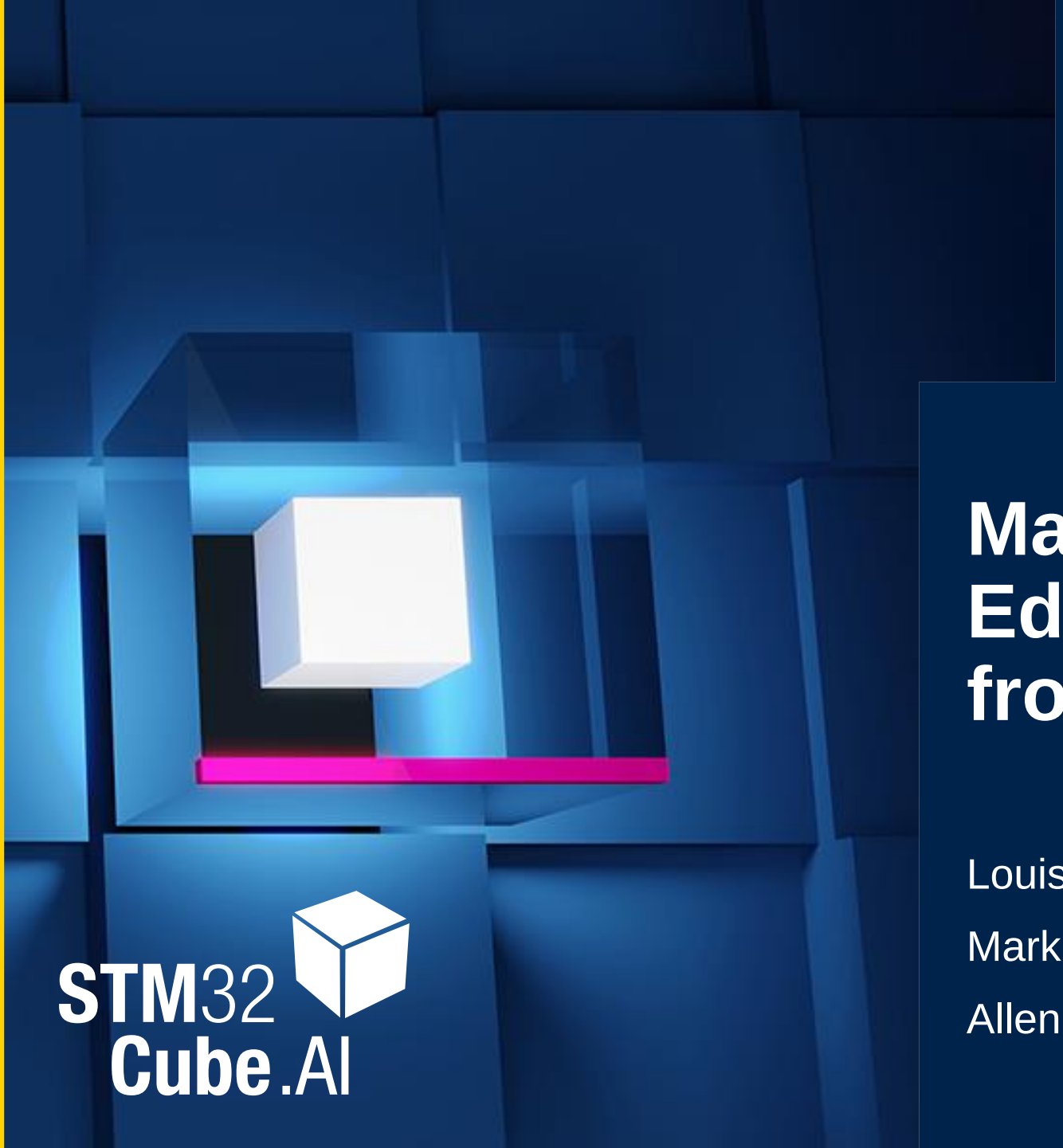


life.augmented

Machine Learning at the Edge: Neural Networks from GPU to MCU

Louis Gobin

Markus Mayr

Allen Ren

STM32 Cube.AI

STM32

Signals turning into data

Embedded applications will collect more data in the future

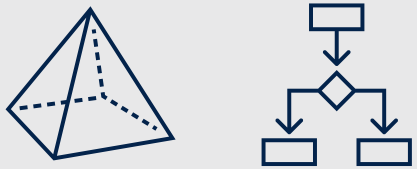
Growing demand for data-driven insights

Increasing use of sensors

Proliferation of IoT devices

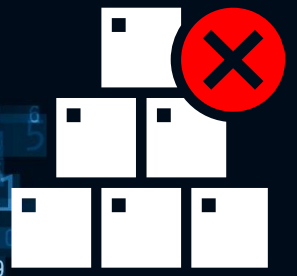
life.augmented

AI is offering the best approach to process this growing amount of data



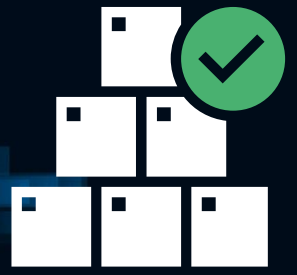
Algorithms and **predefined models** to analyze data and make predictions or decisions

