# OPTIMIZATION AND DEPLOYMENT OF CNNS AT THE EDGE: THE ALOHA EXPERIENCE

Paolo Meloni

**University of Cagliari (UNICA)**

**ALOHA – software framework for runtime-Adaptive and secure deep Learning On Heterogeneous Architectures**
This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 780788
www.aloha-h2020.eu

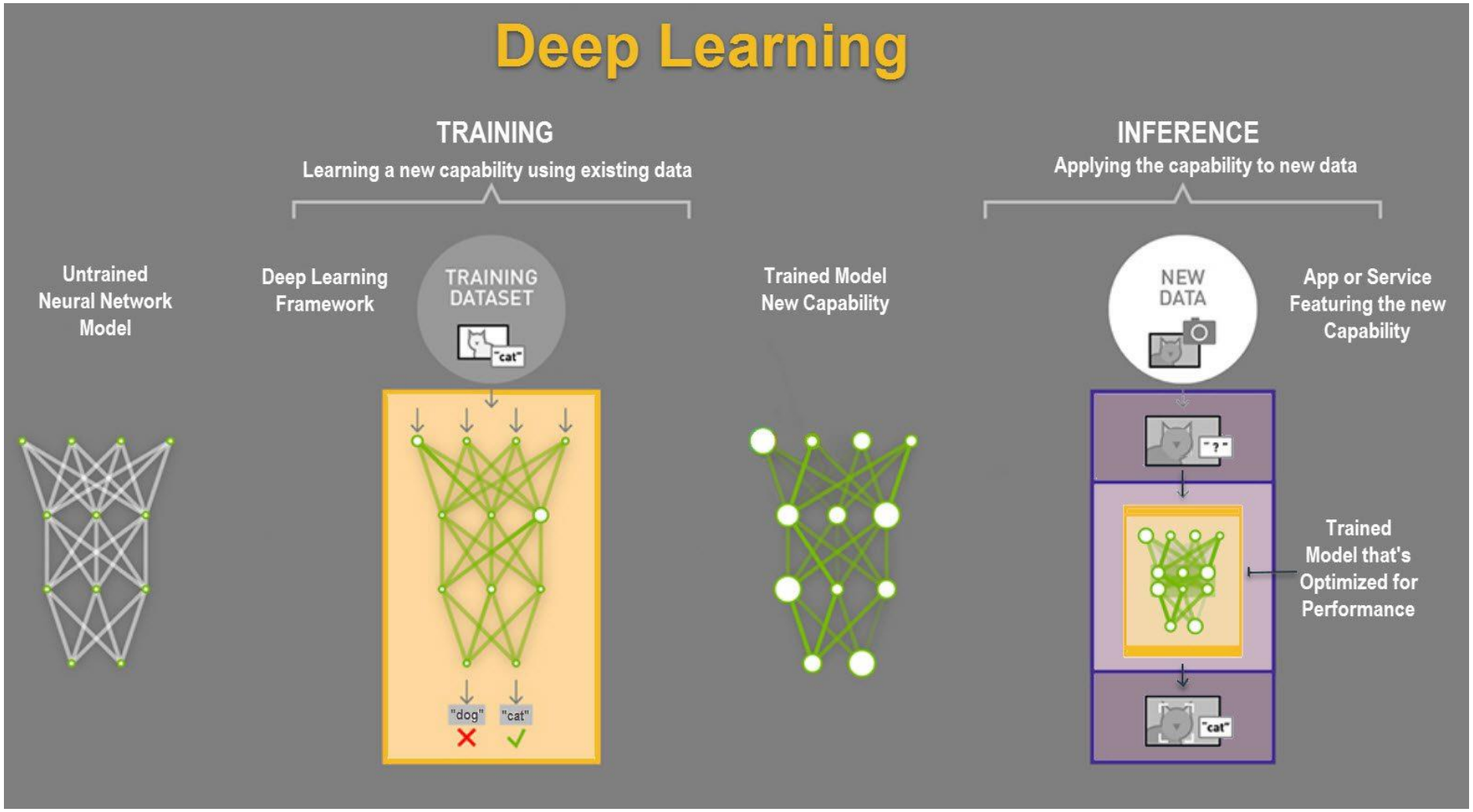
**Coordinator**

Giuseppe Desoli - STMicroeletronics

**Scientific Coordinator**

Paolo Meloni - Università degli Studi di Cagliari, DIEE - EOLAB

www.aloha-h2020.eu

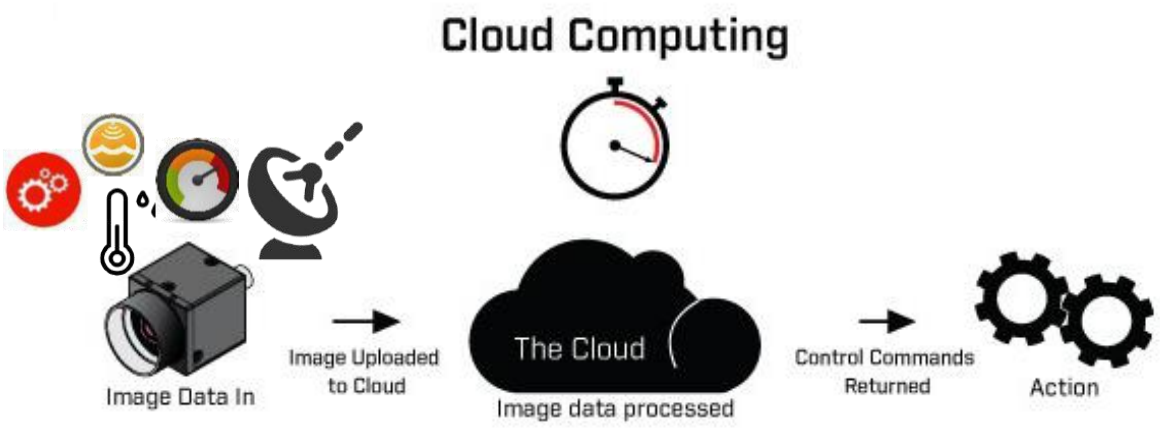


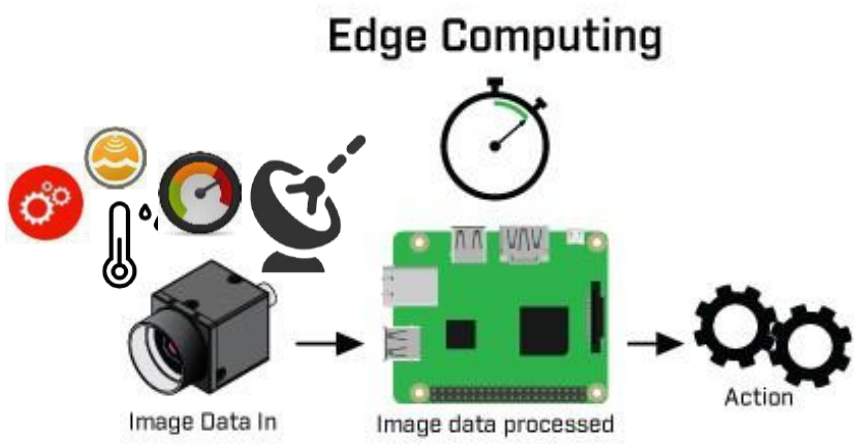



https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/

“Currently, around 10% of enterprise-generated data is created and processed outside a traditional centralized data center or cloud. By 2022, Gartner predicts this figure will reach 50%.”

*Santhosh Rao, Gartner.*




- Latency
- Bandwidth requirement
- Safety
- Resilience
- Security
- Privacy

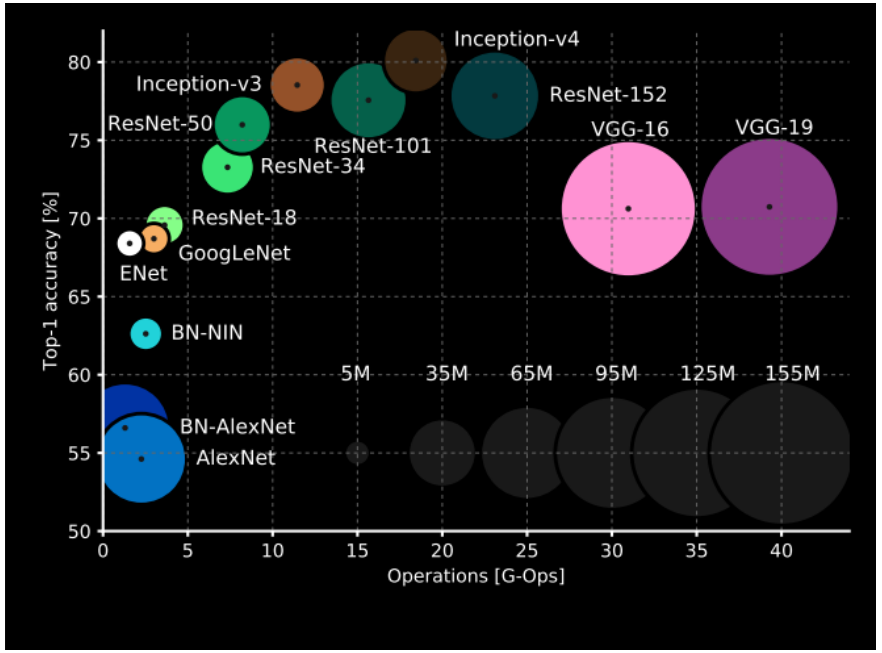
- Resource constrained computing platforms
- Limited power/energy/cost budget



https://eu.ptgrey.com/edge-computing

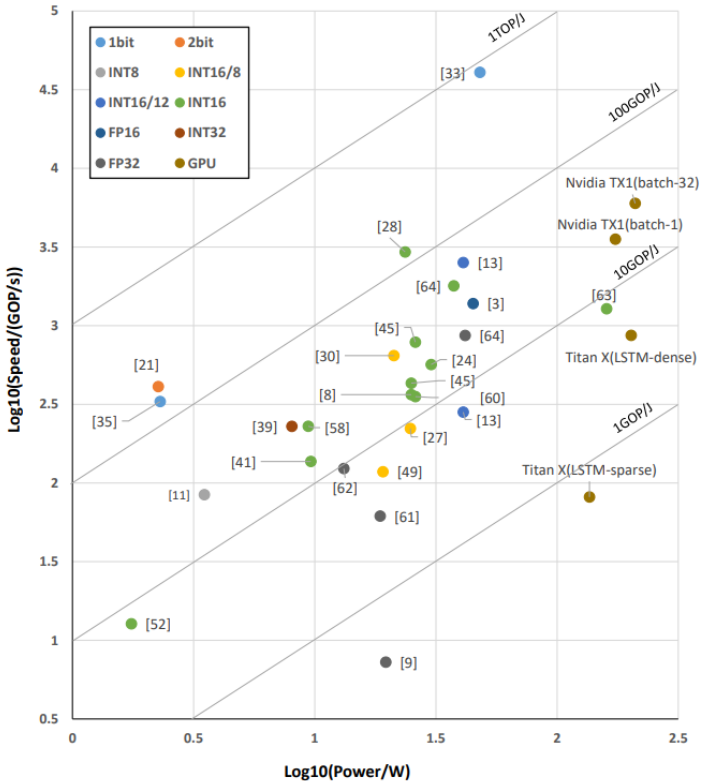
# DEEP LEARNING INFERENCE ON EMBEDDED DEVICES

- Increasing number of novel DL models proposed every year
- Increasing size and complexity (generally)




An Analysis of Deep Neural Network Models for Practical Applications A. Canziani, A. Paszke, E. Culurciello, 2016

