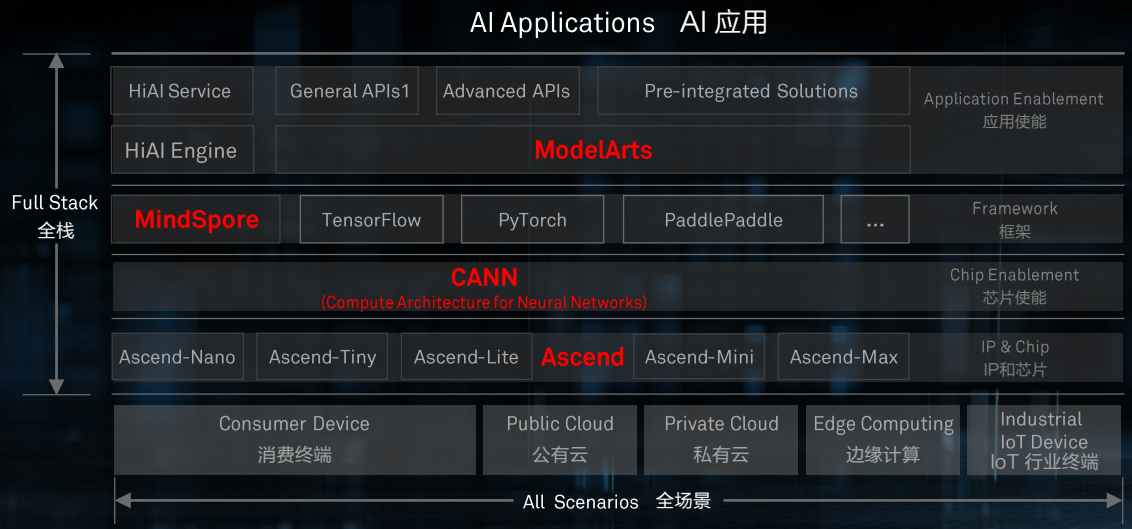
Huawei’s AI Portfolio：Full-Stack, All-Scenario

全栈全场景AI解决方案

Huawei’s AI portfolio
华为AI解决方案

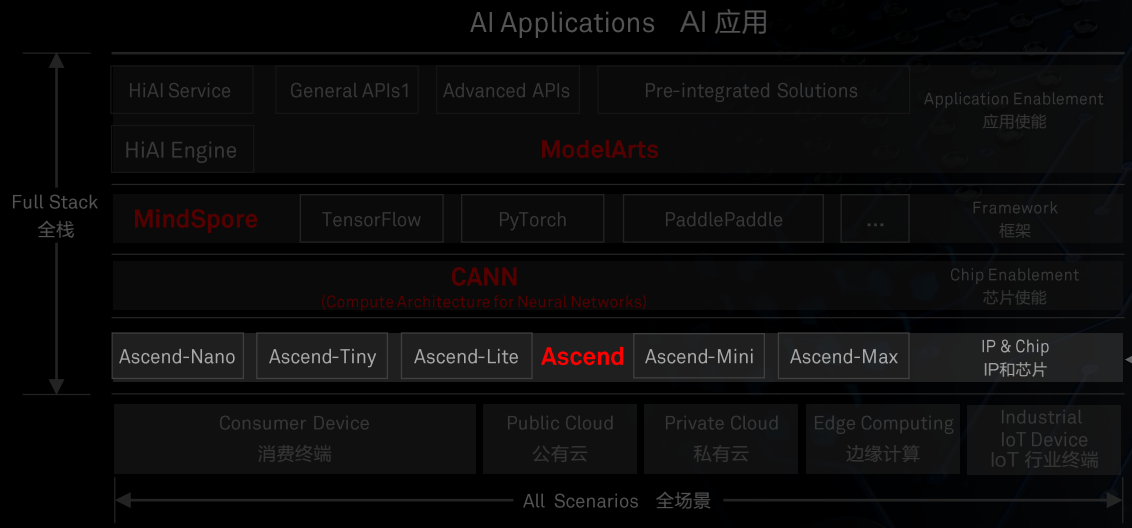


Application enablement
Full-pipeline services(ModelArts), hierarchical APIs, and pre-integrated solutions
应用使能：
提供全流程服务(ModelArts), 分层API和预集成方案

MindSpore：
Unified training and inference framework for device, edge, and cloud (both standalone and cooperative)
支持端、边、云独立的和协同的统一训练和推理框架

CANN：
Chip operators library and highly automated operators development toolkit
芯片算子库和高度自动化算子开发工具

Ascend：
AI IP and chip series based on unified scalable architecture
基于统一、可扩展架构的系列化AI IP 和 芯片



Optimal performance with minimal cost for all scenarios

实现任何场景下以最低成本获得最优性能

Pervasiveness = Diversity
无处不在 = 多样性

	Device 端				Edge 边缘		Cloud 云	
	Earphone 耳机电话	Always-on	Smartphone 智能手机	Laptop 便携机	IPC	Edge Server 边缘服务器	Data Center 数据中心	
Compute 算力	20 MOPS	100 GOPS	1-10 TOPS	10-20 TOPS	10-20 TOPS	10-100 TOPS	200+ TOPS	>10^7 x
Power budget 功耗	1 mW	10 mW	1-2 W	3-10 W	3-10 W	10-100 W	200+ W	>200,000x
Model size 模型大小	10 KB	100 KB	10 MB	10-100 MB	10-100 MB	100+ MB	300+ MB	>30,000x
Latency 延时	< 10 ms	~10 ms	10-100 ms	10-500 ms	10-500 ms	ms ~ s	ms ~ s	>100x
Inference ? 是否推理?	Y	Y	Y	Y	Y	Y	Y	
Training ? 是否训练?	N	N	Y	Y	Y	Y	Y	
Ascend-SKU Ascend-系列	Nano	Tiny	Lite	Mini	Mini	Multi-Mini	Mini	

A unified architecture or not ?
是否统一架构?

