

1.2 Intelligent Sensing: Enabling the Next “Automation Age”

Marco Cassis

President, Analog, MEMS and Sensors Group
Head of STMicroelectronics' Strategy, System Research and Applications,
Innovation Office, Geneva, Switzerland/Tokyo, Japan

Introduction

Sensors and actuators, as well as their associated processing and connectivity functions, are now pervasive across a wide variety of application areas: industrial, medical and healthcare, transportation, consumer and personal devices, urban management, and robotics to name just a few (Figure 1.2.1). The use of sensors and actuators is supporting the evolution of each of these areas thus enabling what can be called the next “Automation Age”:

- The industrial world is moving towards smarter, more efficient and cleaner solutions with initiatives such as industry 4.0/5.0. This means moving from preventive to predictive maintenance and increasing use of robots and cobots to improve efficiency, while preserving humans from repetitive tasks.
- Medical and healthcare is progressing to more personalized, always-on monitoring and the possibility to access medical services from the comfort of home.
- Transportation is evolving to higher levels of driver assistance and autonomous driving, vehicle electrification, predictive maintenance, user customization and increased safety, alongside an evolution of smart bikes, scooters, trains, and airplanes.
- Urban management is moving towards smart cities, using networks of millions of intelligent sensors and IoT nodes to improve monitoring, manage resources, assist citizens, and improve logistics with self-driven drones and vehicles.
- Consumer and personal electronics are evolving to offer a more convenient and personalized experience. Devices such as smartphones, watches, glasses and other wearables interact seamlessly with interconnected objects in the smart home, anticipating our needs and automating routine tasks with home robots.
- Robotics, including drones, are becoming more widespread thanks to technologies developed for automatic navigation based on multiple sensors, detailed environment analysis, object recognition/detection, and tracking.

1.0 Cyber Physical Systems Evolution to Support the Next Automation Age

In all of the applications above, the systems involved are commonly referred to as Cyber Physical Systems (CPS). The general idea of a CPS is to collect sensor data (quantitatively measuring properties of the physical world) from a variety of sources [1], and convert them to information in the cybernetic world, which can process and interpret them. This allows notification of the results to the physical systems, so that decision can be taken, and control commands can be routed back to the physical world for action and/or reconfiguration of the system parameters.

1.1 The CPS consist of 3 domains: The Physical, Object, and Cyber Domains

The Physical Domain is the world we live and act in every day.

The Object Domain interfaces the Physical Domain through nodes consisting of:

- Sensors to perceive the physical domain and acquire meaningful, quantitative information;
- Processors to perform computations on acquired data and make analysis;
- Communication to route the computation results to different information channels;
- Actuators to act on the Physical Domain.

The information computed by the Object Domain is typically transmitted via complex network topologies to the Cyber Domain.

The Cyber Domain (the Cloud) is characterized by virtually unlimited computational and memory resources, so it seems the natural place to deploy intelligent algorithms. However, leaving all intelligence and decision-making in the Cyber Domain would not tap into the processing capabilities in the Object Domain. This is the reason why a CPS

using distributed intelligence exploits system resources optimally (Figure 1.2.2), while contributing to sustainability.

In fact, data center energy consumption will increase from 200-300TWh in 2020 up to around 400TWh lower bound/1000TWh higher bound in 2030 [2]. Having more intelligent objects with local data processing will help minimizing overall CPS energy consumption, thanks to devices exploiting Ultra-Low Power (ULP) architectures and technologies.

It is clear that partial re-deployment of intelligence into the computing assets of sense and act nodes presents many challenges due to constrained computational, memory, and energy resources. For example, a typical microcontroller (MCU) has an operating frequency that can range from 10s to 100s of MHz, memory from a few Kbytes to 1Mbytes of RAM, and 100s Kbytes to 2 Mbytes of flash. These constraints not only call for leaner edge Artificial Intelligence (AI) algorithms, but is also driving the exploration of new hardware acceleration solutions such as neuromorphic and in-memory computing [3], promising improvements from one to two orders-of-magnitude in power and computational efficiency.

To address the challenge of bringing AI on constrained devices at the edge, the TinyML community started in 2018 and now includes several hundreds of companies and universities across the world [4]. Together they are developing Machine Learning (ML) algorithms, architectures, techniques, tools, and system approaches enabling on-device analytics for a variety of sensors (vision, audio, motion, environment, chemical, and so on) at less than mW power range, that is for always-on use and battery-operated devices. Solutions can then be benchmarked thanks to the MLCommons initiative [5].

Although distributed intelligence CPS will shift part of the processing to the object domain, the Cyber Domain will retain an important and diverse role thanks also to the implementation of Federated Learning [6] solutions. In fact, the Cyber Domain can supervise a wide range of sense and act nodes, and can prepare, train, distribute, and aggregate locally trained ML models while coordinating their run-time actions [7].

