End-to-End Deep Learning Compiler Stack

AWS AI

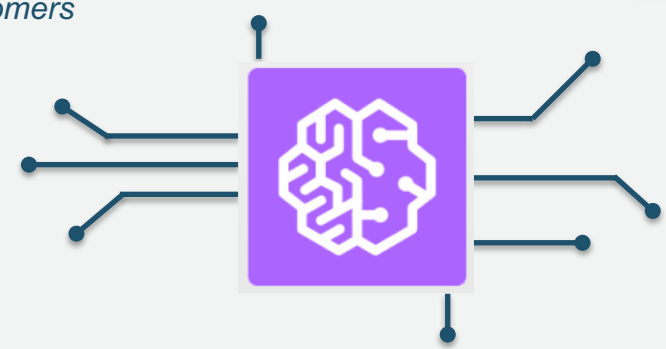
Presenter: Animesh Jain
Amazon SageMaker Neo

Deep Learning is Pervasive

Amazon Lex
Build chatbots to engage customers

Amazon Rekognition
Deep learning-based image and video analysis

Amazon Transcribe
Automatic speech recognition



ONEHOUR TRANSLATION

Amazon Translate
Fluent translation of text

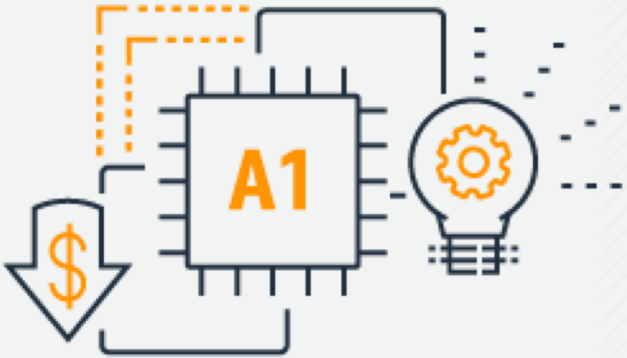
Amazon Polly
Natural sounding text to speech

