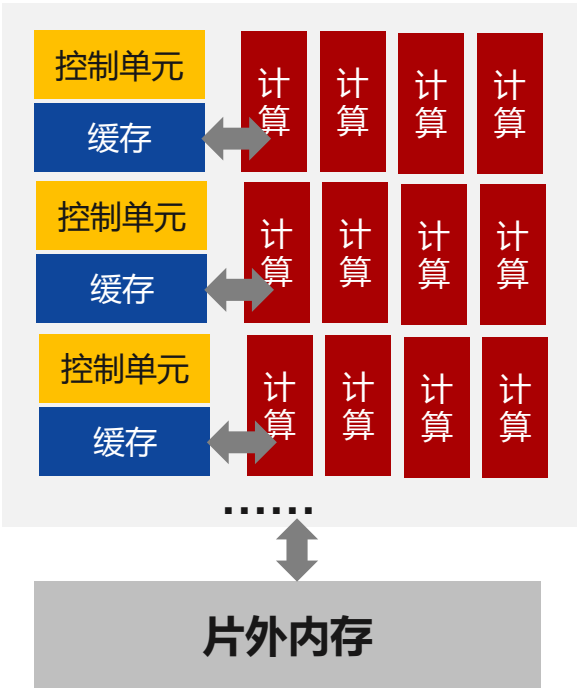
CPU

一般10-100计算核心



GPU

一般万量级计算核心



存算一体

一般百万量级等效计算核心

存算一体

- 存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵运算。

存算一体的应用领域

- 自动驾驶
- 自然语言处理
- 智慧城市
- 商品推荐
- 工业视觉
- 医药计算
- 个性化推荐
- 多语言精准识别

资料来源：知乎《先进存算一体芯片设计》（陈巍等），浙商证券研究所

风险提示

1、AI技术发展不及预期：当前以ChatGPT为代表的NLP模型以及其他类型人工智能模型发展仍不成熟，存在一定缺陷；

2、版权、伦理和监管风险：AIGC生成的内容依赖现有版权素材，另外不当使用或模型自身问题可能导致不良后果；

3、半导体下游需求不及预期：全球芯片行业存在周期性，可能因宏观经济波动导致需求低迷。

行业评级与免责声明

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

1、看好 ：行业指数相对于沪深300指数表现＋10%以上；

2、中性 ：行业指数相对于沪深300指数表现－10%～＋10%以上；

3、看淡 ：行业指数相对于沪深300指数表现－10%以下。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。

建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

行业评级与免责声明

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。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。

本报告仅反映报告作者的出具日的观点和判断，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，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

联系方式

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：http://research.stocke.com.cn