ChatGPT研究框架

——【AIGC算力时代系列报告】

行业评级：看好

2023年2月14日

分析师	陈杭	李佩京	姚天航	张建民	研究助理	安子超
邮箱	chenhang@stocke.com.cn	lipeijing@stocke.com.cn	yaotianhang@stocke.com.cn	zhangjianmin@stocke.com.cn	邮箱	anzichao@stocke.com.cn
证书编号	S1230522110004	S1230522060001	S1230522010001	S1230518060001		

摘要

【芯片算力】▲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；2）价：对高端芯片的需求将拉动芯片均价。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。

相关标的：海光信息、景嘉微、龙芯中科、中国长城、安路科技、复旦微电、紫光国微、寒武纪、澜起科技、德科立、天孚通信、中际旭创。

【深度学习框架】深度学习框架是人工智能算法的底层开发工具，是人工智能时代的操作系统，当前深度学习框架发展趋势是趋于大模型训练，对深度学习框架的分布式训练能力提出了要求，国产深度学习框架迎来发展机遇。

相关标的：百度 、海天瑞声、商汤科技、微软、谷歌、Meta 。

【深度学习大模型】ChatGPT是基于OpenAI公司开发的InstructGPT模型的对话系统，GPT系列模型源自2017年诞生的Transformer模型，此后大模型数量激增，参数量进入千亿时代，国内百度也发布了ERNIE系列模型并有望运用于即将发布的文心一言（ERNIE Bot）对话系统，未来国内厂商有望在模型算法领域持续发力。

相关标的：百度、科大讯飞、商汤科技、谷歌、微软。

【应用】ChatGPT火爆全球的背后，可以窥见伴随人工智能技术的发展，数字内容的生产方式向着更加高效迈进。ChatGPT及AIGC未来有望在包括游戏、广告营销、影视、媒体、互联网、娱乐等各领域应用，优化内容生产的效率与创意，加速数实融合与产业升级。

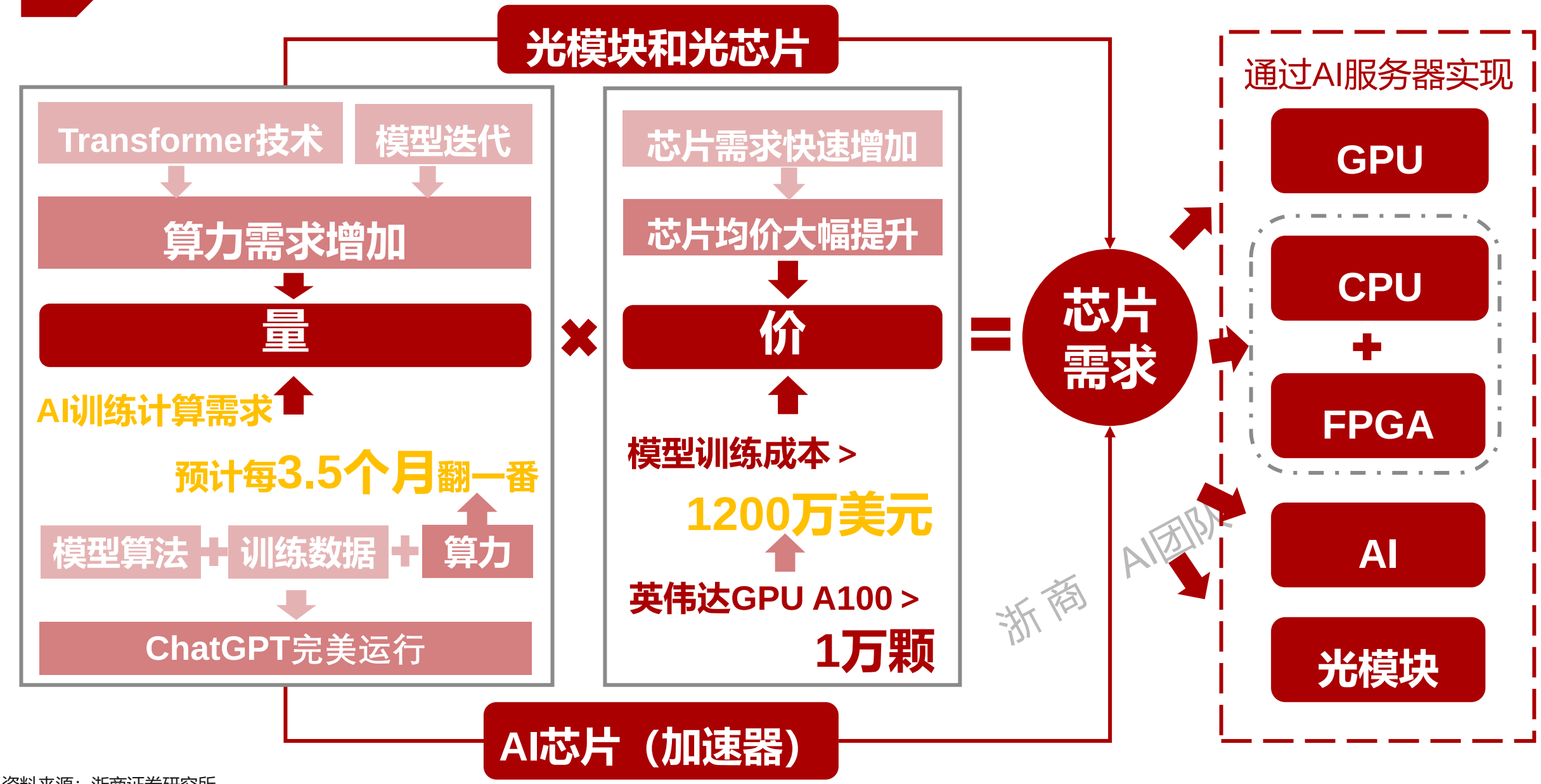
相关标的：百度、腾讯、阿里巴巴、网易、昆仑万维、阅文集团、捷成股份、视觉中国、风语筑、中文在线、三七互娱、吉比特、天娱数科。

【通信】AIGC类产品未来有望成为5G时代新的流量入口，率先受益的有望是AIGC带来的底层基础算力爆发式增长。

相关标的：包括算力调度（运营商）、算力供给（运营商、奥飞数据、数据港）、算力设备（浪潮信息、联想集团、紫光股份、中兴通讯、锐捷网络、天孚通信、光库科技、中际旭创、新易盛）、 算力散热（英维克、高澜股份）。

