ALOHA

A SOFTWARE FRAMEWORK FOR RUNTIME-ADAPTIVE AND SECURE DEEP LEARNING ON HETEROGENEOUS ARCHITECTURE

PAOLO MELONI

UNIVERSITÀ DEGLI STUDI DI CAGLIARI

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

life.augmented

UNIVERSITY OF CAGLIARI

UNIVERSITEIT VAN AMSTERDAM

Universiteit Leiden

University of Sassari

maxQ Artificial Intelligence

ETH Zürich

Reply santer

IBM

ca technologies

scch software competence center hagenberg

Pluribus One seeing one in many

IRIDA LABS

PKE PKE Electronics AG

Coordinator

Giuseppe Desoli - STMicroelectronics

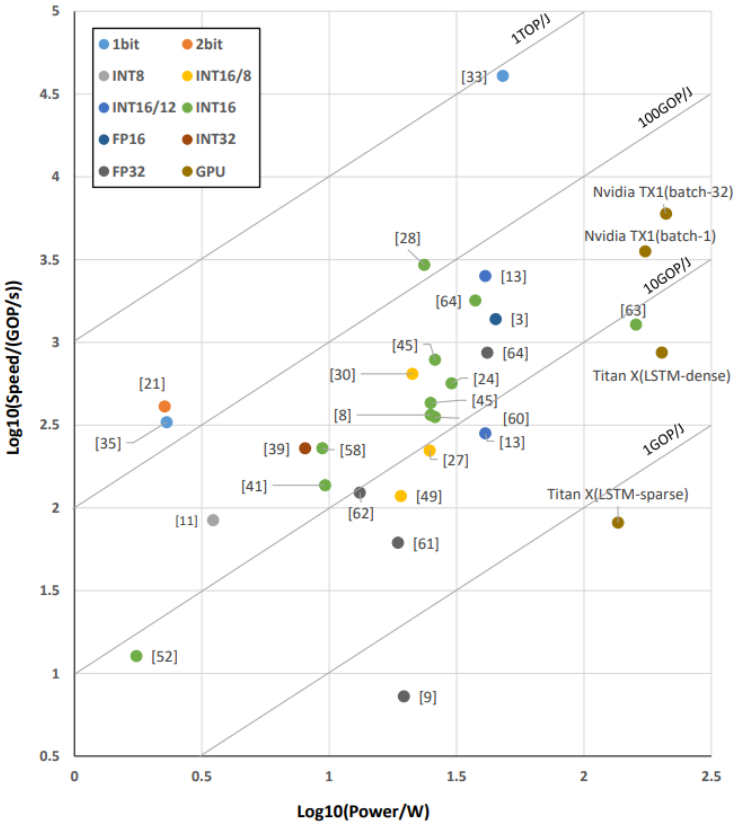
Scientific Coordinator

Paolo Meloni - University of Cagliari, EOLAB

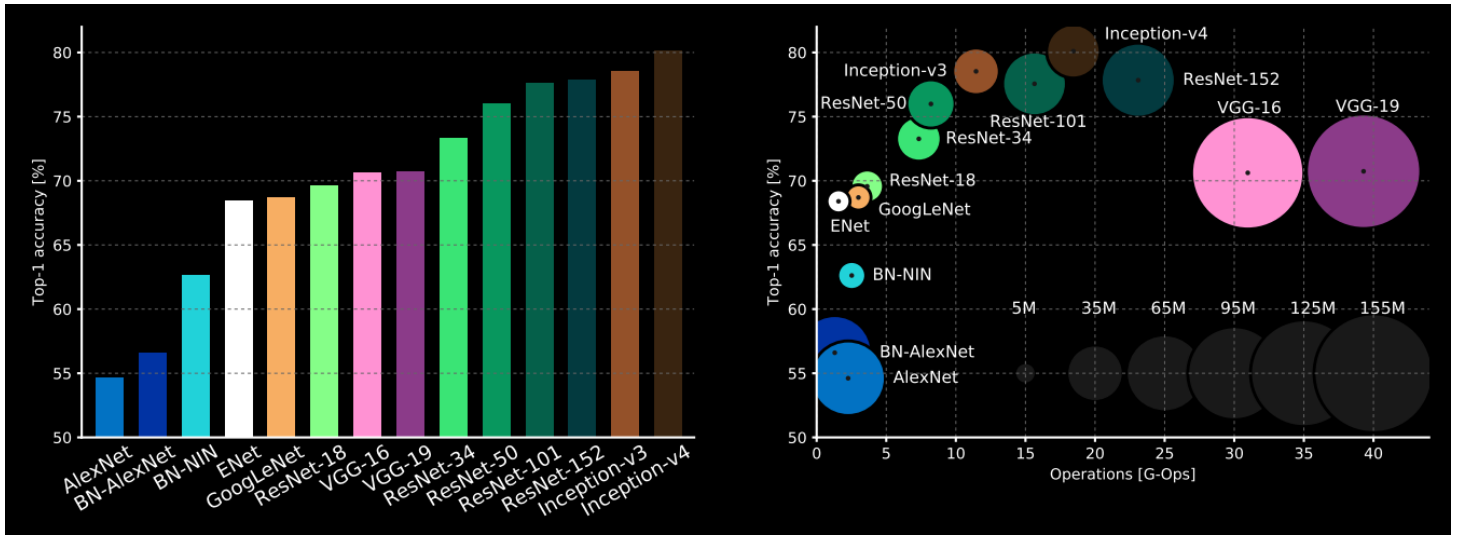
www.aloha-h2020.eu

The landscape of DL on embedded Cognitive edge computing

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



www.aloha-h2020.eu

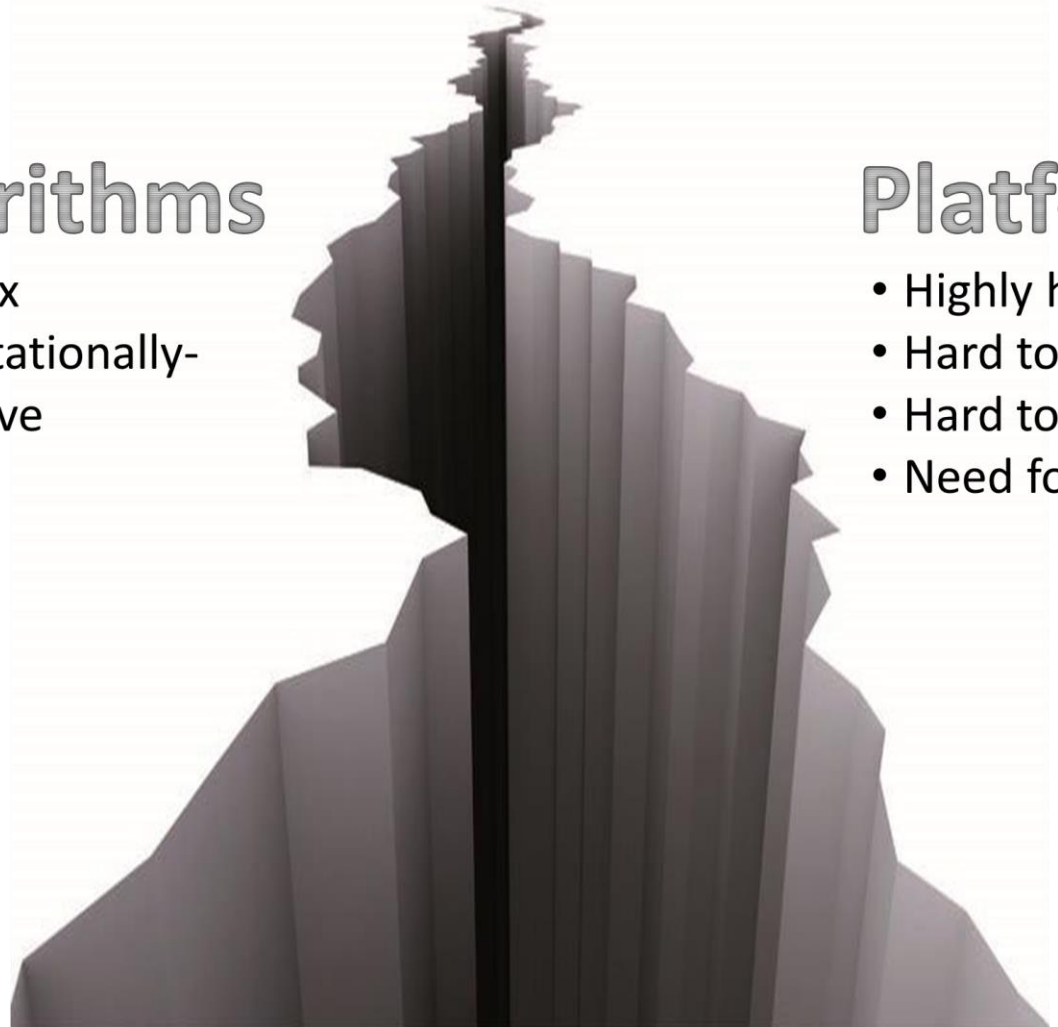


An Analysis of Deep Neural Network Models for Practical Applications A. Canziani, A. Paszke, E. Culurciello, 2016

