HUAWEI

华为云昇腾AI云服务

6A FAMILY 云化算力底座

目录

大模型引发全球算力需求的指数级增长 02

大模型为 AI 产业带来拐点 03
Sora 的出现再次印证 Scaling law，大模型创新需要澎湃算力支撑 04

聚焦业务创新，企业需要全栈算力服务 05

大模型是人类迄今为止最复杂的软件、硬件系统 06
昇腾 AI 云服务，大模型时代的最佳云化全栈算力服务 07
满足多样化算力使用模式 08
满足多样化算力管理模式 09
满足多样化算力部署模式 10

昇腾云服务打造 6A FAMILY 算力沃土 11

昇腾 AI 云服务打造 6A FAMILY 算力沃土，构建百模千态首选云底座 12
故障恢复快 **F**ault recovery Acceleration 13
资源获取快 **A**ccess Acceleration 14
模型迁移快 **M**igration Acceleration 15
云上推理投资优 **I**nvestment Advantage 17
就近服务时延优 **L**atency Advantage 19
云上性能优 **Y**ield Advantage 21

昇腾云服务开放兼容支持百模千态 22

AI Gallery：一站式 AI 社区服务平台，构建百模千态的开放昇腾社区 23
D-Plan：生态伙伴计划 24

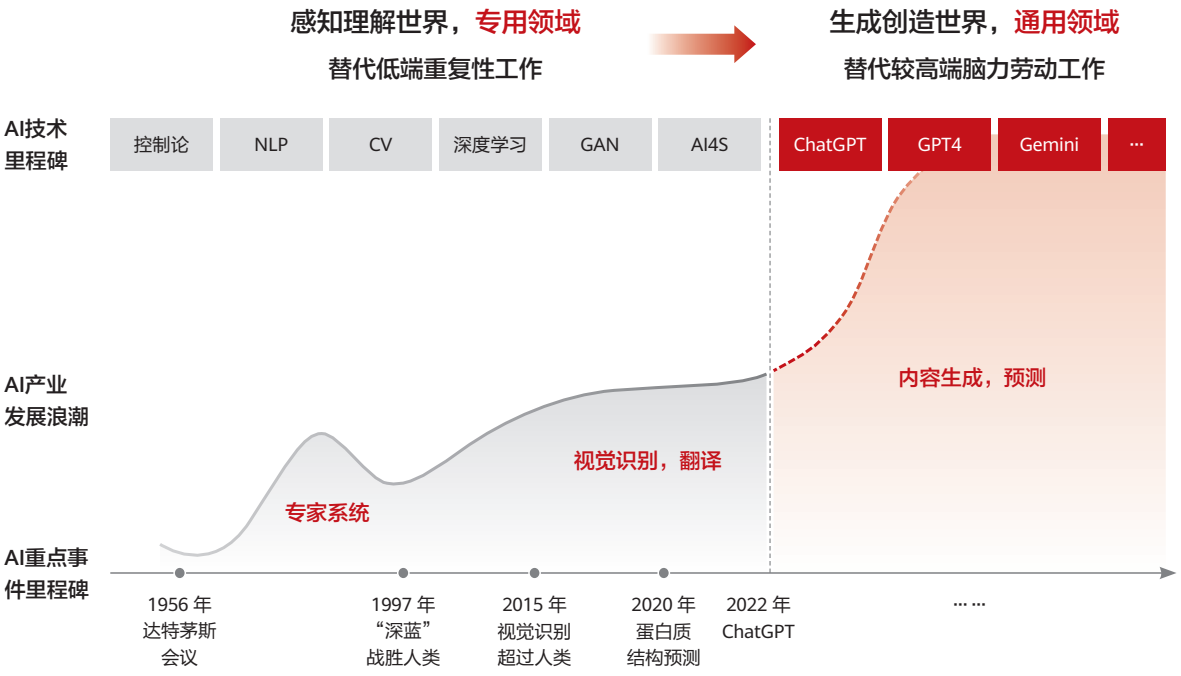
客户案例 26

昇腾 AI 云服务 — 全球行业先行者 26

大模型引发全球算力需求的指数级增长

大模型为 AI 产业带来拐点

从“感知”走向“生成”，从专用走向通用

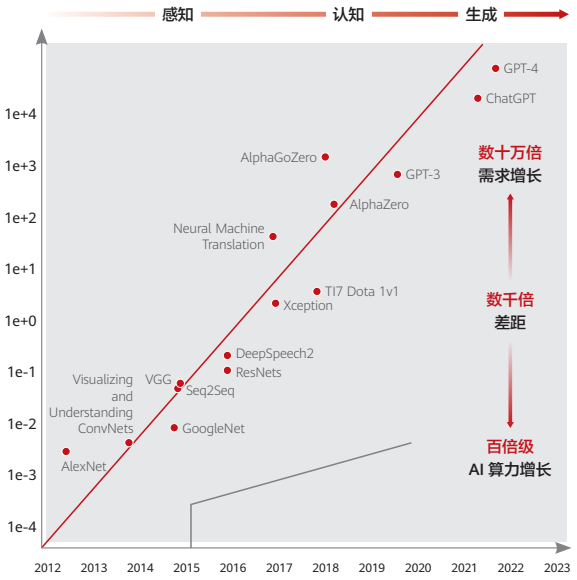


人工智能产业发展经历三次浪潮。最近的一次是以 Transformer 架构为代表的大模型，生成式 AI 的兴起，将我们带入新的 AI 产业浪潮之中。大模型是人工智能历史的分水岭，此前，人们更多关注和讨论的是机器如何感知世界，例如识别日常生活中的各种物体；而现在，人类则进入到通过大模型的生成能力创造数字世界，预测未来趋势。通过对海量数据的预训练，大模型可以在超高维度空间上对人类全部知识进行高度压缩，进行微调就可以完成多个应用场景任务的泛化，模型正在从专用走向通用。

随着人工智能技术的日新月异，AI 将进一步驱动各行各业生产能力、生产效率从“量变到质变”，实现跨越式发展和新质生产力跃升，如何用好 AI 将成为国家、行业、企业的核心竞争力。

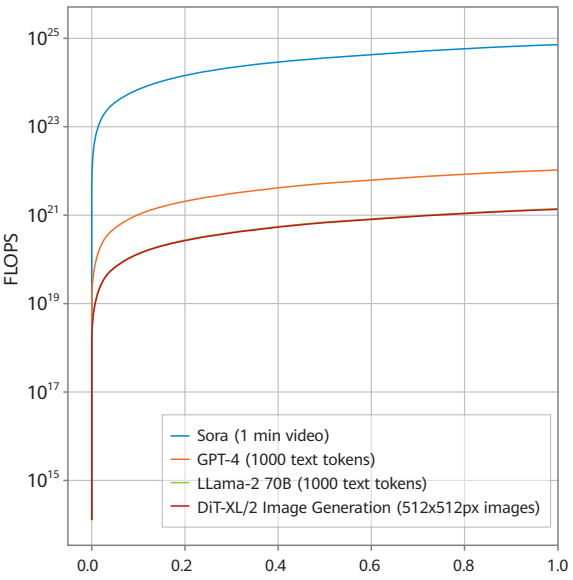
Sora 的出现再次印证 Scaling law，大模型创新需要澎湃算力支撑

2012 年至 2023 年
算力需求翻了数十万倍



模型规模及发布时间

SORA 所消耗的算力相比
LLM 类模型要大数个量级



推理消耗算力对比

Source: Factorial Funds AI inference compute comparison

大模型的爆发引发全球算力需求的指数级增长。2024 年基于扩散的视频生成模型 Sora 的出现，其革命性的视频生成能力，不仅展示了 AI 在视觉内容创造上的突破，更预示着全球算力需求的新一轮激增。数据显示，过去 10 年 AI 算力需求翻了 30 万倍。而未来 10 年 AI 算力将再增长 500 倍。数据集规模将从目前的一两个 T 增长到 100T。此外，大模型还需要理解更长的上下文，Token 长度将从千级发展到十万级。