Amazon Comprehend
Discover insights and relationships in text

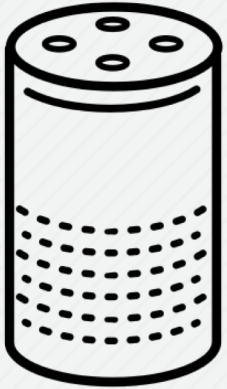
aws

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

ARM – Unique Role in AWS Ecosystem



A1 EC2 Instance
Optimized cost and performance for scale-out workloads

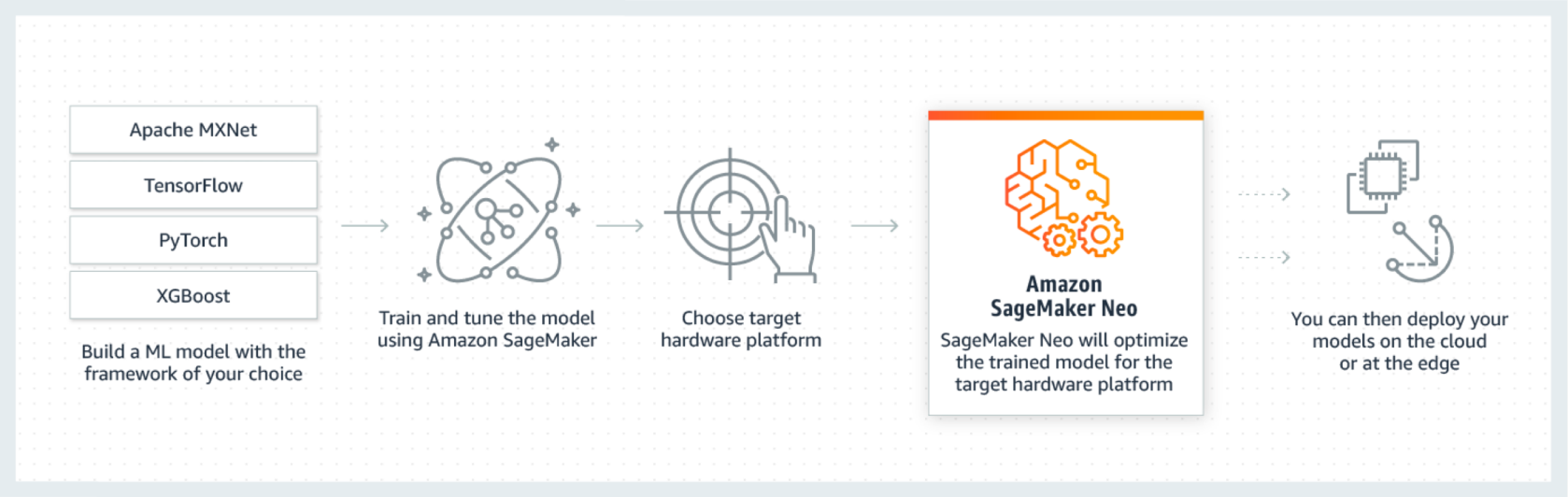


