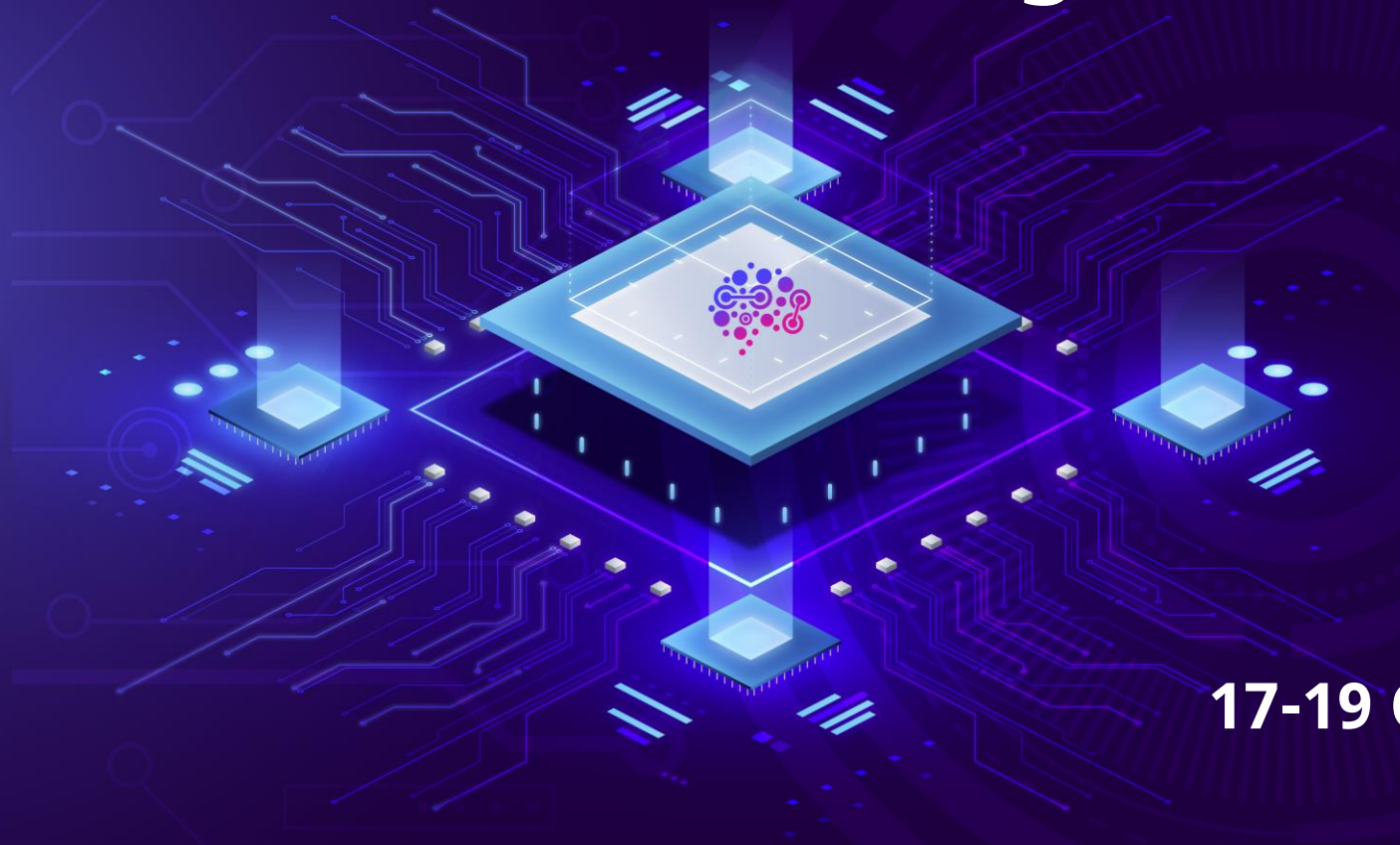
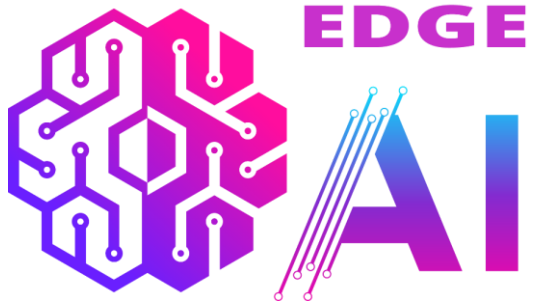
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



Quantization- and HW-aware Neural Architecture Search for EdgeAI solutions

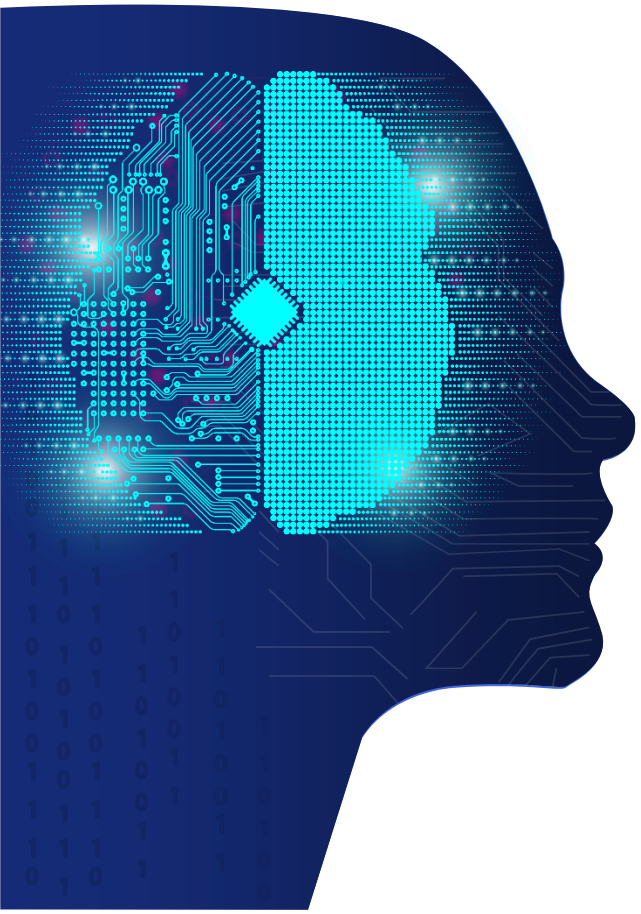
Gerardo Daalderop, Hiram Rayo Torres Rodriguez, Nick van de Waterlaat, Sebastian Vogel, Willem Sanberg et al.

