

TVM: End to End Deep Learning Compiler Stack

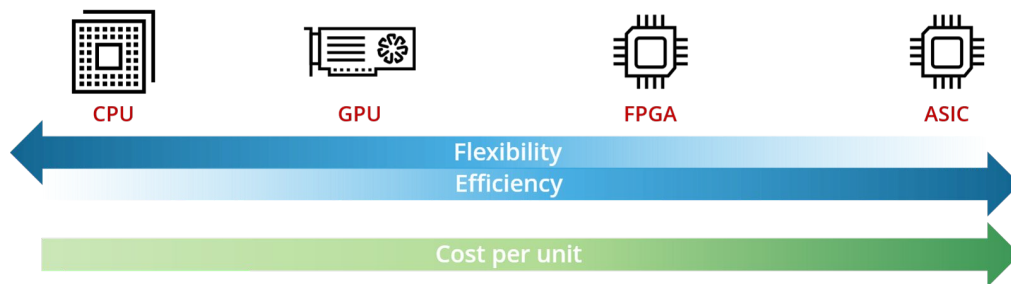
Alina Egorova
Arathy Ajaya Kumar

The challenges that deep learning is facing today



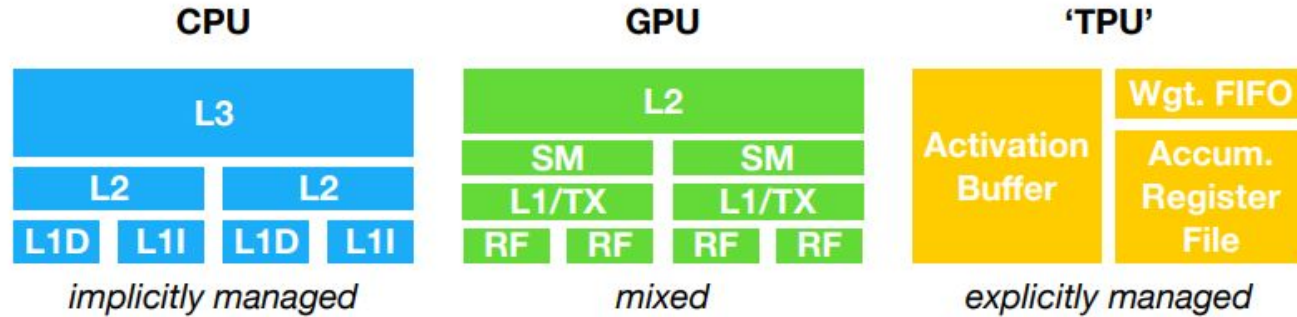
CPU, GPU, FPGA, and ASICs

Tradeoffs

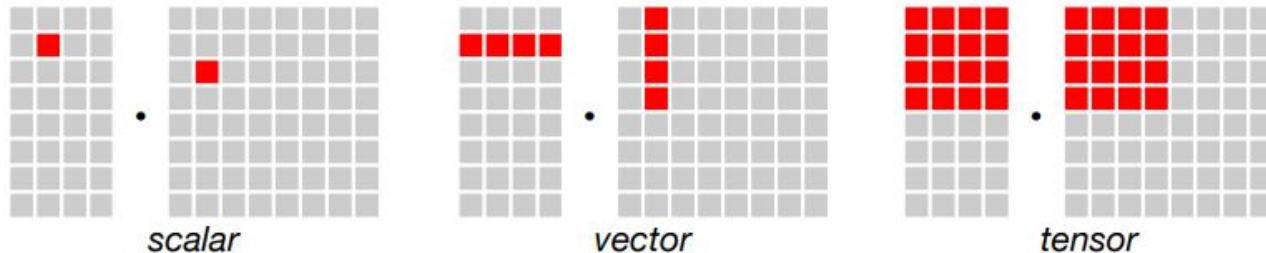


The challenges that deep learning is facing today

Memory Subsystem Architecture



Compute Primitive



TVM: End to End Optimization Stack



Computational Graph Optimization

Tensor Expression Language

Primitives in prior works
Halide, Loopy

Loop
Transformations

Thread
Bindings

Cache
Locality

New primitives for GPU
Accelerators

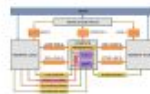
Thread
Cooperation

Tensorization

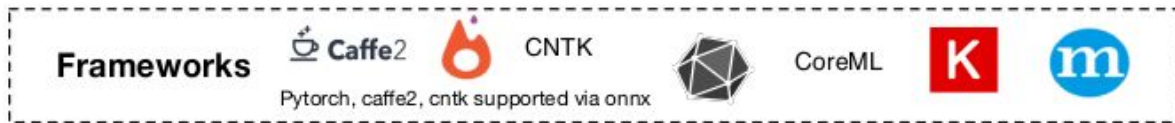
Latency
Hiding



Hardware



TVM: End to End Optimization Stack

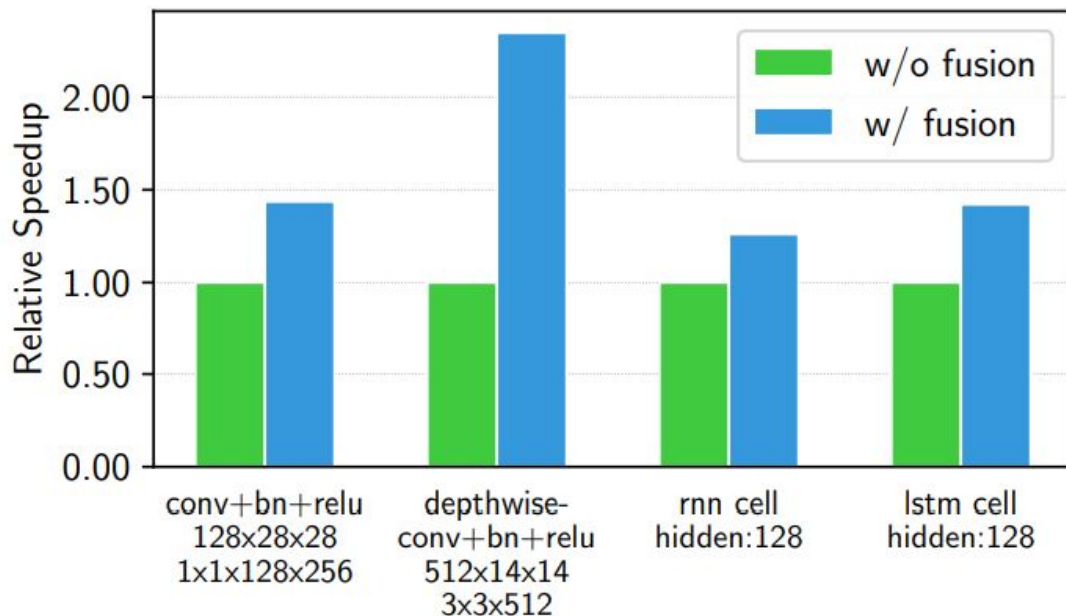


Computational Graph Optimization

- operator fusion
- constant-folding
- static memory planning pass
- data layout transformations

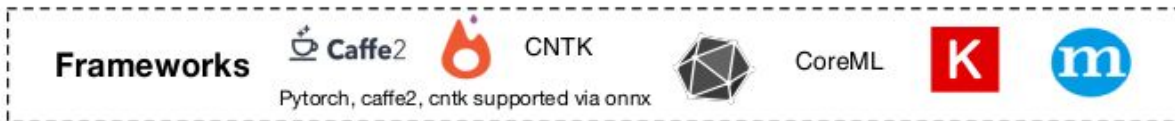


Operator fusion



Performance comparison between fused and non-fused operations. TVM generates both operations. Tested on NVIDIA Titan X

