

life.augmented

## Get to know ST

Danilo PAU

Technical Director

IEEE, ST and AAIA Fellow,

APSIPA Life Member, Sigma XI

System Research and Applications

Alghero, Sept 17, 2024

# Top 5 Semiconductor foundries

intel®



Q1/2024 market share  
89.2%.

• TSMC	61.7%
• Samsung	11%
• GlobalFoundries	5.1%
• UMC	5.7%
• SMIC	5.7%

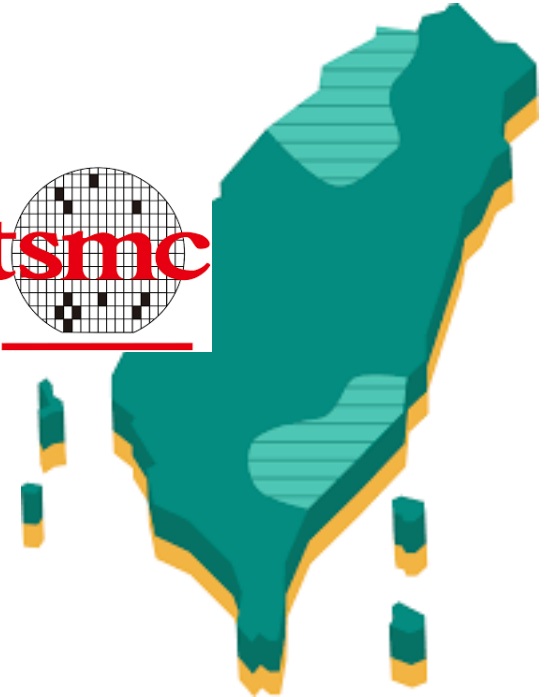
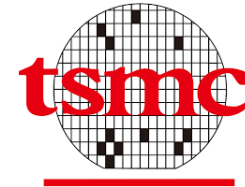
**SMIC**



**ST**  
life.augmented

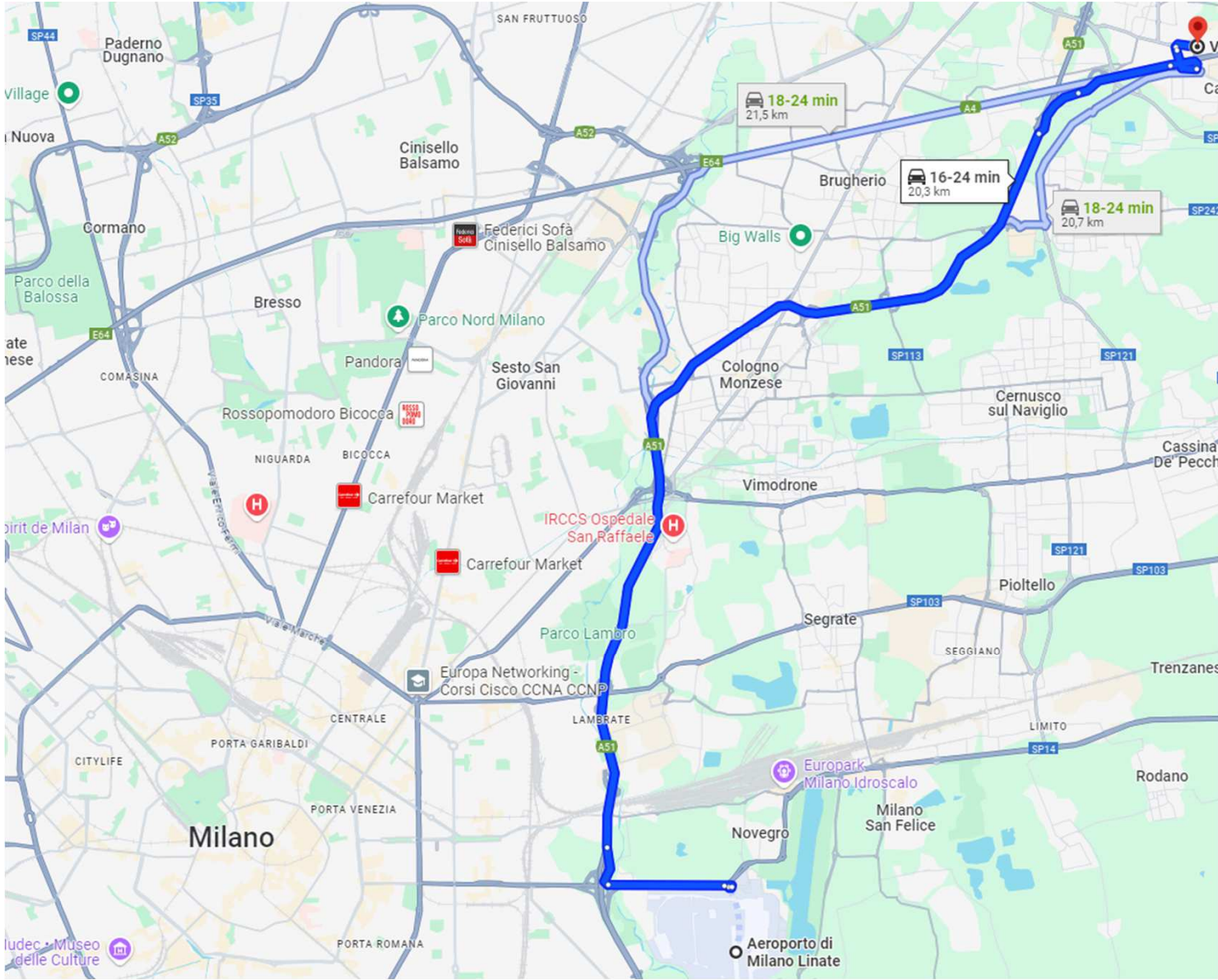
**SAMSUNG**

SEMICONDUCTORS



**UMC**





## IEEE MILESTONE

### Multiple Silicon Technologies on a Chip 1985

SGS (now STMicroelectronics) pioneered the super-integrated silicon-gate process combining Bipolar, CMOS, and DMOS (BCD) transistors in single chips for complex, power-demanding applications. The first BCD super-integrated circuit, named L6202, was capable of controlling up to 60V-5A at 300 kHz. Subsequent automotive, computer, and industrial applications extensively adopted this process technology, which enabled chip designers flexibly and reliably to combine power, analog, and digital signal processing.

May 2021

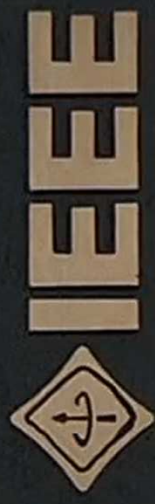


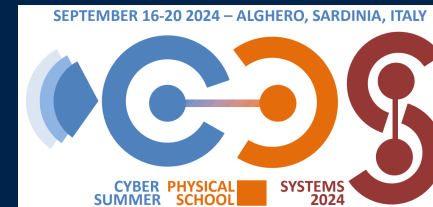
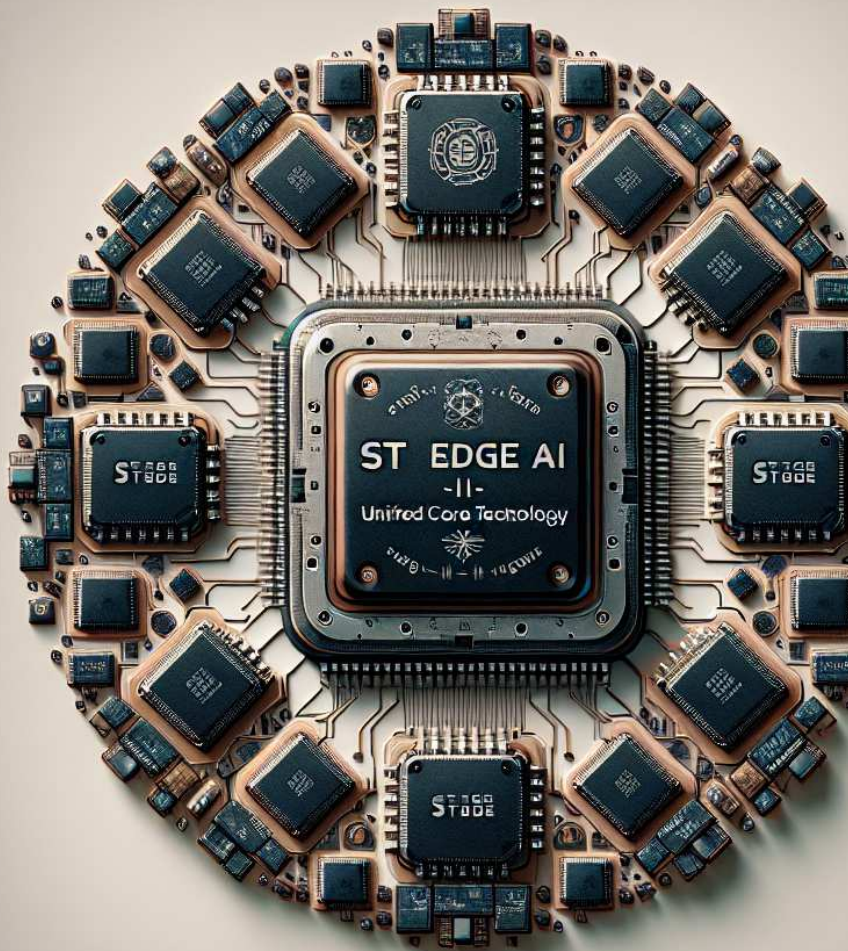
## IEEE MILESTONE

### MPEG Multimedia Integrated Circuits, 1984-1993

Beginning in 1984, Thomson Semiconducteurs (now STMicroelectronics) developed multimedia integrated circuits, which accelerated Moving Picture Experts Group (MPEG) standards. By 1993, MPEG-2 integrated decoders - including innovative discrete cosine transform (developed jointly with ENST, now Telecom ParisTech), bitstream decompression, on-the-fly motion compensation, and display unit - were announced in one silicon die: the ST13500. Subsequent MPEG-2 worldwide adoption made compressed full-motion video and audio inexpensive and available for everyday use.

September 2023





## Tiny Machine Learning through AI Unified Core Technology

Danilo PAU

Technical Director

IEEE, ST and AAIA Fellow,

APSIPA Life Member, Sigma XI

System Research and Applications

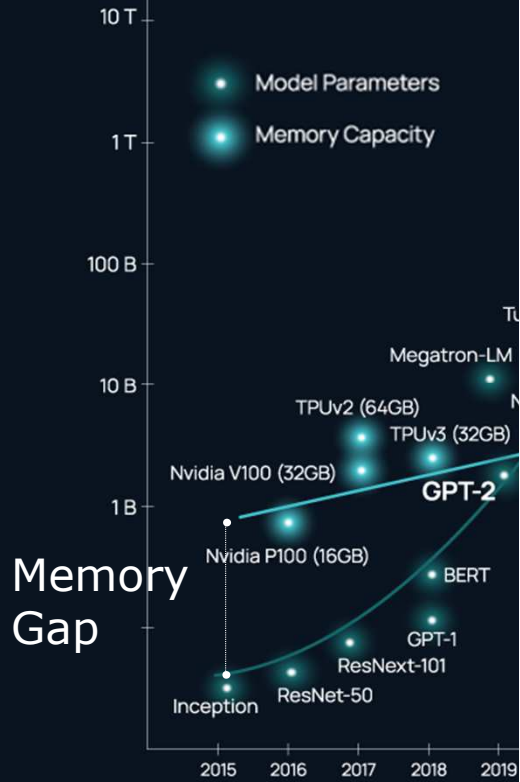
Alghero, Sept 17, 2024

# The Reasons Why !



# Compute Limited Era

Model Parameter

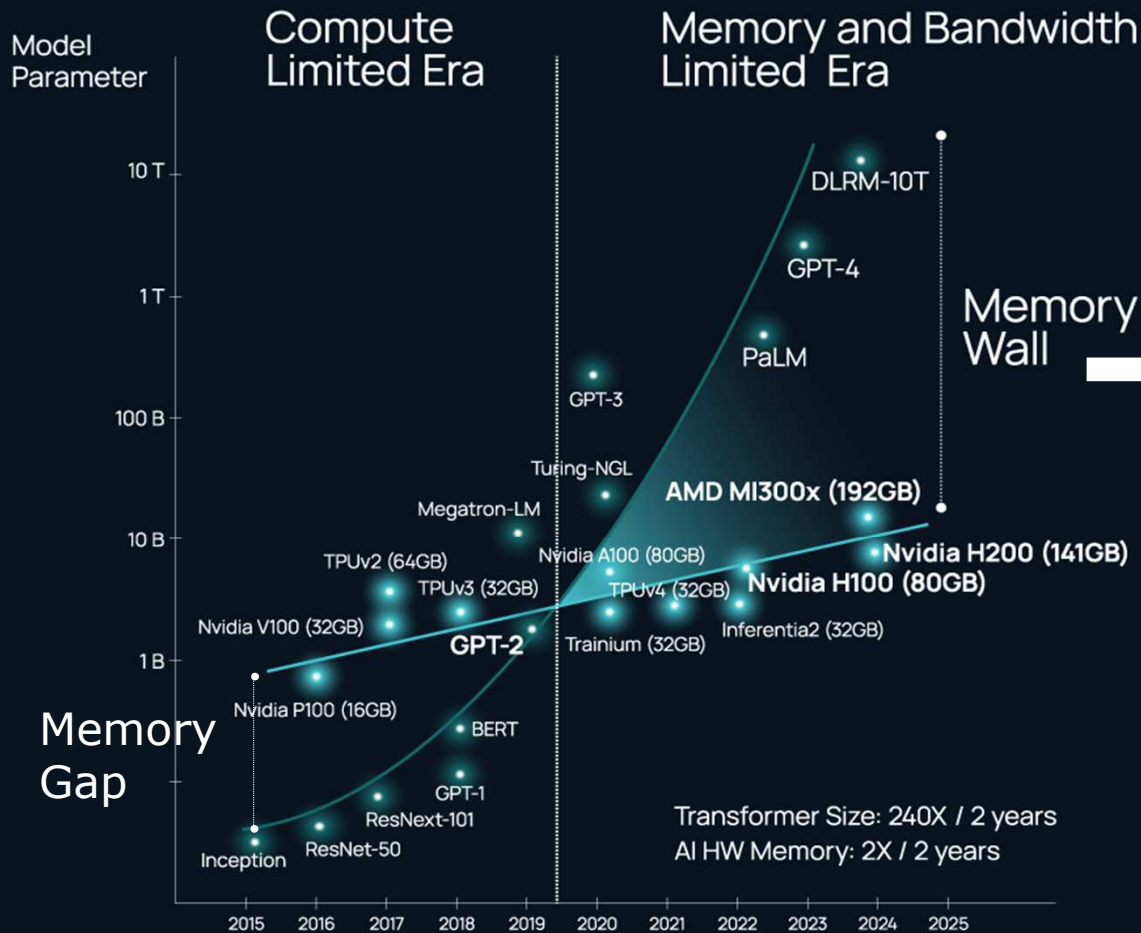


Memory Gap



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





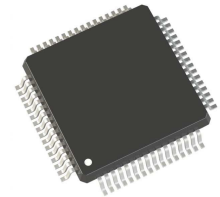
“Process trillions of bytes, billions of times. Distribute workloads on millions of GPUs”

Jensen Huang, CEO nVIDIA  
[https://youtu.be/MwiM\\_nPyx5Y](https://youtu.be/MwiM_nPyx5Y)

# Where to deploy GenAI ?



Data centers ? **OR** At the Edge ?



- The average word length in the English language is 4.7 characters [1].
- In most cases, the text content of a Tweet can contain up to 280 characters [2].
- A Tweet can contain maximum 60 tokens
- iPhone users worldwide are expected to rise to 1.56 billion by the end of 2024 [3].
- Android OS users worldwide are expected to rise to 3.6 billion by the end of 2024 [4].
- By 2024, there will be a potential of 5.16 billion total available users (TAU) for **Multimodal Assistant powered by Generative AI [MAssGenAI]**

[1] <https://www.wyliecomm.com/2021/11/whats-the-best-length-of-a-word-online/>

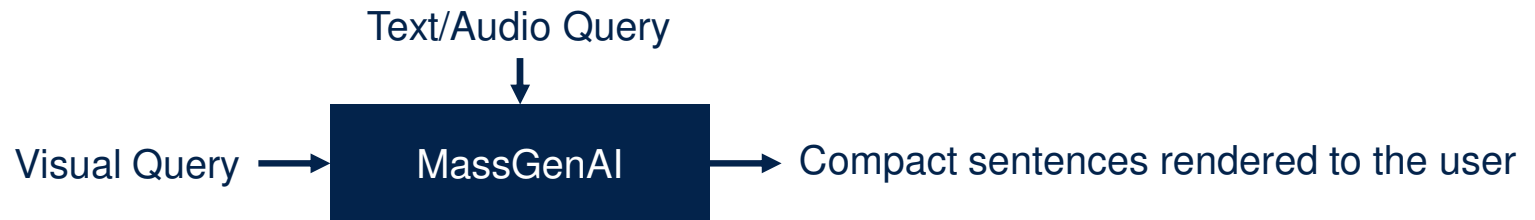
[2] <https://developer.x.com/en/docs/counting-characters#:~:text=In%20most%20cases%2C%20the%20text,as%20more%20than%20one%20character.>

[3] <https://www.coollest-gadgets.com/iphone-statistics#:~:text=In%20the%20first%20three%20months,of%20the%20global%20smartphone%20market.>

[4] <https://www.coollest-gadgets.com/android-statistics#:~:text=By%202023%2C%20there%20will%20be,reach%203.6%20billion%20by%202024.>



# MassGenAI Workload Example



- **Qwen2-VL-7B-Instruct-GPTQ-Int8** [5]:
  - state-of-the-art performance on visual understanding benchmarks, including MathVista, DocVQA, RealWorldQA, MTVQA, etc.
  - can understand videos over 20 minutes for high-quality video-based question answering, dialog, content creation, etc.
  - can be integrated with devices like mobile phones, robots, etc., for automatic operation based on visual environment and text instructions.
  - supports the understanding of texts in different languages inside images, including most European languages, Japanese, Korean, Arabic, Vietnamese, etc.
- **Performances** (NVIDIA A100 80GB)
  - Speed (tokens/s) 31.6 (input length 1)
  - GPU memory (GB) 10.11

# How Much is MAssGenAI Feasible on the Cloud ?

- Hypothetical service condition for GenZ.
  - Monthly subscription → 15 \$/month
  - Maximum acceptable latency → 5 sec

$$\frac{60 \frac{\text{tokens}}{\text{user}}}{31.6 \frac{\text{tokens}}{\text{sec}}} \sim 2 \frac{\text{sec}}{\text{user}}$$



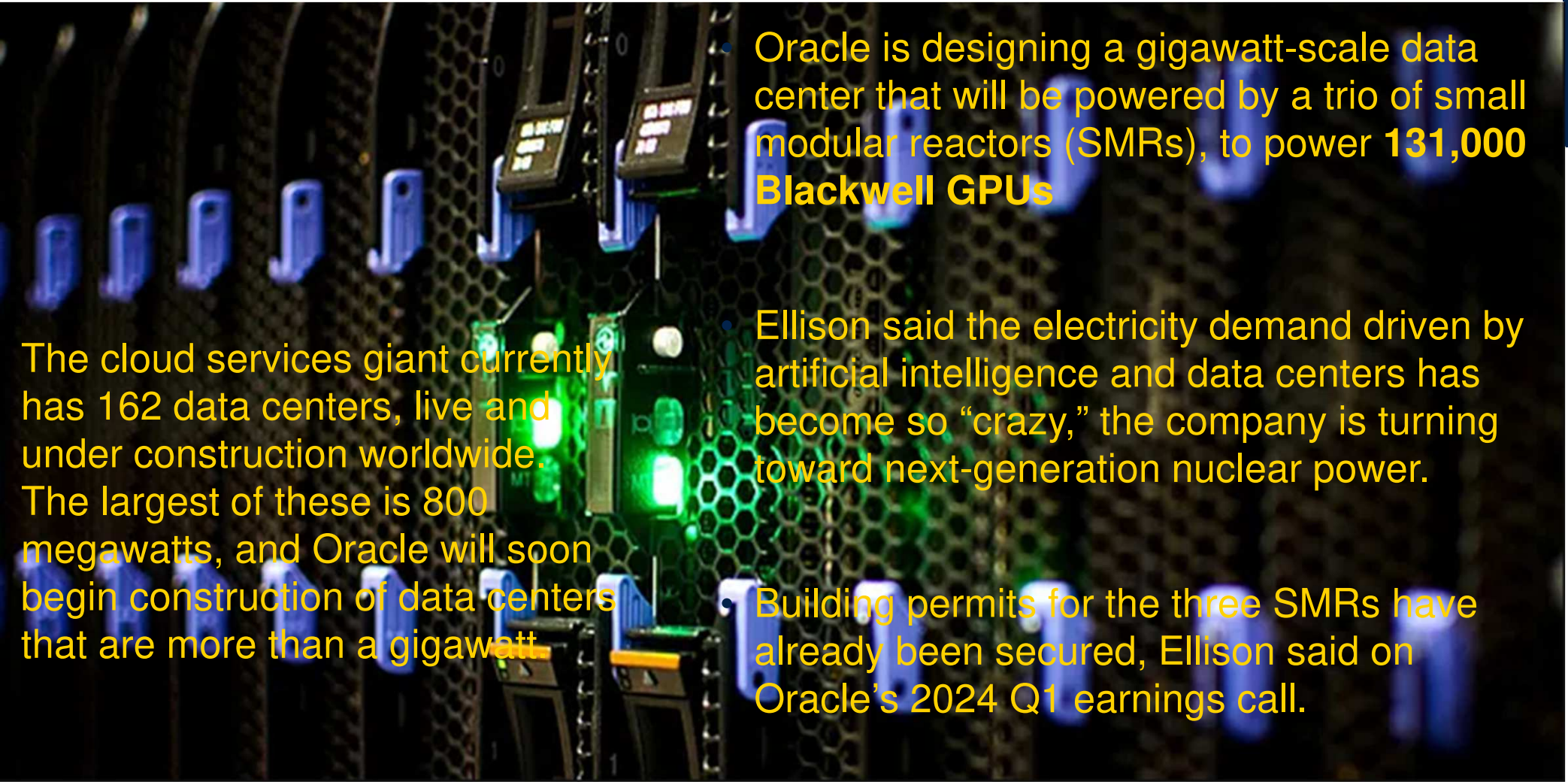
$$5,160,000,000 \text{ users} * 2 \frac{\text{sec}}{\text{user}} \sim 10,320,000,000 \text{ sec } \textit{or} \text{ 119,444.444 days } \textit{or} \text{ 327.3 years}$$



*level of acceleration required to support 5.16 B users simultaneously*

$$\frac{10,320,000,000 \text{ sec}}{5 \text{ sec}} \sim \text{2,064,000,000 } \textit{or 2B times } \textit{or}$$

**41,280 Cortex AI Superclusters (50,000 H100, 130 MW each) [6] !**



The cloud services giant currently has 162 data centers, live and under construction worldwide. The largest of these is 800 megawatts, and Oracle will soon begin construction of data centers that are more than a gigawatt.

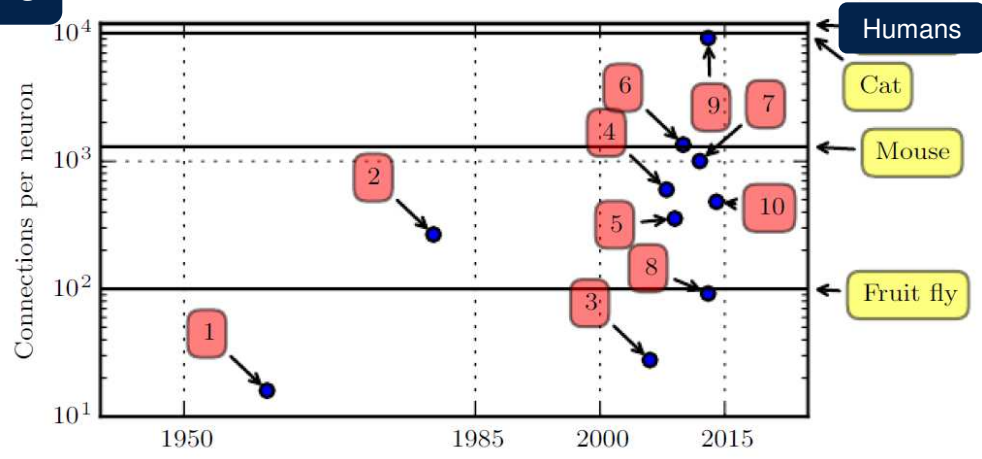
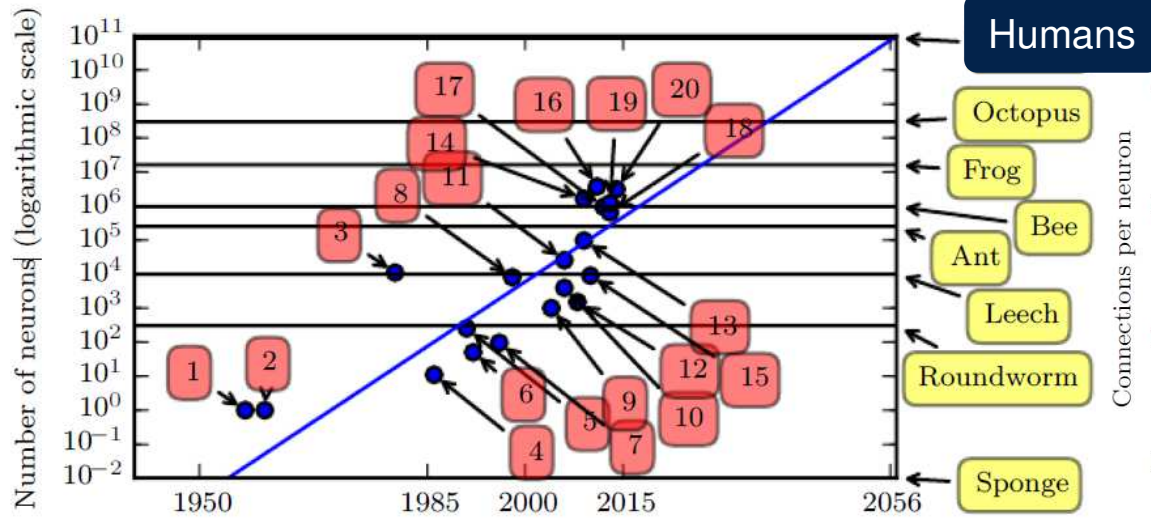
• Oracle is designing a gigawatt-scale data center that will be powered by a trio of small modular reactors (SMRs), to power **131,000 Blackwell GPUs**

• Ellison said the electricity demand driven by artificial intelligence and data centers has become so “crazy,” the company is turning toward next-generation nuclear power.