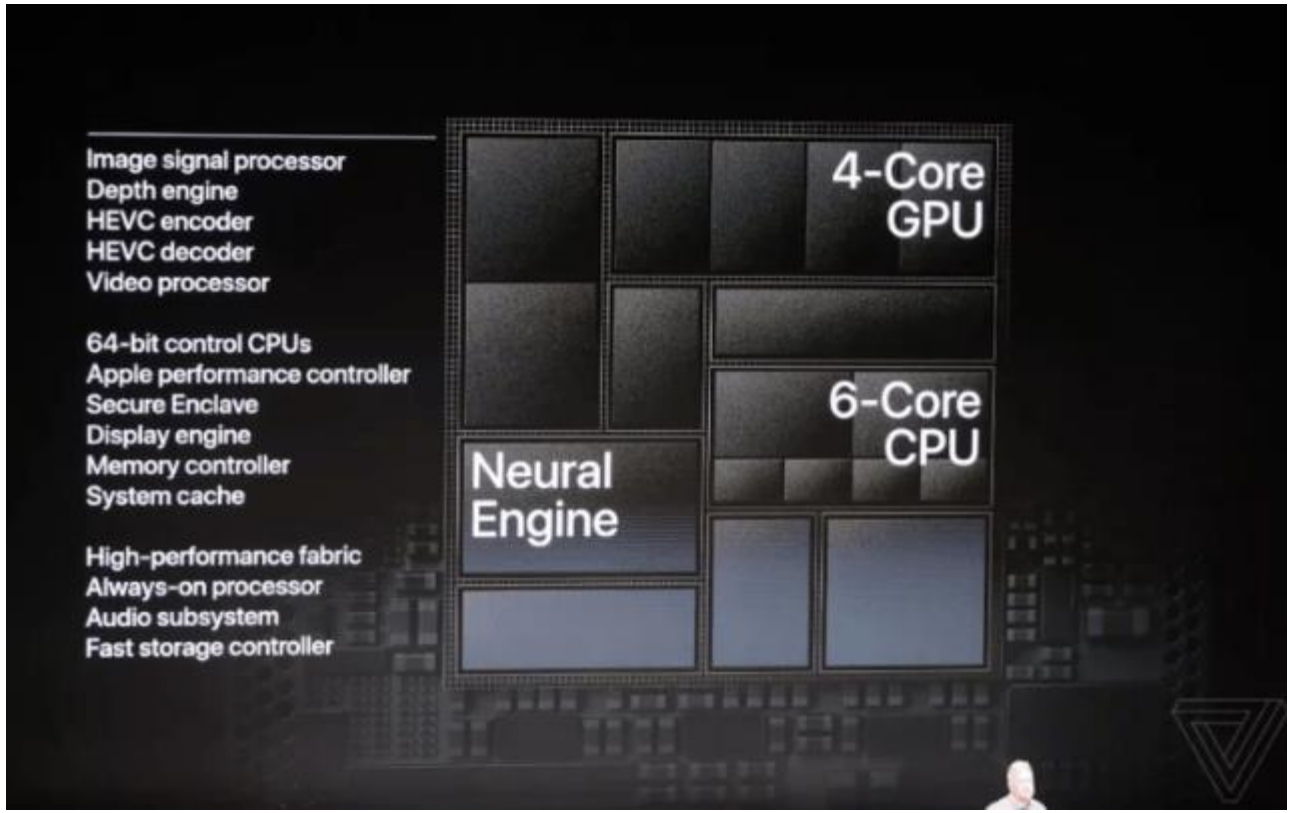
- Increasing number of DL-supporting computing architectures
- Increasing complexity, parallelism and heterogeneity




Guo et al. - https://arxiv.org/pdf/1712.08934.pdf

# DEEP LEARNING INFERENCE ON EMBEDDED DEVICES

| Designer | NPU |
|---|---|
| Amazon | AWS Inferentia |
| Alibaba | Ali-NPU |
| Baidu | Kunlun |
| Bitmain | Sophon |
| Cambricon | MLU |
| Google | TPU |
| Graphcore | IPU |
| Groq | |
| Intel | NNP, Myriad, EyeQ |
| Nvidia | NVDLA |
| Huawei | Ascend |
| Apple | Neural Engine |
| Samsung | Neural Processing Unit (NPU) |



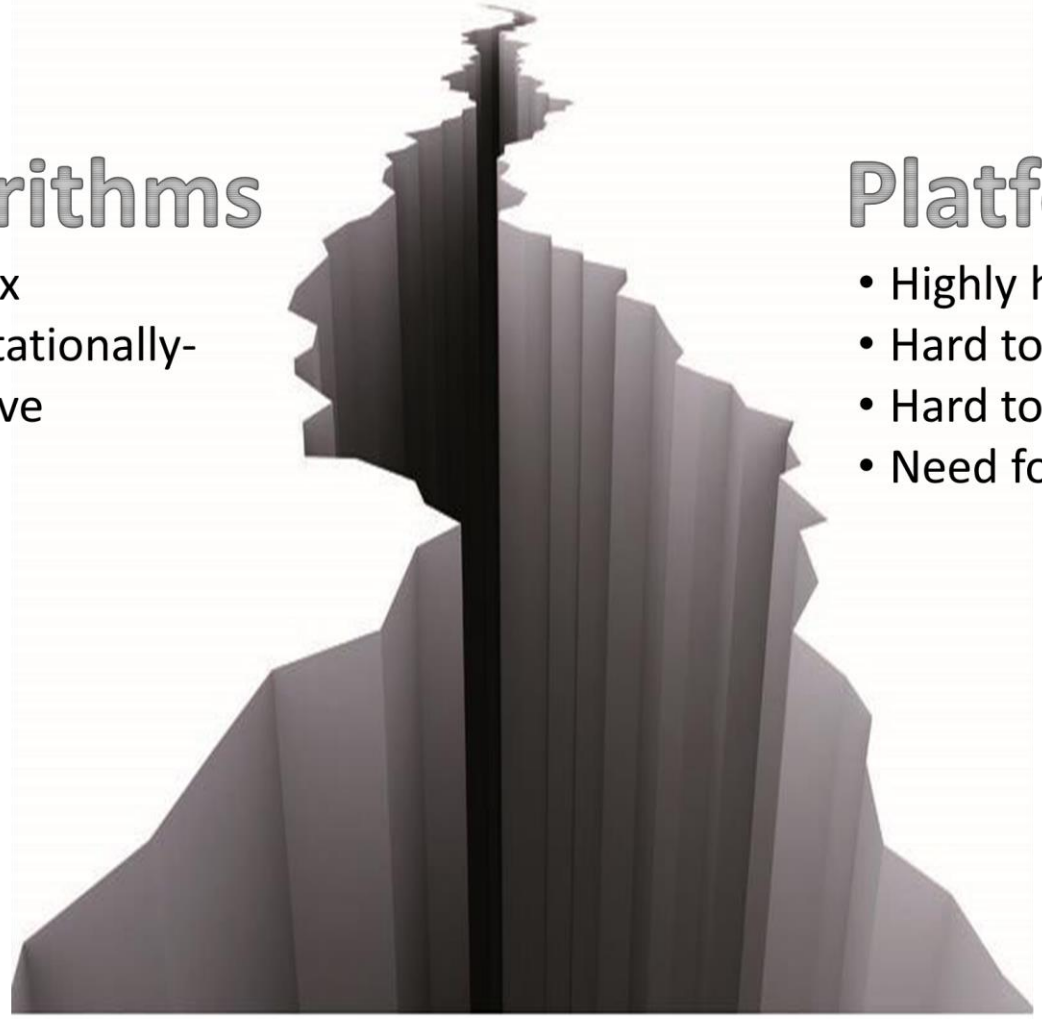



https://en.wikichip.org/wiki/neural_processor

# THE DL DICHOTOMY

## Algorithms

- Complex
- Computationally-expensive

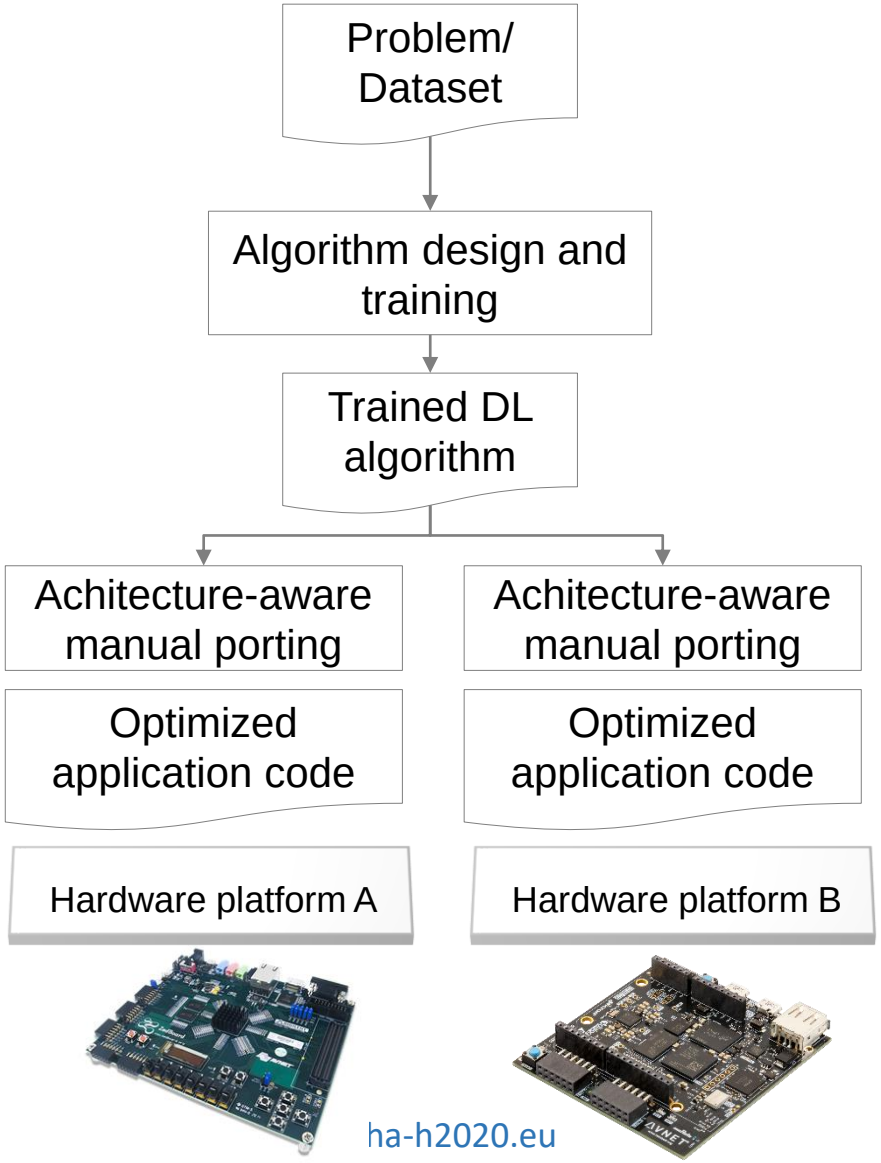


## Platforms

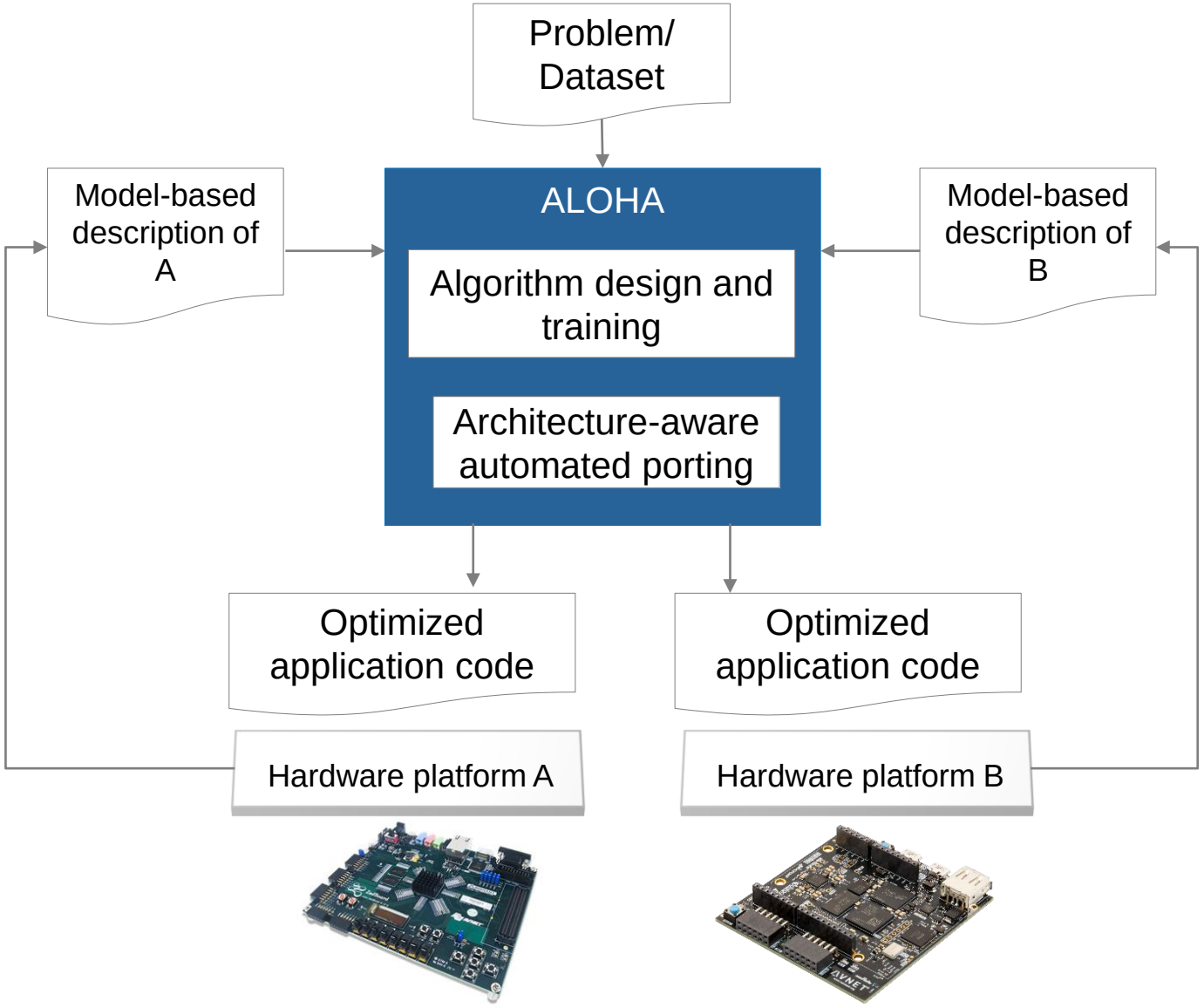
- Highly heterogeneous and parallel
- Hard to program
- Hard to manage at runtime
- Need for low-power

