ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Thesis Defense

Ph.D. Program in Computer and Control Engineering (32.nd cycle)

Optimization Tools for ConvNets on the Edge

Valentino Peluso

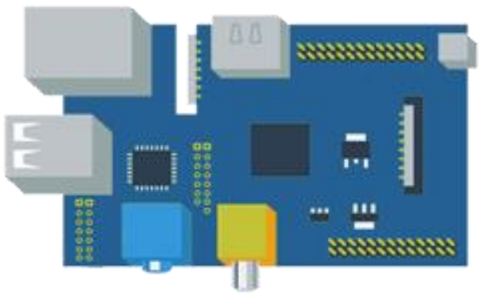
Supervisors

Prof. Enrico Macii, Supervisor

Prof. Andrea Calimera, Co-supervisor

Sensing and Sensemaking

Microphones

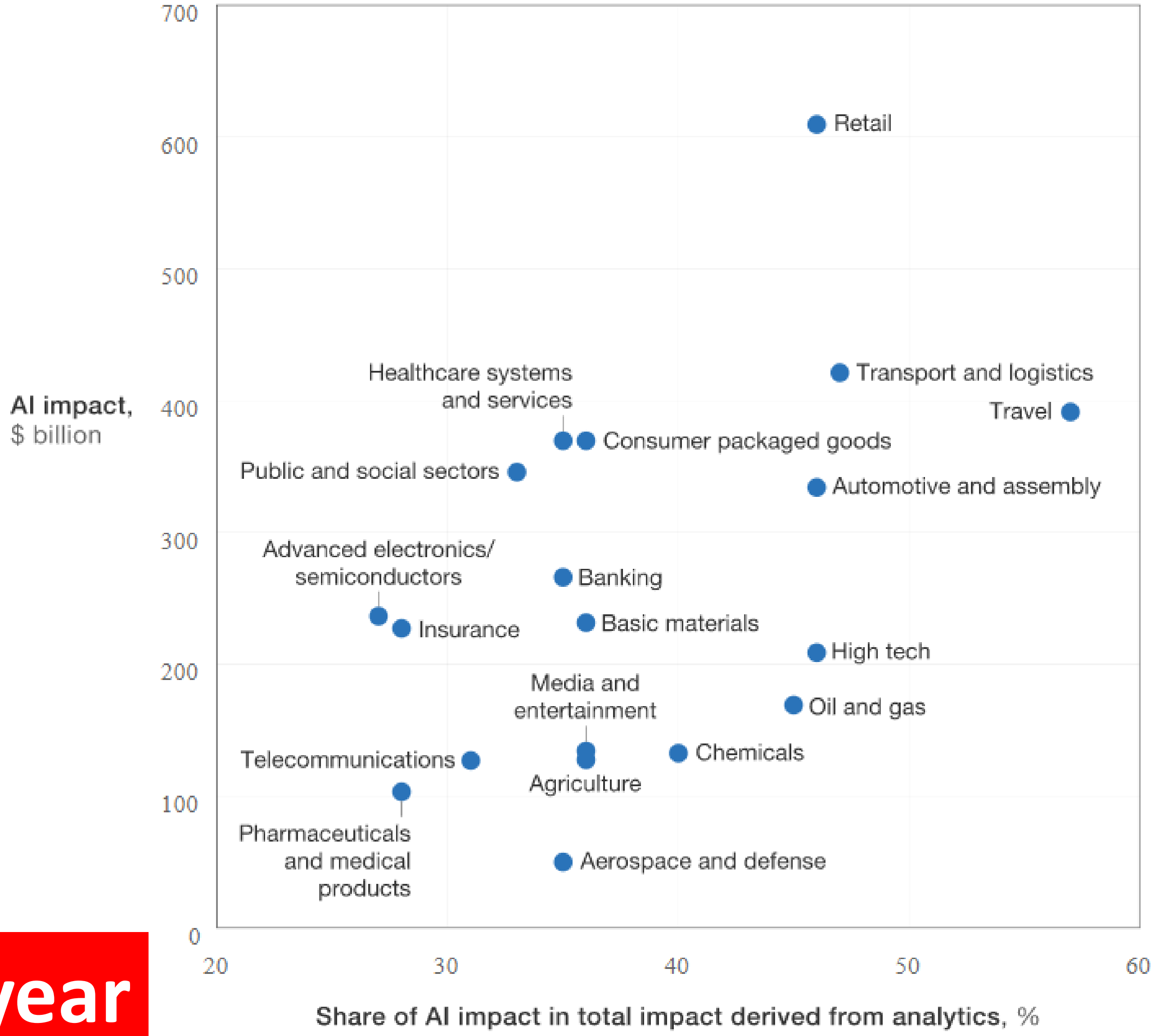


Imaging
- Low-Power Cameras
- Time-of-Flight Sensors
- Infrared Cameras

Motion MEMS
- Gyroscopes
- Accelerometers
- Magnetometers

IoT: Good in sensing, Poor in sensemaking

The value of AI



$5.8 trillion/year

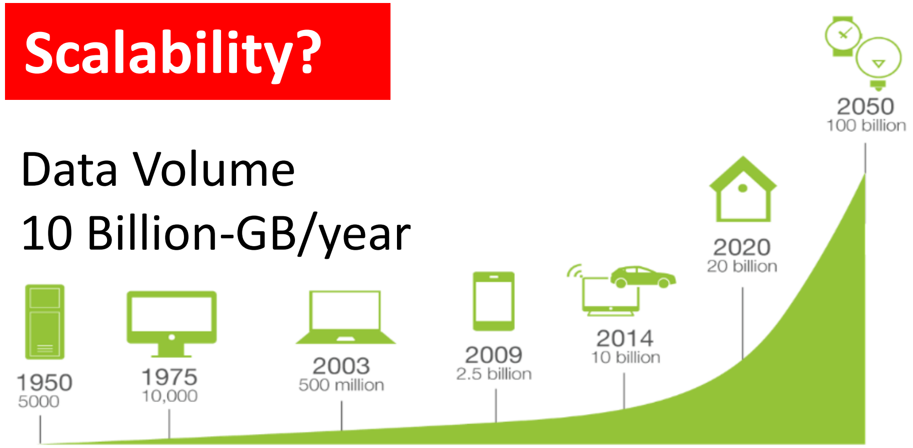
[Source] *McKinsey*

Edge-AI for the IoT

- **Sense-making:**
  - Present: in-the-cloud
  - Future: at-the-edge

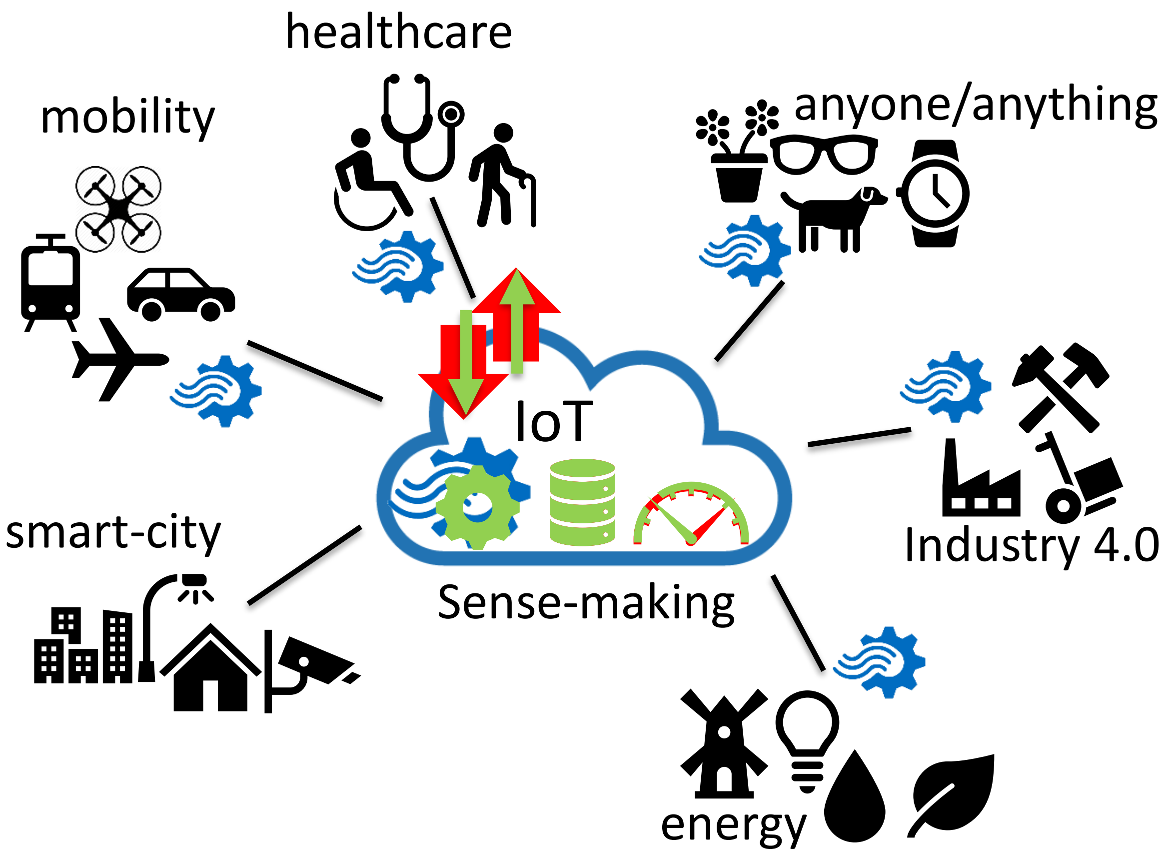
Scalability?

Data Volume
10 Billion-GB/year



Edge Computing

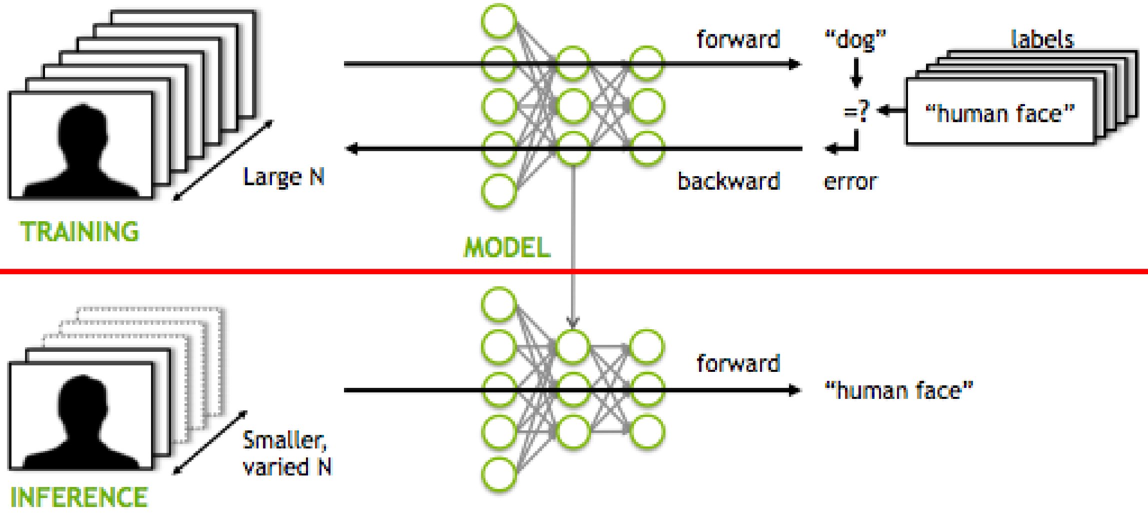
- ✓ Reduce response time
- ✓ Save transmission energy
- ✓ Improve privacy&security



Making sense of data

- **Convolution Neural Networks (ConvNets) achieved human-level accuracy**
  - End-to-end learning, i.e. automatic features selection

- **Designing ConvNets:**
  - Training: learn a proper set of parameters (W, b) using Back-Propagation
  - Inference: Feed-forward execution of the net



in-the-cloud ~ 500W

Goal:

