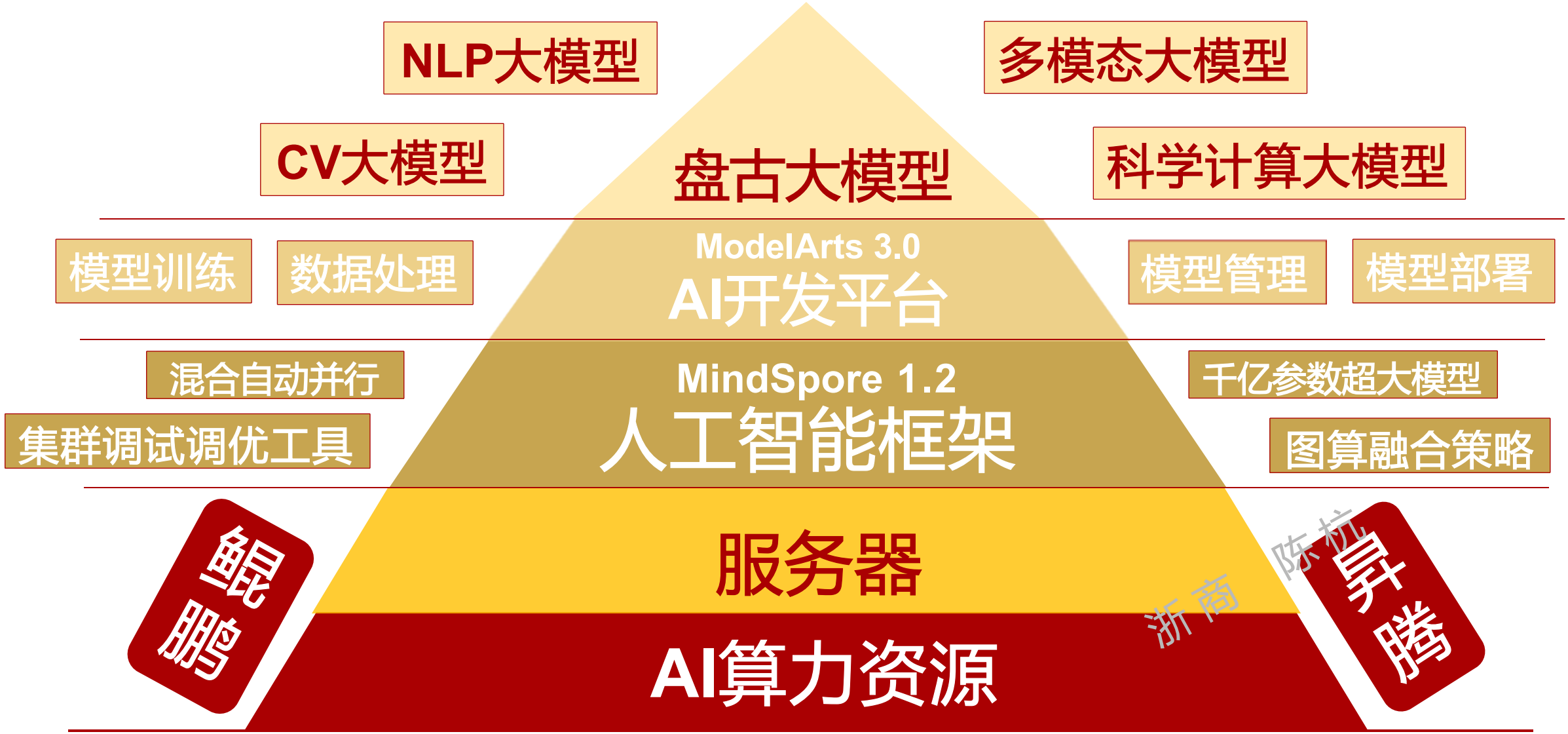
华为AI盘古大模型研究框架

华为产业链深度系列研究

01 盘古大模型金字塔架构，算力为底层支撑



资料来源：

摘要

目前我们将迎来科技的重大转折点： ChatGPT时刻。 而在ChatGPT背后， 不断迭代的GPT系列使得大模型成为当下科技企业核心竞争力的重要体现， 未来， 大模型将成为AIGC时代的核心支撑。 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。 我们将从： 昇腾/鲲鹏→ MindSporeAI框架→ ModelArts→盘古大模型四层架构进行分析：

