# tinyML® Summit

*Miniature dreams can come true…*

**March 28-30, 2022 | San Francisco Bay Area**

TINY ML

www.tinyML.org

life.augmented

# Brains into sensors with AI in the Edge

Andrea Onetti
Executive Vice President MEMS Sensors Sub-Group
Analog, MEMS and Sensors Group

STMicroelectronics

# Is it a far-off future?

life.augmented

# The MEMS journey

## Offline era

**2000**

**A paradigm change in the man-machine interface**

MEMS technology:
from a concept to a product

## Online era

**2010**

**Sensor's proliferation and connections to Cloud**

Performance improvement and technology fusion

## Onlife era

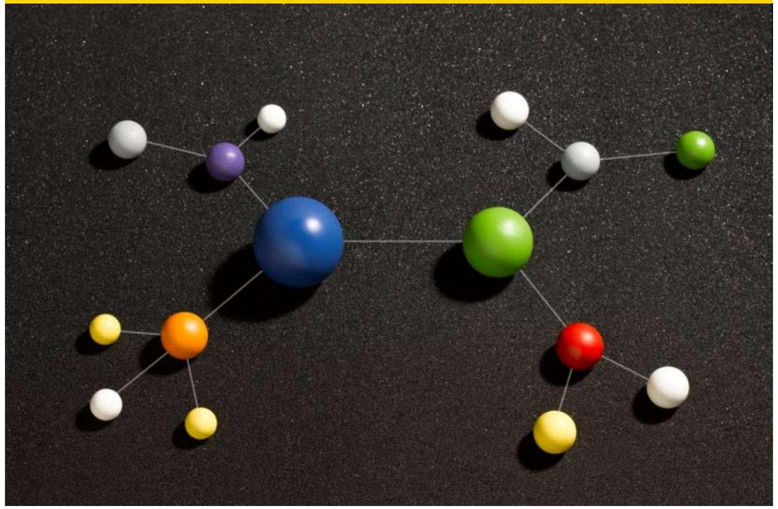
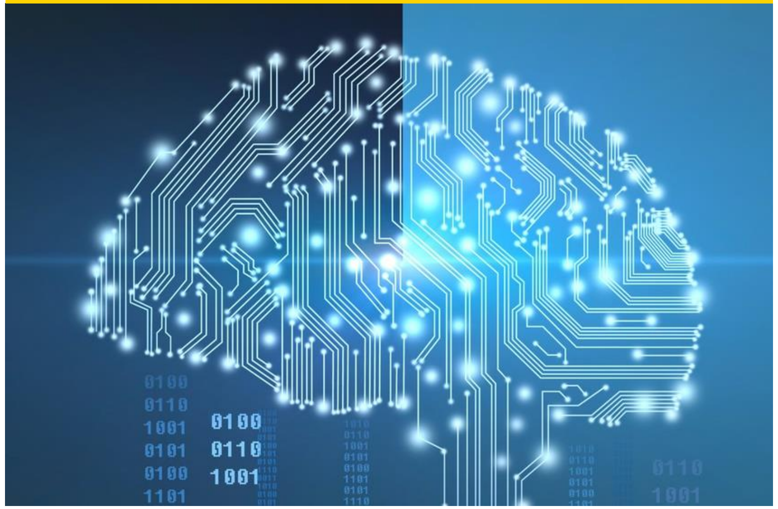
**2020**

**The fusion of technology and life**

Standalone devices able to sense, process and take action

life.augmented

# Systems where sensors live: the evolution

| Offline era | Online era | Onlife era |
|---|---|---|
| **Fragmented** | **Connected** | **Trained** |
|  |  |  |
| The simplest configuration: independent systems | Intertwined nodes enable efficient data exchange | Edge AI local decision making with maximum privacy |

life.augmented

# Industry 5.0 challenges


Source: A futuristic perspective on human-centric assembly - ScienceDirect

# Sensor's semiconductors challenges for Edge AI

| SUPER TINY SILICON | |
|---|---|
| | OPTIMAL POWER (uW) PER PERFORMANCE CAPABILITIES |
| | LIMITED LOGIC AND MEMORY STORAGE |
| | TECHNOLOGY ARCHITECTURE |
| | ECOSYSTEM COMPATIBILITY |
| | ADVANCED TOOLS FOR INCREASED PRODUCTIVITY |