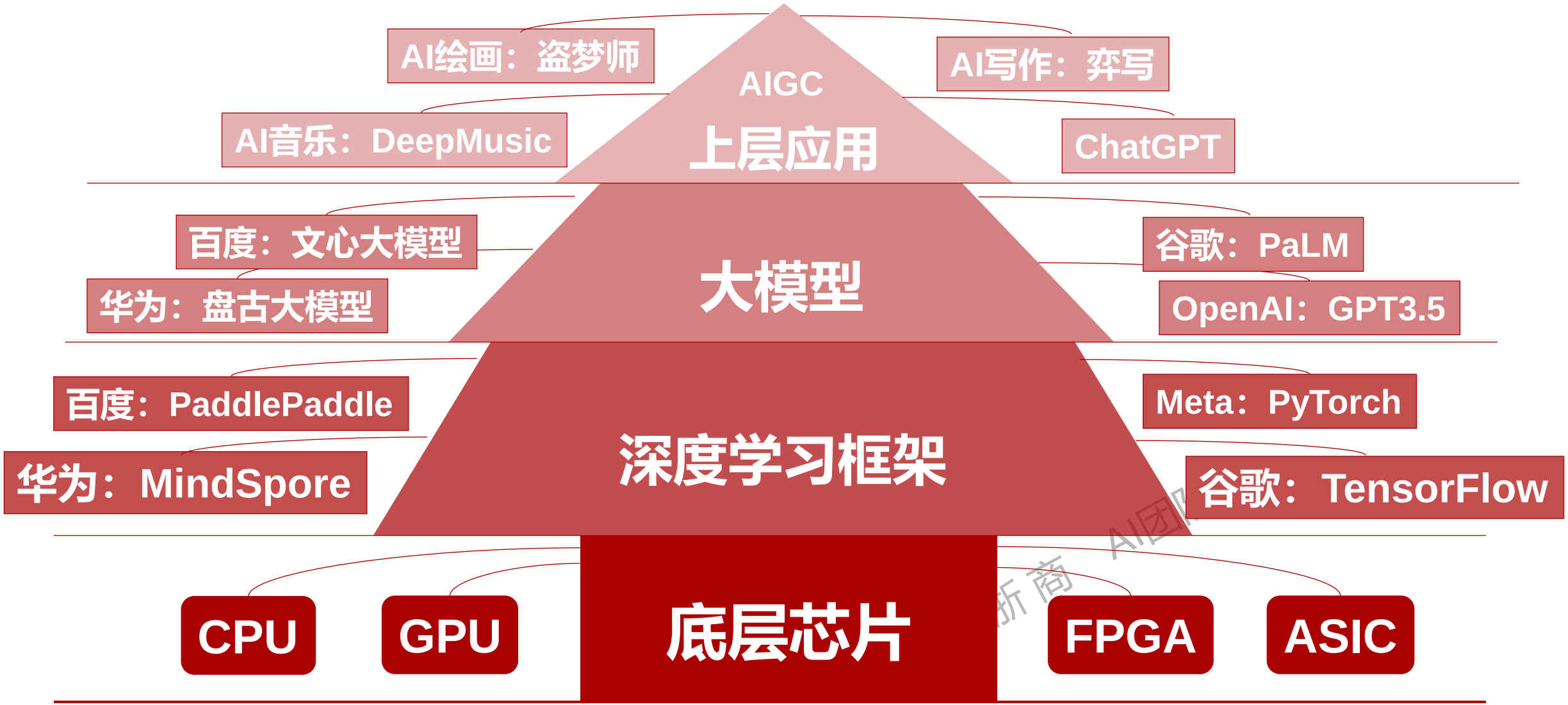
1、芯片算力

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

人工智能四层架构，芯片为底层支撑



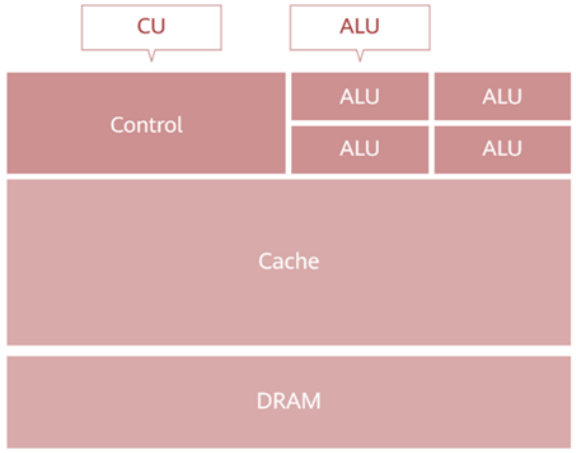
资料来源：澜舟科技，IDC，量子位，电子发烧友等，浙商证券研究所

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

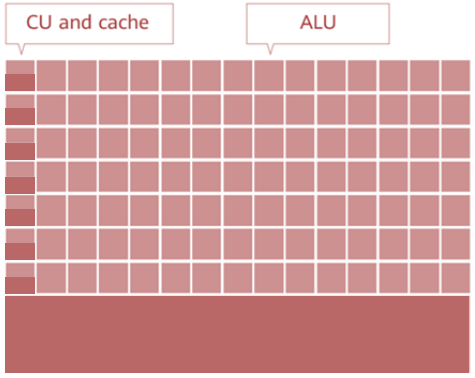
CPU



✓ 逻辑判断
✓ 任务调度与控制

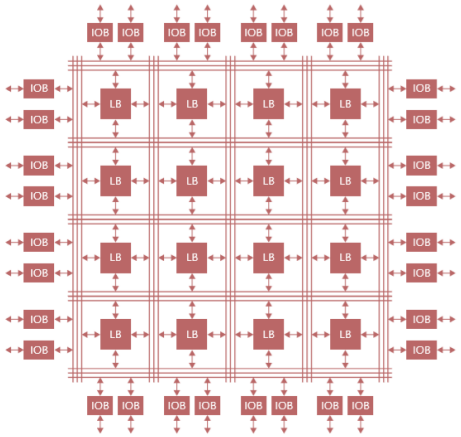
- 并行架构
- 计算单元多
- 适合大量逻辑确定的重复计算

GPU



✓ 模型训练

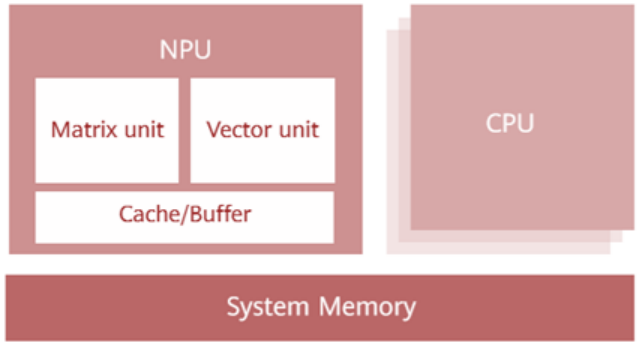
- 低延时
- 开发周期短
- 硬件可根据需求调整
- 成本和壁垒高



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

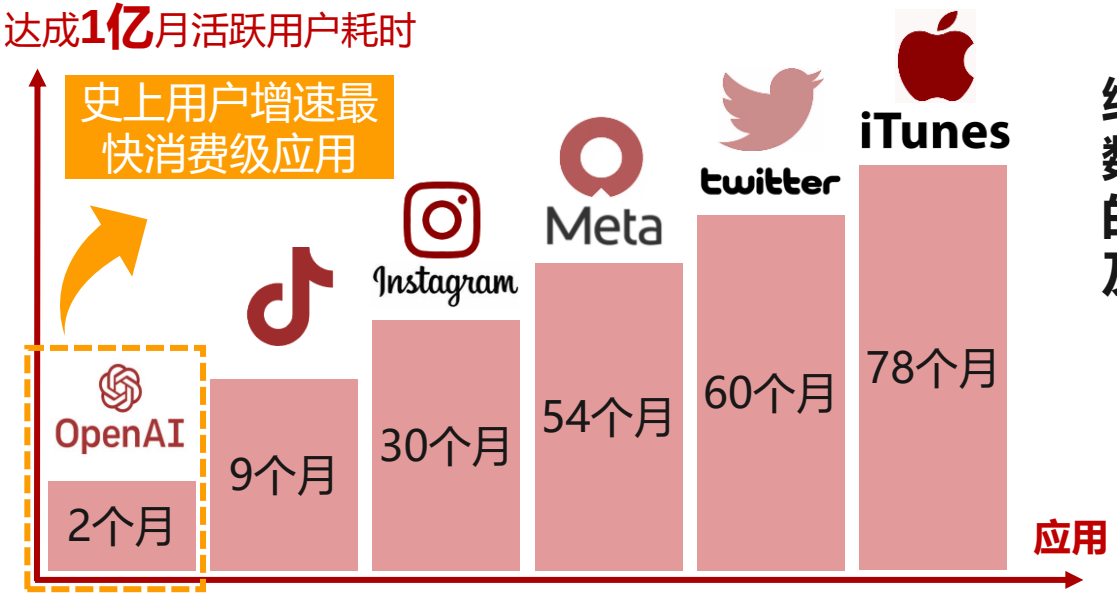
资料来源：Huawei forum，浙商证券研究所

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

↓

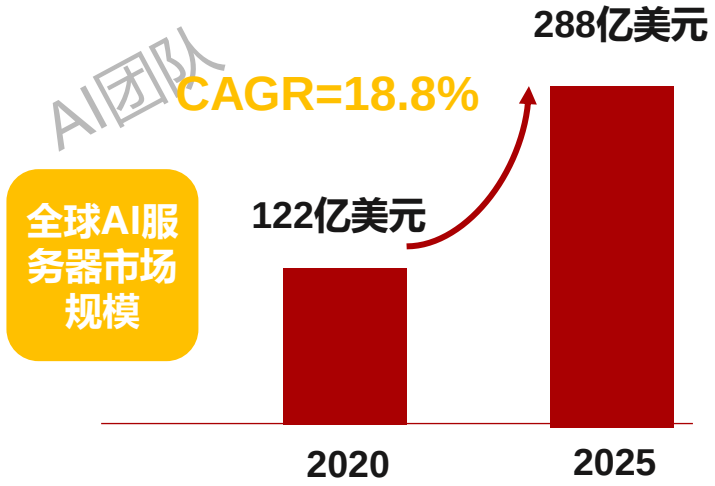
AI服务器需求

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
- ……

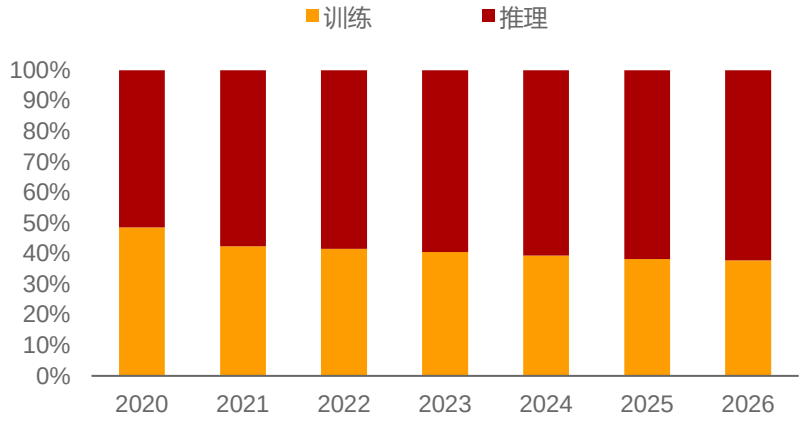


资料来源：Sensor Tower，World of Engineering，头豹研究院，华经产业研究院，浙商证券研究所

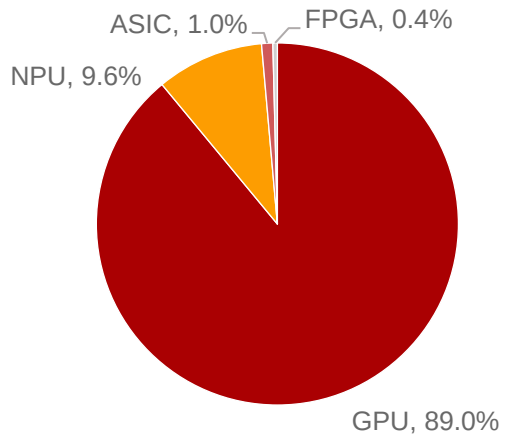
AI服务器快速增长，大力拉动芯片需求

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

中国人工智能服务器工作负载预测



2022年中国人工智能芯片市场规模占比



AI服务器=？

异构形式

CPU

+

GPU

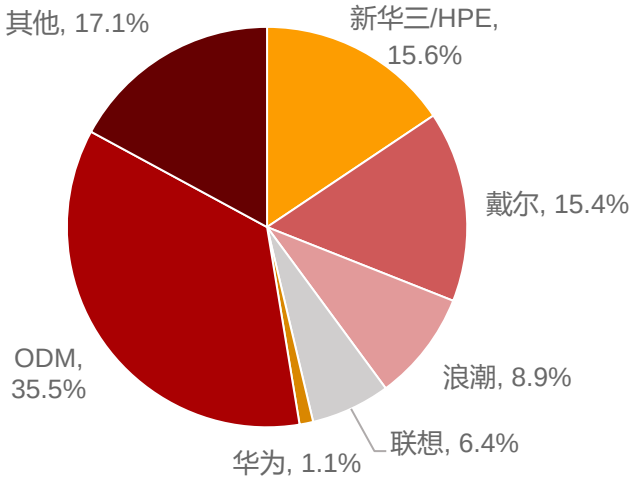
或

FPGA

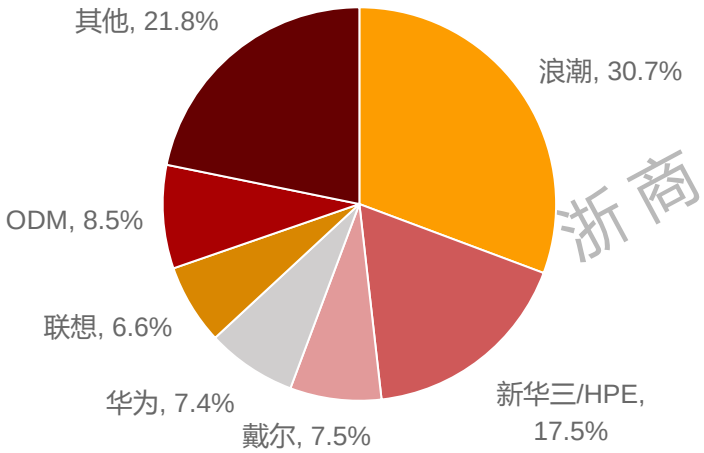
或

ASIC

2021年全球服务器市场格局



2021年中国服务器市场格局



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

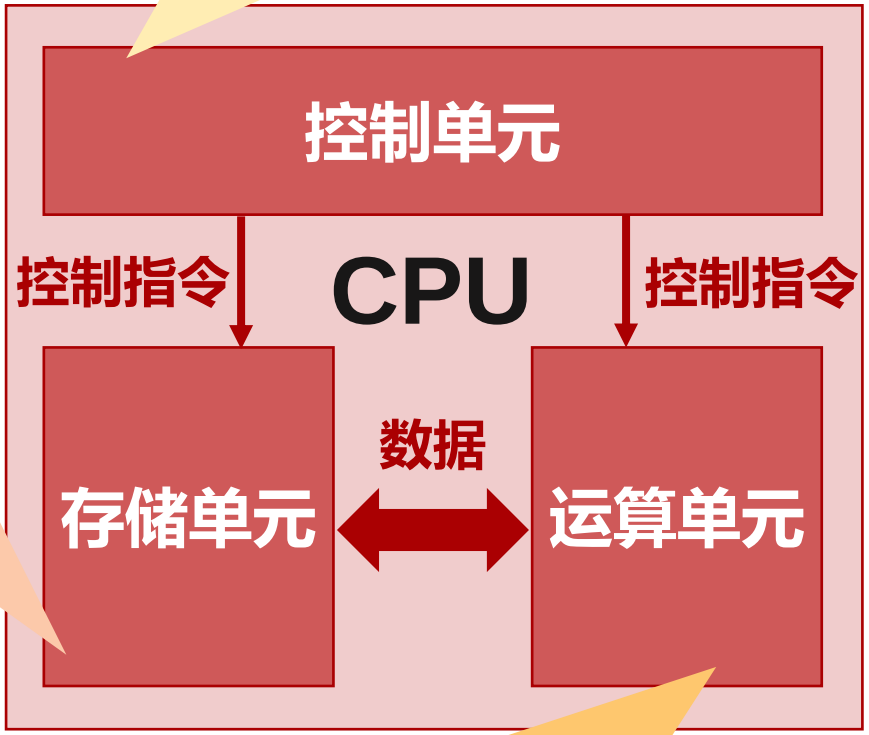
高度适配

- **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

服务器CPU向多核心发展，满足处理能力和速度提升需要

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

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

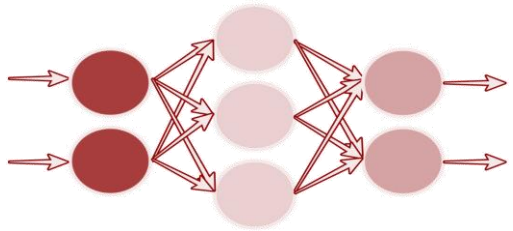
资料来源：芯论语，半导体产业纵横，AMD官网，浙商证券研究所

GPU高度适配AI模型构建

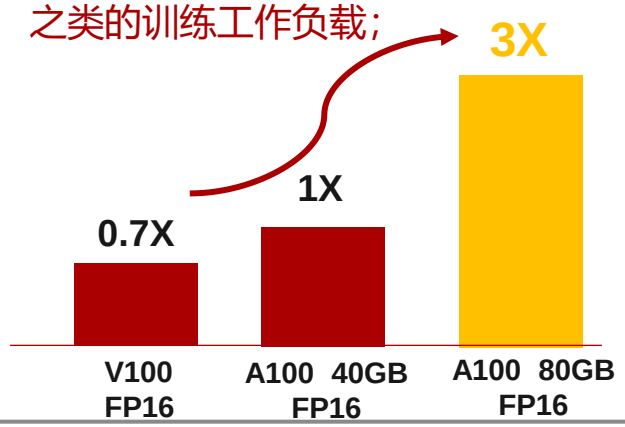
AI模型构建（以英伟达A100为例）

训练过程

GPU的并行计算高度适配神经网络

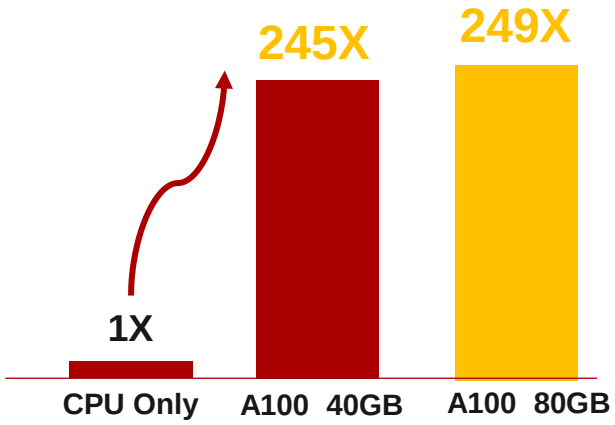


- **GPU帮助高速解决问题：**2048 个 **A100** GPU 可在一分钟内成规模地处理 BERT 之类的训练工作负载；



