浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

华为昇腾服务器研究框架

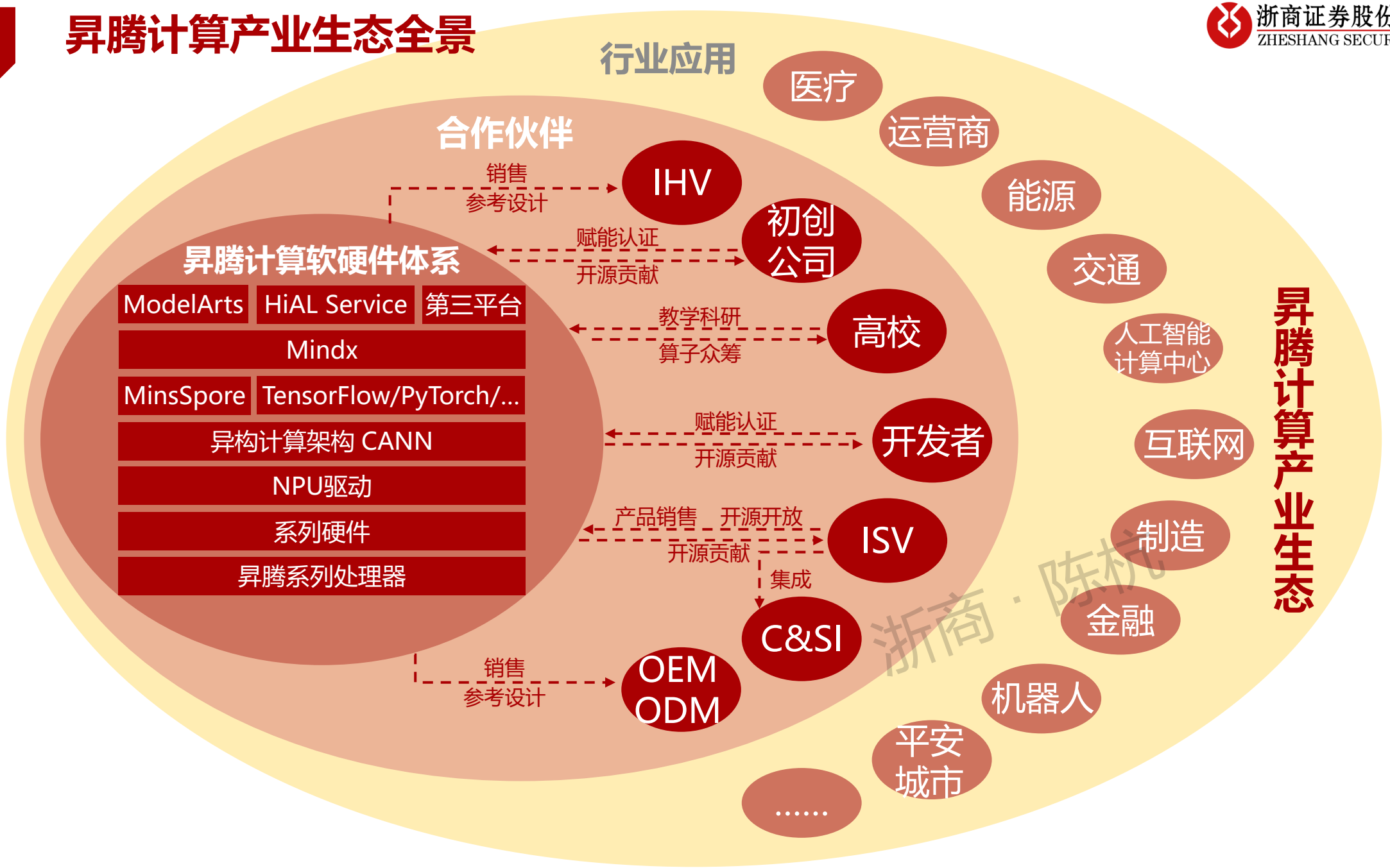
华为算力产业链深度系列研究

行业评级：看好

2023年03月29日

分析师	陈杭	研究助理	安子超
邮箱	chenhang@stocke.com.cn	邮箱	anzichao@stocke.com.cn
证书编号	S1230522110004	电话	18611396466

01 昇腾计算产业生态全景



资料来源：昇腾社区官网，浙商证券研究所

摘要

站在当前科技重大转折点——ChatGPT时刻，科技企业的竞争日益聚焦于AIGC领域，大模型成为核心竞争力的重要体现。华为作为国内科技龙头，2021年发布的盘古大模型有望在AIGC时代中引领潮流。其中，昇腾全栈AI软硬件平台构筑智能世界的基石，昇腾计算产业生态不断发展完善，为盘古大模型提供了底层算力支撑。

1、昇腾系列硬件：AI处理器+多产品形态

- 以昇腾AI处理器为基础，通过模块、标卡、小站、服务器等丰富的产品形态，打造面向“端、边、云“的全栈解决方案，为整个昇腾AI产业的底层核心支撑。

2、CANN：针对AI场景提出的异构计算架构

- 作为华为昇腾AI基础软硬件平台的核心，CANN向上支持多种AI框架，向下服务AI处理器与编程，以极致性能、极简开发、开放生态为目标，助力昇腾构建全场景人工智能平台。