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

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

The DL dichotomy



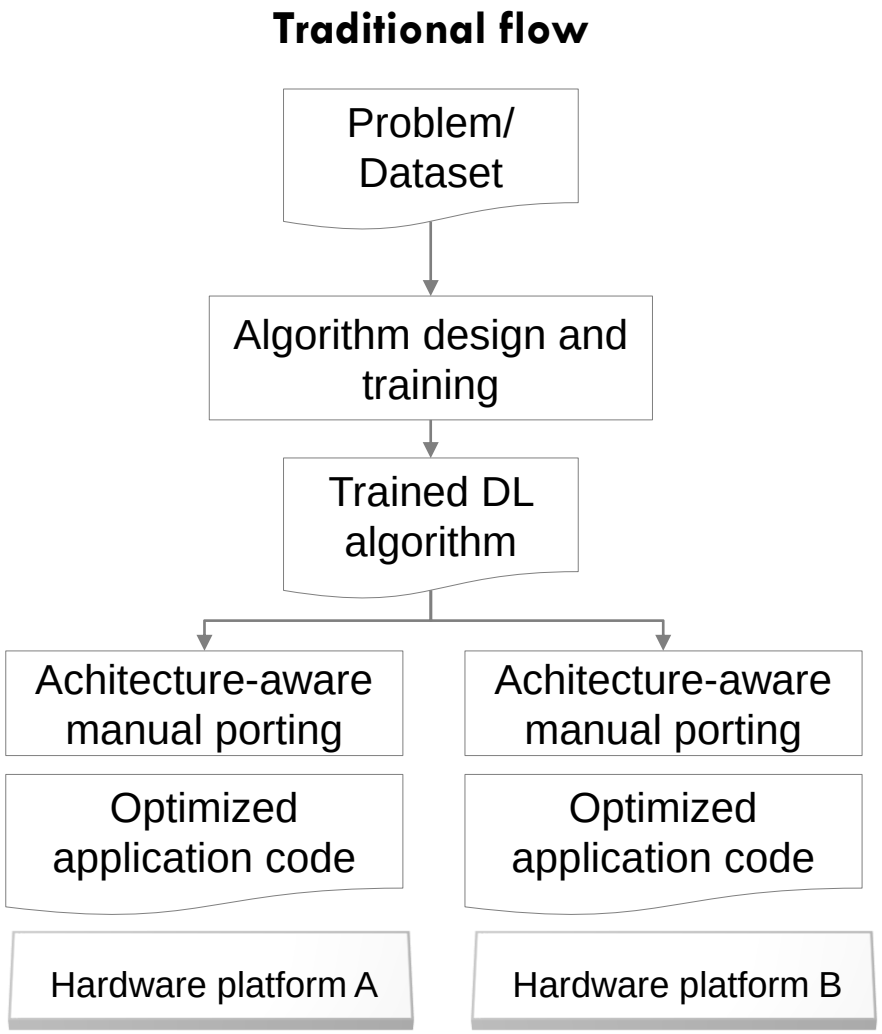
Algorithms

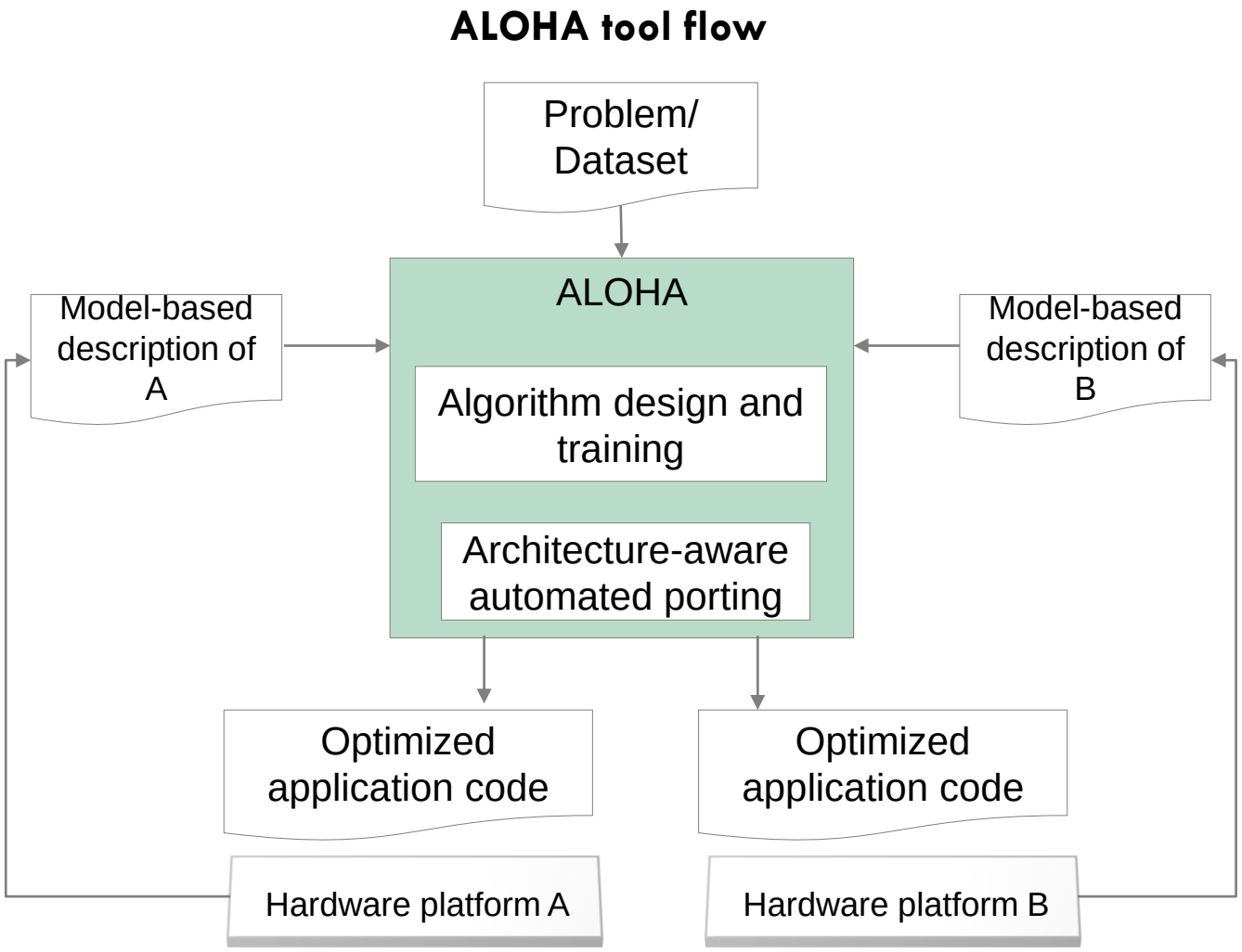
- Complex
- Computationally-expensive

Platforms

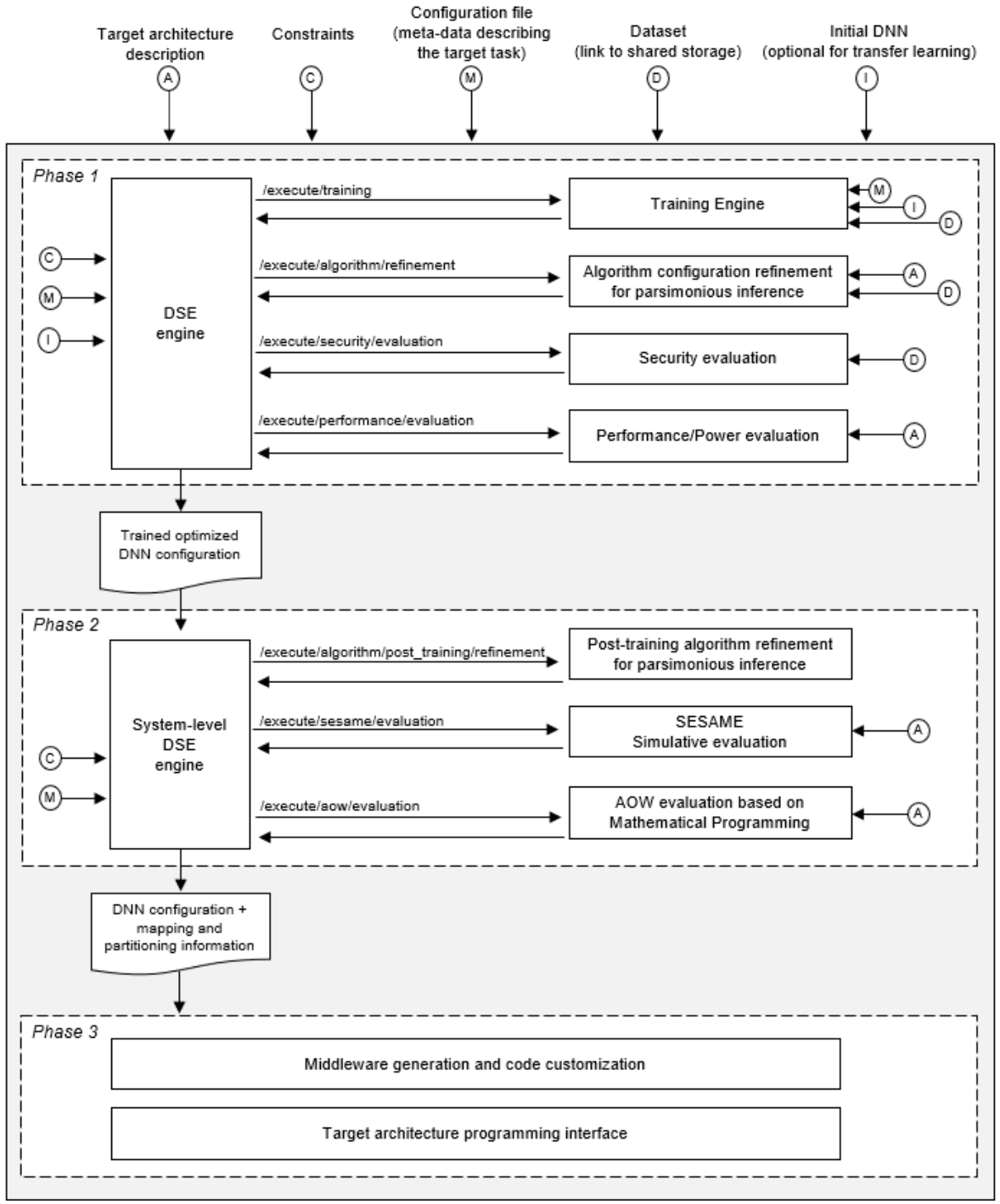
- Highly heterogeneous and parallel
- Hard to program
- Hard to manage at runtime
- Need for low-power

