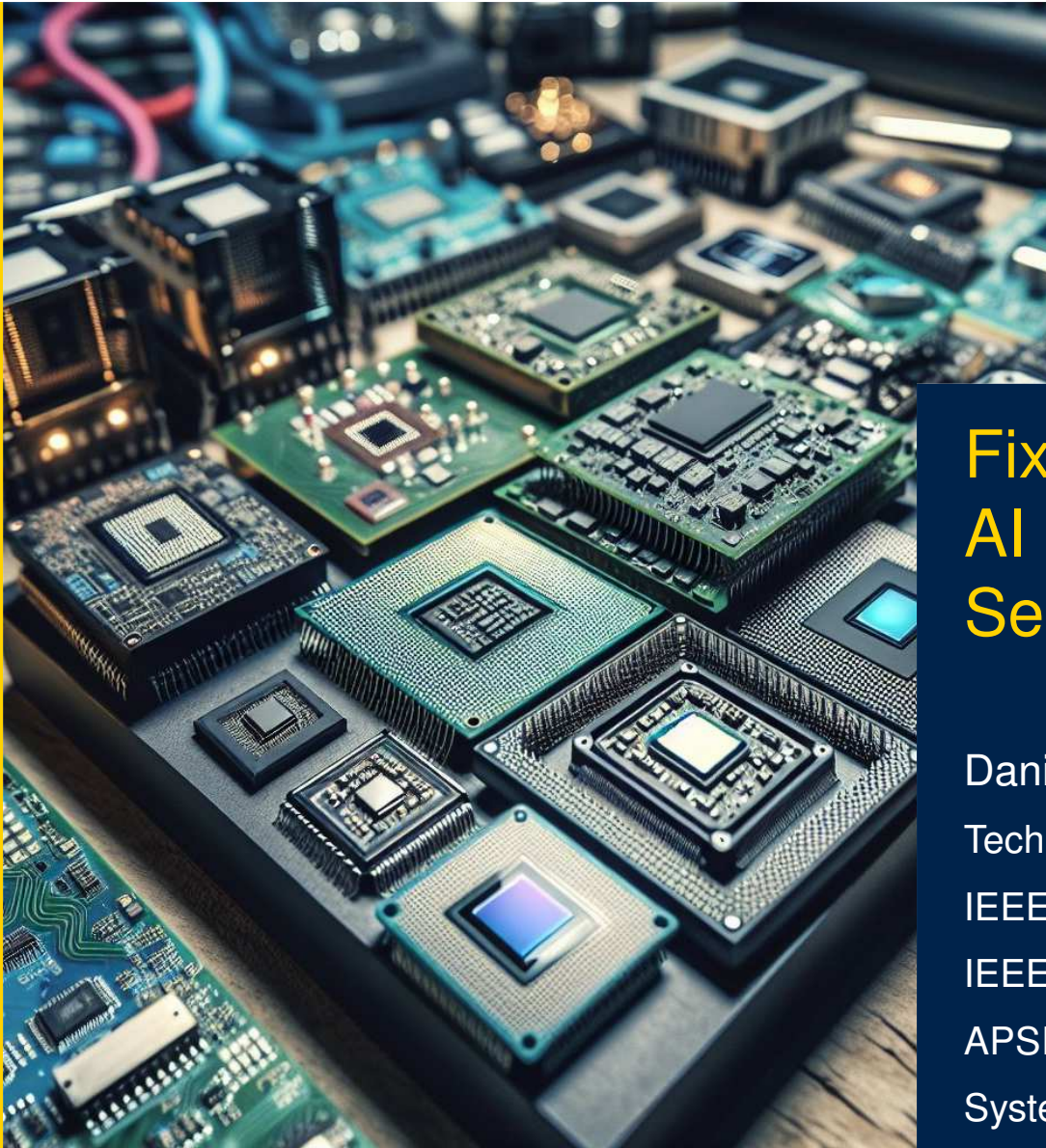




life.augmented

Fixed AI: Practical Use of Edge AI Unified Core Technology on Sensors and MCU

Danilo PAU

Technical Director

IEEE, ST and AAIA Fellow,

IEEE Distinguished Industrial Lecturer

APSIPA Life, Sigma XI, NAAI member

System Research and Applications

Motivations to hear this talk



The origins of Fixed AI and the dawn of TinyML

The Heterogeneity challenge at the EDGE

Solutions to this challenge

Practical use of the solutions on tiny sensor and MCUs

The cloud returns and Gen EdgeAI



Fixed Function AI

- Low power processing (< mW)
- React on data availability
- Provide assessments, insights

Reducing the Dimensionality of Data with Neural Networks
G. E. Hinton*

Krizhevsky, Alex et al.
"ImageNet classification with deep convolutional neural networks".

AI Is Taking Over the Cloud (MIT Tech Review)

