PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

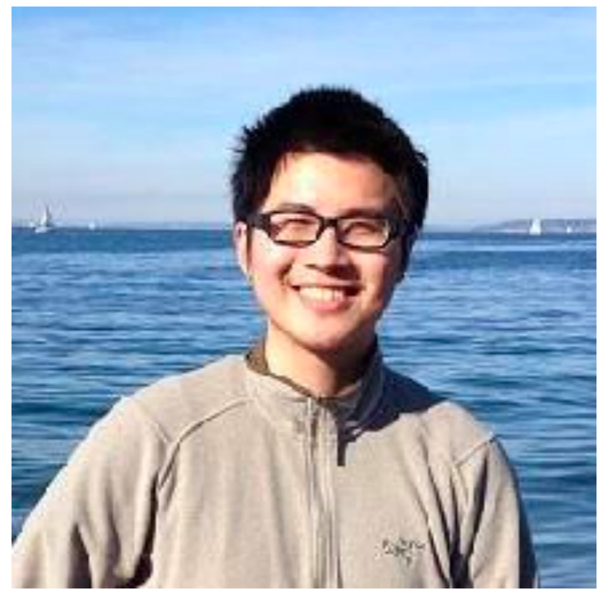
End to End Optimization Stack for Deep Learning

Presenter: Tianqi Chen

Paul G. Allen School of Computer Science & Engineering
University of Washington

Collaborators

University of Washington

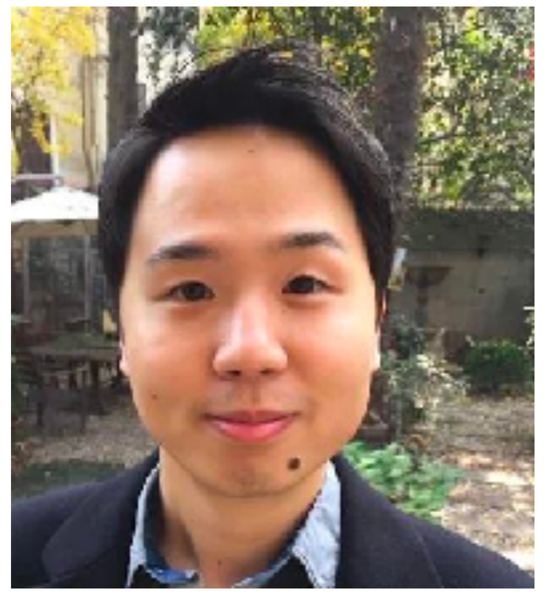


Tianqi Chen

ML, Software Stack

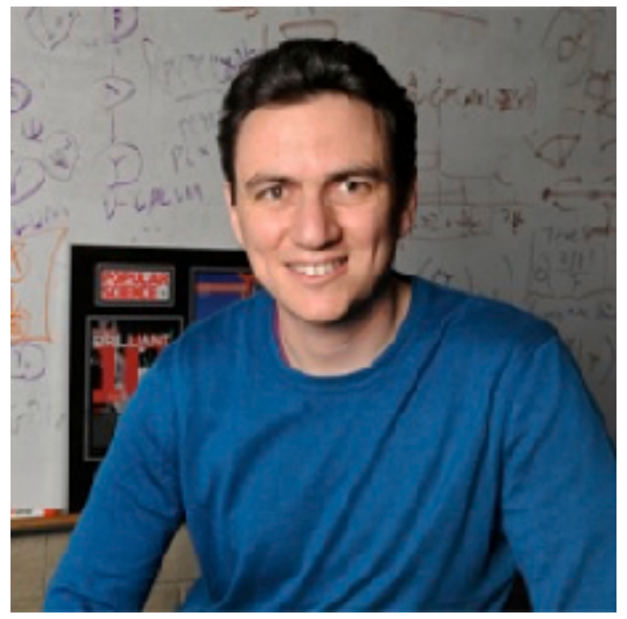
Thierry Moreau

Hardware Stack



Haichen Shen

GPU

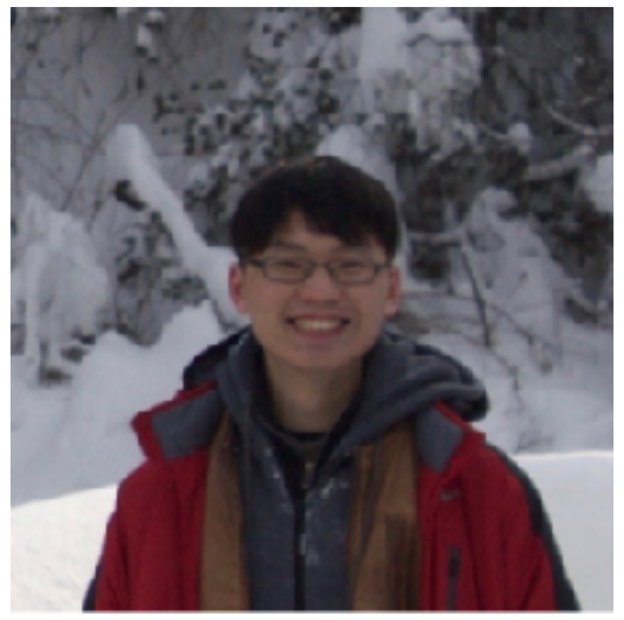


Carlos Guestrin

Luis Ceze

Arvind Krishnamurthy

AWS AI Team



Ziheng Jiang

ARM, NNVM pipeline

and many more contributors in the **DMLC** community

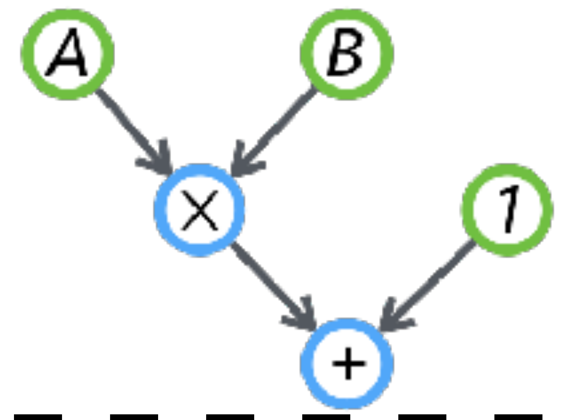
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

Caffe2

Computational graph



Operator Libraries

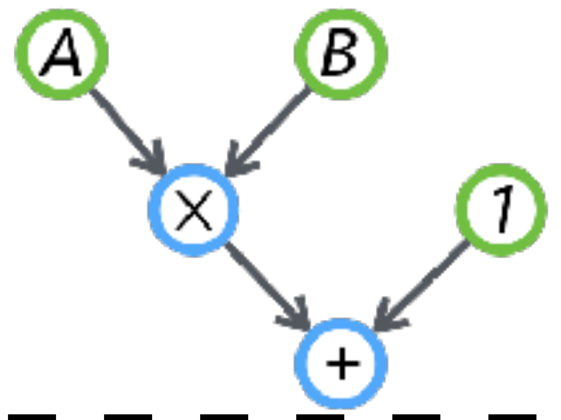
cuDNN, NNPack, MKL-DNN

Hardware

Deep Learning System Research is Exciting but Hard

Frameworks CNTK Caffe2

Computational graph


$$\begin{tikzcd}
A \arrow[dr] & & B \arrow[dl] & \\
& \times \arrow[dr] & & 1 \arrow[dl] \\
& & + &
\end{tikzcd}$$

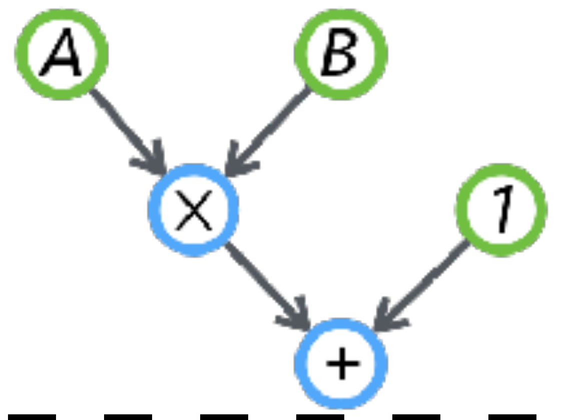
Operator Libraries cuDNN, NNPack, MKL-DNN

Hardware

Deep Learning System Research is Exciting but Hard

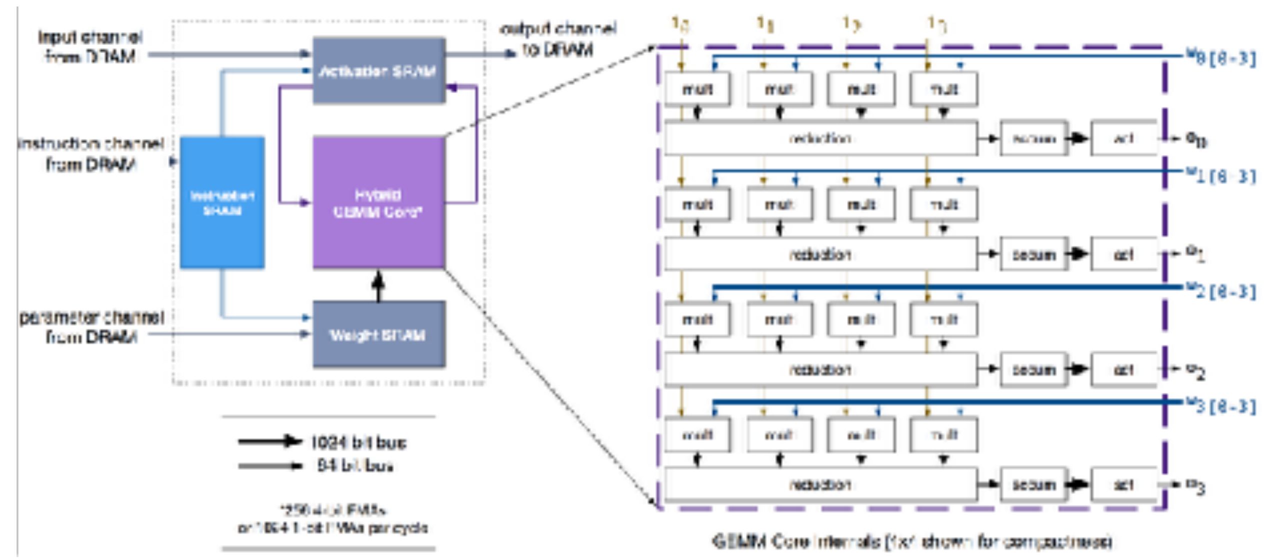
Frameworks CNTK Caffe2

Computational graph



Operator Libraries cuDNN, NNPack, MKL-DNN

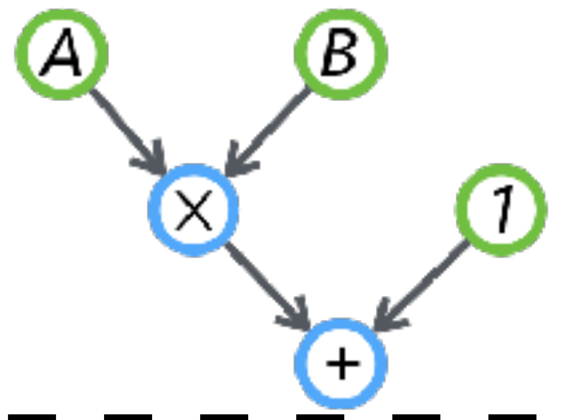
Hardware



Deep Learning System Research is Exciting but Hard

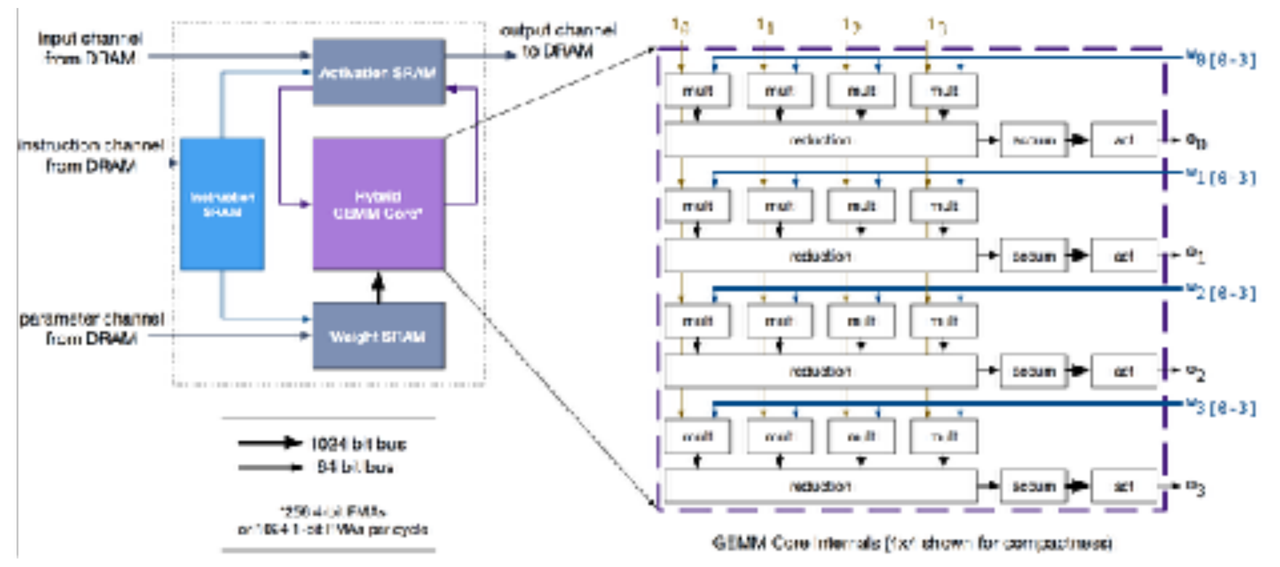
Frameworks CNTK Caffe2

Computational graph



Operator Libraries **cuDNN, NNPack, MKL-DNN**

Hardware



Built a new accelerator

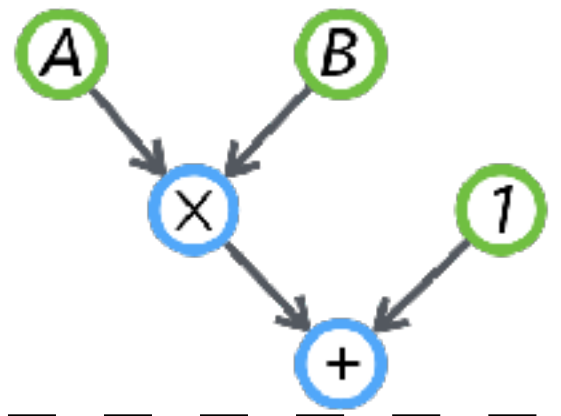
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

Caffe2

Computational graph

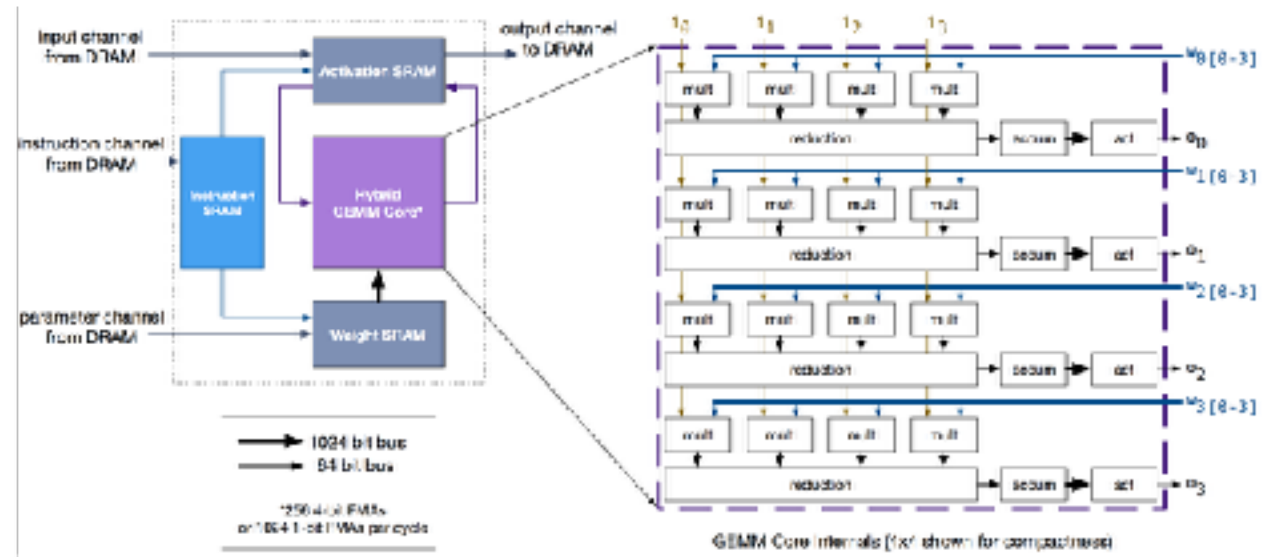


Operator Libraries **cuDNN, NNPack, MKL-DNN**

Hardware

Need entire software stack on top of it!

Layout transformation
Quantization
Operator kernel optimization
Benchmarking
….