ALOHA: the approach

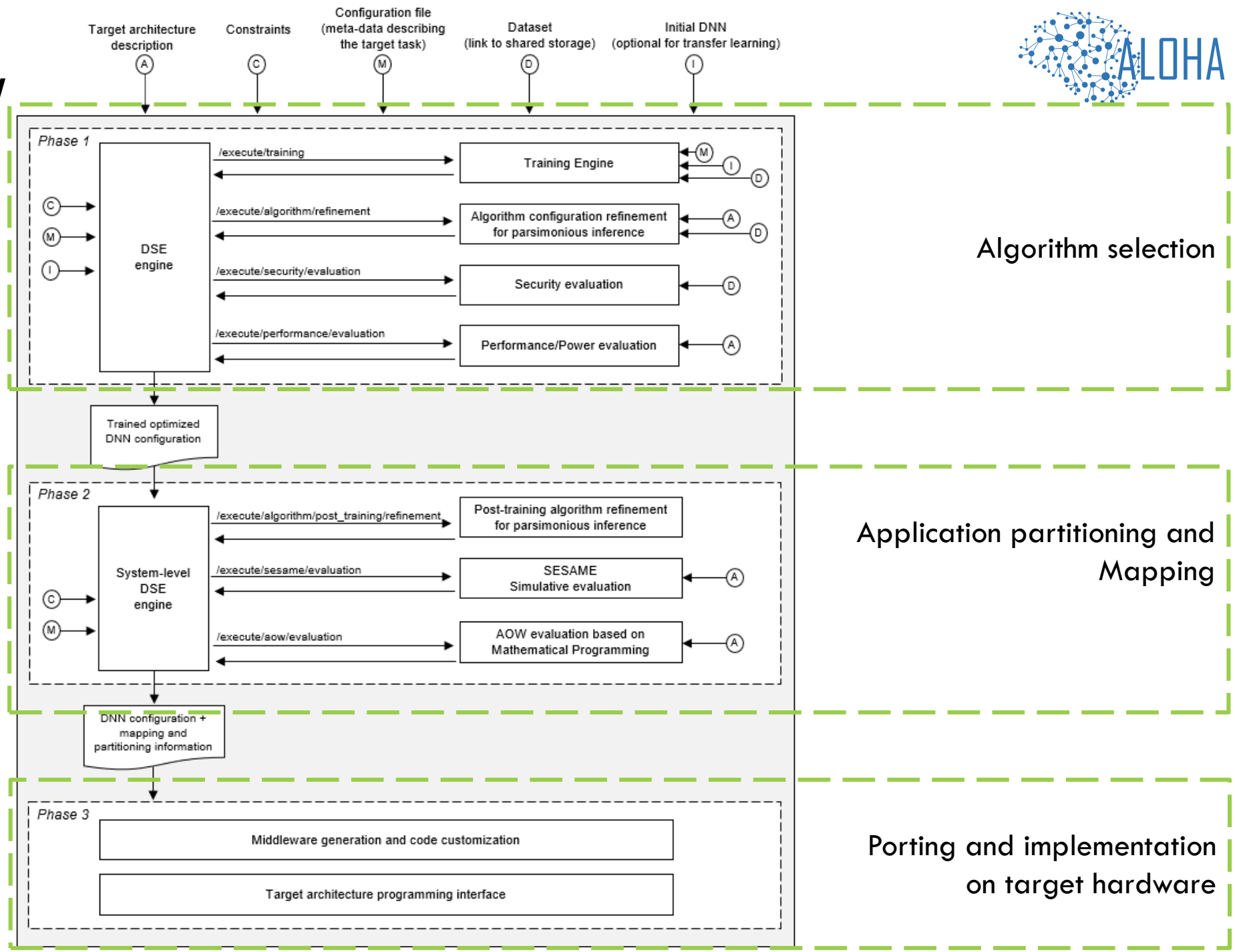




ALOHA Tool flow
General overview



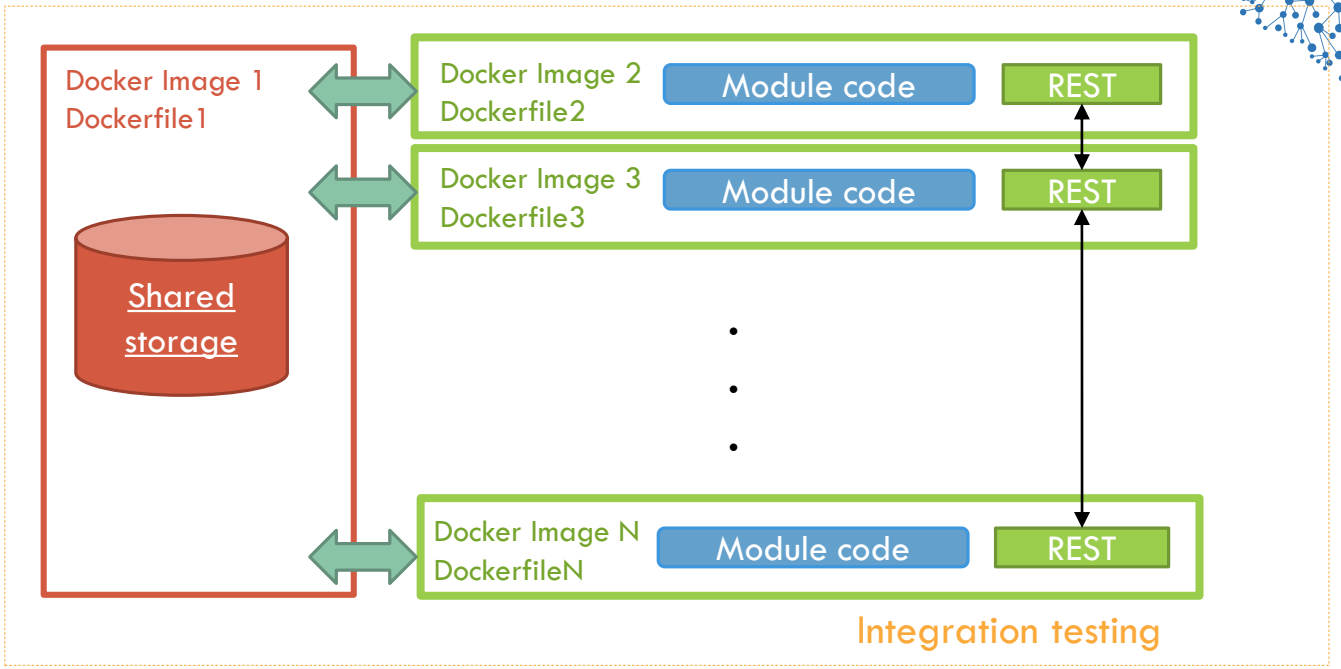
ALOHA Tool flow
General overview

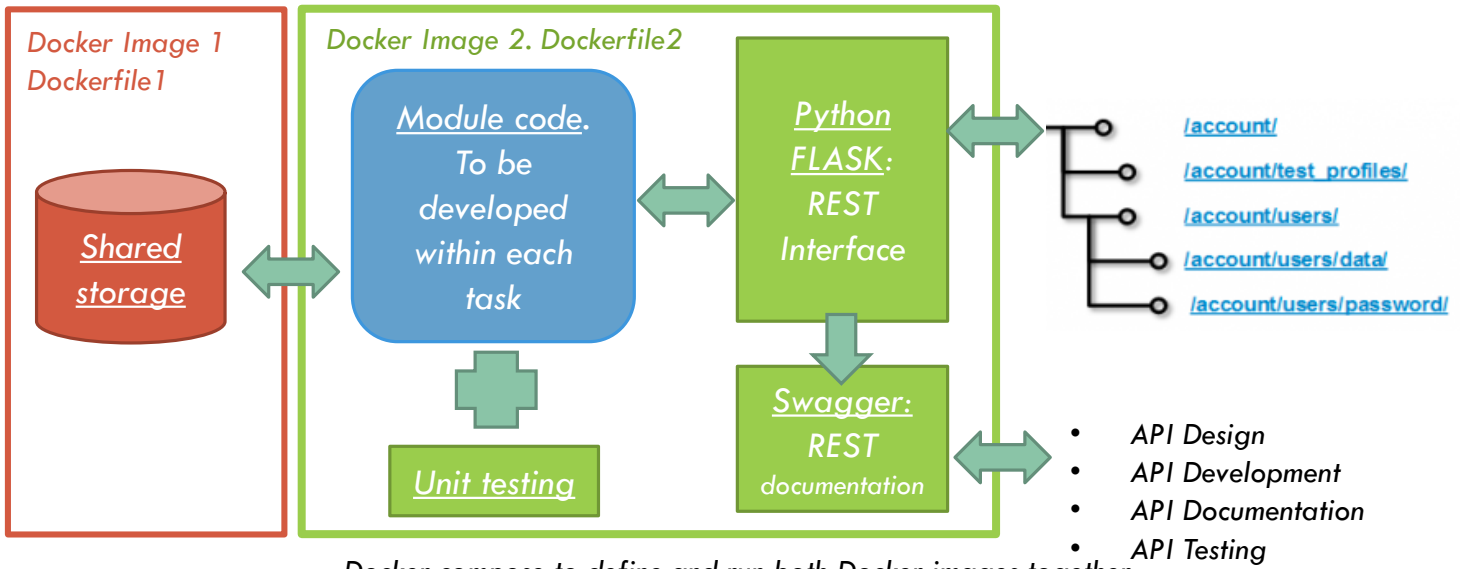


www.aloha-h2020.eu

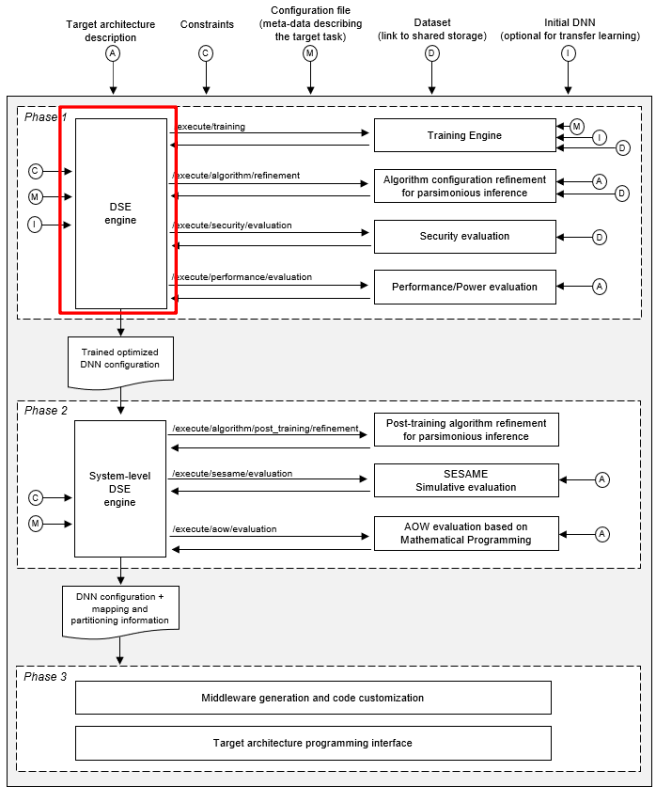
ALOHA Tool flow
Integration methodology

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

