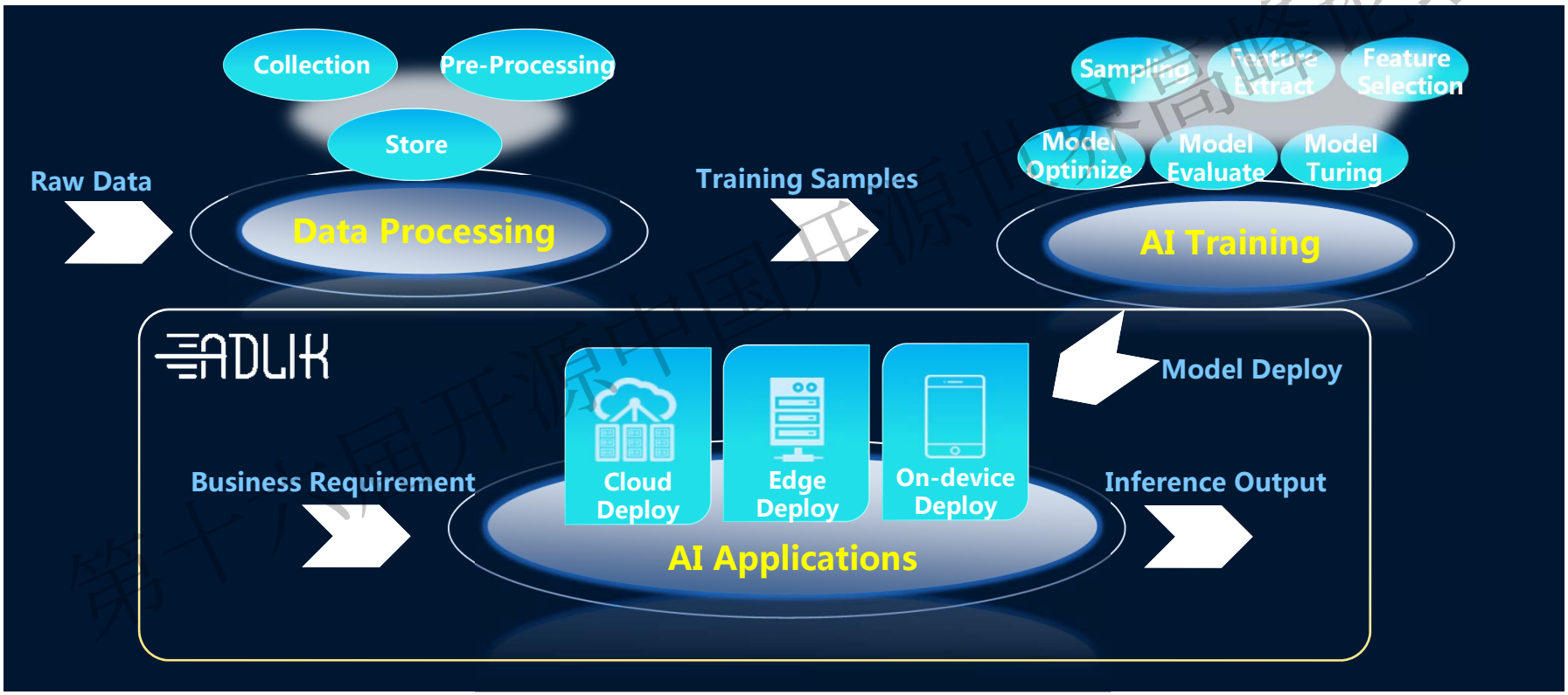
ADLIK

Adlik：

加速模型推理，助力AI落地

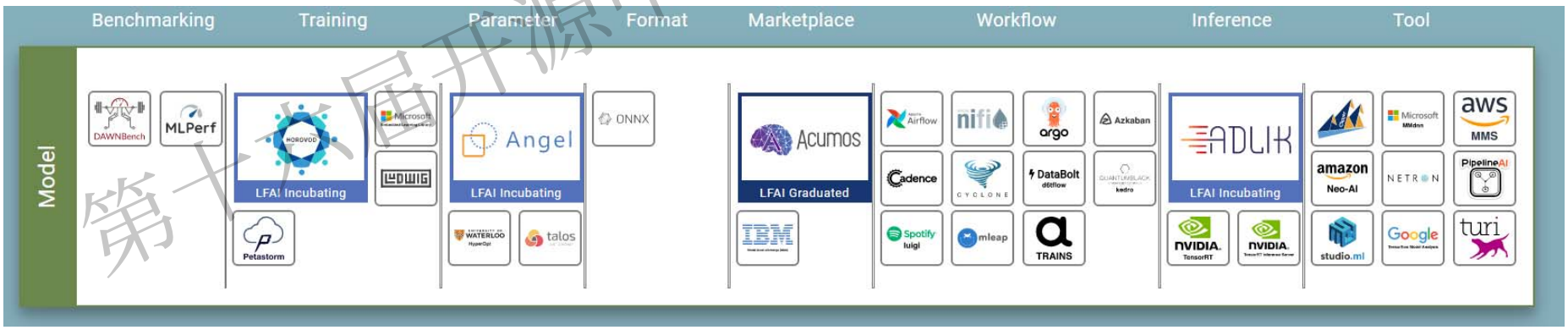
中兴通讯 袁丽雅

