MicroNets

Neural Network Architectures for Deploying TinyML Application on Commodity Microcontrollers

Colby Banbury*[1,2], Chuteng Zhou*[1], Igor Fedorov*[1], Ramon Matas Navarro[1], Urmish Thakker[3], Dibakar Gope[1], Vijay Janapa Reddi[1], Matthew Mattina[1], Paul N. Whatmough[1]

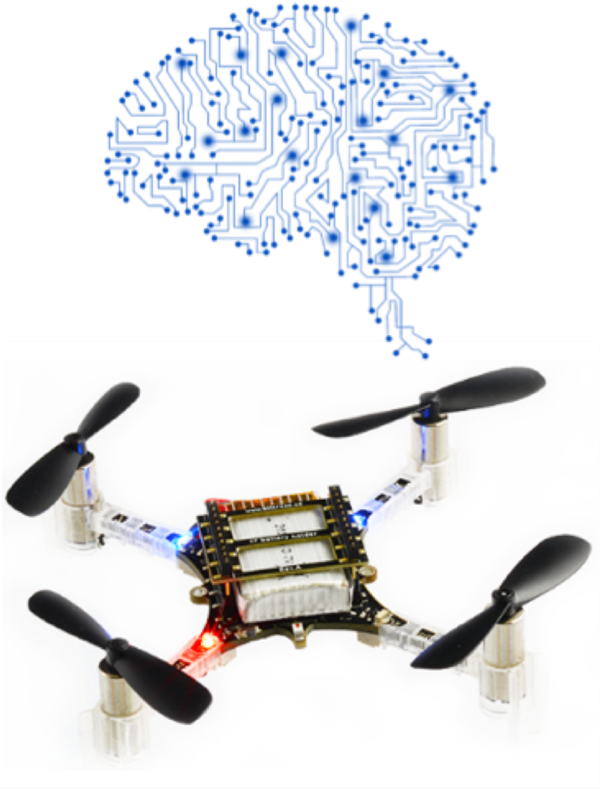
1 arm

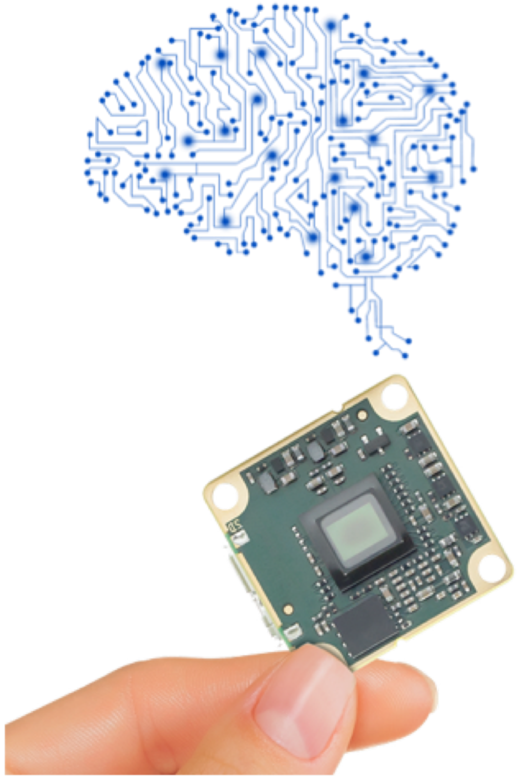
2 HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

3 SambaNova
SYSTEMS

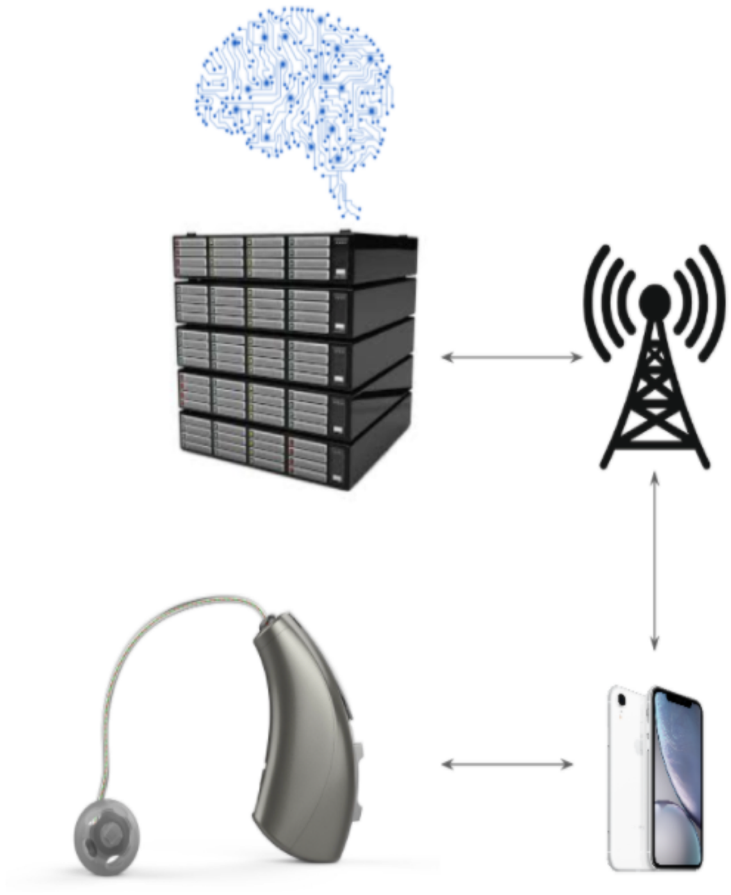
What is TinyML?

ML Inference at <1mWatt





IoT Paradigm

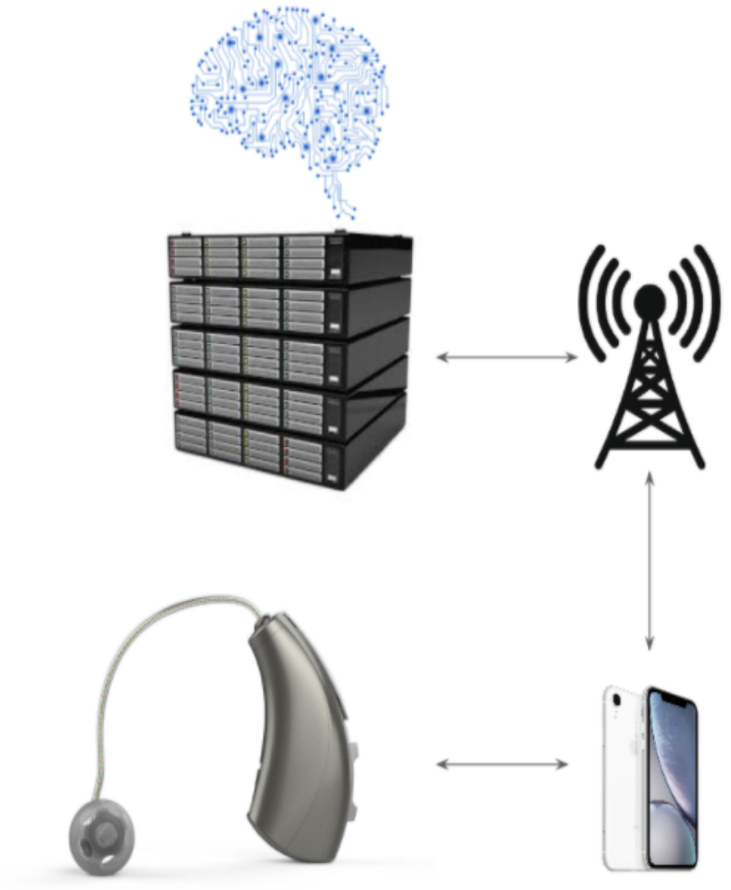


"Smart" Devices

=

Everything is collecting data

IoT’s Fatal Flaw

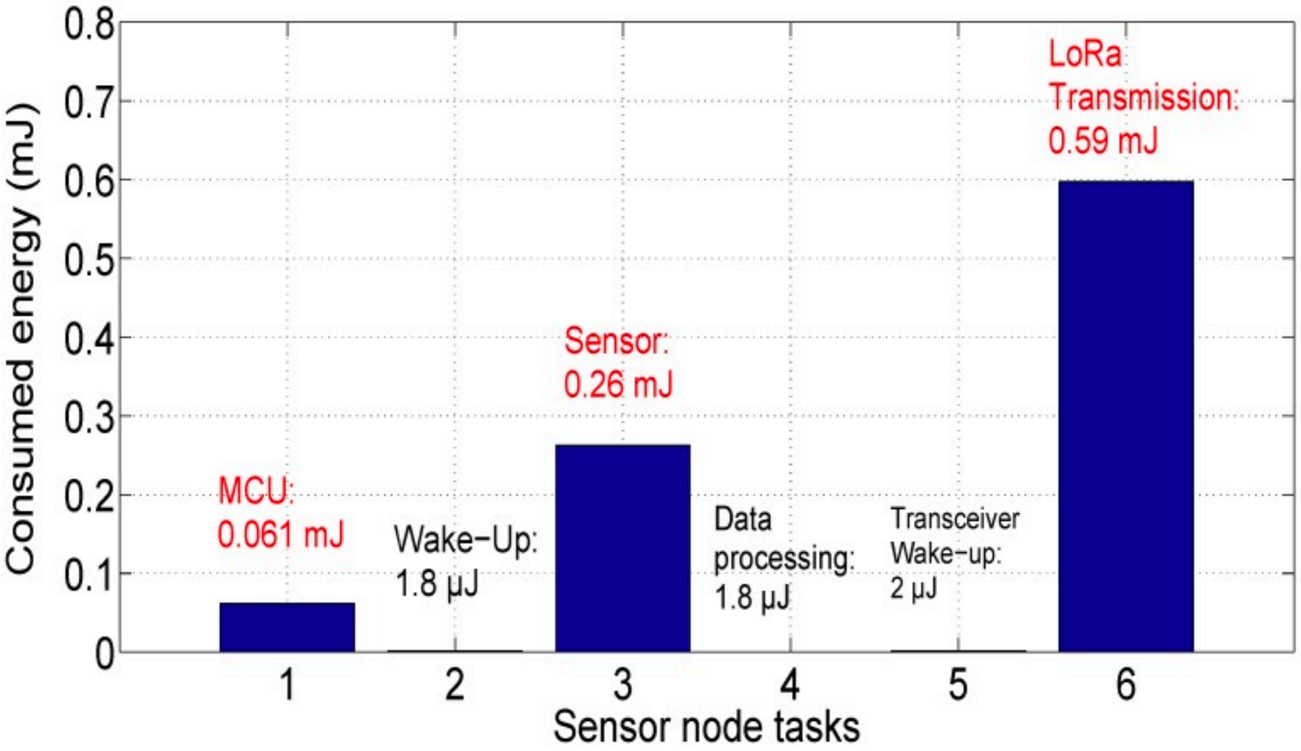


<1%

of IoT data is analyzed or used at all

Source: Mckinsey Global Institute. “The Internet of Things: Mapping the Value Beyond the Hype.” mckinsey.com

IoT’s Fatal Flaw



Bouguera, Taoufik et al. “Energy Consumption Model for Sensor Nodes Based on LoRa and LoRaWAN.” *Sensors (Basel, Switzerland)* vol. 18,7 2104. 30 Jun. 2018, doi:10.3390/s18072104

Transmission is Energy Hungry

The On-Device Advantage

+ Energy Efficiency

+ Responsiveness

+ Privacy

+ Mobility

The On-Device Advantage

+ Energy Efficiency

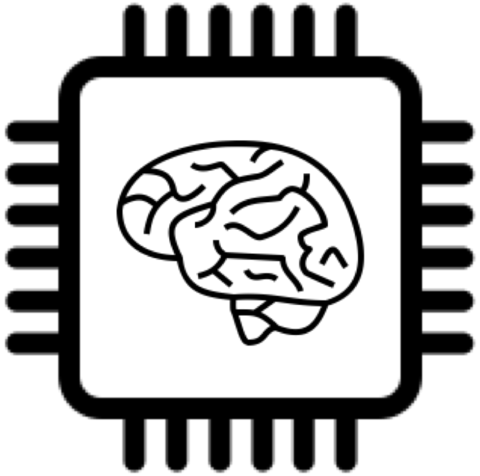
Goal

MicroNets

Create **efficient** and **deployable** model architectures for **tinyML** applications

Executive Summary

TinyML



TinyML systems have **severe constraints** and require **highly tuned** model architectures

SRAM

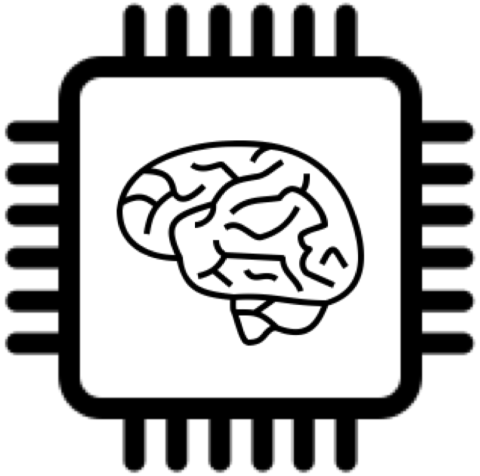
Flash

Latency

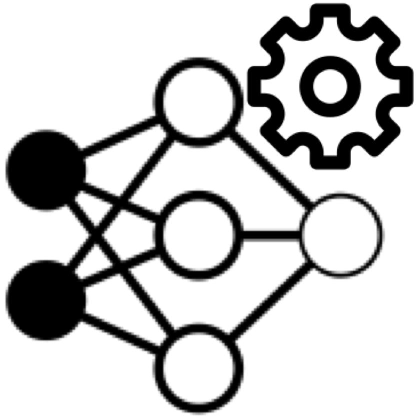
Energy

Executive Summary

TinyML



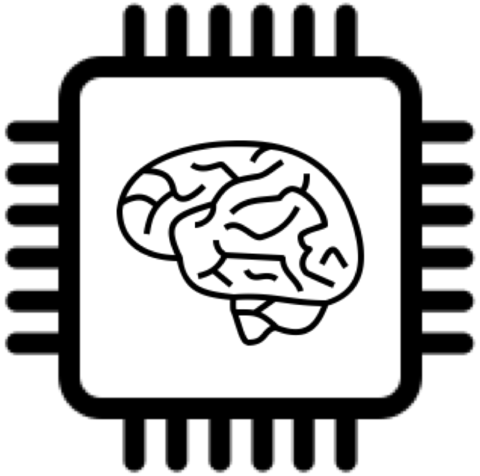
Differentiable
Neural Architecture
Search



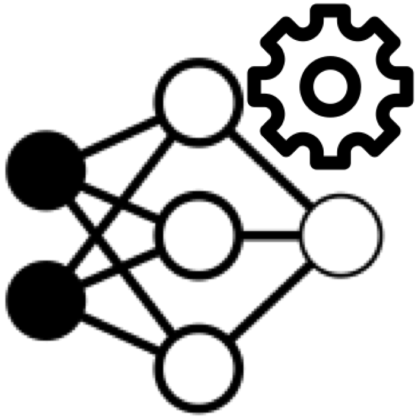
Differentiable Neural Architecture Search (DNAS) can **rapidly** find models that **meet the constraints** given **viable proxies**

Executive Summary

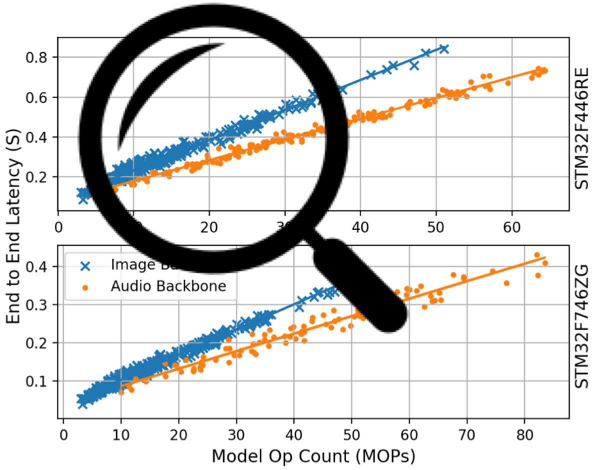
TinyML



Differentiable Neural Architecture Search



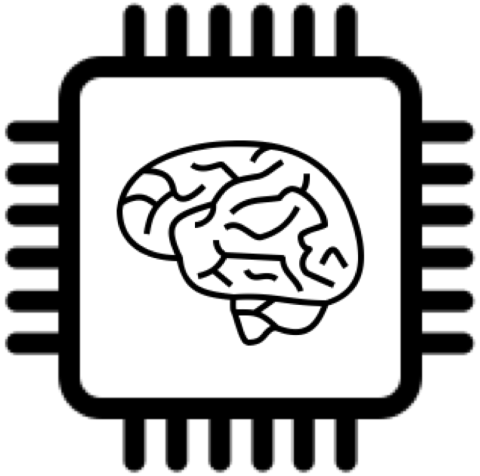
Hardware Characterization



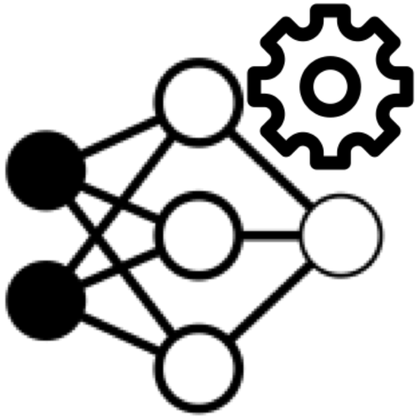
SRAM and Flash are easily calculated while **Op count is a viable proxy** latency and energy

Executive Summary

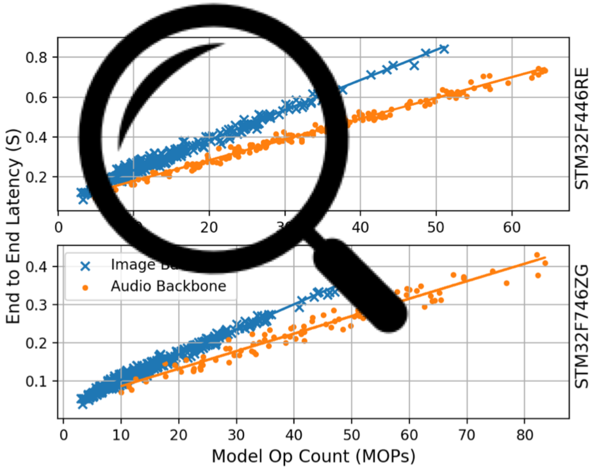
TinyML



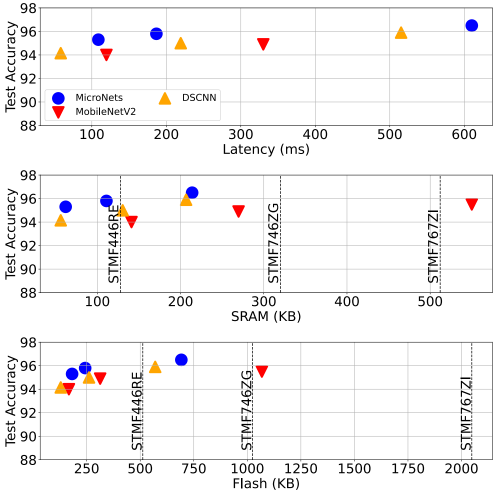
Differentiable Neural Architecture Search



Hardware Characterization



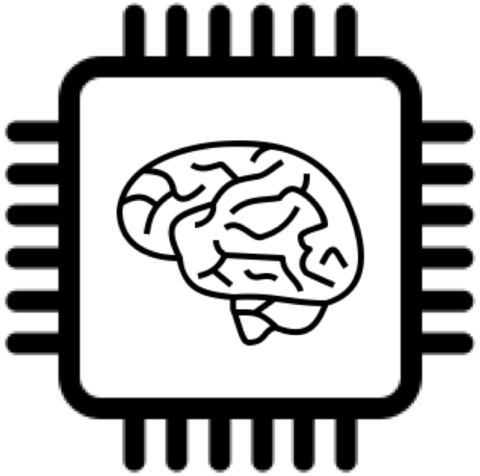
MicroNets



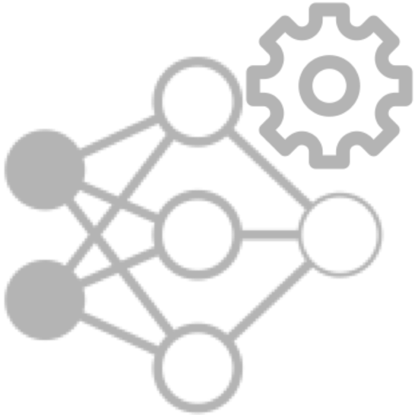
We achieve **state of the art performance** on three TinyML tasks

