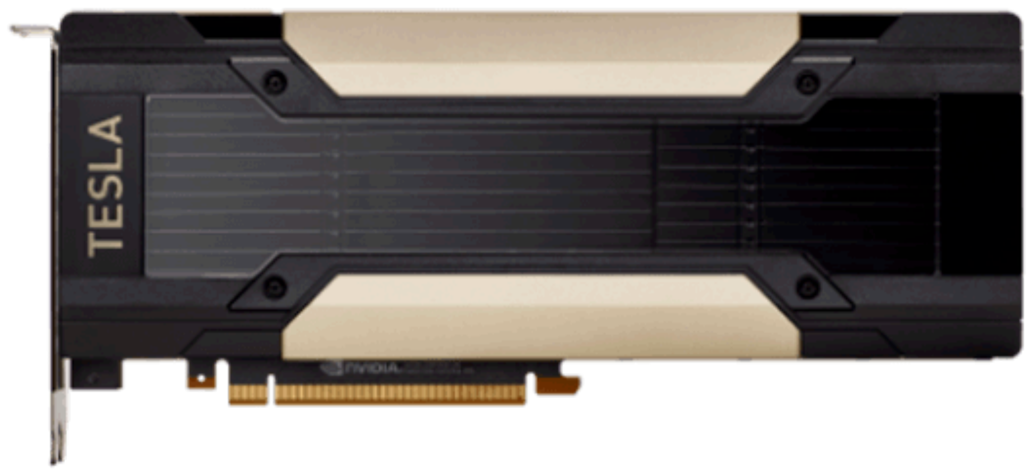
Putting AI on a Diet:

TinyML and Efficient Deep Learning

Song Han
Assistant Professor
Massachusetts Institute of Technology

https://songhan.mit.edu

Deep Learning Going "Tiny"

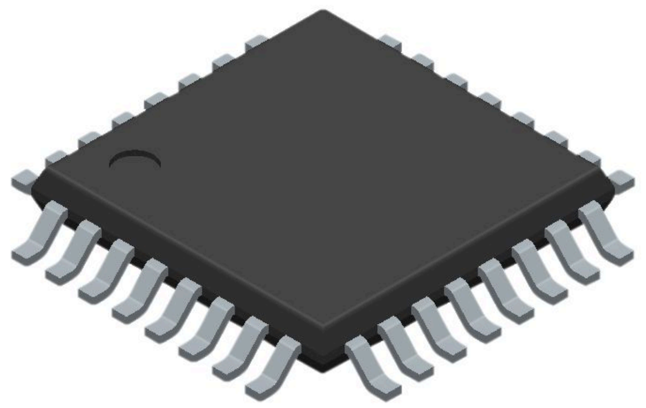


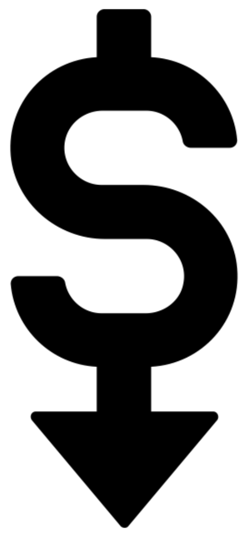
Cloud AI (ResNet)	**Mobile AI (MobileNet)**	**<u>Tiny AI (MCUNet)</u>**
Data centers Expensive Connection required Privacy issue	Smartphones Accessible Process locally	IoT Devices/ Microcontrollers Cheap, small, low-power Rapid growth

- The future belongs to Tiny AI.
- There are billions of IoT devices around the world based on microcontrollers
- Much cheaper, much smaller, almost everywhere in our lives.
- If we can enable powerful AI algorithms on those IoT devices, we can greatly democratize AI and extend the applications of deep learning.

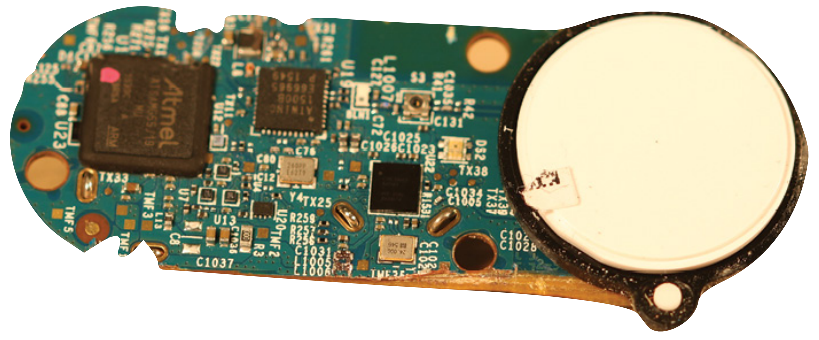
Background: The Era of AIoT on Microcontrollers (MCUs)

Microcontrollers

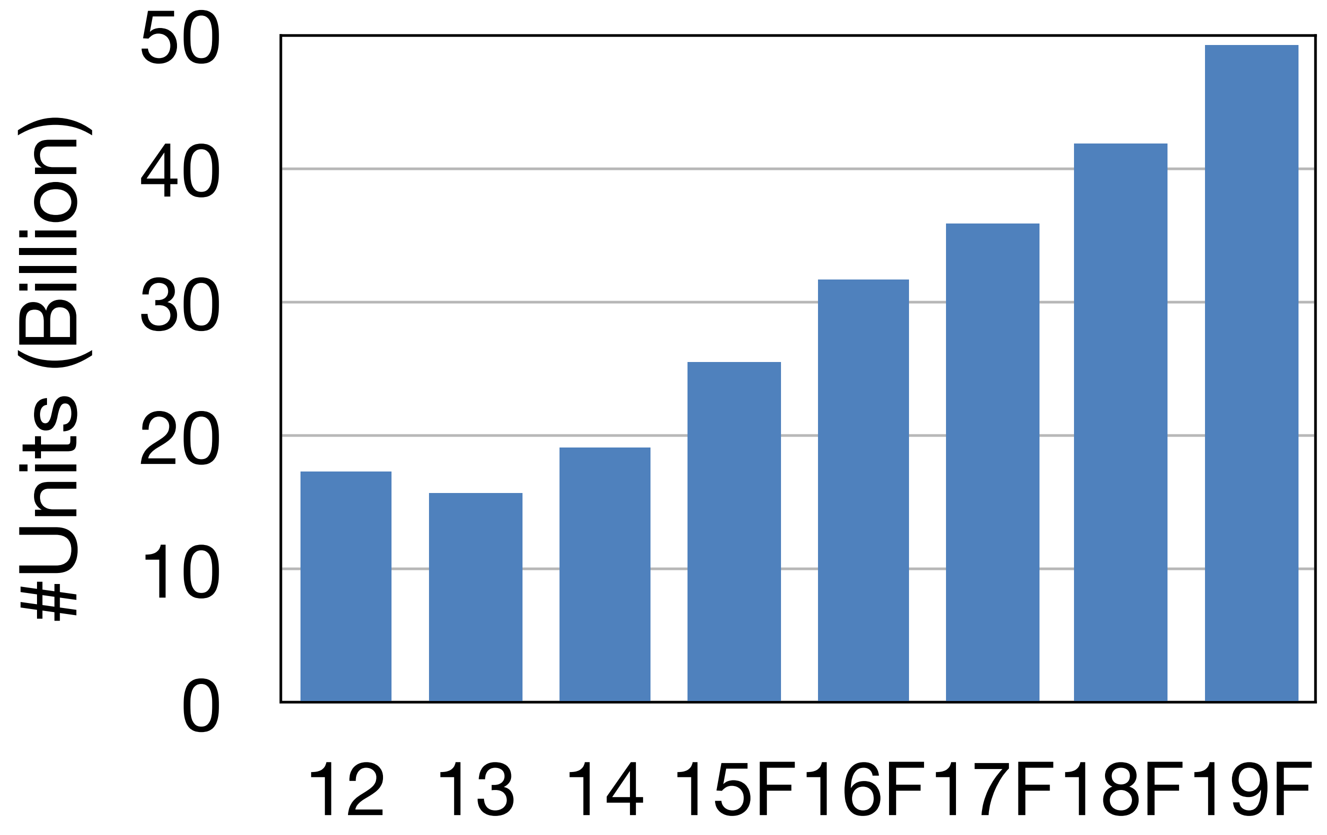




Low-cost
($0.1 - $10)



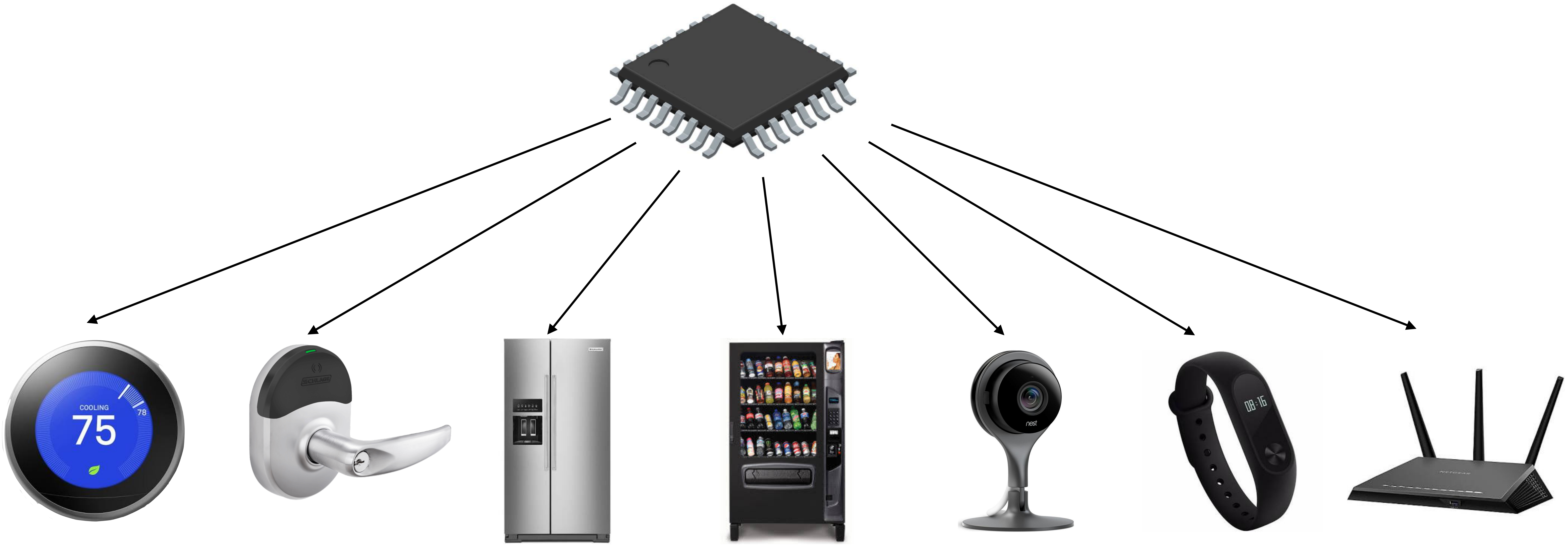
Low-power



Fast growth

Background: The Era of AIoT on Microcontrollers (MCUs)

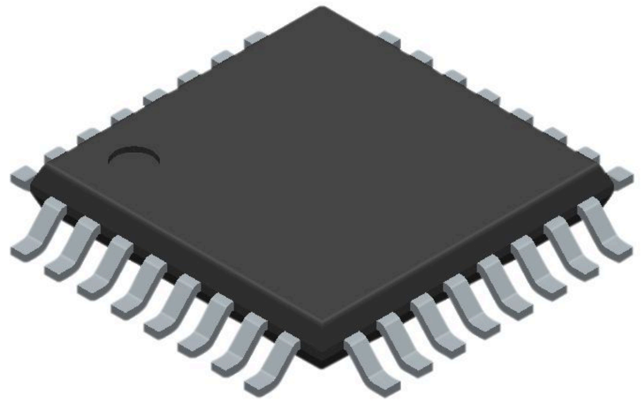
Microcontrollers



Widely deployed

Background: The Era of AIoT on Microcontrollers (MCUs)

