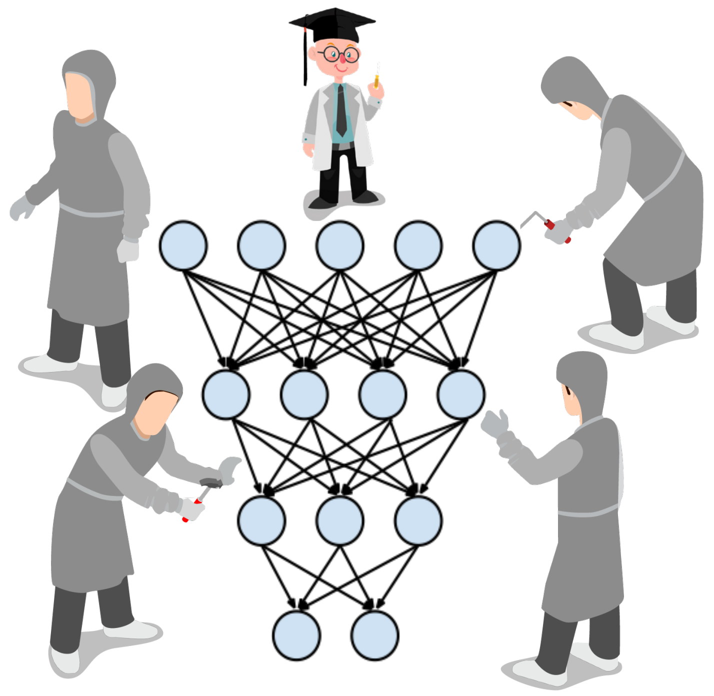
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning

Han Cai, Tianzhe Wang, Zhanghao Wu, Kuan Wang, Ji Lin, Song Han

Massachusetts Institute of Technology

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search
and Quantization-Aware Fine-tuning, ICCV Workshop'2019

From Manual Design to Automatic Design

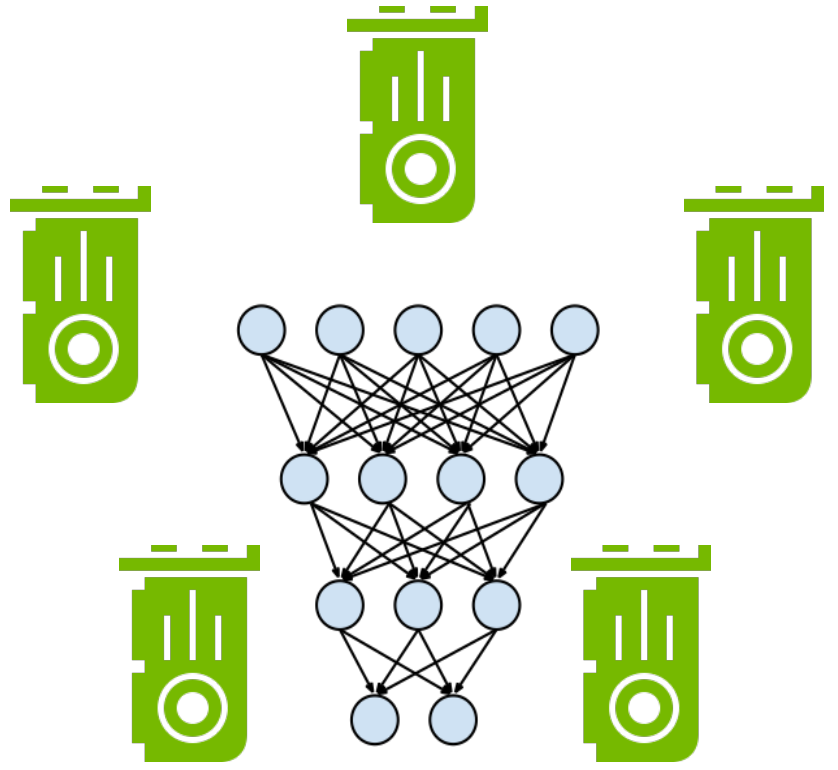


Use Human Expertise

Manual Architecture Design

VGGNets
Inception Models
ResNets
DenseNets
....

Computational Resources



Use Machine Learning (NAS)

Automatic Architecture Search

Reinforcement Learning
Neuro-evolution
Bayesian Optimization
Monte Carlo Tree Search
...

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR’19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop’2019

From General Design to Specialized CNN

Previous Paradigm:
One CNN for all platforms.