Executive Summary

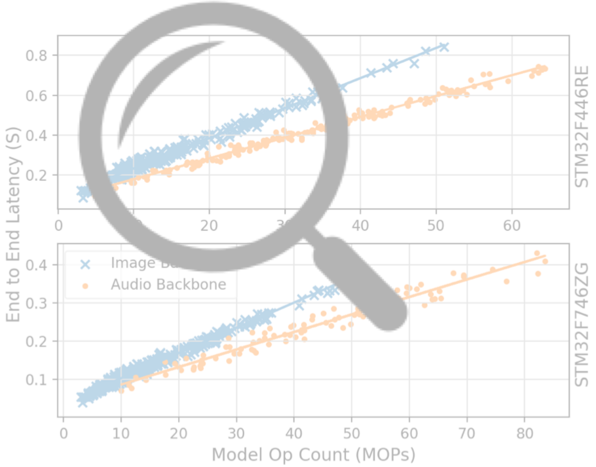
TinyML



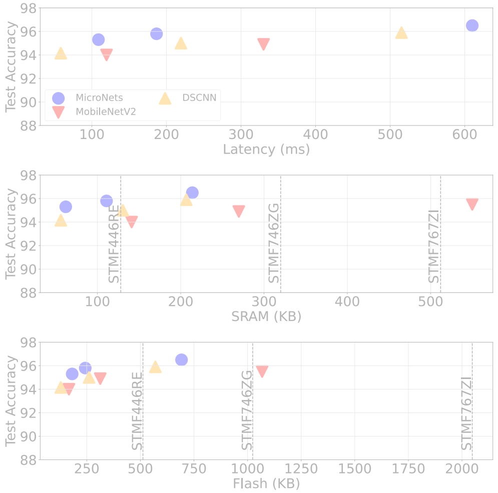
Differentiable Neural Architecture Search



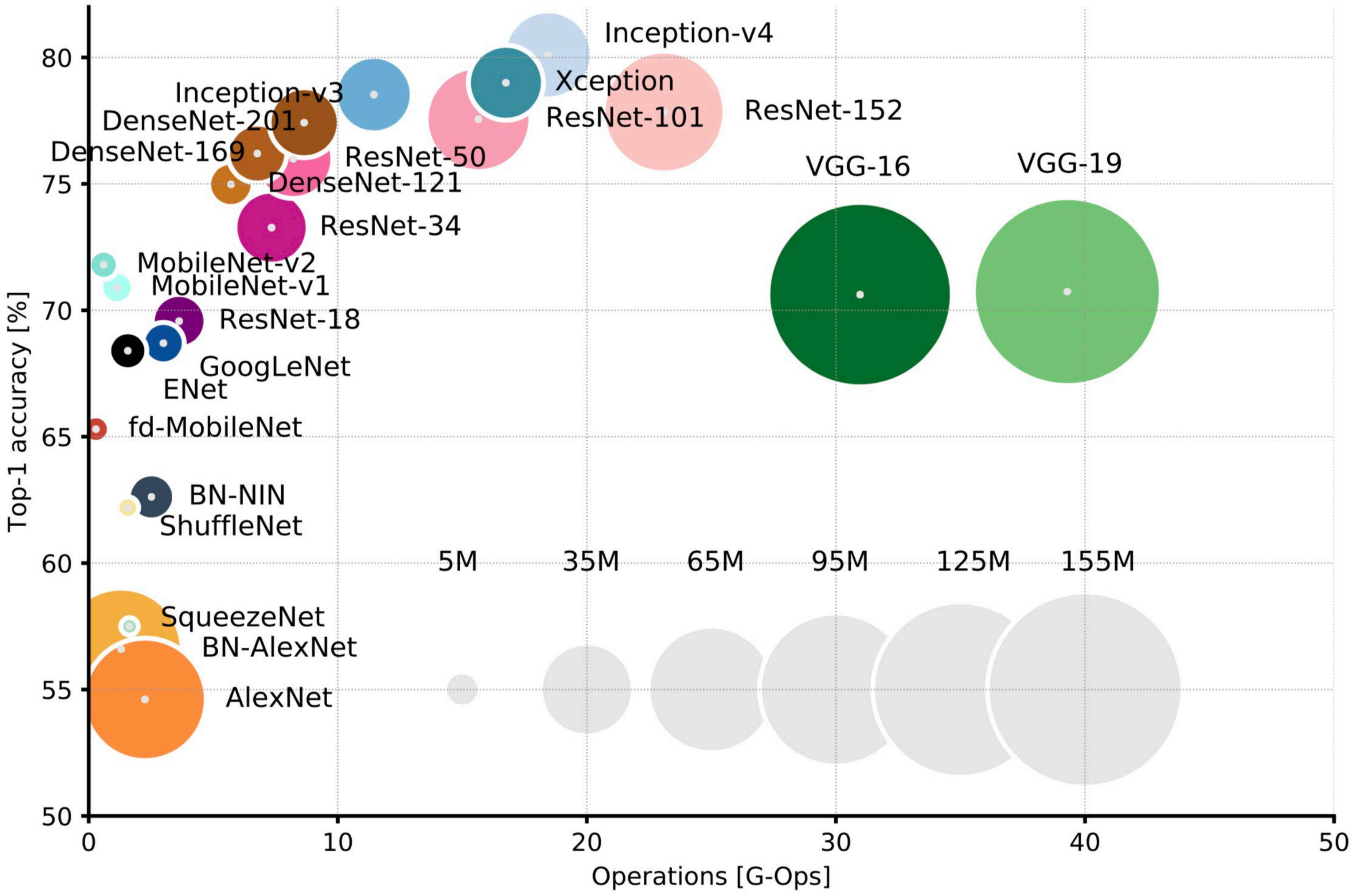
Hardware Characterization



MicroNets

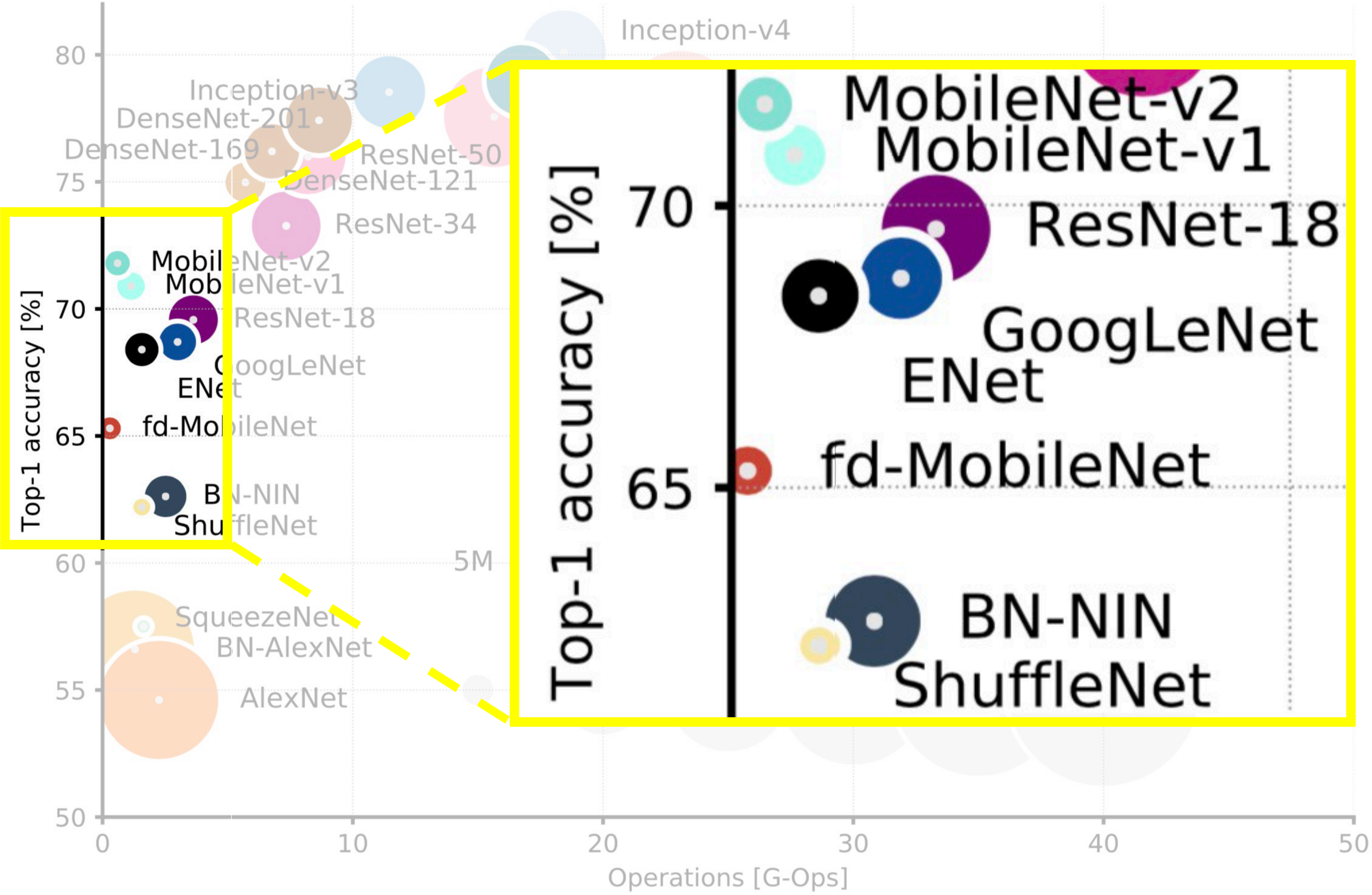


The Challenge



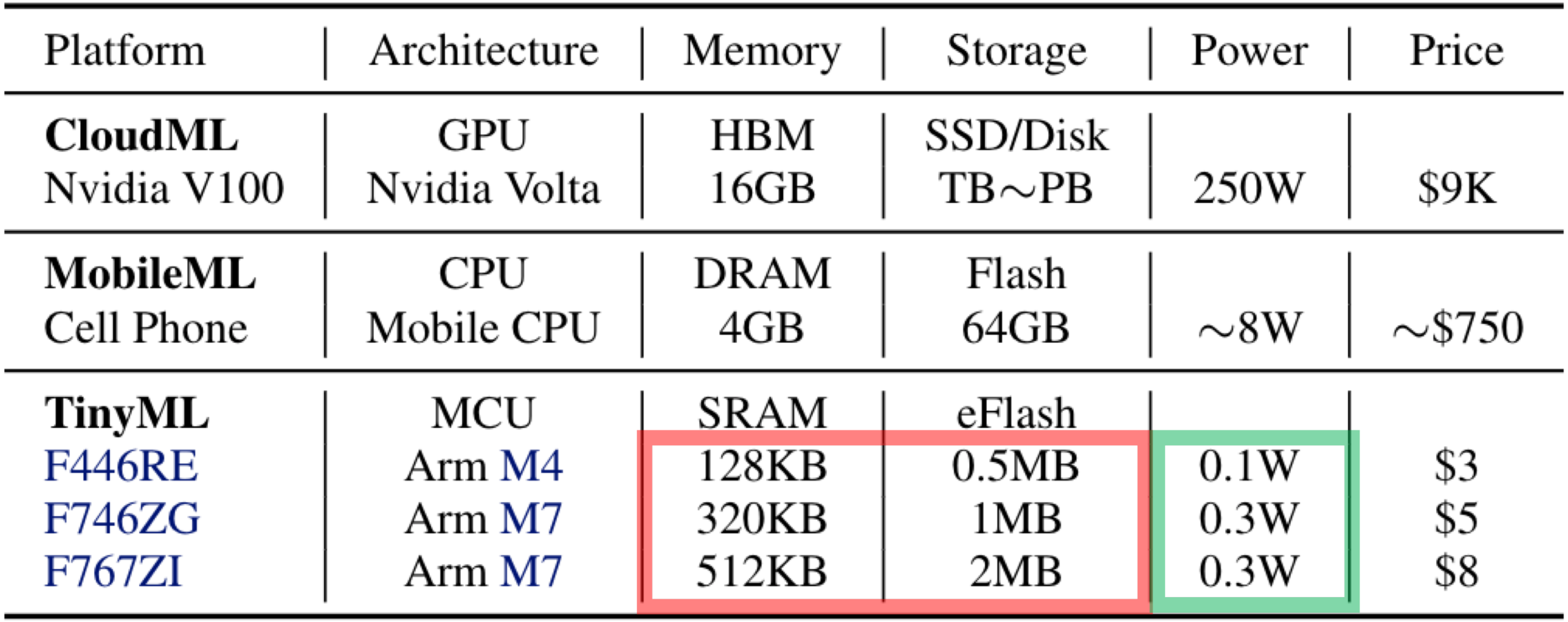
Source: https://culurciello.medium.com/analysis-of-deep-neural-networks-dcf398e71aae

The Challenge



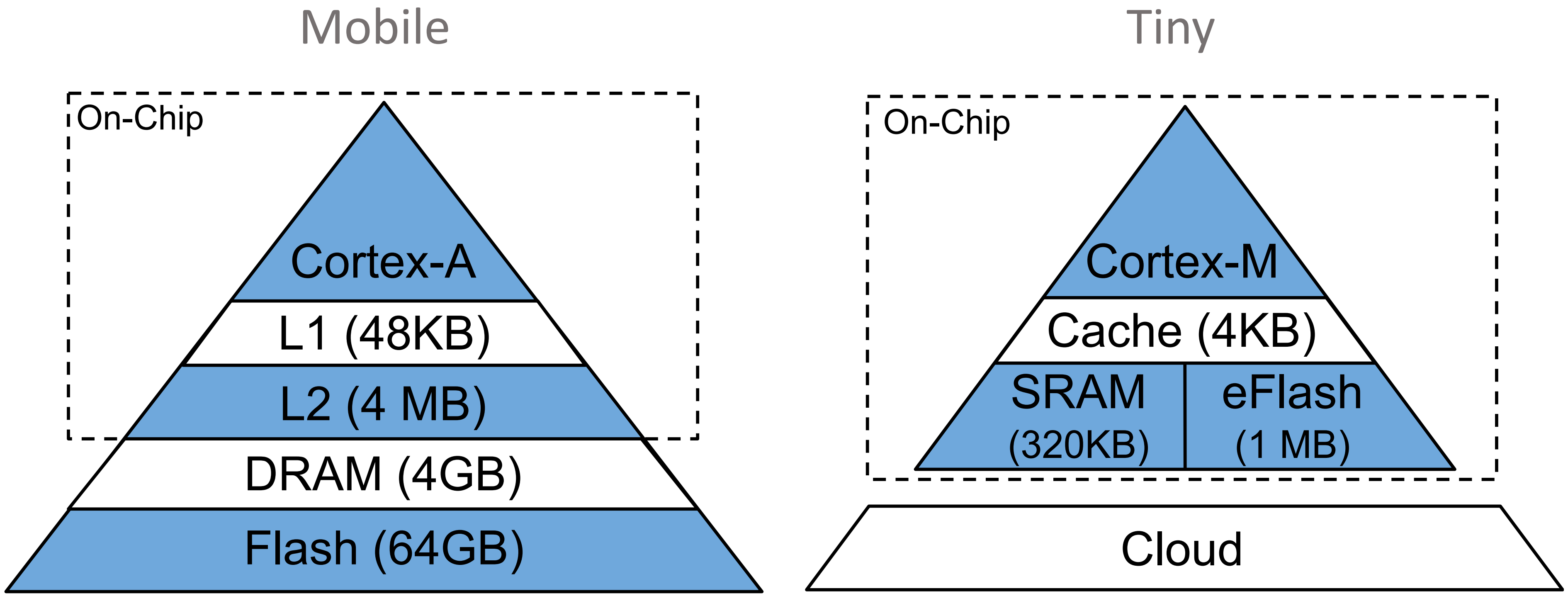
Source: https://culurciello.medium.com/analysis-of-deep-neural-networks-dcf398e71aae

The Constraints

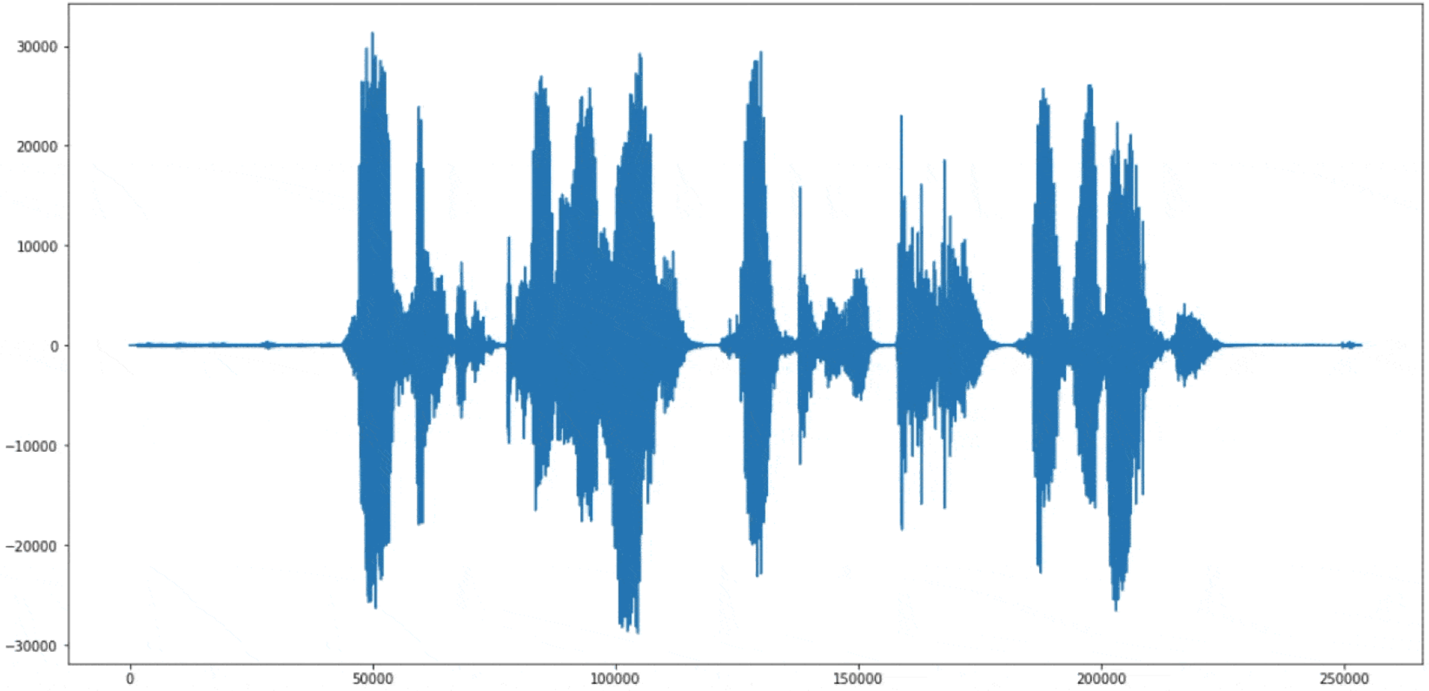


Platform	Architecture	Memory	Storage	Power	Price
CloudML Nvidia V100	GPU Nvidia Volta	HBM 16GB	SSD/Disk TB∼PB	250W	$9K
MobileML Cell Phone	CPU Mobile CPU	DRAM 4GB	Flash 64GB	∼8W	∼$750
TinyML F446RE F746ZG F767ZI	MCU Arm M4 Arm M7 Arm M7	SRAM 128KB 320KB 512KB	eFlash 0.5MB 1MB 2MB	 0.1W 0.3W 0.3W	 $3 $5 $8