• Building permits for the three SMRs have already been secured, Ellison said on Oracle’s 2024 Q1 earnings call.

The human brain contains about  $10^{11}$  neurons

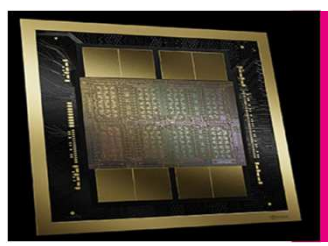
Each of neuron can have up to  $1.5 * 10^4$  connections with other neurons via the synapses



Goodfellow, Bengio, Courville - «Deep Learning» (2016).

Goodfellow, Bengio, Courville - «Deep Learning» (2016).

$\sim 1.5 * 10^{15}$  weights;  $\sim 3 * 10^{15}$  Rosenblatt 1958 ops



Blackwell  $10 * 10^{15}$  FLOPS FP16  
 $\sim 6.7x$  computational power

# Exascale Computing = Frontier

**Location:** Oak Ridge National Laboratory — Tennessee, U.S.

**Performance:** 1.2 exaFLOPS

**Components:**

9,472 AMD Epyc 7713 "Trento" 64 core 2 GHz CPUs

37,888 Instinct MI250X GPUs

**System:** 8,699,904 combined CPU and GPU cores

**Power efficiency rating:** 52.59 GFlops/Watt

**Total power consumption:** 150-500 MW



50K Blackwell  $\rightarrow 5 * 10^{20}$  FLOPS FP16  
~ 1.5-2 B\$ worth, ~ 35  $700W \div 50$   $1KW$  MWatts  
*each* *each*



HGX B100



NVLINK Switch



GB200 Superchip  
Compute Node



Quantum X800 Switch  
ConnectX-8 SuperNIC



Spectrum X800 Switch  
BlueField-3 SuperNIC



# Smaller and Modular Nuclear Power ... is better

<https://www.energy.gov/ne/advanced-small-modular-reactors-smrs>.



These advanced reactors, envisioned to vary in size from tens of megawatts up to hundreds of megawatts, can be used for power generation, process heat, desalination, or other industrial uses.

50K Blackwell GB200  $5 * 10^{20}$  FLOPS FP16  
~ 1.5÷2 B\$ worth, ~ 35  $700W$  ÷ 50  $1KW$  MWatts  
*each* *each*

# Sustainable ?

On 2023 Microsoft and Google respectively consumed 24 and 25 TWh.

CO2 emissions      Google: +48% since 2019.  
Microsoft: + 30% since 2020.

Data centers are projected to 1 PetaWh on 2026  
This is due to AI training workloads

Stay tuned about Oracle and Tesla datacenters !



30 years of human learning<sup>(†)</sup>  $\sim 10^9$  *seconds*

† (G. Hinton) <https://www.youtube.com/watch?v=N1TEjTeQeg0>



world first **digital** satellite service in the US  
launched in **1994**  $\rightarrow 6 \times 10^7$  *pixels/sec.*

To avoid overfitting of a  $10^{15}$  weights AI,  
we need  $\gg 10^{16}$  data [30 years]

Not Enough Data

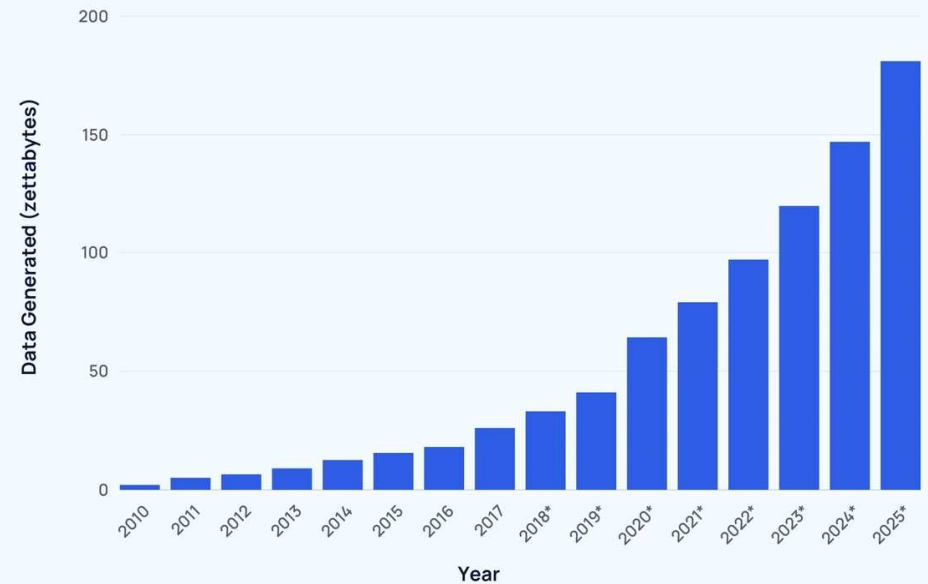
# Generating a tsunami of data...

>300 million TeraBytes data created each day

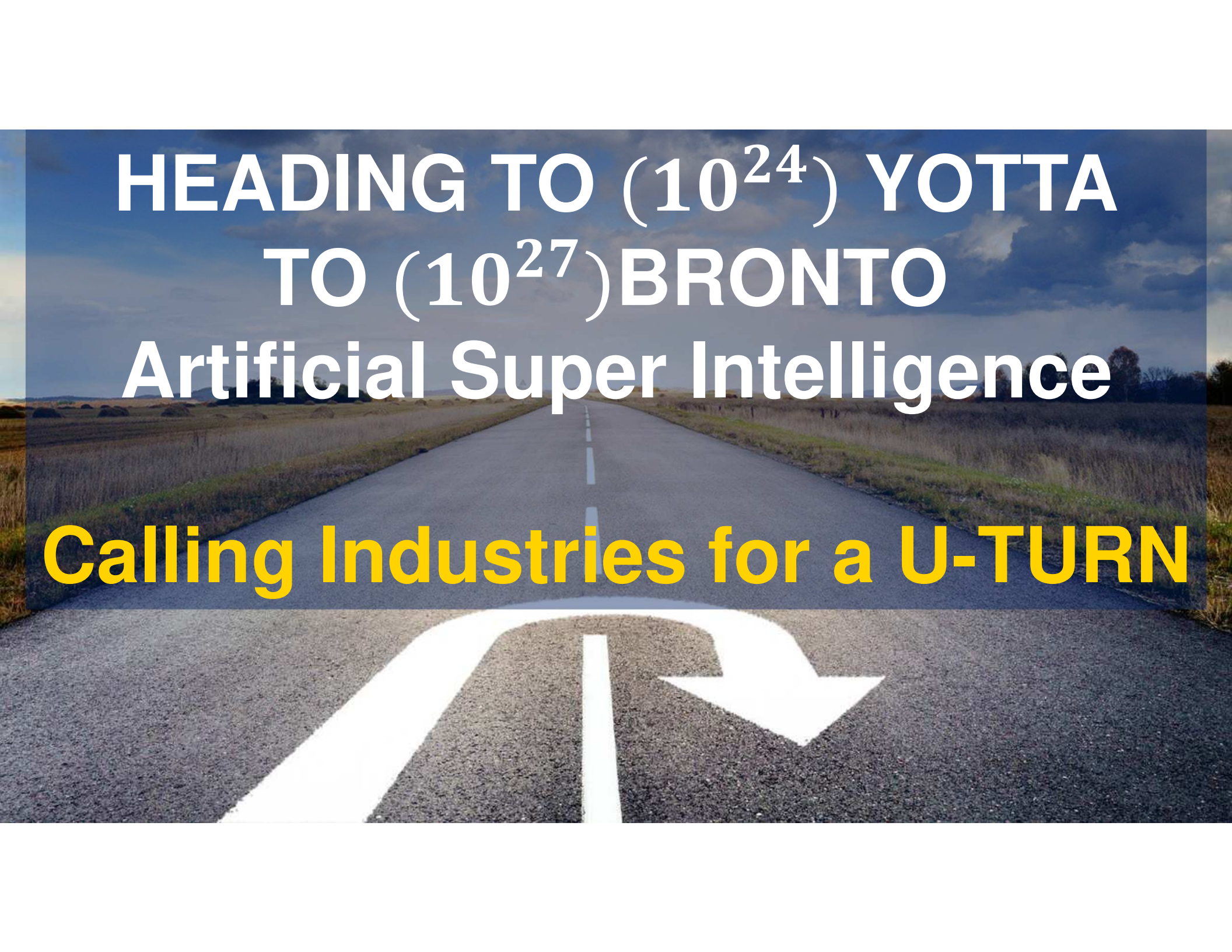
120 ZetaBytes data generated in 2024  
> 180 ZetaBytes in 2025

US alone hosts over 2,700 data centers

## Global Data Generated Annually



Source: explodingtopics.com



**HEADING TO ( $10^{24}$ ) YOTTA  
TO ( $10^{27}$ ) BRONTO  
Artificial Super Intelligence**

**Calling Industries for a U-TURN**

# Human brain vs AI devices capabilities

The human brain consumes about **20 W**

The top500 HPC\*\*, Frontier, delivers  
14,054 Tera flops/s (ref HPCG benchmark)

**Human brain estimated to be ~50 Pops/Watt.  
10<sup>8</sup> more energy efficient than the #1 HPC Frontier**

Source: \*NIST, \*\*top500.org/

**Challenge: How to bridge this chasm?**

# The Change in Perspective



# U-turn to Tiny AI

## Tiny Resources

Devise AI workloads to minimize computational and memory (RAM, FLASH) requirements

## Toward Zero Power

Reduce energy needs from any sources by harvesting and scavenging.

## Achieve High Accuracy

Ensure accuracy and confidence is kept at high levels w.r.t. large models

## Live without floating point numbers

Why shall single and double precision be mandatory ?

## Automated deployments

Import, analyze, optimize and map AI workloads on sensors, MCU, MPU, actuators at the highest productivity level



# Tiny Assets so far

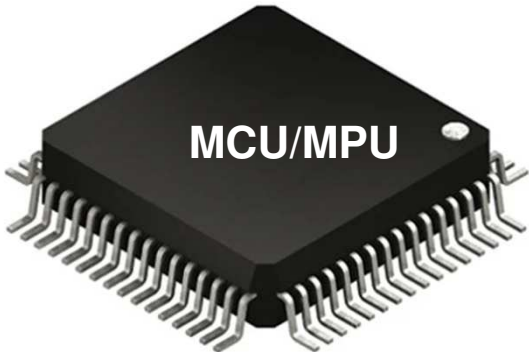
## Sensors



Data



## De-centralized/ scalable/distributed tiny embedded processing



Control



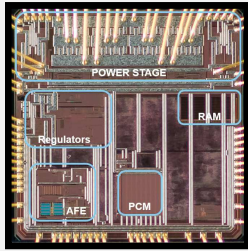
## Actuators



Very few 10s KiB RAM  
No FLASH  
KHz to MHz  
INT1, INT8, INT16; FP32

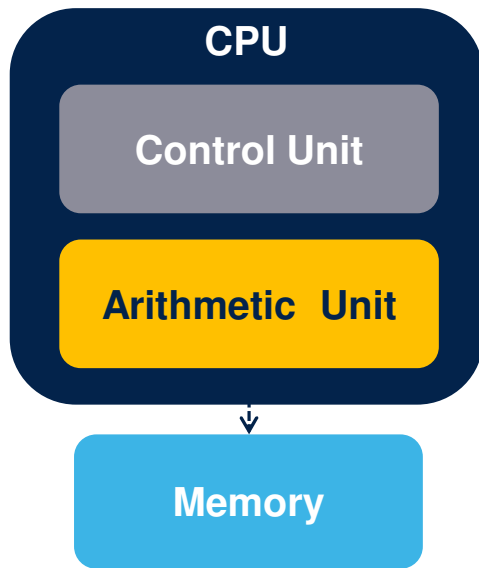
≤ 1(on)vs16(off) MiB RAM  
≤ 2(on)vs64(off) MiB FLASH  
≤ 550 MHz  
NPU (optional)  
INT8, INT16; FP32

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

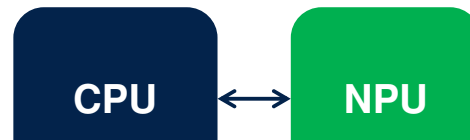


# The Quest for Energy Efficiency

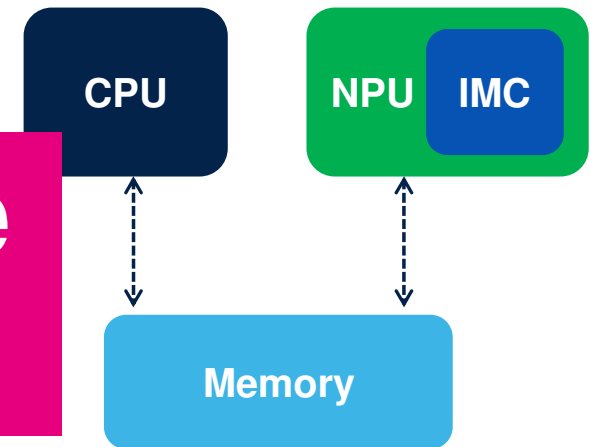
## Von Neumann Architecture



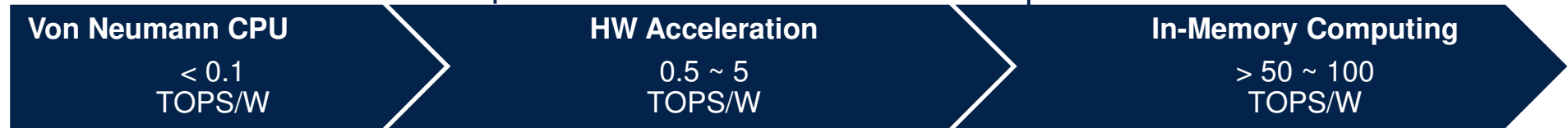
## Digital NPU



## IMC-based NPU



**How to fill the 2x gap?**



**STEPS FORWARD .....**

# Listen to Machine Learning and Embedded Engineers Needs



## Interoperability

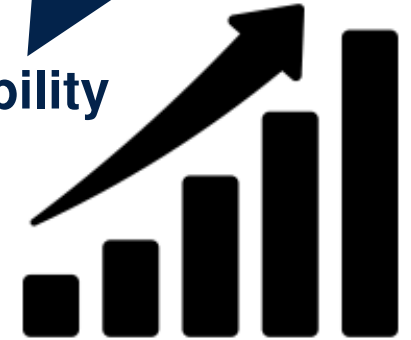


“The documented agreement reached by a group of individuals who recognize the advantage of all doing certain things in an agreed way”.  
Leonardo Chiariglione

## The needs

to trillions of sensors

## Scalability



## Automation



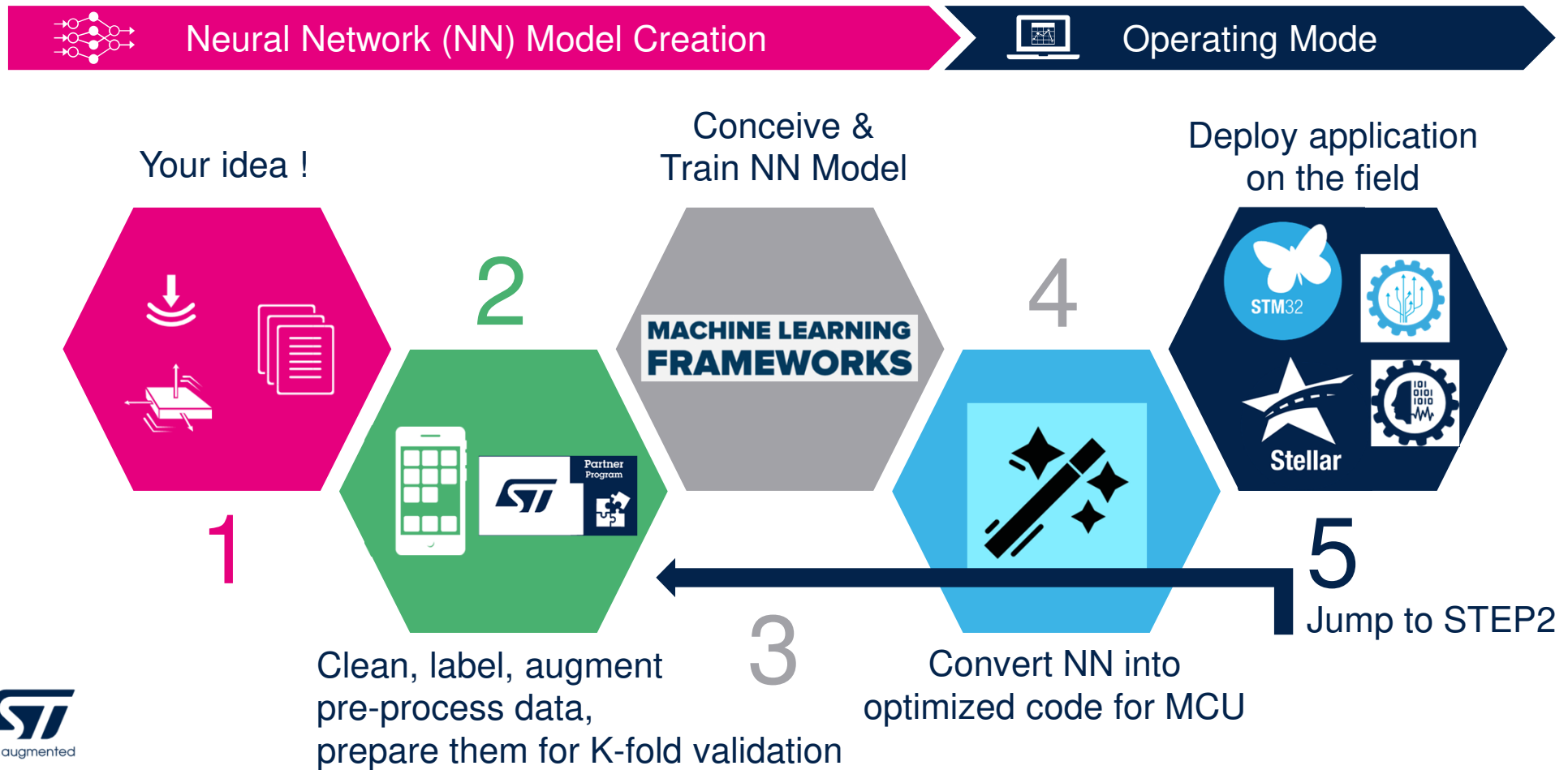
“everything that can be automated will be automated”.  
First law  
Shoshana ZUBOFF

## Productivity



Don't keep calm and don't hand-craft ML

# 5 Key steps for Supervised Deep Learning



## Interoperability

Pre-trained Neural Network models  
Deep Learning framework dependent

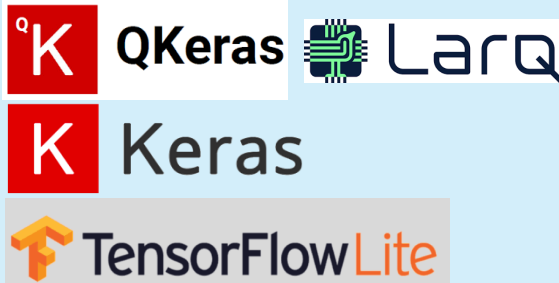
Machine Learning



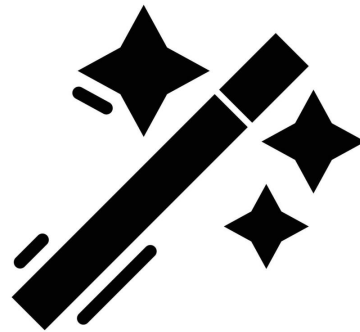
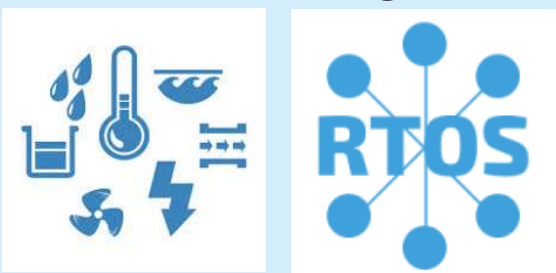
### Everybody else



### Google



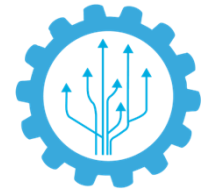
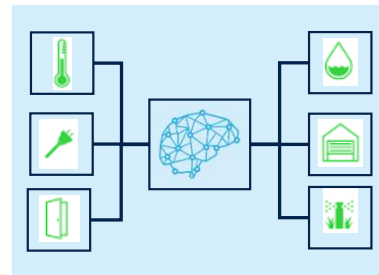
### Sensors and OS Agnostic



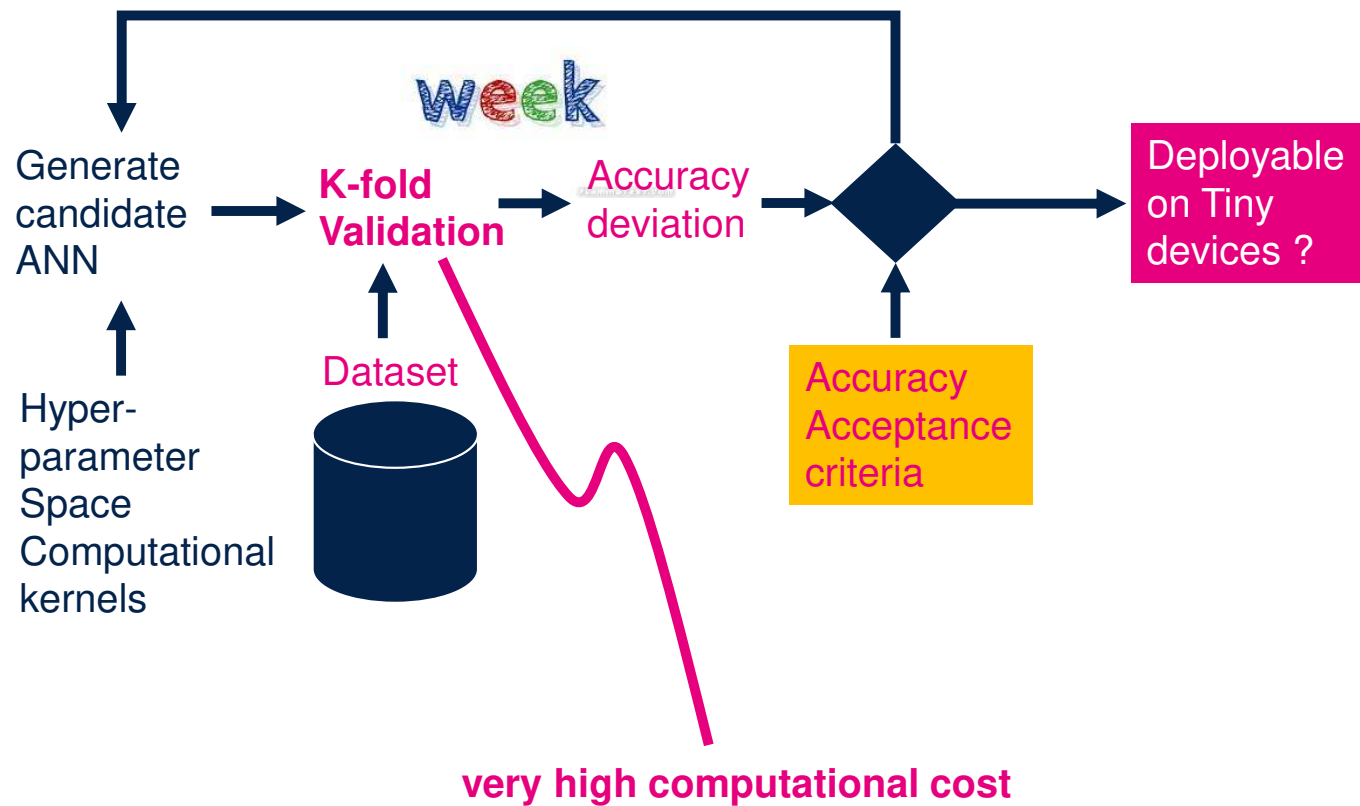
Choose your IDE  
Compiler and Debugger  
Framework Independent

### AI Hub

Multiple Neural Networks

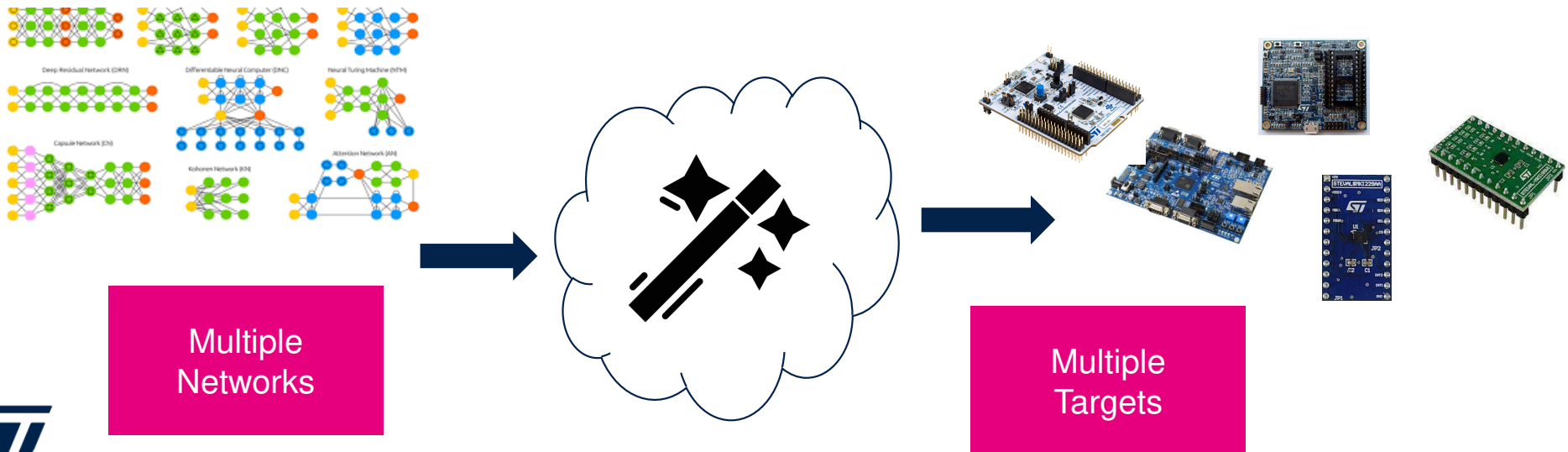


# Deployment un-aware NAS/HPO



# Machine Learning Heterogeneity

- Two types must be tackled:
  - **Source Model:** Machine learning models can be very different in terms of algorithms, topologies, size, representation formats, etc.
  - **Execution Target:** processing elements can be very different in terms computational capabilities, available memory, optimized instruction, etc.