Enable Edge Inference

at-the-edge ~ 10mW – 10W

Applications and Hardware

- Activity recognition
- Anomaly detection
- Keyword Spotting

- Image classification
- Face recognition
- Style Transfer

- Object Detection
- Segmentation
- Autonomous navigation

Microcontrollers (MCUs)

10—100mW
<1MB

✓ Low Cost
✓ Low Energy
× Low Memory
× Low Performance

Embedded CPUs

~3.5W
~2GB

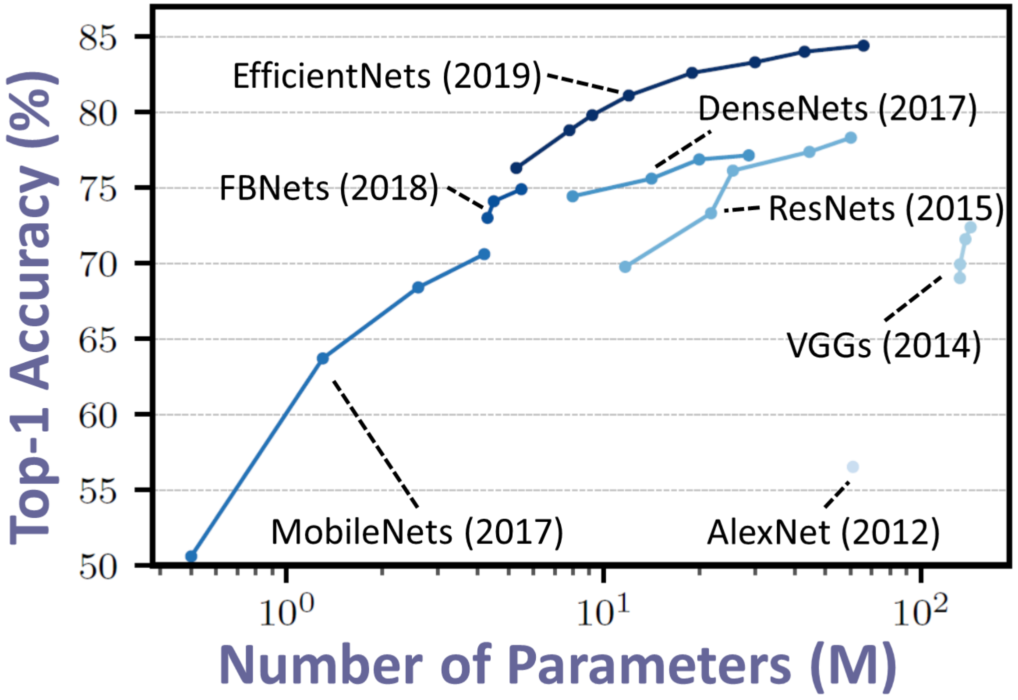
✓ Large diffusion
✓ Stable toolchains
× Low Thermal Design Power

ASICs/DSPs

~10W
~4GB

✓ Power/Performance stability
× High Cost
× Unstable toolchains

ConvNets are huge!

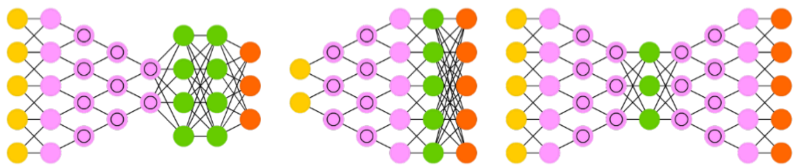


Neural Network Optimization

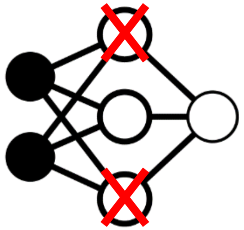
Enable Edge Inference

Existing tools for Neural Network Optimization

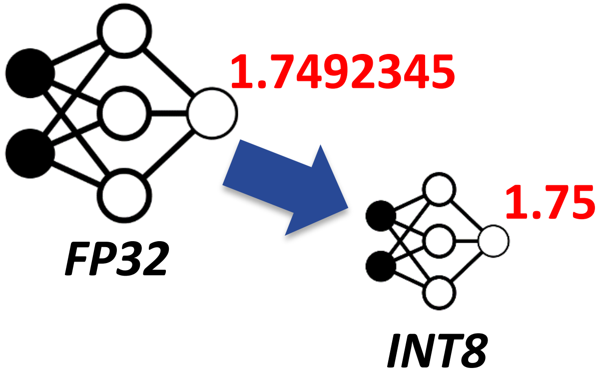
1) **Topology Optimization**
   - Manual or Automatic (NAS)



2) **Pruning**
   - Filter Pruning
   - Weight Pruning



3) **Quantization**
   - Floating-Point → Fixed-Point
   - Bit-width (1-, 2-, 3-, 4-, 8-bit)



→**Joint application to maximize savings**

Challenges

Multi-objective optimization

Accuracy Memory

Energy Power

2012

AlexNet: 1st place on ImageNet

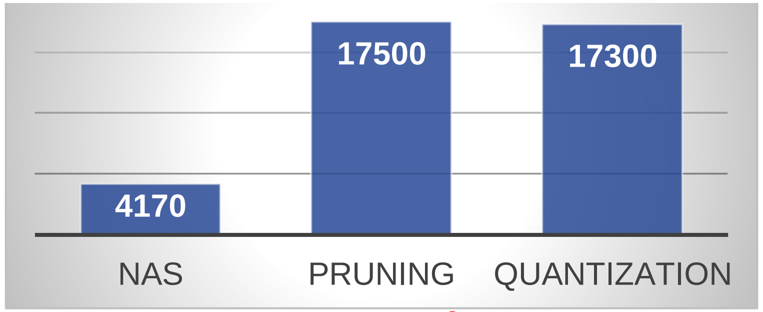
~5 years

2017

NEMO: Neuro-Evolution with Multi-Objective

# of research papers

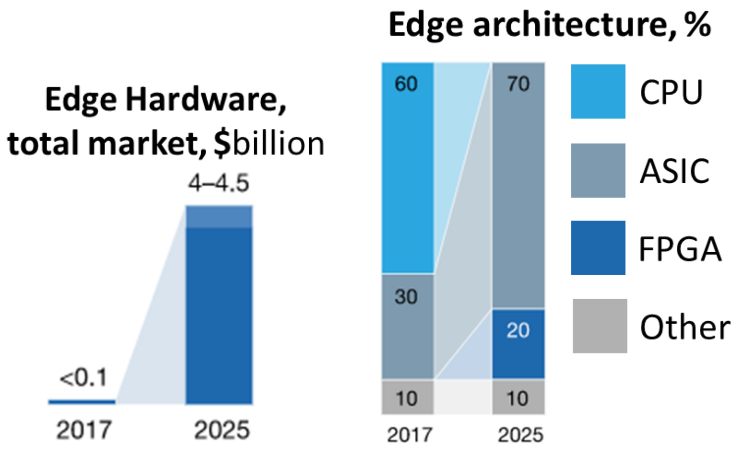
2012-today



>13 paper/day!