One-time Ops development
一次性算子开发

Consistent development & debugging experience
一致的开发和调试体验

Smooth migration across device, edge, and cloud
开发一次，跨端、边和云的平滑迁移

Compute scalability　算力可扩展

- Scale out: Unacceptable power dissipation and area
- Scale in: Complicated scheduling and software
- Scale out: 难以承受的功耗和面积
- Scale in: 复杂的任务调度和软件

Memory wall　内存墙

- Ultra-high bandwidth　超高带宽
- Extremely low latency　极低延时

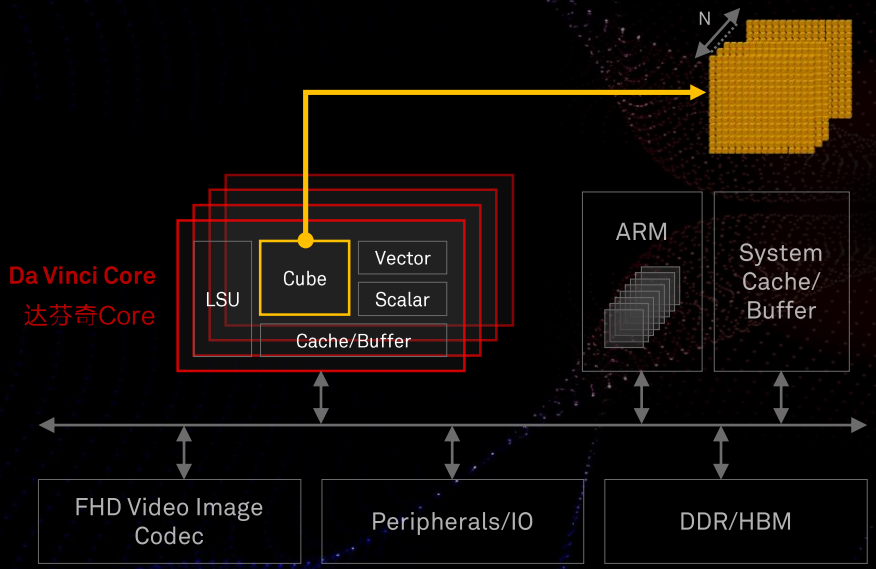
Interconnection　互联

- Power and area constraints　功率和面积约束

Benefits

Challenges

Unified Da Vinci architecture for all Ascend chips
Ascend 芯片，统一达芬奇架构！



Scalable Compute：
Scalable cube: 16x16xN， N=16/8/4/2/1
Multiple precision: int8/int32/FP16/FP32
Multiple Compute units: Tensor/Vector/Scalar

Cube：4096(16^3) FP16 MACs + 8192 INT8 MACs
Vector：128 (8*16) FP16 vector

Current control in picoseconds
Hardware-assisted task scheduler

可扩展计算：
可扩展Cube: 16x16xN， N=16/8/4/2/1
多精度支持: int8/int32/FP16/FP32
多种计算单元: Tensor/Vector/Scalar

Cube：4096(16^3) FP16 MACs + 8192 INT8 MACs
Vector：128 (8*16) FP16 vector

皮秒级电流控制
硬件辅助的任务调度

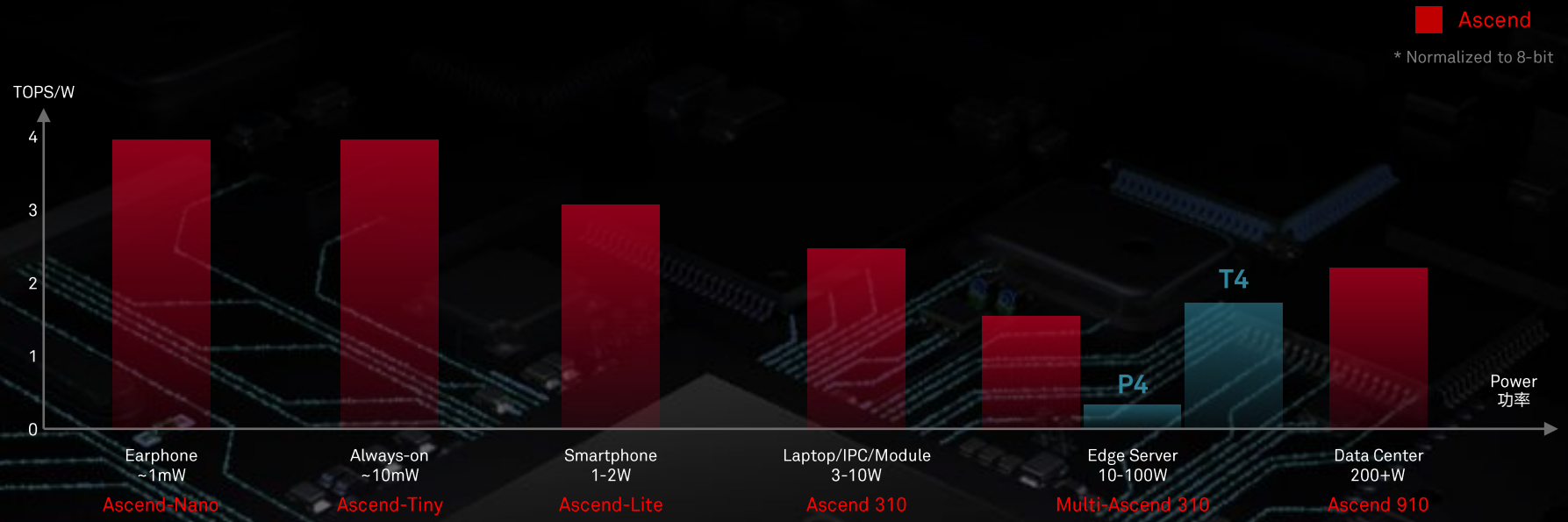
Scalable Memory：
Dedicated & distributed, tiling-friendly, explicit memory design
4 TByte/s L2 buffer
1.2 TByte/s HBM

