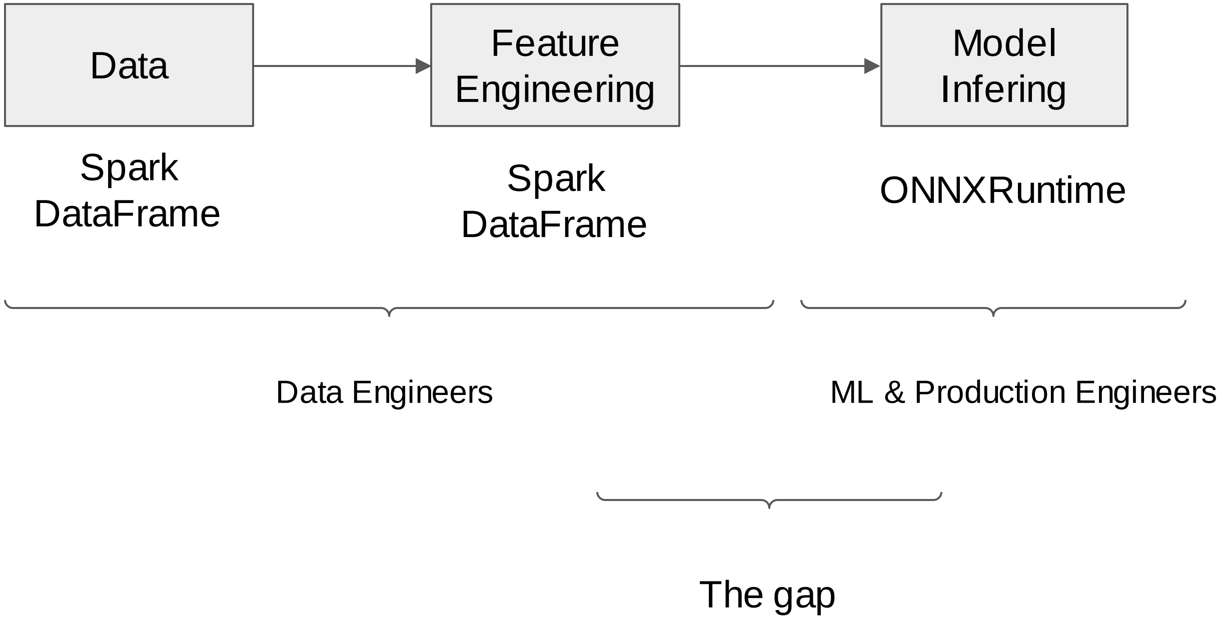
Bring the power of ONNX to Spark as it never happened before

Huawei
Yikun Jiang, Xiyuan Wang, Zhipeng Huang

A Simplest Workflow of Spark + ONNX (Infering)



Data

Feature Engineering

Model Infering

Spark DataFrame

Spark DataFrame

ONNXRuntime

Data Engineers

ML & Production Engineers

The gap

Spark SPIP: Simplified API for DL Inferencing

A new Spark Project Improvement Proposal (SPIP) is being discussed by the community to offer a simplified API for deep learning inference, including built-in integration with popular DL Frameworks:

Goal:

- Simplify the deployment of DL models to Spark Ineference
- Enable integrtions with 3rd-party DL Frameworks

Target Personas:

- Data Engineer who need to deploy DL models on Spark
- Developers who need to deploy DL models on Spark