Three Big Stages in Machine Learning Pipeline



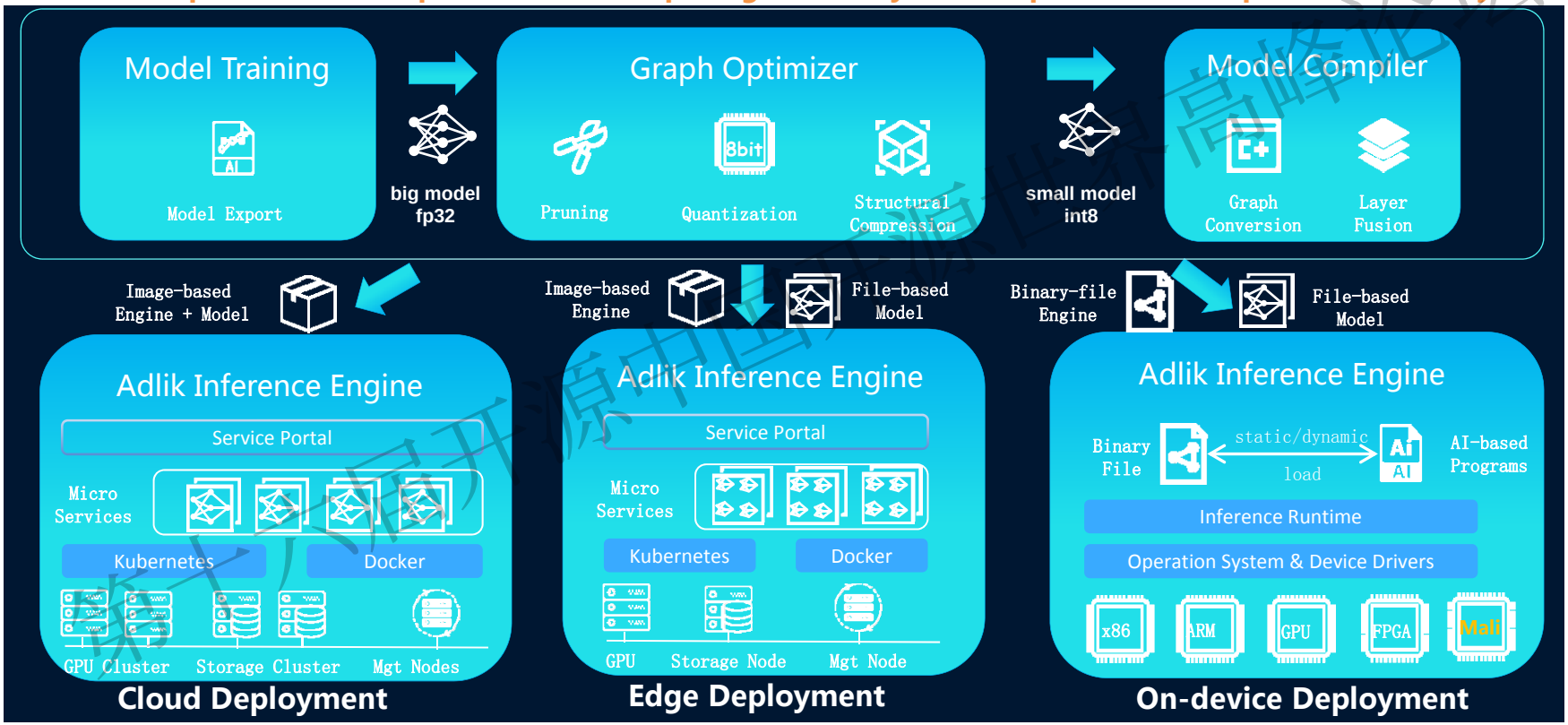
Adlik: A toolkit for accelerating DL inference on specific hardware