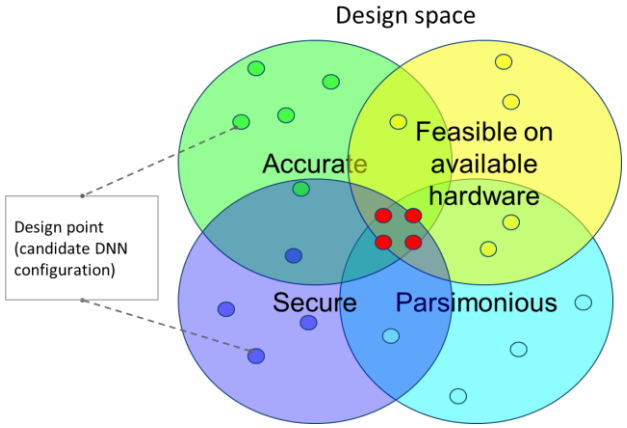




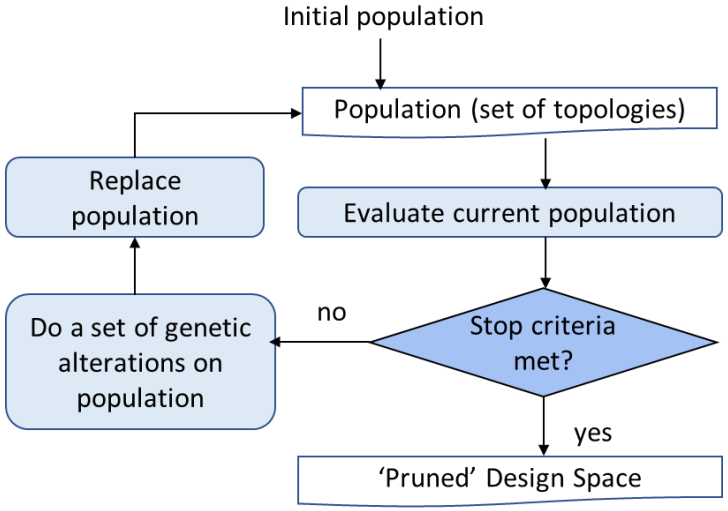
ALOHA Tool flow
The DSE engine



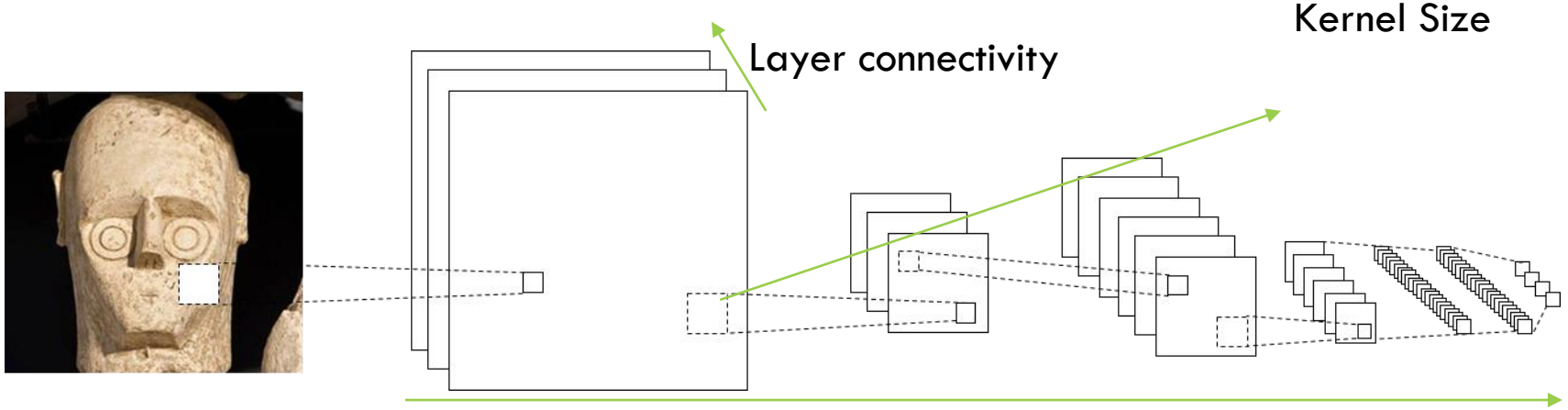
- Multi-objective exploration



- Genetic algorithm DS surfing



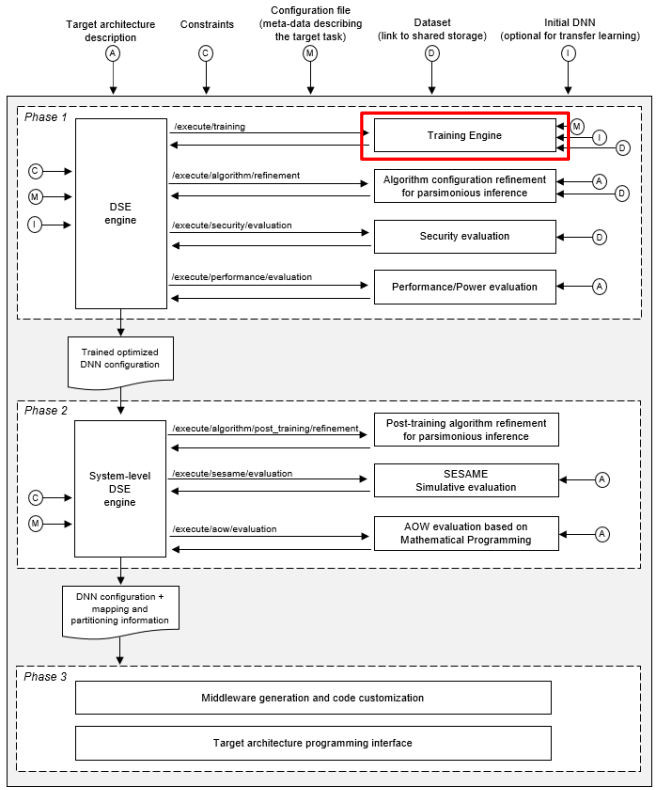
- Explore algorithm configuration degrees of freedom