Alexa Devices
Build natural voice experiences

3rd party Edge Devices
Portable Deep Learning

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How to Accelerate Deep Learning?



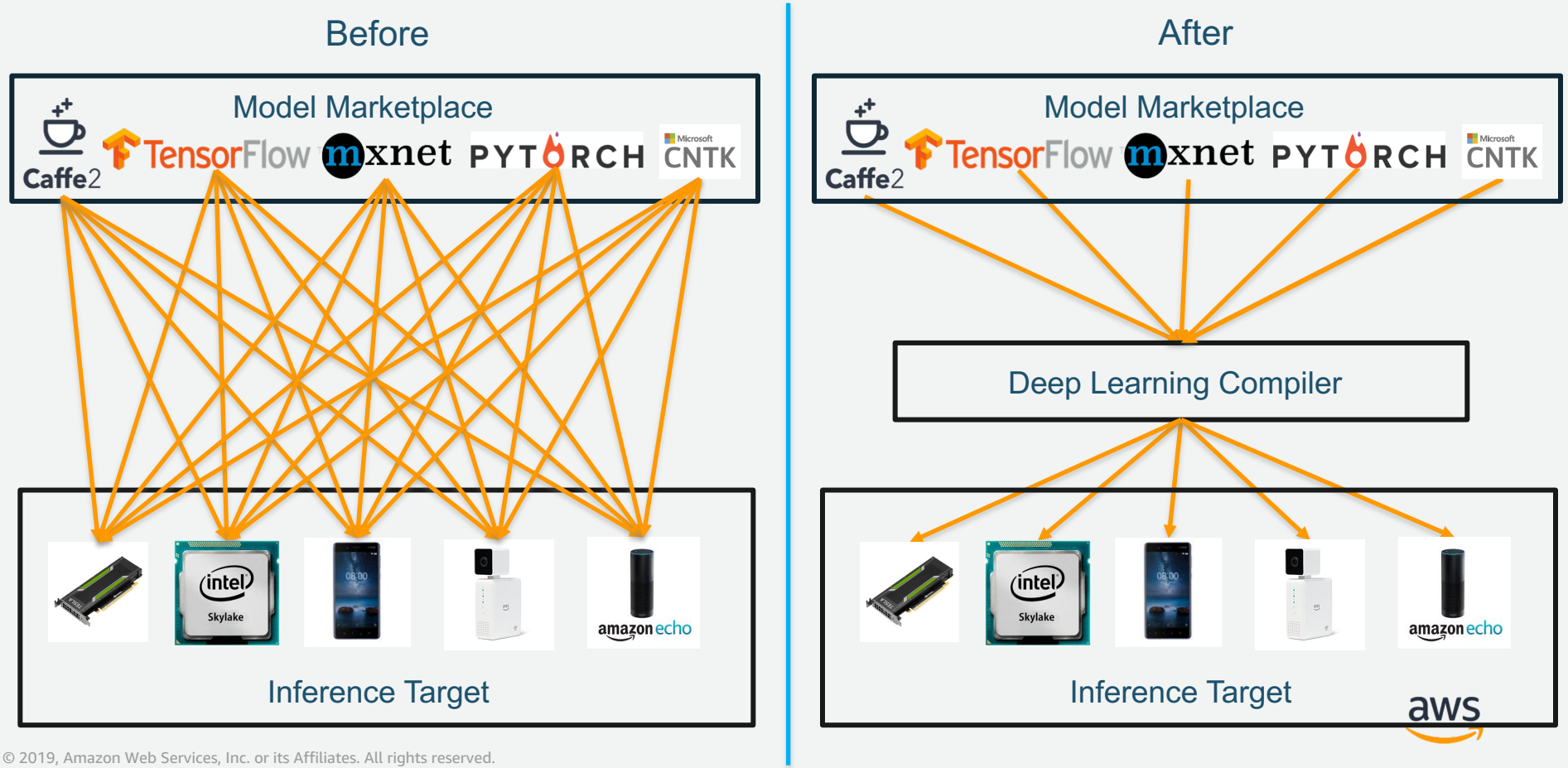
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Agenda

