Open Source at OctoML

TVM Meetup 11/8/2019

Jared Roesch

OctoML is a new company building DL deployment solutions using the Apache (incubating) TVM project.

A goal is to nurture the TVM community and contribute new infrastructure and features.

octoml.ai
@octoml

Founding Team - The Octonauts

Luis Ceze
Co-founder, CEO
PhD in Computer Architecture and Compilers
Professor at UW-CSE
Venture Partner, Madrona Ventures
Previously: IBM Research, consulting for Microsoft, Apple, Qualcomm

Jason Knight
Co-founder, CPO
PhD in Computational Biology and Machine Learning
Previously: HLI, Nervana, Intel

Tianqi Chen
Co-founder, CTO
PhD in Machine Learning
Professor at CMU-CS

Thierry Moreau
Co-founder, Architect
PhD in Computer Architecture

Jared Roesch
Co-founder, Architect
(soon) PhD in Programming Languages

40+ years of combined experience in computer systems design and machine learning

tvm mxnet dmlc XGBoost nGraph

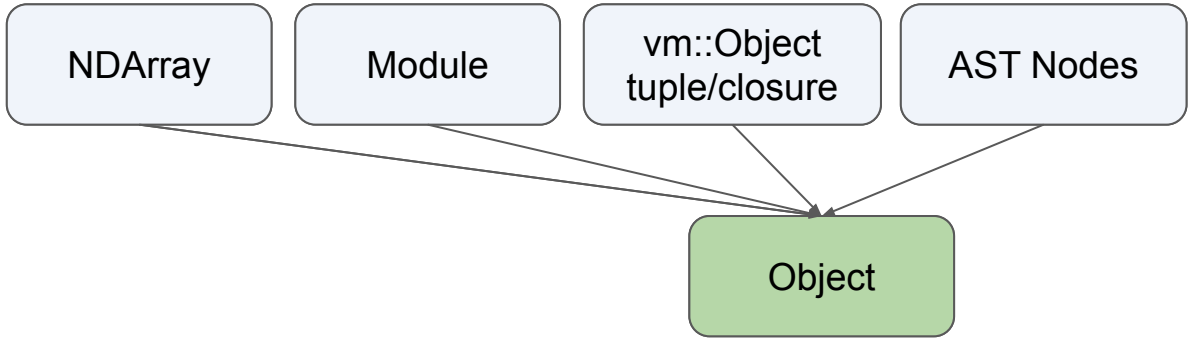
Open Source at OctoML

- We are big believers in the power of open source
  - Sponsoring multiple employees to contribute to TVM.
- Today we'll touch on a few of those contribution areas:
  - Core Infrastructure Improvements to TVM
  - uTVM: support for microcontrollers in TVM
  - Virtual Machine and dynamic NNs support (w/ AWS folks)
  - Improved NLP support, with focus on transformers

Core Infrastructure Refactors

- New Integer Analysis Infrastructure
  - Supports the ability to handle nested division and modulus
  - Improves the ability to reason about and optimize loops
- Support for different integer division modes, floor division and truncating division.
- Unified Object and Node system for TVM runtime
  - Lays groundwork for improved multi-language support for exposing runtime, and IRs.

Unified Object Protocol



Cross language support

Easy to introduce new runtime objects (trees, graphs)

Direct access from other languages

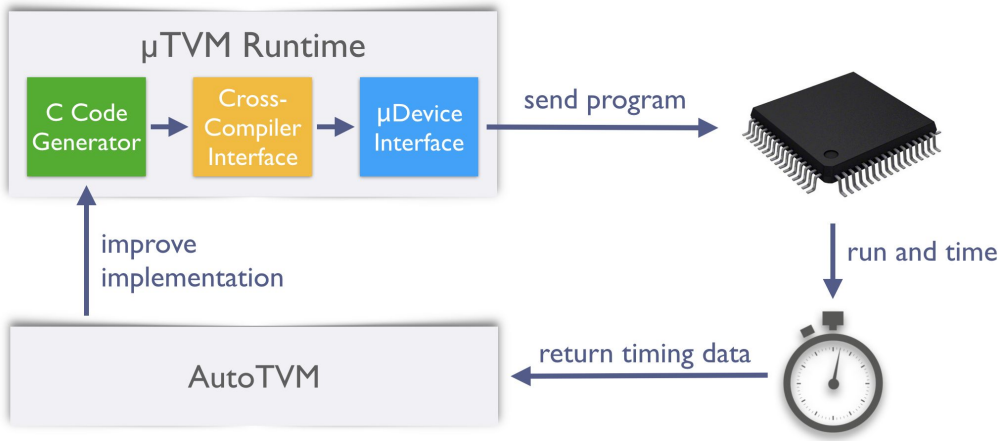
µTVM Overview

- Plug directly into TVM as a backend
- Target C to emit code for microcontrollers that is device-agnostic