3、MindSpore：面向全场景的AI计算框架

- 国内首个支持千亿参数大模型训练AI框架，支持终端、边缘、云全场景灵活部署，开创全新的AI编程范式，致力于实现开发态友好、运行态高效、全场景按需协同三大目标，降低AI开发门槛。

4、MindX：应用使能=2个组件+1个模型库+多行业SDK

- 深度学习组件MindX DL以提供参考架构的方式，供业界平台伙伴快速开发商永版本的深度学习系统；智能边缘组件MindX Edge可实现快速将云端模型推送至边缘端部署，并将边缘侧未识别数据上传至云端进行增量训练；ModelZoo将AI开发模型提前调优后提供给开发者进行选择。2个组件及1个模型库支撑行业AI应用落地，SDK凝聚行业知识、结合AI实践。

5、生态构建及应用落地：多领域合作，10+行业应用，多方面解决方案

- 携手赋能各领域合作伙伴和开发者，目前围绕昇腾计算体系，已在能源、金融、公共、交通、电信、制造、教育等多行业实现应用，提供城市智能中枢、昇腾智巡、昇腾智行、昇腾制造等解决方案。

建议关注标的：神州数码、拓维信息、麒麟信安、软通动力、常山北明、海量数据、润和软件、英方软件

资料来源：昇腾社区官网，昇腾计算产业发展白皮书，浙商证券研究所

风险提示

1、宏观经济下行风险

2、上游晶圆紧缺加剧的风险

3、市场发展不及预期的风险

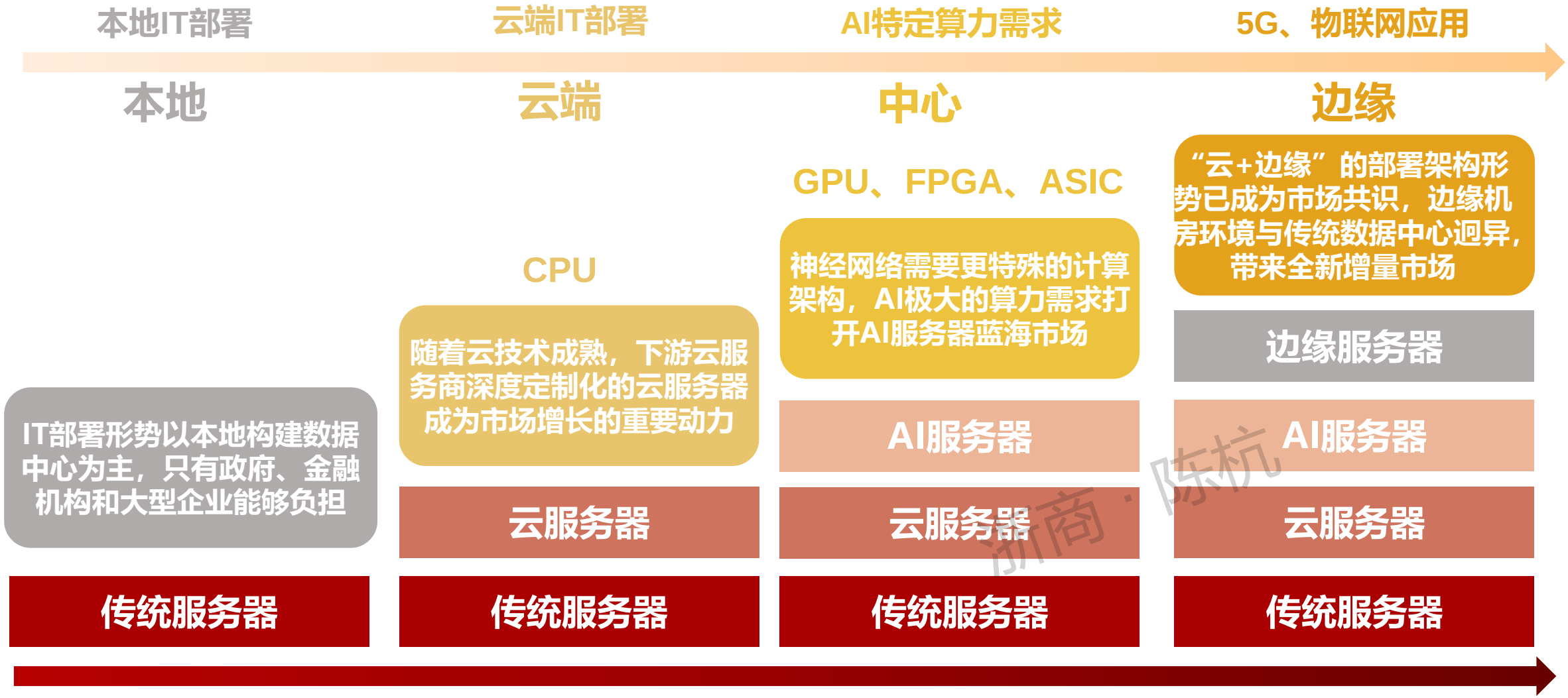
4、行业竞争超预期风险

目录

CONTENTS

01 昇腾AI概览

服务器行业演进历程



资料来源：行行查研究中心，浙商证券研究所

传统服务器

处理器模块　存储模块
网络模块　电源、风扇等

高性能吞吐计算能力

关注工作量总和

云计算服务器

云处理器模块　存储处理模块
网络处理模块　系统件模块

传统服务器业务 VS 云服务器租用业务

	传统服务器业务	云服务器租用业务
投入成本	高额的综合信息化成本投入	按需付费，有效降低综合成本
产品性能	难以确保获得持续可控的产品性能	硬件资源的隔离+独享带宽
管理能力	日趋复杂的业务管理难度	集中化的远程管理平台+多级业务备份
扩展能力	服务环境缺乏灵活的业务弹性	快速的业务部署与配置、规模的弹性扩展能力

资料来源：搜狗百科，浙商证券研究所

昇腾全栈AI软硬件平台，构筑智能世界的基石

行业应用：能源、金融、公共、交通、电信、制造、教育等更多行业应用

应用使能：
- ModelArts
- HiAI Service
- 第三方平台
- MindX DL：深度学习使能
- MindX Edge：智能边缘使能
- ModelZoo：优选模型库
- MindX SDK：行业SDK

全流程开发工具链
- MindStudio：算子、模型、应用一站式高效开发和极简部署

管理运维工具
- FusionDirector：服务器全生命周期智能运维的管理软件
- SmartKit：华为IT产品运维服务工具平台

AI框架：
- MindSpore：匹配昇腾AI处理器算力的全场景深度学习框架
- TensorFlow/Py Torch：可基于第三方框架开发的模型进行二次开发、训练和推理

