

MindSpore架构介绍



目录

1. 整体介绍
2. 软件架构
3. 关键技术

华为AI：构筑业界最强AI算力平台，使能千行百业智能化转型



MindSpore开源深度学习框架

MindSpore

全场景统一API
Unified APIs for all scenarios

自动微分
Automatic
differentiation

自动并行
Automatic parallelization

自动调优
Automatic tuning

MindSpore IR 计算图表达
MindSpore intermediate representation (IR) for computation graphs

On-Device执行
On-device execution

Pipeline并行
Pipeline parallelism

深度图优化
Deep graph optimization

端-边-云按需协作分布式架构（部署、调度、通信等）
Device-edge-cloud cooperative distributed architecture (for deployment, scheduling, communications, etc.)

处理器：Ascend、GPU、CPU

开发态 友好：AI 算法即代码

通过自动微分、自动并行、自动调优等技术，减少代码，提升开发效率

运行态 高效：面向昇腾芯片优化

通过图/算子编译加速以及神经网络并行执行，协同昇腾芯片实现性能优势

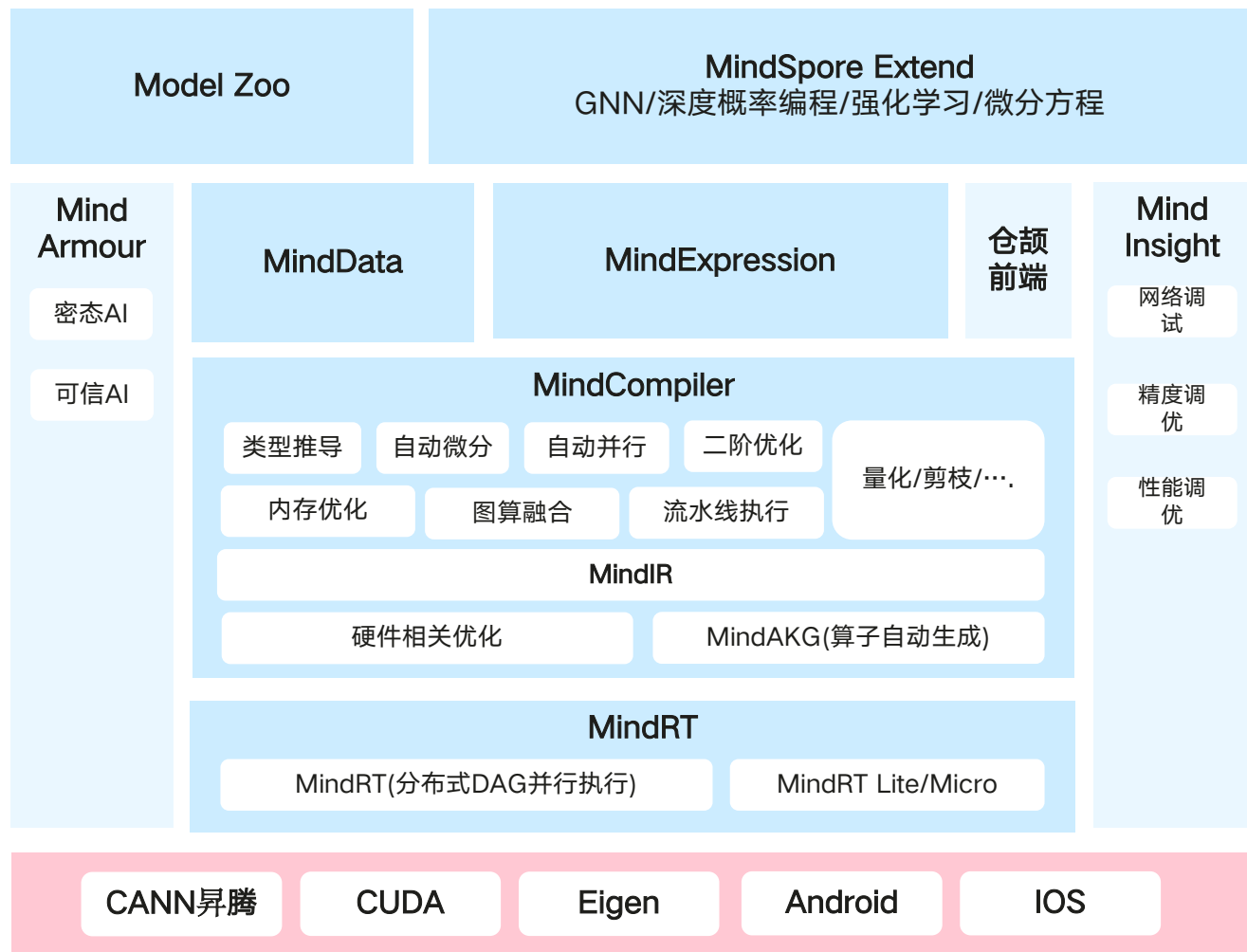
部署态 灵活：全场景按需协同

通过模型自适应、异构计算等技术，实现从IoT设备到云的灵活部署

目录

1. 整体介绍
2. 软件架构
3. 关键技术

MindSpore逻辑架构

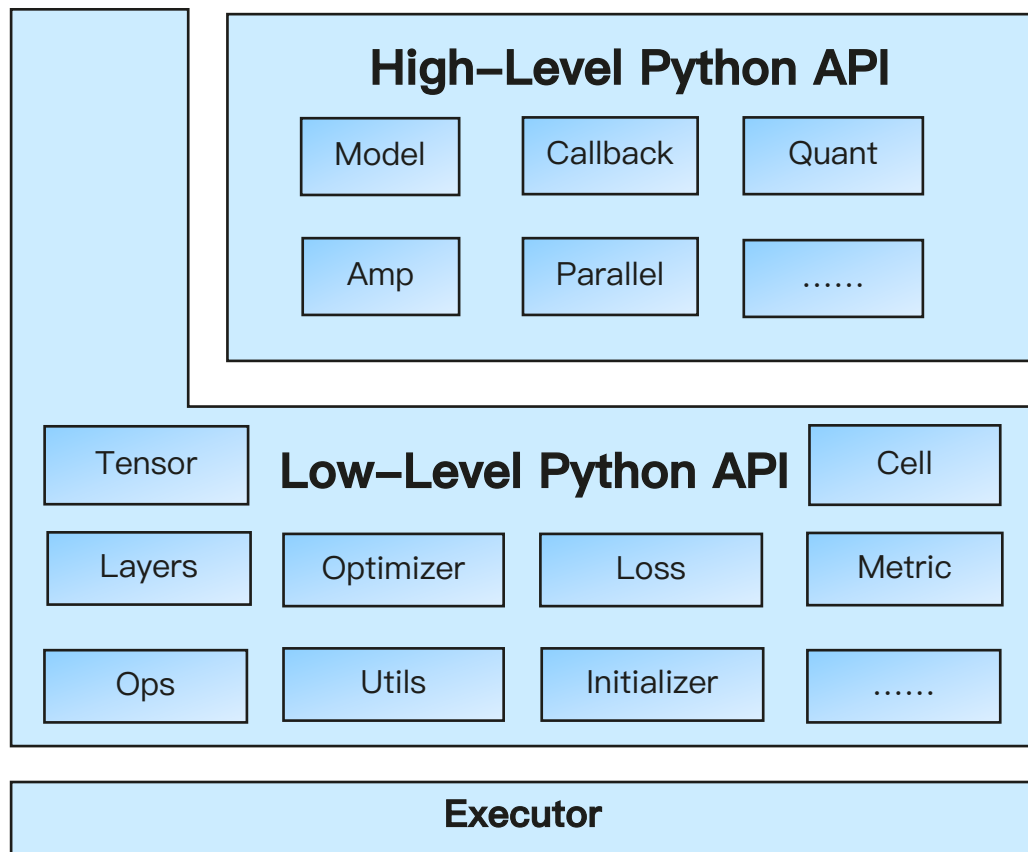