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

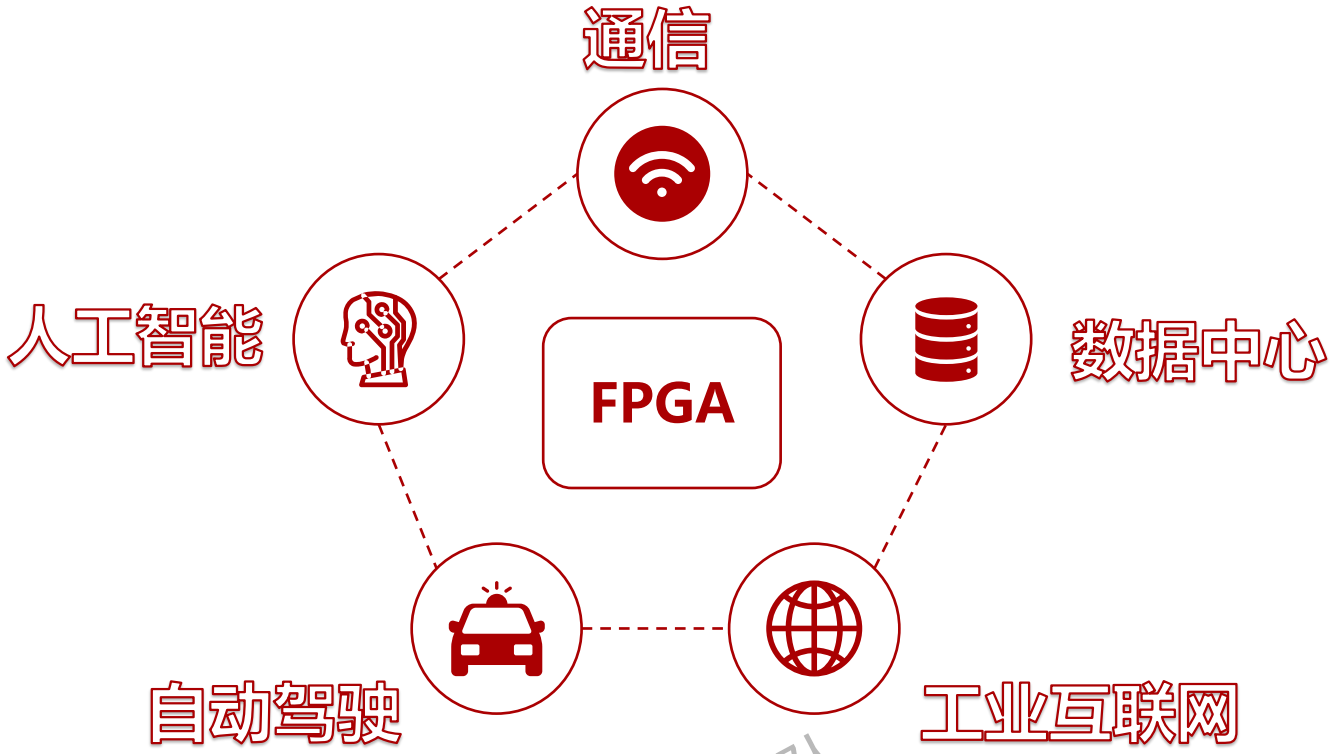
可编程灵活性高：半定制电路，理论上可以实现任意ASIC和DSP的逻辑功能

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

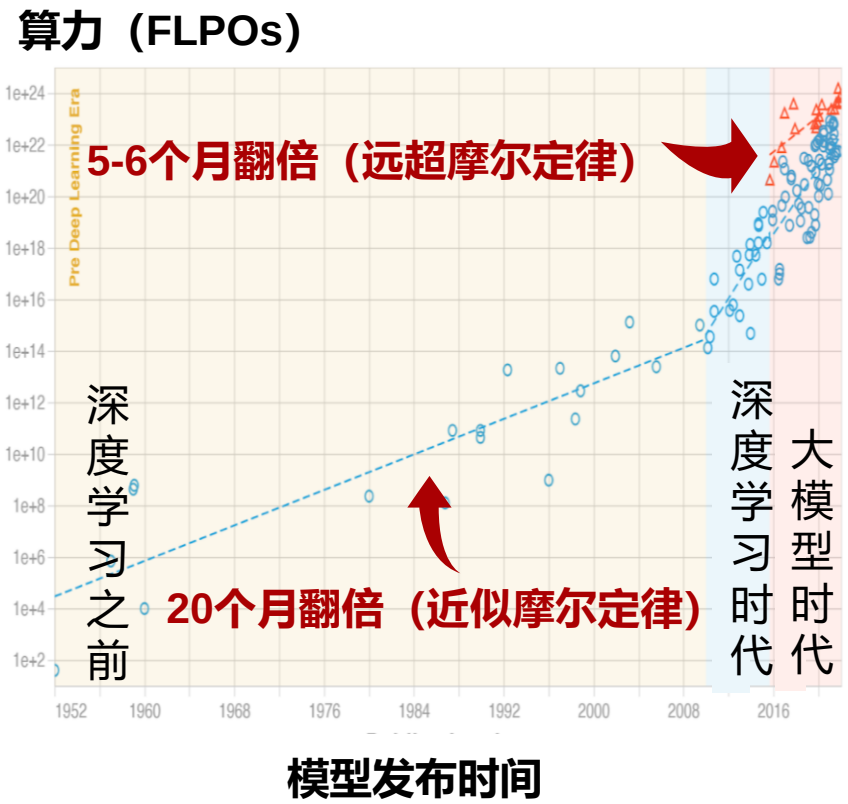
- 业界算力最强的AI处理器，基于自研华为达芬奇架构3D Cube技术；

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

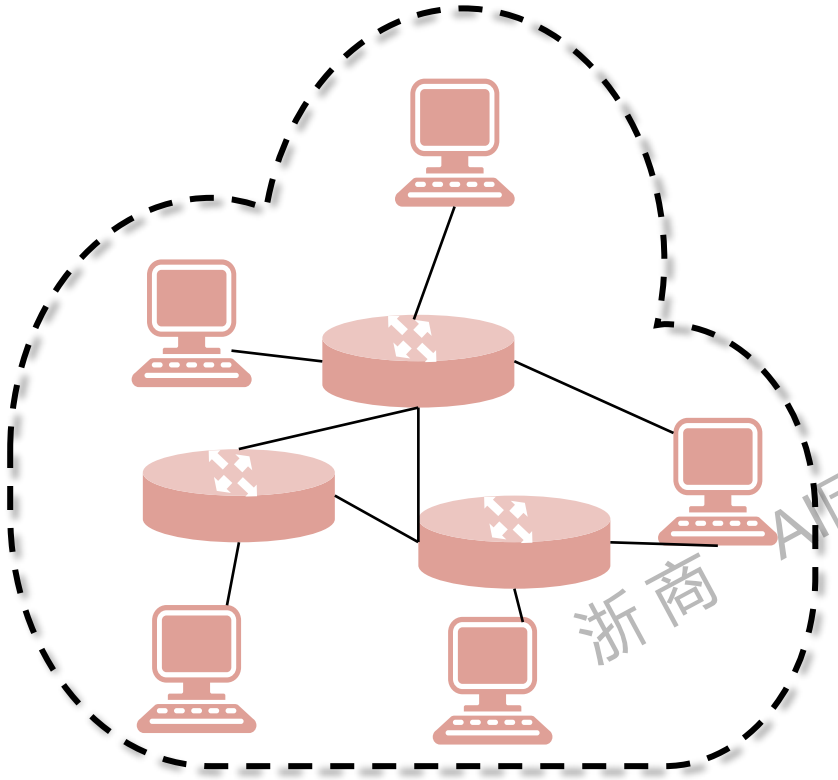
数据传输速率：容易被忽略的算力瓶颈

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**

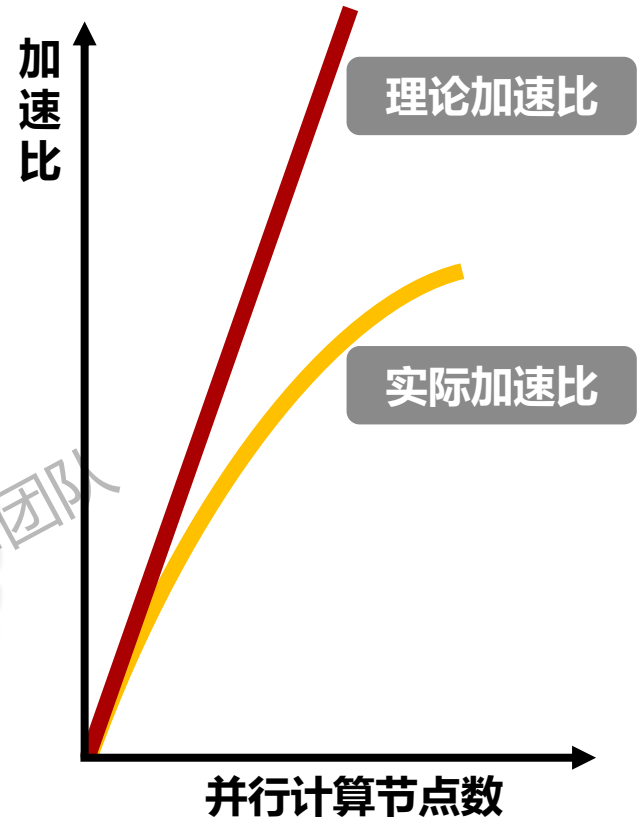
AI时代模型算力需求以超过摩尔定律增长



数据中心通过交换机网络实现设备互联



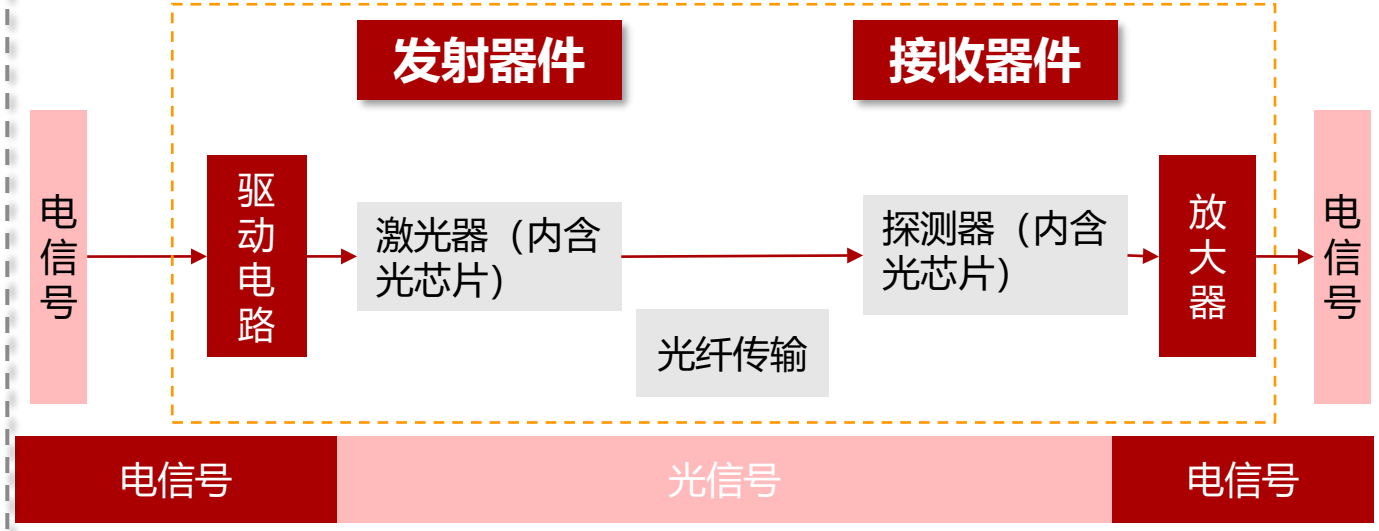
通信延时导致加速放缓



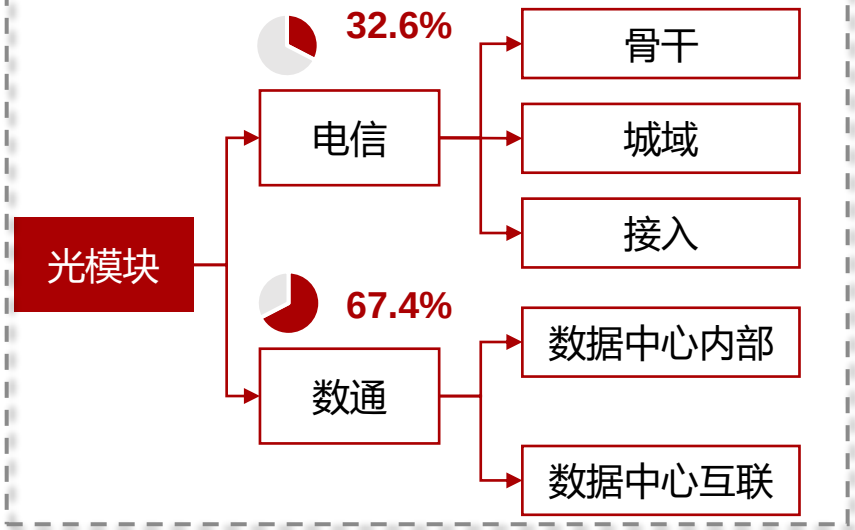
资料来源：Google Scholar，并行处理中节点间通信对加速比的影响，浙商证券研究所

数据传输核心器件：光模块

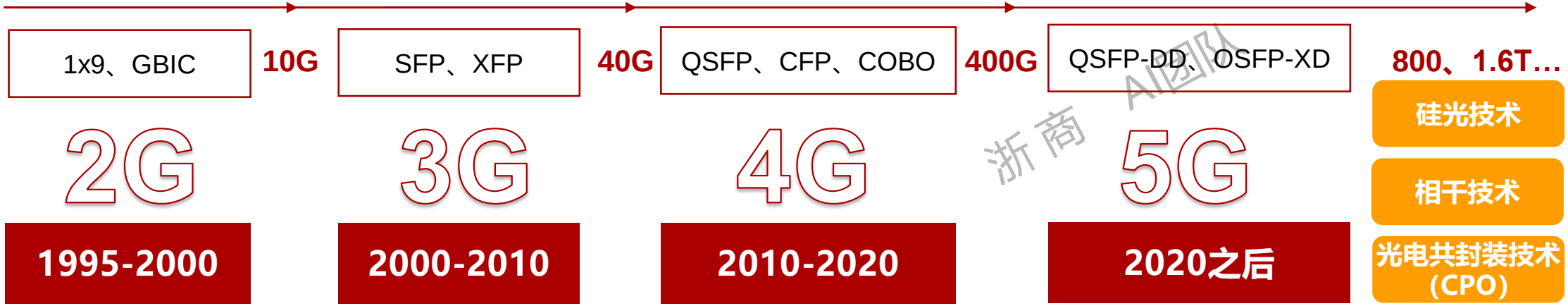
光模块包含发射器件和接收器件，光纤通信时的光电转换



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

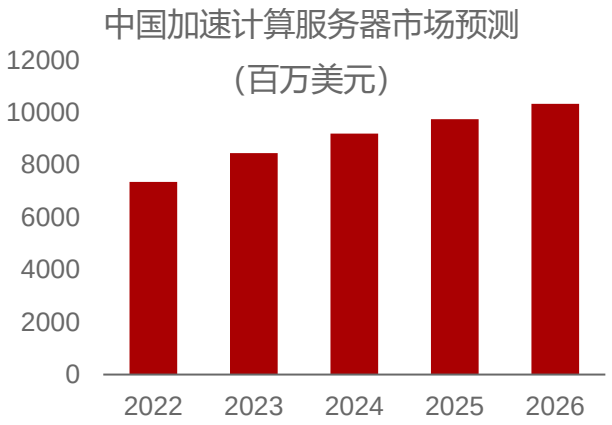


资料来源：中际旭创年报，光纤在线，电子发烧友，华经情报网，浙商证券研究所

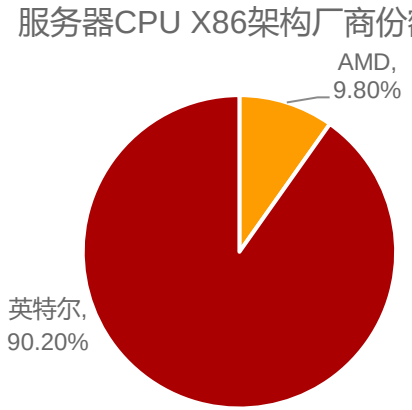
服务器CPU需求增长，国化产三条发展路线

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

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	■ **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
			海光信息	
指令集架构授权	RISC	ARM架构	华为鲲鹏	■ **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
			天津飞腾	
授权+自主研制指令集		MIPS架构	龙芯中科	■ **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
		MIPS架构	申威科技	

资料来源：IDC，wind，浙商证券研究所整理

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)

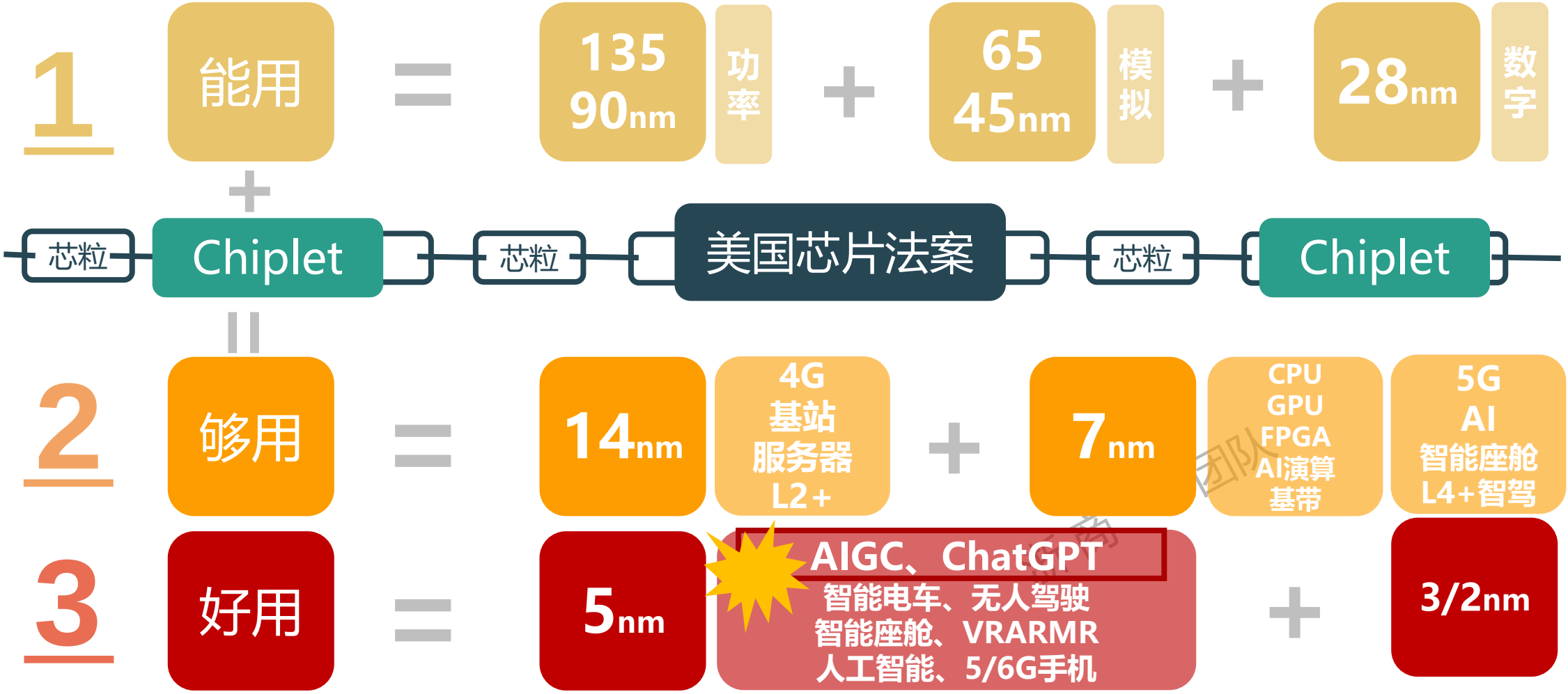
大模型　通用AI　大算力

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

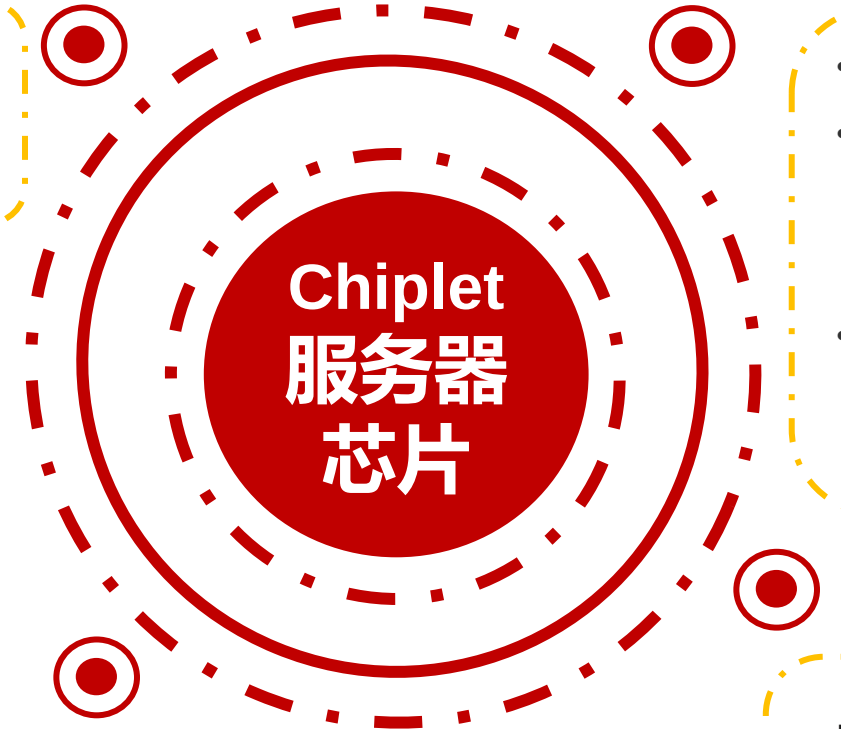
CHIPLET是布局先进制程、加速算力升级的关键技术

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。



数据来源：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）

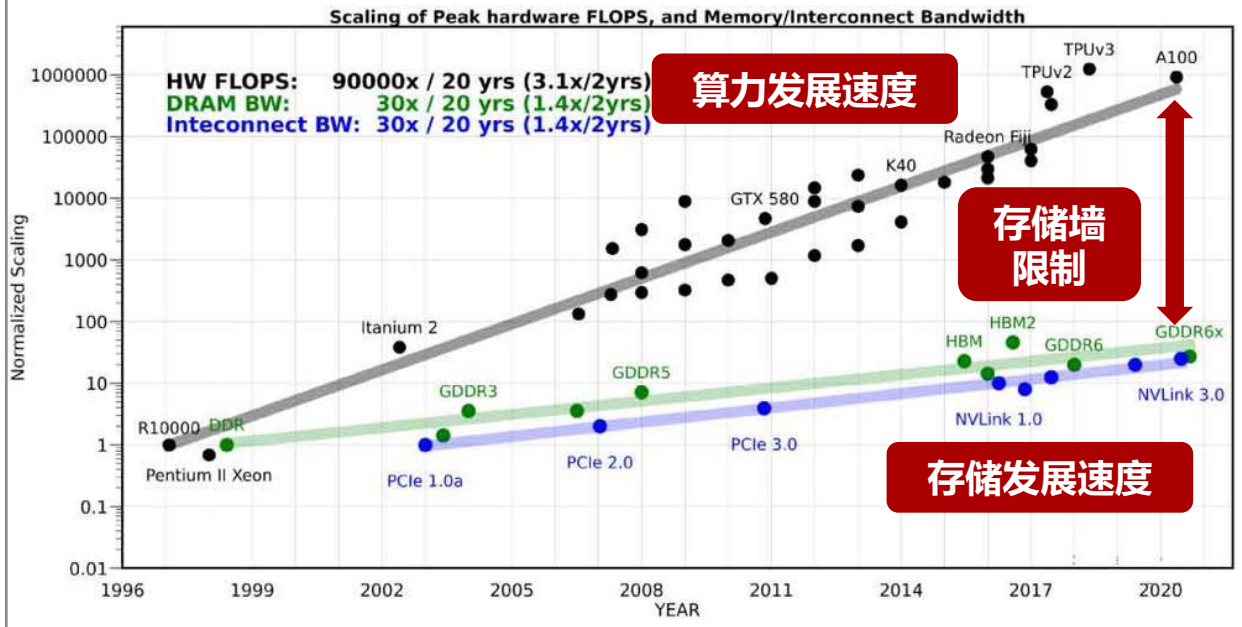
- **面向服务器市场的 32 核 CPU 产品**，通过Chiplet技术把两个 3C5000 硅片封装在一起， 集成了32 个 LA464 处理器核和 64MB 片上共享缓存，22年末初样试验成功