# Migrating intelligent processing From "on the Edge" to "in the Edge"

| Sensor + MCU | rPU | ISPU |
|---|---|---|
| **Microcontroller** | **reconfigurable Processing Unit** | **Intelligent Sensor Processing Unit** |
| Sensor — MCU | rPU+ Sensor — MCU | ISPU + Sensor — MCU |
| MCU standalone or hosted in the sensor package | rPU integrated in the sensor ASIC | ISPU integrated in the sensor ASIC |
| **Standard** MCU runs the algorithms | **Optimized** reconfigured through register setting | **Programmable** dedicated instruction set |
| **Runs any kind of SW** provided it fits the MCU specs | **Constrained** runs same model/mapping | **Runs several AI algorithms** Full precision to 1-bit NN |

life.augmented

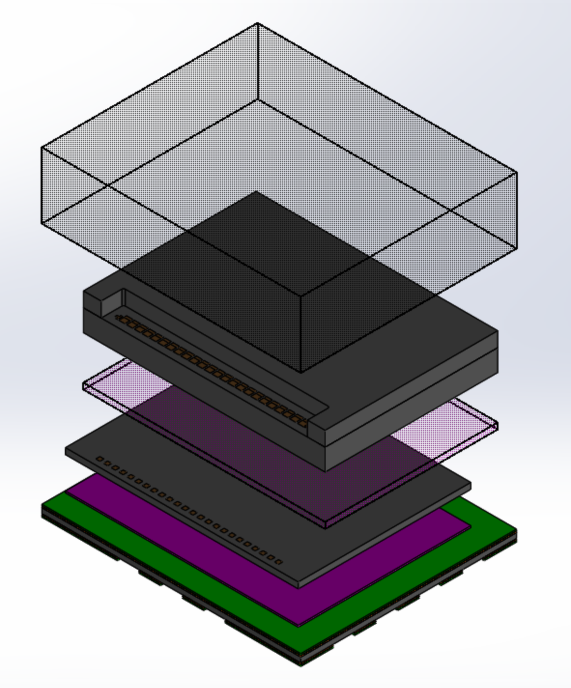
# ISPU: winning the challenges

## DSP for real-time processing and Artificial Intelligence

**Small area**
down to 8 kgates

**Standard package**
3 x 2.5 x 0.83 mm

**RAM based**
40 kiloBytes (program + execution)



**Full precision**
Floating Point Unit

**Binary Neural Network**
convolution acceleration

**Optimization**
Power consumption vs. performance

# Optimization: power consumption vs performance

**5x less current consumption for sensor fusion on ISPU than on M0**
**Below 600 µA for sensor fusion in the edge**

| | GP MCU based on Cortex-M0 | | ISPU | |
|---|---|---|---|---|
| 6x sensor fusion | **1300 µA** @ 4 MHz | + Sensor* 370 µA | **200 µA** @ 5 MHz | + Sensor* 370 µA |
| | **1000 µA** @ 8 MHz | + Sensor* 370 µA | **200 µA** @ 10 MHz | + Sensor* 370 µA |
| Run mode [µA] | **1600 µA** @ 4 MHz | | **600 µA** @ 5 MHz | |
| | **2600 µA** @ 8 MHz | | **1200 µA** @ 10 MHz | |

*Accelerometer + Gyroscope low-power mode @ ODR 104 Hz

life.augmented

# Binary Neural Network (BNN) on ISPU

## BNN on ISPU delivers over 10x better performance than floating point

| Microbenchmark* [single dense layer 128x64] | Floating point | Full BNN | BNN Improvement |
|---|---|---|---|
| **Size**** | **106,324** Bytes | **10,372** Bytes | **10.3x** |
| **Execution time** | **74,210** Cycles | **4,366** Cycles | **17.0x** |

* Kernel size = 128 and number of kernels = 64. ISPU set at 5MHz frequency
** Full Application size: data + code + internal buffers + system libs