Machine learning algorithms to automatically **learn** patterns and relationships **from the data**



Traditional approaches show their limitations:

- when dealing with **large datasets**
- when the **phenomena are too complex**



AI-based data processing offers a more flexible and powerful approach to analyzing and making decisions from large data collection

life.augmented

The raise of Edge AI

Ultra-low latency
Real-time applications

Reduced data transmission
Generate meaningful information

Enhanced privacy and security
No data sharing in the cloud

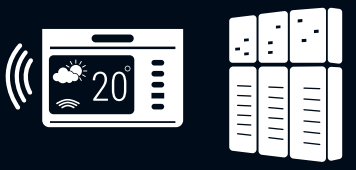
Sustainable on energy
Low-data / Low-power

Improved accuracy
analyze data from a wide range of sensors and sources

Edge AI will benefit to many application domains:

Industrial maintenance
Condition monitoring
Predictive maintenance

Control systems
From home heating systems to industrial machines



Internet of Things (IoT)
smart cities, smart buildings, connected homes, and industrial automation

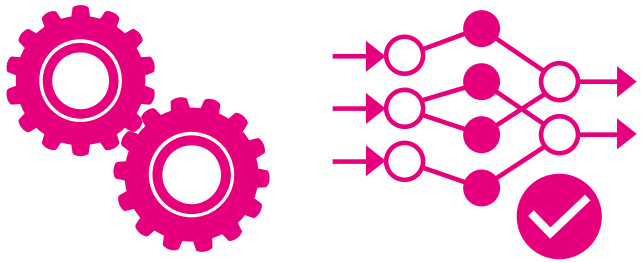
life.augmented

AI development workflow – ST software offering

1 Data Preparation

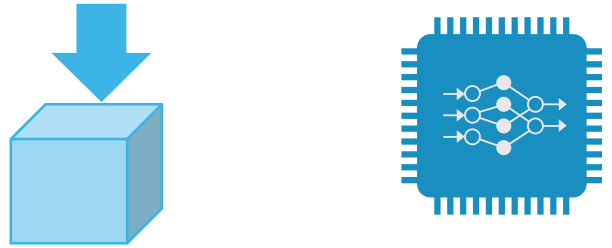
Data acquisition | Data processing

2 Data Science



Model selection and training | Model validation

3 Model Implementation



Model library creation | Model inference

NANOEDGE AI STUDIO — Automated Edge AI Software

STM32 Cube.AI — Edge AI toolkit

STM32

All STM32 MCUs

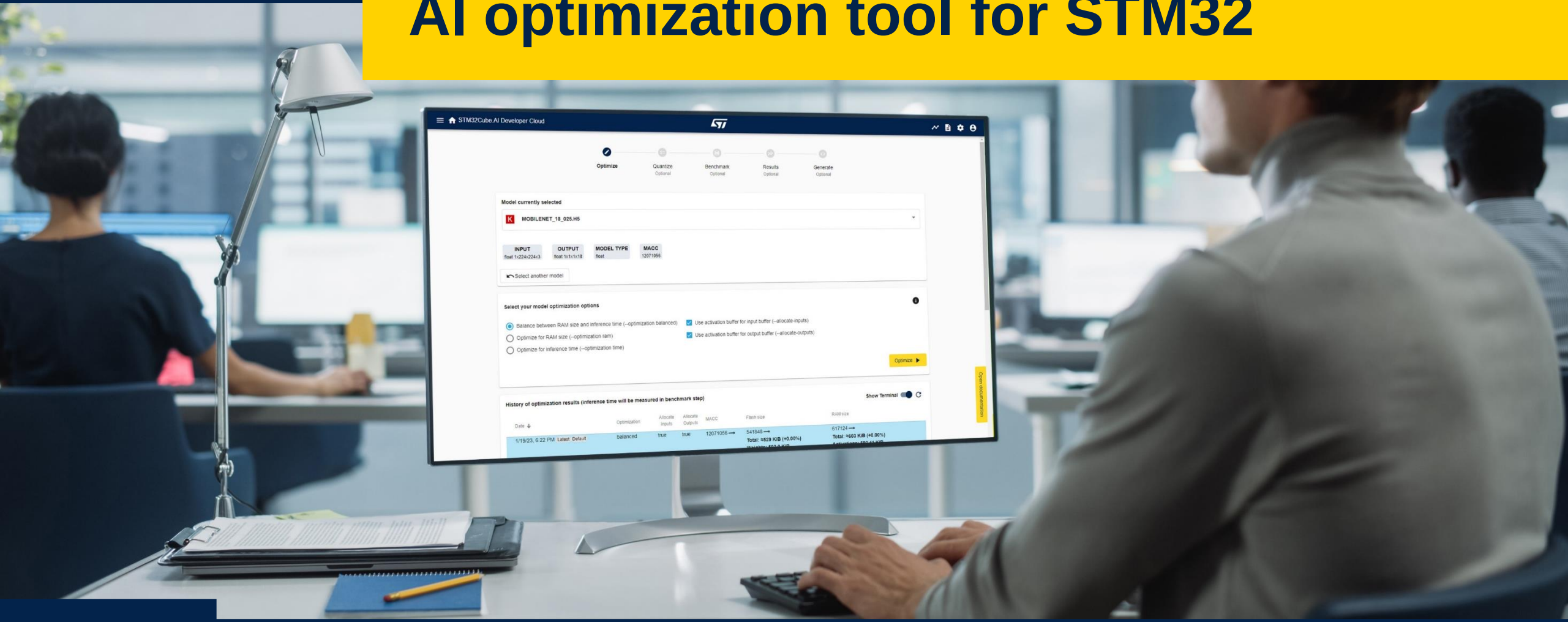
life.augmented

ST ecosystem ease your AI to reach production level

	STM32 Cube.AI **Edge AI toolkit** for model optimization on STM32	NANOEDGE AI STUDIO **Automated ML Software** for end-to-end Edge AI solution design on STM32
Key benefits	✓ Get optimized C-code from your trained model ✓ Desktop and online versions ✓ Benchmark service on remote hardware (online version) ✓ On-device performance validation	✓ The easiest way to integrate AI into your system ✓ Save resource and development cost ✓ Reach the highest performance with the automated model finder embedded in the tool
Application domain	All	All, except voice and vision
Business model	Free of charge	Free for prototyping on STM32 dev boards Production requires right of use