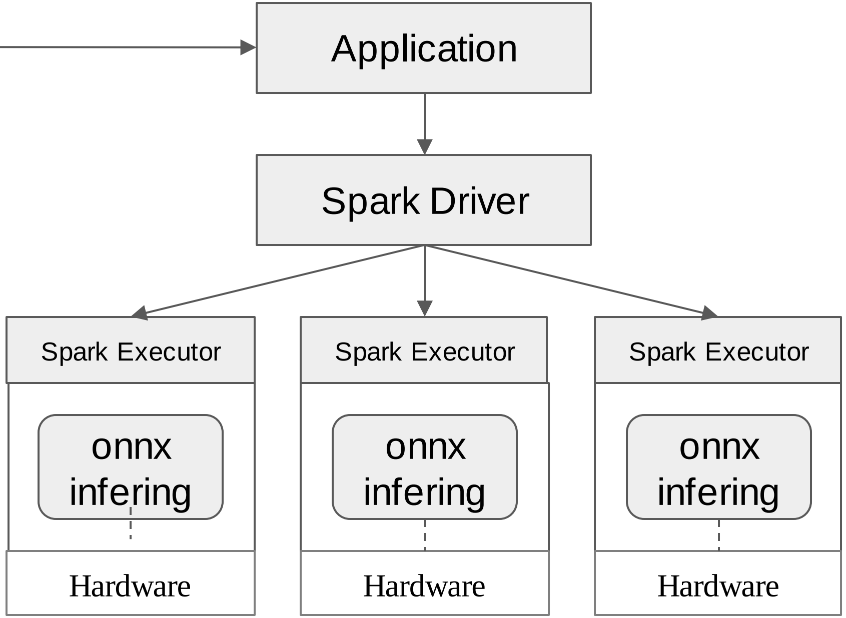
JIRA: SPARK-38648

A complete view for Spark + ONNX

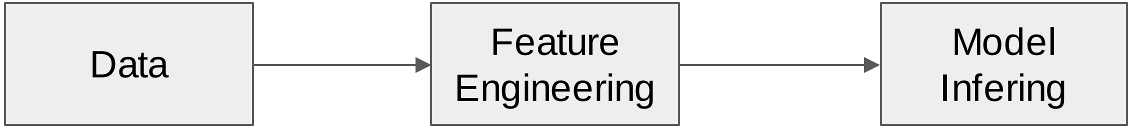
```
from spark.xxx.onnx_runtime import model_udf

predict = model_udf(model_url)

df = DataFrame(data_path)
df.withColumn("preds", predict(col("data")))
```

You just give the model and use model_udf

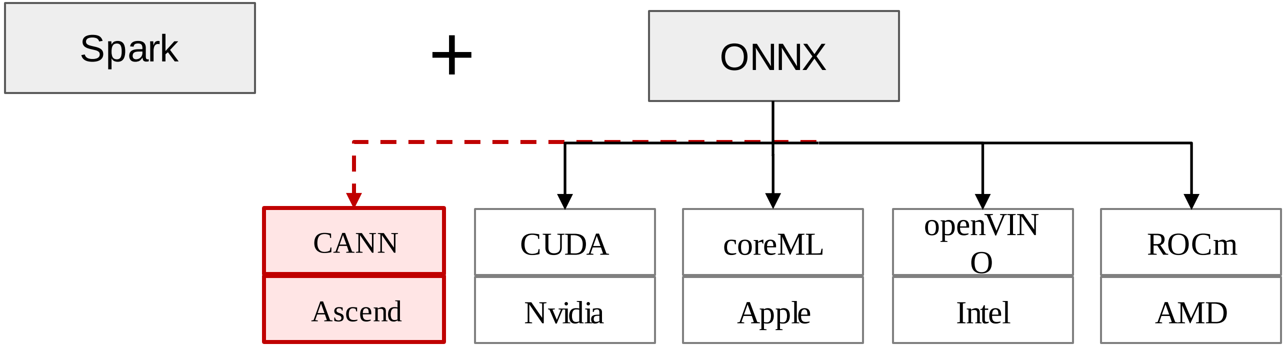
DL on Spark will do the reset.



Spark + ONNX + Hardware !