# STM32 portfolio



## MPU

**STM32MP1**  
Up to 1 GHz Cortex-A7  
209 MHz Cortex-M4

**STM32MP2**  
Dual 1.5 GHz Cortex-A35  
400 MHz Cortex-M33

## High-performance MCUs

**STM32F2**  
Up to 398 CoreMark  
120 MHz Cortex-M3

**STM32F4**  
Up to 608 CoreMark  
180 MHz Cortex-M4

**STM32F7**  
1082 CoreMark  
216 MHz Cortex-M7

**STM32H7**  
Up to 3224 CoreMark  
Up to 550 MHz Cortex -M7  
240 MHz Cortex -M4

**STM32N6**  
MCU with neural processing unit

**STM32H5**  
Up to 1023 CoreMark  
250 MHz Cortex-M33

**STM32H7RS**  
Up to 3174 CoreMark  
600 MHz Cortex -M7

## Mainstream MCUs

**STM32F3**  
245 CoreMark  
72 MHz Cortex-M4

**STM32G4**  
569 CoreMark  
170 MHz Cortex-M4

*Mixed-signal MCUs*

**STM32C0**  
114 CoreMark  
48 MHz Cortex M0+

**STM32F0**  
106 CoreMark  
48 MHz Cortex-M0

**STM32G0**  
142 CoreMark  
64 MHz Cortex-M0+

**STM32F1**  
177 CoreMark  
72 MHz Cortex-M3

## Ultra-low-power MCUs

**STM32L0**  
75 CoreMark  
32 MHz Cortex-M0+

**STM32L4**  
273 CoreMark  
80 MHz Cortex-M4

**STM32L4+**  
409 CoreMark  
120 MHz Cortex-M4

**STM32L5**  
443 CoreMark  
110 MHz Cortex-M33

**STM32U5**  
651 CoreMark  
160 MHz Cortex-M33

## Wireless MCUs

**STM32WL**  
162 CoreMark  
48 MHz Cortex-M4  
48 MHz Cortex-M0+

**STM32WB**  
216 CoreMark  
64 MHz Cortex-M4  
32 MHz Cortex-M0+

**STM32WBA**  
407 CoreMark  
100 MHz Cortex-M33



Latest product generation
● Radio co-processor only
 New series introduced in 2024
Pre-announcement

# STM32N6x AI Product Line

# 600x

ML performance uplift\*



life.augmented

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

## Dedicated Embedded Neural Processing Unit

- 600GOPS NPU
- 3TOPS/W power consumption

## Arm cortex®- M55 core

- 1280 DMIPS / 3360 CoreMark
- New DSP extensions (MVE)

## Embedded RAM

- 4.2 MB embedded RAM

## Computer vision pipeline

- Parallel and MIPI CSI-2 camera I/F
- Dedicated image processor (ISP)

## Extended multimedia capabilities

- 2.5D Graphics accelerator
- H264 encoder, JPEG encoder/decoder

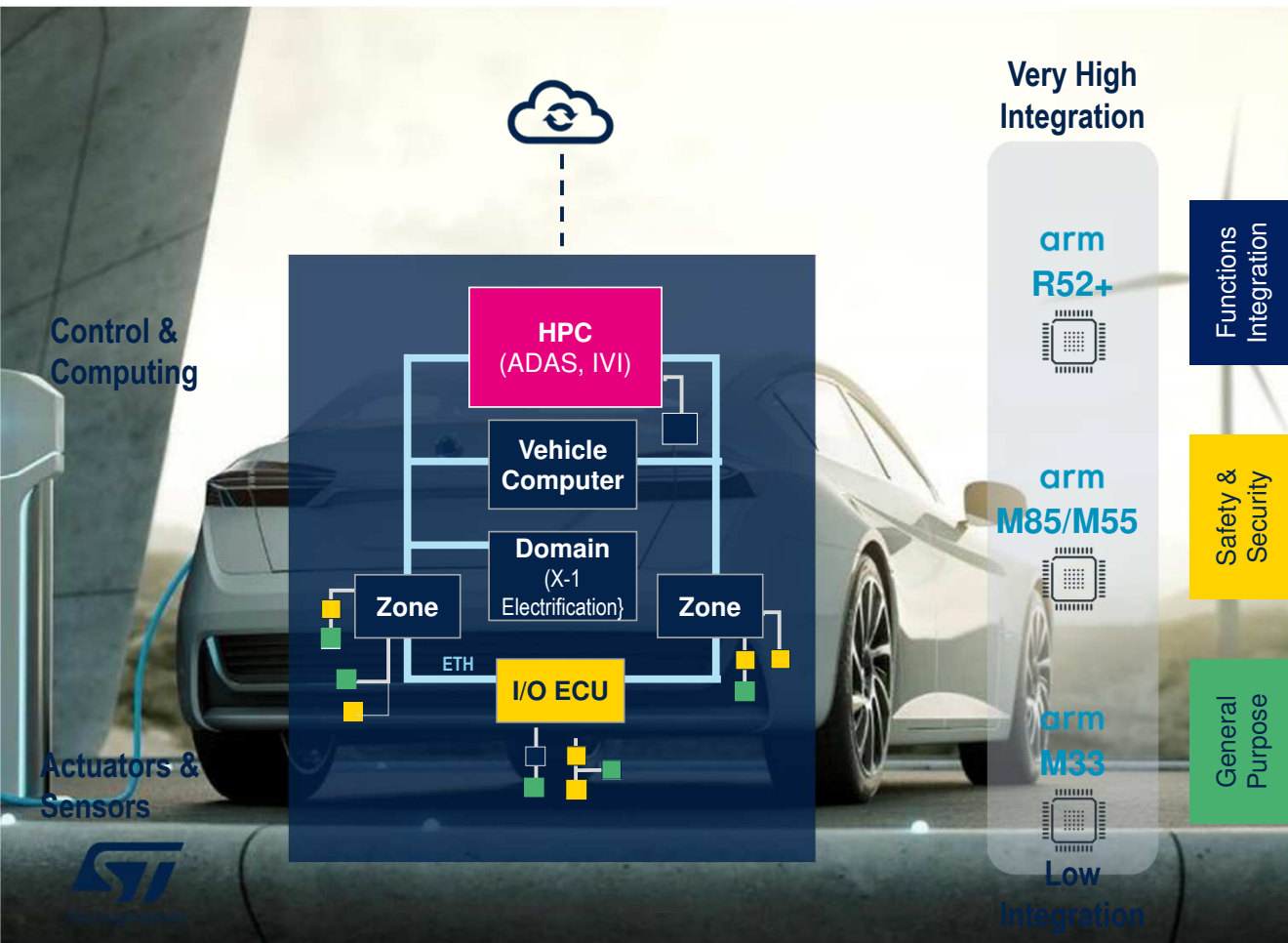
## Extended security features

- arm TrustZone for the Cortex-M55 core and the NPU



# ST Automotive Microcontrollers & Processor

from actuation to centralization



## Integration/Aggregation

Functions Integration	<b>Stellar P / G</b> Real-time integration platform	28nm	Zone Ctrl (Mid/High)	Safety Domains
			Body integration	X-in-1 Electrification

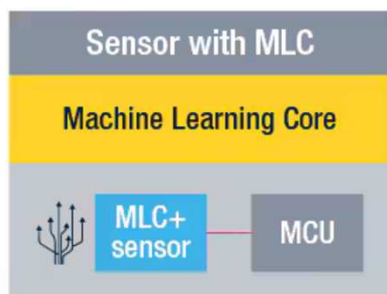
## Highly Safe Control

Safety & Security	<b>Stellar X</b> Actuation & Precise controlling ASIL-D	18nm	Chassis & Safety	Zone Ctrl (Low/Mid)
			AVAS & Audio	Standalone xEV

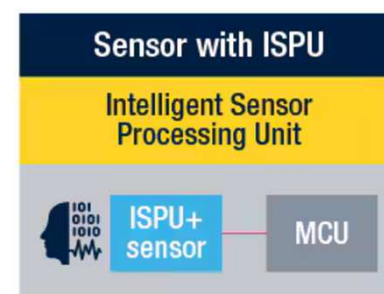
## Value Optimization

General Purpose	<b>STM32 A</b> ASIL-B	40nm	Lighting	Key Entry
			Seats	Doors

# In Sensor AI in Single Package

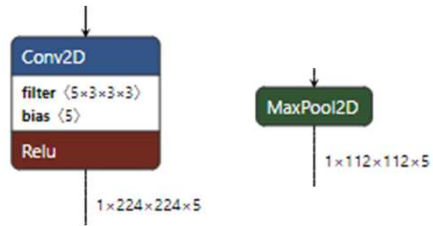


Part number	Application	Power consumption	Part number	Application	Power consumption
<a href="#">LIS2DUX12</a>	Consumer	2.7 $\mu$ A	<a href="#">LSM6DSOX</a>	Consumer	0.55 mA combo
<a href="#">LIS2DUXS12</a>	Consumer	2.7 $\mu$ A	<a href="#">LSM6DSO32X</a>	Consumer	0.55 mA combo
<a href="#">LSM6DSV16X</a>	Consumer	0.65 mA combo	<a href="#">LSM6DSRX</a>	Consumer	1.2 mA combo
<a href="#">LSM6DSV16BX</a>	Consumer	0.95 mA combo	<a href="#">ISM330DHCX</a>	Industrial	1.2 mA combo
<a href="#">LSM6DSV32X</a>	Consumer	0.65 mA combo	<a href="#">IIS2ICLX</a>	Industrial	0.42 mA
<a href="#">ASM330LHB</a>	Automotive	0.8 mA combo	<a href="#">ASM330LHHX</a>	Automotive	0.8 mA combo
<a href="#">ASM330LHBG1</a>	Automotive	0.8 mA combo			
<a href="#">ASM330LHHXG1</a>	Automotive	0.8 mA combo			
<a href="#">ISM330BX</a>	Industrial	0.6 mA combo			



Part number	Application	Power consumption
<a href="#">ISM330IS</a>	Industrial	0.59 mA (combo mode)
<a href="#">ISM330ISN</a>	Industrial (anomaly detection)	0.59 mA (combo mode)
<a href="#">LSM6DSO16IS</a>	Consumer	0.59 mA (combo mode)

