推理引擎- Kernel 优化

基本介绍

ZOMI

BUILDING A BETTER CONNECTED WORLD

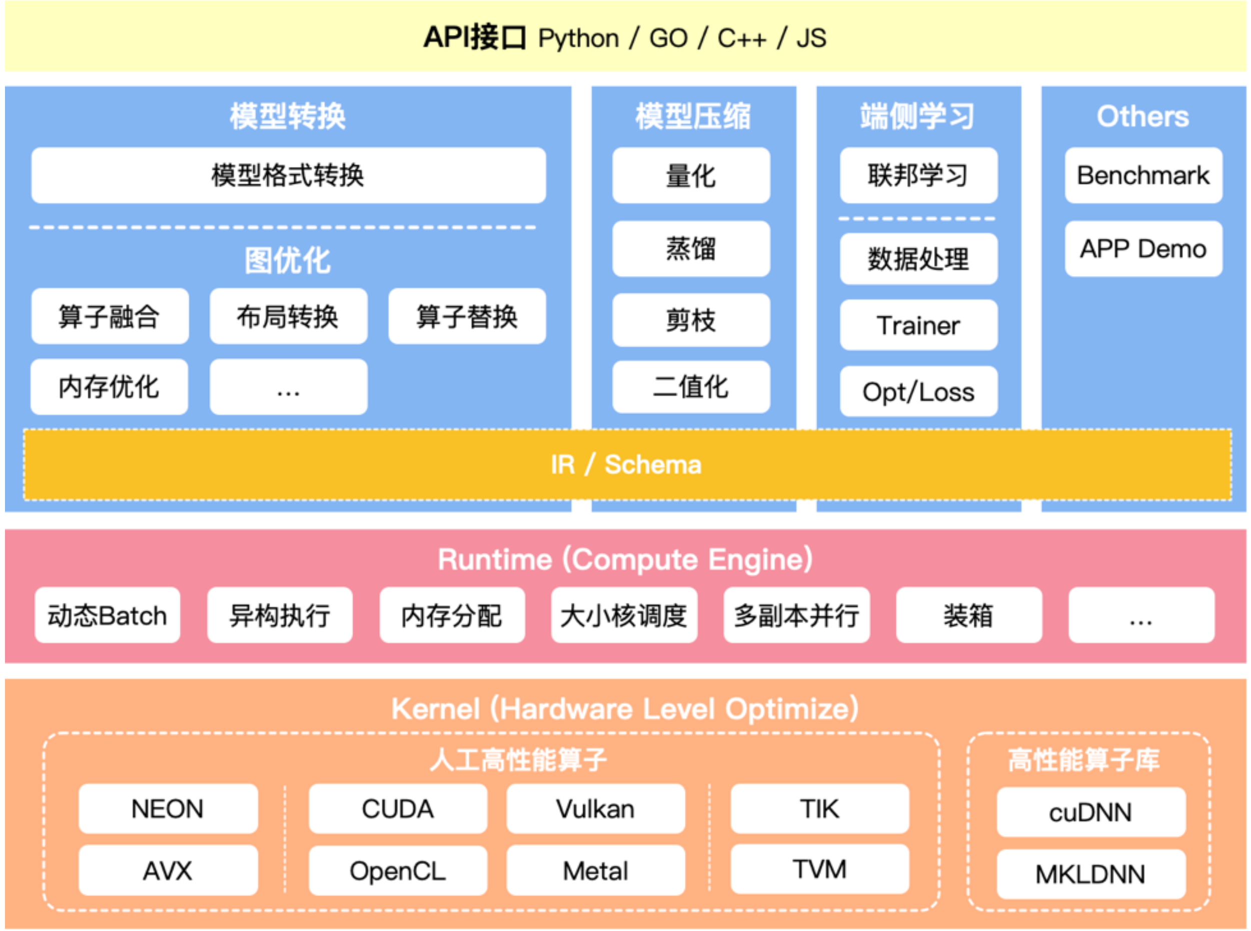
Ascend & MindSpore

www.hiascend.com
www.mindspore.cn

Talk Overview

1. **推理系统介绍**：推理系统架构 - 推理引擎架构
2. **模型小型化**：CNN小型化结构 - Transform小型化结构
3. **离线优化压缩**：低比特量化 - 模型剪枝 - 知识蒸馏
4. **模型转换与优化**：模型转换 - 计算图优化
5. **Kernel 优化**
   - 算法优化 (Winograd / Strassen)
   - 内存布局 (NC1HWC0 / NCHW4)
   - 汇编优化 (指令与汇编)
   - 调度优化
6. **Runtime 优化**