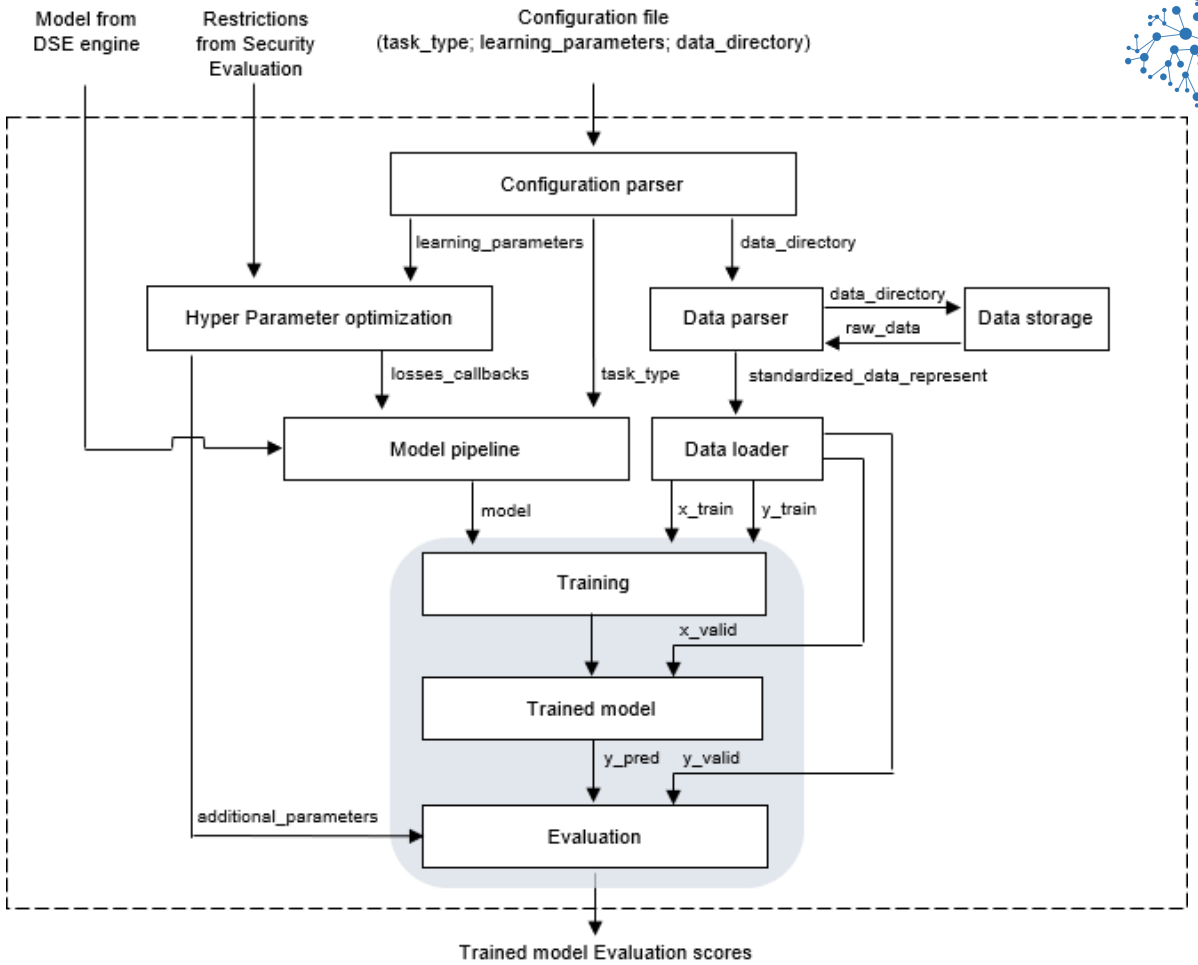


… and others!

www.aloha-h2020.eu

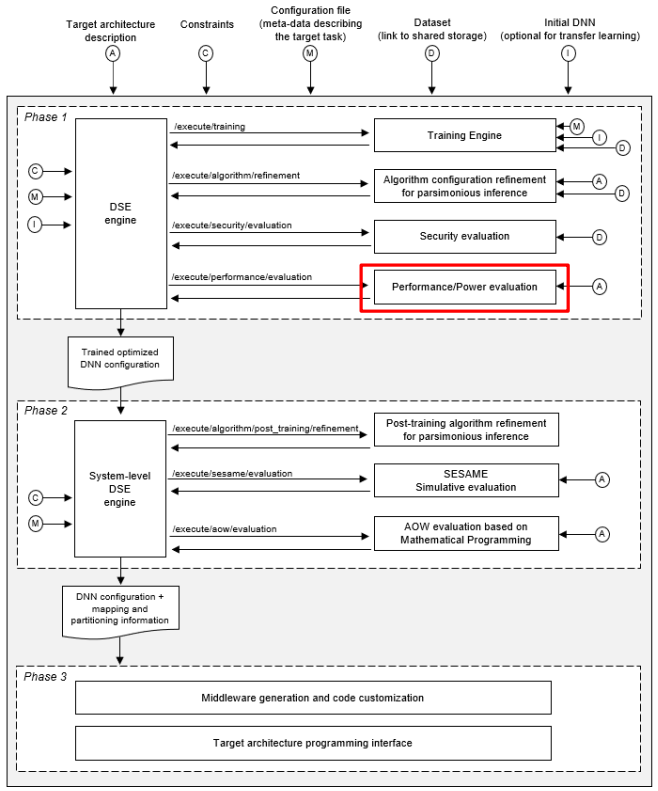
ALOHA Tool flow
The training engine



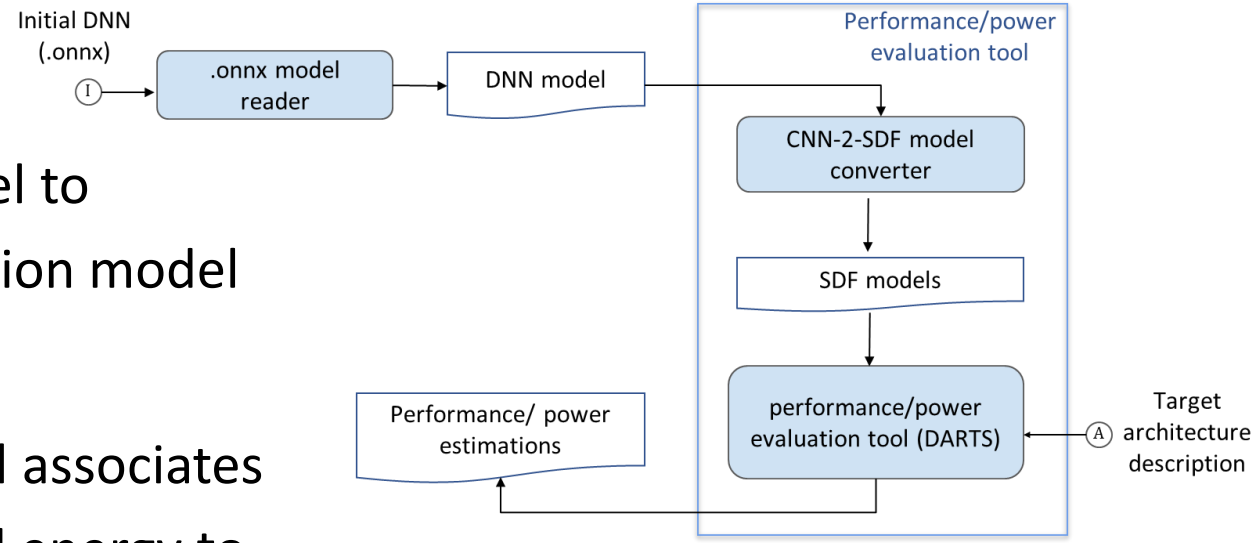


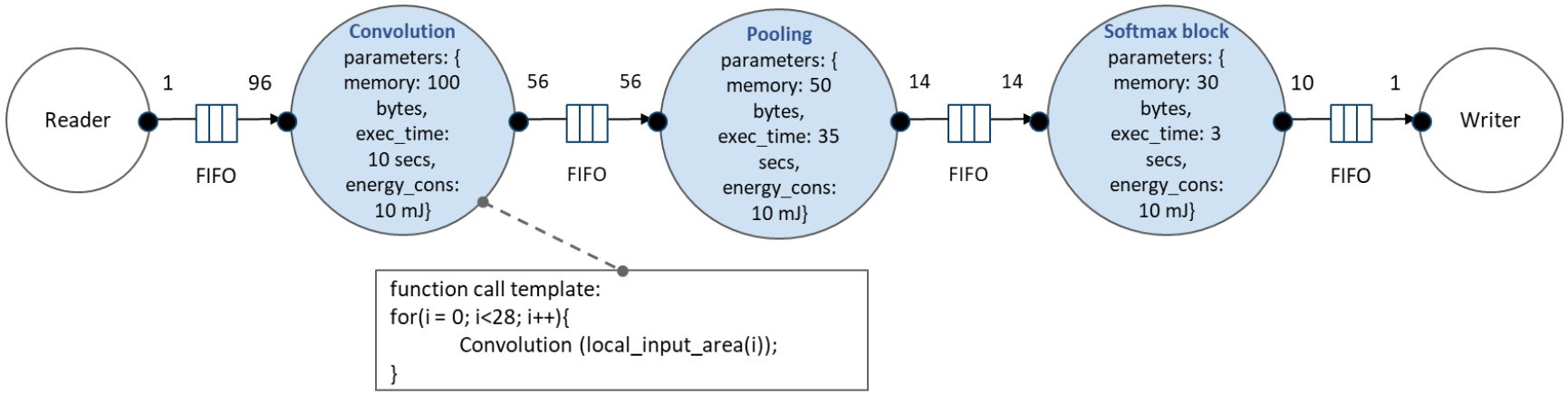
- Transfer learning enabled (reuse of the model in different domains with reduced dataset)
- Local hyperparameter exploration
- Flexible data parsing (multiple input formats)
- Flexible use-case configuration (multiple AI tasks: classification, detection, tracking etc…)