Microcontrollers



+

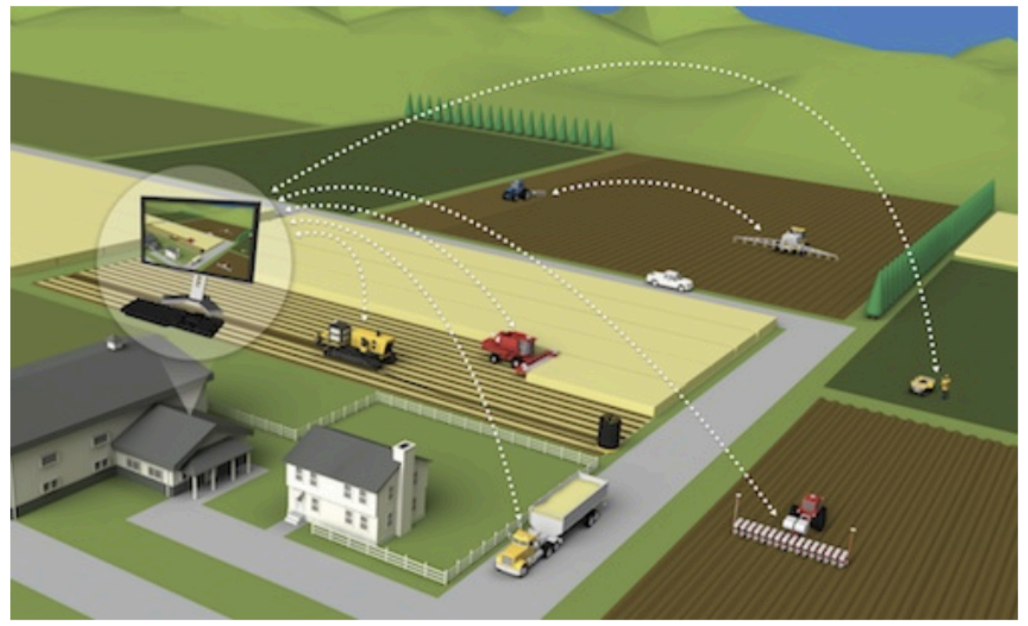
Smart Retail

Personalized Healthcare

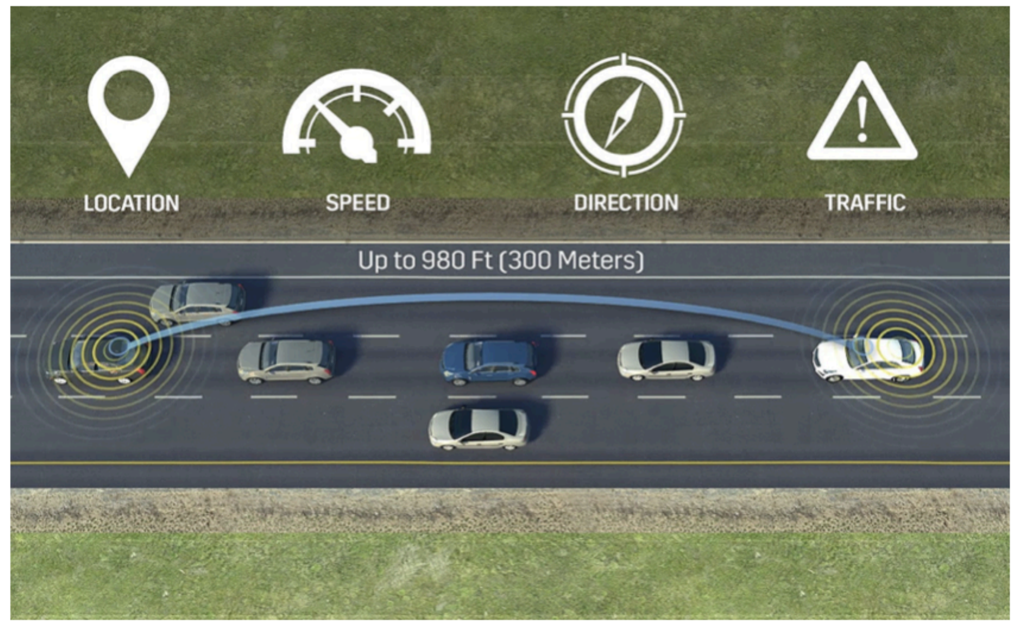
Smart Home

Smart Manufacturing

Precision Agriculture

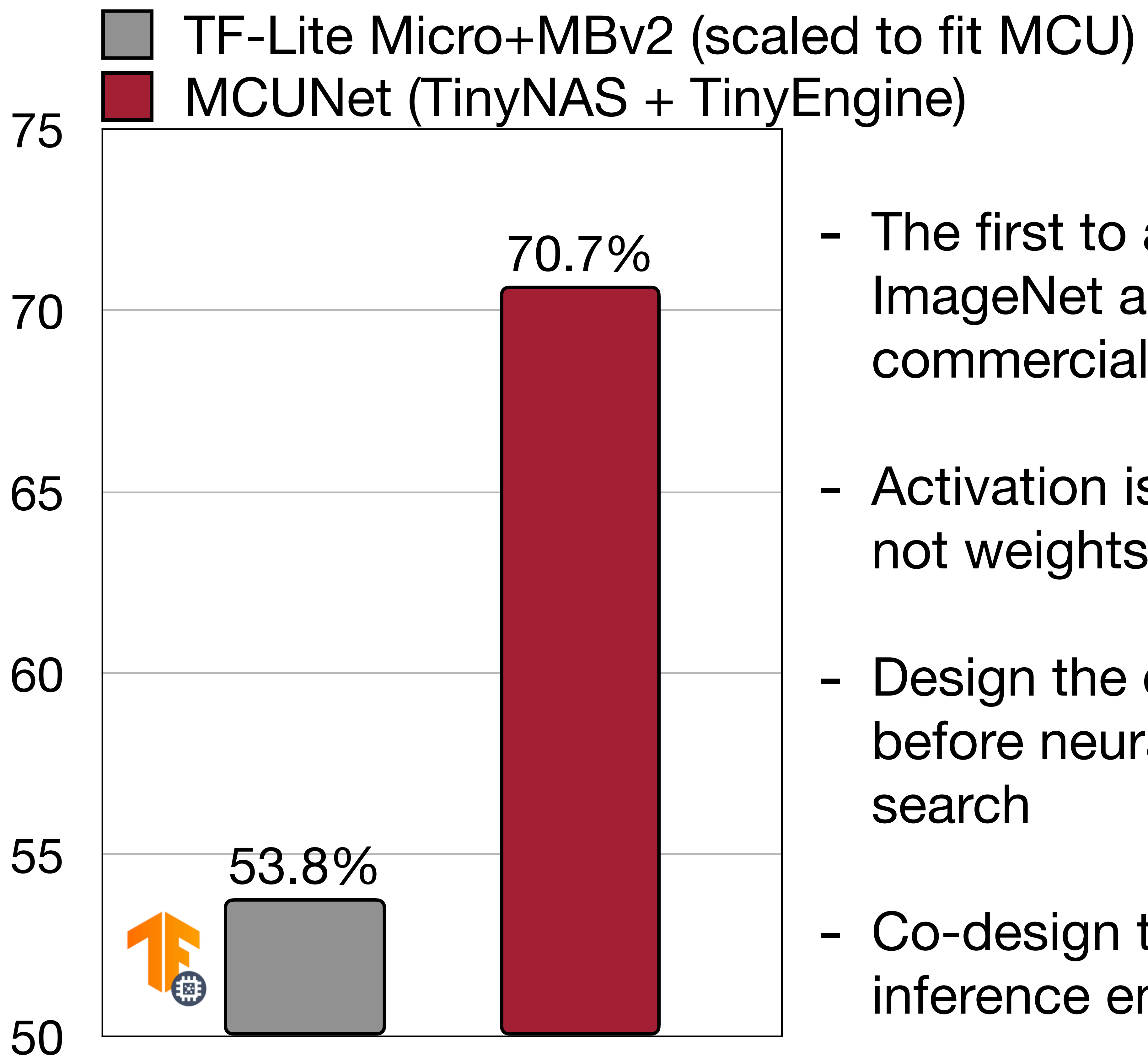


Autonomous Driving



TinyML: Bring AI to IoT Devices

MIT researchers have developed a system, called MCUNet, that brings machine learning to microcontrollers. The advance could enhance the function and security of devices connected to the Internet of Things (IoT). ——MIT News

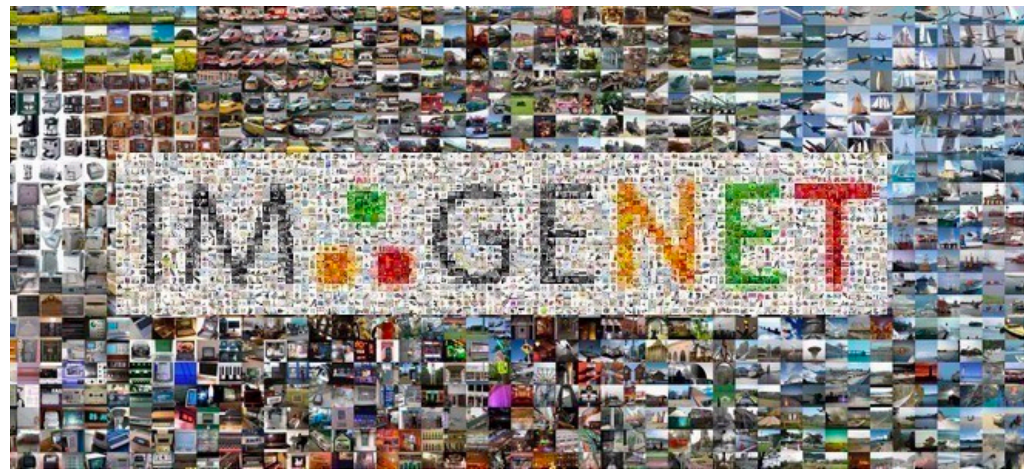


ImageNet Top-1 Accuracy

- The first to achieve >70% ImageNet accuracy on commercial MCU.
- Activation is the bottleneck, not weights
- Design the design space, before neural architecture search
- Co-design the NN with the inference engine

toy IoT applications

MCUNet



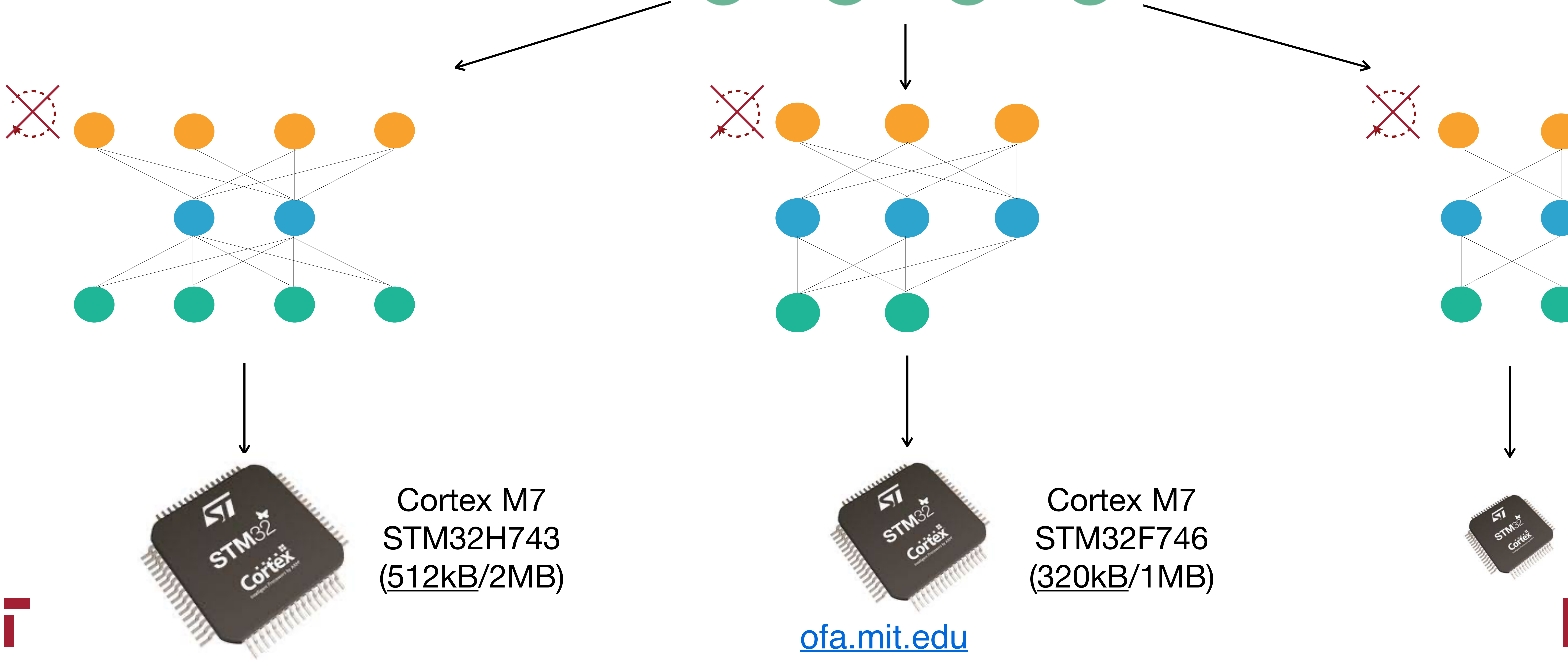
large scale, powerful AI

Once-for-All Network

Train once, get many

Fit diverse hardware constraints

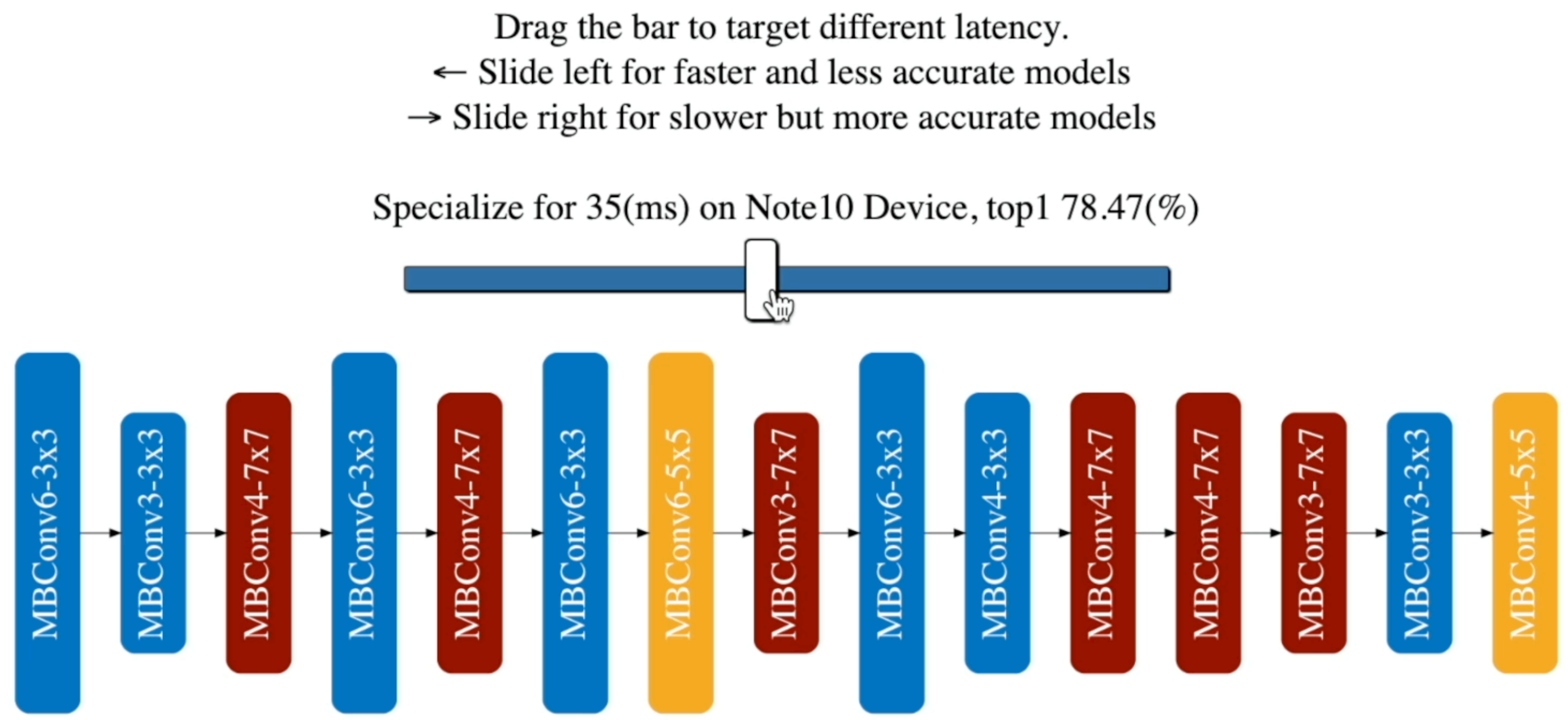
Smaller child networks are nested in larger ones