[Source] Google Scholar

Hardware diversity



[Source] *McKinsey*

Modular collection of optimization tools

Cross-layer optimizations

1. MEMORY OPTIMIZATION

- Prune and Quantize MCU
- Encoding-Aware Sparse Training MCU

2. ENERGY OPTIMIZATION

- On-line Precision Scaling ASIC
- Scalable-Effort ConvNets ASIC

3. POWER OPTIMIZATION

- Voltage-Scaled ConvNets CPU
- FINE-VH ASIC

Dynamic

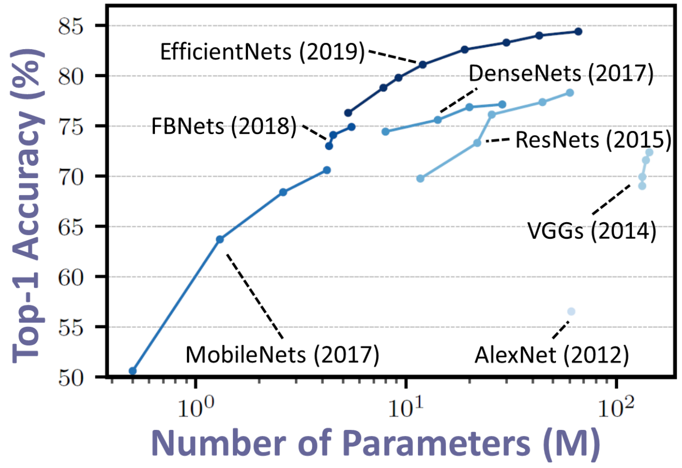
▷ SW

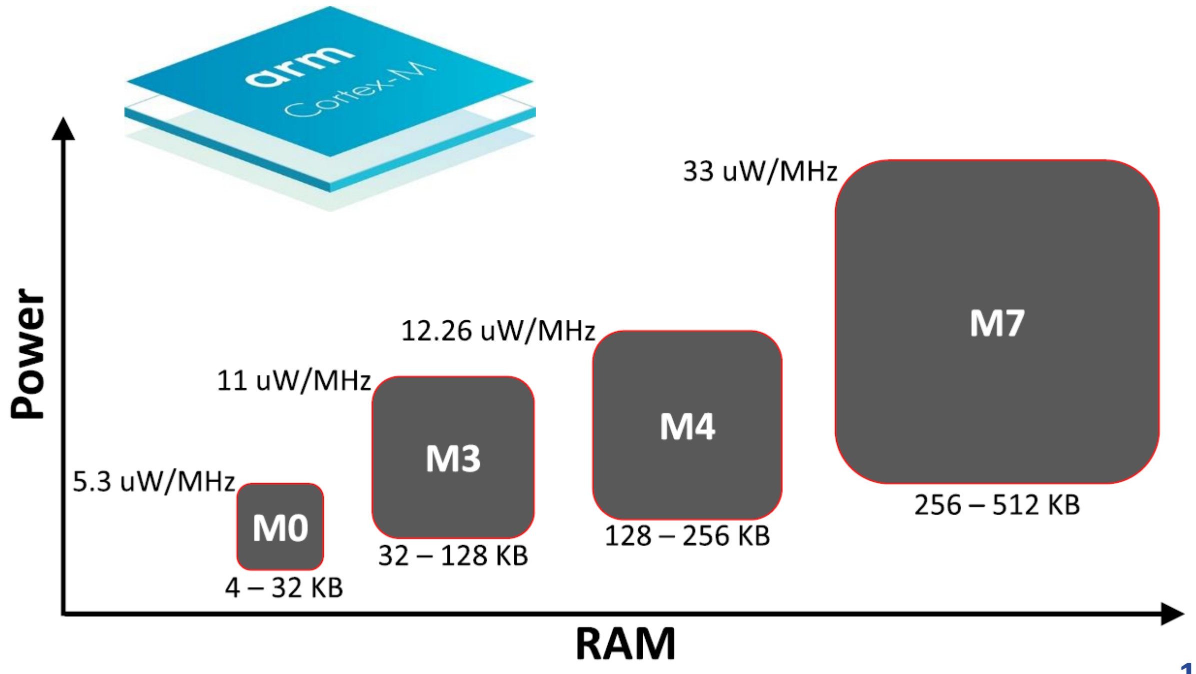
▷ HW

1. MEMORY OPTIMIZATION

Challenges

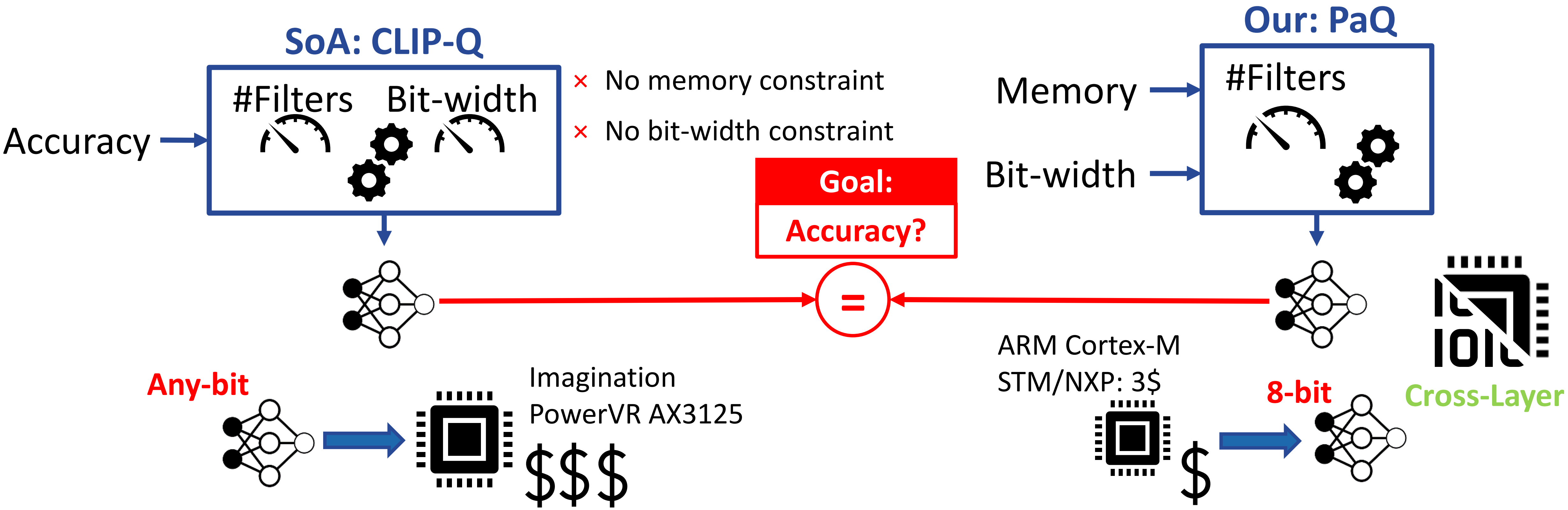
- **Goal:** Edge inference on ultra low-power MCUs.
- **Challenge:** Extreme memory constraints
  - ConvNet Parameters (Flash and RAM)
    - 500K to 100M of parameters
  - ConvNet intermediate results (RAM)
- **Limitation:**
  - × Limited ISA: minimum bit-width is 8-bit





Prune and Quantize (PaQ)

- **Motivation:** Identify the best combination of pruning and quantization for memory-constrained applications.



[Source] *Clip-q: Deep network compression learning by in-parallel pruning-quantization*, F. Tung et al., CVPR18

Prune and Quantize: Results

- **Parametric design-space exploration**
  - Bit-width: 16- down to 2-bit
    - 8-bit tested on-device
    - Other bit-widths via emulation
  - Memory (Mem.)

Image Classification on CIFAR-10

Mem. (KB)	Optimal Bit-width	Optimal Top-1	ARM Bit-width	ARM Loss
245	15	83.10	8	0.25
115	7	82.64	8	0.20
98	7	81.99	8	0.59
82	6	81.49	8	0.70
66	6	80.42	8	1.57
49	5	78.17	8	6.53
33	5	71.85	8	17.17

3x compression
<1% accuracy loss

For most solutions
8-bit has marginal loss

We need custom HW
at extreme constraints

Not supported by MCUs

Encoding-Aware Sparse Training

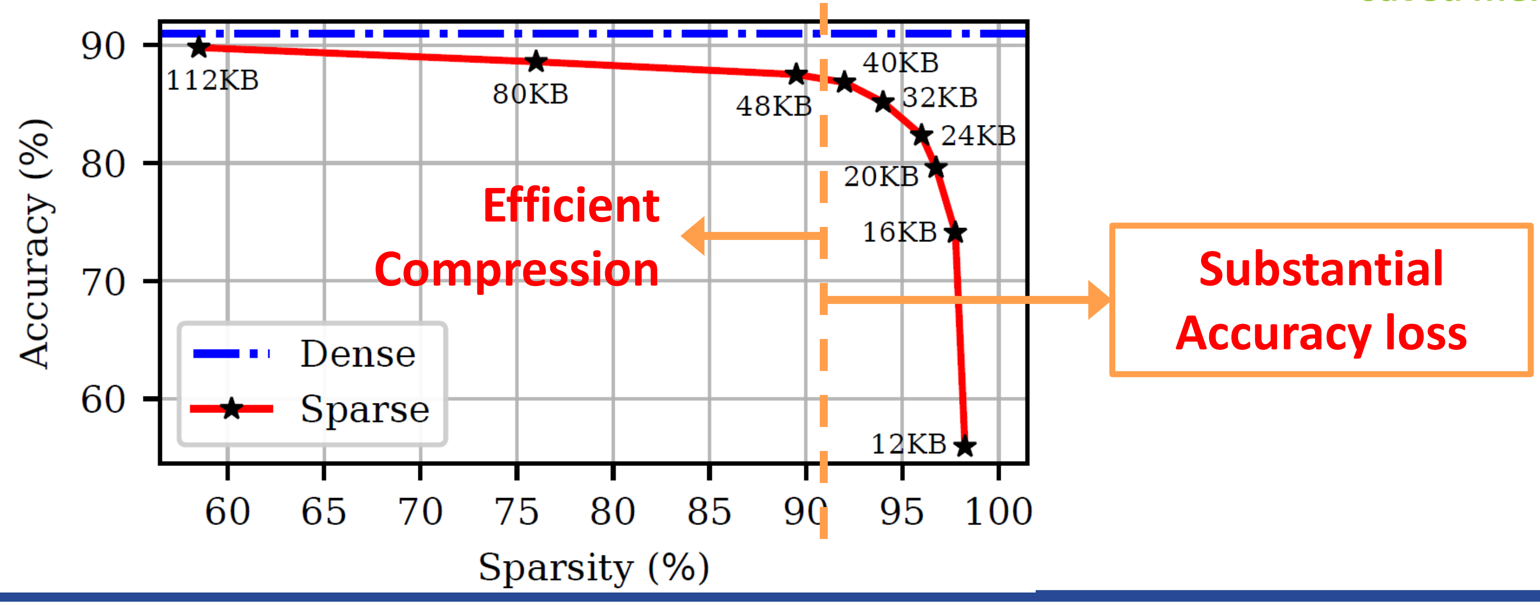
- **Goal:** Reduce size of ConvNet Parameters
- **SoA: Sparse Training + Weight Encoding**

Dense Weights	3	~~1~~	~~2~~	4	4	~~2~~	~~1~~	~~2~~	9	~~1~~	~~2~~	~~1~~	~~1~~	7	~~2~~	~~1~~	~~2~~	0	5	5	5	5	~~2~~
Sparse Weights	3	0	0	4	4	0	0	0	9	0	0	0	0	7	0	0	0	0	5	5	5	5	0

(3,1) (0,2) (4,2) (0,3) ...

Coding result	3	1	0	2	4	2	0	3	9	1	0	4	7	1	0	4	5	4	0	1

saved memory



Encoding-Aware Pruning

Weight Pruning

Sparse Weights	3	0	0	4	4	0	0	0	9	0	0	0	0	7	0	0	0	0	5	5	5	5	0

Size=23

#0s = 15

(3,1) (0,2) (4,2) (0,3) …

Coding result	3	1	0	2	4	2	0	3	9	1	0	4	7	1	0	4	5	4	0	1

Size=20

Group Pruning (Size=6)

Sparse Weights	3	0	0	0	0	0	0	2	3	0	0	0	0	0	0	7	9	5	6	6	2	2	2

#0s = 12

Coding result	3	1	0	6	2	1	3	1	0	6	7	1	9	1	5	1	6	2	2	3

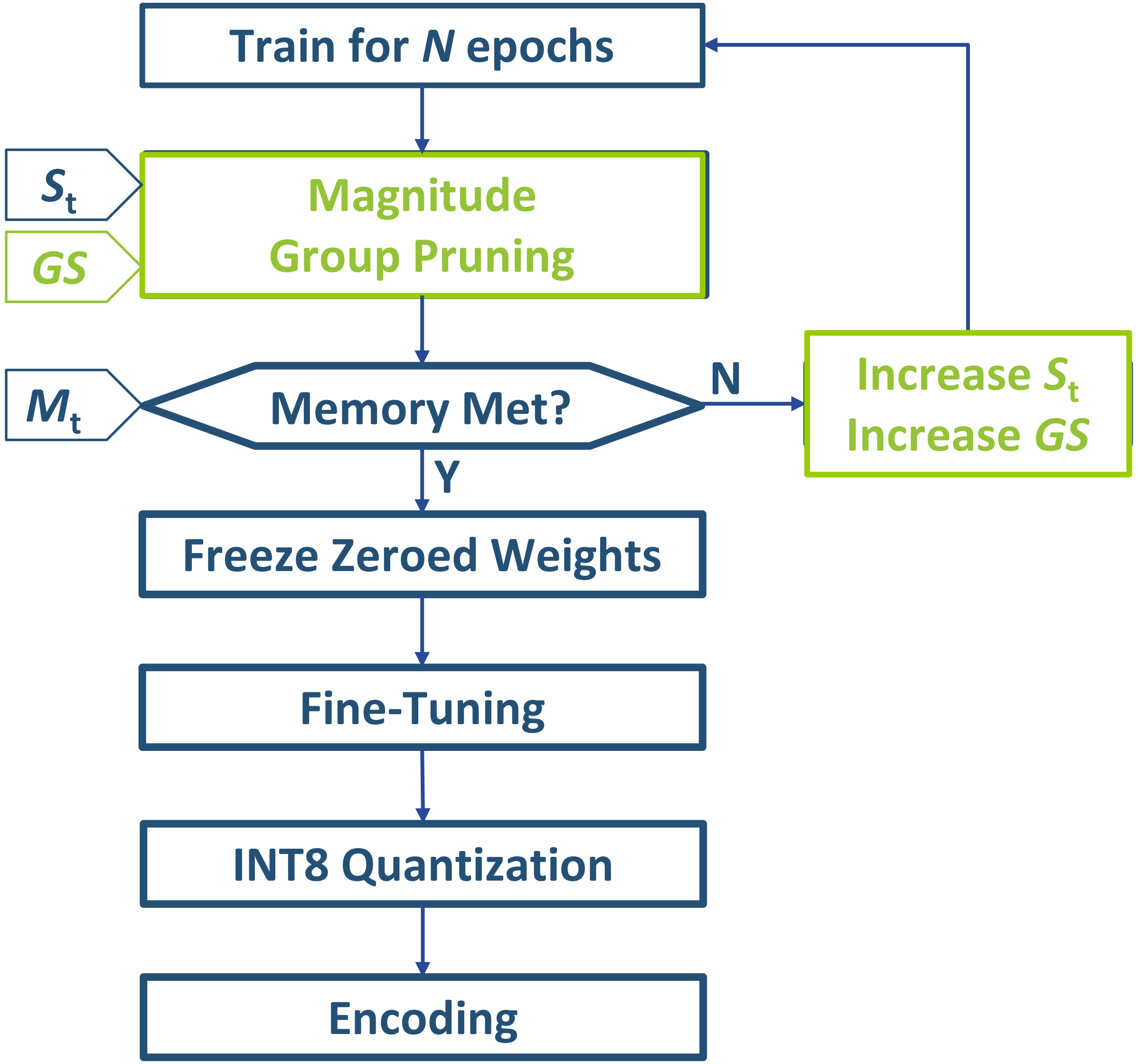
Size=20

more information!