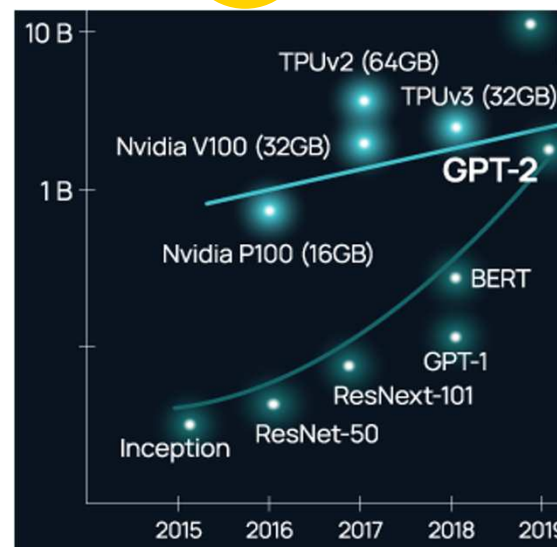


TinyML Products proliferation



Compute Limited Era

Source <https://www.celestial.ai/technology>



Tiny Edge

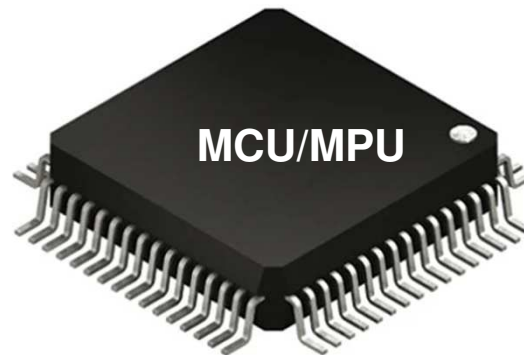
Sensors



Data



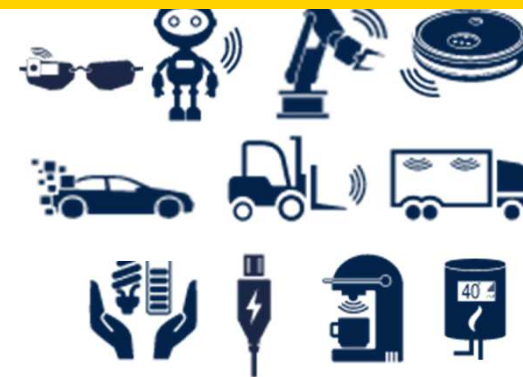
Embedded Computing



Control

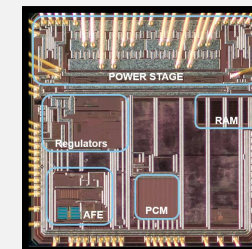


Actuators



10s KiB RAM
No eFLASH
KHz to 10 MHz CPU
INT1, INT8, INT16; FP32

≤ 1 (o/chip) vs 16 (off/chip) MiB RAM
 ≤ 2 (o/chip) vs 64 (off/chip) MiB FLASH
 ≤ 800 MHz
NPU (1 GHz, 4.2 MB o/chip, 600 8b GOPS)
INT8, INT16; FP32



≤ 32 KiB RAM
 ≤ 512 KiB ePCM
160-200 MHz
INT8, INT16; FP32





ISPU

Standard package:
2.5 x 3 mm²

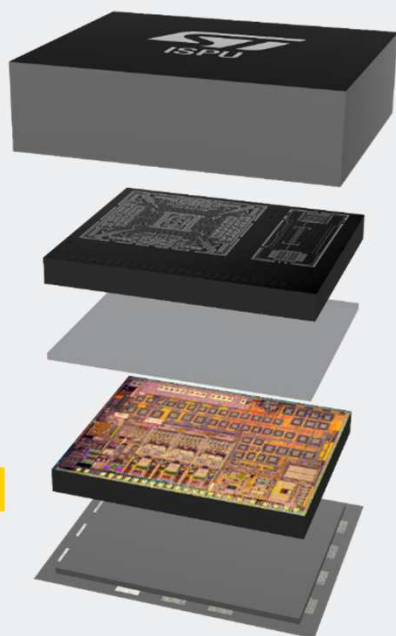
Raw data from
6 DOF MEMS
sensor



ISPU processes AI
(in a few μ W)



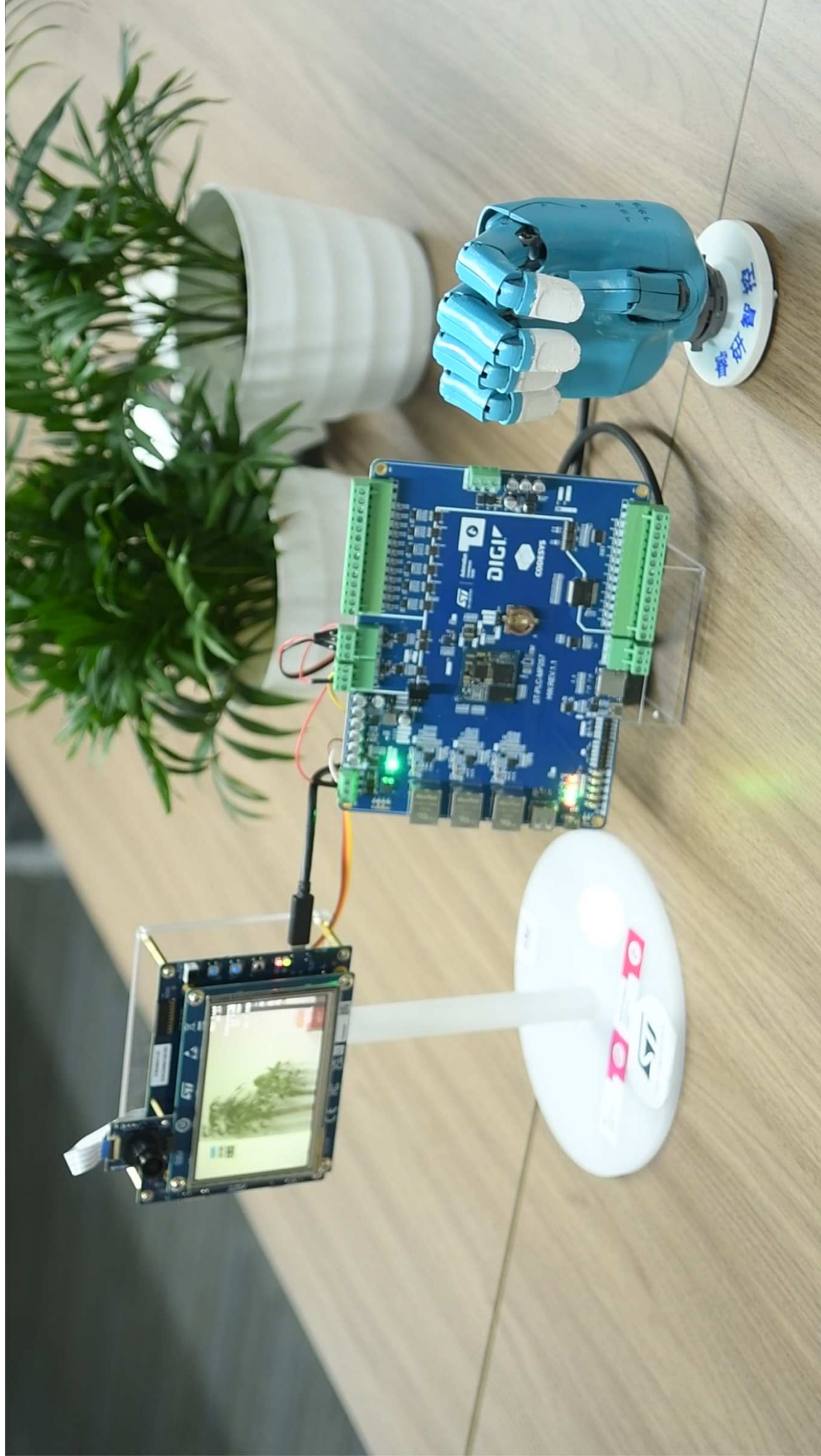
Sensor Hub (up to 4)



