life.augmented

Introducing
ST Neural-ART Accelerator

Enabling high-end, power-efficient edge AI performance on MCUs.

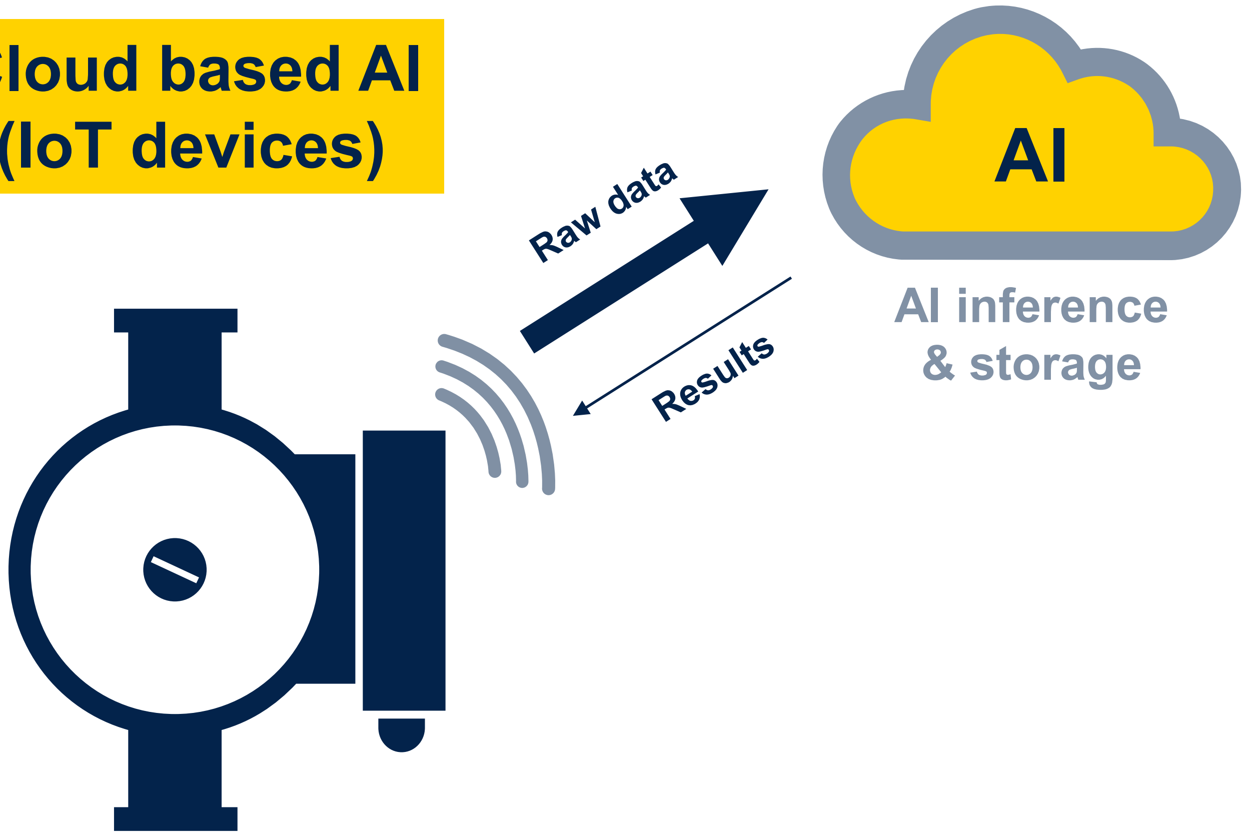
Redefining product greatness

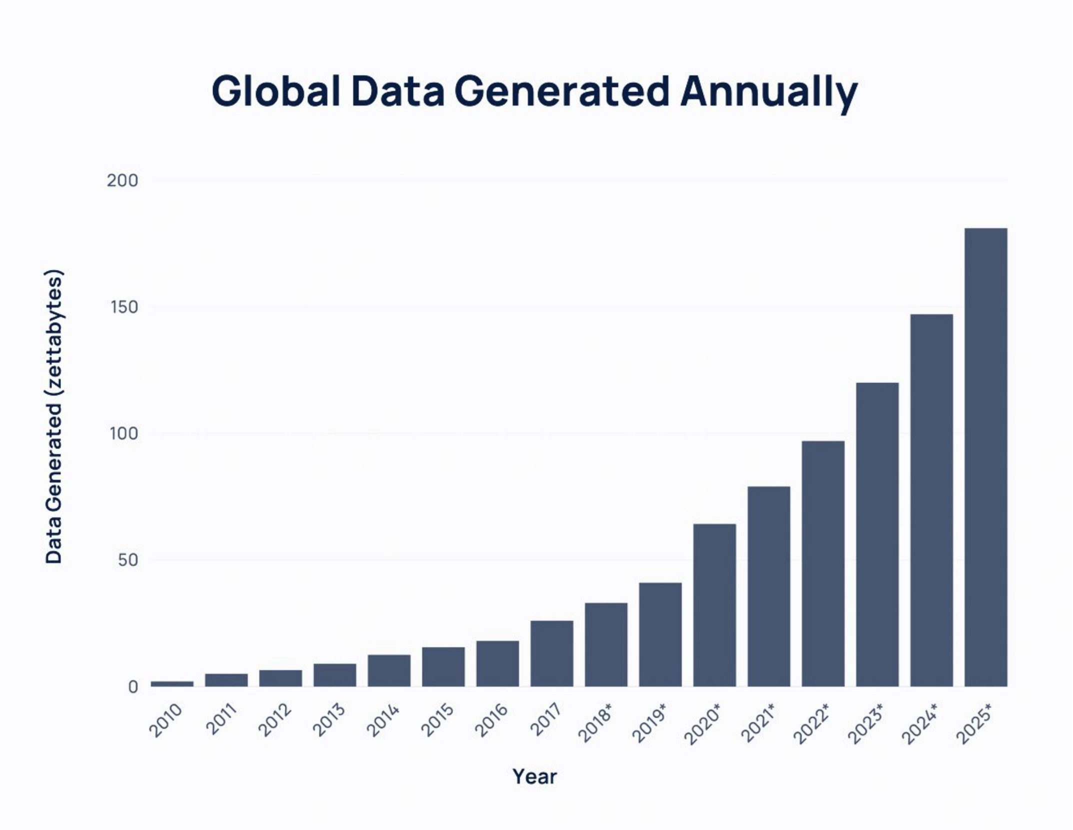
AI drives **smarter products** and **new business models**