关键特性

- 1. 自动并行:** 动整图切分, 感知集群拓扑, 实现通信开销最小, 融合数据并行与模型并行;
- 2. 二阶优化:** 利用二阶计算修正梯度更新方向, 找到训练梯度最优下降路径, 从而加速训练收敛过程;
- 3. 动静态图结合:** 统一自动微分引擎支持动静态图, 一行代码完成模式切换, 兼顾模型开发和执行效率;
- 4. 全场景部署协同:** 超轻量技术, 模型压缩、自适应模型生成、IoT超轻量运行时, 基于异构的混合同行;
- 5. 全栈协同加速:** 通过多级并行、图算融合等深度图优化, 以及跨层内存复用技术, 发挥昇腾极致性能;
- 6. 迁移及调试工具:** 覆盖生态迁移、精度调优、性能调优全场景、可视化, 脚本迁移自动化, 降低开发门槛, 提高调试效率;
- 7. AI+科学计算, 场景应用创新, 拓展MindSpore的边界**
- 8. 安全可信, 首个企业级AI可信计算框架 (Trusted AI)**



MindExpress子系统



设计目标:

- 两层用户API设计（包括High-Level与Low-Level），支撑用户进行网络构建、整图执行、子图执行以及单算子执行
- 向用户提供统一的模型训练、推理和导出等接口，满足端、边、云等不同场景
- 动态图和静态图统一的编码方式
- 单机和分布式训练统一的编码方式

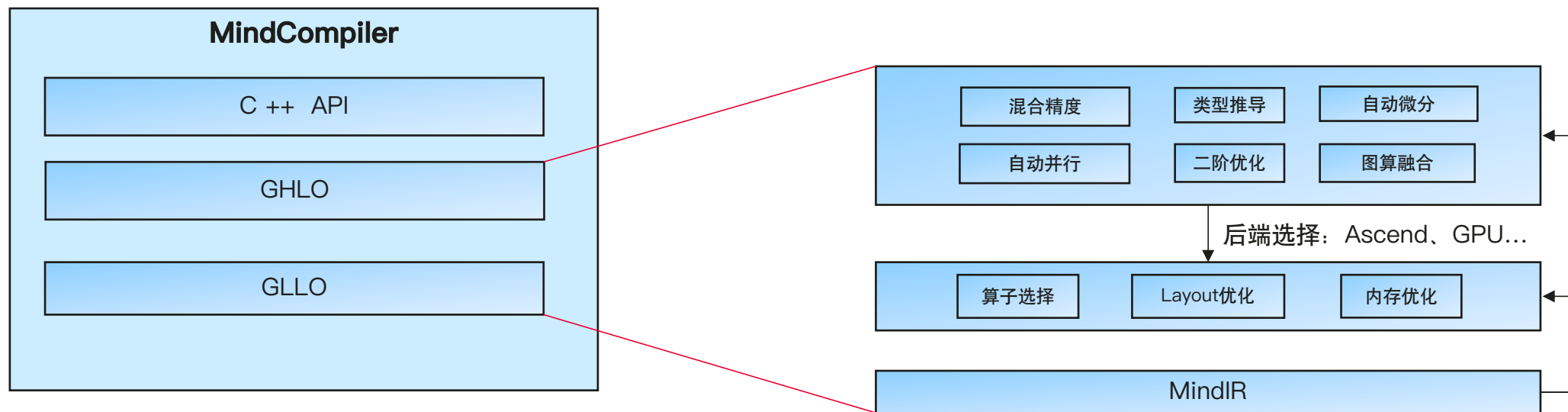
功能模块:

- High-Level API提供训练推理的管理接口、Callback、量化、混合精度、并行等控制接口，易于用户实现整网流程的控制
- Low-Level API提供基础的Tensor、Cell、NN-Layers、优化器、初始化等，易于用户灵活构建网络和控制执行流程
- Executor提供计算的执行控制，与MindSpore backend交互