**ISPU Wand with
ST Edge AI Core Technology**

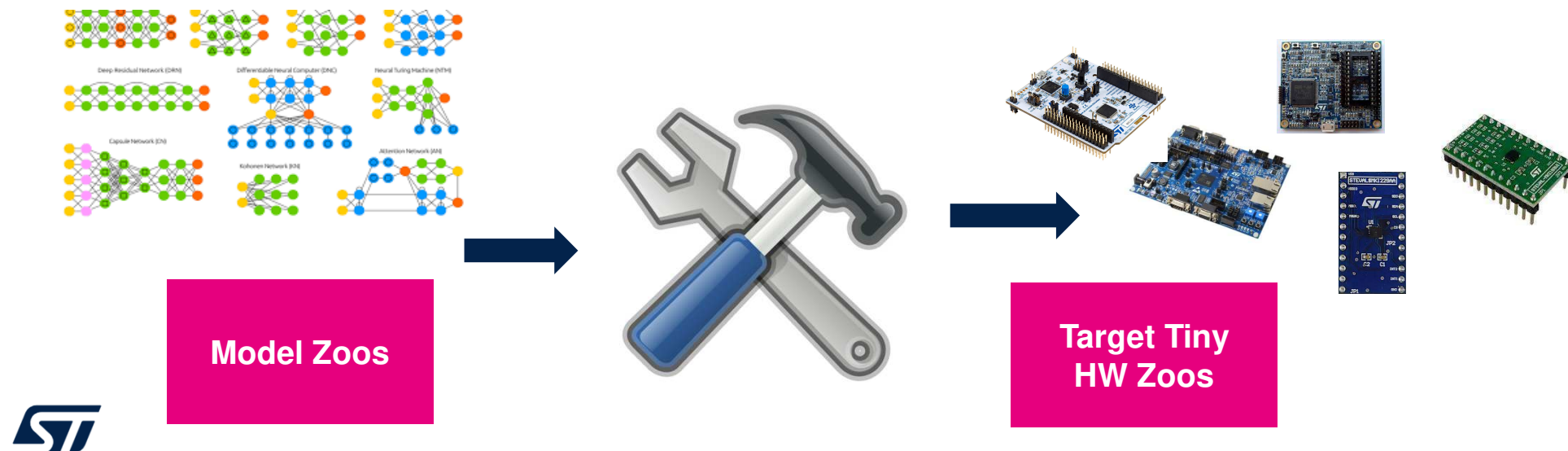
STMicroelectronics



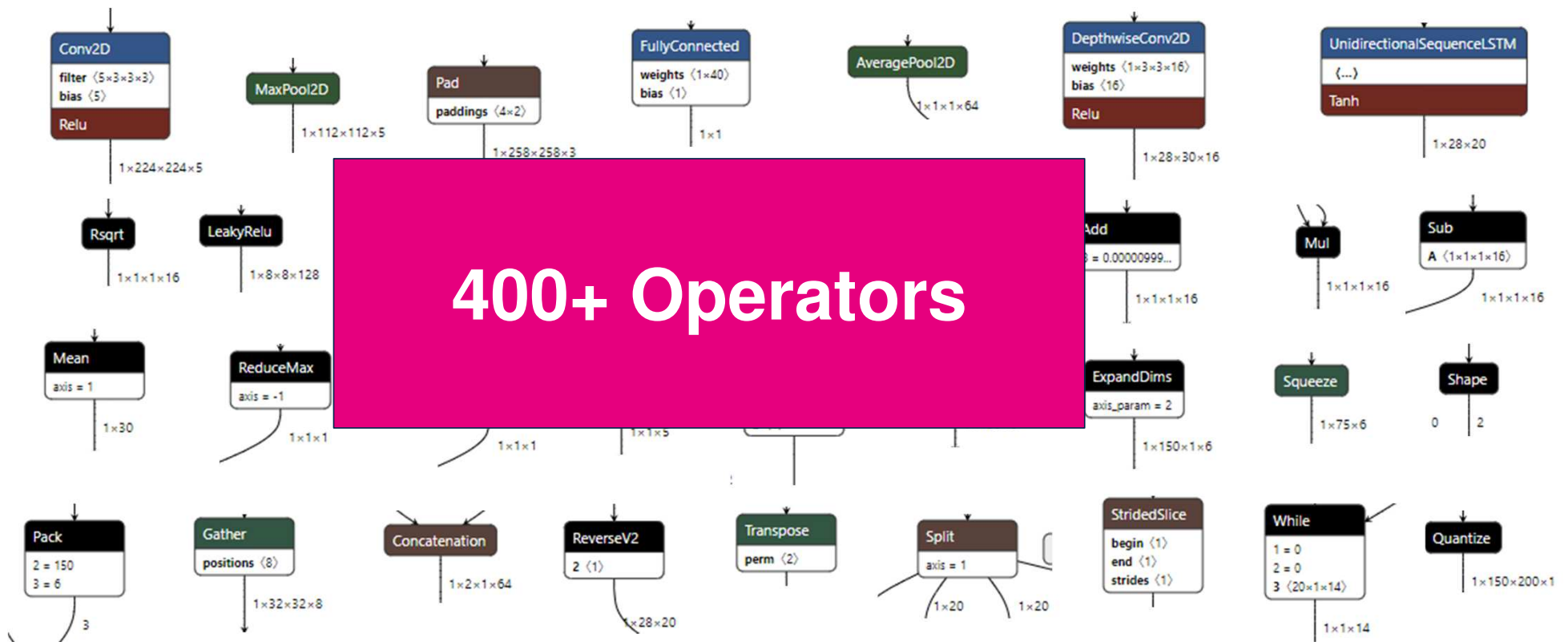


Machine Learning Heterogeneity

- Very different w.r.t.:
 - **Source Model**: operators, topologies, footprint, representation formats, etc.
 - **Execution Target**: computational capabilities, available memory, optimized instructions, etc.