Built a new accelerator

Deep Learning System Research is Exciting but Hard

Frameworks CNTK Caffe2

Computational graph

$$\begin{tikzcd}
A \arrow[dr] & & B \arrow[dl] & \\
& \times \arrow[dr] & & 1 \arrow[dl] \\
& & + &
\end{tikzcd}$$

Operator Libraries **cuDNN, NNPack, MKL-DNN**

Hardware

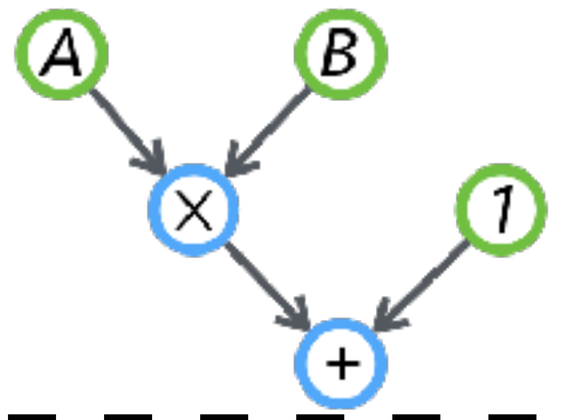
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

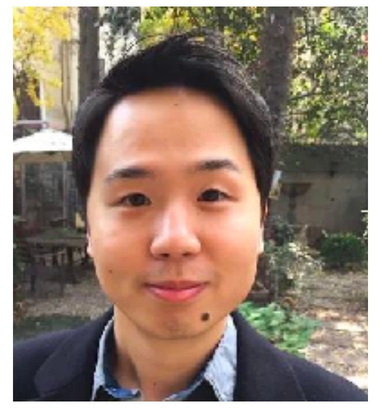
Caffe2

Computational graph



Operator Libraries cuDNN, NNPack, MKL-DNN

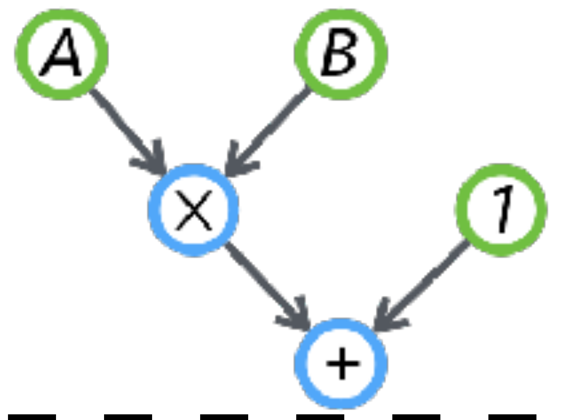
Hardware



Deep Learning System Research is Exciting but Hard

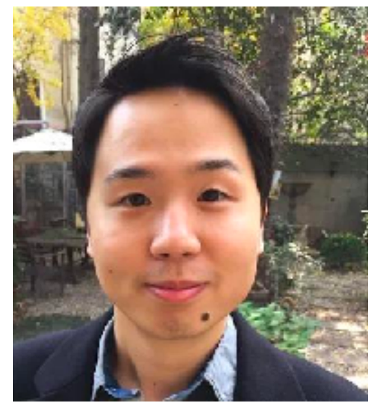
Frameworks — CNTK, Caffe2

Computational graph



Operator Libraries — cuDNN, NNPack, MKL-DNN

Hardware



Data Layout Optimization

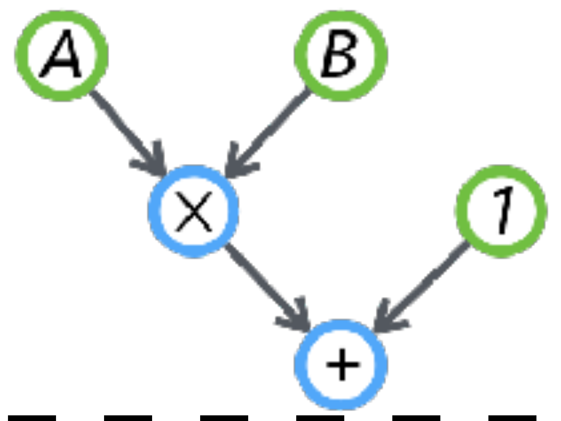
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

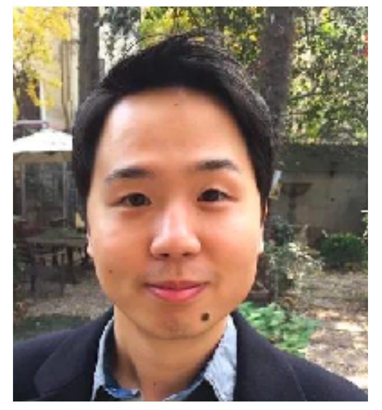
Caffe2

Computational graph



Operator Libraries cuDNN, NNPack, MKL-DNN

Hardware



Data Layout Optimization

Operator Fusion

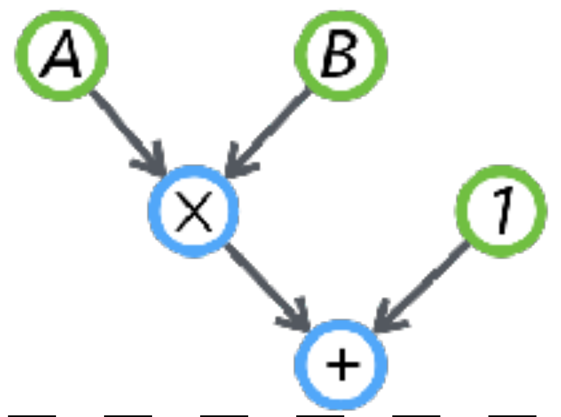
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

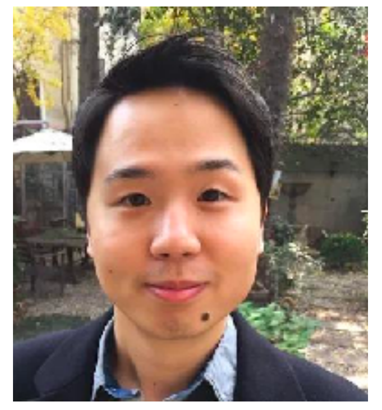
Caffe2

Computational graph



Operator Libraries cuDNN, NNPack, MKL-DNN

Hardware



Data Layout Optimization

Operator Fusion

Need optimized hardware kernel for each variant, on each hardware!

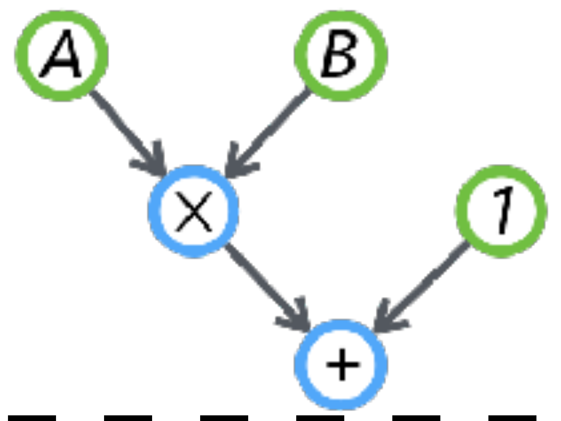
Deep Learning System Research is Exciting but Hard

Frameworks

CNTK

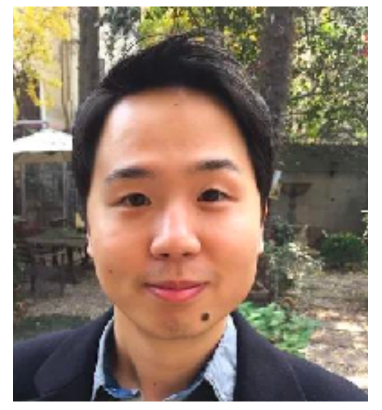
Caffe2

Computational graph



Operator Libraries cuDNN, NNPack, MKL-DNN

Hardware



Data Layout Optimization

Operator Fusion

Serving

Need optimized hardware kernel for each variant, on each hardware!