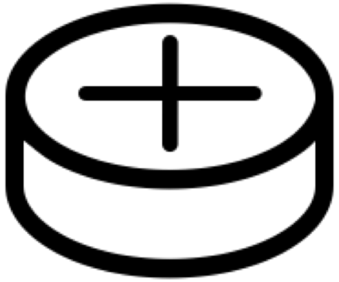
The Constraints



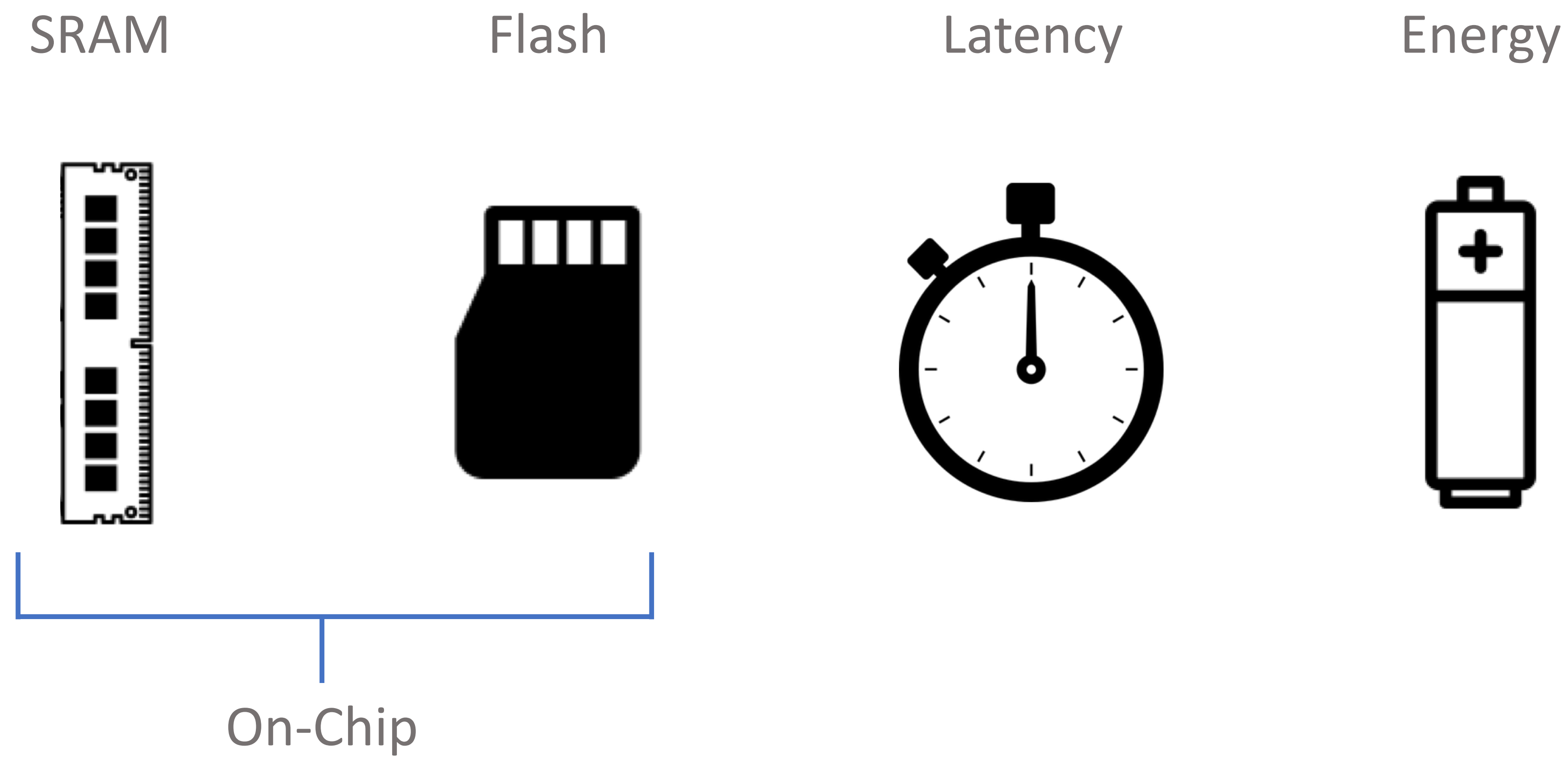
The Constraints



The Constraints



TinyML Constraints



SRAM

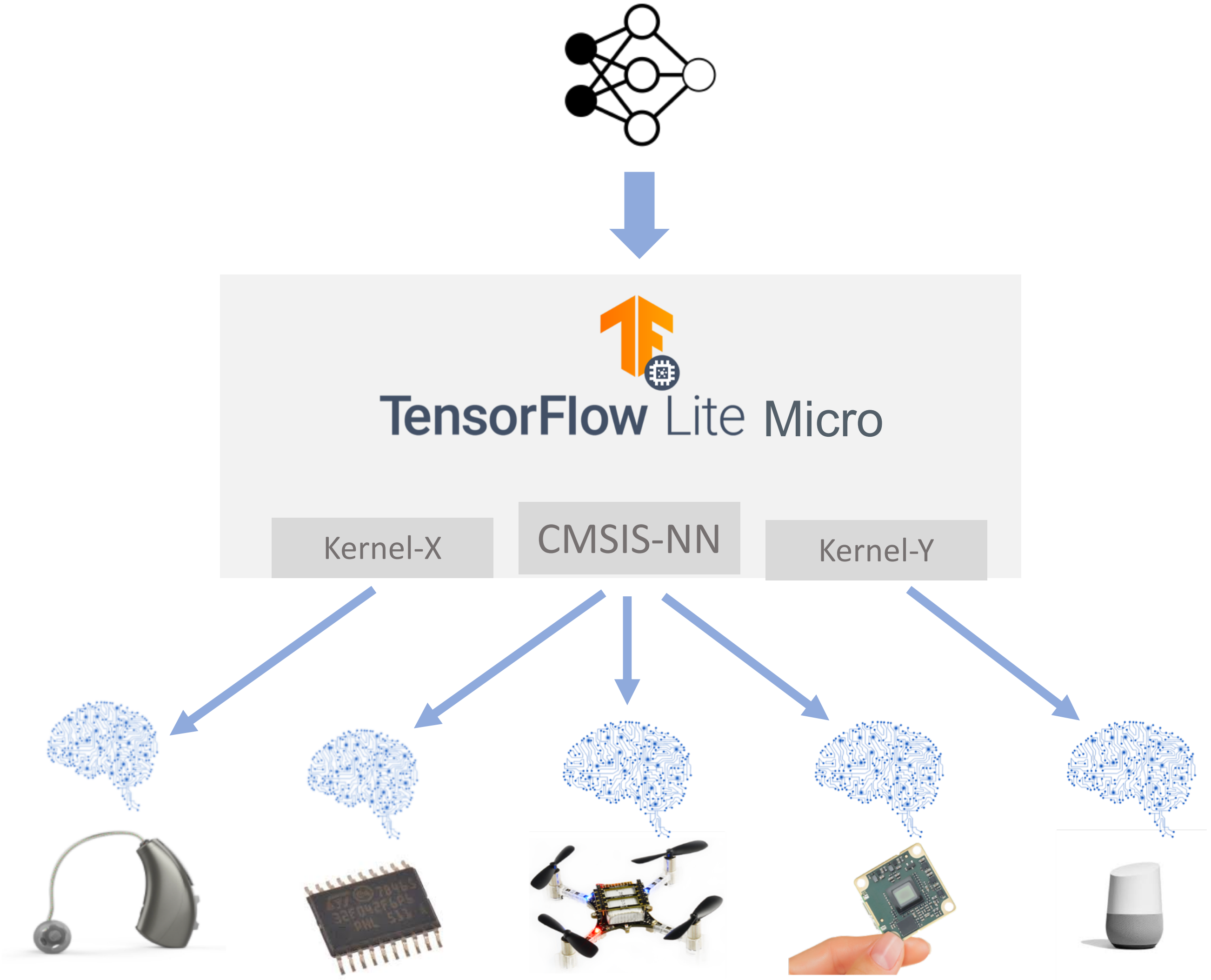
Flash

Latency

Energy

On-Chip

TensorFlow Lite for Microcontrollers



TensorFlow Lite Micro

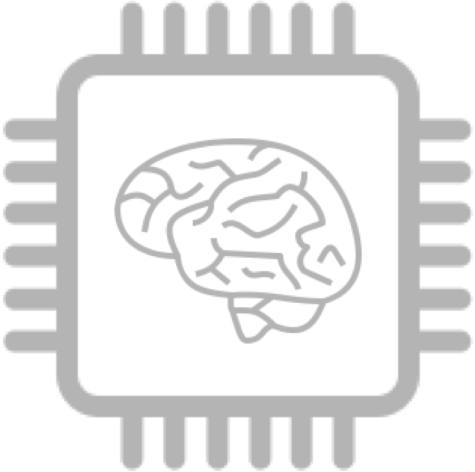
Kernel-X

CMSIS-NN

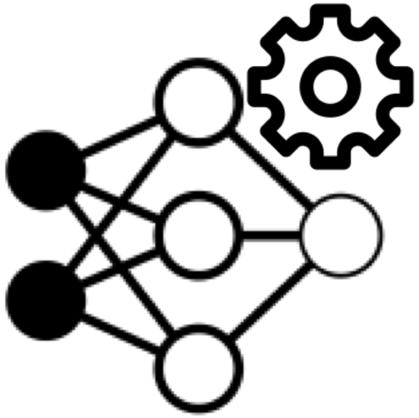
Kernel-Y

Executive Summary

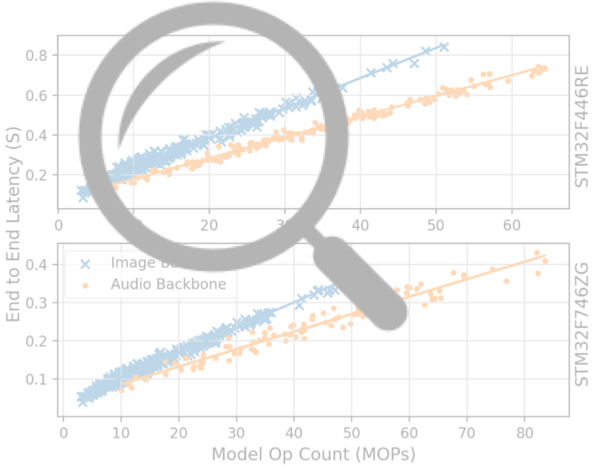
TinyML



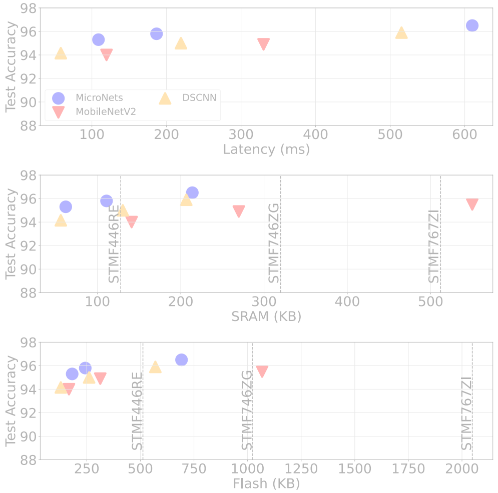
Differentiable Neural Architecture Search



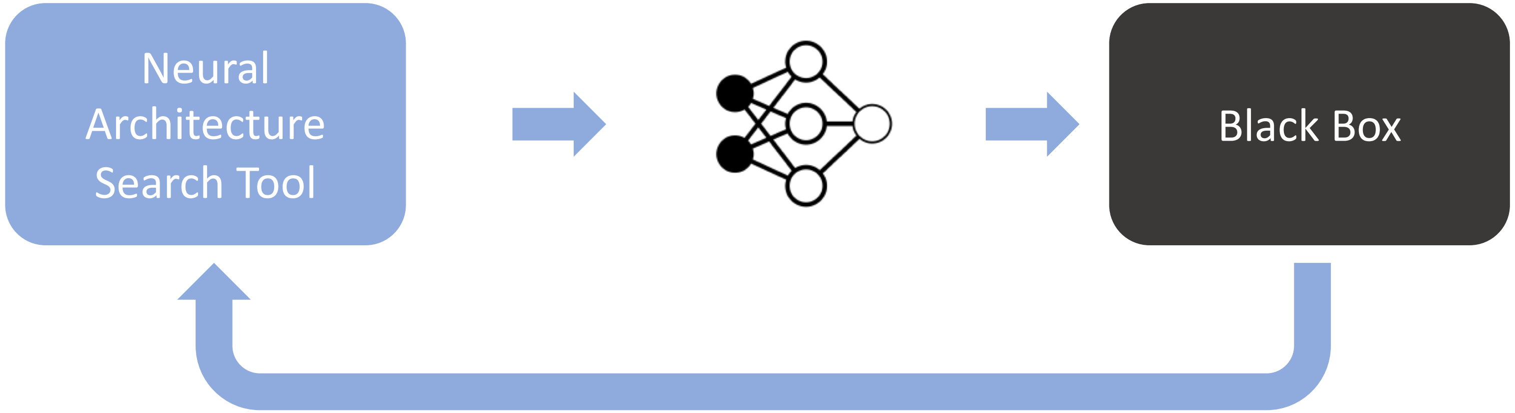
Hardware Characterization



MicroNets

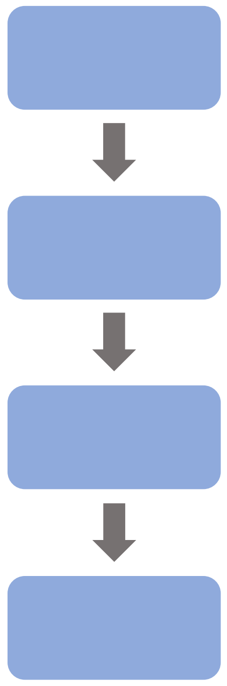


Neural Architecture Search (NAS)

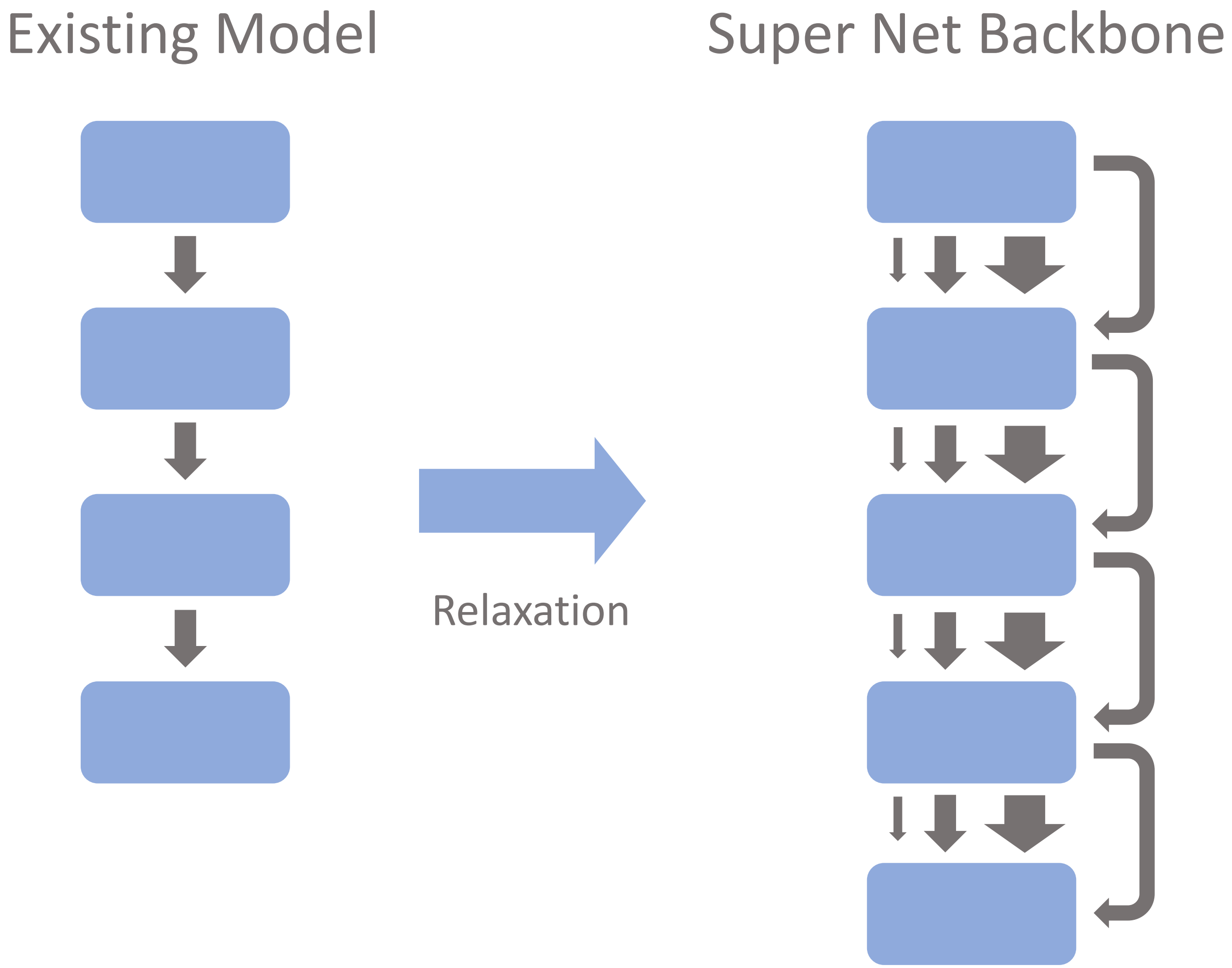


Differentiable Neural Architecture Search (DNAS)

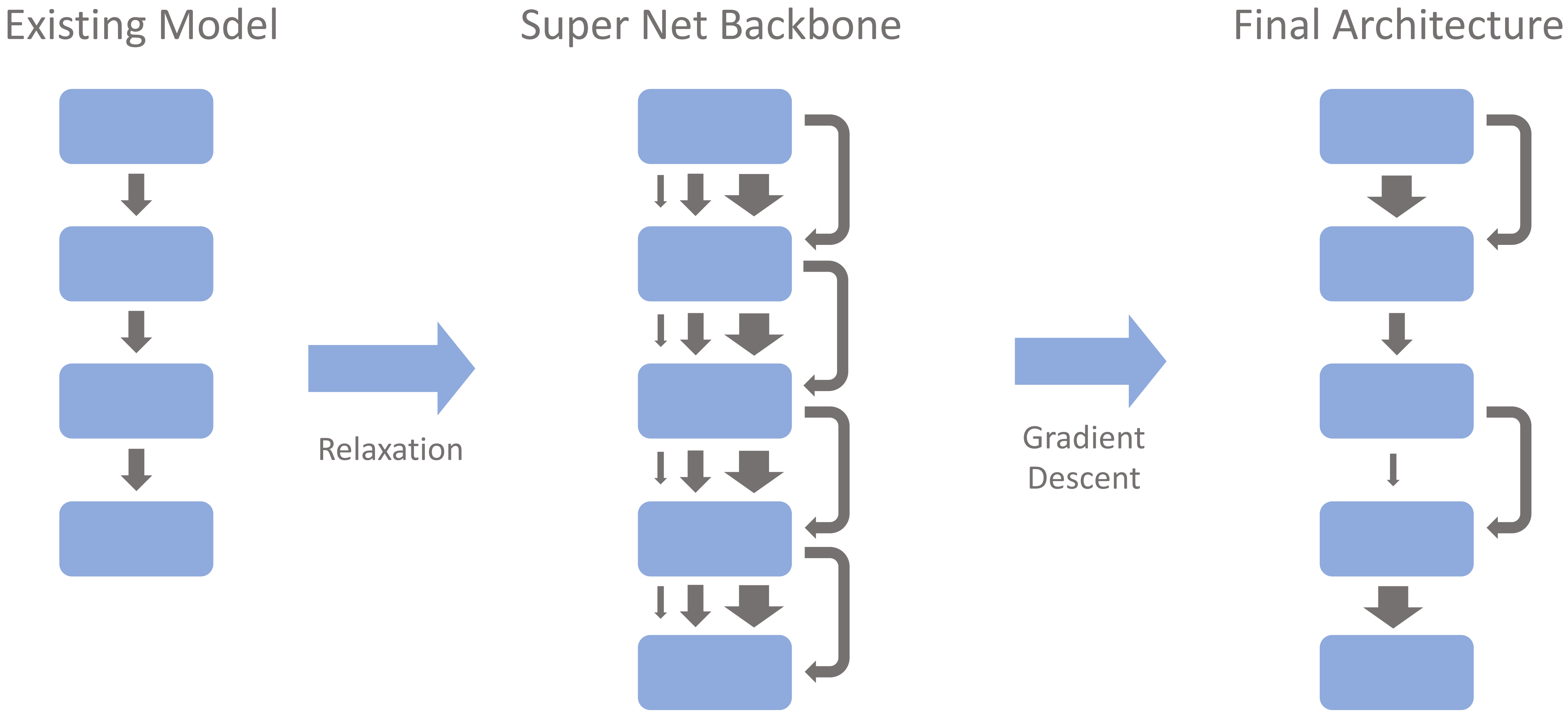
Existing Model



Differentiable Neural Architecture Search (DNAS)