life.augmented

Cloud processing for AI & IoT: Generating a tsunami of data

Cloud based AI (IoT devices)



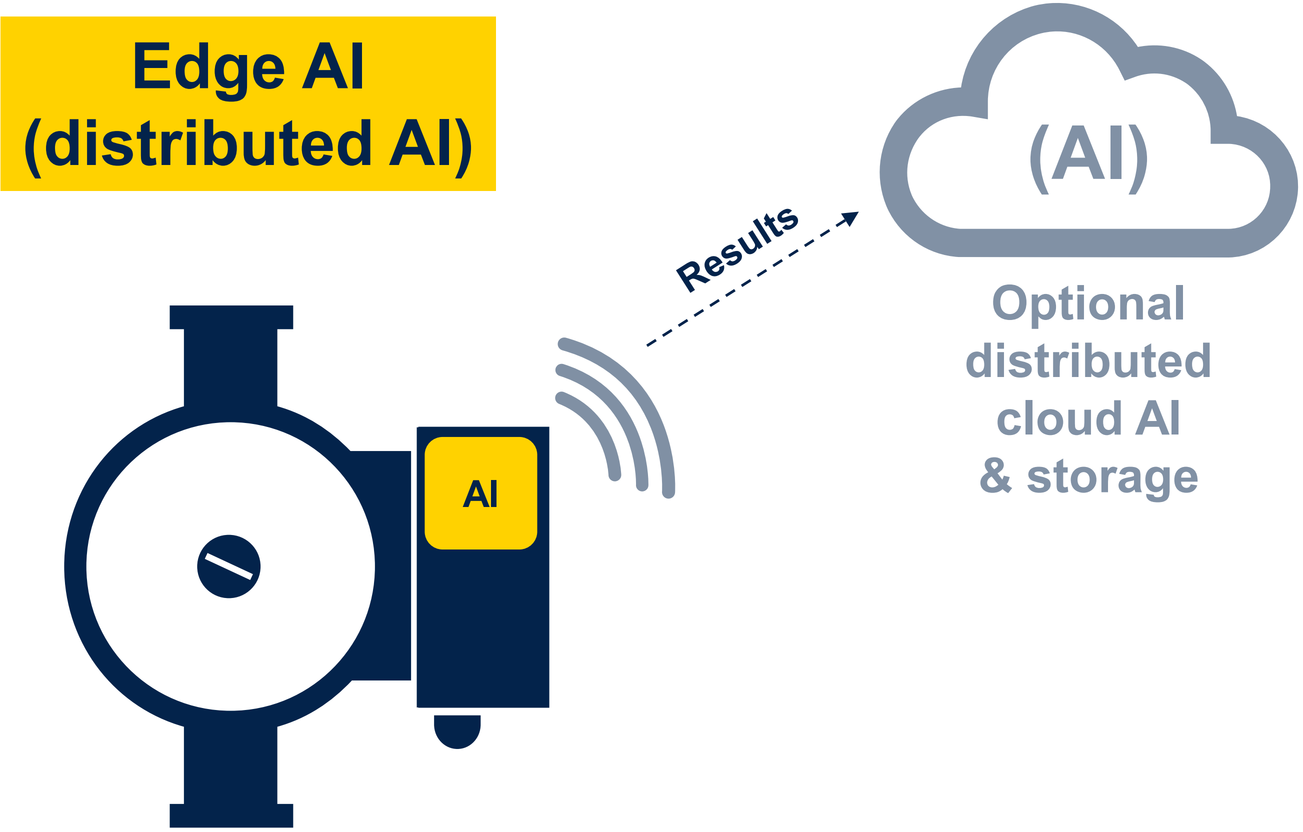


120 ZetaBytes data generated in 2024
> 180 ZetaBytes in 2025

Source: explodingtopics.com

life.augmented

The rise of edge AI: AI at device level

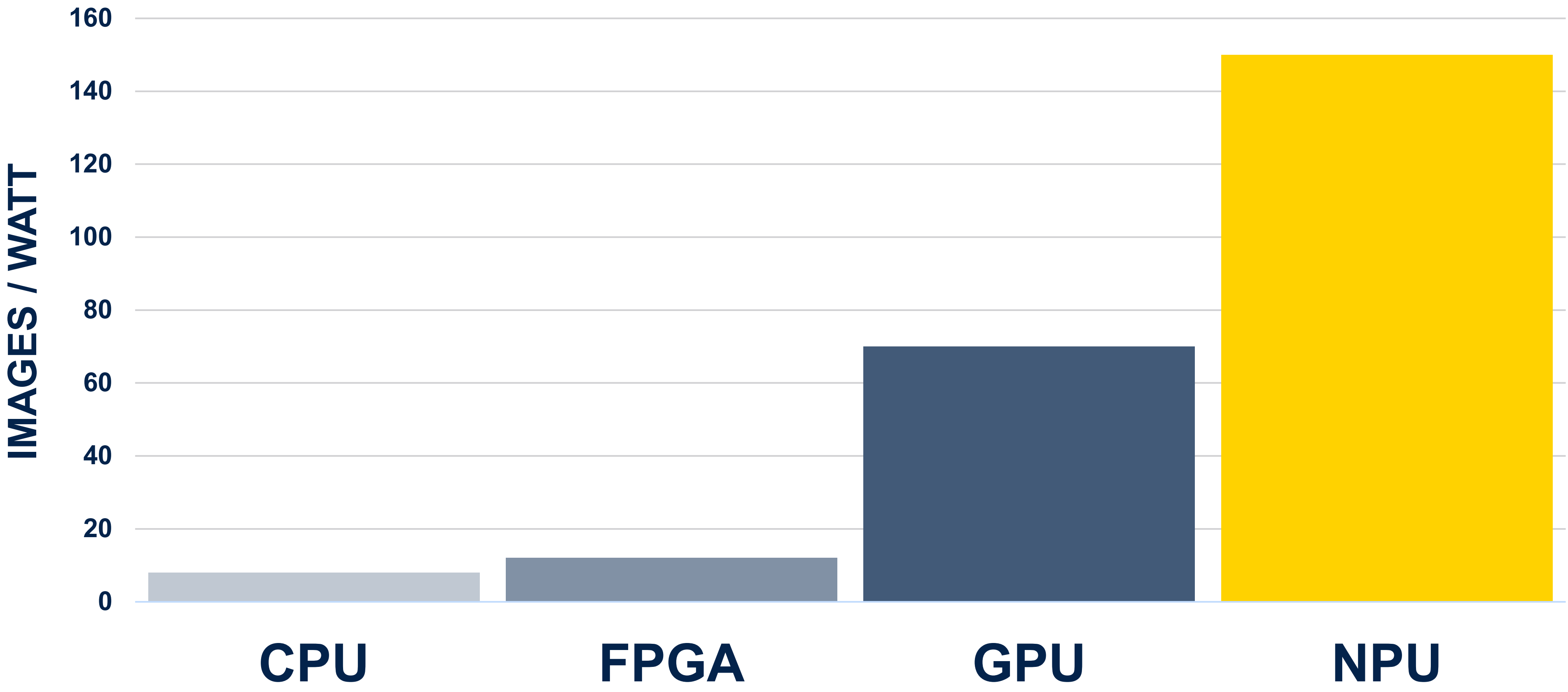


Edge AI
(distributed AI)