- **Integrate existing solutions (TFLite、TensorRT、OpenVINO et.) and provide a universal entrance**
  - Easy to migrate from one hardware to another
  - Easy to expand to support new inference frameworks
- **Automatically decide optimal engineering parameters (backend、batch size etc.)**
  - No learning curve required to select the appropriate solution for model deployment
  - No tedious tuning work to meet performance reqirements (latency、thoughput、resource constrains).



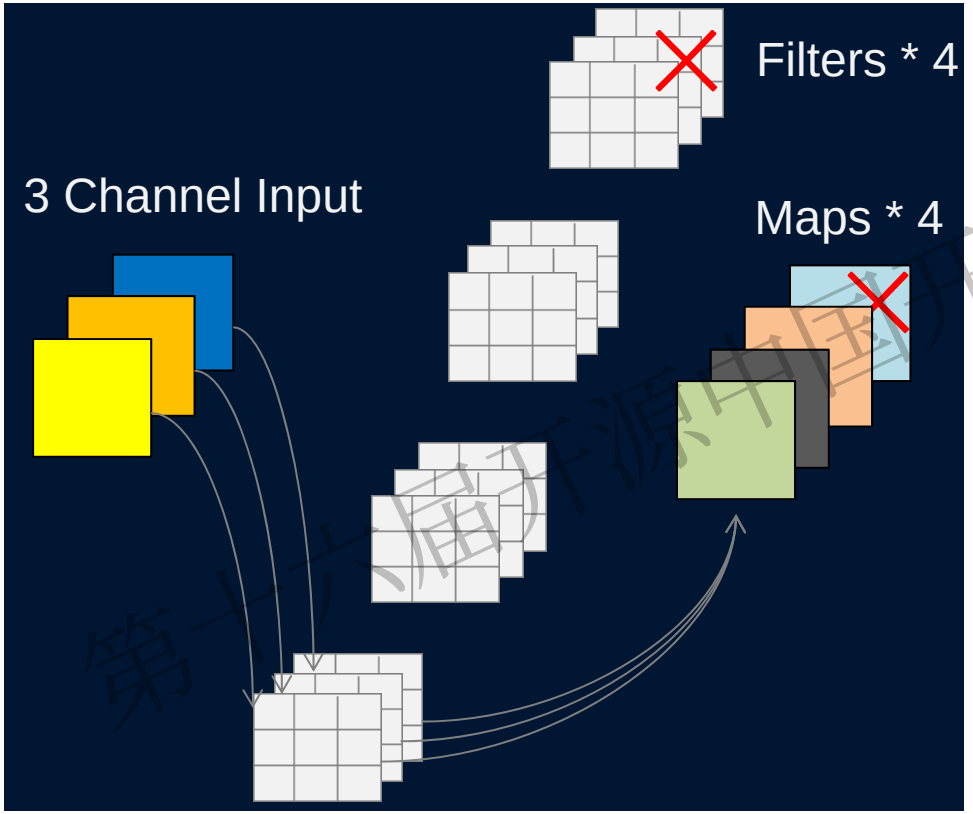
Adlik Architecture

Model Optimizer & Compiler : boost computing efficiency, reduce power consumption and latency