Heterogeneity: Operators



700+ ICs



Heterogeneity: Execution Targets

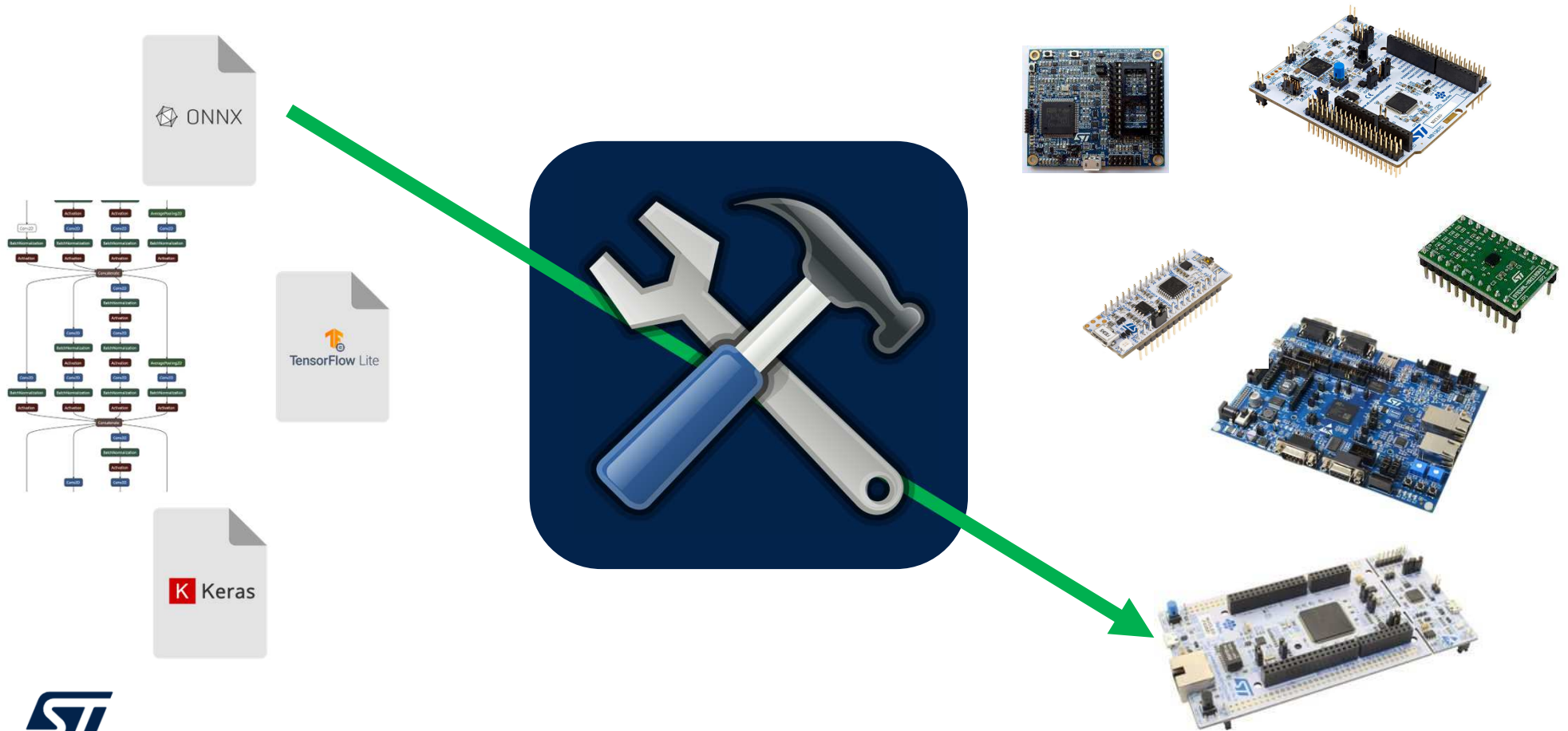
Instruction Sets (ARM Cortex M 0/4/7/33/55/85, R52, STRed)

Tiny chips, heterogeneous memory schemes, e.g.,
Sensors: MLC, ISPU
STM32 MCUs, STM32N6, STM32MP2, Stellar MCUs, x86-64