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。



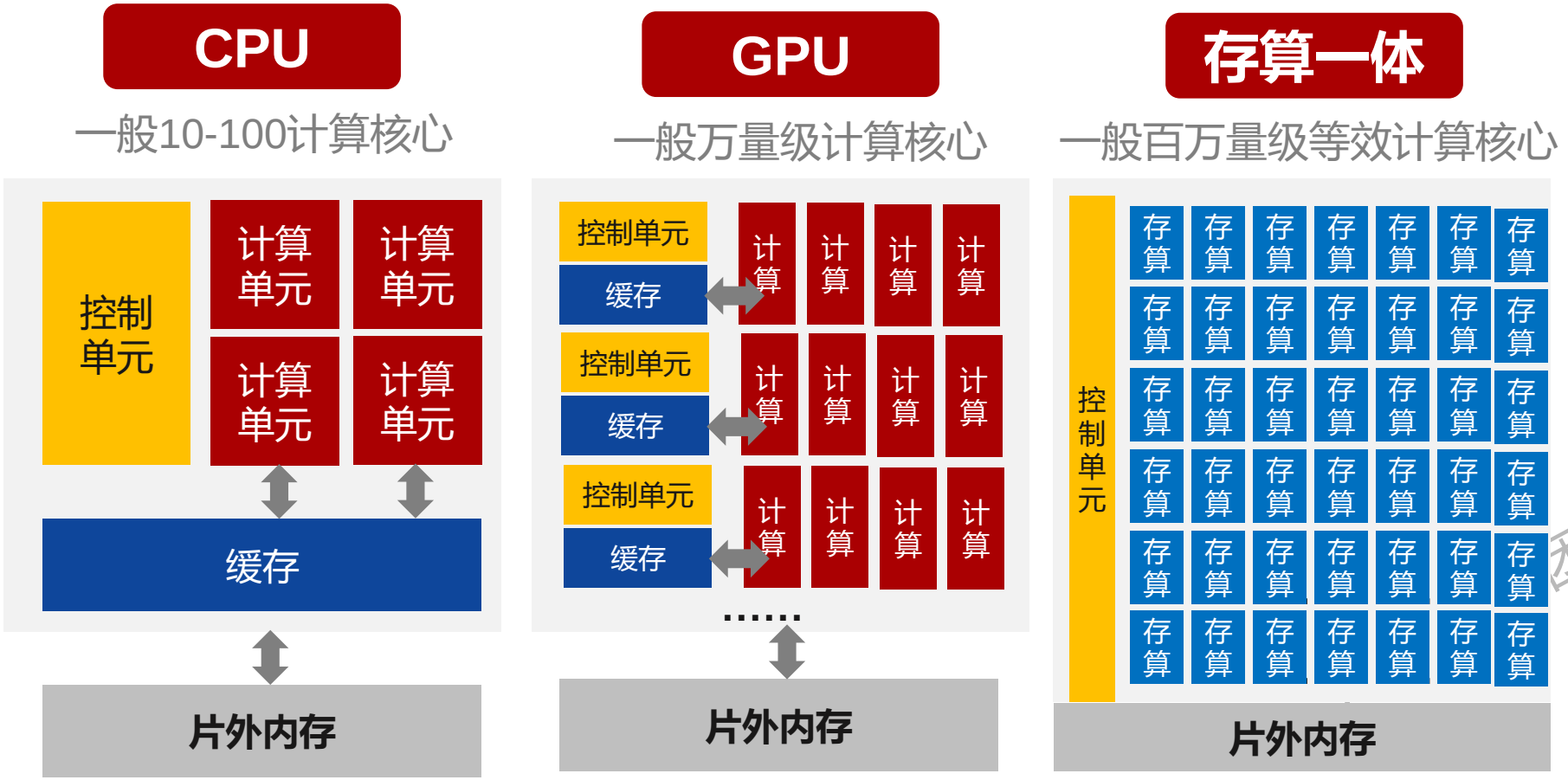
存算技术演进路线

- **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）



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

2、深度学习 框架

深度学习框架：人工智能时代的操作系统

人工智能开发
链条长且复杂

训练 → 数据准备 → 环境安装 → 算法实现 → 模型训练 → 精度调试 → 模型验证

推理 → 环境安装 → 模型转换 → 模型部署 → 效果验证 → 数据收集

数据

使用深度学习框架工具进行开发

人工智能基础设施

深度学习框架

物理资源调度、I/O设备管理

分布式硬件资源

通用模型架构

支持计算机视觉应用

使用流行的编程语言

GPU加速拓展包

使用深度学习框架工具进行开发

模型

开发者使用简单

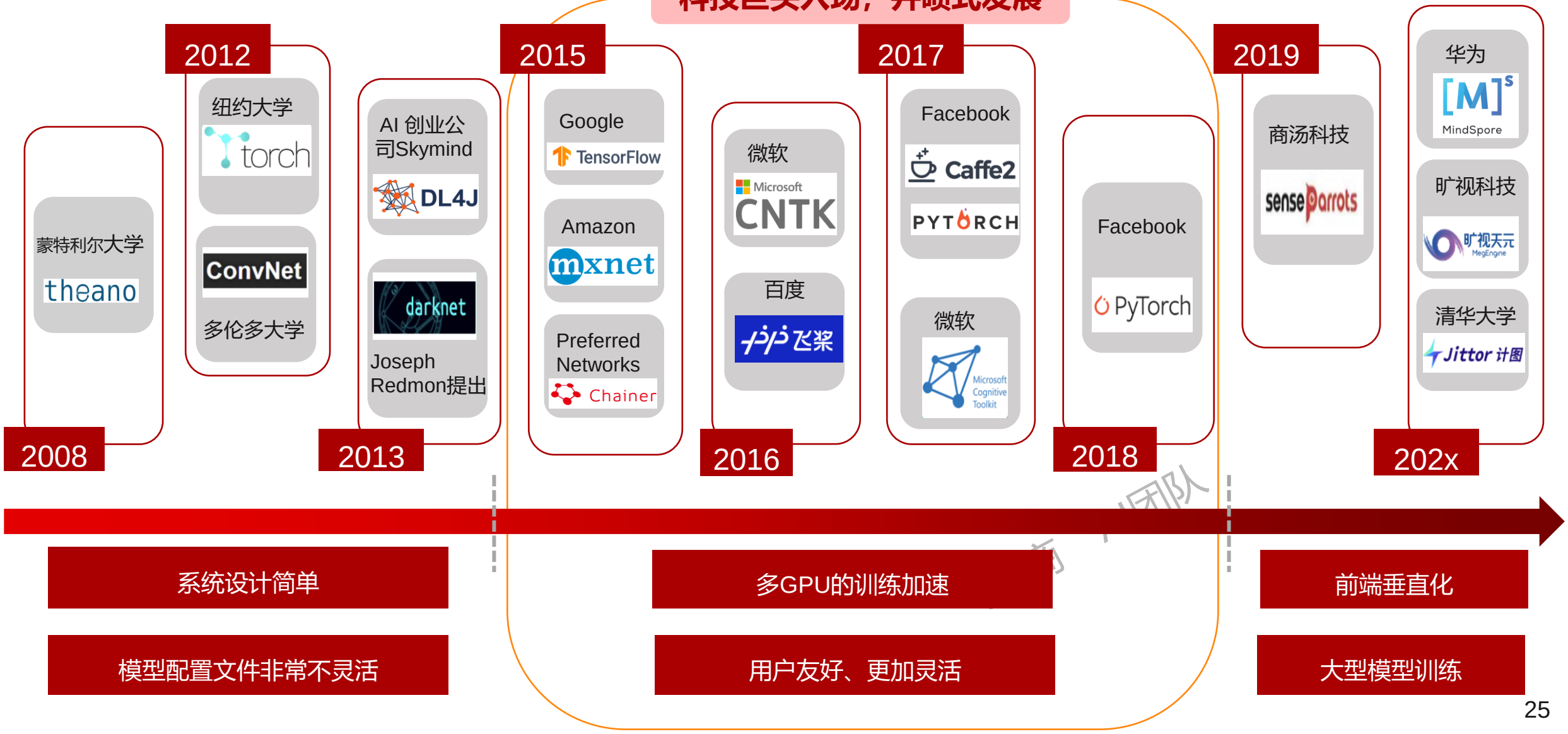
支持AI领域的快速变化

资料来源：北京日报、认知计算与云安全公众号、华为云

02 深度学习框架：大模型训练和前端垂直化成为趋势

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

科技巨头入场，井喷式发展



资料来源：CSDN、 伯禹人工智能学院、机器之心、中国计算机学会、知乎、简书、MAREK REI、磐创AI、浙商证券研究所

国际主流深度学习框架：互联网巨头主导开发

国内外深度学习框架

发布时间	开发公司	深度学习框架	语言	是否开源	计算图	是否是分布式框架	特点/优点
2013	BVLC	Caffe	Python	√	静态	×	速度快、使用方便、社区好
2014	Microsoft	CNTK	Lua，Python (new)	√	静态	√	性能高、适合做语音任务
2015	Google	TensorFlow	C++	√	动态	√	高效灵活、易用
2016	Baidu 百度	飞桨 PaddlePaddle	Python	√	动静兼容	√	容易上手
2017	Meta	PyTorch	C++	√	静态	√	简单清晰
2020	腾讯优图	TNN	Lua，Python (new)	√			移动端高性能、通用轻便
2020	HUAWEI	昇思 MindSpore	Python	√	不依赖计算图	√	高效灵活、易用
202x	MEGVII 旷视	天元 MegEngine	C++、CUDA、Python	√	动静合一		灵活高效

资料来源：机器之心，知乎 ,CSDN

Open AI：从多种框架的使用到专注于Pytroch

深度模型框架转变

2015年

大多数情况使用 TensorFlow + 特殊情况使用 Theano

2020年

Pytroch成为统一的深度学习框架

核心功能

- 支持GPU加速的张量计算
- 方便优化模型的自动微分机制

原因

- 简单易懂：PyTorch 具有用户友好的 API；
- 便于调试：对 Python 生态的良好支持；
- 强大高效：Pytorch提供了非常丰富的模型组件。

极大缩短研究周期

资料来源：OpenAI官网

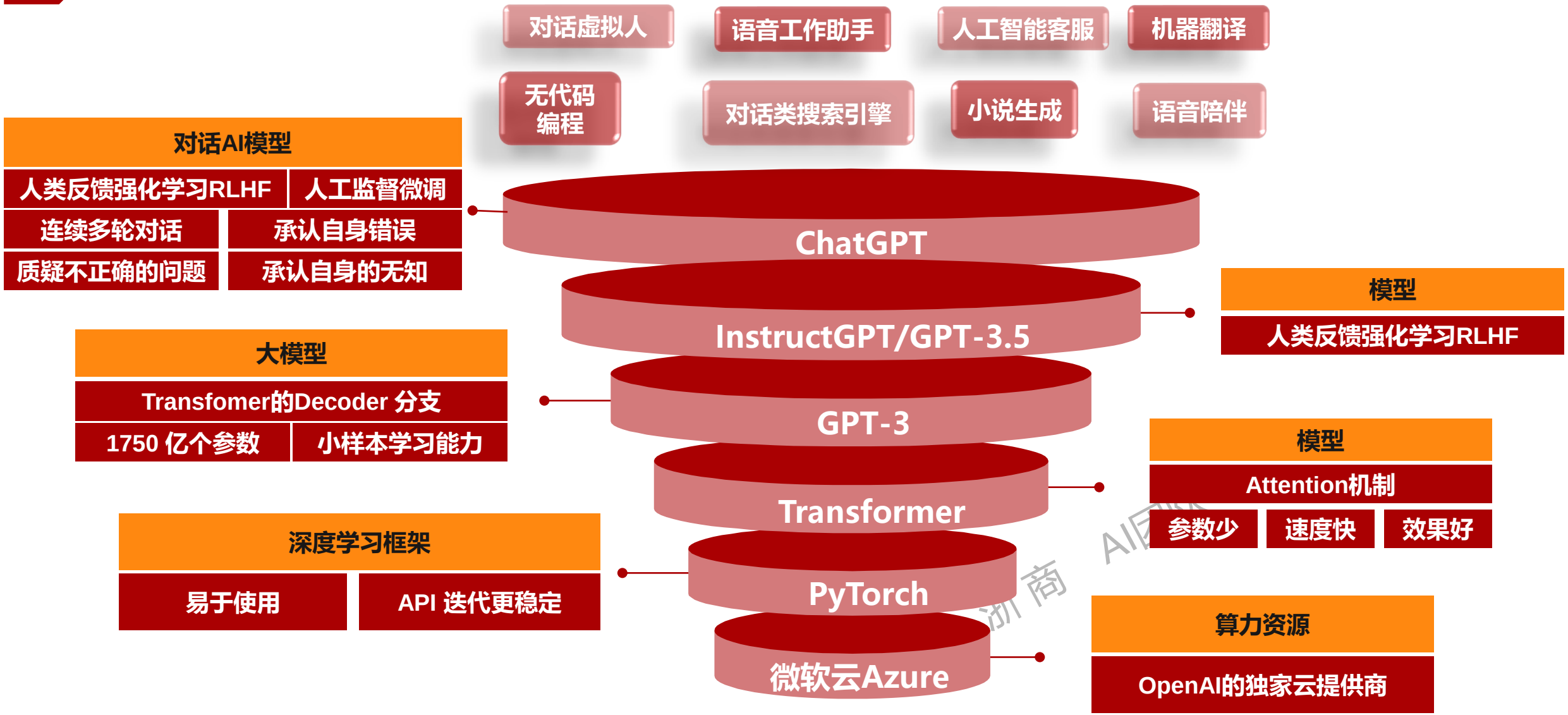
Tensorflow：谷歌开源的向更加易用发展的主流学习框架

Tensorflow从0.1到2.0的发展历程

	2015年	2017年	2019年	202x年
缺点		缺点：调试困难、API混乱、入门困难	缺乏调度能力，需手动配置	
优点	使得神经网络能够有效计算	优点：更快、更灵活、随时就绪	优点：简化的模型开发流程、强大的跨平台能力、强大的研究发现	
改进之处	采用CPU+GPU计算模型	引入更高级的API，可在移动设备上运行	运用更简单的模型进行构建、简化PI	添加两种新机制，解决输入管道瓶颈并节约资源
发布版本	开源Tensorflow 0.1版本	Tensorflow1.0.0发布，稳定版诞生	Tensorflow2.0发布	Tensorflow2.3发布