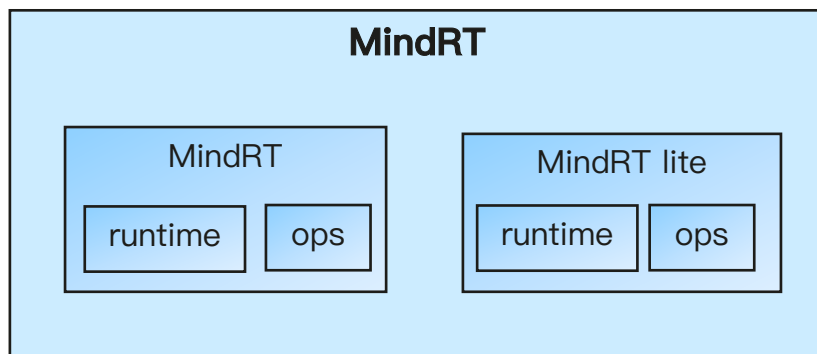
MindCompiler子系统

MindCompiler提供面向Mind IR的图级即时编译能力：

- 1) Graph High Level Optimization (GHLO)面向应用，进行偏前端的优化和功能，如类型推倒、自动微分、二阶优化、自动并行等
- 2) Graph Low Level Optimization (GLLO)面向硬件，进行偏底层的优化，如算子融合、layout优化、冗余消除、内存优化等



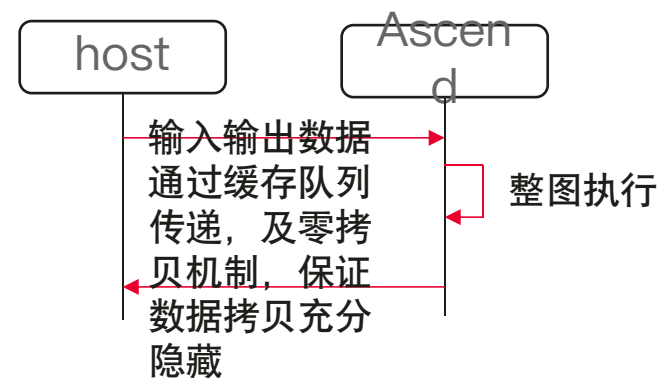
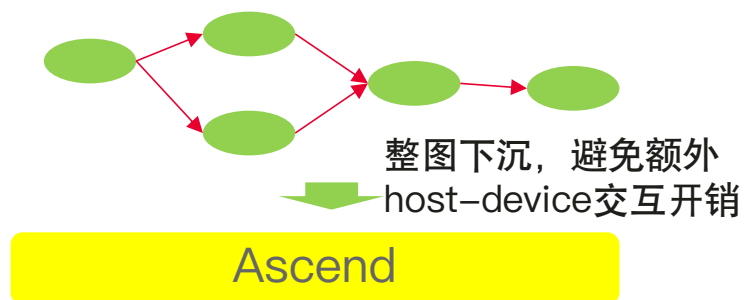
MindRT子系统



统一的运行时（RunTime）系统：

- 1、支持端、云多种设备形态要求
- 2、支持多种硬件设置的调度管理，如Ascend、GPU、CPU
- 3、内存池化管理，高效内存复用
- 4、算子异步、异构执行，多流并发