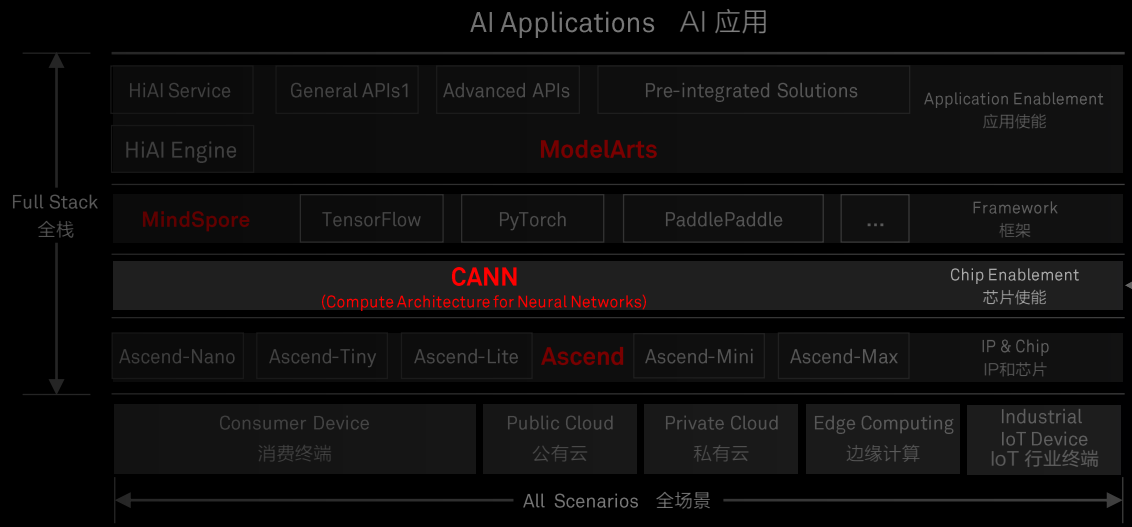
可扩展内存：
专用的和分布的，Tiling-Friendly，显式控制的内存分布设计
4 TByte/s L2 Buffer 缓存
1.2 TByte/s HBM 高带宽内存

Scalable On-chip Interconnection：
Ultra-high bandwidth mesh network on chip

可扩展片上互联：
片上超高带宽Mesh网络

Ascend: Optimal TOPS/W across all scenarios
Ascend系列，横跨全场景的最优TOPS/W

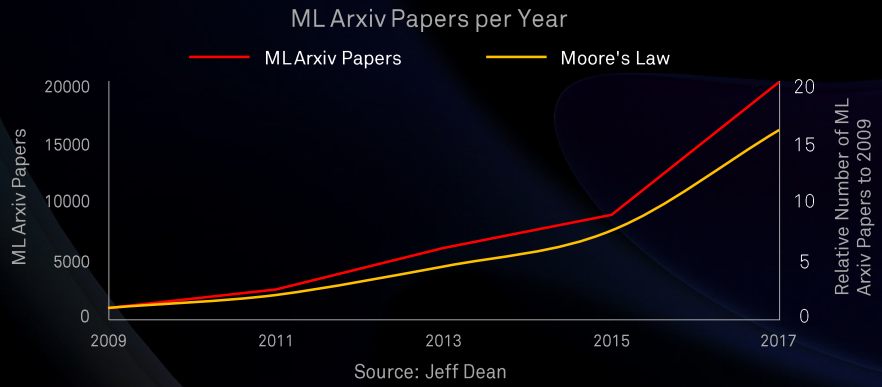




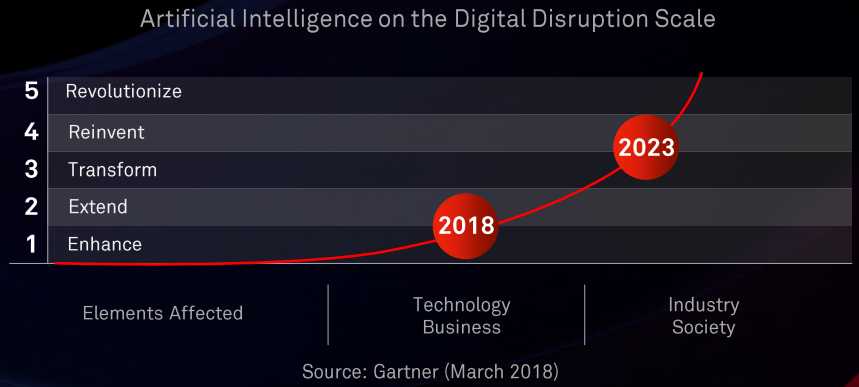
Deliver optimal performance and development efficiency of Ops simultaneously to cope with booming academic research and industrial applications