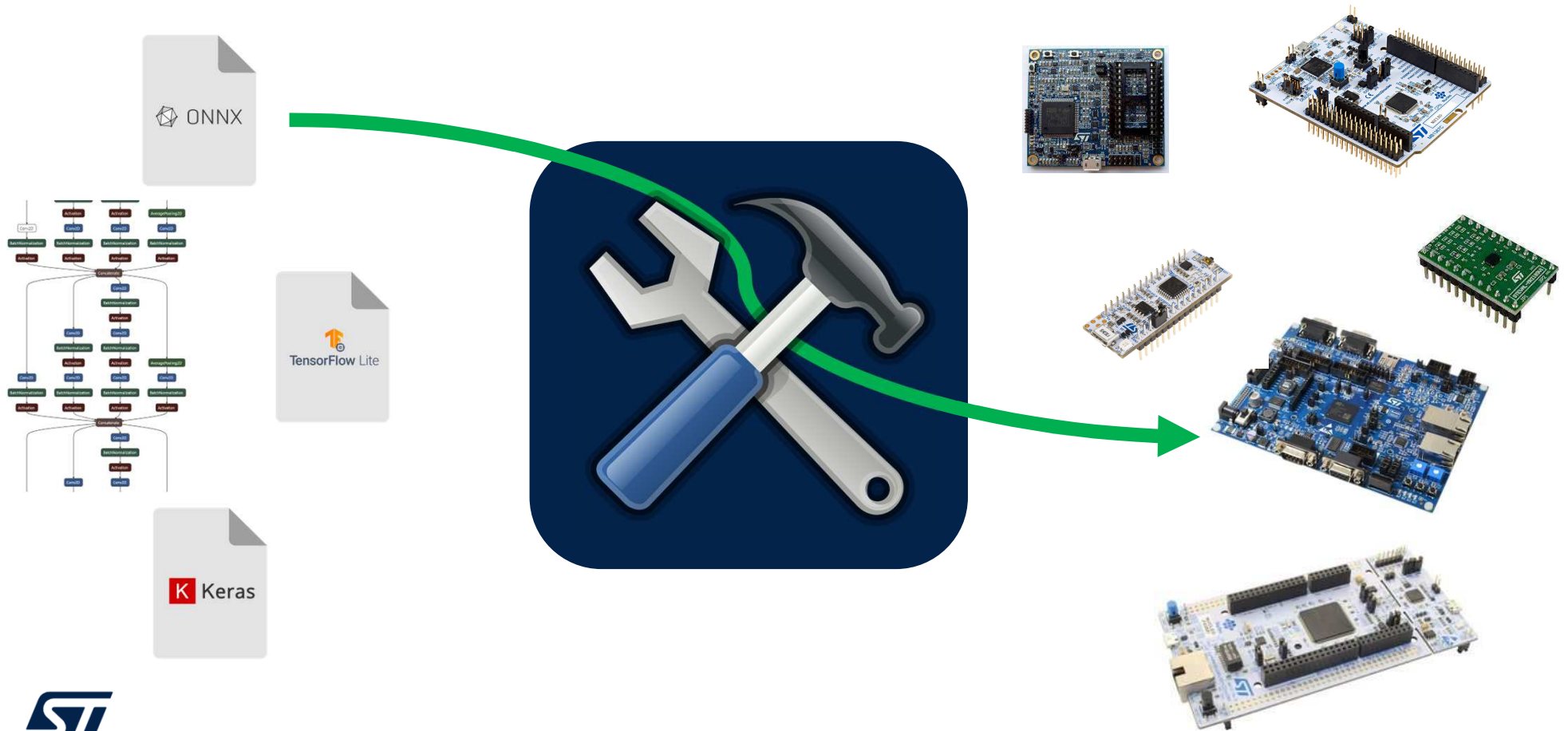
Machine models, e.g.,
Binary instructions in ISPU
Integer SIMD/Vector instructions in ARM
Convolutional co-processors and DMAs in STM32N6

Compilers, e.g.,
gcc, IAR, Hightec, Keil, etc
Neural-ART compiler (STM32N6)
ISPU compiler