Optional
distributed
cloud AI
& storage

life.augmented

Edge AI acceleration requires new architectural solution: the NPU



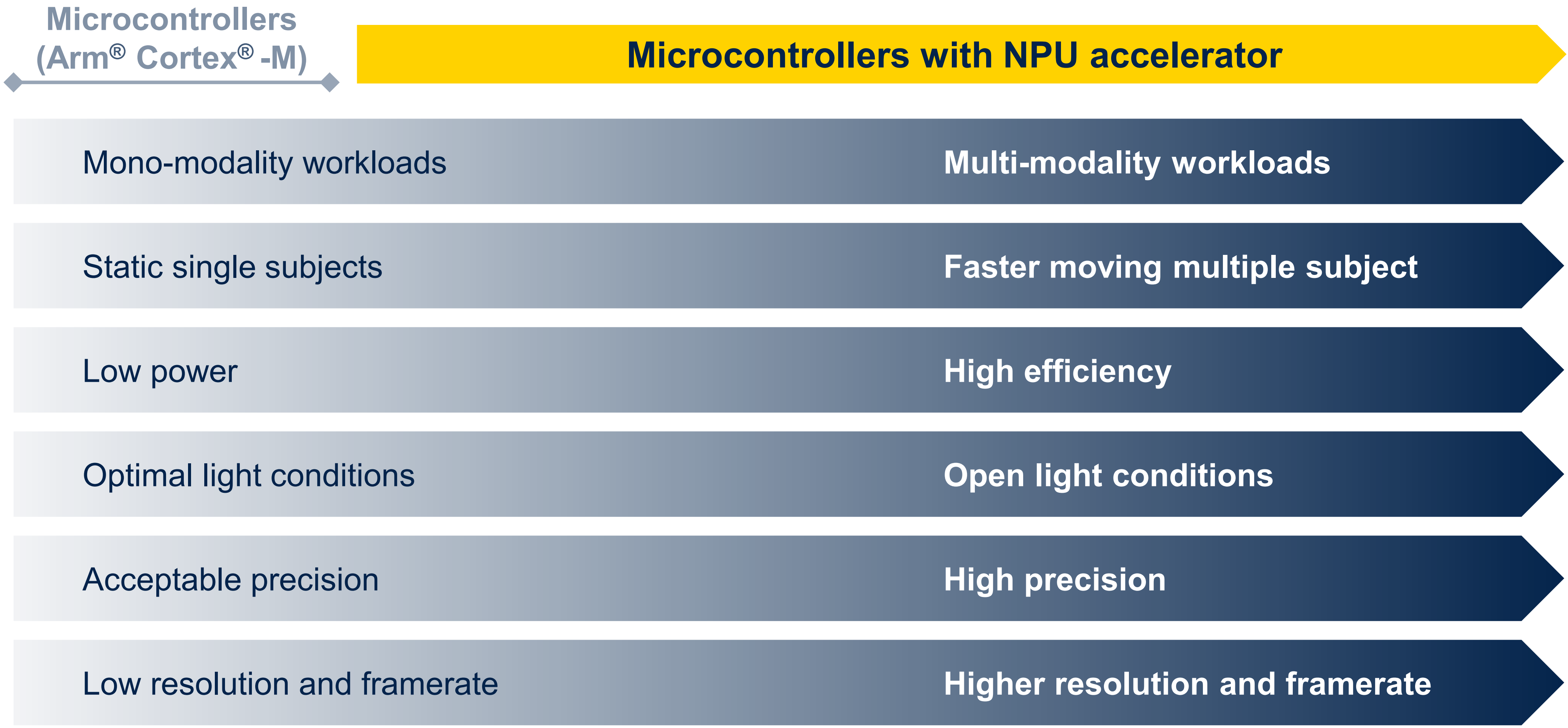
GPU: graphic accelerator

NPU: neural processing unit (AI accelerator)

Source W. Dally

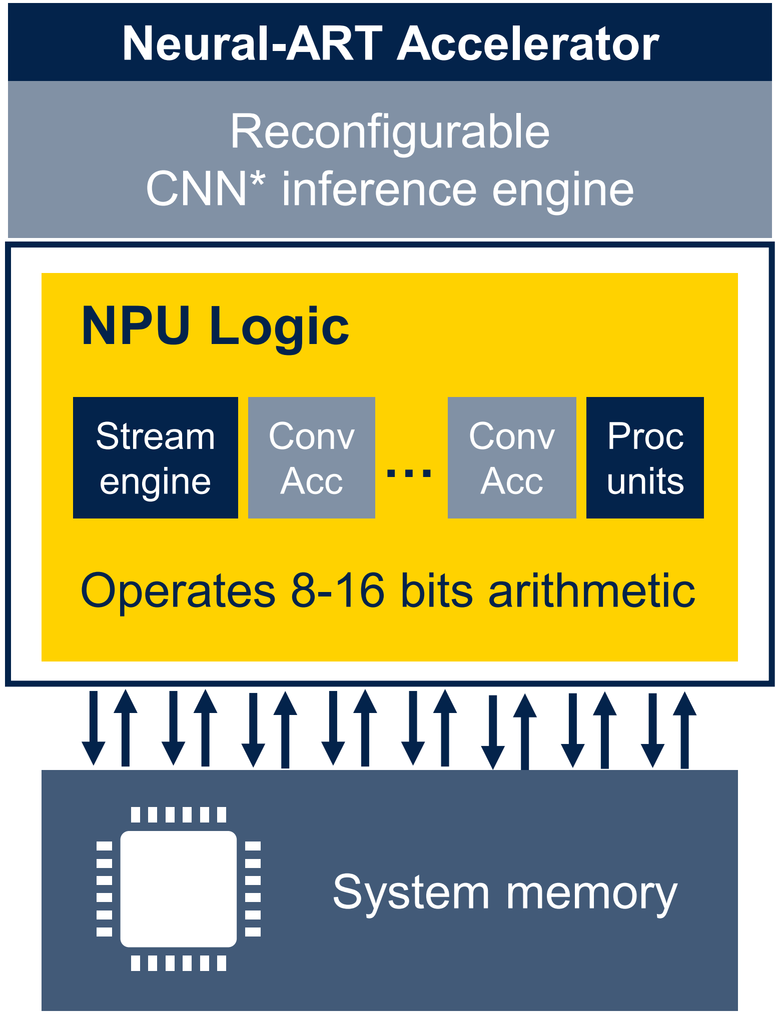
From DMIPS to TOPS, the paradigm shift
Opening a new range of embedded AI applications

- Object segmentation localization
- Pose estimation
- Object classification
- Speech recognition
- Sound analysis
- Face/people detection
- Wake word
- Time series classification
- Anomaly detection



Microcontrollers (Arm® Cortex® -M)	Microcontrollers with NPU accelerator
Mono-modality workloads	**Multi-modality workloads**
Static single subjects	**Faster moving multiple subject**
Low power	**High efficiency**
Optimal light conditions	**Open light conditions**
Acceptable precision	**High precision**
Low resolution and framerate	**Higher resolution and framerate**

Neural-ART Accelerator architecture overview



- **A paradigm shift** from the Von Neumann architecture towards a flexible, dedicated dataflow **stream processing engine.**
- Hardware acceleration for a **wide range of neural network architectures.**
- **Embedded security** to protect assets.
- **Seamless integration** into the MCU backbone via **two 64-bit AXI** interfaces.
- **Configurable** from 72 MACs to 2304 MACs.
- Achieves **up to 4.6 TOPS** at **1 to 5 TOPS/W****