NXP Semiconductors, the Netherlands

17-19 October 2023 Athens, Greece

Copyright © 2023

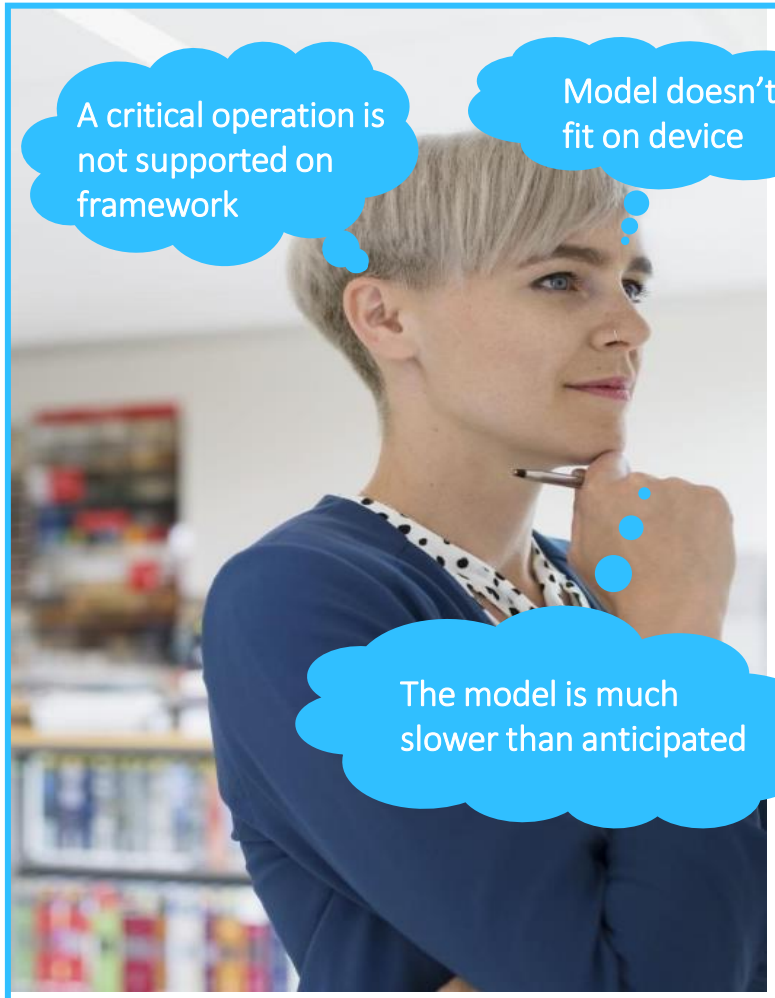
Presentation Outline



- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction: AI design versus Deployment across the Edge

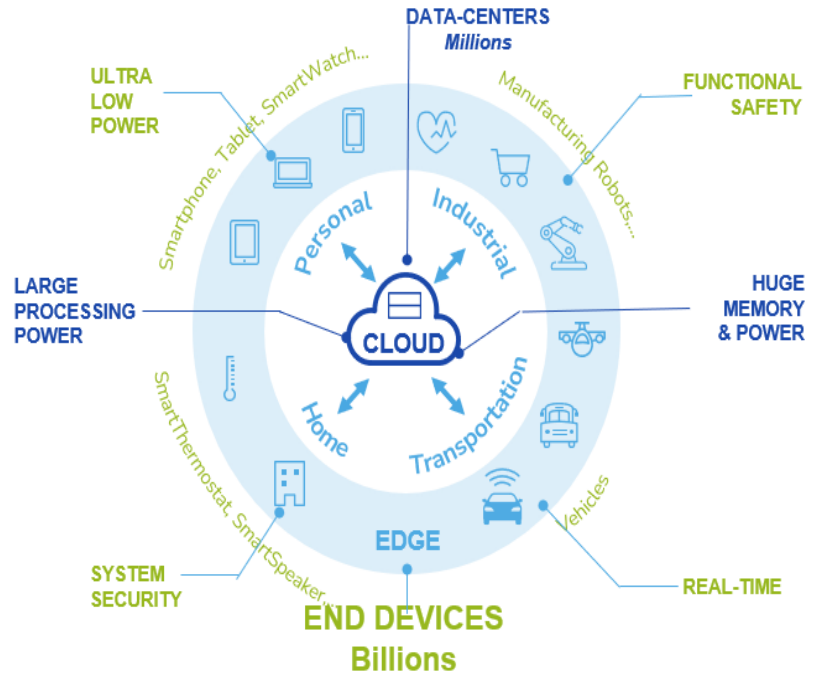


AI design vs deployment dilemma

- **Deployment is infrequently considered**
- **When deployment is not considered during model design, problems occur**

AutoML

CONTEXT: LARGE VARIETY OF SMART EDGE NODES DEMANDS OPTIMIZATION OF MULTIPLE CONCURRENT REQUIREMENTS



Neural Networks increasingly provide the 'Smartness'

