GOTC 2023

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

OPEN SOURCE，INTO THE FUTURE

「AI is everywhere」专场

本期议题：模型推理优化，探索AI落地潜能

嘉宾姓名 2022年00月00日

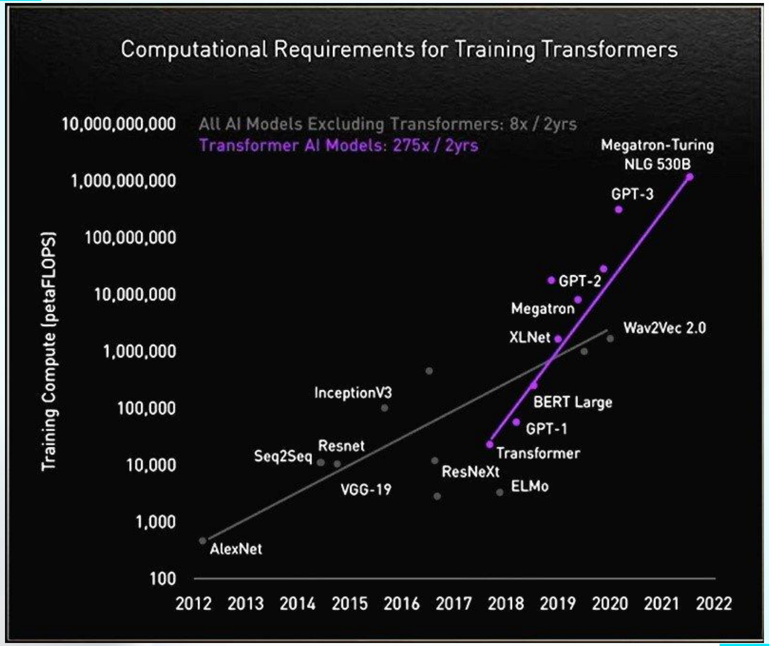
01

趋势与挑战

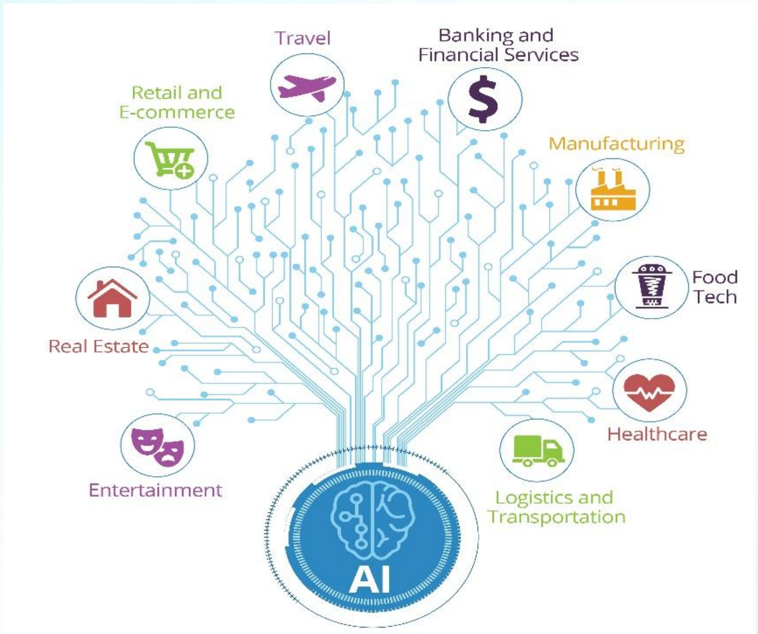
AI总体趋势

GOTC

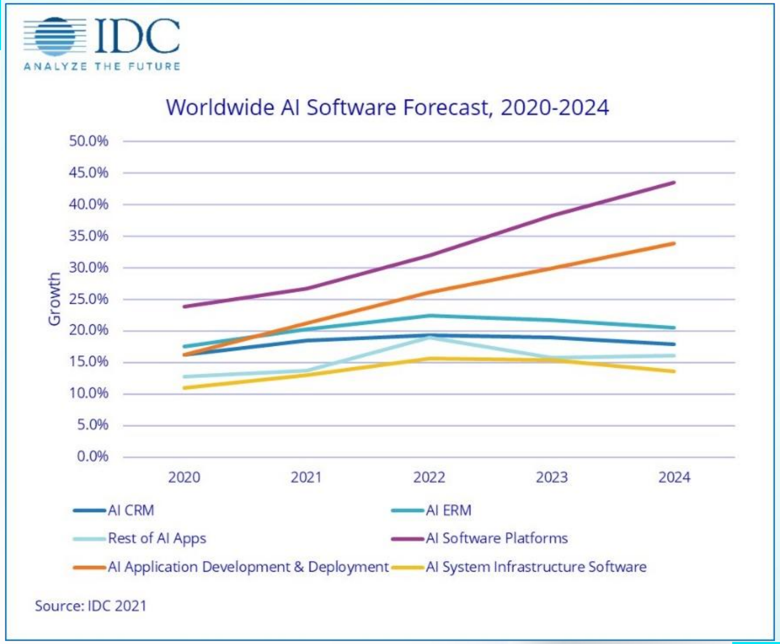
大模型研究如火如荼



AI和产业的融合日渐加深



大规模部署已成趋势



为了将AI前沿的理论研究在实际中发挥价值，AI的产业化势在必行

AI产业化挑战

数据工程 + 模型工程 + 代码工程

- 各工程间的协作困难
- 人工操作繁琐、易错
- 上线时间漫长
- 模型更新、维护困难