兼具最优开发效率和算子性能，
以应对学术研究和行业应用的蓬勃发展

Entering an era of dual prosperity
迈向双繁荣时代

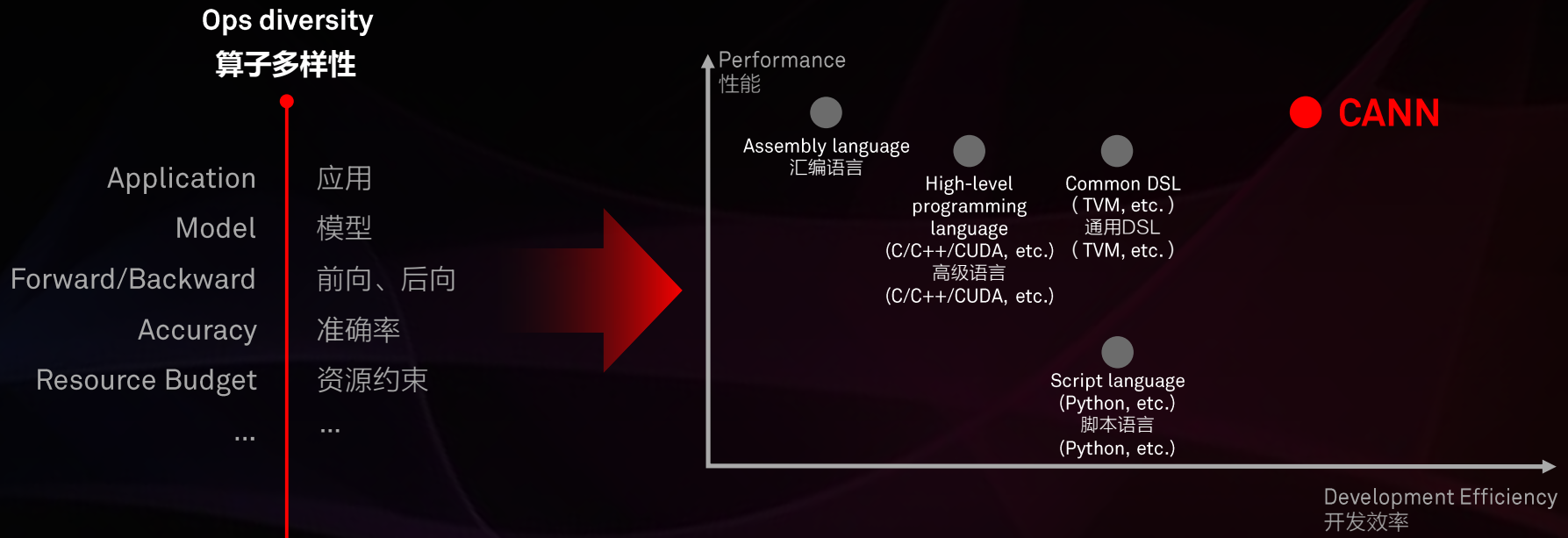


Source: Jeff Dean

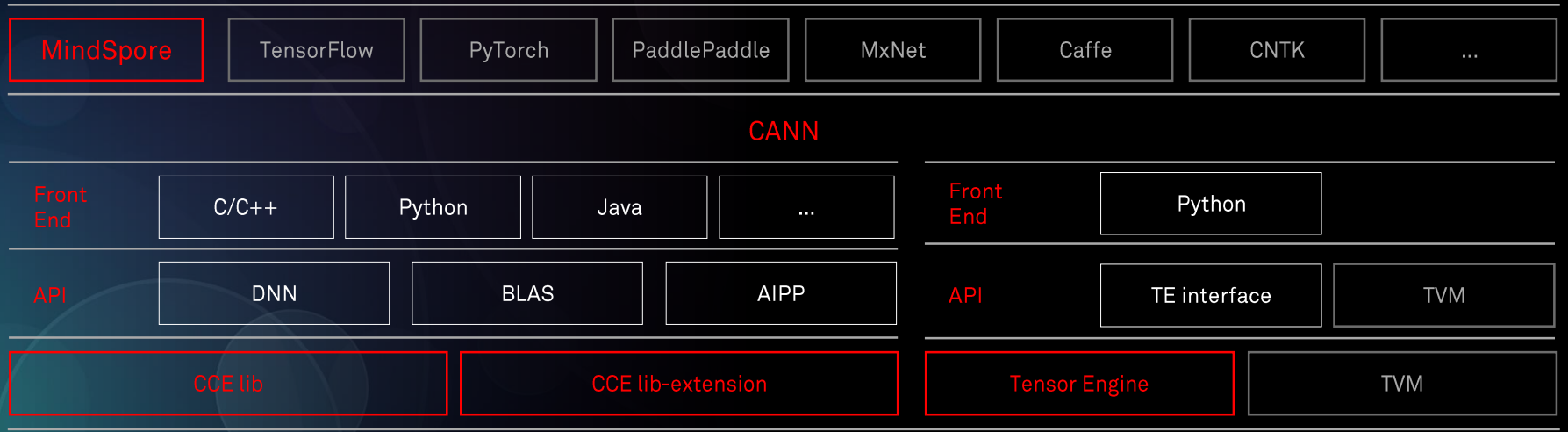


Source: Gartner (March 2018)