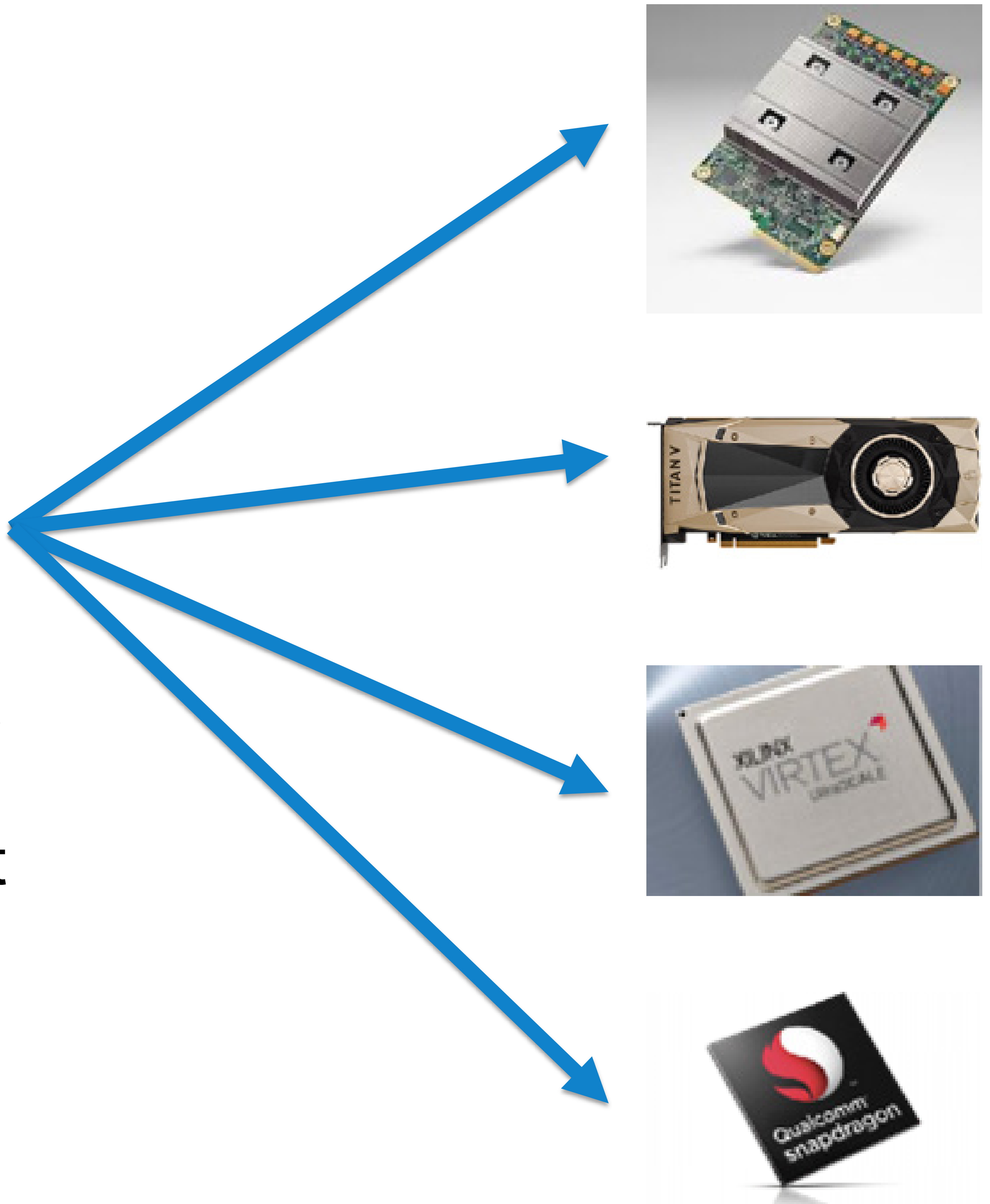
ResNet

Inception

DenseNet

MobileNet

ShuffleNet



Sub-optimal, different in:

- Degree of parallelism
- Cache size
- Memory BW
- ...

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search
and Quantization-Aware Fine-tuning, ICCV Workshop'2019

From General Design to Specialized CNN

Previous Paradigm:
One CNN for all platforms.