1、AI算力资源： “鲲鹏+昇腾”， 打造盘古算力底座

- 鲲鹏： 华为自主芯片→鲲鹏芯片→鲲鹏服务器→ 欧拉操作系统→高斯数据库→行业应用向外拓展，构建鲲鹏生态， 提供算力支撑。
- 昇腾 昇腾AI处理器 →CANN异构计算架构→ MindSporeAI框架→应用使能→行业应用，助力打造华为昇腾全栈AI软硬件平台，构筑智能世界基石

2、人工智能框架： ◦MindSpore高效易开发， 可实现全场景覆盖

- CANN： 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， 助力芯片使能。
- MindSpore： 是国内首个支持千亿参数大模型训练AI计算框架， 最佳匹配昇腾处理器算力， 支持终端、 边缘、 云全场景灵活部署，开创全新的AI编程范式， 降低AI开发门槛。

3、AI开发平台： ModelArts强势赋能开发者， 精度效率双提升

- 为机器学习与深度学习提供海量数据预处理及交互式智能标注、 大规模分布式训练、 自动化模型生成， 及端-边-云模型按需部署能力， 帮助用户快速创建和部署模型， 管理全周期AI工作流。

4、盘古大模型： AI落地的重要途径

- 由NLP大模型、 CV大模型、 多模态大模型、 科学计算大模型等多个大模型构成， 目前已实现等AI场景落地。

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

数据来源： 华为云官网，

目录

CONTENTS

01

算力资源

鲲鹏　　昇腾

01 AI算力资源

华为鲲鹏生态： 华为自主芯片→鲲鹏芯片→鲲鹏服务器→欧拉操作系统→高斯数据库→行业应用

1、鲲鹏芯片： 鲲鹏920作为低功耗、 高性能的Arm处理器， 为鲲鹏服务器主板及整机产品提供芯片支撑， 是鲲鹏生态发展壮大的核心所在， 在此基础上， 华为进一步开启自主研发芯片， 为鲲鹏生态发展奠定坚实基础。

2、鲲鹏服务器： 华为凭借多年积累的硬件工程能力， 打造TaiShan服务器， 使能整个产业链， 进一步构建完整鲲鹏生态。

3、 欧拉操作系统： 作为面向B端的电脑服务器操作系统， 华为自主研发的EulerOS， 以Linux稳定系统内核为基础， 南向支持多样性设备， 北向覆盖全场景应用， 横向对接鸿蒙， 通过能力共享实现生态互通。