- **Overview of Neo**
- Relay Graph Optimizations
- TVM Tensor IR Optimizations
- Evaluation

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deep Learning Inference



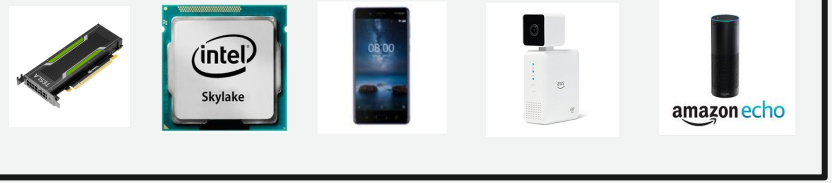
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Models and hardware targets are far away!

Caffe2 TensorFlow mxnet PYTORCH Microsoft CNTK

tvm

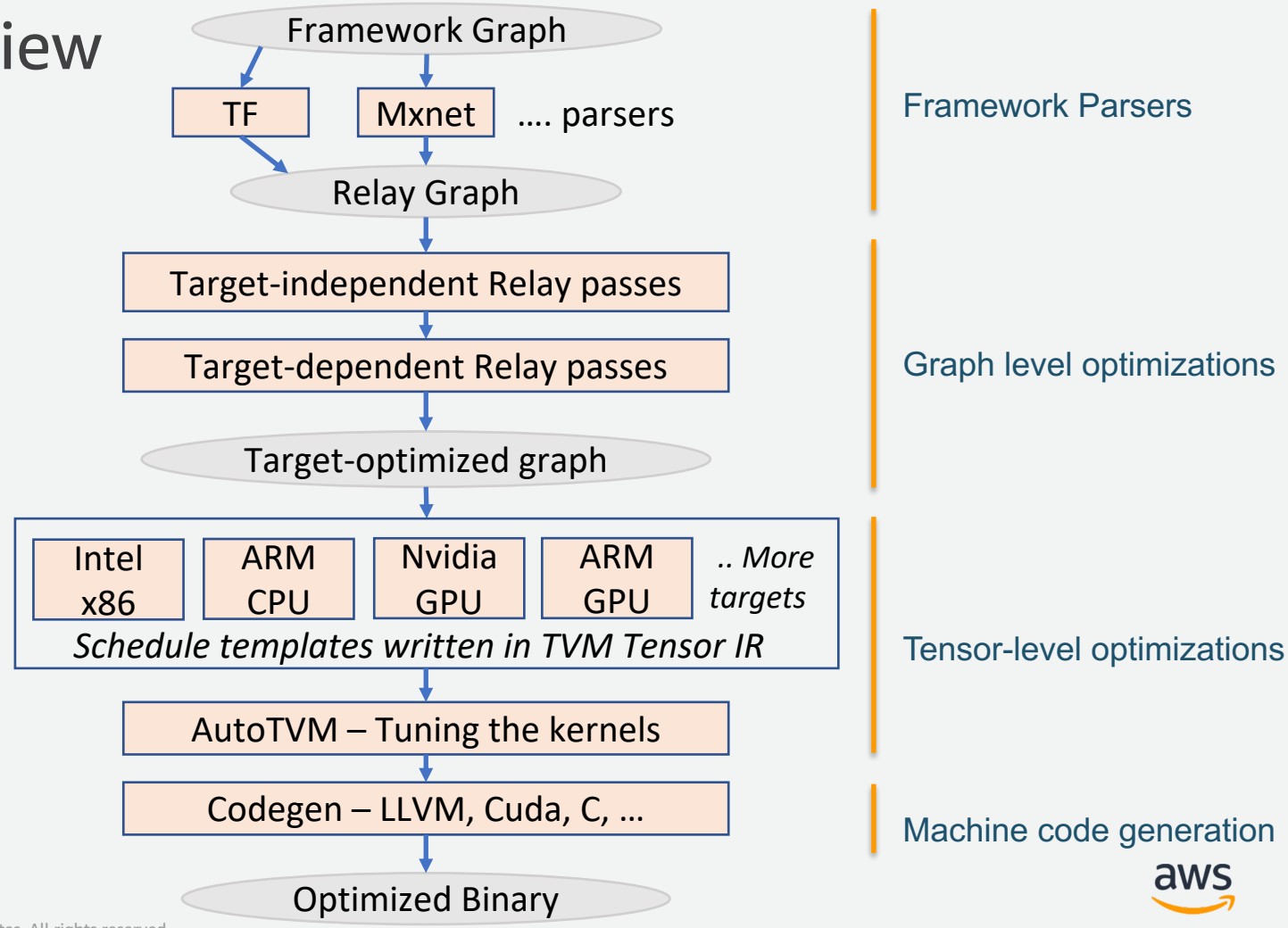
TVM: end-to-end optimization stack



aws

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TVM Overview



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

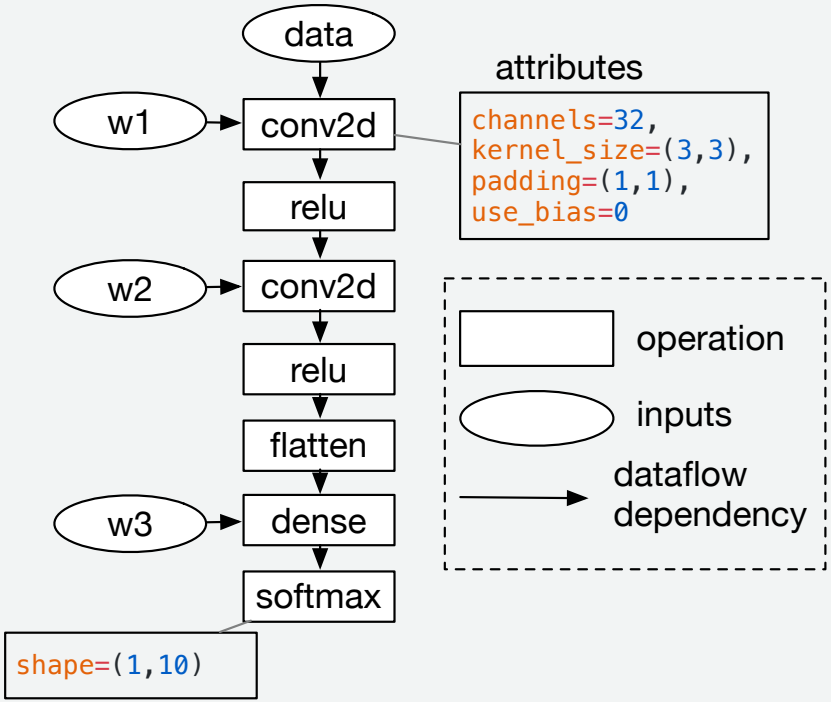
Agenda

- Overview of TVM
- **Relay Graph Optimizations**
- TVM Tensor IR Optimizations
- Evaluation

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Computation Graph Optimization