CANN: Coping with growing Ops diversity
CANN: 应对蓬勃发展的多样性



CANN: High performance and efficiency
CANN: 高性能+高效率



Huawei-developed
Extreme performance

华为开发
极致性能

Non-Huawei developers

非华为开发

One DSL interface
Auto optimization
Auto generation
Auto tuning

统一DSL接口
自动算子优化
自动算子生成
自动算子调试

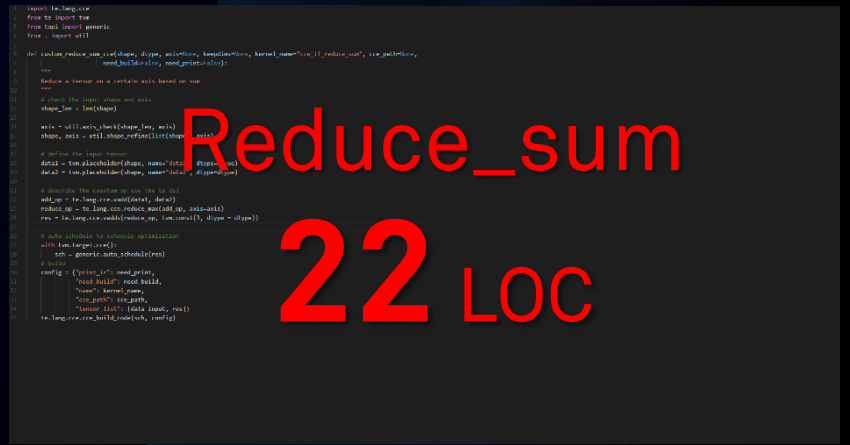
TVM-supported

支持TVM

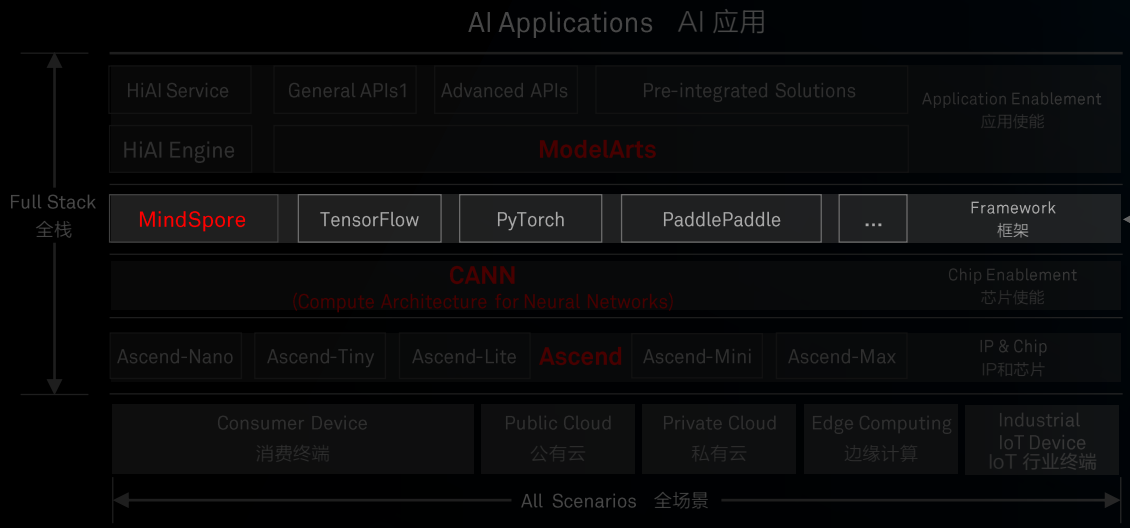
CANN: ~3x greater development efficiency
CANN，提升开发效率~3x

Common DSL

Tensor Engine DSL



Example using “Reduce_sum” ops



An AI framework that is:

Design-time-friendly:
Such as dramatically reducing training time and costs
Runtime-efficient:
Such as using the least amount of resources with highest OPS/W
Adaptable to all scenarios:
Including device, edge, and cloud

AI 框架必须：

设计态友好：如大幅减少训练时间和成本…
运行态友好：如最小资源需求和最高能效比…
适应所有场景：包括端、边和云

The future of AI will be highly dynamic
迎接AI剧烈变化的未来

Trends

- Mission-critical AI
- Personalized AI
- AI across organizations
- AI demands outpacing Moore's Law
- ...

Challenges & Research

Acting in dynamic environments:

- R1: Continual learning
- R2: Robust decisions
- R3: Explainable decisions

Secure AI:

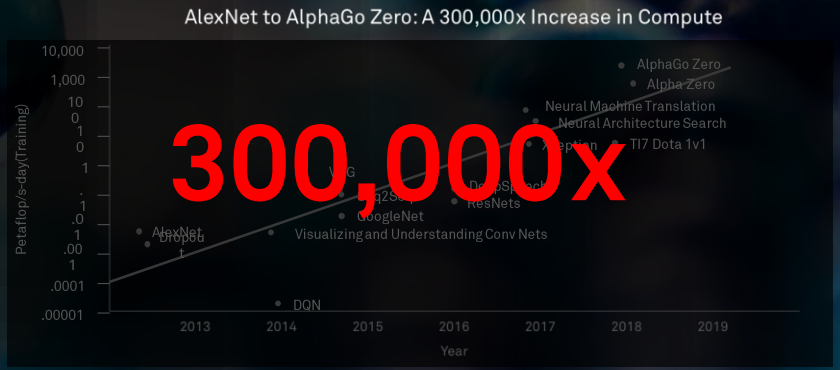
- R4: Secure enclaves
- R5: Adversarial learning
- R6: Shared learning on confidential data

AI-specific architectures:

- R7: Domain specific hardware
- R8: Composable AI systems
- R9: Cloud-edge systems

...

Source: M.Jordan



Source: OpenAI

Unified training and inference framework
统一训练/推理框架

Consistent Development Experience 一致开发体验

Cooperative Training/Inference

协作训练/推理

Device 端
Light 轻量

Cooperative Training/Inference

协作训练/推理

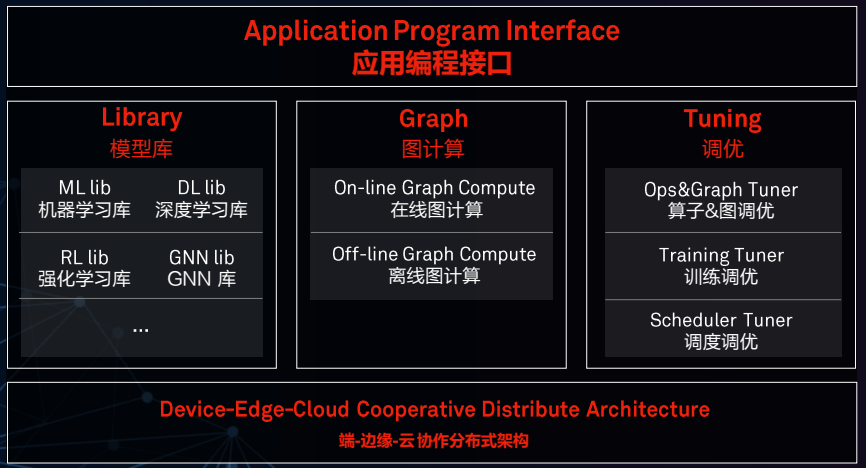
Edge 边缘
Local 本地

Cooperative Training/Inference

协作训练/推理

Cloud 云
Large-scale 大规模

MindSpore architecture
MindSpore 架构



Big or small: Flexible deployment for different resource budget environments	可大可小，灵活适应不同资源预算的部署环境
Device-Edge-Cloud cooperative training and inference	端 – 边 – 云协同的训练/推理
Unified distributed architecture for machine learning, deep learning, reinforcement learning, etc.	以统一分布式架构支持多种学习，ML/DL/RL..
Flexible API decoupled from core system	与核心系统解耦的编程接口，灵活适应多种语
2019 Q2	2019 Q2