推理引擎架构



API接口 Python / GO / C++ / JS

模型转换
- 模型格式转换
- 图优化
  - 算子融合
  - 布局转换
  - 算子替换
  - 内存优化
  - …

模型压缩
- 量化
- 蒸馏
- 剪枝
- 二值化

端侧学习
- 联邦学习
- 数据处理
- Trainer
- Opt/Loss

Others
- Benchmark
- APP Demo

IR / Schema

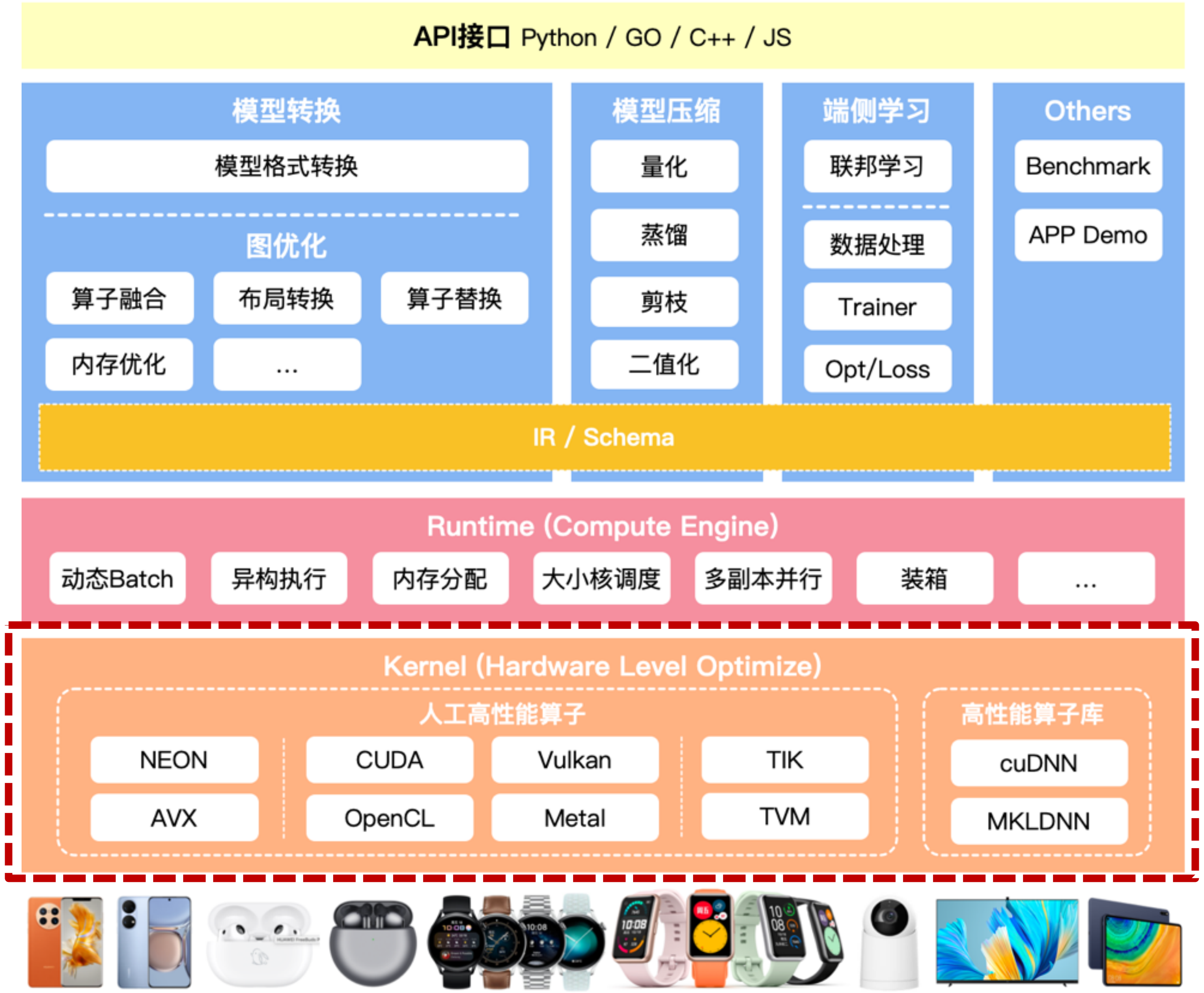
Runtime (Compute Engine)
- 动态Batch
- 异构执行
- 内存分配
- 大小核调度
- 多副本并行
- 装箱
- …