Represent high-level
deep learning computations



Target Independent

- Constant propagation
- Dead code elimination
- Operation fusion

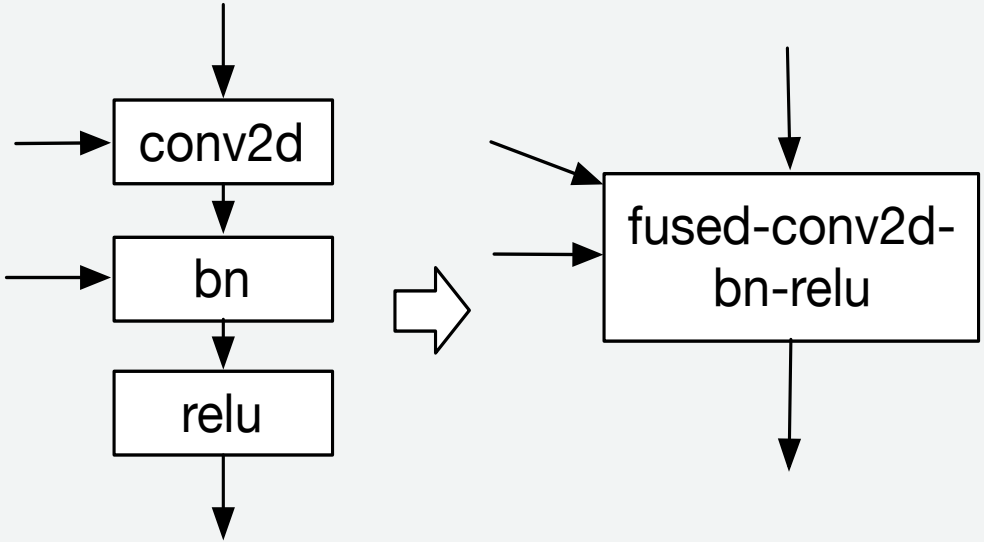
Target Dependent

- Data layout transform
- Graph partitioning
- Legalization

aws

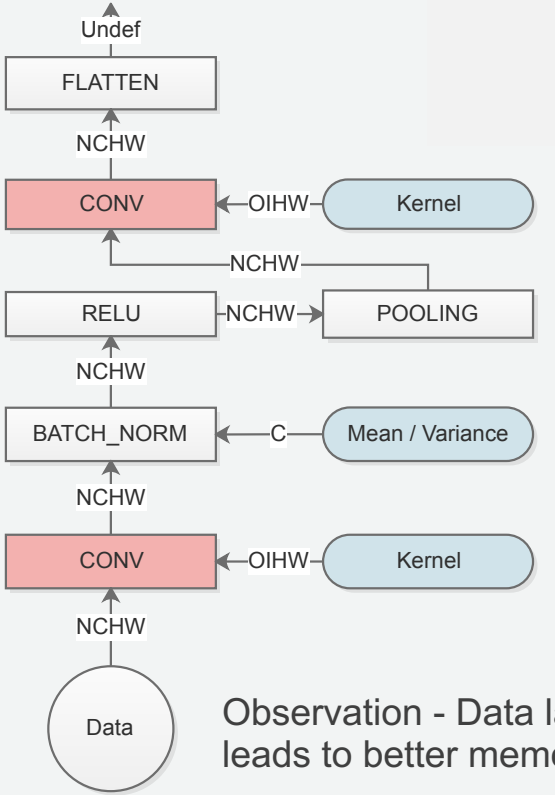
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Operation Fusion Example



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Target Dependent Layout Transformation



Observation - Data layout NCHWc leads to better memory accesses

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Agenda

- Overview of TVM
- Relay Graph Optimizations
- **TVM Tensor IR Optimizations**
- Evaluation

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TVM Tensor IR

- Compute definition

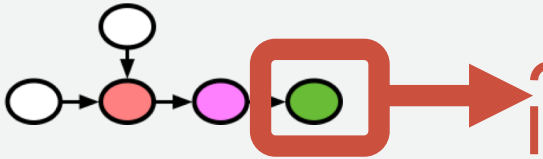
```
C = tvm.compute((m, n),
    lambda i, j: tvm.sum(A[i, k] * B[k, j], axis=k))
```

- TVM Schedule – *Developer-friendly loop transformations*
  - *Do not need hardware ISA knowledge to perform loop optimizations*

```
s = tvm.create_schedule(C.op)
➢ xo, yo, xi, yi = s[C].tile(C.op.axis[0], C.op.axis[1], bn, bn)
➢ ko, ki = s[C].split(k, factor=4)
➢ s[C].reorder(ko, xi, ki, yi)
➢ s[C].unroll(ki)
➢ s[C].vectorize(yi)
➢ s[C].parallel(xo)
...
```

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Large Search Space for Schedule



Compute Description

```
C = tvm.compute((m, n),
    lambda y, x: tvm.sum(A[k, y] * B[k, x], axis=k))
```