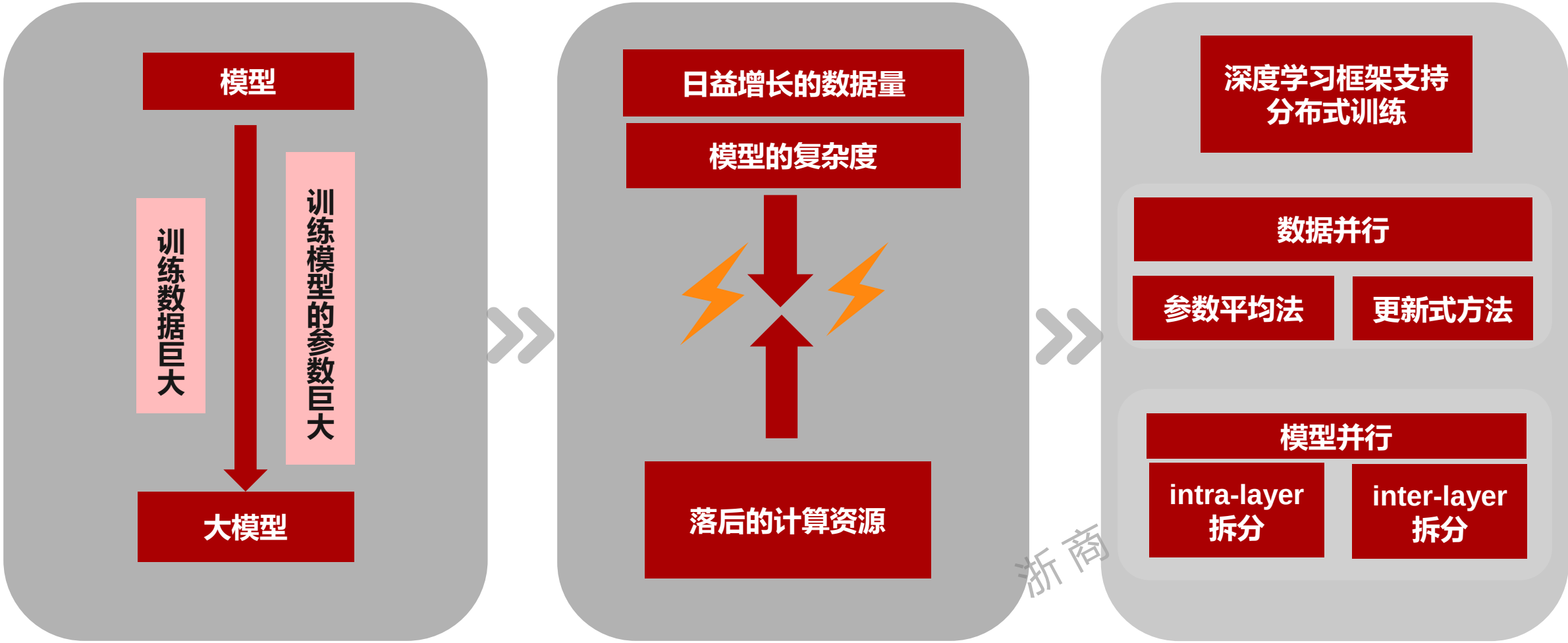
资料来源：腾讯云、CSDN、helloword、城市经济网

02 ChatGPT实现路径：算力与框架支持，应用百花齐放



资料来源：CSDN、电子工程世界、量子位、机器之心、专知、浙商证券研究所

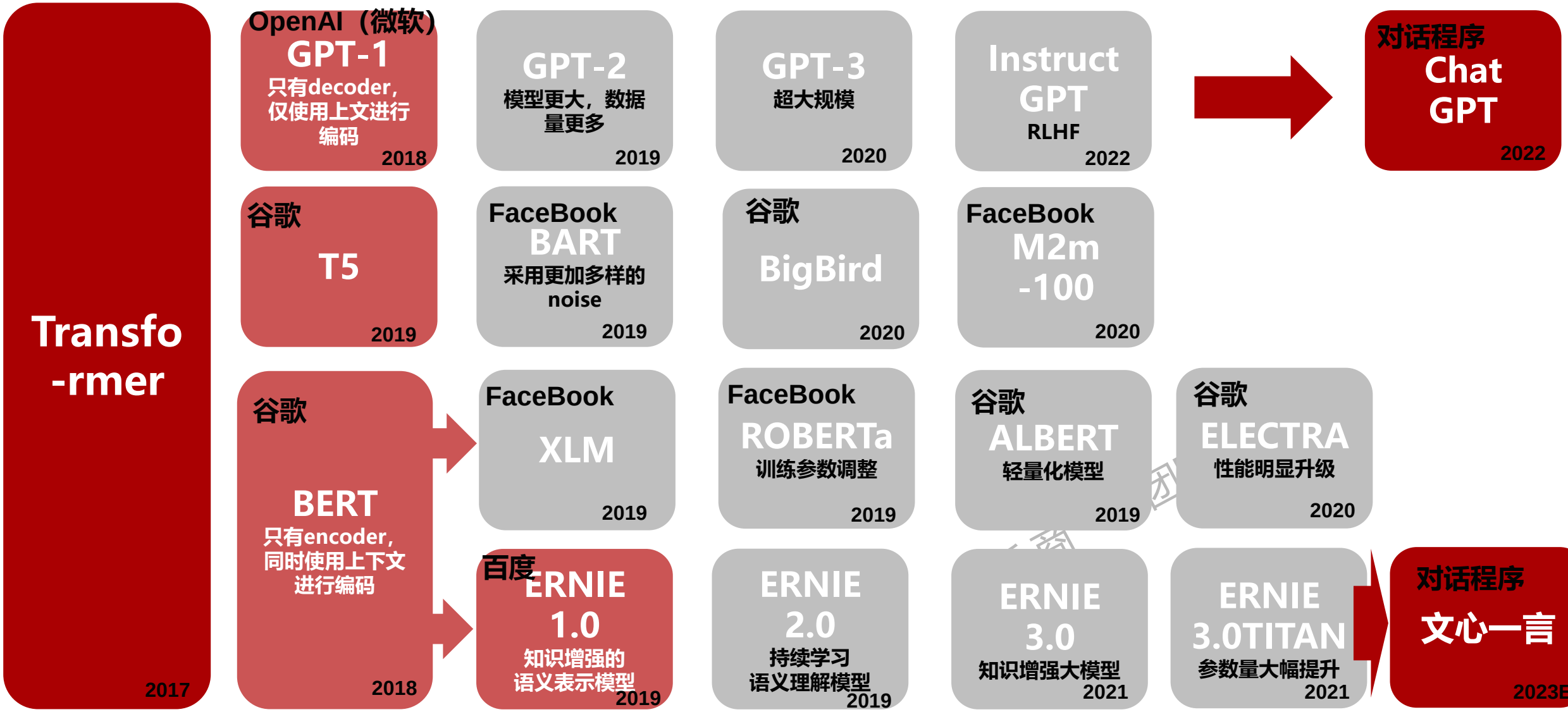
分布式训练需求对深度学习框架提出更高要求



资料来源：CSDN、腾讯云开发者社区、知乎、浙商证券研究所

3、深度学习大模型

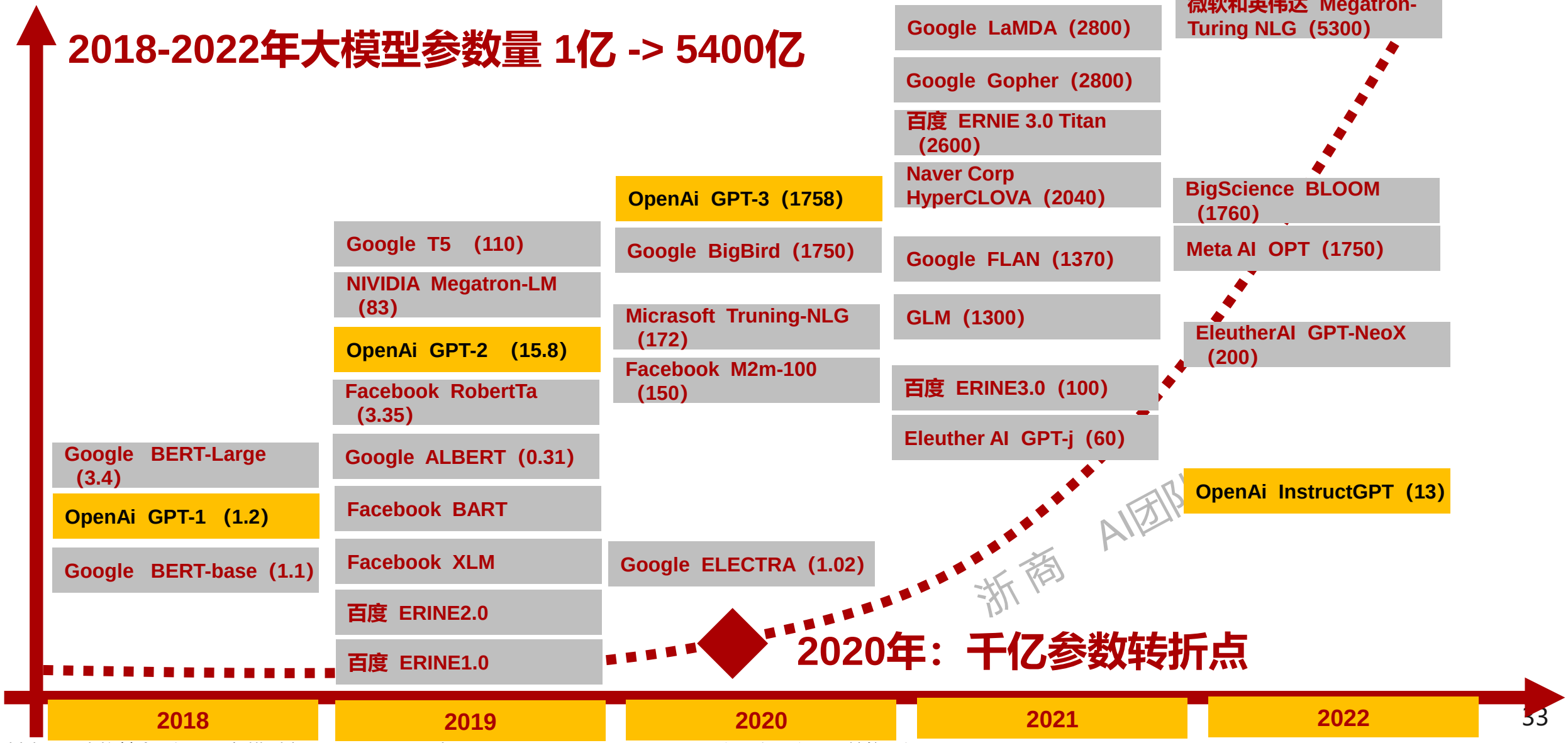
Transformer延伸的四条线



资料来源：真格基金、知乎、各模型官网、arxiv.org 、浙商证券研究所，注：先后顺序按照发布时间进行了重新排序

03 大模型参数迈向千亿时代

2018-2022年大模型参数量 1亿 -> 5400亿

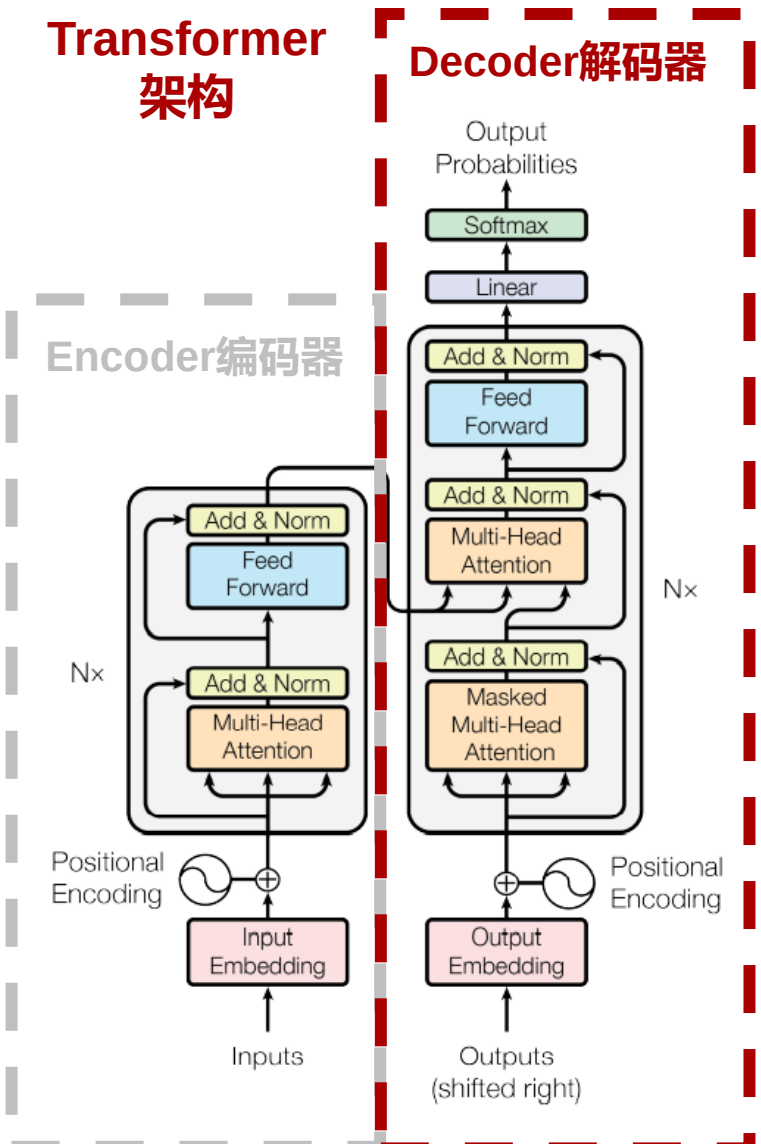


资料来源：真格基金、知乎、各模型官网、arxiv.org、电子工程世界、HuggingFace、浙商证券研究所，单位：亿

03 GPT模型仅保留Transformer的解码器，逐步升级层数和参数量

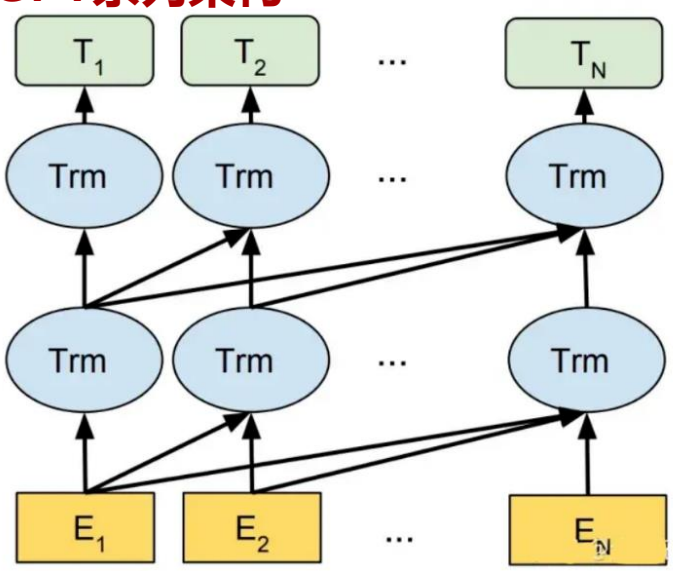
浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

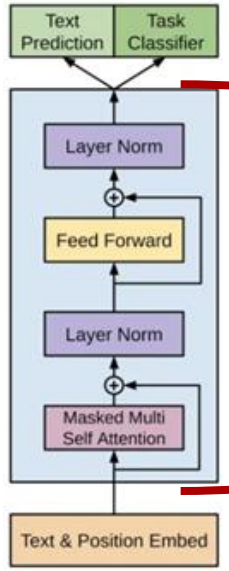
Transformer架构



仅保留 Decoder 解码器

GPT系列架构

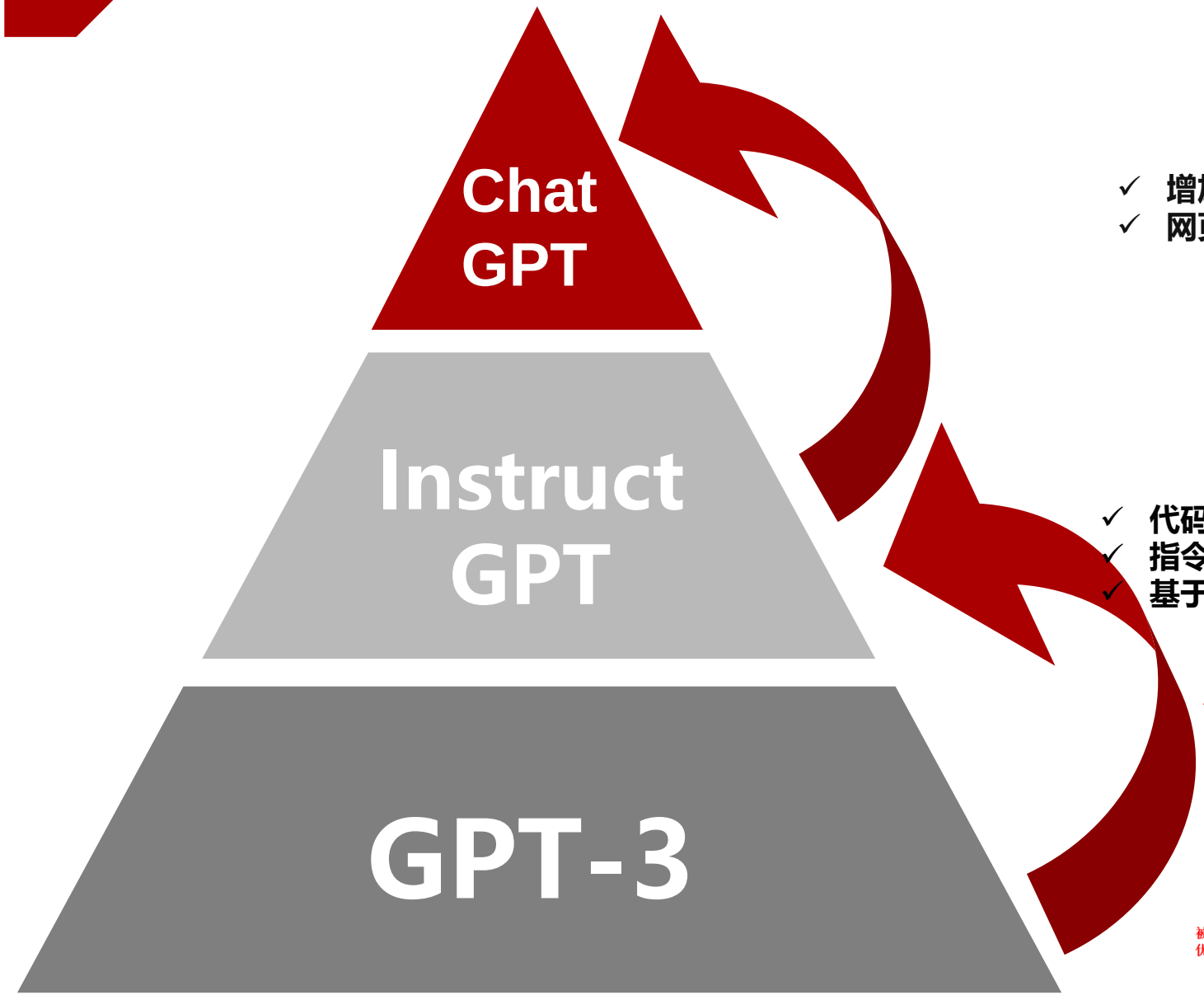




	GPT-1	GPT-2	GPT-3	Instruct GPT
论文年份	2018	2019	2020	2022
Transformer层数	12	48	96	-
参数量	1.2亿	15.8亿	1750亿	13亿
预训练数据量	5GB	40GB	45TB	-