#0s ↘

Size ==

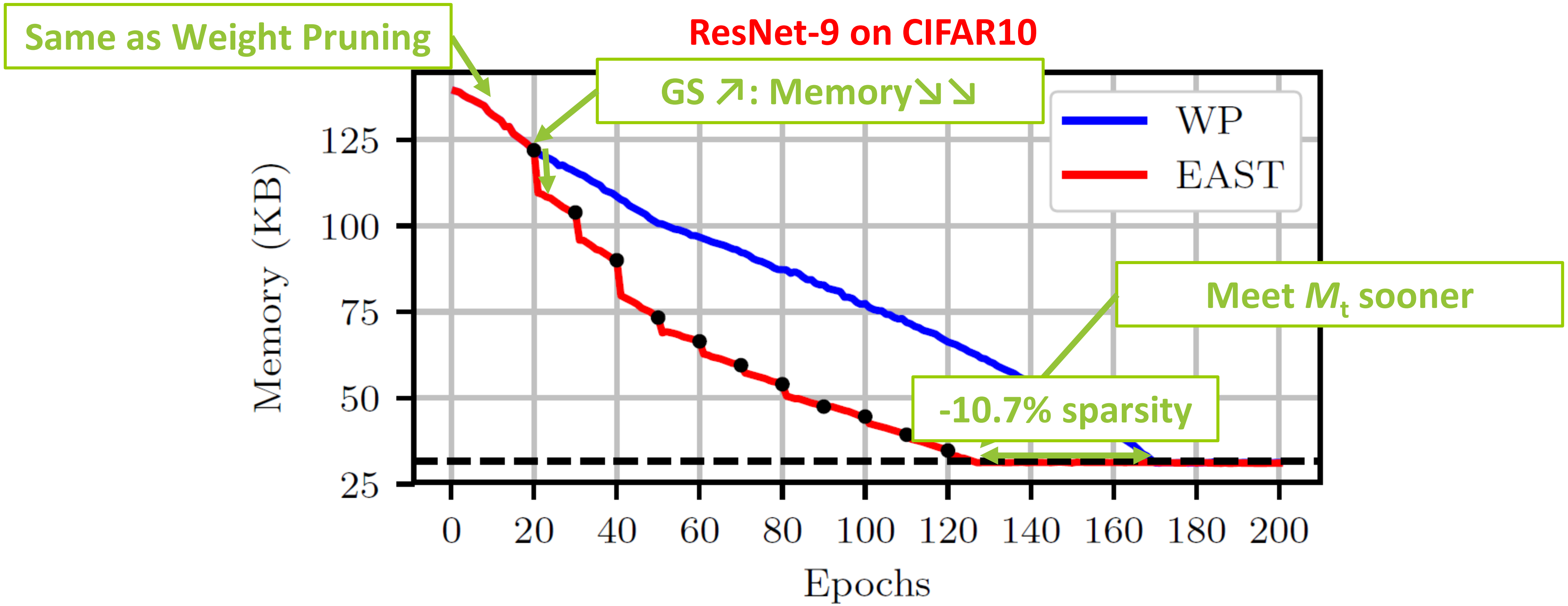
Sparse Training



Hyper-parameters	***Notation***	***Initial value***
Target Memory	M_t	12—112KB
Pruning Frequency	N	*1*
Target Sparsity	S_t	*30%*
Group Size	*GS*	*1*

[Source] *To prune, or not to prune: exploring the efficacy of pruning for model compression,* M. Zhu et al., arXiv 2017

Weight Pruning vs. EAST: Memory



Lower Sparsity = Higher Accuracy?

Yes! Lower sparsity = Higher accuracy

M_t: Target Memory
CR: Compression Ratio
S_x: Sparsity
A_x: Accuracy
ΔA: Accuracy difference

ResNet-9 on CIFAR10

M_t	CR	S_{WP}	S_{EAST}	A_{WP}	A_{EAST}	ΔA
112	5.0×	58.5%	49.5%	89.80%	89.46%	-0.34%
80	7.0×	76.0%	60.5%	88.67%	88.61%	-0.06%
48	11.6×	89.5%	74.8%	87.51%	87.44%	-0.07%
40	14.0×	92.0%	79.0%	86.80%	86.82%	**0.02**%
32	17.4×	94.0%	83.3%	85.30%	86.11%	**0.81**%
24	23.3×	96.0%	87.8%	82.33%	83.65%	**1.32**%
20	27.9×	96.8%	90.0%	79.63%	81.11%	**1.48**%
16	34.9×	97.5%	91.8%	74.16%	78.45%	**4.29**%
12	46.5×	98.3%	94.0%	55.59%	64.32%	**8.73**%

Similar accuracy for larger memory

Better accuracy for tighter memory

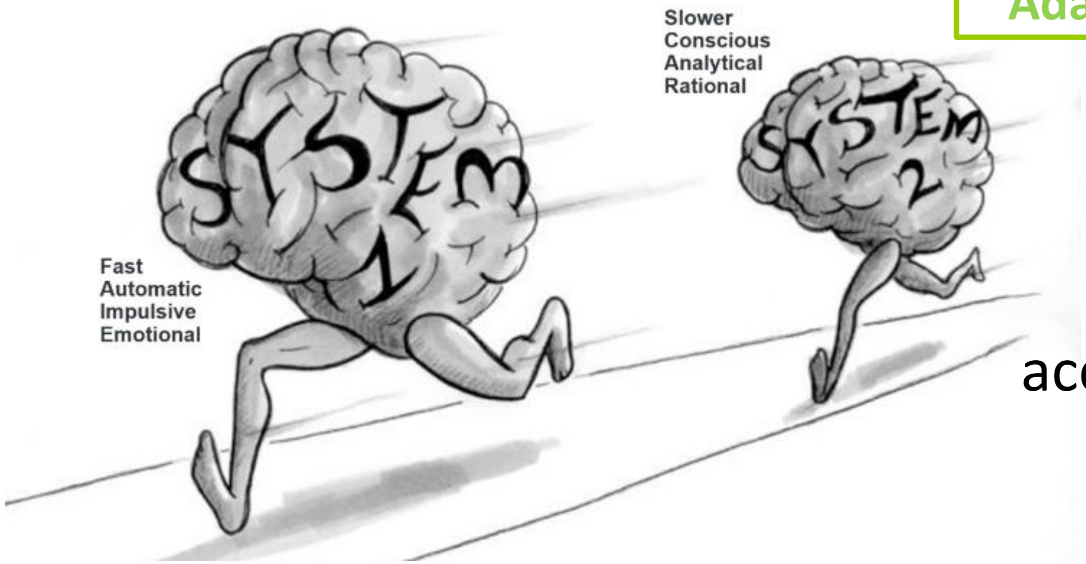
2. ENERGY OPTIMIZATION

Adaptive ConvNets

- **Motivation:** SoA ConvNets are designed and deployed as static graphs
- **Goal: Adaptive ConvNets**
- **Contributions:**
  1) Online Precision Scaling
  2) Scalable-Effort ConvNets

Goal:

Dynamic SW:
Adaptive ConvNets

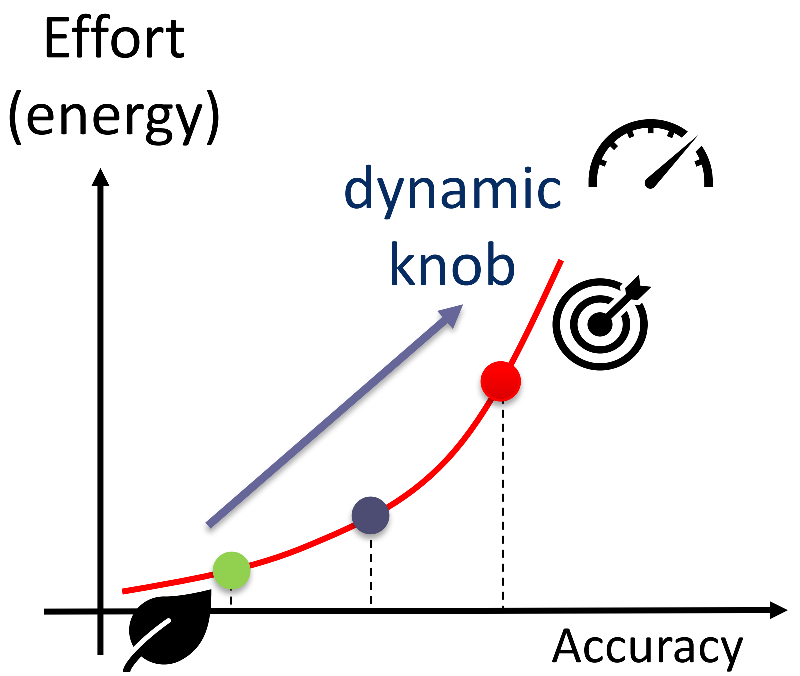


accuracy = more time

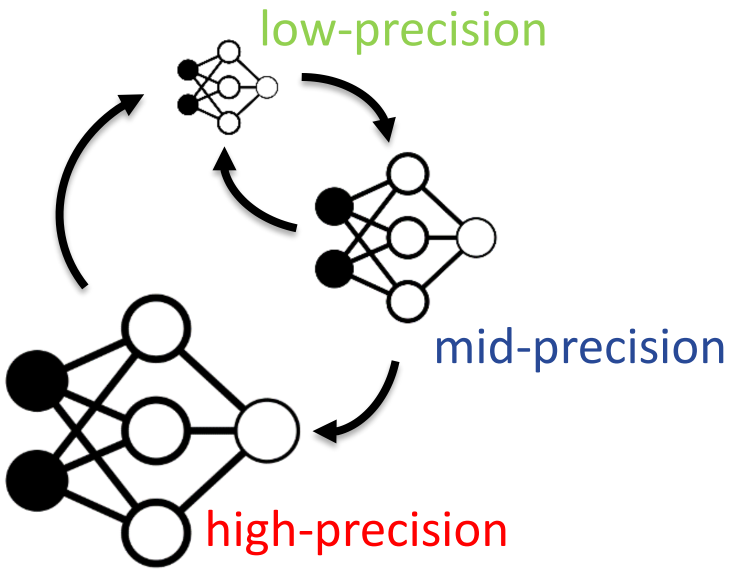
less time = less accuracy

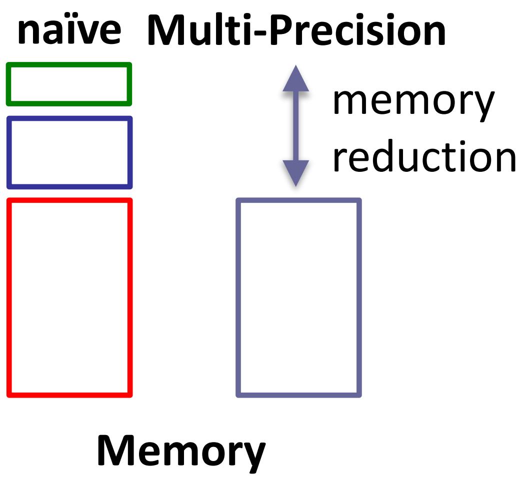
Enable Effort-Accuracy Scaling

- **Improve/Reduce accuracy → Reduce/Increase effort, hence energy**
  - Knob: dynamic precision scaling
  - Granularity: per-layer
  - Key Feature: single weight-set



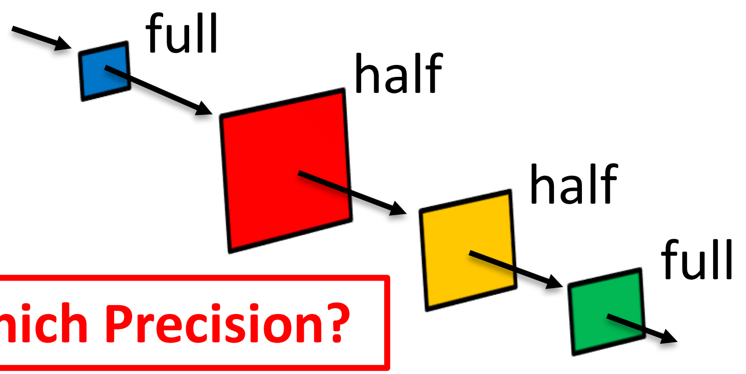
Software-Programmable
Multi-Precision





Per-layer precision assignment

- **Why per-layer?**
  - Define multiple operating points
  - Fine-grain control on effort-accuracy trade-off
- **Objective:**
  - Identify Pareto optimal configurations in the energy-accuracy space



Which Precision?