Differentiable Neural Architecture Search (DNAS)



Existing Model

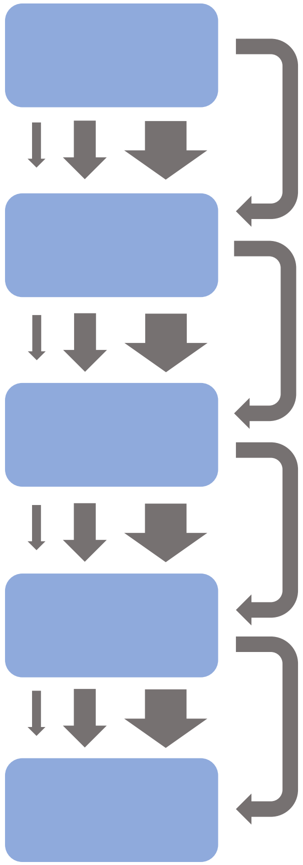
Relaxation

Super Net Backbone

Gradient Descent

Final Architecture

Differentiable Neural Architecture Search (DNAS)



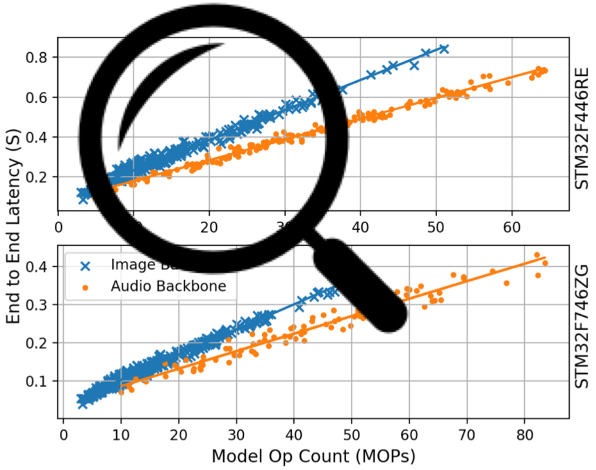
DNAS is **fast** but needs **continuous functions** for the hardware objectives

Executive Summary

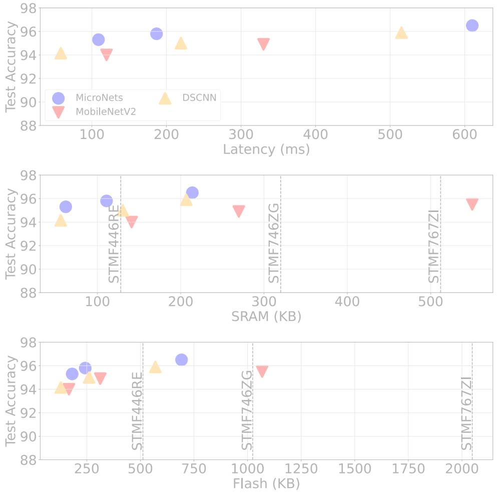
TinyML

Differentiable Neural Architecture Search

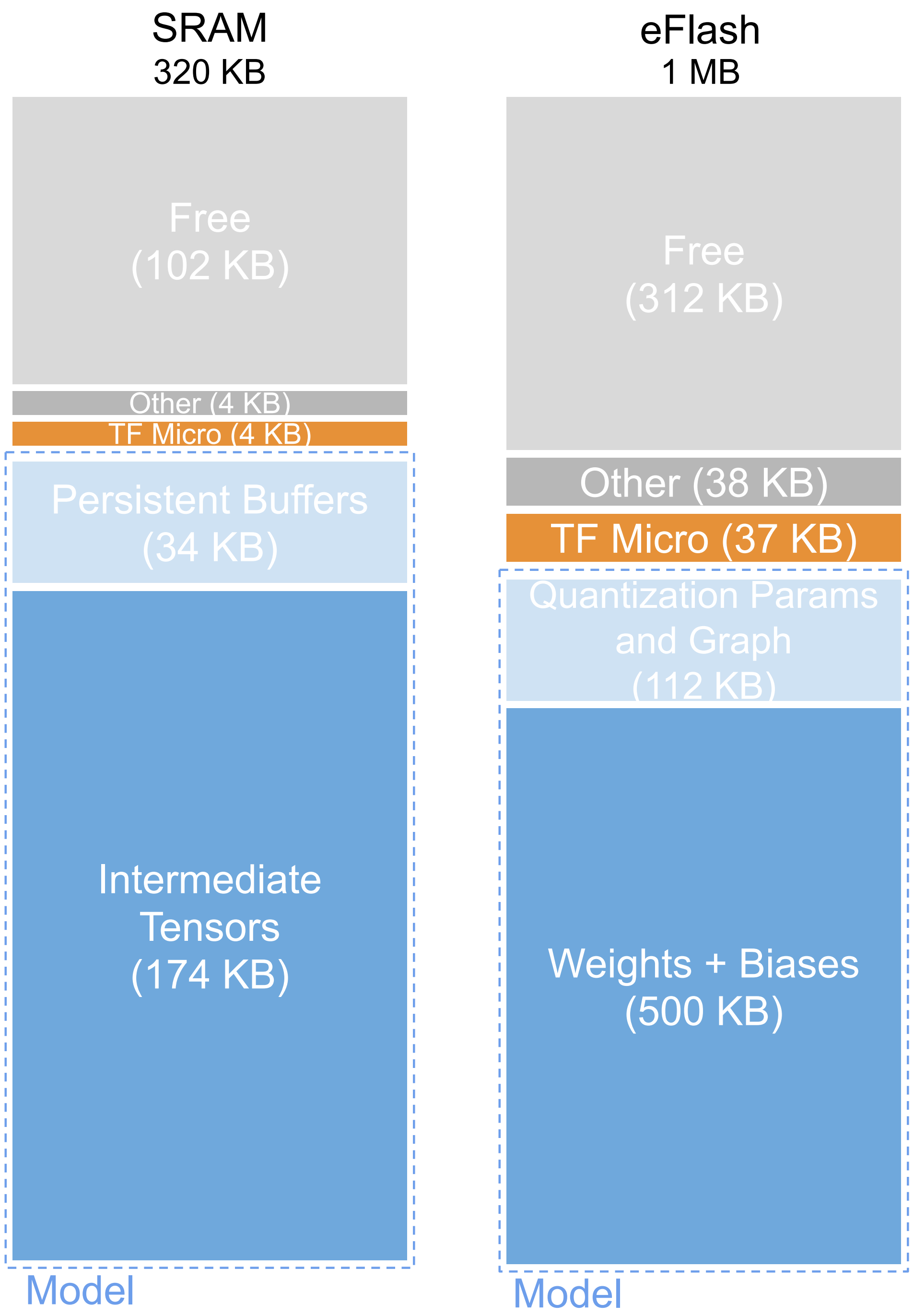
Hardware Characterization

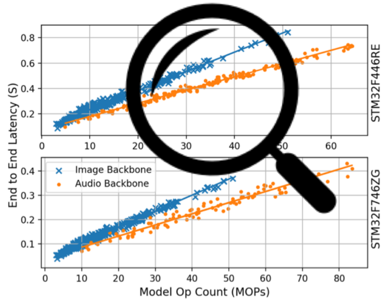


MicroNets

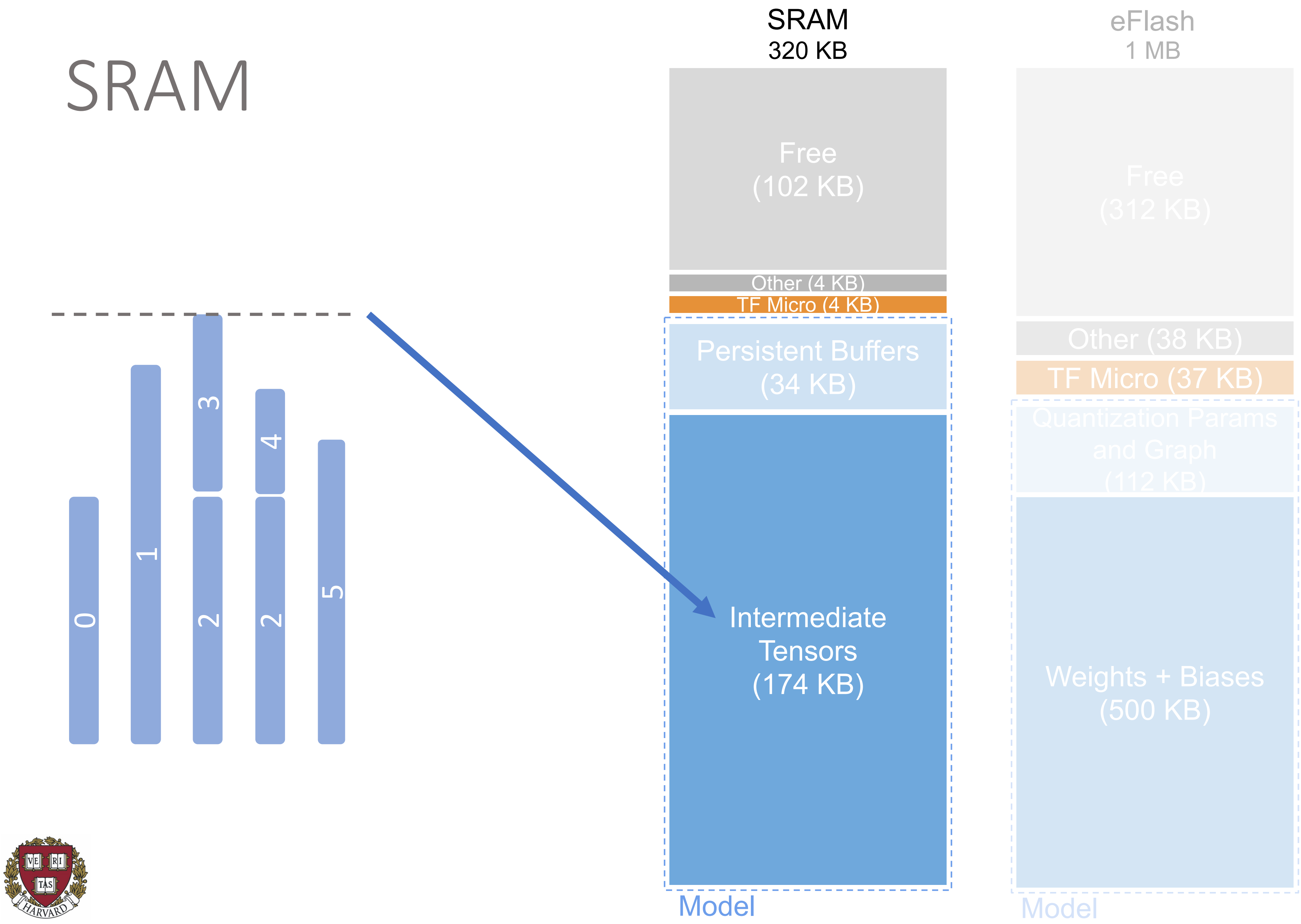


SRAM and Flash





SRAM



SRAM
320 KB

Free
(102 KB)

Other (4 KB)

TF Micro (4 KB)

Persistent Buffers
(34 KB)

Intermediate
Tensors
(174 KB)

Model

eFlash
1 MB

Free
(312 KB)

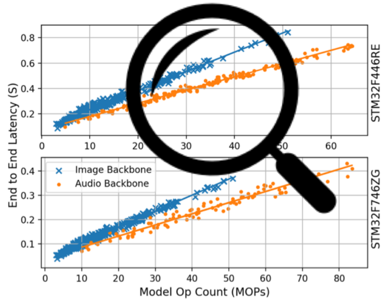
Other (38 KB)

TF Micro (37 KB)

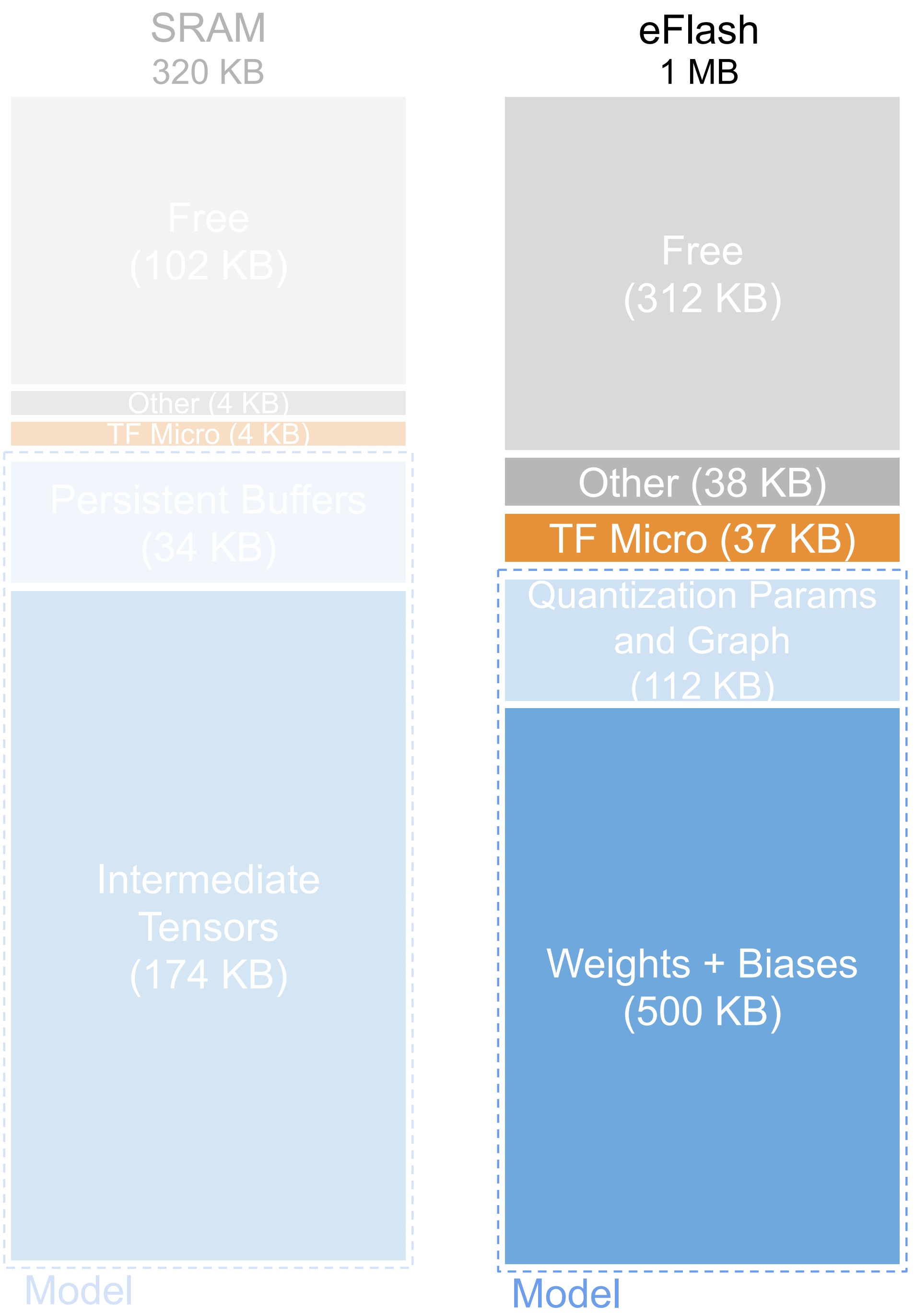
Quantization Params
and Graph
(112 KB)

Weights + Biases
(500 KB)

Model



Flash



SRAM
320 KB

Free
(102 KB)

Other (4 KB)

TF Micro (4 KB)

Persistent Buffers
(34 KB)

Intermediate
Tensors
(174 KB)

Model

eFlash
1 MB

Free
(312 KB)

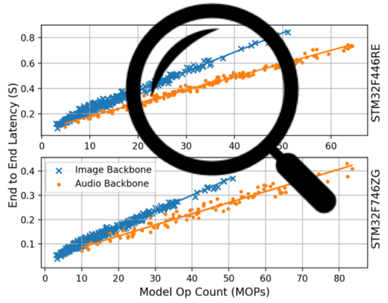
Other (38 KB)

TF Micro (37 KB)

Quantization Params
and Graph
(112 KB)

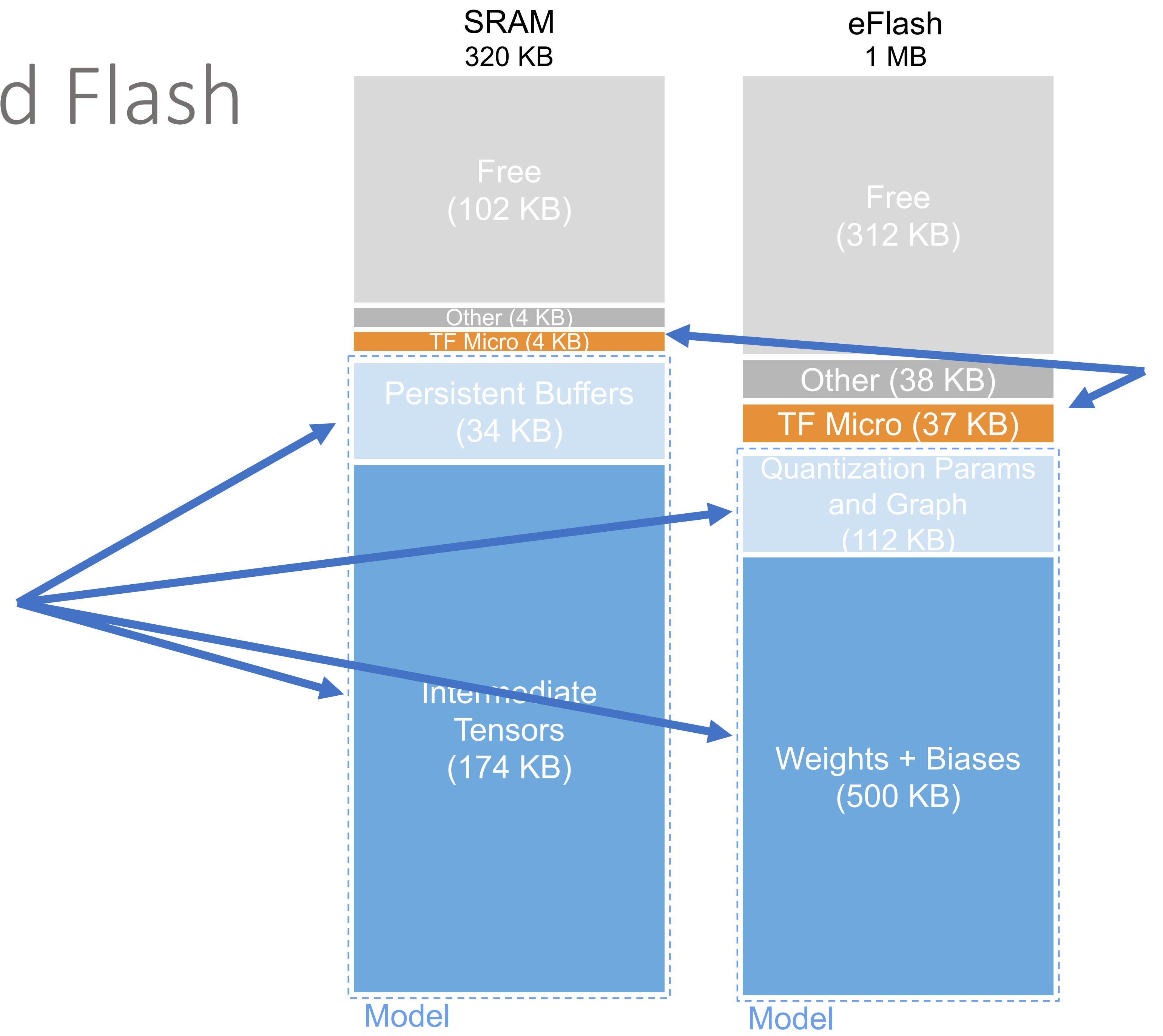
Weights + Biases
(500 KB)