特色技术点

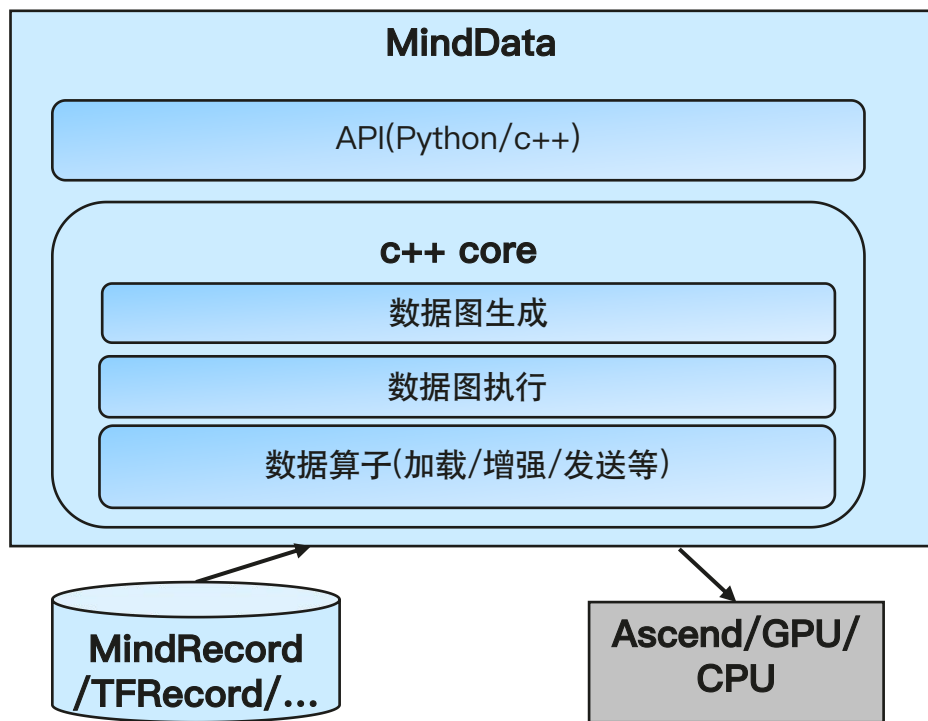


MindData子系统

MindData负责高效执行训练数据处理pipeline，与计算形成流水，数据及时导入训练。



典型训练数据处理pipeline



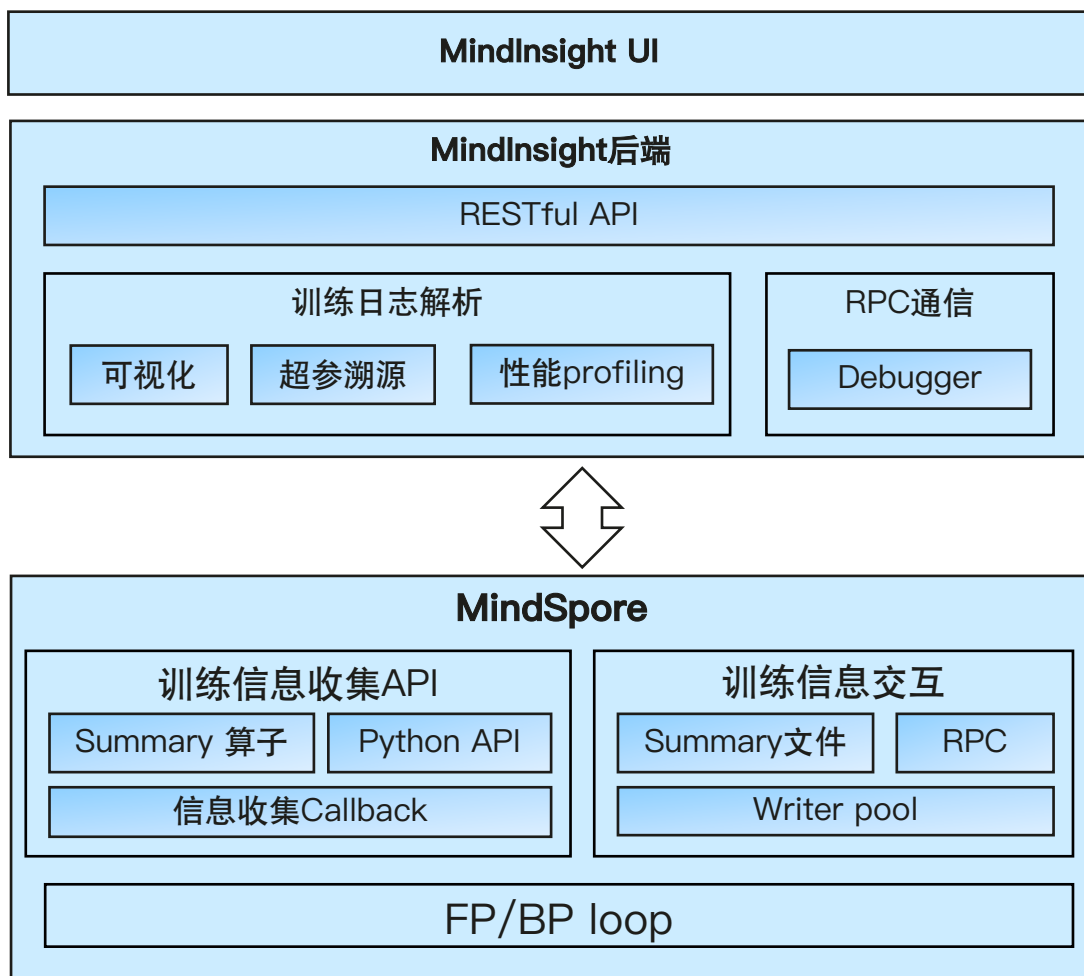
关键功能:

- 流水线+并行方式执行，提高数据处理吞吐量；
- 丰富的数据算子；
- 用户自定义python算子，以及用户灵活定制pipeline（数据加载、采样、增强等）；
- 异构硬件加速(Ascend/GPU/CPU)；
- MindRecord：自带元数据、聚合存储；

运行流程:

1. 数据图生成：根据用户的Python API调用，生成数据图；
2. 数据图执行：Pipeline并行执行数据图中的数据算子完成数据集加载、shuffle、数据增强、batch等处理；
3. 数据导入Device：处理后数据导入Device训练；

MindInsight子系统



MindInsight是MindSpore的调试调优子系统，提供训练过程可视化、模型溯源、debugger和性能profiling功能。

关键功能：

- 易用的API接口，在训练过程中，用户可以方便的收集训练过程指标，包括计算图、标量数据(loss/accuracy...)、直方图数据(梯度/权重...)、性能数据等，并通过Web UI界面进行展示。
- 通过收集训练的超参，数据集、数据增强信息实现模型溯源，并可在多次训练间进行对比

运行流程：

1. 训练信息收集：用户可通过callback接口，收集常用训练指标。用户也可以按需要收集自定义信息，如通过summary算子收集计算图中信息，通过Python接口收集Python层信息。
2. 训练日志生成：用户在训练过程中收集到的过程信息，最终会生成训练日志。
3. 训练信息展示：MindInsight通过打开并解析训练日志，以图形化方式为用户展示训练过程信息。

MindArmour子系统

MindArmour针对可信AI的各个领域提供全面、有效、易用的评测工具和增强方法。

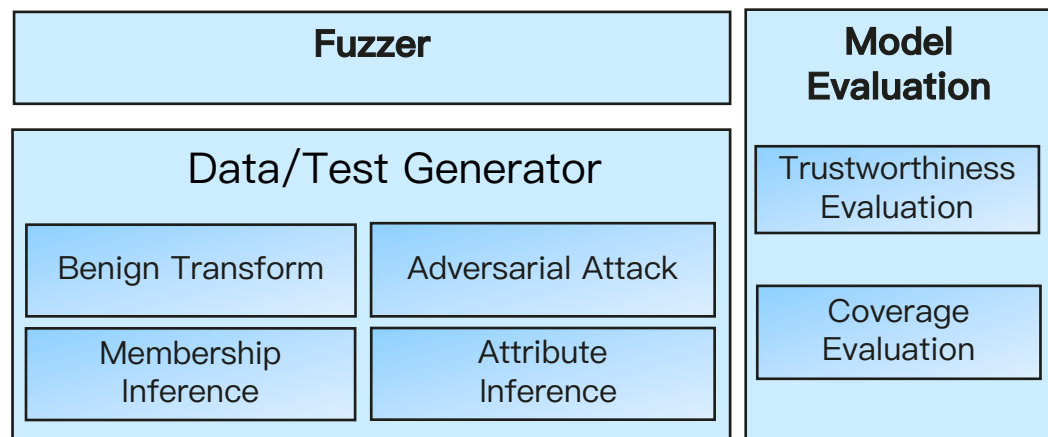


Where we from

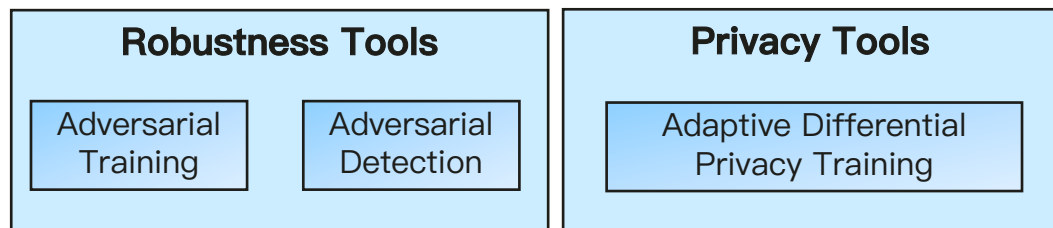
We are HERE

Where we go

AI Model Trustworthiness Test



AI Trustworthiness Enhancement



关键功能:

- 涵盖黑白盒对抗攻击、成员/属性推理攻击、数据漂移等测试数据产生方法，覆盖场景全面；
- 基于覆盖率的Fuzzing测试流程，灵活可定制的测试策略和指标；
- 包括对抗训练、输入重建在内的常见对抗样本检测和模型鲁棒性增强方法；
- 高效自适应差分隐私训练和预算统计算法，数学上可证明的模型隐私泄露约束；

运行流程:

1. 配置策略：根据威胁向量、可信需求定义测试策略，选择合适的测试数据产生方法；
2. Fuzzing执行：根据模型覆盖率和配置策略启发式地产生可信测试数据；
3. 产生评估报告：可以基于自带的或自定义的可信指标；
4. 可信增强：使用预置的方法增强AI模型可信程度。

目录

1. 整体介绍
2. 软件架构
3. 关键技术

关键特性1：自动并行

挑战：

超大模型与超大数据集的分布式训练，需通过数据并行+模型并行的混合并行方式，才能高效训练网络；

高性能：

- 传统graph-level模型切分，计算资源利用率不高，需通过operator-level模型切分提高并行加速比；
- 选择一种高效的模型切分方式需要专家经验；

易用：

- 混合并行复杂度高，传统API难以编写混合并行代码，算法与并行逻辑耦合，修改并行策略，就要重新修改编码；
- 算法科学家需要关注系统（集群拓扑、网络带宽等）和并行的实现细节，才能写出高性能算法；

方案

前端手自一体并行表达

(单行代码完成模型并行)

并行策略生成

算子切分

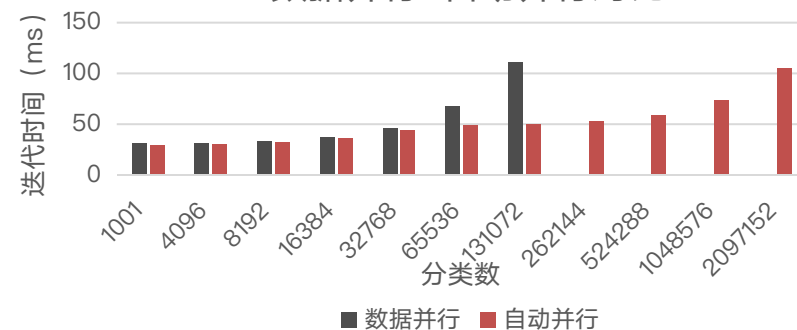
整图切分

拓扑感知子图调度

进展：

- **ResNet**: ResNet50网络在131072分类下，32卡混合并行线性度超过**92%**；
- **ReID**: 数据并行切换到混合并行，**手工代码143行 vs 自动并行1行**
- **推荐**: Wide&Deep单卡串行代码Host-Device混合并行，支持十亿特征数百GB模型推荐网络

数据并行-自动并行对比



- 通过**Optimzier**模型并行，PipeLine并行提升性能；
- 联合鹏城实验室发布业界首个**2000亿参数中文模型**【鹏程·盘古】

关键特性2：二阶优化

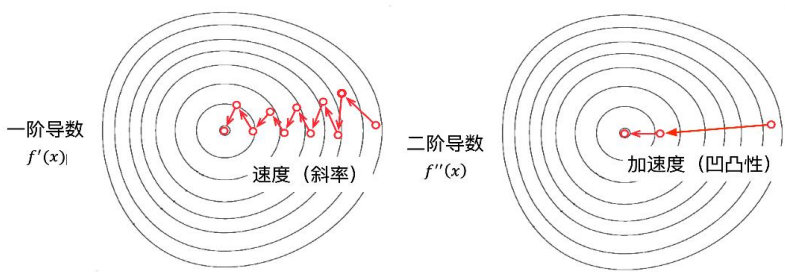
挑战

- 训练深度学习模型需要进行大量的计算，训练收敛时间长；
- 二阶优化方法可以有效加速模型收敛，减少迭代次数，同时会引入大量复杂计算，限制其在深度模型训练中广泛应用；
- 二阶优化器参数更新：

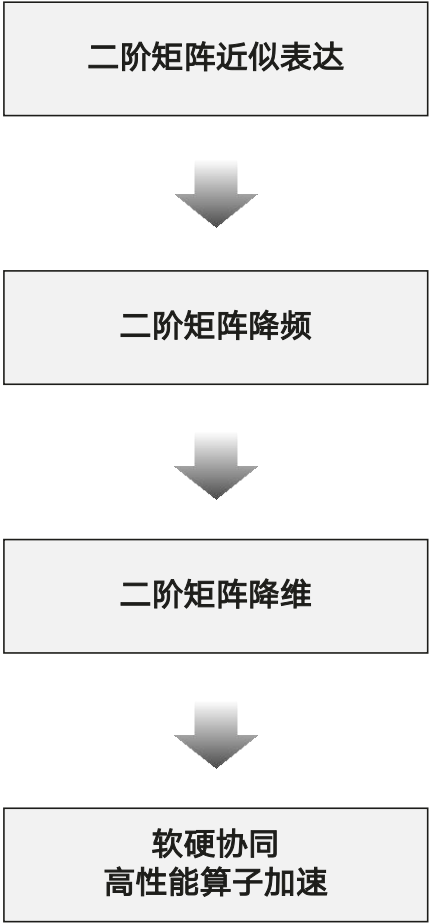
$$\theta^{(t+1)} = \theta^{(t)} - \epsilon I M_{\theta}^{-1} \Delta g(\theta^{(t)})$$