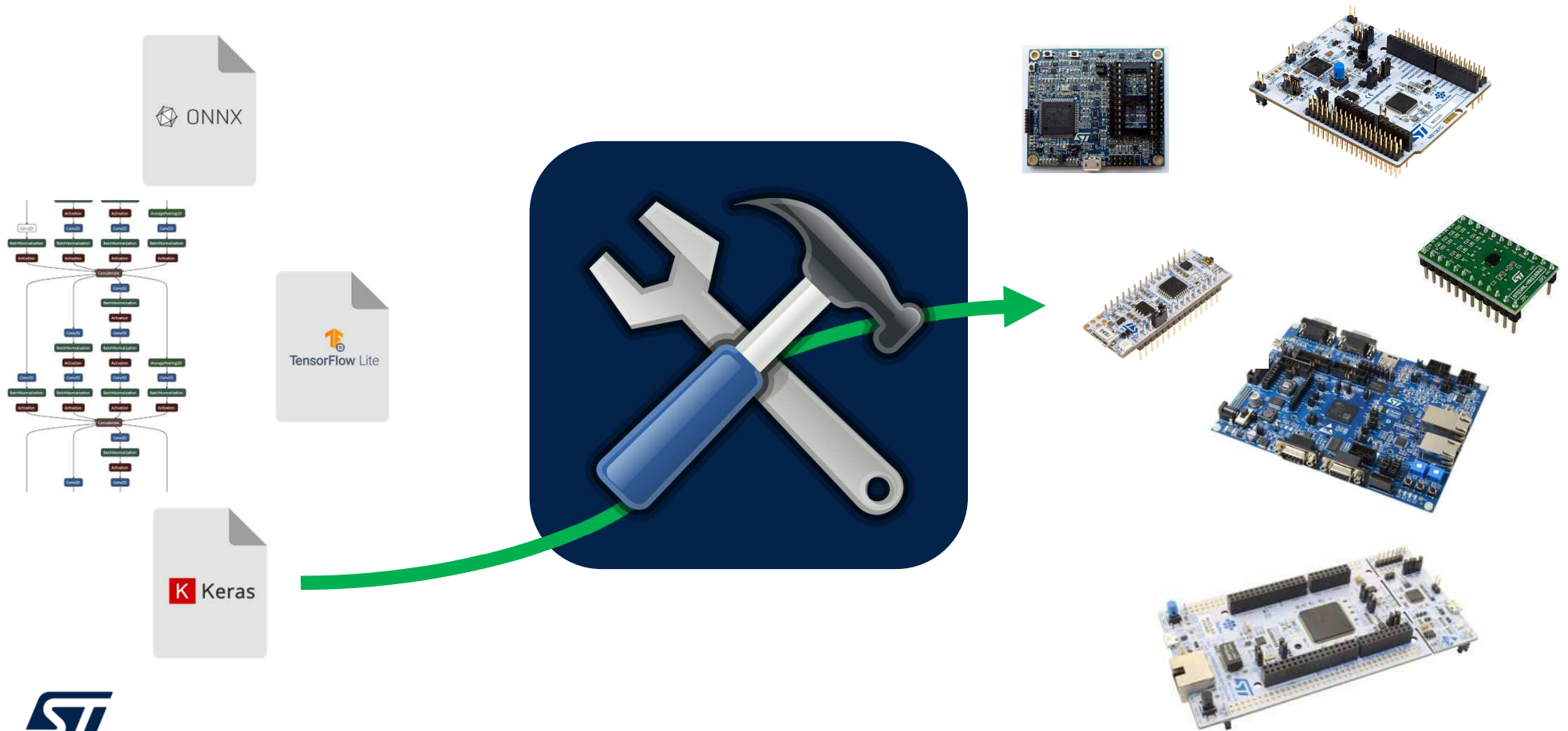
Paths from Source to Target



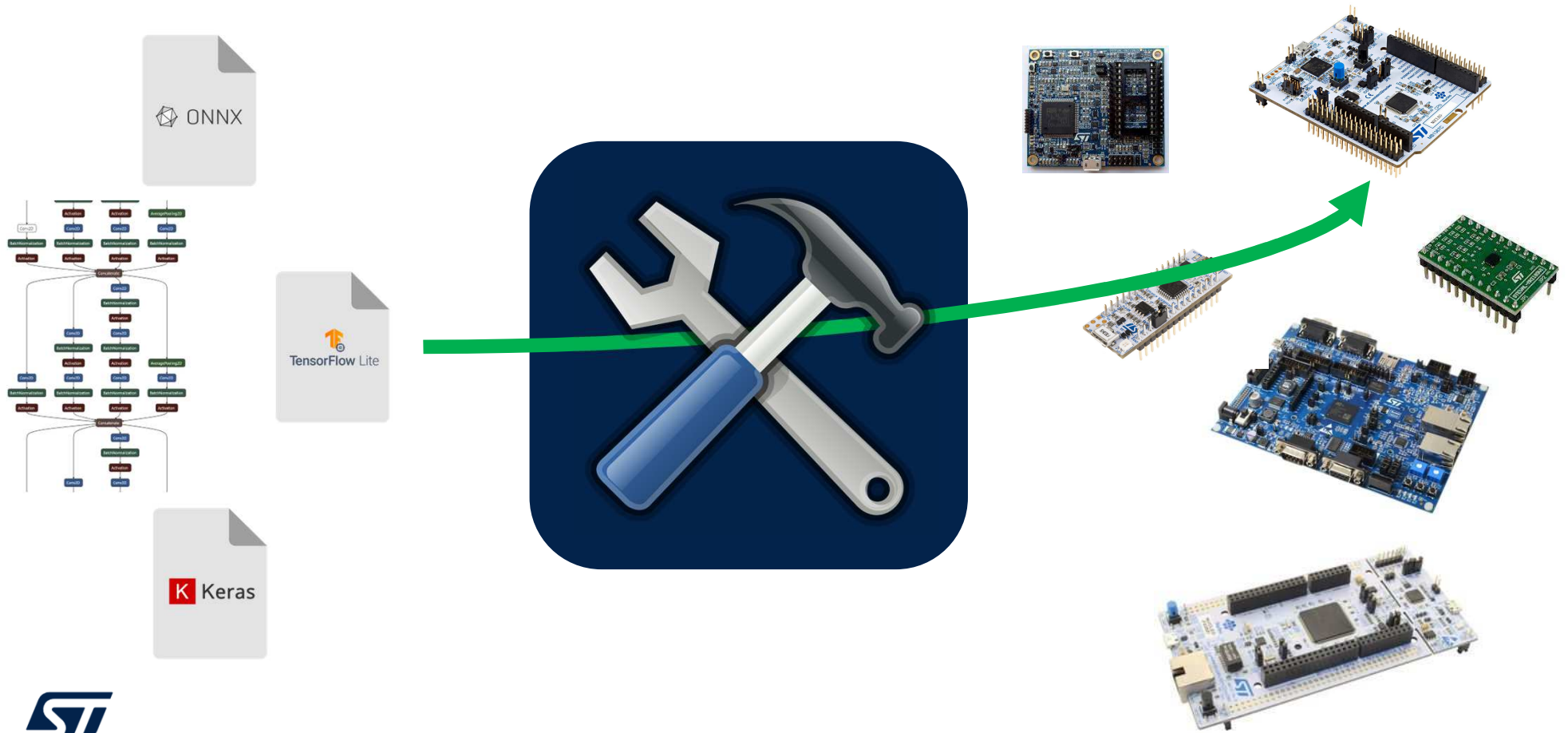
Paths from Source to Target



Paths from Source to Target



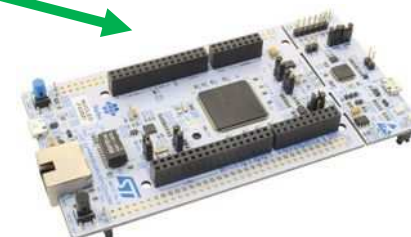
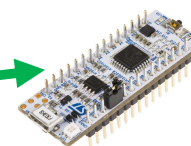
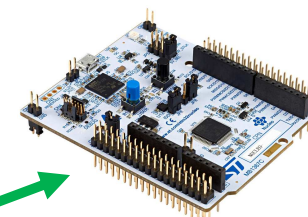
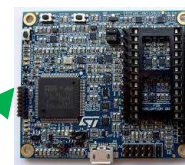
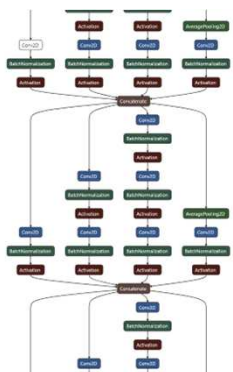
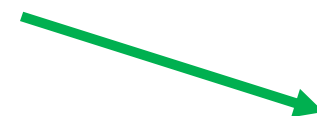
Paths from Source to Target