ofa.mit.edu

Once-for-All Network

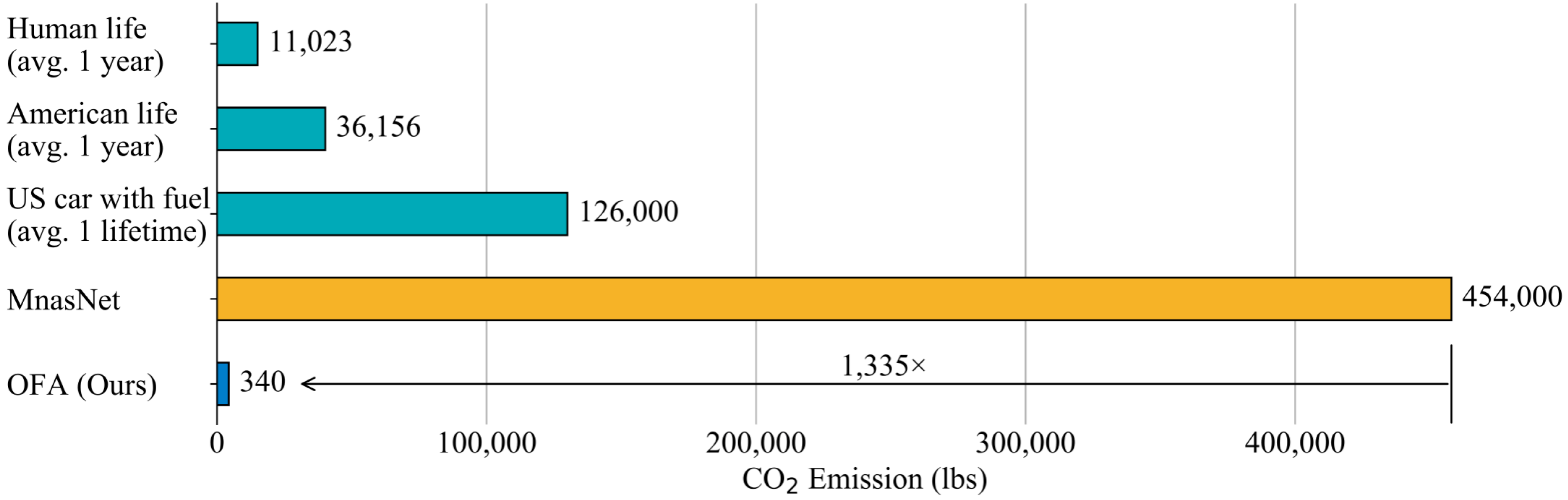
Contains 10^{19} sub-networks, trained at the same time



Once-for-All Network

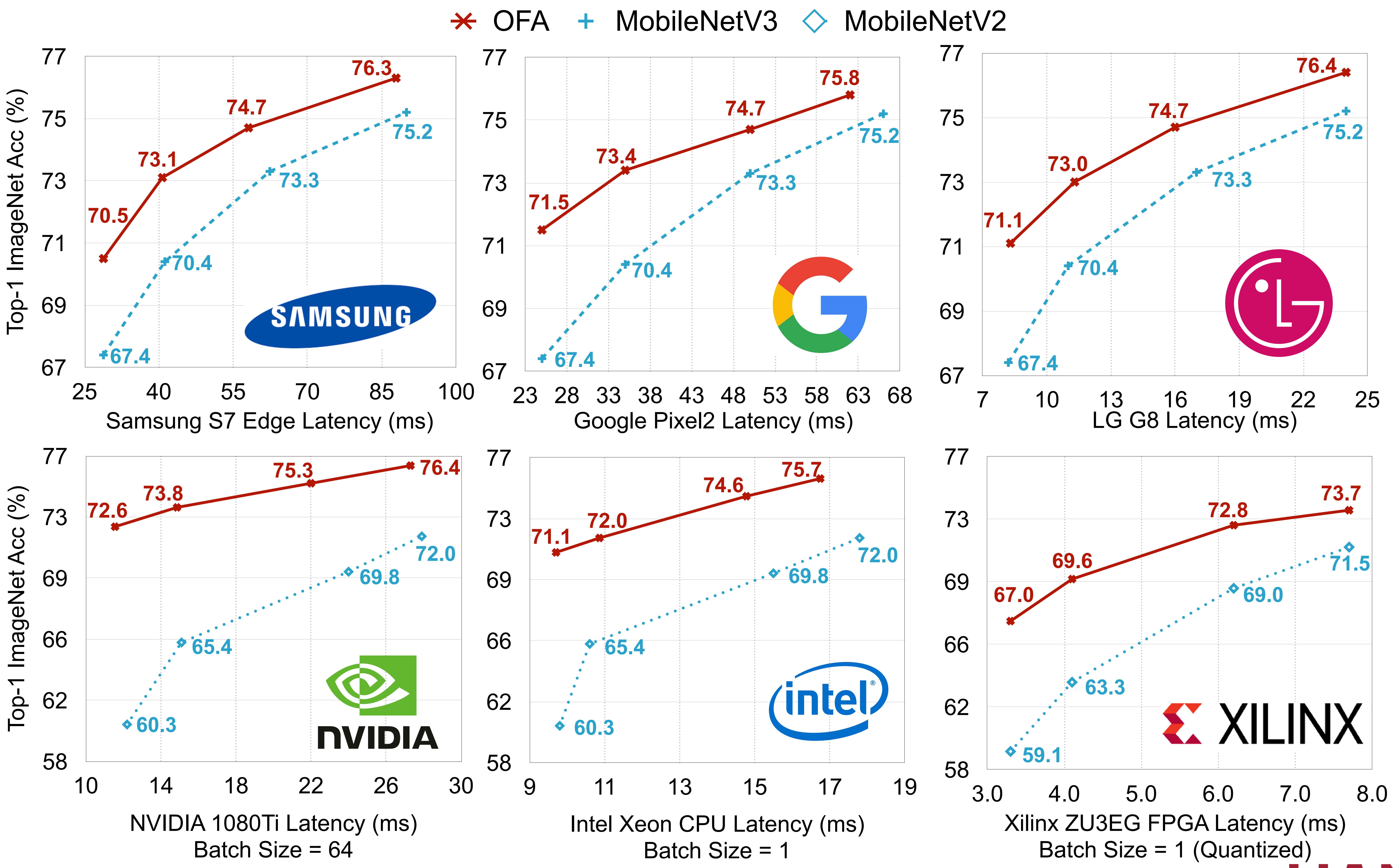
Today's NAS is too expensive w.r.t. carbon emission

Low marginal cost given new hardware platforms: CPU/GPU/DSP/FPGA…



- **Six first-place finishes in top competitions in efficient AI**

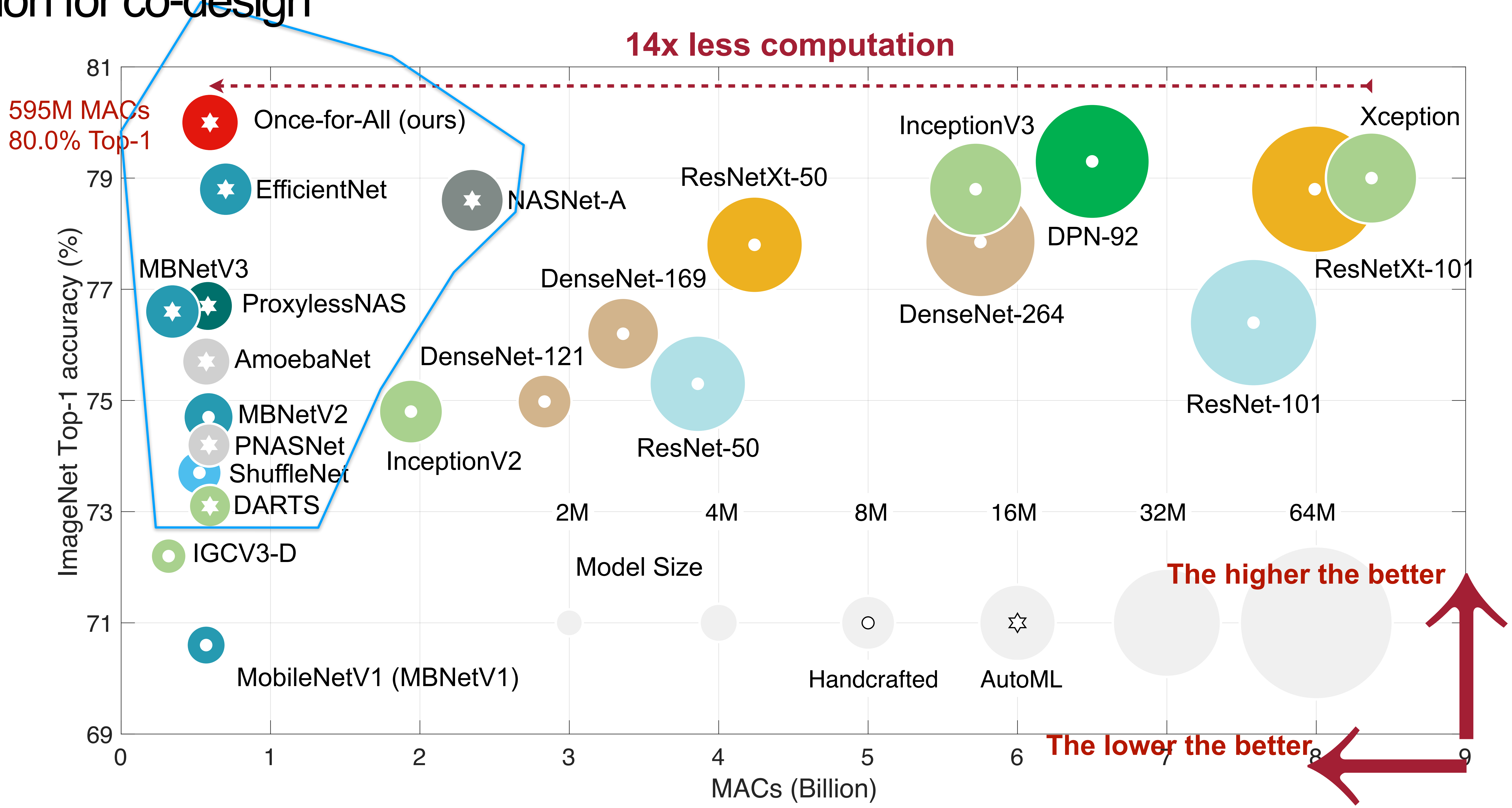
Once-for-All Network



Once-for-All Network

[Once-for-All](), ICLR'20
[ofa.mit.edu]()

Consistently outperforms human baselines, world-record on MLPerf
Turn-key solution for co-design



- OFA sets a new state-of-the-art **80% ImageNet top-1 accuracy** for mobile vision (< 600M MACs).
- OFA sets a world-record in the open division of [MLPerf Inference Benchmark](): 1.078M images per second on eight A100 GPUs

Award Winning Technology

CPU detection
FPGA detection

5th Low-Power Computer Vision Challenge

CPU classification CPU detection

4th Low-Power Computer Vision Challenge

DSP Recognition

3th Low-Power Computer Vision Challenge

Visual Wake Words
on TF-lite

Visual Wake Words Challenge @CVPR 2019

3D Semantic
Segmentation

SemanticKITTI

NLP track
Language Model

MicroNet Challenge @NeurIPS 2019

Industry Adoption

MLPerf
Alibaba.com

Once-for-All (OFA) Network adopbed by Alibaba received a world-record in the open division of MLPerf Inference Benchmark, achiveing 1.078M images per second on eight A100 GPUs

maxim integrated

Once-for-All (OFA) Network adopbed by Maxim Integrated provides 6% accuracy increase in image recognition and 2% accuracy increase in speech command recognition, with >100x energy efficiency compared to Cortex-M4.

PyTorch
aws AutoGluon

Proxyless Neural Architecture Search, an efficient neural architecture search algorithm with light-weight model for mobile AI is integrated by AWS AutoGluon and Facebook PyTorch.