# ALOHA APPROACH

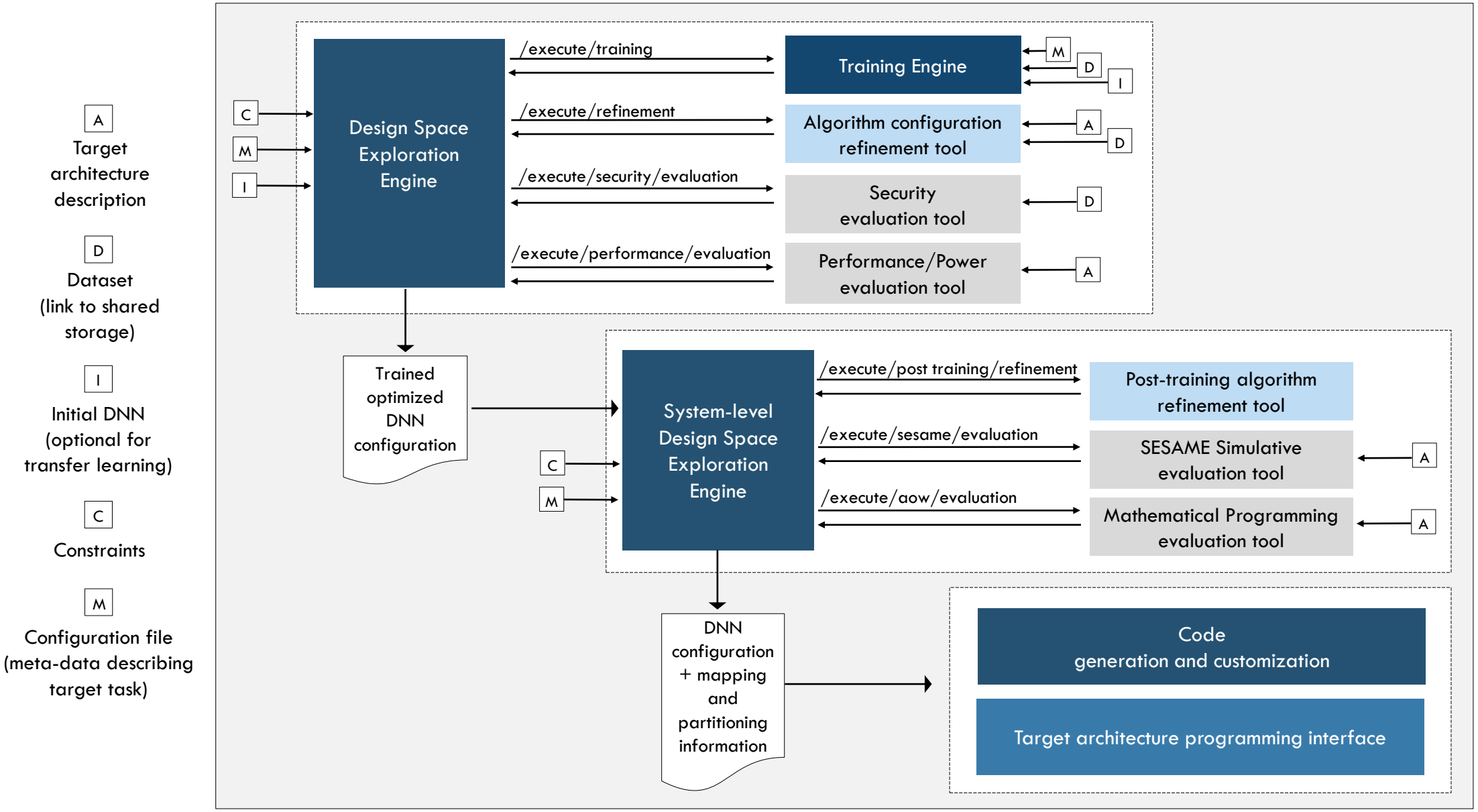
## Traditional flow



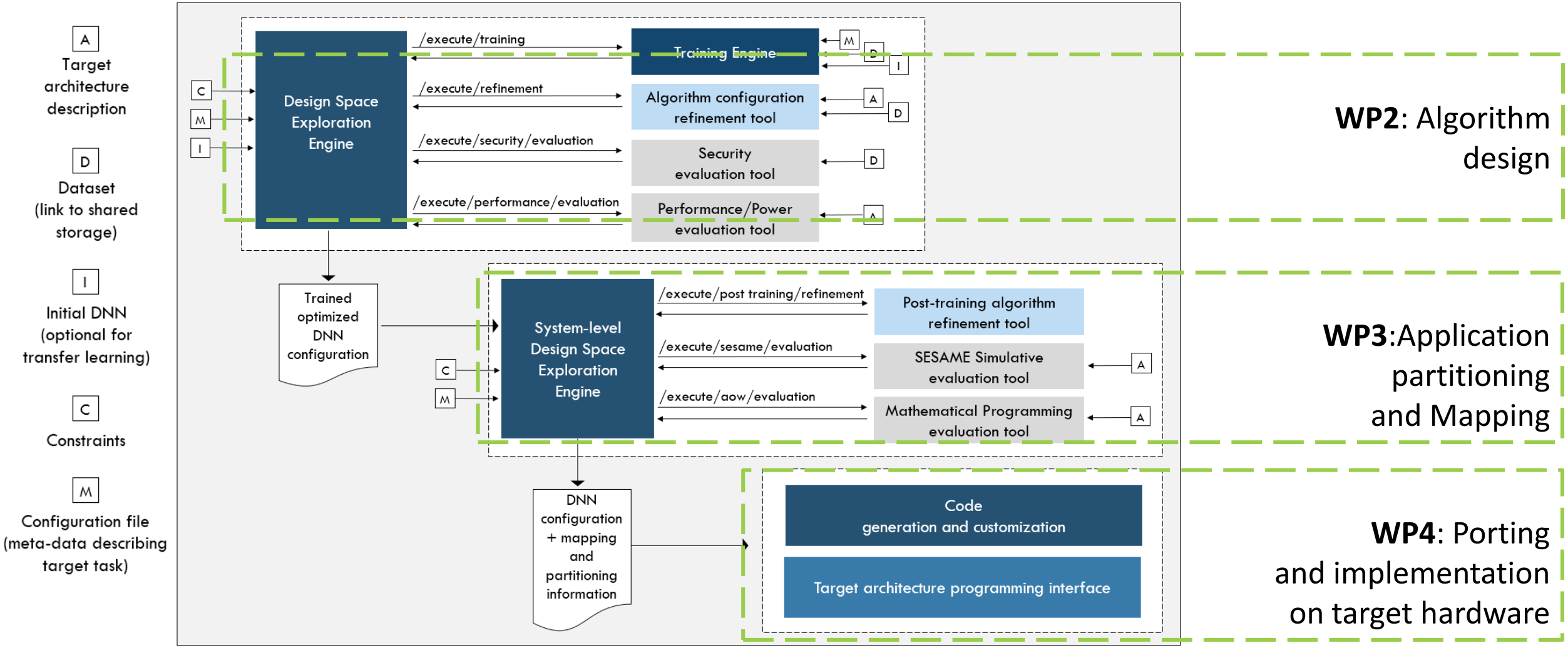

## ALOHA tool flow



# ALOHA TOOLFLOW ARCHITECTURE

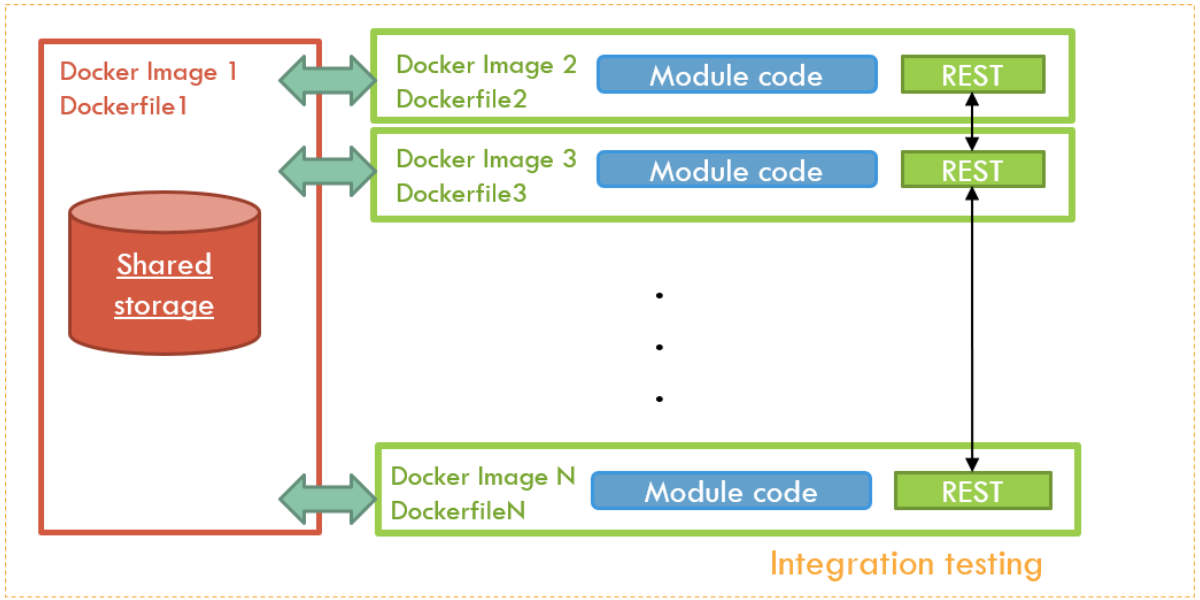


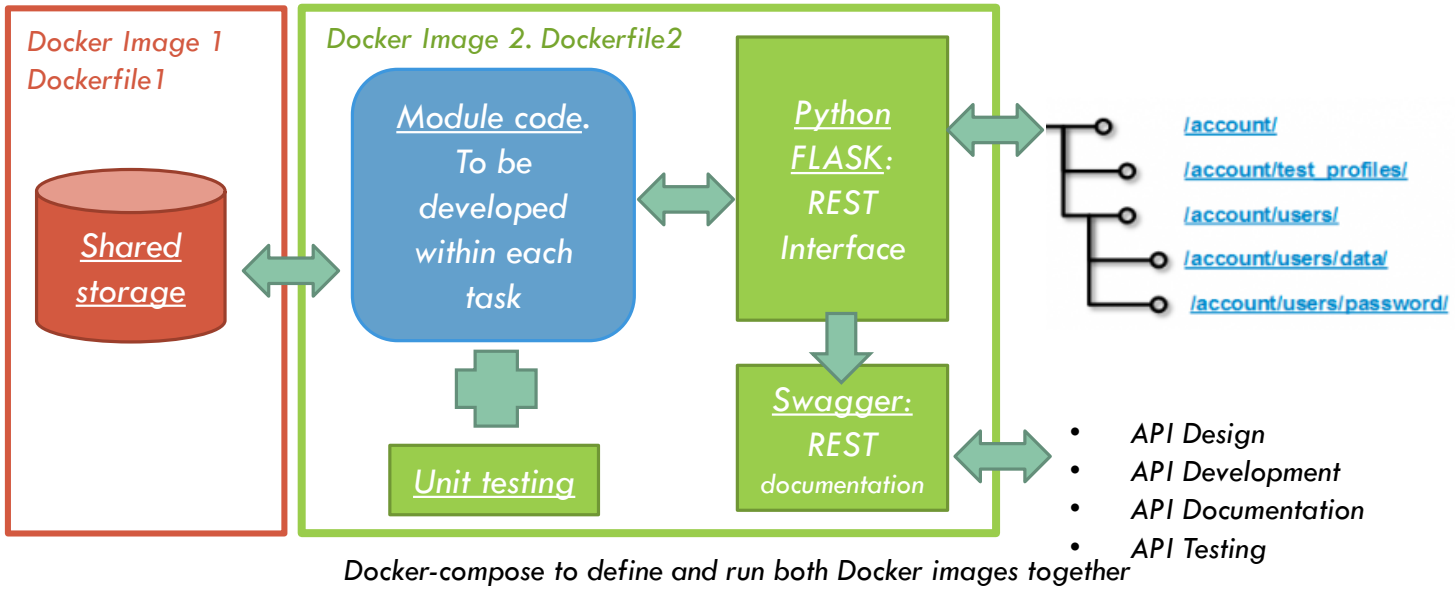
# ALOHA TOOLFLOW ARCHITECTURE



# ALOHA TOOLFLOW INTEGRATION METHODOLOGY

- Components communicate through REST APIs
- Independent containers
- Modularity
- Agile development methodology