The End to End System Challenge

Frameworks

Hardware
Back-Ends

The End to End System Challenge

Frameworks

Hardware
Back-Ends

The End to End System Challenge

Frameworks

Hardware
Back-Ends

The End to End System Challenge

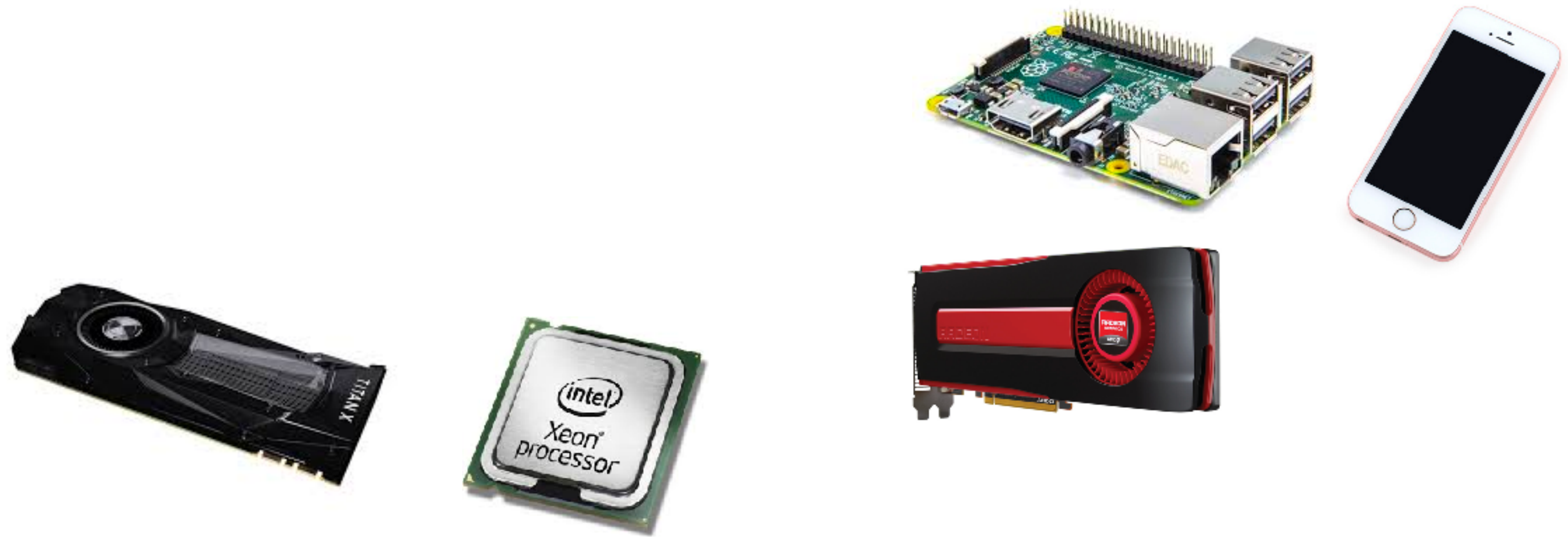
Frameworks

Hardware
Back-Ends

The End to End System Challenge

Frameworks

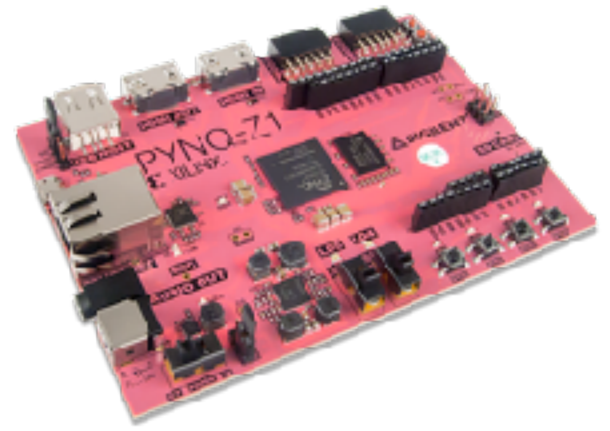
Hardware
Back-Ends



The End to End System Challenge

Frameworks

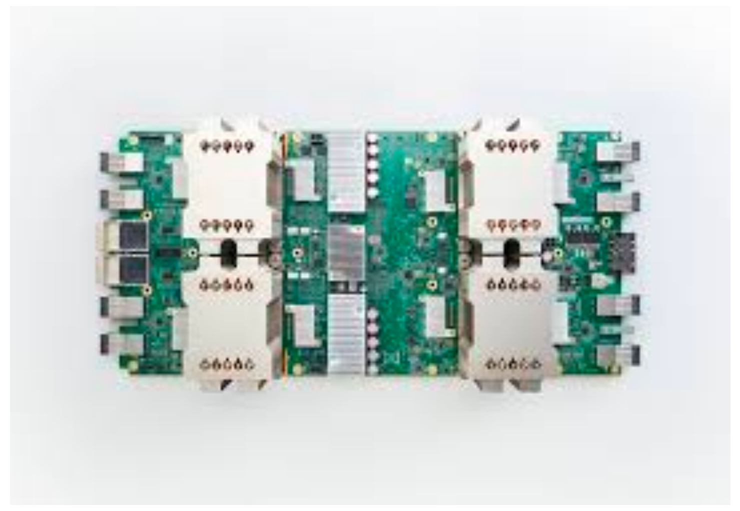
Hardware Back-Ends

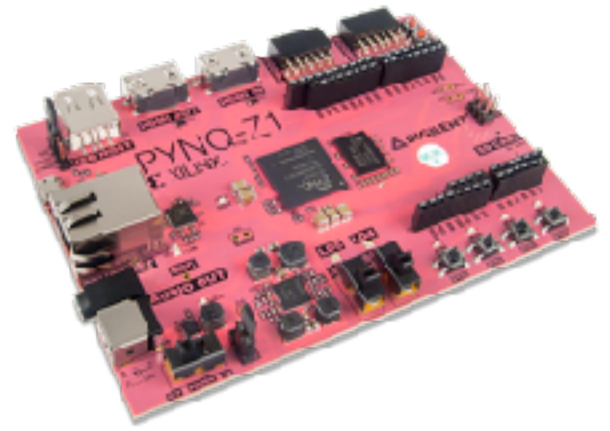


The End to End System Challenge

Frameworks

Hardware
Back-Ends





The End to End System Challenge

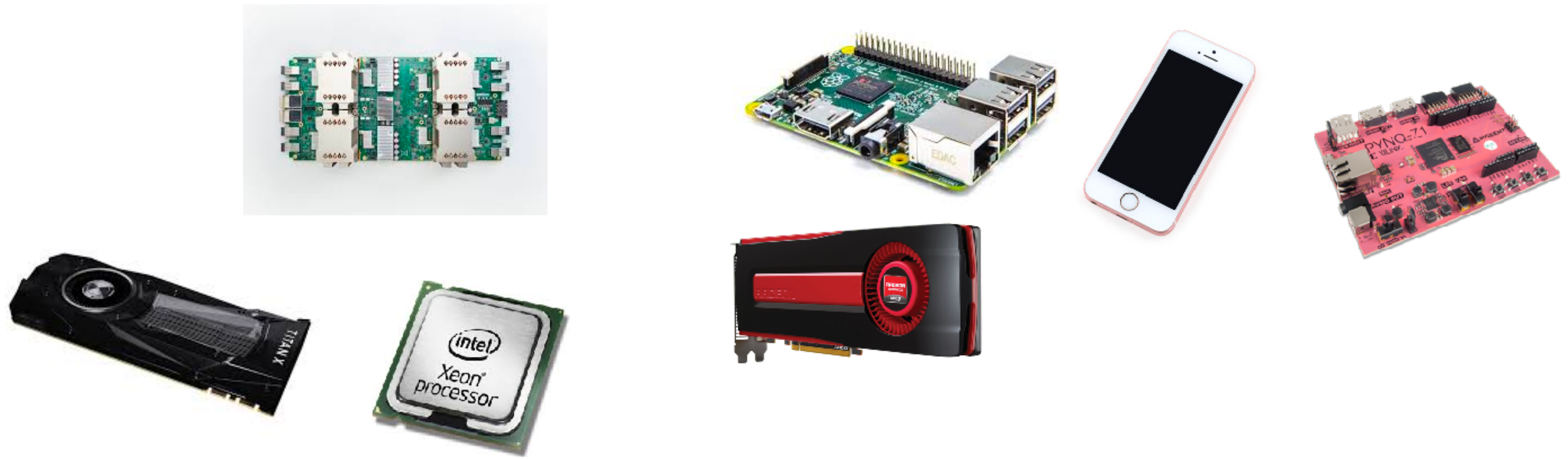
Frameworks

↓

Intermediate representation

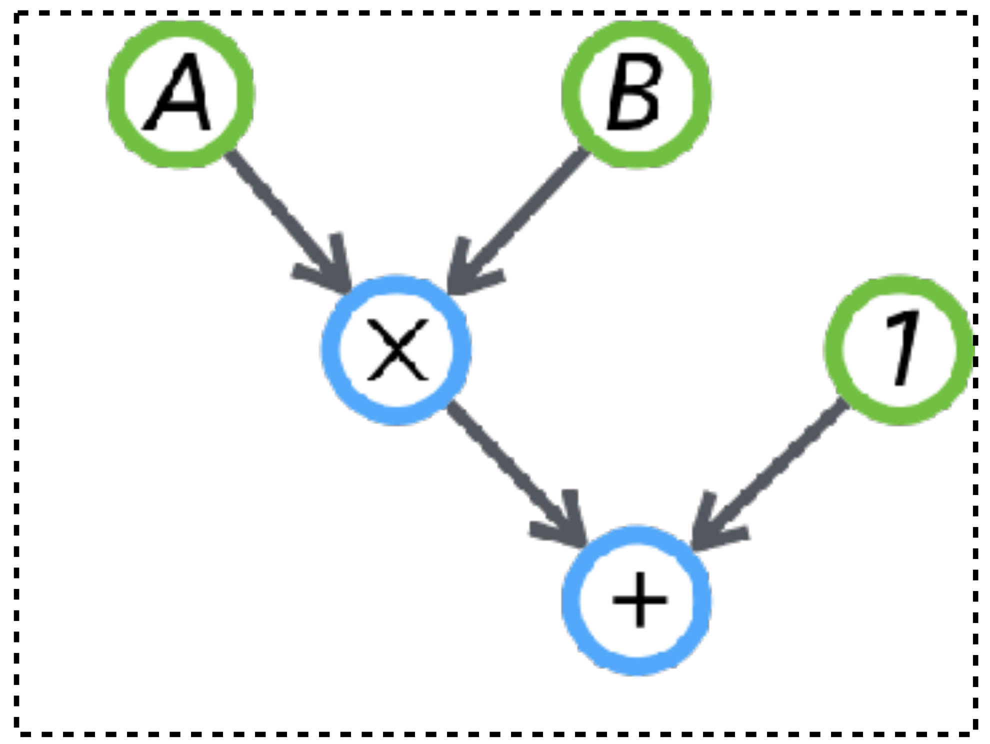
↓

Hardware
Back-Ends



Computational Graph IR and Remaining Gap

Examples: NGraph, XLA, NNVM, DLVM …



Computational Graph

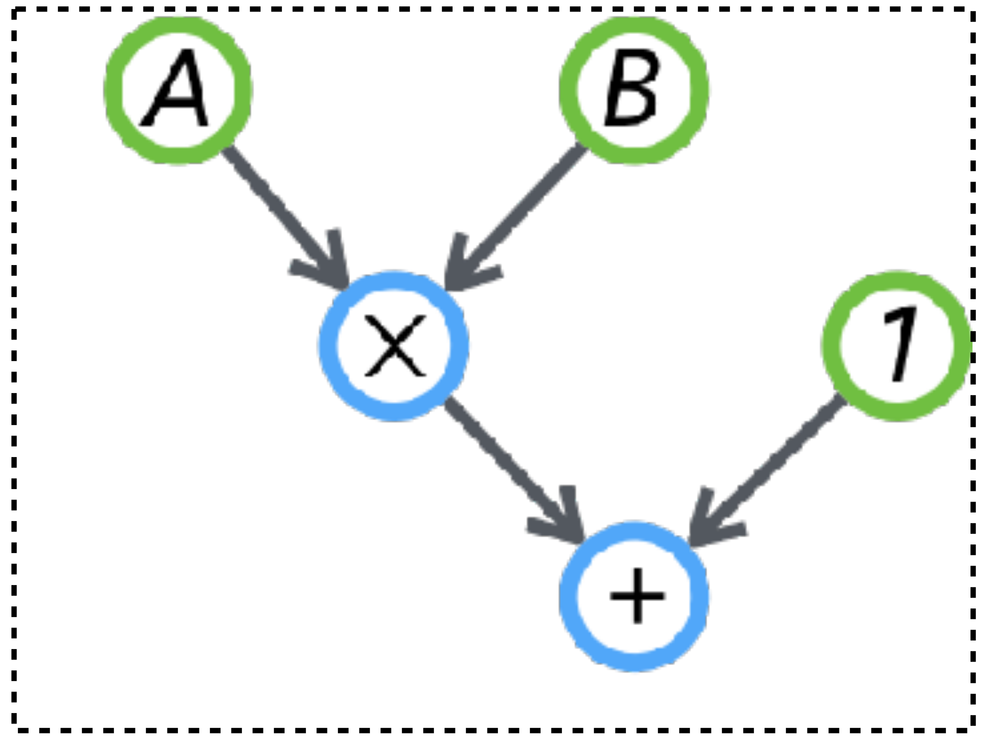
Auto Differentiation

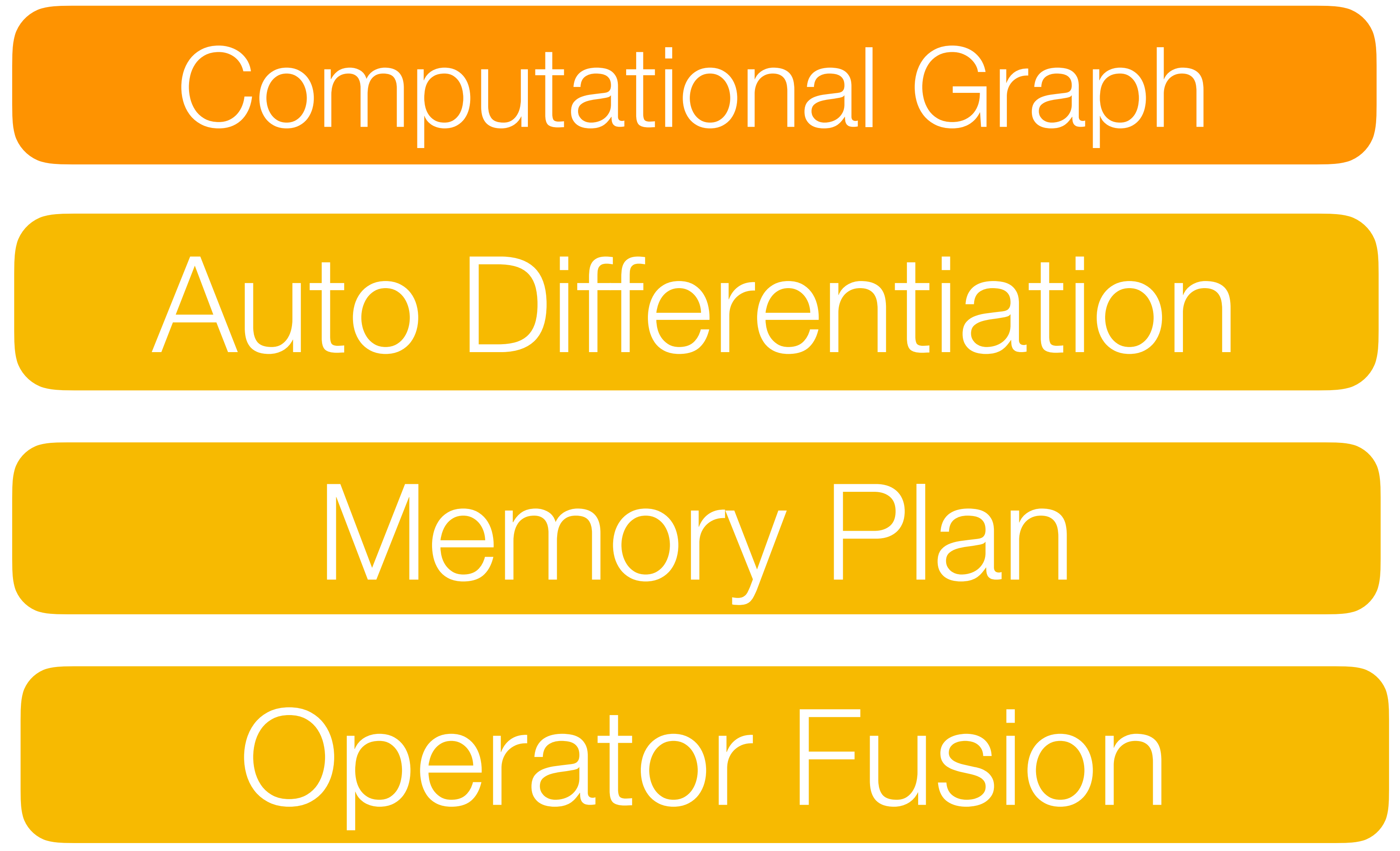
Memory Plan

Operator Fusion

Backends

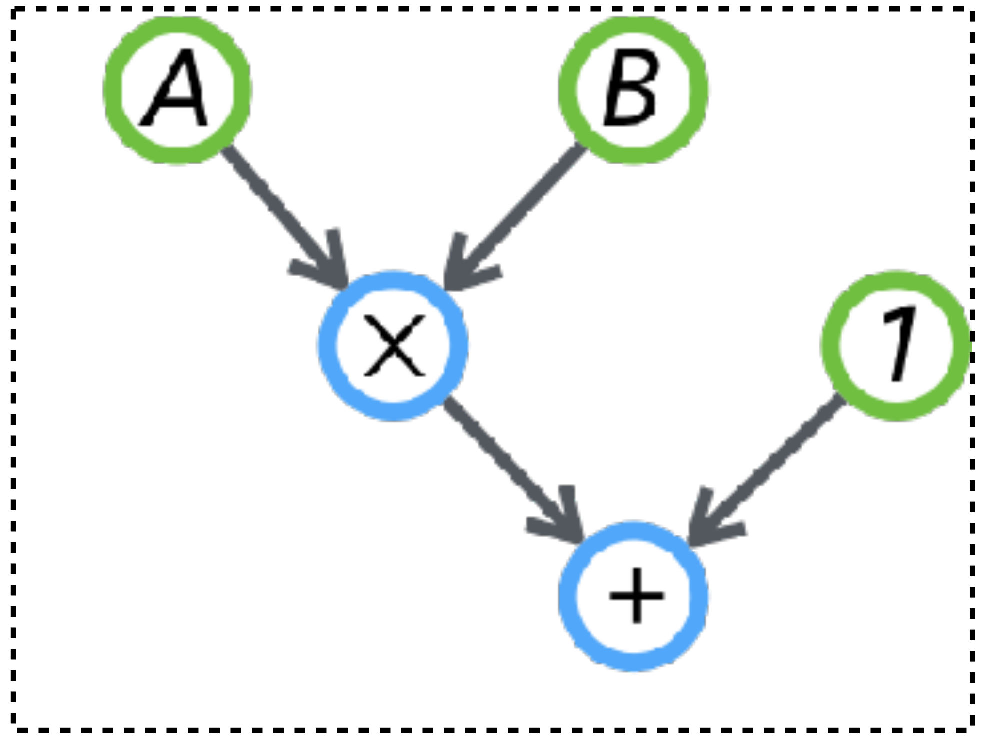
Computational Graph IR and Remaining Gap





Backends

Computational Graph IR and Remaining Gap



Computational Graph

Auto Differentiation

Memory Plan

Operator Fusion

too many possible choices:

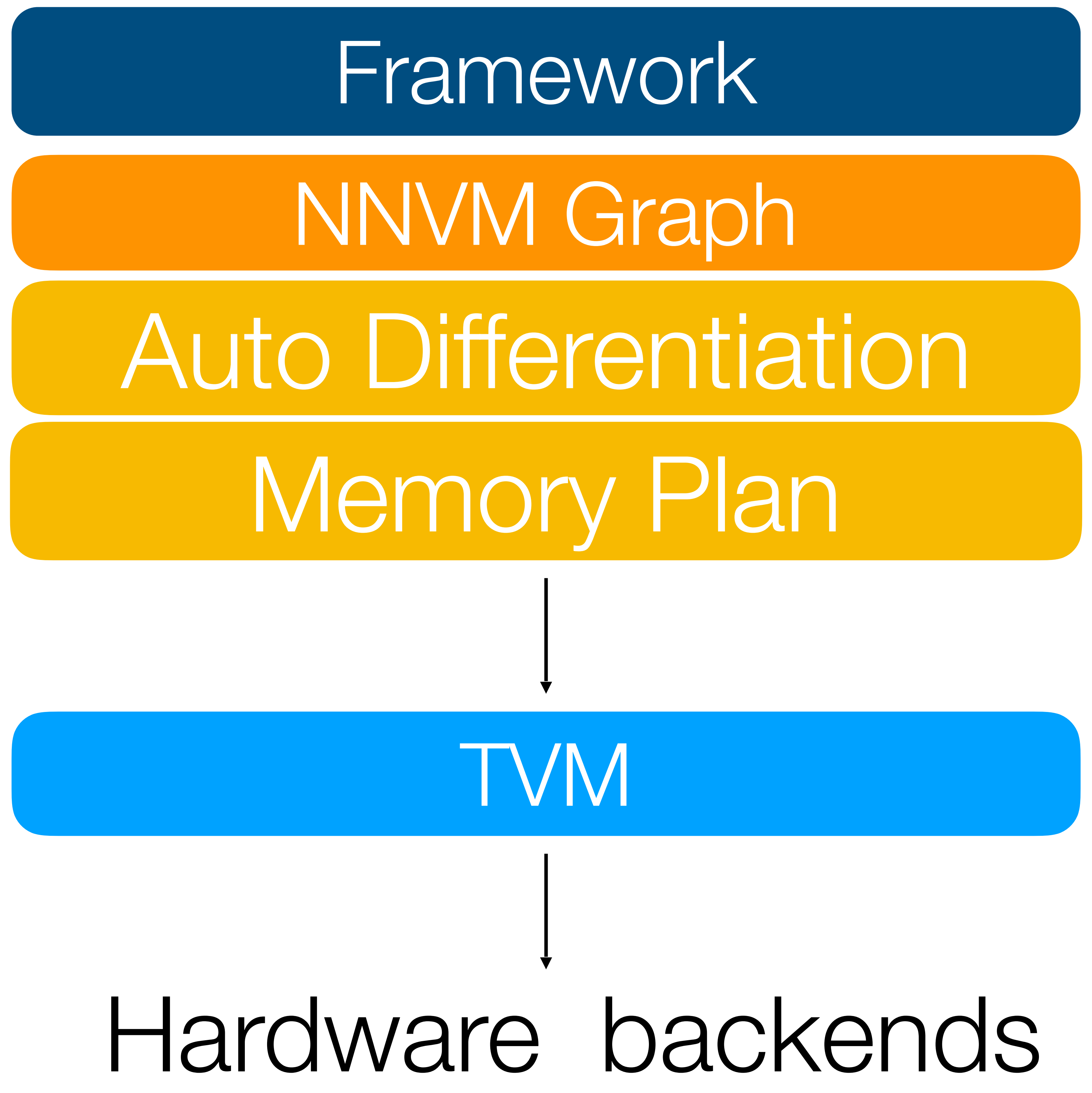
precision, layout, fused pattern, device, threading …

Need a low level IR to express them explicitly

Backends

TVM: Low Level IR

- Concise and compact description
- Explicit control on codegen
- Ease of deployment
- Support new hardware backends



Tensor Index Expression Declaration

Compute C = dot(A, B.T)

```
import tvm
m, n, h = tvm.var('m'), tvm.var('n'), tvm.var('h')
A = tvm.placeholder((m, h), name='A')
B = tvm.placeholder((n, h), name=‘B')
k = tvm.reduce_axis((0, h), name=‘k')
C = tvm.compute((m, n), lambda i, j: tvm.sum(A[i, k] * B[j, k], axis=k))
```