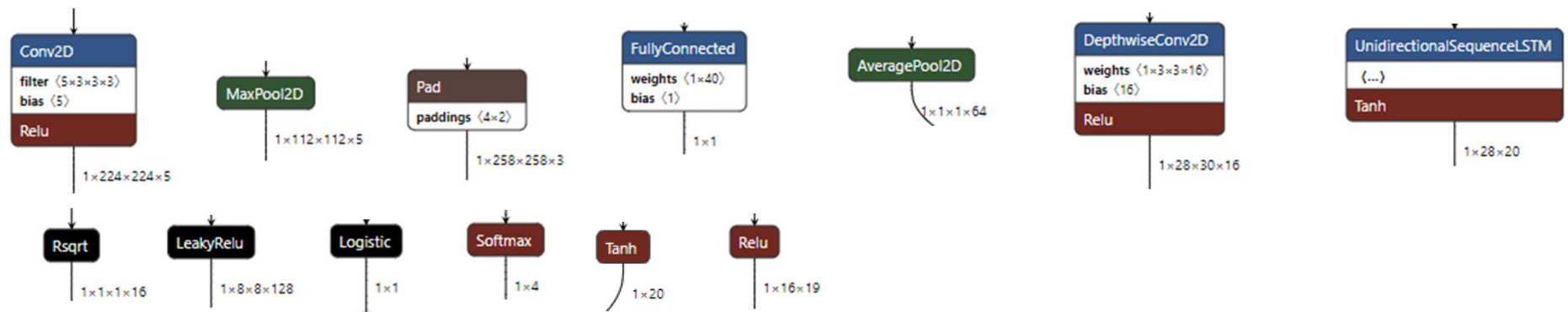
# Heterogeneity: Operators



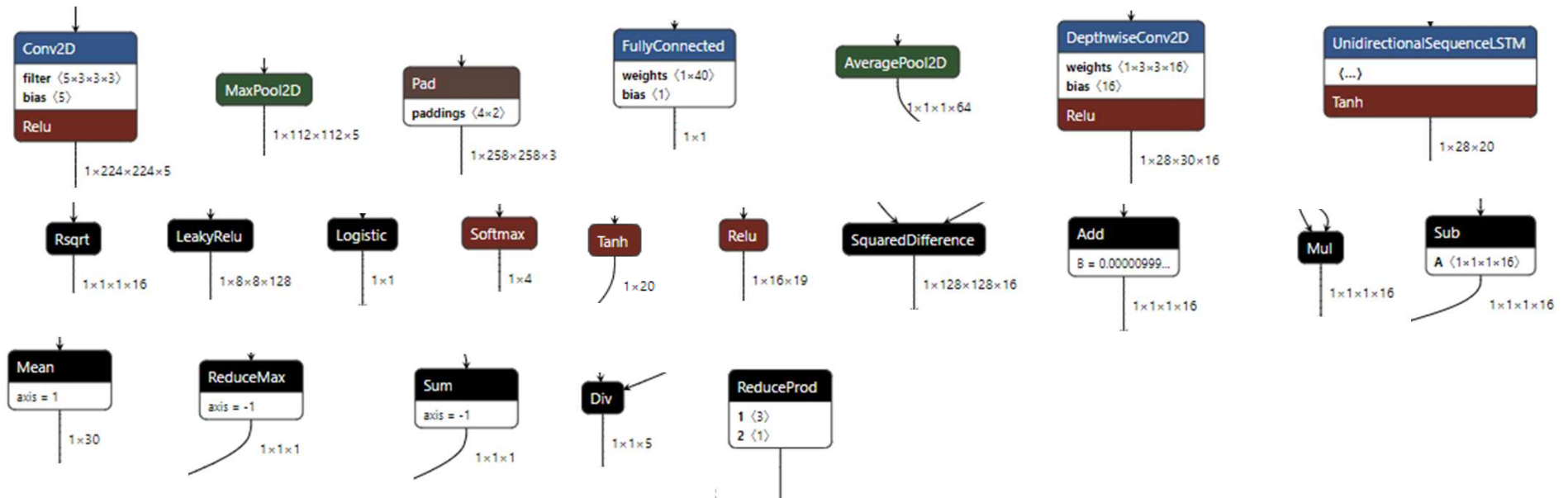
# Heterogeneity: Operators



# Heterogeneity: Operators

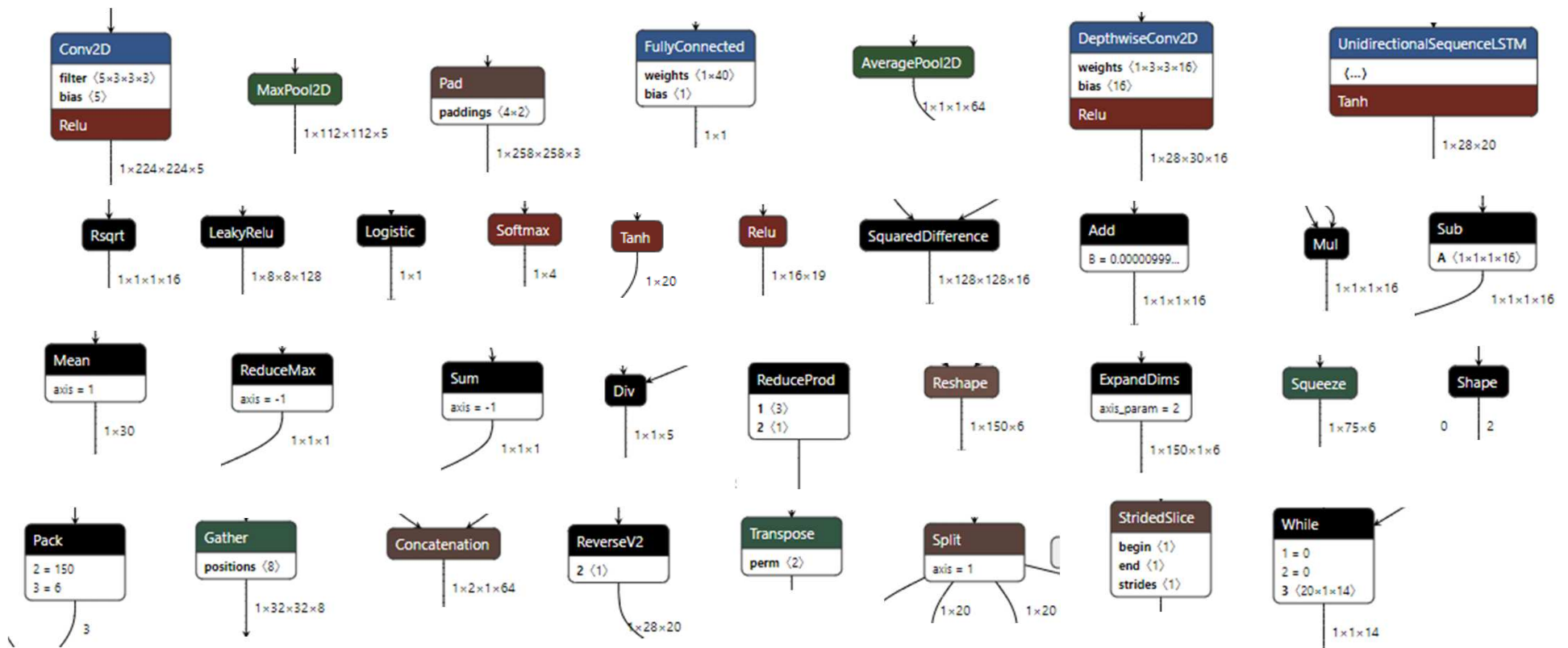


# Heterogeneity: Operators

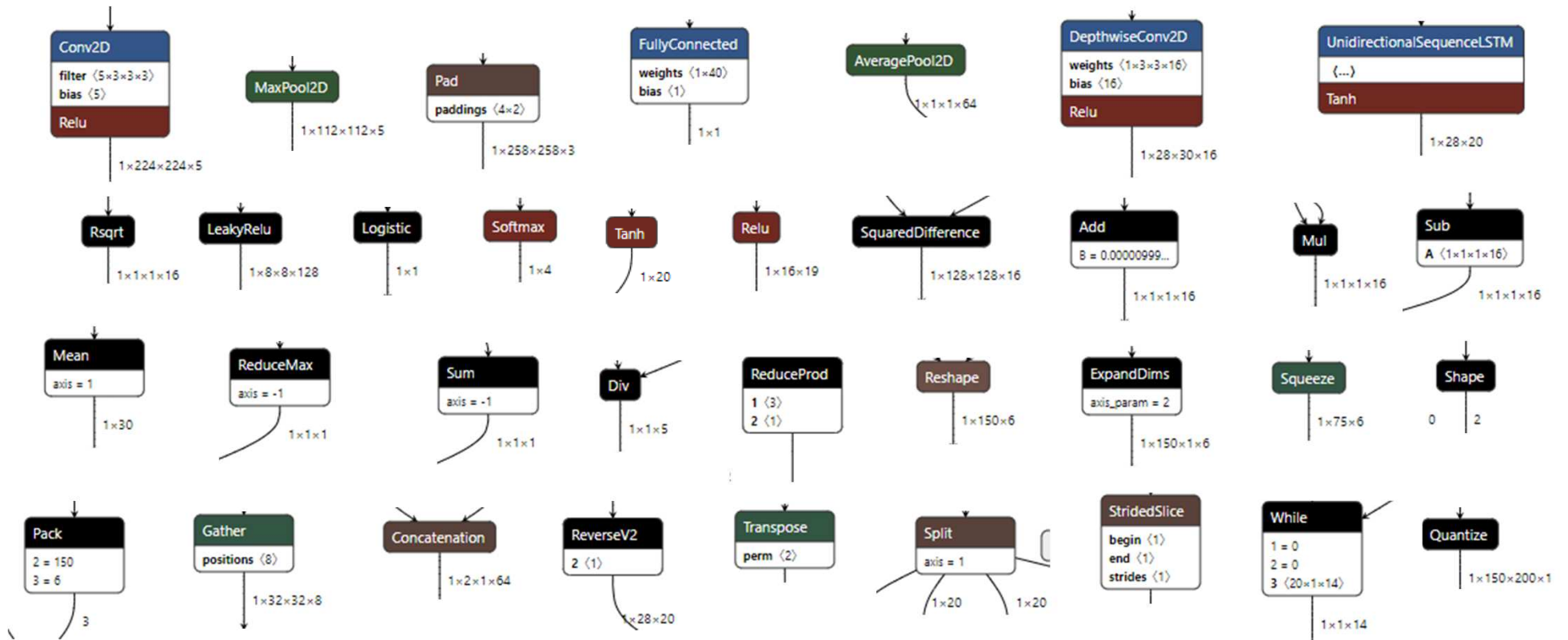




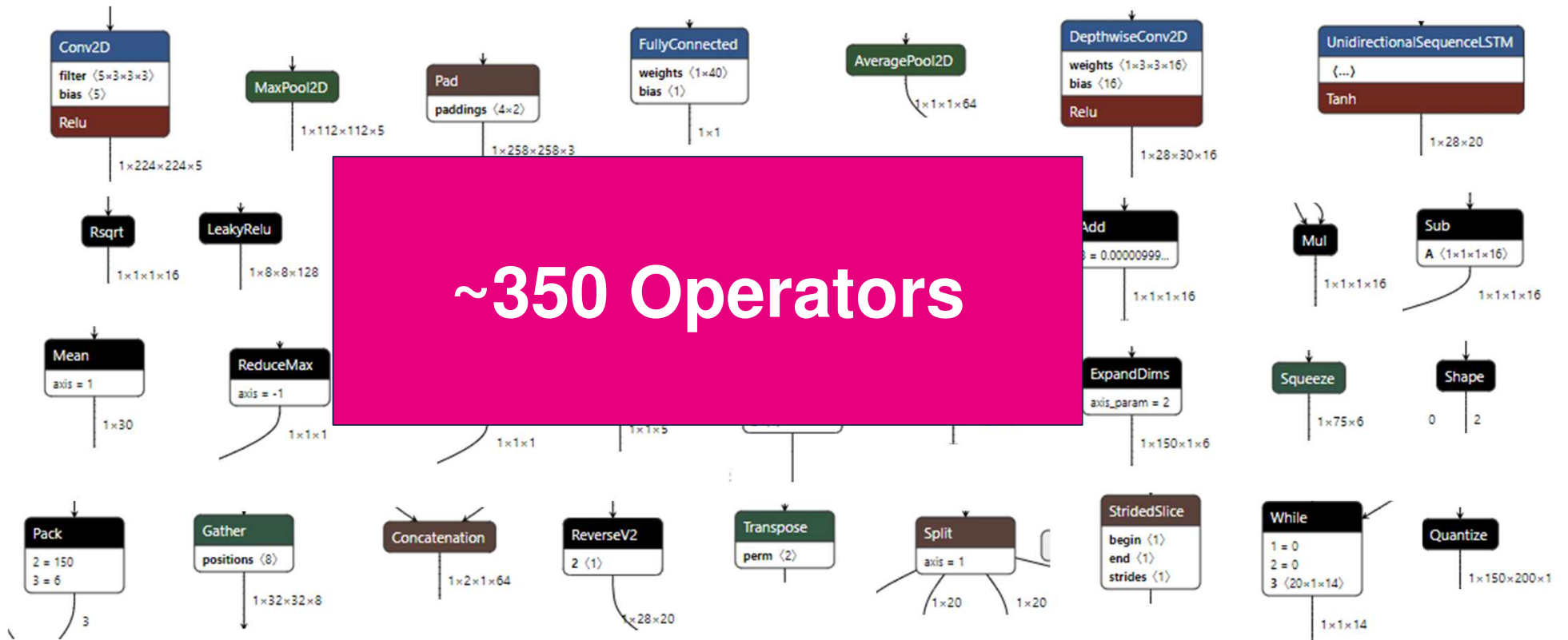
# Heterogeneity: Operators



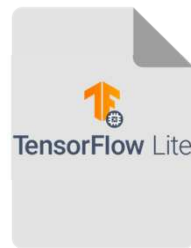
# Heterogeneity: Operators



# Heterogeneity: Operators

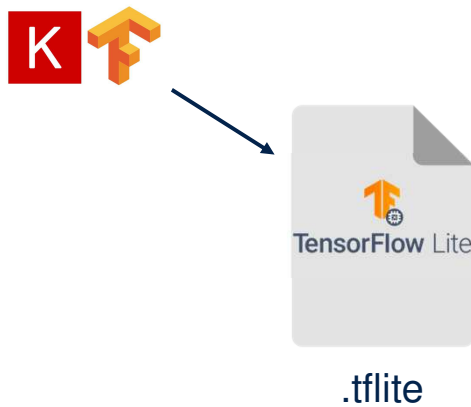


# Heterogeneity: DL Formats

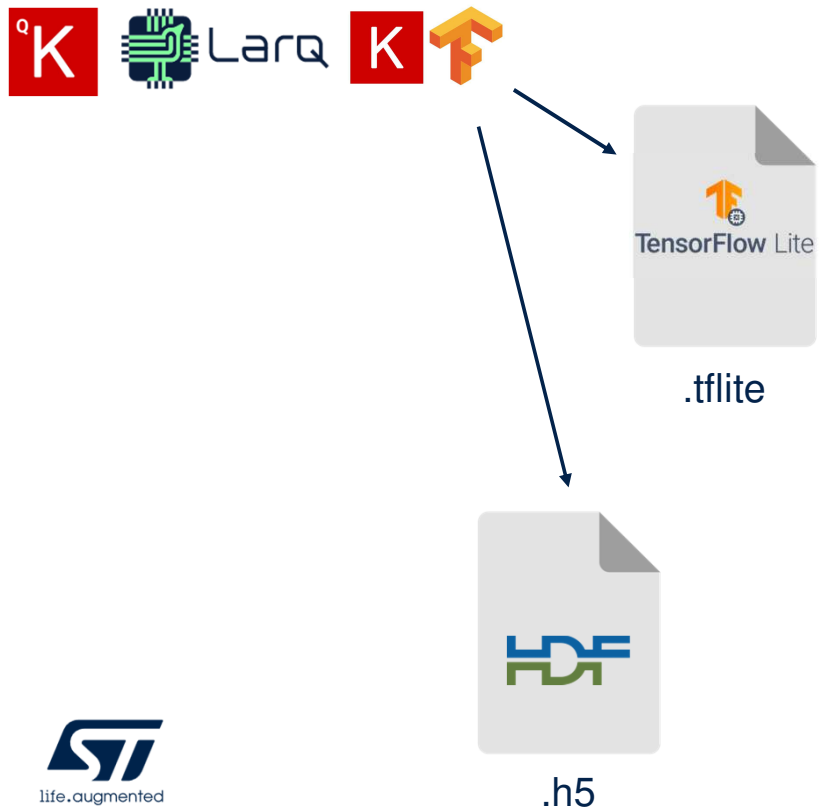


.tflite

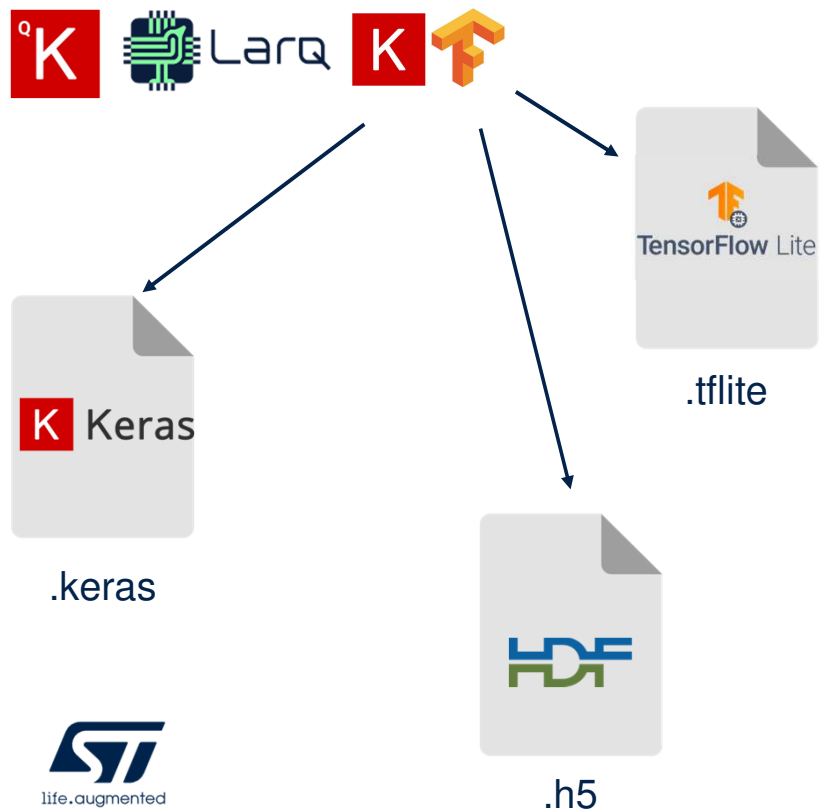
# Heterogeneity: DL Formats



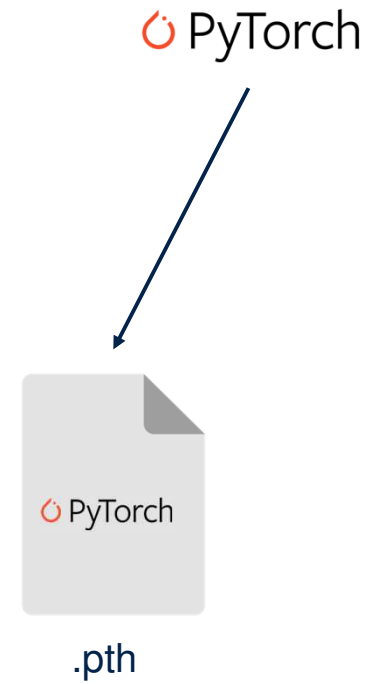
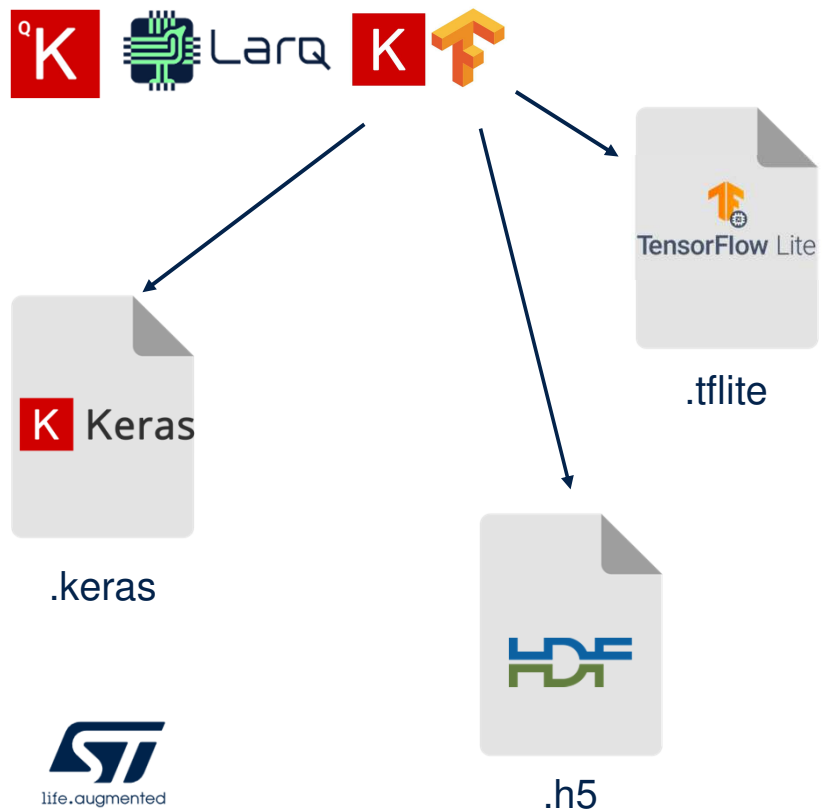
# Heterogeneity: DL Formats



# Heterogeneity: DL Formats

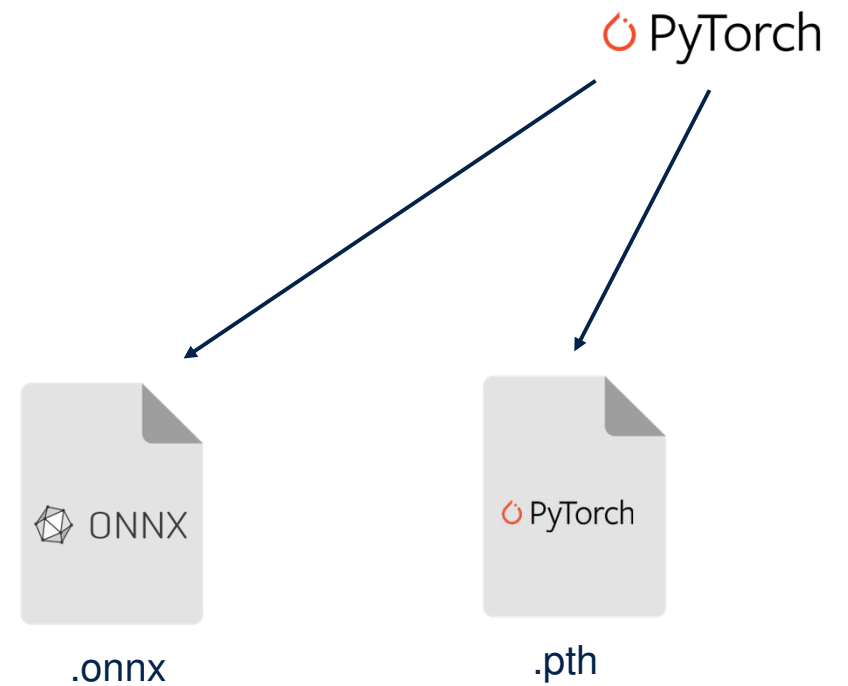
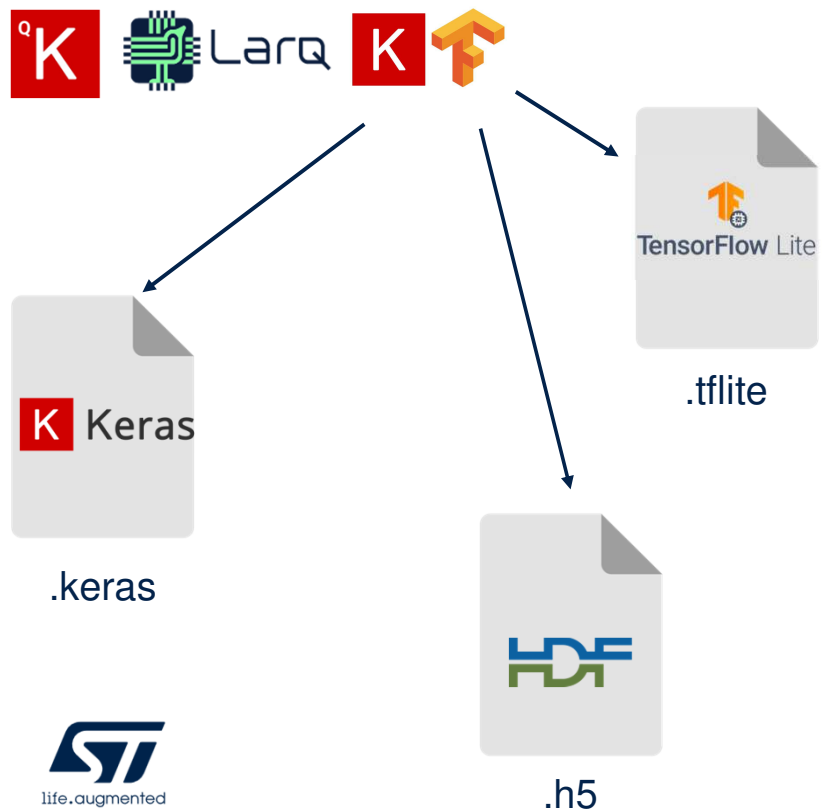


# Heterogeneity: DL Formats

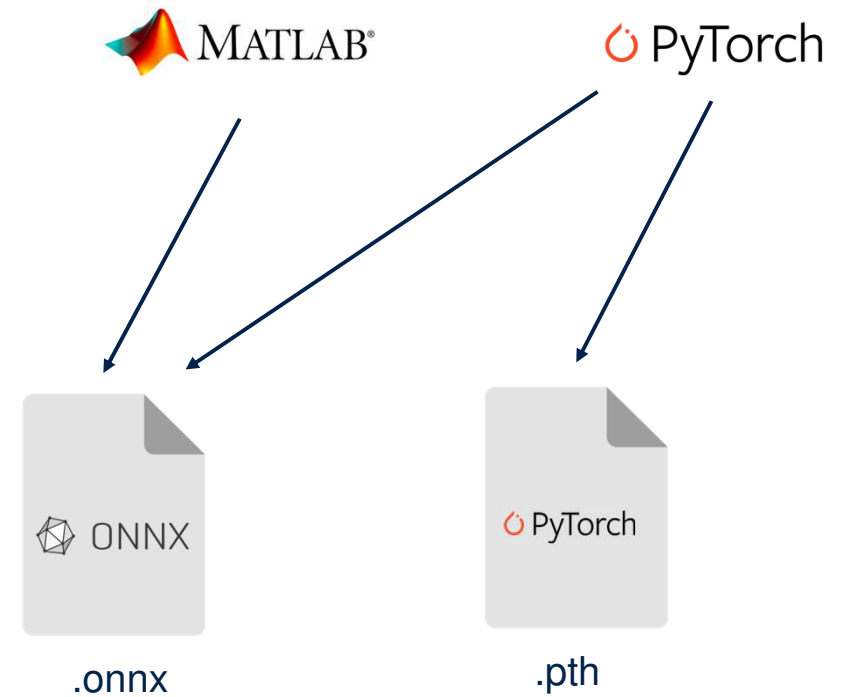
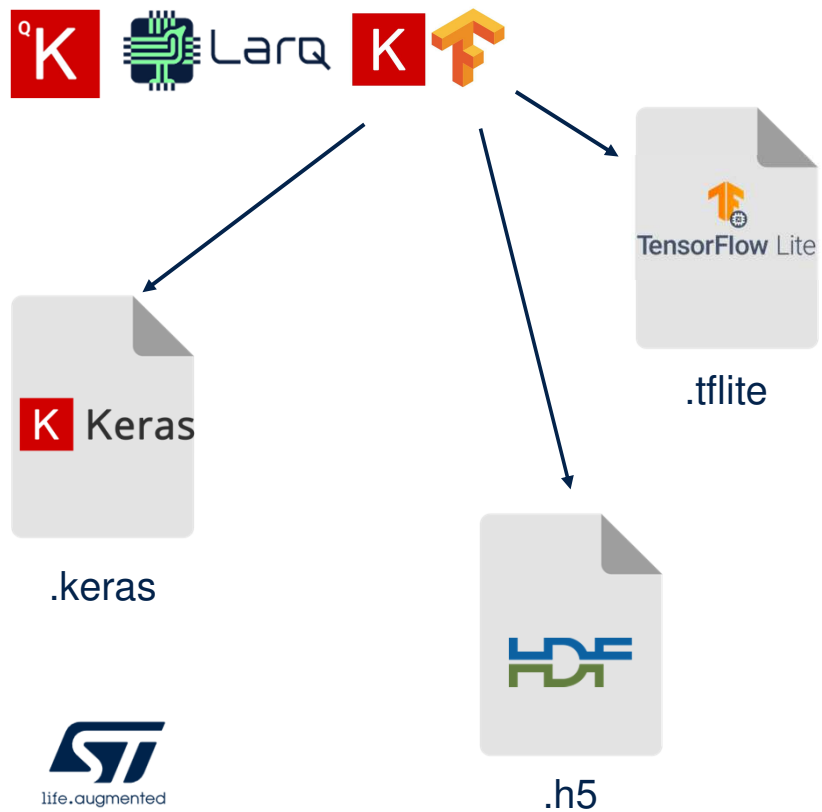




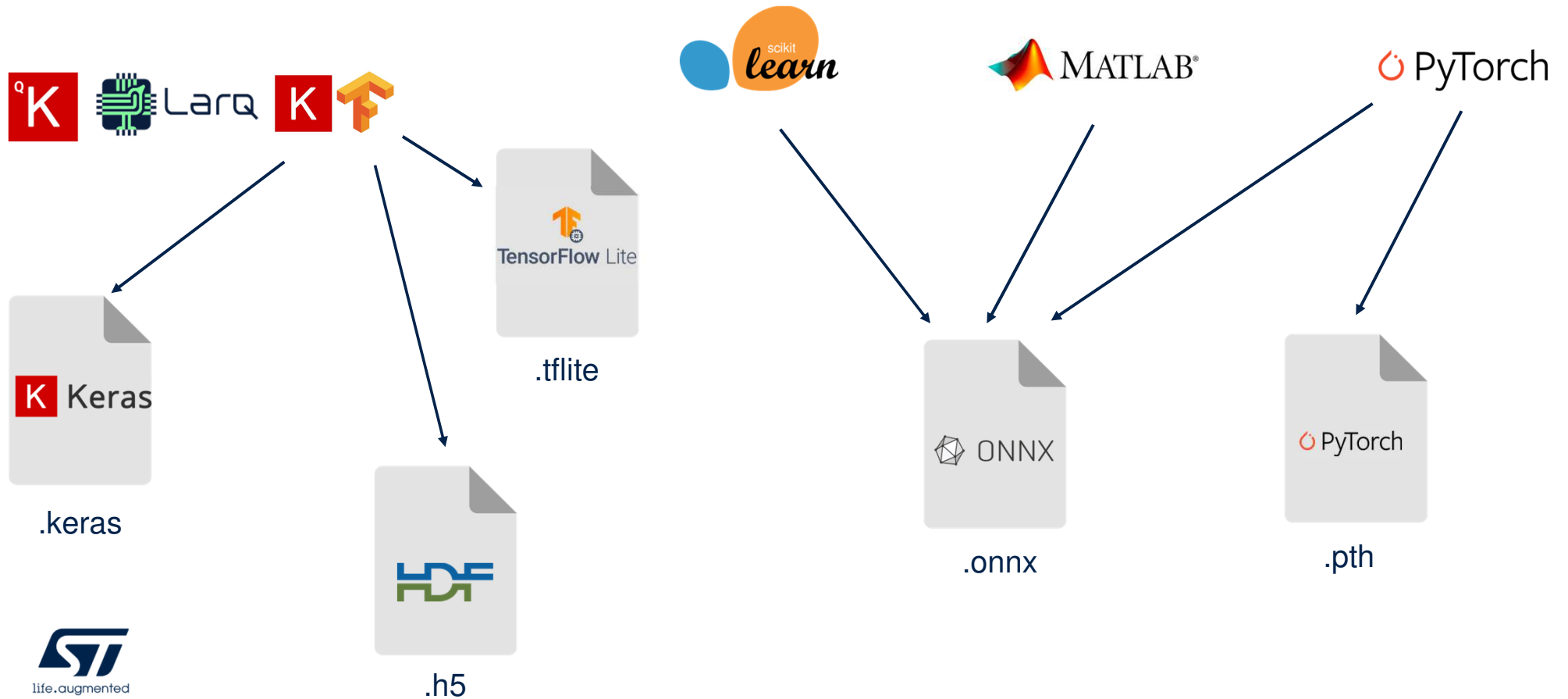
# Heterogeneity: DL Formats



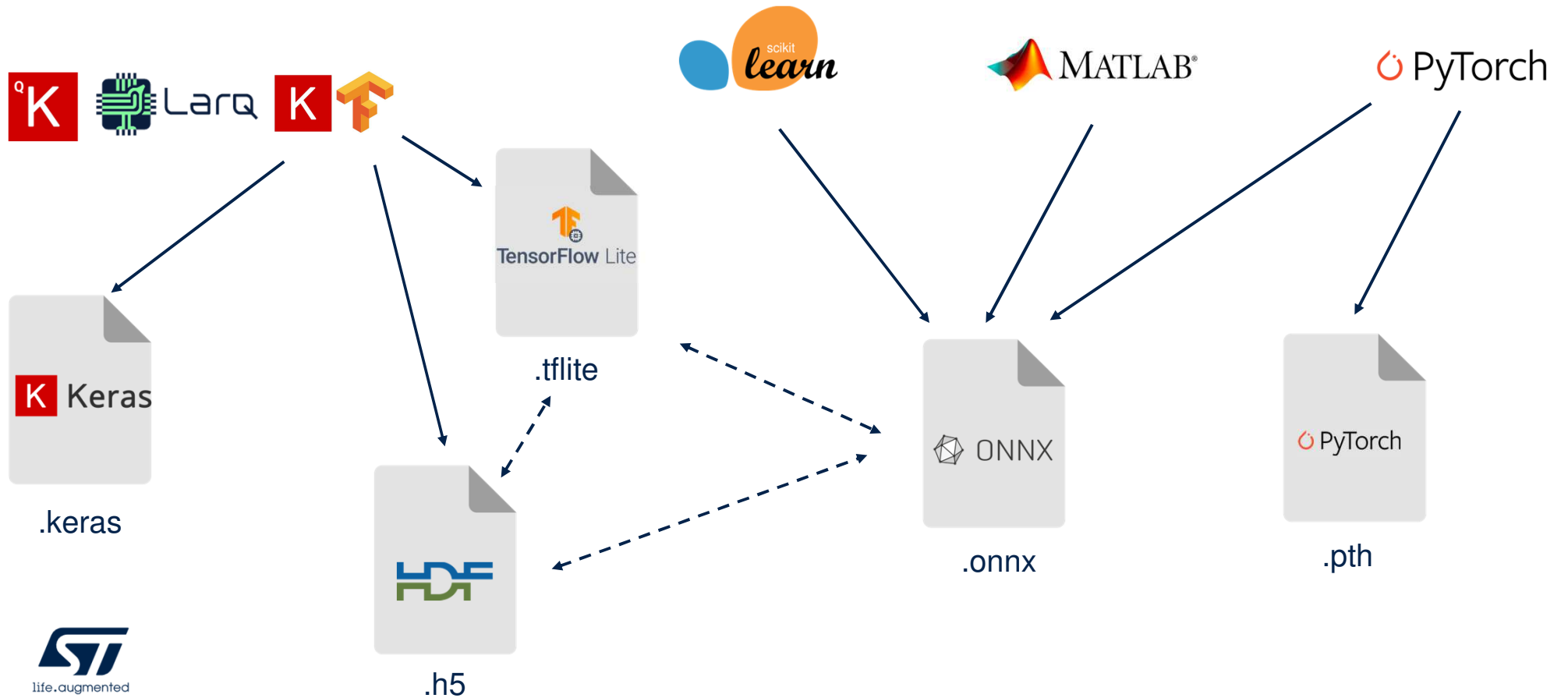
# Heterogeneity: DL Formats



# Heterogeneity: DL Formats



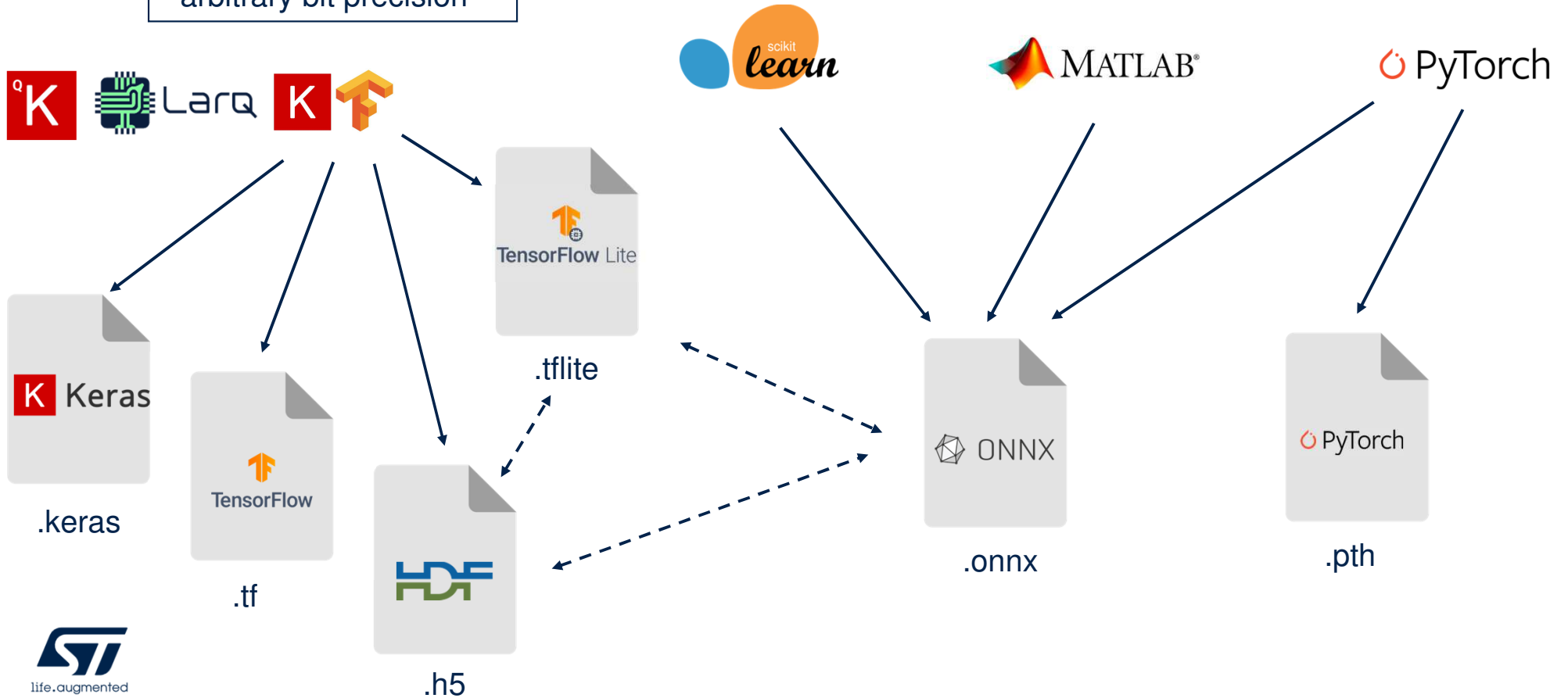
# Heterogeneity: DL Formats



# Heterogeneity: DL Formats

Different type/precisions:

- float32
- int8 scale/offset
- binary
- arbitrary bit precision



# Heterogeneity: DL Differences



TFLite  
MEAN

ONNX  
Mean

# Heterogeneity: DL Differences



TFLite  
MEAN

≠

ONNX  
Mean

# Heterogeneity: DL Differences



TFLite  
MEAN

≠

ONNX  
Mean

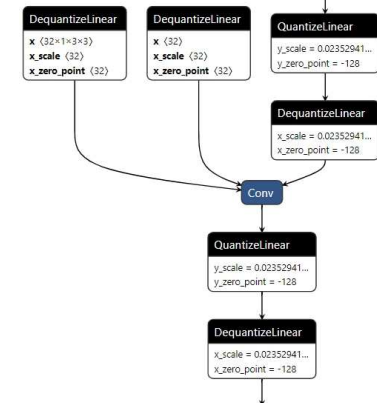
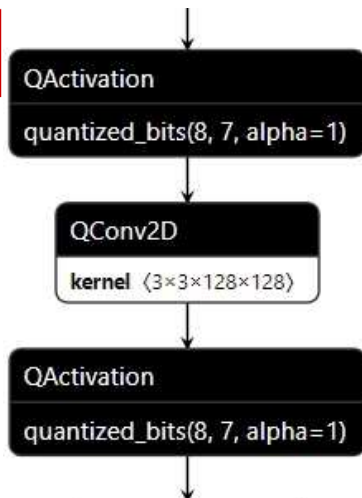
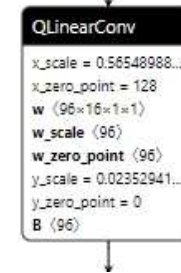
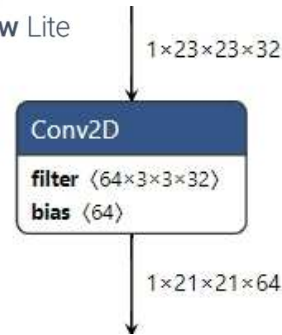
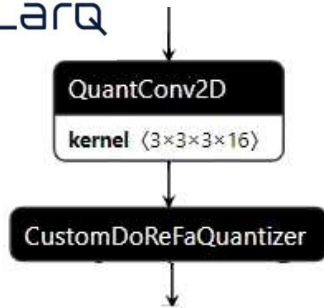
TFLite  
MEAN

=

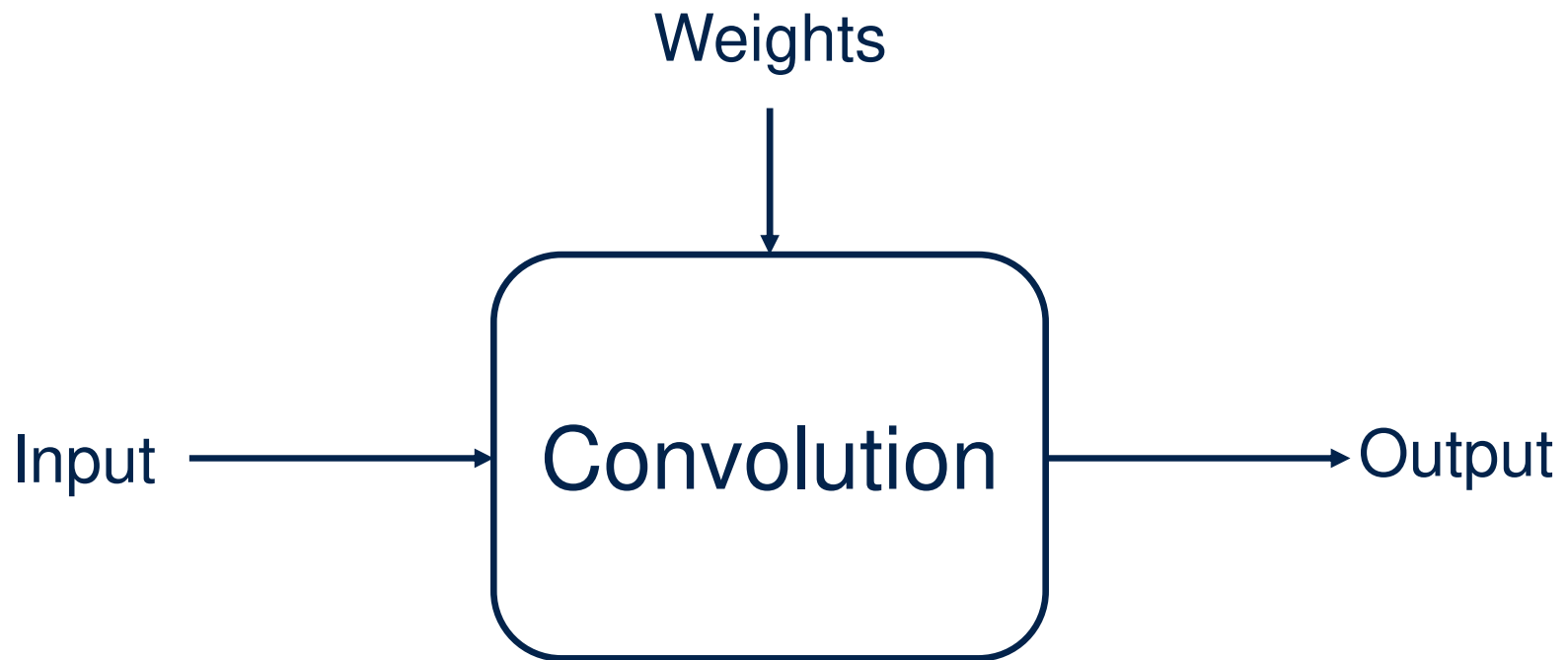
ONNX  
ReduceMean



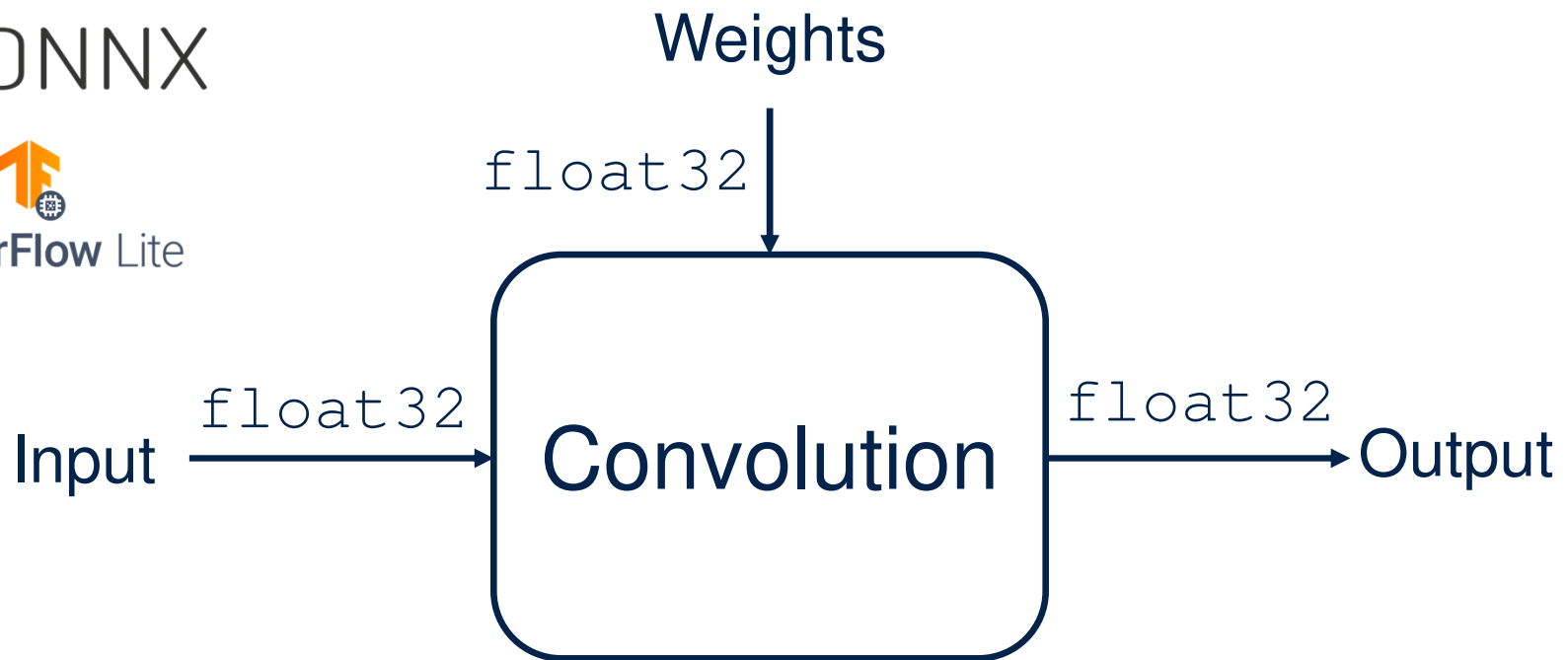
# Heterogeneity: DL Formats



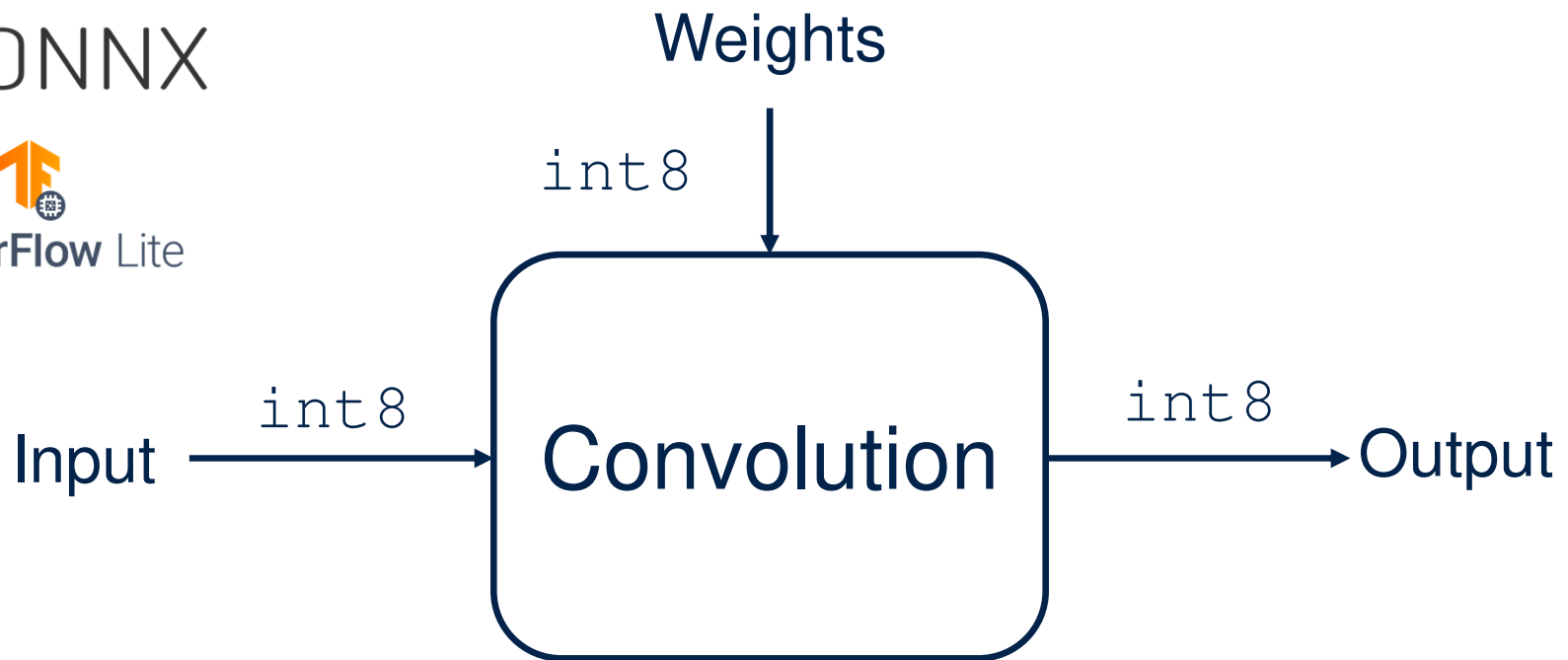
# Heterogeneity: Quantization



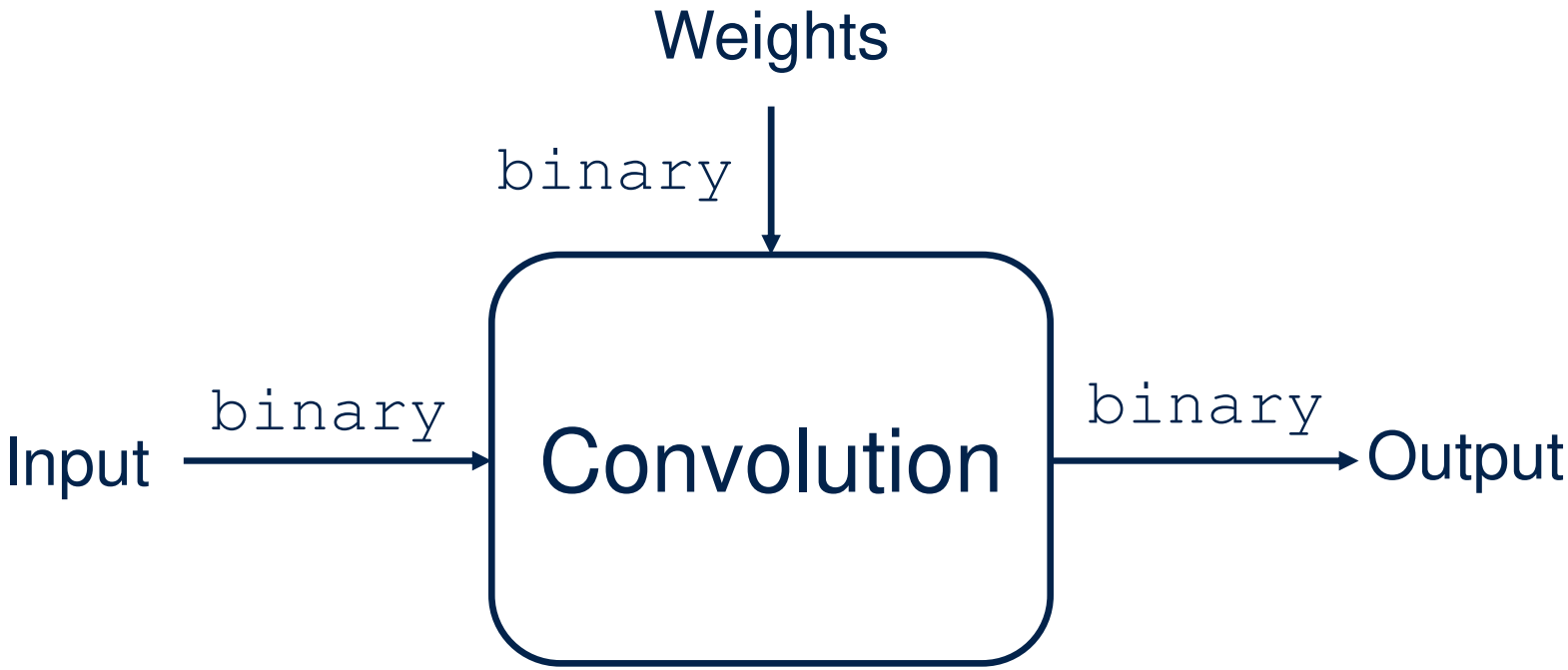
# Heterogeneity: Quantization



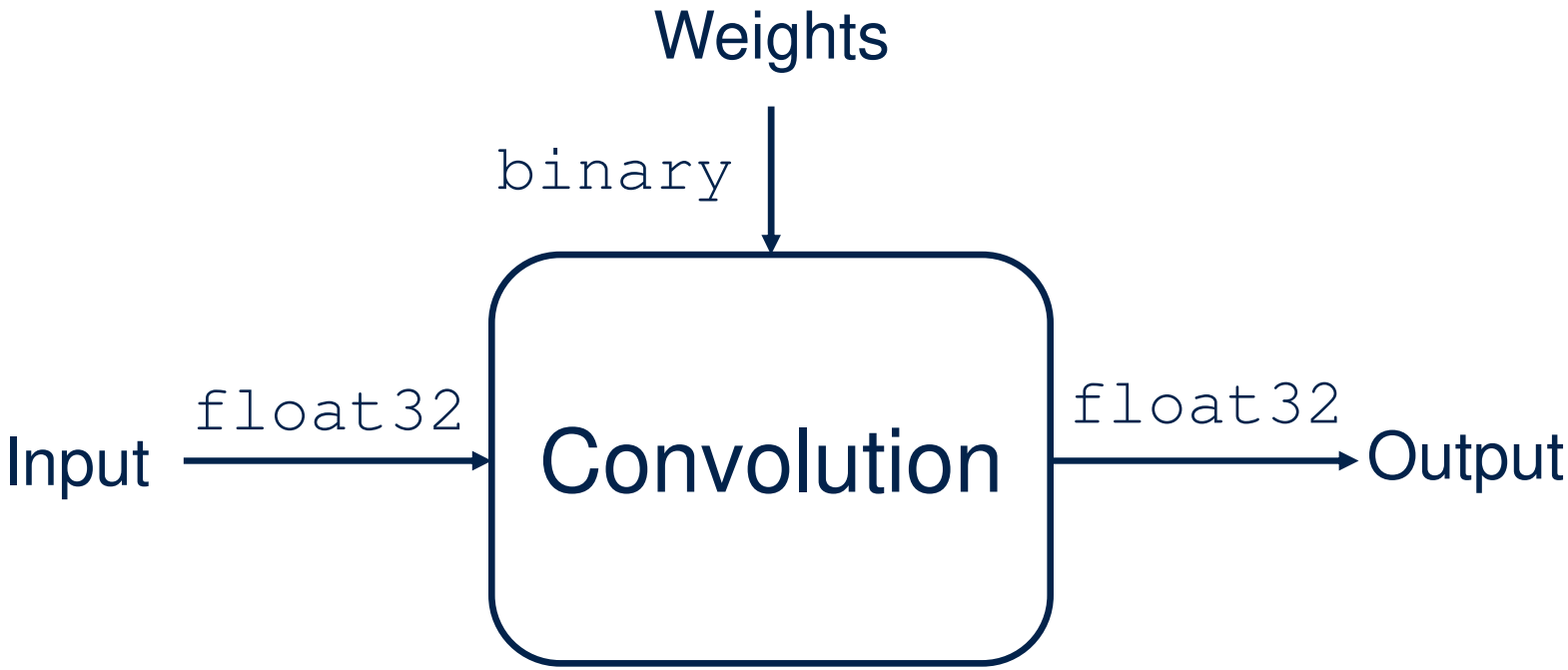
# Heterogeneity: Quantization



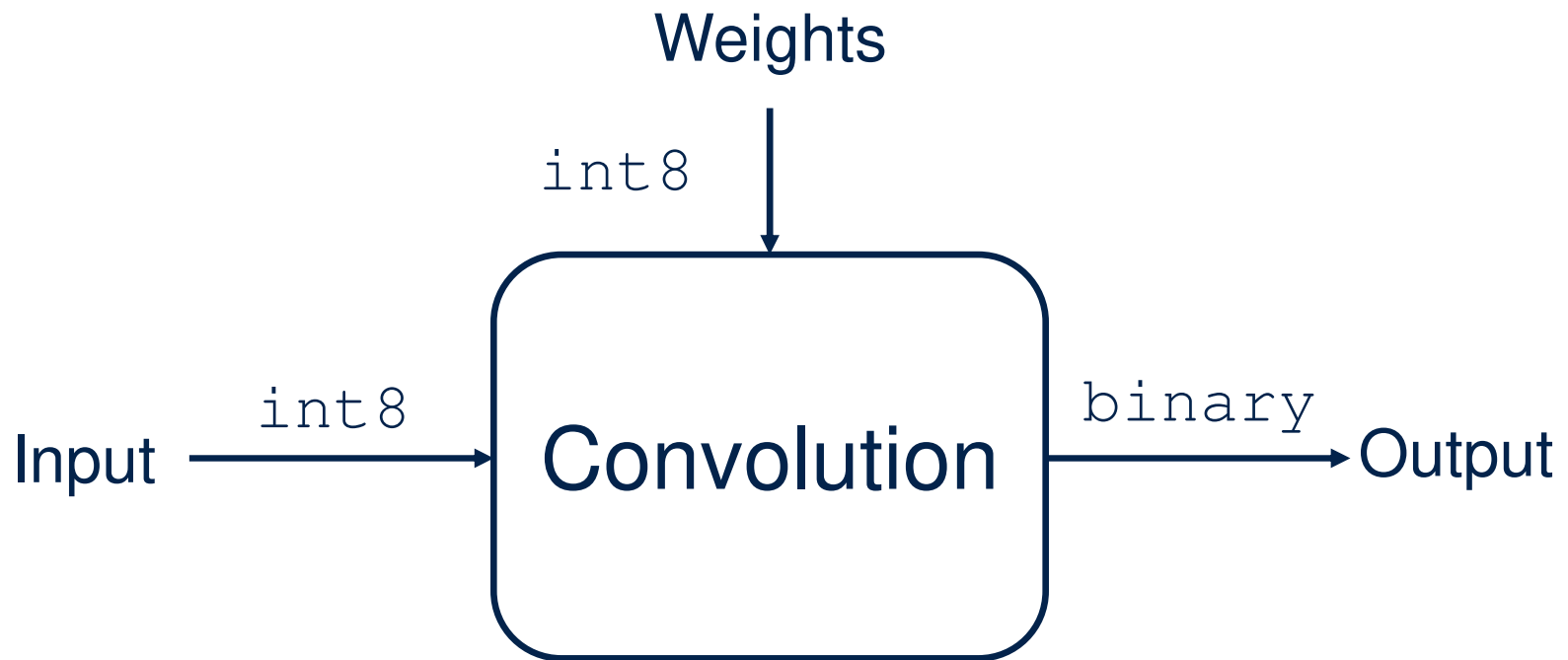
# Heterogeneity: Quantization



# Heterogeneity: Quantization



# Heterogeneity: Quantization



# Heterogeneity: Quantization



Limiting to binary, int8,  
float32 there are  
 $3^3=27$  combinations  
to be supported for a  
single layer