life.augmented

# Hybrid Binary Neural Network (BNN) on ISPU

**ISPU makes solutions ready for Onlife with faster and smaller algos**

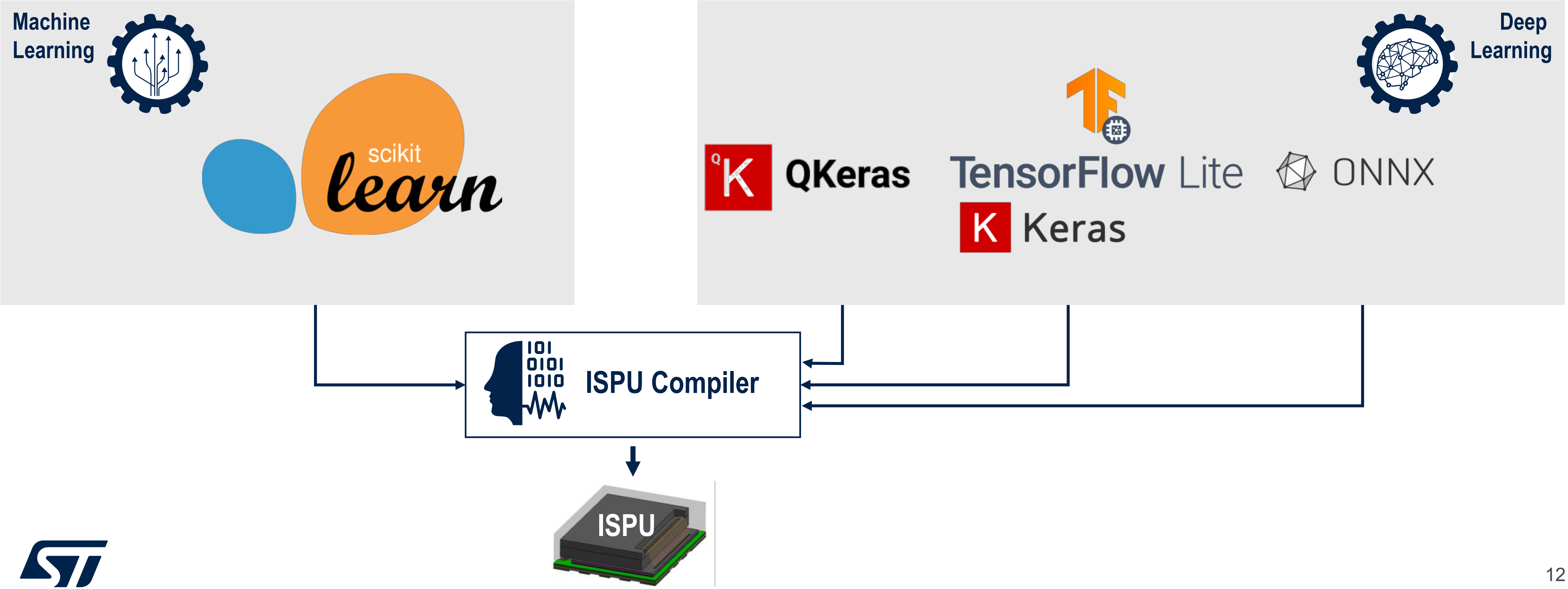
| Fan blade condition monitoring algorithm | Floating point | Hybrid* BNN QKeras | Hybrid* BNN Improvement |
|---|---|---|---|
| **Size**** | **107,200** Bytes | **11,404** Bytes | **9.4x** |
| **Execution time** | **246,170** Cycles | **194,380** Cycles | **1.3x** |

* Some layers are floating point activations with binary weights, some are fully binarized (weights and activations). ISPU set at 5MHz frequency

** Full Application size: data + code + internal buffers + system libs

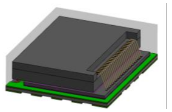
# In-sensor Machine Learning & Deep Learning

## ST ISPU delivers more options and greater freedom



# ISPU in the Edge AI

**Very constrained silicon area for logic and RAM**
**No Flash memory**



**Ultra-low power consumption (µW envelope)**

**Easily programmable with AI commercial models**

NANOEDGE AI

**Interoperates with**

Keras
QKeras
TensorFlow Lite
ONNX
learn

**?**

**And?**

# The B.E.T. benchmark

## Bytes

Amount of data transferred from sensor to cloud

## Energy

Total system power consumption

## Time

From event to reaction: make local analysis cuts reaction time

life.augmented

# An example: the robotic arm handling

| Parameters | Offline – Benefits | Online – Benefits | Onlife – Benefits |
|---|---|---|---|
| **Byte saving** (transferred from sensor to cloud) | No data transfer | | No data to be stored or transferred |
| **Energy saving** (total consumption) | | Wafer waste reduced but data stored and processed on cloud | AI/ML processing in the edge |
| **Time saving** (from event to reaction) | | Time to reaction reduced but still too long | Machine Learning approach for failure detection / prediction |
| **OUTCOME** | 1 lot (25 wafers) wasted + machine calibration time | 1 or 2 wafers wasted + machine calibration time | No wafer wasted |