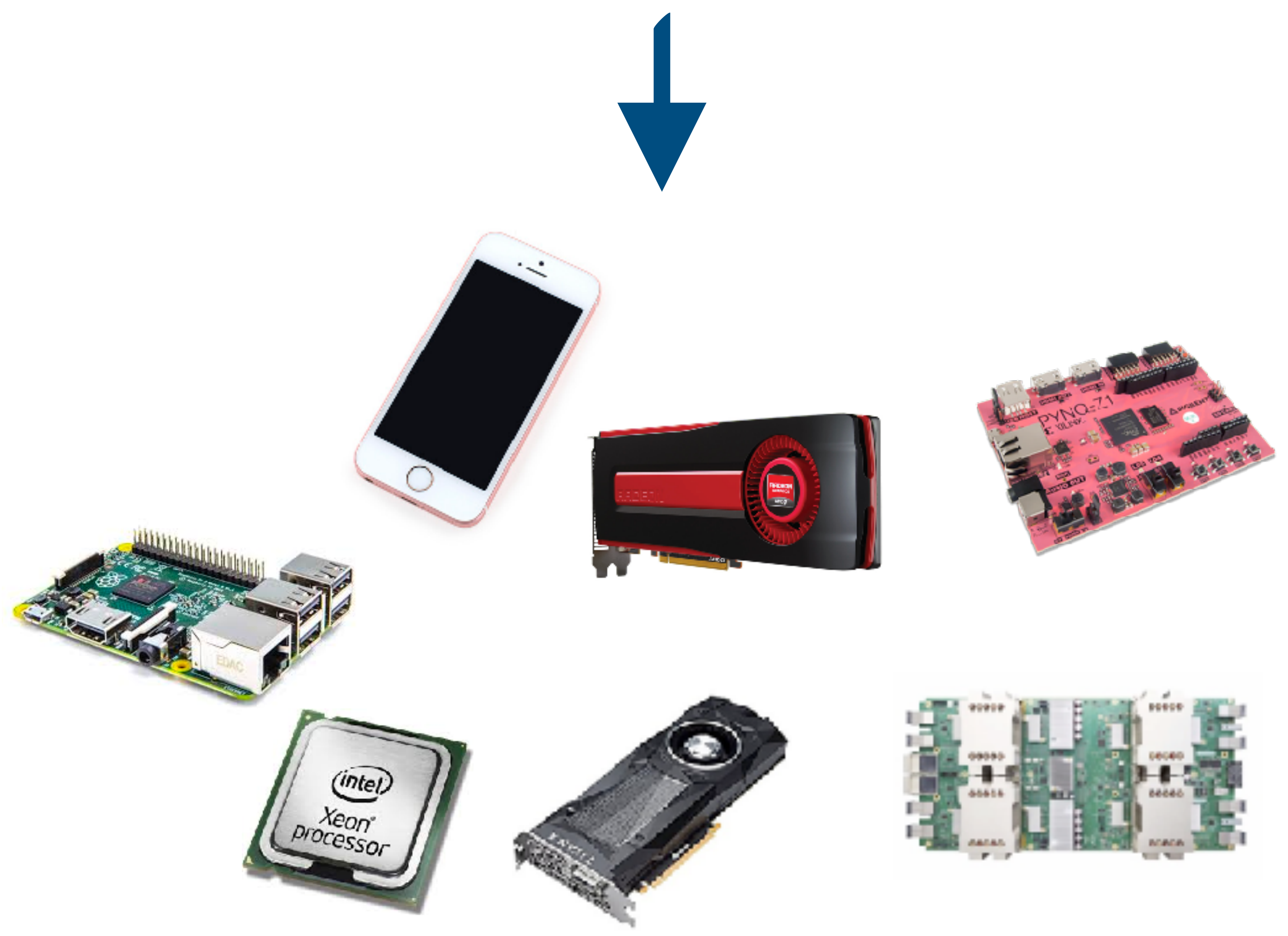
Inputs

Shape of C

Computation Rule

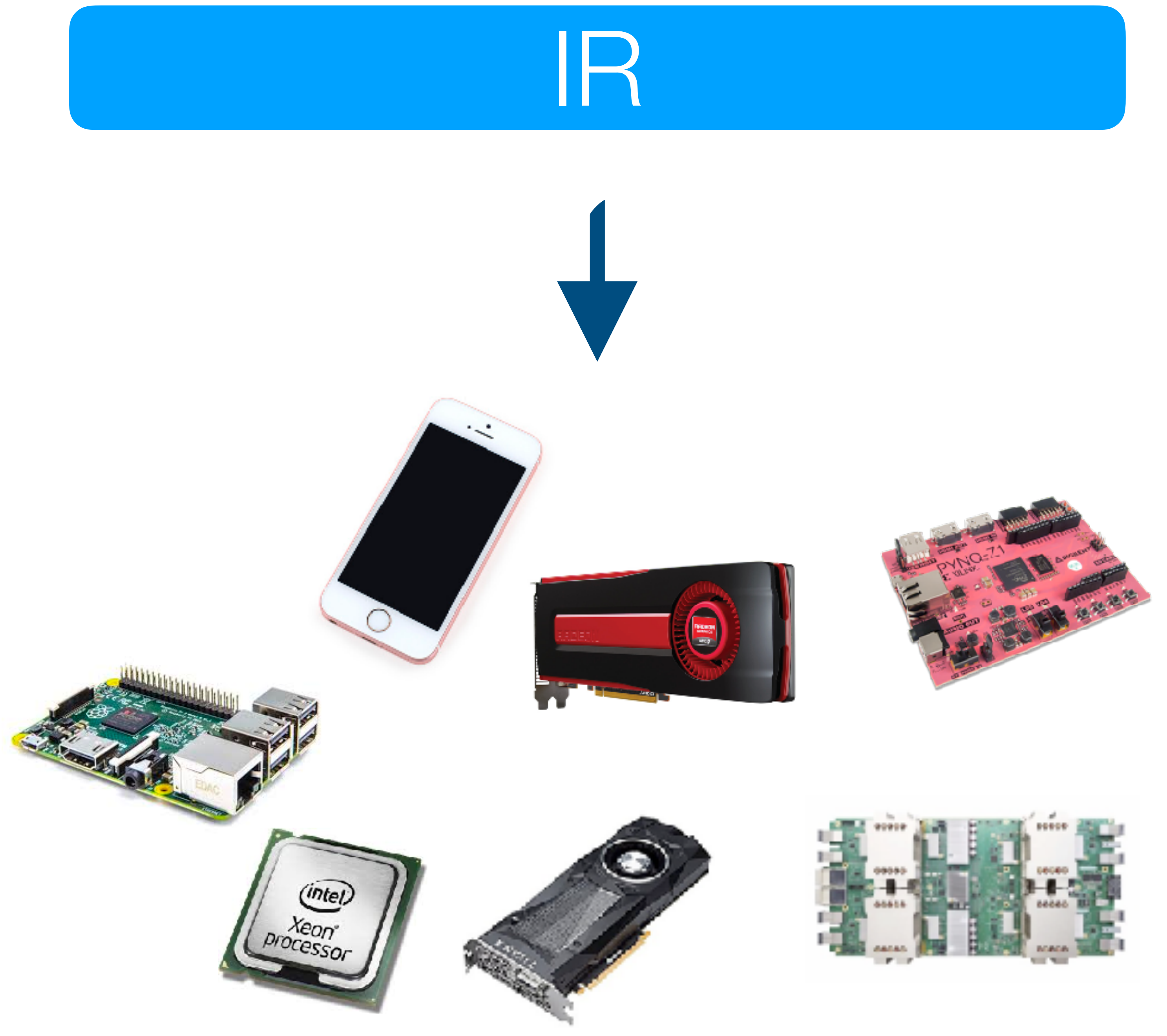
Challenge: Hardware Diversities

IR

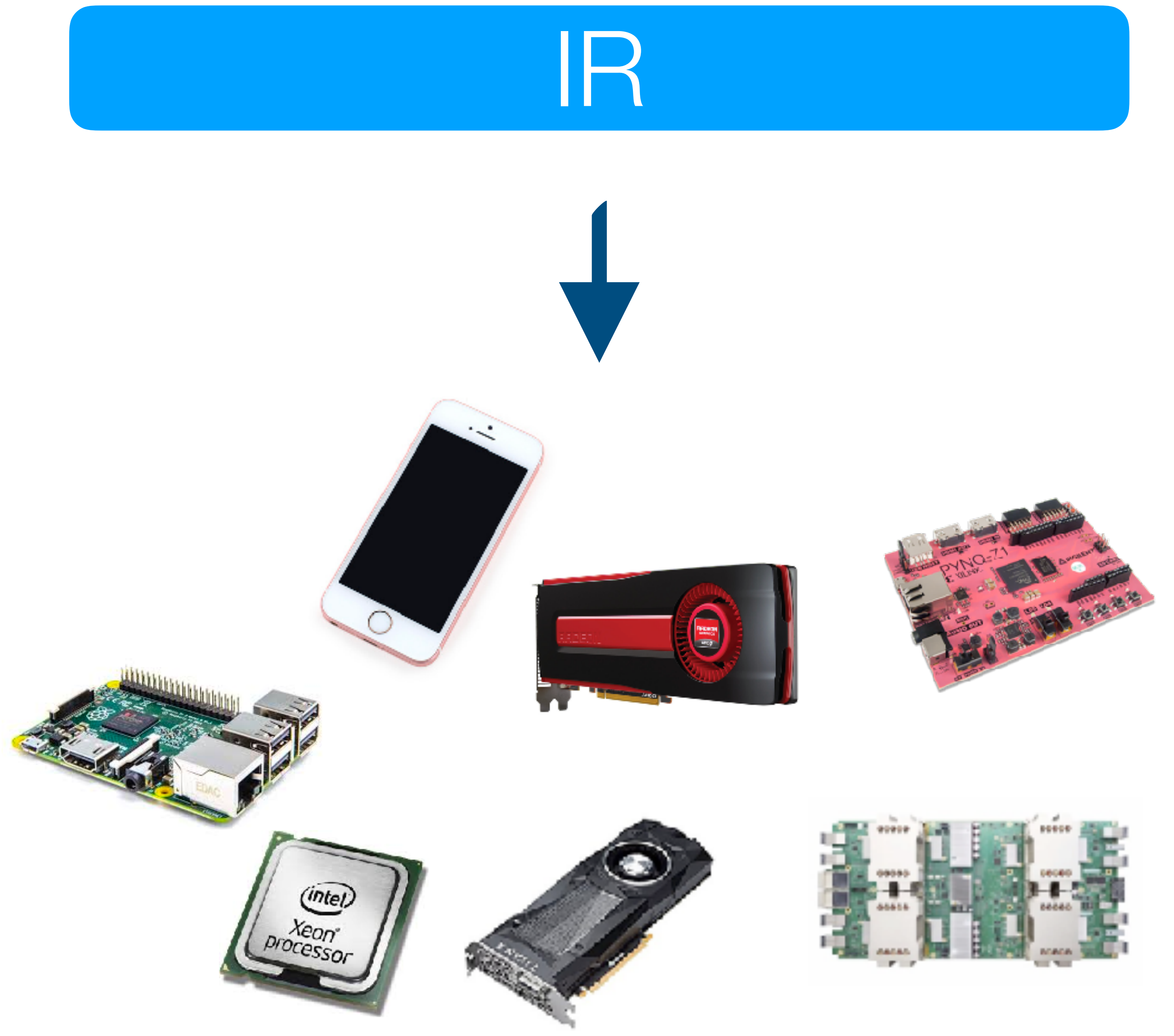


Challenge: Hardware Diversities

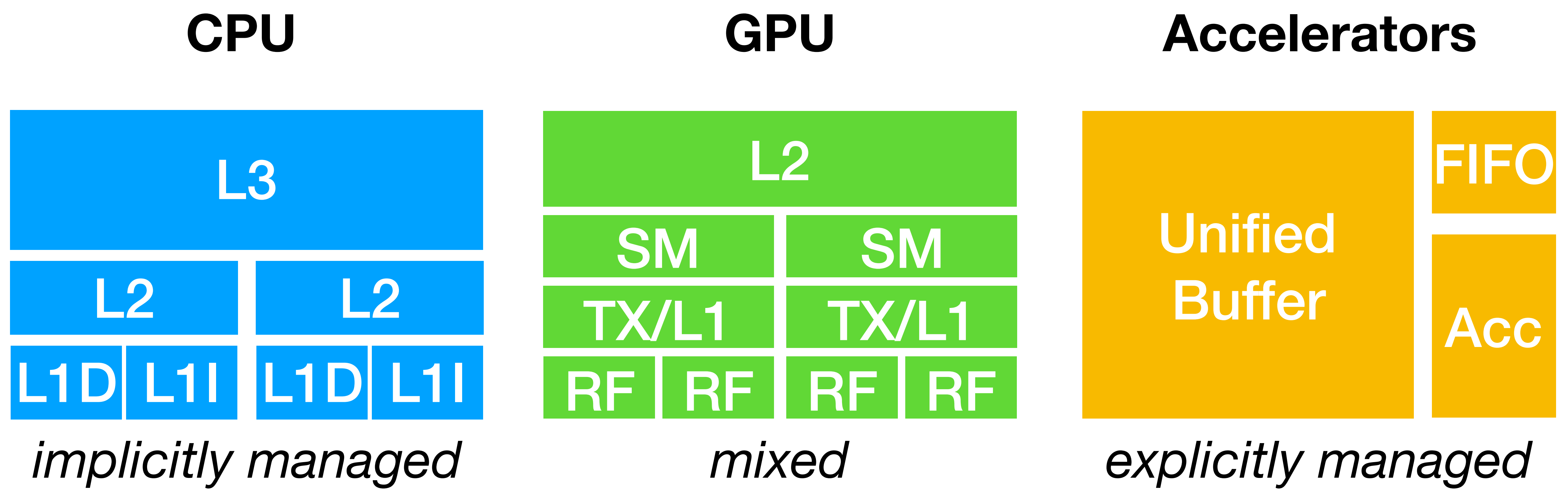
CPU **GPU** **Accelerators**



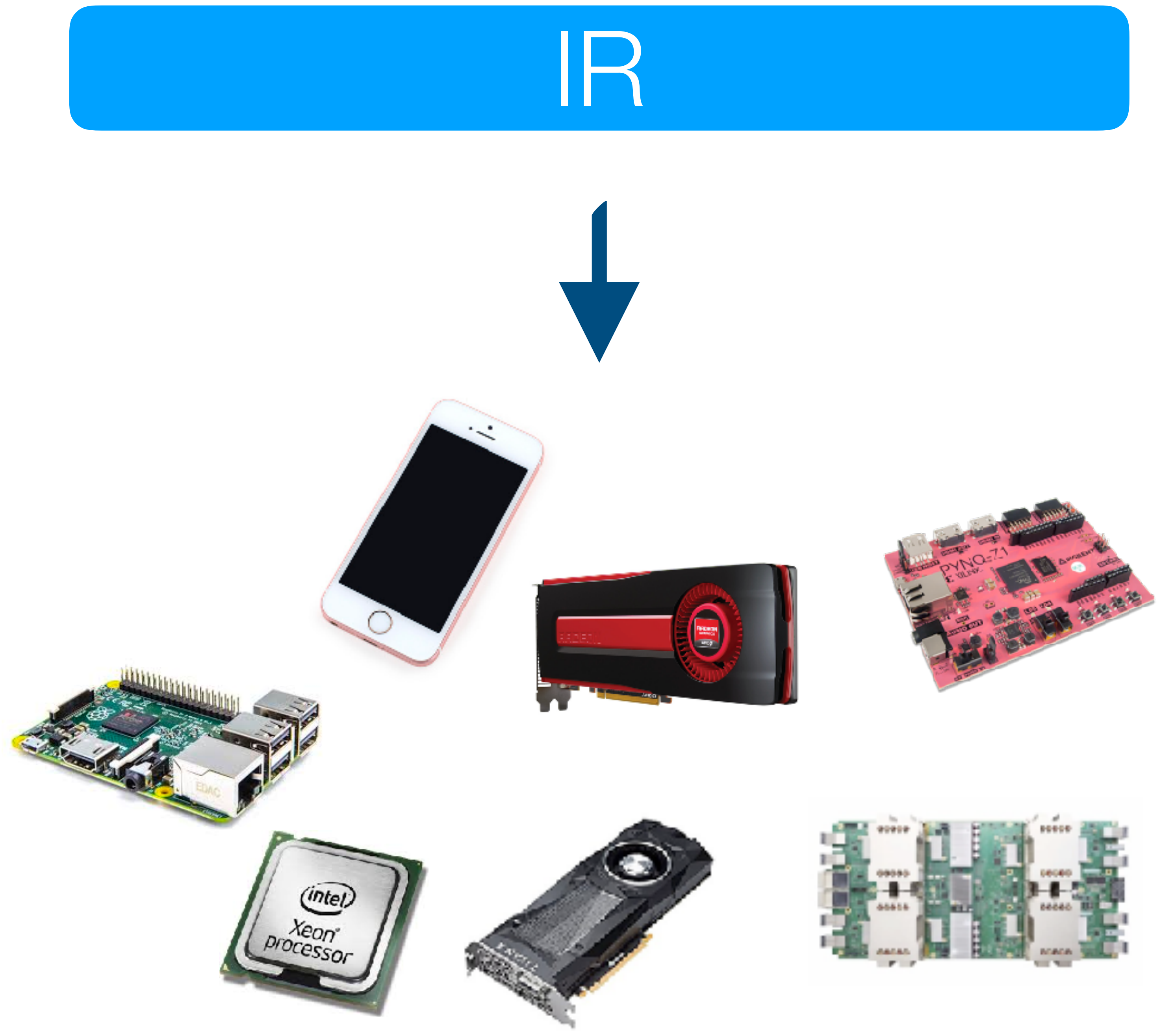
Challenge: Hardware Diversities



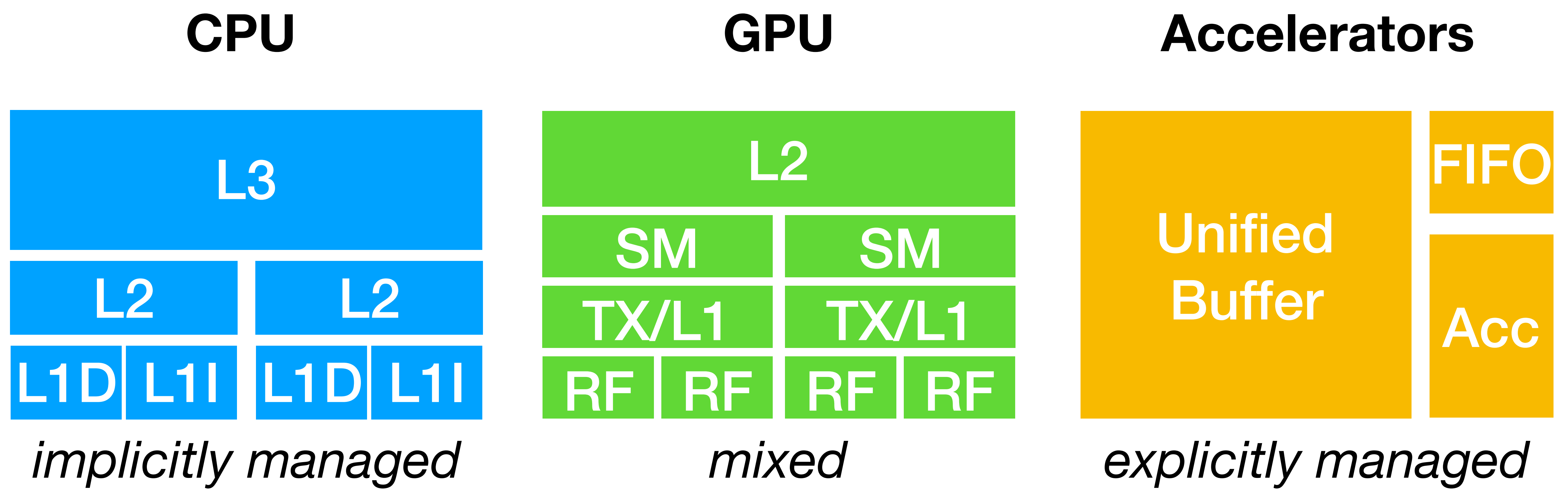
Memory subsystem