Adlik Engine: support three kinds of deployment environment

Model Optimizer: Pruning

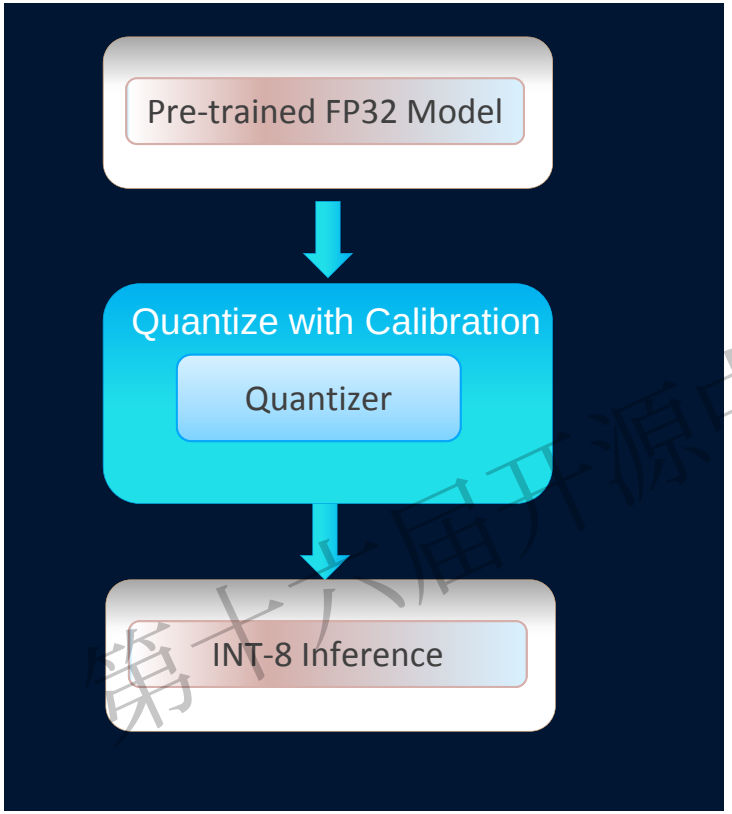


- Supporting multi-nodes and multi-GPU pruning and tuning.
- Supporting channel pruning and filter pruning, reducing the number of parameters and flops.

ResNet-50	Top-1	Parameters	Size
baseline	76.19%	25.61M	99MB
pruned	75.50%	17.43M	67MB

ResNet-50	MACs	Inference speed
baseline	$5.10*10^7$	7.2 pcs/s
pruned	$3.47*10^7$	9.57 pcs/s

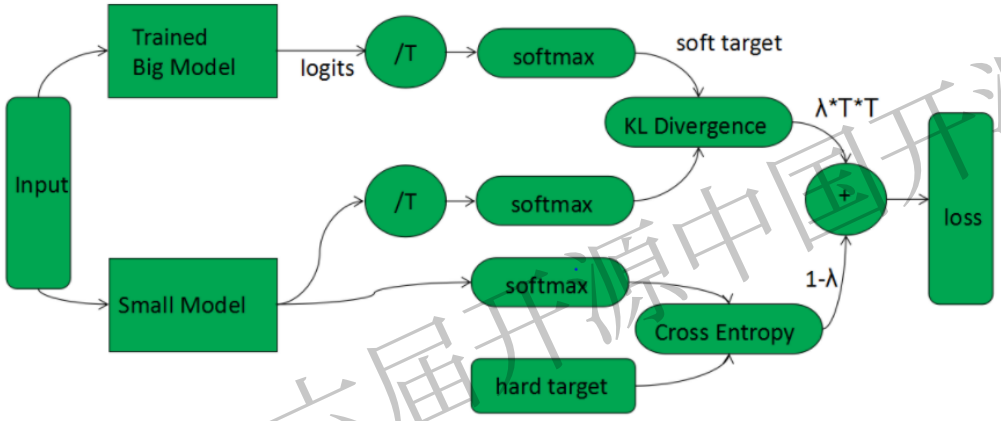
Model Optimizer: Quantizing



- Supporting 8-bit Calibration Quantization.
- Quantizing process needs only a small batch of datasets and few minutes.