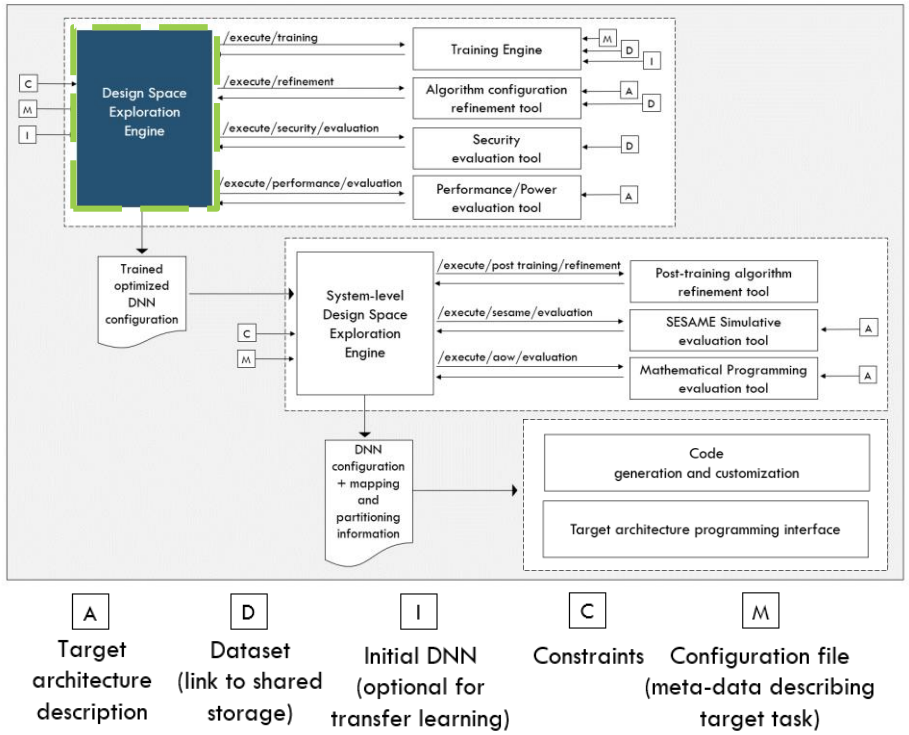

ONNX

www.aloha-h2020.eu

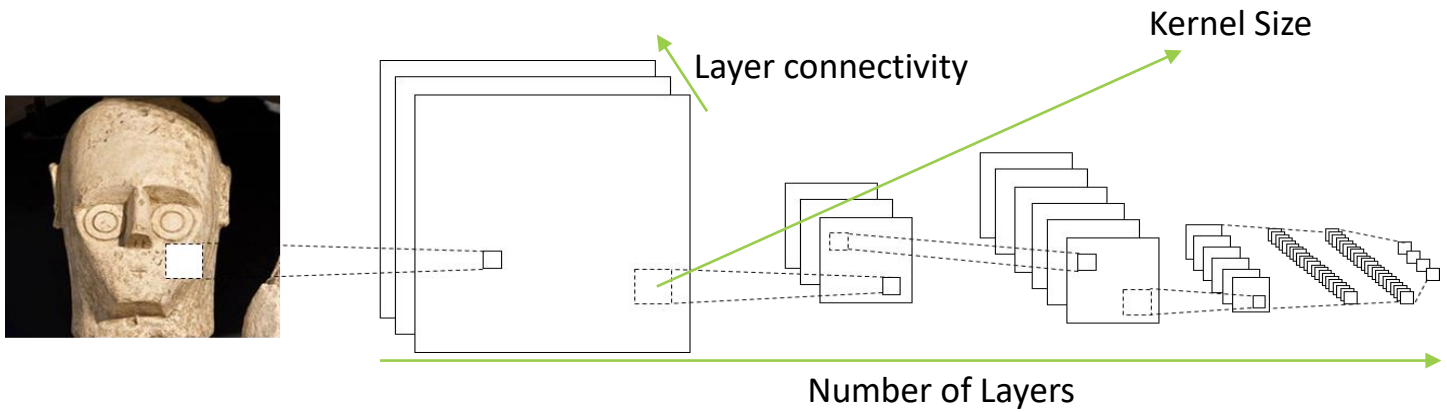
# FIRST STEP: JOIN THE USER COMMUNITY

https://www.aloha-h2020.eu/project/get-involved

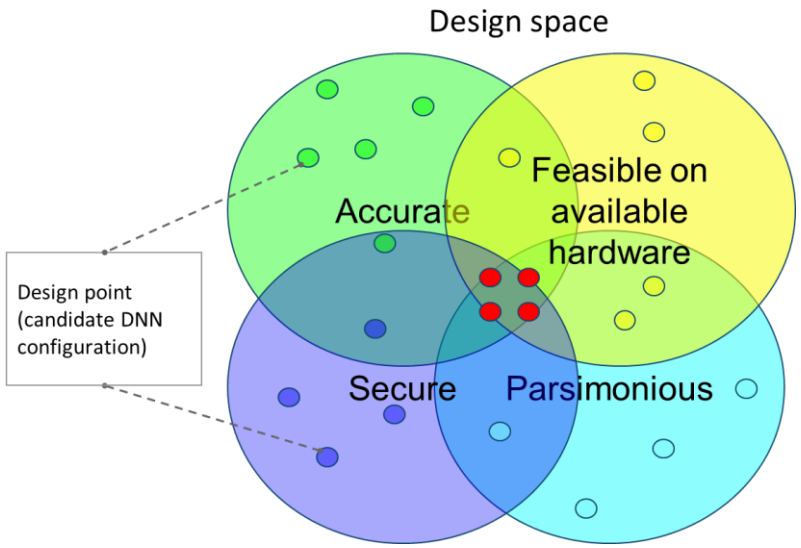
# ALOHA TOOLFLOW: THE DSE ENGINE




- Explore algorithm configuration degrees of freedom

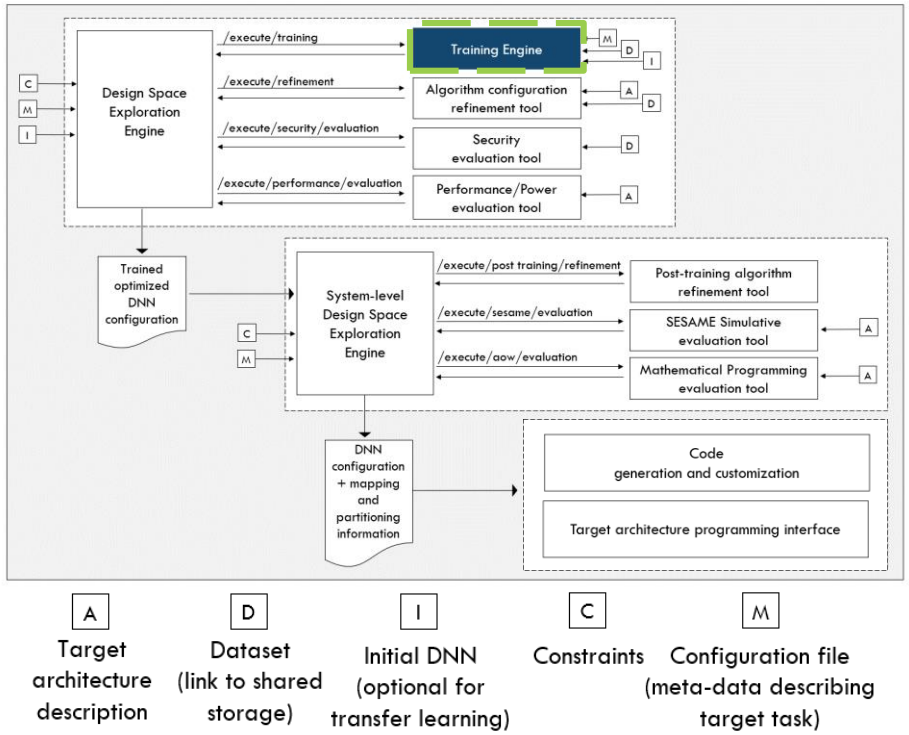


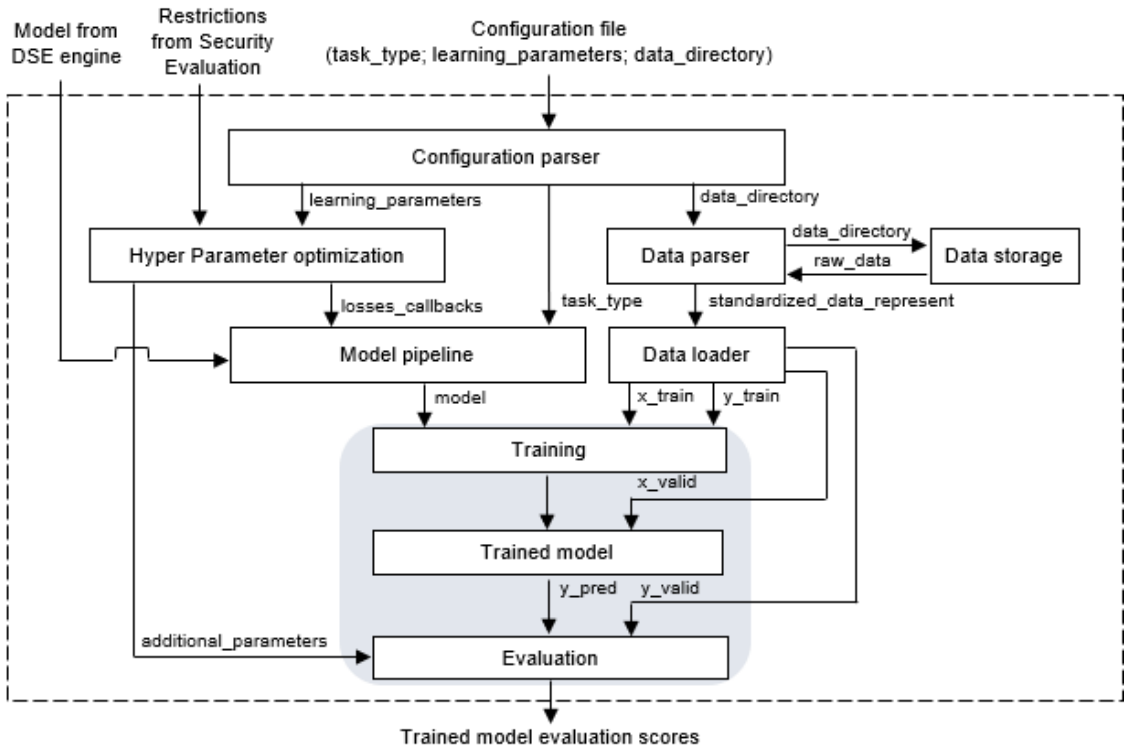

- Multi-objective exploration
- Genetic algorithm for design space surfing




www.aloha-h2020.eu

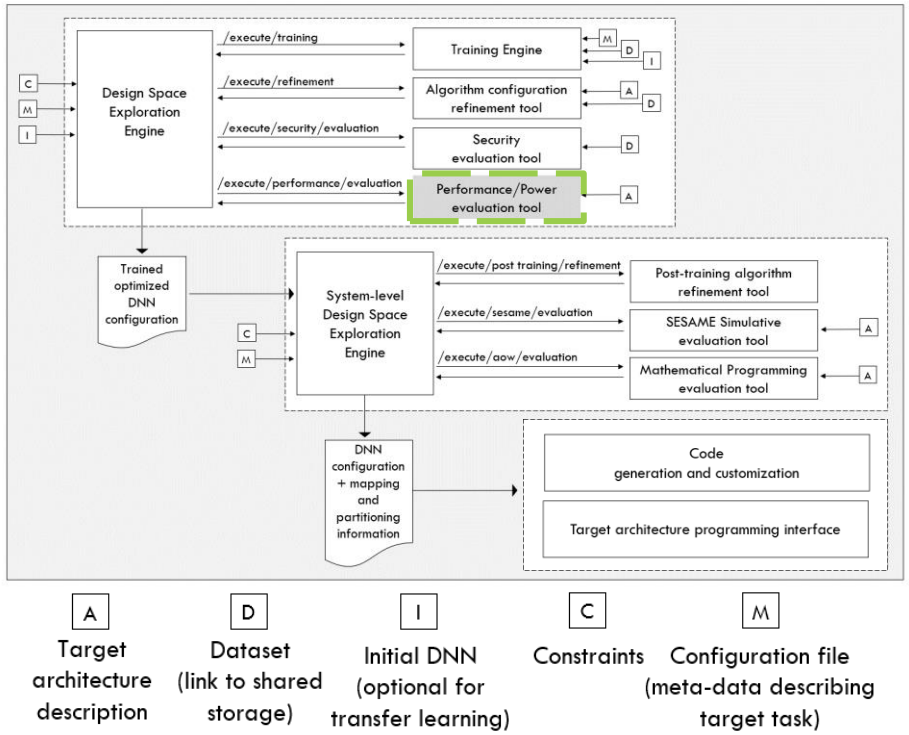
# ALOHA TOOLFLOW: THE TRAINING ENGINE

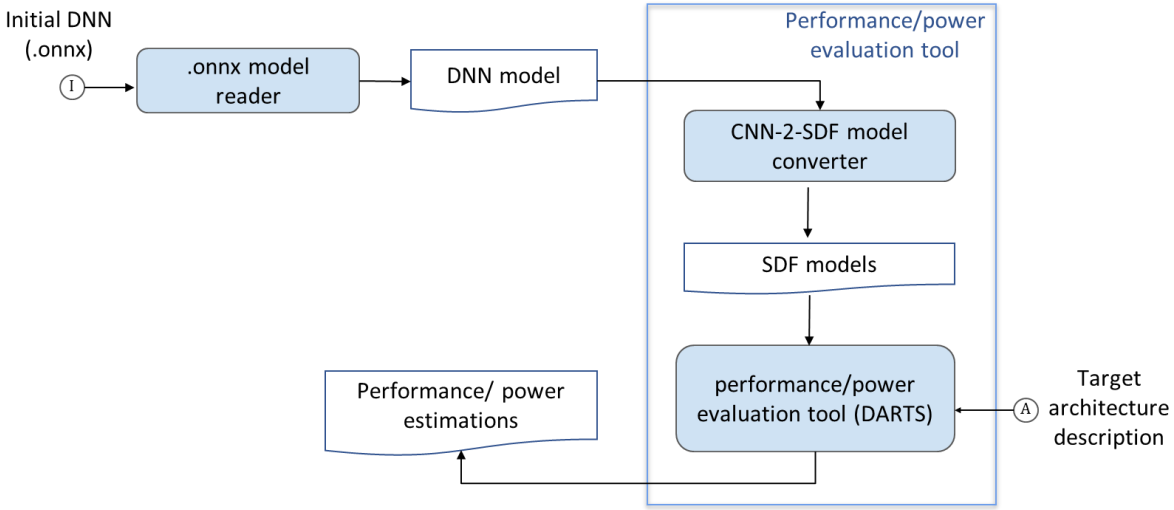


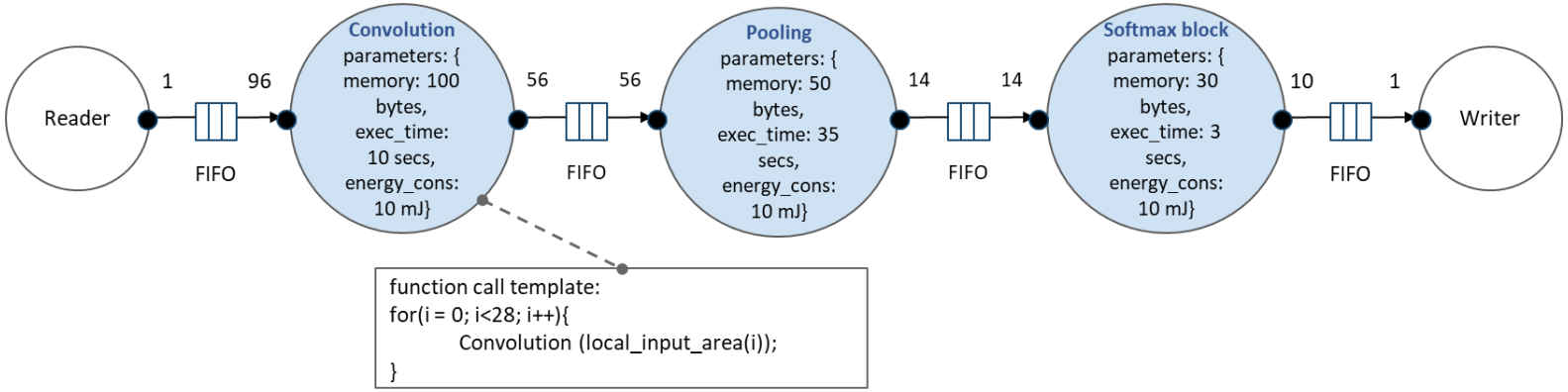