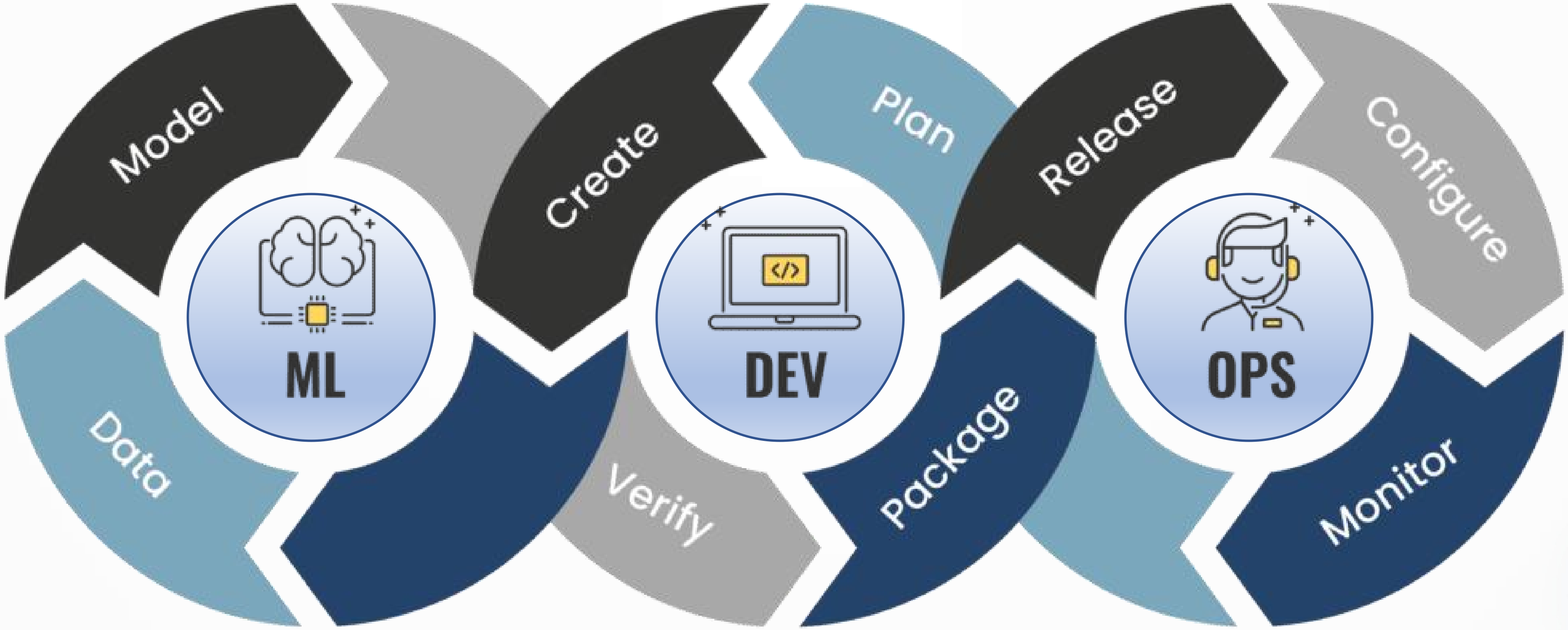
→ MLOps

02

MLOps概览

MLOps：Machine Learning Operations

MLOps=DevOps for ML



模型构建、训练、
调优

ML工作流创建，模型
持续集成、持续部署

模型大规模持续交付、
持续监控

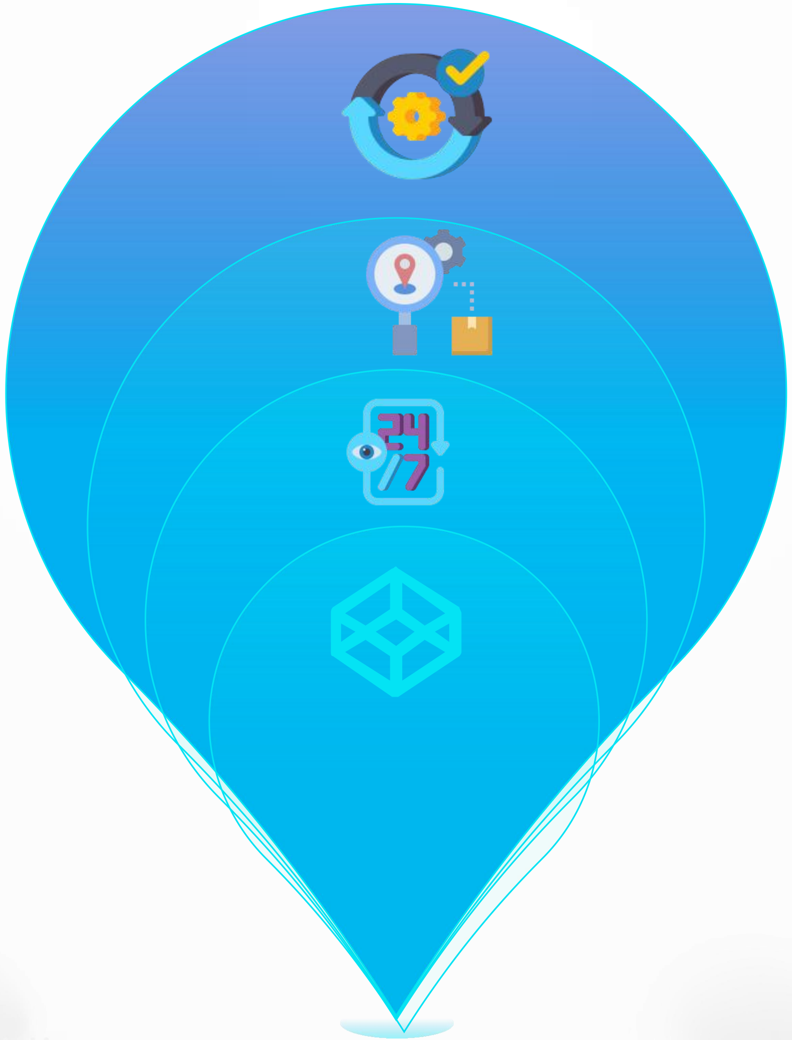
MLOps原则

自动化 & DevOps

- 自动化ML的pipeline
- Continuous CI，CD，CM，CT

实验跟踪

- 数据、模型、代码的版本控制
- 代码构建的相关信息的追踪



监控

- 传统系统监控：日志、系统资源
- ML监控：输入输出、模型性能

可靠性

- 可复现性
- 预测结果的正确性
- 系统的可靠性

MLOps能力

Dataset & feature repository

核心MLOps技术能力