参数 学习率 二阶信息矩阵 一阶梯度

- 核心问题：二阶优化器需要额外计算二阶信息矩阵的逆矩阵，计算量大，直接求解时间可达小时级，如何**高效求解**二阶矩阵是技术难点。

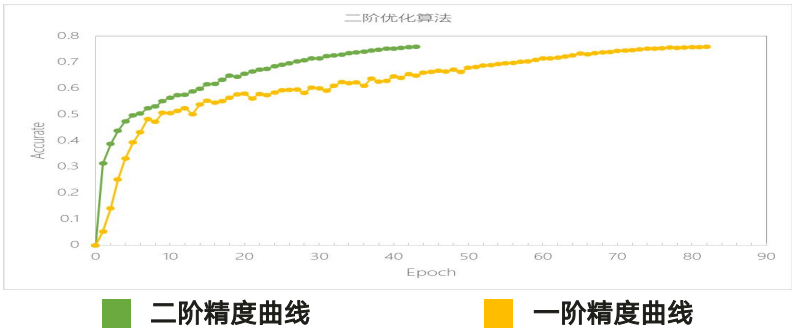


方案



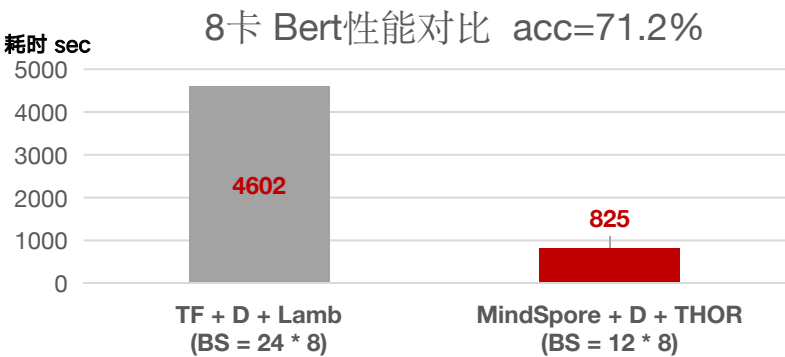
进展

CNN：基于ResNet-1.5@ImageNet2012 训练收敛提速**20.6%**

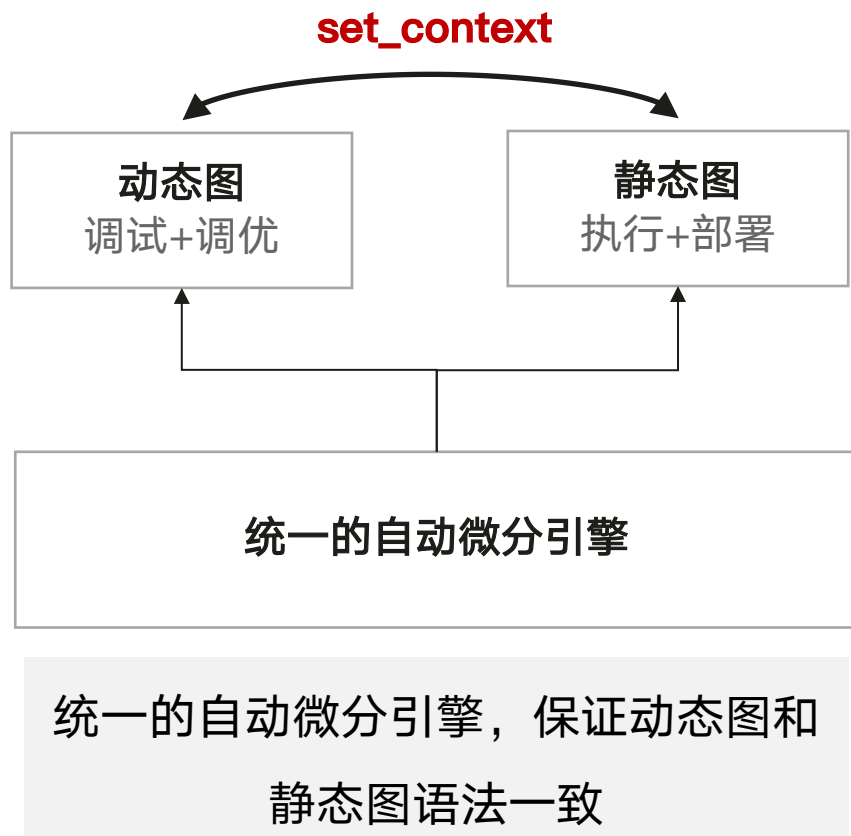


优化器	epoch	收敛时间	测试规模
SGD+MOMENTUM	约60	90min	8卡910
MindSpore二阶优化	42	71.5min	8卡910

- NLP：Bert网络二阶优化能力复制；支持异构硬件



关键特性3：动静态图结合，模式灵活切换，兼顾开发效率和执行性能



灵活切换：一行代码完成动静态图模式切换

```
#切换为动态图模式
context.set_context(mode=context.PYNATIVE_MODE)
#切换为静态图模式
context.set_context(mode=context.GRAPH_MODE)
```

