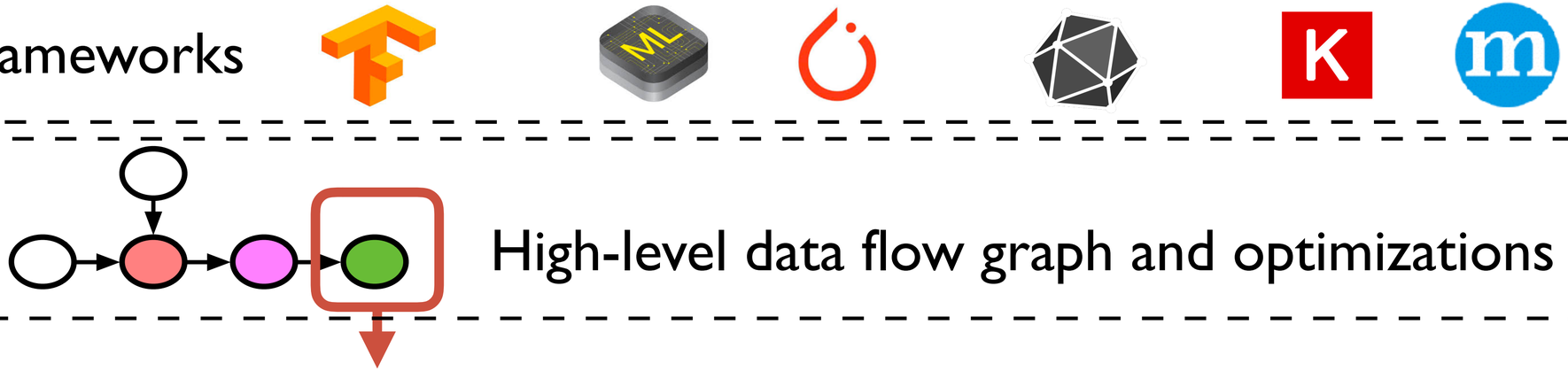
AutoTVM & Device Fleet

`

Learning to Optimize Tensor Programs

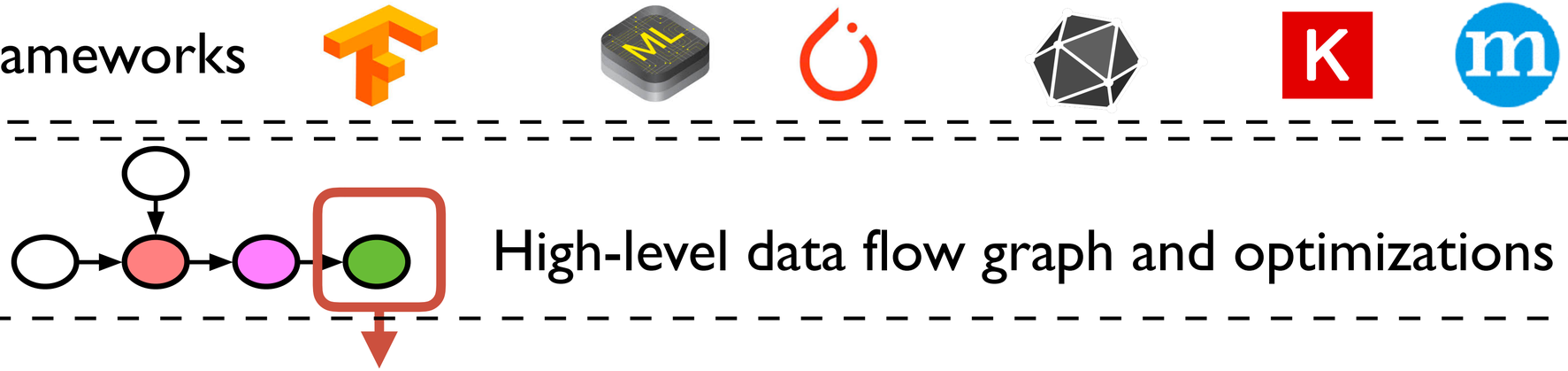
Frameworks



Hardware

Learning to Optimize Tensor Programs

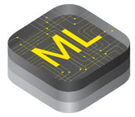
Frameworks



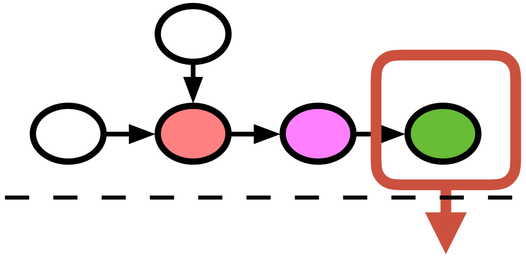
Hardware

Learning to Optimize Tensor Programs

Frameworks





High-level data flow graph and optimizations

Machine Learning based Program Optimizer

Hardware

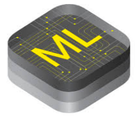




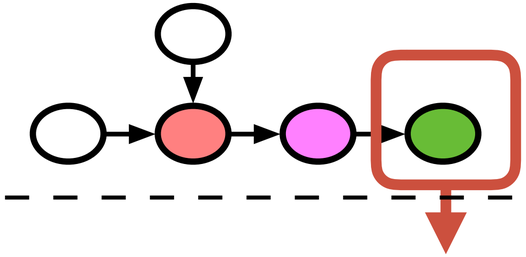



Learning to Optimize Tensor Programs

Frameworks





High-level data flow graph and optimizations

Machine Learning based Program Optimizer

Learning to generate optimized program
for new operator workloads and hardware

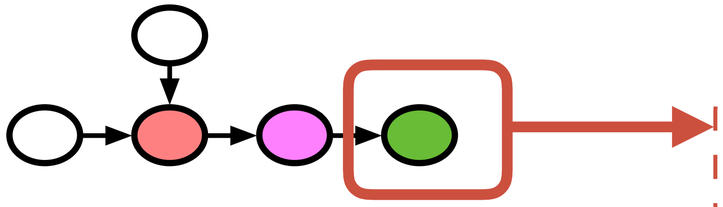
Hardware








Search over Possible Program Transformations



Compute Description

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

Loop Transformations

Thread Bindings

Cache Locality

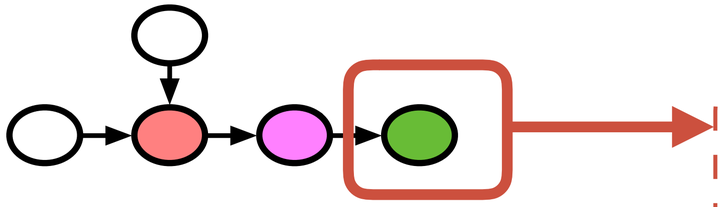
Thread Cooperation

Tensorization

Latency Hiding

Hardware

Search over Possible Program Transformations



Compute Description

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

Loop Transformations

Thread Bindings

Cache Locality

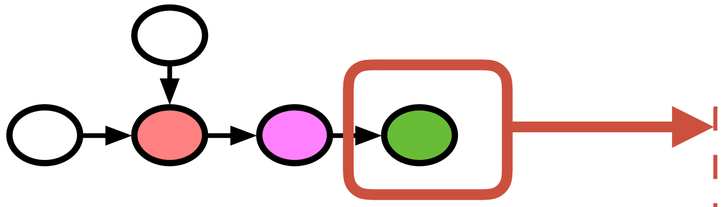
Thread Cooperation

Tensorization

Latency Hiding

Hardware

Search over Possible Program Transformations



Compute Description

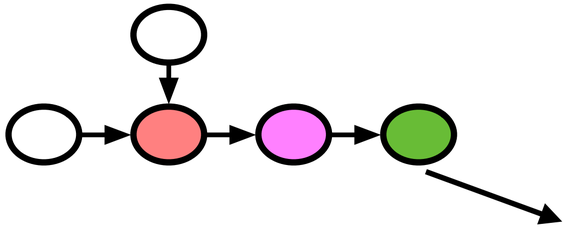
```
C = tvm.compute((m, n),
    lambda y, x: tvm.sum(A[k, y] * B[k, x], axis=k))
```