资料来源：CSDN、电子工程世界、浙商证券研究所，注：GPT（Gererate Pre-Training Model）

03 基于InstructGPT形成ChatGPT对话系统

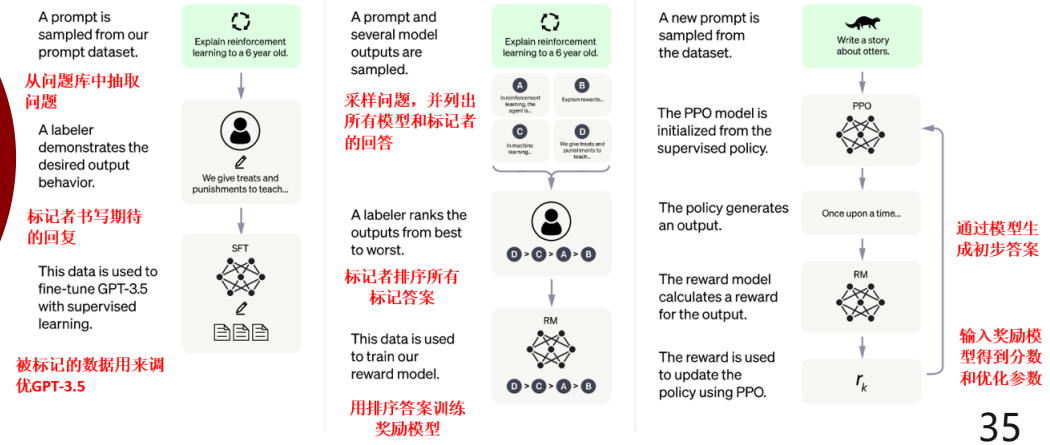


✓ 增加Chat属性
✓ 网页公众测试入口

➢ 略微降低参数量

✓ 代码训练
✓ 指令微调 （instruction tuning）
✓ 基于人类反馈的强化学习 （RLHF）

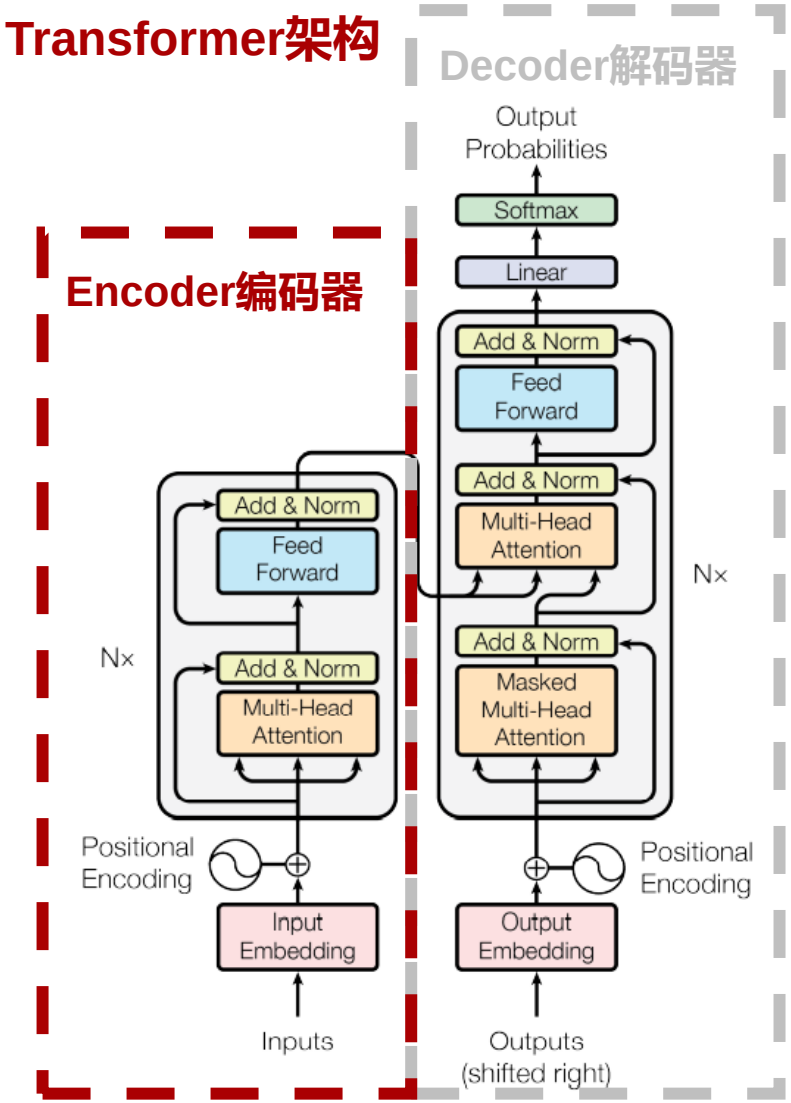
➢ 参数数量降低了100倍（1750亿->13亿）



资料来源：CSDN、电子工程世界、新智元、浙商证券研究所

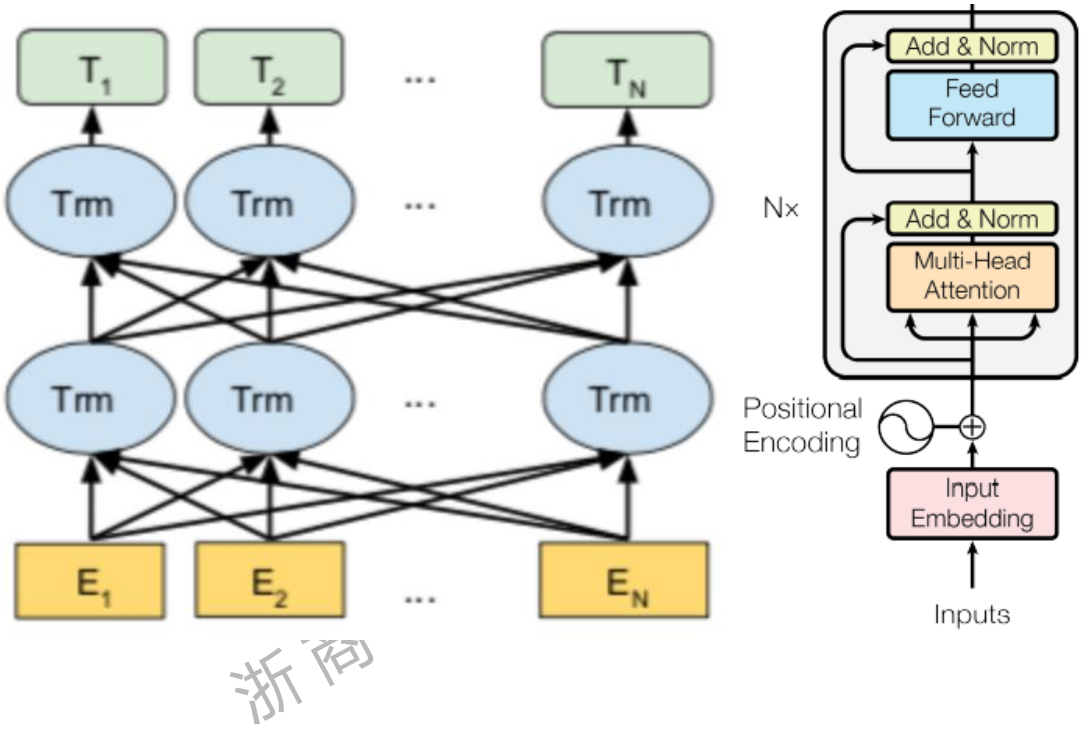
基于BERT衍生百度文心大模型，料将推出对话系统文心一言

Transformer架构



仅保留 Encoder 编码器

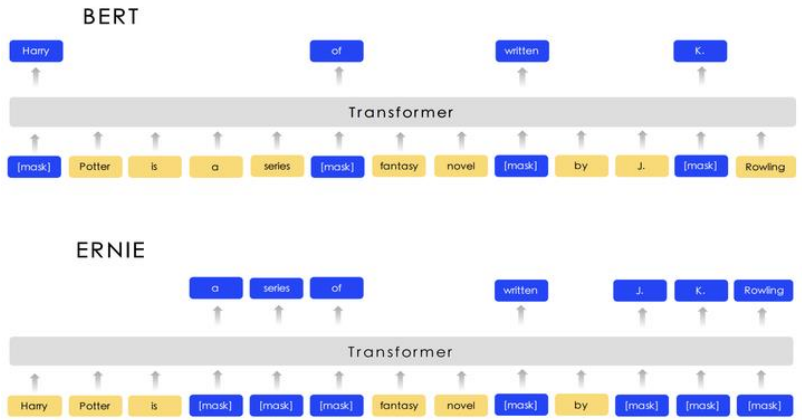
BERT架构



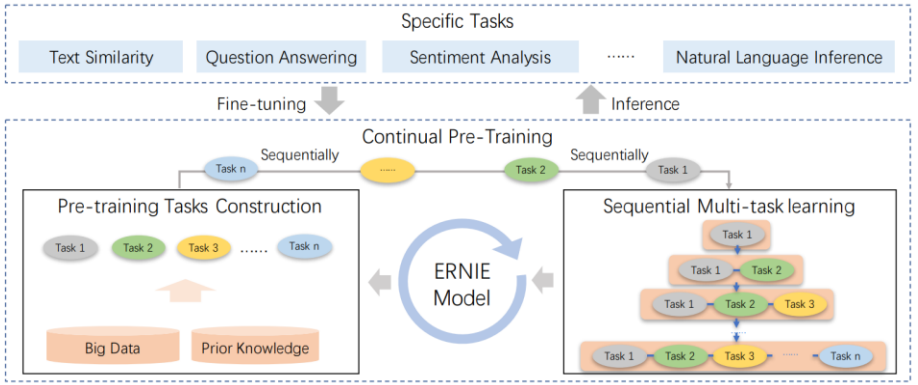
资料来源：CSDN、电子工程世界、浙商证券研究所，注：BERT（Bidirectional Encoder Representation from Transformers）

03 基于BERT衍生百度文心大模型，料将推出对话系统文心一言

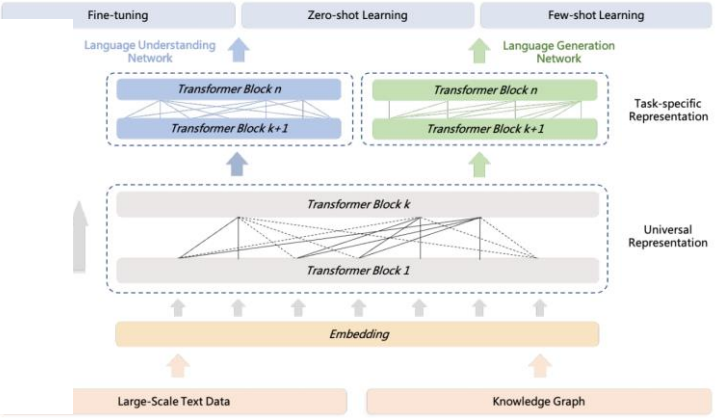
浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD



ERNIE 1.0架构：改进了MLM任务



ERNIE 2.0：+持续学习框架



ERNIE 3.0、3.0TITAN：+参数量

ERNIE版本	1.0	2.0	3.0	3.0 TITAN
论文年份	2019	2019	2021	2021
参数量	参考bert base(1.1亿)	参考bert base(1.1亿)，bert large（3.4亿）	100亿	2600亿
预训练数据量	Wiki，baike，news，tieba	wiki，news，dialogue，IR，discourse relation	4TB	-

资料来源：黑马程序员、知乎、BaiduResearch、浙商证券研究所，注：ERNIE（Enhanced language Representation with Informative Entities）

4、应用

ChatGPT的背后：技术驱动的数字内容新生产方式——AIGC

人工智能技术驱动下，数字内容生产方式向更高效迈进

内容生产总量

内容消费体验

专家生成内容

- PGC：Professionally-Generated Content
- 内容生产总量较低
- 单人/多人专业体验
- 举例：长视频

用户生成内容

- UGC：User-Generated Content
- 内容生产总量变多
- 小规模多人交互
- 举例：小红书、微博

AI辅助生成内容

- AIUGC：AI-assisted Generated Content
- 内容生产总量大
- 大规模沉浸交互
- 举例：语音转文字协助纪要生成、百家号TTV协助新闻图文生成视频

AI生成内容

- AIGC：AI-Generated Content
- 内容生产总量巨大
- 元宇宙式体验
- 举例：ChatGPT、AI绘画、AI生成视频等

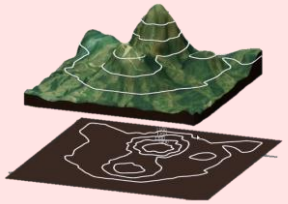
数据来源：浙商证券研究所

04 按技术场景，AIGC内容可分为文本、音频、图像、视频等

01 文本生成

02 音频生成

03 图像生成



04 视频生成

05 跨模态生成

06 策略生成

07 Game AI

08 虚拟人生成

数据、浙商证券研究所

产业链逐步形成，玩家百花齐放，商业模式持续探索

以ChatGPT为代表的AIGC行业，上游主要包括数据供给方、算法/模型机构、创作者生态以及底层配合工具等，中游主要包括文字、图像、音频、视频等数字内容的处理加工方，下游主要是各类数字内容分发平台、消费方及相关服务机构等。

上游

- **数据供给方**（业务数据联通、素材数据收集等，如第三方营销公司、版权图库方等）
- **数据分析及标注**
- **创作者生态**：生物资产及内容素材
- **底层配合工具**：渲染引擎、混音设备等
- **相关算法/模型研究机构**
- **相关开源算法**
- **嵌入/结合能力的业务平台/业务生态**

中游

- **内容设计**：视觉设计、文案设计、结构设计等
- **内容制作工具**：音频/视频编辑工具、新闻采编、游戏制作等
- **周边售卖**：NFT、个性化产品等
- **运营增效**：智能客服、简单决策等
- **个性化市场营销**：广告植入、品牌虚拟形象打造、营销内容生成等
- **数据梳理**：内外部数据分析及更新、数据分析及算法训练
- **自动实时交互**

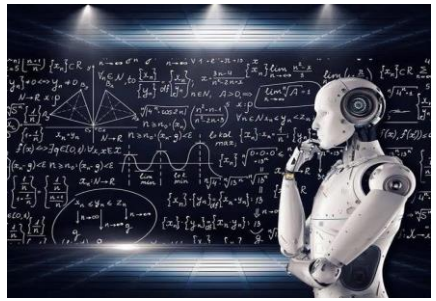
下游

- **各内容创作及分发平台**：基于AICG激发PGC及UGC活力
- **第三方分发渠道**
- **内容终端生产厂商**：新闻媒体机构、金融机构等
- **消费品厂商**
- **AIGC内容检测**
- **第三方内容服务机构**：MCN公司、公关公司等

数据来源：量子位、浙商证券研究所

04 应用领域广泛，核心价值在于优化内容生产的效率与创意

ChatGPT



AIGC

+

01 游戏
02 广告营销
03 影视
04 媒体
05 互联网
06 娱乐
07 其他

=

提升内容生产效率

让创作者拥有一个更加高效的智能创作工具，优化内容创作，大幅提升效率并降低成本；提升创作效率的同时，同样提升了反馈生成效率，有助于实现实时交互内容。

降低内容生产成本

AIGC能够代替人工完成声音录制、图像渲染、视频创作等工作，从而降低内容生产的成本与门槛，使更多用户能够参与到高价值的内容创作流程中。

捕捉激发创作灵感

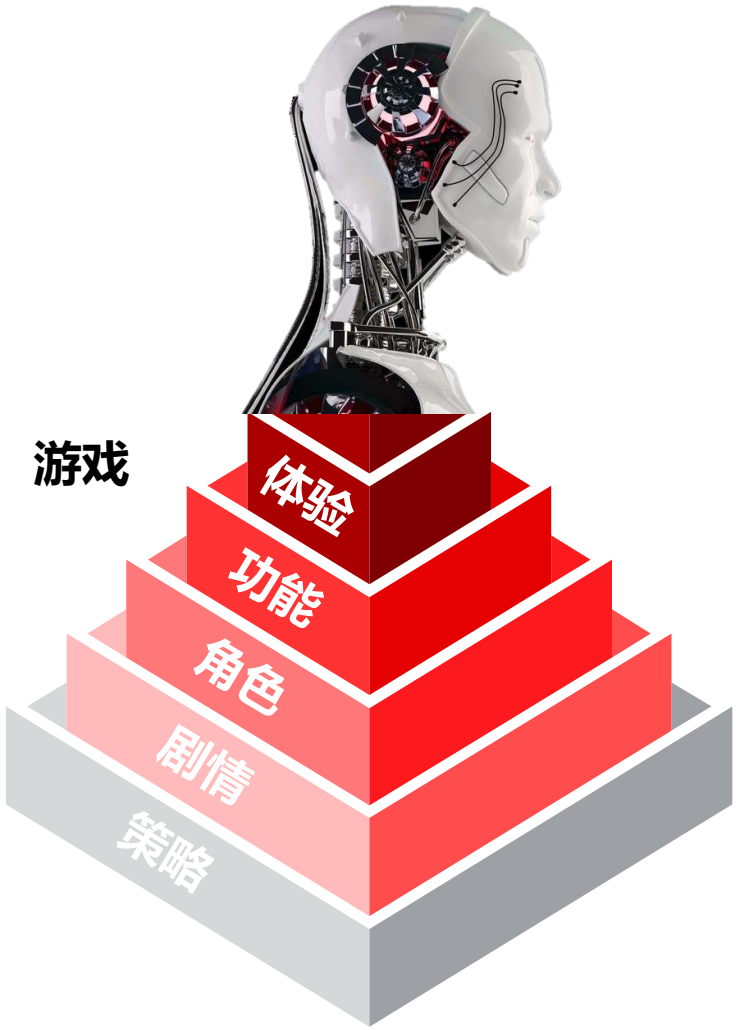
帮助有经验的创作者捕捉灵感，在设计初期生成大量草图，更好的理解创作需求并寻找创作灵感。海量数据提高创造性和开放性，激发创意认知、提升生产多样性。

联动实现数据优化

在与其他特定的数据库（例如实时更新数据、特定主体数据等）或AI系统进行联动后，AIGC能够实现更精准的未来预测或更个性化预测基础上调整生成内容。

数据来源：量子位、浙商证券研究所

ChatGPT/AIGC+游戏：推动游戏生产范式升级，行业新篇章



增加玩家游戏体验

1）对局陪伴。可陪伴玩家进行游戏，包括平衡匹配、冷启动、玩家掉线接管等。
2）特定风格模拟。AI通过模仿职业选手，玩家则感觉像在与真实的职业选手对抗。
3）玩法教学。与玩家在真实对战环境中交流协作，并在过程中向玩家传授职业级的策略与操作技术，帮助玩家迅速熟悉英雄操作与游戏玩法，提高游戏的可玩性。

游戏性能测试

1）前期平衡性测试。AI bot可充分地模拟玩家在某一套数值体系下的游戏体验，提出优化策略，为玩家带来更加平衡的多样性游戏交互。
2）游戏功能测试。通过AI bot针对性的找出游戏中所有交互的可能性，通过发现潜在漏洞辅助游戏策划。

NPC角色AI生成

AI可以创造不同的面孔、服饰、声音甚至性格特征，甚至可同步驱动嘴型、表情等面部变化，达到高度逼真；并通过大量数据模拟人类运动，完成行走、跑步等一系列动作反应。

NPC逻辑及剧情AI生成

AI智能NPC能够分析玩家的实时输入，与玩家动态交互，构建几乎无限且不重复的剧情，增强用户体验并延长游戏生命周期。

游戏策略生成

让AI感知环境、自身状态并基于特定目标决定当下需要执行的动作，基于特定问题和场景，自主提出解决方案。

数据来源：量子位、浙商证券研究所

04 ChatGPT/AIGC+广告营销：24小时虚拟客服，专业可靠智能

01 全天候24小时在线

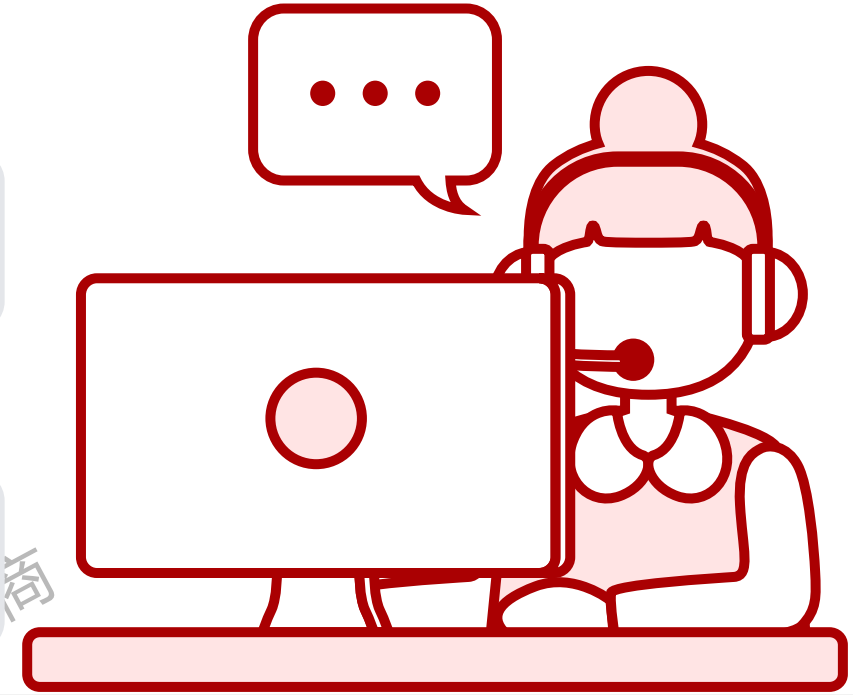
ChatGPT可作为AI驱动的虚拟客服，在广告营销领域为客户提供24小时全天候的客服服务，同时亦能减轻商家人工客服的营销成本。