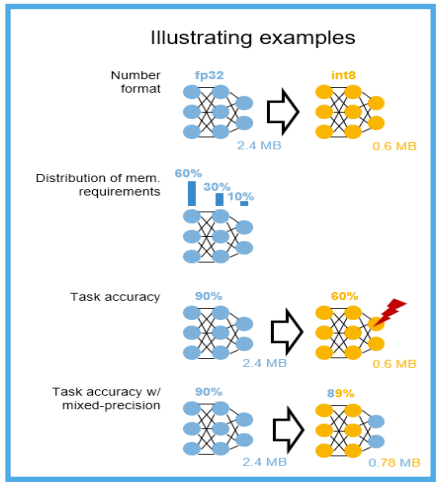
Yet today, neural networks are trained largely irrespective of other embedded requirements and specifications

Copyright © 2023

State-of-the-art Optimization for the Edge

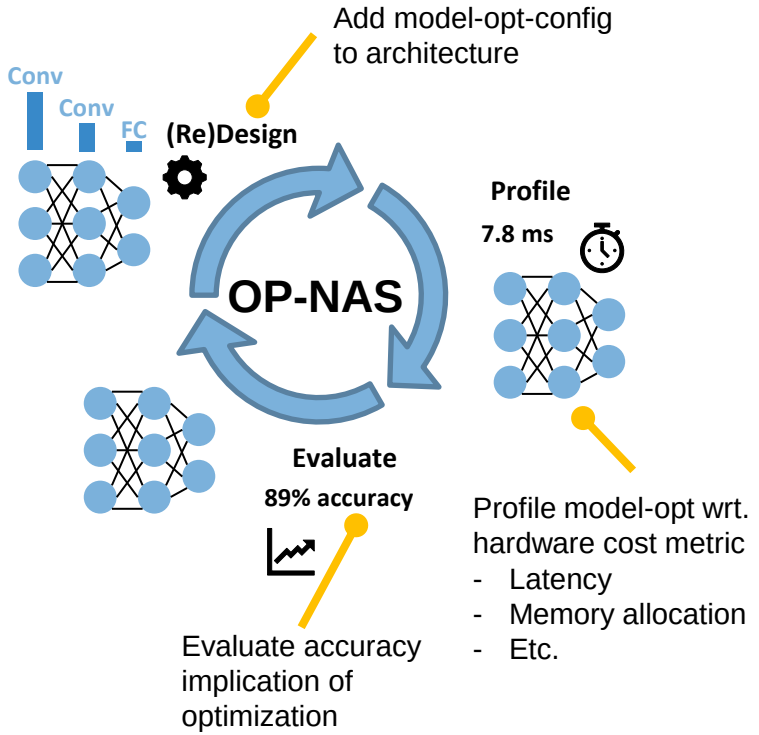
- AutoML
  - Methods and processes for accelerating research and development in ML
    - Preprocessing/cleaning data
    - Selecting/Constructing model features
    - Hardware-aware optimization of network design and training-related hyperparameters (NAS + HPO)
    - Postprocessing ML Outputs
  - Reduces complexity of tasks that are beyond non-ML experts
  - Automated deployment, but not tied to model optimization

- Traditional model optimization
  - Quantization, Mixed Precision
  - Pruning
  - Weight Clustering
  - Knowledge Distillation
  - Low rank NN Compression
  - ……



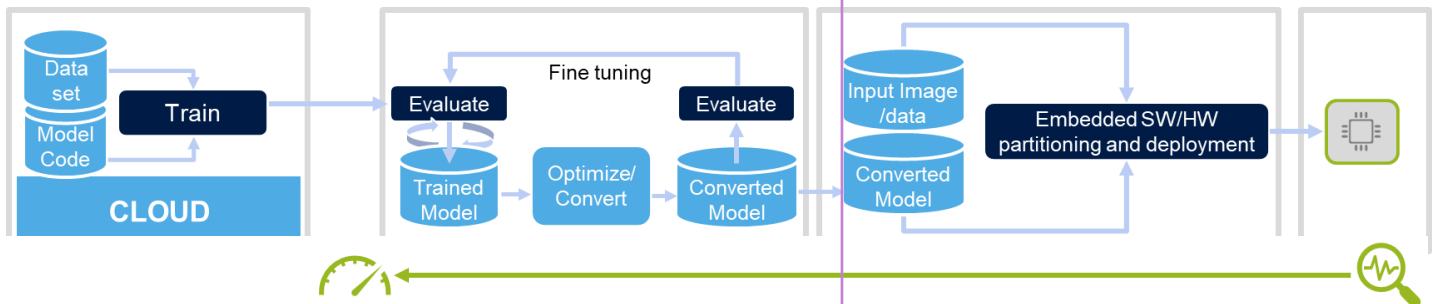
Integrated Approach

Model-Opt & SW/HW-aware OP-NAS



Copyright © 2023

Methodology Towards scalable, collaborative optimization- and hw-aware automated design