Billions of possible optimization choices

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

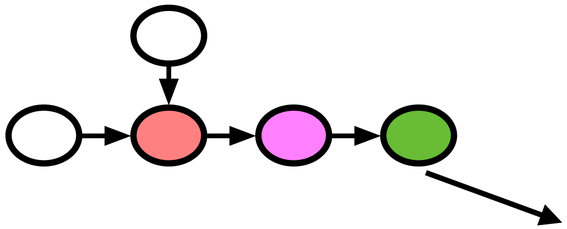
Hardware

Learning-based Program Optimizer



Program Optimizer → Code Generator → Program

Learning-based Program Optimizer



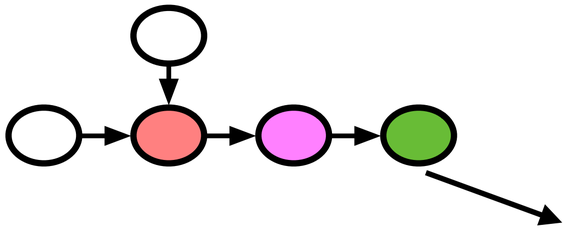
Program Optimizer

Code Generator

Program

Runtime Measurements

Learning-based Program Optimizer



Program Optimizer

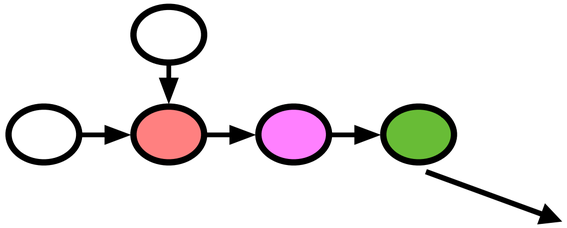
Code Generator

Program

Runtime Measurements

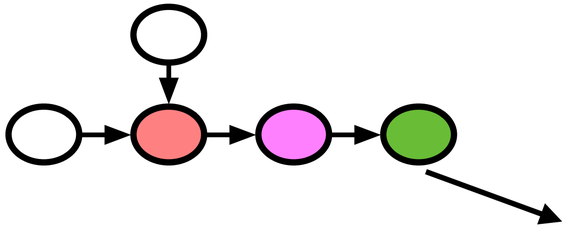
High experiment cost, each trial costs ~1second

Learning-based Program Optimizer



Program Optimizer → Code Generator → Program

Learning-based Program Optimizer



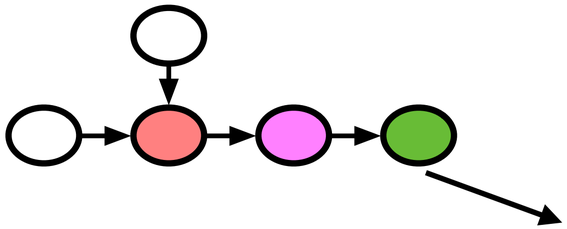
Program Optimizer

Code Generator

Program

Cost Model

Learning-based Program Optimizer



Program Optimizer

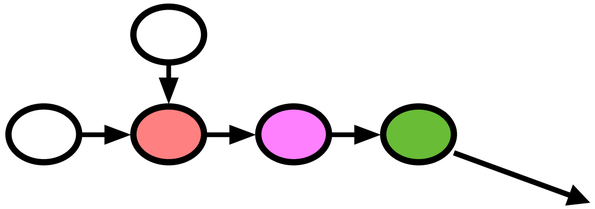
Code Generator

Program

Cost Model

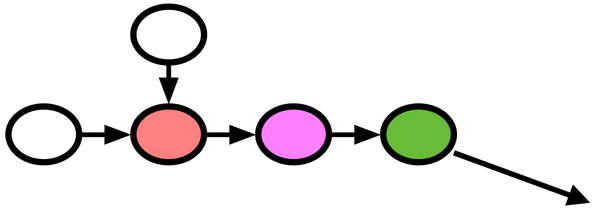
Need reliable cost model per hardware

Learning-based Program Optimizer



Program Optimizer → Code Generator → Program

Learning-based Program Optimizer



Program Optimizer

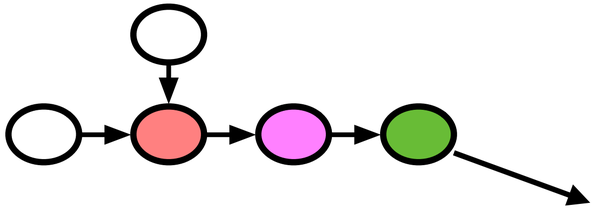
Code Generator

Program

$\mathcal{D}$

Training data

Learning-based Program Optimizer



Program Optimizer

Code Generator

Program

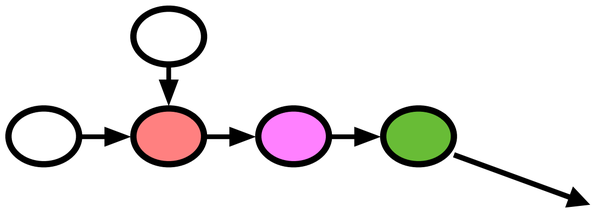
Learning

Statistical Cost Model

$\mathcal{D}$

Training data

Learning-based Program Optimizer



Program Optimizer

Code Generator

Program

Learning

Statistical Cost Model

$\mathcal{D}$

Training data

Unique Problem Characteristics

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

Program-aware Cost Modeling

High-Level Configuration

Program-aware Cost Modeling

High-Level Configuration

```
for y in range(8):
  for x in range(8):
    C[y][x]=0
    for k in range(8):
      C[y][x]+=A[k][y]*B[k][x]
```

Low-level Abstract Syntax Tree
(shared between tasks)

Program-aware Cost Modeling

High-Level Configuration

```
for y in range(8):
  for x in range(8):
    C[y][x]=0
    for k in range(8):
      C[y][x]+=A[k][y]*B[k][x]
```

Low-level Abstract Syntax Tree
(shared between tasks)

statistical features

	touched memory		
	C	A	B
y	64	64	64
x	8	8	64
k	1	8	8

outer loop length	
y	1
x	8
k	64

Boosted
Tree Ensembles

Program-aware Cost Modeling

High-Level Configuration

```
for y in range(8):
  for x in range(8):
    C[y][x]=0
    for k in range(8):
      C[y][x]+=A[k][y]*B[k][x]
```

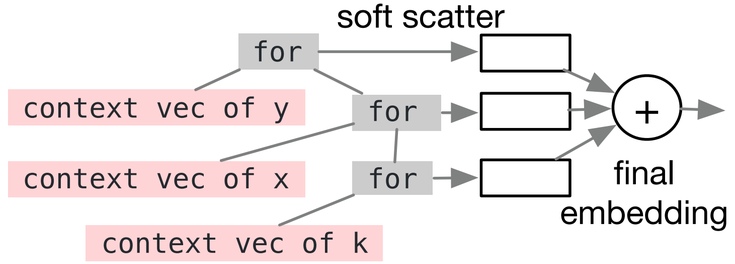
Low-level Abstract Syntax Tree
(shared between tasks)