02 稳定可靠，快速解答

ChatGPT作为虚拟客服相比人工客服更加稳定可靠，能够快速解答客户问题、传递标准化营销话术等，并提升问题解答的准确程度。

03 千人千面，个性化营销推荐

ChatGPT可结合数据及客户的诉求，进行个性化推荐系统的应用给出用户的营销线索，实现更标准、更贴心的用户服务。



数据来源：浙商证券研究所

04 ChatGPT/AIGC+广告营销：内容智能生成，精准个性化营销

多模态广告智能制作

AI可按广告主要求自动生成广告文案；亦可根据广告文案自动生成为广告海报、广告视频，大大降低了广告的制作成本。

多套广告营销解决方案生成

AI可根据目标人群，进行素材分析、抠图、配色等项目，制作多种类型的广告文案/海报/视频，生成多套设计解决方案。

营销内容个性化

AI生成系统与底层的客户数据系统进行数据联通，实时根据数据的反馈，对需求进行针对性调整，由AI快速迭代对营销内容进行更新，提升个性化营销的效率和精准性。

数据来源：量子位、浙商证券研究所

04 ChatGPT/AIGC+影视：赋能影视制作，提升作品质量

影视剧本文稿创作

通过对海量剧本数据进行分析归纳，并按照预设风格快速生产剧本，创作者再进行筛选和二次加工，激发创作者的灵感，缩短创作周期。

提升影视剪辑、后期制作水平

1）实现对影视图像进行修复、还原，提升影像资料的清晰度，保障影视作品的画面质量。2）实现影视预告片自动生成。3）实现将影视内容维度转制，从2D向3D自动转制。

扩展角色和场景创作空间

1）通过AI人脸合成、声音合成实现数字复活已故演员、替换“劣迹艺人”、演员角色年龄的跨越、高难度动作合成等，减少演员自身局限对影视作品的影响；
2）通过人工智能合成虚拟物理场景，将无法实拍或成本过高的场景生成出来，拓宽影视作品想象边界，带来更优质的视觉效果和听觉体验。

影视发行智能审核、用户端个性化推荐

1）通过自然语言处理NLP和深度学习DL实现视频审核和视频传播技术；
2）用户端实现视频自主互动、弹幕防挡。

数据来源：中国信息通信研究院、量子位、1905电影网、浙商证券研究所

04 ChatGPT/AIGC+媒体：推进人机协作共生，媒体效率提升

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

新闻采编环节 提高内容制作效率

实现采访录音语音转写

借助语音识别技术将录音语音转写成文字，有效压缩重复工作，保障新闻时效性。

实现智能新闻写作

提升新闻资讯的时效。基于算法自动编写新闻，将工作自动化，更快、更准、更智能化地生产内容。

实现智能视频剪辑

提升视频内容的价值。通过使用视频字幕生成、视频锦集、视频拆条、视频超分等视频智能化剪辑工具，节省成本，最大化版权内容价值。

新闻传播环节 播报高效智能化

应用范围不断拓展

目前新华社、中央广播电视总台、人民日报社、湖南卫视等积极布局，推出“新小微”、“小C”等虚拟新闻主持人。

应用场景不断升级

除了常规的新闻播报，AI合成主播开始陆续支持多语种播报和手语播报，不断升级应用场景。

应用形态日趋完善

在形象方面，逐步向3D拓展；在驱动范围上，向面部表情、肢体、手指、背景内容素材延伸；在内容构建上，向智能化生产探索。

新闻主体影响 智媒影响产业及生活

对传媒机构产生深刻营销

AIGC大幅提高生产效率，带来新的视觉化、互动化体验，推动传媒向智媒转变。

对传媒从业者产生深刻影响

AIGC助力生产更具人文关怀、社会意义和经济价值的新闻作品，并将部分劳动性的采编播工作自动化。

对传媒受众产生深刻影响

AIGC使其在更短时间内获得以更丰富多元的形态呈现的新闻内容，也降低了传媒行业的技术门槛，极大增强其参与感。

数据来源：中国信息通信研究院、浙商证券研究所

04 ChatGPT/AIGC+互联网：重构搜索引擎，更便捷高效多元

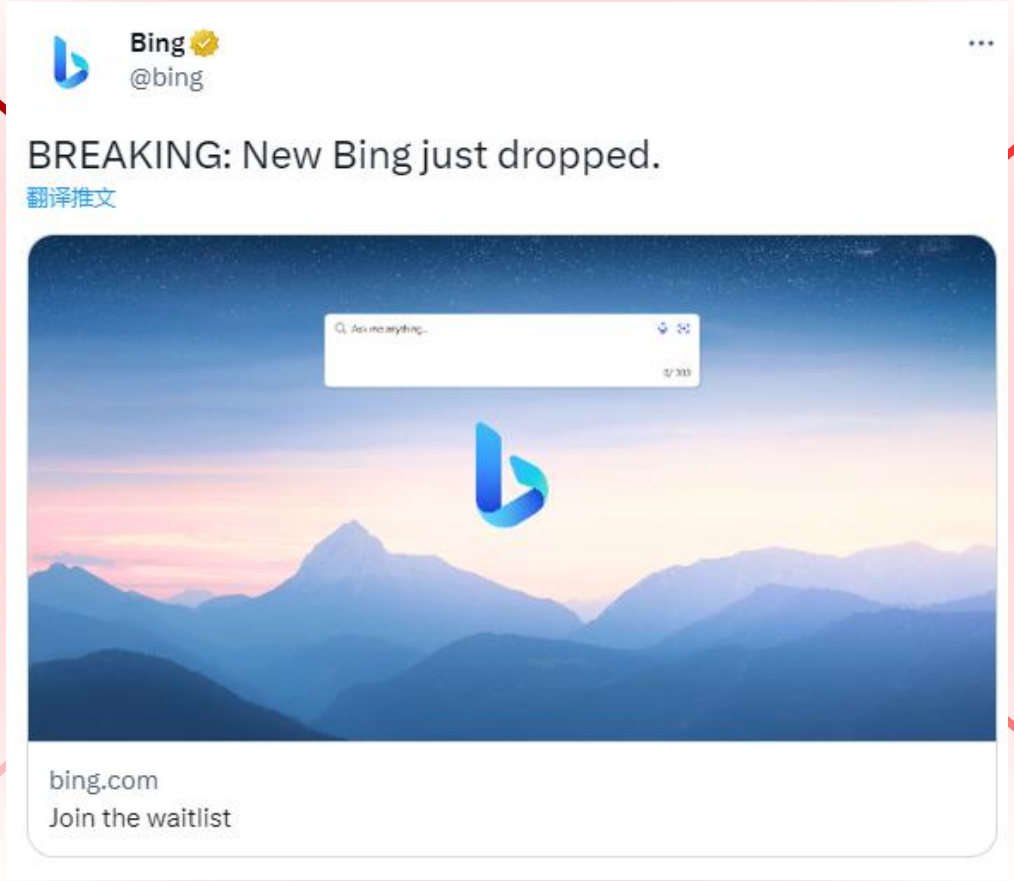
更好的搜索体验

改进用户搜索体验，在一些简单的事情如体育比分、股票价格和天气等，新必应会提供更相关结果，同时显示更全面的答案。

全新的交互式聊天体验使用户能够通过询问更多细节、清晰度和想法来优化搜索，直到获得正在寻找的完整答案，并提供可用链接。

全新的聊天体验

图：微软旗下搜索引擎集成ChatGPT



更完整的答案

审查从网络上搜索到的结果找到并总结答案。例如问题「如何用鸡蛋代替蛋糕中的另一种成分」，新版必应能够给出详细说明。

可帮助用户获得灵感，例如可以帮助用户编写电子邮件、规划旅游行程、准备工作面试等；还引用了信息所有来源，用户可详细查看链接。

激发创造性火花

数据来源：机器之心、Bing、浙商证券研究所

ChatGPT/AIGC+互联网：电商购物虚实结合，沉浸式体验

01 生成商品3D模型 用于商品展示和虚拟试用

基于不同角度的商品图像，借助视觉生成算法自动化生成商品的3D几何模型和纹理，辅以线上虚拟“看、试、穿、戴”，提供接近实物的差异化网购体验，助力高效提升用户转化。

图：优衣库4D试衣间

02 打造虚拟主播 提升直播带货效率

1）为观众提供24小时不间断的货品推荐介绍，增加商品商铺曝光度；

2）推进店铺品牌年轻化科技化进程；

3）虚拟主播稳定性强，行为言谈可根据品牌方要求个性化定制，失误率低。

图：快手虚拟主播与真人主播场景互动

03 线上线下商城加速演变 打造全新购物场景

通过二维图像的三维重建，实现虚拟货场快速、低成本、大批量的构建，有效降低商家搭建3D购物空间的门槛及成本，为消费者提供新消费体验。

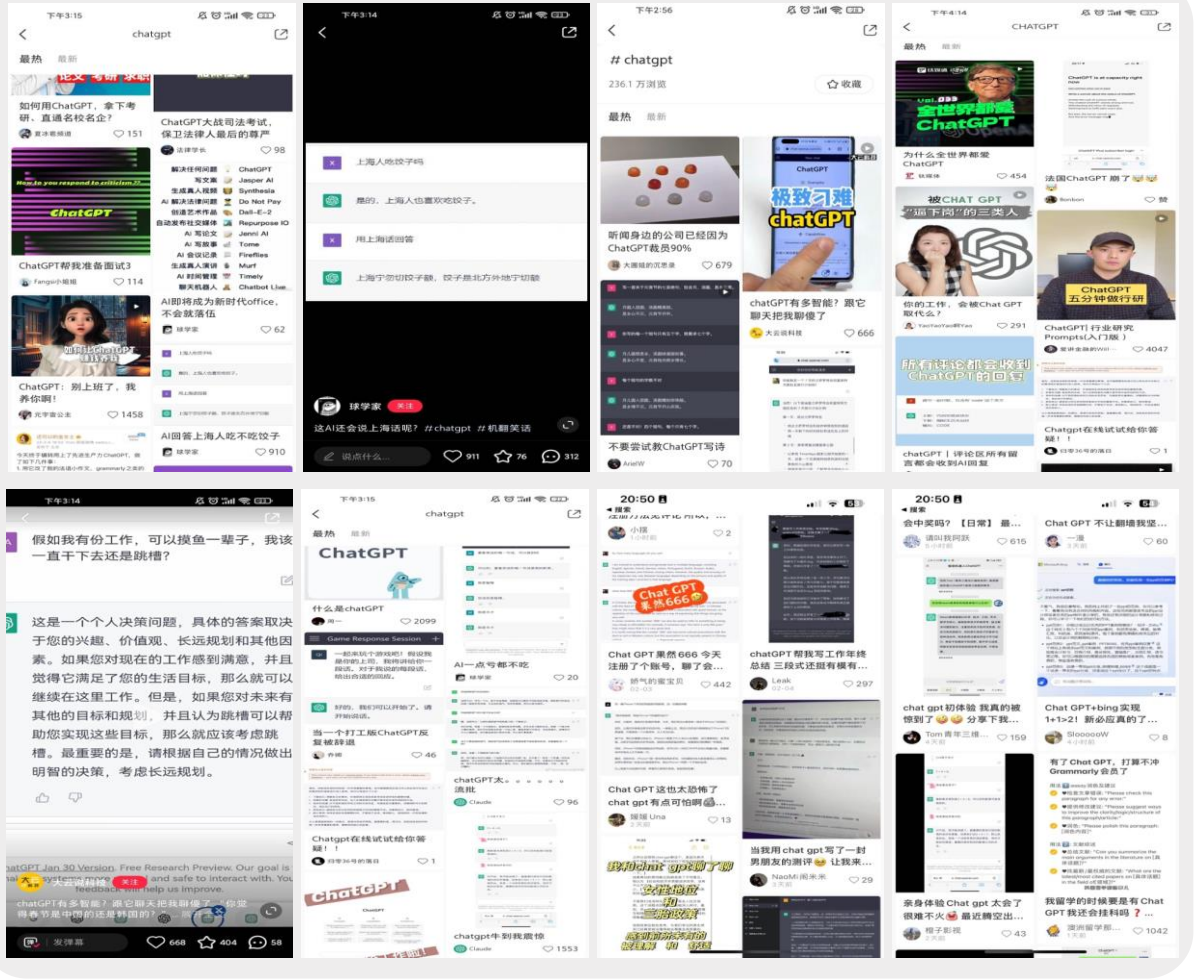
图：潮牌Vans在游戏Roblox的店铺展览

数据来源：中国信息通信研究院、优衣库APP、电商报、Roblox官网、浙商证券研究所

04 ChatGPT/AIGC+娱乐：人机交互新台阶，有趣有料高热度

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

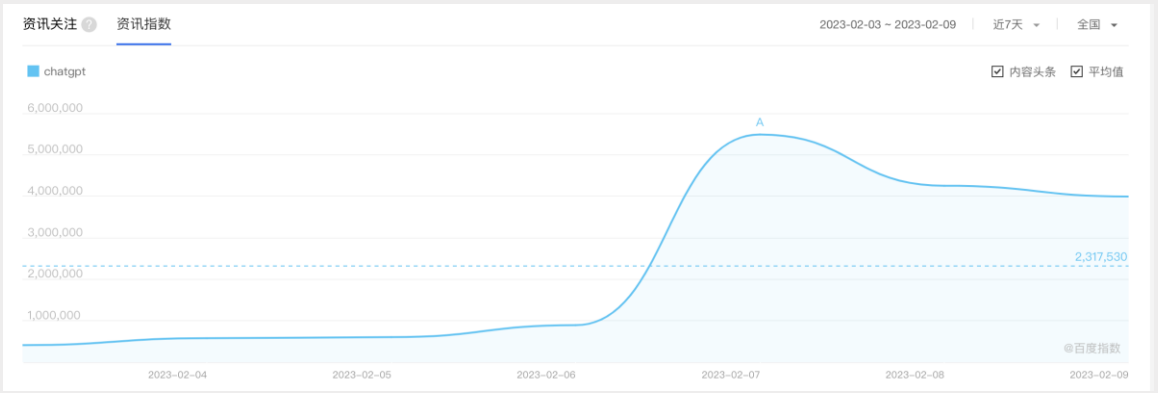
ChatGPT有趣有料，人机交互娱乐迈入新台阶



ChatGPT火爆全网高热度，2个月活跃用户破亿

图：ChatGPT百度搜索指数趋势图（2022/02/02-2023/02/09

图：ChatGPT百度资讯指数趋势图（2022/02/02-2023/02/09



数据来源：微博、小红书、百度指数、浙商证券研究所

04 ChatGPT/AIGC+娱乐：降低内容生产门槛，助力元宇宙加速

实现趣味性图像或音视频生成，激发用户参与热情

- 1）图像视频生成，极大满足用户**猎奇需求**；
- 2）语音合成，变声增加**互动娱乐性**。

打造虚拟偶像，释放IP价值

- 1）与用户共创合成歌曲，加深粉丝黏性；
- 2）合成音视频动画，支撑虚拟偶像在更多元的场景进行内容变现。

开发C端用户数字化身

- 各大科技巨头积极探索与加速布局“虚拟数字世界”与现实世界大融合的“未来”。

数据来源：中国信息通信研究院、浙商证券研究所

04 ChatGPT/AIGC+其他：加快数实融合，产业升级提速

教育+

AIGC赋予教育材料新活力，为教育工作者提供了新的工具，使原本抽象、平面的课本具体化、立体化。

金融+

AIGC助力实现降本增效。1）实现金融资讯、产品介绍视频内容的自动化生产，提升效率；2）塑造视听双通道的虚拟数字人客服。

医疗+

AIGC赋能诊疗全过程。1）辅助诊断，可用于改善医学图像质量、录入电子病历等；2）康复治疗，为失声者合成语言音频，为残疾者合成肢体投影等。

工业+

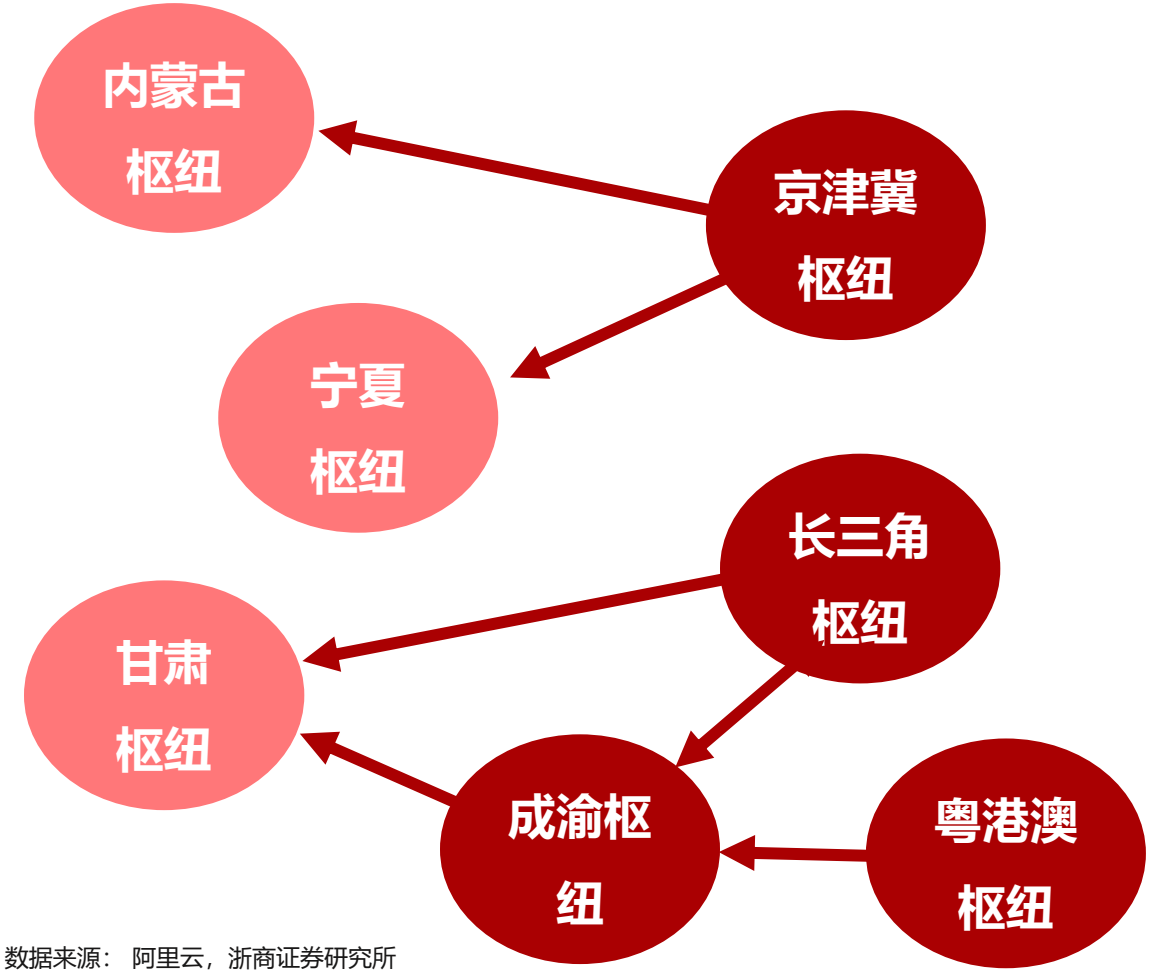
AIGC提升产业效率和价值。1）融入计算机辅助设计CAD，极大缩短工程设计周期；支持生成衍生设计，实现动态模拟；2）加速数字孪生系统的构建，高效创建数字孪生系统。