Experimentation	Data processing	Model training
Model evaluation	Model serving	Online experimentation
Model monitoring	ML pipeline	Model registry

ML metadata & artifact repository

Source and artifact repositories & CI/CD

Security & privacy

Infrastructure

03

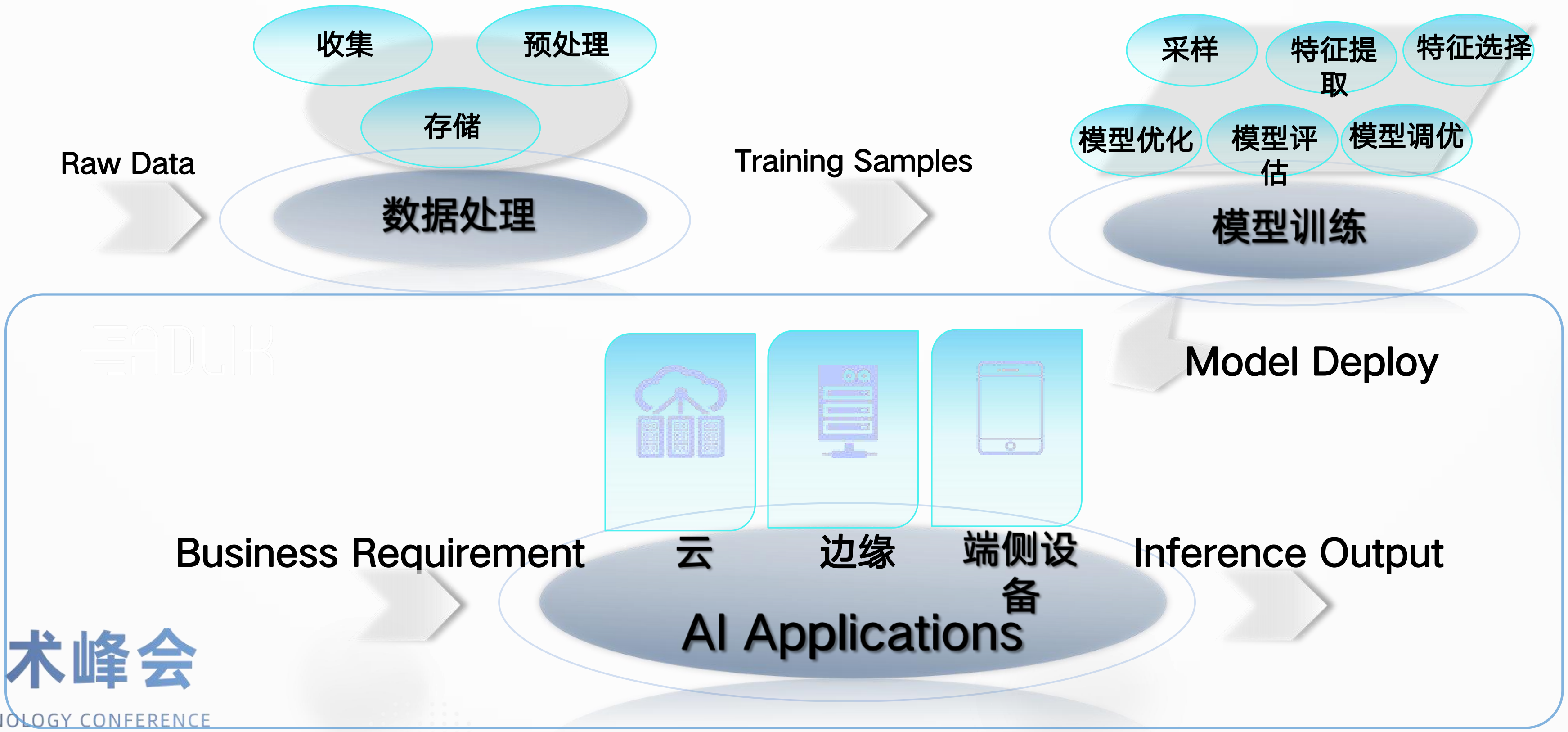
Adlik推理优化实践

模型部署对推理工具链提出的需求

- 广泛支持主流训练框架
- 高效的模型压缩，优化能力
- 支持多种AI硬件
- 具备完善的工业应用特性，高可用、易运维
- 高性能，可以发挥硬件最佳性能
- 轻量化，可以满足边缘、终端各类部署环境
- 易用性，快速上手解决问题