Model



SRAM and Flash

Determined by the Model Architecture



SRAM
320 KB

Free
(102 KB)

Other (4 KB)

TF Micro (4 KB)

Persistent Buffers
(34 KB)

Intermediate
Tensors
(174 KB)

Model

eFlash
1 MB

Free
(312 KB)

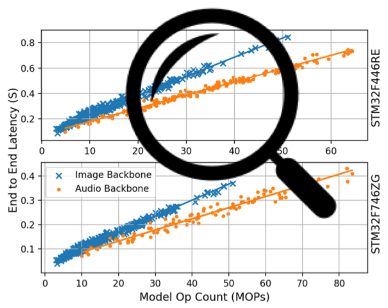
Other (38 KB)

TF Micro (37 KB)

Quantization Params
and Graph
(112 KB)

Weights + Biases
(500 KB)

Model



Overhead

TinyML Constraints

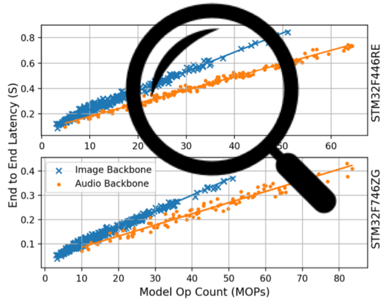
✓ SRAM

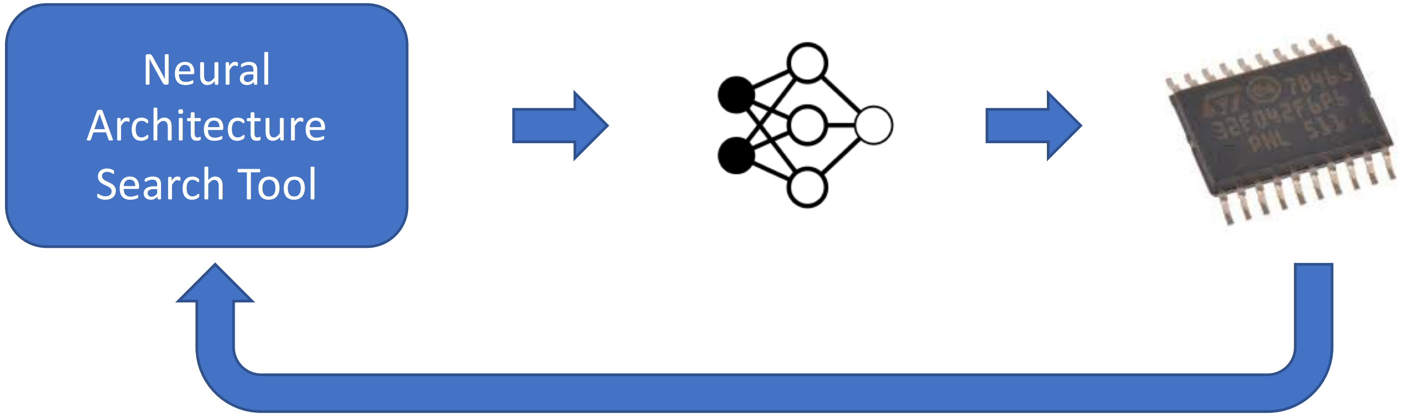
✓ Flash

Latency

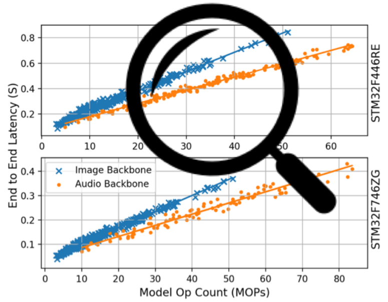
Energy

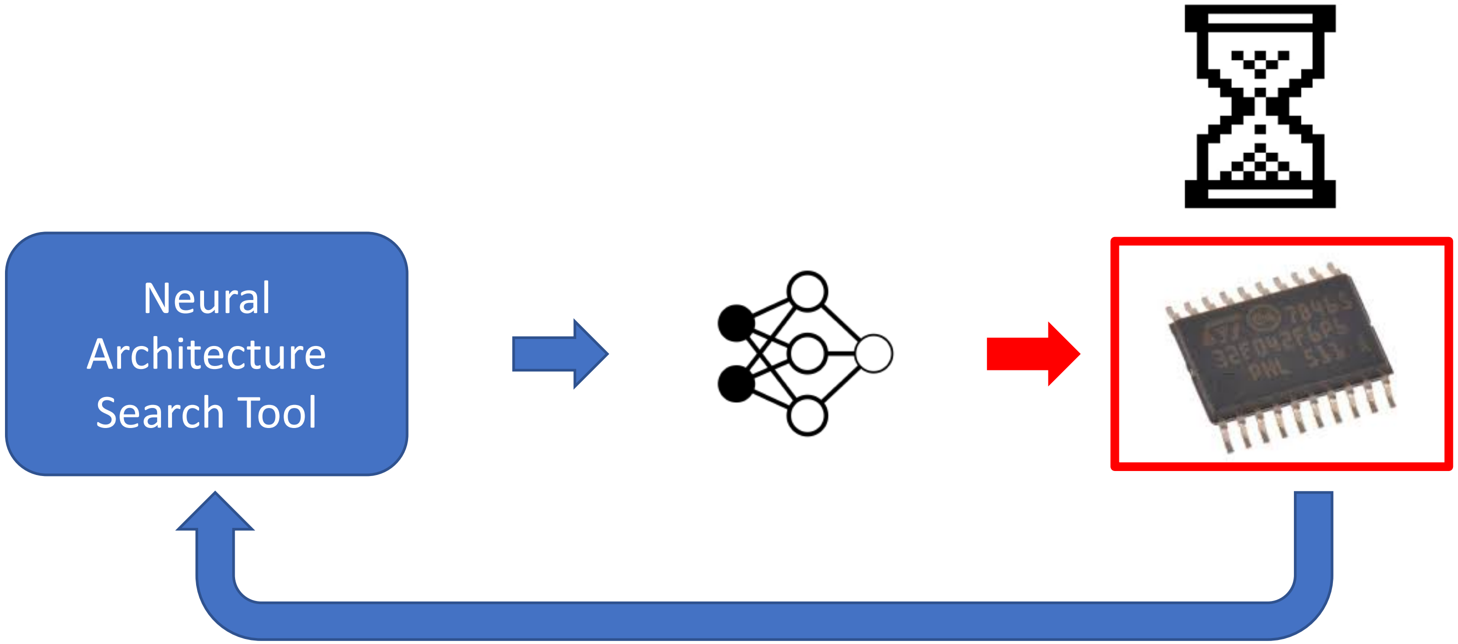
Direct Latency Benchmarking





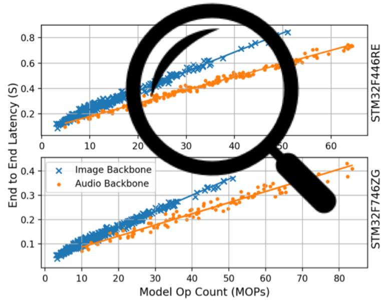
Direct Latency Benchmarking