Input

Output



# Heterogeneity: Quantization & Topology

```
x = x_in = Input(X_train.shape[1:])
x = Reshape(x.shape[1:]+(1,))(x)
x = QActivation("quantized_bits(16, 15, alpha=1)", name="act_0")(x)
x = QConv2D(32, (1, 3),
           kernel_quantizer="quantized_bits(16, 15, alpha=1)",
           use_bias = False,
           name="conv2d_1")(x)
x = BatchNormalization()(x)
x = QActivation("binary(alpha=1)", name="act_2")(x)
x = QDepthwiseConv2D((3, 3),
                    depthwise_quantizer="binary(alpha=1)",
                    padding="same",
                    use_bias = False,
                    name="depthconv_1")(x)
x = BatchNormalization()(x)
x = QActivation("binary(alpha=1)", name="act_3")(x)
x = Flatten(name="flatten")(x)
x = QDense(16,
          kernel_quantizer="binary(alpha=1)",
          bias_quantizer="binary(alpha=1)",
          name="dense_1")(x)
x = Activation("relu")(x)
x = Dense(y_train.shape[-1],
          name="dense_out")(x)
x = Activation("softmax", name="softmax")(x)
```

# Heterogeneity: Execution Targets

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

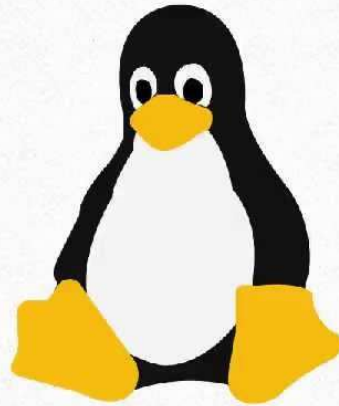
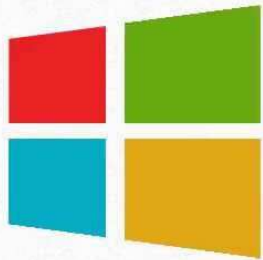
Different computational power, e.g.,  
MLC, ISPU  
STM32 std MCU family, STM32N6, STM32MP2, Stellar-E  
x86-64

Different specialized functional units, e.g.,  
Binary accelerator in ISPU  
Integer SIMD/Vector ISA in ARM  
Convolution accelerators in STM32N6

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

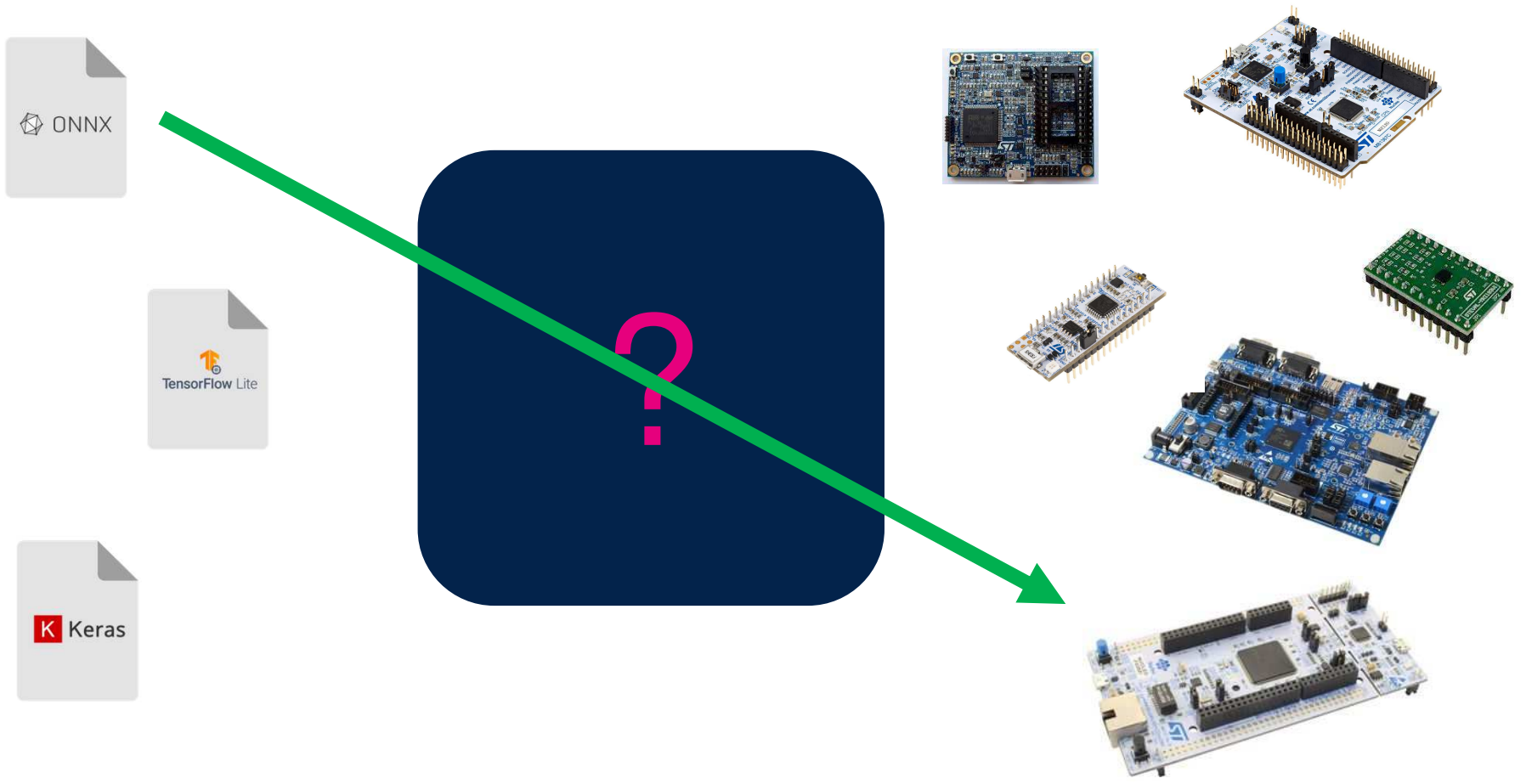
# Heterogeneity: Operating Systems

MacOSX (x86, ARM)

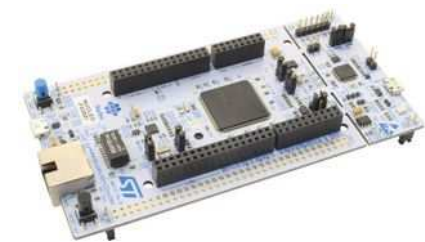
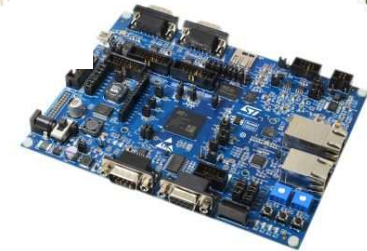
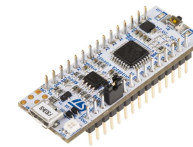
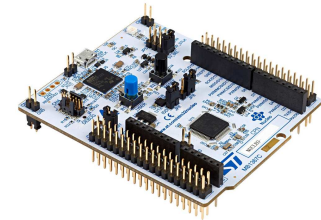


MacOSX ARM based requires specific build  
Emulation platform does not support execution of x86  
version because of TensorFlow

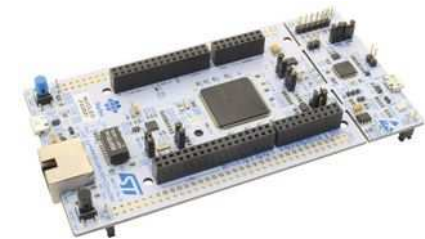
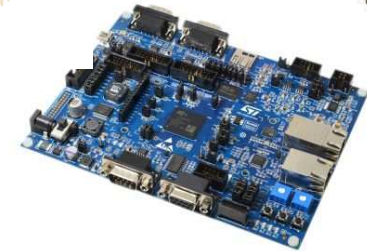
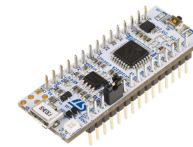
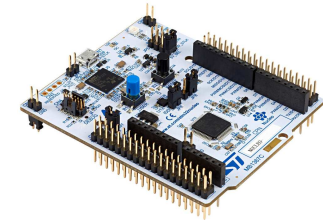
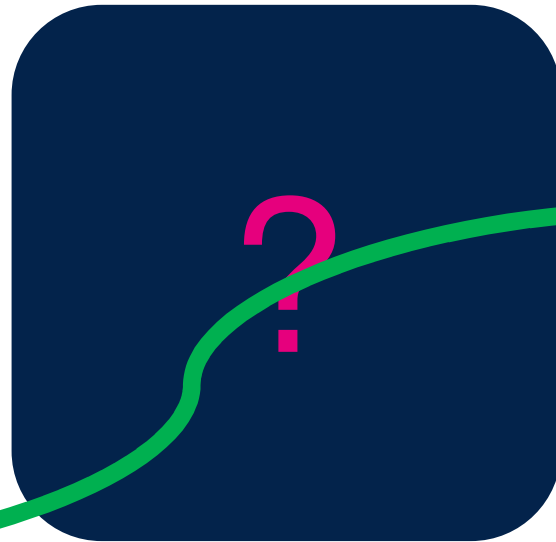
# Paths from Source to Target



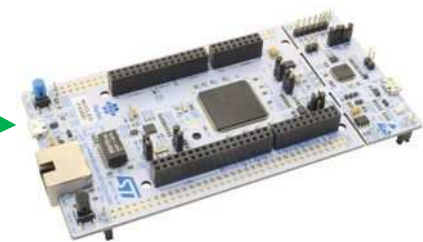
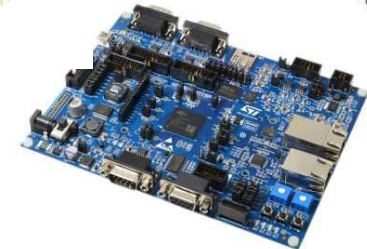
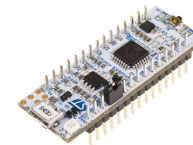
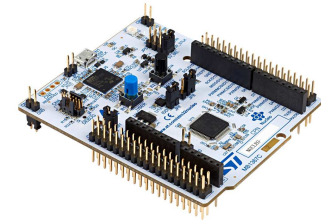
# Paths from Source to Target



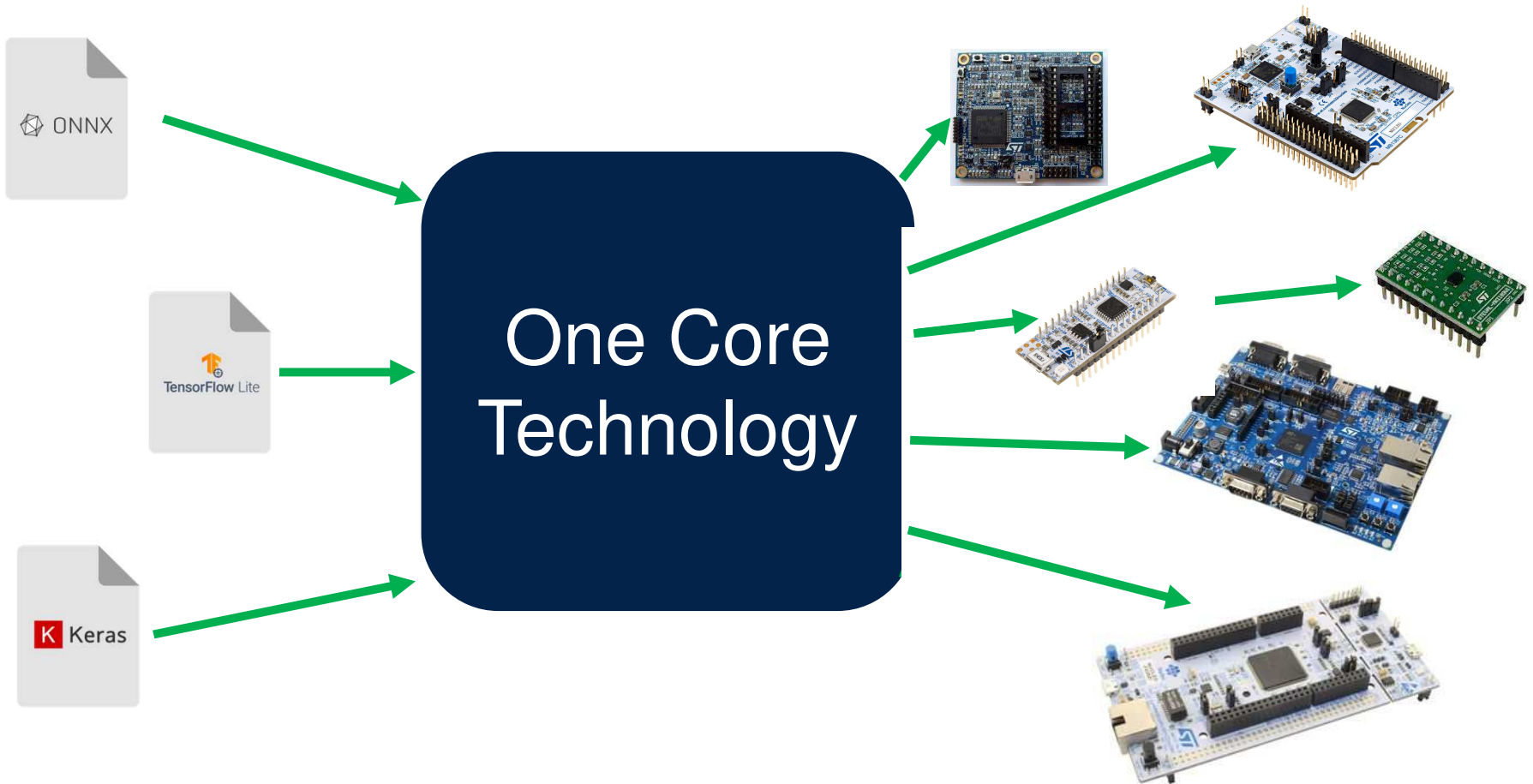
# Paths from Source to Target



# Paths from Source to Target

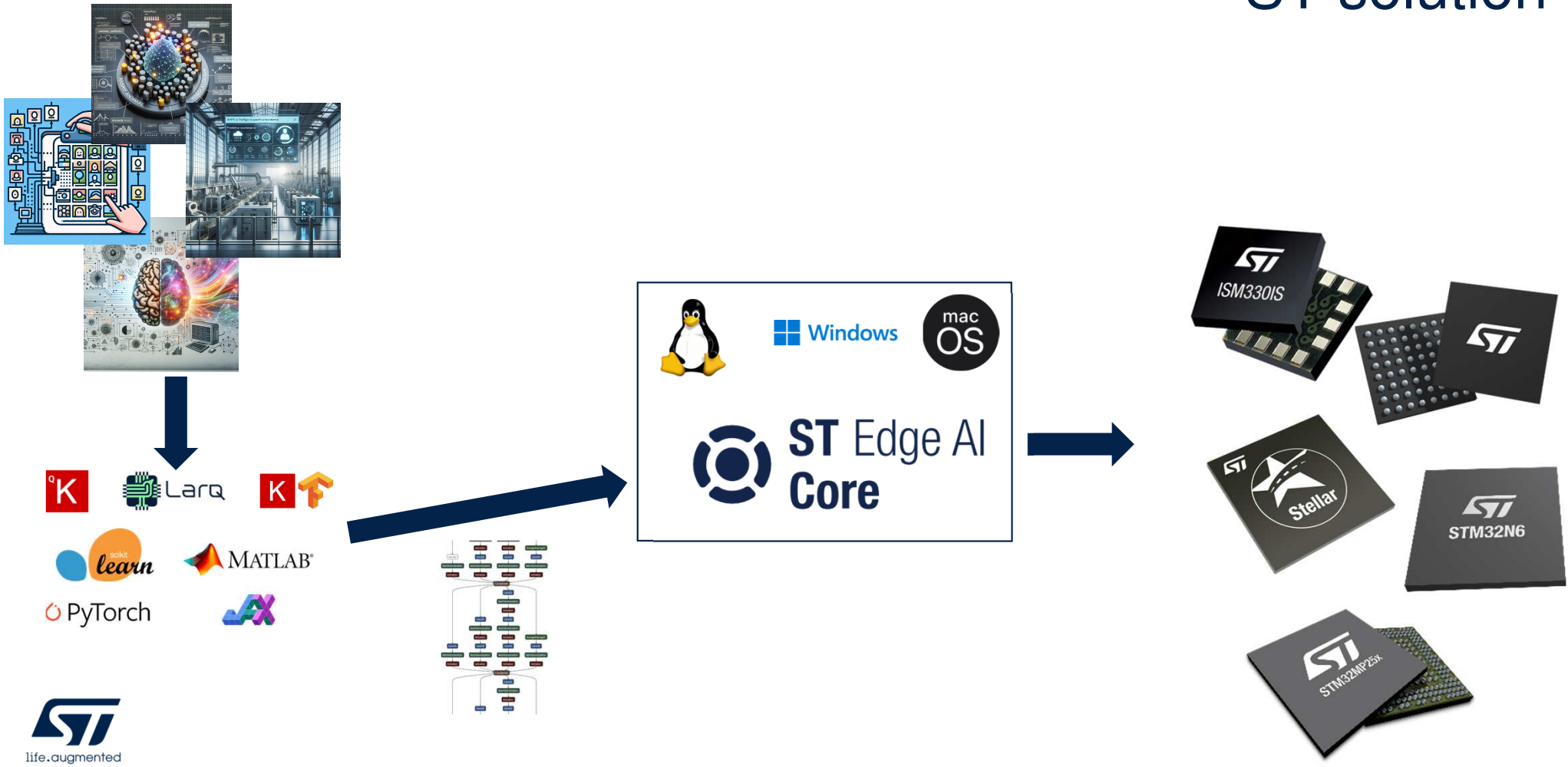


# Homogenize Developer Experience

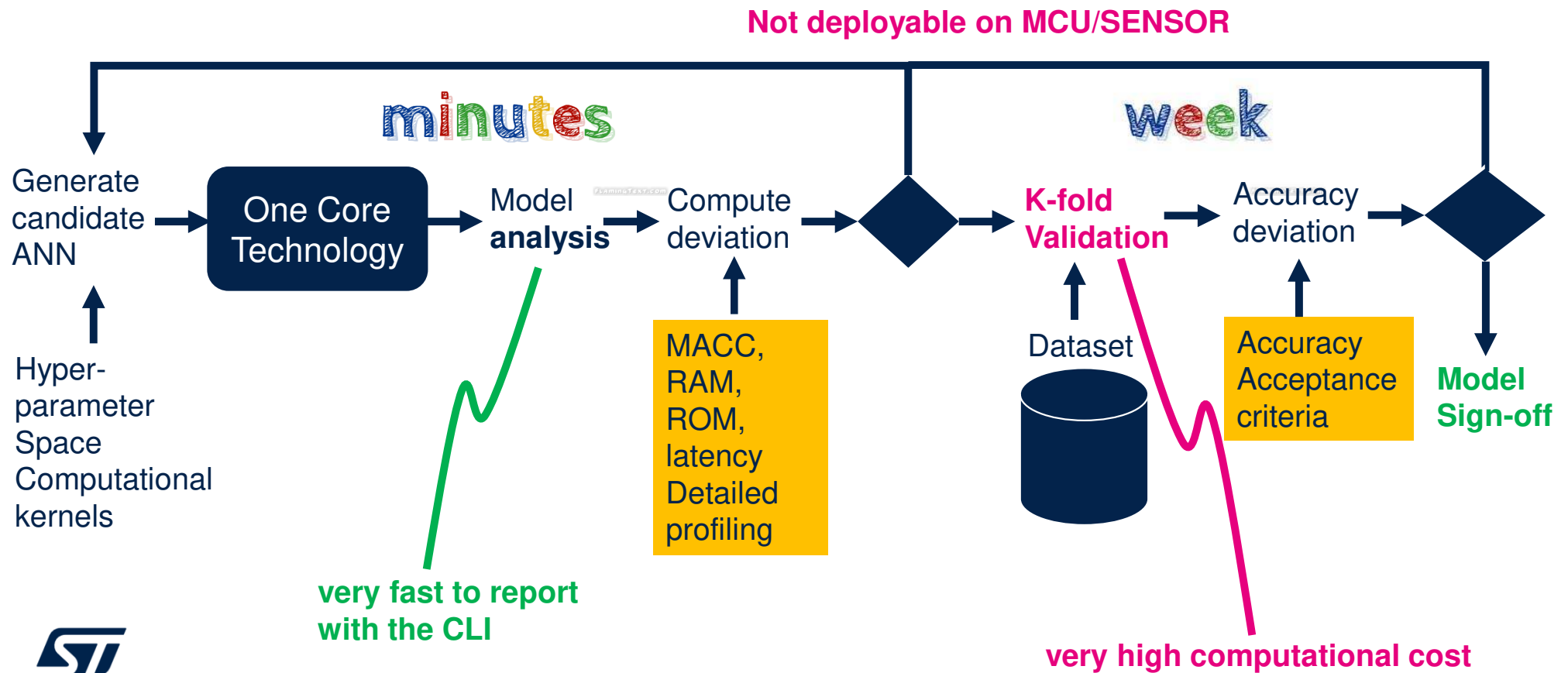




# Artificial intelligence on the edge ST solution



# Deployment aware NAS/HPO

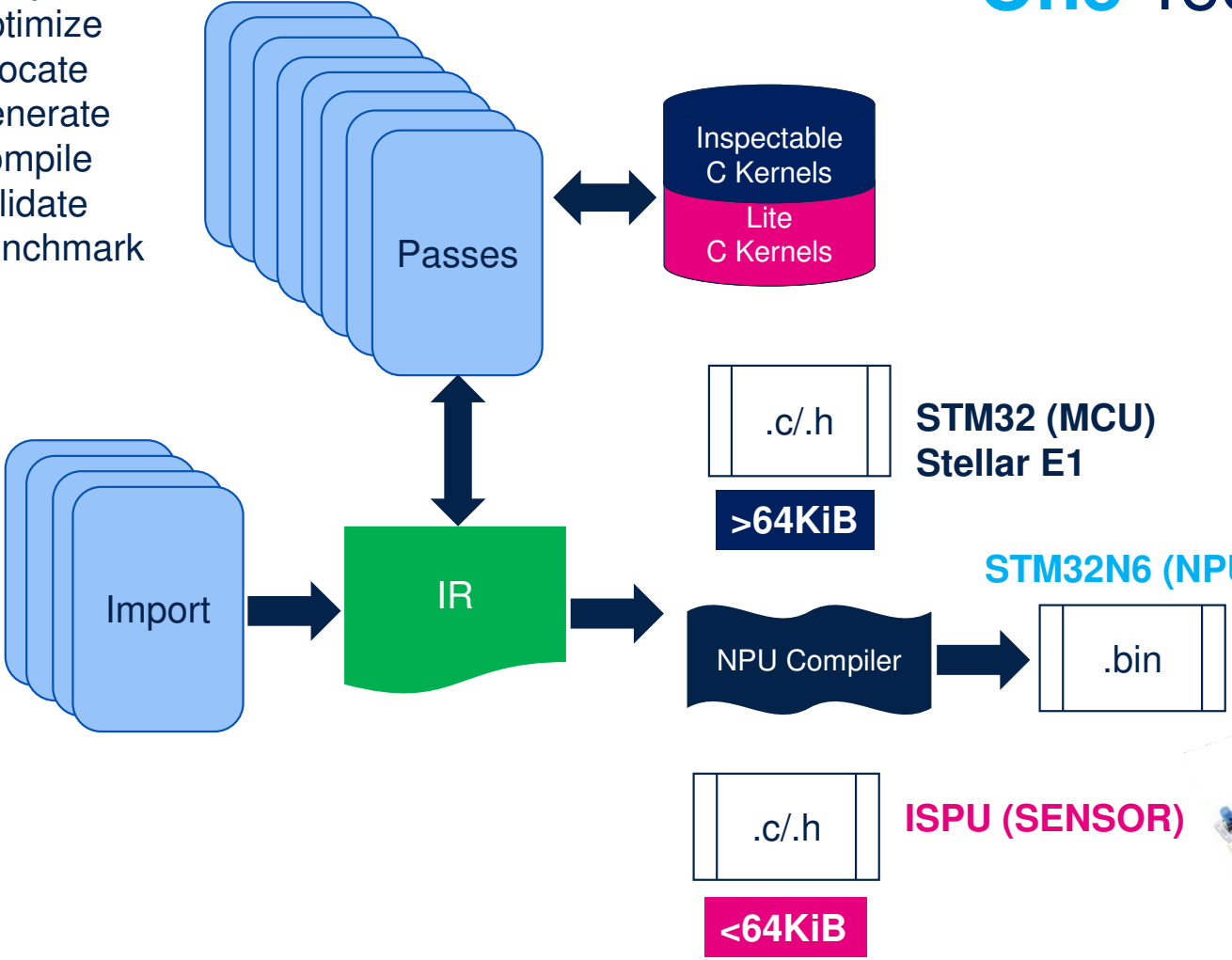


# One Technology

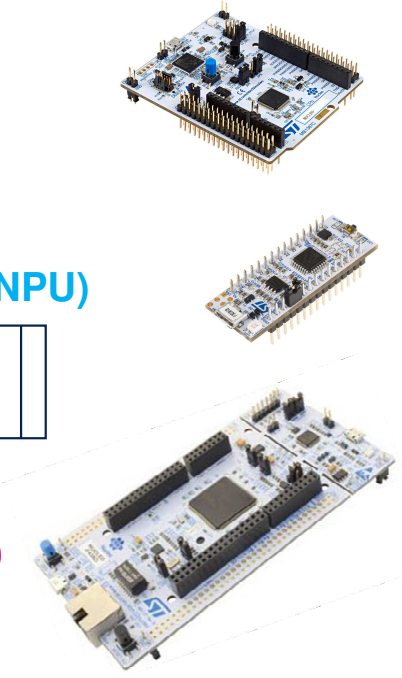
- Analyze
- Optimize
- Allocate
- Generate
- Compile
- Validate
- Benchmark