2 precision options:

- **full (16-bit)**
- **half (8-bit)**

ImageNet Classification

ConvNet	FP32 Acc.	#Params	#Cycles	#Layers
ResNet18	69.13%	11.68M	29.35M	21
ResNet34	72.69%	21.78M	57.28M	37
ResNet50	74.10%	25.50M	74.40	54
SqueezeNet	56.36%	1.23M	6.45M	32
MobileNet v2	69.98%	3.47M	12.13M	54

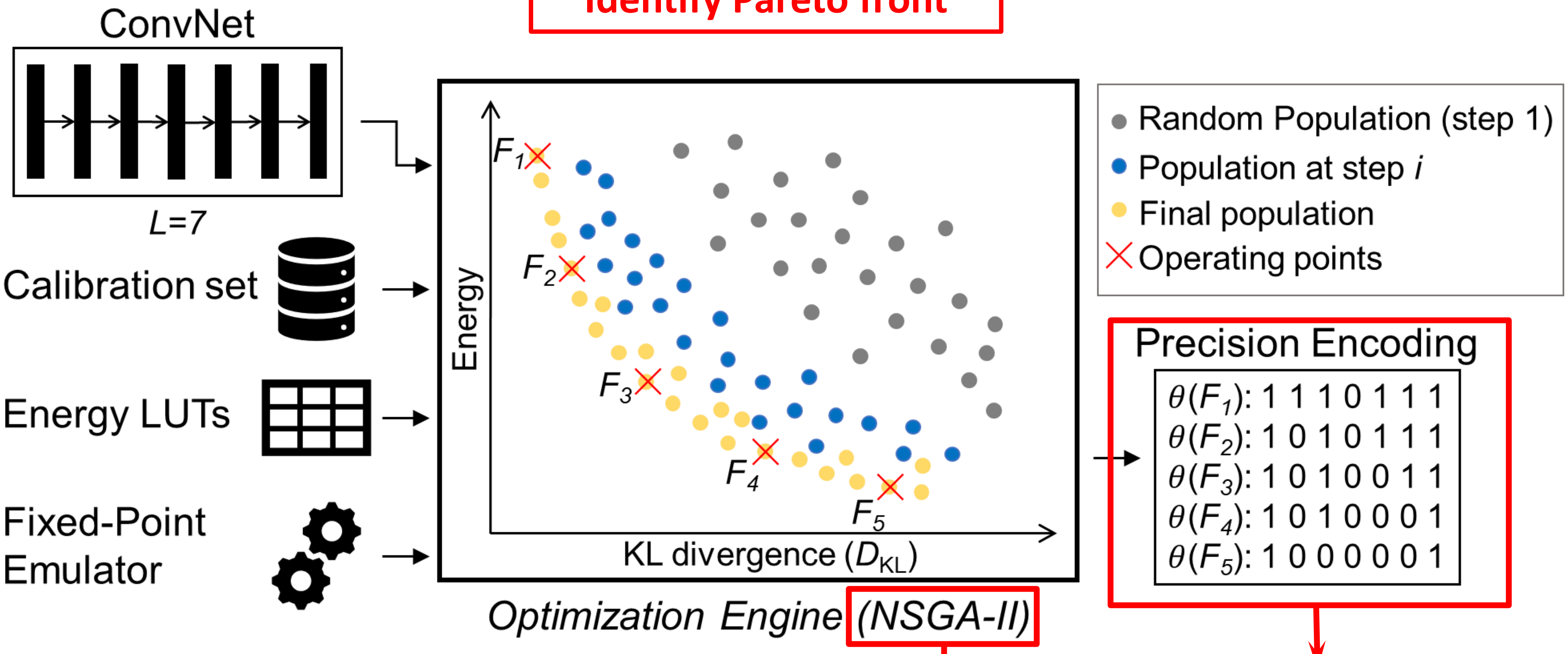
$\mathbf{2^{21} = 2.1 \times 10^6}$

We need heuristics!

$\mathbf{2^{54} = 1.8 \times 10^{16}}$

Online Precision Scaling: Design

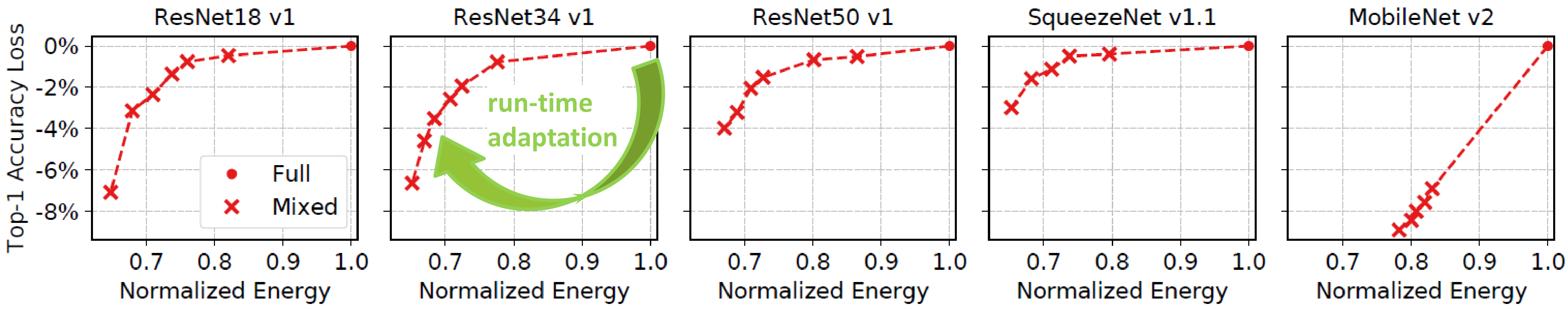
Goal:
Identify Pareto front



Multi-Objective Optimization

Output

Online Precision Scaling: Results

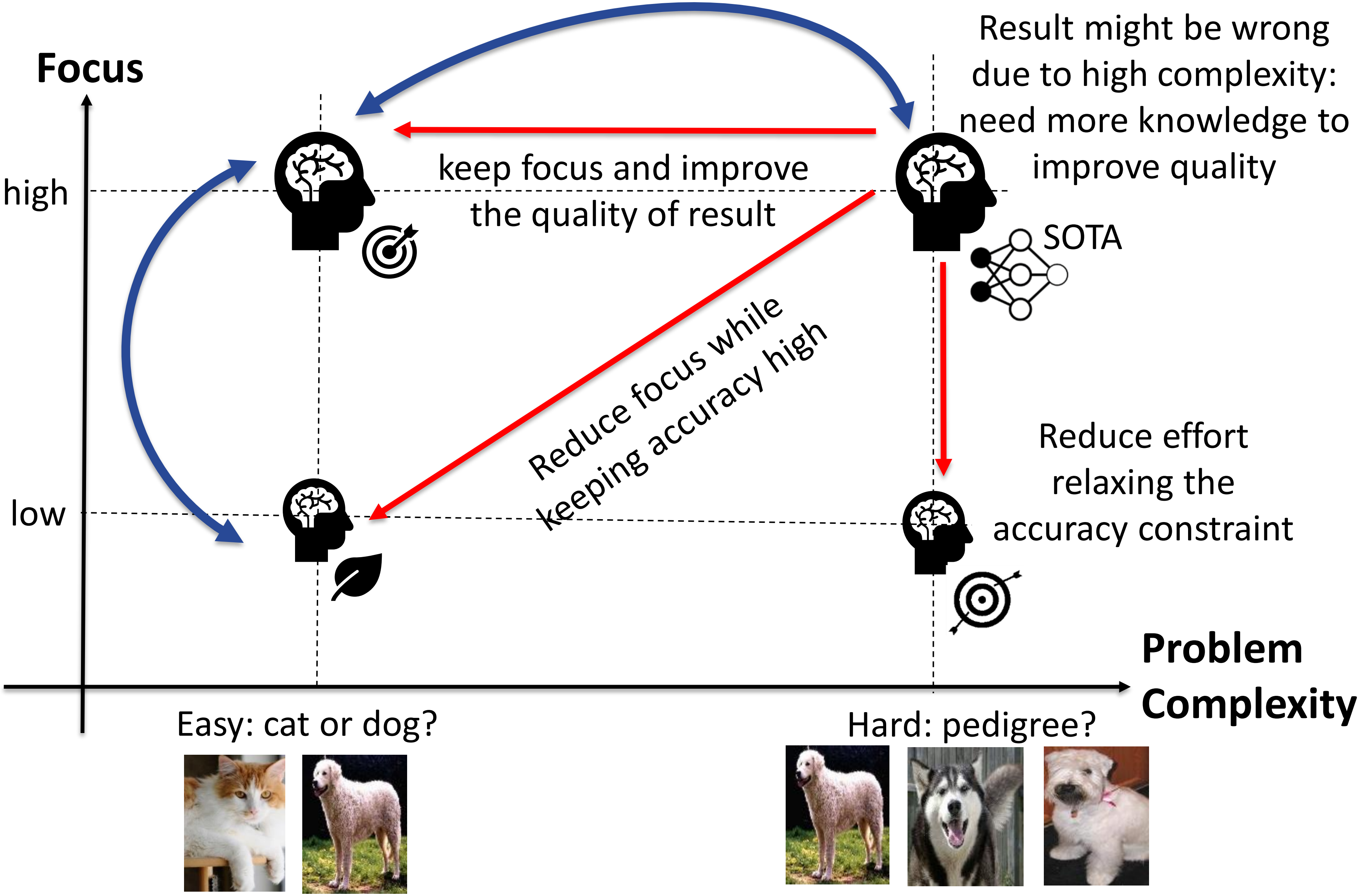


Benchmark	#Points	Δ Top-1	Savings	Ex. Time
ResNet18 v1	6	2.5%	27.4%	8 min 32 s
ResNet34 v1	6	3.3%	29.8%	12 min 36 s
ResNet50 v1	6	2.0%	25.6%	25 min 17 s
SqueezeNet v1.1	5	1.3%	28.4%	6 min 19 s
MobileNet v2	5	8.0%	19.2%	14 min 33 s

Depthwise Convolution need high-precision

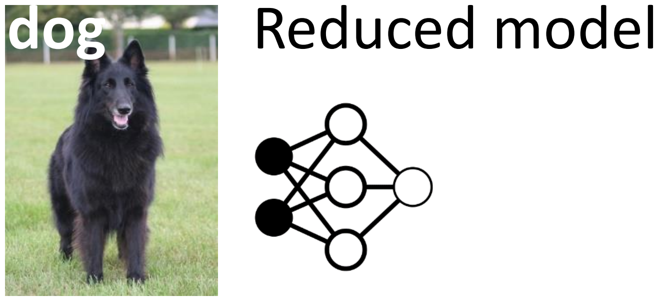
ConvNet↗ Ex. Time ↗

Beyond Energy-Accuracy Scaling: Brain Teaching

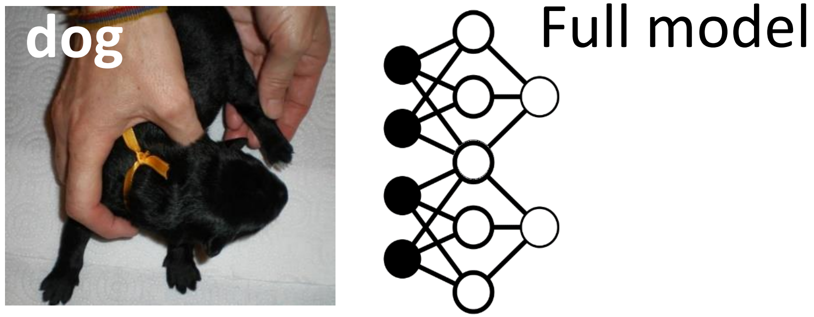


Static vs Dynamic

- **SoA: Hierarchical ConvNets**
  - Tune the computational effort depending on the complexity of the input
    - E.g. drop some filter/layer at run-time