Large search space for optimization choices

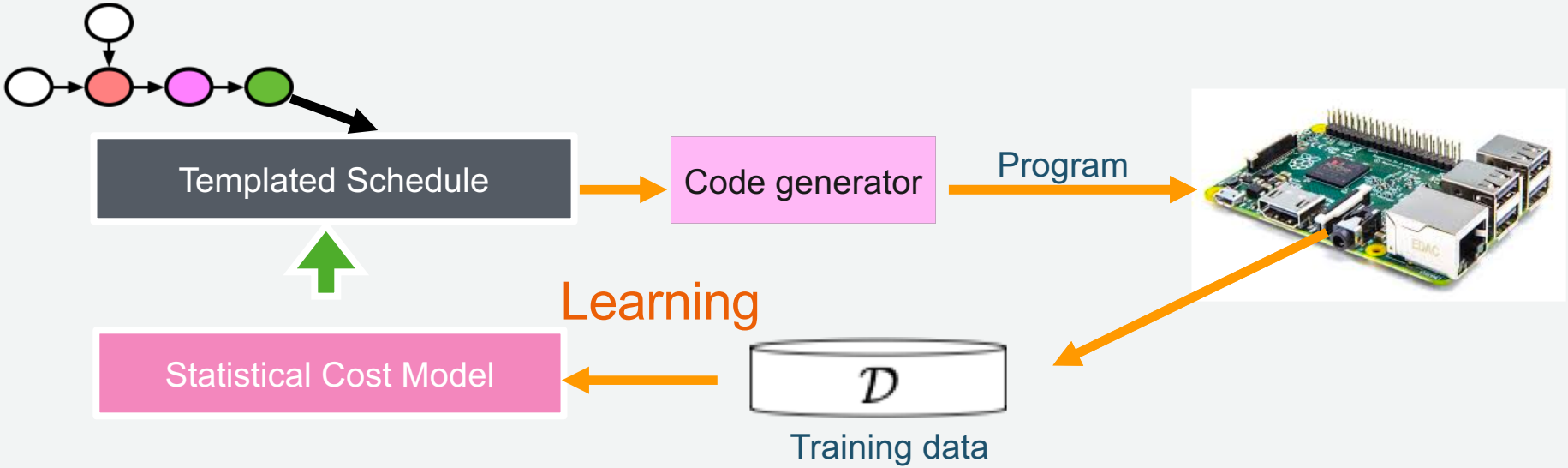
- Loop Transformations
- Thread Bindings
- Cache Locality
- Thread Cooperation
- Tensorization
- Latency Hiding

Hardware

aws

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AutoTVM - Learning-based Program Optimizer



- Relatively low experiment cost
- Domain-specific problem structure
- Large quantity of similar tasks

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Agenda

- Overview of TVM
- Relay Graph Optimizations
- TVM Tensor IR Optimizations
- **Evaluation**