- Transfer learning enabled (reuse of the model in different domains with reduced dataset)
- Local hyperparameter exploration
- Flexible data parsing (multiple input formats)
- Flexible use-case configuration (multiple AI tasks: classification, detection, tracking etc…)

# ALOHA TOOLFLOW: PERFORMANCE/POWER EVALUATION TOOL

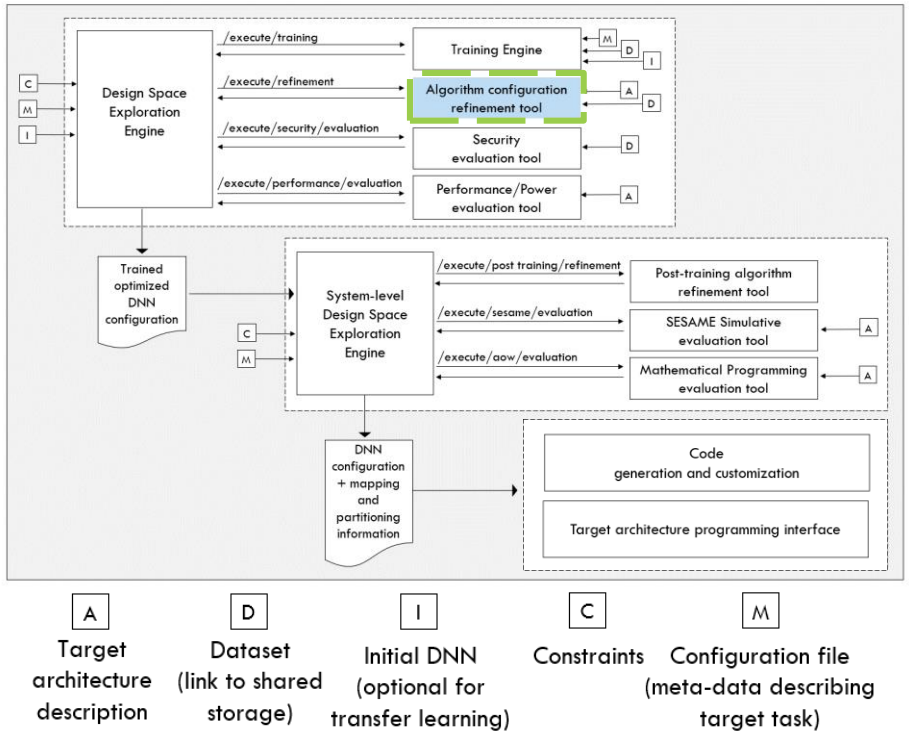


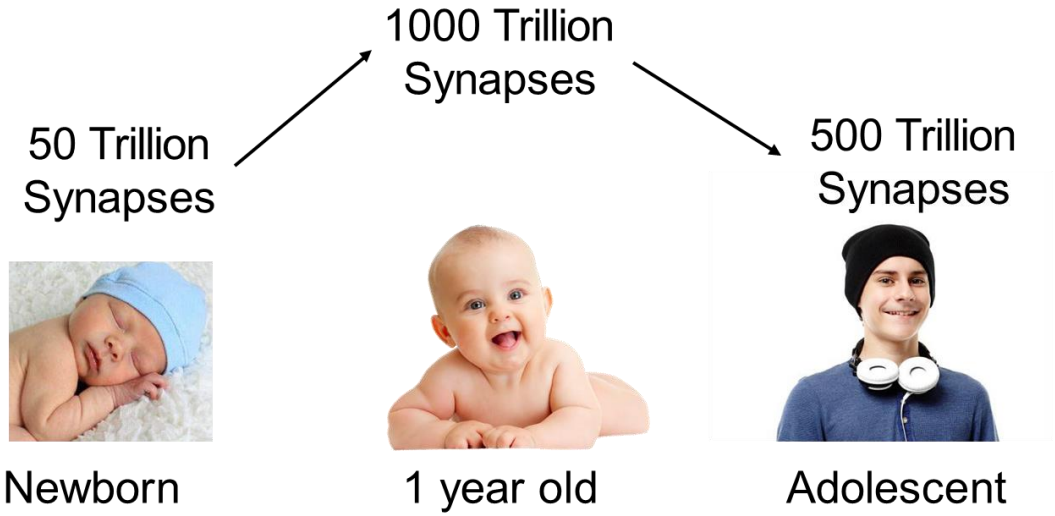




- Convert DNN model to analyzable application model (CSDF)
- Architecture model associates exec. time and energy to CSDF actors
- Analyze/transform/evaluate

# ALOHA TOOLFLOW: REFINEMENT FOR PARSIMONIOUS INFERENCE



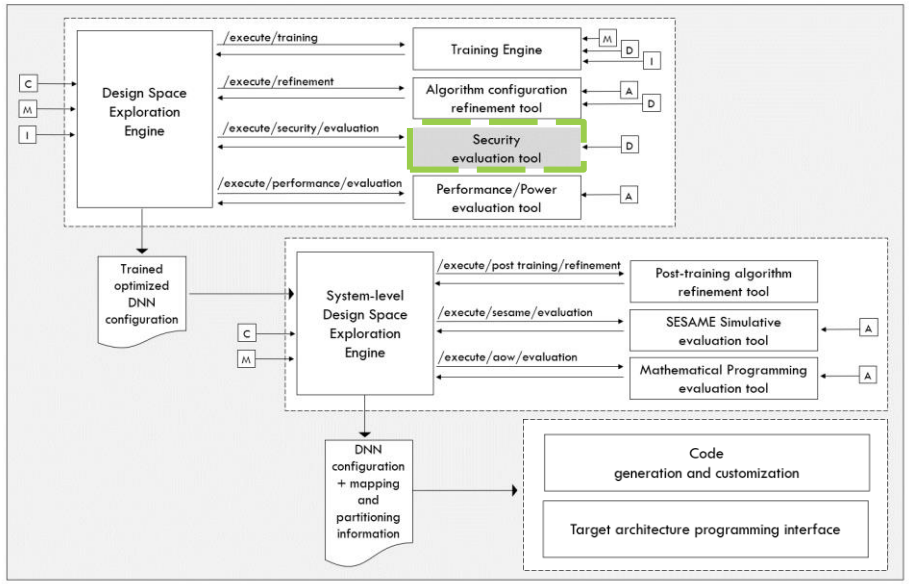




Christopher A Walsh. Peter Huttenlocher (1931-2013). Nature, 502(7470):172–172, 2013.