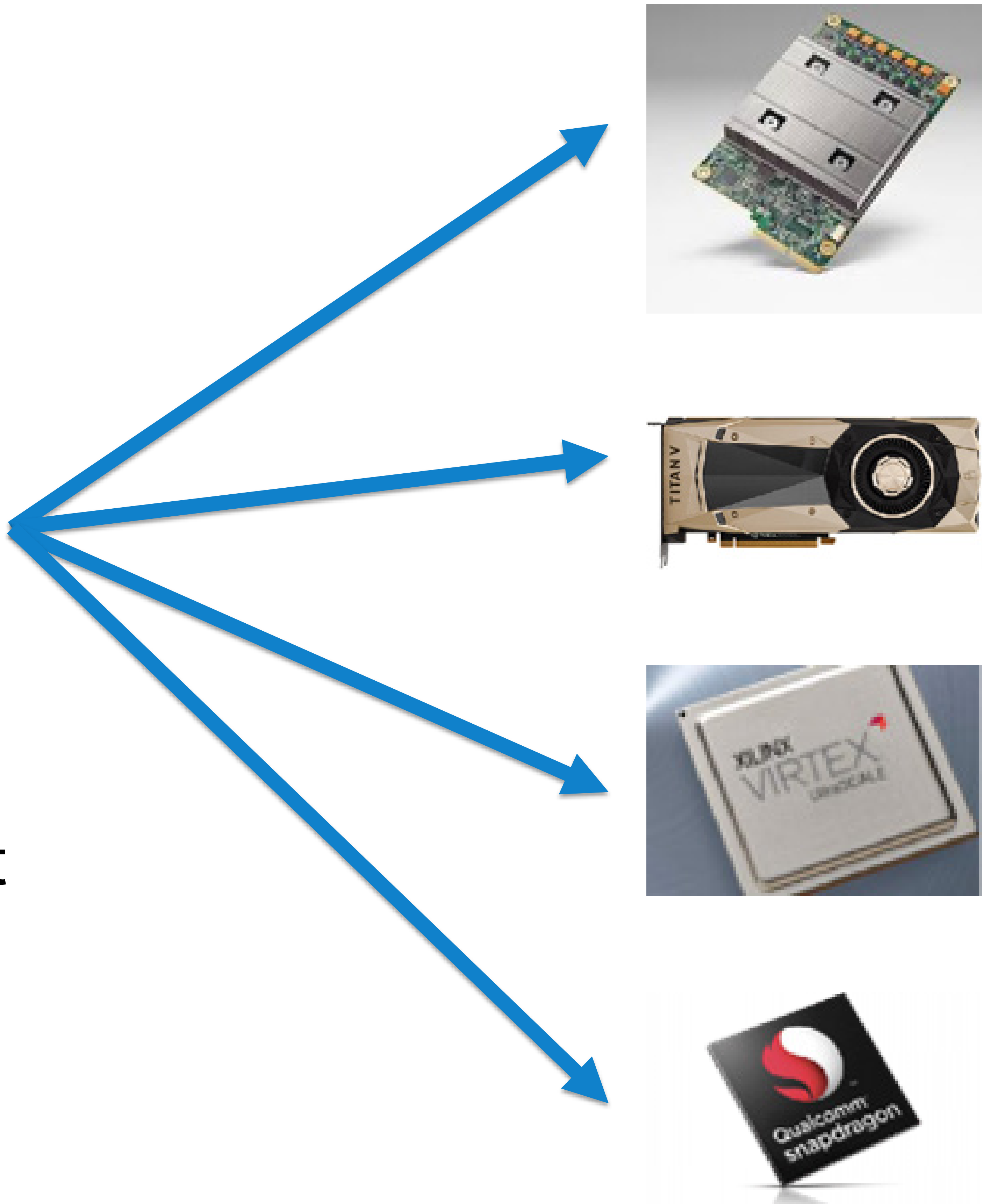
ResNet

Inception

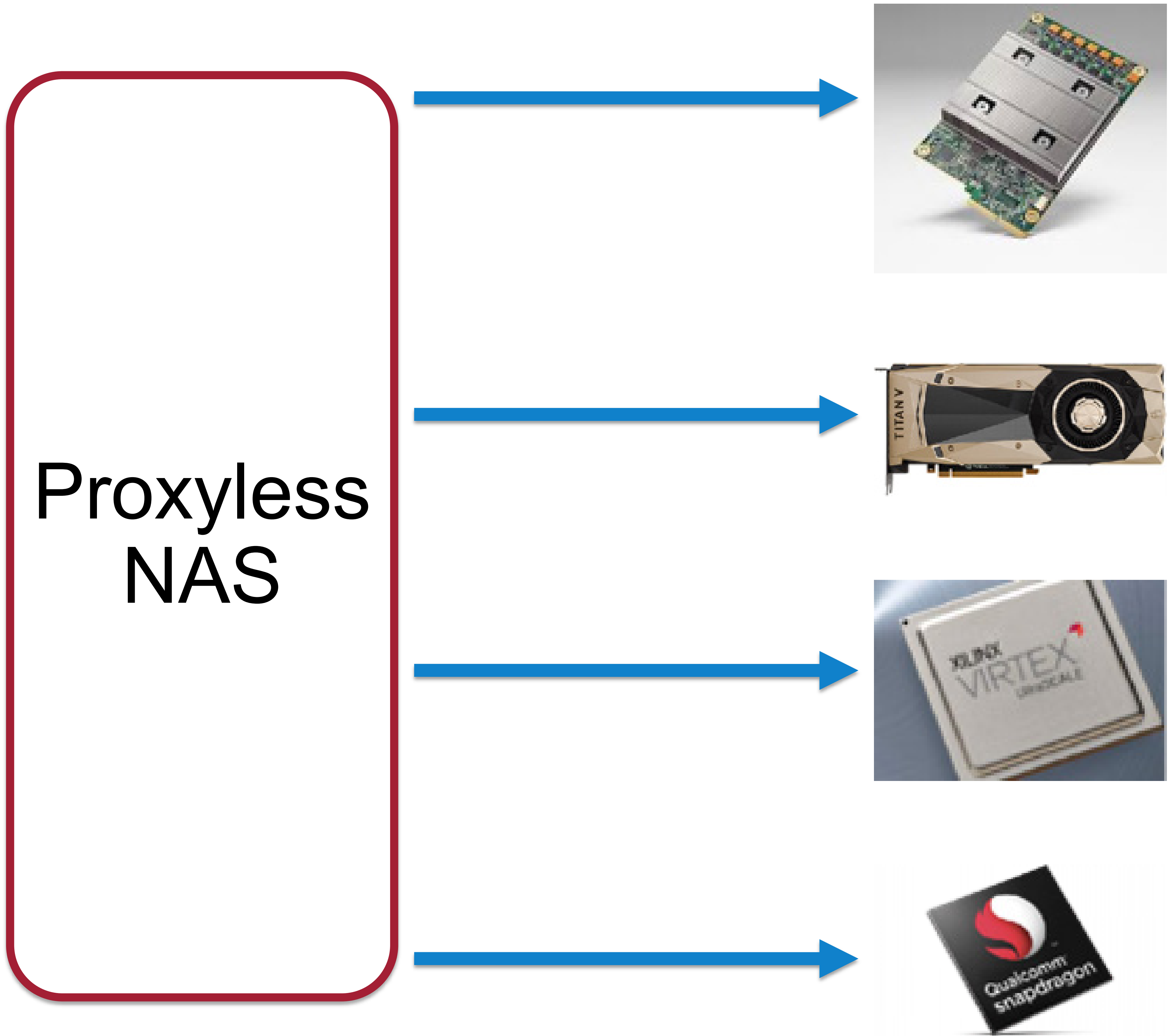
DenseNet

MobileNet

ShuffleNet

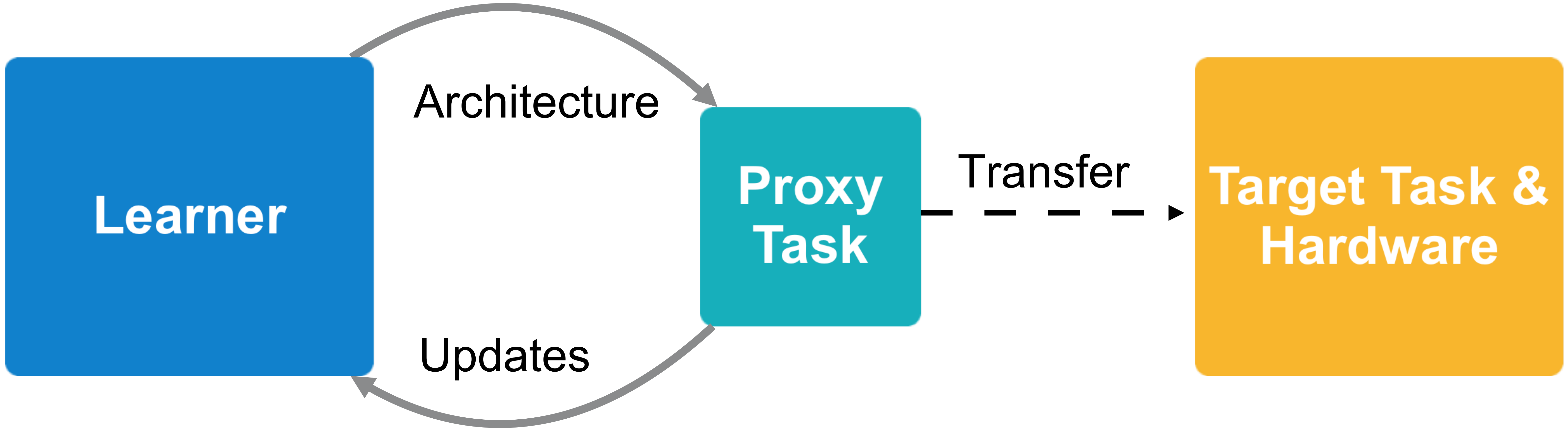


Proxyless NAS:
Customize CNN for each platform.



ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

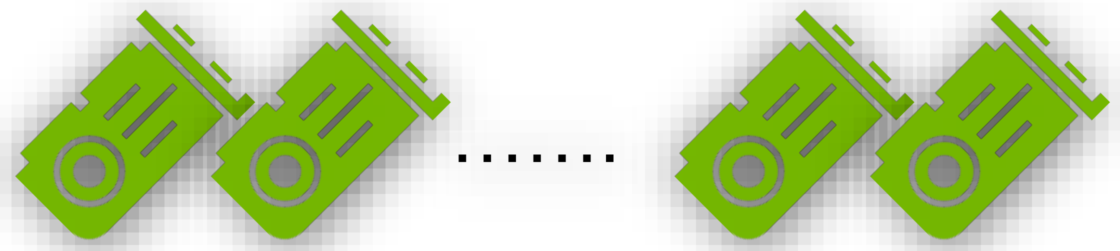
Conventional NAS: Computation Expensive, thus Proxy-Based



Current neural architecture search (NAS) is **VERY EXPENSIVE**.

- NASNet: 48,000 GPU hours ≈ 5 years on single GPU
- DARTS: 100Gb GPU memory* ≈ 9 times of modern GPU

*if directly search on ImageNet, like us

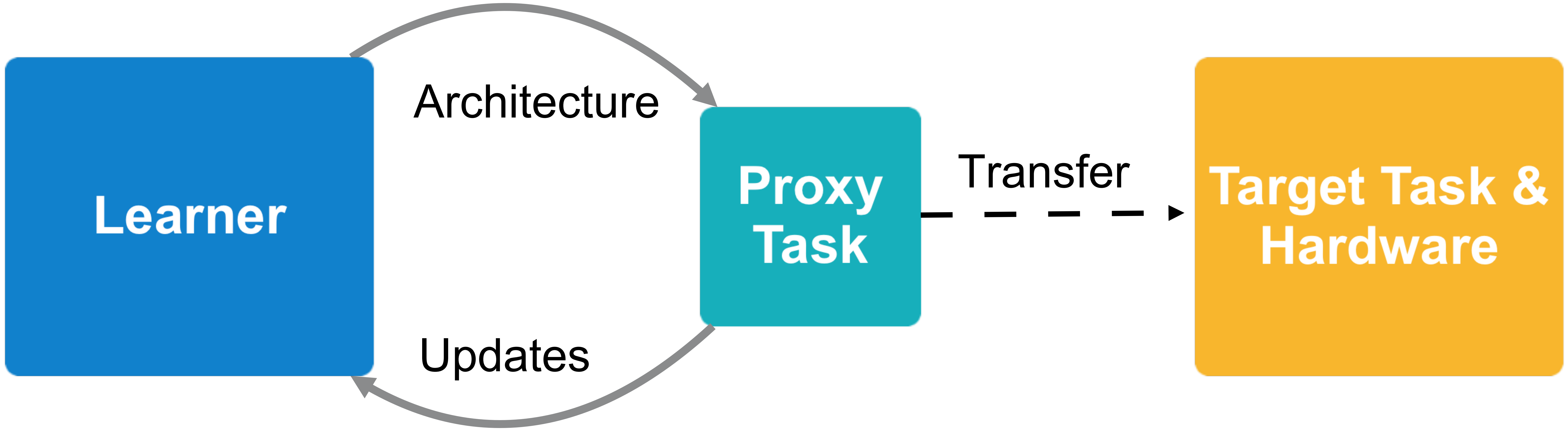


Therefore, previous work have to utilize **proxy tasks:**

- CIFAR-10 -> ImageNet
- Search a block -> stack to build a full net
- Fewer epochs training -> full training

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Conventional NAS: Computation Expensive, thus Proxy-Based

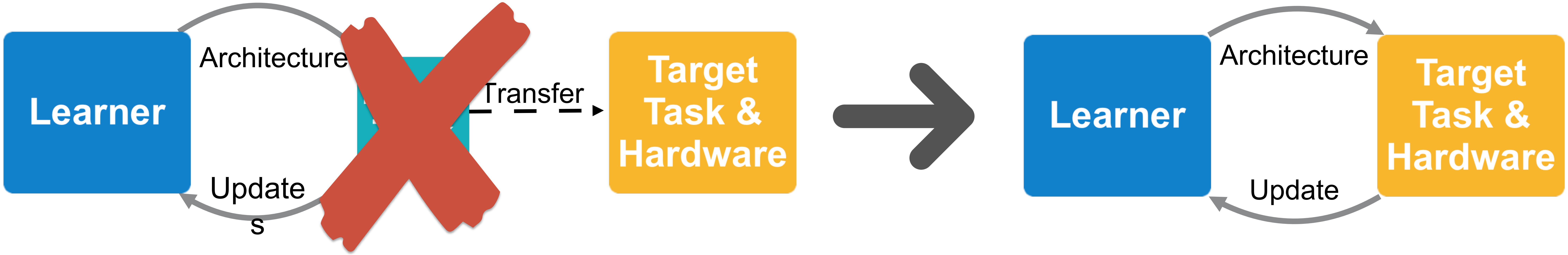


Limitations of Proxy

- **Suboptimal** for the target task
- Blocks are forced to **share the same structure**.
- Cannot optimize for **specific hardware**.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Proxyless, Save GPU Hours by 200x

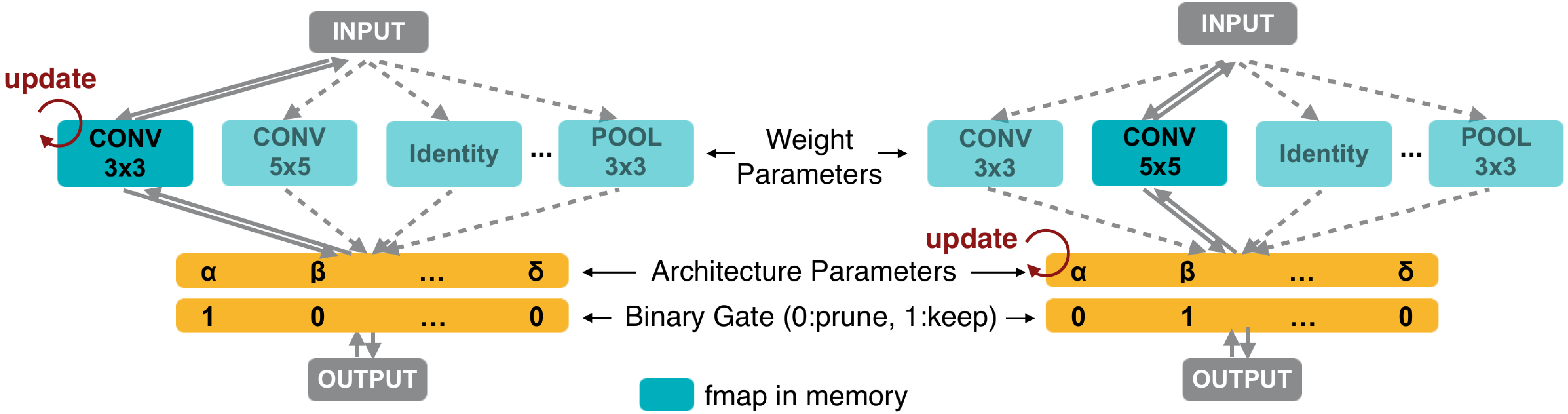


Goal: Directly learn architectures on the **target task** and **hardware**. We achieved by