** Convolutional neural network*
*** May vary according to technology node*

life.augmented

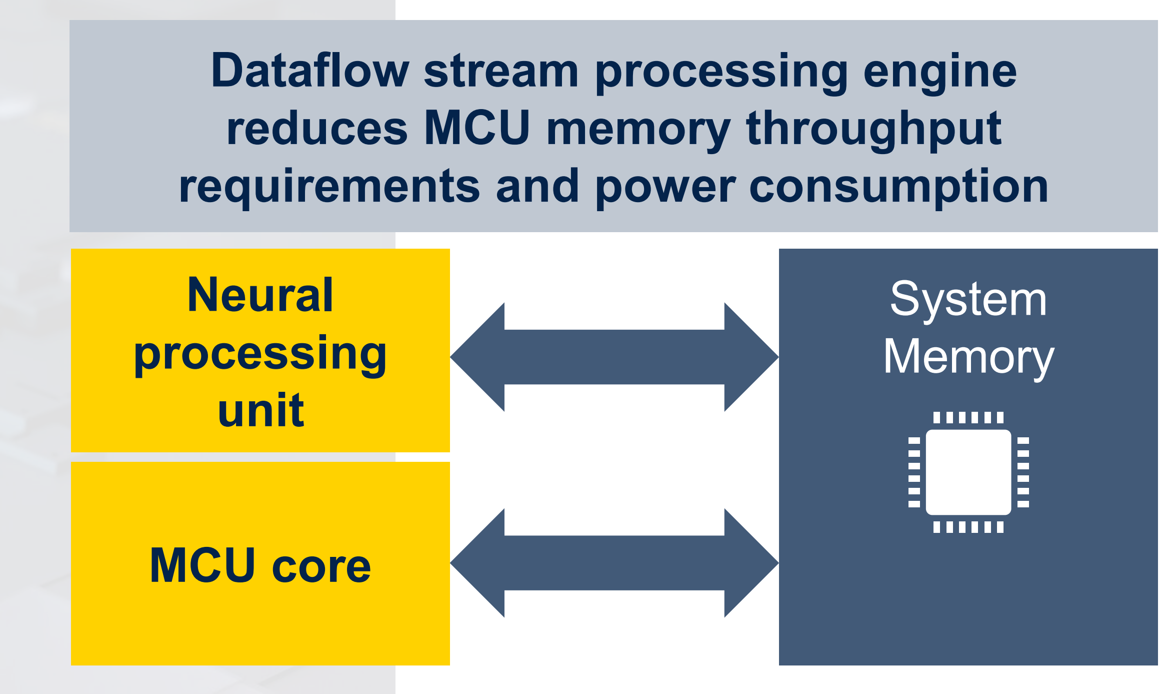
Neural-ART Accelerator in STM32N6 MCU

600x

ML performance uplift*

Dedicated embedded neural processing unit

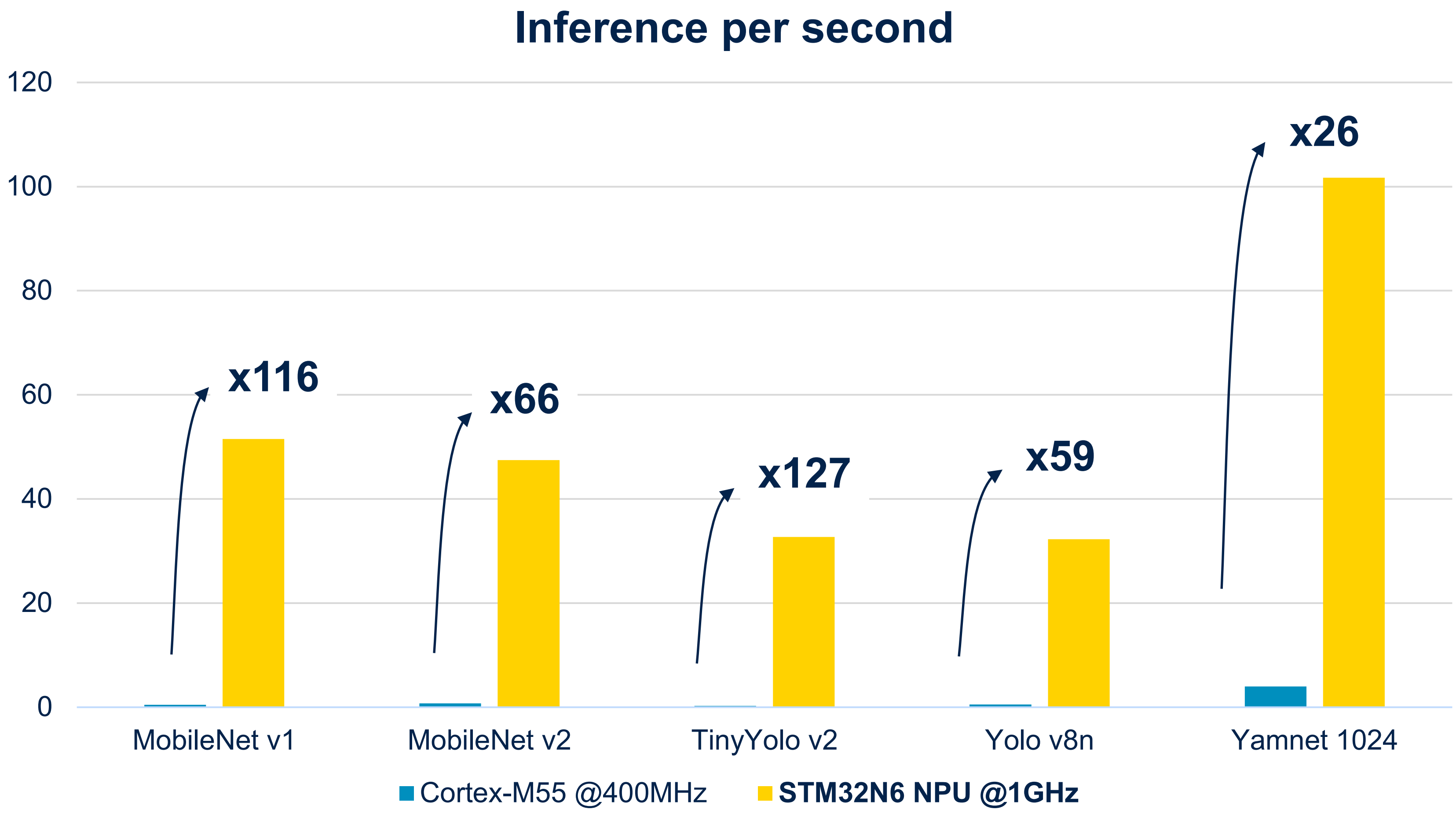
- **600 GOPS**
- **3 TOPS/W** power consumption
- Cache memory to optimize external memory access