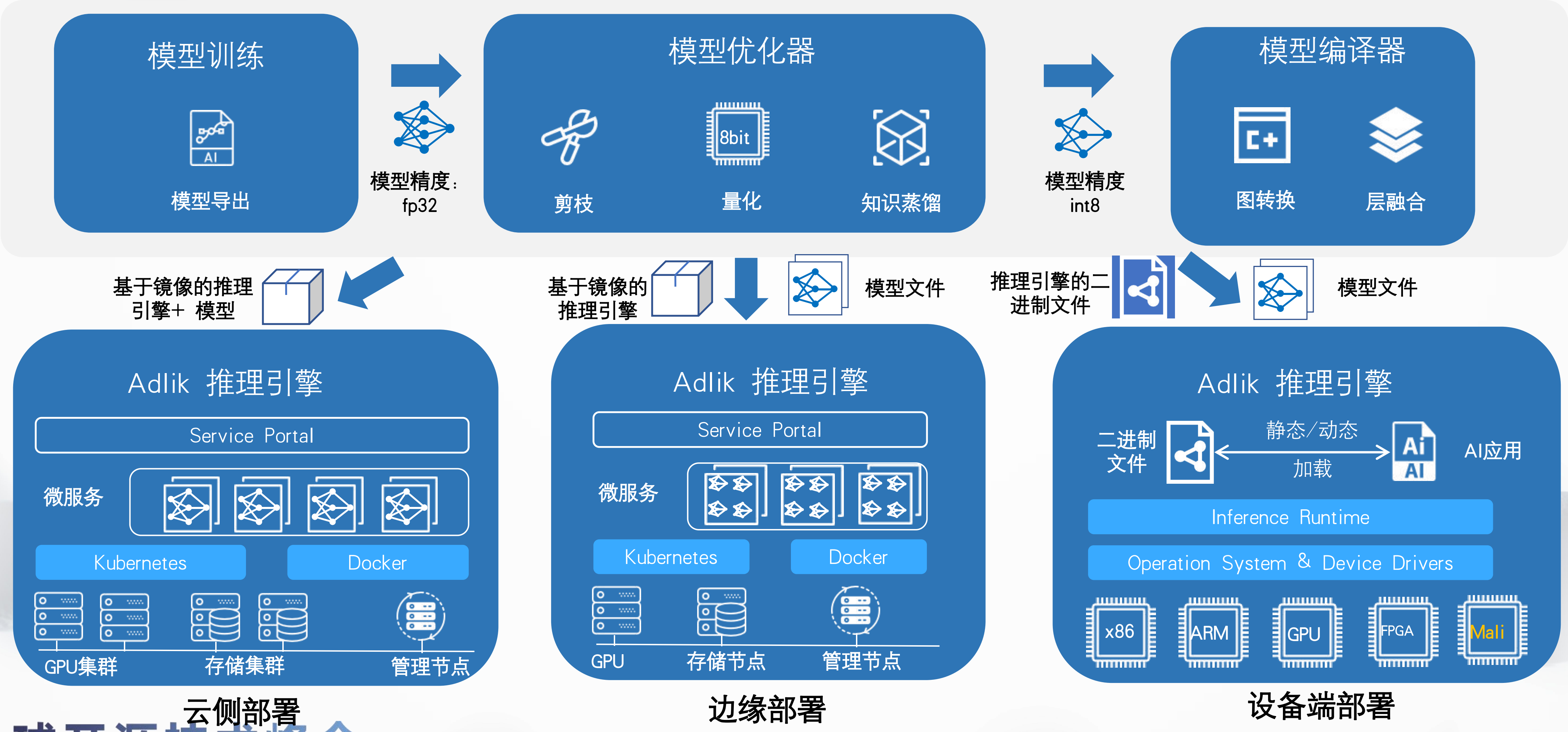
Adlik深度学习推理工具链

- 一种将深度学习模型从训练完成，到部署到特定硬件并提供应用服务的端到端工具链，实现模型从研发状态到生产应用环境的高效切换
- 与多种推理引擎协作，提供灵活的模型加速、部署、推理方案，助力用户构建高性能AI应用
- 在Linux基金会AI和数据基金会（LF AI & Data）开源



Adlik架构

模型优化器&模型编译器：提升模型计算效率，减少能耗，降低推理时延



Adlik Engine：支持云边端等多种环境的模型部署

Adlik：解决模型部署过程中挑战性问题

GOTC

更省

- 工程化参数自适应优化，降低部署复杂度，更省部署人力
- 简单方便的模型部署pipeline，缩短模型上线周期，更省部署时间
- 统一的模型推理和管理接口，更省模型迁移成本