QKeras  
Larq  
Keras  
{JSON}  
ONNX  
TensorFlow Lite  
ST  
life.augmented

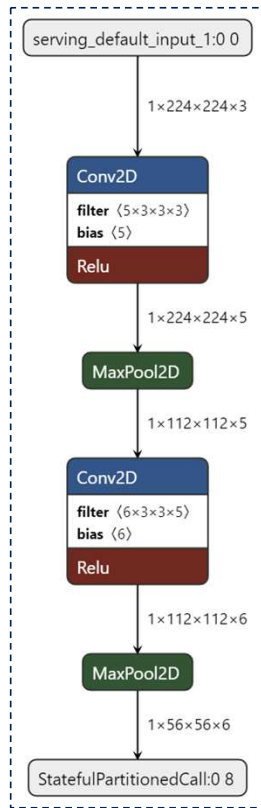


Add More Targets



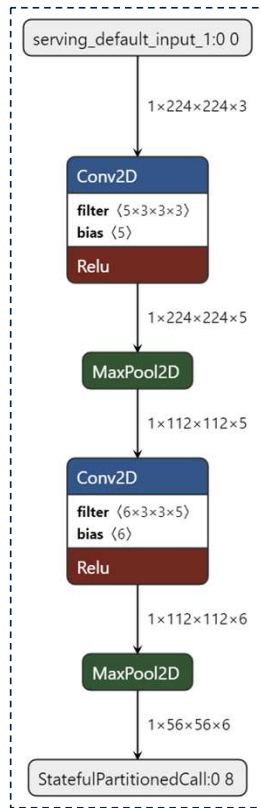
# From the ML model to its deployment

python

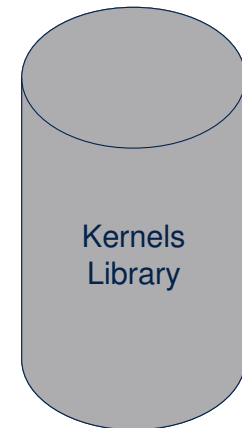


# From the ML model to its deployment

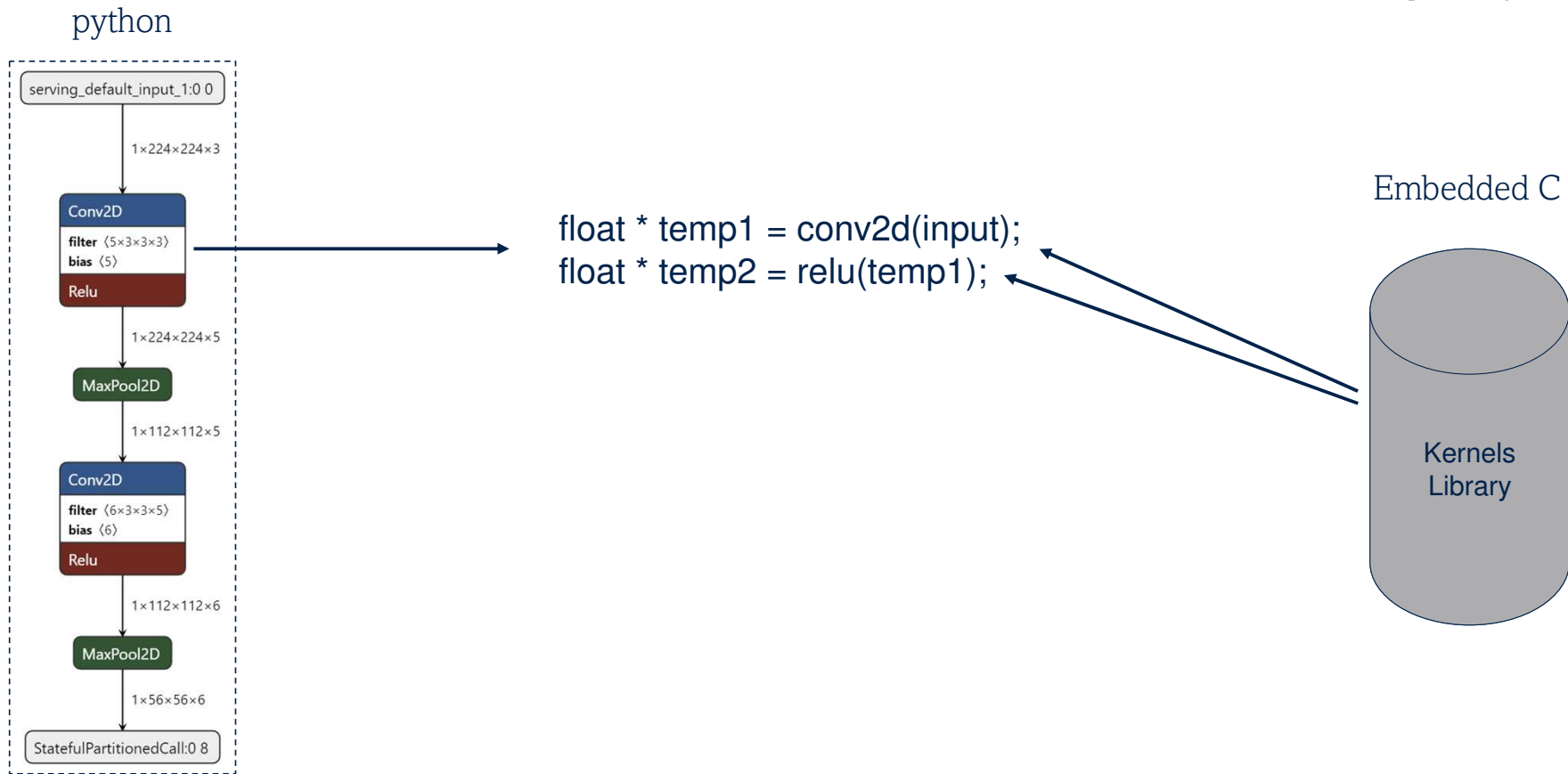
python



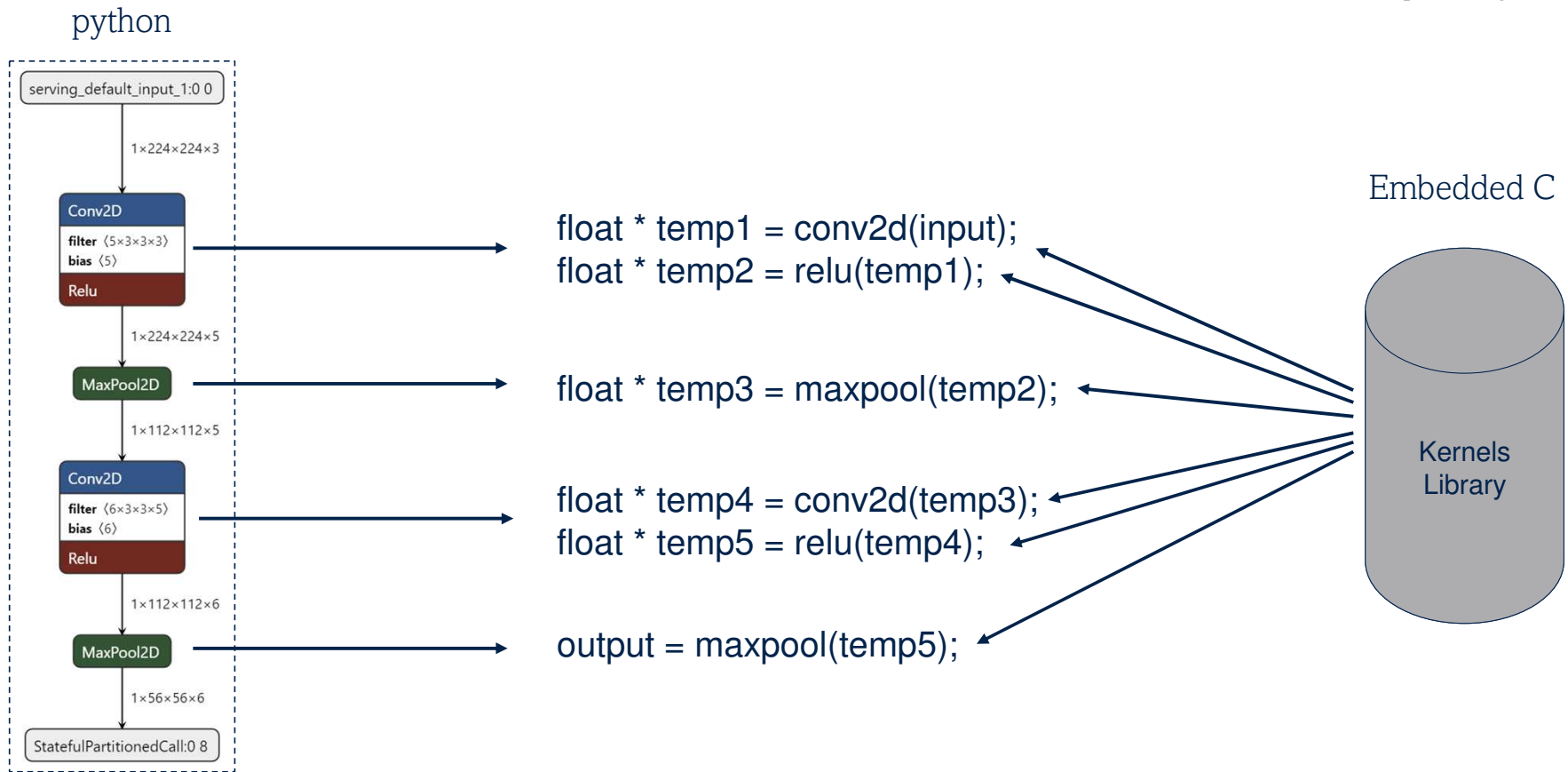
Embedded C



# From the ML model to its deployment

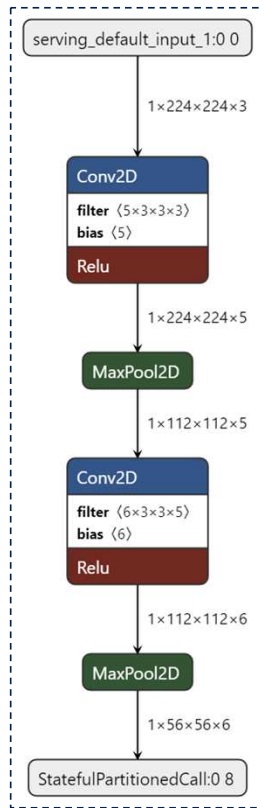


# From the ML model to its deployment



# From the ML model to its deployment

python



```
void network(float * input, float * output) {
```

```
float * temp1 = conv2d(input);  
float * temp2 = relu(temp1);
```

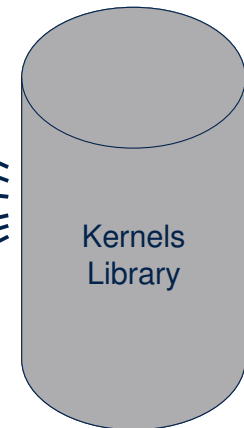
```
float * temp3 = maxpool(temp2);
```

```
float * temp4 = conv2d(temp3);  
float * temp5 = relu(temp4);
```

```
output = maxpool(temp5);
```

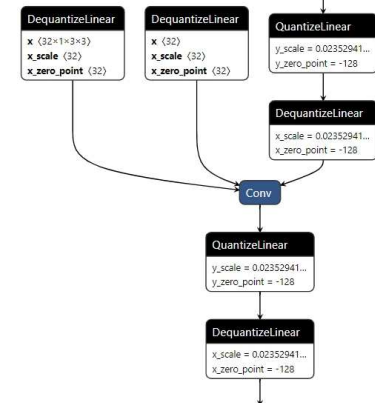
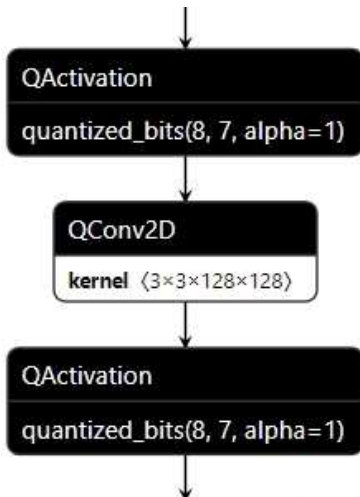
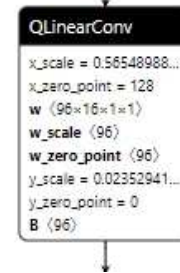
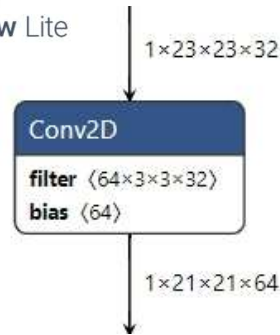
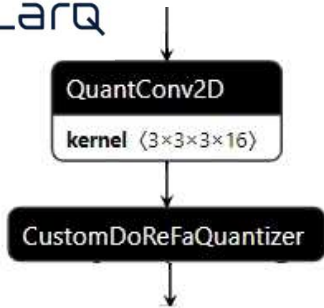
```
}
```

Embedded C

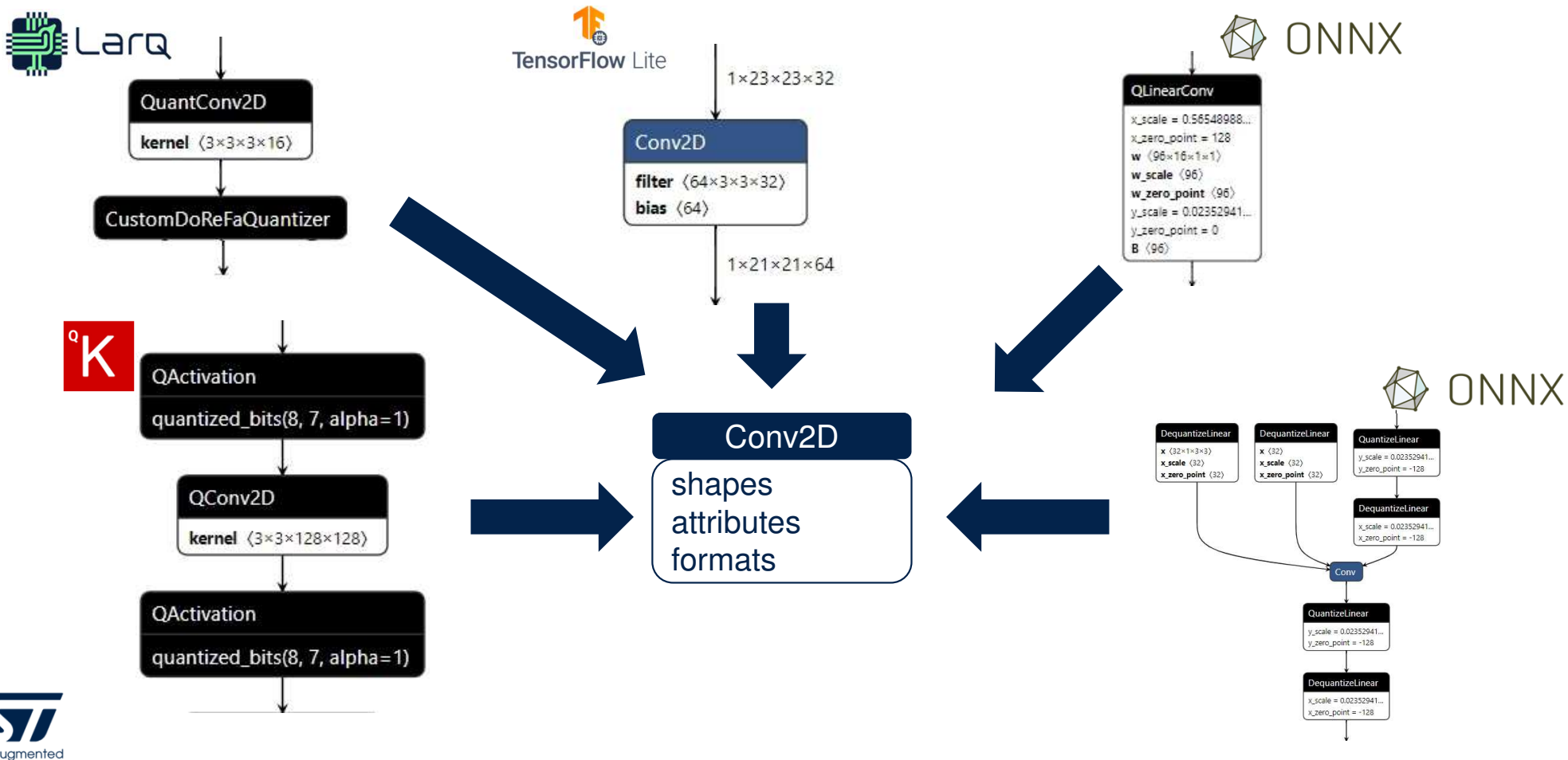




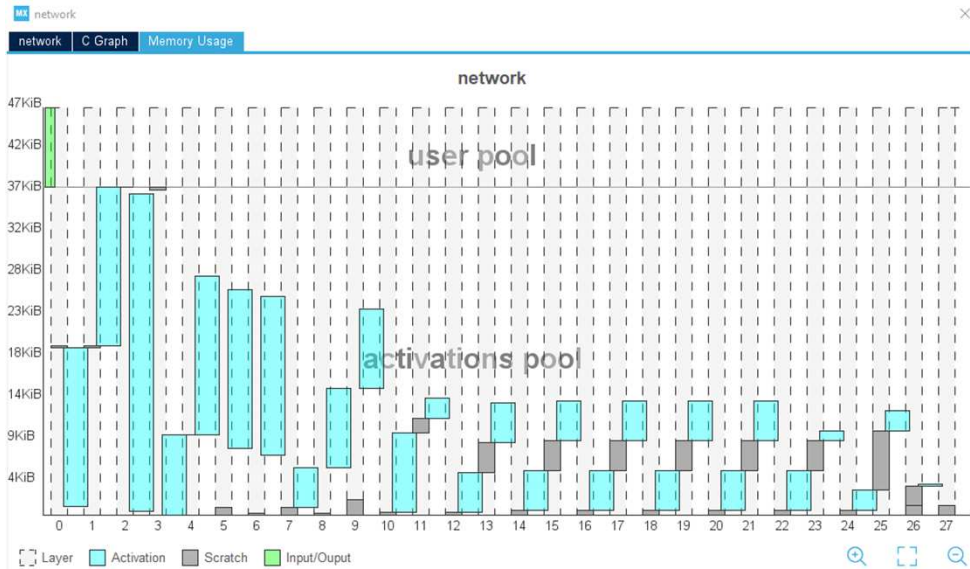
# Removing Deep Learning Frameworks Differences



# Removing Deep Learning Frameworks Differences



# Heterogeneity: Execution Targets



Memory optimizer

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

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

- Different memory infrastructure e.g.,
  - Single memory component
  - Multiple and homogeneous memories
  - Multiple and heterogeneous memories

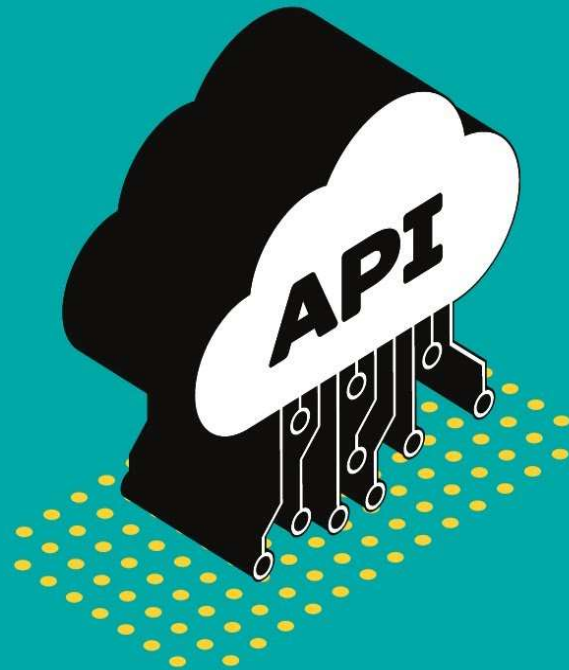
# Hiding Target Differences



- Different Optimization Objectives
  - **Time:** to minimize inference time
  - **RAM:** to minimize use of memory
  - **Balanced:** trade-off between inference time and memory usage
- Multiple networks instancing in the same application