4、高斯数据库： 华为GaussDB是主打政企核心业务负载的金融级分布式数据库， 目前已实现助力部分保险及车企数字化转型。

5、行业应用： 华为以行业聚合应用， 通过平台和生态双轮驱动， 形成行业应用矩阵， 为众多行业客户提供解决方案。 并陆续成立五大军团， 不断开发全新应用场景。

华为昇腾AI产业： 昇腾AI处理器→CANN异构计算架构→ MindSporeAI框架→应用使能→行业应用

1、Ascend： 昇腾AI处理器作为基础， 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。

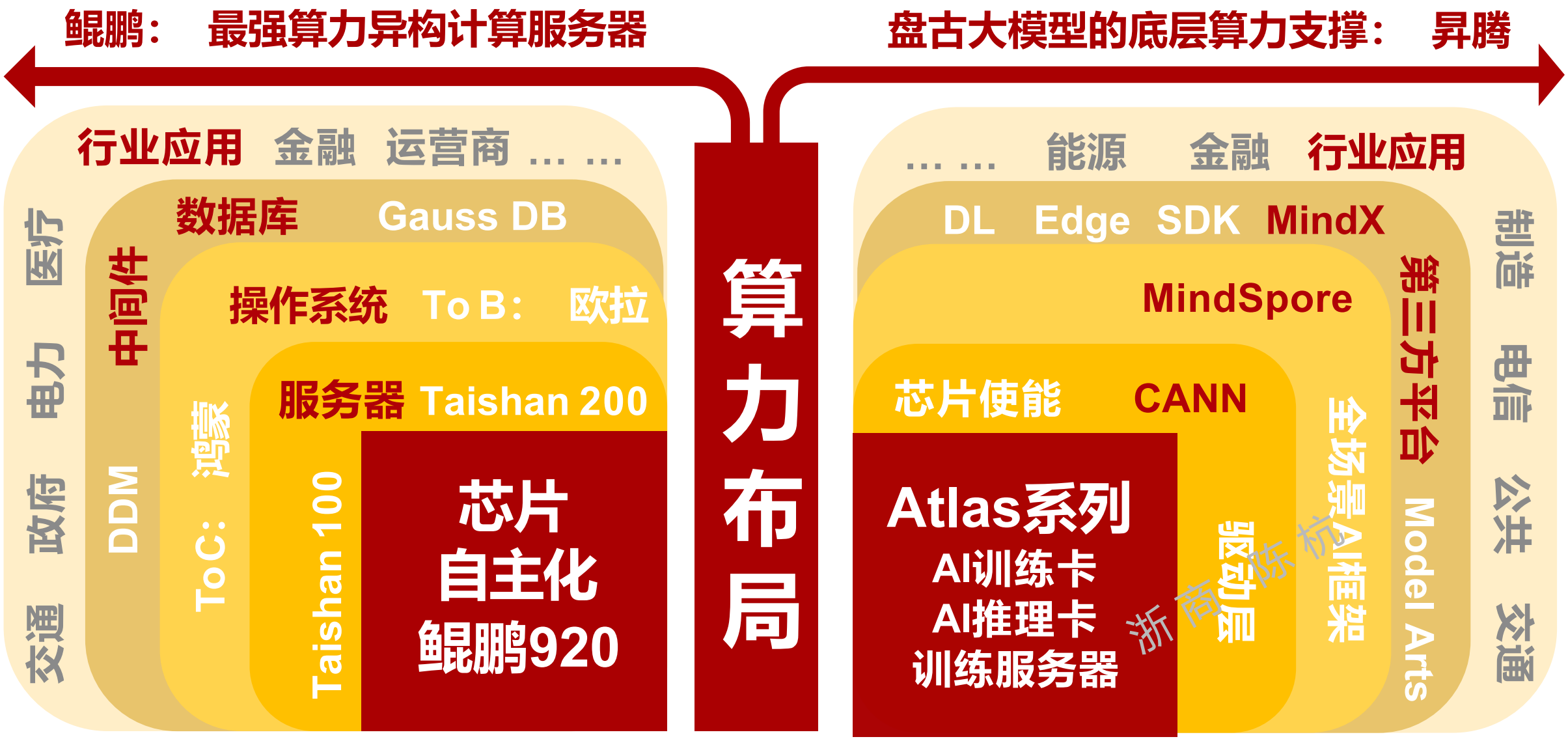
2、 CANN： 作为华为昇腾AI基础软硬件平台的核心， CANN向上支持多种AI框架， 向下服务AI处理器与编程， 以极致性能、 极简开发、 开放生态为目标， 助力昇腾构建全场景人工智能平台。