视频生成类模型的算力消耗相比 LLM 提升 20 倍，意味着训练集群规模要扩大一个数量级。万卡训练集群将成为训练下一代生成式模型的必备条件。由于算力规模扩大，算力的调度和管理的难度将大幅提升，需要有一个算力平台可以整合管理，调度，自动故障隔离，checkpointing，自动任务恢复的任务。这些挑战相互影响、环环相扣。

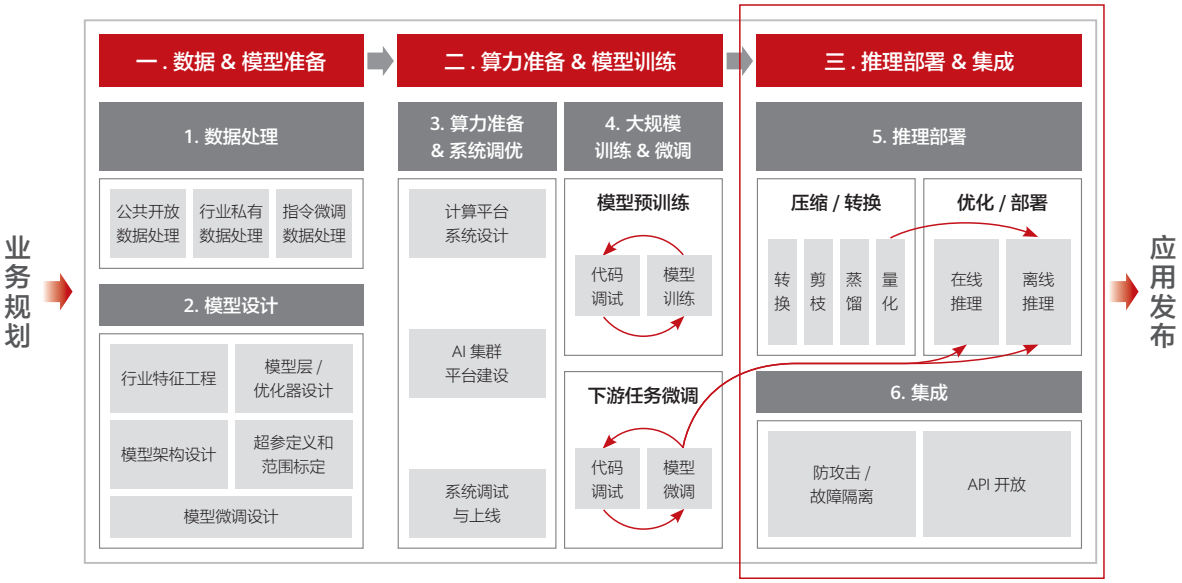
针对 AI 时代的这些挑战，华为云提出了软硬件结合的系统性创新，华为云昇腾 AI 云服务整合集群算力、计算引擎 CANN、AI 开发框架 MindSpore 和 ModelArts AI 开发生产线。为大模型的训练，推理，AI 应用的开发、运行提供稳定可靠的全栈算力保障。

聚焦业务创新，企业需要全栈算力服务

大模型是人类迄今为止最复杂的软件、硬件系统

大模型是一个复杂系统工程，大模型开发的每一步都存在着大量的工程化技术挑战。算力系统也并非算力的简单堆积，需要解决诸如低时延数据交换，节点之间均衡计算避免冷热不均，消弭算力堵点。避免出现单点硬件故障导致的全面训练中断、梯度爆炸、算法重训等一系列的问题，是一项复杂的系统工程，需要从算力效率、线性扩展、长效稳定等多个方面进行系统设计。而云化的全栈算力服务由于积累了足够多的模型训练，运维经验，以服务的方式让企业使用到最新的经验，技术成果，避免重复解决问题，让企业得以聚焦创新。

大模型不仅需要算法，而且需要数据处理，
软硬件优化、模型开发、应用创新、推理部署的系统工程能力



- 参数面无损网络
- 多级存储优化
- 计算集群密度设计
- 液冷设计
- 多样化算力调度
- 集群稳定性设计
- ……

- 开源数据集选择
- 预训练数据清洗
- 数据质量测试
- 稠密稀疏混合架构
- RLHF 算法设计
- RLHF 数据集标注
- ……

- 多种并行策略设计
- 通信链路加速
- 多任务可视化 profiling
- 断点续训设计
- 算子融合调优
- 多样化算力调度
- 多任务权重融合
- ……

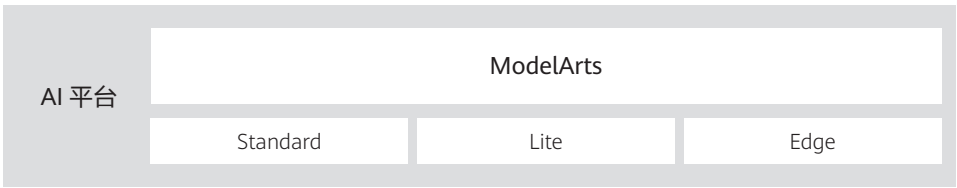
- 大模型分布式推理切分
- 在线推理框架
- 模型剪枝和蒸馏技术
- 模型 INT 量化
- 下游多任务效果测试
- 微调算法优化
- 推理性能调优
- ……

- 推理集群设计
- 推理集群调度系统
- 多应用 Load Balance
- API 接口设计
- 防攻击设计
- 故障恢复和隔离
- ……

昇腾 AI 云服务，大模型时代的最佳云化全栈算力服务

昇腾 AI 云服务：包括云化算力、AI 开发生产线 ModelArts 和 AI 开发者生态 AI Gallery。为支持大模型的“百模千态”创新，昇腾 AI 云服务提供触手可及的澎湃 AI 算力服务，独有的多级恢复机制和完备的工具链可实现千卡训练连续 30 天不中断，任务恢复时长小于 30 分钟，为大模型和 AI 应用的开发、运行、运维提供最佳算力云底座。

昇腾AI云服务官网



澎湃算力
即开即用
无需自建或
改造数据中心

高效易用
全栈平台能力
无需投资
通用 AI 技术

集群训练
故障自动恢复
无需担心
运维和安全

打造百模千态的
黑土地
无需担心
模型开发应用难

云网边端芯
算力协同
无需担心
端侧算力瓶颈

满足多样化算力使用模式

自研大模型

需要数千卡算力

拥有超级 APP

offering

- 提供大规模算力集群
- 提供分布式加速库
- 提供大模型适配和优化
- **技术栈开放，高度自主可控**

增量训练大模型

需要数百卡算力

拥有较强行业背景和大量行业数据

offering

- 提供主流三方大模型
- 提供完善的 SFT 训练框架
- 提供参考案例
- 提供易用的大模型应用开发工具链

智能应用开发

需要数十卡算力

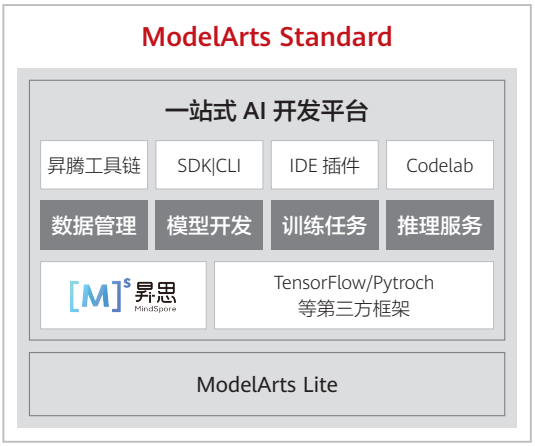
拥有较深的行业理解

offering

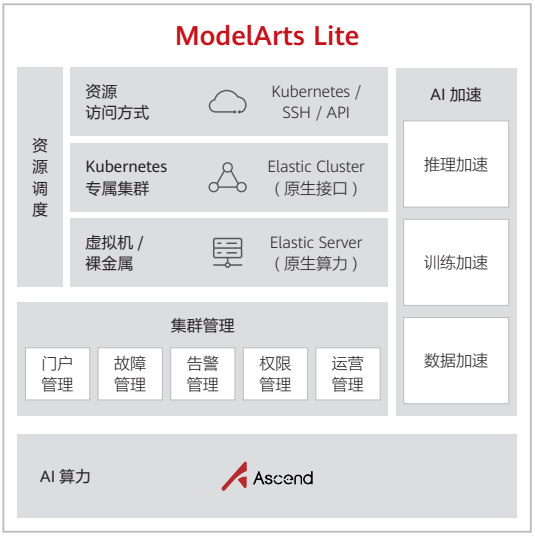
- 提供开箱即用的开源大模型，支持微调，快速上手
- 提供端到端应用开发工具链、向量数据库等
- 提供丰富的预制应用模板