	touched memory		
	C	A	B
y	64	64	64
x	8	8	64
k	1	8	8

	outer loop length
y	1
x	8
k	64

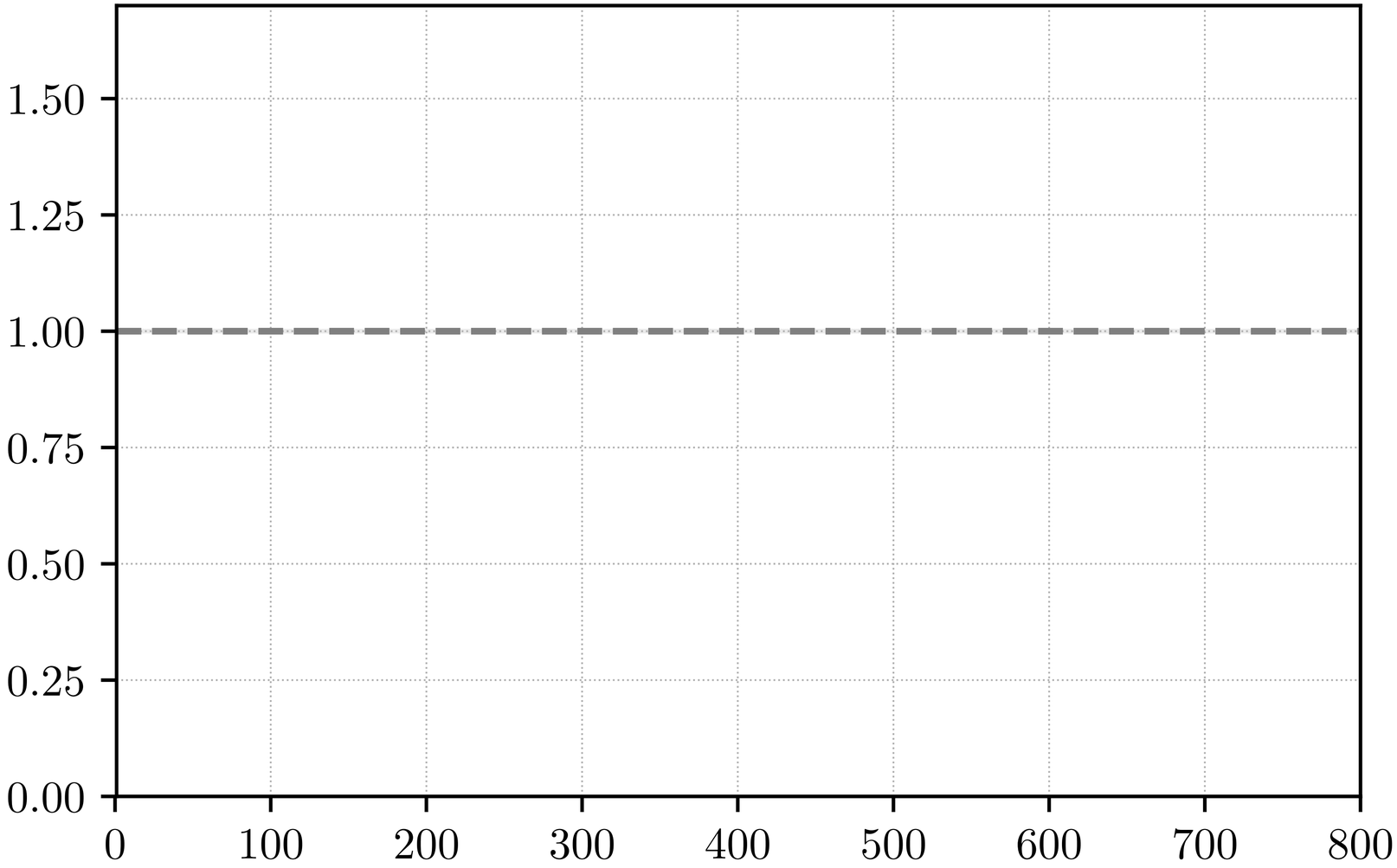
statistical features

Boosted Tree Ensembles

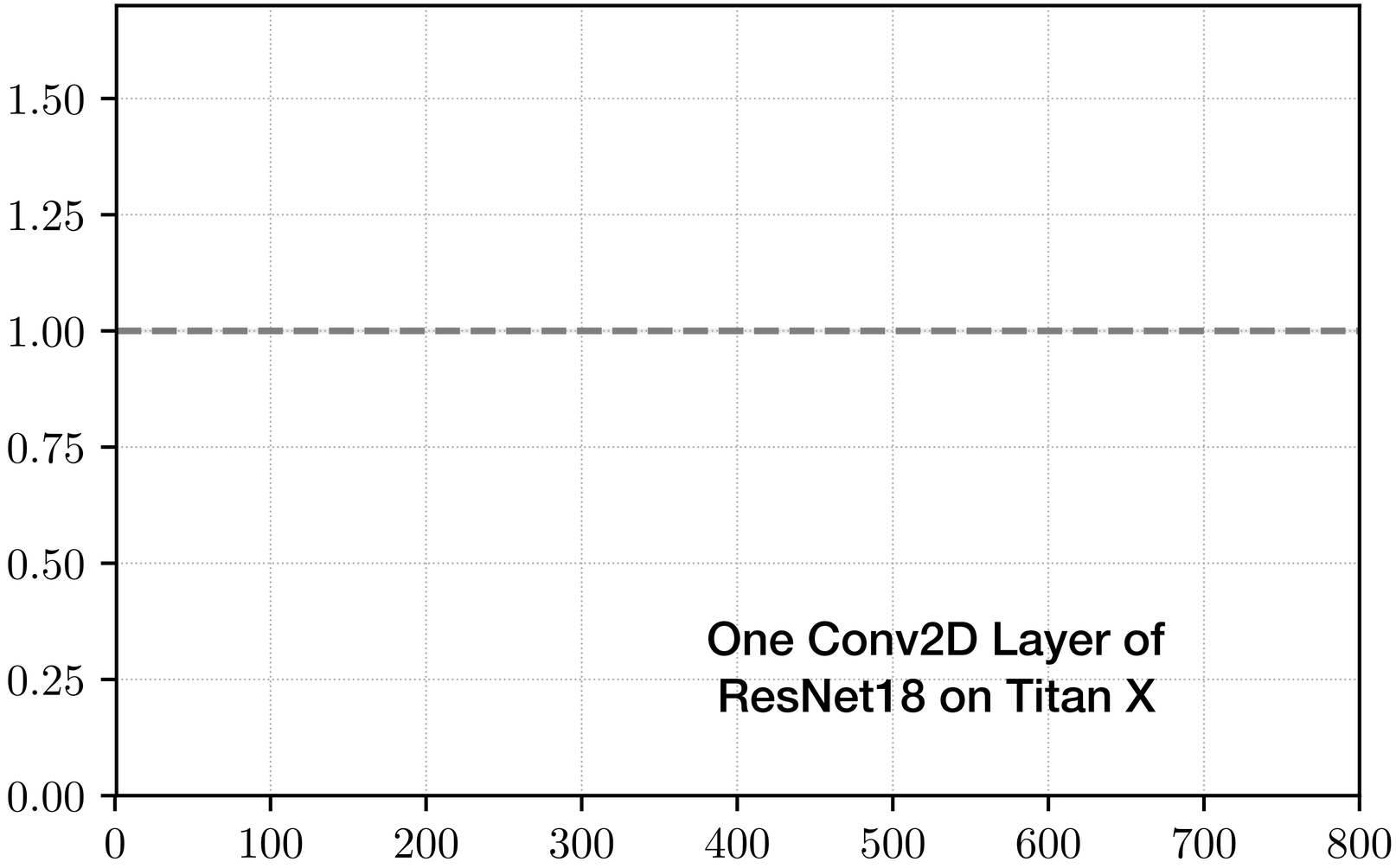


TreeGRU

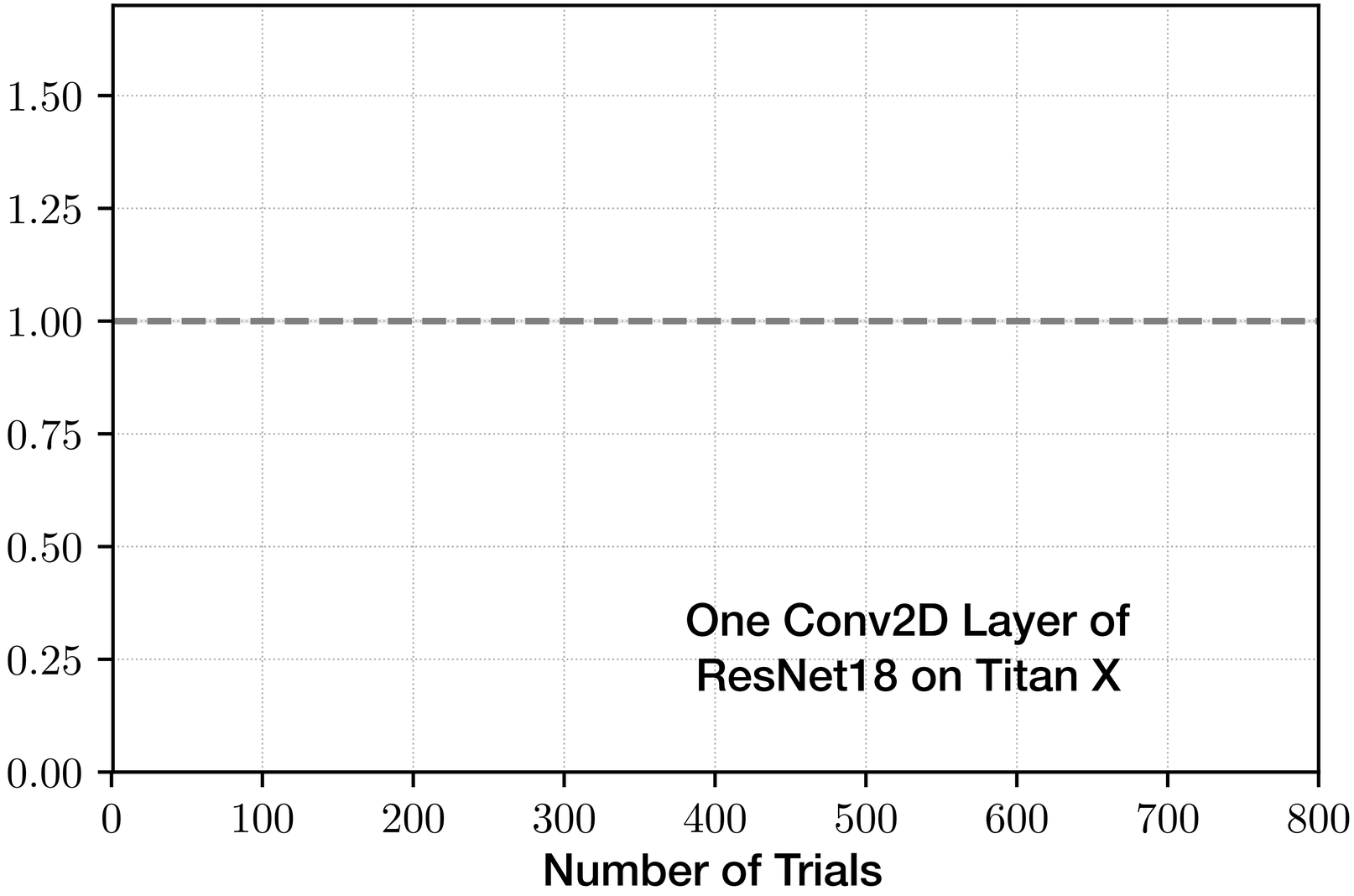