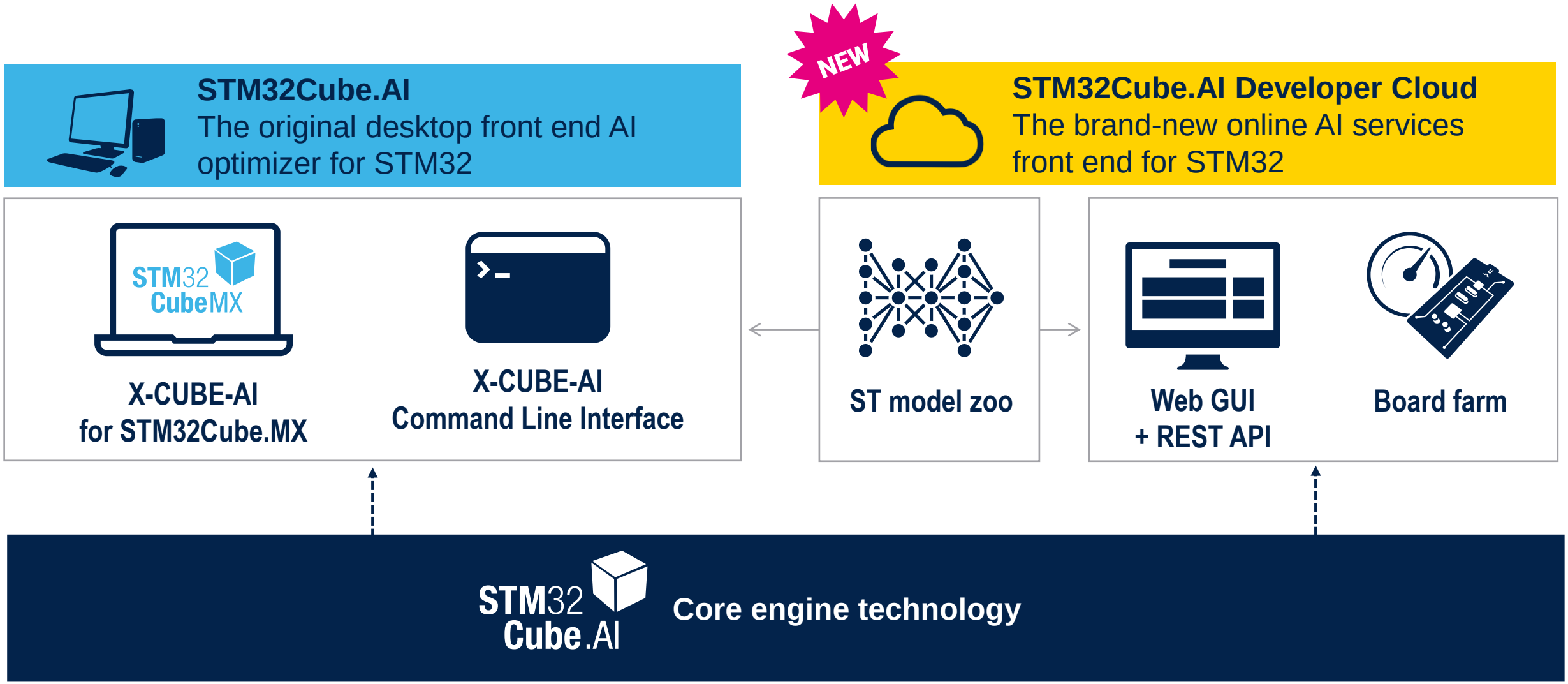
life.augmented

STM32Cube.AI
AI optimization tool for STM32



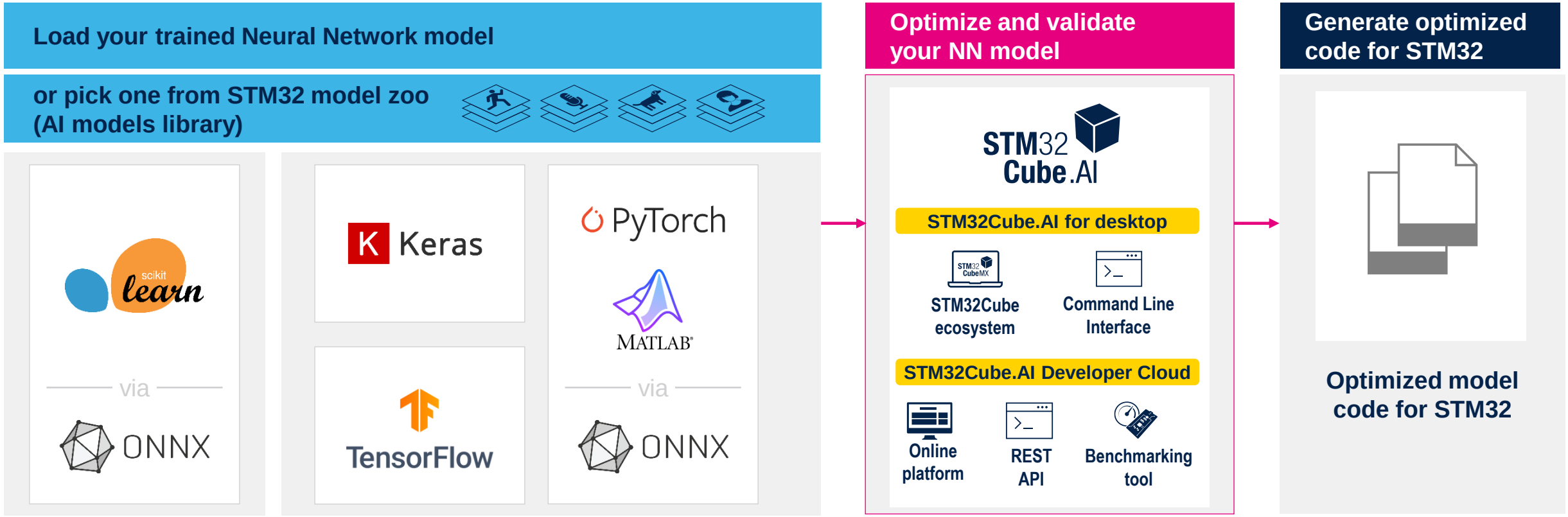
life.augmented

STM32Cube.AI overview



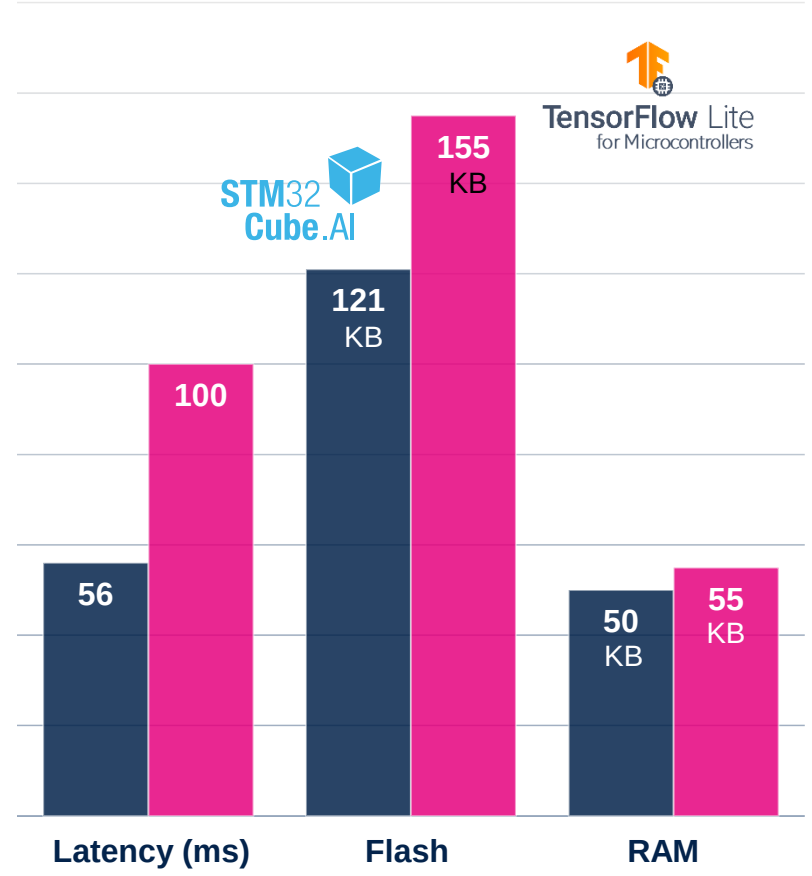
life.augmented

One tool – two versions to deploy AI on STM32

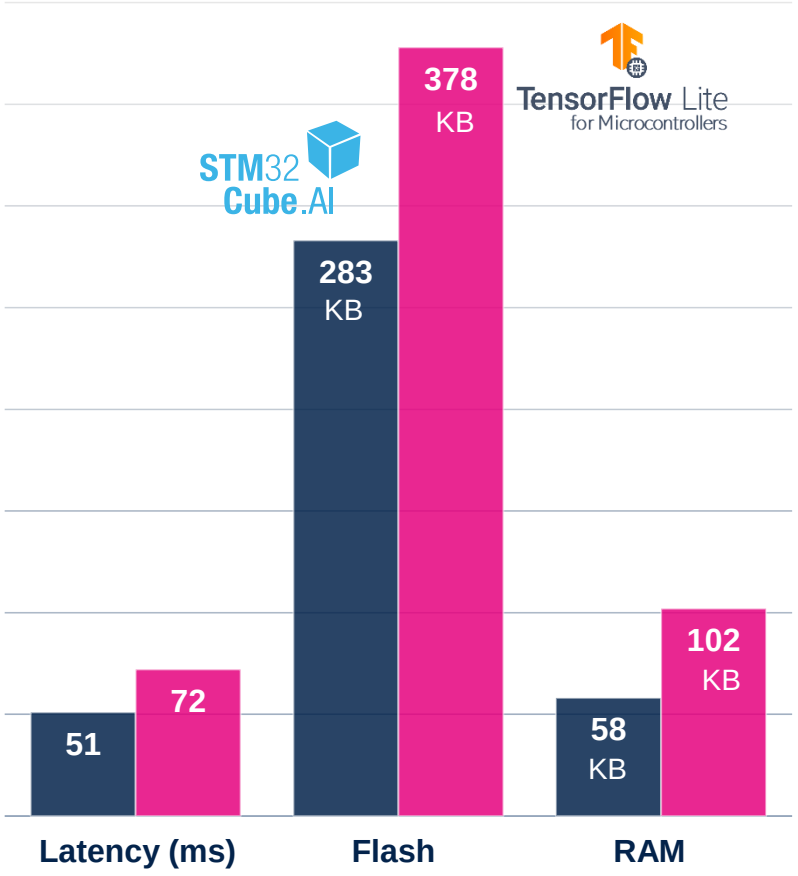


STM32Cube.AI
Get the best AI performance on STM32

Image Classif v1.0 MLPerf Tiny



Visual Wake Word v1.0 MLPerf Tiny