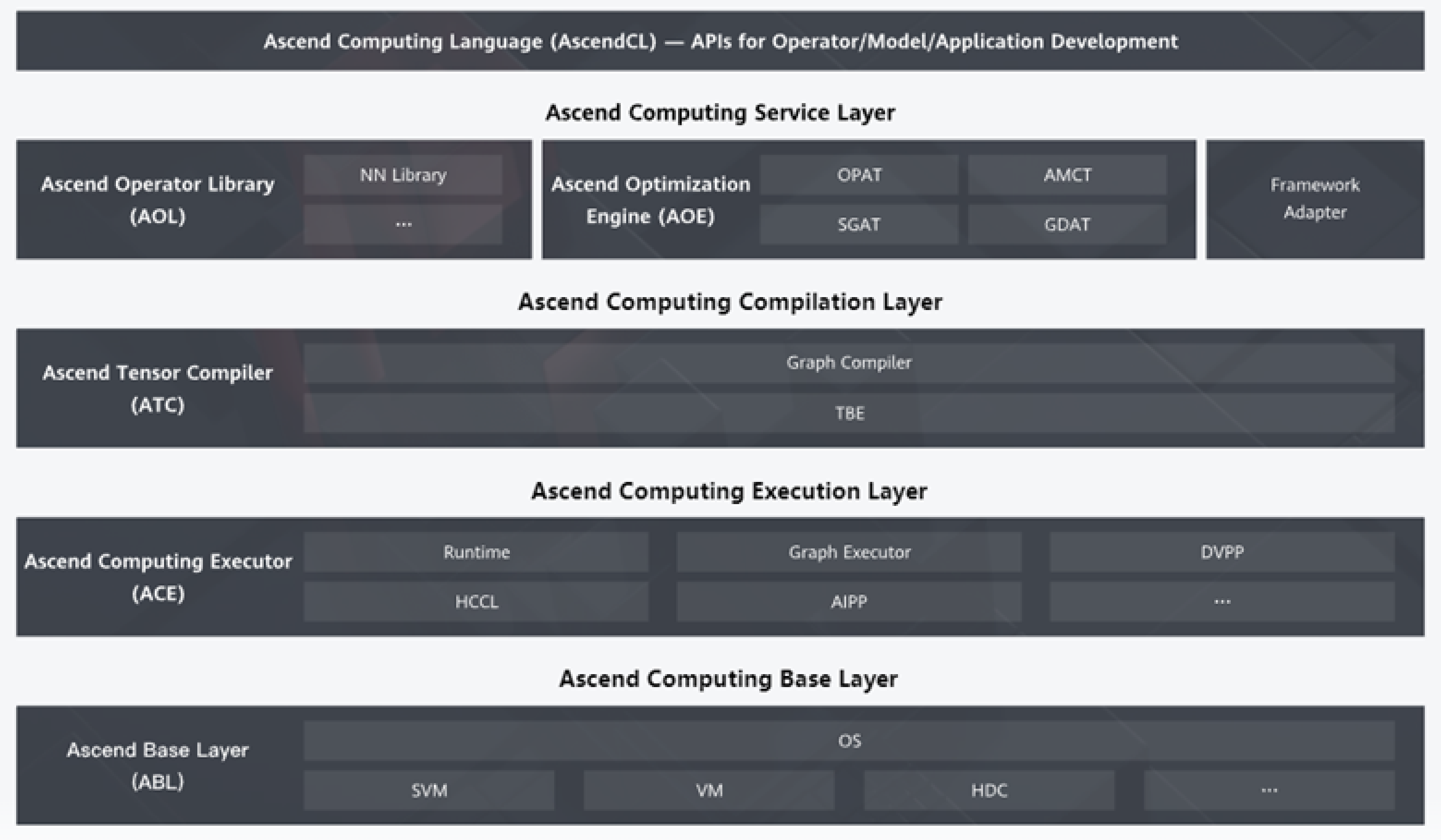
Data → Feature Engineering → Model Infering



Spark + ONNX

CANN	CUDA	coreML	openVINO	ROCm
Ascend	Nvidia	Apple	Intel	AMD

Ascend CANN Technical Stack



Ascend Computing Language (AscendCL) — APIs for Operator/Model/Application Development

Ascend Computing Service Layer

Ascend Operator Library (AOL)
NN Library
...