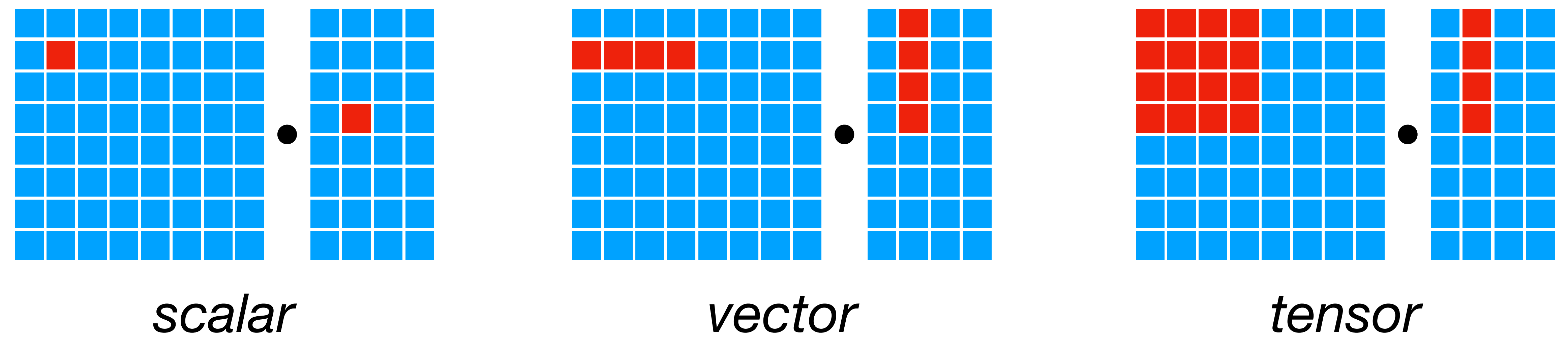
Challenge: Hardware Diversities



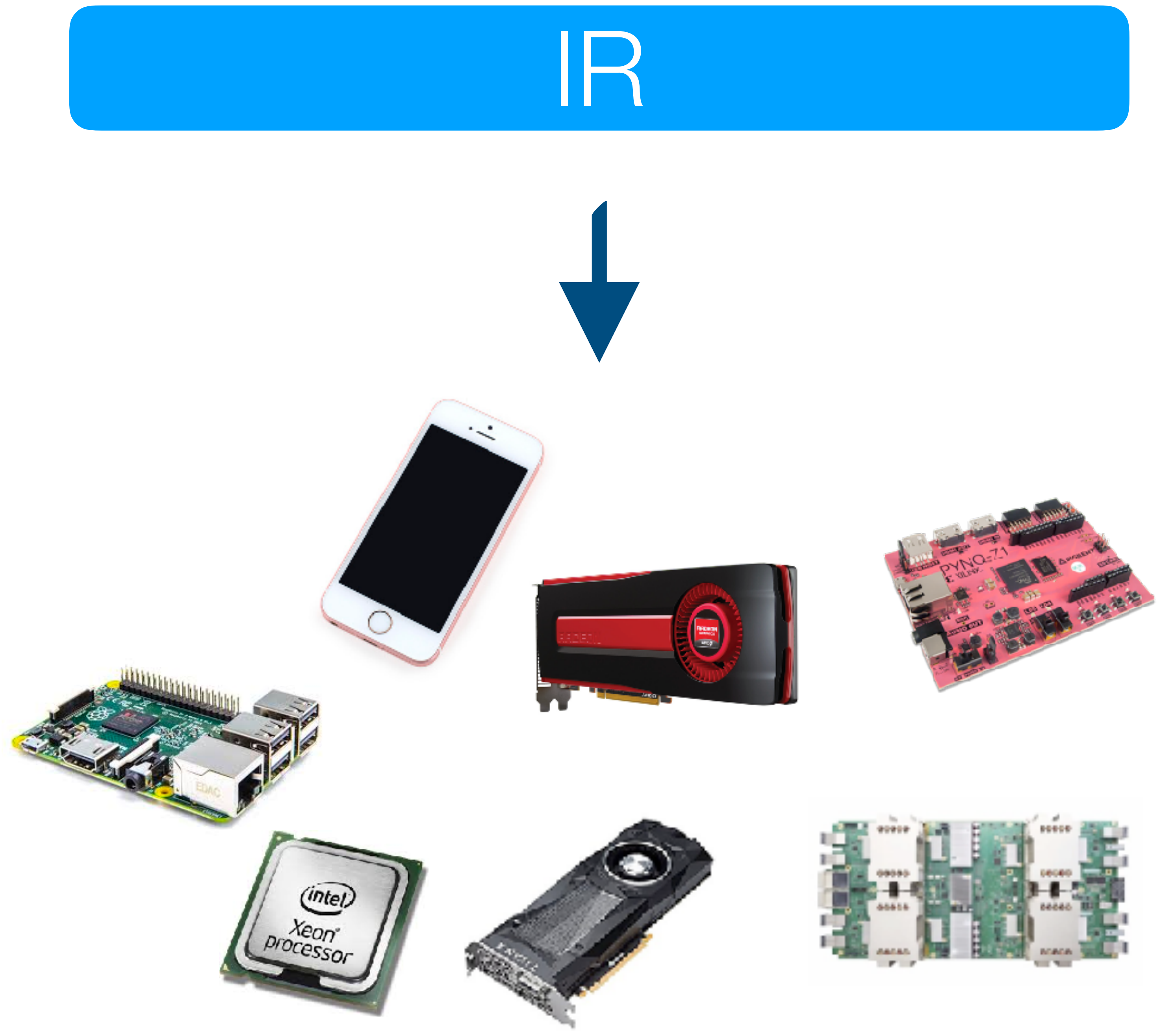
Memory subsystem

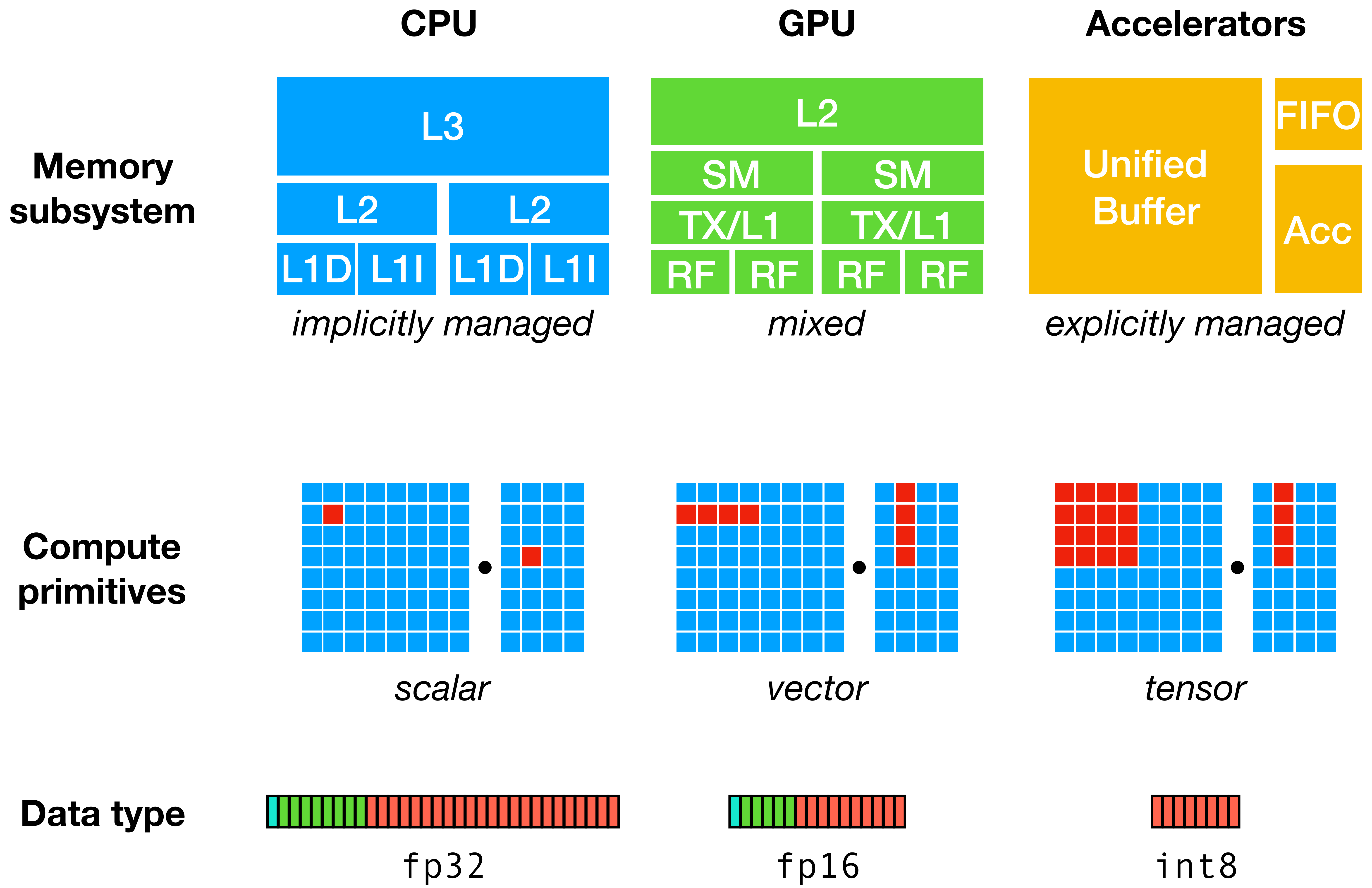


Compute primitives

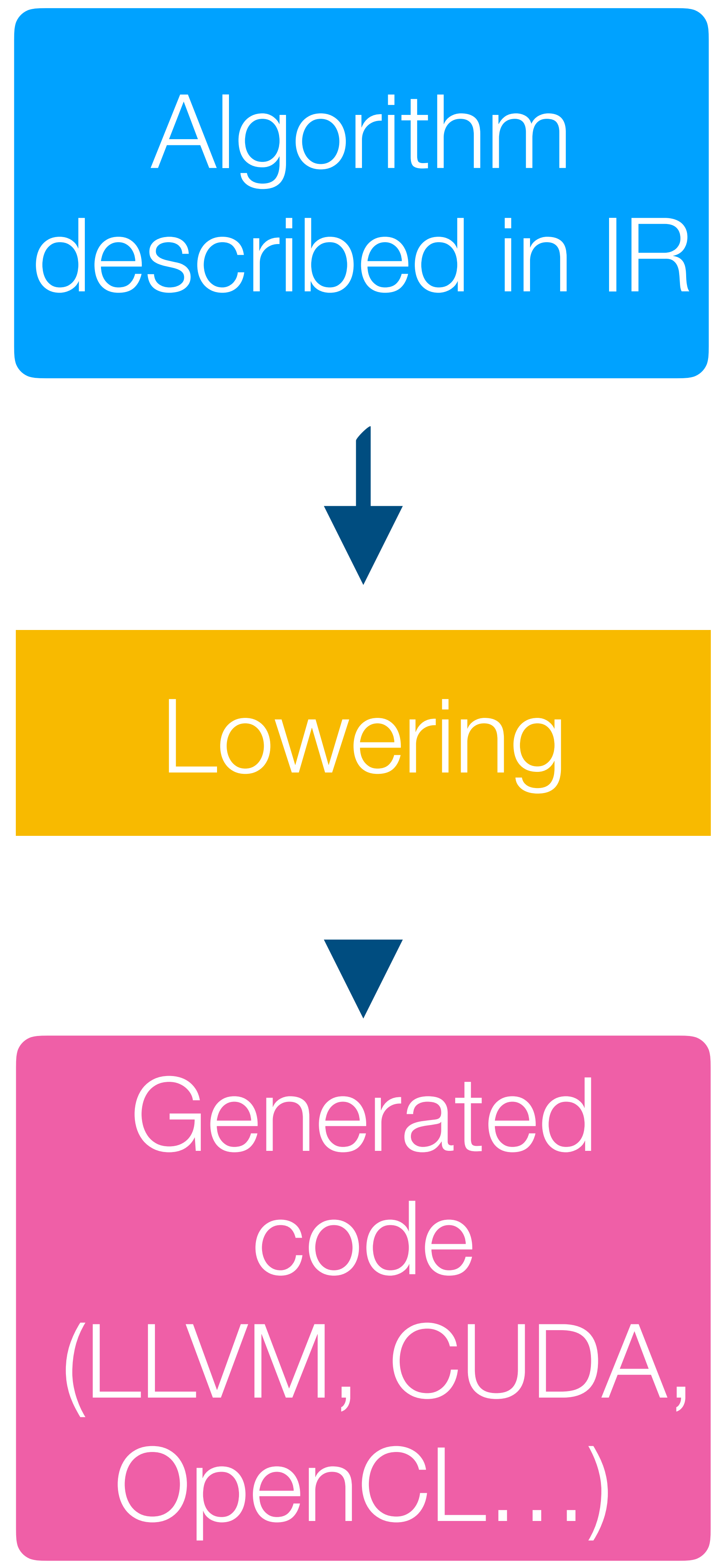


Challenge: Hardware Diversities

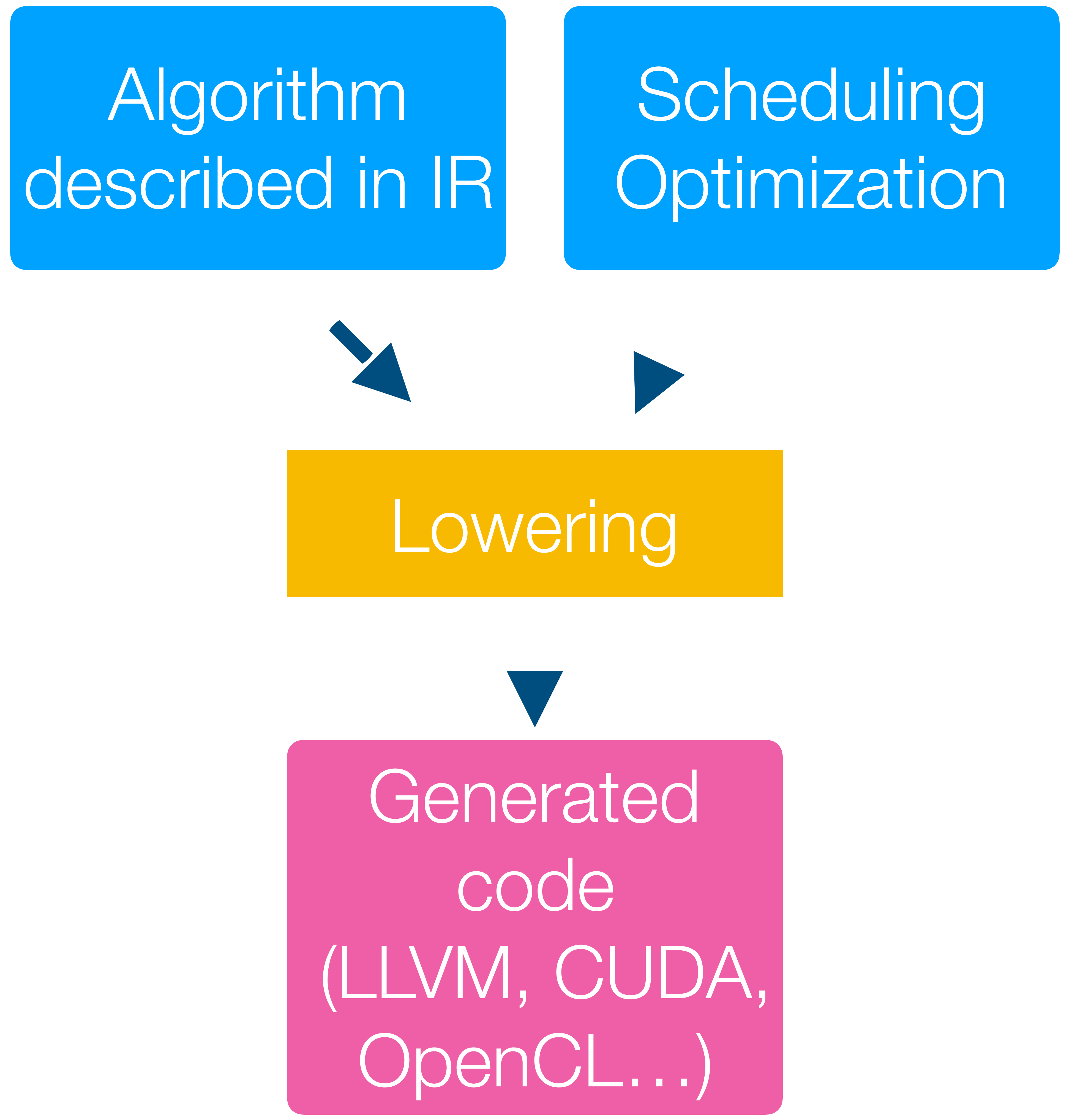




Unified Schedule Optimizations for Hardwares


Algorithm described in IR
Lowering
Generated code (LLVM, CUDA, OpenCL…)