Effectiveness of ML based Model



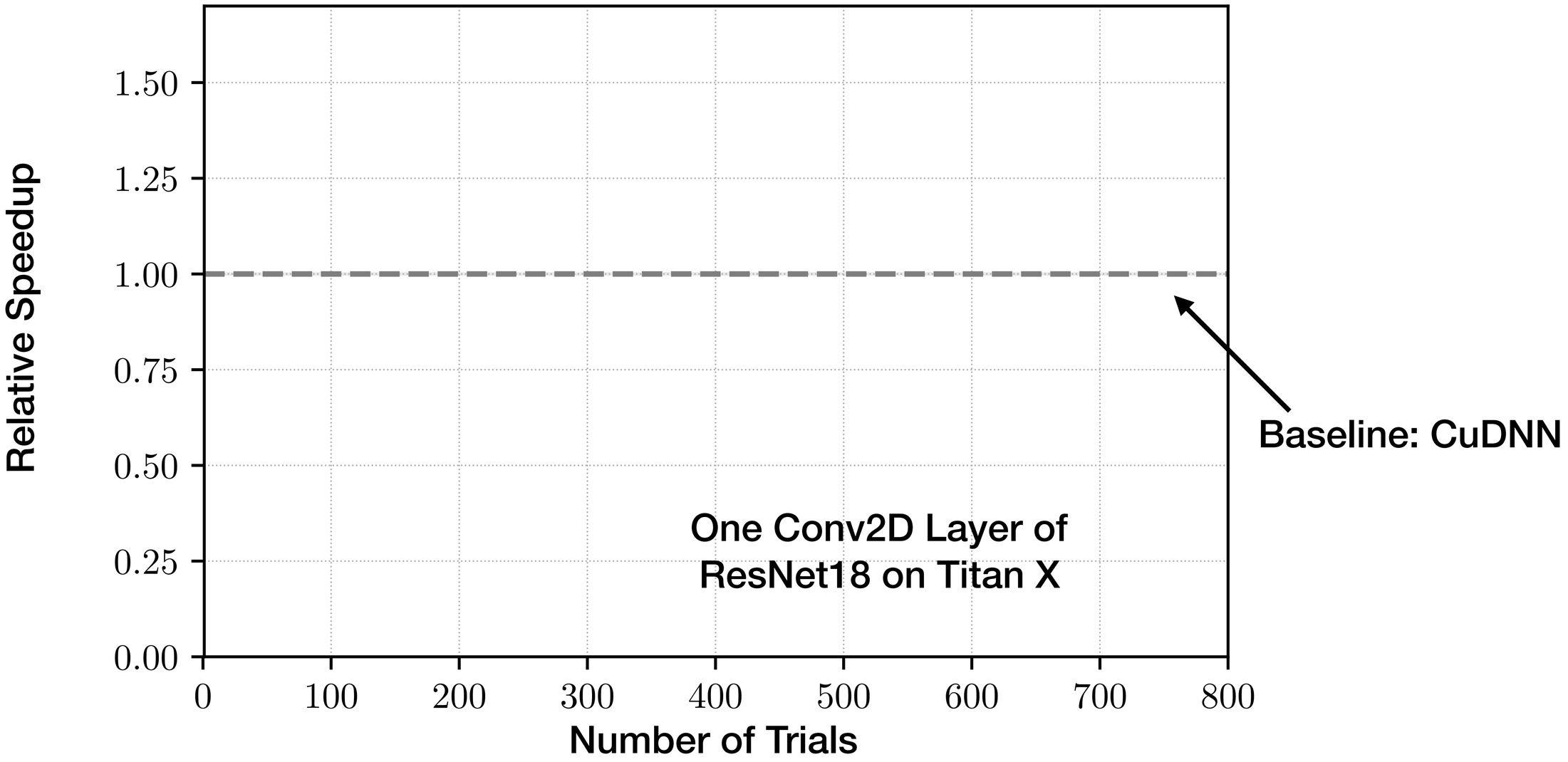
Effectiveness of ML based Model



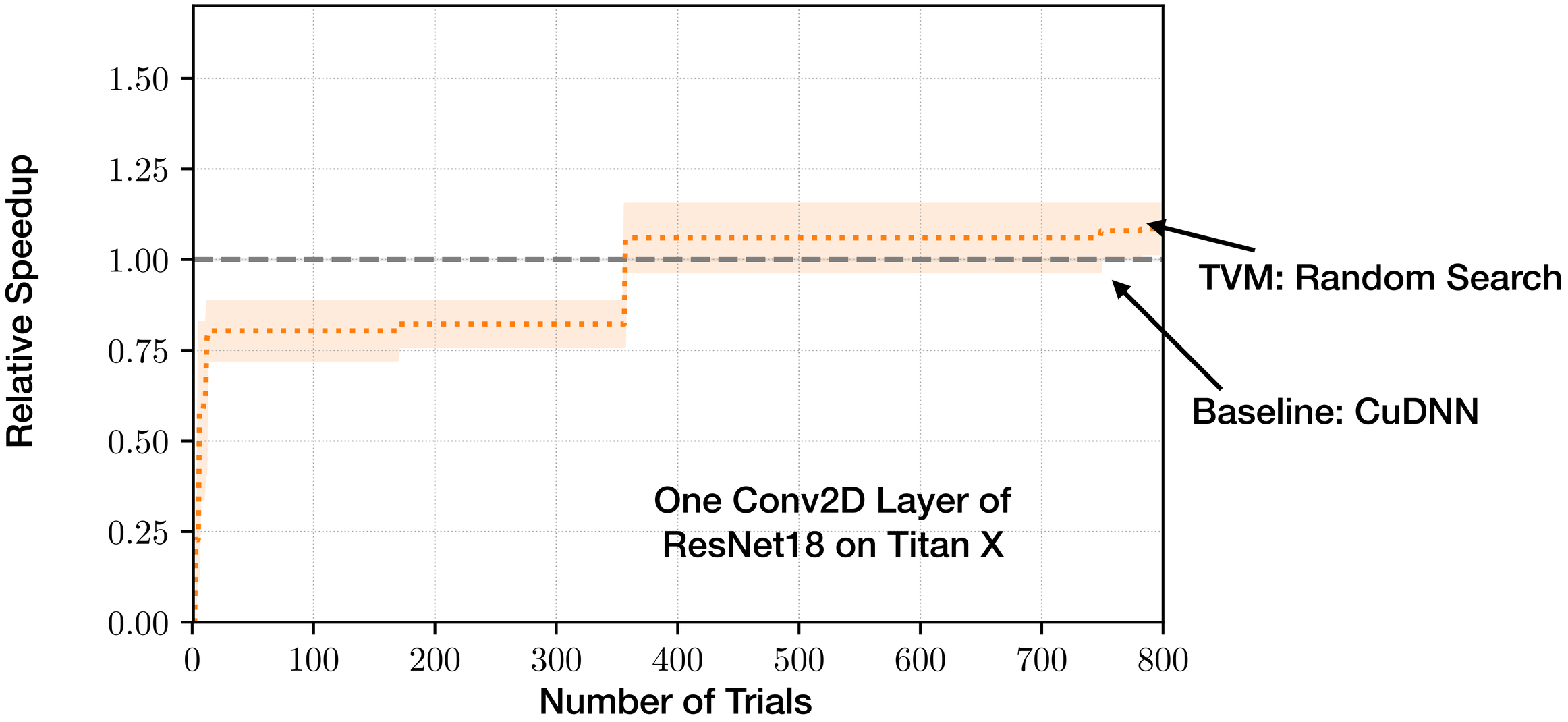
Effectiveness of ML based Model



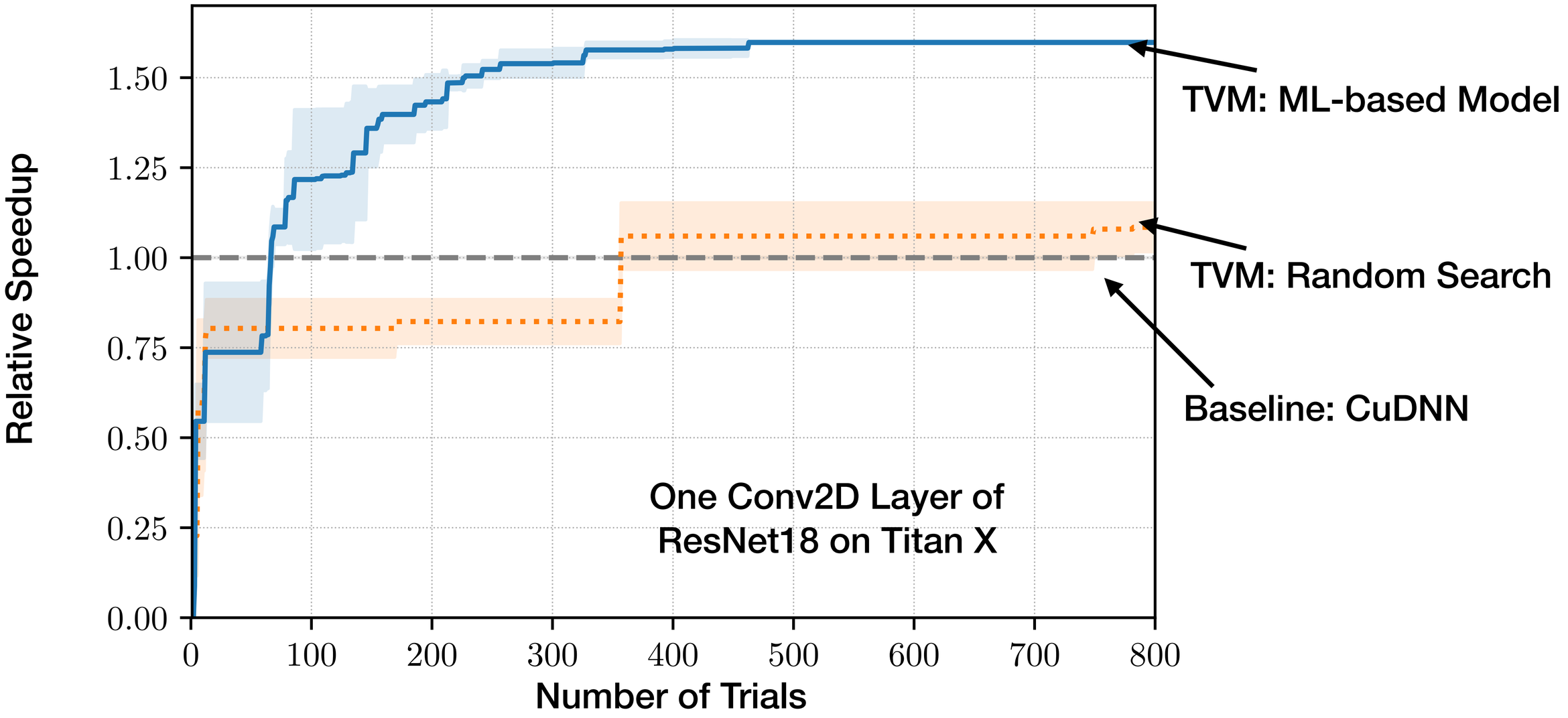
Effectiveness of ML based Model



Effectiveness of ML based Model