数据来源：中国信息通信研究院、浙商证券研究所

5、通信

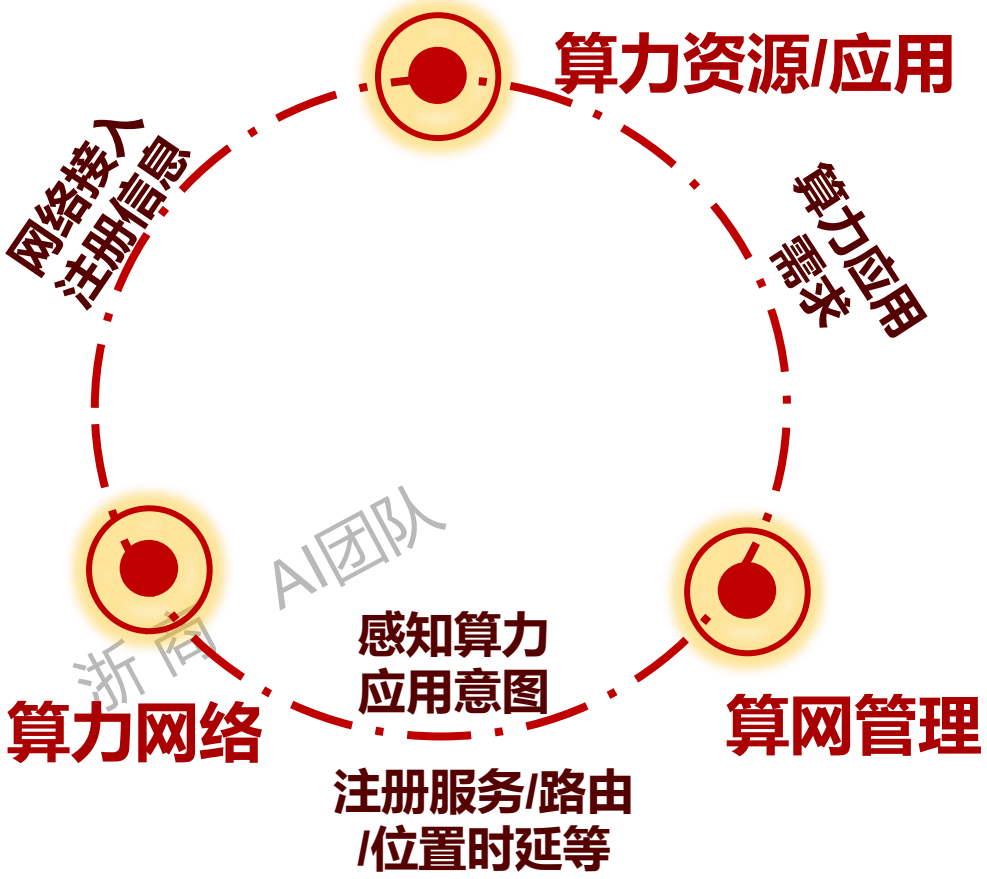
05 算力调度：提高算力使用效率，电信运营商打造算力统一调度平台

智能调度，提高算力使用效率



数据来源： 阿里云，浙商证券研究所

算力调度平台



05 算力供给：为应对不同应用场景需求，AIGC将推动智能算力中心大发展

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

1	数据中心	X86、ARM服务器	奥飞数据	■ **基础算力：**当前主流通用计算模式 ■ **应用场景：**电商、短视频等
			数据港	
2	智算中心	AI、GPU型服务器	运营商	■ **智能算力：**80%以上非结构化数据处理需要多样化算力 ■ **应用场景：**AIGC、无人驾驶等
			华为	
3	超算中心	超级计算机	中科曙光	■ **超算算力：**超级计算机供给算力，算力规模极高 ■ **应用场景：**科学计算、AI
			联想	

资料来源：浙商证券研究所

05 算力设备：向高速率多样化演进，AIGC有望催动国产化替代进程

服务器

- **方向：AI、GPU型服务器**
- **代表厂商：**浪潮信息、紫光股份、中兴通讯、中科曙光等

交换机

- **方向： 400G/800G高速率交换**
- **代表厂商：**锐捷网络、紫光股份、中兴通讯等

以太网芯片

- **方向： 25G/200G等国产芯片**
- **代表厂商：**裕太微等

光模块

- **方向：硅光、 CPO新型技术**
- **代表厂商：**天孚通信、中际旭创等

光芯片

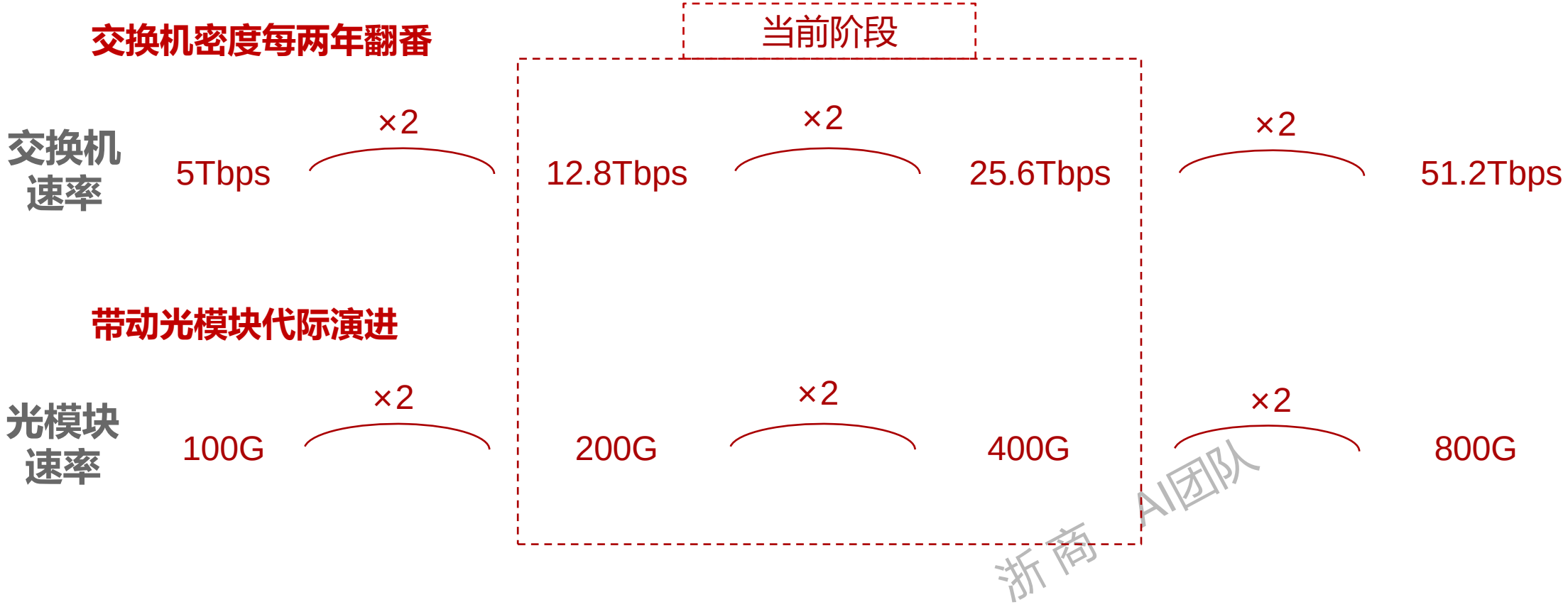
- **方向： 25G/50G/100G等国产光芯片**
- **代表厂商：**源杰科技、光迅科技等

光器件

- **方向：新型调制解调器**
- **代表厂商：**天孚通信、光库科技等

数据来源： 浙商证券研究所

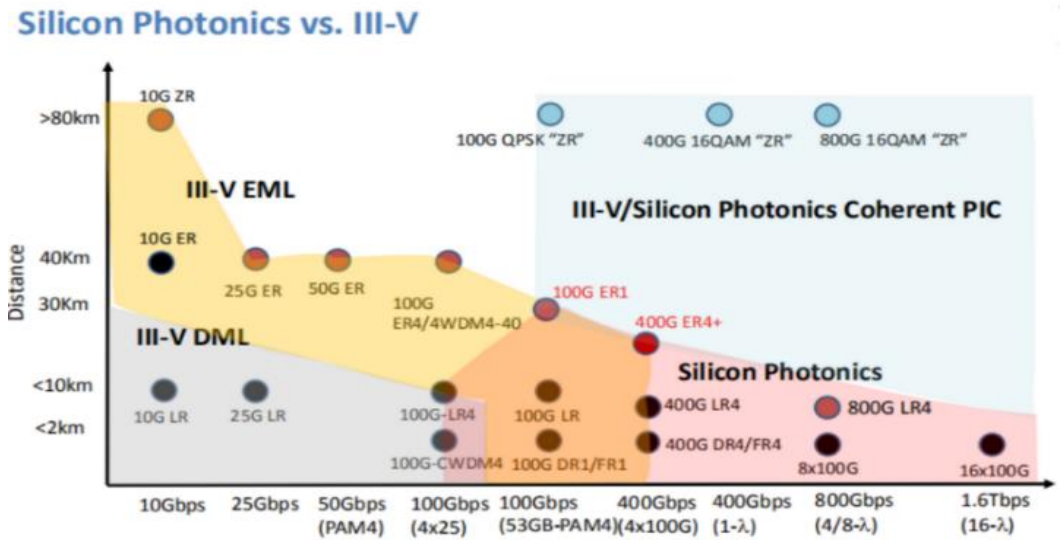
05 算力设备：光模块作为算力核心器件，有望明显受益

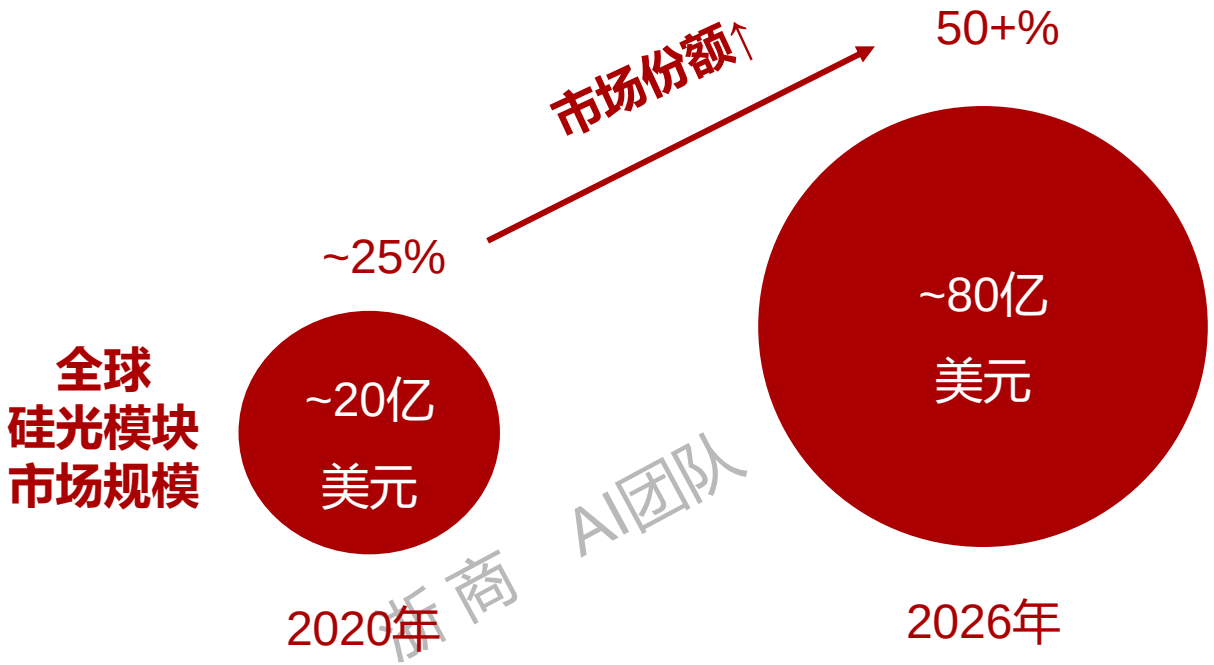


资料来源：Yole，浙商证券研究所

算力设备：光模块作为算力核心器件，有望明显受益

硅光模块：将光学器件与电子元件整合在一个独立微芯片中，硅片上用光取代铜线作为信息传导介质，高集成/低成本/低功耗，在高速率场景具备优势





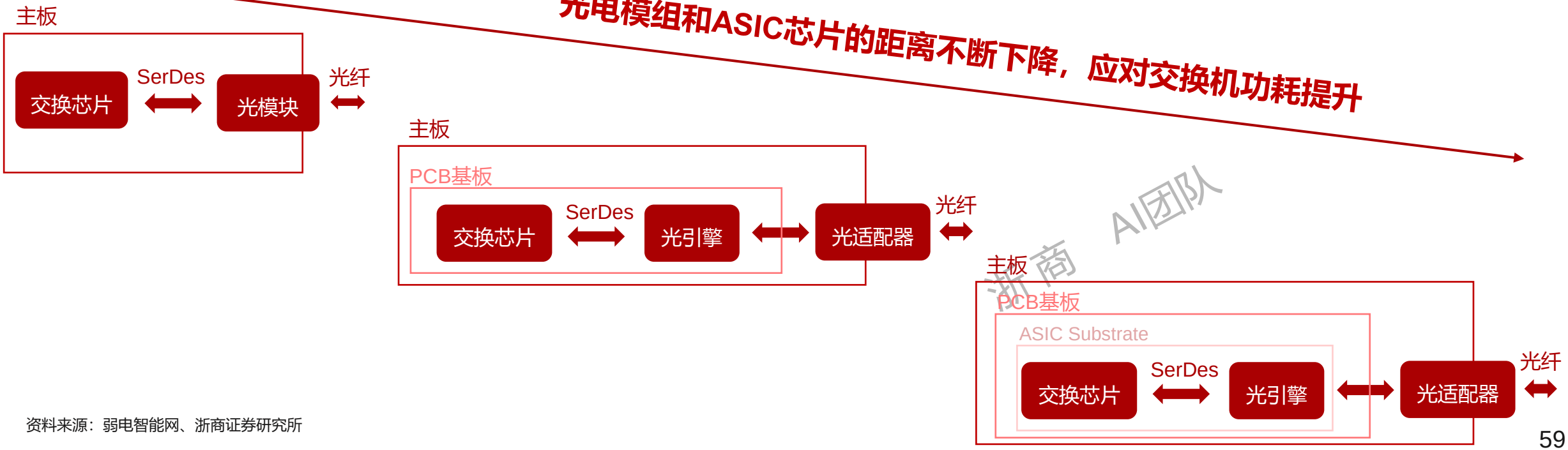
资料来源：讯石光通讯，SiFotonics Technologies，Lightcounting，通信产业报，浙商证券研究所

算力设备：光模块作为算力核心器件，有望明显受益

光模块封装工艺演进：CPO有望成为主流

可插拔 → NPO → CPO

光电模组和ASIC芯片的距离不断下降，应对交换机功耗提升



资料来源：弱电智能网、浙商证券研究所

算力配套：散热向芯片级液冷演进

设备越来越贴近核心发热源

房间级 → 机柜级 → 芯片级

更加高效的冷却介质

风冷 → 液冷

资料来源：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。

板块	建议关注的公司
芯片算力	海光信息、景嘉微、龙芯中科、中国长城、安路科技、复旦微电、紫光国微、寒武纪、澜起科技、德科立、天孚通信、中际旭创
深度学习框架	百度 、海天瑞声、商汤科技、微软、谷歌、Meta
深度学习大模型	百度、科大讯飞、商汤科技、谷歌、微软
应用	百度、腾讯、阿里巴巴、网易、昆仑万维、阅文集团、捷成股份、视觉中国、风语筑、中文在线、三七互娱、吉比特、天娱数科
通信	底层基础算力设施：算力调度（运营商）、算力供给（运营商、奥飞数据、数据港）、算力设备（浪潮信息、联想集团、紫光股份、中兴通讯、锐捷网络、天孚通信、光库科技、中际旭创、新易盛）、算力配套（英维克、高澜股份）

行业评级与免责声明

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

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

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