UP TO
60 %
faster inference time*

UP TO
20 %
space freed-up in FLASH and RAM*

HW Target: STM32H7A3
Flash: 2Mbyte
RAM: 1.4 Kbytes
Freq: 280 MHz

SW Version:
X-Cube.AI v 7.3.0
TFLm v2.10.0

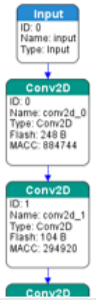
** versus TensorFlow Lite for microcontroller*

life.augmented

The 3 pillars of STM32Cube.AI

Graph optimizer

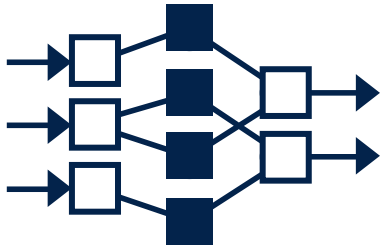
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding…
- Operator-level info to fine-tune memory footprint and computation

Quantized model support

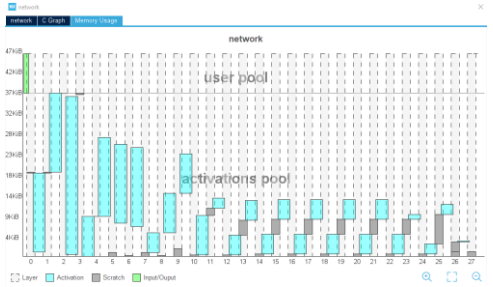
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8 or mixed-precision
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory footprint

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design



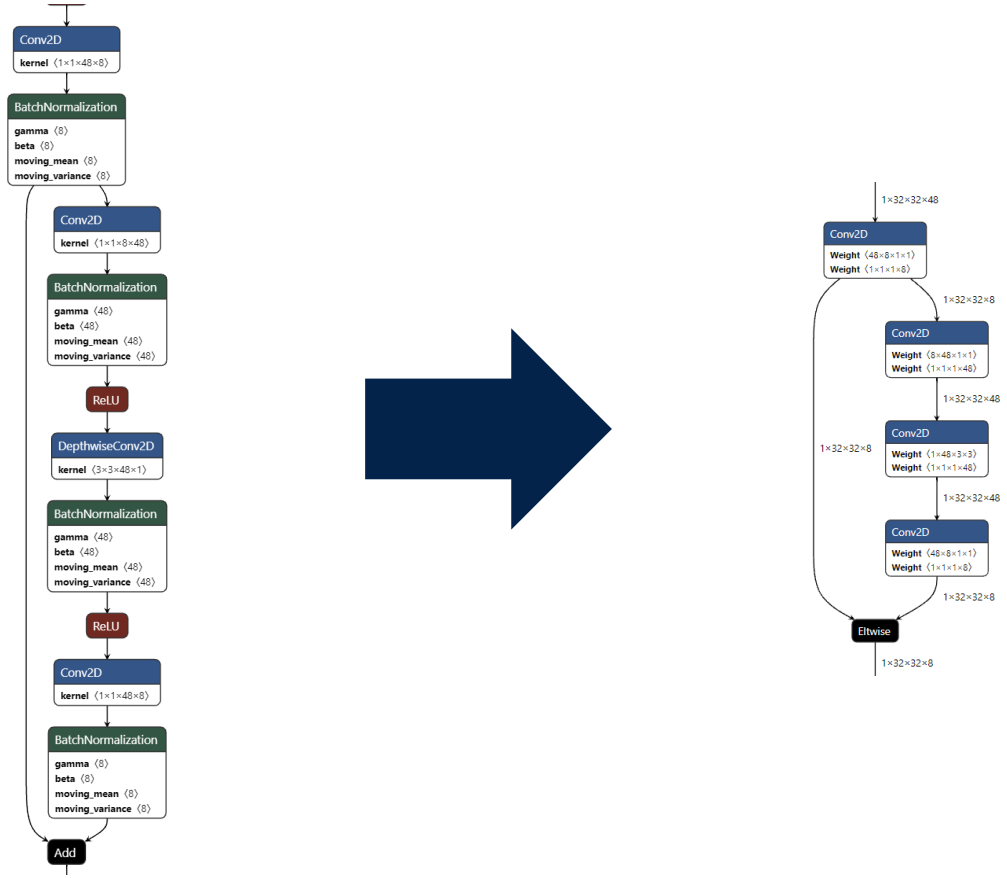
- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