1. Reducing the cost of NAS (GPU hours and memory) to the **same** level of regular training.
2. Cooperating **hardware feedback** (e.g. latency) into the search process.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search
and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Save GPU Hours



(1) Update weight parameters

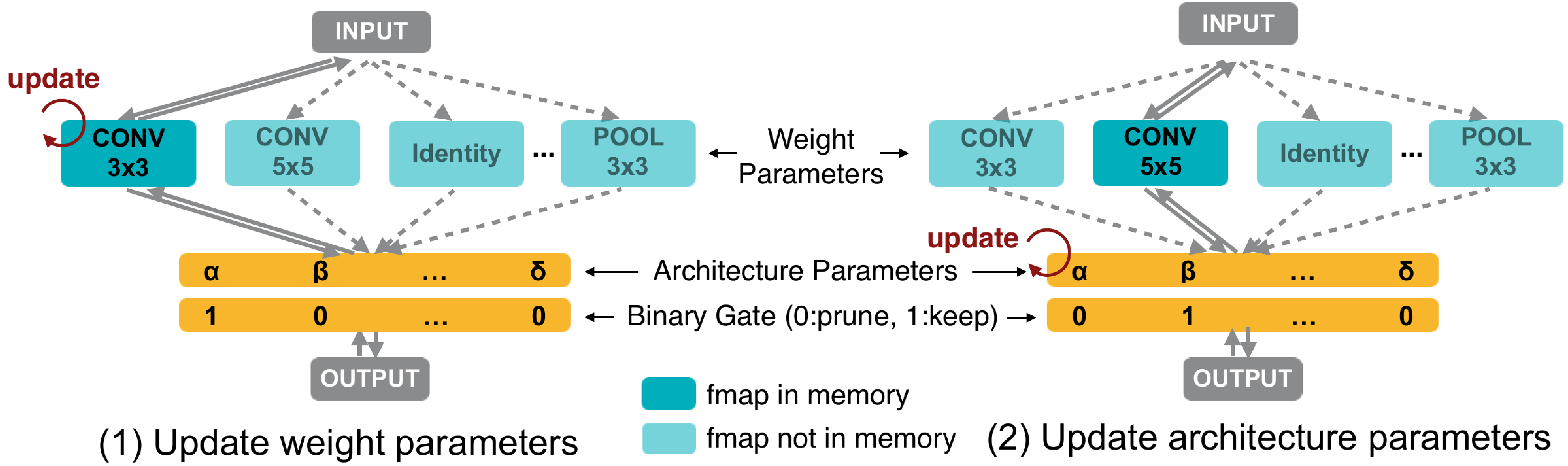
(2) Update architecture parameters

Simplify NAS to be a **single training process** of a over-parameterized network.

Build the cumbersome network **with all candidate paths**. No meta controller.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Save GPU Memory



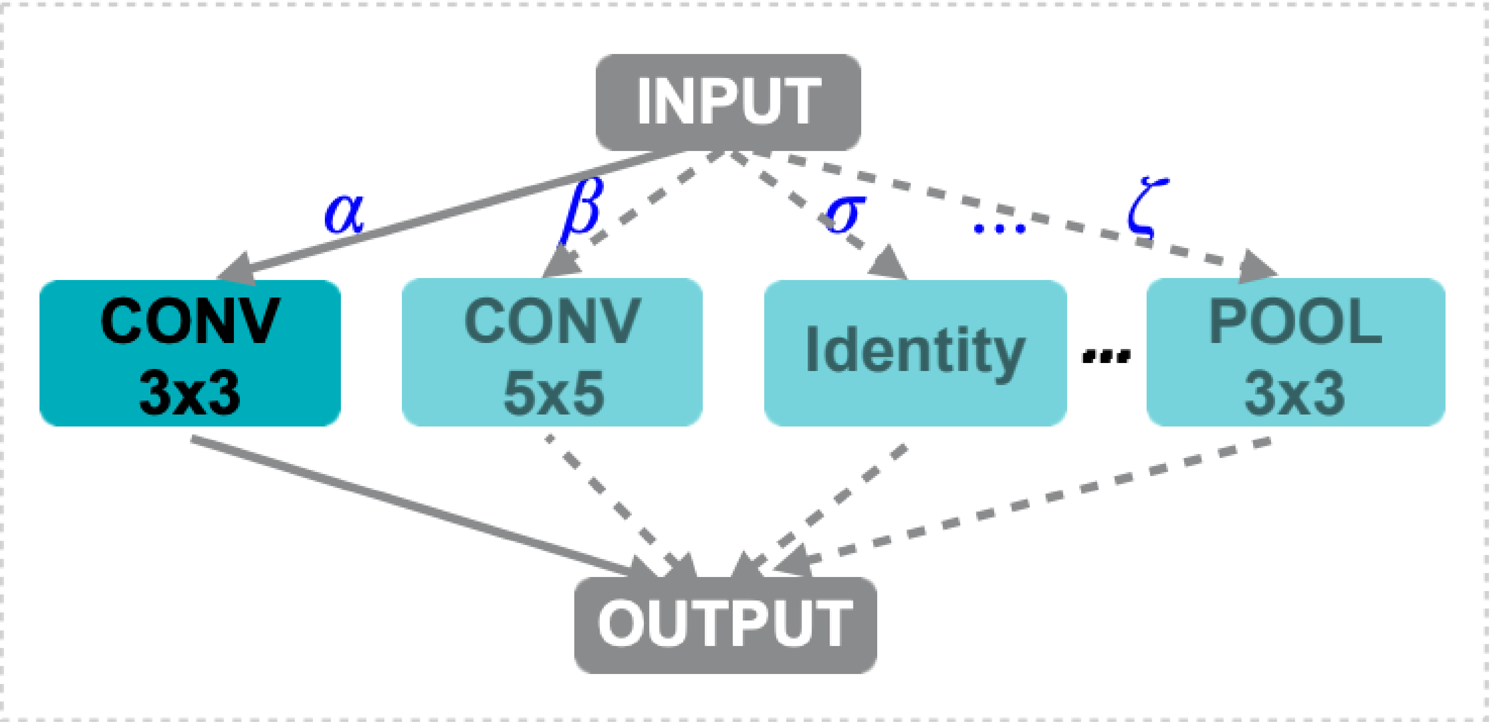
Binarize the architecture parameters and allow only **one path of activation** in memory.