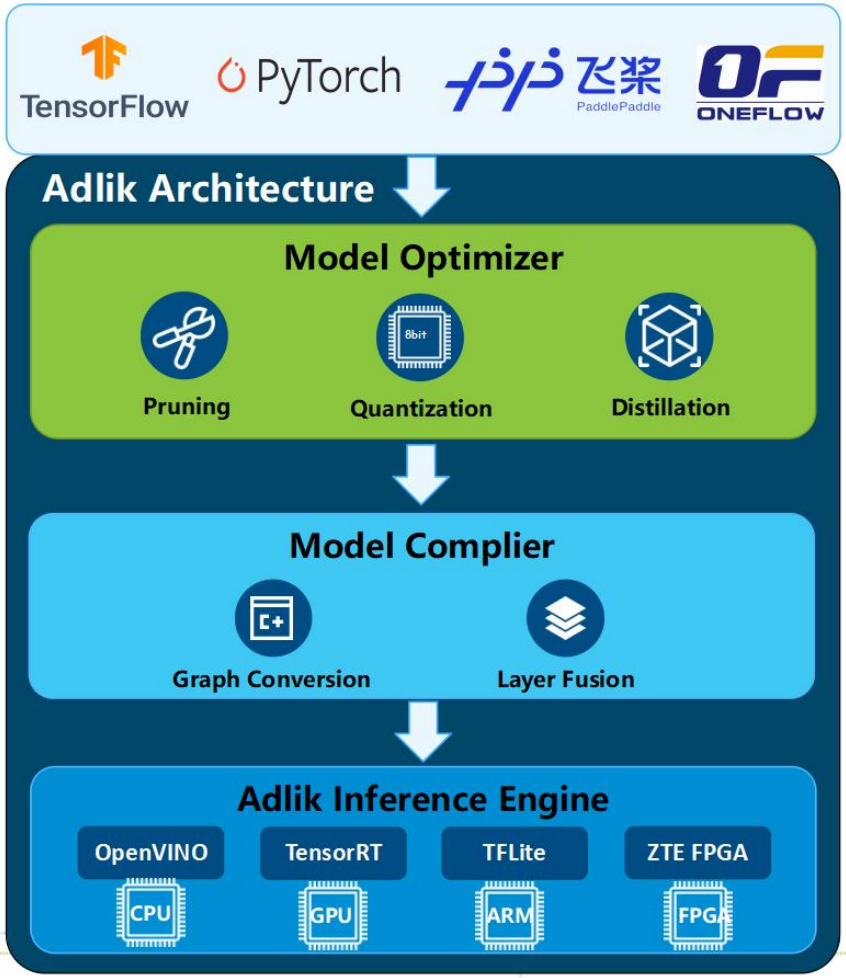
更快

- 内置多种高性能运行时，以供用户更快地按需选用
- 提供可拓展性强的Serving SDK，可更快集成自定义推理运行时
- 提供灵活易用的推理API，更快实现AI应用的构建、迭代

更优

- 多种模型压缩、优化算法在实践中表现出出色性能
- 面对异构的部署硬件、不同的应用场景，提供更优的端到端方案，保障更优的推理性能
- 更优的模型管理、完善的模型运行监控

Adlik模型优化器



- 支持多种结构化剪枝方法，并可**通过多节点，多GPU进行剪枝**；支持**自动剪枝**，根据网络类型（如ResNet-50 等）和限制条件（如FLOPs，Latency），就能自动决定模型每一层的通道数，得到在限制条件下最优的模型结构
- 支持多种蒸馏方法，使用剪枝和多教师蒸馏**组合优化**，进一步压缩模型，同时提升模型精度，完成图像分类、目标检查等任务精度提升
- 利用少量校准数据快速实现8-bit PTQ量化；支持QAT量化算法，提升量化模型精度
- 支持In4量化

Adlik模型优化器组合优化效果

ResNet50

- 使用自动剪枝工具剪枝，剪枝后吞吐量提升2.74倍
- 经过蒸馏，精度提升4.2%
- 使用INT8 PTQ量化，吞吐量再提升2.96倍，端到端加速13.82倍

模型优化方法	Throughput (OpenVINO)	Accuracy
Baseline	432	76.80%
自动剪枝	1615	73.30%
自动剪枝+蒸馏	1615	77.50%
自动剪枝+蒸馏+INT8量化	6401	77.00%