life.augmented

Graph optimizer

Squeeze your graph to fit into an MCU!



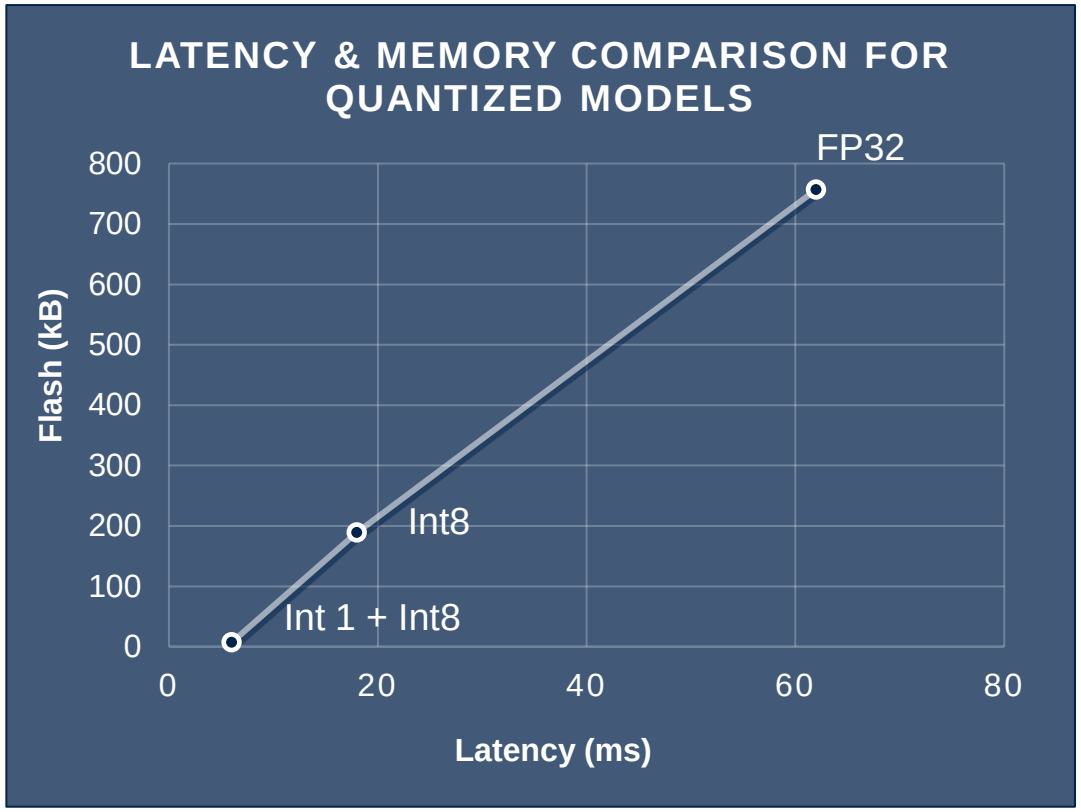
Fully automated process in the STM32Cube.AI workflow

- Your original graph is optimized at the very early stage for optimal integration into STM32 MCU/MPU
- Loss-less conversion

life.augmented

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time



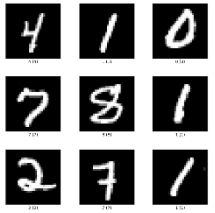
STM32Cube.AI support quantized Neural Network models with **all parameter formats**:

- FP32
- Int8
- Mixed binary Int1 to Int8 (Qkeras*, Larq.dev*)

**Please contact edge.ai@st.com to request the relevant version of STM32Cube.AI*

HW Target: NUCLEO-STM32H743ZI2
Model: Low complexity handwritten digit reading
Freq: 480 MHz
Accuracy: >97% for all quantized models