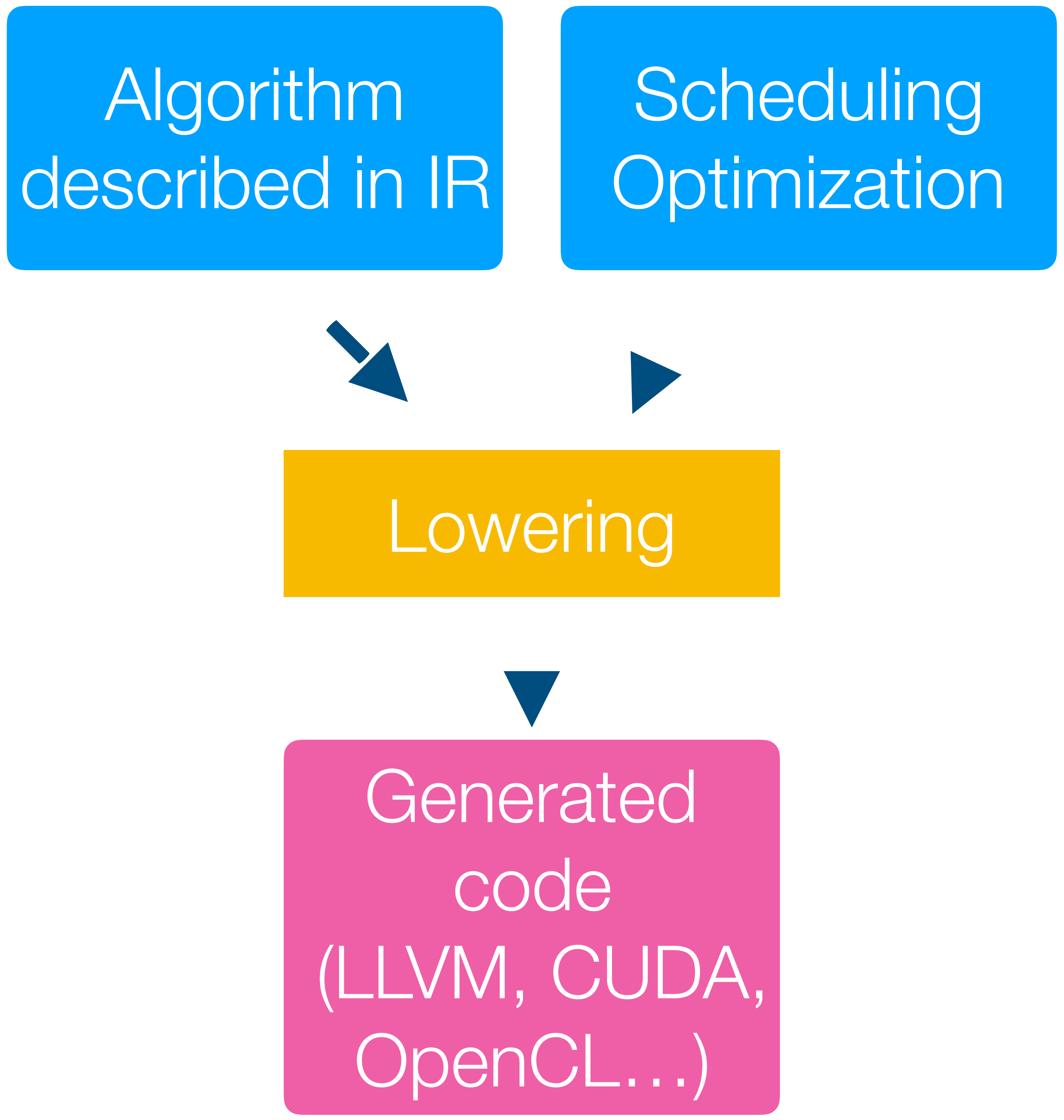
Unified Schedule Optimizations for Hardwares


Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

Unified Schedule Optimizations for Hardwares

Scheduling Optimizations

(✓) Data layout



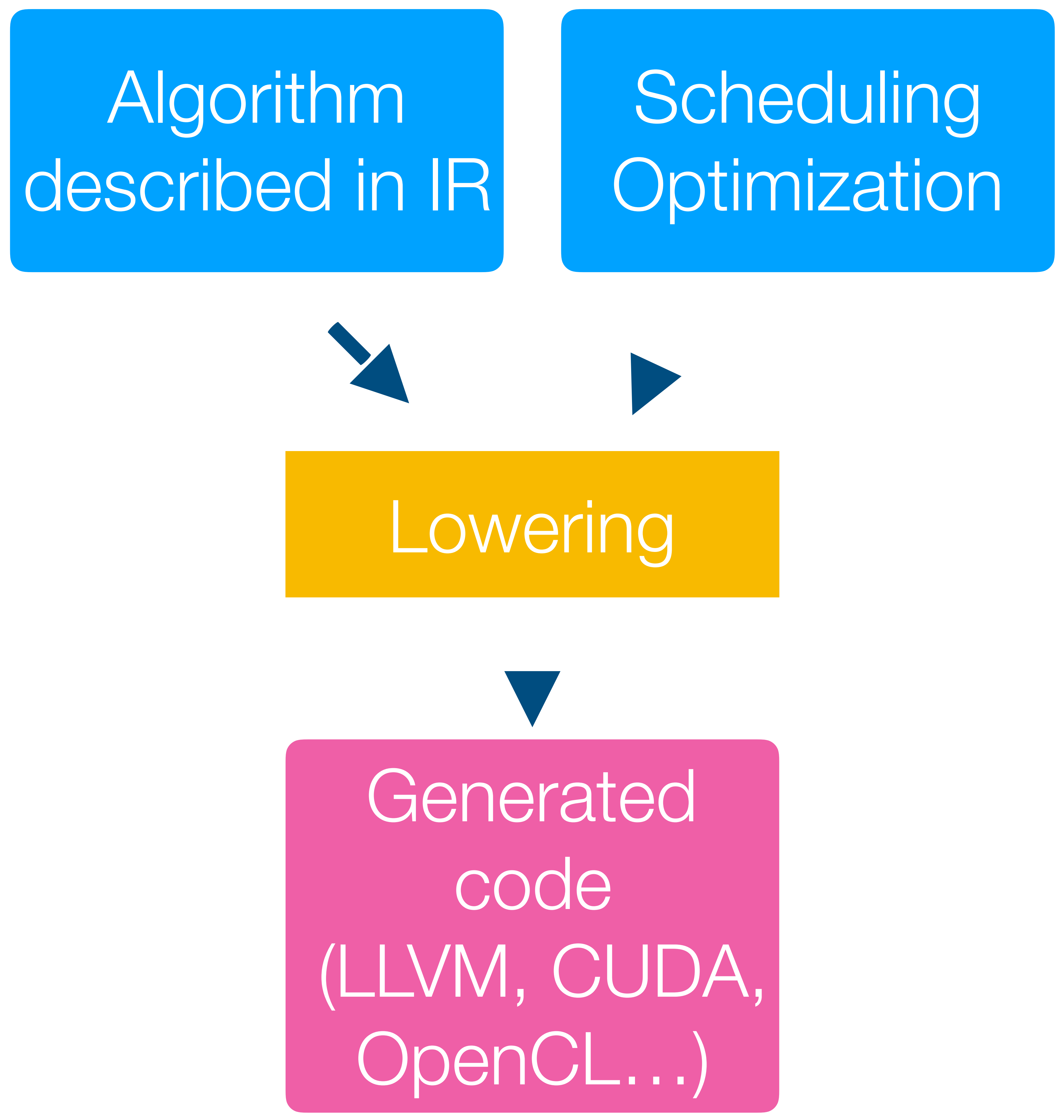
Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

Unified Schedule Optimizations for Hardwares

Scheduling Optimizations

(✓) Data layout

(✓) Tiling


Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

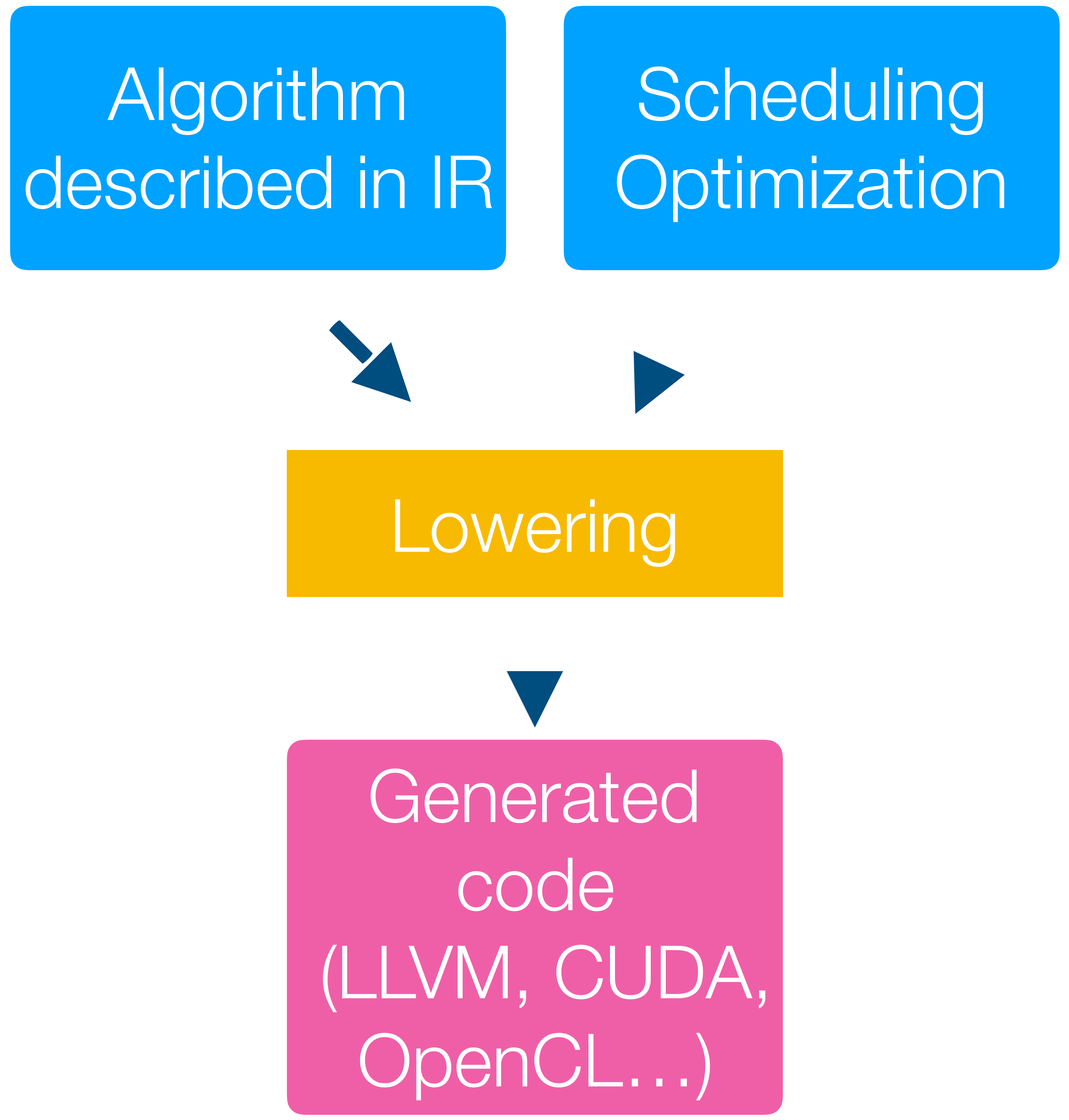
Unified Schedule Optimizations for Hardwares