On-device learning framework
On-device 学习框架

Framework < 2 MB
RAM < 50 MB
5x less ROM
2019 Q1

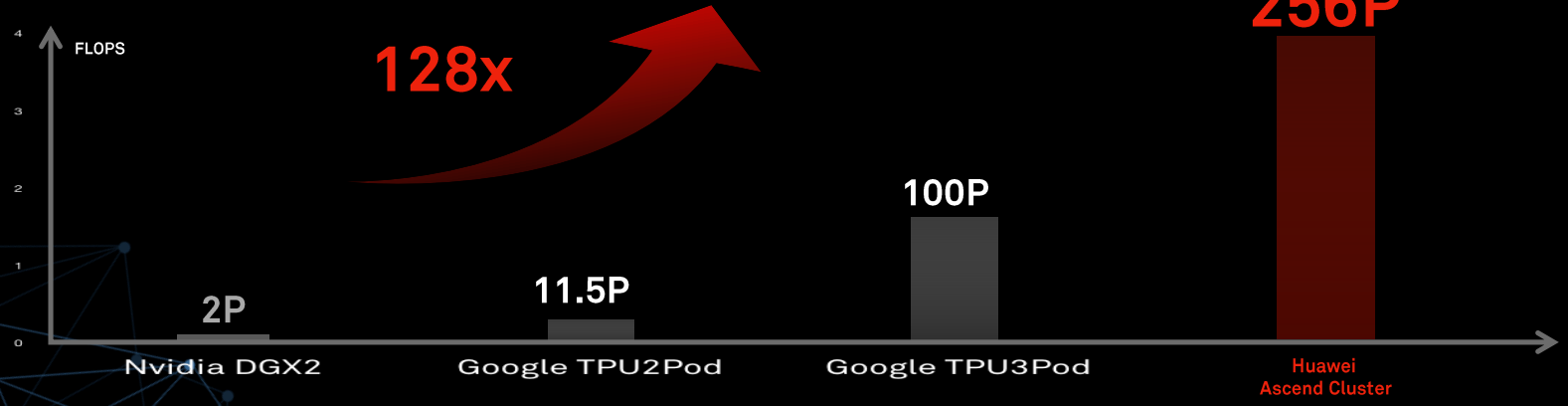
框架大小 < 2 MB
占用内存 < 50 MB
5x 或更小的存储空间需求
2019 Q1

Large-scale distributed training system
大规模分布式训练系统

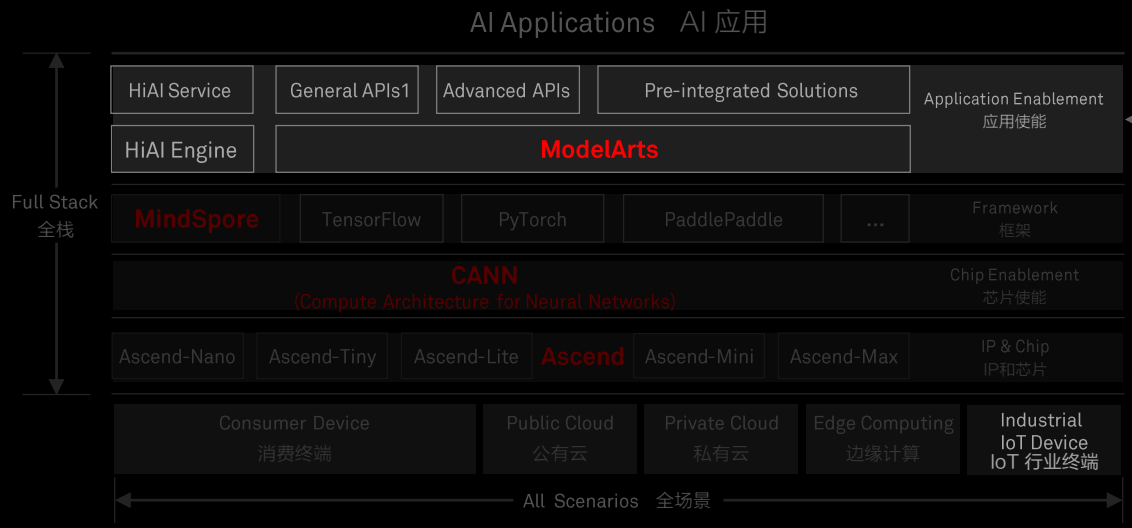
Ascend Cluster

256 PetaFLOPS
HBM: 32 TB
HBM BW: 8 Pbps
Network: 100 Tbps
Linearity:>90%
Parallelism: Data/Model/Hybrid
2019 Q2

256 PetaFLOPS
HBM: 32 TB
HBM 带宽: 8 Pbps
内部互联网络: 100 Tbps
线性度:>90%
并行模式: 数据/模型/混合
2019 Q2



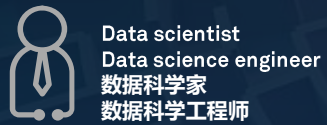
Supports all models using major frameworks
支持基于主要框架的模型



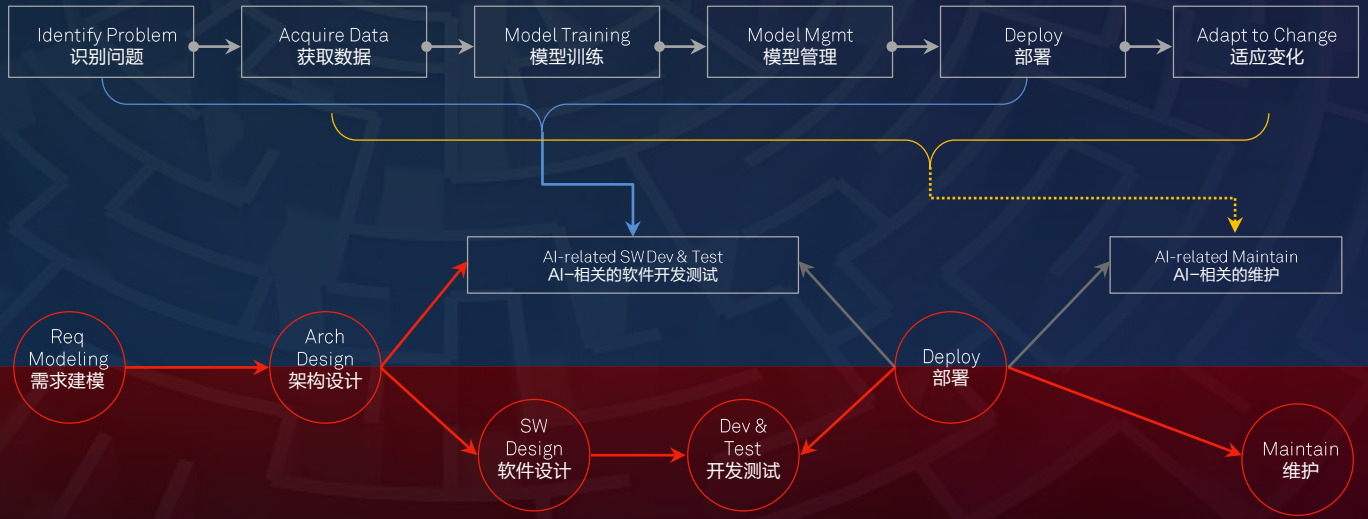
An ML PaaS to facilitate AI adoption by enabling developers of different needs with full-pipeline services, hierarchical APIs, and pre-integrated solutions