异构计算框架：
- CANN：统一异构计算架构，释放昇腾硬件澎湃算力

Atlas 系列硬件：昇腾系列AI处理器 基于 模块 / 标卡 / 小站 / 服务器 / 集群

资料来源：鲲鹏社区官网，昇腾计算产业发展白皮书，浙商证券研究所

Atlas人工智能计算解决方案，覆盖深度学习全流程

超强算力
达芬奇架构
最强AI算力底座

更优效能
极致散热技术
更高能效比

开放易用
端边云协同
软硬件开放

安全可信
从芯片到系统
构建可信AI平台

关键特性

昇腾系列AI处理器 → **Atlas人工智能计算解决方案** ← **基础软件**

“开放、简单、可信”的AI解决方案

昇腾计算产品

AI模块

AI加速卡

智能边缘

AI服务器

AI集群

资料来源：华为官网，浙商证券研究所

昇腾AI夯实智能根基，生态蓬勃向上

2020 → 2022

硬件开放

5 → 20+

硬件合作伙伴

软件开源

10.8万+ → 320万+

MindSpore社区下载量

使能伙伴

159 → 1000+

合作伙伴

解决方案

222 → 1600+

解决方案

发展人才

4.5万+ → 100万+

开发者

资料来源：华为全联接大会2022，浙商证券研究所

02 硬件产品

昇腾AI处理器

AI模块+AI加速卡+智能边缘+AI服务器+AI集群

昇腾AI处理器构建底座，提供强大算力支持

昇腾 310

华为首款全栈全场景人工智能芯片

自研华为达芬奇架构NPU

在8W数据精度下算力可达16TOPS

高性能 3D Cube

架构	• 华为达芬奇
设计	• 3D Cube
性能	• 16TOPS@INT8 和 8TOPS@FP16
最大功耗	• 8W
工艺	• 12nm FFC

昇腾 910

算力最强AI处理器

自研华为达芬奇架构NPU

640 TOPS@INT8，320TFLOPS@FP16

最大功耗310W

架构	• 华为达芬奇
设计	• 3D Cube
性能	• 320TFLOPS@FP16 和 640TOPS@INT8
最大功耗	• 310W
工艺	• N7+

资料来源：海思官网，浙商证券研究所

昇腾AI硬件体系完善，提供多样化算力选择

AI模块

开发者套件

芯片：昇腾310

最高算力：22 TOPS

AI加速模块

芯片：昇腾310

最高算力：22 TOPS

加速卡

推理卡

芯片：昇腾310

最高算力：88 TOPS

训练卡

芯片：昇腾910

最高算力：280 TFLOPS

智能边缘

智能小站

芯片：昇腾310