满足多样化算力管理模式

AI 开发生产线 ModelArts，是面向 AI 开发者的一站式开发平台，提供海量数据预处理及半自动化标注、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。为满足客户多样化的算力管理模式，ModelArts 提供 Standard 和 Lite 两种模式。ModelArts Standard 包含端到端的 AI 开发生产线 + 算力持续运维平台。ModelArts Lite 仅包含算力持续运维平台。



» 提供端到端的 AI 开发生产线 + 算力持续运维平台

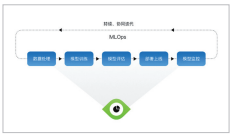


» 算力持续运维平台

ModelArts Standard 服务的介绍

端到端生产工具链，一致性开发体验

» 线上线下协同开发，开发训练一体化架构，支持大模型分布式部署及推理



AI 工程化能力，支持 AI 全流程生命周期管理

» 支持 MLOps 能力，提供数据诊断、模型监测等分析能力，训练智能日志分析与诊断

容错能力强，故障恢复快

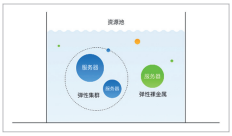
» 故障检测覆盖度 95%，故障 30 分钟内恢复，恢复成功率大于 95%，保障千卡作业稳定训练数周以上，训练有效卡时大于 95%

ModelArts Lite 服务的介绍

零改造迁移

» 提供业界通用的 k8s 接口使用资源，业务跨云迁移无压力

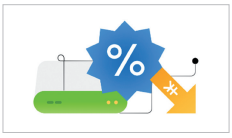
» SSH 直达节点和容器，一致体验



多种资源形态

» 集群模式，开箱即提供好 Kubernetes 集群，直接使用，方便高效

» 节点模式，客户可采用开源或自研框架，自行构建集群，更强的掌控力和灵活性



极致性价比

» 提供高性价比国产算力

» 多年软硬件经验沉淀，AI 场景极致优化

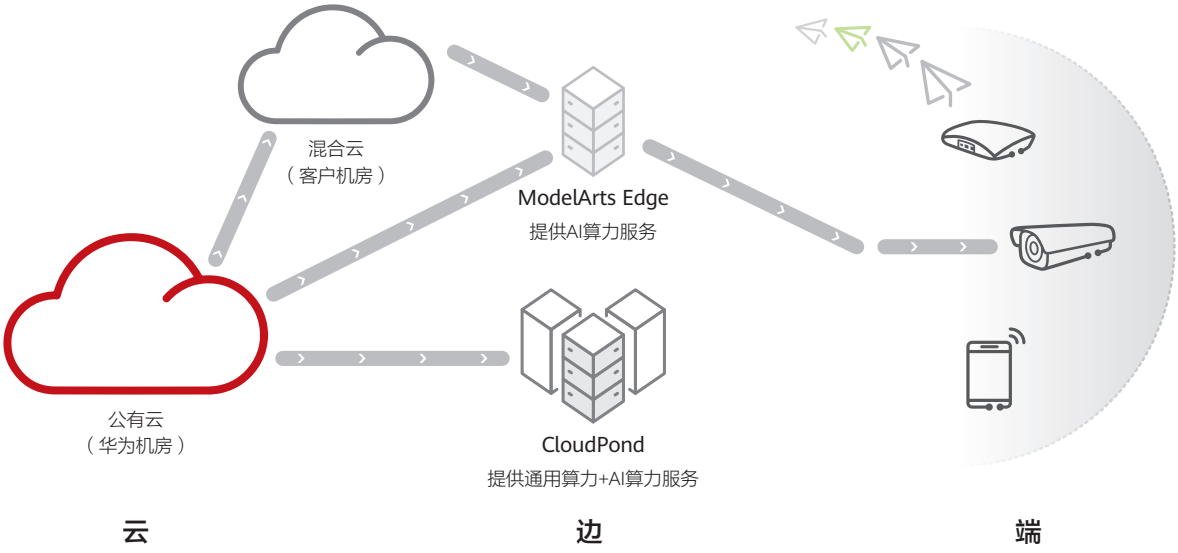
» 加速套件，训练、推理、数据访问多维度加速

故障恢复

» 机柜、节点、加速卡、任务多场景故障感知

» 节点级、作业级、容器级，多级故障恢复

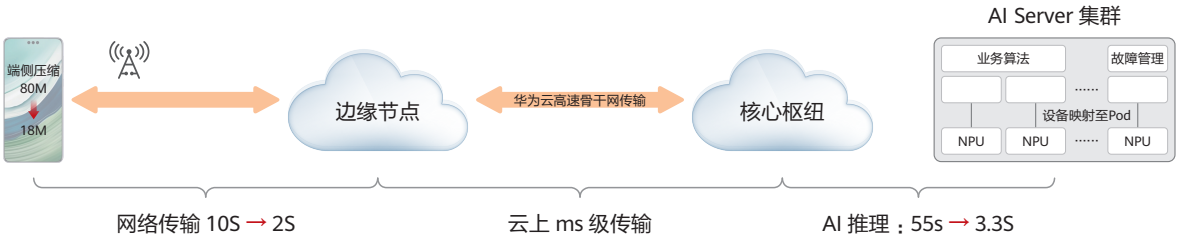
满足多样化算力部署模式



端云协同，以云助端的案例

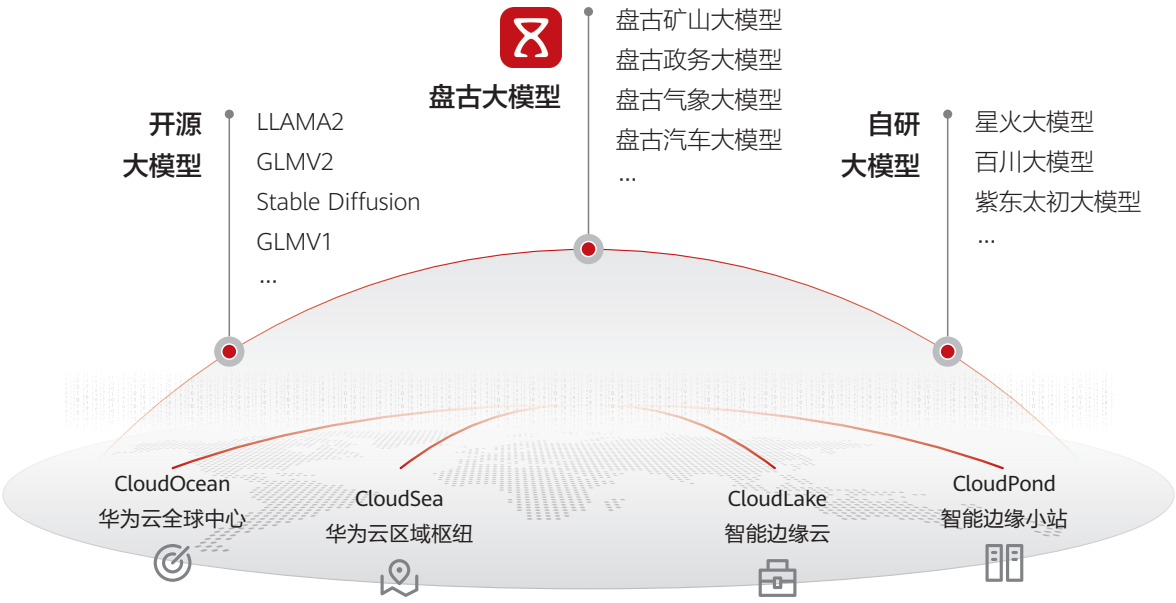
昇腾 AI 云服务通过云网边端芯算力协同，为端侧提供更充沛算力，让终端应用更智能。