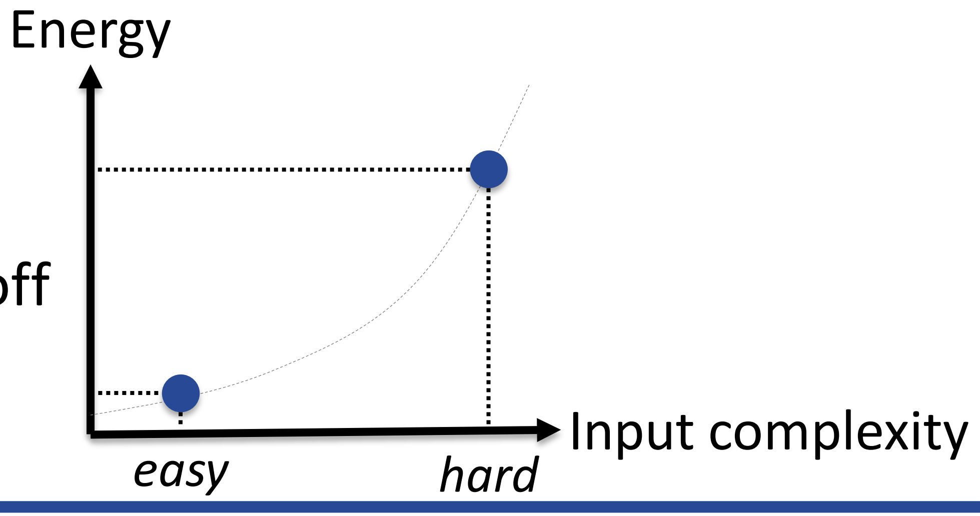


features clearly visible



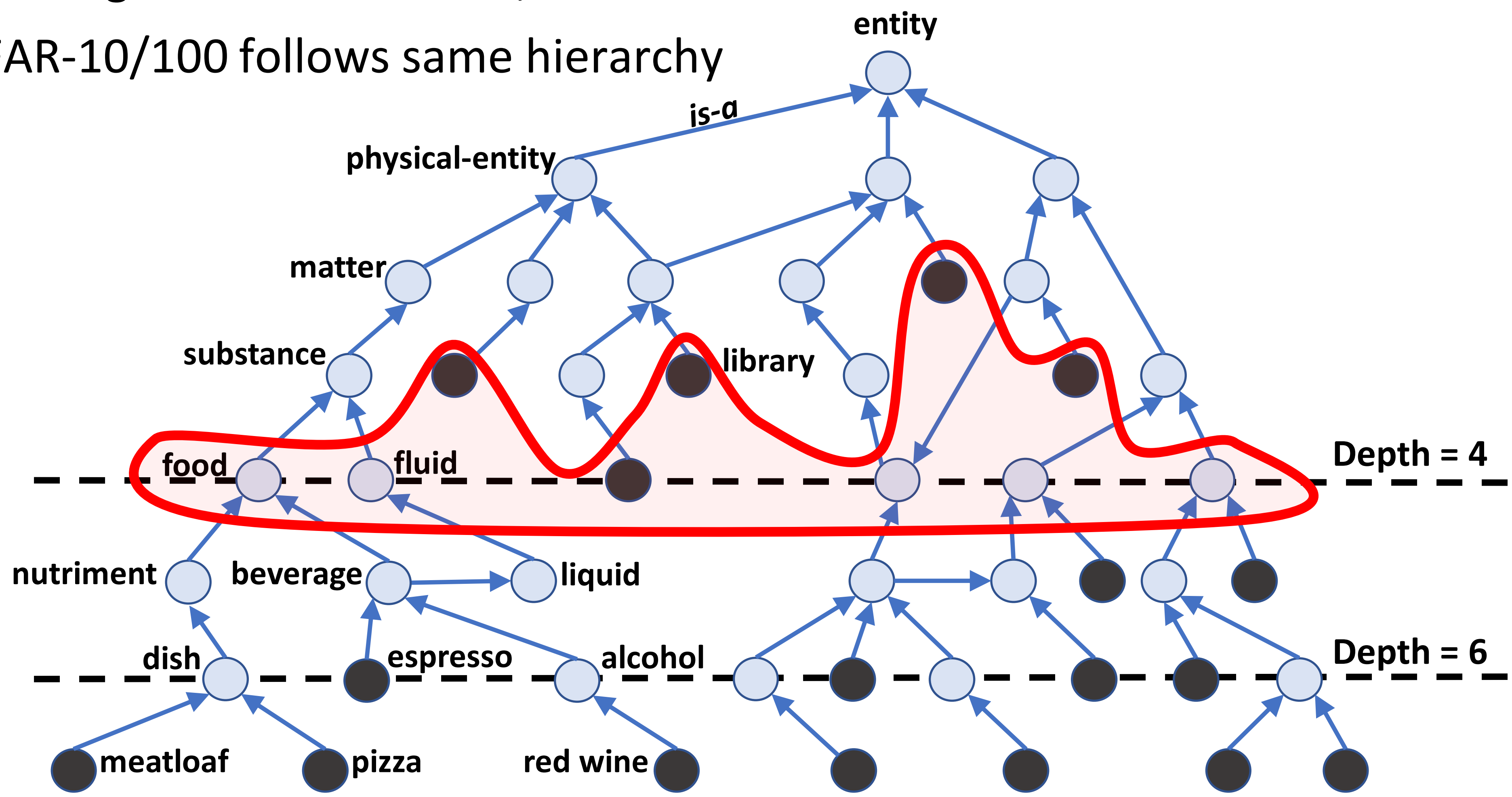
features are “masked”

- No abstraction level
- No Energy/Accuracy trade-off



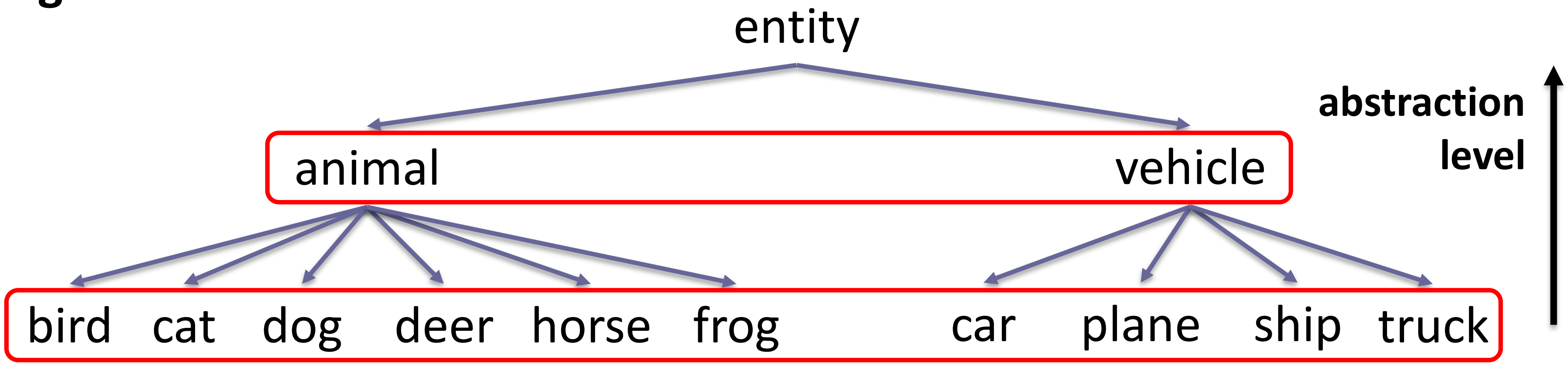
Training Data-Sets are Hierarchical

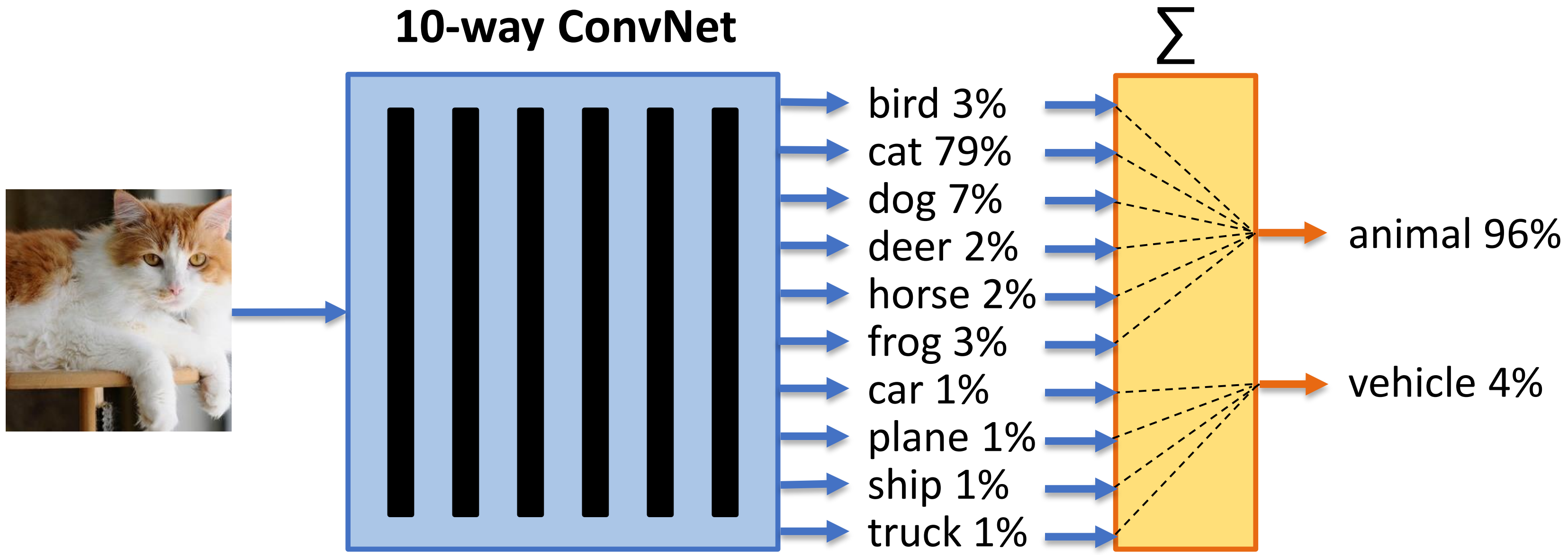
- **Common datasets reflects the semantic abstraction of human reasoning**
  - E.g. ImageNet: 1000 classes, 16 levels of abstraction
  - CIFAR-10/100 follows same hierarchy