最高算力：22 TOPS

边缘服务器

芯片：鲲鹏920

最高算力：352 TOPS

AI服务器

推理服务器

2*鲲鹏920

最高算力：704 TOPS

训练服务器

8*昇腾910+4*鲲鹏920

最高算力：2.24 PFLOPS

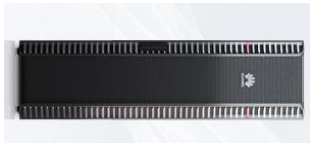
AI集群

AI集群

数千颗昇腾910

算力：256P～1024P FLOPS

AI集群基础单元



64*昇腾910+32*鲲鹏920

形态：47U机柜

数据来源：华为昇腾官网，浙商证券研究所

03

软件框架

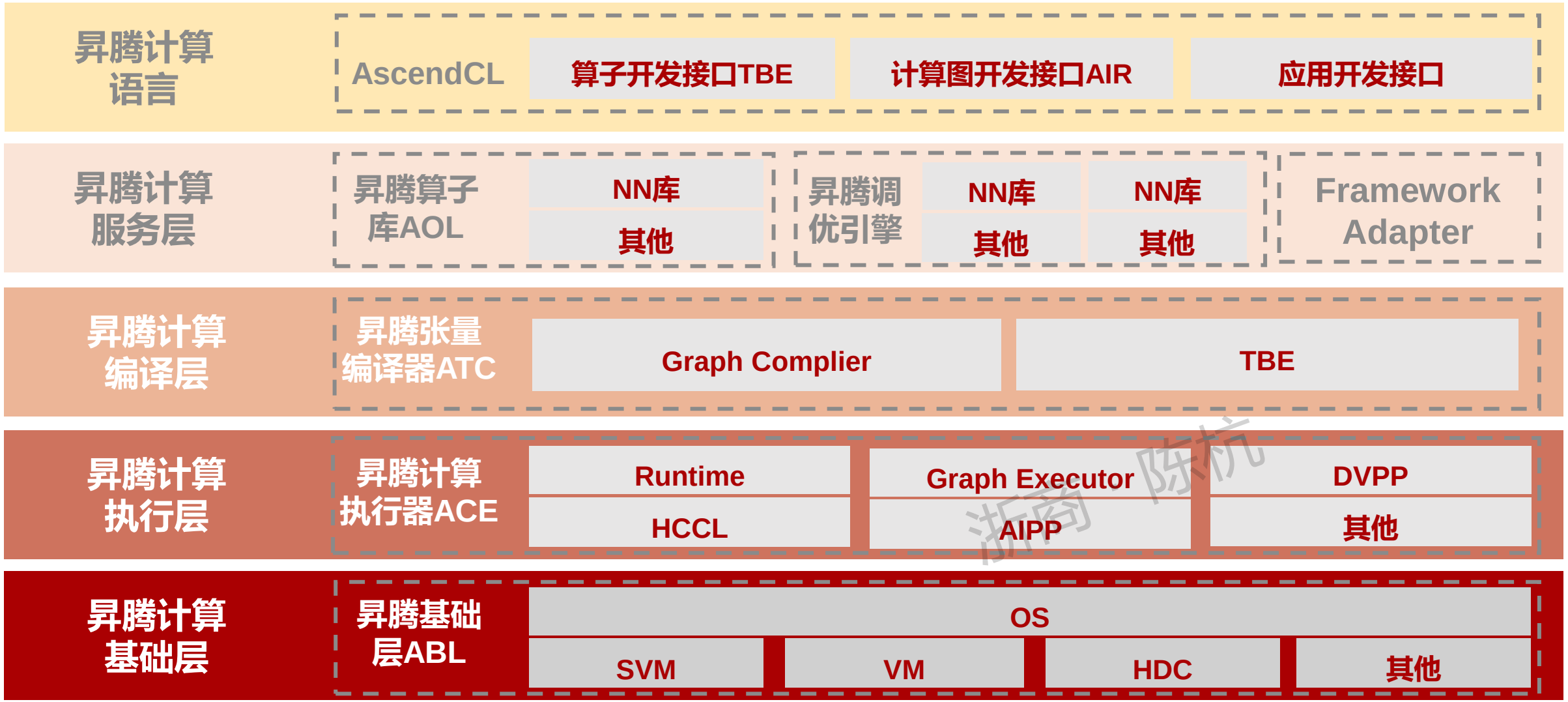
CANN异构计算架构

MindSpore人工智能框架

MindStudio开发工具链

MindX应用使能

03 CANN：软硬协同释放澎湃算力



数据来源：华为昇腾官网，浙商证券研究所

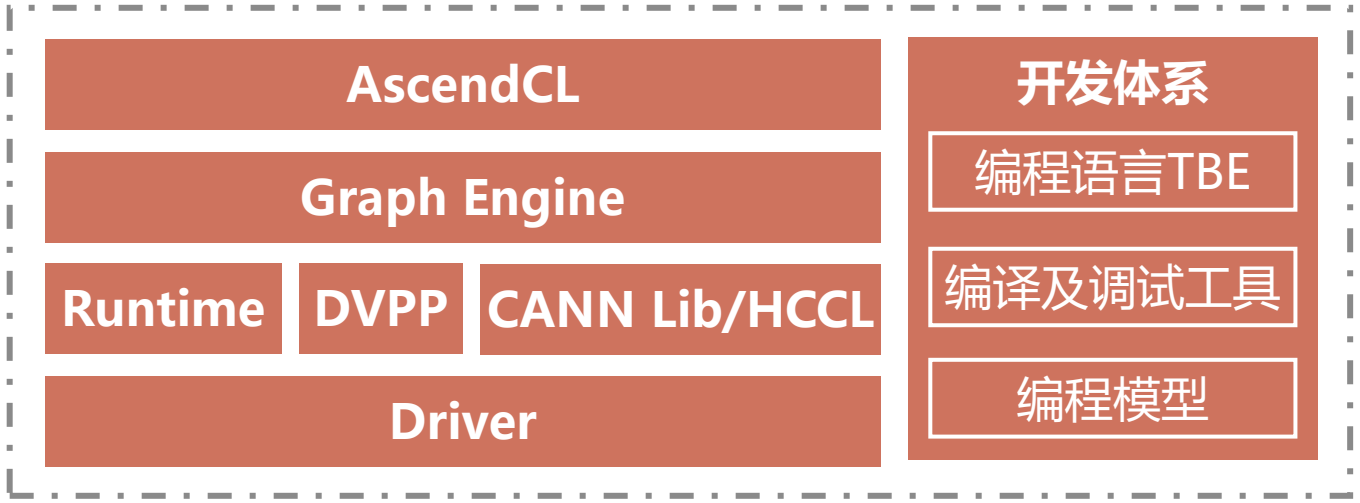
03 CANN：针对AI场景提出的异构计算架构

CANN

AI异构计算架构 开放平台

编程语言 | **编译及调试工具** | **编程模型**

- 支持端边云全场景协同，支持超过10种设备形态、EMUI、Andriod、openEuler等超过14种操作系统和多种AI计算框架
- 支持多种计算架构和计算框架，一套体系支持CPU、NPU等架构和多种AI计算框架
- 支持向后兼容和演进，向后兼容是为了保护开发者的已有开发投资
- 具有极强的伸缩性和适应性，适应不同算力和内存变化



资料来源：昇腾计算产业发展白皮书，浙商证券研究所

MindSpore：总体架构

昇腾计算语言

Model Zoo模型库
- Vision
- NLP
- Audio
- Rec
- Model Hub

Extend
强化学习/GNN/深度概率编程…

Sience科学计算
电磁仿真/分子模拟/量子计算

Insight
- 网络调试
- 性能调优
- 精度调优

开发态友好

Expression 全场景统一API
- 函数/OOP编程范式融合
- AI+数值计算表达融合
- 动静表达统一
- 单机分布式表达统一
- NN
- Ops
- Dataset
- Train/Infer
- Numpy

第三方前端
- 仓颉
- Julia

运行态高效

Data
- 数据加载
- 数据格式
- 异构加速
- 数据增强

Complier AI编译器
- 类型推导
- 自动微分
- 自动并行
- 二阶优化
- 量化/剪枝/蒸馏