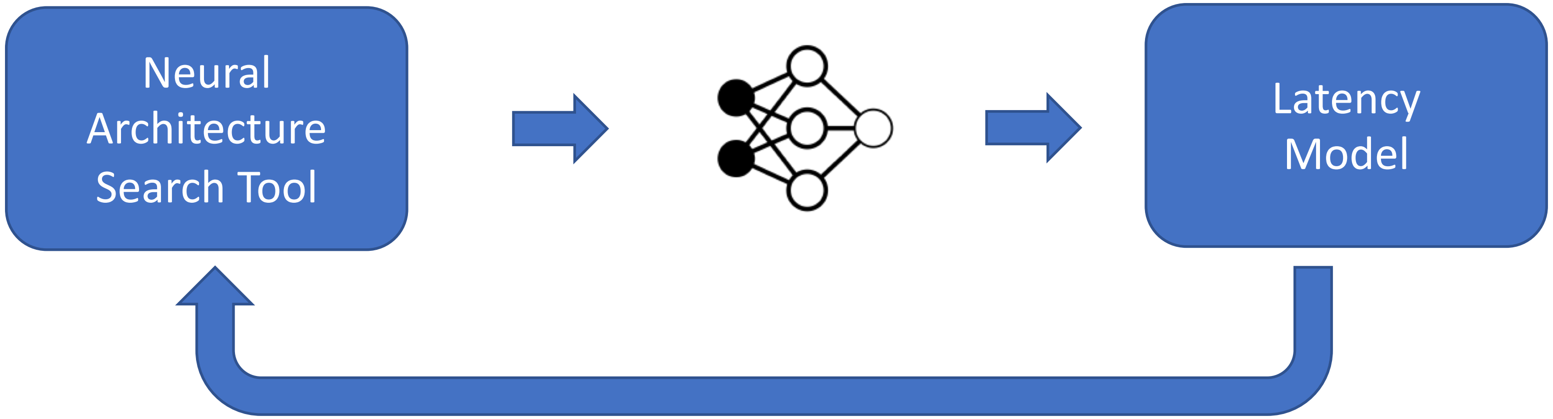




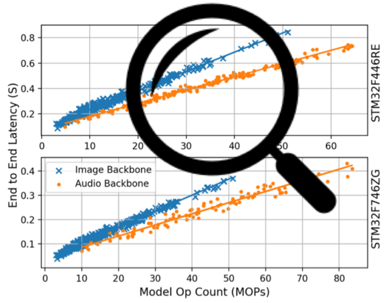
Neural Architecture Search Tool

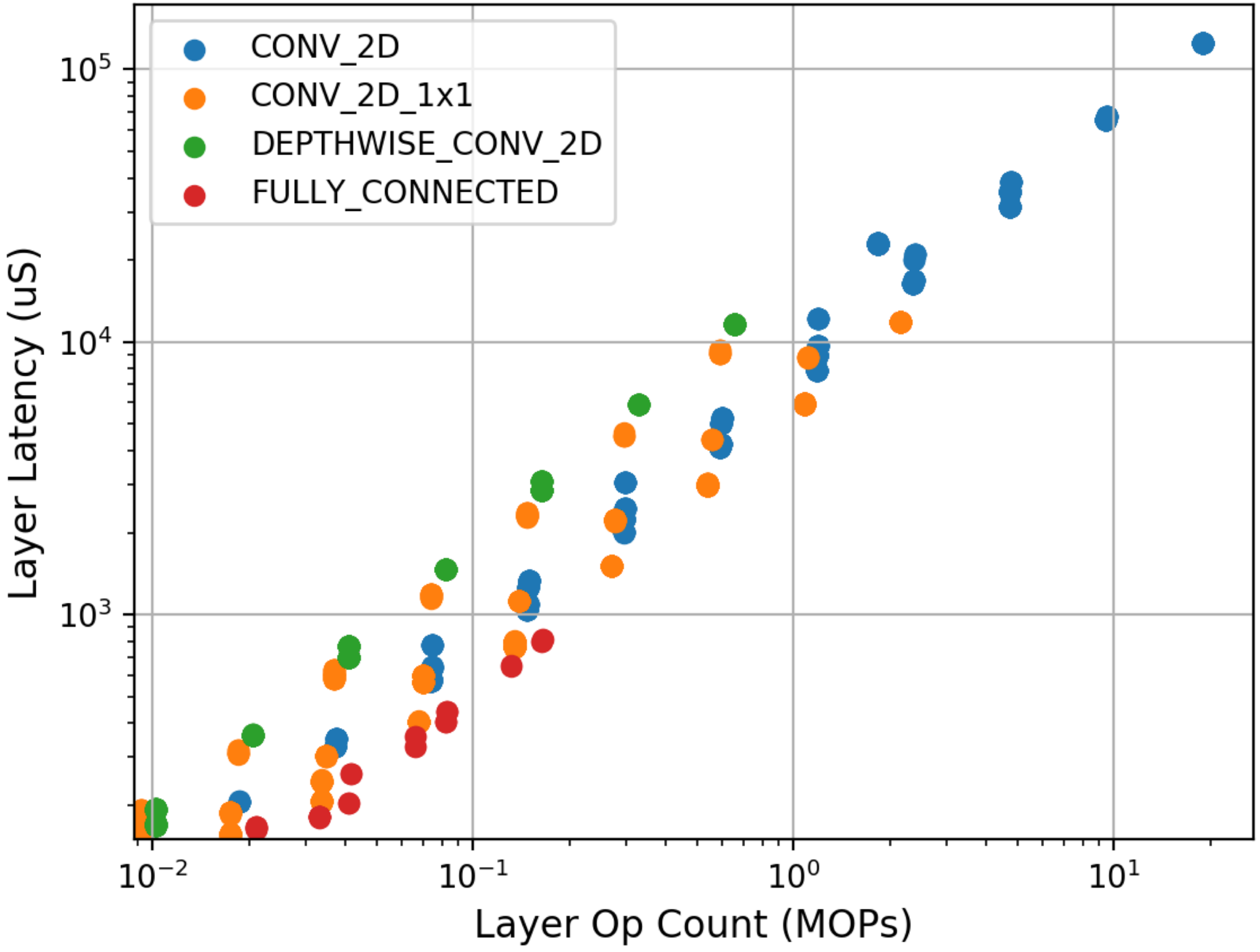
Latency Model



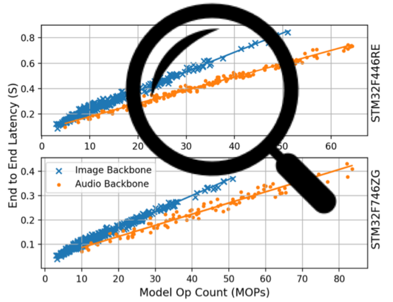


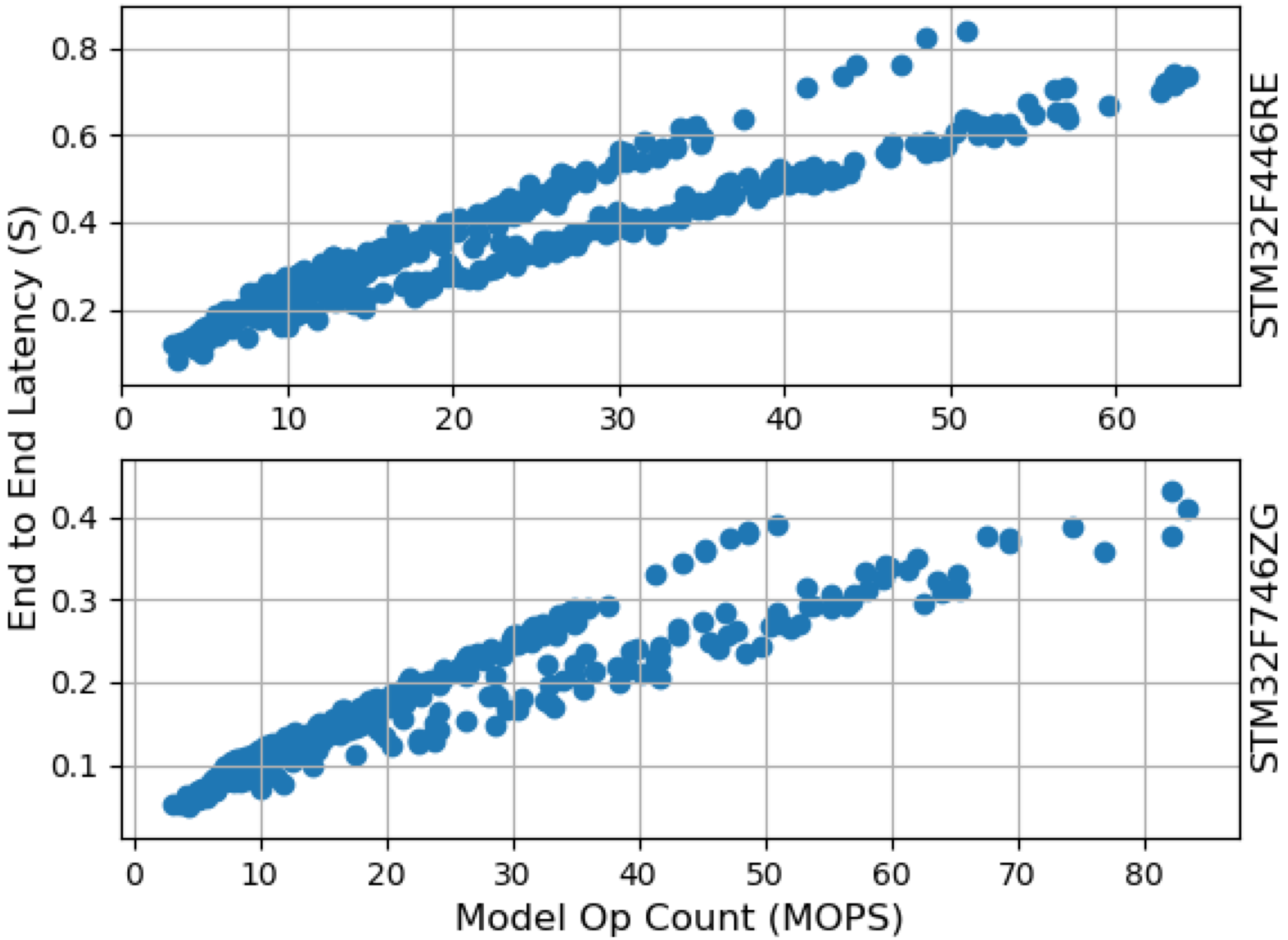
Per Layer Latency



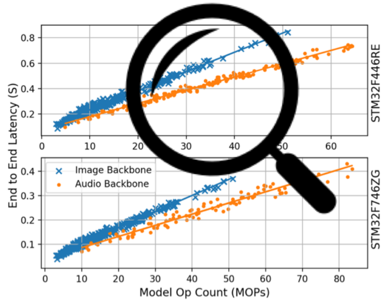


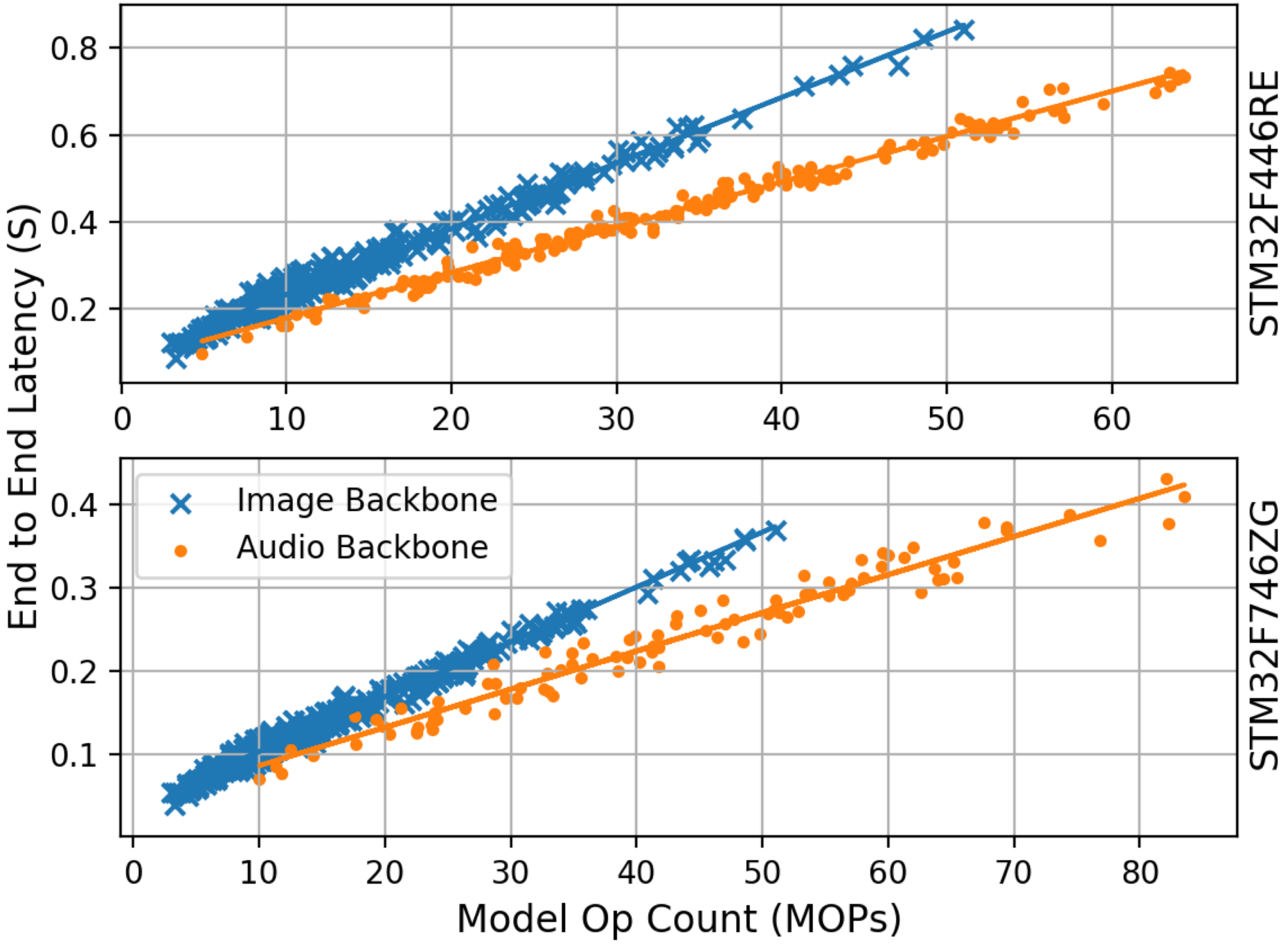
Model Latency



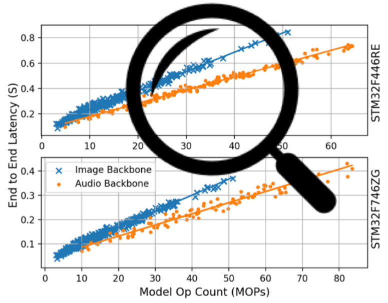


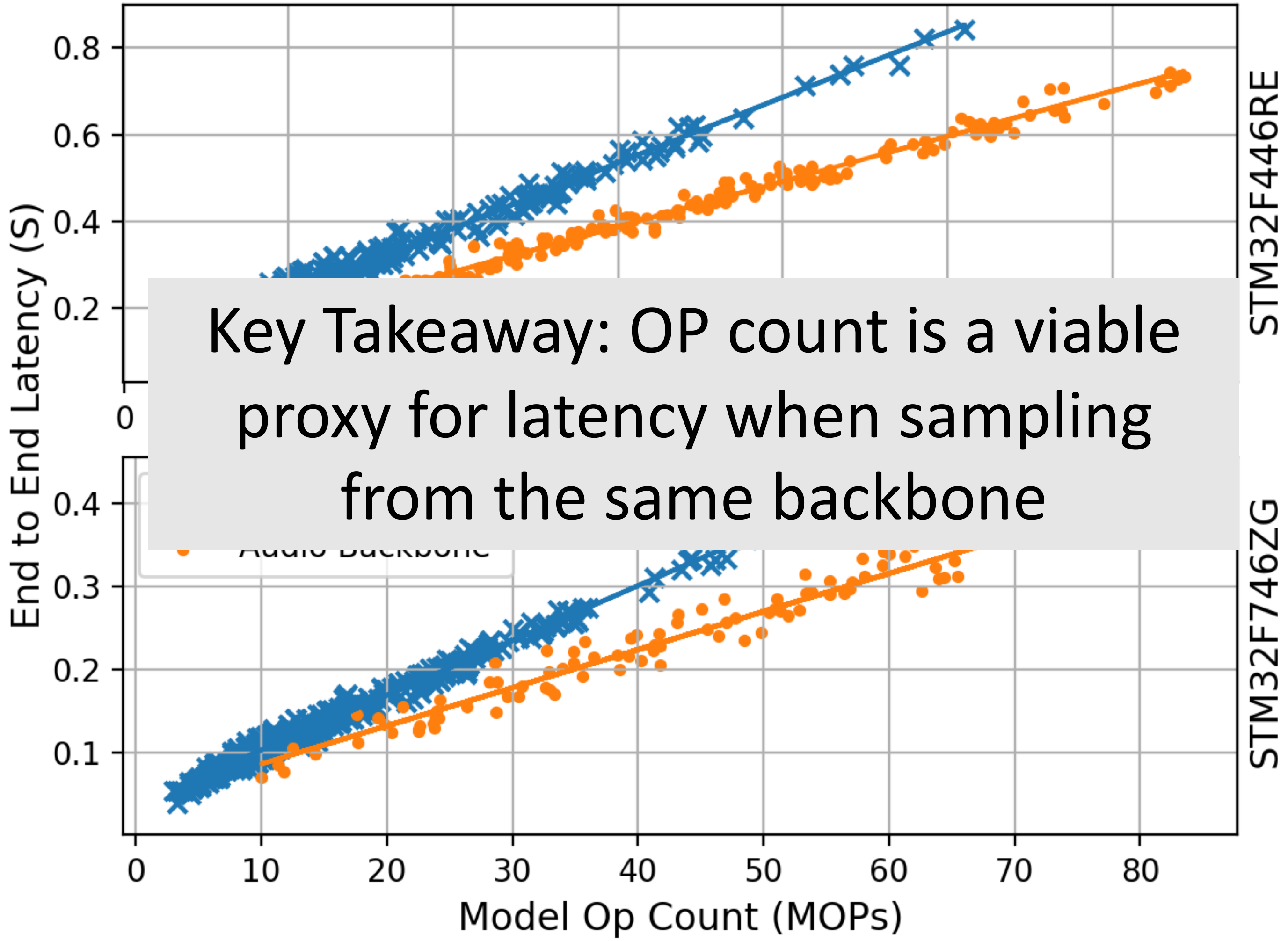
Model Latency





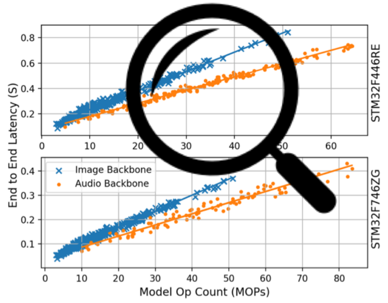
Model Latency

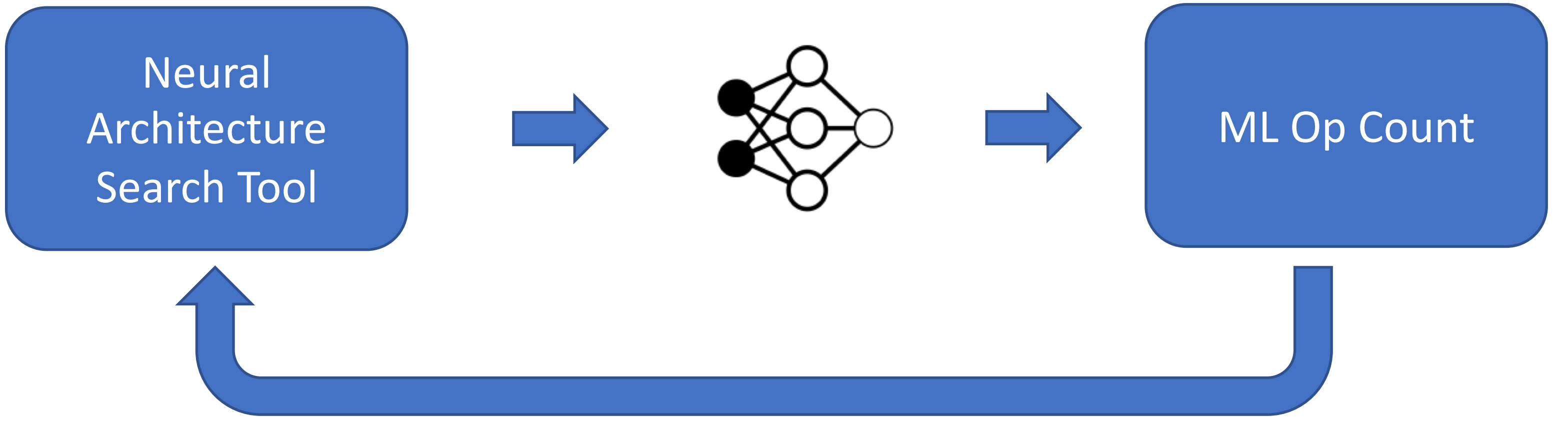




Key Takeaway: OP count is a viable proxy for latency when sampling from the same backbone