Kernel (Hardware Level Optimize)
- 人工高性能算子
  - NEON
  - AVX
  - CUDA
  - OpenCL
  - Vulkan
  - Metal
  - TIK
  - TVM
- 高性能算子库
  - cuDNN
  - MKLDNN

推理引擎架构

高性能算子层

- 算子优化
- 算子执行
- 算子调度



API接口 Python / GO / C++ / JS

模型转换

模型格式转换

图优化

算子融合 布局转换 算子替换

内存优化 …

模型压缩

量化

蒸馏

剪枝

二值化

端侧学习

联邦学习

数据处理

Trainer

Opt/Loss

Others

Benchmark

APP Demo

IR / Schema

Runtime (Compute Engine)

动态Batch 异构执行 内存分配 大小核调度 多副本并行 装箱 …

Kernel (Hardware Level Optimize)

人工高性能算子

NEON CUDA Vulkan TIK

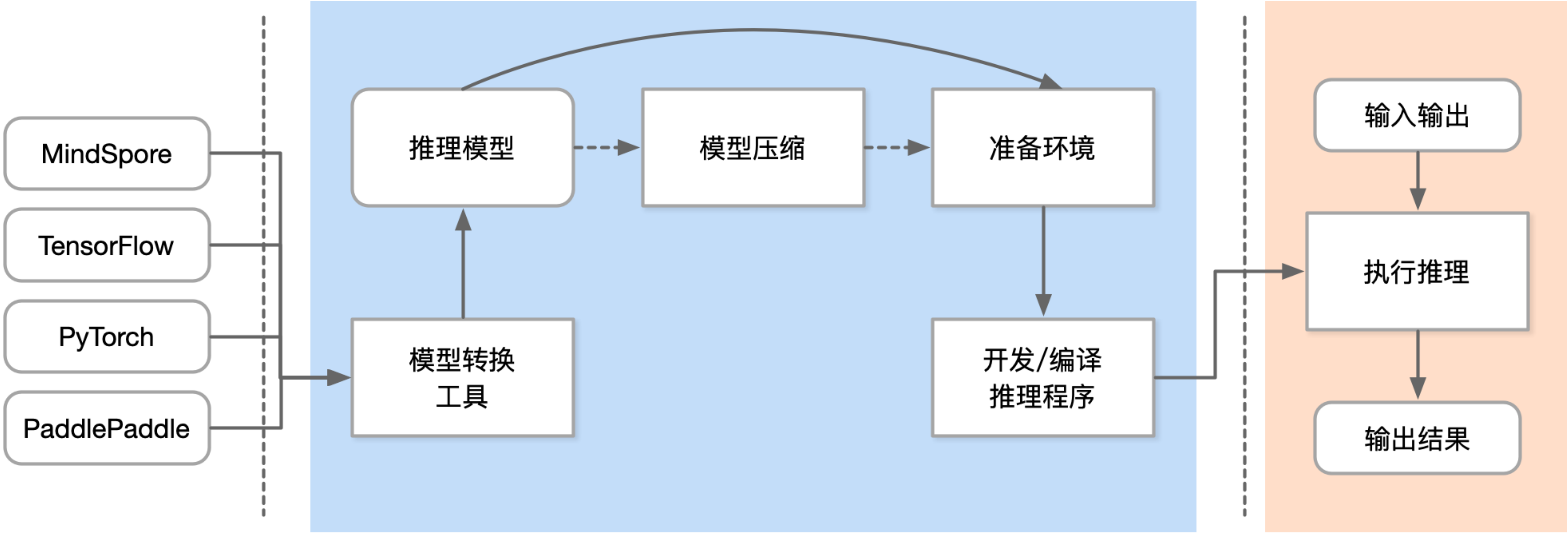
AVX OpenCL Metal TVM

高性能算子库

cuDNN

MKLDNN

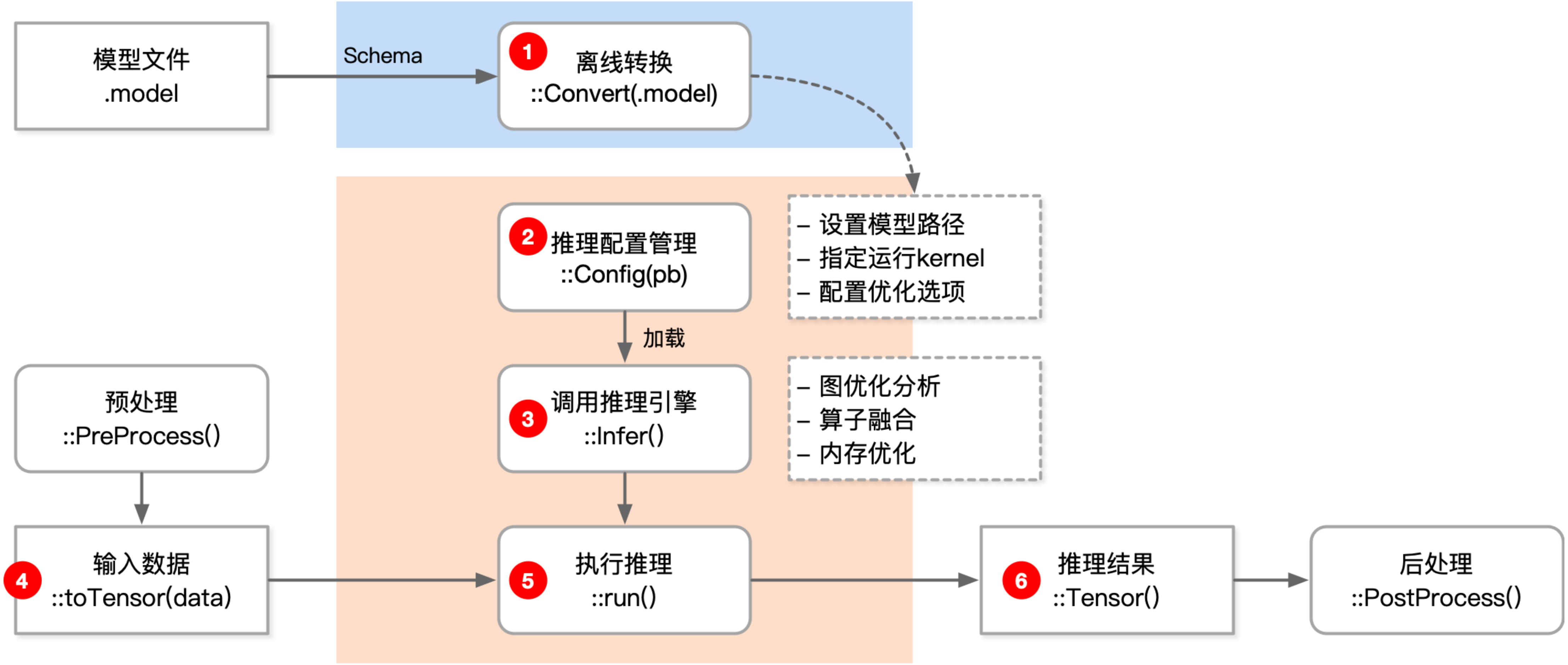
推理流程



离线模块

在线执行

开发推理程序



模型文件
.model

Schema

1 离线转换
::Convert(.model)

- 设置模型路径
- 指定运行kernel
- 配置优化选项

2 推理配置管理
::Config(pb)

加载

3 调用推理引擎
::Infer()

- 图优化分析
- 算子融合
- 内存优化

预处理
::PreProcess()

4 输入数据
::toTensor(data)

5 执行推理
::run()

6 推理结果
::Tensor()

后处理
::PostProcess()

Talk Overview

1. **推理系统介绍**：推理系统架构 - 推理引擎架构
2. **模型小型化**：CNN小型化结构 - Transform小型化结构
3. **离线优化压缩**：低比特量化 - 模型剪枝 - 知识蒸馏
4. **模型转换与优化**：模型转换 - 计算图优化
5. **Kernel 优化**
    - 算法优化 (Winograd / Strassen)
    - 内存布局 (NC1HWC0 / NCHW4)
    - 汇编优化 (指令与汇编)
    - 调度优化
6. **Runtime 优化**

BUILDING A BETTER CONNECTED WORLD

THANK YOU

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.