- 内存优化
- 图算融合
- 流水线执行
- IR中间表达层
- 硬件相关优化
- 量子自动生成

Armour AI安全
- 密态AI
- 可信AI

全场景部署

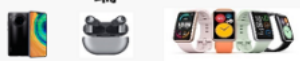
Runtime
端-边-云全场景运行时系统

多样性硬件

- CNN(NPU)
- CUDA(GPU)
- Eigen(CPU)
- oneDNN(CPU)

云 

边 

端 

数据来源： 华为昇腾官网，浙商证券研究所

MindSpore：执行流程

ModelZoo 模型库

MindSpore DevKit开发套件

MindSpore Extend高阶扩展库

MindSpore Hub 预训练模型

下载模型

MindSpore 训练和验证

模型导出

MindIR 端云统一IR

模型导入

MindSpore Servinig 云测推理

MindSpore Lite 端侧推理

端边增量学习 Fine-tuning训练等

调试调优

数据处理

安全机制

MindSpore Insight 可视化

MindSpore Data数据

MindSpore Armour AI安全

端云协同（联邦学习、端云AutoML等）

数据来源： 昇腾计算产业发展白皮书，浙商证券研究所

MindSpore：三方面创新

昇思 MindSpore

全场景深度学习框架

动静态图转换

采用SCT的AD机制，实现对静态图和动态图的支持，高效易用，静态图和动态图模式切换只需要一行代码。帮助开发者提升网络调试调优效率，并获得训练性能的收益。

自动并行

Auto Parallel特性，用于实现自动的数据并行和模型并行的混合并行训练，最大优势是易用和高性能。此外MindSpore可以结合内存、计算和通信开销，为用户选择一个性能较优的并行切分策略实现大规模网络的线性加速、自动扩展。

端边云协同

针对“端、边、云”全场景提供一致的开发和部署能力，以及按需协同能力，让开发者能够实现AI应用在云、边缘和手机上的快速部署，全场景互联互通，更好利用资源和保护隐私，创造更加丰富的AI应用。

- **快速多处部署：**针对特定应用场景，搜索满足性能约束的模型，拿来即用，无需重训。
- **全栈性能优化：**神经架构搜索、模型压缩、编译优化等手段优化精度、大小、时延。
- **灵活易用：**支持多种策略组合使用，打通云到端全流程，集中管理全流程策略和配置。
- **多种学习形态：**比如当前业界常用的端侧推理形态，满足开发者各种各样的场景需求。