life.augmented

# “In” the Edge: towards a new ecosystem

## What’s need to be explored together for ISPU?

- Ecosystem revision for tools, algorithms, and quantization procedures in sensors assets
- Development of new benchmarks, and design tools to serve this innovation
- Raise productivity and achieve synergies within the embedded developer community

life.augmented

# Takeaways

ISPU is real: global launch in 2022

ISPU is sustainable

ISPU empowers 10M+ C language developers in using AI in the Edge

ISPU makes Onlife possible

life.augmented

Our technology starts with You

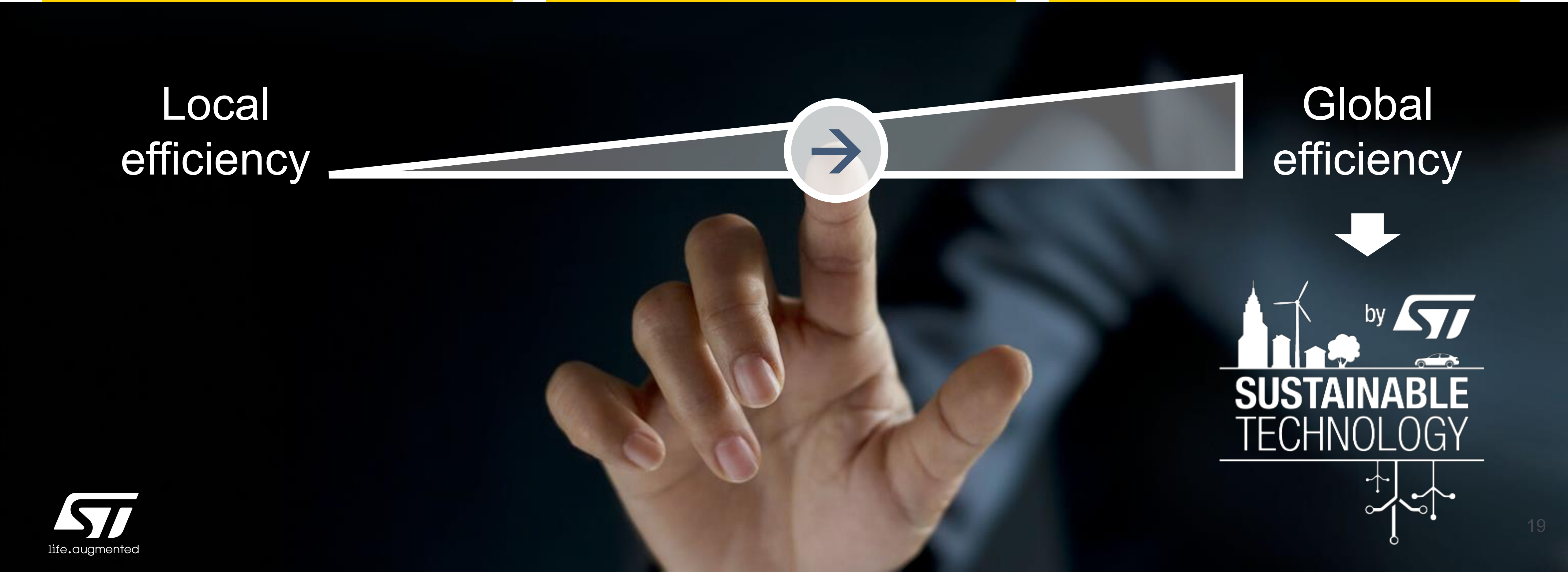
Find out more at www.st.com



life.augmented

# Systems where sensors live

| Offline era | Online era | Onlife era |
|---|---|---|
| Fragmented | Connected | Trained |




life.augmented

# From “on the Edge” to “in the Edge”

## Sensor + MCU

**Microcontroller**

Sensor — MCU

MCU standalone or hosted in the sensor package

**Standard**
MCU runs the algorithms

**Runs any kind of SW**
provided it fits the MCU specs

## rPU

**reconfigurable Processing Unit**

rPU+ Sensor — MCU

rPU integrated in the sensor ASIC

**Optimized**
reconfigured through register setting

**Constrained**
runs same model/mapping

## ISPU

**Intelligent Sensor Processing Unit**

ISPU + Sensor — MCU

ISPU integrated in the sensor ASIC

**Programmable**
dedicated instruction set

**Runs several AI algorithms**
Full precision to 1-bit NN

# One solution cannot fit all, but ISPU comes close

| Parameters | Sensor + MCU<br>(Sensor — MCU) | Benefits | rPU<br>(rPU+ Sensor — MCU) | Benefits | ISPU<br>(ISPU + Sensor — MCU) | Benefits |
|---|---|---|---|---|---|---|
| **Flexibility** | ■■■■■ | ad-hoc coding | ■□□□□ | | ■■■■□ | Optimized for inertial data |
| **Low Power** | ■■□□□ | | ■■■■■ | rPU adds few uA | ■■■■□ | Integrated computing cell, MCU in standby with sensor wakeup |
| **Data transfer optimization Sensor - MCU** | ■□□□□ | | ■■■□□ | rPU sends pre-processed data | ■■■■■ | Intelligent local processing |