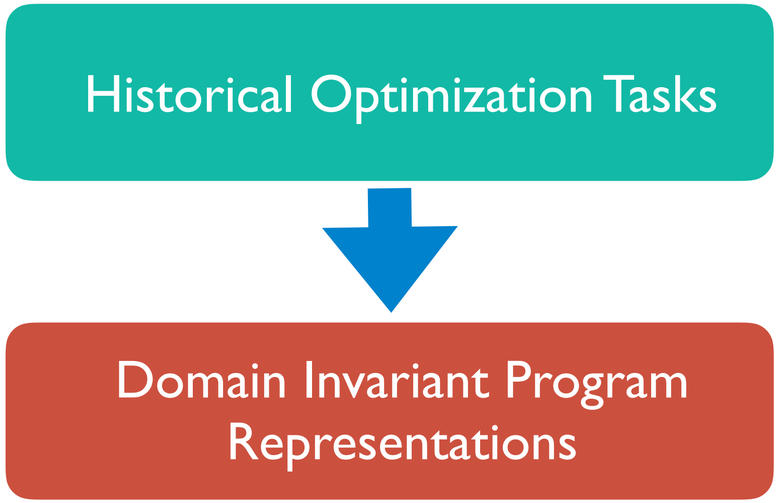
Effectiveness of ML based Model



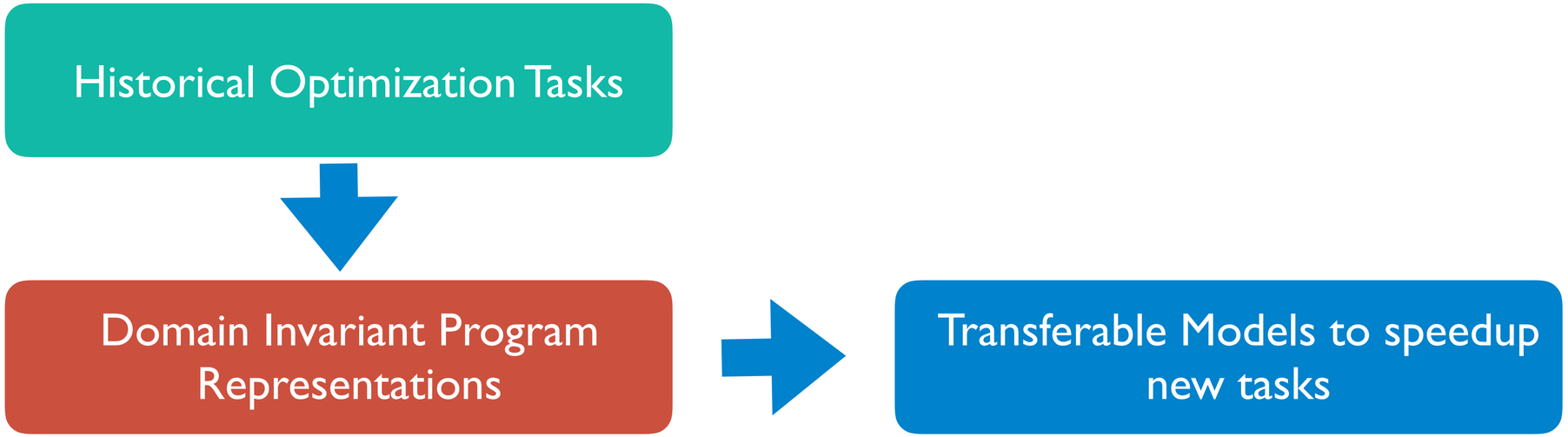
Transfer Learning Among Different Workloads

Historical Optimization Tasks

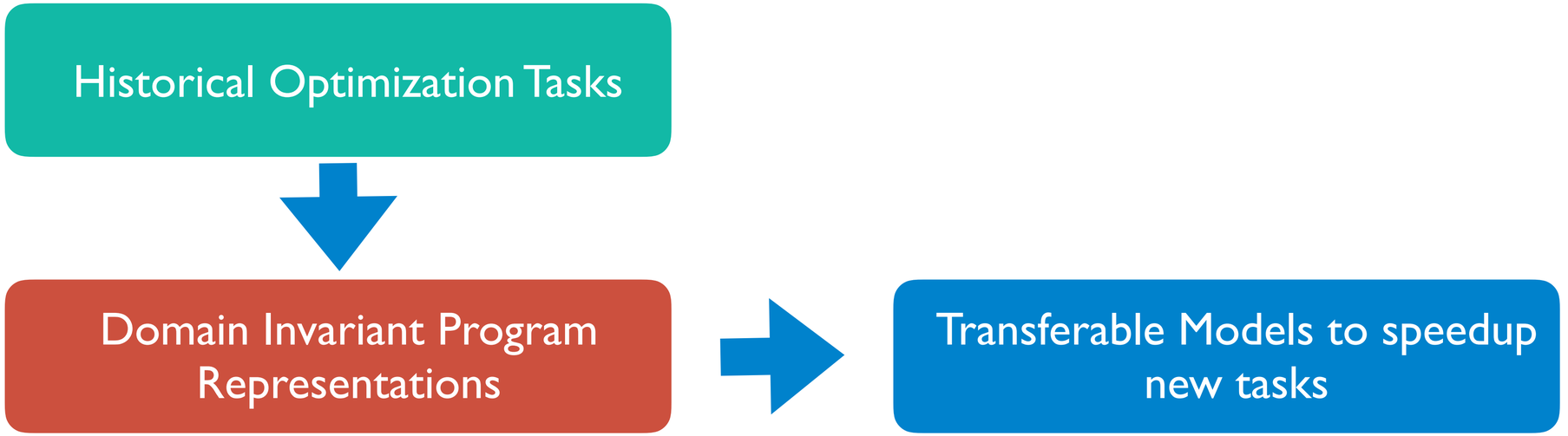
Transfer Learning Among Different Workloads