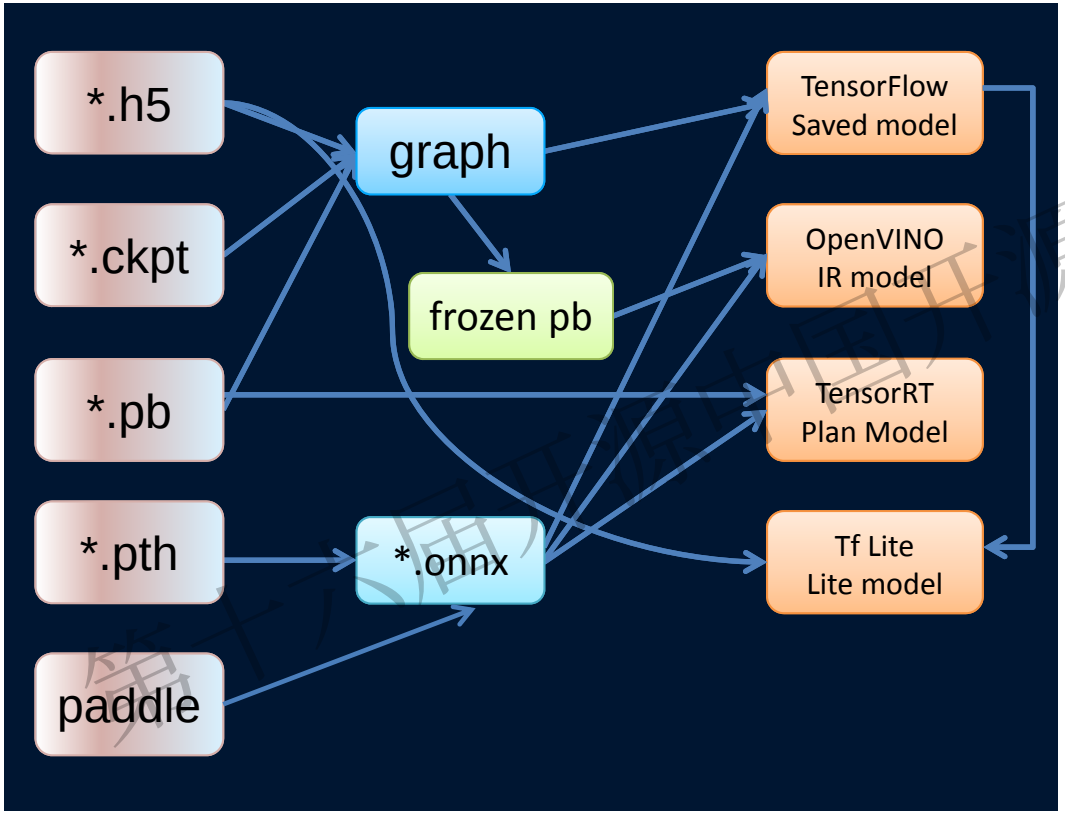
ResNet-50	Top-1	Parameters	MACs	Size
baseline	76.19%	25.61M	$5.10*10^7$	99MB
pruned	75.50%	17.43M	$3.47*10^7$	67MB
pruned+quantized(TF-Lite)	75.3%	17.43M	$3.47*10^7$	18MB

Model Optimizer: Knowledge Distillation



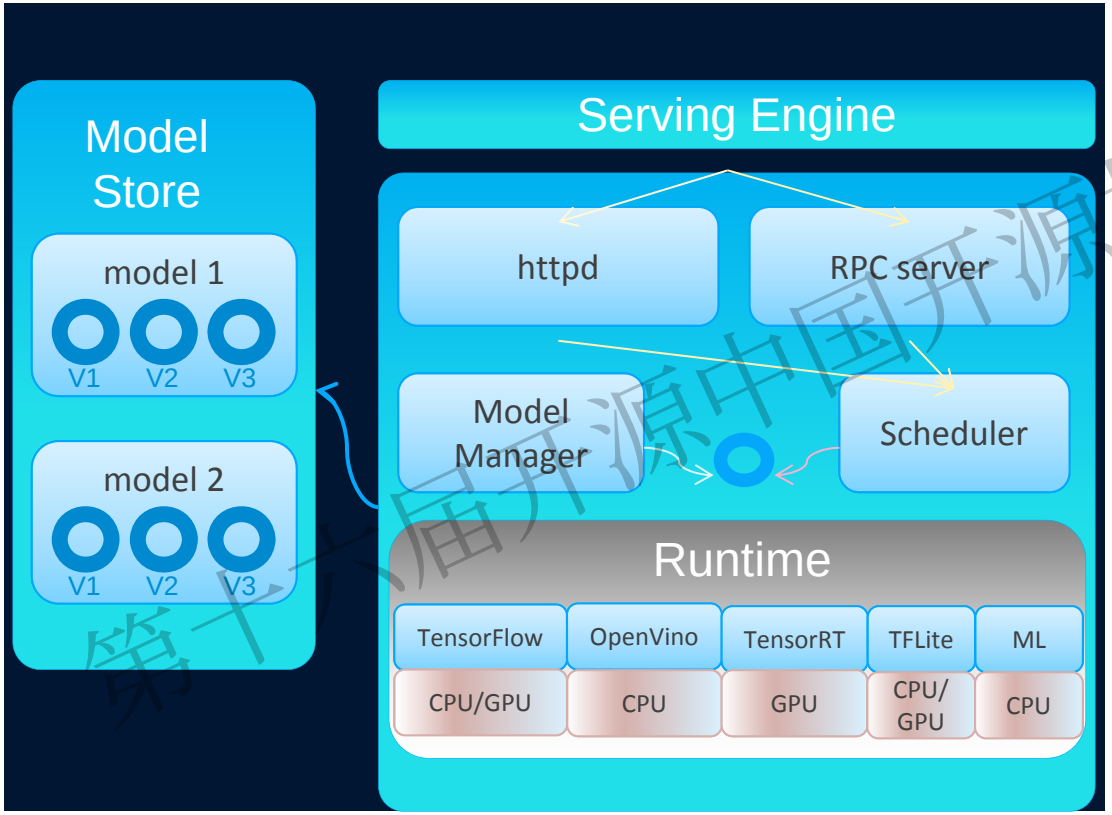
- Reduce the scale of the small model, and decrease the number of parameters and flops.
- Increase the performance of the small model.