ALOHA Tool flow
Performance/power evaluation

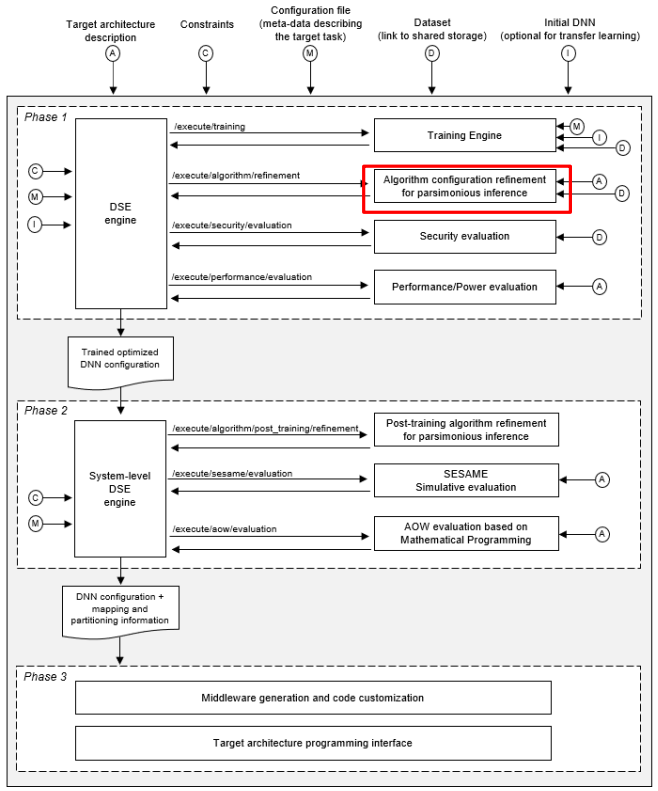


- Convert DNN model to analyzable application model (SDF)
- Architecture model associates execution time and energy to each SDF actor (iteration)
- Analyze/transform/evaluate



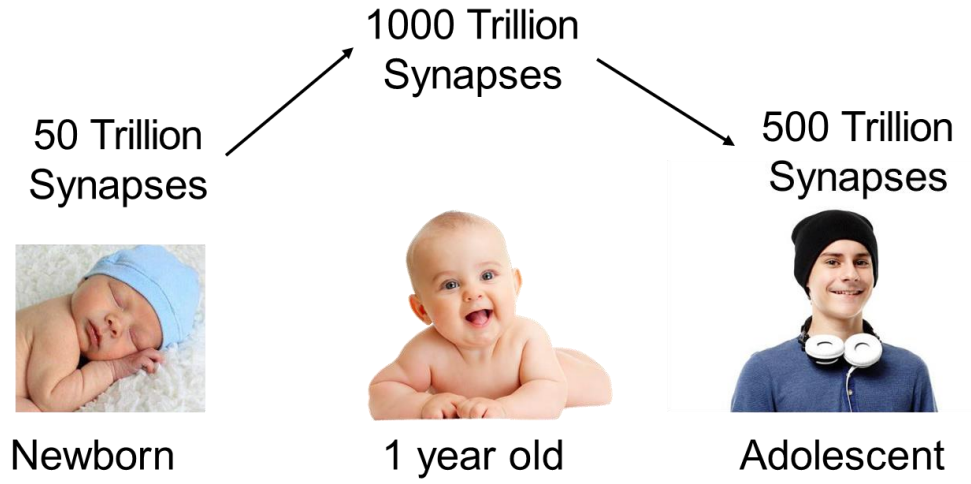


ALOHA Tool flow
Refinement for parsimonious inference



Refine algorithm selection to:

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



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

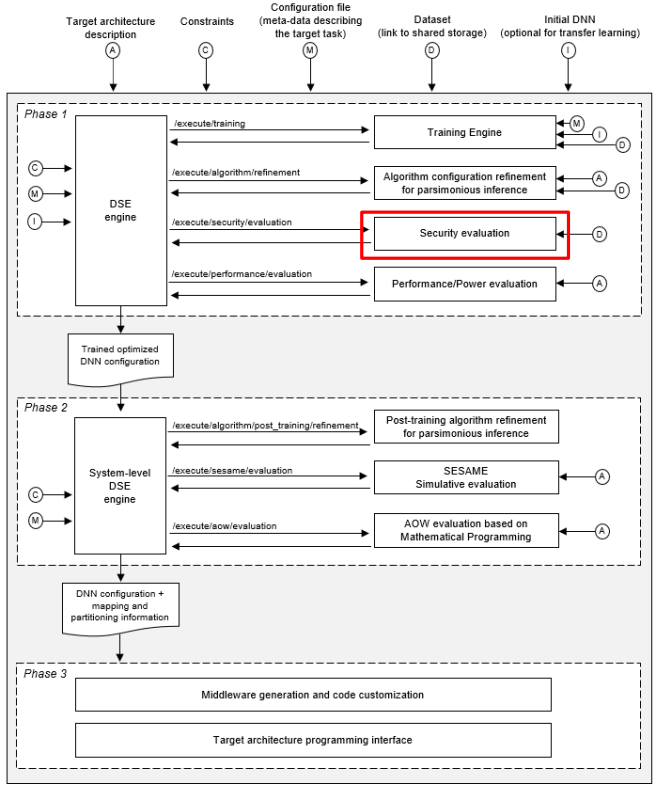
QUANTIZE

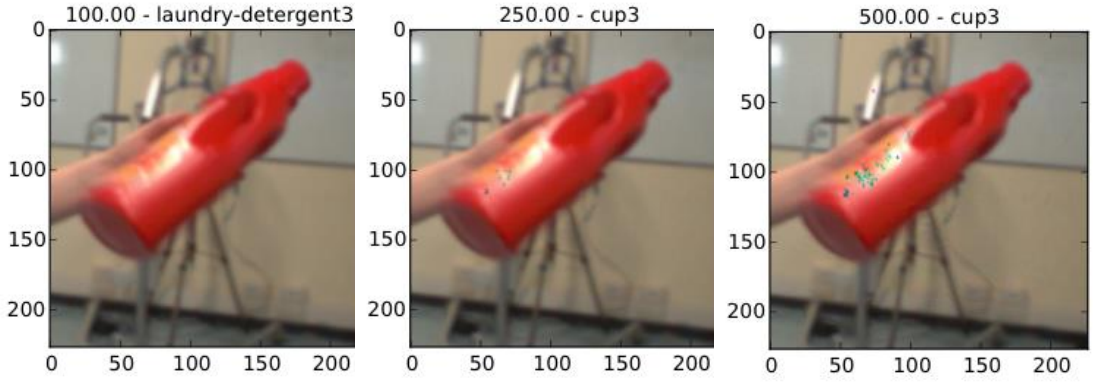
- ***fixed***: low-precision calibration to 16/8 bits
- ***quant_thresh***: quantization to 8/4/2 bits
- *inq*: incremental network quantization
- ***bin***: binarization with XNOR-Net or BNN
- ***abcnet***: binarization with ABC-Net

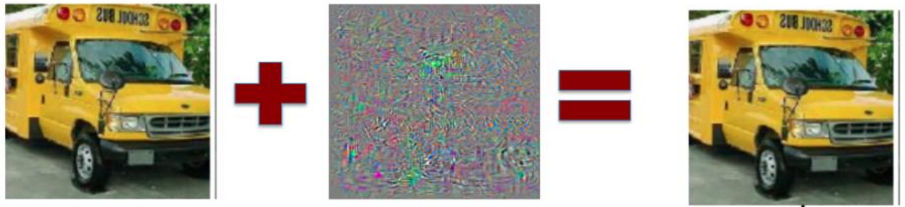
PRUNE

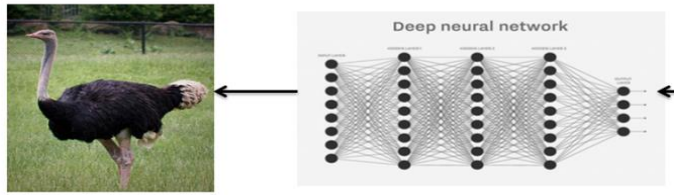
- *prune:* Han et al. iterative pruning of least used connections
- *inq:* pruning together with incremental network quantization