Homogenize Developer Experience

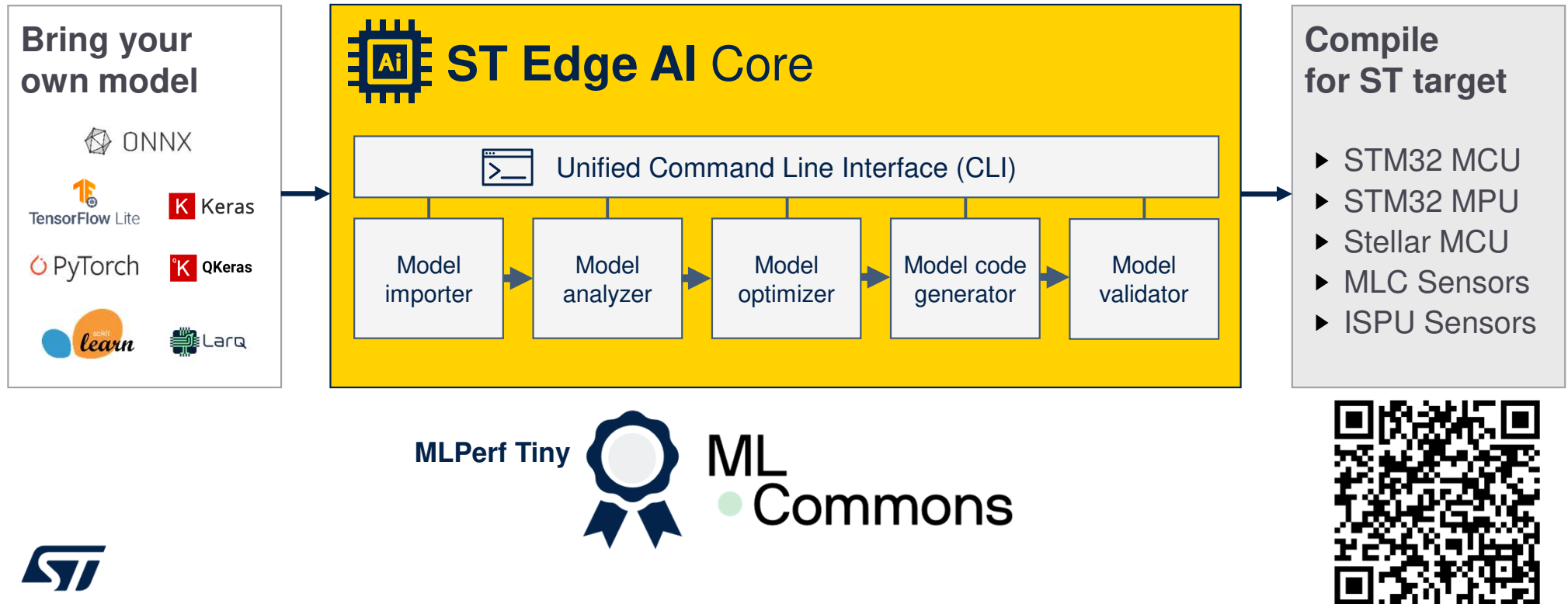


One Unified
Core
Technology

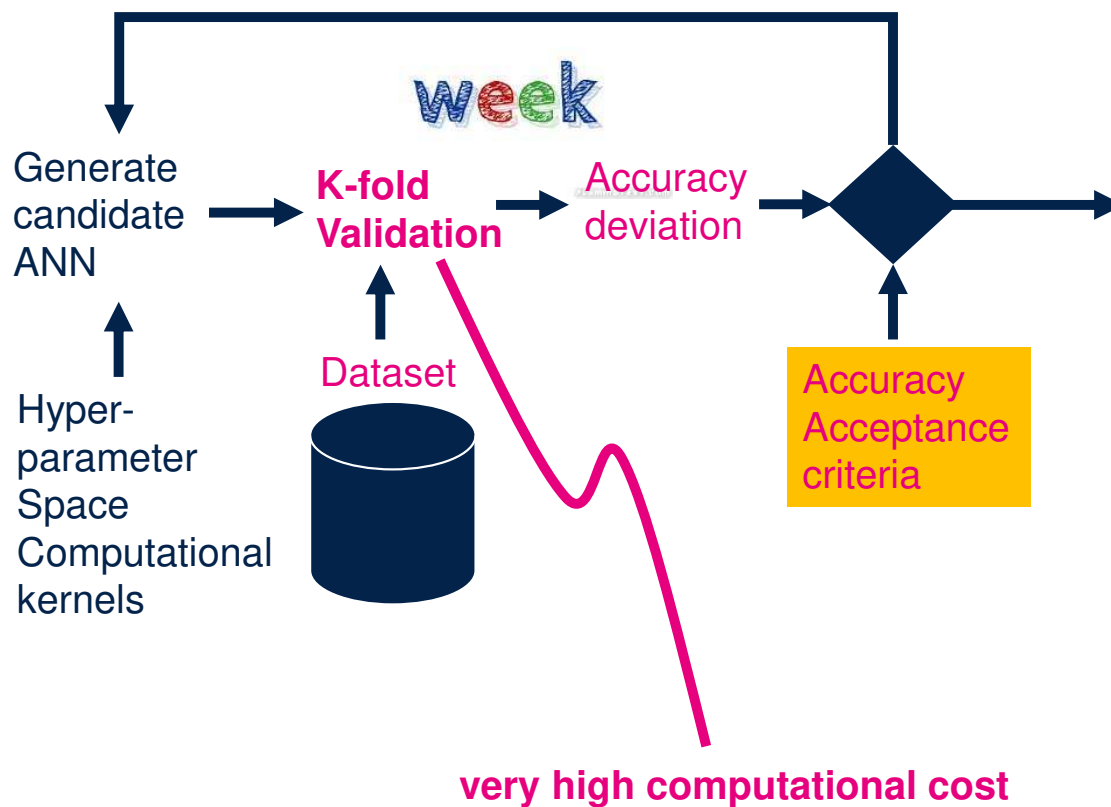


ST Edge AI Core technology

Common state-of-the-art optimizer technology for ALL ST devices

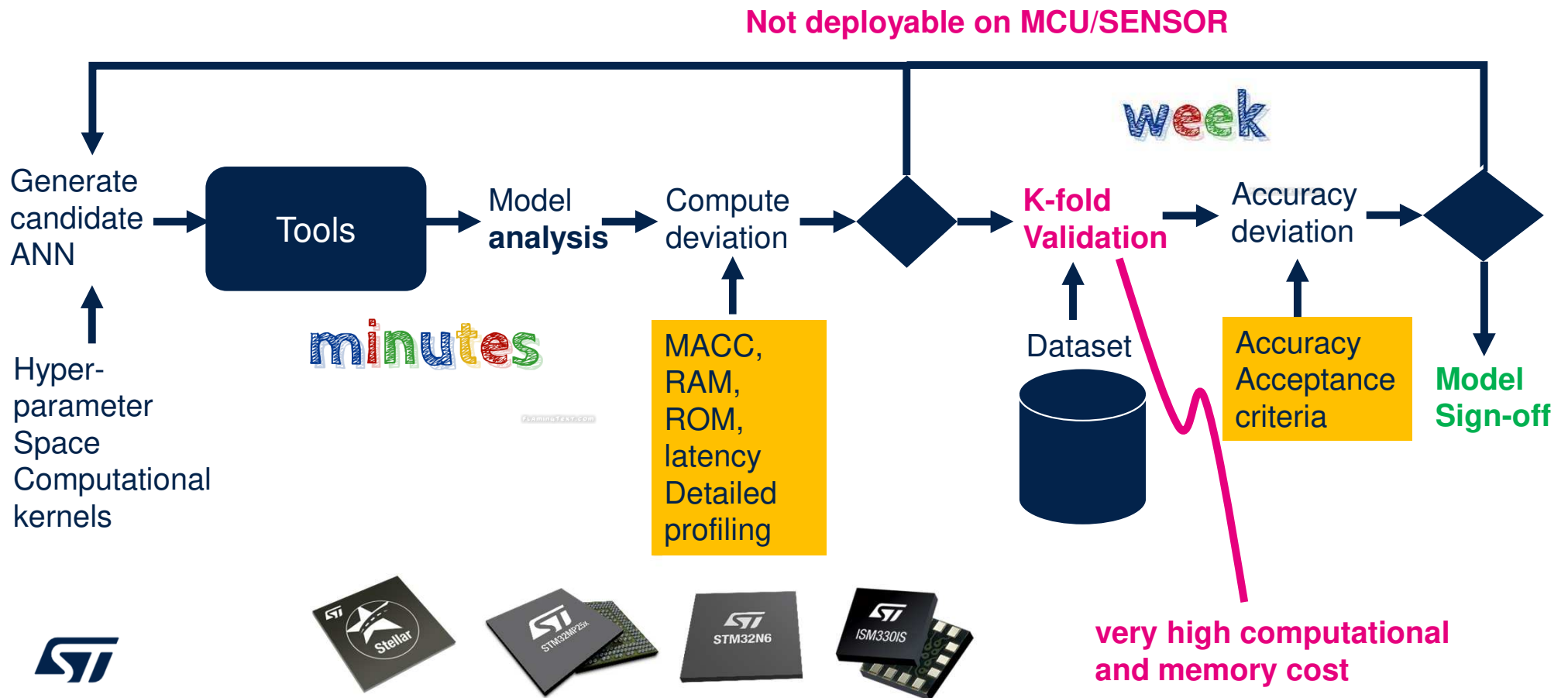


Deployment un-aware NAS/HPO



The implicit assumption

Deployment aware NAS/HPO



MCU Edge-AI Developer Cloud



NEW



STM32Cube.AI Dev Cloud

The brand-new online AI services front end for STM32



ST model zoo



Web GUI
+ REST API



STM32
Cube.AI

Core engine technology





Fixed Function AI

- Low power processing (< mW)
- React on data availability
- Provide assessments, insights

Reducing the Dimensionality of Data with Neural Networks
G. E. Hinton*

Krizhevsky, Alex et al.
"ImageNet classification with deep convolutional neural networks".

AI Is Taking Over the Cloud (MIT Tech Review)