一个机器学习PaaS，提供全流程服务、分层分级API以及预集成方案，满足不同开发者的不同需求，使AI的采用更加容易

AI is redefining application development
AI 重新定义应用开发



Data scientist
Data science engineer
数据科学家
数据科学工程师



App developer
应用开发者

ModelArts: Full-pipeline model production
ModelArts 全流程模型生产

Acquire Data
获取数据

- Process
- Labeling
- Common data set
- Versioning

- 处理
- 标注
- 公共数据集
- 版本管理

Model Training
模型训练

- Online script coding
- AI frameworks
- Pre-loaded algorithms
- Auto hype-parameter tuning
- Distributed cluster
- Model visualization
- **ExeML**

- 在线脚本编码
- 常用AI框架
- 内置算法
- 超参自动调优
- 分布式集群
- 模型可视化
- **ExeML**

Model Mgmt
模型管理

- Model warehouse
- Model tracing
- Precision tracking
- Model marketplace

- 模型仓库
- 模型溯源
- 精度追踪
- 模型市场

Deploy
部署

- Real-time cloud service
- Batch service
- Cooperative Device-Edge-Cloud
- API market

- 云端实时服务
- 云端批量服务
- 端边云协同
- API市场

Adapt to Change
适应变化

- Data distribution monitoring
- Accuracy monitoring
- Smart re-training data generation
- Auto model adaption
- Local learning

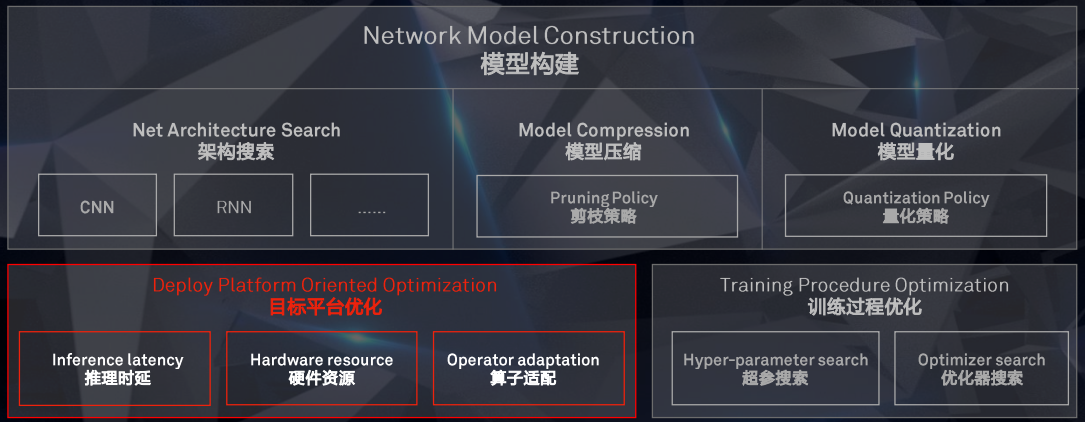
- 样本分布变化监控
- 预测精度偏差监控
- 数据智能推送
- 自动模型调整
- 本地学习

UI Console
UI 控制台

Python
SDK

ExeML: Production-oriented auto optimization
ExeML, 面向生产的自动优化

ExeML



Execution-oriented automatic model generation, and auto optimization that is adapted to different deployment environments