# The 3 design criteria for working in the Edge

| Local | Fast | Intelligent |
|---|---|---|
| In the Edge: data privacy, low power… | Ad-hoc processor customization for real-time execution | Runs complex AI analyses and takes actions |

life.augmented

# Sensor fusion on ISPU consumes far less current

**5x less current consumption for sensor fusion on ISPU than on M0**
**Below 600 µA for sensor fusion in the edge**

| | ISPU | GP MCU<br>based on Cortex-M0 |
|---|---|---|
| **6x sensor fusion** | **200 µA** @ 5 MHz + Sensor* 370 µA | **1300 µA** @ 4 MHz + Sensor* 370 µA |
| | **200 µA** @ 10 MHz + Sensor* 370 µA | **1000 µA** @ 8 MHz + Sensor* 370 µA |
| **Run mode [µA]** | **600 µA** @ 5 MHz | **1600 µA** @ 4 MHz |
| | **1200 µA** @ 10 MHz | **2600 µA** @ 8 MHz |

*Accelerometer + Gyroscope low-power mode @ ODR 104 Hz

life.augmented

# Running Hybrid Binary Neural Network (BNN) for condition monitoring

## ISPU makes solutions ready for Onlife with faster and smaller algos

| Fan coil condition monitoring algorithm | Hybrid* BNN | Floating point | Hybrid* BNN Improvement |
|---|---|---|---|
| **Size**** | **11,404** Bytes | **107,200** Bytes | **9.4x** |
| **Execution time** | **194,380** Cycles | **246,170** Cycles | **1.3x** |

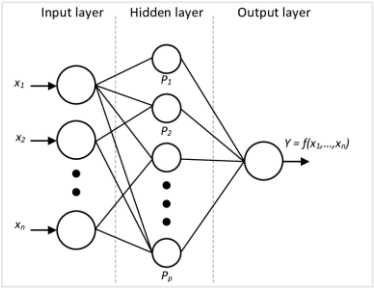
* Some layers are floating point activations with binary weights, some are fully binarized (weights and activations). ISPU set at 5MHz frequency

** Full Application size: data + code + internal buffers + system libs

life.augmented

# BNN on ISPU far outperforms floating point

**BNN on ISPU delivers over 10x better performance than floating point**
**ISPU can now run dense SW layers in the Edge**