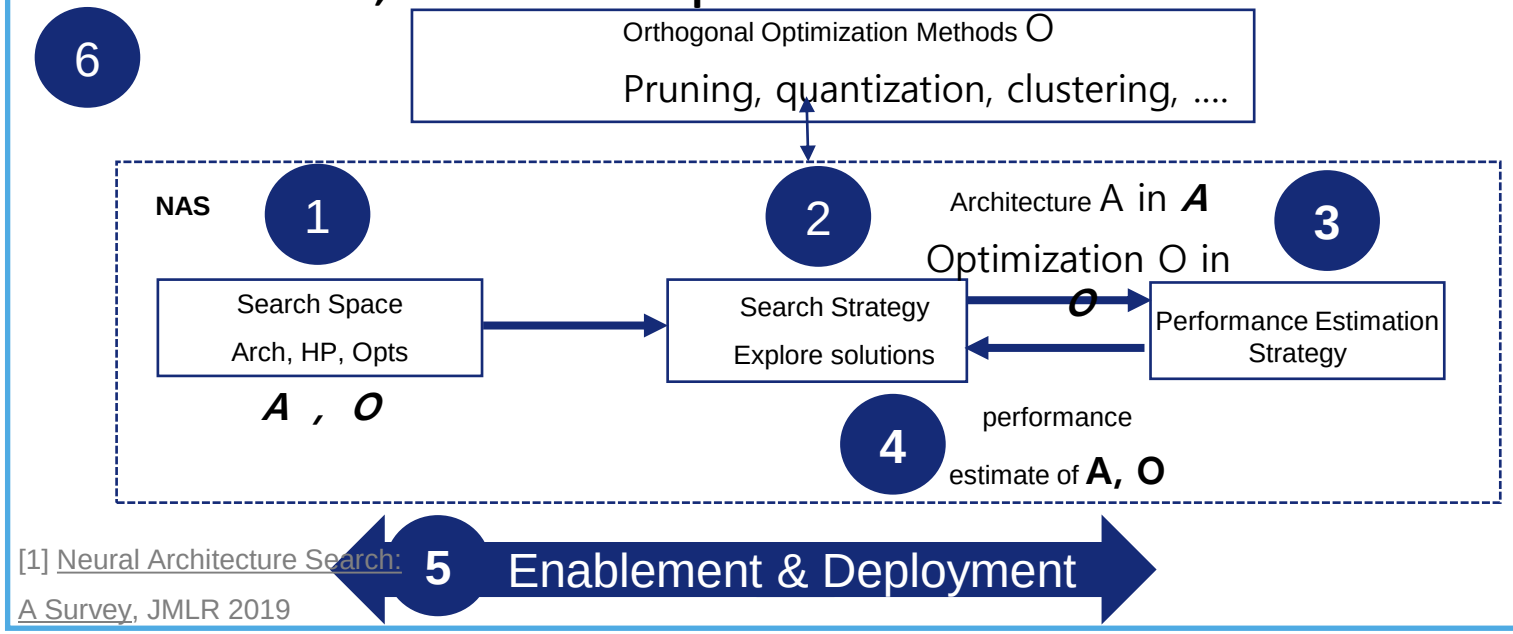


Multi-objective DNN design by Neural Architecture Search

1) Optimize task accuracy
2) Taking edge-node properties into account

NXP Profiling API
providing latency, energy, etc.

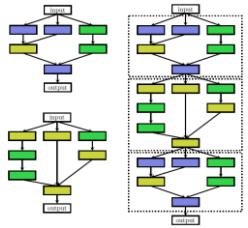
Towards scalable, collaborative optimization- and hw-aware automated design



[1] Neural Architecture Search: A Survey, JMLR 2019

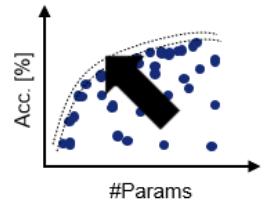
Scalability

1 Search Space (Smaller space)



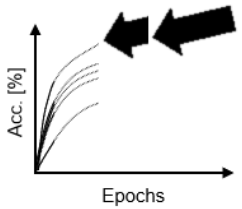
- Multi-branch networks incl. HPs*
- Chain structured DNNs**
- Cell-based
- Macro-architecture/One-shot model

2 Search strategy (Faster convergence)



- Random search
- Reinforcement learning
- Evolutionary/Genetic Algorithms
- Bayesian optimization esp. for HPs*
- Differentiable architecture search

3 Performance Estimation (Faster estimation)



- Full training
- Learning curve extrapolation
- *Additionally for 2ndary objectives:*
- *Profiling, MIL#, HIL##*
- *Surrogate-model-based evaluation*

Copyright © 2023

Research Goals

- Integrated approach:
  - HW-aware Multi-objective Optimization
  - Automated Design (Neural Architecture Search)
  - Scalable Collaborative Optimization Framework → Experience, Guidelines and Best Practices
  - Evaluation of Cost and Benefits

Benefits:

- **Ease-of-Use** → Easier for customers to develop **optimal AI for Inference nodes**
- **Quality of Service gain** at equal or better accuracy → 'Smaller' HW beats 'bigger' HW
- **HW** and **SW insights** → Guide customers & derive **insights in design requirements**
- **Differentiator** → e.g. **SW tooling, consultative selling**

Copyright © 2023

Design and Implementation on the basis of 4 examples

Use Case	Dataset	HW Target	Network
Audio Denoising	NXP collected & public datasets	Minsky (Arm Ethos & Coolflux DSP)	LSTM GRU
Image based Person detection	Pascal VOC	i.MX RT117x	CNN
Image based Semantic Segmentation	Cityscapes & MS-COCO	i.MX8M+	CNN
Image based Classification	CIFAR10/100	NXP internal	MobileNetV2

Search Space

- Multi-branch networks incl. HPs*
- Chain structured DNNs**
- Cell-based
- Macro-architecture/One-shot model

Search strategy

- Random search
- Reinforcement learning
- Evolutionary/Genetic Algo
- Bayesian optimization esp. for HPs*
- Differentiable architecture search

Performance Estimation

- Full training
- Learning curve extrapolation
- *Additionally for 2ndary objectives:*
- *Profiling, MIL#, HIL##*
- *Surrogate-model-based evaluation*

Copyright © 2023

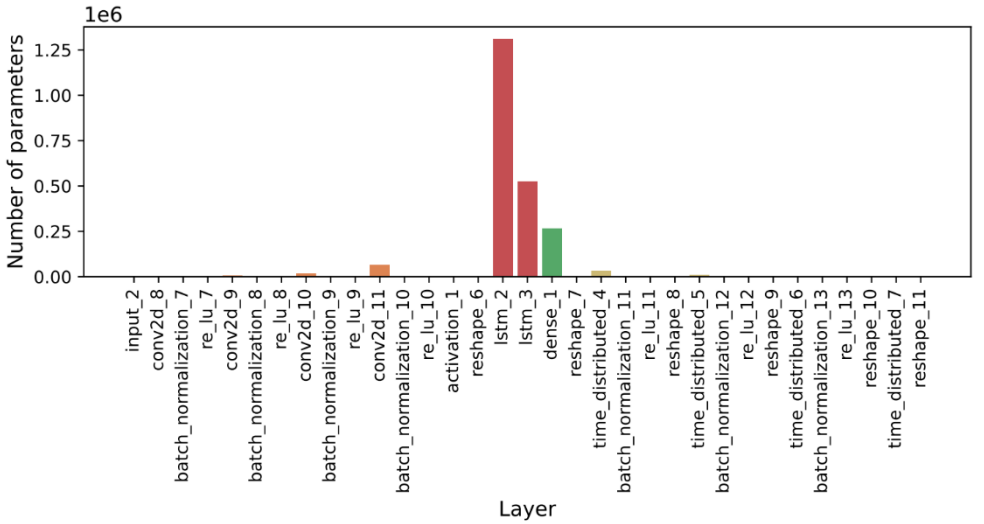
Design and implementation (1) We analyze seed networks layer-wise statistics w.r.t secondary metrics to identify bottlenecks and design a search space around