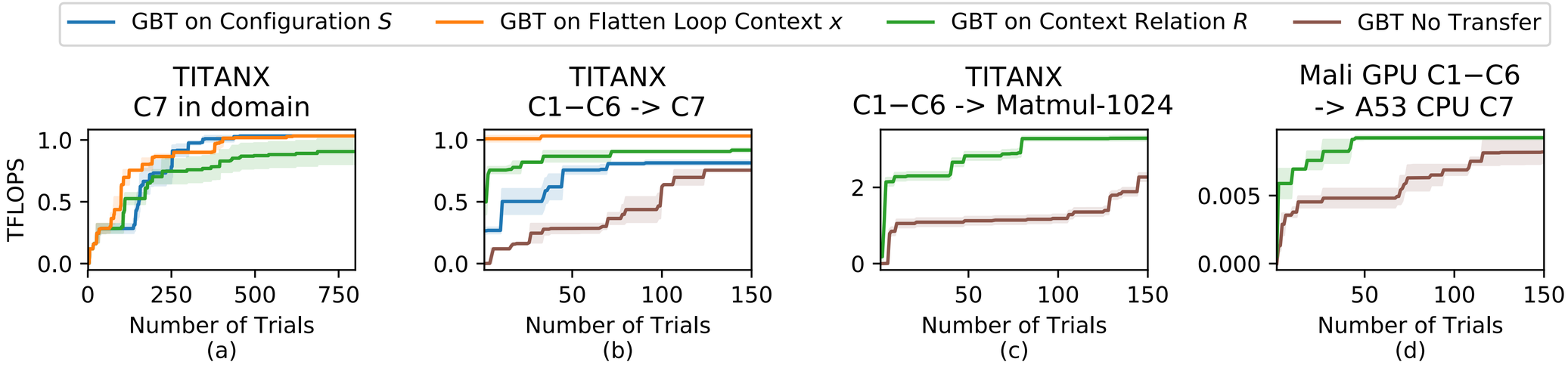


Transfer Learning Among Different Workloads

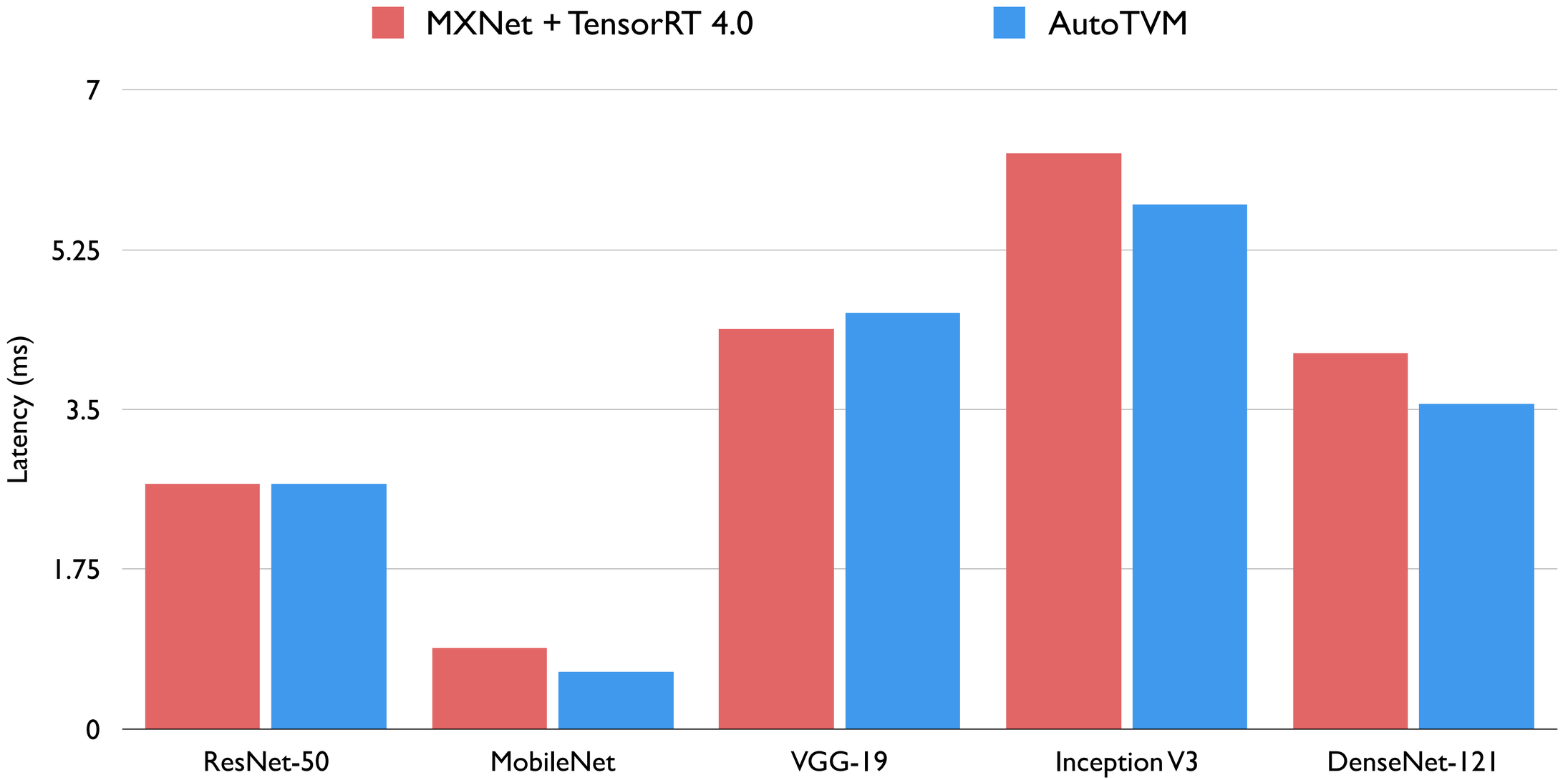


Transfer Learning Among Different Workloads