T I N Y

TinyML products proliferation

2006

2012

2014

2017

2018

2019

2024

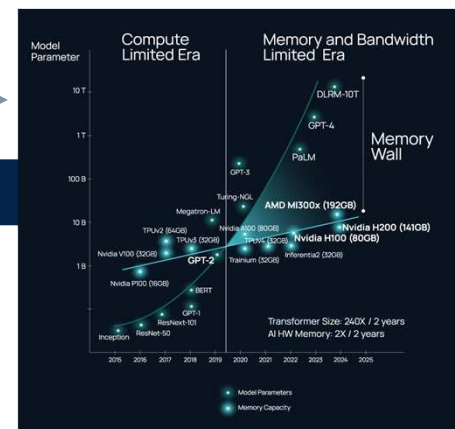
Generative Adversarial Networks
Ian J. Goodfellow
arXiv:1406.2661

Memory Wall Era starts
Gen AI drives the cloud

Edge AI scope expanded

Attention Is All You Need
arXiv:1706.03762

- Reason, plan, act upon the queries
- Multi modal
- Conversational HMI
- Energy efficient processing (100sT/W)



Generative AI

danilo.pau@st.com

Our technology starts with You



Find out more at www.st.com

© STMicroelectronics - All rights reserved.

ST logo is a trademark or a registered trademark of STMicroelectronics International NV or its affiliates in the EU and/or other countries.

For additional information about ST trademarks, please refer to www.st.com/trademarks.

All other product or service names are the property of their respective owners.