- In order to find networks that outperform the seed, we carefully consider the design of the **search space.**

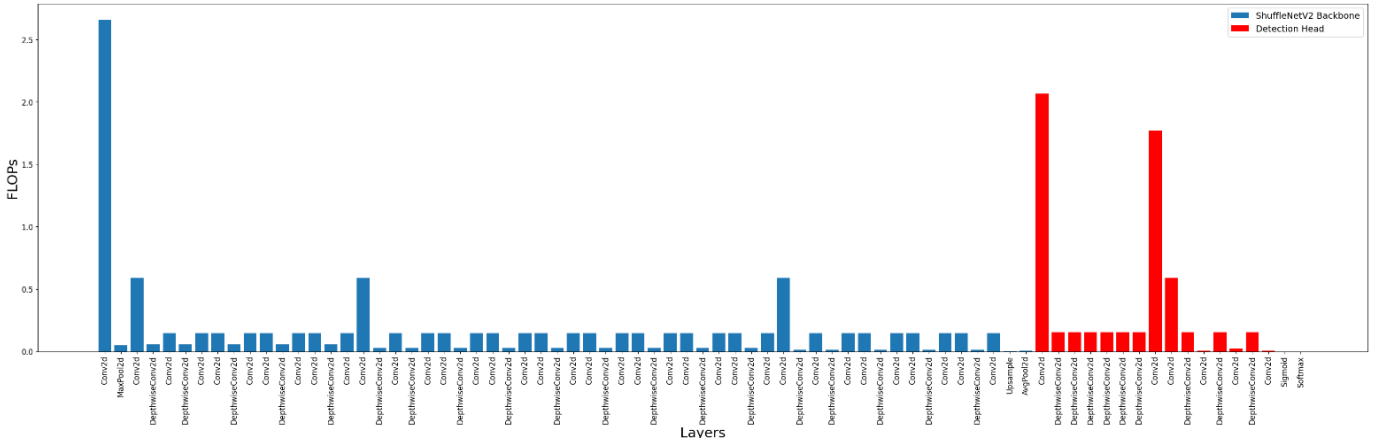
- We collect the **layer-wise statistics** corresponding to the **secondary** metric of interest (e.g., memory, latency) to find bottlenecks.

- We introduce **searchable operations** to optimize the bottlenecks in these layers.

Audio denoising



Object detection



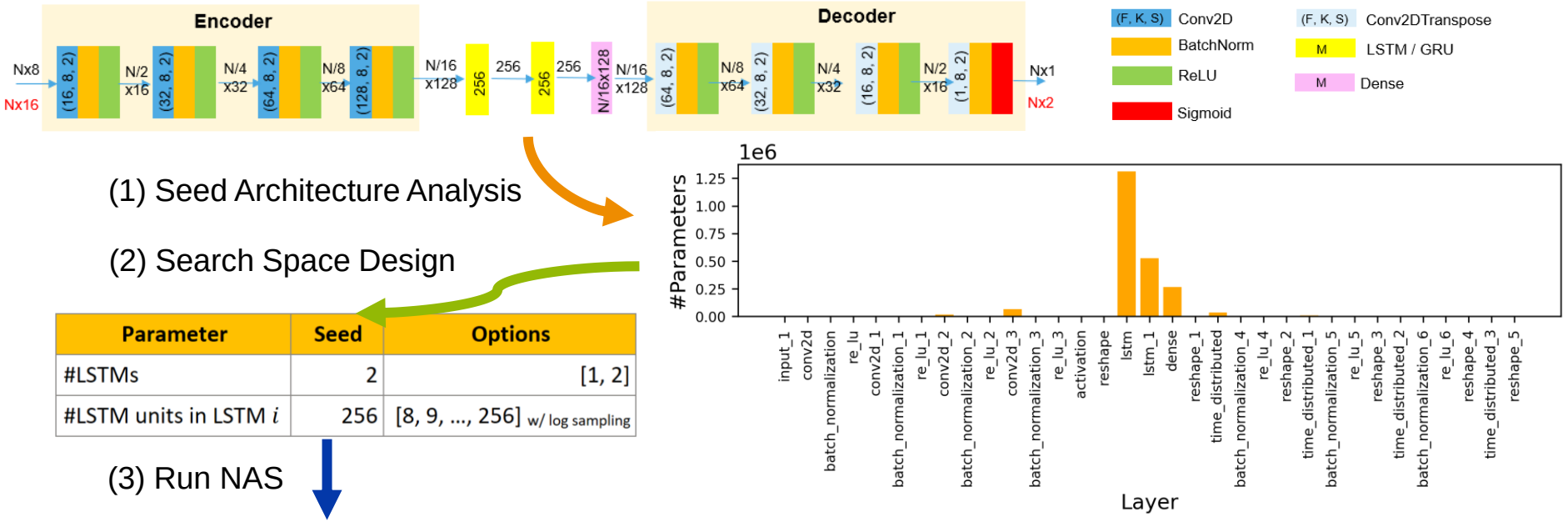
Copyright © 2023

Design and implementation (2) Optimizing an audio denoising network for high task performance and reduced model size