Tested database: MNIST dataset

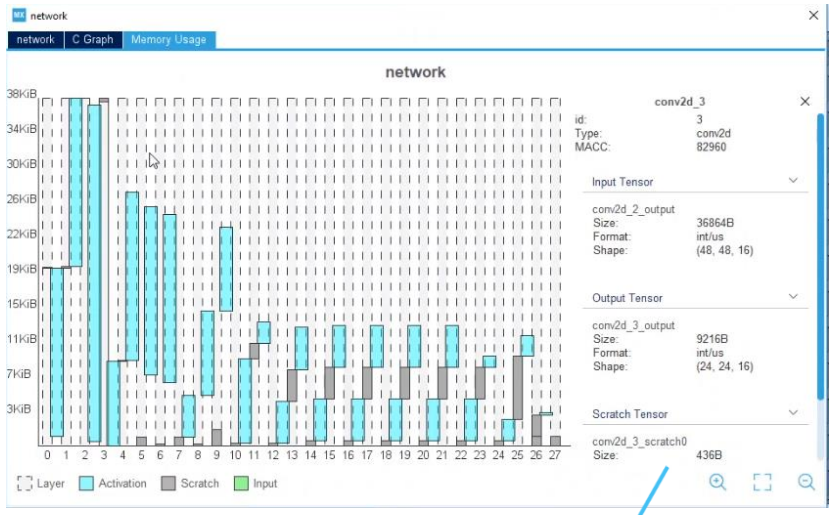


MNIST dataset

life.augmented

Memory optimizer

Optimize performance easily with the memory allocation tool



Model RAM consumption per layer

- Easily identify most critical layers

Model memory allocation

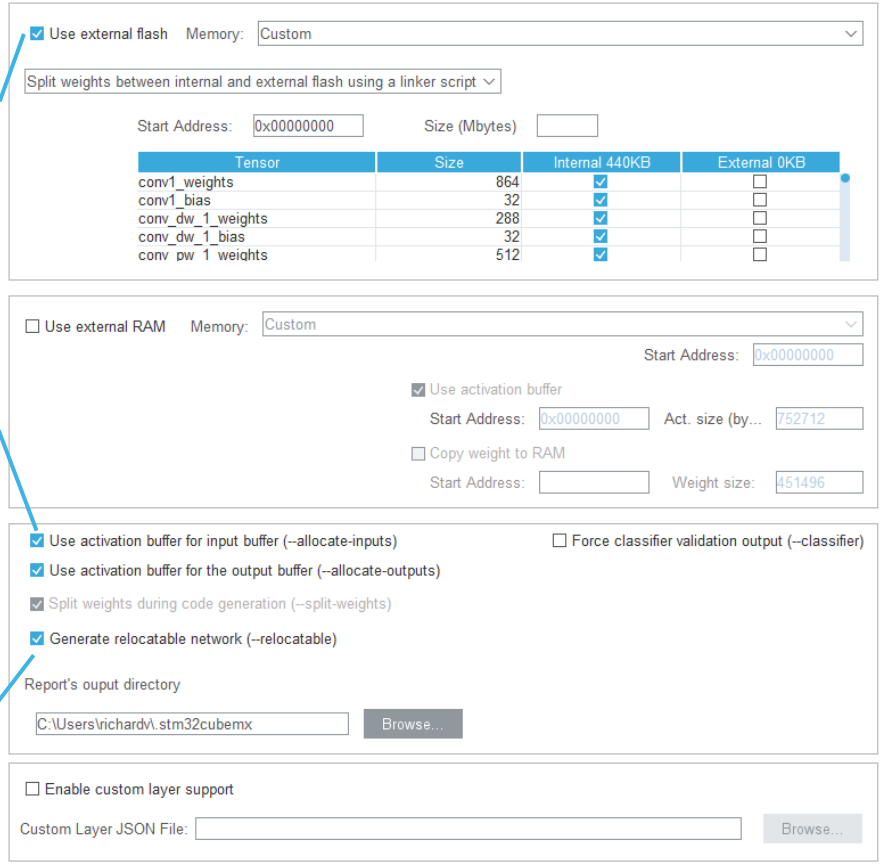
- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

Re-use model input buffer to store activation data*

- Minimize RAM requirements

Relocatable network

- A separate binary is generated for the library and the network to enable standalone model upgrade



** Requires input and activation buffers in same memory*

life.augmented

Performance benchmarking made simple
STM32Cube.AI Developer Cloud

The unique possibility to evaluate the performance of models remotely, on real STM32 boards