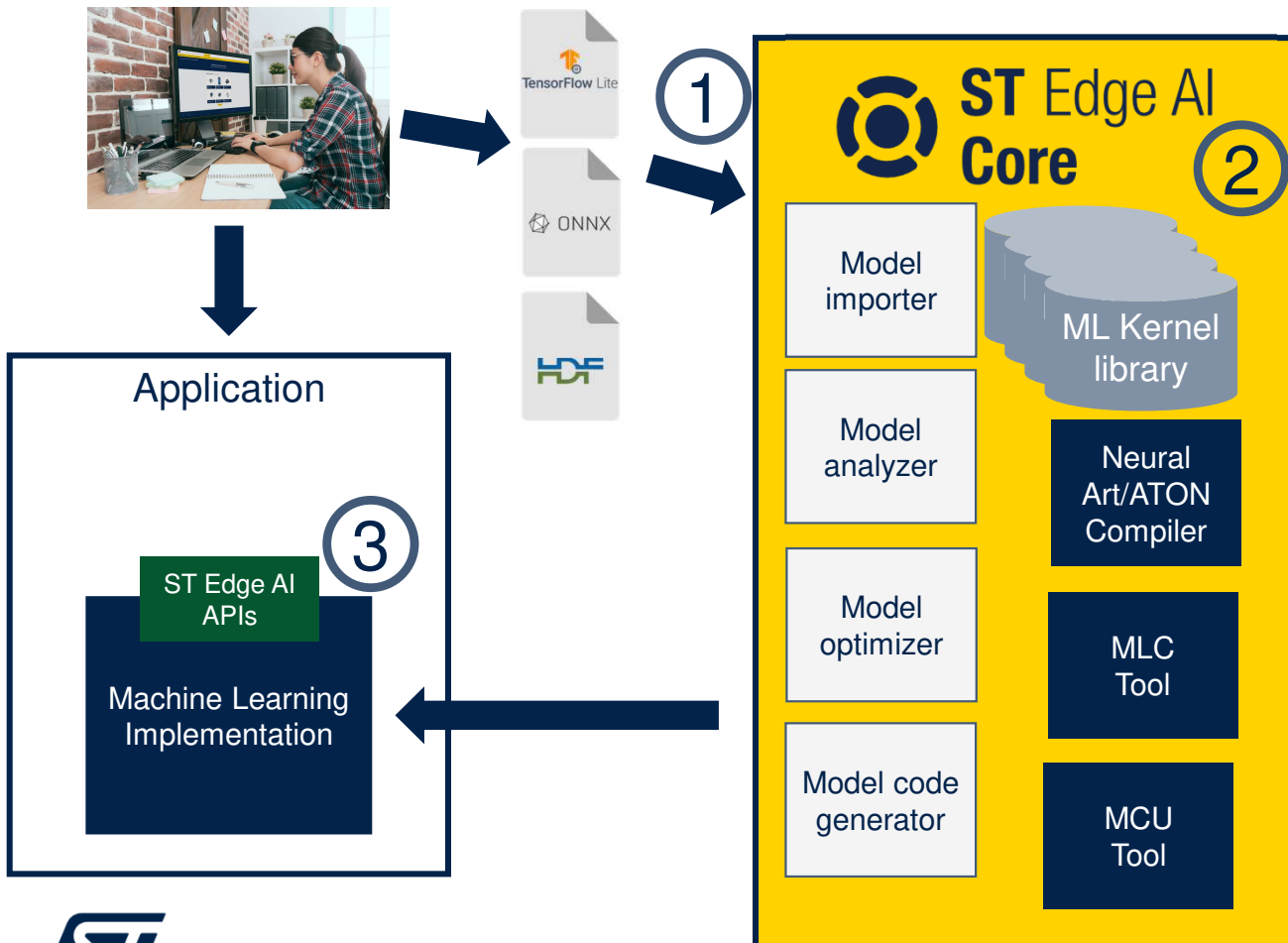
# Hiding Target Differences

## Public APIs



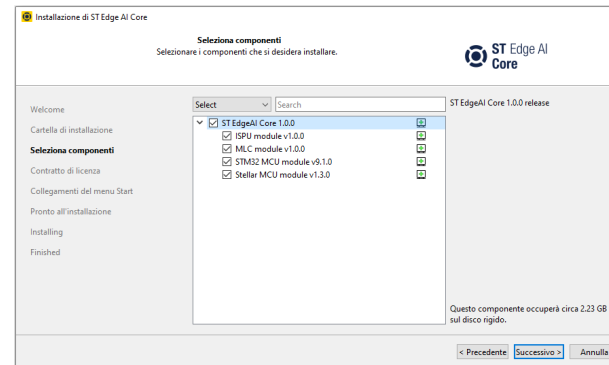
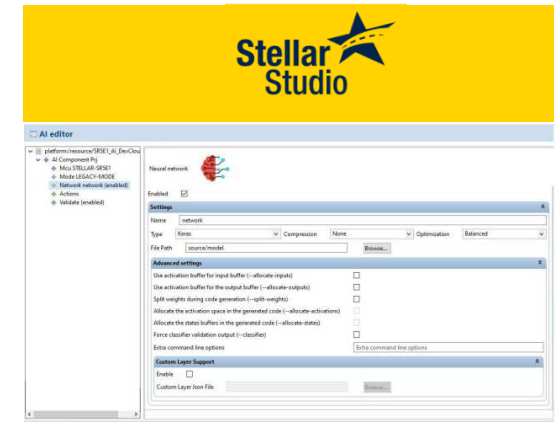
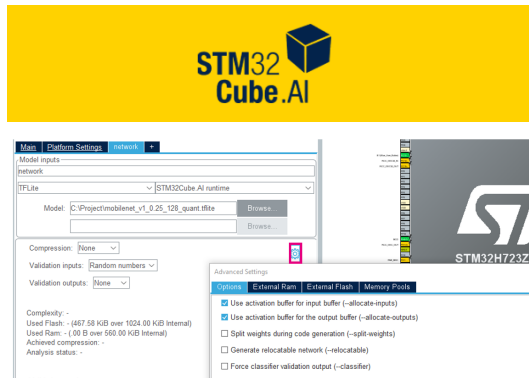
- Simplified **public APIs**: Init + run + deinit
- **Same** for all the execution targets
  - Same application code means portability on all the supported target
- Internal code is optimized for the different targets exploiting the heterogeneous hardware capabilities
- **Extra APIs** for NPU

# Unified View



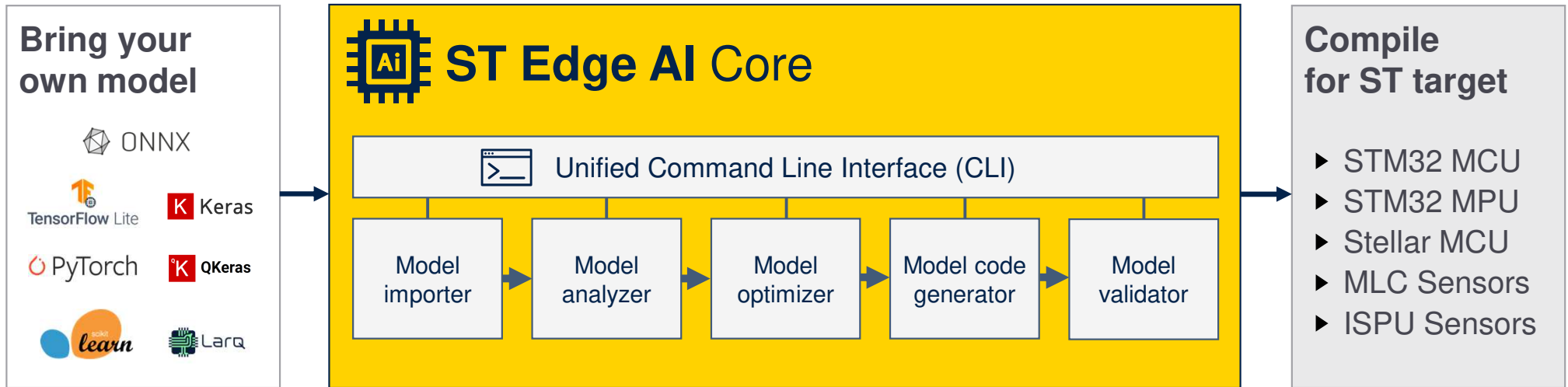
1. Single entry point for end-users: provide machine learning models + options
2. Multiple modules and tools included to optimize solutions for different targets
3. All solutions have the same interface to simplify integration in applications

# Integration



# ST Edge AI Core technology

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



MLPerf Tiny



ML Commons

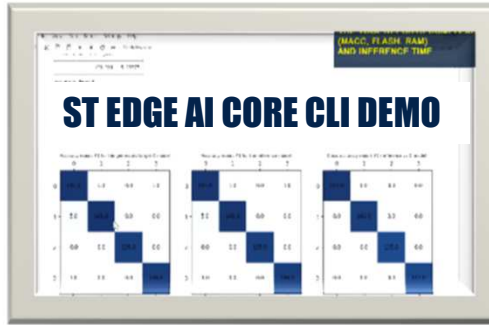
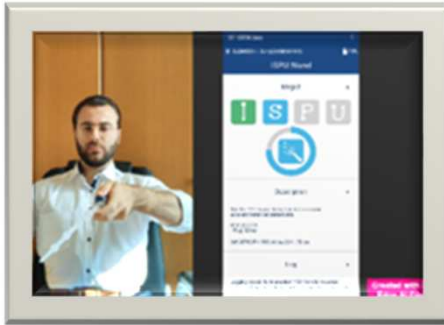


# ST Edge AI Core technology

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

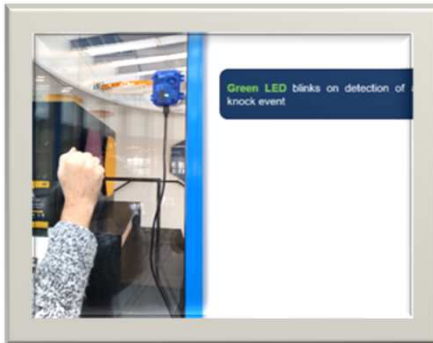


**ISPU-  
Based  
Magic  
Wand**



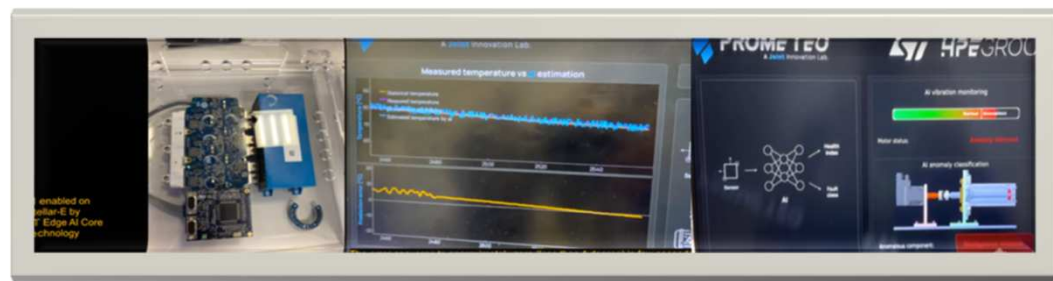
**Camera  
Based Hand  
Gesture  
Recognition  
on STM32N6**

**Door Knock  
Recognition  
On ISPU**



**TOF Based  
Hand  
Gesture  
Recognition  
on  
STM32F4**

**Vehicle  
Monitoring  
on MLC**



**eMotor  
Monitoring  
on  
Stellar-E  
(by HPE)**



life.augmented

# ISPU Wand with ST Edge AI Core Technology

STMicroelectronics



life.augmented

# MLC Truck with ST Edge AI Core Technology

STMicroelectronics



life.augmented

# ST EDGE AI CORE TECHNOLOGY enables TOF-based Hand Pose Recognition

STMicroelectronics





life.augmented

# ST EDGE AI CORE TECHNOLOGY enables HW accelerated Camera-based Hand Pose Recognition

STMicroelectronics

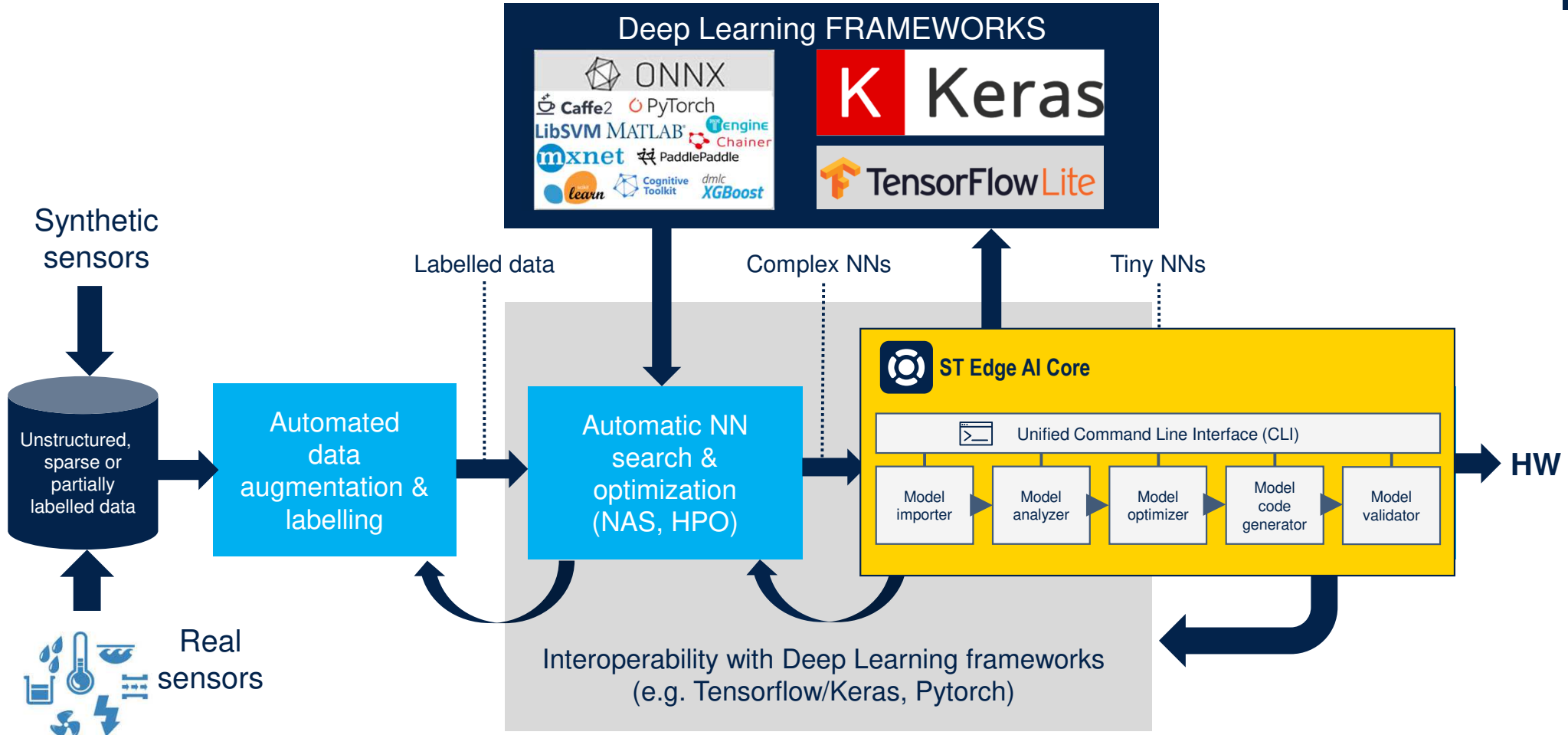


life.augmented

# AI enhancing e-Motor efficiency powered by ST Edge AI Core Technology

STMICROELECTRONICS

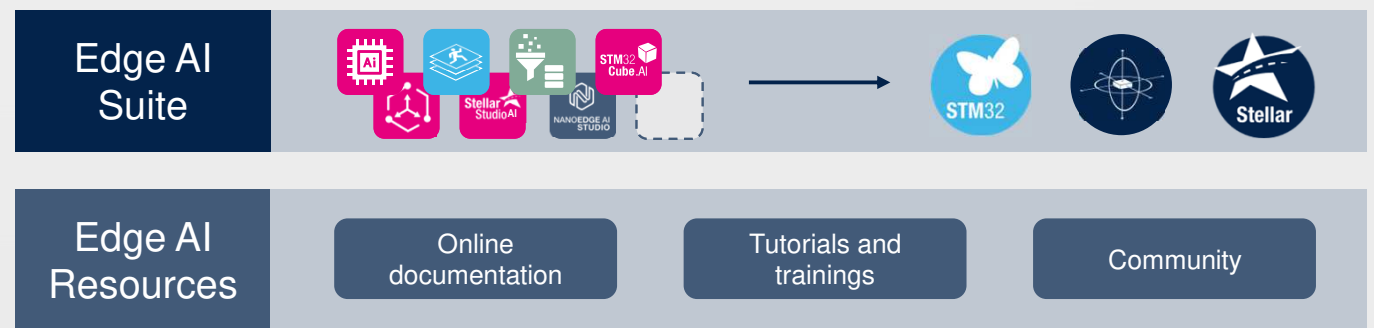
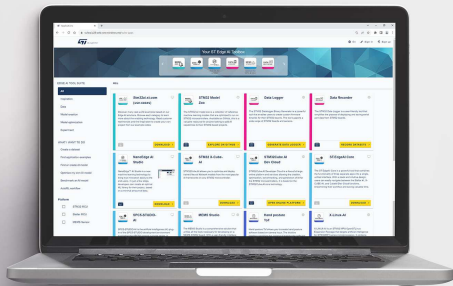
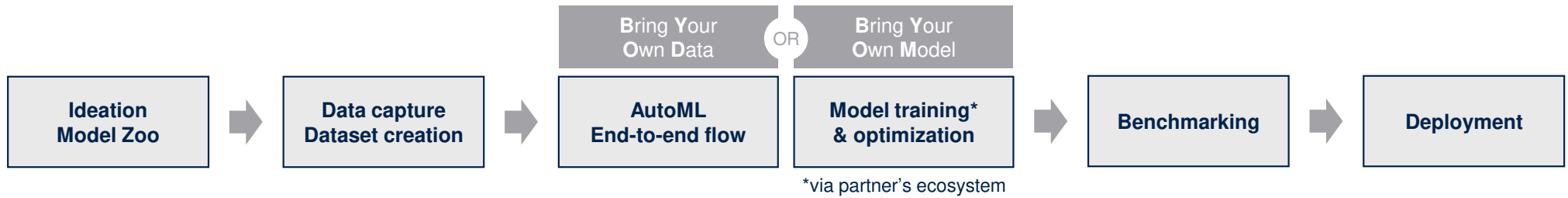
# Tiny ML means an ecosystem too





# ST Edge AI Suite at a Glance

The most complete developer-centric approach to accelerate the deployment of edge AI



# ST Edge AI Suite at a Glance

The most complete developer-centric approach  
to accelerate the deployment of edge AI



**SCAN ME**

# Next Challenges for the Community

On-Device Learning



# Next Challenges for the Community

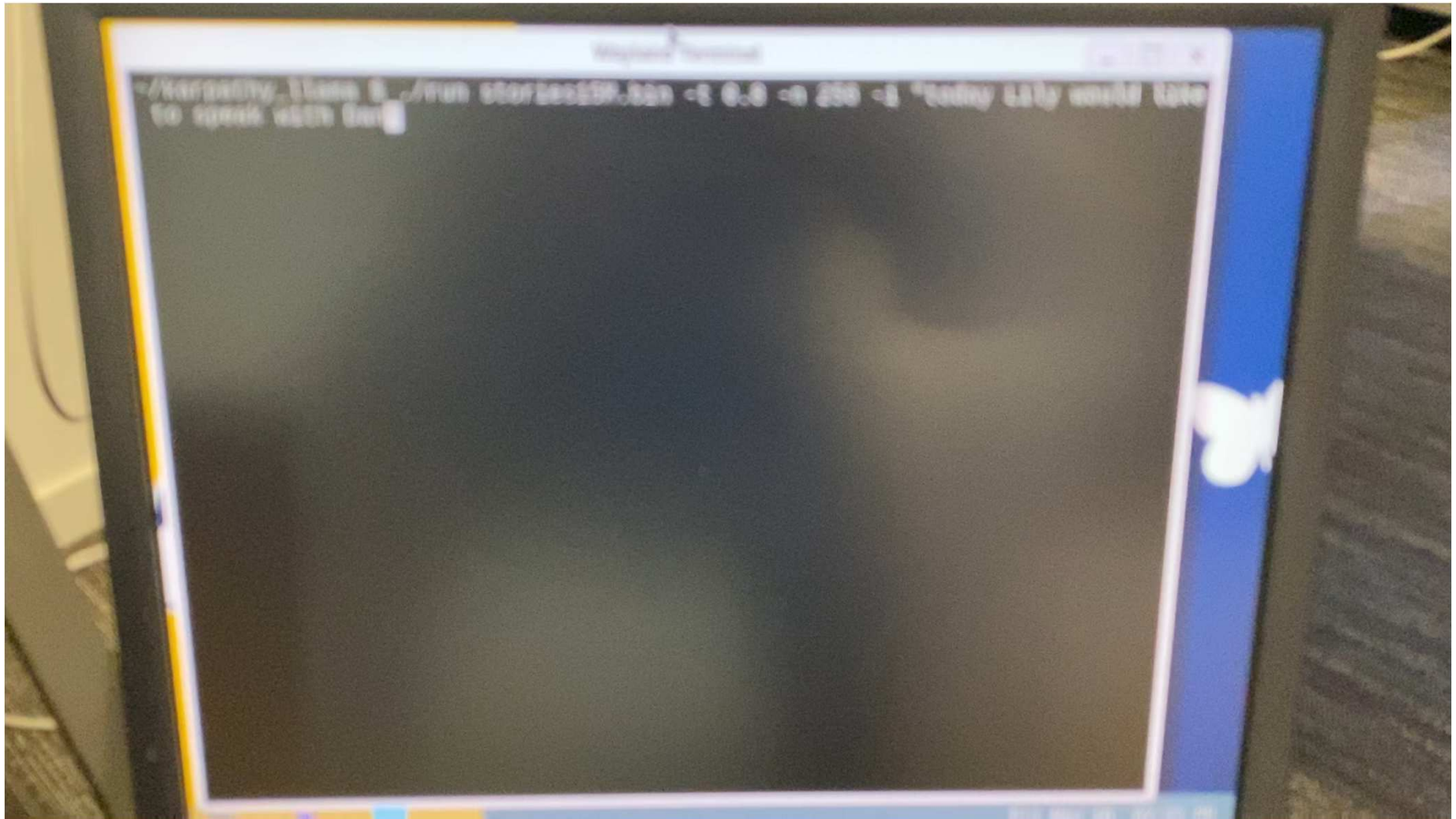
On-Device Learning

Edge Generative AI

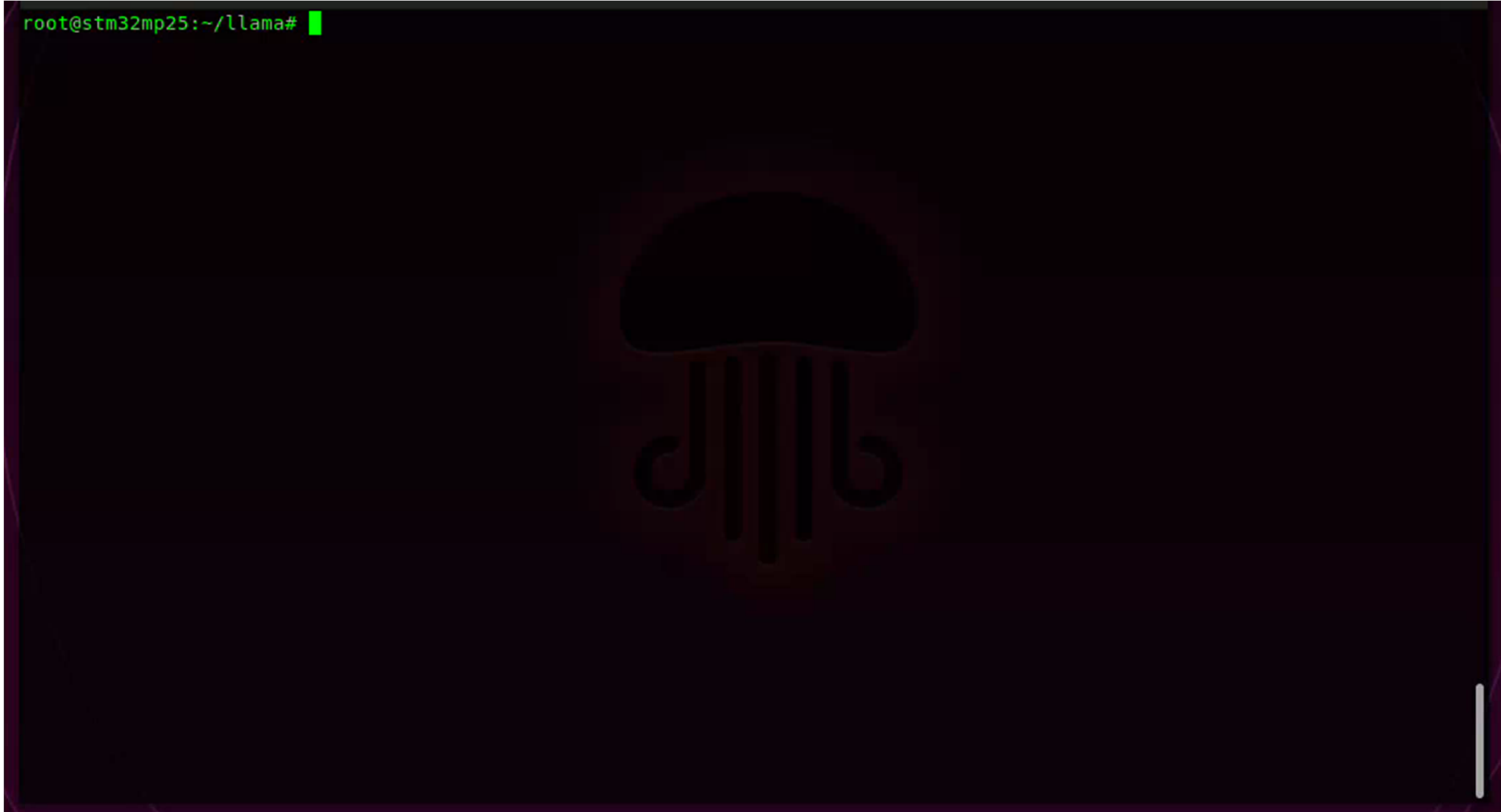
**NEW**



# Tiny generative AI (15M-fp32) on STM32MP1



# Tiny generative AI (1.1B-2bits) on STM32MP2



## LLava on STM32MP2



*The man is sitting on the side of a road, which is covered in debris. He appears to be in a state of distress. It seems like he is also in a state of pain. The ground he is sitting on is made of concrete, and there are other people standing in the distance. It looks like there are also orange lights visible in the distance, which could be part of a rescue operation. The scene is quite dramatic and tense.*

# Timeline

Reduced memory region based deep Convolutional Neural Network detection (1)

A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm (2)

ML for visual search (ISO/IEC/SC29/WG11 CDVS)

2011

2016

2017

2018

2019

2020

2021

2022

2023



edge AI for sensors (MLC)

edge AI on Automotive MCUs

Introduction of autoML

edge AI Developer Cloud

ST Edge AI Core

edge AI on STM32 at CES

TinyML membership

edge AI for ST Linux

edge AI for sensors (ISPU\*)

edge AI on latest Automotive MCU



- (1) <https://ieeexplore.ieee.org/document/7684706>  
(2) <https://ieeexplore.ieee.org/document/7870349>

\*Intelligent Sensor Processing Unit



# Our technology starts with You



Find out more at [www.st.com](http://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](http://www.st.com/trademarks).

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



life.augmented