Model Optimizer& Compiler



- Supporting several original trained model formats and target runtime formats with unified compiling request.
- Support DAG generation for end-to-end compilation of models with different representation.
- Support model quantization for TfLite and TensorRT.

Adlik Inference Engine



- Model upload, upgrade, versioning, inference and monitoring
- Unified inference interface
- Unified management and scheduling of multi-runtime, multi-model and multi-instance
- Supporting hybrid scheduling of ML and DL inference runtime
- Easy to expend

Adlik in Cloud Native Environment

①

```
docker run -it --rm -v /media/B/work/keras:/model 10.233.170.2:5000/adlik/model-compiler:7.0_10.0 bash
        root@ecaf2fd16421:/# cd model/
        root@ecaf2fd16421:/model# python3 compile_model.py
Source type: ONNXModelFile.
Target type: OpenvinoModel.
Compile path: ONNXModelFile -> OpenvinoModel.
{'status': 'success', 'path': 'model_tf_yolov3_608_128/yolov3_1.zip'}
```

②

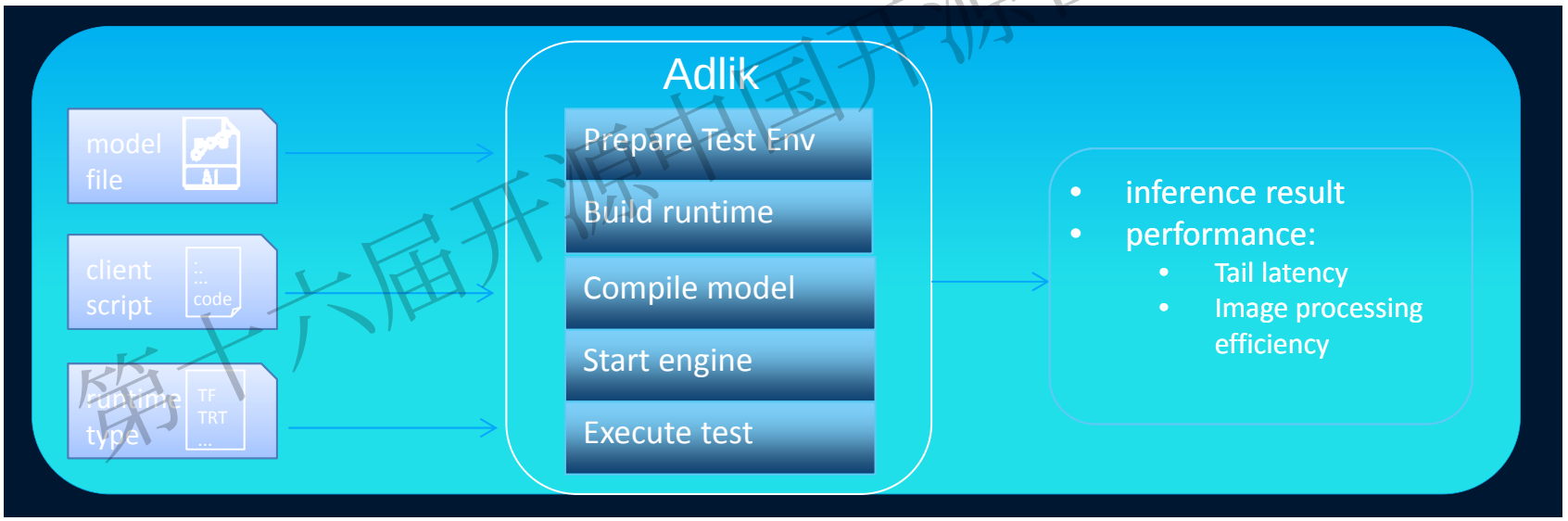
```
docker run -it --rm -v /home/t630/zkl:/model -p 31000:8500 10.233.170.2:31000/00253486/adlik_serving-openvino:latest bash
/# adlik-serving --model_base_path=/model/yolov3_repos/ --grpc_port=8500 --http_port=8501
I adlik_serving/server/core/server_core.cc:54] Adlik serving is running...
I adlik_serving/server/grpc/grpc_options.cc:88] grpc server port: 8500
I adlik_serving/server/grpc/grpc_server.cc:24] grpc server is serving...
I adlik_serving/server/http/http_options.cc:35] http server port: 8501
```