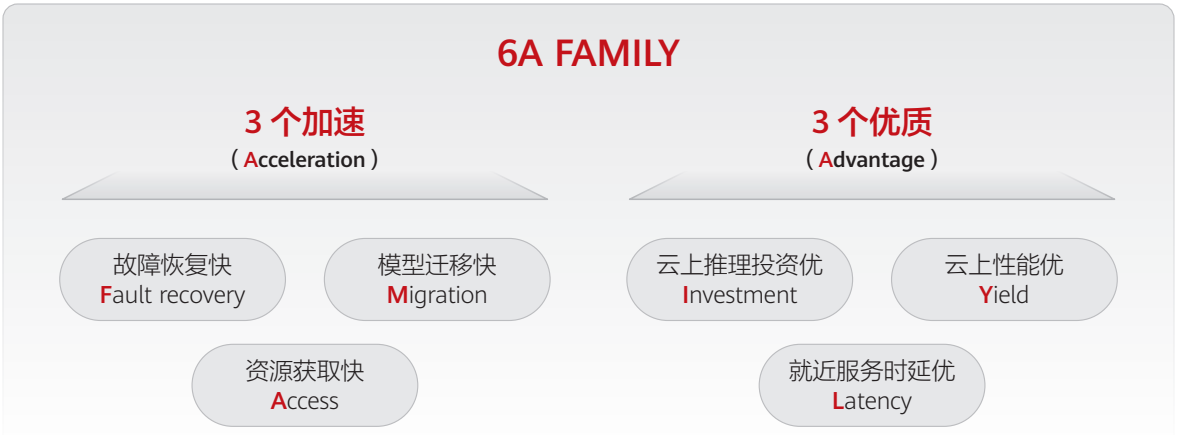
受限于体积和成本等因素，手机硬件很难做到高清拍摄，也无法支撑超分修图的算力要求。通过以云助端，调用云端强大的算力，利用枢纽节点大规模算力来进行超分修图，突破手机硬件的限制，为用户的手机拍照体验带来了全新的突破，使得用户能够在手机上轻松获得专业级的照片效果。



昇腾云服务
打造 6A FAMILY 算力沃土

昇腾 AI 云服务打造 6A FAMILY 算力沃土，构建百模千态首选云底座





大模型时代的 AI 算力对数据中心的基础设施要求极高。以散热为例，AI 服务器的功率密度远超通用服务器，单机柜的功耗是过去的 6-8 倍，并需要专用的液冷系统进行散热。大模型训练动辄需要百卡、千卡甚至万卡，自建 AI 数据中心面临 AI 研发人员稀缺，硬件建设周期长、集群运维团队经验少、推理服务时延高等诸多挑战。

故障恢复快 Fault recovery Acceleration

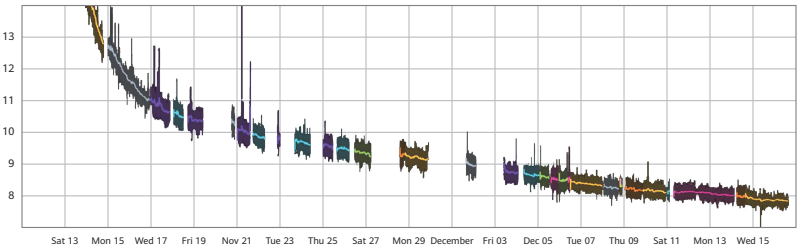
在云上，可以获得更快速的运维保障，集群故障可以做到 1 分钟发现，5 分钟诊断，10 分钟恢复。

传统方式是被动响应集群故障，重启范围广，作业恢复慢

业界实践：

» 业界大模型训练平均 **2.8 天出现一次中断**

» 业界故障处理时间约 **1~30 天**，严重拉低大模型训练效率



（图示为训练过程中的意外中断情况，横坐标为训练时间，纵坐标为困惑度 PPL）

训练时间变长

在 1000 个 80G A100 上训练 3000 亿个单词，需要 33 天。实际训练了 90 天，期间出现 112 次故障。

硬件故障占比高

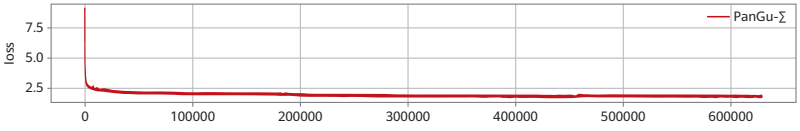
大模型训练期间碰到的主要问题是硬件故障、导致任务手动重启 35 次，自动重启约 70 多次，严重影响模型的训练进程。

业界

昇腾 AI 云服务主动诊断故障，避免训练中断，确保集群长稳运行

华为实践：

» 盘古 -200B 在非故障停机前**连续稳定训练 30 天**



昇腾 AI 云服务

1 分钟故障检测，5 分钟诊断

» **全链路故障感知**，覆盖不同层次的故障感知；

» **故障诊断引擎：**训练任务分层分级诊断能力；

» **丰富的诊断类型：**支持 300+ 通用种故障类型诊断，覆盖度 95%+。

10 分钟故障恢复

» 通过三级故障恢复，减少 50% 故障恢复耗时；

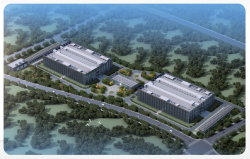
» 硬件故障不影响业务，10 分钟故障恢复；

» CKPT、图编译、建链、调度协同优化，缩短恢复时间。

资源获取快 Access Acceleration

在云上，模型训练可一键接入贵安、乌兰察布、芜湖，香港 AI 算力中心，支撑万亿参数大模型、百 P 数据训练。

贵安
AI 算力中心



乌兰察布
AI 算力中心

芜湖
AI 算力中心

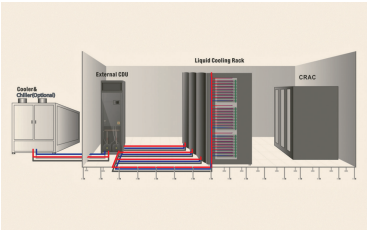
香港
AI 算力中心

支持 6+ 主流 AI 框架，90%+ 算子

PyTorch
昇思 MindSpore
飞桨
TensorFlow
Caffe
Jittor 计图
.....

澎湃算力：超大集群，30 天训练不中断

绿色：全液冷，PUE 低至 1.1

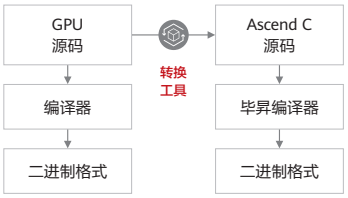


模型迁移快 Migration Acceleration

昇腾云服务支持业界各类框架、加速库及三方社区生态，可快速、无损实现模型和应用的迁移适配。

第三方算子

支持算子 Kernel 级
源码迁移



算子迁移开发周期

2 人月 » 2 人周

第三方模型

已支持三方社区
数百个模型

OpenMMLab

机器视觉领域主流方向

支持 OpenMMLab 社区算子库、套件等

- MMcv CV算子库
- MMClassification 图像分类套件
- MMdetection 图像检测套件
- MMSegmentation 语义分割套件

Hugging Face

自然语言处理类模型套件

各类 Transformer 模型

- HuggingFace Transformers
- .....

