TVM in the Arene AI Platform of Woven Planet

Dec/16/2021 1:00 pm

Ryo Takahashi
Senior Engineer

Srushti Rashmi Shirish
ML Engineer

Agenda

About my company 3
About my project 6
Why Woven Planet focuses on TVM ? 8
Experiment: Automated Mapping 9

Future works 10

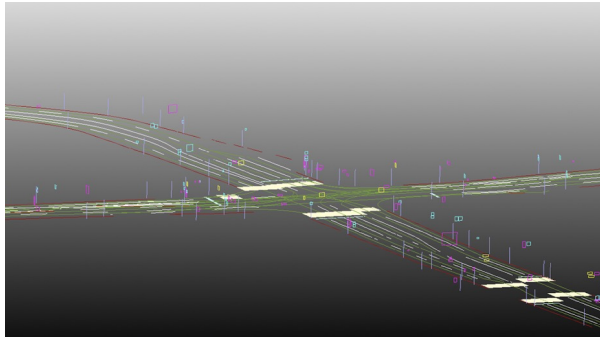
What company is Woven Planet Holdings, Inc. ?

Software company for future mobility solutions

Automated Driving

Robotics

Automated Mapping Platform



Smart City

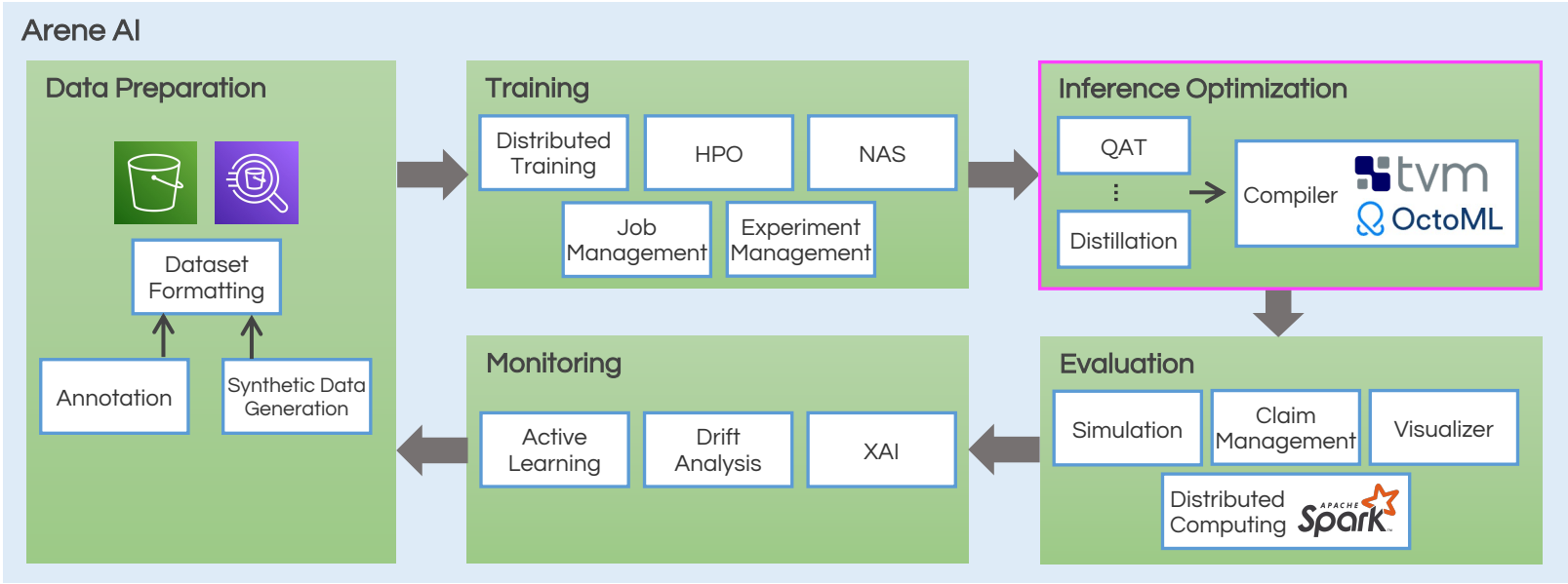
We are still a new company …

Growing together with world-class talent !!

Machine Learning (ML) is the key technology for all of Woven Planet's businesses.
Woven Planet is a treasury of ML applications !!