Seed Architecture

Search Space

Neural Architecture Search (NAS)

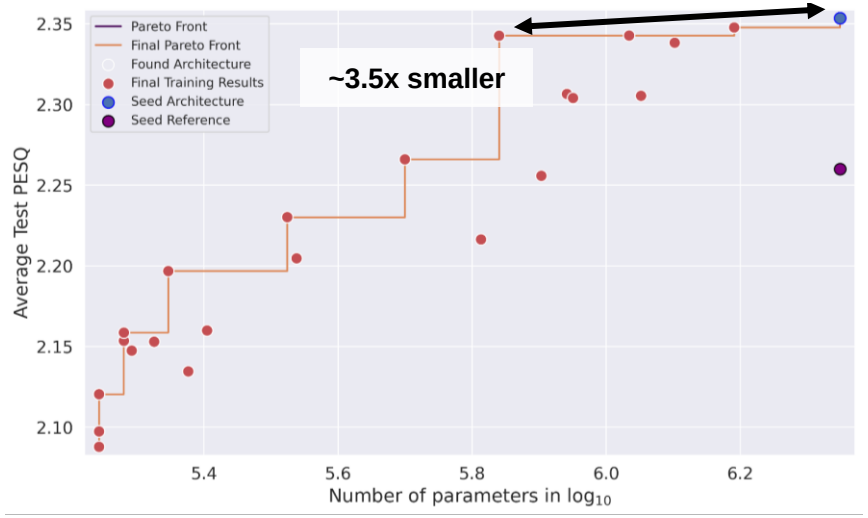


(1) Seed Architecture Analysis

(2) Search Space Design

Parameter	Seed	Options
#LSTMs	2	[1, 2]
#LSTM units in LSTM *i*	256	[8, 9, …, 256] w/ log sampling

(3) Run NAS

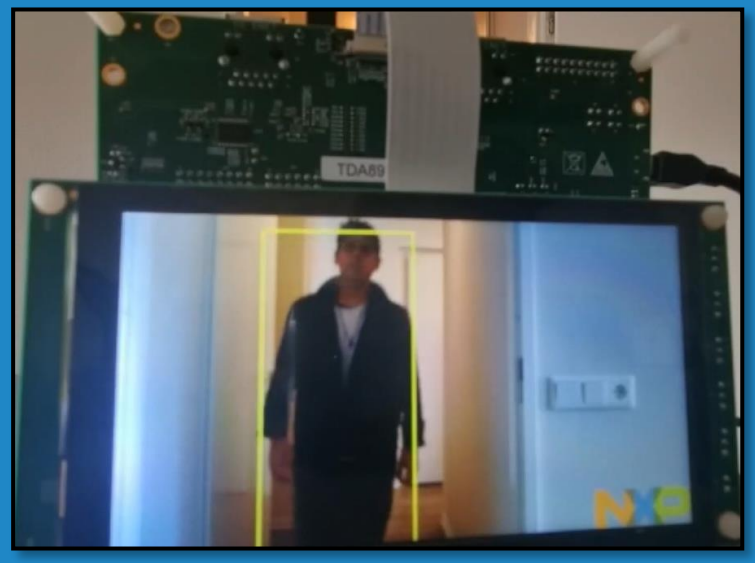


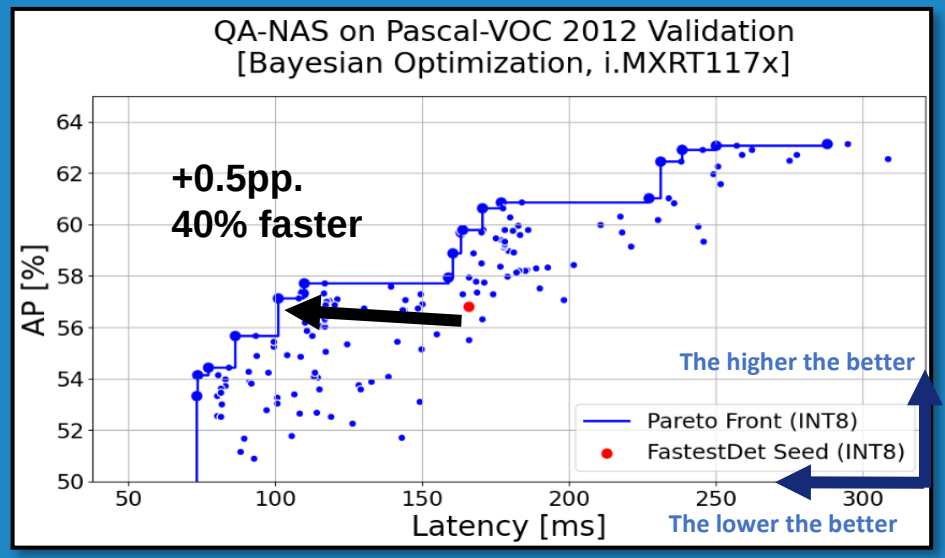
Targeted search space design through a **network analysis** enables **NAS** to find a network **~3.5x smaller** at **near iso-performance** without exploding the number of possible architectures.