第三方 AI 框架

支持并兼容各版本
高阶特性

PyTorch

全面兼容

1.8、1.11、1.13、2.0 主流版本图模式、分布训练、量化等高阶特性

TensorFlow

全面支持

1.5、2.X 主流版本，300+ 模型

飞桨

已适配 70+ 模型

正与百度深度合作，共同推进模型适配

第三方加速库

跟随版本
支持最新特性

DeepSpeed

分布式并行训练加速库

支持混合精度、MoE、通信优化等特性

NV MegatronLM

Transformer 加速库

支持多维混合并行、跨节点预训练等特性

第三方推理服务

支持“0 代码”
快速对接

ONNX RUNTIME

业界推理模型标准

100+ 基础模型，定制模型零成本迁移

NV Triton

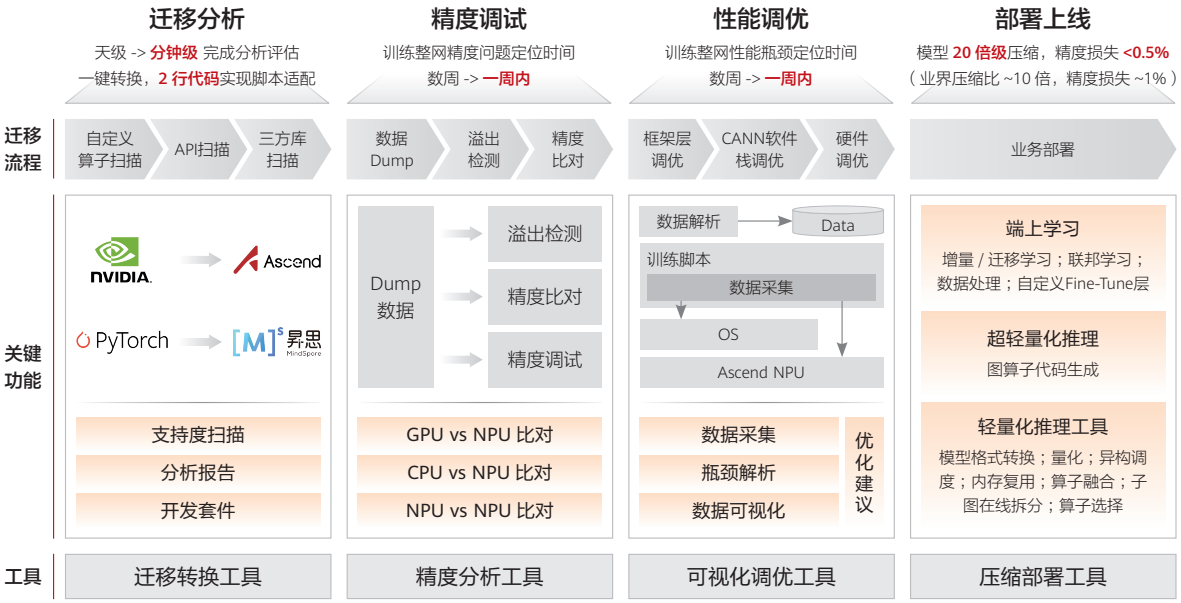
业界推理部署平台

支持并行推理、动态调度等关键特性

提供端到端昇腾迁移工具链，自动化迁移工作可从 4 周缩减至 2 周。

大模型迁移工具

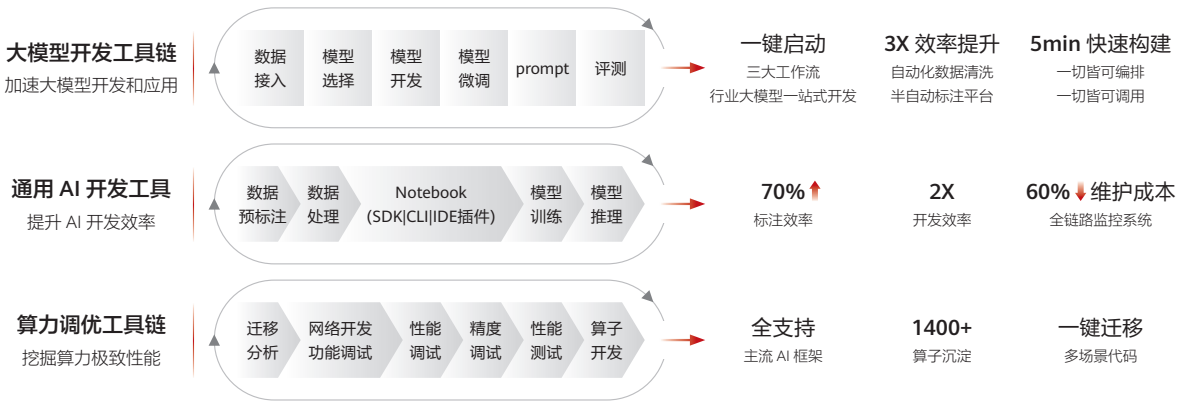
提供工具化端到端迁移调优服务，包括 30+ 可视化调优部署工具、自动化迁移工具，典型场景迁移至生产环境 <2 周，助力客户业务快速上线



大模型开发工具

华为云昇腾 AI 云服务提供从云化算力、模型开发、模型托管到生态的全栈服务，企业无需再次投资 AI 相关的通用技术，可以一键链接云上的开发平台，获取开发所需要的工具的套件。