Scheduling Optimizations

(✓) Data layout

(✓) Tiling

(✓) Thread cooperation


Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

Unified Schedule Optimizations for Hardwares

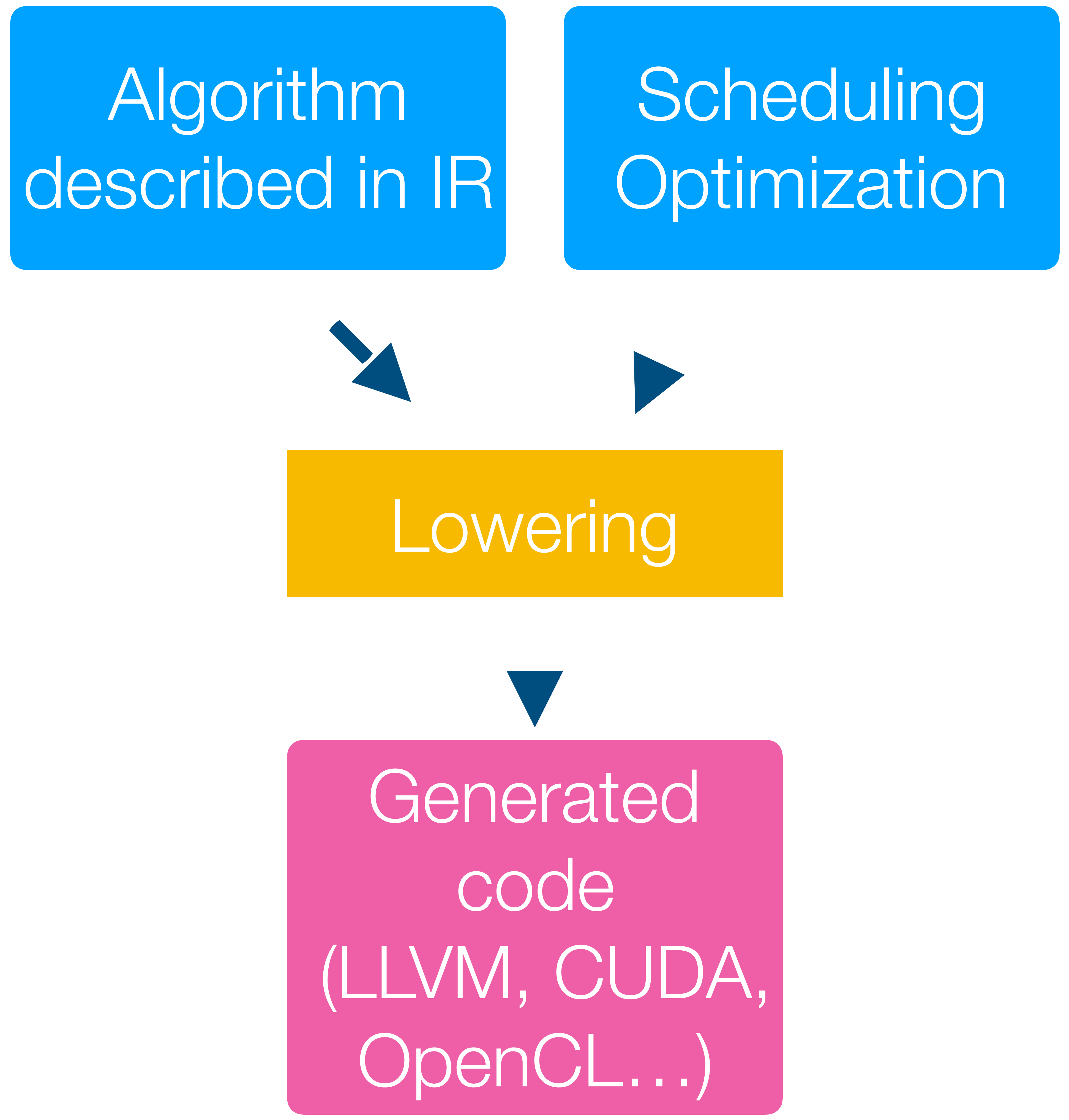
Scheduling Optimizations

(✓) Data layout

(✓) Tiling

(✓) Thread cooperation

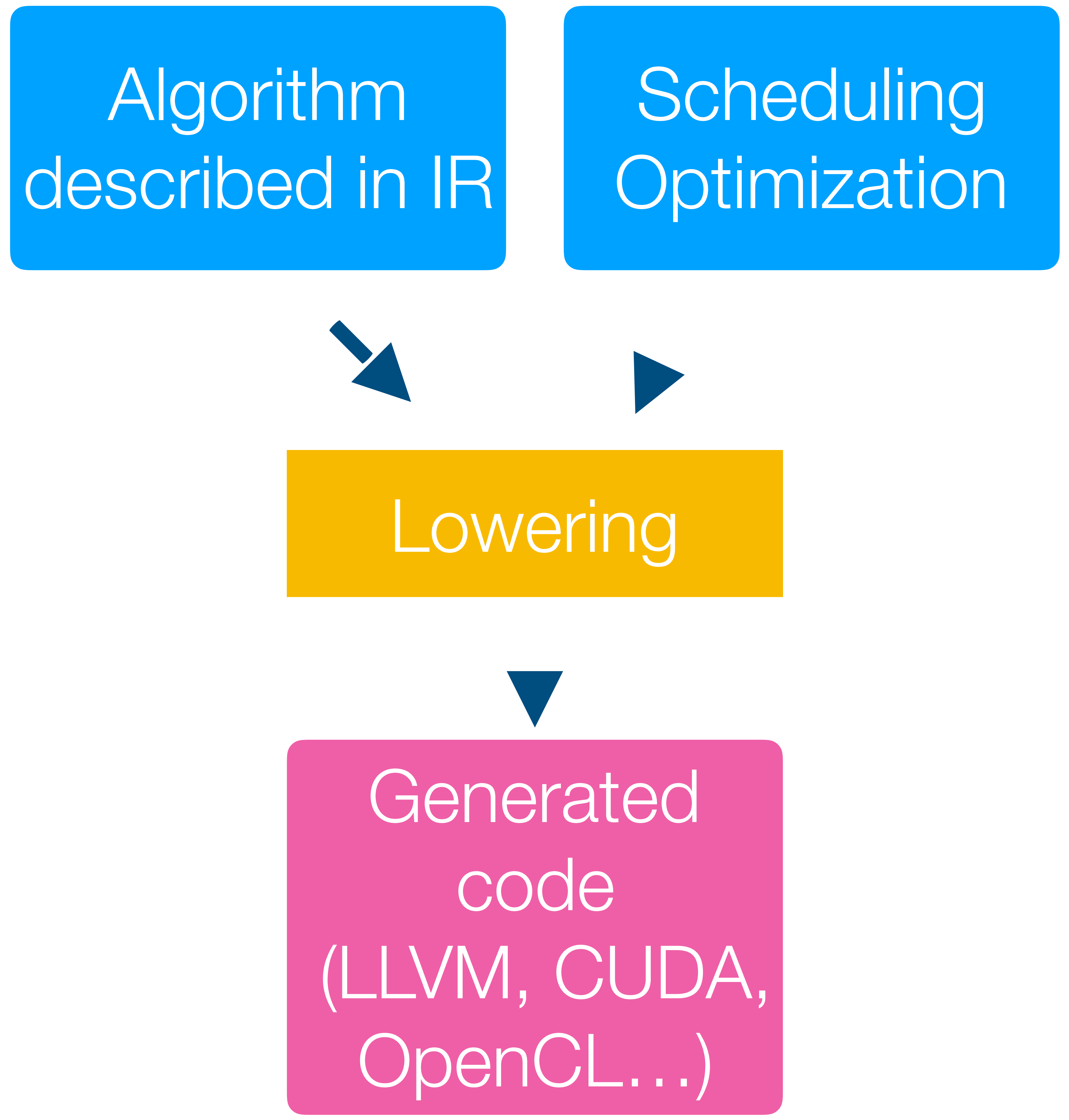
(✓) Latency hiding


Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

Unified Schedule Optimizations for Hardwares

Scheduling Optimizations

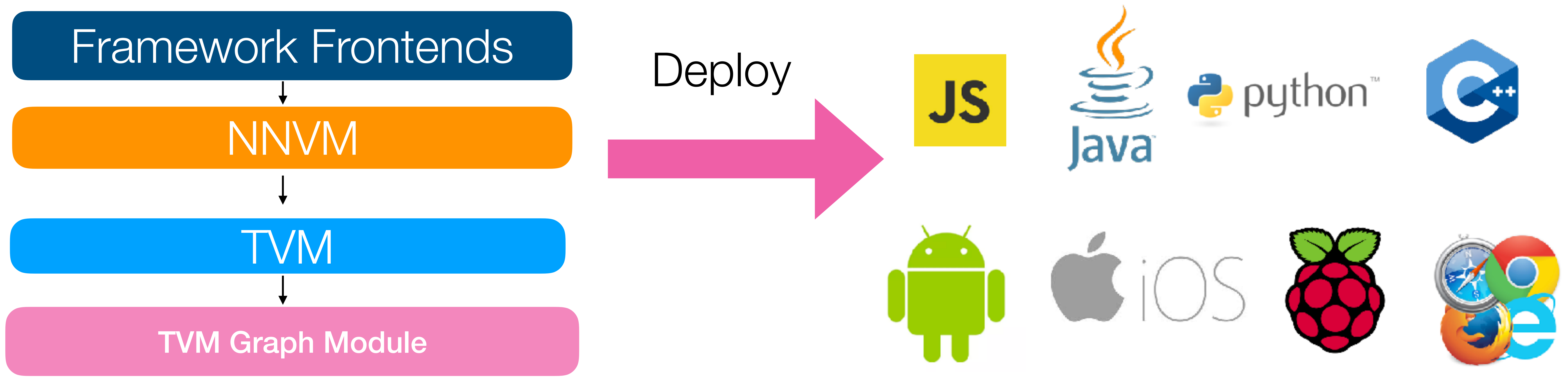
- (✓) Data layout
- (✓) Tiling
- (✓) Thread cooperation
- (✓) Latency hiding
- (✓) Tensorization


Algorithm described in IR
Scheduling Optimization
Lowering
Generated code (LLVM, CUDA, OpenCL…)

Separation of Compilation and Deployment

Compilation Stack

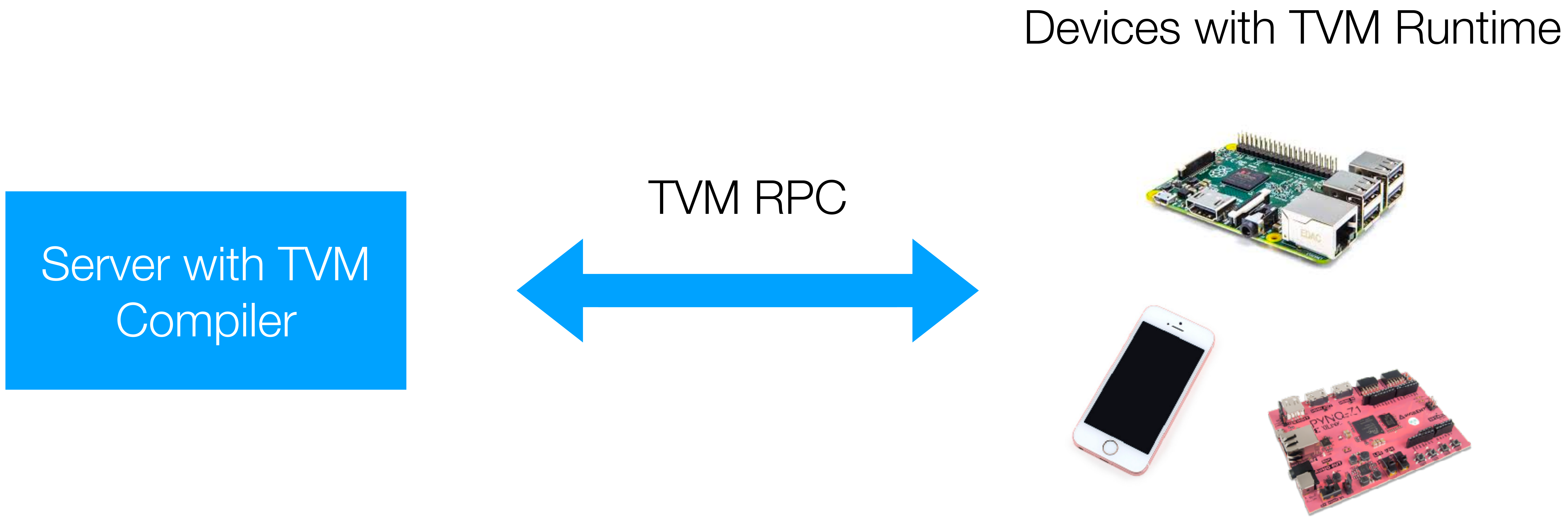
TVM Runtimes



Heavy optimizations

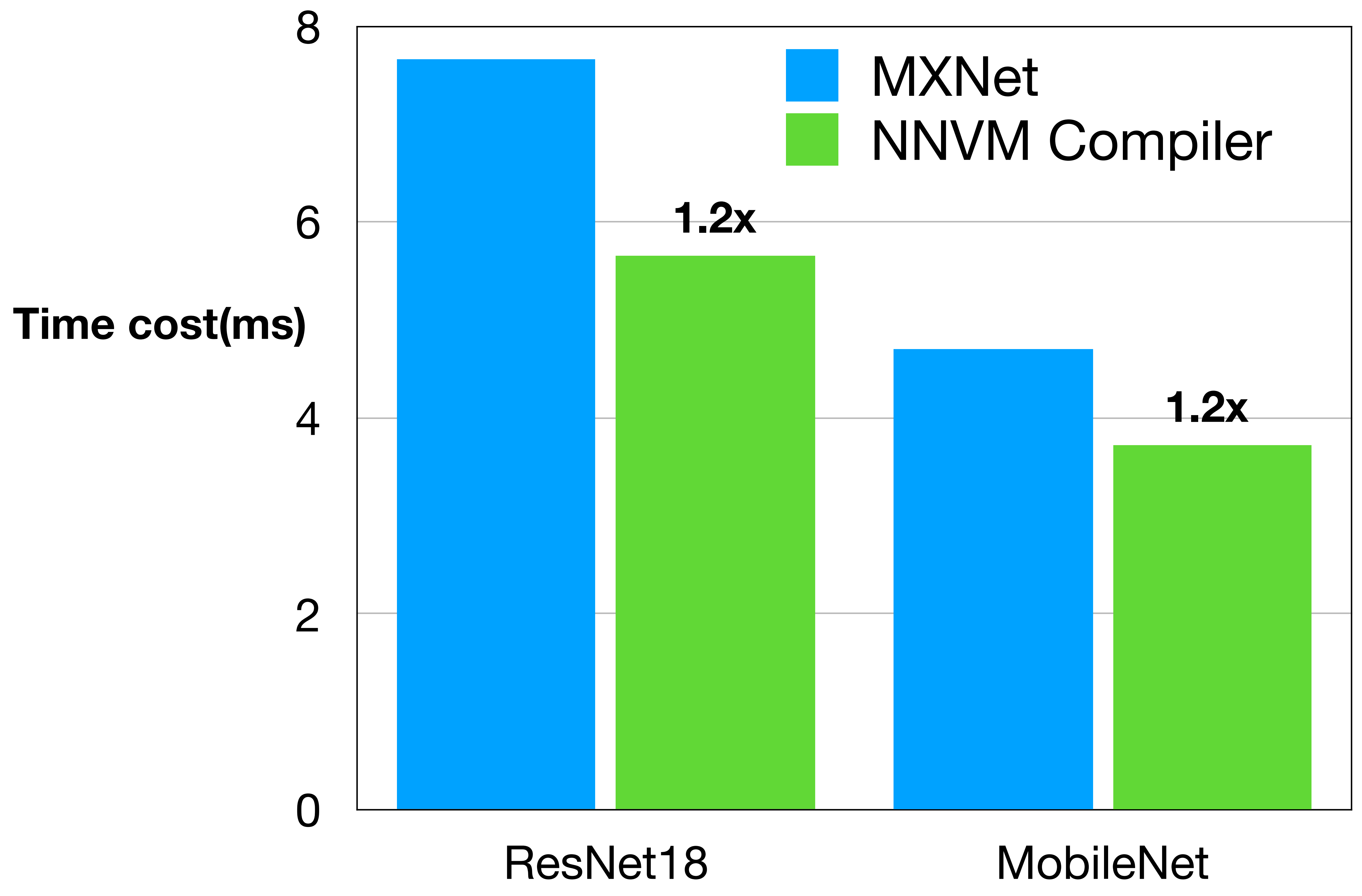
Lightweight, 300 to 600 KB

Remote Execution and Profiling

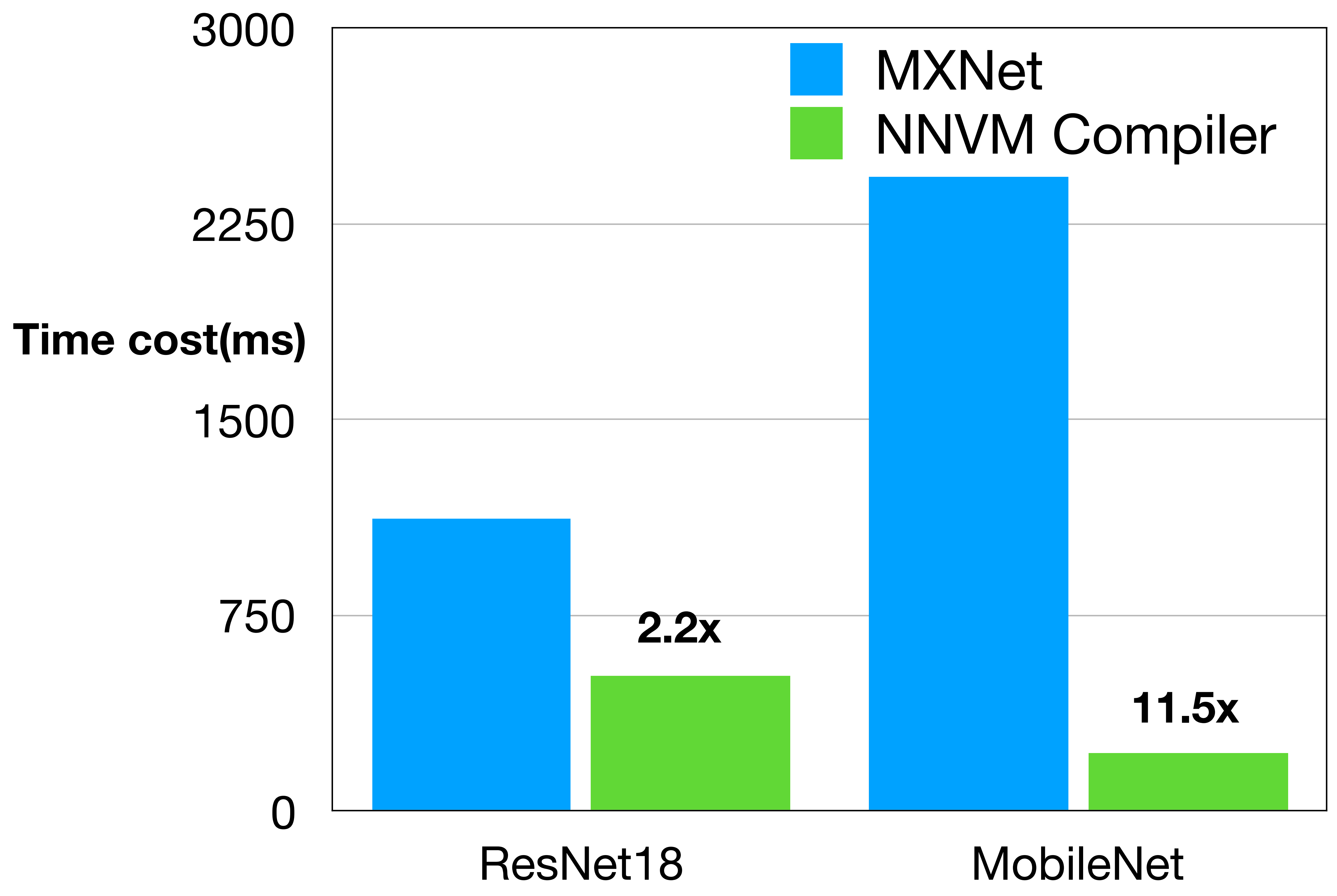


Performance Portable against state of art

K80, Baseline
MXNet with cuDNN auto tune enabled
One grad student month



Raspberry Pi 3
Baseline: MXNet with OpenBLAS and NNPack
Two undergrad weeks



Credit: Leyuan Wang(AWS/UCDavis), Yuwei Hu(TuSimple), Zheng Jiang(AWS/FDU)