Dataflow stream processing engine reduces MCU memory throughput requirements and power consumption

* 600 GOPS NPU vs 1 GOPS NN peak processing capabilities on STM32H7

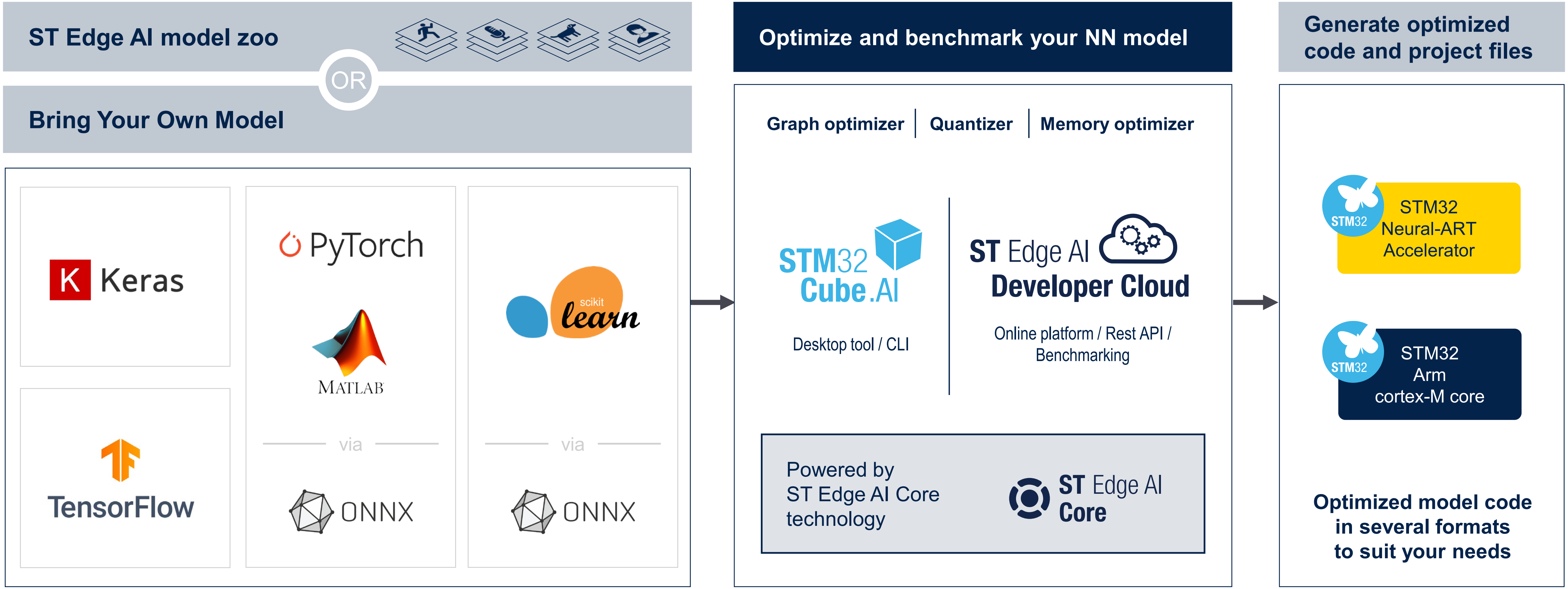
Neural-ART Accelerator provides a huge performance leap for AI inference



- **MobileNet v1:** image classification
- **MobileNet v2:** image classification
- **TinyYolo v2:** object detection
- **Yolov 8n :** object detection
- **Yamnet 1024:** audio recognition