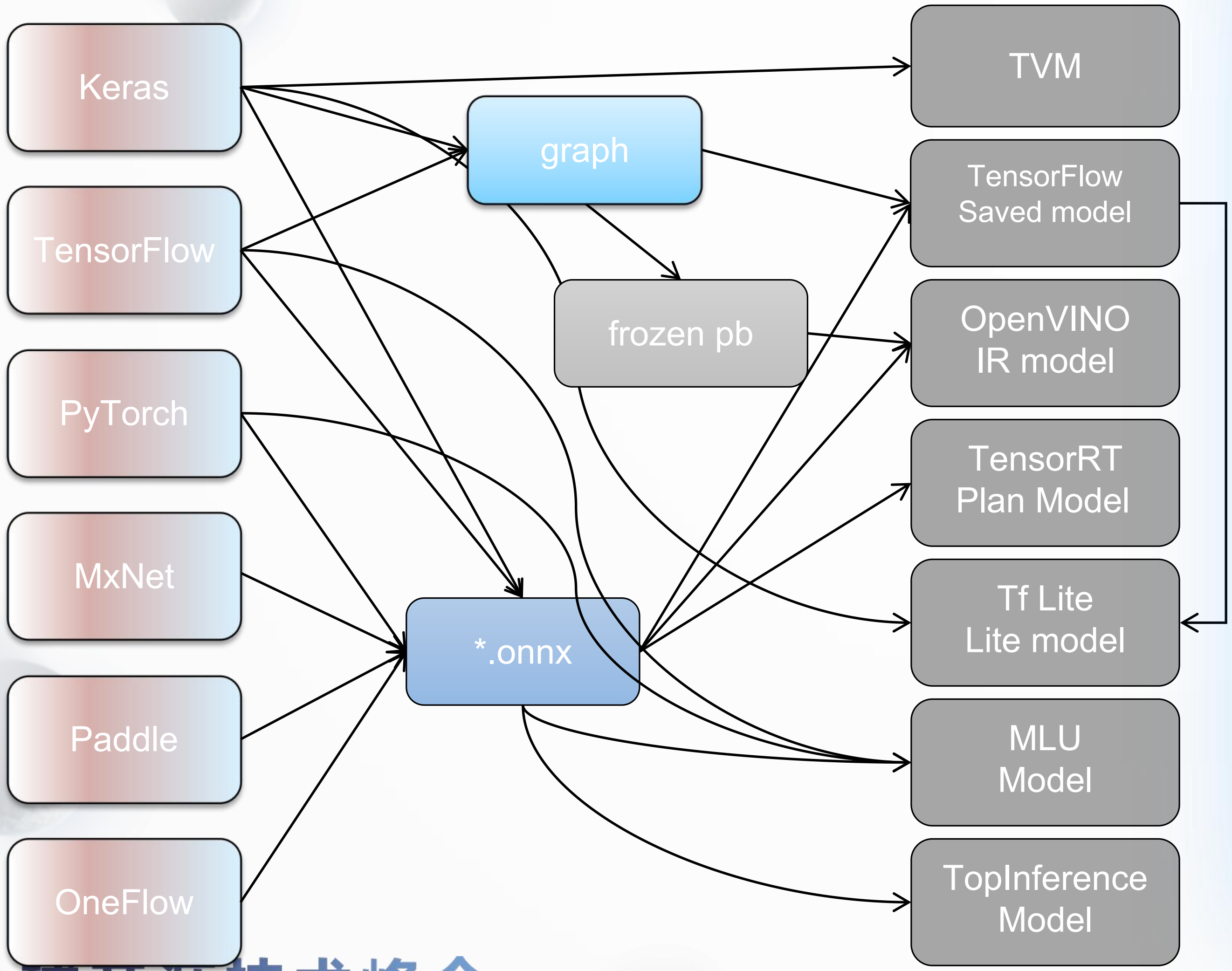
YOLOv5m

- 通过自动剪枝，提升25%吞吐量
- 经过蒸馏，精度提升2.8%
- 通过INT8 PTQ量化，吞吐量再提升3.1倍，端到端加速3.33倍

模型优化方法	Throughput (OpenVINO)	Accuracy
Baseline	73	44.70%
自动剪枝	95	41.80%
自动剪枝+蒸馏	95	44.60%
自动剪枝+蒸馏+INT8量化	316	43.90%

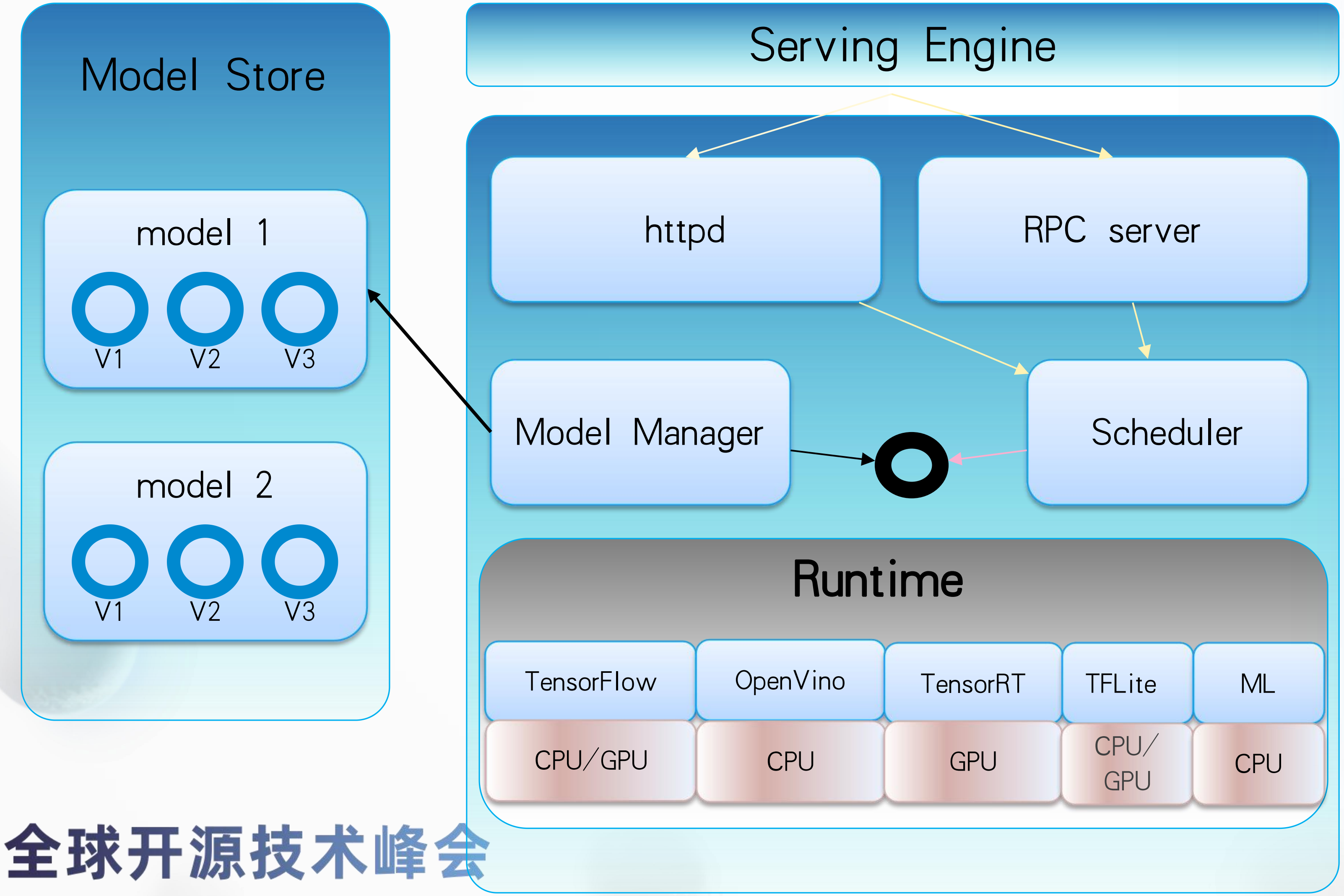
Test environment： Intel® Xeon® Platinum 8378C CPU @ 2.80GHz * 2