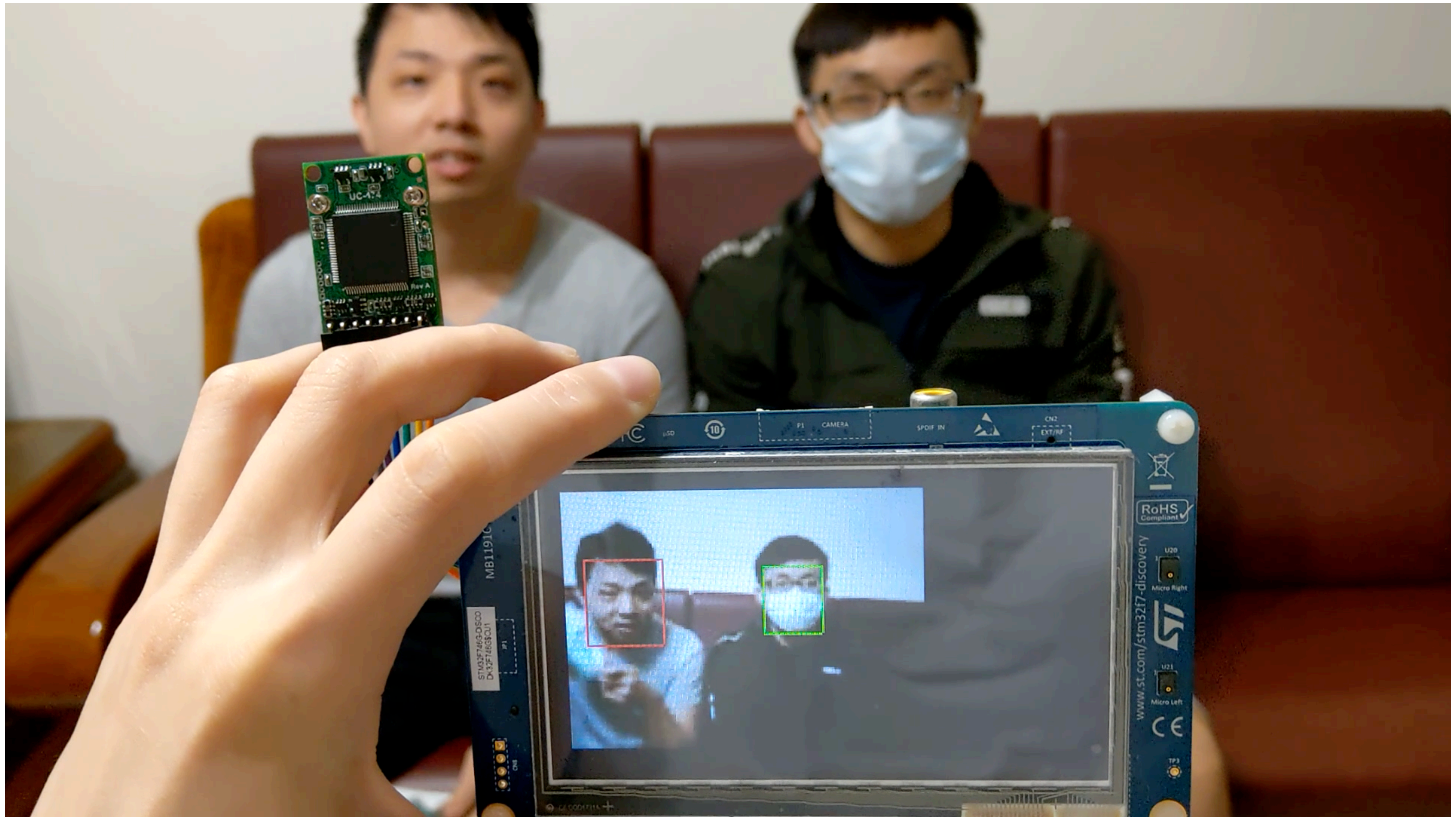
intel

HAQ: Hardware-Aware Automated Quantization with Mixed Precision is integrated by Intel OpenVINO Toolkit. Efficiently search over the bitwidth space for mixed-precision machine learning inference (2, 4, 8 bits)

XILINX

Deep Compression takes the performance of AI inference on Xilinx FPGA to the next level. Reduce model complexity by 5x to 50x with minimal accuracy impact.

TinyML Demo: Face Mask Detection on MCU



- Detecting faces & masks

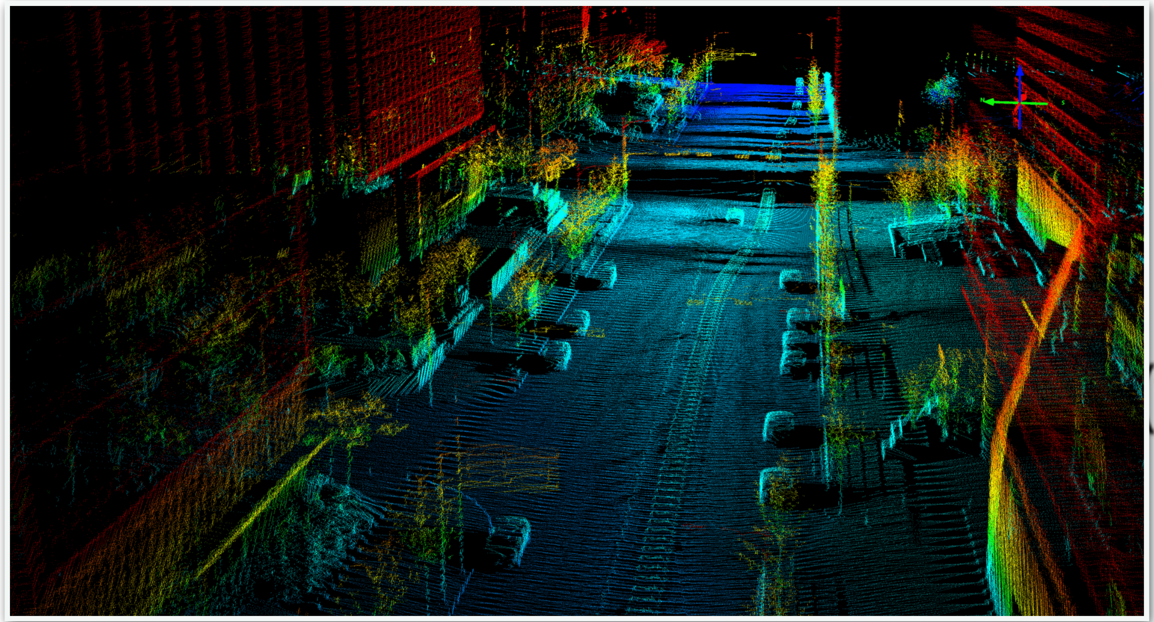
- STM32F746
- 320KB SRAM
- 1MB Flash
- ARM Cortex-M7 @216MHz

TinyML for Driving

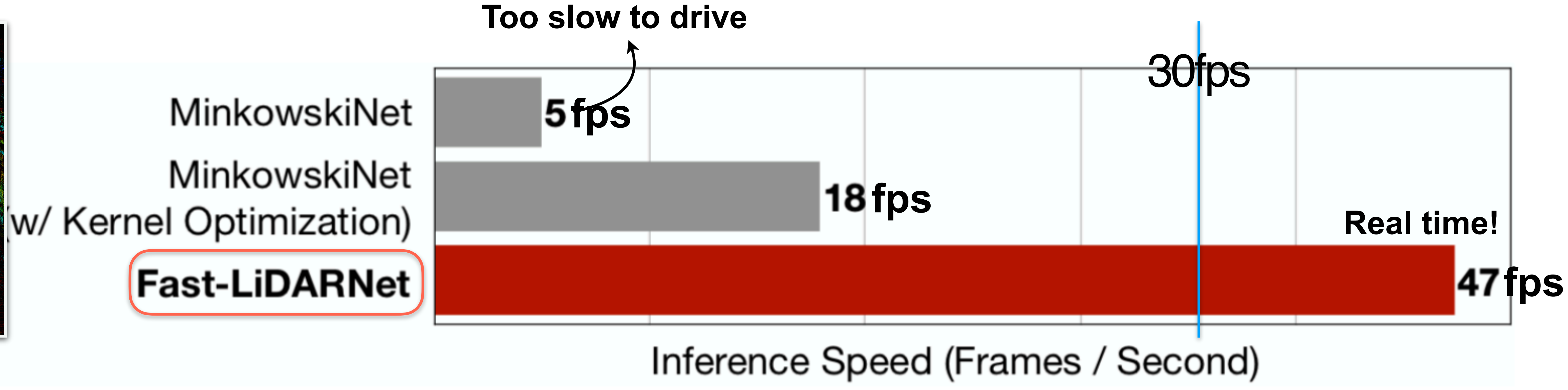
[Liu et al. ICRA'21]

In collaboration with Daniela Rus

3D LiDAR Sensor

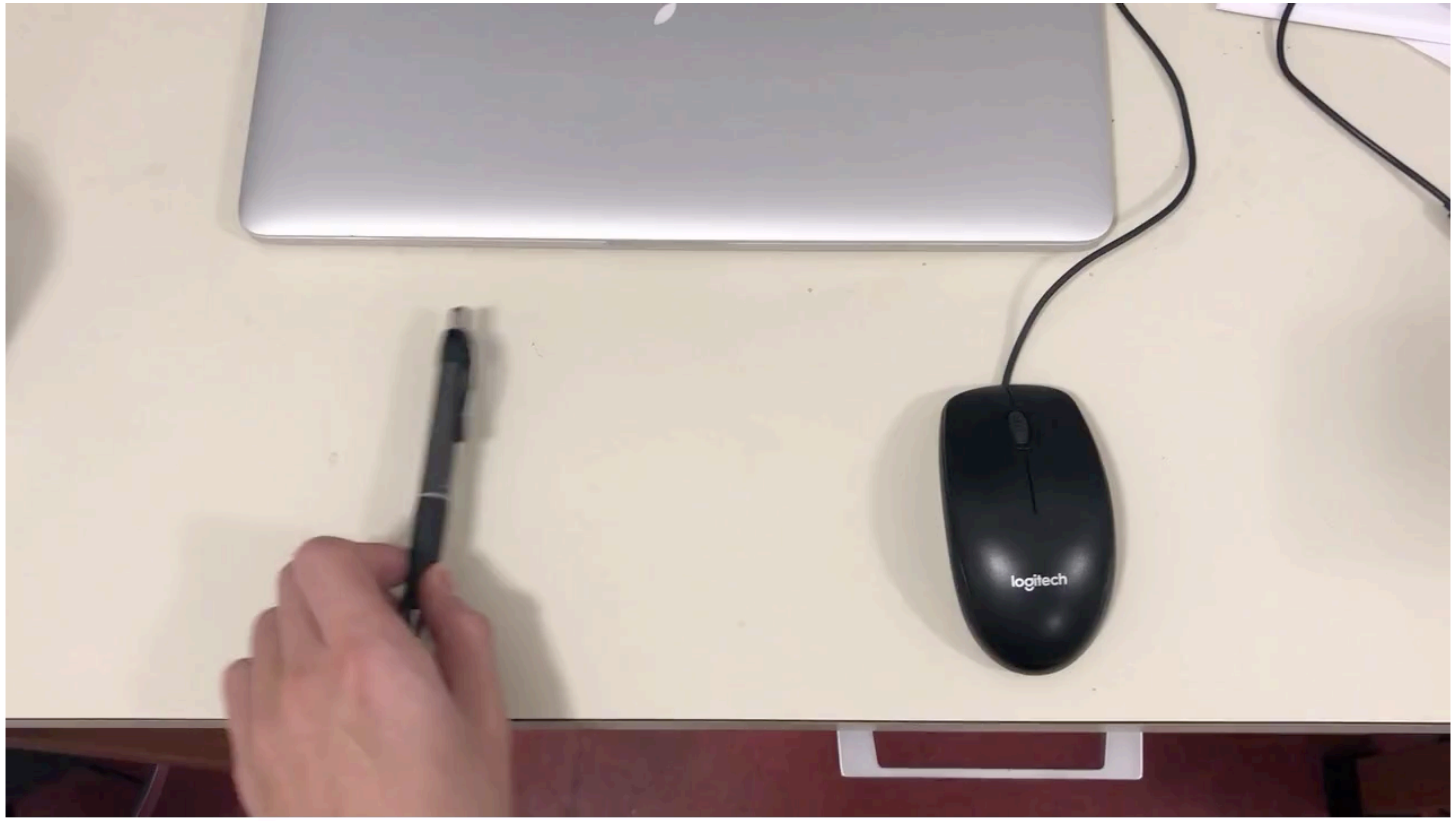


3D Point Cloud: 2M points/s

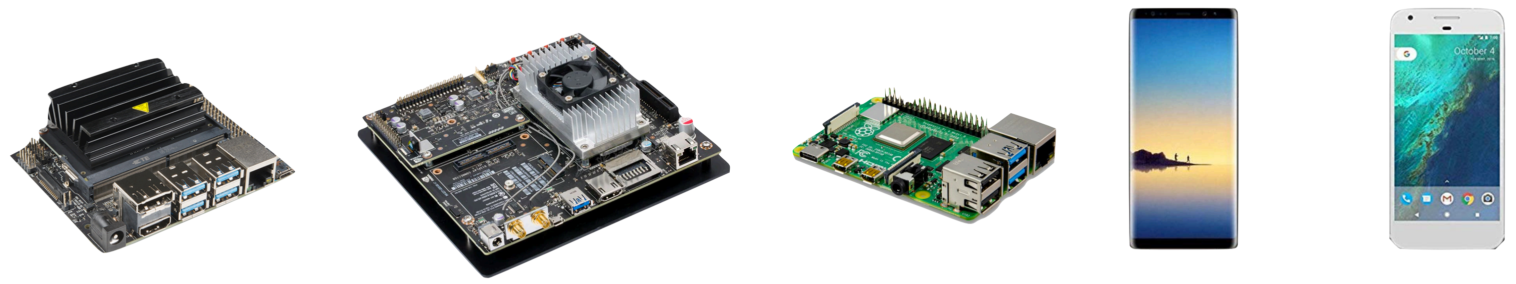


Demo:

TinyML for Video Recognition



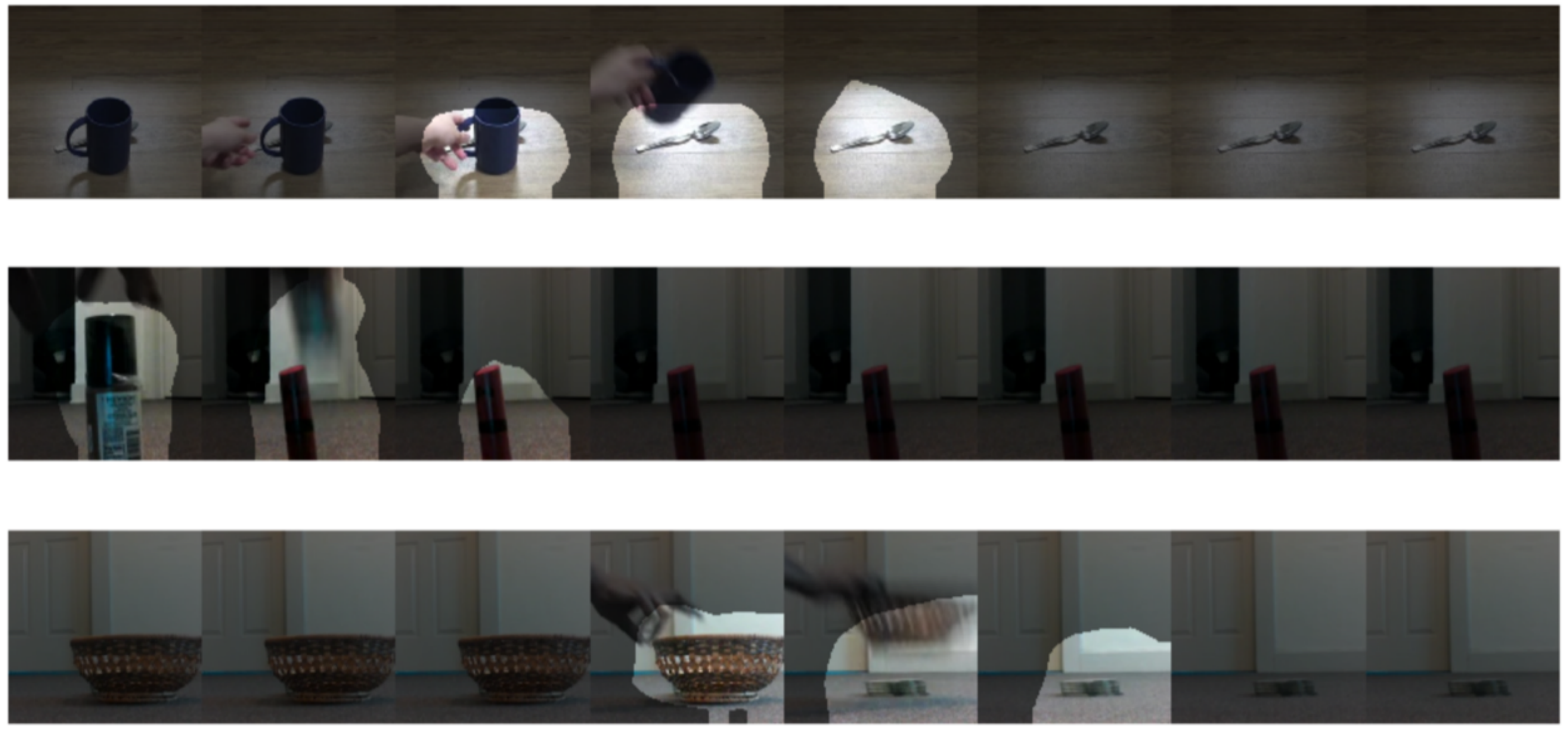
Prediction: Moving something closer to something



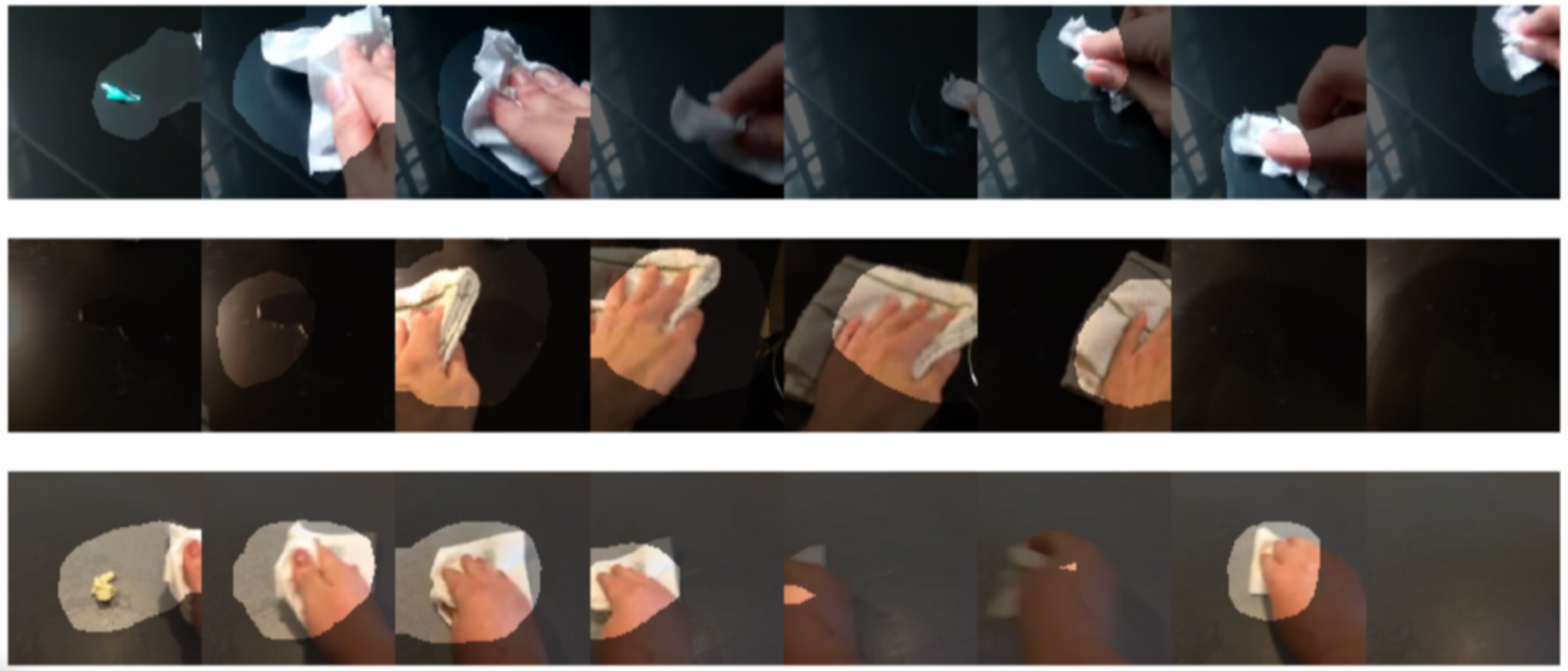
Devices	Jetson Nano		Jetson TX2		Rasp.	Note8	Pixel1
	CPU	GPU	CPU	GPU			
FPS	20.9	74.6	27.5	117.6	14.4	29.0	21.1
Power (watt)	4.8	4.5	5.6	5.8	3.8		

- **LED Bulb Level!**

- Each channel learns different semantics
- *Channel 5: Move something away*



- *Channel 162: Wiping*

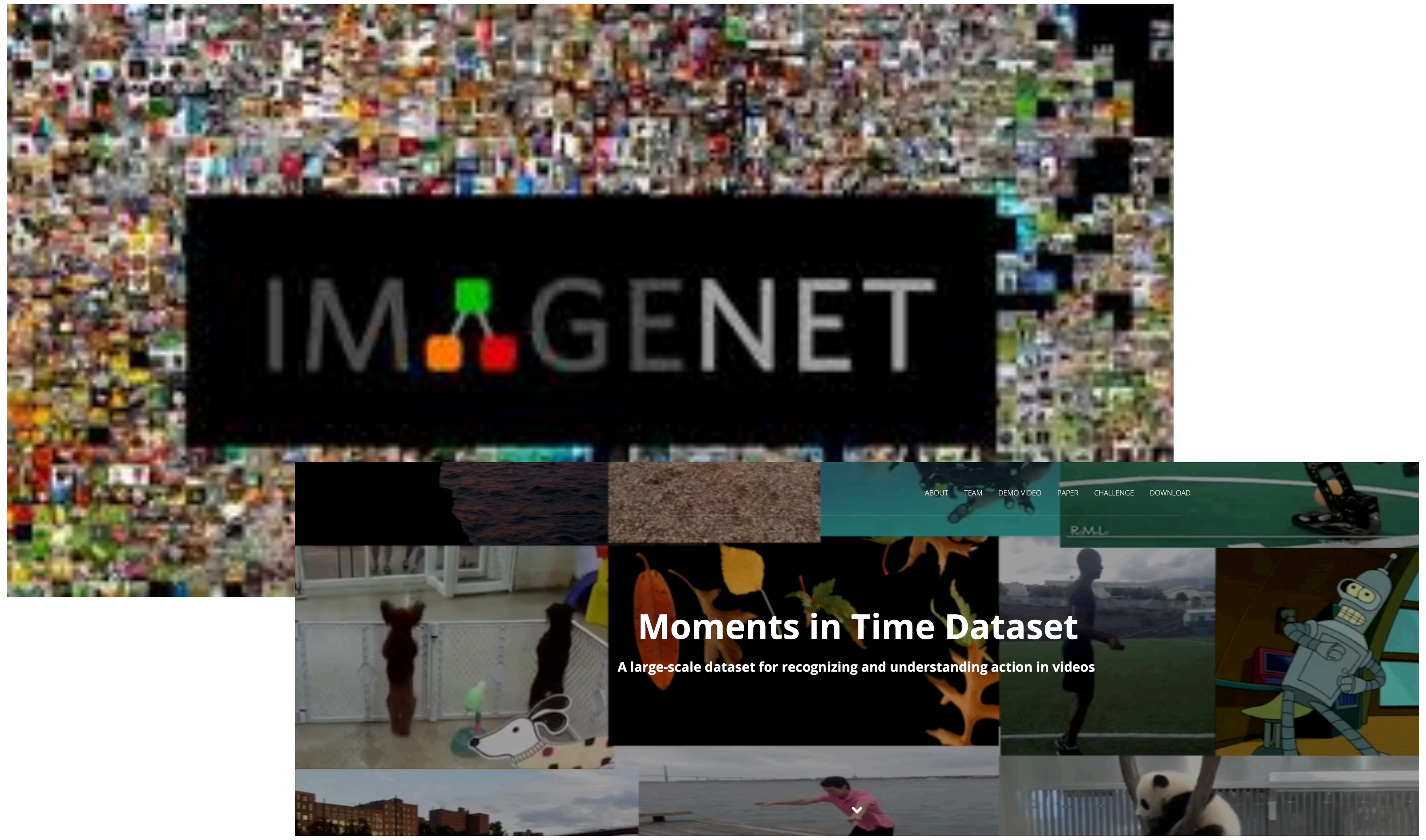


Tiny Transfer Learning

- Customization: AI systems need to continually adapt to new data collected from the sensors.
- Security: Data cannot leave devices because of security and regularization.
- We can reduce the training memory **from 300MB to 16MB**

Data Is Expensive

FFHQ dataset: **70,000** selective post-processed human faces



ImageNet dataset: **millions** of images from diverse categories

"in artificial intelligence, the focus would not be on further refining current algorithms, but rather on developing profoundly new approaches that would enable machines to "learn" using much smaller data sets — a fundamental advance that would eliminate the need to access immense data sets. Success in this work would have a double benefit: seeding economic benefits for the U.S. while reducing the pressure to weaken privacy and civil liberties in pursuit of more "training" data."