We propose **gradient-based** and **RL** methods to update the architecture parameters.

Reduce the memory footprint from **O(N) to O(1)**.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

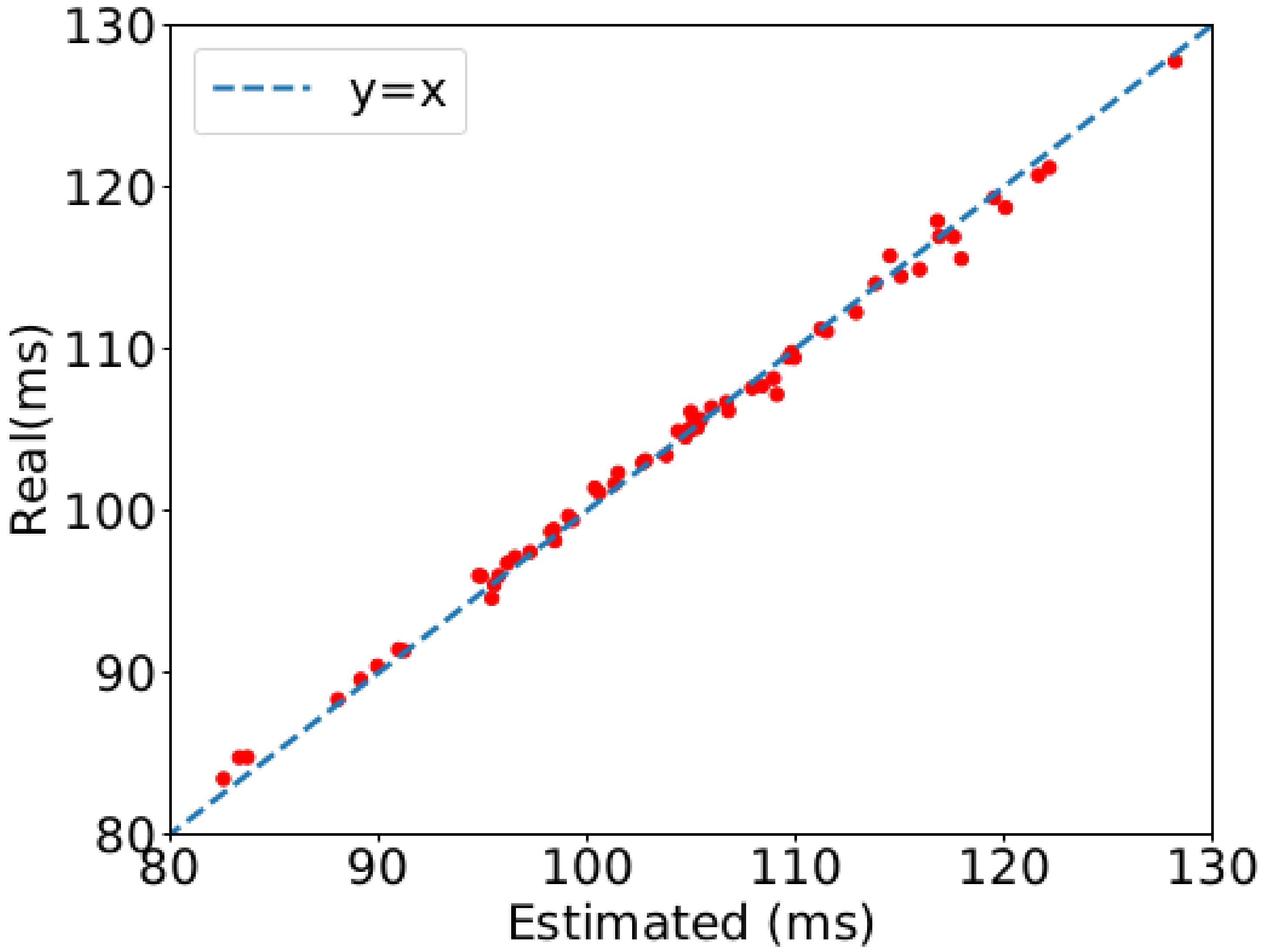
Direct Search on Target Hardware:
Making Latency Differentiable



$$\mathbb{E}[\text{Latency}] = \alpha \times F(\text{conv\_3x3}) + \beta \times F(\text{conv\_5x5}) + \sigma \times F(\text{identity}) + \ldots\ldots + \zeta \times F(\text{pool\_3x3})$$