Adlik模型编译器



- 支持使用统一接口方案将多种原始训练模型格式转换到目标运行时模型格式
- 支持构建DAG 完成端到端的不同模型表达格式的转换
- 支持TfLite, TensorRT, OpenVINO模型量化
- 支持Paddle，OneFlow等国产训练框架，寒武纪，燧原等国产AI芯片

Adlik推理引擎



- 支持模型上载，升级，版本管理，监控等功能
- 使用统一推理接口完成各类运行时的推理
- 多模型，多模型实例和多运行时多统一调度管理
- 支持用户自定义运行时
- 支持机器学习算法

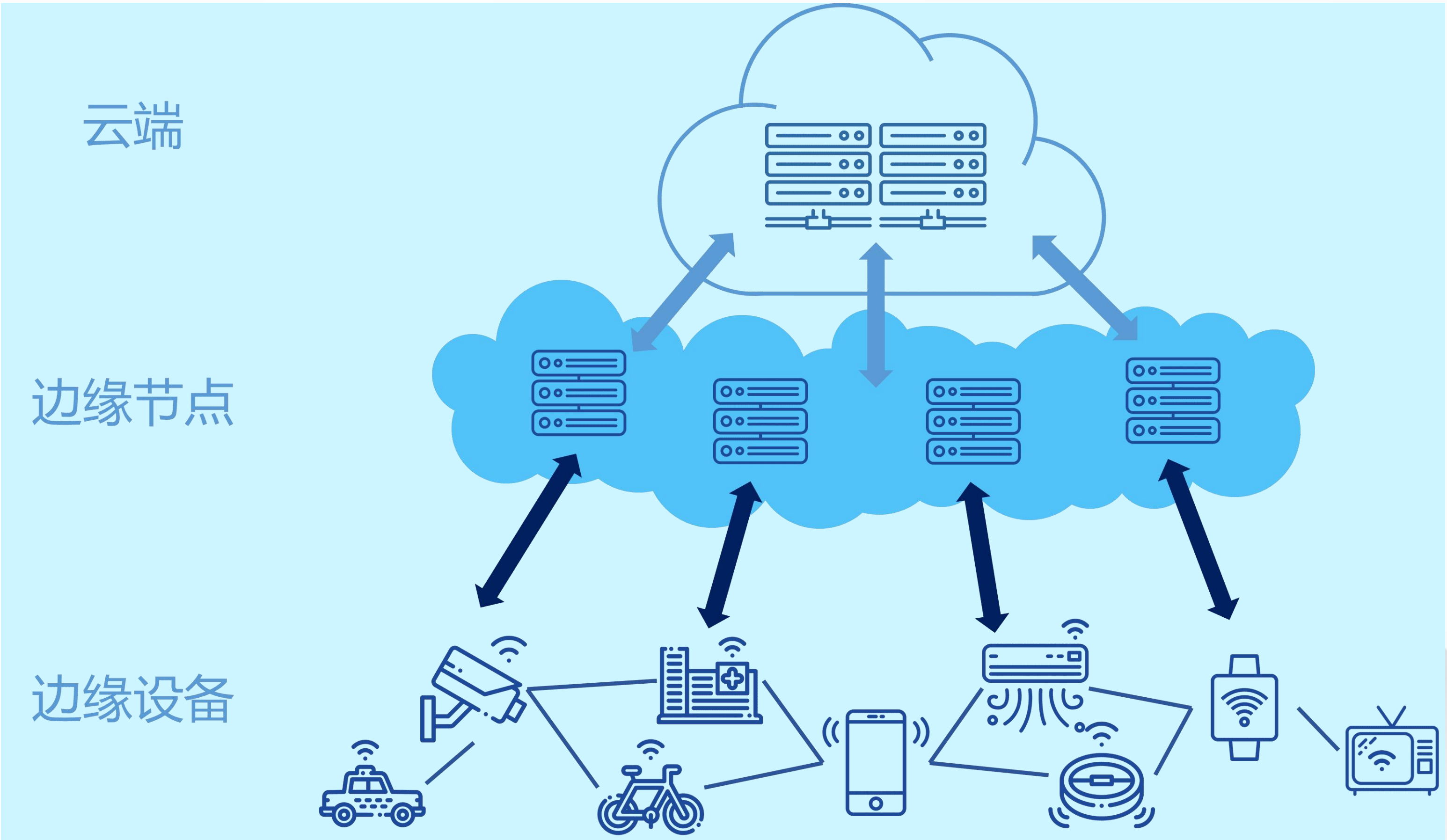
Adlik使用场景

CV模型

LSTM模型

NLP模型

深度推荐模型



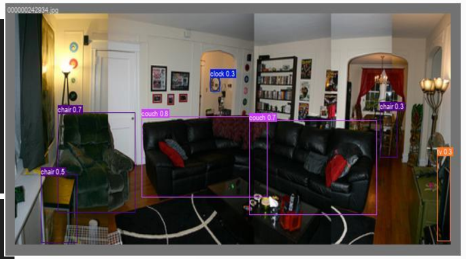
Adlik快速上手

1. 从阿里云获取Adlik镜像

```
docker pull docker_image_name:tag
```

2. 运行模型编译器镜像，根据自定义的配置文件实现模型的编译

```
docker run -it --rm  -v source_model:/mnt/model
registry.cn-beijing.aliyuncs.com/adlik/model-compiler:v0.5.0_trt7.2.1.6_cuda11.0 bash