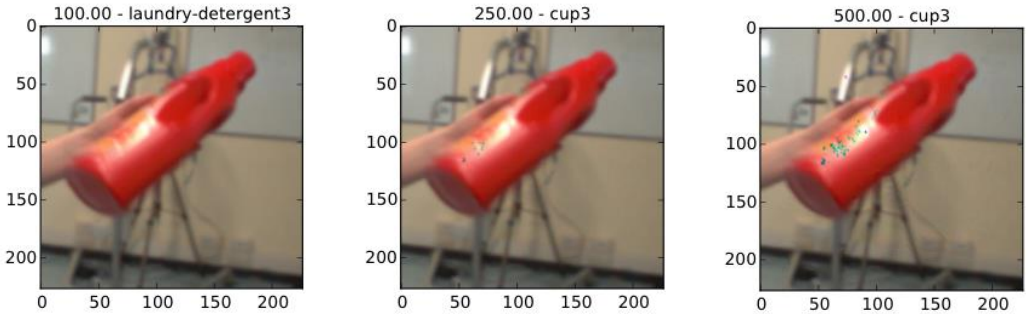
- QUANTIZATION
- PRUNING

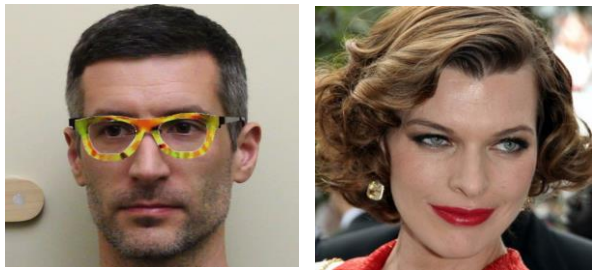
- Reduce computational workload
- Reduce memory footprint
- Reduce IO bandwidth requirements

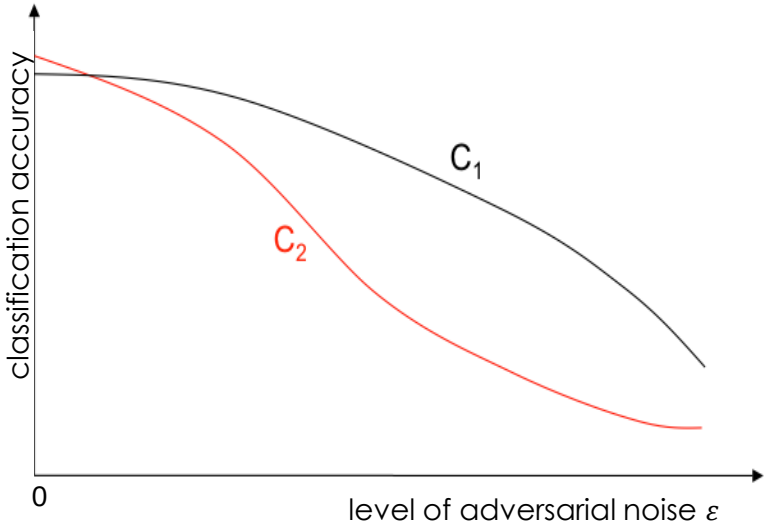
# ALOHA TOOLFLOW: SECURITY EVALUATION




| A | D | I | C | M |
|---|---|---|---|---|
| Target architecture description | Dataset (link to shared storage) | Initial DNN (optional for transfer learning) | Constraints | Configuration file (meta-data describing target task) |

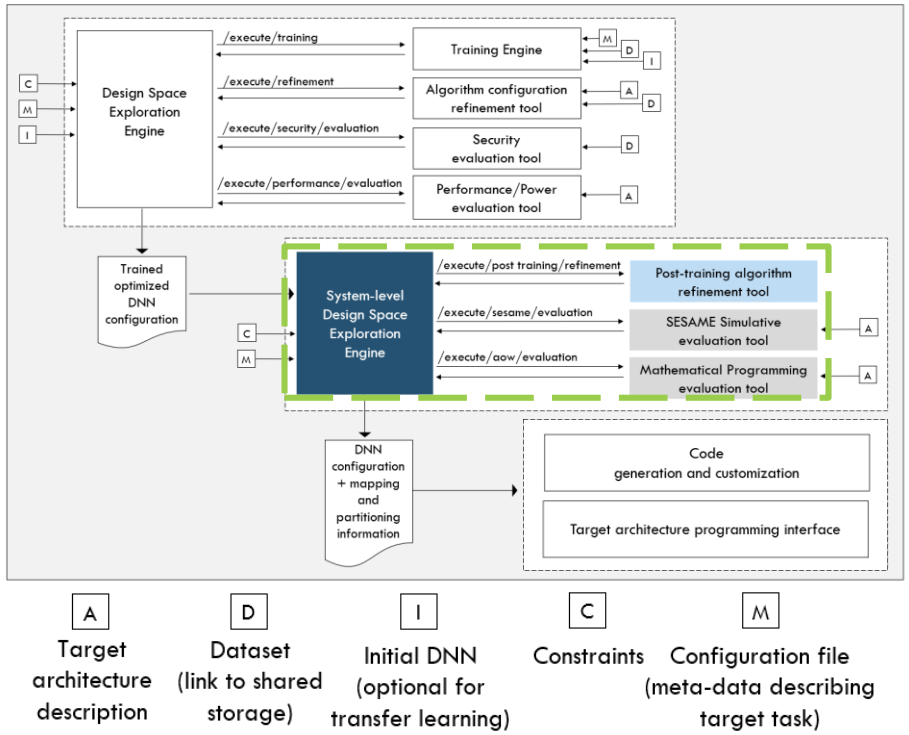


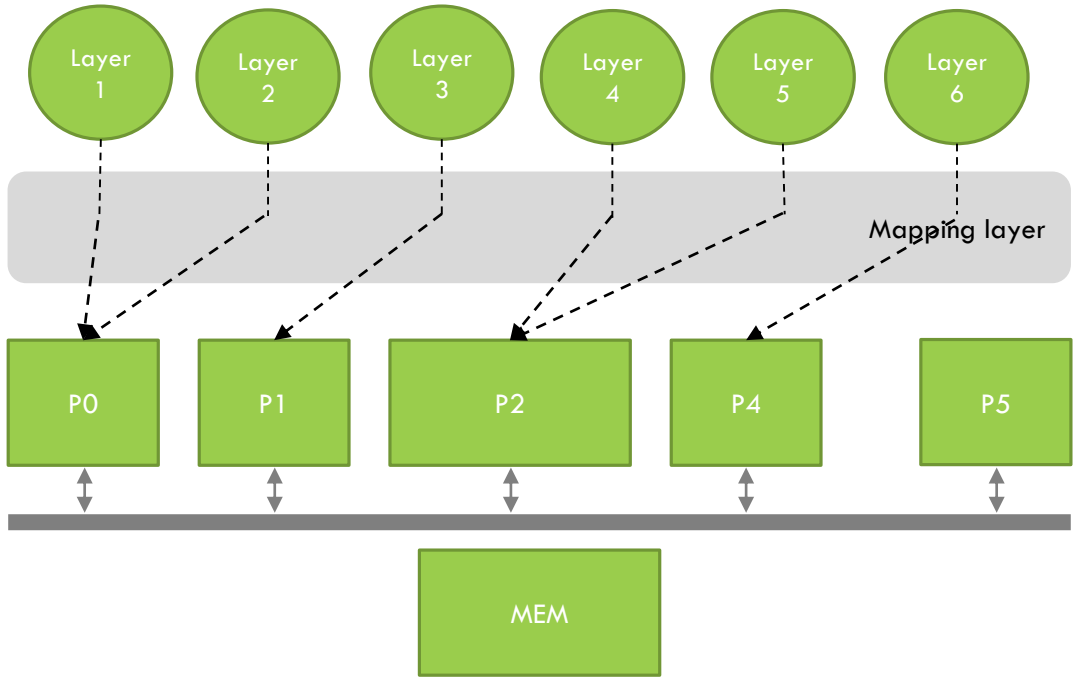






www.aloha-h2020.eu

# ALOHA TOOLFLOW: SYSTEM LEVEL DESIGN

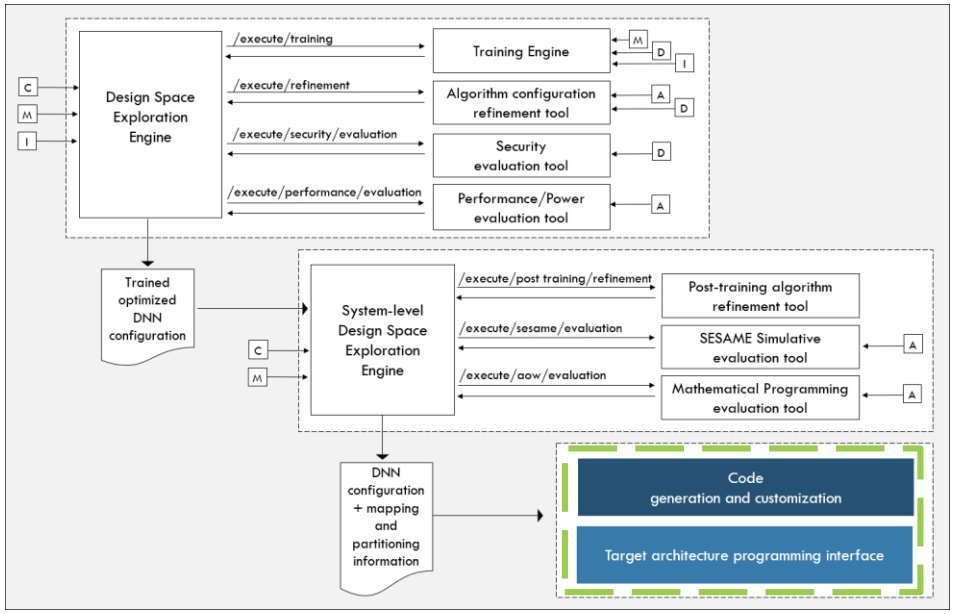



- DSE instrument: Genetic Algorithm
- Design point evaluation
  a) High-level simulative: SESAME by UvA
  b) Mathematical Programming and analytic modeling: Architectural Optimization Workbench by IBM
- Post-training parsimonious inference enabled: runtime workload reduction

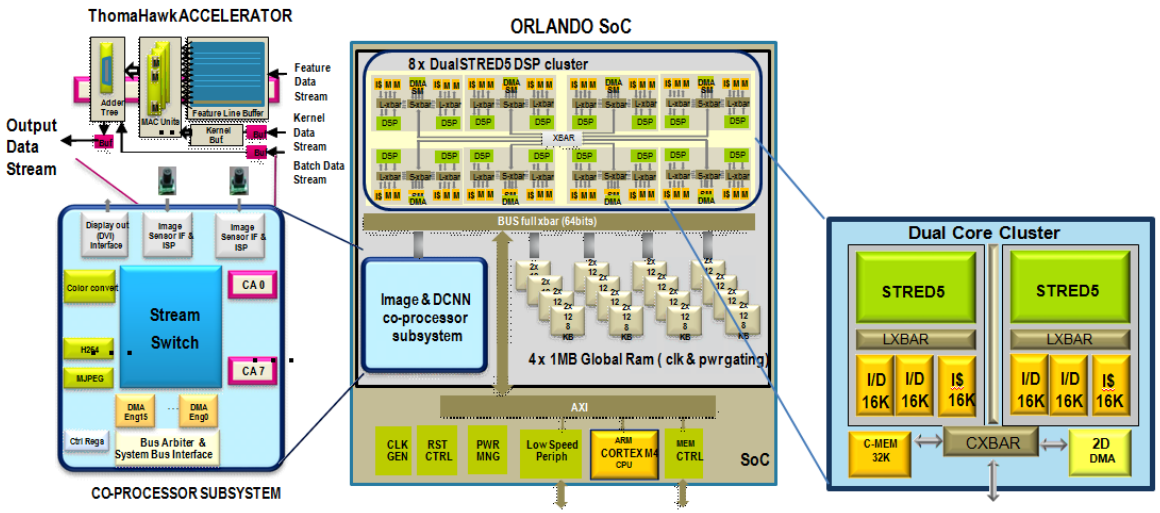
www.aloha-h2020.eu

# ALOHA TOOLFLOW: TARGET COMPUTING PLATFORMS

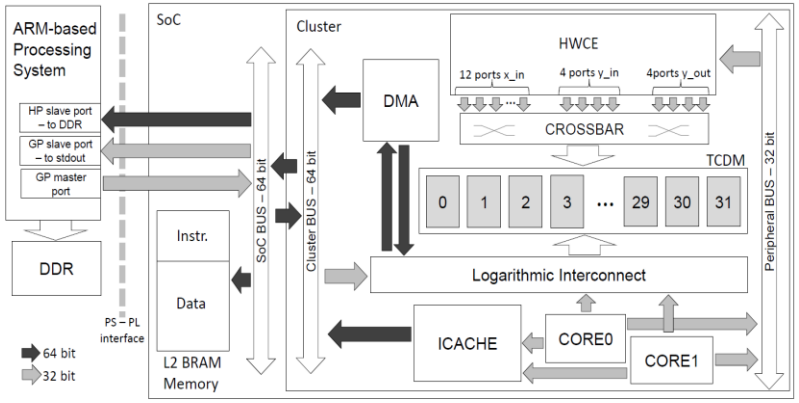



| A | D | I | C | M |
|---|---|---|---|---|
| Target architecture description | Dataset (link to shared storage) | Initial DNN (optional for transfer learning) | Constraints | Configuration file (meta-data describing target task) |

- STMicro's Orlando




- NEURAghe




www.aloha-h2020.eu

# NEURAGHE: CNN INFERENCE ACCELERATOR FROM SARDINIA

nuraghe (Sardinian pronunciation: [nuˈɾaɣe], Italian: [nuˈraːge]