- Inference latency
- Hardware resource
- Operator adaptation

Designed for optimal production performance with minimal cost

2019 Q2

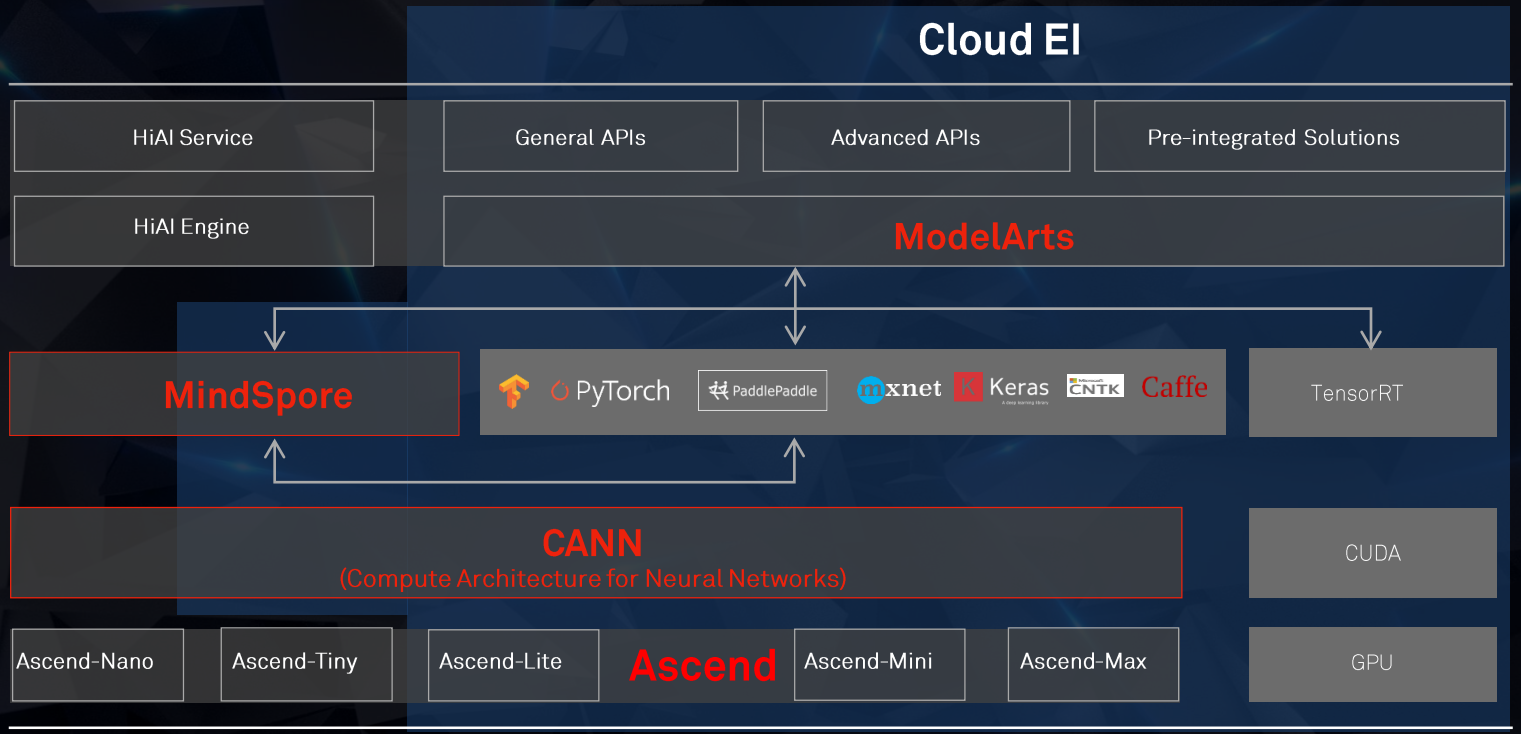
面向执行的自动模型生成和适应部署环境的自动优化

- 推理时延
- 硬件资源
- 算子适配

从一开始就通过设计以最低的成本实现最佳的生产性能

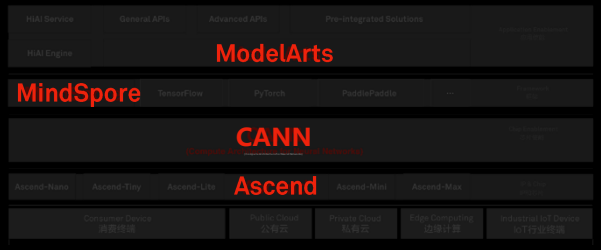
2019 Q2

Cloud EI platform supports GPU
Cloud EI 平台支持 GPU



Pre-integrated solutions for public and hybrid clouds
预集成解决方案，公有云/混合云

Pre-integrated solutions for private clouds
预集成解决方案，私有云



FusionMind

Atlas

AI training appliance
AI 训练一体机

Out-of-the-Box
开箱即用

20+ training models
20+ 训练模型

MindSpore / TensorFolw / Pytorch...

Video analysis appliance
视频分析一体机

256 real-time video channel analyses per node
256 通道实时视频分析/每节点

100 million picture analyses per node per day
1亿图片分析/每天每节点

Off-site enforcement appliance
外场执法一体机

TrafficGO

14 pre-integrated enforcement algorithms
14 执法算法预集成

Multi-algorithm shared computing power
多算法算力共享

OCR appliance
OCR 一体机

99.3% – Chinese character recognition accuracy
99.3% 汉语字符识别准确率

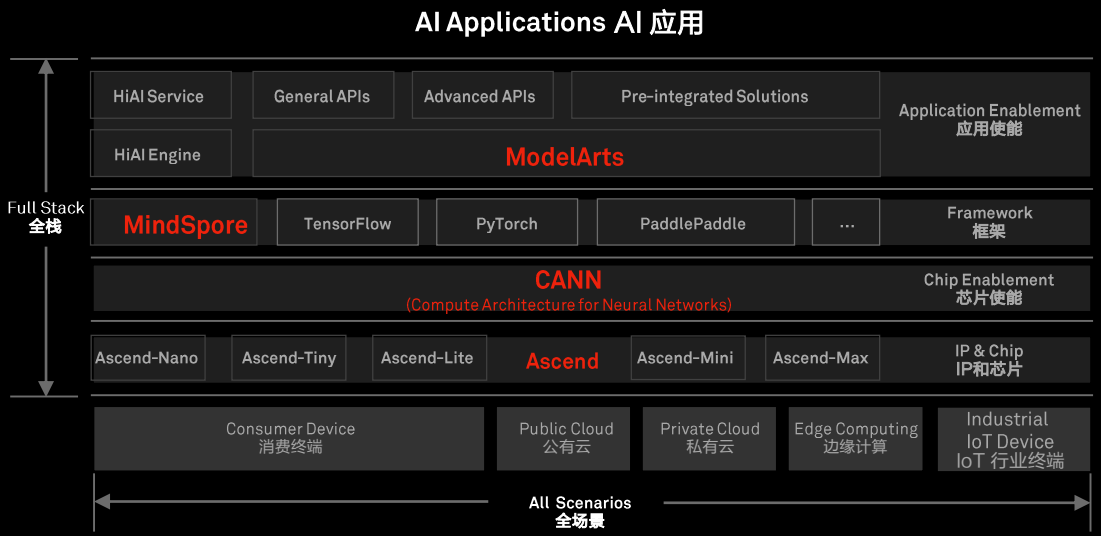
99.8% – Digit recognition accuracy
99.8% 数字识别准确率

< 1 second to recognize single A4 page
< 1 秒，单页A4识别

Summary
总结

Drive AI to new horizons with all-scenario native, full-stack solutions

以原生全场景的全栈解决方案把人工智能推向新高度



Application enablement
Full-pipeline services(ModelArts), hierarchical APIs, and pre-integrated solutions
应用使能：
提供全流程服务(ModelArts), 分层API和预集成方案

MindSpore：
Unified training and inference framework for device, edge, and cloud (both standalone and cooperative)
支持端、边、云独立的和协同的统一训练和推理框架

CANN：
Chip operators library and highly automated operators development toolkit
芯片算子库和高度自动化算子开发工具

Ascend：
AI IP and chip series based on unified scalable architecture
基于统一、可扩展架构的系列化AI IP 和 芯片

Thank you