

# 推理引擎-模型压缩

# 模型压缩介绍



# ZOMI

# Talk Overview

## 1. 推理系统介绍

- 推理系统与推理引擎区别
- 推理工作流程
- 推理系统介绍
- 推理引擎介绍

## 2. 模型小型化

- 基础参数概念
- CNN小型化结构
- Transform小型化结构

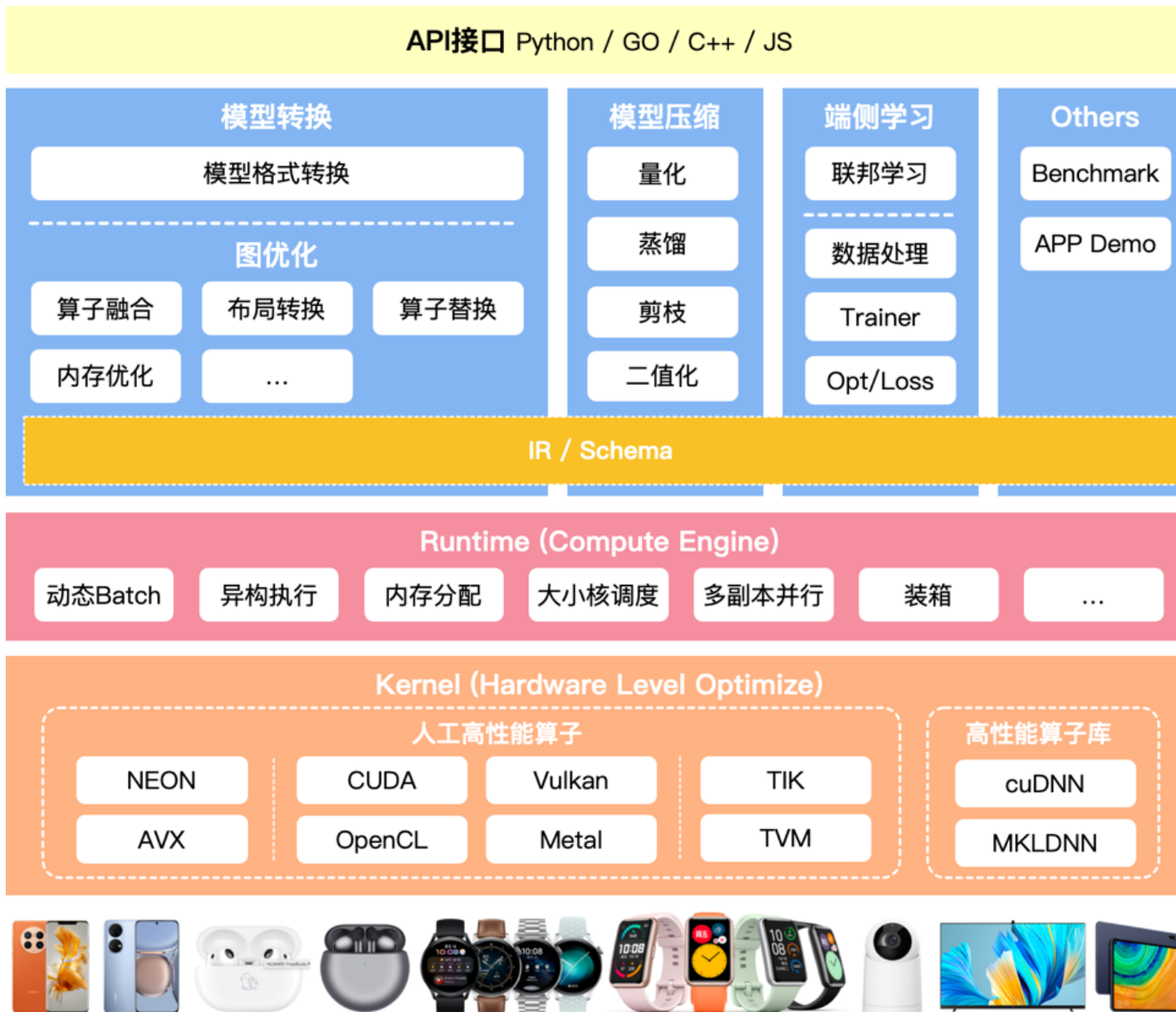
## 3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型剪枝
- 知识蒸馏