$$\mathbb{E}[\text{latency}] = \sum_i \mathbb{E}[\text{latency}_i]$$

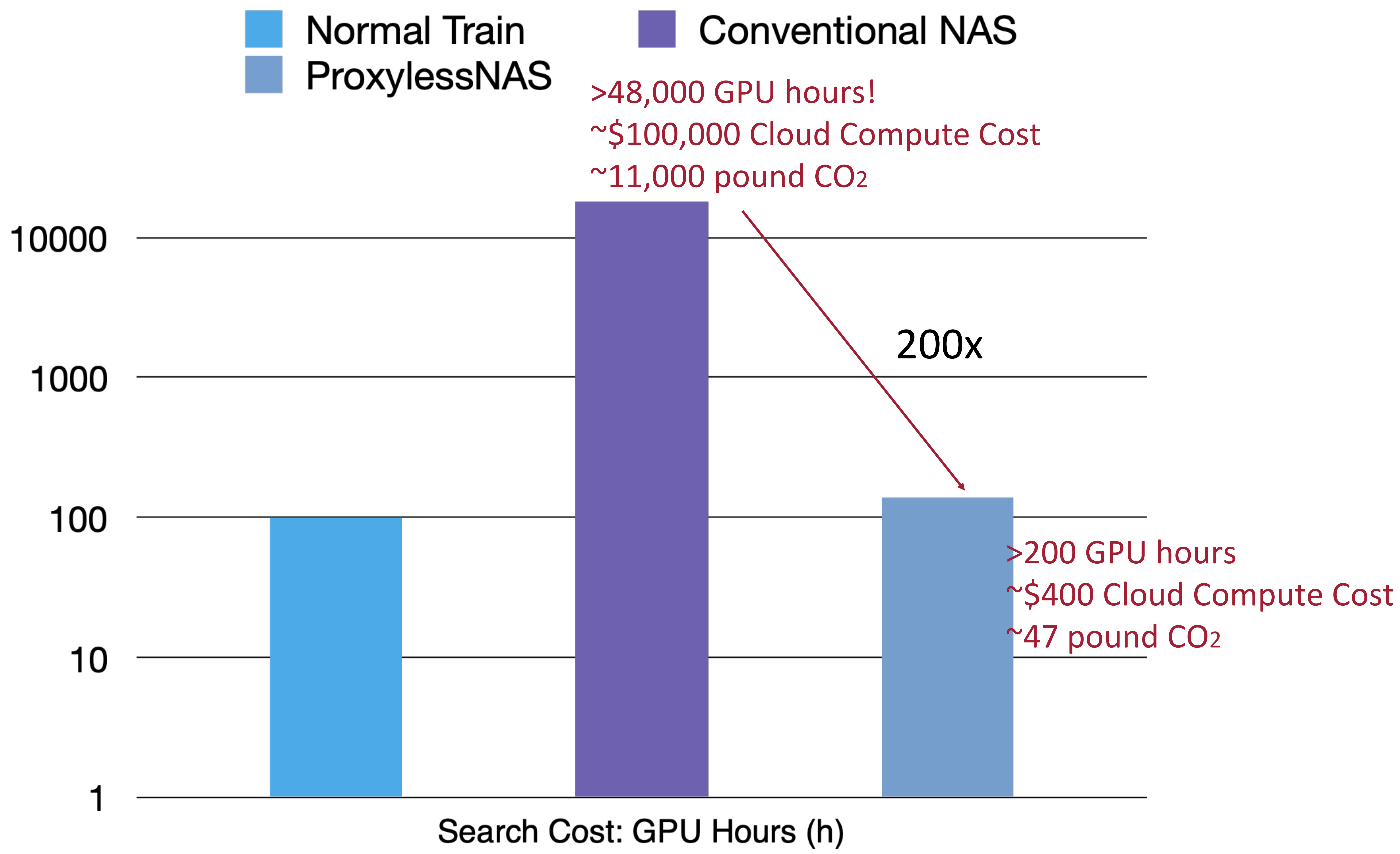
$$Loss = Loss_{CE} + \lambda_1 ||w||_2^2 + \lambda_2 \mathbb{E}[\text{latency}]$$



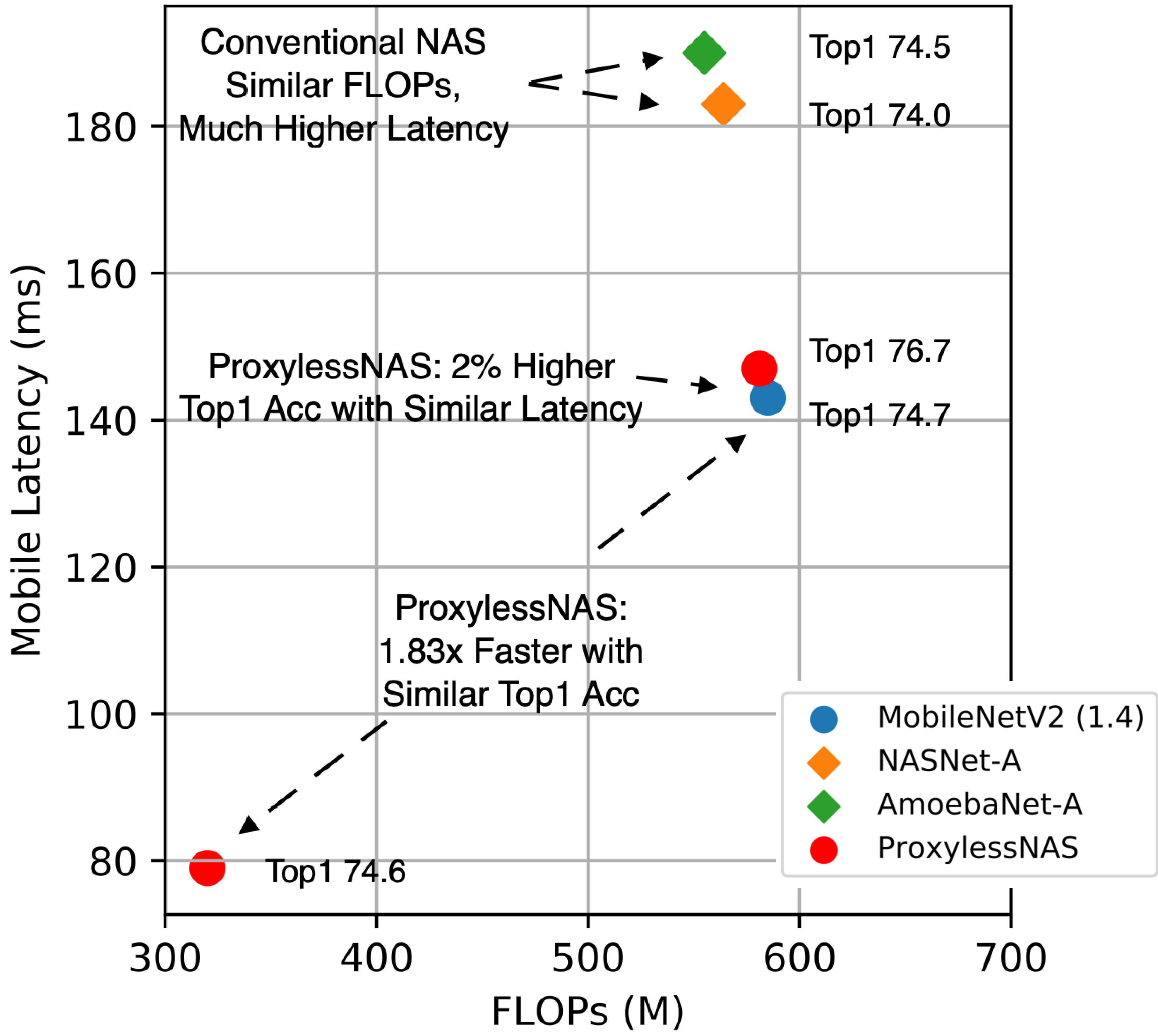
- **Mobile farm infrastructure** is expensive and slow.
- Use the **latency estimation model** as an economical alternative
- Optimize during search stage use **Gradient**.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Efficiently search a model