ALOHA Tool flow
Security evaluation

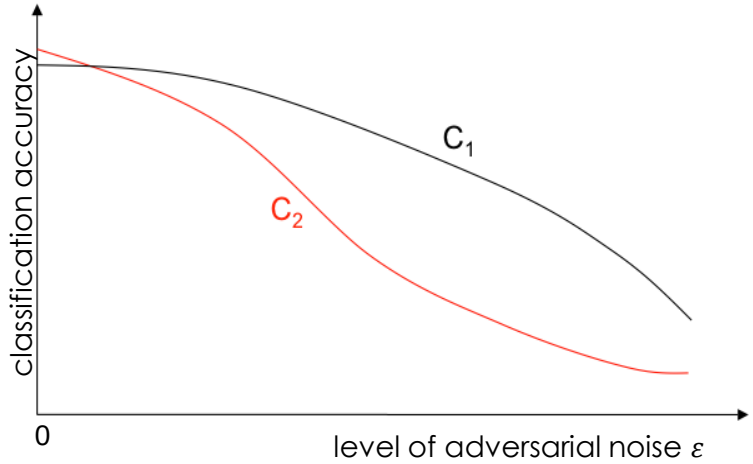






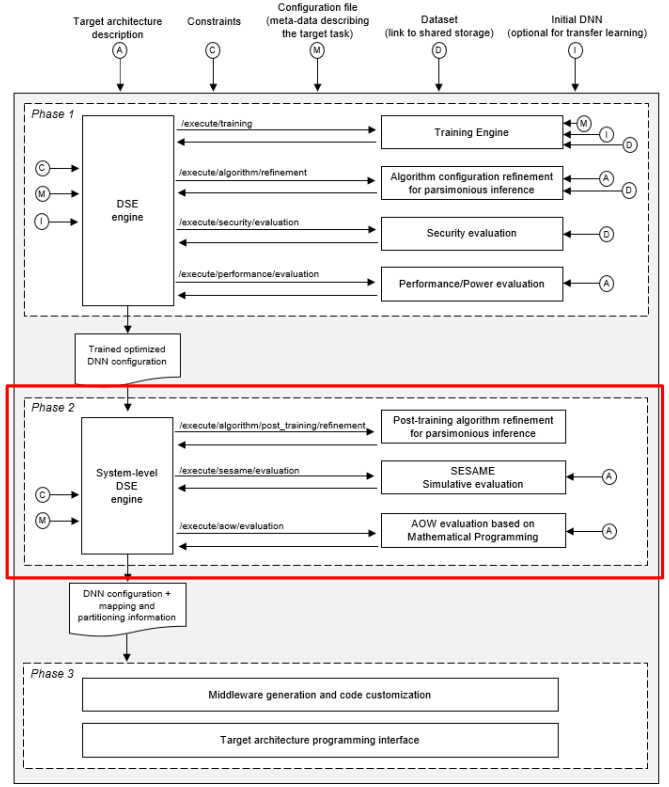


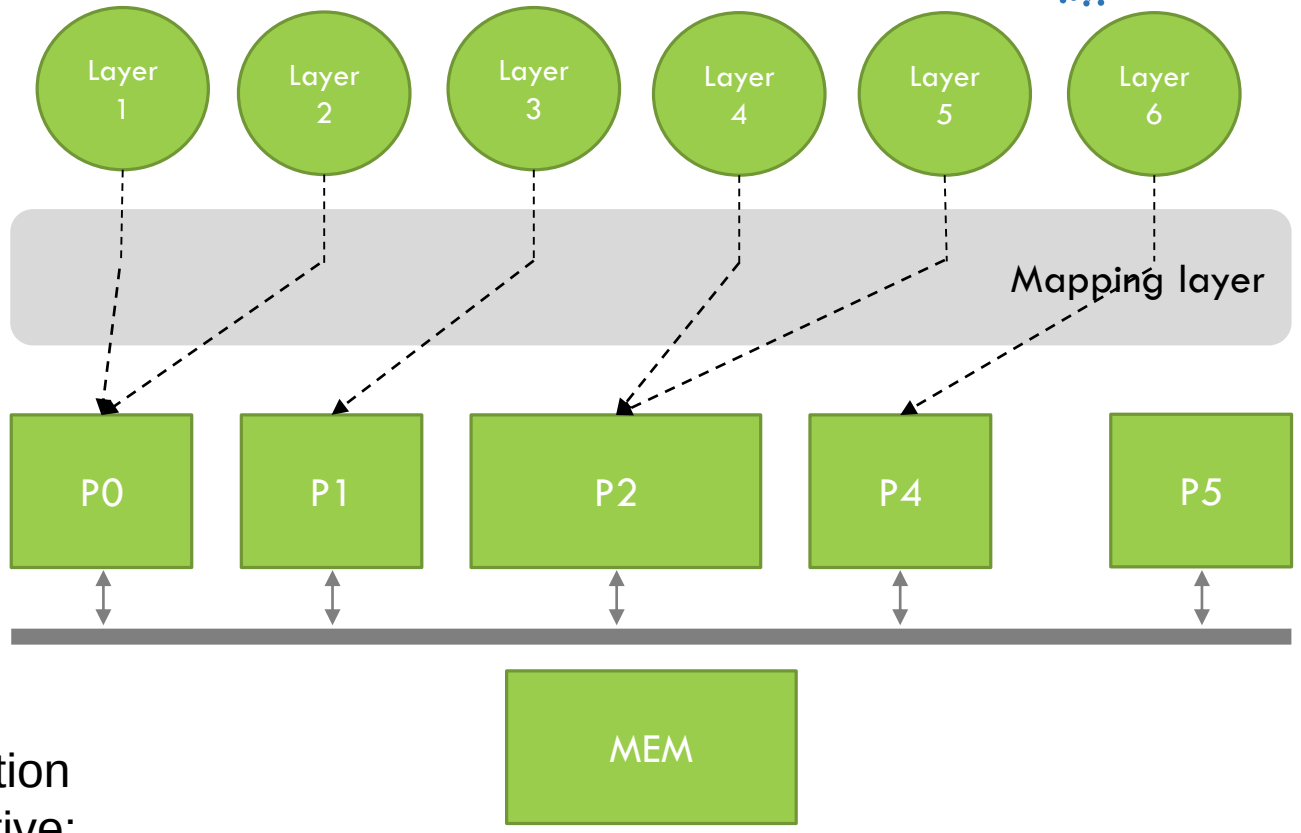
Biggio et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**



www.aloha-h2020.eu

ALOHA Tool flow
Phase 2: system-level design





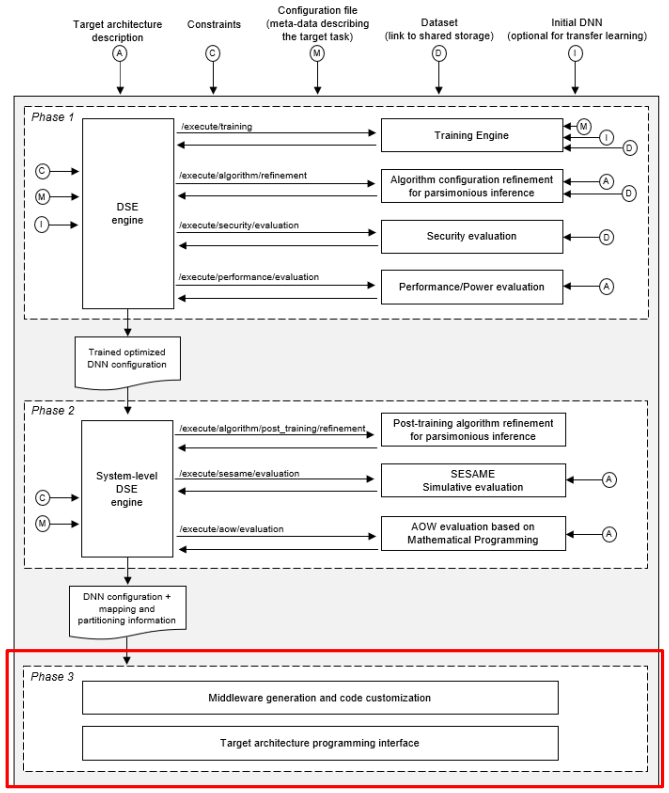
- DSE instrument: Genetic Algorithm

- Design point evaluation
  a) High-level simulative:
  SESAME by UvA (previously on www.madnessproject.org)
  http://sesamesim.sourceforge.net/

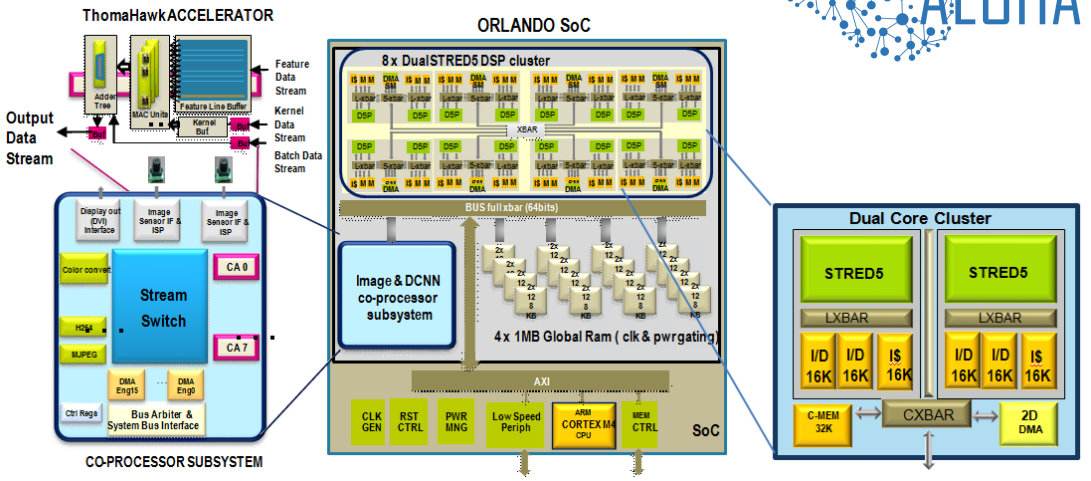
  b) Mathematical Programming and analytic modeling:
  Architectural Optimization Workbench by IBM
  Michael Masin et al. -Pluggable Analysis Viewpoints for Design Space Exploration, 2013