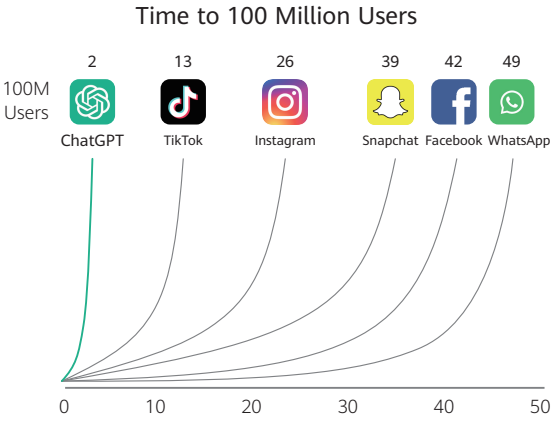
三大全流程工具链，一站式加速大模型敏捷开发



云上推理投资优 Investment Advantage

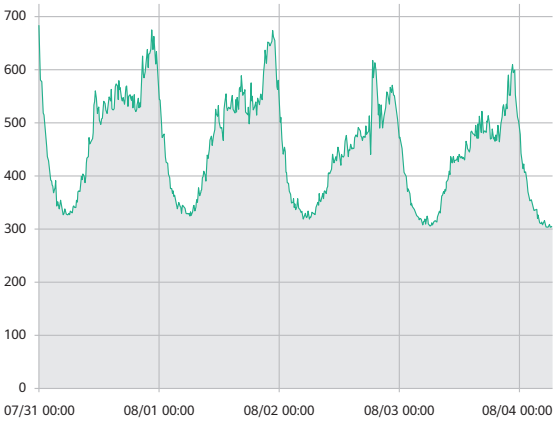
在云上，云计算弹性扩缩容支持业务快速增长的同时避免业务波谷时资源闲置。

新应用一旦顺利渡过孵化期拐点，
用户规模爆发式增长



单位：月

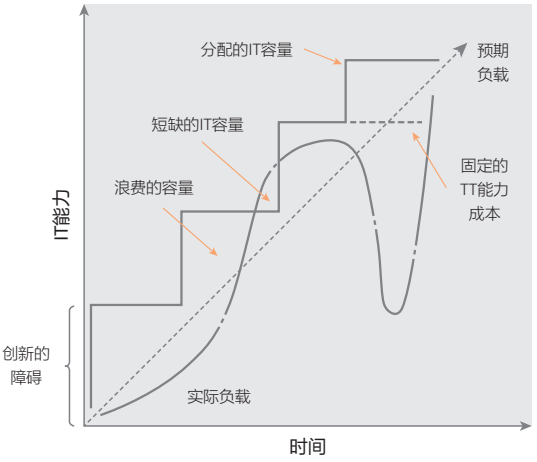
ToC 业务的推理，
存在明显的波峰波谷现像



单位：天

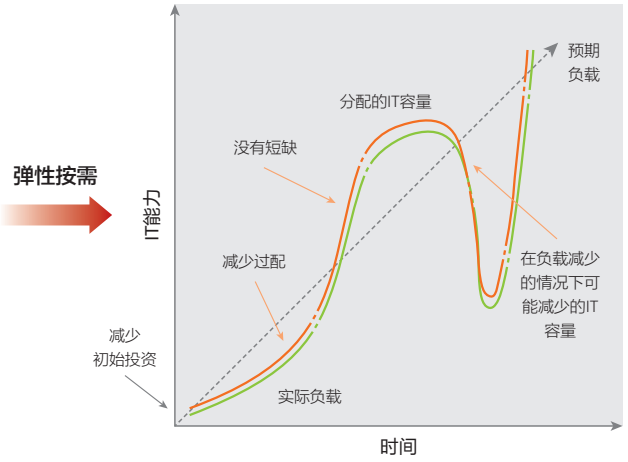
资源按需付费，让不确定的推理资产投资变得可控，成本控制更优。

业务曲线与资源曲线有 GAP
起步投资大，容易产生资源短缺与浪费



自建IDC

云资源可以根据业务情况灵活增减
起步投资小，更快进行开发和部署，提升利用率



云计算

云上，资源按需付费，让不确定的推理资产投资变得可控，成本控制更优。

1 公有云模式 分钟级开通

Z 客户：华为云 20 分钟开通 1000 卡（自建需 3 个月），可随时弹性扩容，TCO 节省 30% 以上

2 专属云模式 1 个月内开通

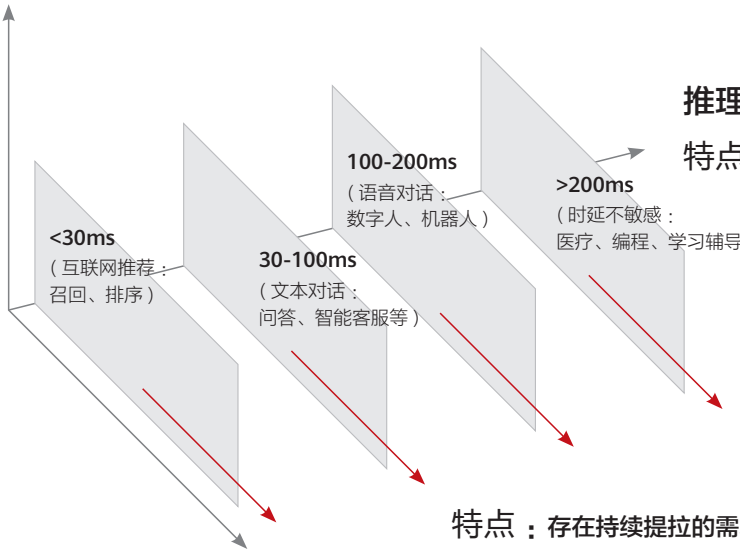
M 客户：购买贵安机房内 1000 卡，符合金融监管要求，华为云服务、DC 运营运维服务

3 自采自建云 3-6 个月开通

G 银行：线下自建昇腾云周期太长，考虑转向专属云模式

就近服务时延优 Latency Advantage

在云上，推理服务可以就近接入，实现超低时延优质服务体验。



推理时延（与应用场景强相关）

特点：达到场景诉求后无需持续提升

推理时延档位：

- <30ms：用户界面实时交互
- 30-100ms：文本交互
- 100-200ms：语音交互
- >200ms：非实时类业务

特点：存在持续提拉的需求

并发数（同等资源，并发越高，性价比越高）

推荐业务主要诉求是低时延 & 高精度

Case：

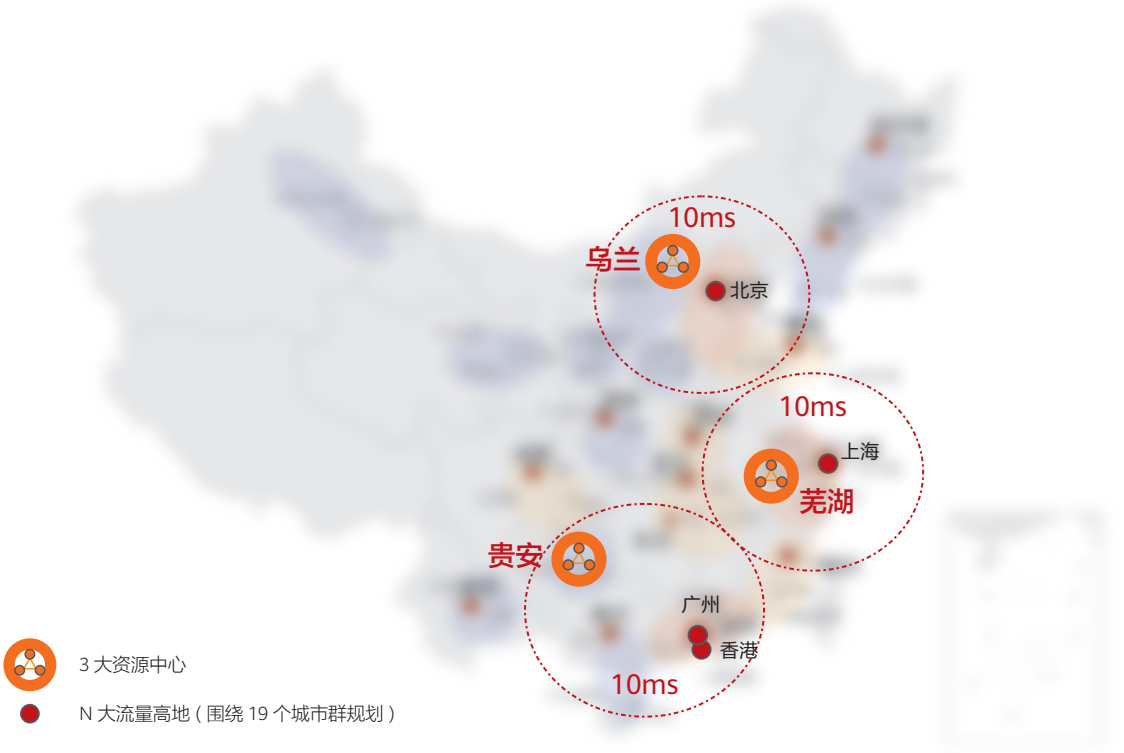
用户登录 **APP

基于用户特征推理

** 首页商品广告推荐

端到端应用时延：以 500 公里范围为例

在云上，推理服务可以就近接入，实现超低时延优质服务体验。



围绕三大数据中心构建核心的训练推理大集群 10ms 可达

流量高地（支撑 X 十万级规模）

北京：规模最大

- 7万+服务器，400万+核资源

上海：金融高地

- A类机房，金融等保4级，70+柜金融基础设施，最大金融专区

广州：出海桥头堡

- 广州&深圳双POP&AZ就近接入
- 跨境电商基地，出海时延低至5ms

香港：覆盖亚太

- 4AZ部署，50ms覆盖亚太区域
- 跨境出海首选

资源中心（支撑 X 百万级规模）

乌兰察布：算力 &AI 中心

- 70 万核，全球最大渲染超算基地
- 大规模集群支持千亿、万亿参数大模型训练与推理

芜湖：全新技术加持

- 华东枢纽节点，规划百万级服务器
- UB 网络、Grid 架构、IPv6、管理区云原生化

贵安：东数西算枢纽，PUE 最低

- 东数西算中心，国家 8 大枢纽节点之一
- PUE1.12，国家节能示范基地

云上性能优 Yield Advantage

在云上，通过持续的算子优化，显存优化，通讯优化可以显著提升集群性能，线性度 >90%。

Model Flops Utilization 用来衡量 AI 集群的算力利用率

通讯优化

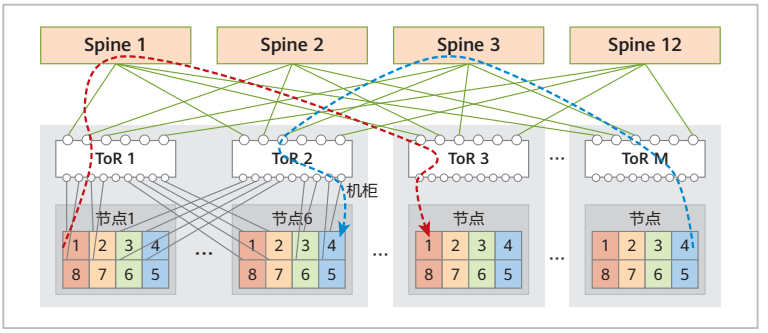
√ **MFU** = (1-AllReduce 占比 -All2All 占比 -Bubble 占比 - 无法掩盖的内存转移占比) × Mac 利用率

