tinyML Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org
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?
The MEMS journey

**Offline era**
- A paradigm change in the man-machine interface
- MEMS technology: from a concept to a product

**Online era**
- Sensor's proliferation and connections to Cloud
- Performance improvement and technology fusion

**Onlife era**
- The fusion of technology and life
- Standalone devices able to sense, process and take action
Systems where sensors live: the evolution

- **Offline era**
  - Fragmented
  - The simplest configuration: independent systems

- **Online era**
  - Connected
  - Intertwined nodes enable efficient data exchange

- **Onlife era**
  - Trained
  - Edge AI local decision making with maximum privacy
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**
- 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
- Optimized: reconfigured through register setting
- Constrained: runs same model/mapping

**ISPU**
- Intelligent Sensor Processing Unit
- ISPU + Sensor + MCU
- Programmable: dedicated instruction set
- Runs several AI algorithms
  - Full precision to 1-bit NN
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

<table>
<thead>
<tr>
<th>GP MCU based on Cortex-M0</th>
<th>ISPU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>6x sensor fusion</strong></td>
<td></td>
</tr>
<tr>
<td>1300 μA @ 4 MHz</td>
<td>200 μA @ 5 MHz</td>
</tr>
<tr>
<td>1000 μA @ 8 MHz</td>
<td>200 μA @ 10 MHz</td>
</tr>
<tr>
<td><strong>Run mode [μA]</strong></td>
<td></td>
</tr>
<tr>
<td>1600 μA @ 4 MHz</td>
<td>600 μA @ 5 MHz</td>
</tr>
<tr>
<td>2600 μA @ 8 MHz</td>
<td>1200 μA @ 10 MHz</td>
</tr>
</tbody>
</table>

*Accelerometer + Gyroscope low-power mode @ ODR 104 Hz
### Binary Neural Network (BNN) on ISPU

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

<table>
<thead>
<tr>
<th>Microbenchmark*</th>
<th>Floating point</th>
<th>Full BNN</th>
<th>BNN Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>[single dense layer 128x64]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size**</th>
<th>106,324 Bytes</th>
<th>10,372 Bytes</th>
<th>10.3x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Execution time</td>
<td>74,210 Cycles</td>
<td>4,366 Cycles</td>
<td>17.0x</td>
</tr>
</tbody>
</table>

---

* Kernel size = 128 and number of kernels = 64. ISPU set at 5MHz frequency
** Full Application size: data + code + internal buffers + system libs
## Hybrid Binary Neural Network (BNN) on ISPU

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

<table>
<thead>
<tr>
<th>Fan blade condition monitoring algorithm</th>
<th>Floating point</th>
<th>Hybrid* BNN</th>
<th>Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>107,200 Bytes</td>
<td>11,404 Bytes</td>
<td>9.4x</td>
</tr>
<tr>
<td><strong>Execution time</strong></td>
<td>246,170 Cycles</td>
<td>194,380 Cycles</td>
<td>1.3x</td>
</tr>
</tbody>
</table>

* 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
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
- 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
An example: the robotic arm handling

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Offline</th>
<th>Online</th>
<th>Onlife</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Byte saving</strong>&lt;br&gt;(transferred from sensor to cloud)</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
</tr>
<tr>
<td>Benefits</td>
<td>No data transfer</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
</tr>
<tr>
<td><strong>Energy saving</strong>&lt;br&gt;(total consumption)</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
</tr>
<tr>
<td>Benefits</td>
<td>![Gray square]</td>
<td>Wafer waste reduced but data stored and processed on cloud</td>
<td>![Gray square]</td>
</tr>
<tr>
<td><strong>Time saving</strong>&lt;br&gt;(from event to reaction)</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
<td>![Gray square]</td>
</tr>
<tr>
<td>Benefits</td>
<td>![Gray square]</td>
<td>Time to reaction reduced but still too long</td>
<td>![Gray square]</td>
</tr>
<tr>
<td><strong>OUTCOME</strong></td>
<td>1 lot (25 wafers) wasted + machine calibration time</td>
<td>1 or 2 wafers wasted + machine calibration time</td>
<td>No wafer wasted</td>
</tr>
</tbody>
</table>
“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
ISPU makes Onlife possible
ISPU is sustainable
ISPU empowers 10M+ C language developers in using AI in the Edge
ISPU is real: global launch in 2022
Our technology starts with You