- Post-training parsimonious inference enabled – runtime network modification

ALOHA Tool flow
Reference computing platforms

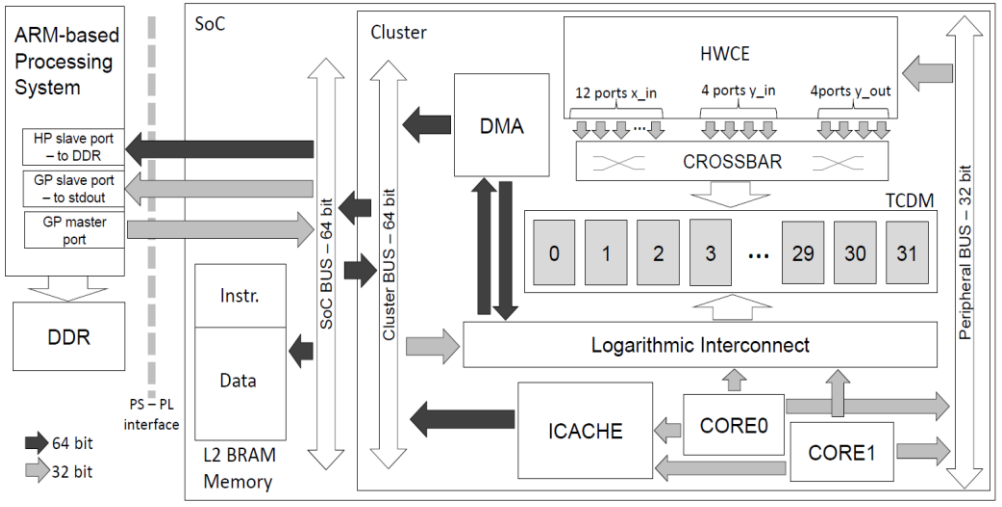


STMicro's Orlando



G. Desoli/et al/., **"14.1 A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems"**, 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 238-239. doi: 10.1109/ISSCC.2017.7870349

NEURAghe

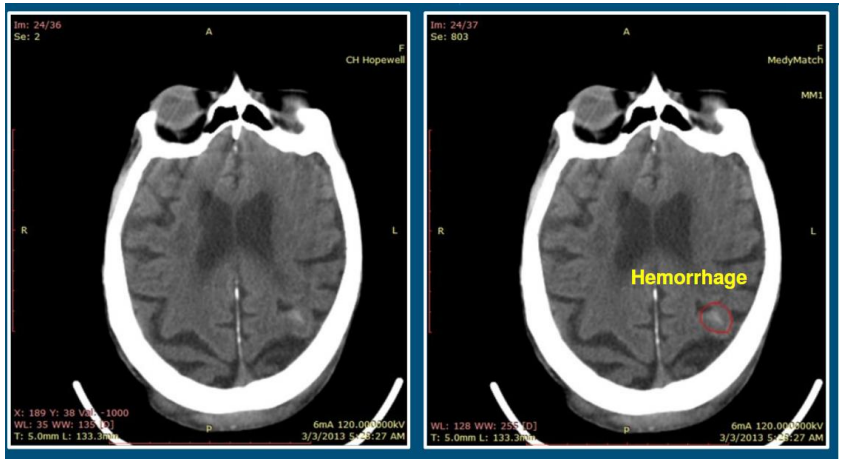


P. Meloni, A. Capotondi, G. Deriu, M. Brian, F. Conti, D. Rossi, L. Raffo, L. Benini, **"NEURAghe: Exploiting CPU-FPGA Synergies for Efficient and Flexible CNN Inference Acceleration on Zynq SoCs"**, 2017, https://arxiv.org/abs/1712.00994

www.aloha-h2020.eu

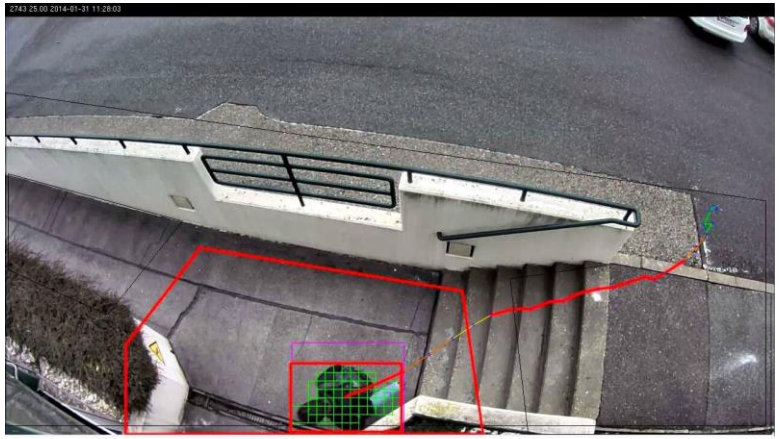
ALOHA use-cases

Demonstration on three application use-cases



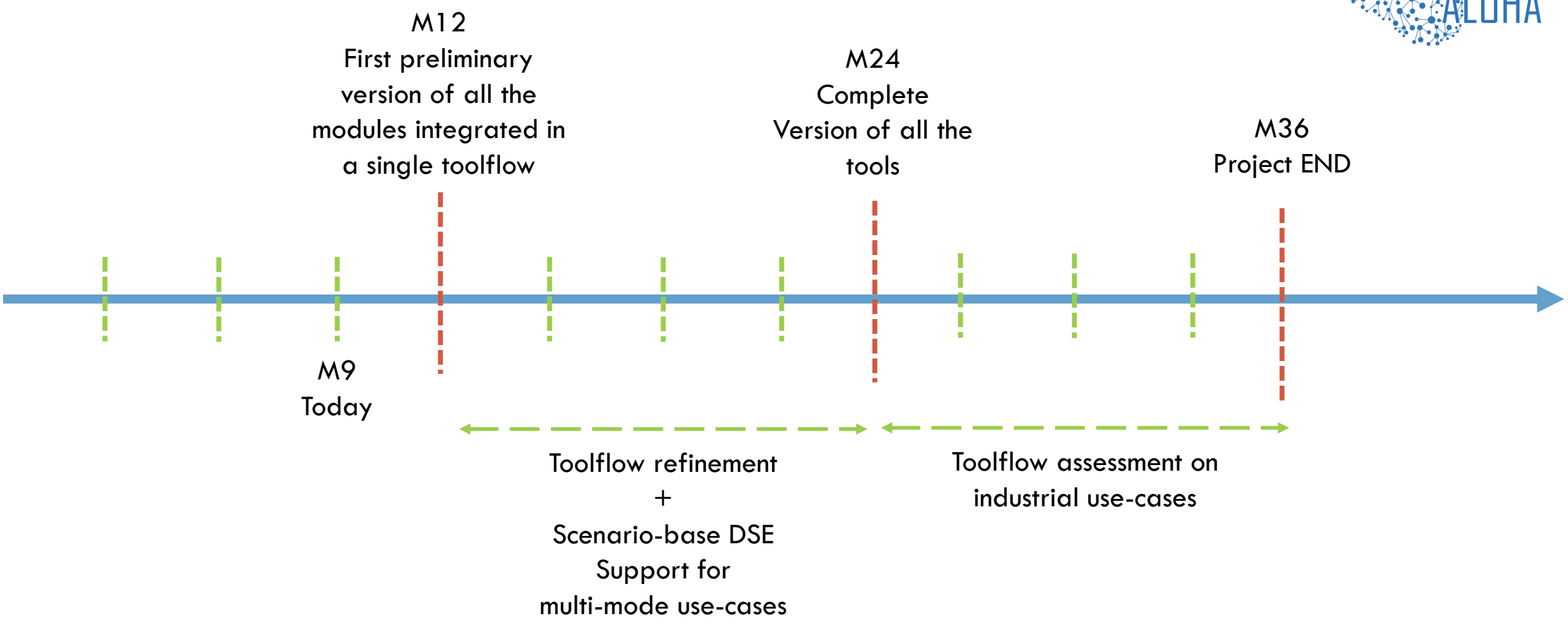
Cost- and power-effective medical decision assistant

Command Recognition in Smart Industry Applications



Surveillance of Critical Infrastructures

ALOHA status



THANK YOU FOR YOUR ATTENTION!

www.aloha-h2020.eu

@ALOHA_H2020

ALOHA project

http://www.aloha-h2020.eu/index.php/project/get-involved

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