算子优化 + 显存优化

通讯优化

通信零冲突、零拥塞

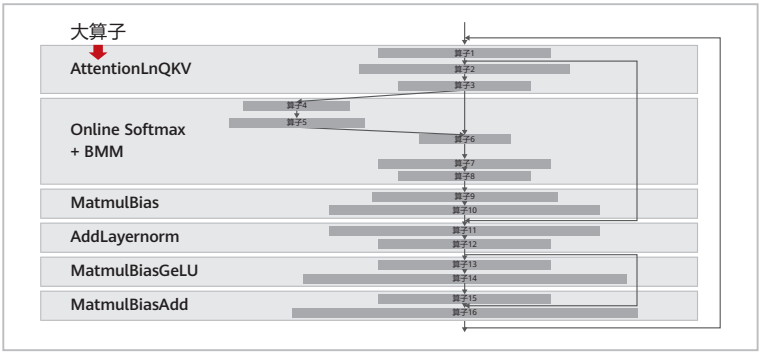
动态路由算法
智能编排通信路径



算子优化

小算子融合成大算子

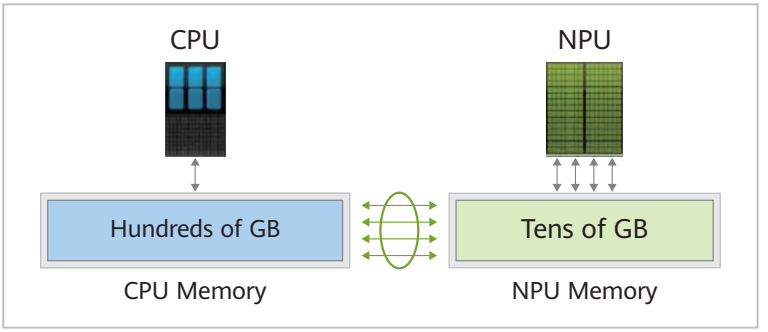
减少 launch 时间和
内存访问



显存优化

ZeRO-Offload

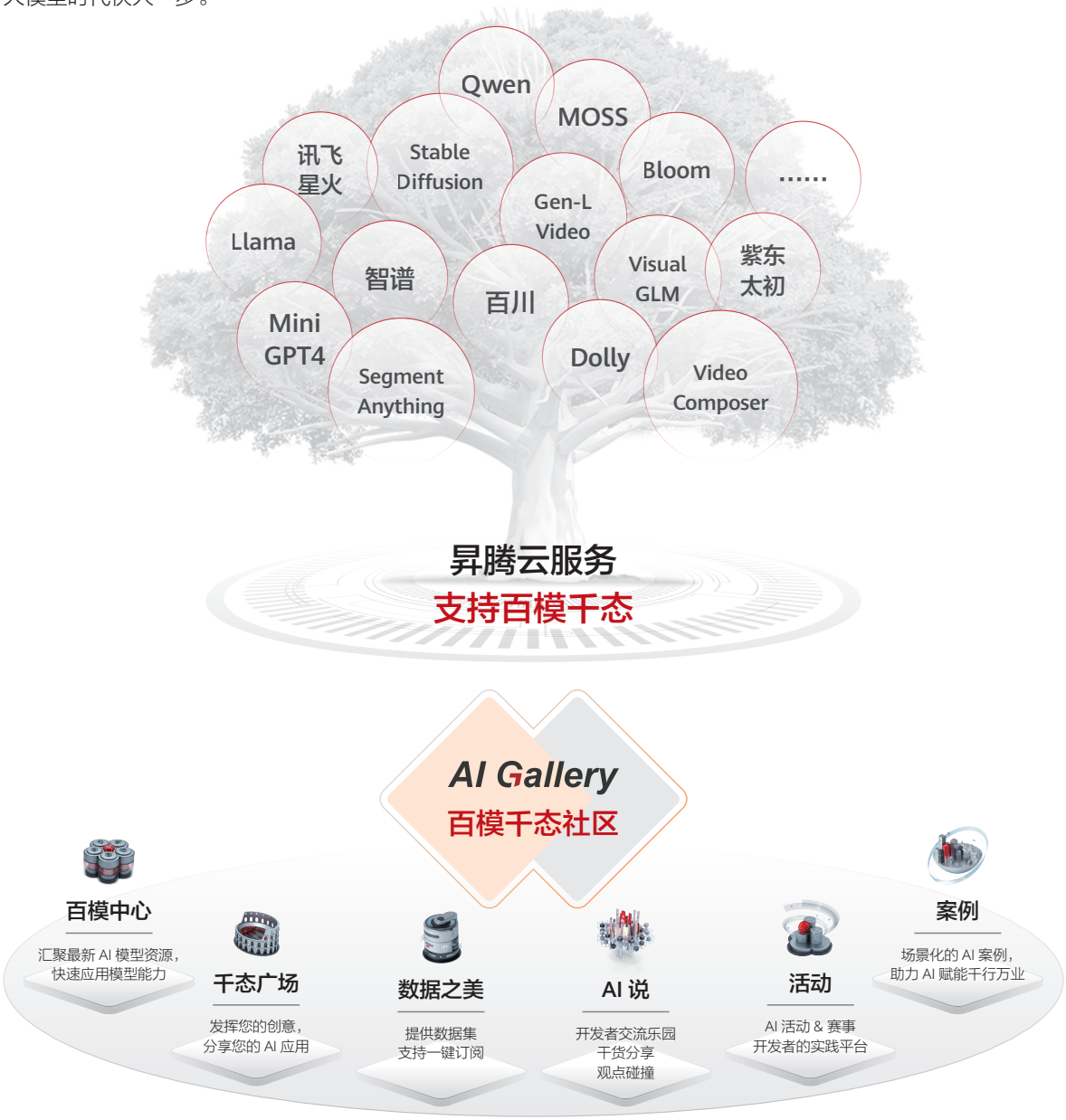
在显存中直接进行通信和
同步，不再需要通过网络
或主机内存



昇腾云服务
开放兼容支持百模千态

AI Gallery：一站式 AI 社区服务平台，构建百模千态的开放昇腾社区

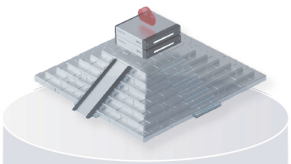
AI Gallery 百模千态社区，基于昇腾云服务算力底座，致力于构建一站式 AI 社区服务平台，包含丰富 AI 资产、服务、解决方案。适配业界主流开源大模型，易用开发工具和超强算力，助力企业和开发者快速创建模型应用，在大模型时代快人一步。