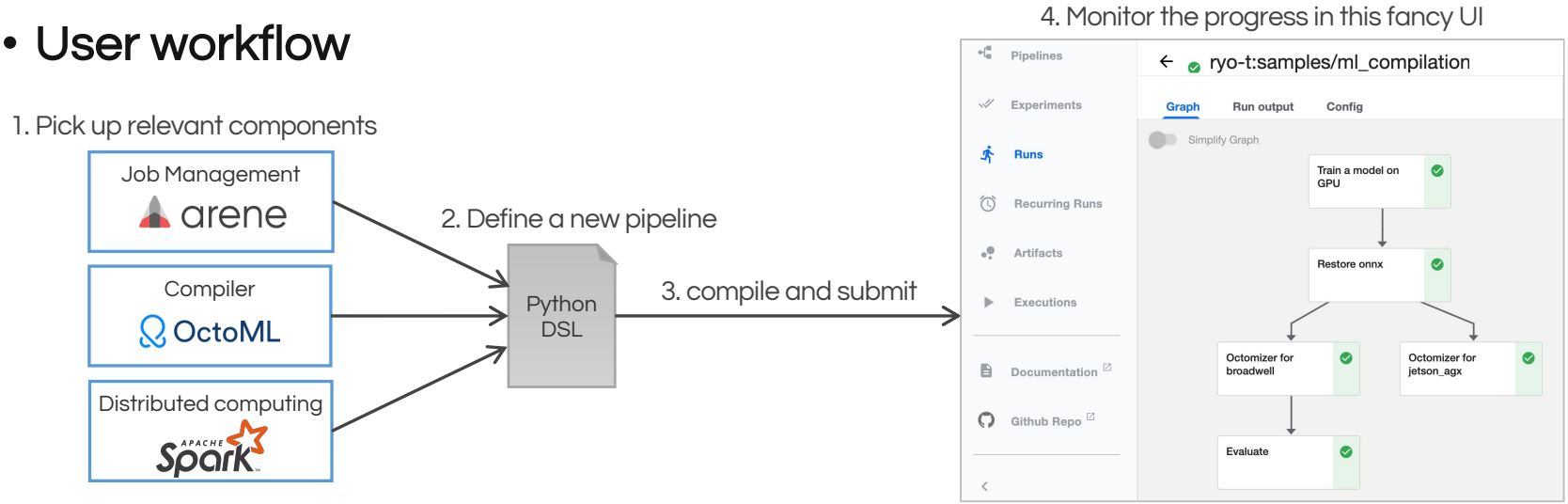
Arene AI Platform

- Provide a cloud-agnostic, common MLOps platform for all the ML applications
- Automate and parallelize all ML workflows to deploy SOTA models timely
- Accumulate/distribute reusable code as *components*



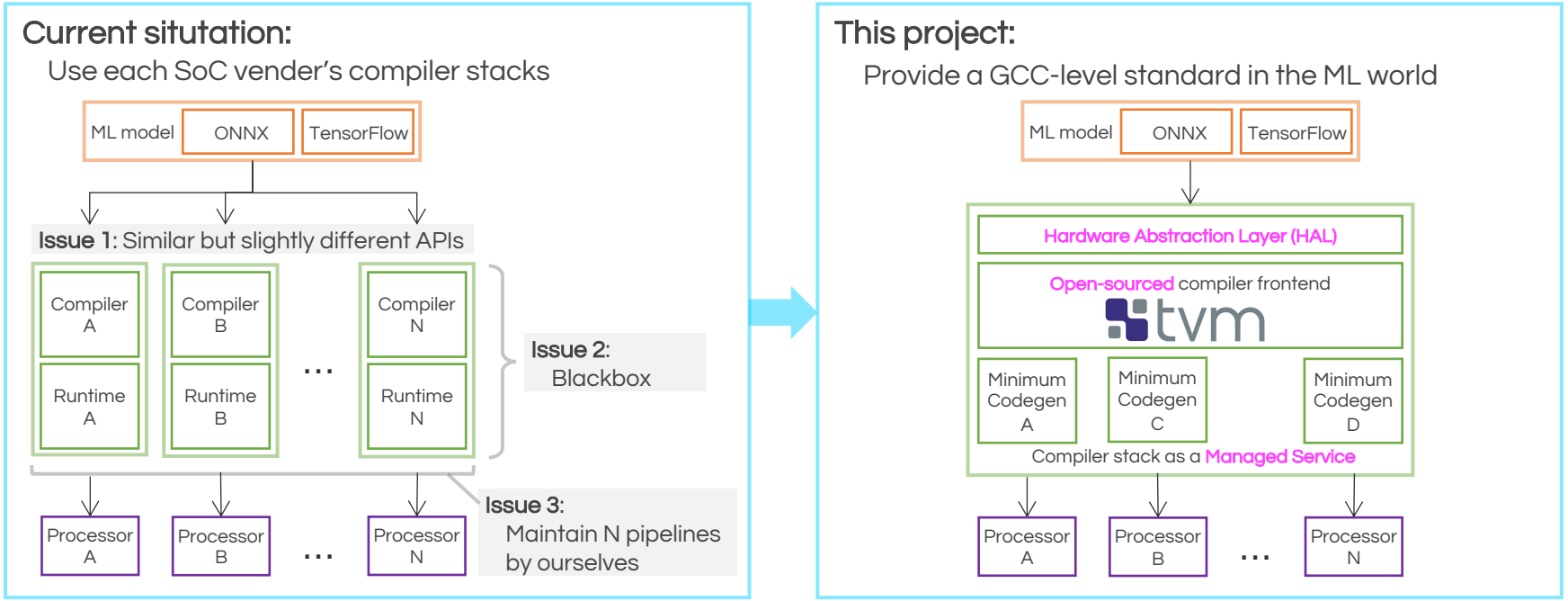
Example: Workflow using Jetson AGX Xavier™