Latency Model



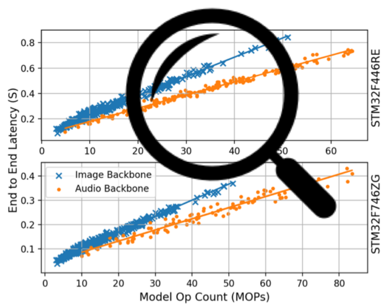


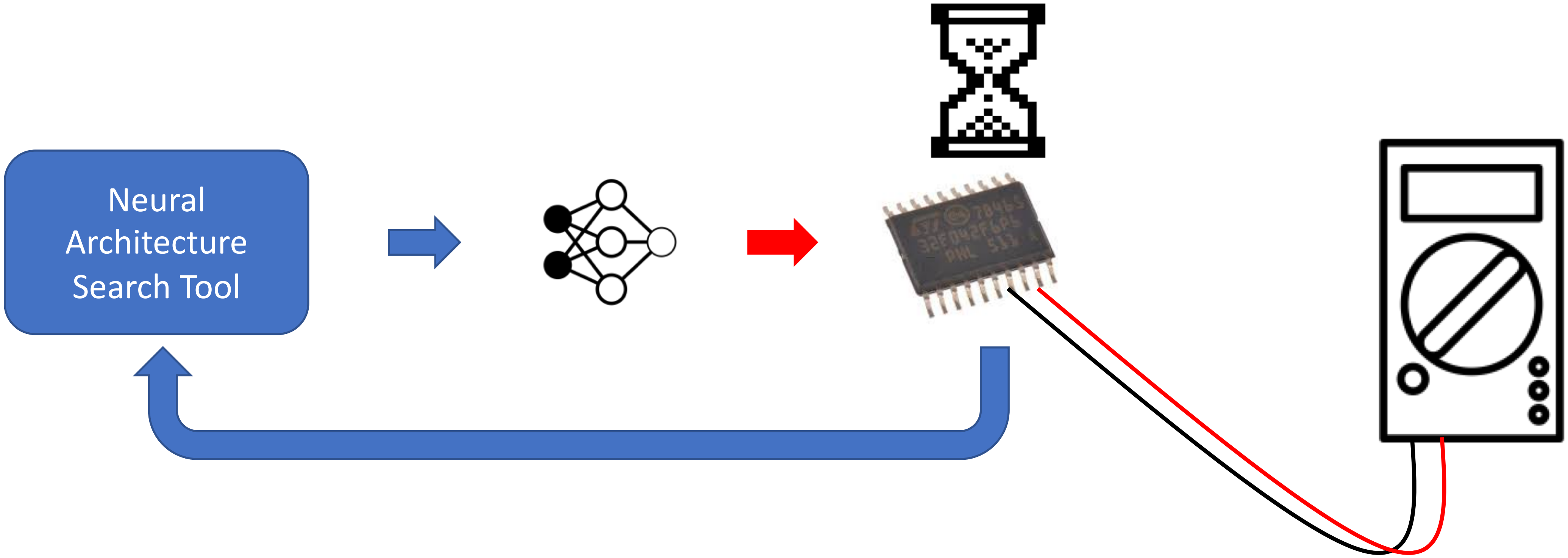
Latency Estimate

TinyML Constraints

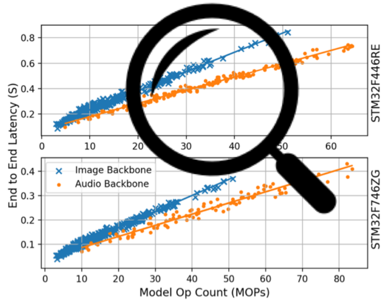
- ✓ SRAM
- ✓ Flash
- ✓ Latency
- Energy

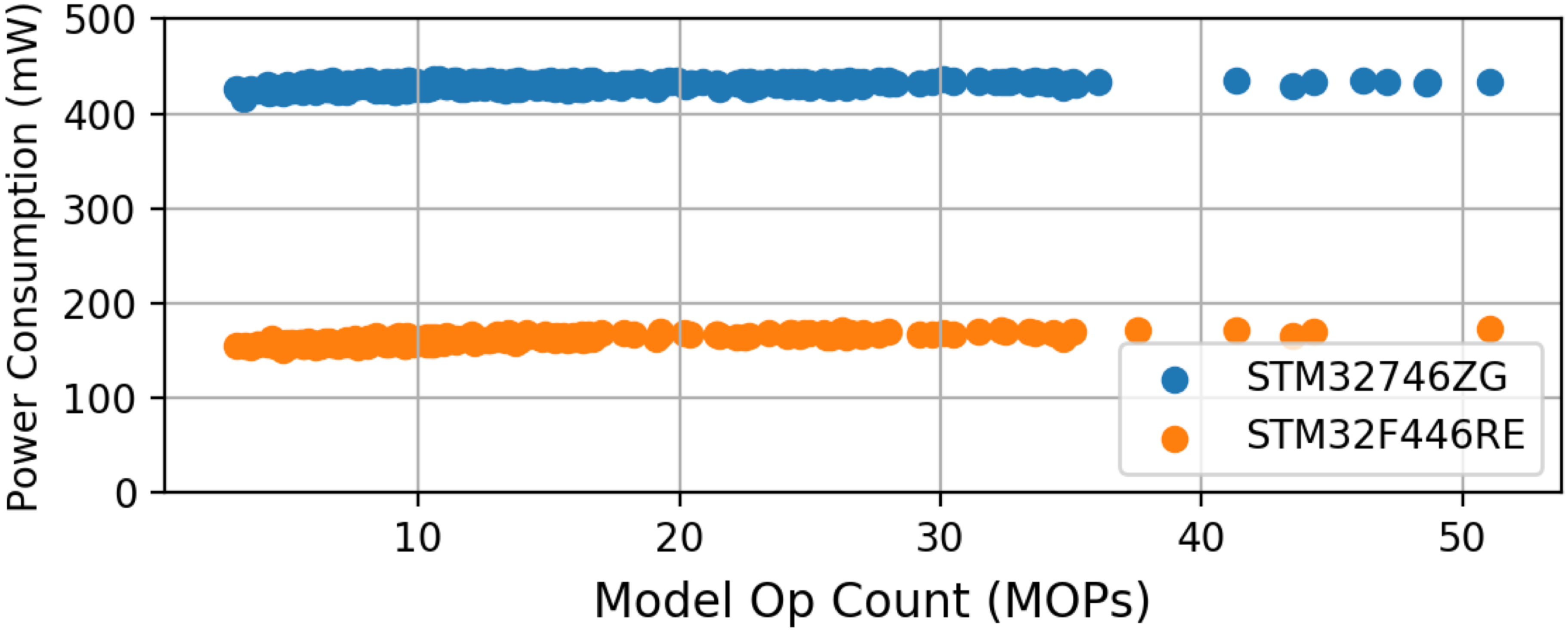
Direct Energy Benchmarking



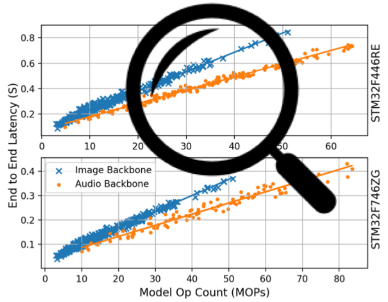


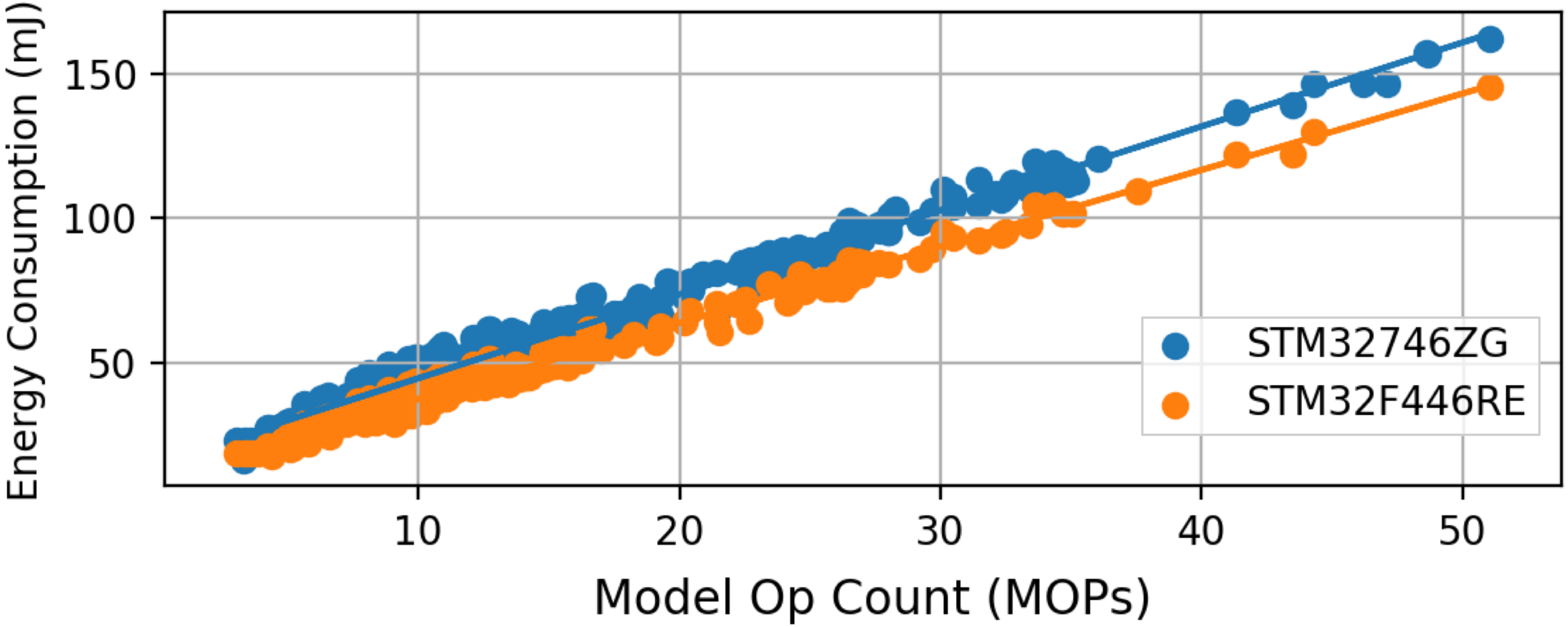
Model Energy



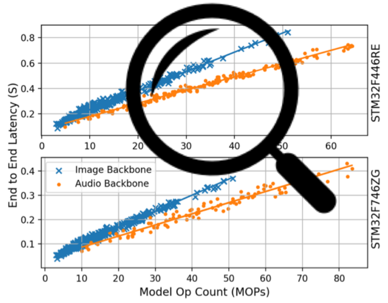


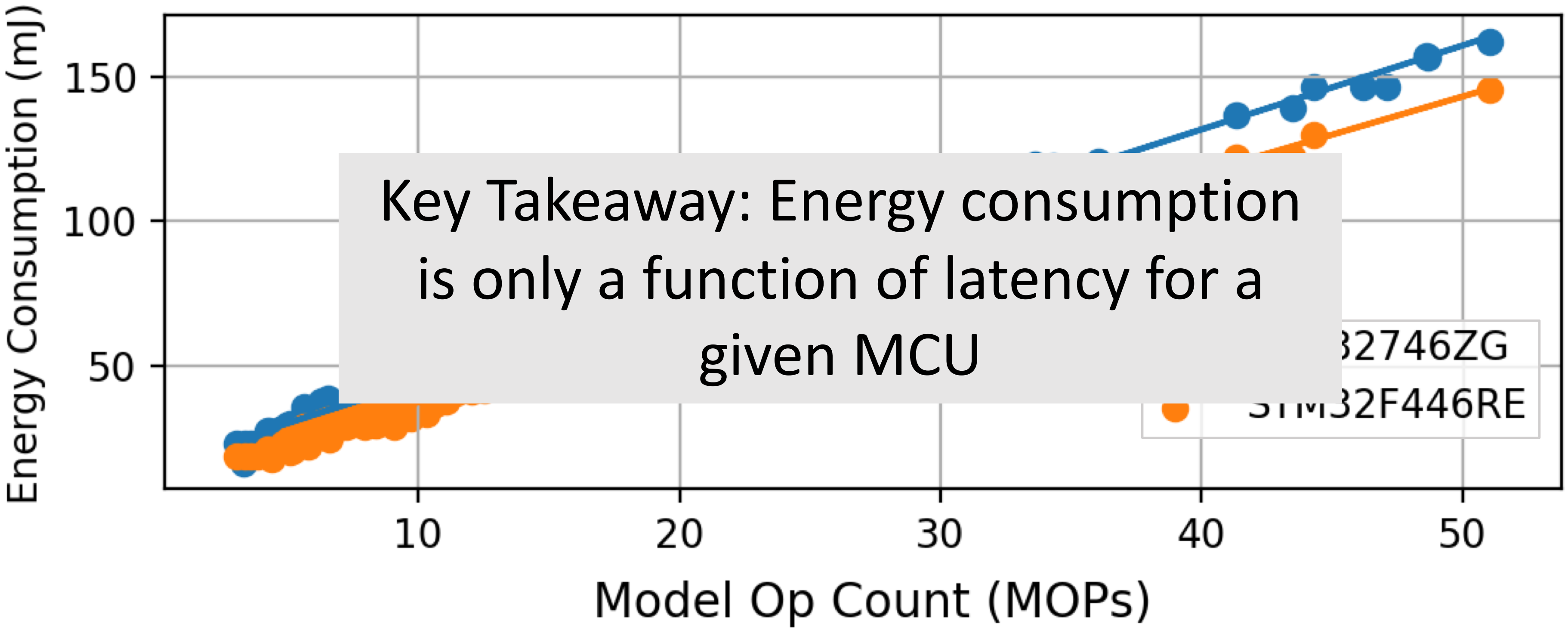
Model Energy





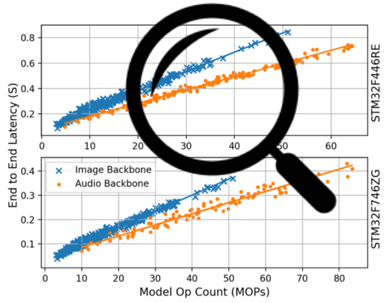
Model Energy

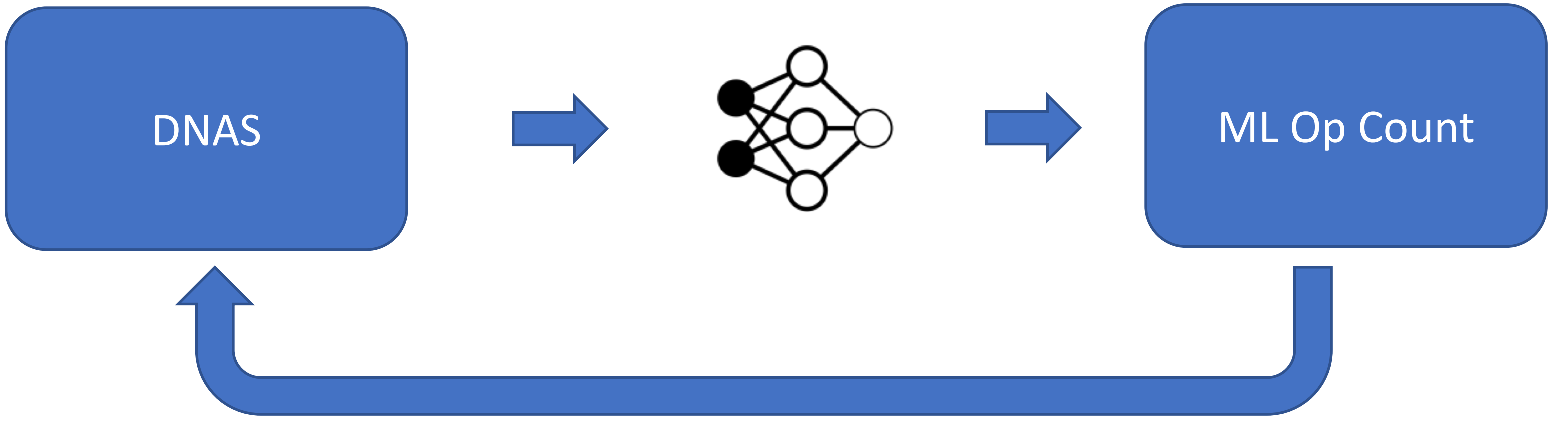




Key Takeaway: Energy consumption is only a function of latency for a given MCU

Latency & Energy Model





DNAS

ML Op Count

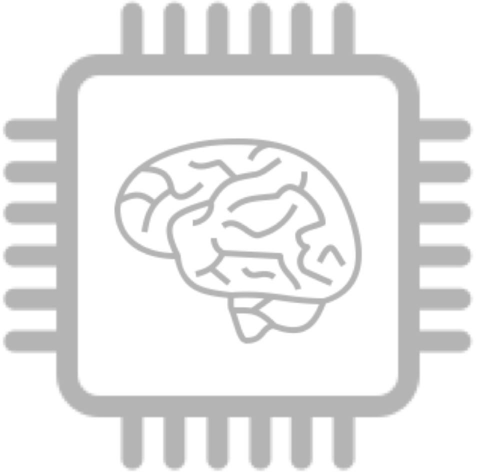
Latency & Energy
Estimate

TinyML Constraints

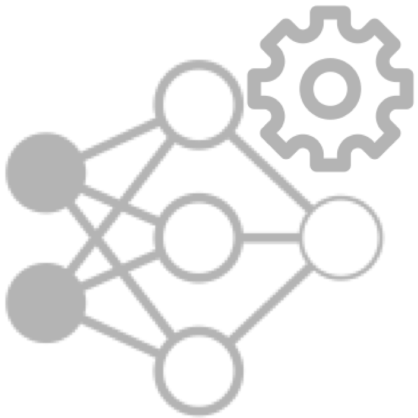
- ✓ SRAM
- ✓ Flash
- ✓ Latency
- ✓ Energy

Executive Summary

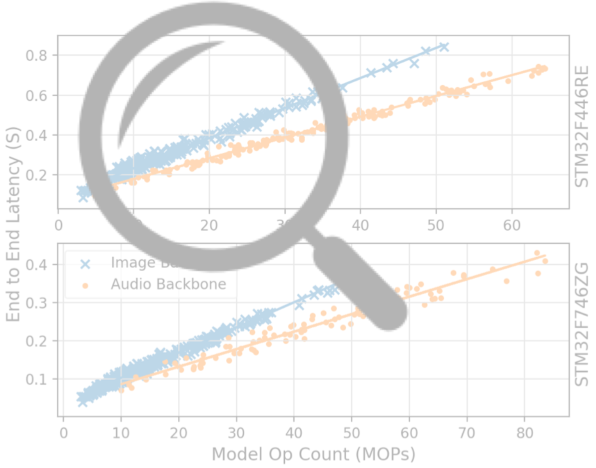
TinyML



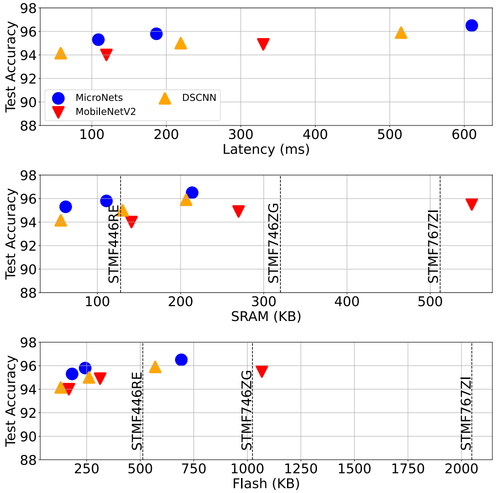
Differentiable Neural Architecture Search



Hardware Characterization

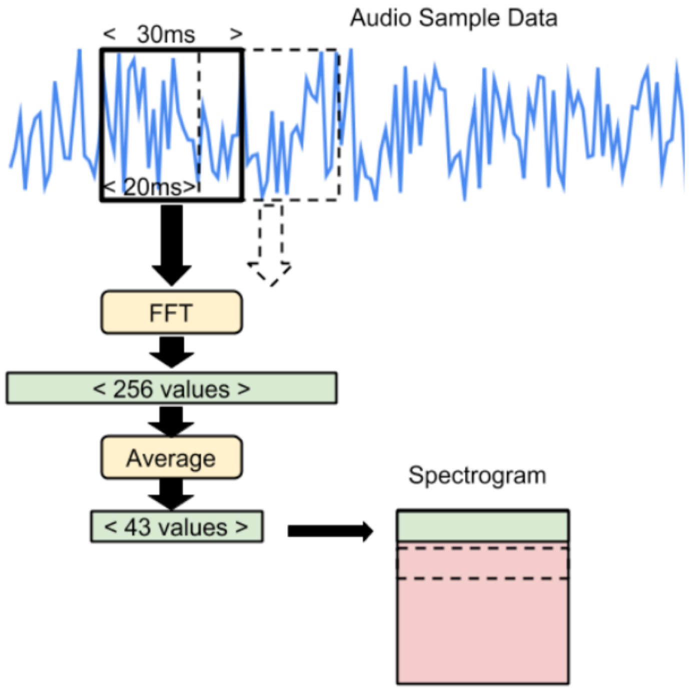


MicroNets



TinyMLPerf Use Cases

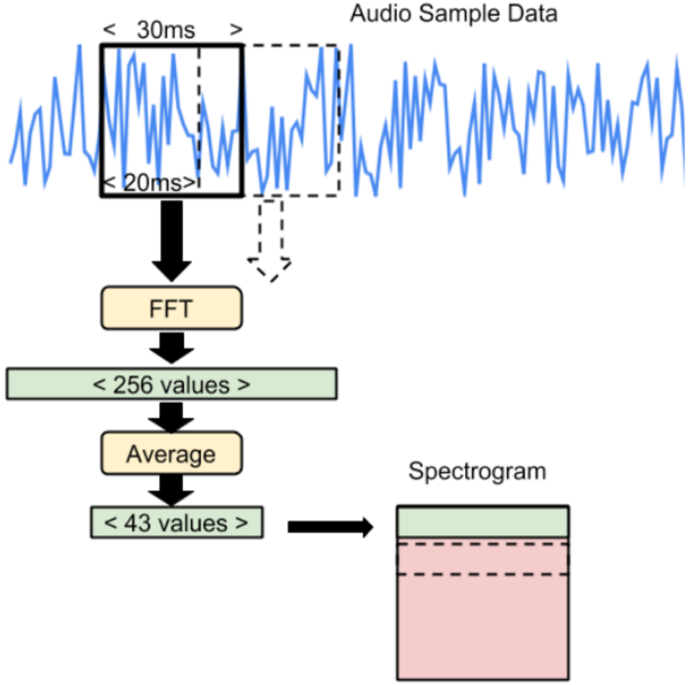
Keyword Spotting



Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

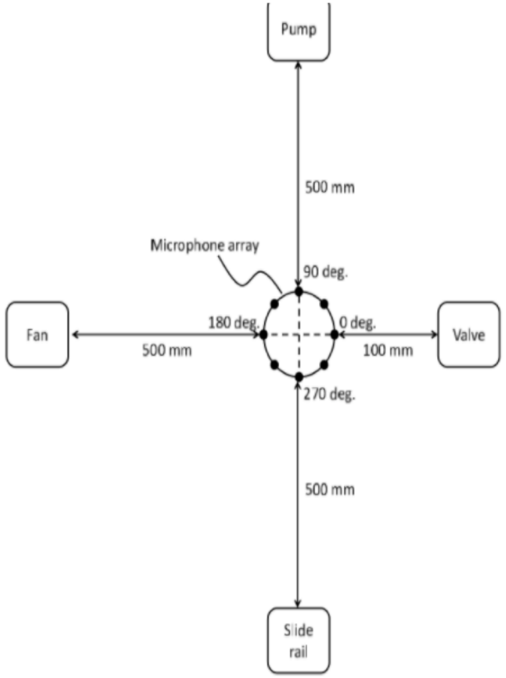
TinyMLPerf Use Cases

Keyword Spotting



Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

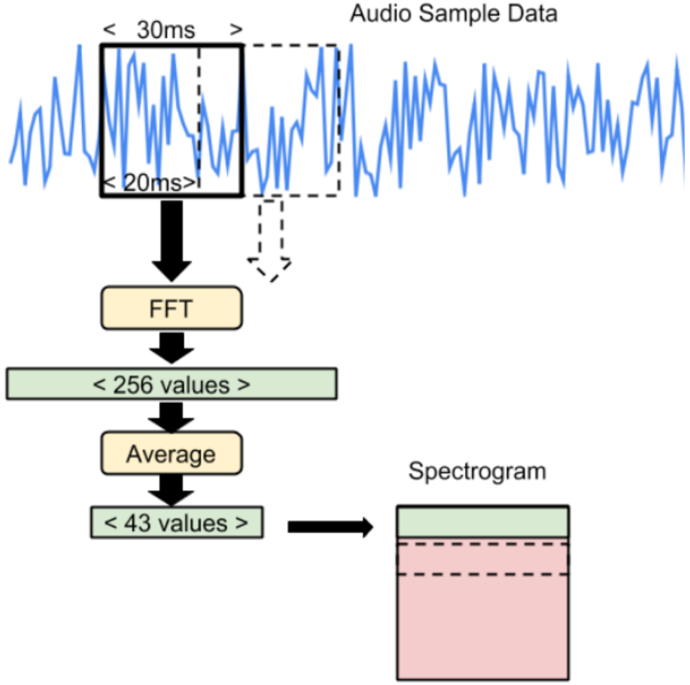
Anomaly Detection



Purohit, Harsh, et al. "MIMII dataset: Sound dataset fi malfunctioning industrial machine investigation and inspection." *arXiv preprint arXiv:1909.09347* (2019).

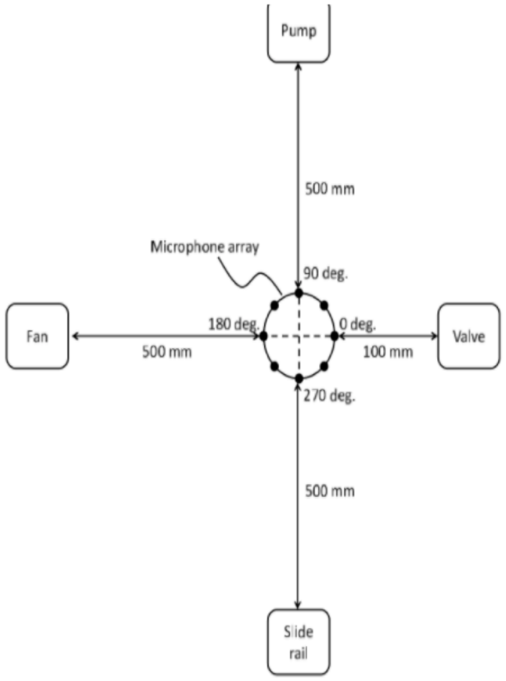
TinyMLPerf Use Cases

Keyword Spotting



Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

Anomaly Detection



Purohit, Harsh, et al. "MIMII dataset: Sound dataset f malfunctioning industrial machine investigation and inspection." *arXiv preprint arXiv:1909.09347* (2019).

Visual Wake Words

(a) 'Person'

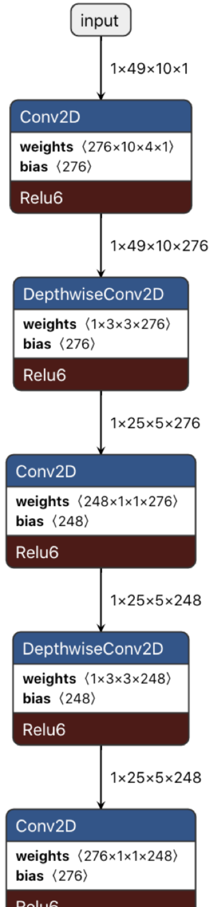
(b) 'Not-person'

Chowdhery, Aakanksha, et al. "Visual wake words dataset." *arXiv preprint arXiv:1906.05721* (2019).