Multilevel classification with ConvNets

- **E.g. Image Classification in CIFAR-10**

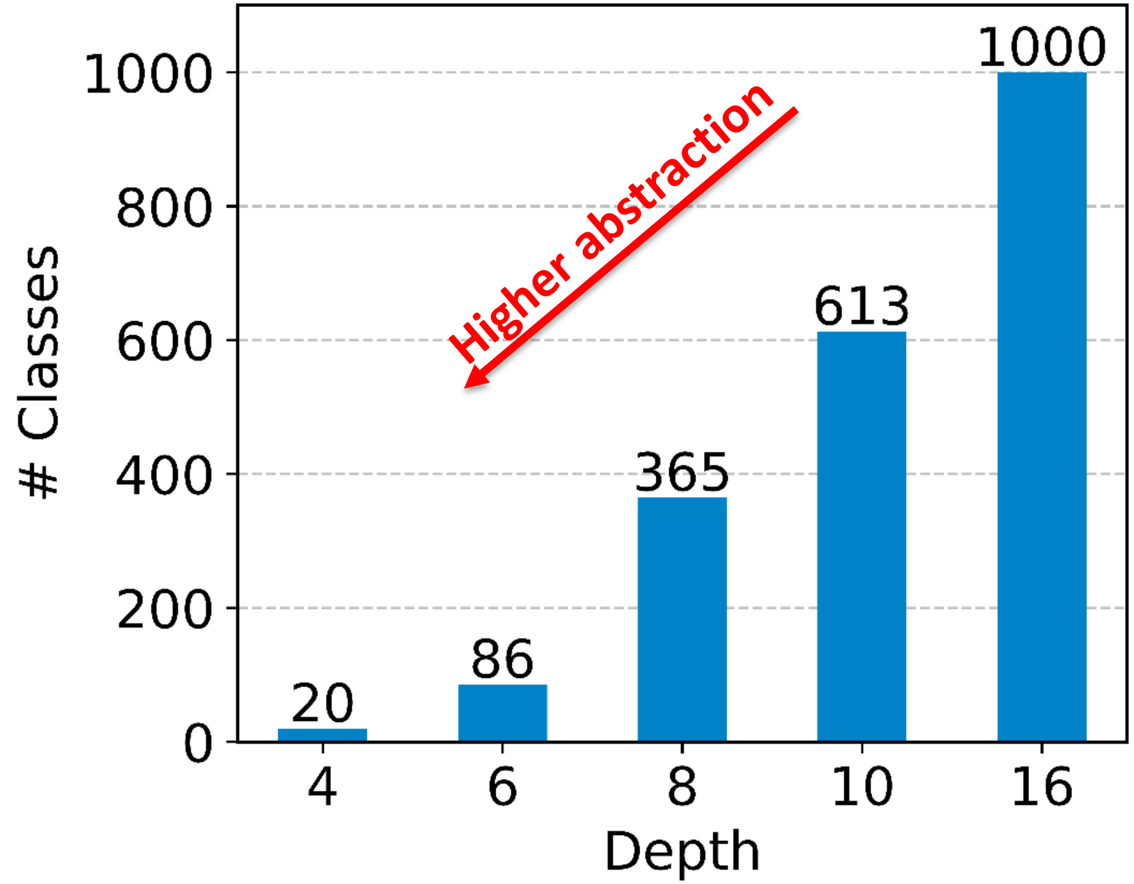




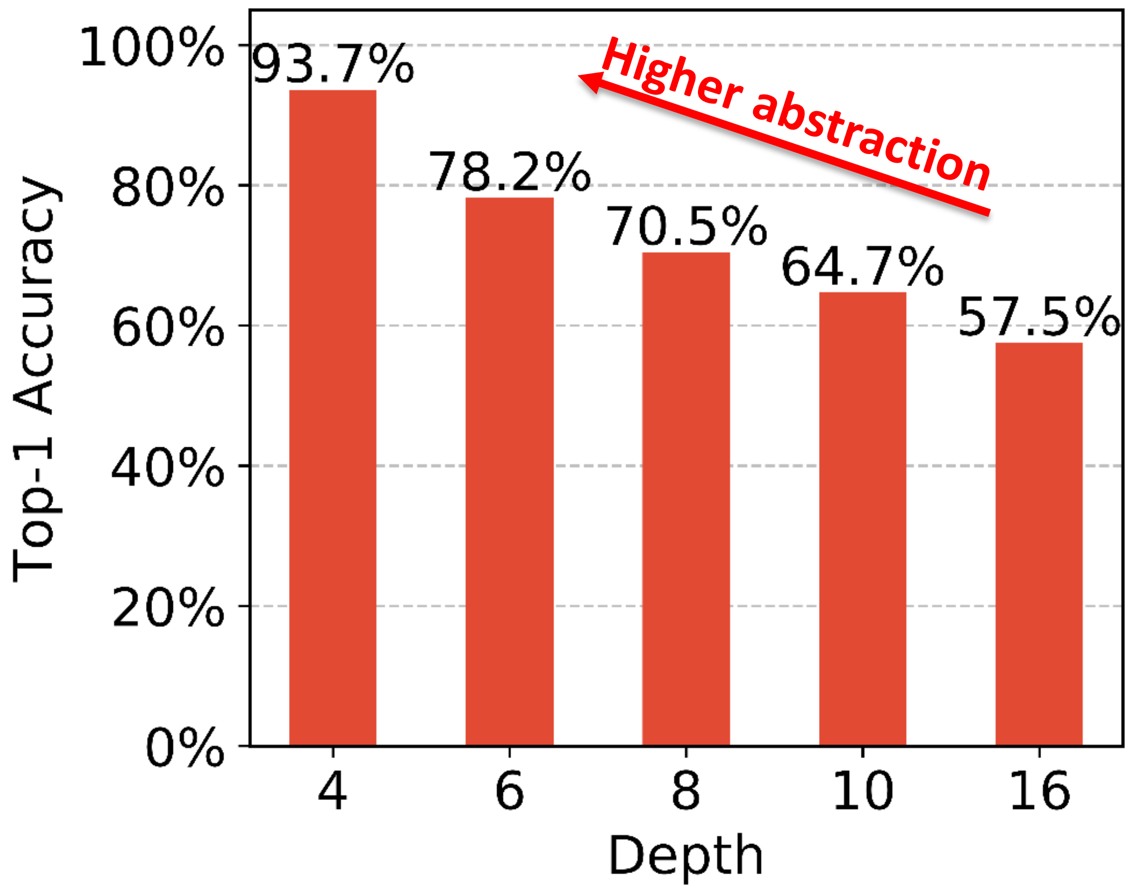
P(animal) = P(bird) + P(cat) + P(dog) + P(deer) + P(horse) + P(frog)

Scalable Effort ConvNets: Results (1)

- **Multi-level Classification on ImageNet**

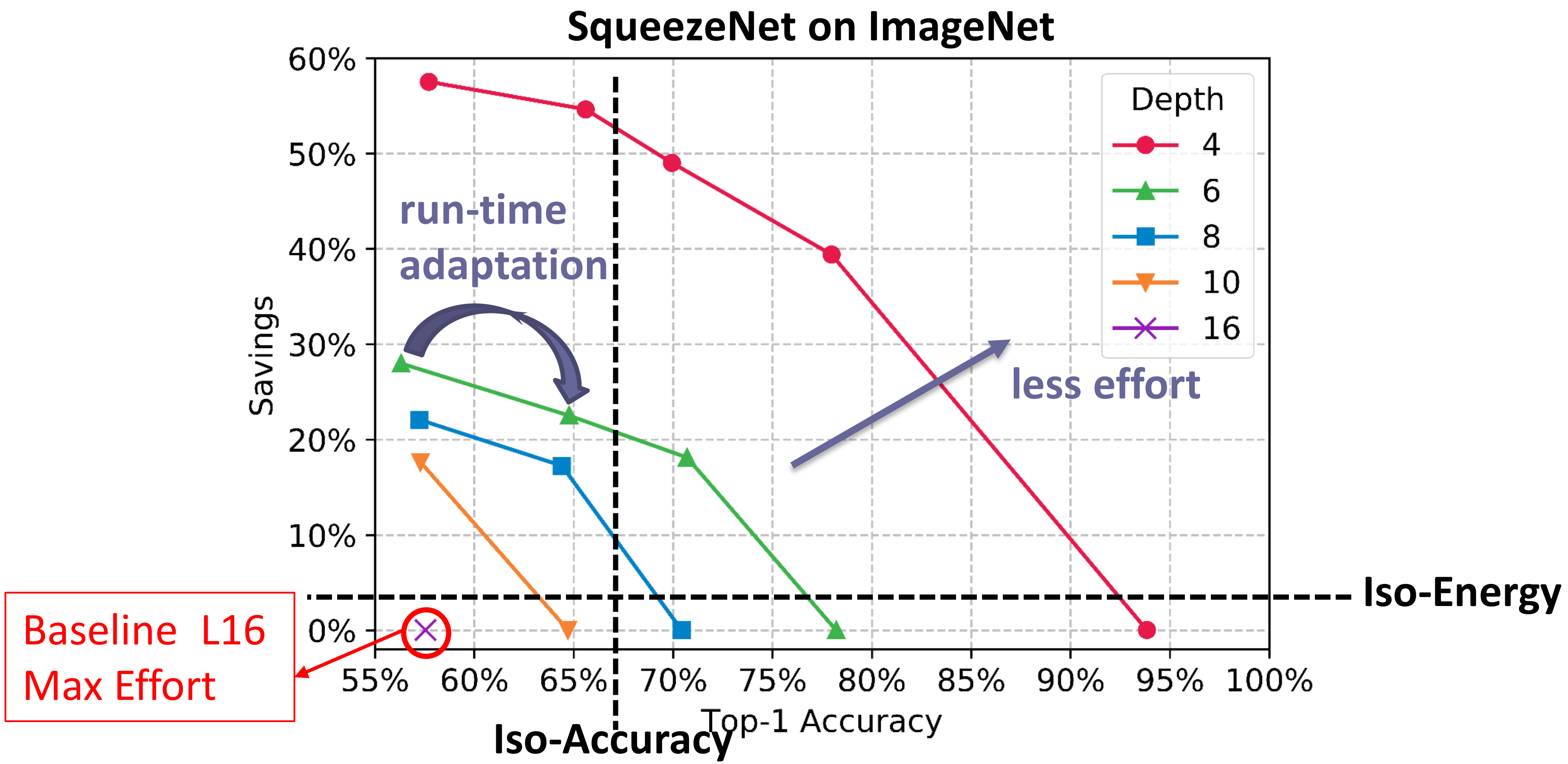


SqueezeNet



Scalable Effort ConvNets: Results (2)

- **Adaptive ConvNets**
  - Multilevel Classification → increase accuracy with same effort
  - Per-layer Precision Scaling →define multiple points in the energy-accuracy space



3. POWER OPTIMIZATION

Motivations

- **1. Temperature**
  - Embedded SoCs have limited TDP
    - →High temperature when running intensive workloads (e.g. inference)
    - →Peak-performance for short run-time windows.

Neglected by SoA NN optimization

- **2. Energy**
  - Energy reduction via power minimization

Goal:

Dynamic HW: DVFS

Dynamic Voltage Frequency Scaling (DVFS)

Voltage-Scaled ConvNets — CPU

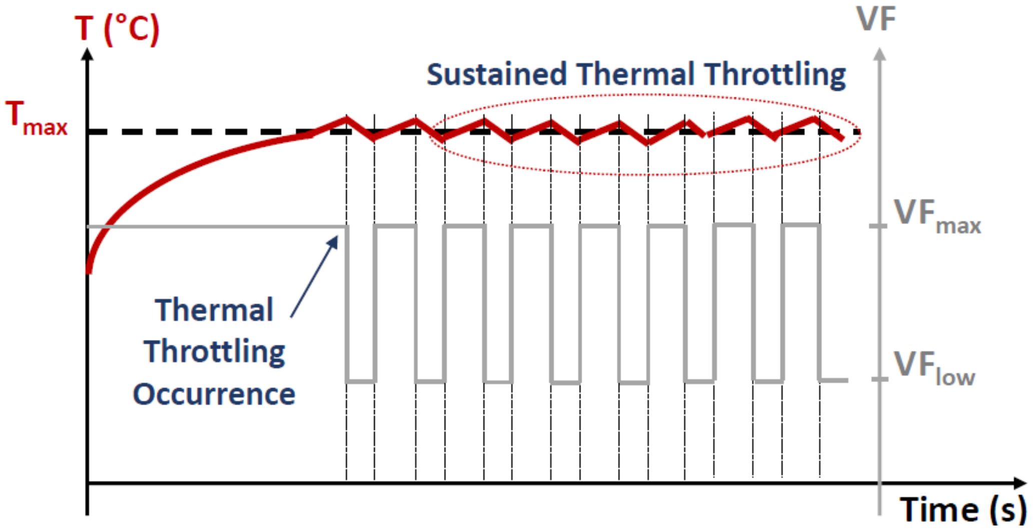
Performance Profiling of Embedded ConvNets under thermal-aware DVFS