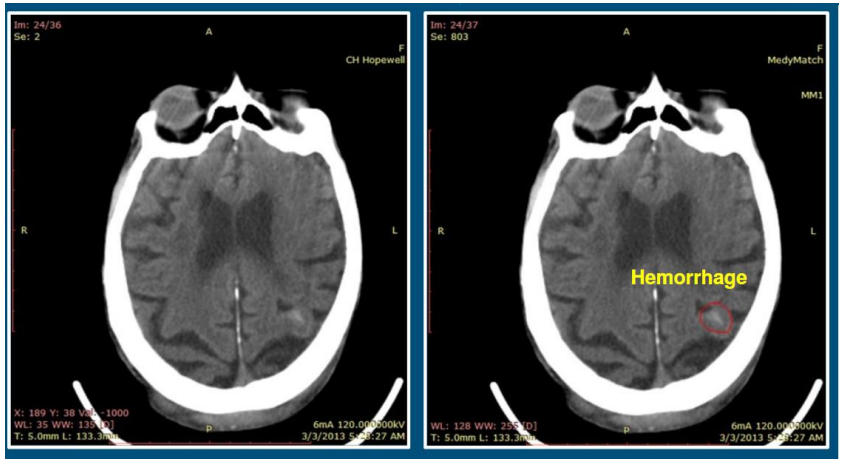
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

P. Meloni, G. Deriu, F. Conti, I. Loi, L. Raffo and L. Benini, "A high-efficiency runtime reconfigurable IP for CNN acceleration on a mid-range all-programmable SoC", *ReConFig* 2016
Paolo Meloni, Alessandro Capotondi, Gianfranco Deriu, Michele Brian, Francesco Conti, Davide Rossi, Luigi Raffo, Luca Benini, "NEURAghe: Exploiting CPU-FPGA Synergies for Efficient and Flexible CNN Inference Acceleration on Zynq SoCs", TRETS 2018

www.aloha-h2020.eu

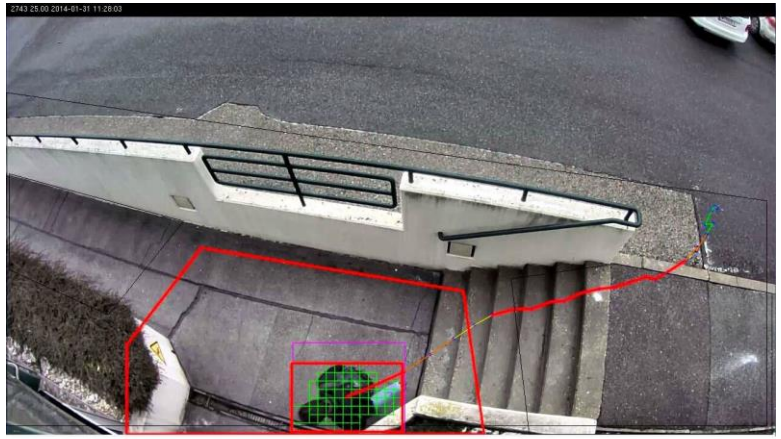
# ALOHA USE-CASES

Demonstration on three application use-cases




Cost- and power-effective
medical decision assistant

Command Recognition in
Smart Industry Applications



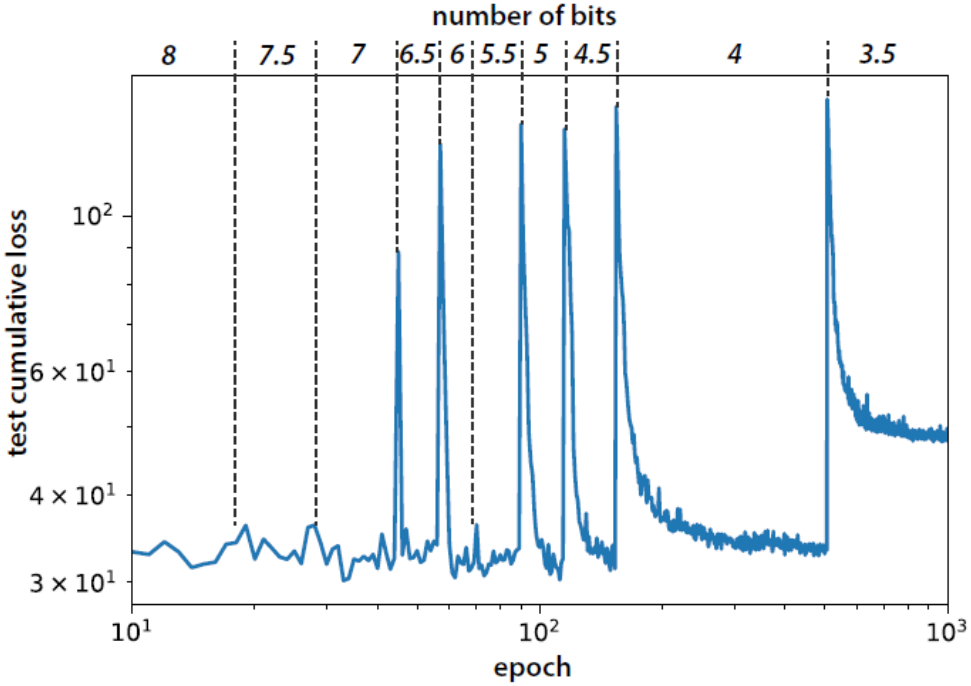
Surveillance
of Critical Infrastructures

# ALOHA – a glimpse on results

## Architecture-aware algorithm selection

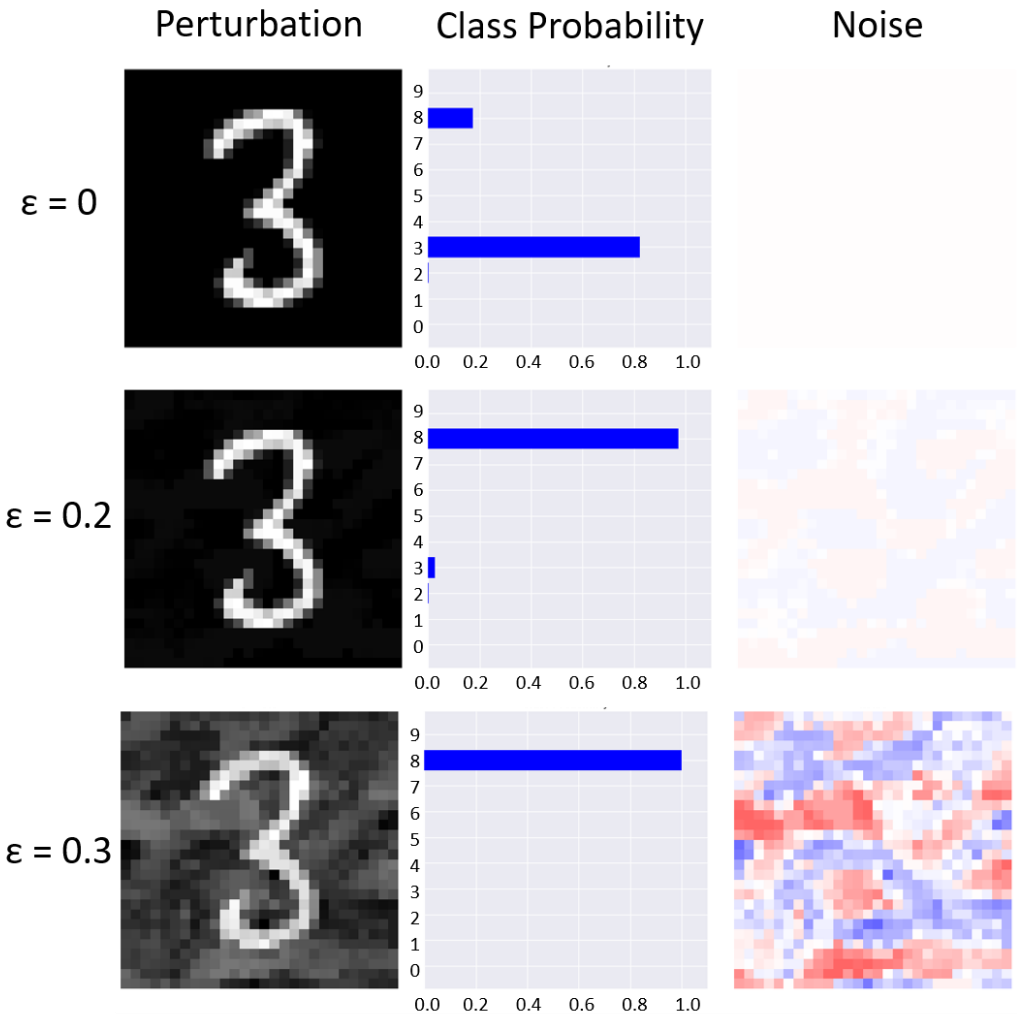
| Benchmark | Performance (GOps/s) | Accuracy (%) |
|---|---|---|
| VGG-16 | 172.67 | 88.4 |
| NEURAghe-aware VGG-16 (over) | 182.43 | 89.6 |
| NEURAghe-aware VGG-16 (under) | 183.75 | 79.7 |

## Architecture-aware quantization

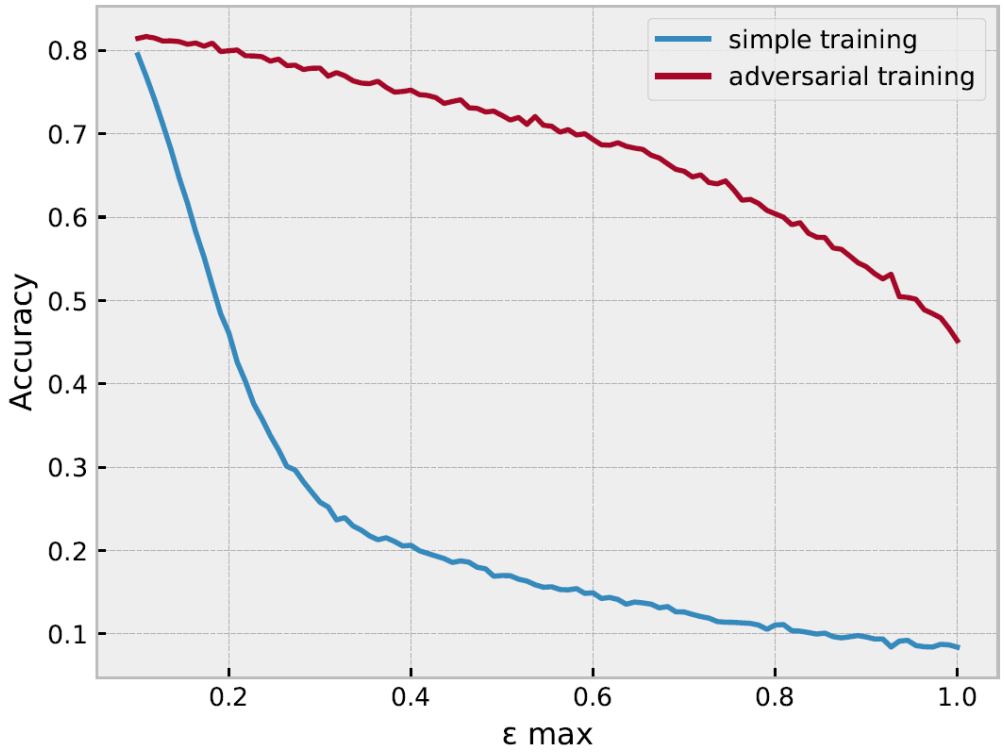



| | LOSA Single 4x4 | ARRUBIU Dual 2x4 | SABINA Single 2x2 | LOSA Single 4x4 | ARRUBIU Dual 2x4 | SABINA Single 2x2 | BANZOS Single 1x1 |
|---|---|---|---|---|---|---|---|
| Device | Z-7045 | Z-7045 | Z-7020 | Z-7045 | Z-7045 | Z-7020 | Z-7007s |
| DSP [#]; Freq [MHz] | 864; 140 | 864; 140 | 216; 80 | 864; 140 | 864; 140 | 216; 80 | 54; 80 |
| Benchmark Net | ResNet 18 | ResNet 18 | ResNet 18 | VGG-16 | VGG-16 | VGG-16 | SqueezeNet |
| GOps/s (16 bit)<br>GOps/s per Watt (16 bit)<br>GOps/s per k$ (16 bit) | 61.91<br>6.19<br>25 | 103<br>10.3<br>41 | 18.34<br>5.24<br>41 | 172.67<br>17.26<br>69 | 184<br>18.4<br>74 | 29.32<br>8.37<br>65 | 3.84<br>1.53<br>43 |
| GOps/s (8 bit) | 111.12 | - | 33.07 | 335.09 | - | 56.51 | - |