Ascend Optimization Engine (AOE)
OPAT
AMCT
SGAT
GDAT

Framework Adapter

Ascend Computing Compilation Layer

Ascend Tensor Compiler (ATC)
Graph Compiler
TBE

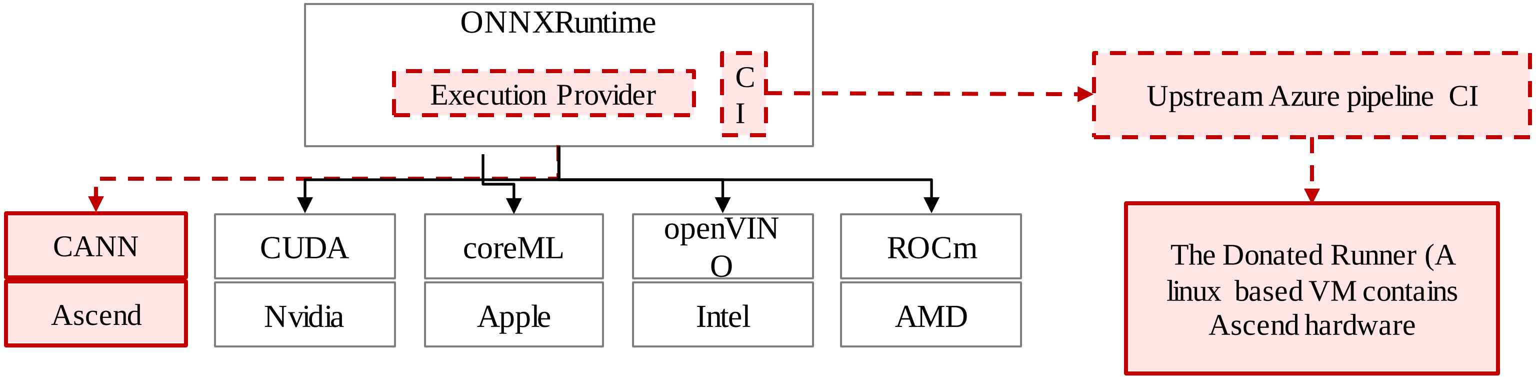
Ascend Computing Execution Layer

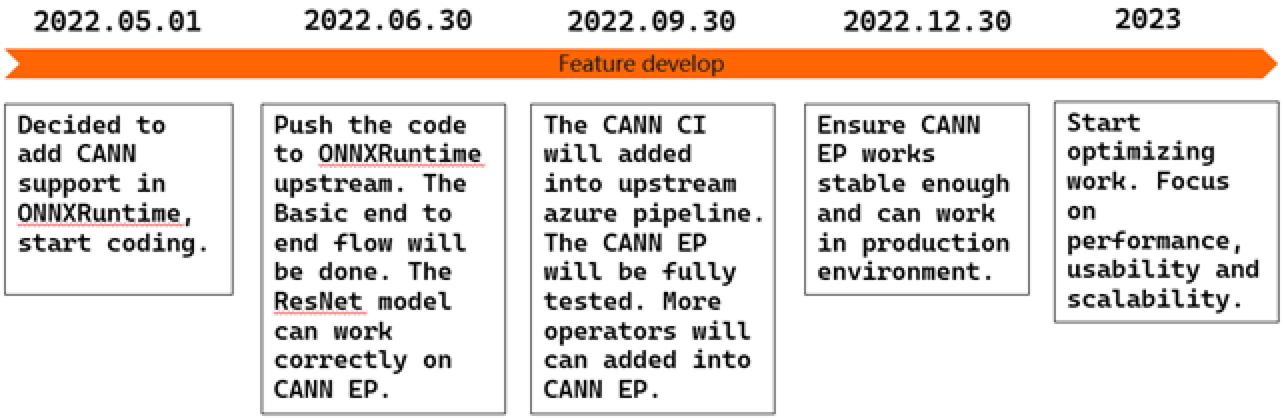
Ascend Computing Executor (ACE)
Runtime
Graph Executor
DVPP
HCCL
AIPP
...

Ascend Computing Base Layer

Ascend Base Layer (ABL)
OS
SVM
VM
HDC
...

ONNXRuntime CANN execution provider support


ONNXRuntime
Execution Provider
CI
Upstream Azure pipeline CI
CANN
Ascend
CUDA
Nvidia
coreML
Apple
openVINO
Intel
ROCm
AMD
The Donated Runner (A linux based VM contains Ascend hardware



2022.05.01: Decided to add CANN support in ONNXRuntime, start coding.