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

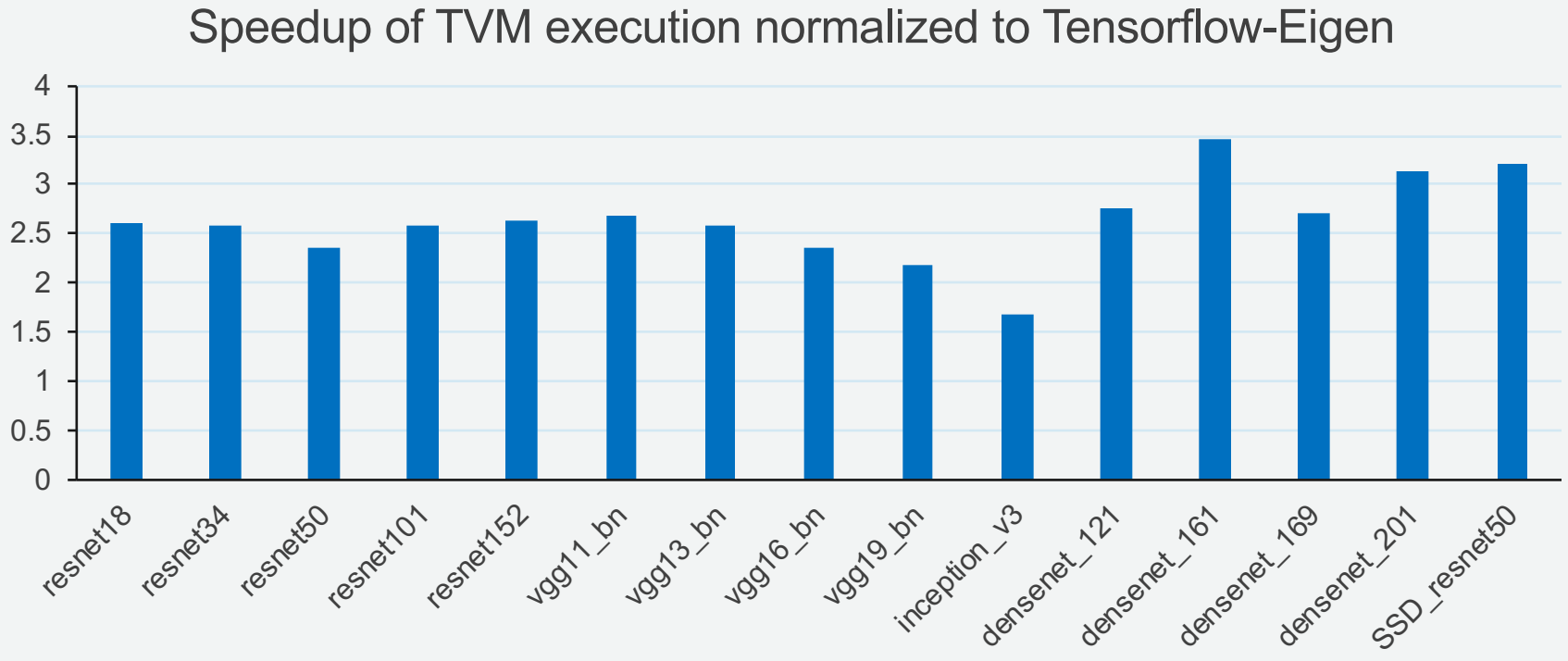
Experiment Setup

- Server - EC2 A1 Instance
  - 16-core ARMv8

- Edge device - Acer aiSage
  - Rockchip RK3399 SoC + ARM Mali GPU T-860

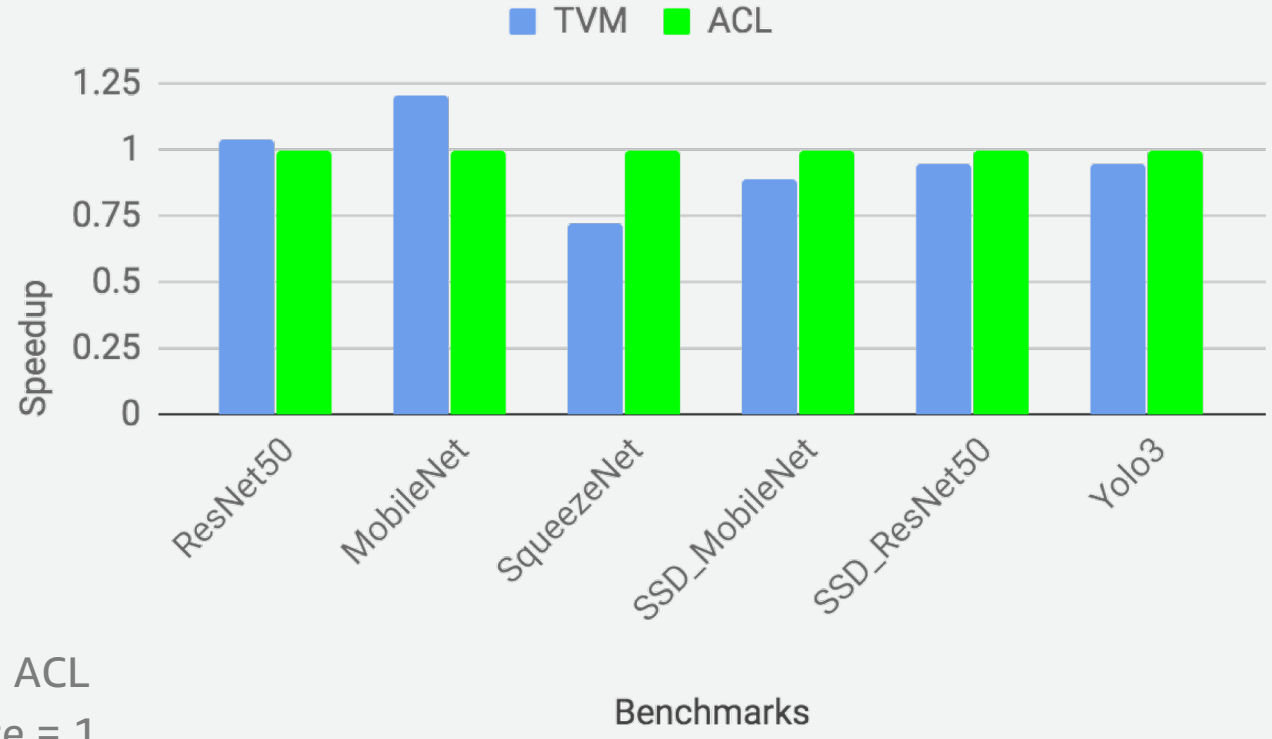
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TVM – Evaluation on ARM A1 Server