D-Plan：生态伙伴计划

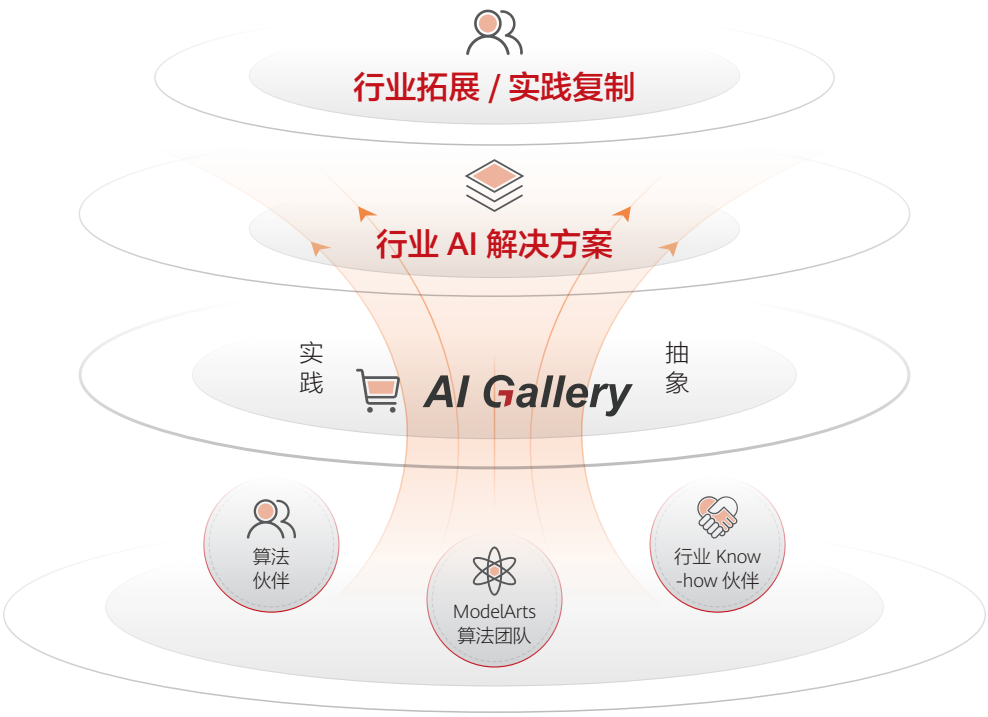
D-Plan AI 生态伙伴计划是围绕华为云 AI 开发生产线 ModelArts 推出的一项合作伙伴计划，旨在与合作伙伴一起构建合作共赢的 AI 生态体系，加速 AI 应用落地，华为云向伙伴提供培训、技术、营销和销售的全面支持。

以行业实际场景
需求出发践行
普惠 AI

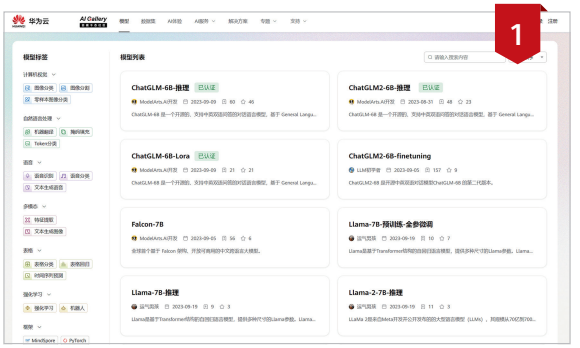


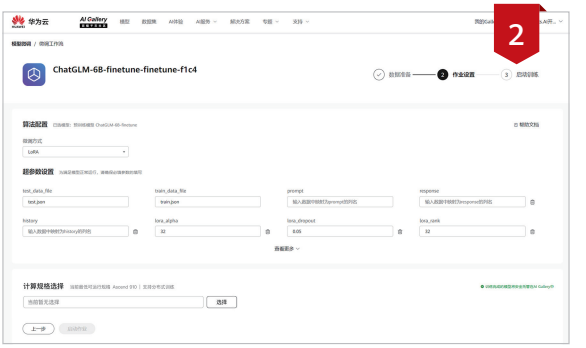
与行业 Know
-how 一起构建
行业竞争力

联接供需两端共建
百模千态社区



简单易上手的开发流程，帮助企业和开发者快速创建模型应用



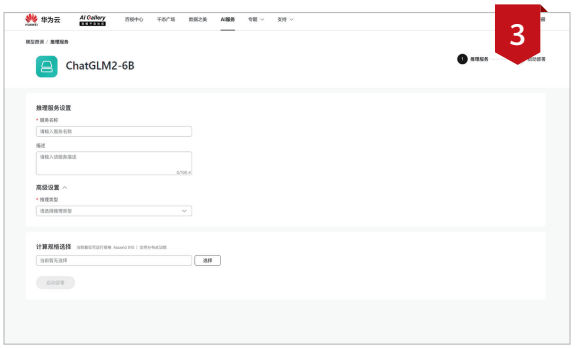


选择模型

选择经昇腾适配优化后的模型

启动微调训练

选择预置数据训练模型



进行推理部署

将模型部署为在线服务

验证模型效果

调用模型服务验证效果

AI Gallery官网二维码

昇腾 AI 云服务 — 全球行业先行者

华为小艺

大模型给小艺带来体验的全面提升，大模型加持下的小艺日人均使用时长相比之前提升了 15 倍，人均对话次数提升了 1.8 倍。小艺大模型能力的升级根植于华为云昇腾 AI 云服务算力黑土地，通过华为云 ModelArts 管理大规模算力集群，提升可靠性与性能，降低成本，打造训推一体资源底座，支撑小艺的日常预训练在线推理，支撑千万用户在线使用。

科大讯飞

科大讯飞是全球知名的智能语音和人工智能头部企业，通过华为提供昇腾集群进行讯飞星火大模型训练，训练性能整体提升 17%。昇腾 AI 云服务不仅可以快速提供数百卡的推理资源，也可以根据业务上线情况随时调整资源使用量，不会导致投资浪费。华为云全球算力布局也支持了科大讯飞业务出海，共同服务全球企业智能化。

华为云数字人

华为云 MetaStudio 数字人，依托昇腾 AI 云服务的澎湃算力，提供数字人快速生成及定制服务，具备数字人视频制作、视频直播、智能交互、企业代言等多种服务能力，可大幅提升视频制作、直播效率，重塑数字内容生产。

网易伏曦

网易伏羲与华为云进行技术联合创新，依托云原生技术构建了 AI 多云平台，并进一步适配华为云昇腾 AI 云服务，在算子层和框架层进行大量性能优化，满足交互场景的秒级时延要求，保障玩家流畅丝滑的互动体验。

此外，云原生技术可实现游戏服分钟级部署，4000 容器分钟级扩容，轻松应对玩家流量洪峰，让智能 NPC“忙时不慌，闲时不废”。

美图

美图自研 AI 视觉大模型 MiracleVision（奇想智能），广泛应用于电商、广告、游戏、动漫、影视五大行业，帮助细分领域设计场景提升效率。

在华为云昇腾 AI 云服务的助力下，将文生图、图生图等场景使用到的模型迁移到了昇腾 AI 云服务上，双方共同进行了 30 多个算子的优化以及流程的并行加速，迁移后，美图 AI 绘画等业务推理提升 30%，帮助企业更好地实现降本增效。

HKGAI

HKGAI 于 2023 年 10 月成立，是创新香港研发平台下唯一专注于生成式人工智能的研究及开发的中心，开发了香港本地首个自主训练的基础大模型。华为云为 HKGAI 提供云原生服务，保障线上应用的平稳运行，并将支持多元化、高效、稳定的算力选择，实现可持续的高效创新。面向未来 HKGAI 也在昇腾云等领域和华为云探讨合作，携手为香港人工智能产业创新注入新动能。

合合科技

华为云与软件伙伴合合信息构建联合解决方案，基于昇腾云服务在香港提供的 AI 算力及跨境可信网络与合规框架，承载智能文档解晰、商业大数据及合规审计、风控管理等多个子功能，可为企业提供资质的验真、分类、识别等 AI 辅助能力，加快基金申请审批速度，优化端到端流程服务，大幅降低相关人力投入。


与我们联系

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com