③

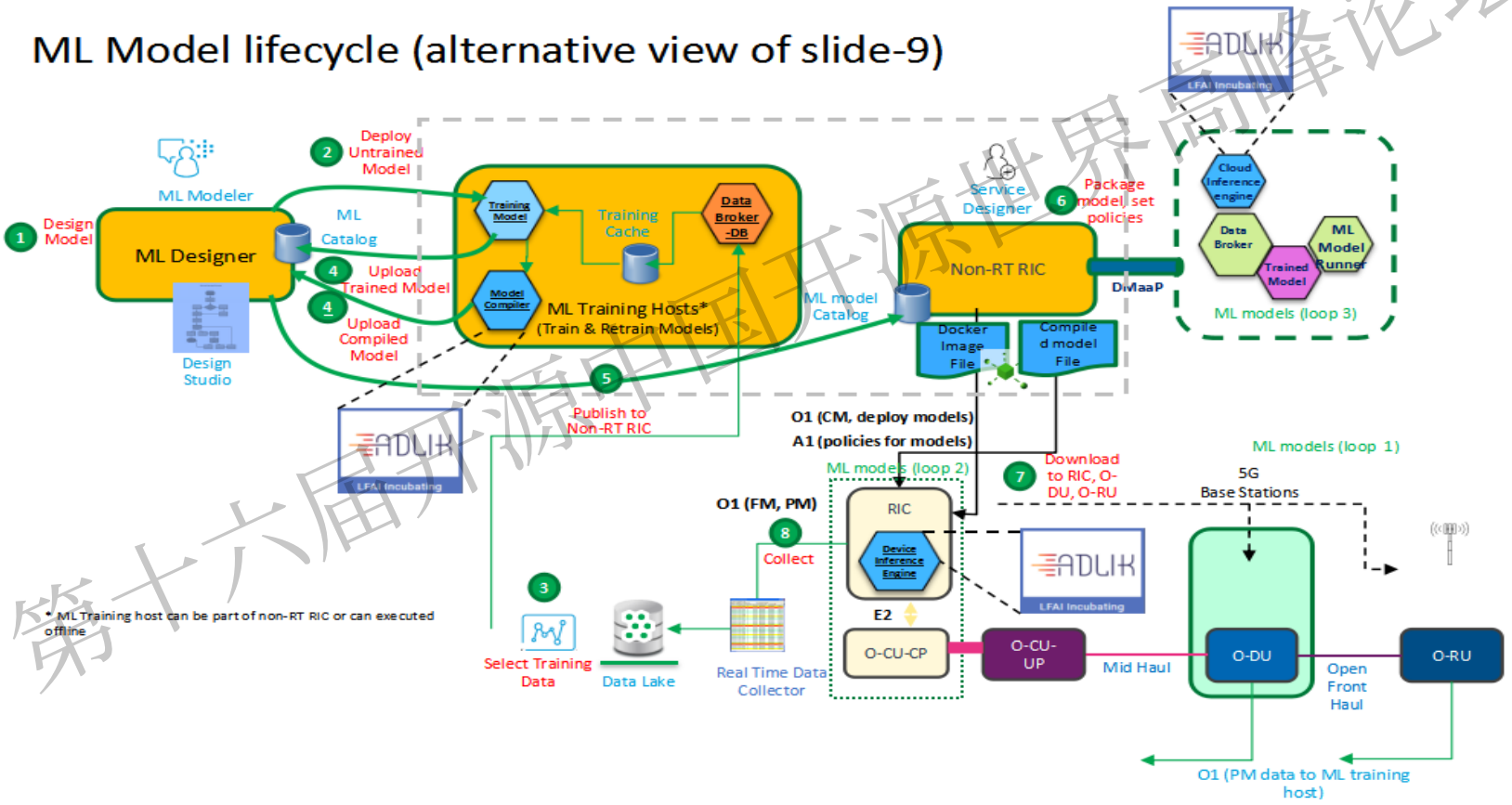
```
python3 yolov3_client.py -n yolo416 -b 1 dog.jpg
```