3、 MindSpore： 是国内首个支持千亿参数大模型训练AI计算框架， 覆盖包含生物医学在内的多个领域。

4、应用使能： 以昇腾AI处理器→CANN异构计算架构→ MindSporeAI框架的传导机制， 为深度学习、 智能边缘以及行业应用解决方案等强势赋能。

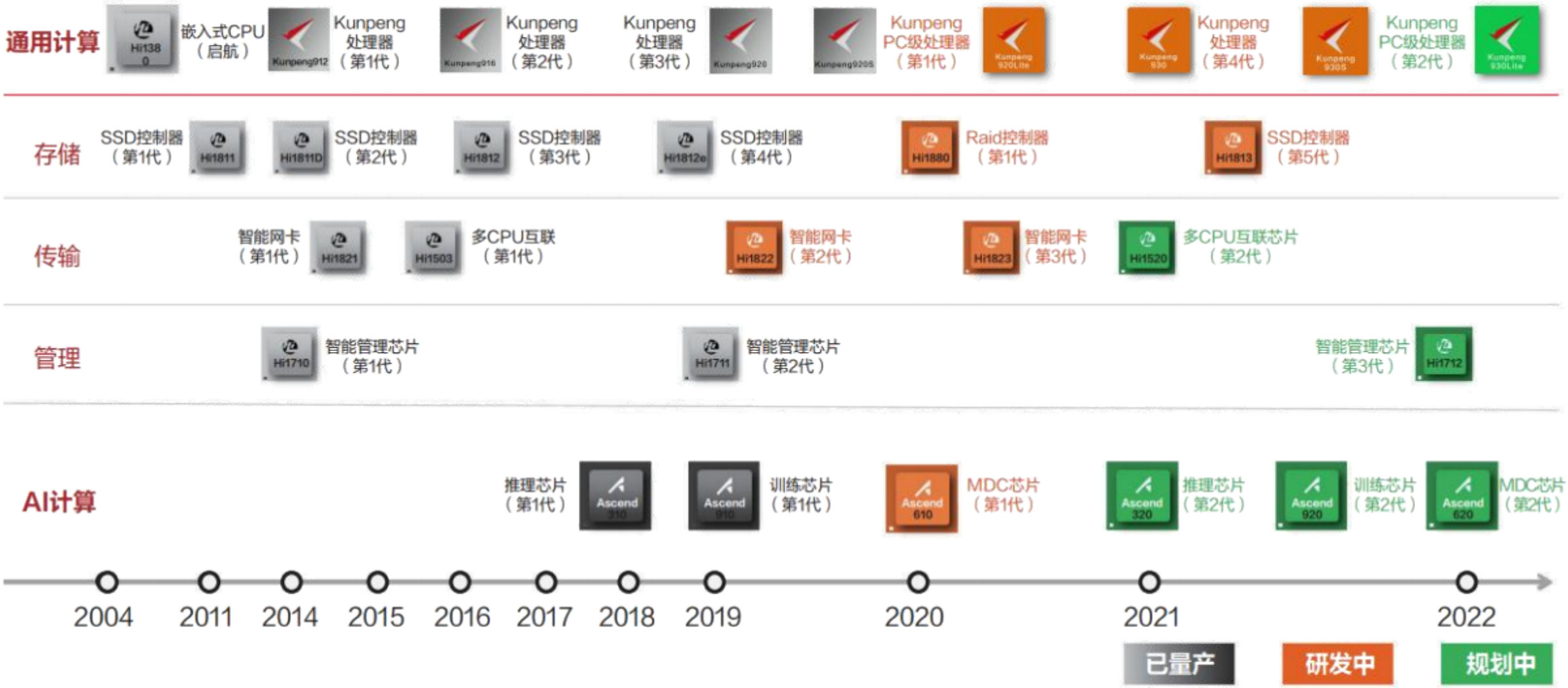
数据来源： 昇思官网，

01 算力布局 = 鲲鹏+异腾



数据来源： 昇思官网，

01 Kunpeng处理器支持鲲鹏计算产业长期演进



数据来源： 鲲鹏计算产业发展白皮书，

01 硬件：鲲鹏服务器实现数据中心高效能计算

鲲鹏服务器

TaiShan100

- 包含2280均衡型和5280存储型等产品型号。
- 基于鲲鹏916处理器的数据中心服务器， 具有多核高并发、 低功耗等计算优势， 适合为大数据、分布式存储等应用高效加速。

TaiShan200

- 包含2280E边缘型、 1280高密型、 2280均衡型、 2480高性能型、 5280存储型和X6000高密型等产品型号。
- 基于华为鲲鹏920处理器， 旨在满足数据中心多样性计算需求。

TaiShan200 Pro

- 包含2480、 2280和1280等三款高端产品型号。
- 基于鲲鹏920 3.0GHz高主频处理器， 同时集成三大创新RAS特性， 获得权威安全可信认证。

高效能计算

- 搭载具有超强算力的鲲鹏处理器
- 多核计算架构
- 高效加速应用

安全可靠

- 处理器及服务器芯片全自研
- 17年计算工程能力铸就稳如泰山品质

开放生态

- 开放计算平台
- 支持业界主流软件
- 携手合作伙伴， 共赢计算新生态

数据来源： 鲲鹏社区官网，

01 澎湃算力携手合作伙伴，共同拓展鲲鹏计算产业

数据来源： 鲲鹏社区官网，

01 昇腾AI底层支撑系列产品

AI模块

开发者套件

芯片： 昇腾310

最高算力： 22 TOPS

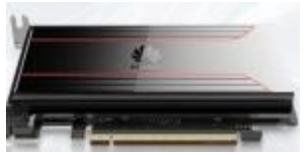
AI加速模块

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，

01 昇腾生态伙伴网络遍及主流厂商

IHV硬件伙伴

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

一体机解决方案伙伴

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
i0i0百信 TRUST
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华耀智科
YISA 以萨
DEEPGLINT 格灵深瞳
云从科技 一定义智慧生活一
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，

02

人工智能框架

CANN　　Mindspore