life.augmented

Seamless integration with existing software ecosystem



ST Edge AI model zoo

OR

Bring Your Own Model

Keras

TensorFlow

PyTorch

MATLAB

via

ONNX

scikit learn

via

ONNX

Optimize and benchmark your NN model

Graph optimizer | Quantizer | Memory optimizer

STM32 Cube.AI

Desktop tool / CLI

ST Edge AI Developer Cloud

Online platform / Rest API / Benchmarking

Powered by ST Edge AI Core technology

ST Edge AI Core

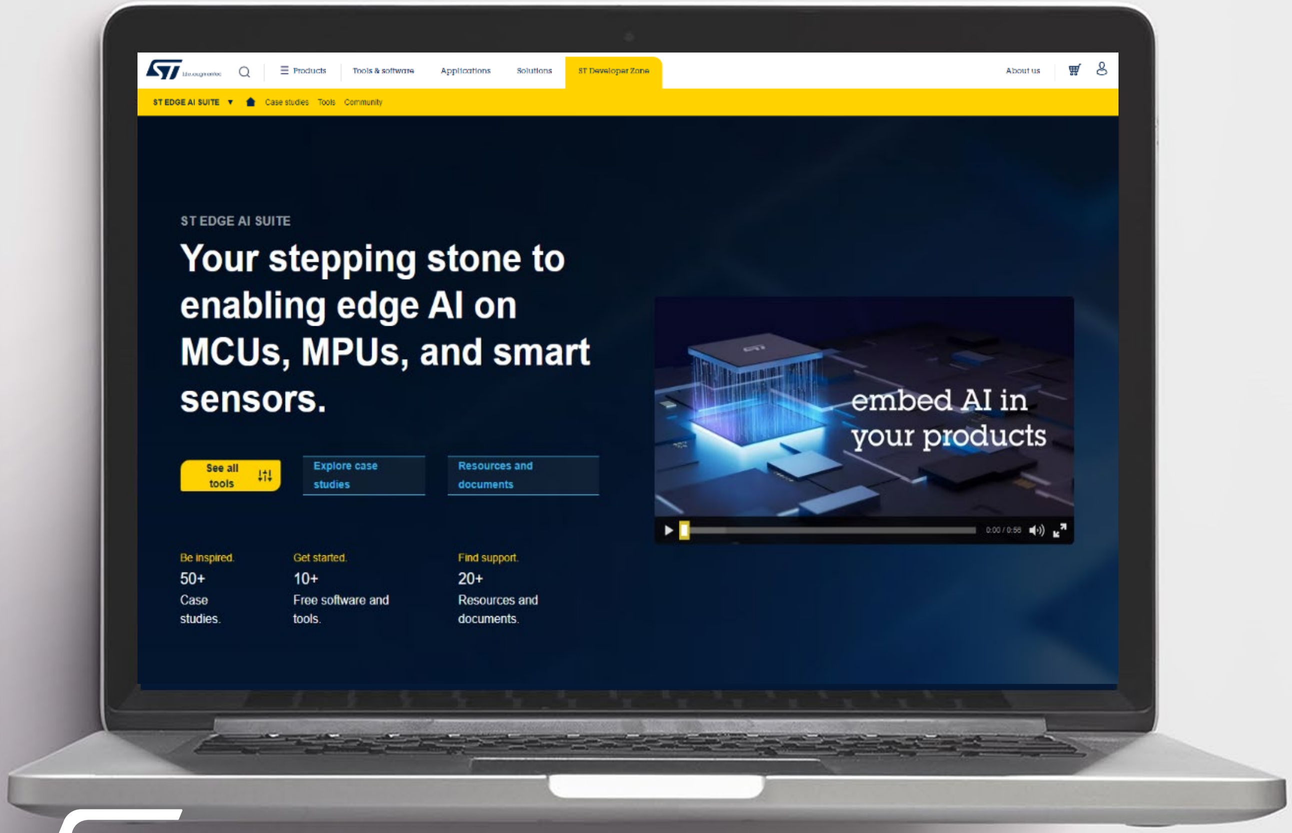
Generate optimized code and project files

STM32 Neural-ART Accelerator

STM32 Arm cortex-M core

Optimized model code in several formats to suit your needs

Reach the full potential of your application



ST EDGE AI SUITE

Your stepping stone to enabling edge AI on MCUs, MPUs, and smart sensors.

Be inspired.
50+
Case studies.

Get started.
10+
Free software and tools.

Find support.
20+
Resources and documents.