— L. Rafael Reif

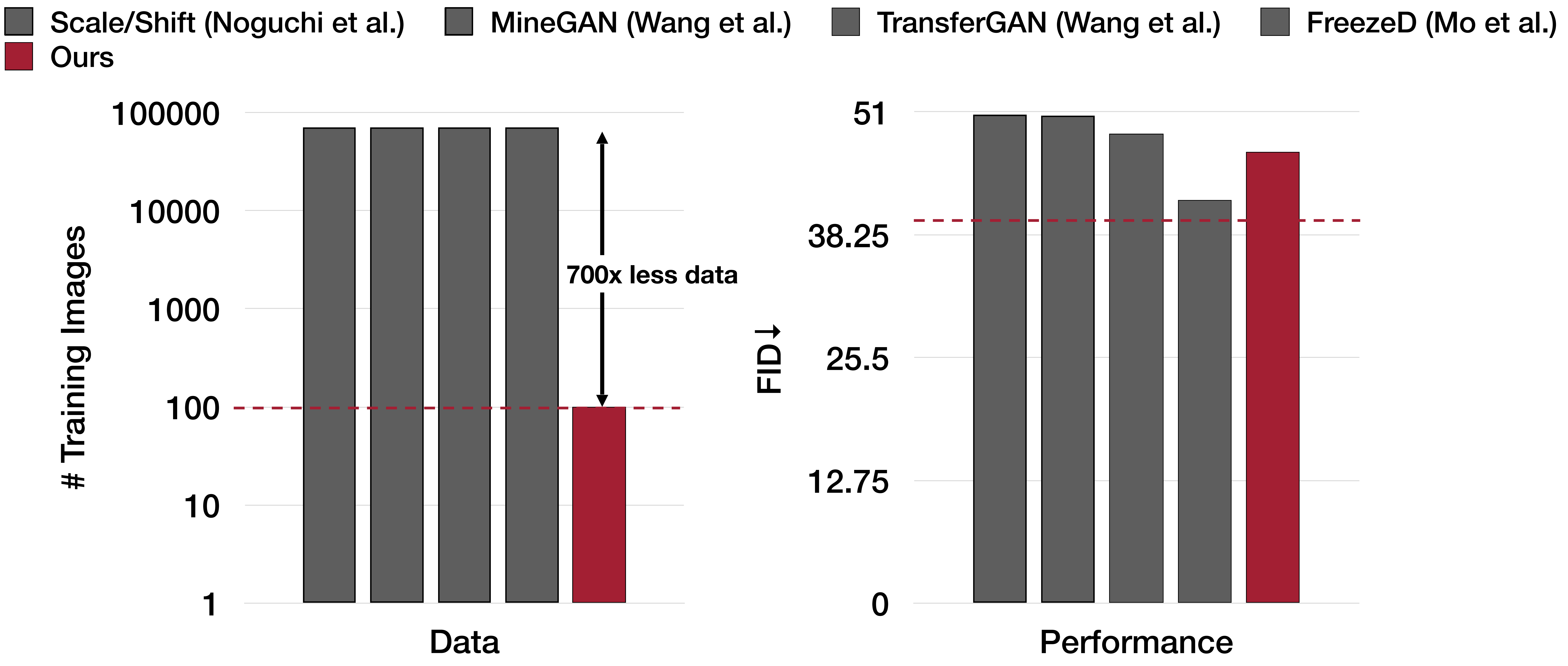
Improve Data-Efficiency

Train GAN with only 100 Images (used to require 70,000 images)

Without our technique:

With our technique:

Improve Data-Efficiency



100-shot Obama

Train GANs with only 100 Images

Smooth interpolation, generalize well

https://github.com/mit-han-lab/data-efficient-gans

ML is Revolutionizing Hardware Design

- Fast:
  - Inference can be accelerated by GPUs and AI accelerators
- Data-Driven:
  - The more data, the higher accuracy; exceed traditional methods
  - Continuous learning

ML is Revolutionizing Hardware Design

ML for Physical Design & Manufacture

"DreamPlace"[1] for placement
"LithoGAN"[2] for lithography modeling
"Google's Chip Design AI"[3] for floorplaning

ML for Circuits Design

"Circuits-GNN"[4] for RF circuits
"Learning to Design Circuits"[5] for Analog IC
"Analog and Digital Circuits Classifier"[6] for sub-circuits classification

ML for System-Level Modeling & Optimization

"PowerNet"[7] for power modeling
"Resource Management with RL"[8] for many-core resources management
"Combine Evolutionary with Deep Learning"[9] for Interface Optimization

[1] Lin, Y., Dhar, S., Li, W., Ren, H., Khailany, B., & Pan, D. Z. DREAMPlace: Deep learning toolkit-enabled GPU acceleration for modern VLSI placement. In DAC 2019
[2] Ye, W., Alawieh, M. B., Lin, Y., & Pan, D. Z. Lithogan: End-to-end lithography modeling with generative adversarial networks. In DAC 2019.
[3] Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J., Songhori, E., Wang, S., ... & Nazi, A. (2020). Chip Placement with Deep Reinforcement Learning. arXiv preprint arXiv:2004.10746.
[4] Zhang, G., He, H., & Katabi, D. (2019, May). Circuit-GNN: Graph Neural Networks for Distributed Circuit Design. In *International Conference on Machine Learning* (pp. 7364-7373).
[5] Wang, H., Yang, J., Lee, H. S., & Han, S. (2018). Learning to design circuits. *NeurIPS 2018, ML for System Workshop*.
[6] Liou, G. H., Wang, S. H., Su, Y. Y., & Lin, M. P. H. (2018, July). Classifying Analog and Digital Circuits with Machine Learning Techniques Toward Mixed-Signal Design Automation. In *SMACD* 2018
[7] Chen, J., Alawieh, M. B., Lin, Y., Zhang, M., Zhang, J., Guo, Y., & Pan, D. Z. (2020). Powernet: SOI Lateral Power Device Breakdown Prediction With Deep Neural Networks. *IEEE Access*
[8] Mao, H., Alizadeh, M., Menache, I., & Kandula, S.. Resource management with deep reinforcement learning. In*15th ACM Workshop on Hot Topics in Networks*.
[9] Servadei, L., Mosca, E., Werner, M., Esen, V., Wille, R., & Ecker, W. Combining Evolutionary Algorithms and Deep Learning for Hardware/Software Interface Optimization.

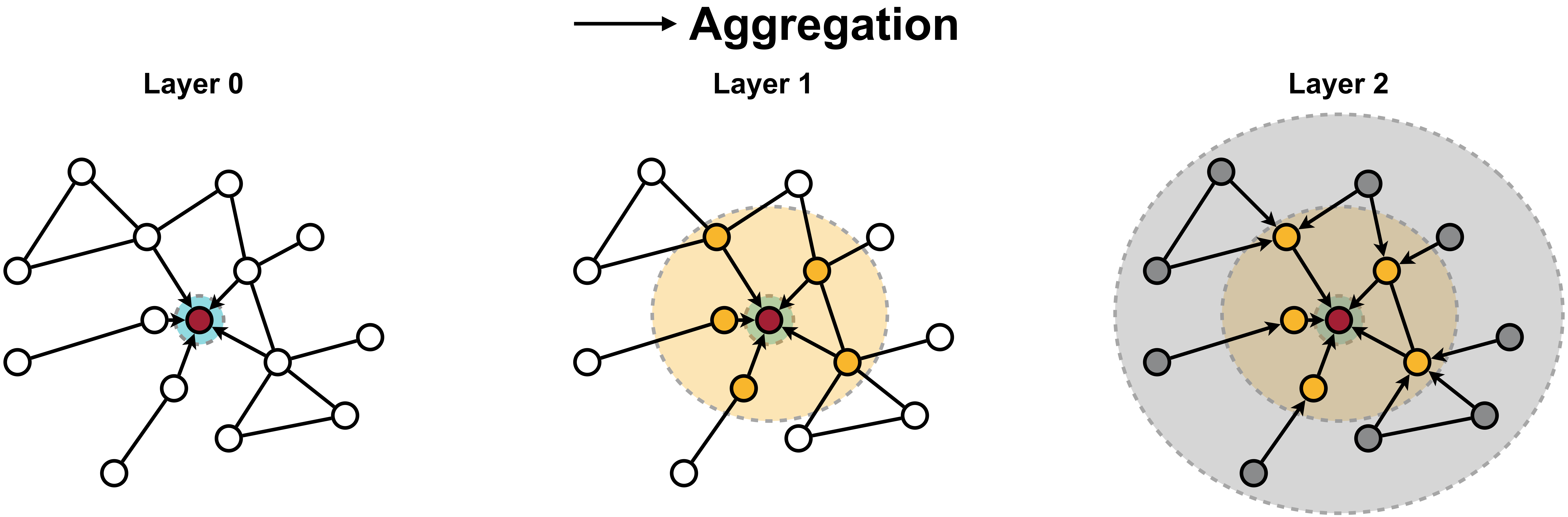
[Wang et al. DAC'20]

GCN-RL Circuit Designer

After many iterations, we get a trained RL agent

GCN RL Agent: Circuit is a Graph

[Wang et al. DAC'20]

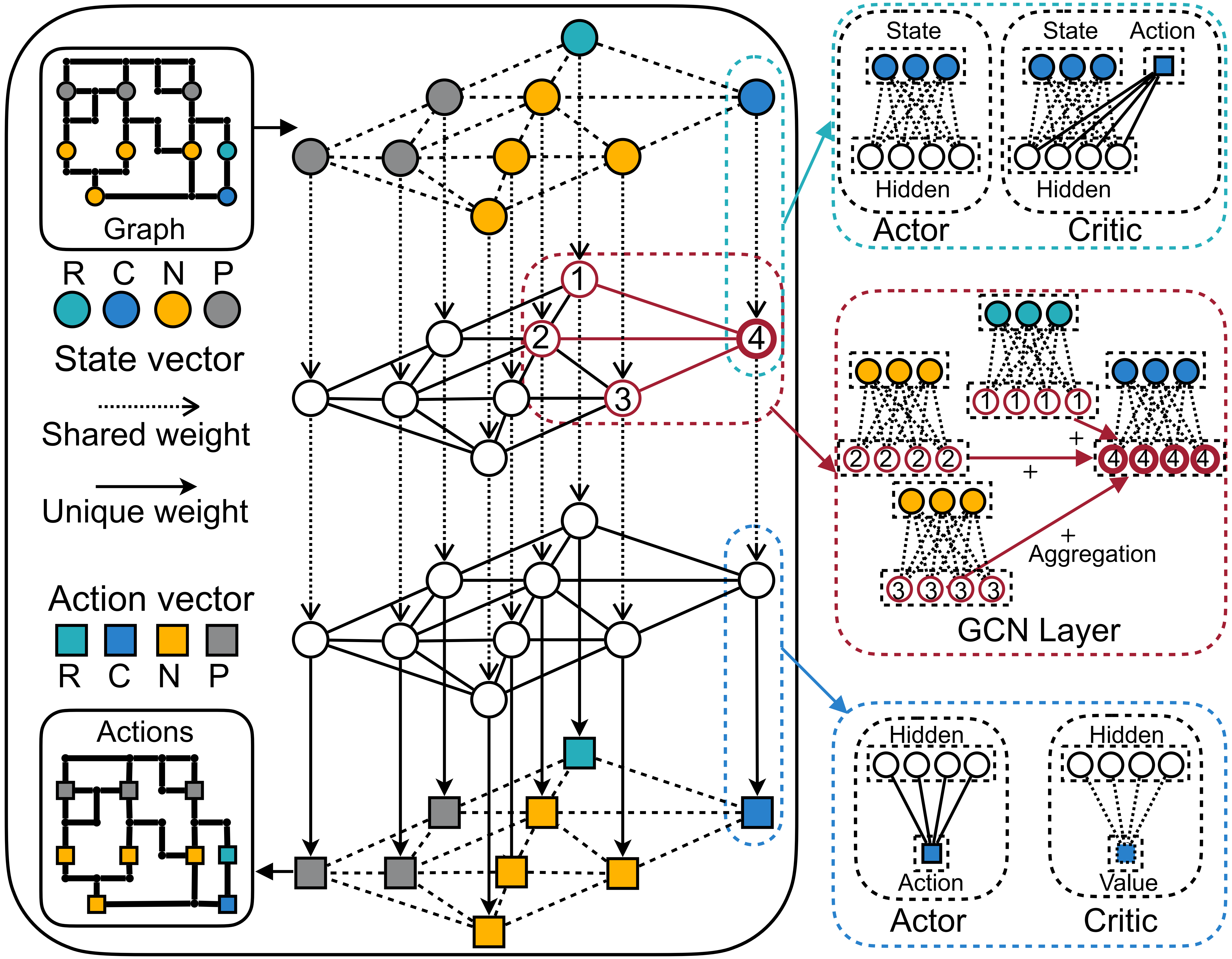


Receptive Field:
Neighbors

Receptive Field:
Neighbors +
Neighbors of neighbors

GCN RL Agent

[Wang et al. DAC'20]



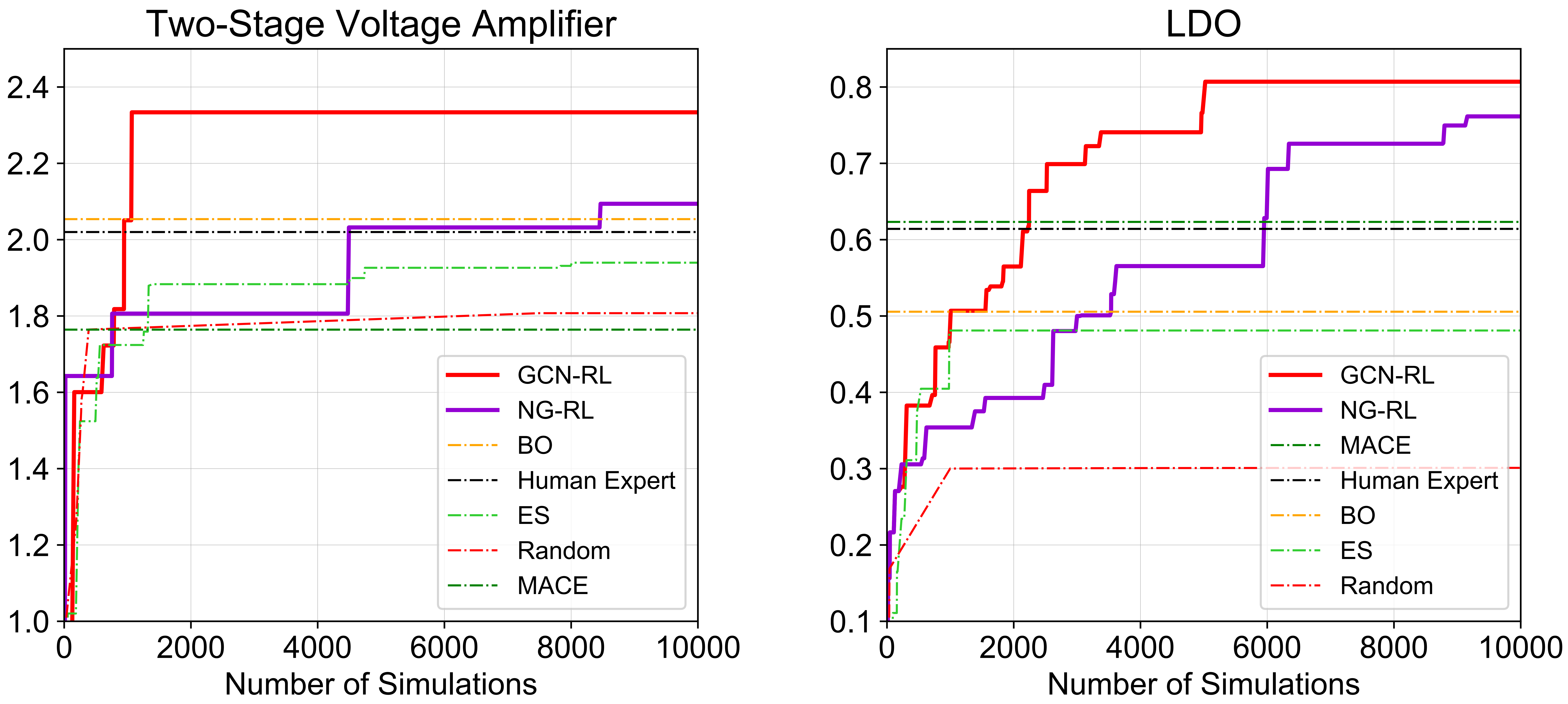
Apply an Actor-Critic RL agent with GCN

Actor: Generates the sizings

Critic: Emulates the real simulator environment.
Estimates the FoM of the sizings
Provides gradients for weights update