## 4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

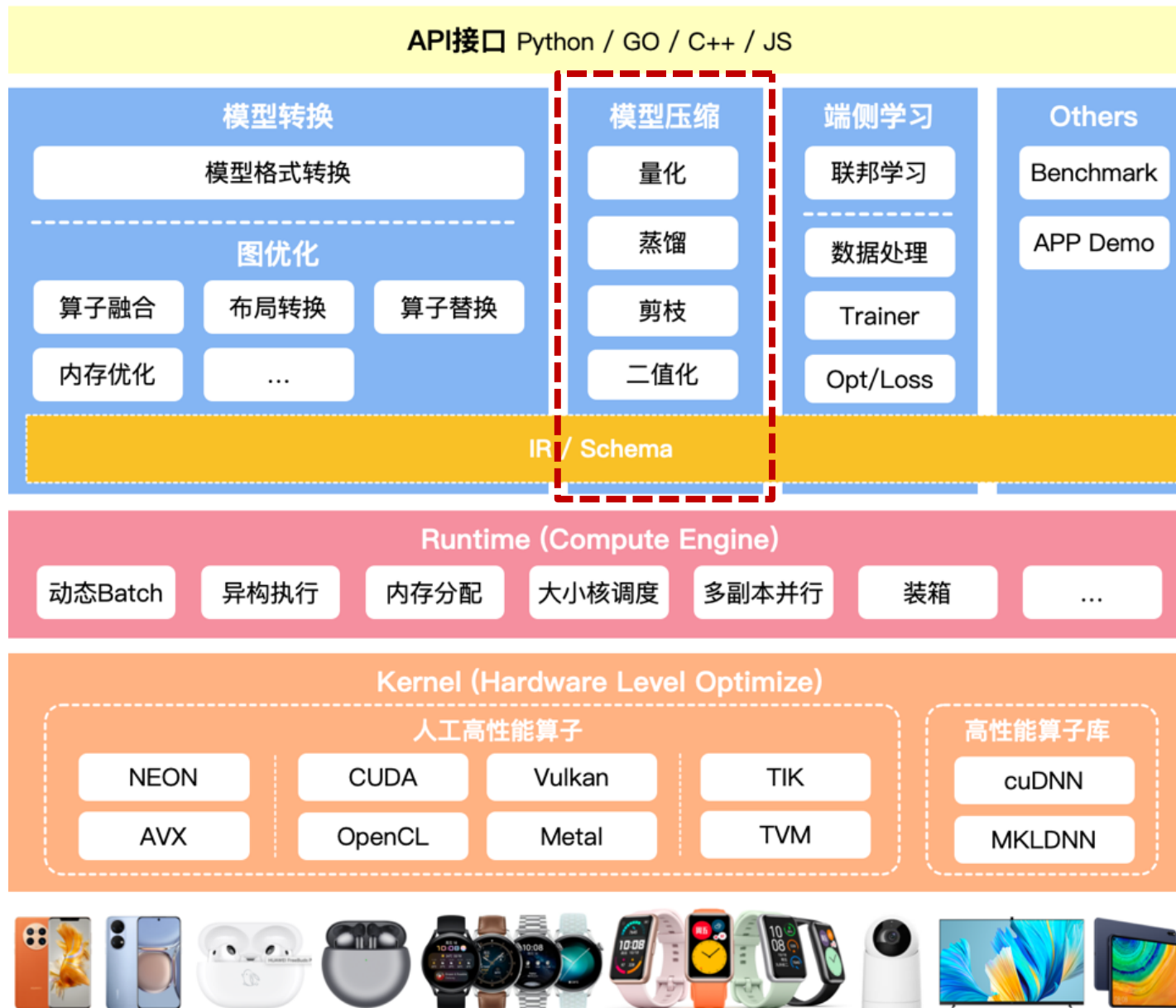
# 推理引擎架构



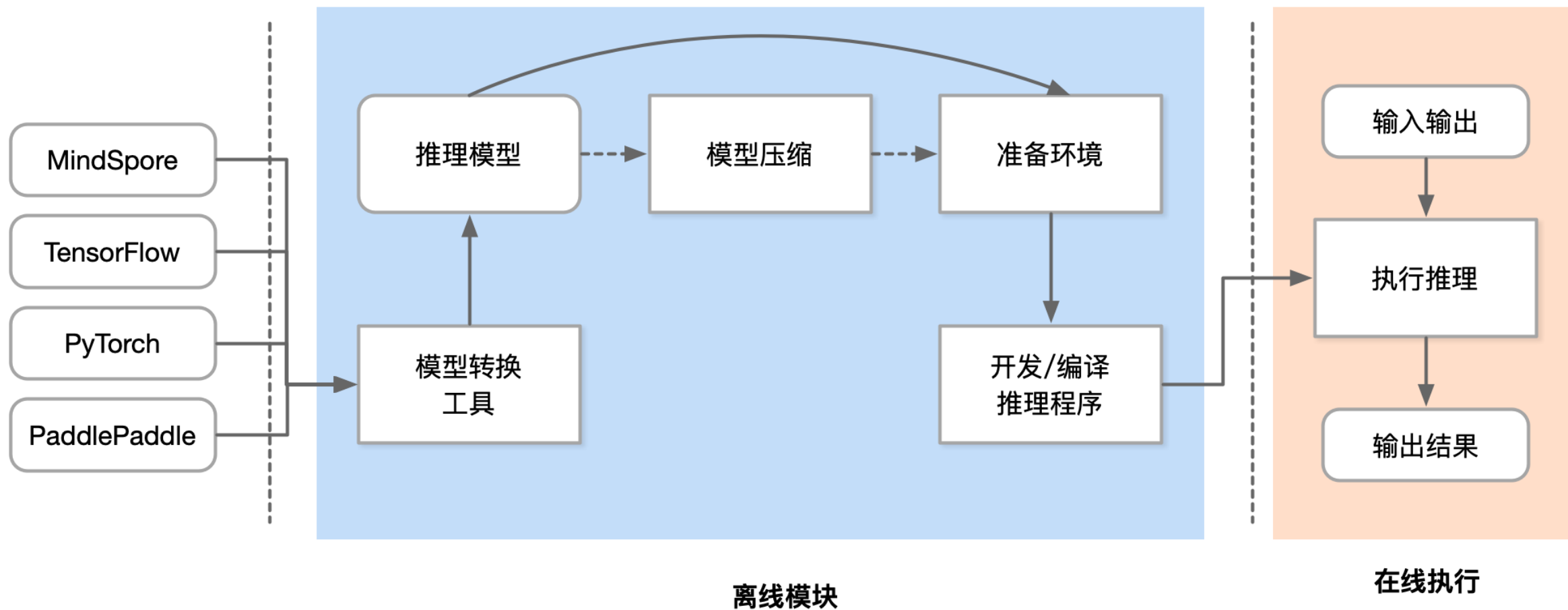
# 推理引擎架构

对模型进行压缩

- 减少模型大小
- 加快推理速度
- 保持相同精度



# 推理流程





BUILDING A BETTER CONNECTED WORLD

THANK YOU

**Copyright©2014 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.