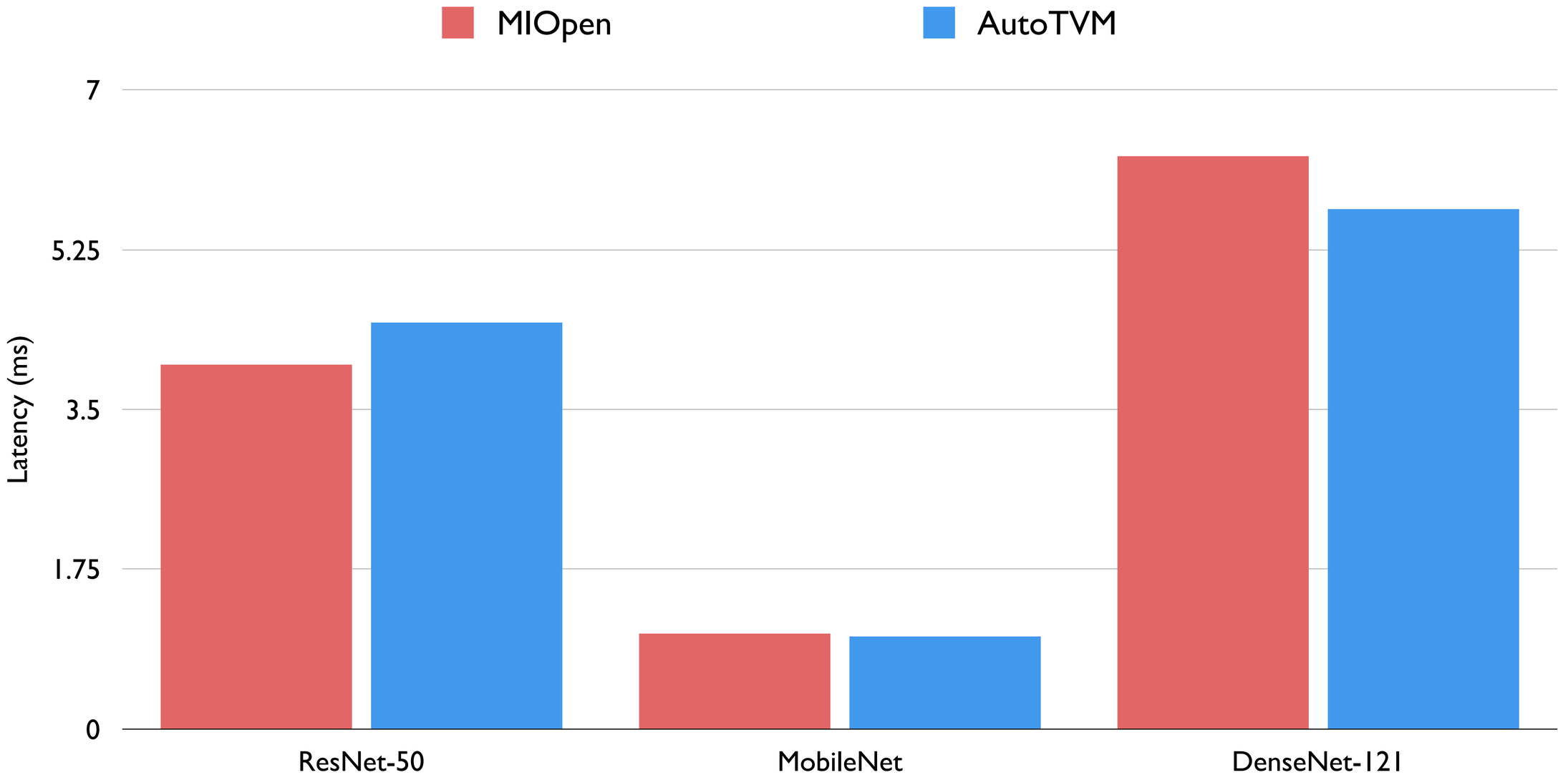




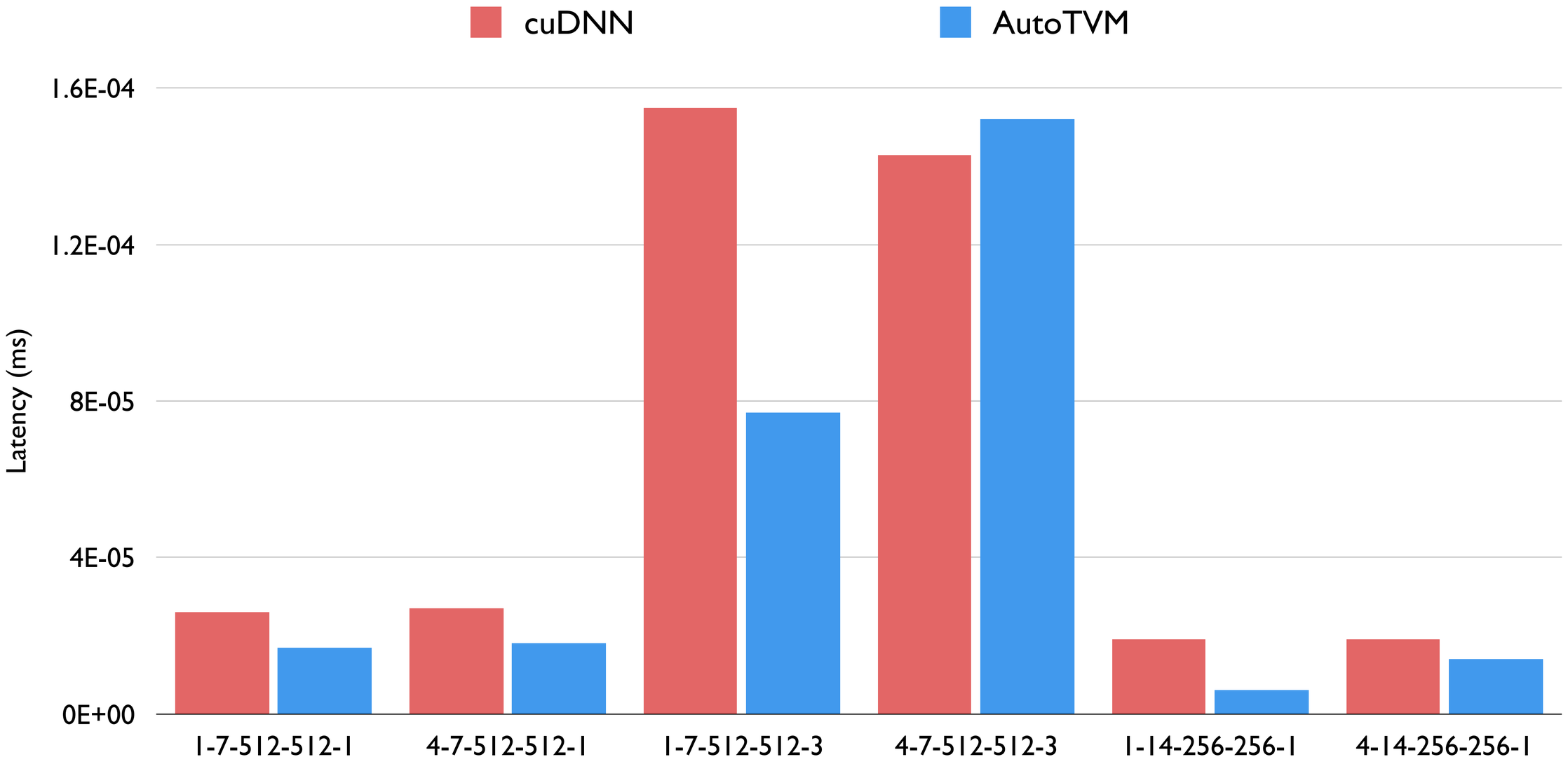
NVIDIA GPU Optimization (GTX 1080 Ti)



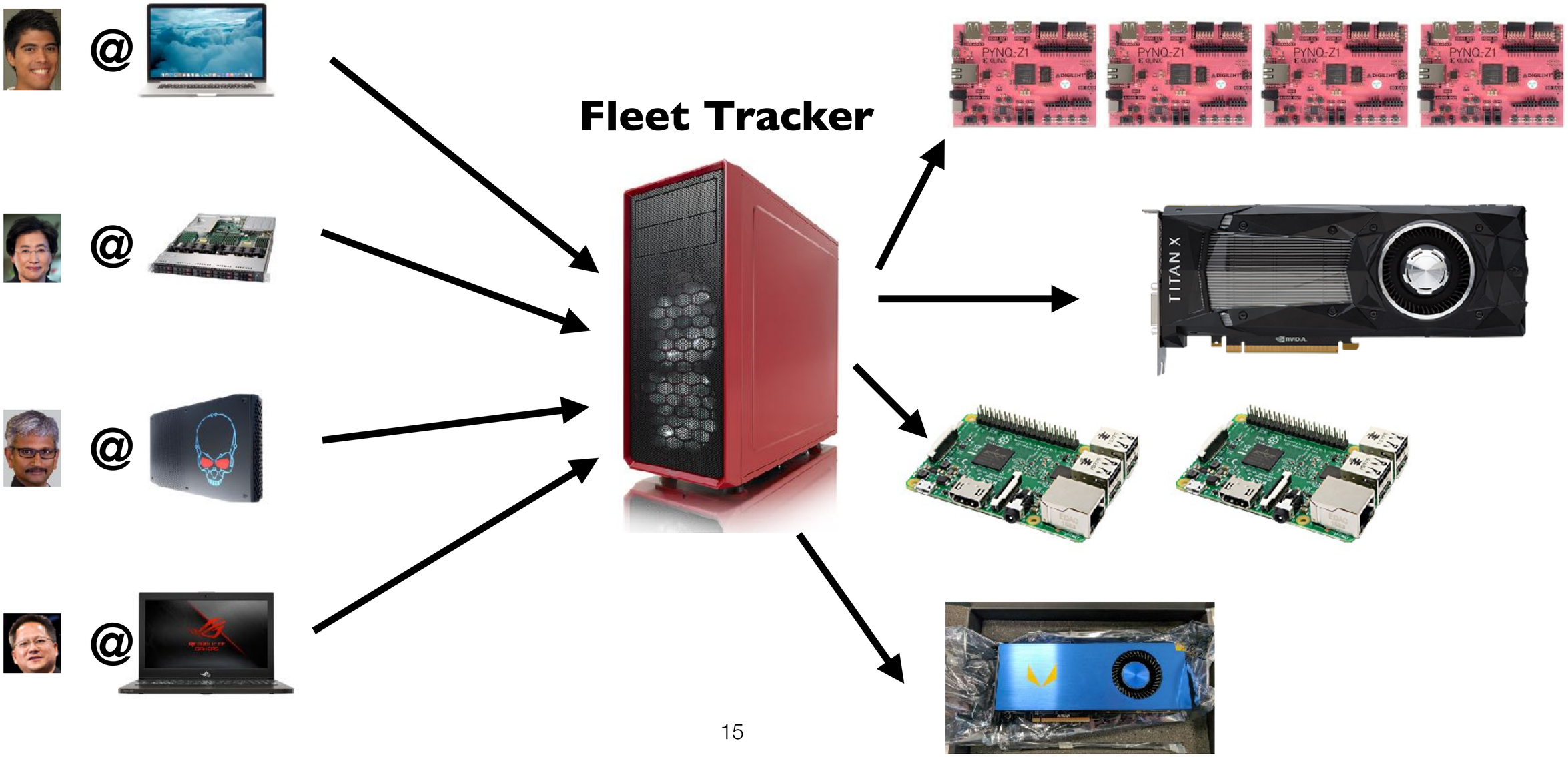
AMD GPU Optimization (Vega FE)