GCN-RL Achieves Highest FoM

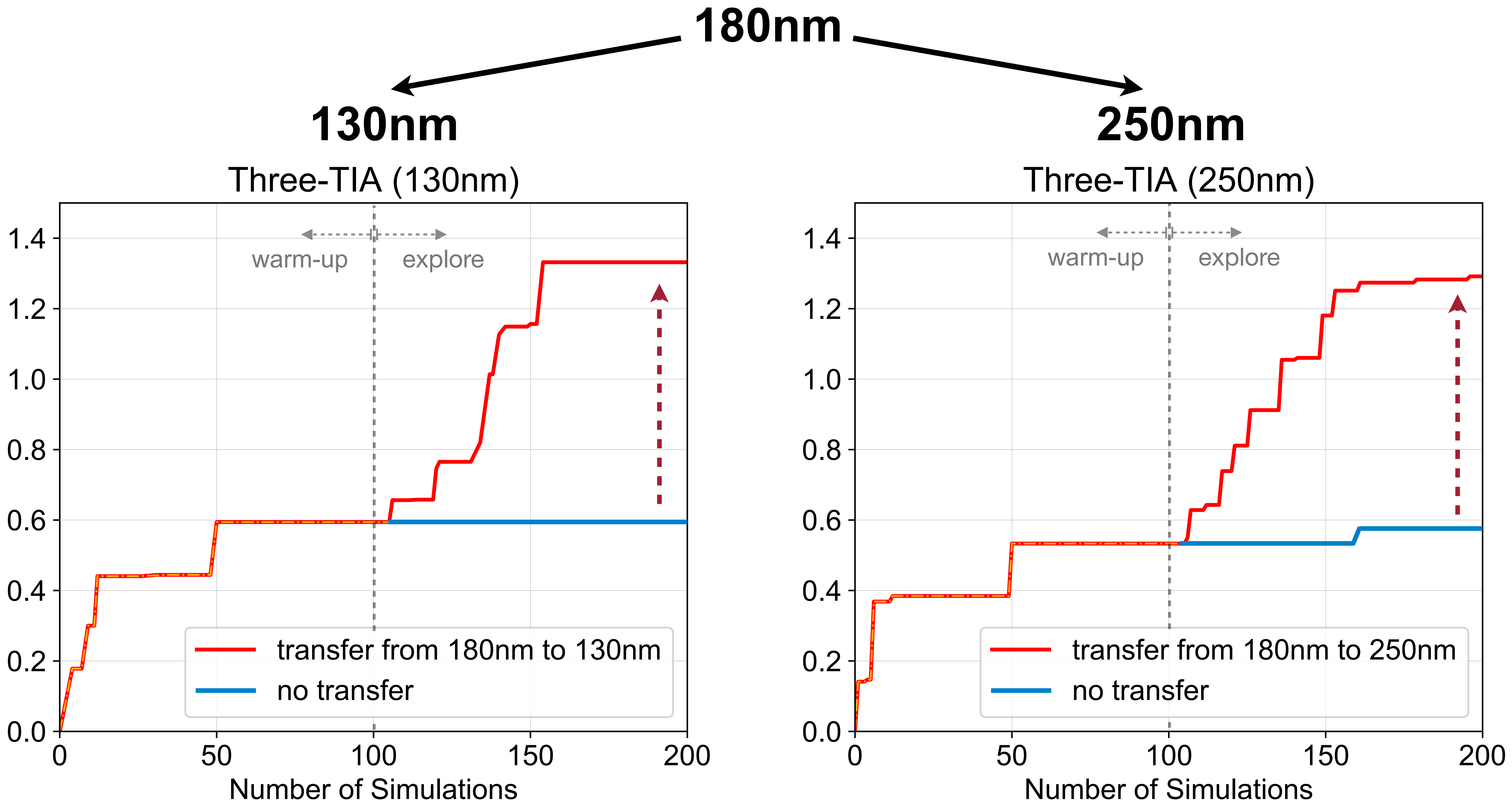
[Wang et al. DAC'20]



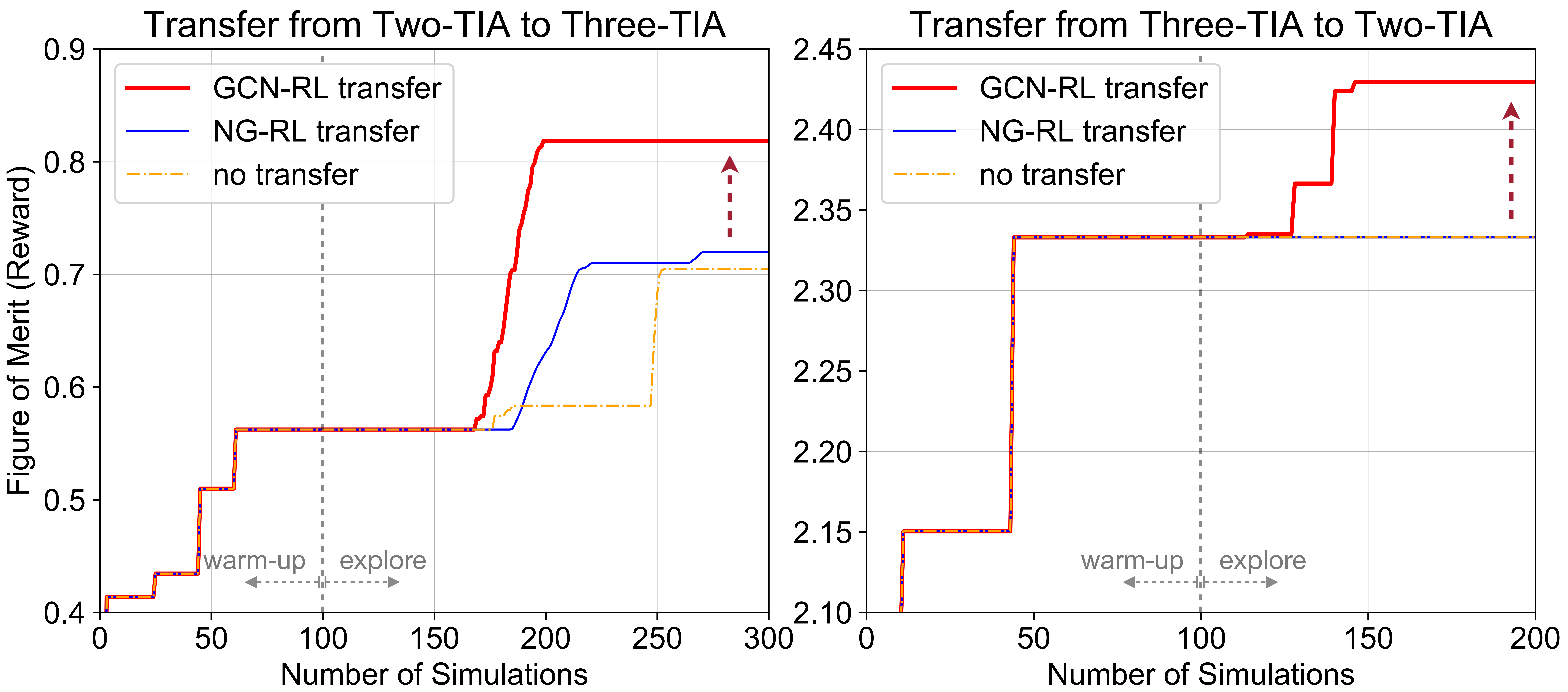
GCN-RL has highest FoM and fast converging speed

Graph info improves FoM (GCN-RL v.s. NG-RL)