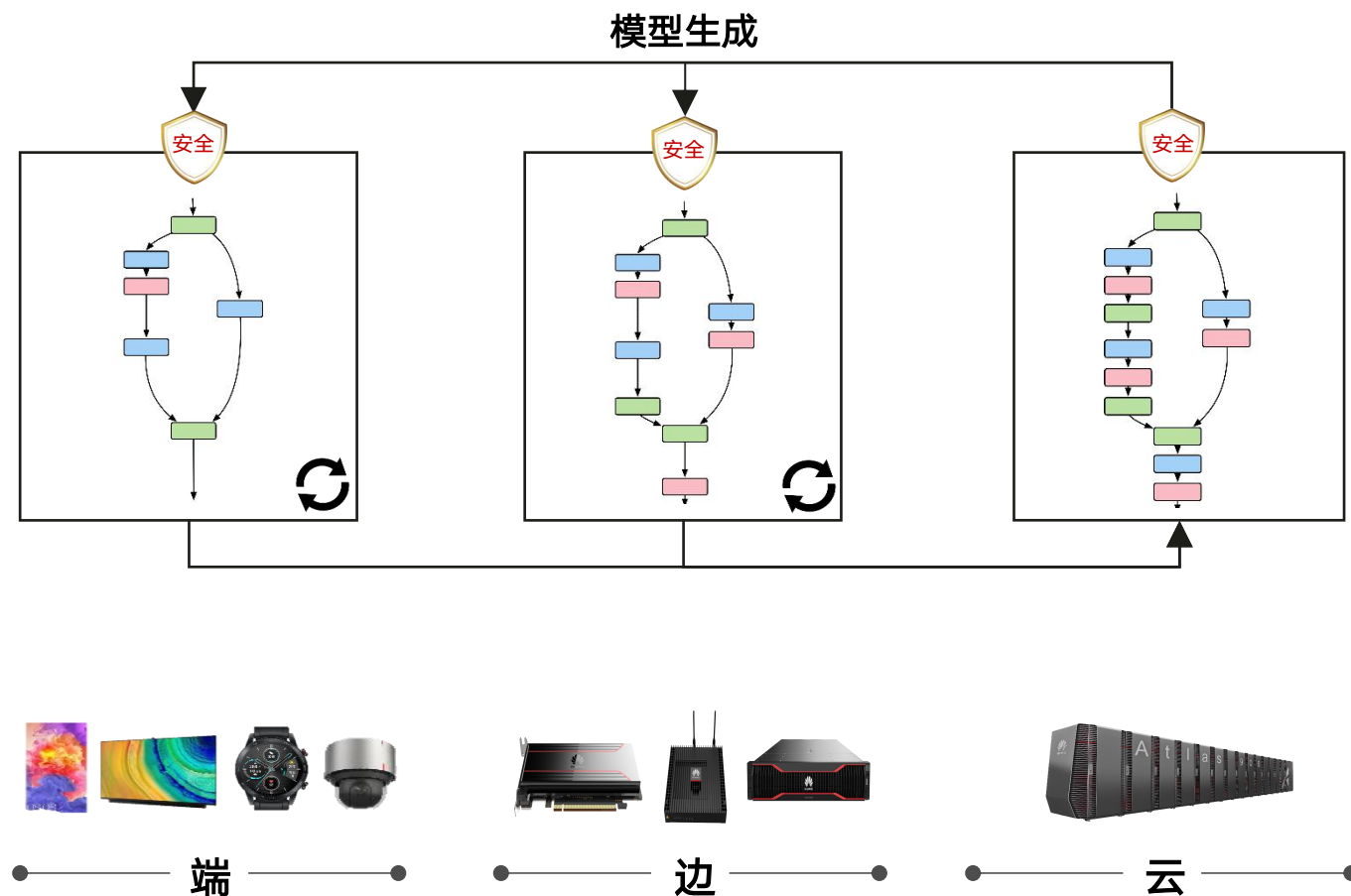
```
@ms_function
def sub_net(self, x):
    x = self.conv(x)
    return x
```

调试通过的代码
静态图模式执行

```
def construct(self, x):
    x = self.sub_net(x)
    x = self.relu(x)
    return x
```

待调试的代码
动态图模式执行

关键特性4：全场景协同，智能匹配端、边、云多样设备



MindIR统一模型

端、边、云共架构，支持统一的模型文件，
一次训练多处部署

端侧轻量学习

轻量化学习，提供个性化AI服务，智能
设备“越用越懂你”

超轻量框架

框架尺寸可变（KB~MB），部署方式
灵活，覆盖全场景设备

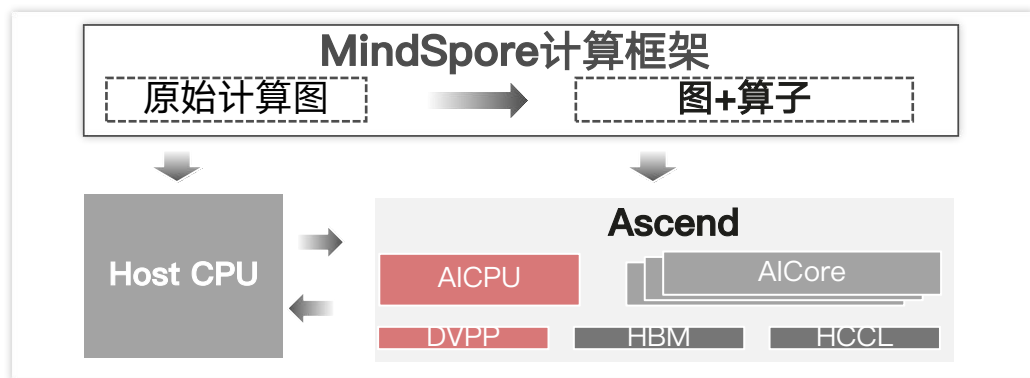
关键特性5：全栈协同加速，最亲和的昇腾执行引擎，发挥极致性能

框架优化

- **Pipeline并行**：数据处理和前向/反向计算并发，反向计算与梯度同步Allreduce并发
- **跨层内存复用**：复用不同Layer的内存，使能更大模型在卡上执行

软硬件协同优化

- **On-Device执行**：整图下沉到Device执行，减少host-device交互开销。
- **深度图优化**：包括整图的格式转换消除、类型转换消除、图算融合
- **多种执行模式**：PyNative执行、图下沉执行、图异步执行

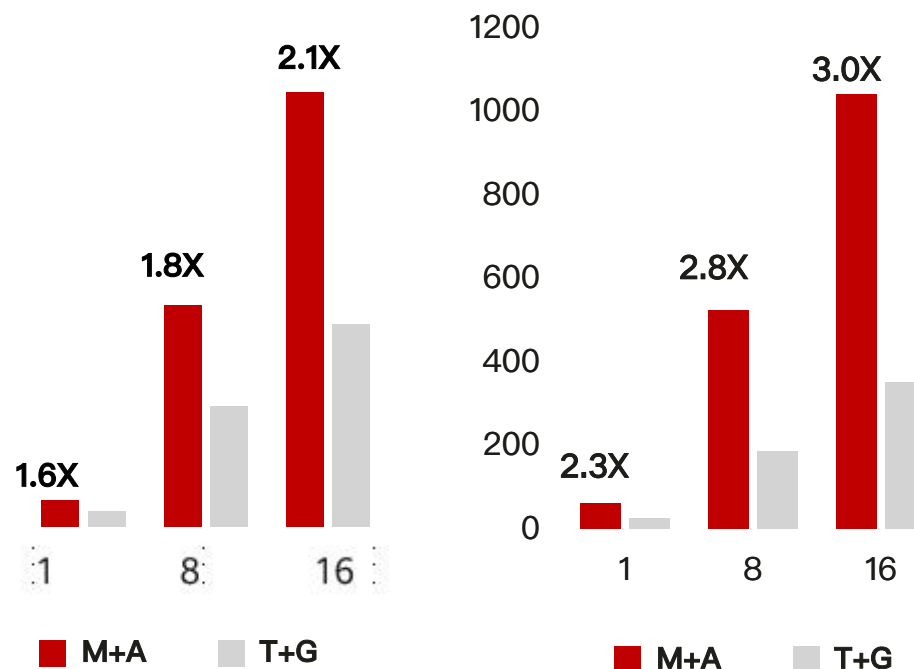


MindSpore+Ascend相比TF+GPU训练时间显著减少

少

ResNet50
ImageNet (images/s)