Automatic Test

- A containerized solution which could automatically execute all test steps.
- Support all compilers and runtimes integrated in Adlik.
- Usage scenarios: DevOps, Benchmark test, etc..



Use Case: Adlik for O-RAN



Adlik Development Status

Initial Commit:
Supporting h5,ckpt to be compiled to Tensorflow serving, TensorRT, OpenVino

2019.9

2019.12

Supporting custom-defined runtime
Supporting custom-defined service core;
Supporting multi-model management and scheduling;

Release 0.1.0:
Supporting ARM cpu
Supporting Optimizer
Supporting ML;
Auto Test framework

2020.6

Release 0.3.0:
Support knowledge ditillation;
Support more framework;
Introduce AutoML into Adlik;

2020.11

Release:
Release 0.2.0:
Adlik Cloud Service;
Hybrid scheduling of ML and DL inference jobs;
New compiler framework;

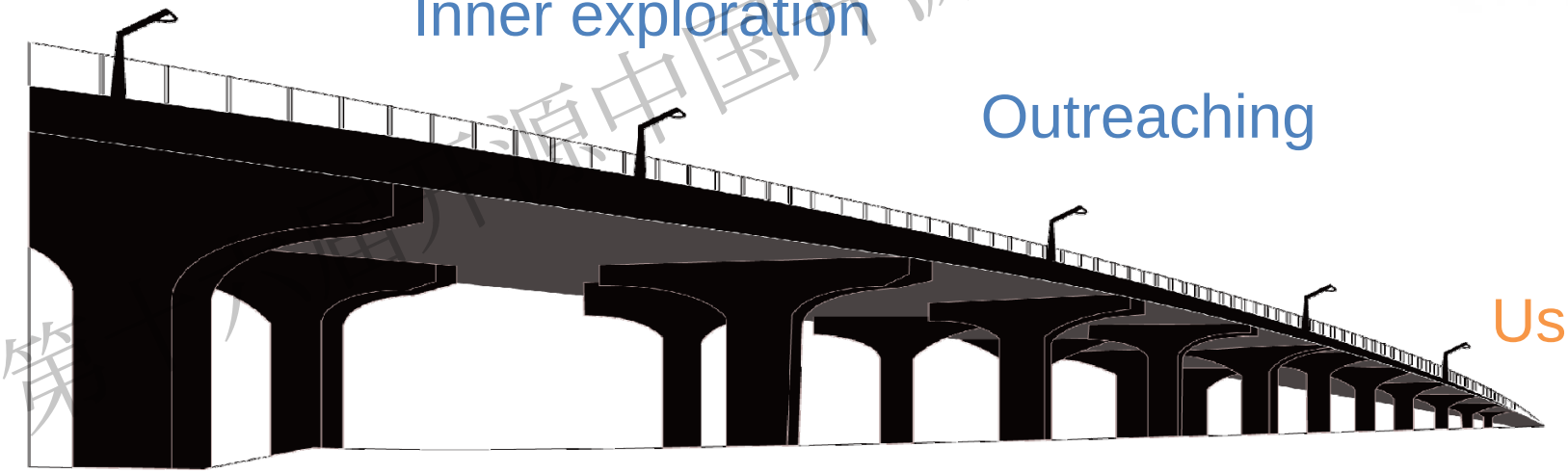
2021.6

Get connected by open source

Project

Inner exploration

Outreaching



User

Welcome to join Adlik!

Let's embrace era of open source!