ST Edge AI Suite

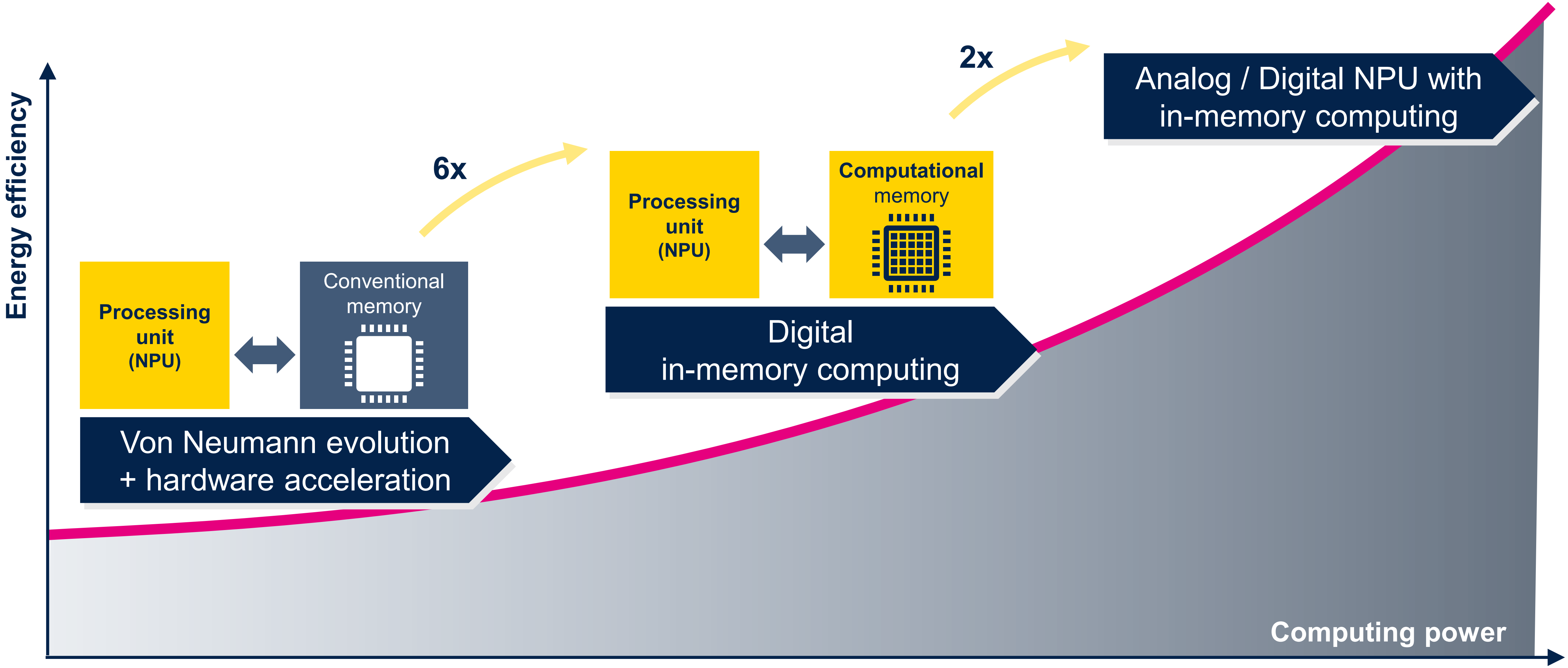
50+ case studies

10+ free software tools

Unified ST Edge AI core technology

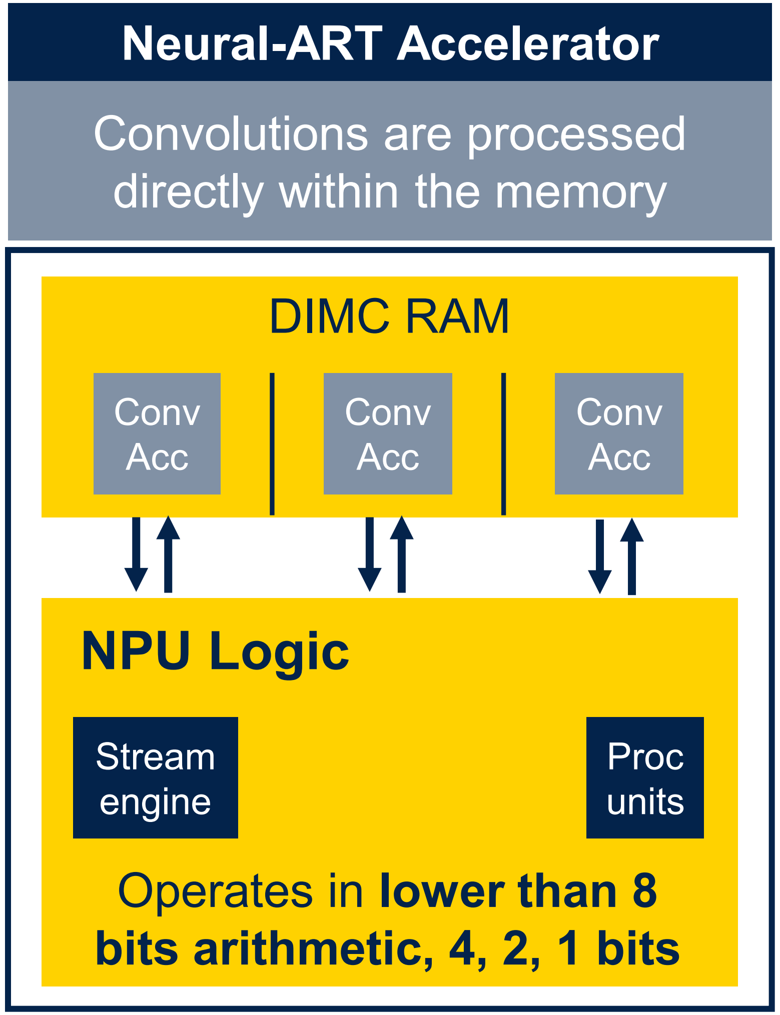
life.augmented

Neural-ART Accelerator outlook


Energy efficiency
Processing unit (NPU)
Conventional memory
Von Neumann evolution + hardware acceleration
6x
Processing unit (NPU)
Computational memory
Digital in-memory computing
2x
Analog / Digital NPU with in-memory computing
Computing power

life.augmented

More about next generations Neural-ART Accelerator


Neural-ART Accelerator
Convolutions are processed directly within the memory
DIMC RAM
Conv Acc
Conv Acc
Conv Acc
NPU Logic
Stream engine
Proc units
Operates in **lower than 8 bits arithmetic, 4, 2, 1 bits**

- **In-memory cell arithmetic** significantly reduces data transfer with memory hence power consumption.
- Achieves up to **6x improvement in TOPS** and **TOPS/W.**
- Support **advanced quantization (4, 2, 1 bit)** for further performance improvements.
- Ensures **seamless workflow integration** in the continuity of Gen 1.

Our technology starts with You

Read the whitepaper to know more

life.augmented