TVM: End to End Optimization Stack



Computational Graph Optimization

Tensor Expression Language

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

Schedule Optimizations

Hardware



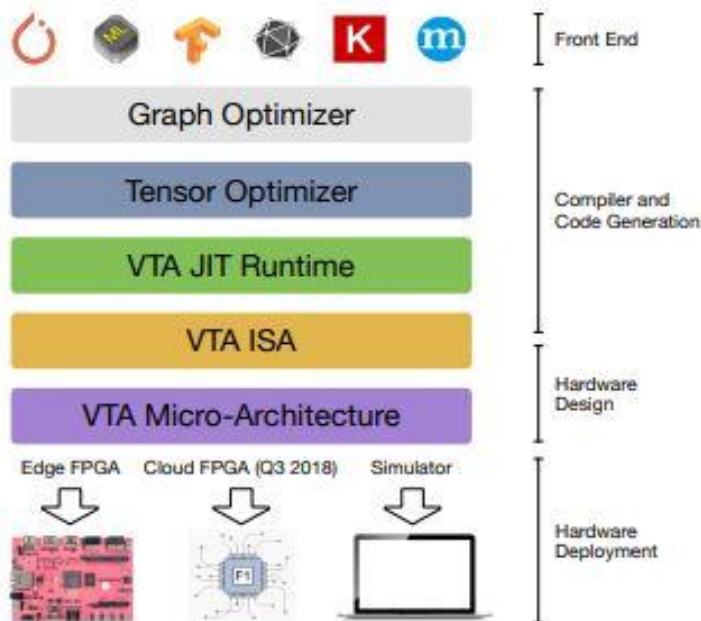
Schedule Optimizations

| Schedule primitives used in various hardware backends | CPU Schedule | GPU Schedule | Accel. Schedule |
|--|-----------------|-----------------|--------------------|
| [Halide] Loop Transformations | ✓ | ✓ | ✓ |
| [Halide] Thread Binding | ✓ | ✓ | ✓ |
| [Halide] Compute Locality | ✓ | ✓ | ✓ |
| [TVM] Special Memory Scope | | ✓ | ✓ |
| [TVM] Tensorization | ✓ | ✓ | ✓ |
| [TVM] Latency Hiding | | | ✓ |

Versatile tensor accelerator (VTA)

- An extension of TVM stack
- Exposes salient features of deep learning accelerators.
- Provides an open deep learning system stack for optimizations.

VTA stack overview



VTA stack overview

1. NNVM (Neural network virtual machine)Intermediate Representation
2. TVM Intermediate Representation
3. VTA JIT Runtime
4. VTA Instruction Set Architecture
 - high-level CISC ISA
 - a low-level, and fixed latency RISC ISA
5. VTA Hardware Micro-Architecture