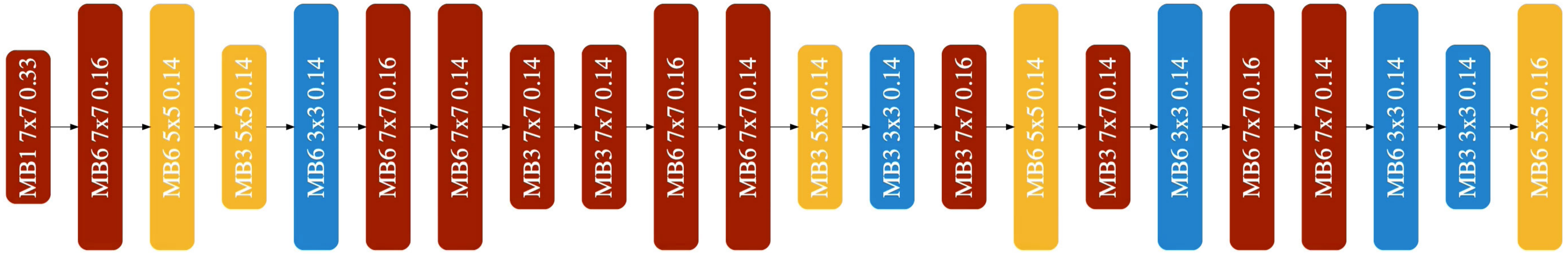


Search an efficient model

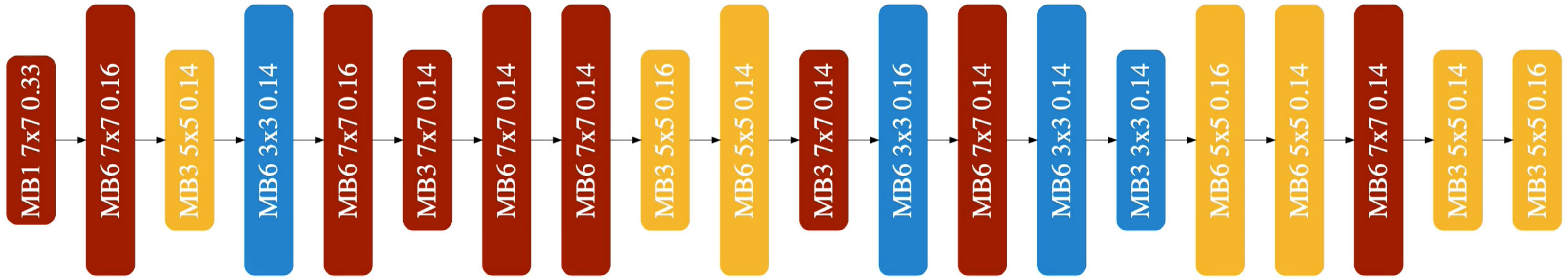


ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

The History of Architectures

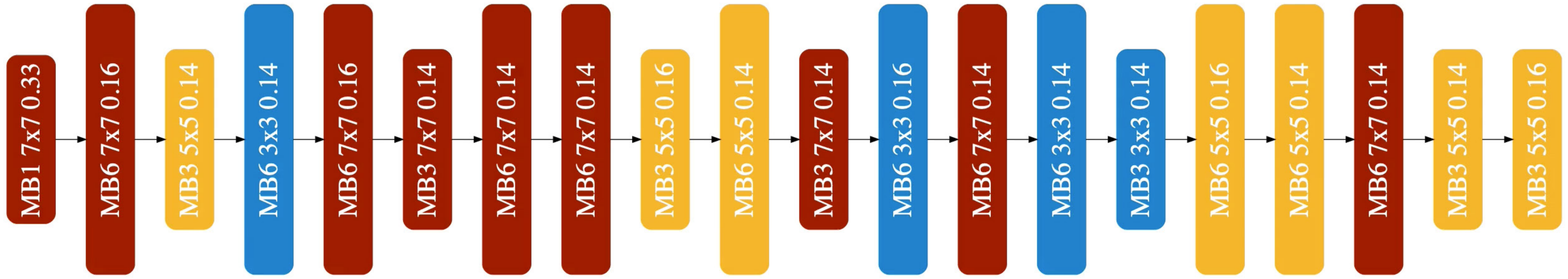


(1) The history of finding efficient Mobile model



Epoch-00

(2) The history of finding efficient CPU model



(3) The history of finding efficient GPU model

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR’19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop’2019

Results for LPIRC

Model	Setting	Accuracy	Latency
MoblieNetV2	224-0.5	63.7%(65.4%)	28ms
MobileNetV2	192-0.75	67.4%(68.7%)	36ms
MobileNetV2	160-1.0	67.4%(68.8%)	31ms
ProxylessNAS	224-0.5	65.7%(67.0%)	31ms
ProxylessNAS	160-1.0	**69.2%**(70.3%)	35ms

Table 1. Results of 8-bit model using different preprocessing, the number in the bracket denotes the full-precision model's top-1 accuracy on ImageNet The latency is directly measured on Google Pixel 2. It takes only 200 GPU hours to find the specialized model with ProxylessNAS in the table.

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning, ICCV Workshop'2019

Open-source

- Both search code and models are released on Github:

```
# https://github.com/MIT-HAN-LAB/ProxylessNAS
from proxyless_nas import *
net = proxyless_cpu(pretrained=True)
net = proxyless_gpu(pretrained=True)
net = proxyless_mobile(pretrained=True)
```

Open-source

- ProxylessNAS is available on PyTorch Hub:

```
# https://pytorch.org/hub/pytorch_vision_proxylessnas
import torch
target_platform = 'proxyless_mobile'
net = torch.hub.load('mit-han-lab/ProxylessNAS',
    target_platform, pretrained=True)
```

Thank you!

MIT HAN LAB

Hardware, **A**I and **N**eural-nets

songhan@mit.edu