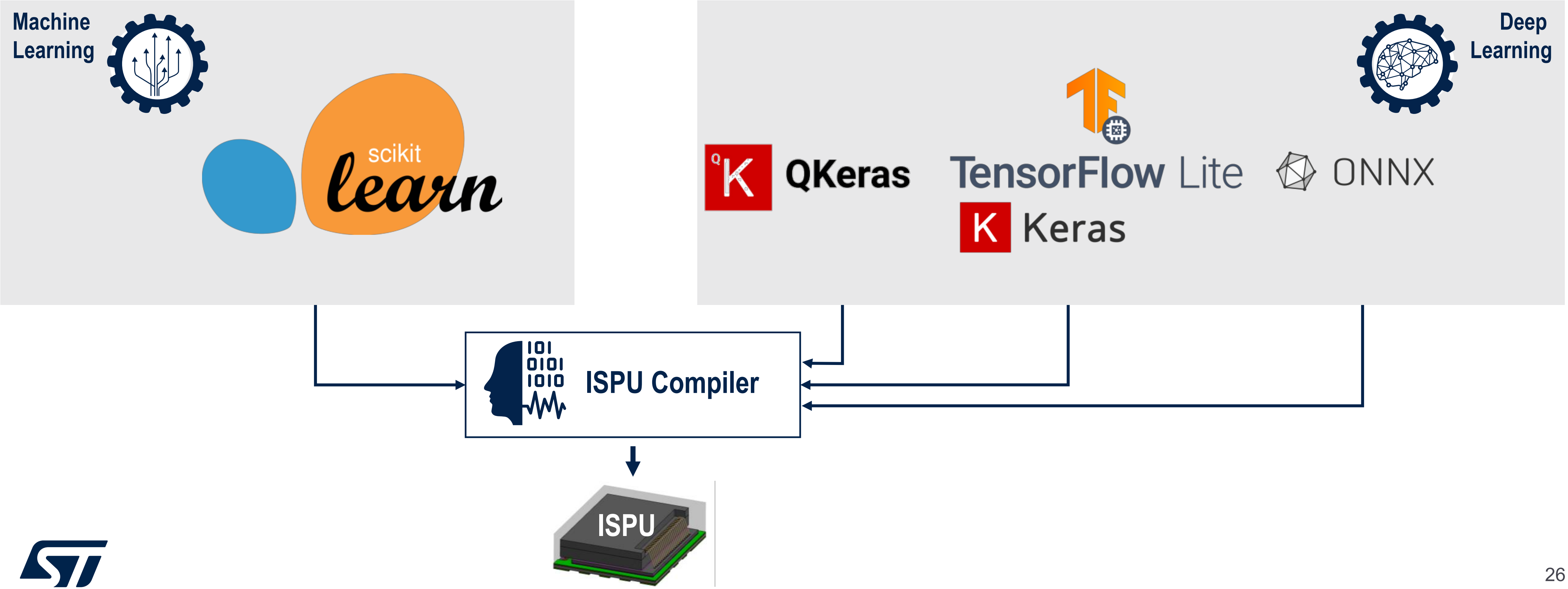

| Microbenchmark* [single dense layer 128x64] | Full BNN | Floating point | BNN Improvement |
|---|---|---|---|
| **Size**** | **10,372** Bytes | **106,324** Bytes | **10.3x** |
| **Execution time** | **4,366** Cycles | **74,210** Cycles | **17.0x** |

* Kernel size = 128 and number of kernels = 64. ISPU set at 5MHz frequency
** Full Application size: data + code + internal buffers + system libs

life.augmented

# In-sensor Machine Learning & Deep Learning

## ST ISPU delivers more options and greater freedom




life.augmented

# ISPU - Toolkit

### Compilation Tool

- Compiler (GNU) / Assembler (GNU) / Linker (GNU)
- Neural network library generation from high level tools (Keras, Tensorflow, etc.)
- Ad hoc optimization for ISPU target

### IDE Tools

- Source-level debugger (GNU) / On-chip debugger
- Simulator (STMicroelectronics)
- Eclipse graphical interface

### Runtime

- Platform SDK / Peripheral drivers
- Platform libraries

# ISPU with NanoEdge™ AI for self-learning solutions

by NANOEDGE AI

**Onlife-ready:**
classify data patterns and detect in the edge

**Commercial libraries**
ready to deploy on ISPU

**Reference design**
with customization and support

## Industrial IoT

## Personal Electronics


# Sensor's semiconductors challenges

**AI @ Edge**

- **ADVANCED TOOLS FOR INCREASED PRODUCTIVITY**
- **ECOSYSTEM COMPATIBILITY WITH AI TOOLS**
- **PROVEN TECHNOLOGY ARCHITECTURE IN SUPER TINY PACKAGE**
- **LIMITED LOGIC AND MEMORY STORAGE FOR EDGE AI**
- **OPTIMAL POWER (uW) PER PERFORMANCE CAPABILITIES**

TINY ML SUMMIT

# tinyML Summit 2022 Sponsors

ANALOG DEVICES AHEAD OF WHAT'S POSSIBLE™
AONdevices
arm
ASPINITY
brainchip The Neuromorphic Computing Company
CEVA®

Deeplite
EDGE IMPULSE
emza visual sense
FotaHub
GREENWAVES TECHNOLOGIES
Grovety Inc.

Himax
HOTG
imagimob
infineon
itemis
KLIKA TECH GLOBAL IOT SOLUTIONS

LatentAI
LATTICE SEMICONDUCTOR
Micro.ai
OmniML
NXP
PXI
Plumerai

PROPHESEE
Qeexo
Qualcomm
Rackner
Reality AI® Engineering Solutions for the Edge
REEXEN technology

RENESAS
SAP
seeed studio The IoT Hardware Enabler
SensiML™
Sony Semiconductor Solutions Corporation
ST life.augmented
SA STREAM ANALYZE

synaptics®
SynSense
SYNTIANT
Tensil.ai
TensorFlow
XMOS

# Copyright Notice