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TVM – Evaluation on Edge Device Acer aiSage



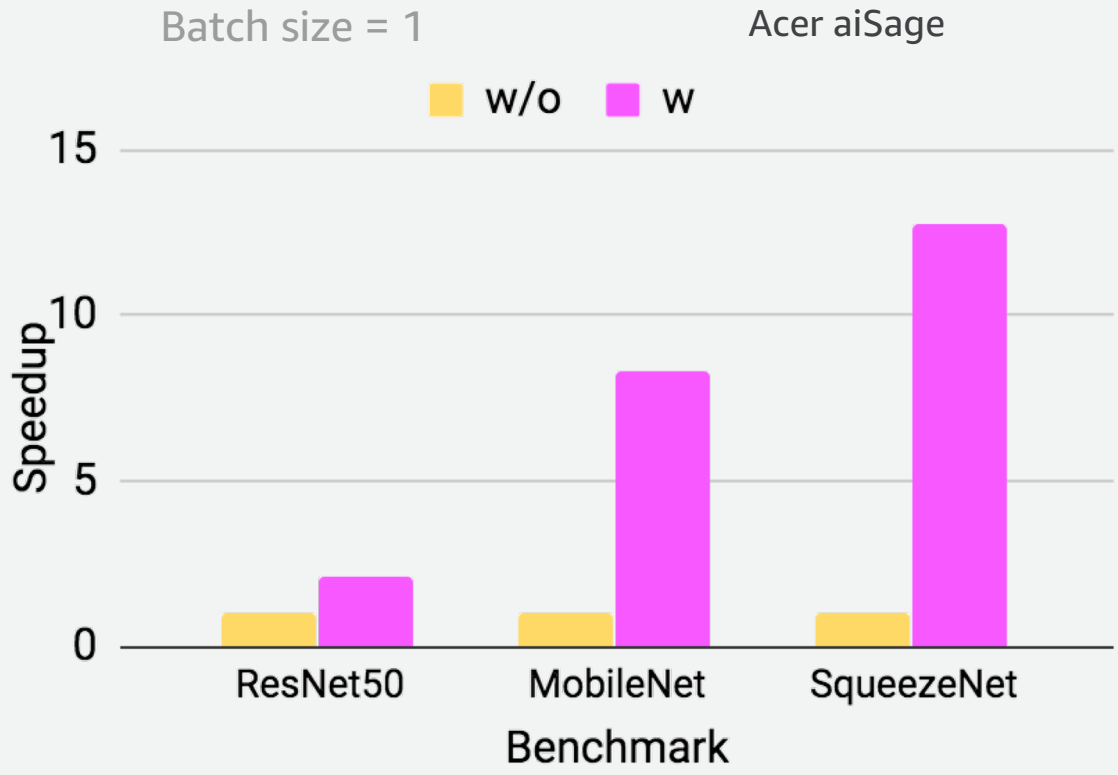
Baseline: ACL

Batch size = 1

ICPP'19 A Unified Optimization Approach for CNN Model Inference on Integrated GPUs

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Effects of Tuning Convolution operators in TVM



ICPP'19 A Unified Optimization Approach for CNN Model Inference on Integrated GPUs

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Takeaways

- Deep learning compilation is essential for portability and performance across a variety of targets.
- Optimizations are important at all levels – graph- and tensor-level.
- Abstracting compute and HW-dependent schedule enables developers to write kernels without extensive knowledge of HW ISA.
- Open-source collaborations are the key to achieve the dream of running deep learning everywhere.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Linaro Community

- TVM enjoys LLVM ARM codegen support
  - Better support for Int8 instructions
  - Better support for different ARM variants
- Better schedules
  - Data layout optimizations are hardware dependent
  - TVM performance of ARM GPUs can be improved
- ACL support for TVM

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Thank you!

Check it out!

Amazon
SageMaker Neo

https://github.com/neo-ai/

https://github.com/dmlc/tvm

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.