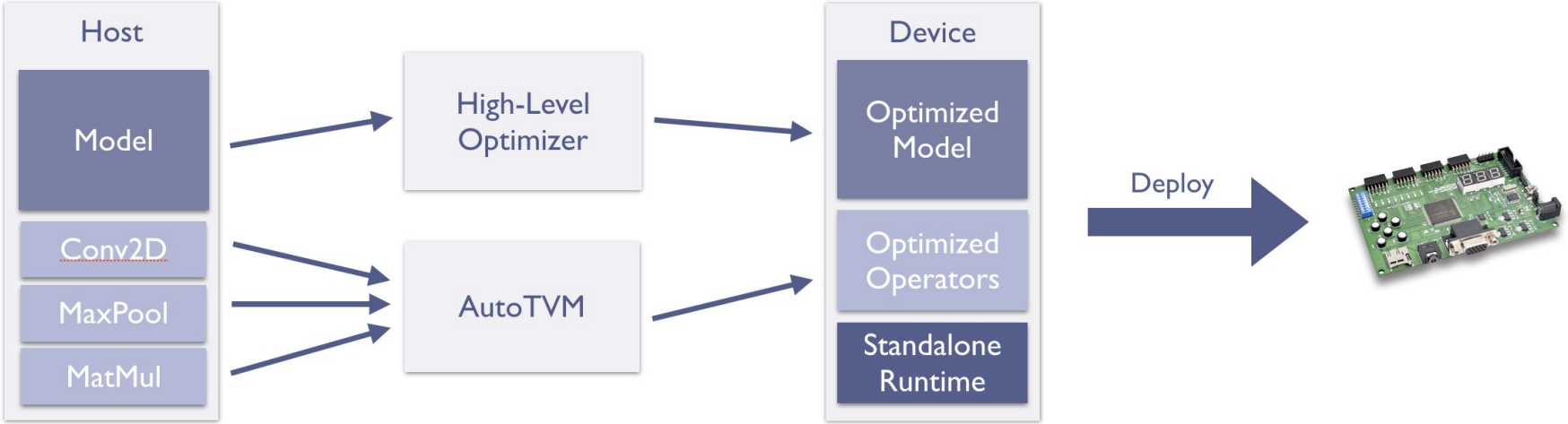
AutoTVM on µTVM

Optimize TVM operators on microcontrollers by making use of AutoTVM

https://github.com/apache/incubator-tvm/pull/4274



Coming Soon to µTVM (Self-Hosted Models)



Transformer Improvements

Transformer based models such as BERT have recently become very popular and require first class support in TVM.

- What we've done:
  - Extend the relay ONNX frontend to support all opset versions of BERT.
    - This enables importing of native ONNX models and those converted from Tensorflow.
  - Improve scheduling of batch matrix multiplies.
    - Early autotuning templates improve performance by ~20%
- What we're working on:
  - BERT has many reshape operations, which are currently implemented using copy.
  - This prevents most compute layers from being fused.
  - Reshape could be implemented as a non-copying view instead.
  - We want to add this form of view as a relay intrinsic to enable highly fused and optimized transformer models.

Virtual Machine

- Many improvements from contributors at UW, AWS, and OctoML.
- Initial implementation is quickly moving towards production quality.
  - VM compiler
  - VM runtime
  - VM serialization
  - Dynamic Shape Support
  - Dynamic Shape Allocation
  - Dynamic Shape Code generation
- Looking for more contributions in this part of the system!
- Haichen and I will discuss more details at TVMConf.