2022.06.30: Push the code to ONNXRuntime upstream. The Basic end to end flow will be done. The ResNet model can work correctly on CANN EP.

2022.09.30: The CANN CI will added into upstream azure pipeline. The CANN EP will be fully tested. More operators will can added into CANN EP.

2022.12.30: Ensure CANN EP works stable enough and can work in production environment.

2023: Start optimizing work. Focus on performance, usability and scalability.

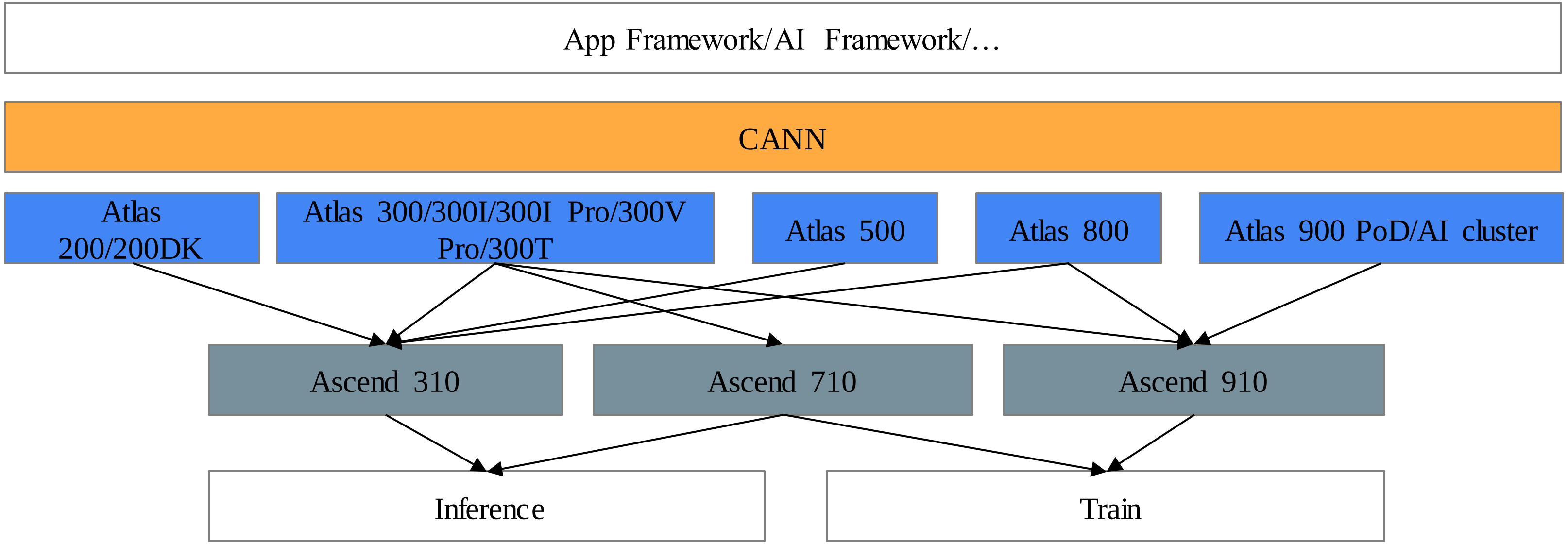
Issue: https://github.com/microsoft/onnxruntime/issues/11477

POC : https://github.com/learningbackup/onnxruntime/tree/add_cann

PR: coming soon

Ascend Stack

Ascend: A series NPU AI Processor from Huawei
Atlas: A series Hardware Powered on Ascend AI Processors
CANN: A heterogeneous compute architecture in AI scenarios provides multi-layer APIs to help you quickly build AI applications and services based on the Ascend platform.


App Framework/AI Framework/…
CANN
Atlas 200/200DK
Atlas 300/300I/300I Pro/300V Pro/300T
Atlas 500
Atlas 800
Atlas 900 PoD/AI cluster
Ascend 310
Ascend 710
Ascend 910
Inference
Train