- **Goals**
  1. Train a CNN model on GPU
  2. Cross-compile the model for Volta
  3. Test the same model with multiple x86-64 nodes
- **User workflow**



Why Woven Planet uses TVM in Arene AI Platform ?

Background: Arene AI has to deal with very various processors…



Experiment: Automated Mapping

Run a super huge pipeline on global-scaled data everyday…

Anonymized
vehicle probe data
(e.g. RGB image, LiDAR)

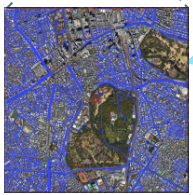
Space maps
(e.g. satellite imagery)

ML Pipeline

ML Pipeline

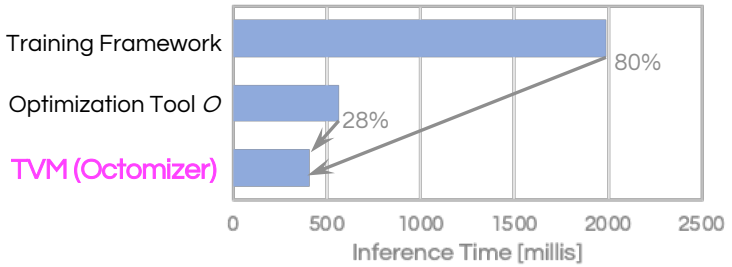
Fusion

High-Definiton Maps



DeepLab-like Segmentation
with hundreds of CPU instances
(AWS c5n.9xlarge [Skylake])

Benchmark



TVM performed the best in the OSS solutions !!

Future works

- Use TVM on actual vehicles (e.g. ADAS/AD ECU)
  - Collaborate with edge NPU venders for TVM integration
- Achieve functional safety with TVM
  - Modify TVM's MISRA-C/C++ compliant runtime [#3494] to meet Woven Planet's safety standards
- Connect TVM to our in-house hardware-aware NAS
  - PAPIMetricCollector [#7983] is what we were looking for !!

Name	perf::CACHE-MISSES	perf::CYCLES	perf::STALLED-CYCLES-BACKEND	perf::INSTRUCTIONS
fused_nn_dense_nn_bias_add_nn_relu	2,494	1,570,698	85,608	675,564
fused_nn_dense_nn_bias_add_nn_relu_1	1,149	655,101	13,278	202,297
fused_nn_dense_nn_bias_add	288	600,184	8,321	163,446
fused_nn_batch_flatten	301	587,049	4,636	158,636
fused_nn_softmax	154	575,143	8,018	160,738

Sum	4,386	3,988,175	119,861	1,360,681
Total	10,644	8,327,360	179,310	2,660,569

Summary

- We, Woven Planet, need various ML accelerators for future mobility solutions
- We expect TVM to become the HAL on all the ML accelerators
- We are moving towards practical uses with TVM

Thanks!

Let's get Apache TVM going in industry and academia :)
(We're hiring !! Find out more in our Medium blog)