资料来源：昇腾计算产业发展白皮书，浙商证券研究所

MindStudio：提供全栈全流程支持

算子开发

- 支持TBE自定义支持TBE自定义算子
- 语法智能纠错与代码自定补全
- 支持算子工程的仿真调试
- 多设备日志即时查看
- 算子性能profiling分析

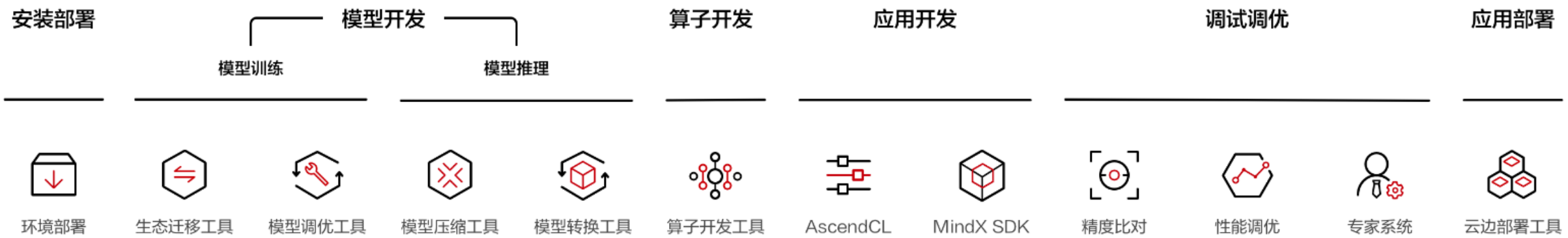
模型开发

- 支持MindSpore/caffe/TF等框架
- 离线模型转换
- 模型算子精度比对
- 模型整网性能profiling分析
- 支持网络模型的可视化

应用开发

- AI应用开发与调试
- 算法代码自动生成
- 模型市场ModelZo0
- 推理结果的图形化展示
- 全系统调优

MindStudio开发流程



资料来源：昇腾社区官网，浙商证券研究所

MindX：2+1+X 应用使能

2个组件

深度学习组件

MindX DL

深度学习平台基础组件

- 集群调度
- 模型保护
- ToolBox

智能边缘组件

MindX Edge

智能边缘平台基础组件

- MindX Edge Framework
- AtlasEdge
- 设备管理OM
- 基础服务

1个模型库

优选模型库

ModelZoo

预训练高性能模型

- 计算机视觉
- 自然语言处理
- 多模态
- 科学技术
- ……

X行业应用

行业 SDK

mxVision

视频分析 SDK

- 质检算法
- 视觉工具

mxManufacture

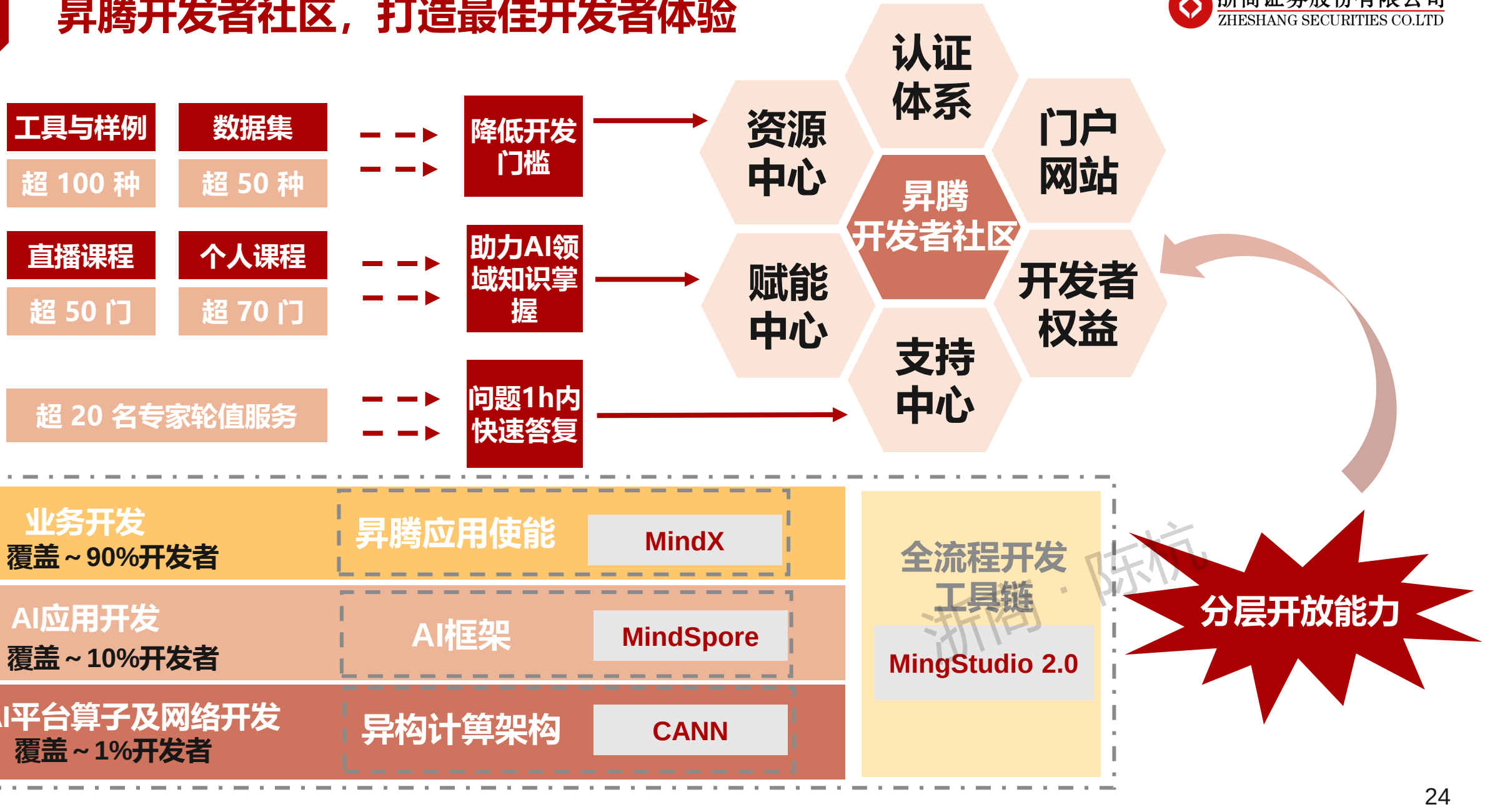
智能制造 SDK

- 模型
- 算法
- 工具