Copyright © 2023

Other experimental results: Quantization-aware NAS
From FP32 to Int8; Object Detection

Our NAS methodology helps in reducing the seed network latency by up to 40% while still improving the accuracy of predictions



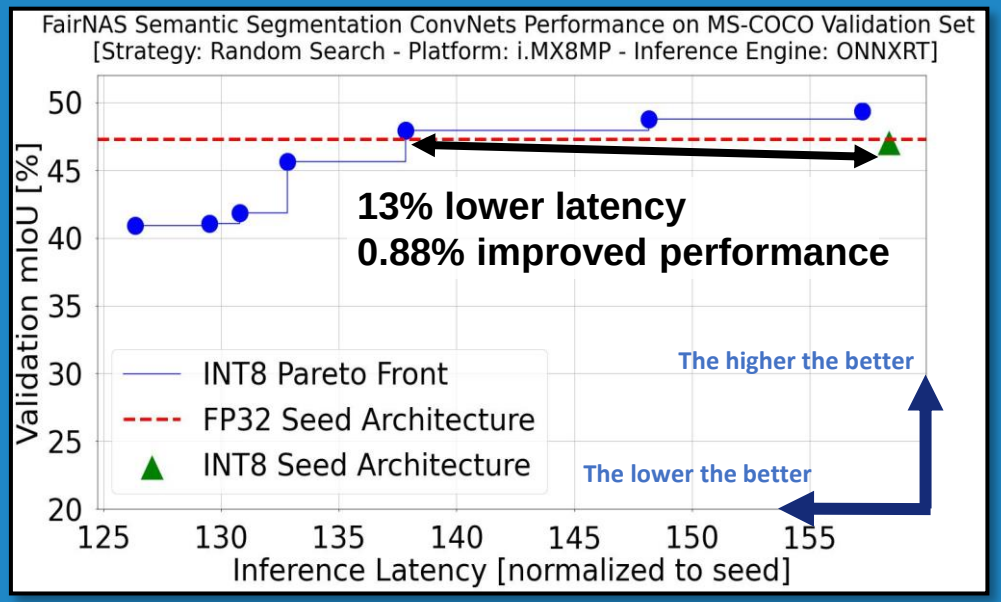


- **NXP Platform:** i.MXRT1170/1160
- **NN Architecture:** FastestDet (ShuffleNetV2)
- **Dataset:** PASCAL-VOC (customer-request)

Copyright © 2023

Experimental results: Quantization-aware NAS
From FP32 to Int8; Semantic Segmentation

Our NAS methodology helps in reducing a seed network latency by up to 13% while still improving accuracy in terms of mIoU by 0.88%

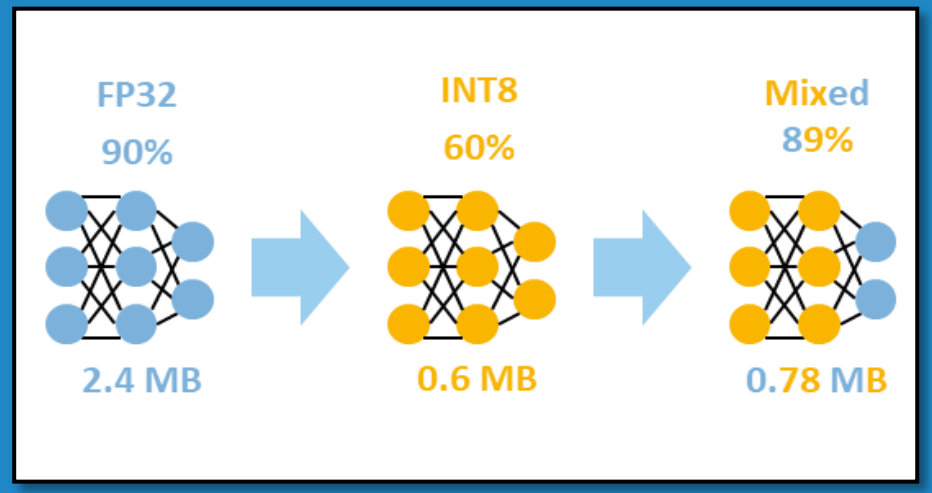


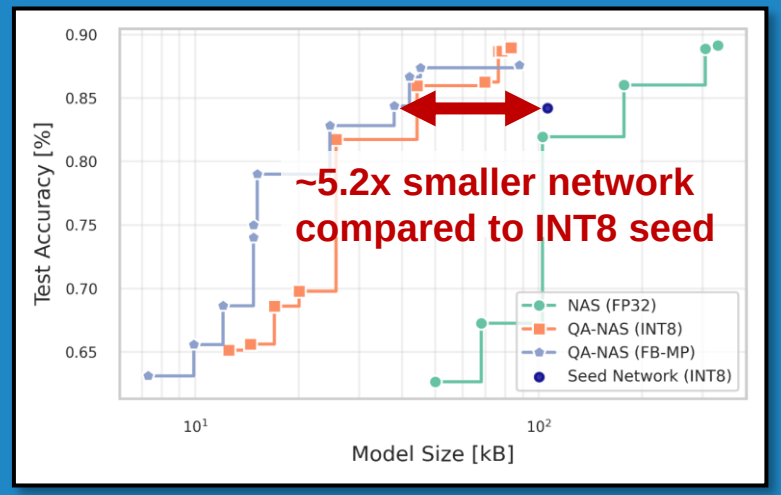
- **NXP Platform:** i.MX8M Plus
- **NN Architecture:** DeepLabV3
- **Dataset:** MS-COCO

Copyright © 2023

Experimental results Quantization-aware NAS
From FP32 to Mixed Precision ; CIFAR 10/100 Image Classification

Our NAS methodology can exploit features like Mixed-Precision computation, finding networks that outperform the seed network (INT8) in terms of accuracy and model size