GCN-RL with Transfer Learning



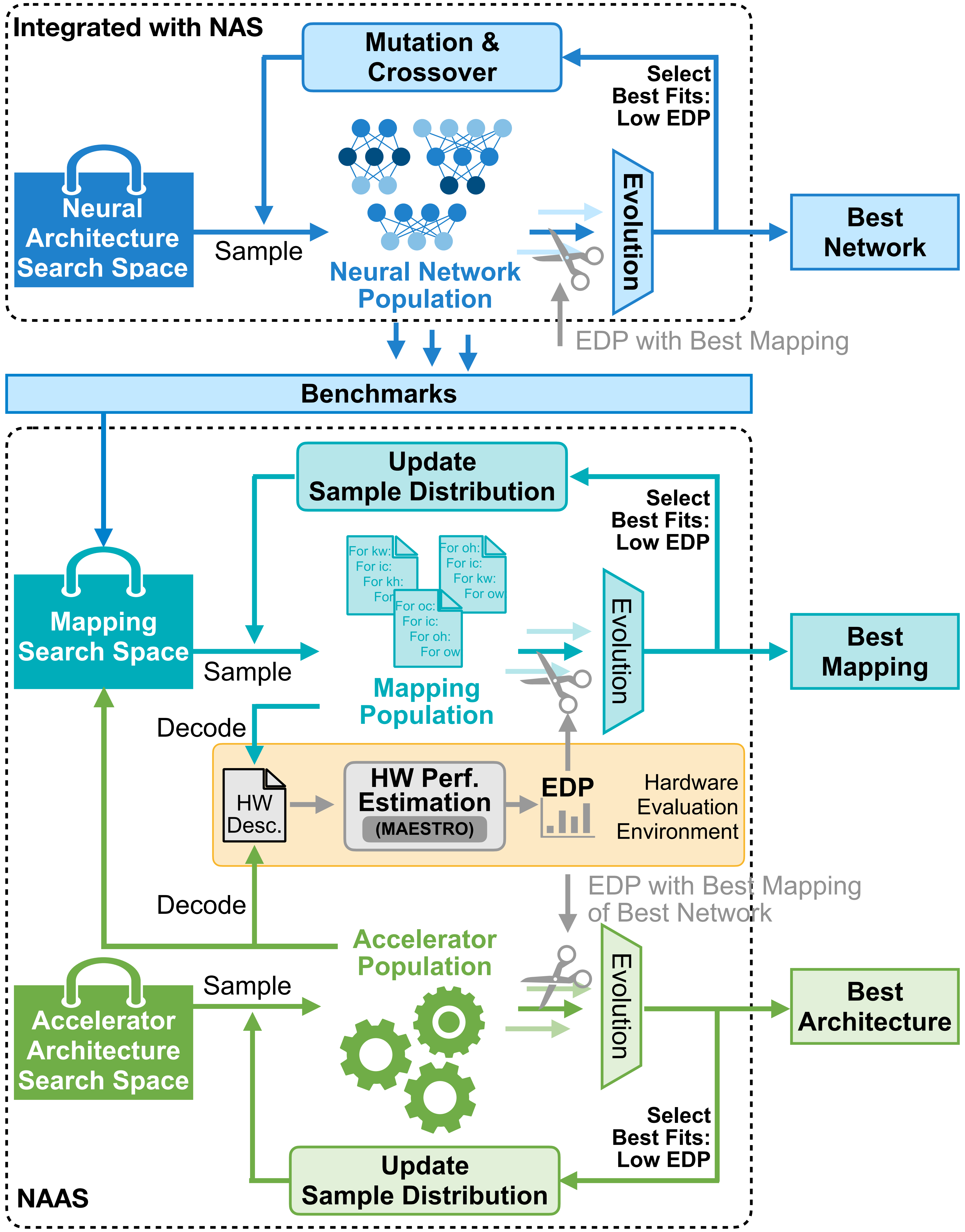
GCN-RL with Transfer Learning

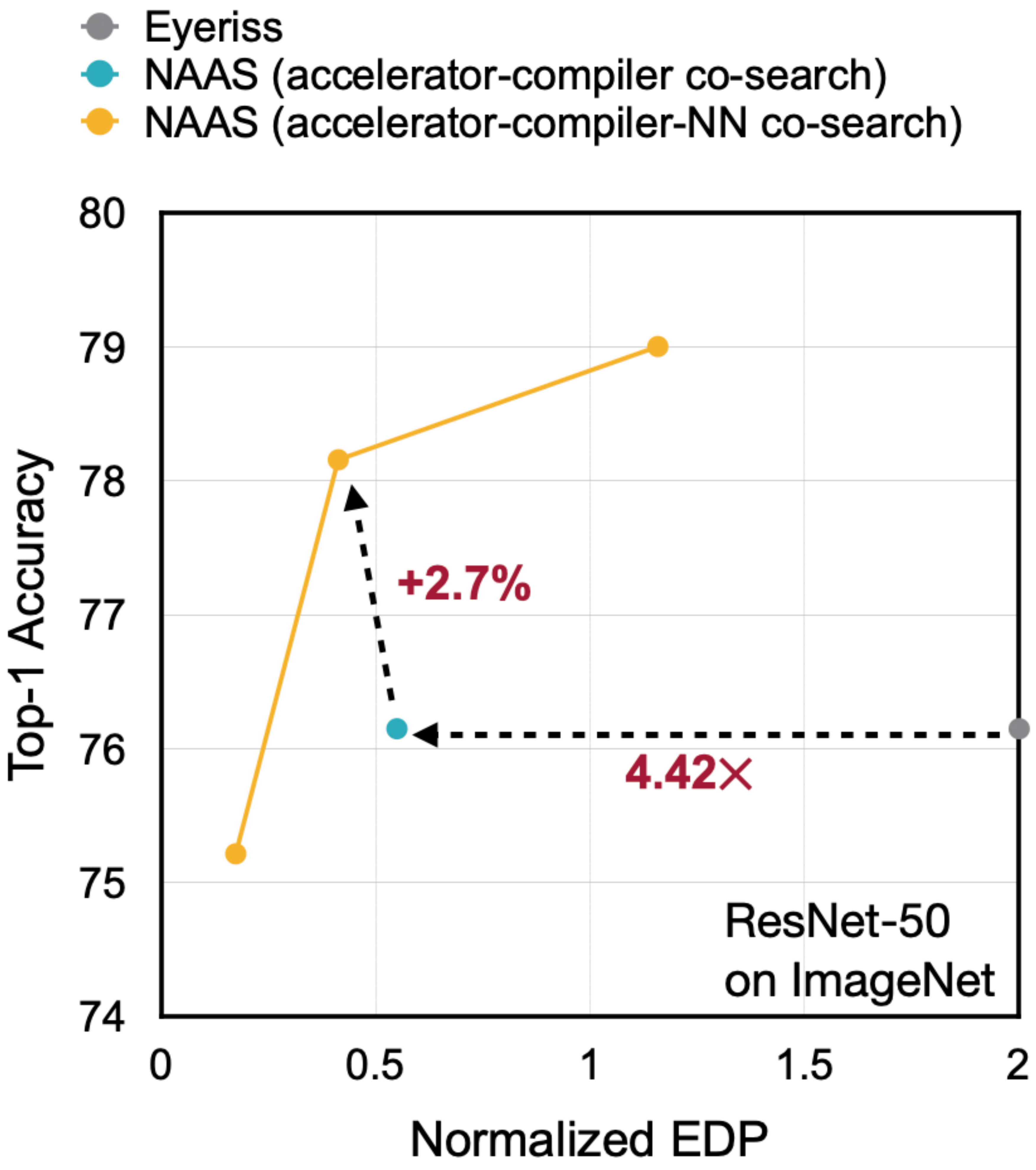


ML for Digital Architecture Design

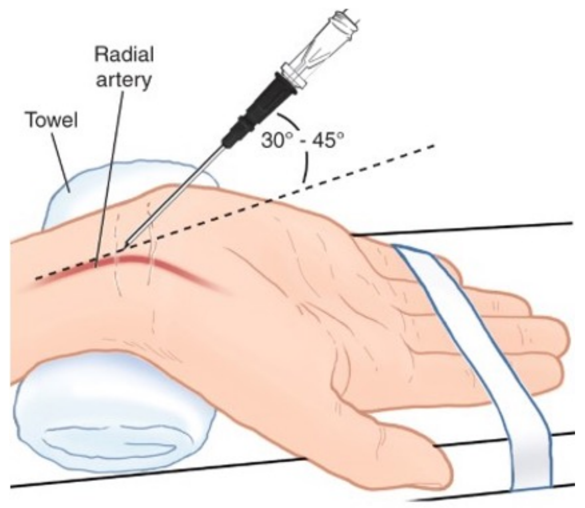
[Lin et al. DAC'21]

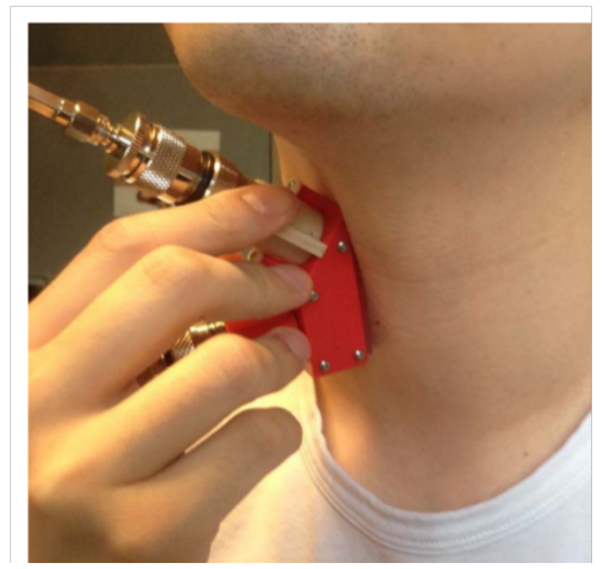
NAAS: Neural Accelerator Architecture Search

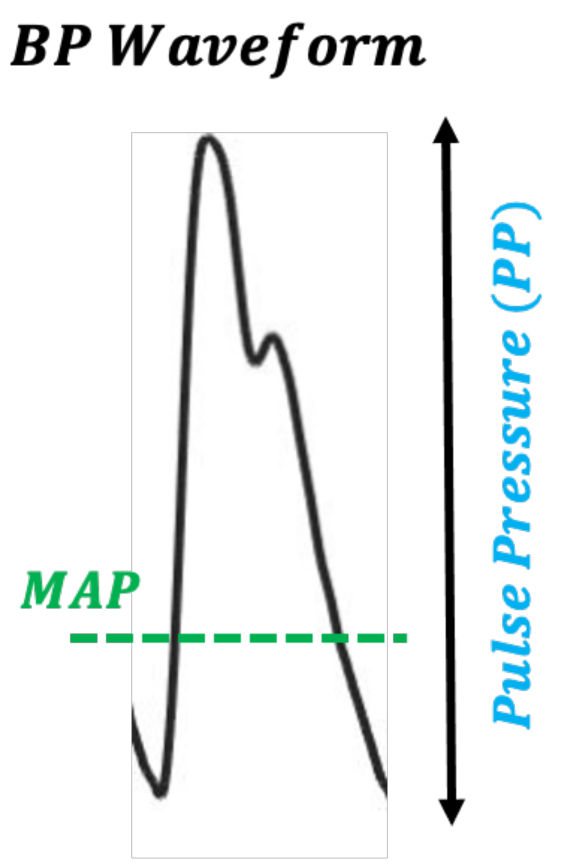




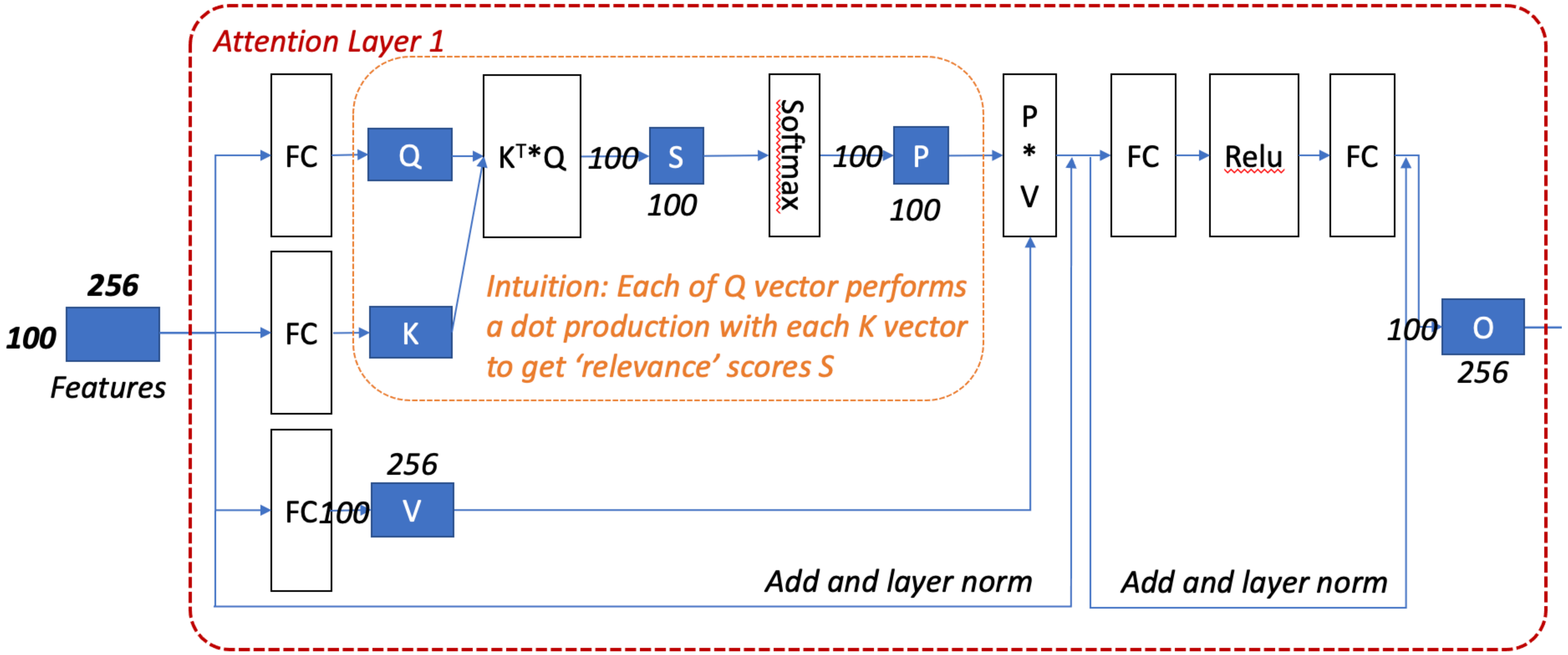
ML for Blood Pressure Measurement





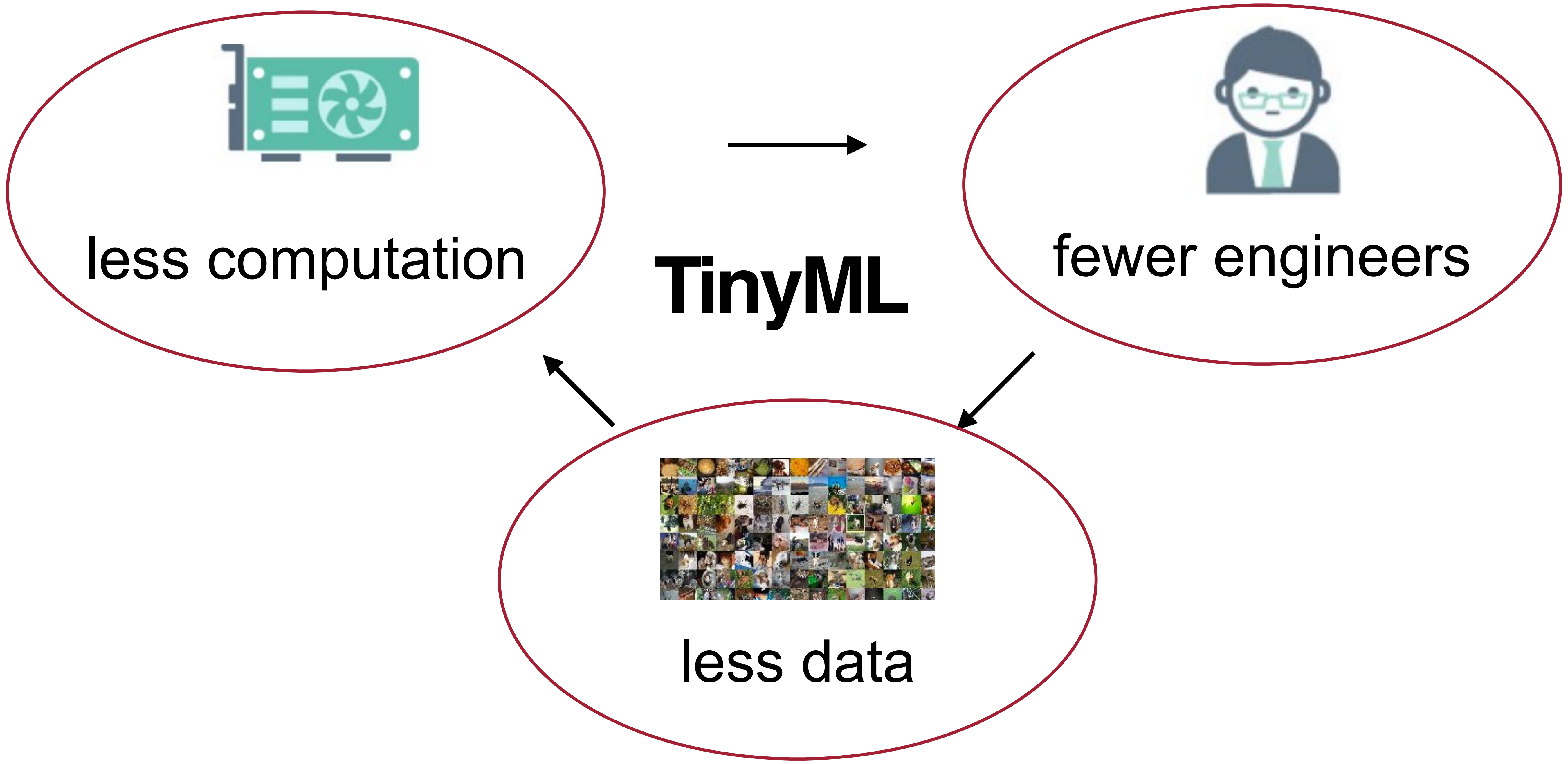


$$BP(t) = \frac{(A(t) - A_\mu)}{\Delta A} * PP + MAP$$



	loss	Bias mean	Bias Std	RMSE
LSTM	**0.096**	1.352	1.575	**2.076** mmHg
Pure attention	**0.169**	2.049	1.840	**2.754** mmHg
Conv	0.174	2.059	1.891	2.795 mmHg
Conv+Attention	**0.172**	2.036	1.894	**2.781** mmHg
FC2D	**0.251**	2.483	2.253	**3.353** mmHg

Better software/hardware for AI



AI for better hardware design