Bert-Large
Wiki (sentences/s)



关键特性6： 迁移及调试工具

技术生态迁移工具包



提升MindSpore技术生态迁移效率

```
class AlexNet(nn.Module):
    def __init__(self, num_classes=1000):
        super(AlexNet, self).__init__()
        self.features = nn.Sequential(
            nn.Conv2d(3, 64, kernel_size=11,
                      stride=4, padding=2),
            nn.ReLU(inplace=True),
            nn.MaxPool2d(kernel_size=3, stride=2),
            ...
        )
        self.avgpool = nn.AdaptiveAvgPool2d((6, 6))
    def forward(self, x):
        x = self.features(x)
        x = self.avgpool(x)
        x = torch.flatten(x, 1)
        x = self.classifier(x)
        return x

class AlexNet(nn.Module):
    def __init__(self, num_classes=1000):
        super(AlexNet, self).__init__()
        self.features = nn.SequentialCell([
            nn.Conv2d(in_channels=3, out_channels=64,
                      kernel_size=11, stride=4,
                      pad_mode='pad', padding=2,
                      has_bias=True),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=3, stride=2,
                          pad_mode='valid'),
            ...
        ])
        self.avgpool = nn.AdaptiveAvgPool2d((6, 6))
    def construct(self, x):
        x = self.features(x)
        x = self.avgpool(x)
        x = P.Flatten()(x)
        x = self.classifier(x)
        return x
```

自动转换

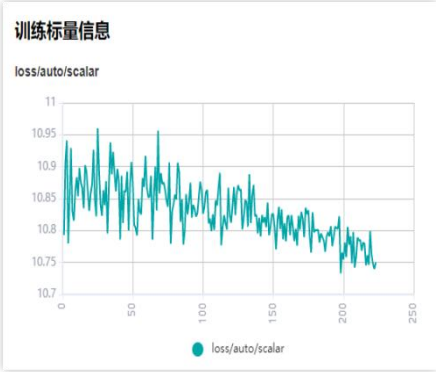
需人工转换

自动转换

精度调优工具包



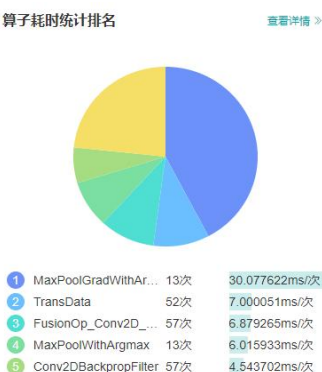
精度快速达标，加速模型上市



性能调优工具包



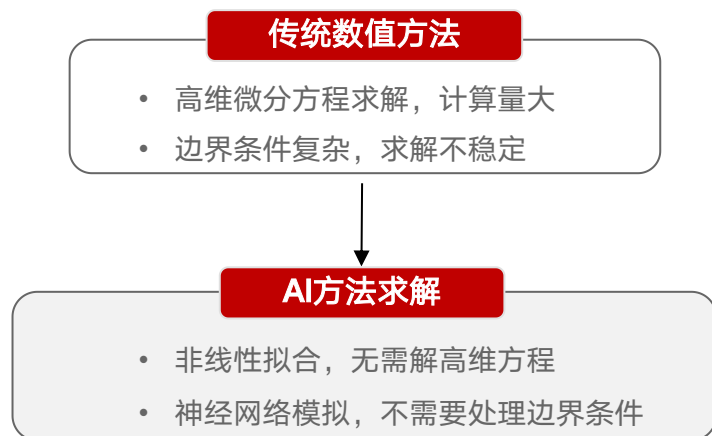
助力MindSpore框架性能算力发挥



关键特性7： AI+科学计算，突破AI求解技术，打造科学计算领域绝对优势

科学计算近况：

- 科学计算核心问题是微分方程求解，算力消耗巨大，大规模求解器垄断历年戈登贝尔奖，近年来结合AI方法成为趋势。



业界AI+科学计算现状：

- TF**：众筹方式构建AI求解模型，典型应用领域取得突破；面向DNN设计的自动微分，计算高阶微分时效率低下；
- Nvidia**：支持高精度计算；构筑cuBLAS、cuFFT等基础数学库；上层框架依赖TensorFlow，继承TF缺点

MindSpore关键优势：

- 大规模高维微分方程AI求解器**：AI建模加速材料、气象等领域；AI求解加速麦克斯韦电磁方程、Burgers等，性能提升**10倍**；AI框架加速支持海洋GOMO模型，性能提升**1.5倍**。
- 自动微分**：前向自动微分、混合自动微分、向量化自动微等，性能提升**10倍**。
- 通用（稀疏）张量代数计算加速**：支持基础BLAS运算(矩阵求逆等)、FFT运算、稀疏矩阵运算等，性能提升**2-3倍**。



关键特性8：安全可信，首个企业级AI可信计算框架（Trusted AI）

全场景安全部署

模型 “偷”不走

- 模型混淆，原创轻量级模型保护
- 密态AI，数据和模型可用不可见

数据安全和隐私保护

隐私 不泄露

- 模型隐私泄露评估，差分隐私训练
- 敏感数据识别和高效数据脱敏
- 安全聚合，端云联合学习隐私保护

AI责任与算法安全

模型 抗攻击

- 模型安全性评估和模糊测试
- 对抗样本识别和安全性增强
- 后门/投毒检测和模型供应链安全

系统级安全

OS级 安全保护

- 高可信执行环境，实现AI模型强保护
- 系统资源强隔离，算子安全执行
- 全虚拟地址运行，组件故障隔离

和德国BSI，法国ANSSI等机构合作开展可信AI标准建设和认证

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

**Copyright©2020 Huawei Technologies Co., Ltd.
All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