Backbone Design

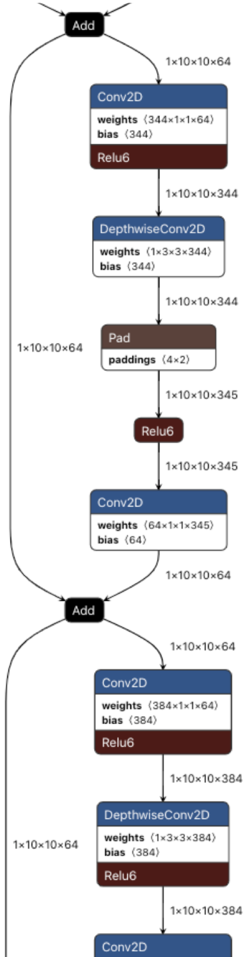
Keyword Spotting & Anomaly Detection

DSCNN-L[1]



Visual Wake Words

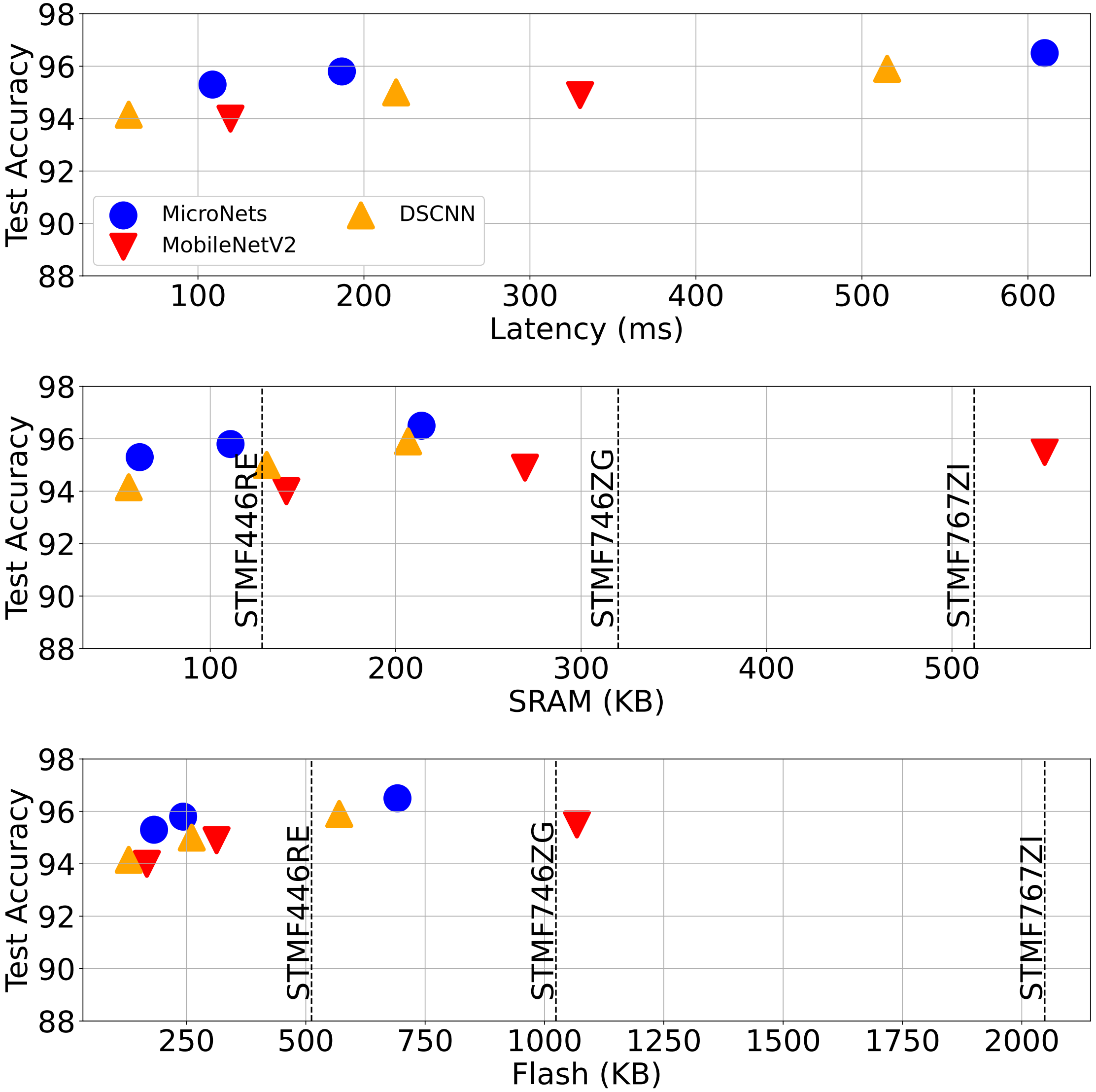
MobileNetV2[2]



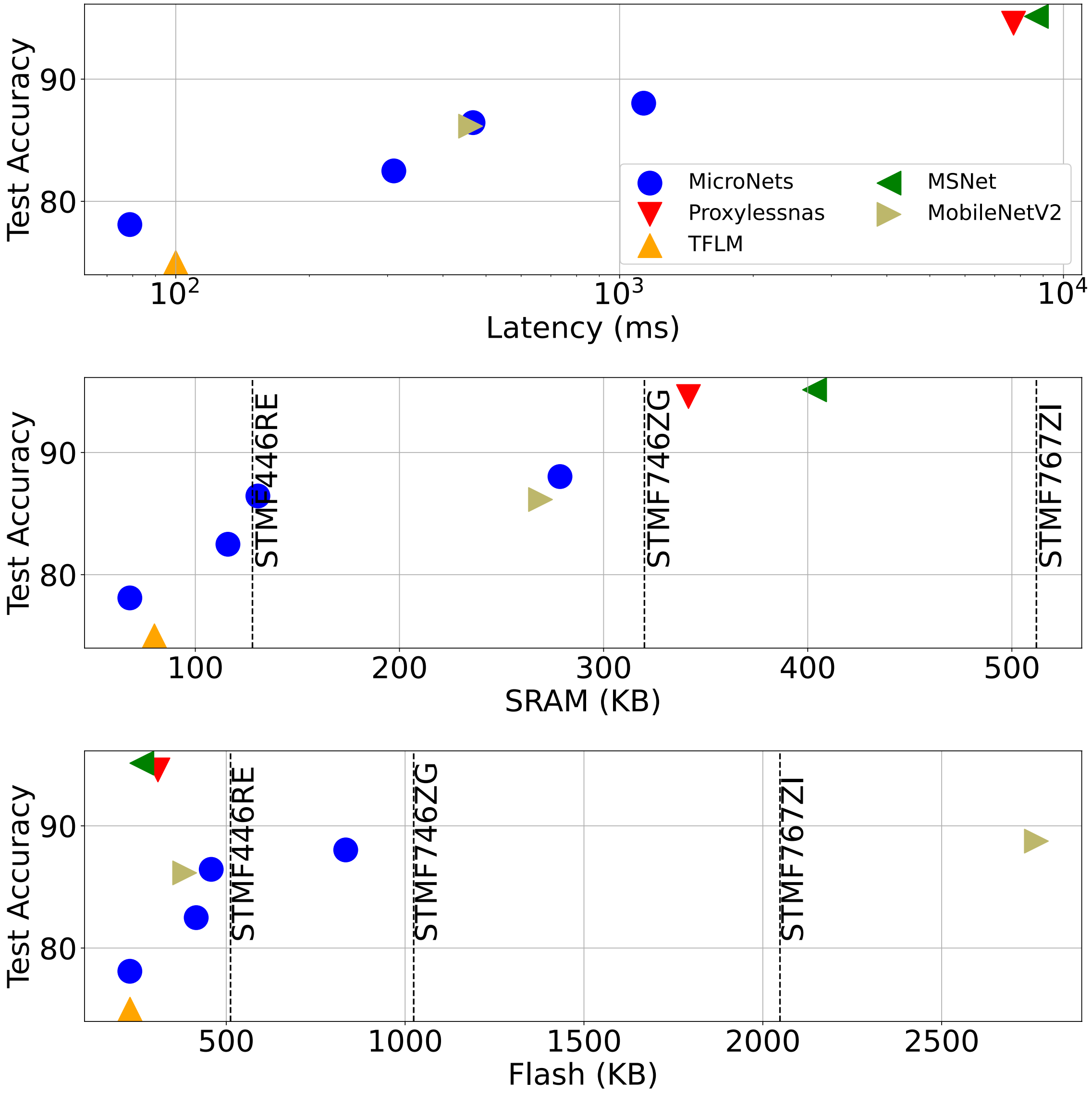
[1] Zhang, Yundong, et al. "Hello edge: Keyword spotting on microcontrollers." *arXiv preprint arXiv:1711.07128* (2017).

[2]Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Keyword Spotting

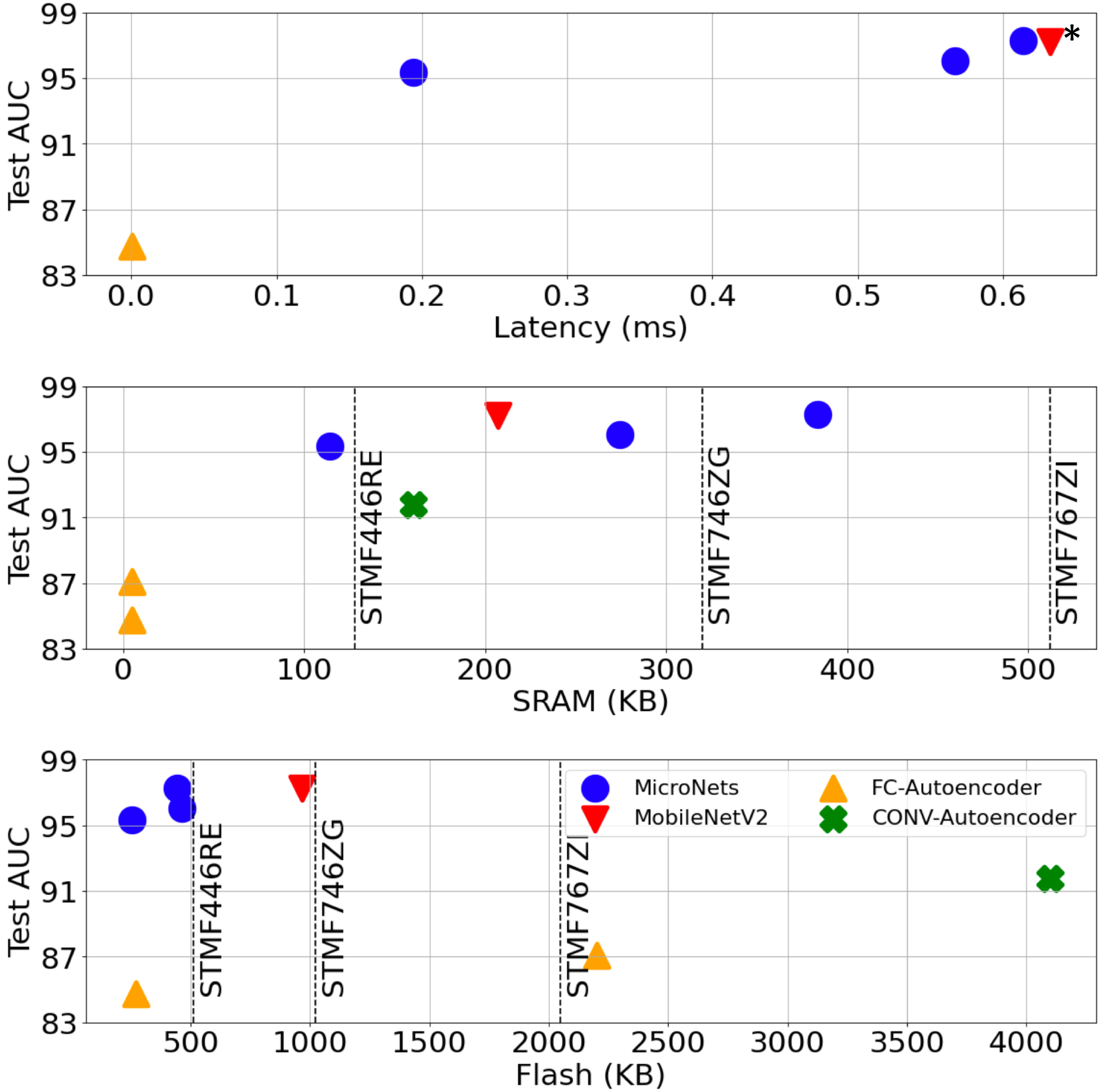


Visual Wake Words



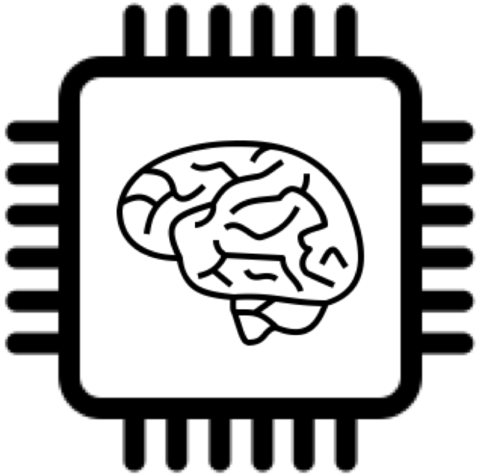
Anomaly Detection

MicroNets



Conclusion

TinyML



TinyML systems have **severe constraints** and require **highly tuned** model architectures

SRAM

Flash

Latency

Energy

Conclusion

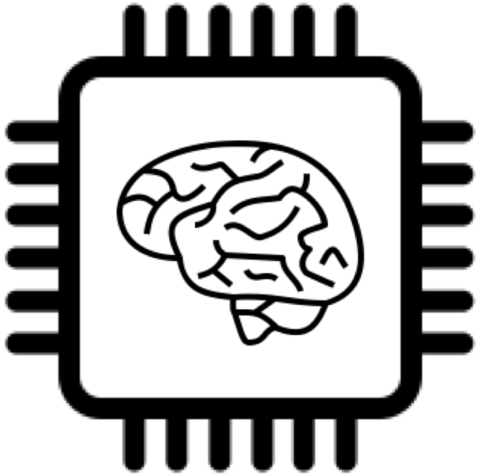
TinyML

Differentiable Neural Architecture Search

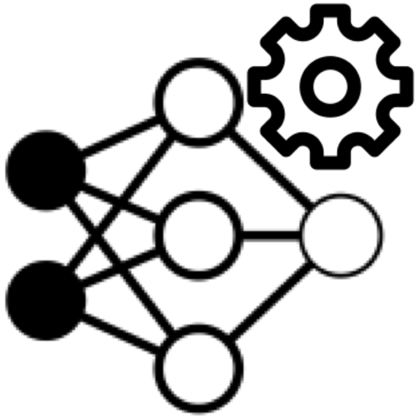
Differentiable Neural Architecture Search (DNAS) can **rapidly** find models that **meet the constraints** given **viable proxies**

Conclusion

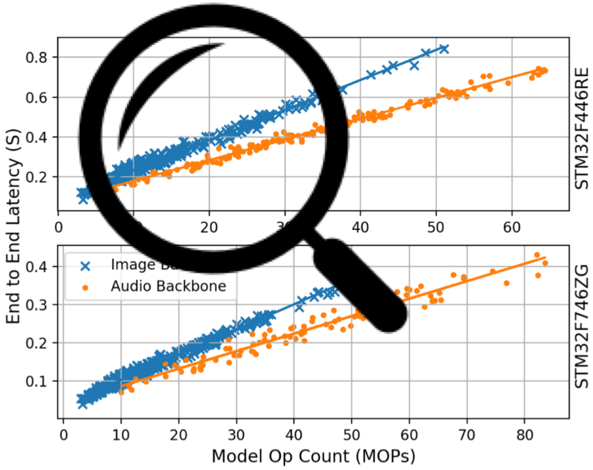
TinyML



Differentiable Neural Architecture Search



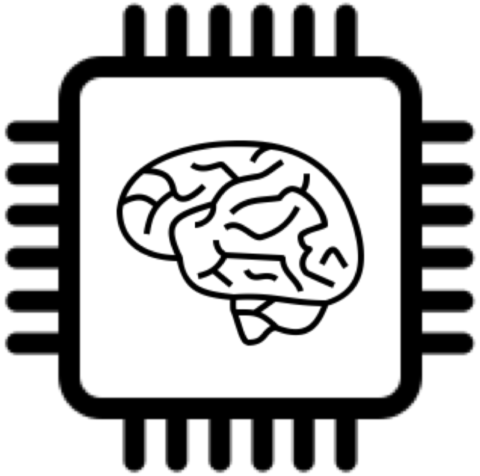
Hardware Characterization



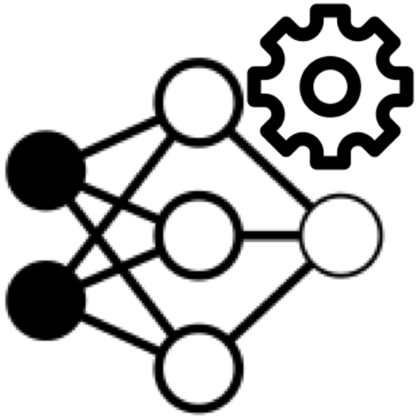
SRAM and Flash are easily calculated while **Op count is a viable proxy** latency and energy

Conclusion

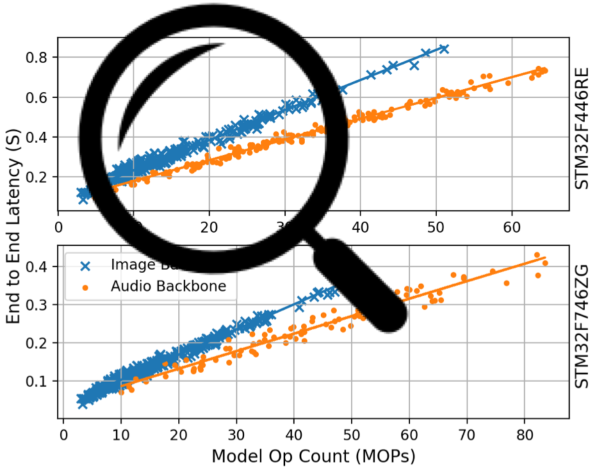
TinyML



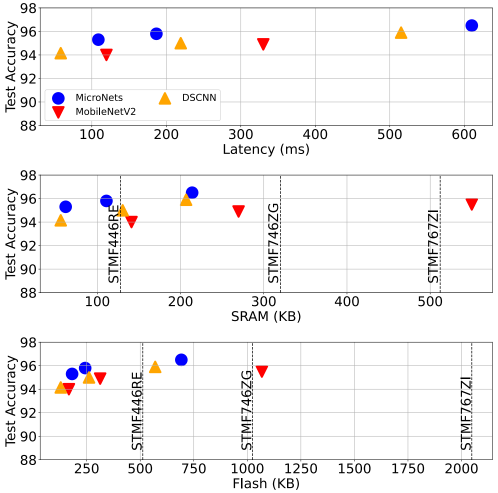
Differentiable Neural Architecture Search



Hardware Characterization



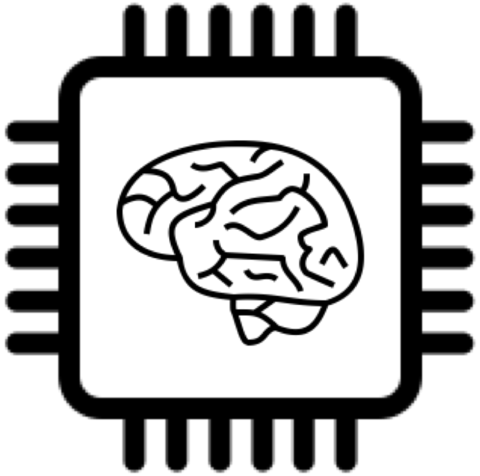
MicroNets



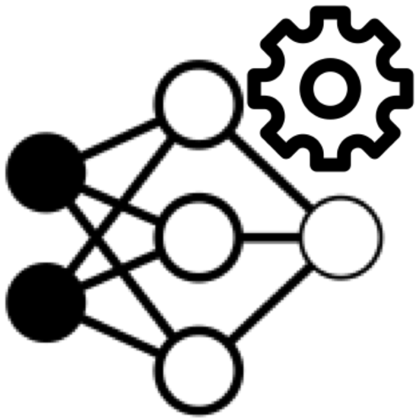
We achieve **state of the art performance** on three TinyML tasks

Conclusion

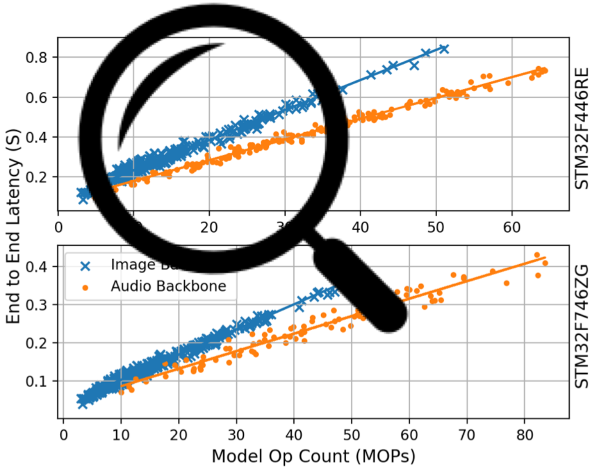
TinyML



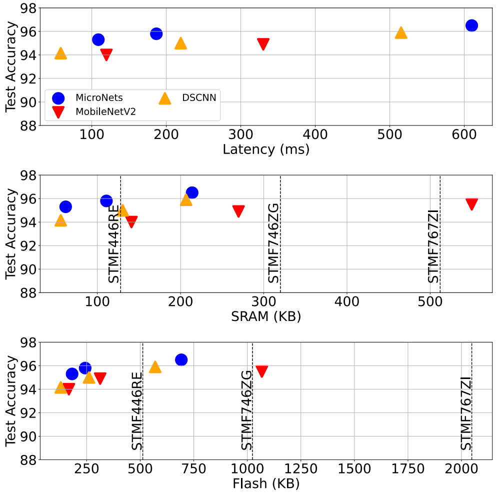
Differentiable Neural Architecture Search



Hardware Characterization



MicroNets



Models and Training Scripts are available:

github.com/ARM-software/ML-zoo