资料来源：昇腾计算产业发展白皮书，浙商证券研究所

04 生态伙伴

04 昇腾开发者社区，打造最佳开发者体验



资料来源：昇腾计算产业发展白皮书，浙商证券研究所

昇腾万里合作伙伴计划

昇腾万里合作伙伴计划

- 伙伴合作计划
- 开发人员成长计划
- 高校合作计划
- 初创扶植计划

合作伙伴独立生态计划

鼓励伙伴发布关于昇腾的单独生态计划权益可叠加

- 咨询与解决方案集成商C&SI
- 独立软件开发商ISV
- 独立硬件开发商IHV
- 云服务提供商XaaS
- 原始设备制造商OEM
- 原始设计制造商ODM
- 机器视觉生态
- 机器视觉生态

全方位激励机制，繁荣昇腾生态

学习阶段 (LDF，学习成长基金)		构建阶段 (PDF，产品开发基金)			上市阶段 (MDF，市场开发基金)		上市阶段 (MDF，市场开发基金)	
人才培训	云资源	样机	NRE	PAE/FAE	Marketplace	MDF	直接业绩激励	间接业绩激励

资料来源：昇腾计算产业发展白皮书，浙商证券研究所

昇腾生态伙伴网络覆盖众多主流厂商

IHV硬件伙伴

ADLINK
ADVANTECH 研華科技
BFA
全爱科技
Dongsheng 东声智能科技

一体机解决方案伙伴

YISA 以萨

整机硬件伙伴

北联国芯 BEILIAN COMPUTING
Wuhan Yangtze Computing Technology 长江计算
华鲲振宇
清华同方 TSINGHUA TONGFANG
H3C
TRUST 可信 百信
KunLun
宝德 PowerLeader
五舟 WUZHOU
湘江鲲鹏
Huanghe
神州数码 Digital China
安擎 NGINETECH