```



```
python3 "-c" "import json; import model_compiler as compiler;
file=open('/mnt/model/serving_model.json','r');
request = json.load(file); compiler.compile_model(request); file.close()"
```

3. 运行推理引擎，开始模型推理

```
docker run -it --rm  -p 8500:8500 -v compiled_model:/model
registry.cn-beijing.aliyuncs.com/adlik/seving_openvino:v0.5.0 bash
```

```
adlik-serving --grpc_port=8500 --http_port=8501 --model_base_path=/model
```

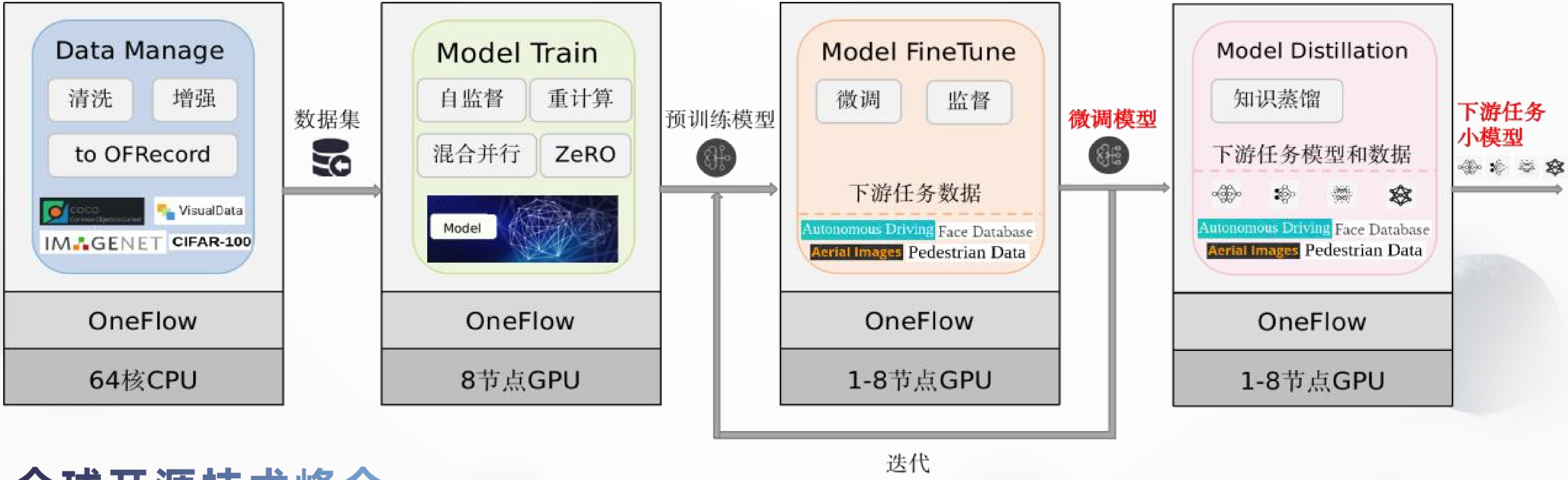
Adlik Next：大模型优化服务

收集大规模无标签图像，并进行清洗和增强

使用并行训练技术，通过自监督训练方式进行模型训练

使用下游任务的带标签数据集对大模型进行监督训练

将大模型作为教师网络，下游任务模型作为学生，进行蒸馏



THANKS

欢迎关注Adlik公众号、代码仓库!