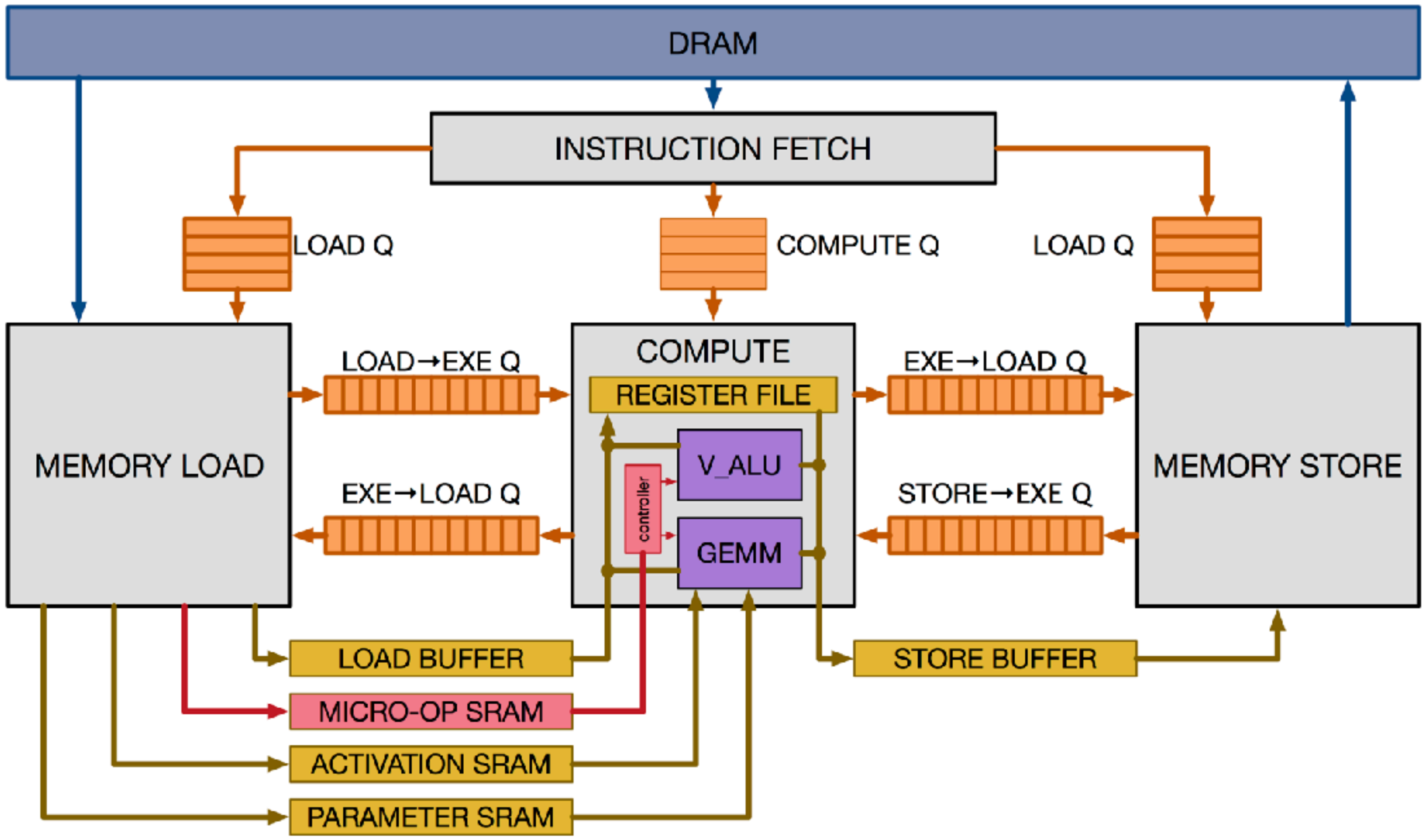
Coming Soon: Target New Accelerators

Tensorization

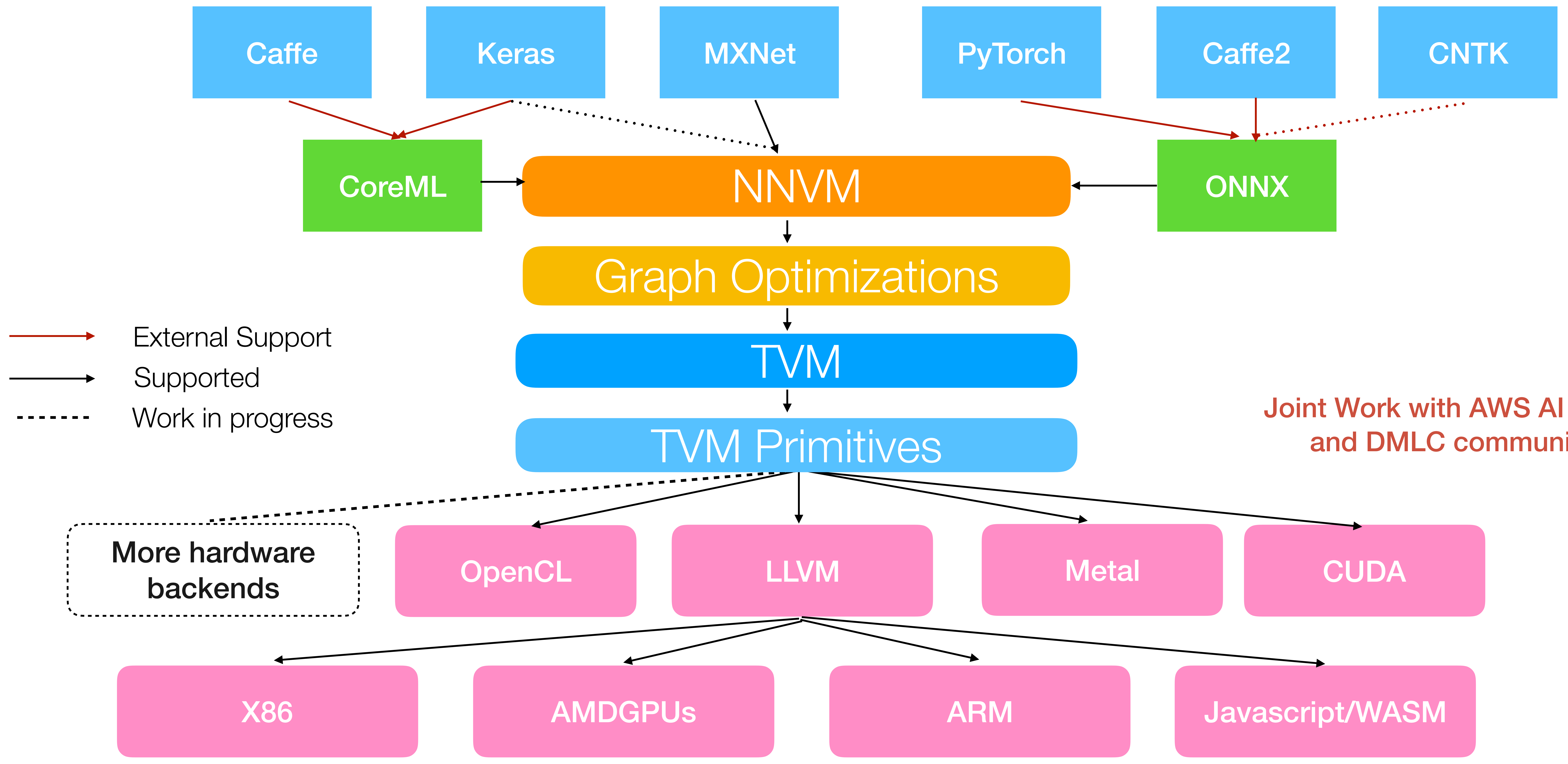
Latency Hiding

FPGA Example for building new hardware backend

Open-source soon



NNVM Compiler: Open Compiler for AI Systems

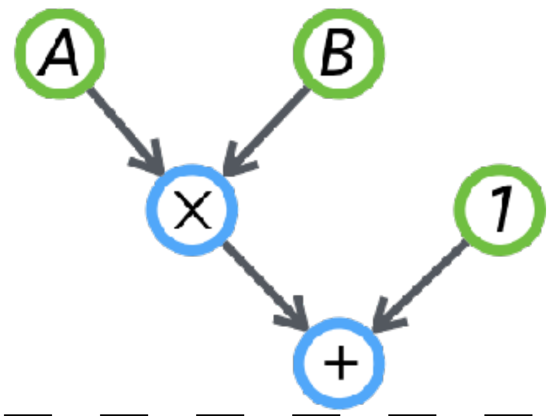


Joint Work with AWS AI Team and DMLC community

Deep Learning System Research is Exciting but Hard

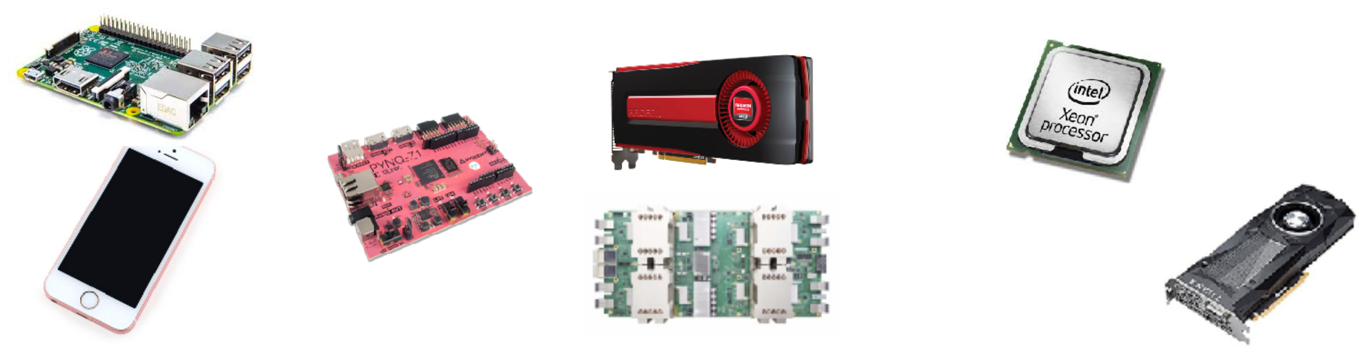
Frameworks CNTK Caffe2

Computational graph

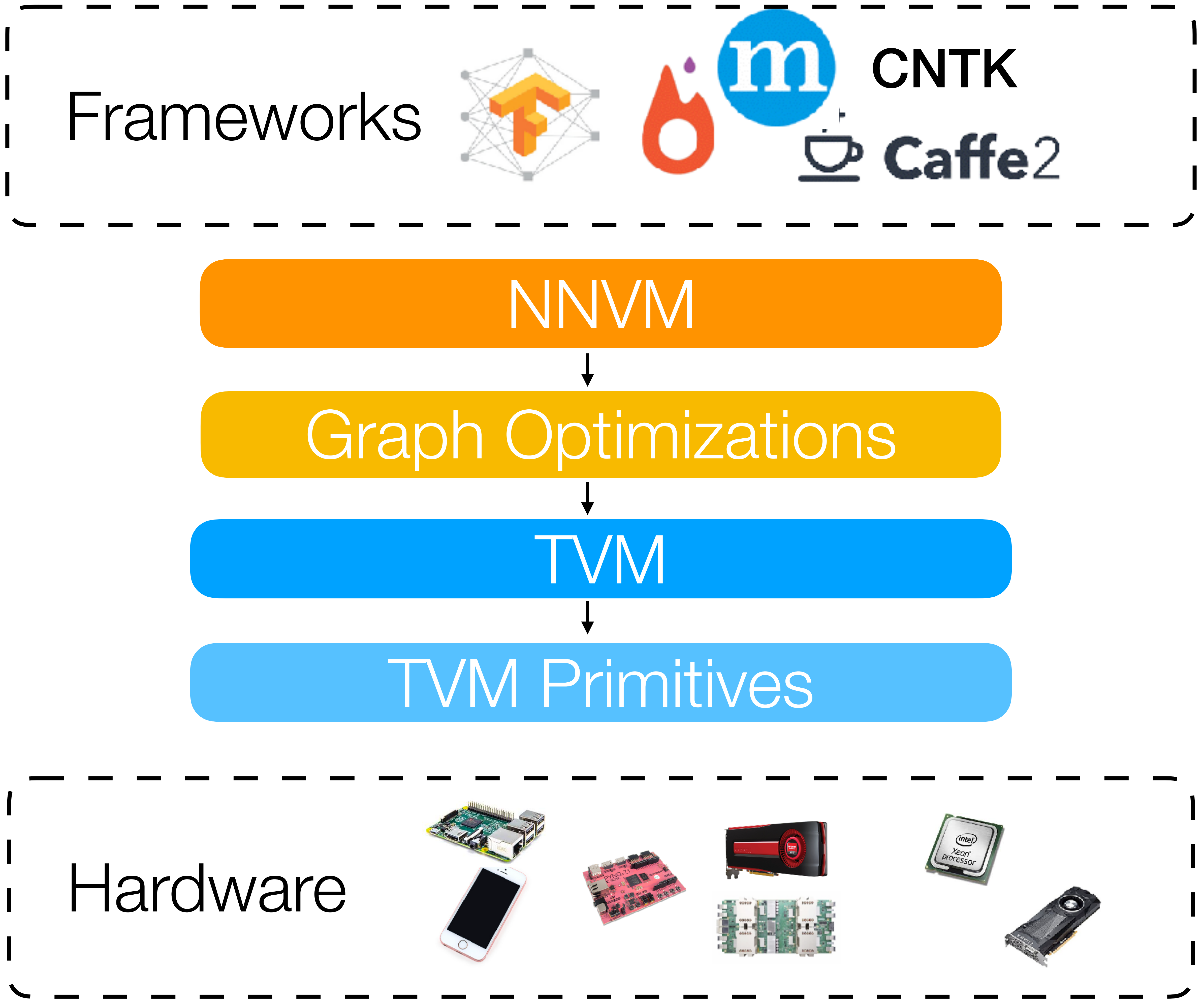


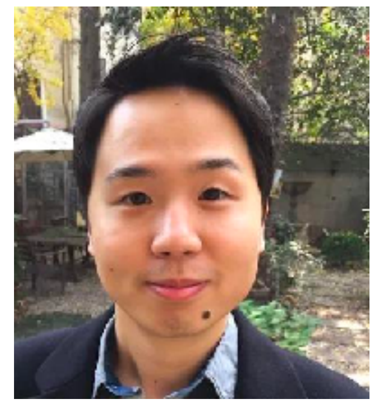
Operator Libraries **cuDNN, NNPack, MKL-DNN**

Hardware



Deep Learning System Research is Just Exciting

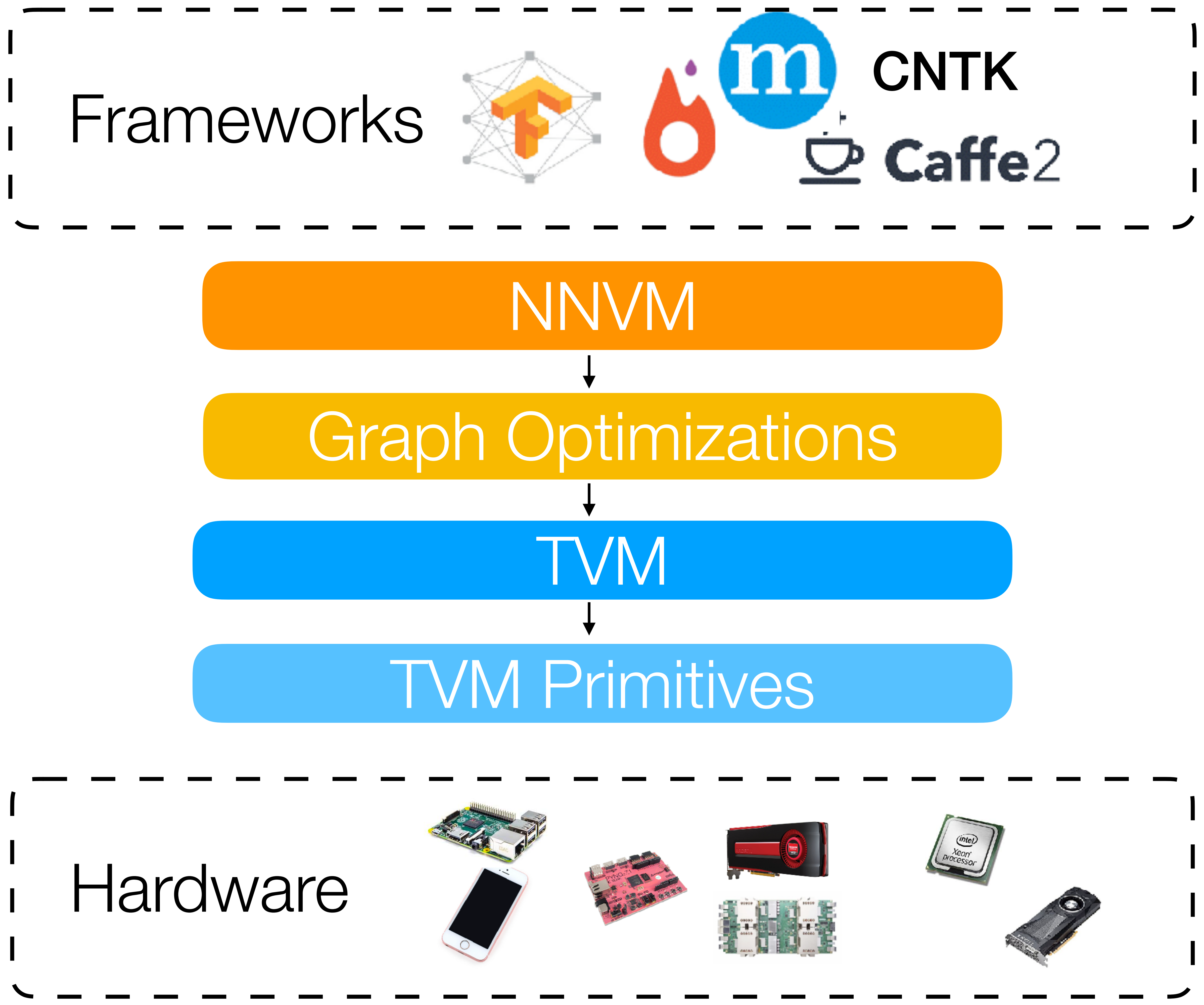


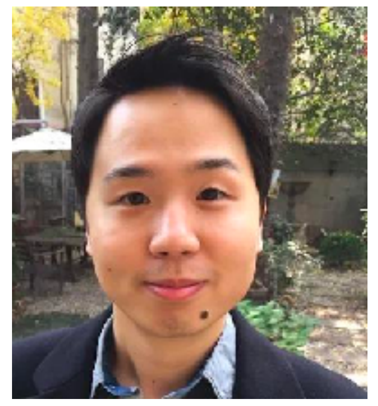


My new optimizations works on all platforms !

I can program my new accelerators from python :)

Deep Learning System Research is Just Exciting





My new optimizations works on all platforms !

I can program my new accelerators from python :)

You can be part of it!