应用软件伙伴

WHAYER 华雁智科
YISA 以萨
DEEPGLINT 格灵深瞳
云从科技 定义智慧生活
intellifusion 云天励飞
ZHY 智洋创新

辅助运营伙伴

极视 开创AI视觉算法
ICS&S 中软国际

数据来源： 昇腾社区官网，浙商证券研究所

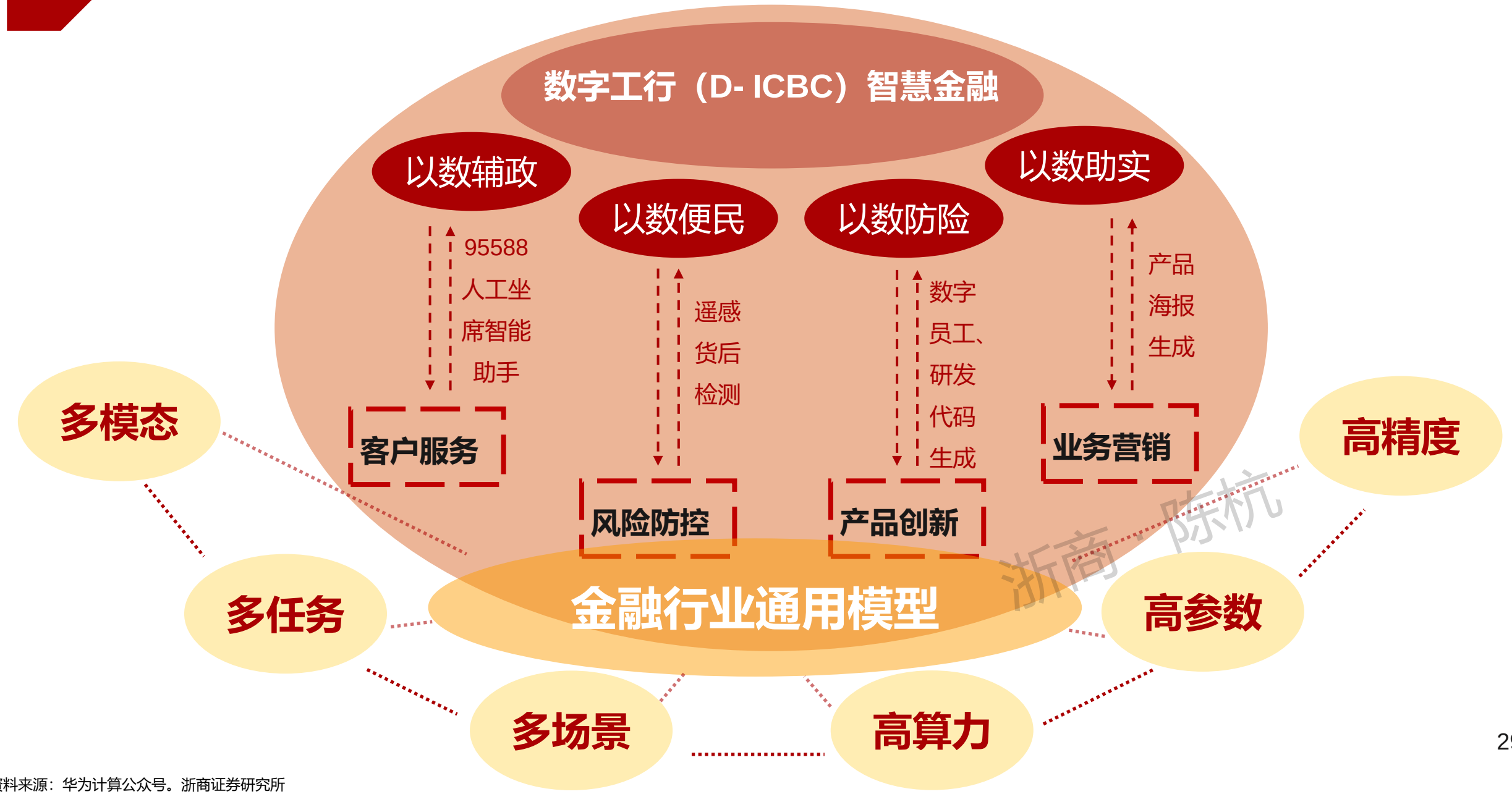
05 应用落地

AI计算中心解决方案

	全栈AI计算中心解决方案	云AI计算中心解决方案	轻量化AI计算中心解决方案
L4应用平台	AI应用（合作伙伴/开发者/高校）		
L3软件平台	应用使能、AI框架、芯片使能	华为云HCSO (ModelArts)	合作伙伴深度学习平台
L2硬件平台	通用计算：鲲鹏、X86 AI计算：昇腾、GPU	Atlas 800 训练服务器 Atlas 900 AI 集群	Atlas 800 训练服务器 Atlas 900 AI 集群
L1机房设施	机柜/电源/冷机/油机	可选	可选
L0楼宇平台	预制式机房	可选	可选

资料来源：昇腾计算产业发展白皮书，浙商证券研究所

金融：昇腾AI训练的金融行业通用模型



资料来源：华为计算公众号。浙商证券研究所

05 生物医药：蛋白质结构预测平台实现降本增效

华为
+
武汉伯生科技

基于昇腾AI

“思符”

可应用于生物医药研发的AI蛋白质结构预测平台

- 一键式蛋白质结构预测
- 超长序列结构预测
- 多序列结构预测
- AI预测功能合作定制
- 进化的AI预测体系

① 位于中间层的昇腾AI异构计算架构CANN为模型推理提供加速引擎，在算子融合方面实现网络中计算单元的优化整合

② 在内存优化方面完成模型特征图的有效内存排布

③ 在混合精度计算方面实现不同精度的计算分配

降本

蛋白质取样费用至少1.5万元，累计费用超过数十万

预测费用下降100倍以上

增效

从蛋白表达、蛋白纯化等步骤，到使用冷冻电镜解析蛋白质结构，至少13天

全步骤压缩到1天以内
结构分析效率平均提升超10倍

资料来源：华为计算公众号，浙商证券研究所

风险提示

1、**宏观经济下行超预期风险**：若宏观经济下行并超出预期，将影响整个产业链；

2、**上游晶圆紧缺加剧的风险**：若本土晶圆厂扩产进度不及预期，将影响芯片供给；

3、**市场发展不及预期的风险**：若AIGC所带来的催化作用不及预期，将影响下游需求；

4、**行业竞争超预期风险**：若行业竞争加剧，可能对公司地位造成不利影响。

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

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