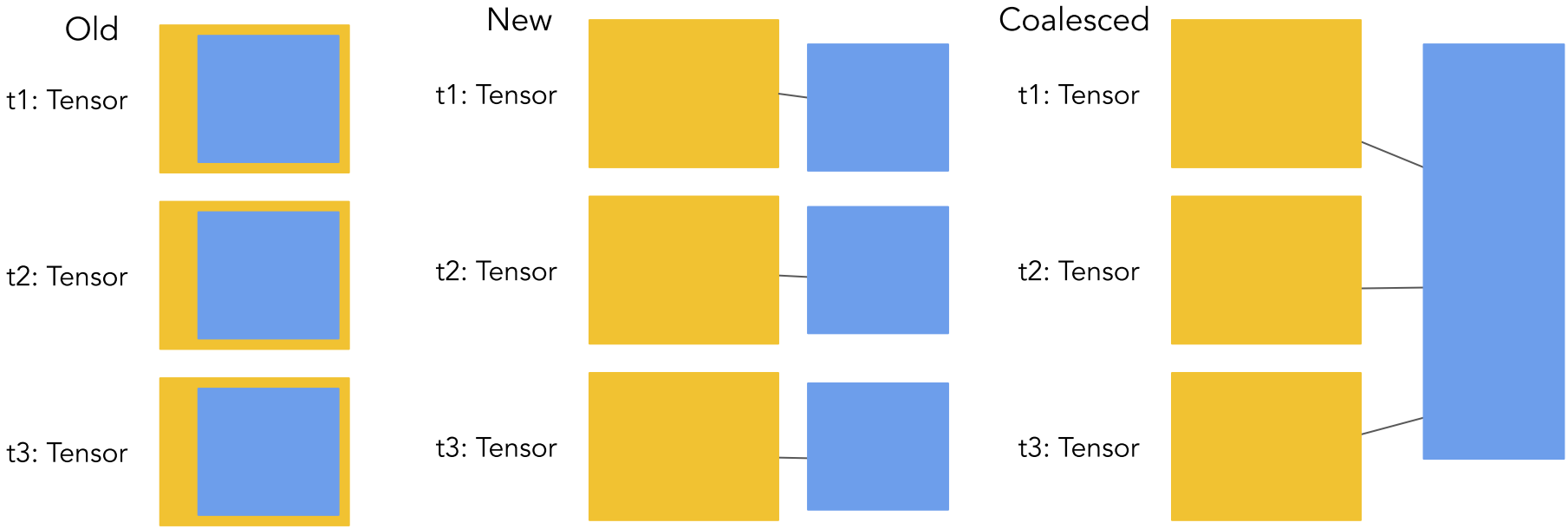
VM Memory Planning

- Recently shipped a first version of dynamic memory planning
  - https://github.com/apache/incubator-tvm/pull/3560
- Enables future optimizations and end-to-end dynamic memory planning, storage coalescing, memory re-use for loops, and offloading dynamic allocation to devices.

```
fn @main() -> Tensor[(k,), f32] {
    let t1: Tensor[(10,), f32] = ...;
    let t2: Tensor[(10,), f32] = ...;
    // Implicitly allocates
    add(t1, t2)
}
```

```
fn @main() -> Tensor[(k,), f32] {
    let t1 = ...;
    let t2 = ...;
    let s = alloc_storage(40, 64, f32);
    let out1 = alloc_tensor(s, (10,), f32);
    invoke_tvm_op(add, (t1, t2), (out1,));
    out1
}
```

VM Memory Abstractions



Acknowledgments

- The Apache(incubating) community members.
- ASF Mentors and PMC members who make this awesome project possible!
- AWS for hosting the first Bay Area meetup

Annual TVM Conference 2019

Organized and participated
by community members

Thursday, December 5th
Seattle WA

Register Today!

tvmconf.org

9:00	Keynote and Community Update
10:00	TVM @ AWS – Yida Wang, Amazon
10:40	TVM @ FB – Andrew Tulloch and Bram Wasti, Facebook
11:10	**break**
11:40	AI Compilers at Alibaba – Yangqing Jia, Alibaba
12:10	Dynamic Execution and Virtual Machine, Jared Roesch and Haichen Shen, UW and AWS
12:30	**Lunch (boxed lunches will be provided), contributors meetup**
13:30	Building FPGA-Targeted Accelerators with HeteroCL – Zhiru Zhang, Cornell
14:00	TVM @ Microsoft – Jon Soifer and Minjia Zhang
14:20	TVM @ ARM
14:40	TVM @ Xilinx – Eliott Delaye
15:00	**break**
15:30	TVM @ OctoML – Jason Knight
15:50	TVM @ Qualcomm
16:10	Talk by Nilesh Jain, Intel Labs
16:30	Talk by Zhihao Jia, Stanford
17:00	**Lightning talks session**
	TensorCore and Tensorization – Siyuan Feng, SJTU
	uTVM: TVM on bare-metal devices – Logan Weber, UW
	TVM for ads ranking stack, opportunities and challenges – Hao Lu and Ansha Yu, Facebook
	TVM for edge computing platforms – Morita Kazutaka, NTT
	Efficient quantized inference on CUDA with TVM – Wuwei Lin, CMU
	Supporting TVM on RISC-V Architectures – Jenq-Kuen Lee, NTHU
	Integrating model pre-processing functionality into TVM – Abelardo Lopez-Lagunas, Latent AI
	Talk by Josh Fromm, OctoML
	More talks to be added
18:15 to 20:00	*Social (drinks, food)*

Questions?

We are hiring see octoml.ai for more details!