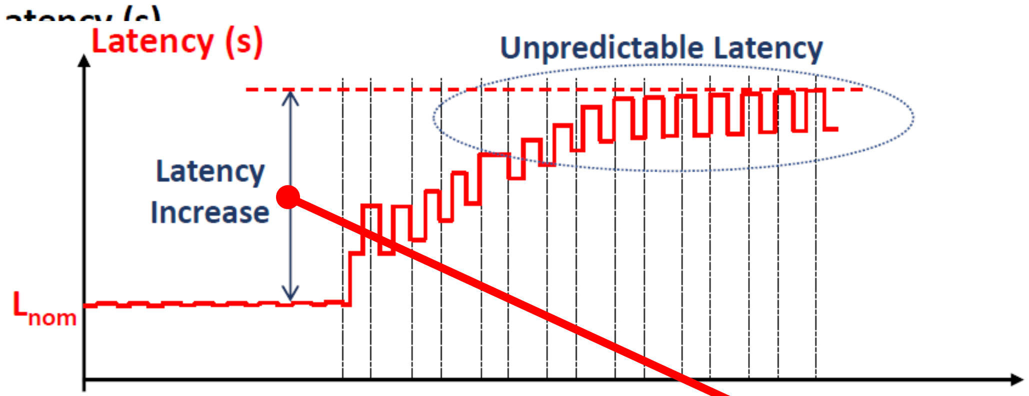
Beyond DVFS — ASIC

FINE-VH: a novel power distribution scheme

Thermal-Aware DVFS

- **<u>Problem:</u>** Data Analytics on a stream-of-data → *Continuous Inference*
- **<u>Challenge:</u>** Mobile SoCs have limited TDP



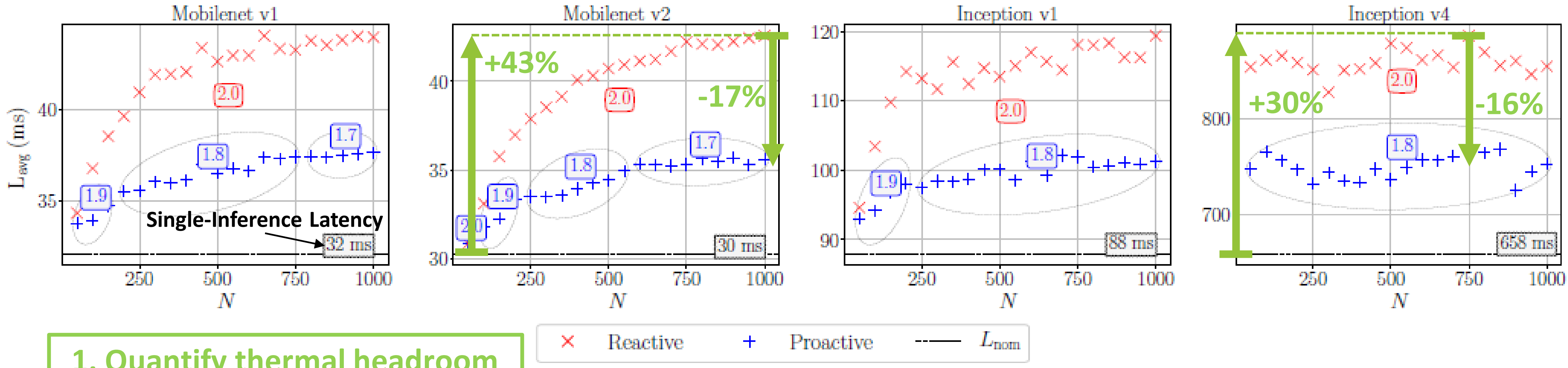


nullify savings brought by NN optimization!

- **<u>Thermal-Aware DVFS:</u>** Reactive vs Proactive
- **<u>Goal:</u>** Identify the optimal VF operating point

What about ConvNets?

Voltage-Scaled ConvNets on ARM Cortex-A15



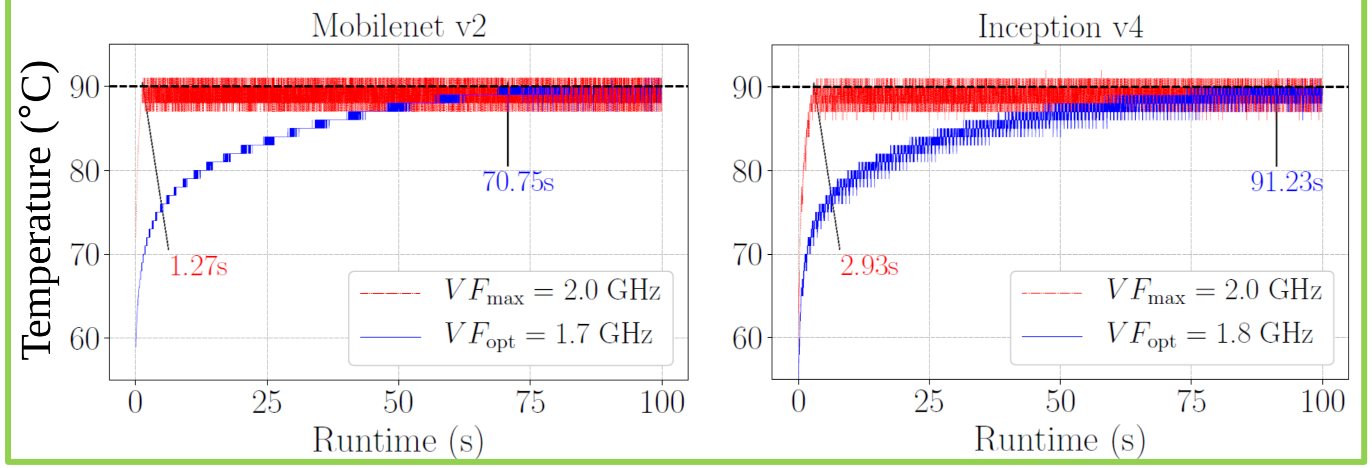
1. Quantify thermal headroom

ConvNet	N_{safe}	t_{safe} (s)
MobileNet v1	39	1.26
MobileNet v2	42	1.27
Inception v1	25	2.21
Inception v4	4	2.93

T<90°C

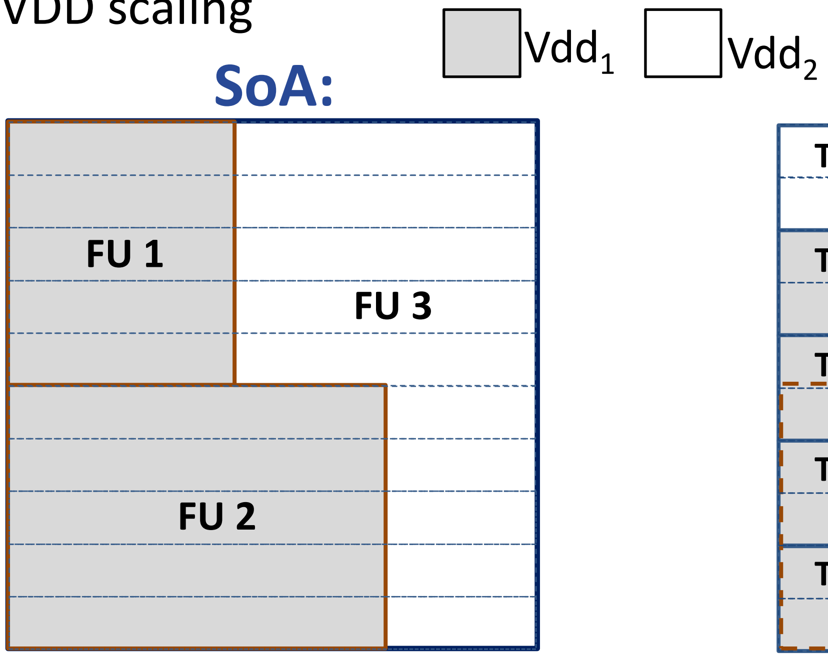
2. Assess latency under thermal-aware DVFS

3. Demonstrate thermal profile depend on topology

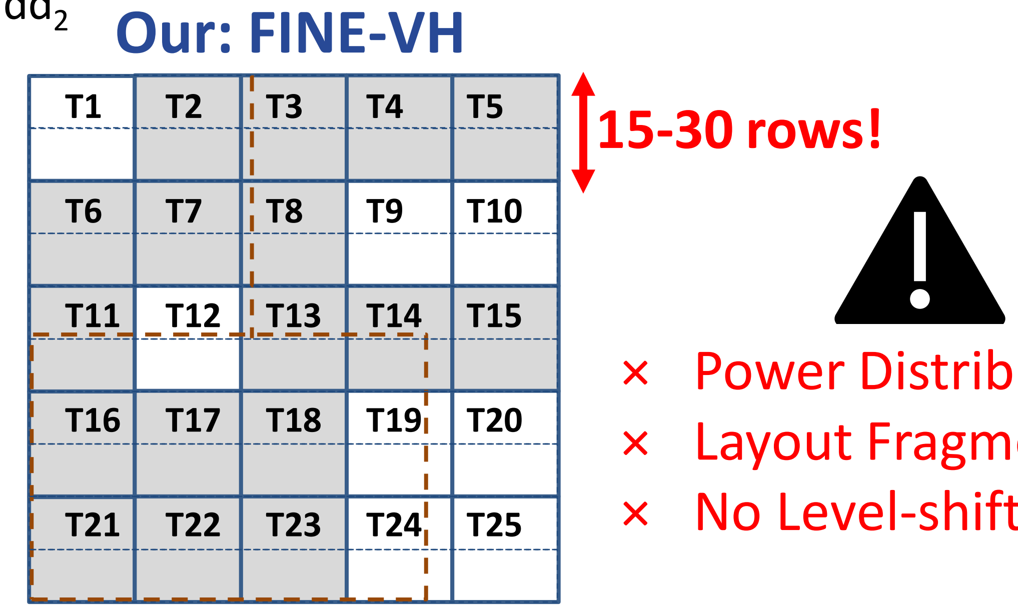


Beyond DVFS: FINE-VH

- **<u>Goal:</u>** apply a finer VDD scaling



FU = Functional Unit



T = Layout Tile

× Power Distribution
× Layout Fragmentation
× No Level-shifters

- **<u>How:</u>** Fully automated design and simulation flow integrated on a standard EDA tool
- **<u>Validated on:</u>**
  - RISC Core
  - Deep Learning Accelerator

FINE-VH outperforms DVFS

- **Limited area overhead: 6% w.r.t. standard flow**
- **From 32.0% to 38.2% w.r.t. ideal-DVFS**

Wrap-up

1. MEMORY

MCU

Prune and Quantize

Memory vs. Accuracy design-space exploration

3$ HW is enough:
3x compression with <1% loss

MCU

Encoding-Aware Sparse Training

Maximize compression of encoding schemes

+8.73% accuracy at 12KB

2. ENERGY

ASIC

Online Precision Scaling

Dynamic Energy-Accuracy Scaling

Up to 35.2% savings
with <8% loss

ASIC

Scalable-Effort ConvNets

Dynamic Energy-Accuracy-Abstraction Scaling

40% more accurate
or 60% more efficient

3. POWER

CPU

Voltage-Scaled ConvNets

Performance profiling under thermal-aware DVFS

Look at Temperature!
Safe latency: 1-3s

ASIC

FINE-VH

Novel power distribution scheme to improve DVFS

Up to 38.2% power savings

The Lesson Learnt

The definitive solution does not exist!

TRAINING

Present: Exploratory Data Analysis

- Data Collection/Cleaning
- Data Visualization
- Assess different hypothesis:
  - Hyper-Param. Optimization
  - Learning Strategy
    - Supervised, Self-Supervised, Transfer Learning etc.
  - …

OPTIMIZATION

Future: Exploratory Optimization Analysis

- Design-Space Exploration
  - Accuracy, Memory, Energy, Power…
- Cost Analysis
  - Which HW?
- Assess different hypothesis:
  - NAS
  - Pruning
  - Quantization
  - Static vs. Dynamic…

Research Activities

4	14	3
journals	conference papers (3 best paper candidates)	book chapters

Technical Speaker at:

- 2 international conferences (ICCAD18 and SNAMS19)
- 1 national workshop (IWES18)

Live Demonstrations at 2 international conferences (DATE19 and ISLPED19)

SENSEI - Sensemaking for Scalable IoT Platforms with In-Situ Data-Analytics: A Software-to-Silicon Solution for Energy-Efficient Machine-Learning on Chip (2 years)

Thank you

Question Time