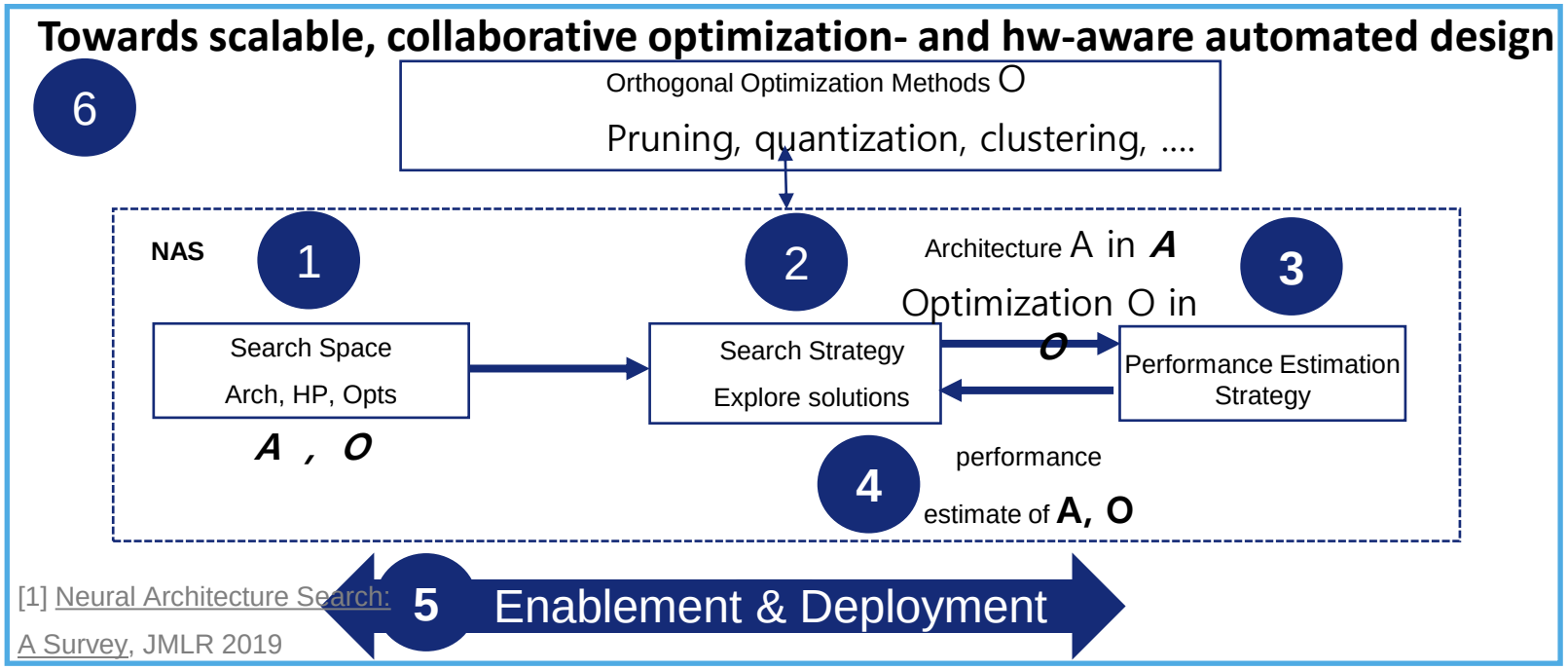
Search space

Block	Parameter	Choices
Inverted bottleneck 1	kernel size	[2, **3**, 4, 5, 6, 7]
	width multiplier	[0.01, 0.05, **0.1**, 0.2, 0.3]
	expansion factor	[**1**]
	repetitions	[**1**]
Inverted bottleneck 2-6	kernel size	[2, **3**, 4, 5, 6, 7]
	width multiplier	[0.01, 0.05, **0.1**, 0.2, 0.3]
	expansion factor	[1, 2, 3, 4, 5, **6**]
	repetitions	[0, **1**, 2, 3, 4, 5]
Inverted bottleneck 7	kernel size	[2, **3**, 4, 5, 6, 7]
	width multiplier	[0.01, 0.05, **0.1**, 0.2, 0.3]
	expansion factor	[1, 2, 3, 4, 5, **6**]
	repetitions	[**1**]
Convolutional 2	number of filters	[128, 256, 512, 1024, **1280**]
	kernel size	[**1**]
	repetitions	[**1**]
Any	bitwidth	[4, 5, 6, 7, **8**]

- **NXP Platform:** internal
- **Dataset:** CIFAR10/100
- **NN architecture:** MobileNetV2

Copyright © 2023

Future Research



- **HW-aware Multi-Objective Optimization**
- **Variety of (combined) Optimization methods**
- **Automated Design through Neural Architecture Search**
- **Scalable Collaborative Optimization Framework and Guidelines**

- Enablement
- Evaluation of Cost and Benefits

Copyright © 2023

Conclusion

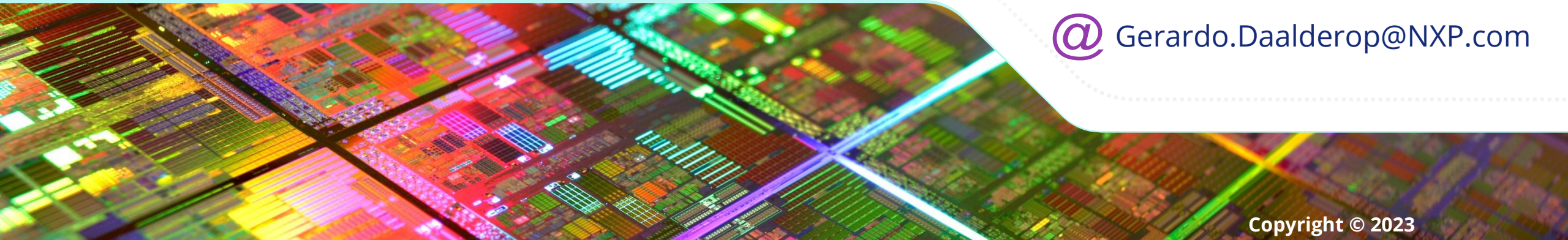
- **NAS is broadly applicable** and can be designed to **not be infeasibly expensive**.
- **NAS is a game-changer for "under-researched" use cases**. Accuracy improvements and order of magnitude efficiency improvements can be reported.
- **Few-bit mixed-precision quantization outperform** homogeneous 8-bit quantization in various use cases. Yet, **non-trivial NAS design decisions** influence outcomes.
- **NXP QA-NAS** methodology enables **efficient edge AI** exploiting current (& future) NXP hardware features.
- We are building up experience on **scalable collaborative optimization strategies through NAS**, and are striving towards (some) **guidelines and best practices**.

Copyright © 2023

EDGE AI

Thank You

For your attention



@ Gerardo.Daalderop@NXP.com

Copyright © 2023