Get the real inference time from optimized models running on STM32

Benchmark models on a large variety of STM32 boards

- Find the most appropriate board for your application

NEW

Get access to the most recent devices

- A board farm is constantly updated with the latest available boards

life.augmented

Start with edge AI optimized models STM32 model zoo

A collection of application-oriented models optimized for STM32

Human activity

Motion Sensing

Image classification

Computer vision

Audio event detection

Audio classification

Object detection

Computer vision

Hosted on Github

Model training scripts

- Scripts to generate and validate

Application code example

- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

life.augmented

We provide everything to kick off your project

Design documentation

Getting started

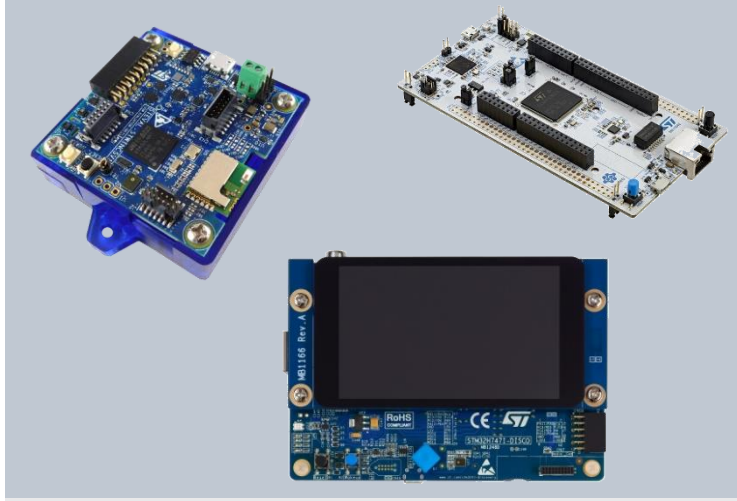
Be guided step-by-step to learn STM32 ecosystem

Development zone

Get started on application development and project sharing

- **Wiki by ST** is a great forum to learn and start developing AI on STM32!
- Videos of application examples
- Massive Open Online Course (MOOC)

Hardware and software tools



- Evaluation platforms for STM32 MCU/MPU
- Extra sensor boards
- Full software suite

Support & Updates

- **ST Community**: STM32 ML & AI group
- Distributor certified FAE
- Support center
- Newsletter

life.augmented

What's new in STM32Cube.AI v8.0.0?

v8.0.0

Bringing higher degree of versatility with STM32Cube.AI

ONNX quantized models support

Introducing the support of **ONNX Tensor-oriented file format (QDQ):**

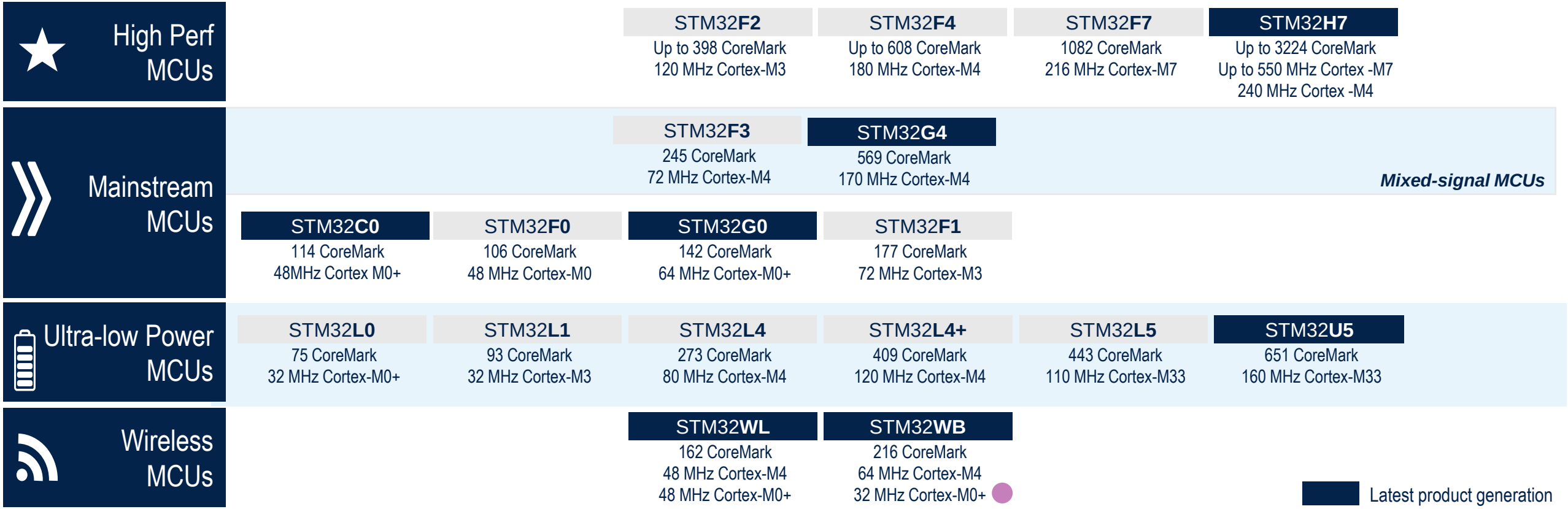
Initial model converted in ONNX QDQ can be:
- ONNX models
- Quantization-Aware training (QAT) models from Tensorflow or exported from PyTorch
- Quantized models converted from TFLite and other frameworks

Up-to-date and improved code generation

- **Support for TensorFlow 2.11 models**
- **Support Keras.io 2.11**
- **Support ONNX Runtime 1.12.1**
- New kernel performance improvements.

Making Edge AI accessible to all STM32 portfolio

Take advantage of STM32Cube.AI on all STM32 series



life.augmented

Our technology starts with You

Find out more at stm32ai.st.com

life.augmented