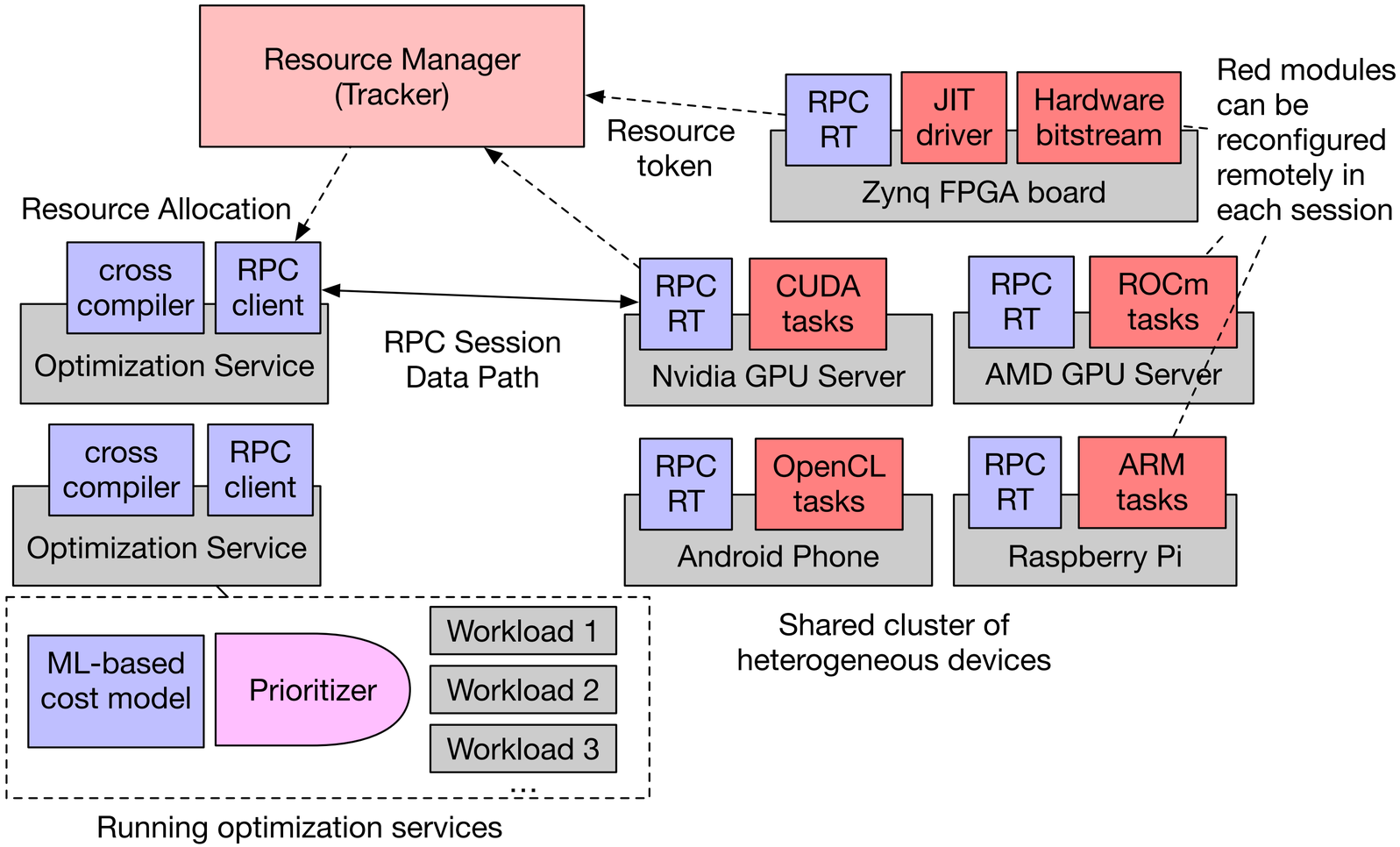
Bonus (INT8, GTX 1080)



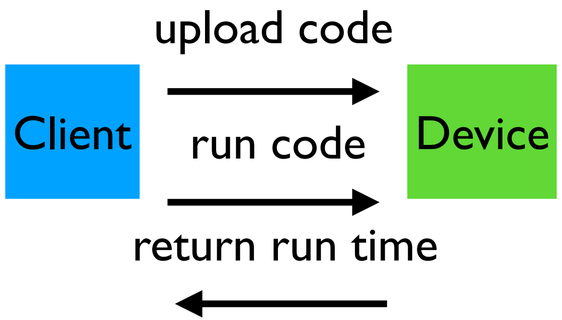
High Level: Scaling Automatic Performance Profiling



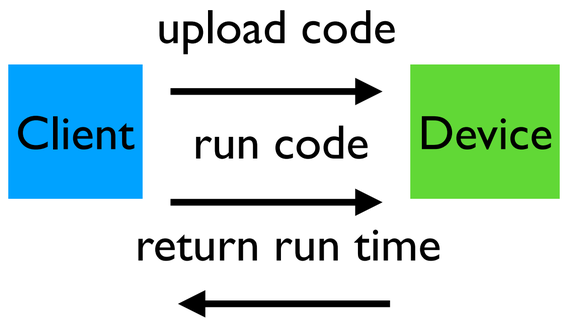
Low Level: Portable RPC Tracker + Server



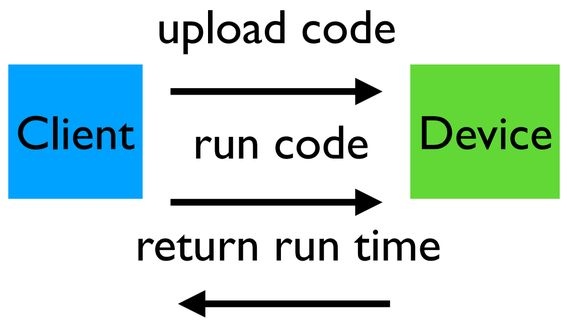
RPC Communication Flow



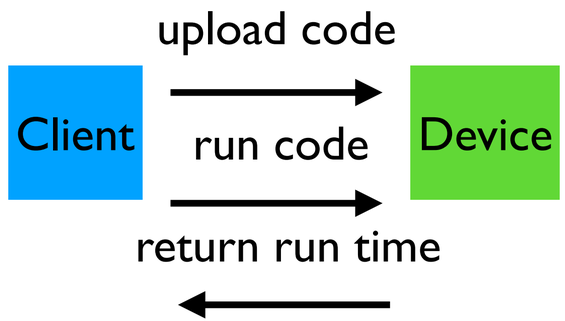
RPC Communication Flow



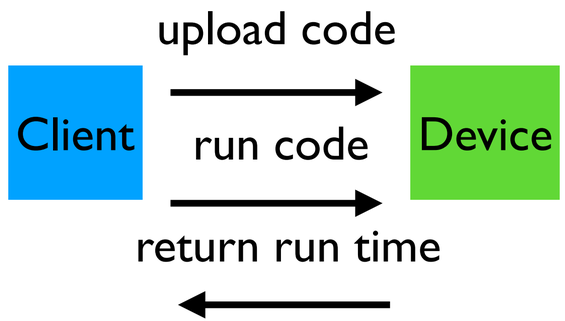
RPC Communication Flow



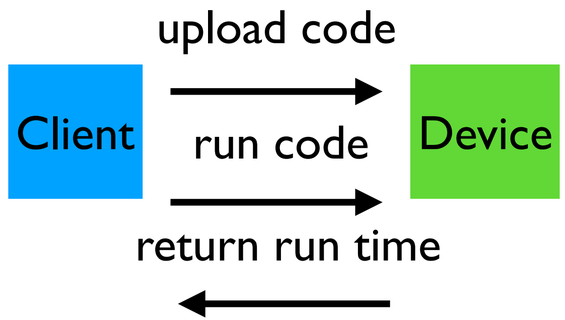
RPC Communication Flow



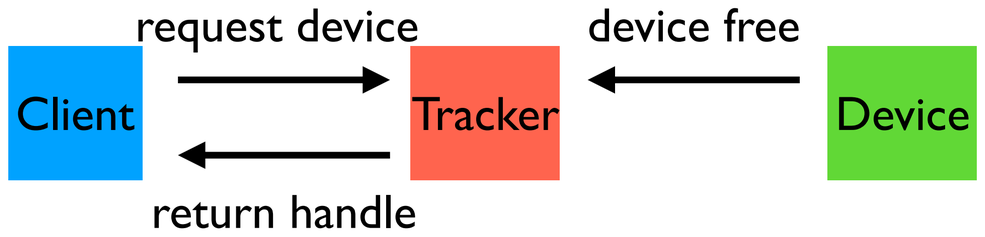
RPC Communication Flow

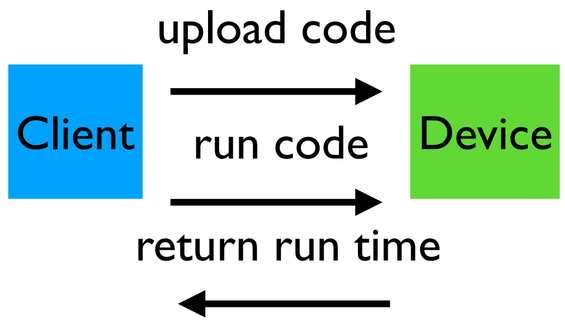


RPC Communication Flow

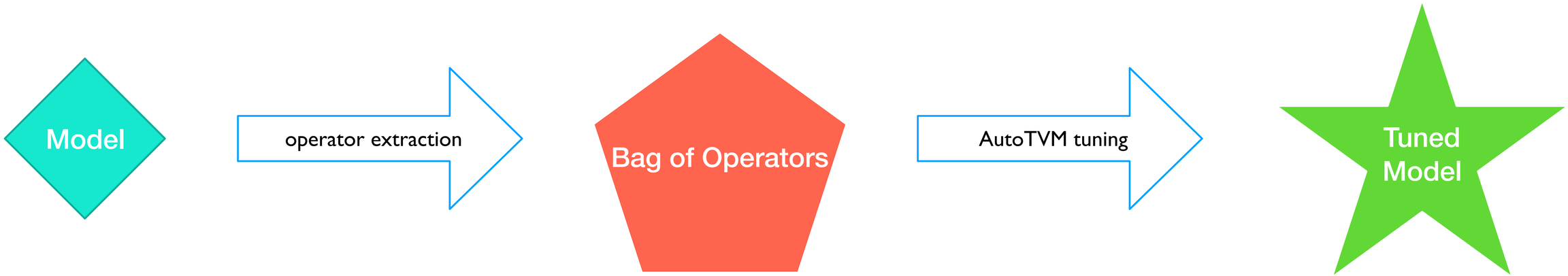


RPC Communication Flow





Model to Tuned Implementation



Next: Autoscheduler, Lianmin @ 16:30