# ALOHA – a glimpse on results



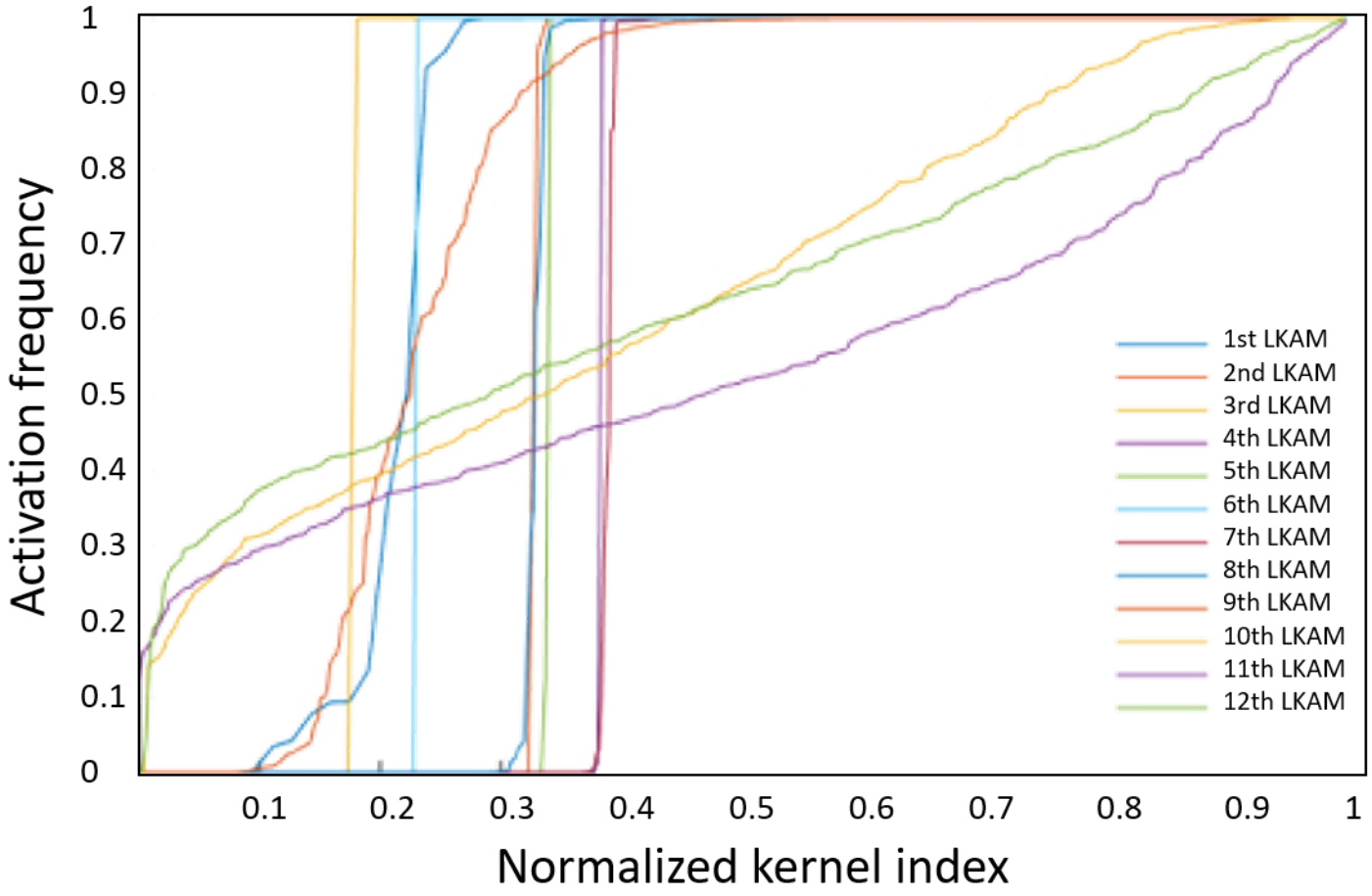

Security evaluation



# ALOHA – a glimpse on results

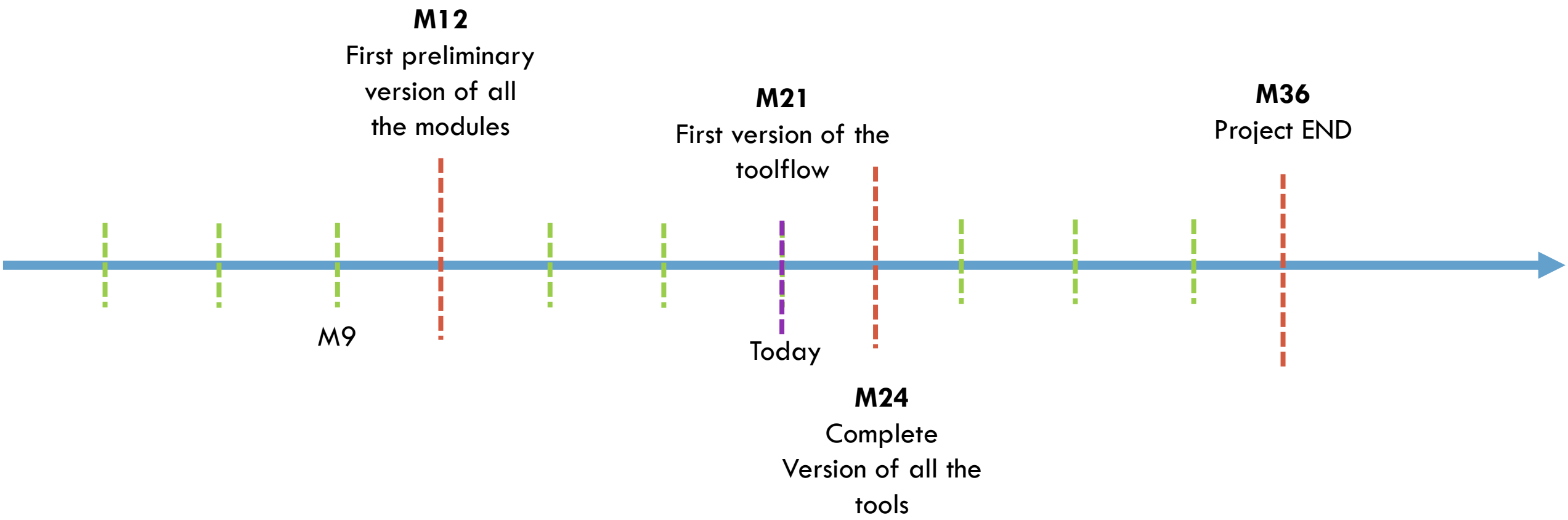
Post-training parsimonious inference

Kernel activity profile –VGG-16




48.31% workload reduction
66.14% average kernel utilization

# ALOHA TIMELINE



alona.net

aloha.net

Create New Project

ALOHA

## Algorithm configuration

**Project two** Ready

*The second example project*

Start

**Title** Finished

*Description*

Get results

## System level configuration

**Example project 1** Ready

*This is the first project*

Start

## Architectural configuration

**Another project** Ready

*a second project*

Start

## Finished

ITA 10:24 07/02/2019

alona.net

aloha.net

Create New Project

**Algorithm configuration**

**Project two** Finished

*The second example project*

Get results

**Title** Finished

*Description*

Get results

**System level configuratio**

**Example project 1**

*This is the first project*

Start

# Create Project

Project Title

Title

Project Description

Description

Architecture Specification File

Scegli file Nessun file selezionato

+ Constraints

## Power consumption

Power value 1

Power priority 1

## Security

Security value 1

Security priority 1

## Execution time

Execution time value 1

Execution time priority 1

## Accuracy

Accuracy time value 1

Accuracy priority 1

ITA 10:25 07/02/2019

alona.net

aloha.net

Create New Project

ALOHA

**Algorithm configuration**

**Project two** Finished

*The second example project*

Get results

**Title** Finished

*Description*

Get results

**System level configuration**

**Example project 1** Ready

*This is the first project*

Start

**Architectural configuration**

**Another project** Ready

*a second project*

Start

**Finished**

# Algorithm Configuration summary

## Training evaluation:

- Training accuracy value: **83.4**
- Training loss value: **0.7**
- Validation accuracy value: **82.2**
- Validation loss value: **0.9**
- Download traing plot

## Security:

- Security level: **medium**
- Download security plot

Title

## Power performance evaluation:

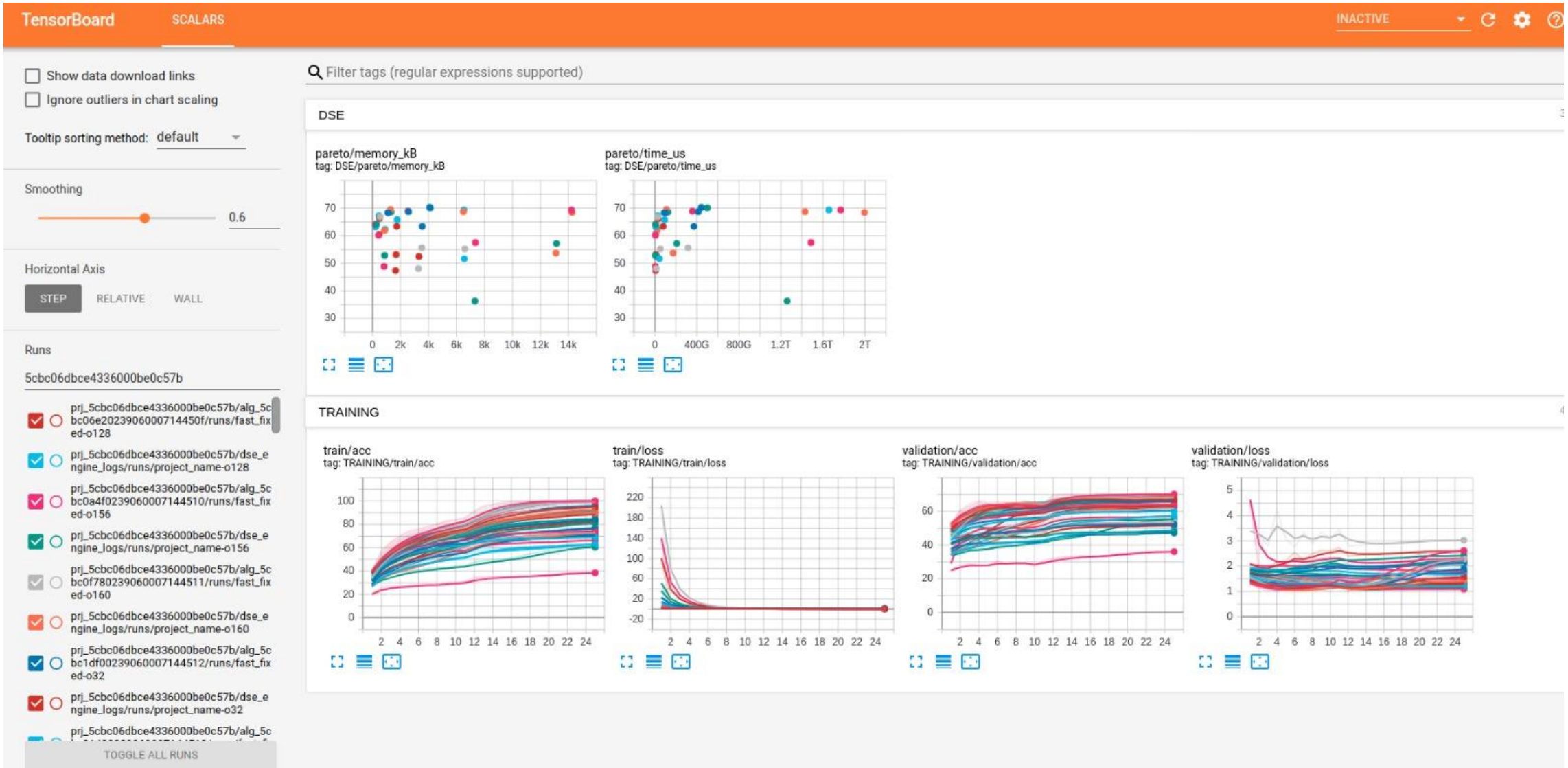
- Performance value: **4.41** seconds
- Energy value: **0.321285** Joule
- Minimum number of processors: **2**
- Memory value: **81248** B

## Refinement for parsimonious inference:

- Download Parsimonious inference plot

ITA 10:27 07/02/2019

# ALOHA – evaluation dashboard



# THANK YOU FOR YOUR ATTENTION!

www.aloha-h2020.eu

@ALOHA_H2020

ALOHA project

https://www.aloha-h2020.eu/project/get-involved

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 780788