Find out more at www.st.com
### Systems where sensors live

<table>
<thead>
<tr>
<th>Era</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offline era</td>
<td>Fragmented</td>
</tr>
<tr>
<td>Online era</td>
<td>Connected</td>
</tr>
<tr>
<td>Onlife era</td>
<td>Trained</td>
</tr>
</tbody>
</table>

- **Local efficiency**
- **Global efficiency**

---

*by STI - SUSTAINABLE TECHNOLOGY*
From “on the Edge” to “in the Edge”

Sensor + MCU
- Microcontroller
- 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
- Optimized: reconfigured through register setting
- Constrained: runs same model/mapping

ISPU
- Intelligent Sensor Processing Unit
- ISPU + Sensor
- MCU
- Programmable: dedicated instruction set
- Runs several AI algorithms: Full precision to 1-bit NN
One solution cannot fit all, but ISPU comes close

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Benefits</th>
<th>Benefits</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexibility</td>
<td>ad-hoc coding</td>
<td>Optimized for inertial data</td>
<td>Integrated computing cell, MCU in standby with sensor wakeup</td>
</tr>
<tr>
<td>Low Power</td>
<td>rPU adds few uA</td>
<td>rPU sends pre-processed data</td>
<td>Intelligent local processing</td>
</tr>
<tr>
<td>Data transfer optimization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensor - MCU</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The 3 design criteria for working in the Edge

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

<table>
<thead>
<tr>
<th>Mode</th>
<th>ISPU</th>
<th>GP MCU based on Cortex-M0</th>
</tr>
</thead>
<tbody>
<tr>
<td>6x sensor fusion</td>
<td>200 µA @ 5 MHz + 370 µA Sensor</td>
<td>1300 µA @ 4 MHz + 370 µA Sensor</td>
</tr>
<tr>
<td></td>
<td>200 µA @ 10 MHz + 370 µA Sensor</td>
<td>1000 µA @ 8 MHz + 370 µA Sensor</td>
</tr>
<tr>
<td>Run mode [µA]</td>
<td>600 µA @ 5 MHz</td>
<td>1600 µA @ 4 MHz</td>
</tr>
<tr>
<td></td>
<td>1200 µA @ 10 MHz</td>
<td>2600 µA @ 8 MHz</td>
</tr>
</tbody>
</table>

*Accelerometer + Gyroscope low-power mode @ ODR 104 Hz
Running Hybrid Binary Neural Network (BNN) for condition monitoring

ISPU makes solutions ready for Onlife with faster and smaller algos

<table>
<thead>
<tr>
<th>Fan coil condition monitoring algorithm</th>
<th>Hybrid* BNN</th>
<th>Floating point</th>
<th>Hybrid* BNN Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>11,404 Bytes</td>
<td>107,200 Bytes</td>
<td>9.4x</td>
</tr>
<tr>
<td>Execution time</td>
<td>194,380 Cycles</td>
<td>246,170 Cycles</td>
<td>1.3x</td>
</tr>
</tbody>
</table>

* 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
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]

<table>
<thead>
<tr>
<th></th>
<th>Full BNN</th>
<th>Floating point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size**</td>
<td>10,372 Bytes</td>
<td>106,324 Bytes</td>
</tr>
<tr>
<td>Execution time</td>
<td>4,366 Cycles</td>
<td>74,210 Cycles</td>
</tr>
</tbody>
</table>

** BNN Improvement

- 10.3x
- 17.0x

* Kernel size = 128 and number of kernels = 64. 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

Machine Learning

Deep Learning

ISPU Compiler

ISPU
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

- 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 (μW) per performance capabilities
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

[www.tinyML.org](http://www.tinyML.org)