As our societal dependence on Cyber Physical Systems increases, so too are the security stakes. It is essential to guarantee the integrity and confidentiality of the gathered data, its processing as well as the resulting actuation functions. It is also important to protect the privacy of users. CPS are much more exposed than conventional systems through a larger “attack surface” at the system’s software and interfaces. Attackers can be remote and operate from anywhere in the world, driven by diverse motivations. It is therefore critical for the industry to adopt state-of-the-art cyber-security techniques and endorse thorough security evaluation and certification processes to achieve a proper level of security assurance. Artificial Intelligence is also introducing new threats such as “adversarial examples” class of attacks allowing the attacker to influence network outputs through network input manipulations [8]. For example, [9] demonstrated the feasibility of tricking a road sign recognition system into misclassifying a STOP sign as a 45 MPH speed limit by placing small black and white stickers on the STOP sign (Figure 1.2.3).

2.0 Adding Intelligence at the Edge

The evolution of all the elements present in sense and act nodes must consider the exploitation of embedded AI (e-AI), one of the new technologies included in the Gartner Hype Cycle [10] (Figure 1.2.4), due to its potential to increase the accuracy, insights and intelligence gained from current and next-generation sensors. The figure indicates that e-AI technology will reach the plateau of productivity in one to four years in various devices targeting different applications and segments.

The next sections will introduce the main elements of innovation in sensors, processing, actuation, and communications that will be seen in the coming years, enabling the next Automation Age.

3.0 Intelligent Sensors

3.1 Optical Sensors for IoT

Photonics sensors will play a key role in the evolution of intelligent sensing providing the “eyes” of Cyber Physical Systems. While the smartphone market continues to drive image sensor pixel scaling, enabling high resolution, and high dynamic range scene capture, other key developments such as dynamic vision sensing, 3D depth mapping, and low-resolution environment sensing (multi-spectral imaging, proximity/presence detection, and so on) will add a wealth of data.

Significant progress has been made in the realization of Dynamic Vision Sensors (DVS). Compared to traditional image sensors the data rates of DVS devices make them well suited as “edge” sensors and they have the potential for lower power consumption.

Image sensors benefit from 3D integration processes using dense digital nodes on the bottom tiers, enabling integration of complex signal processing logic, massive parallel readout structures, embedded DRAM frame buffers [11], and high-speed interfaces. 3D integration is especially important for on-chip image analysis, enabling fully autonomous detection devices through ultra-low power artificial intelligence solutions. This approach:

- Supports power consumption minimization thanks to close coupling of operations between the imager focal plane and the bottom-tier processing resources [12].
- Offers a higher level of privacy since only metadata rather than the full image is shared outside the device.

Two-layer stacked image sensors are high volume products manufactured by several Integrated Device Manufacturers (IDMs) and foundries. This is done using bonding techniques enabling the stacking of 2 wafers (Figure 1.2.5). The top wafer is dedicated to a CMOS Image Sensor (CIS) with a full pixel matrix. The bottom wafer is used for Image Signal Processing (ISP) using a CMOS technology dense enough to match the optical format selected for the CIS resulting in a top and bottom die of the same area. This system integration approach results in product area saving with respect to a 2D conventional approach.

3D integration presents a number of challenges as products have to be fully redefined taking into account the 3rd dimension. This impacts: architecture partitioning, design, physical implementation, test strategy, package and thermal management. CAD tools must be developed to enable the simulation and optimization of such design architectures. Technology-wise the bonding connection density increase can be achieved by connection pitch decrease down to 1 μ m or below. Bonding of 3 wafers (or more) can be achieved by mastering both bonding and Through-Silicon-Via (TSV) processes with aggressive dimensions (critical dimensions and pitch) or relying on sequential layering technology. Product test strategy optimization is also a challenge. The final product can be tested only after full product process completion with the risk of yield loss due to issues on any one of the wafers within the stack. The smaller footprint of the stacked product helps balance this risk but does not fully eradicate it.

Enlarging the trend analysis to other optical sensors, it is evident that in recent years Single Photon Avalanche Diodes (SPADs) and fast photodiodes have moved from R&D and academic studies to mainstream products [13], [14], [15]. Such sensors can be part of complex systems aimed at producing 3D depth map images for applications such as LiDAR, AR/VR and Simultaneous Localization And Mapping (SLAM) or simple single or low-resolution range sensors. These active depth sensing solutions will be complemented by other techniques such as structured light and stereoscopic imaging.

The data produced by these arrays of sensors is ideally suited to local processing, enabling features such as object/person recognition and tracking, object speed and acceleration, gesture recognition, and extraction of key data from the scene. For all these devices, the access to high density nodes through 3D integration enables the development of devices in line with stringent requirements, such as those for automotive safety, and ensures the security of the device as well, in terms of privacy, authentication and data integrity.

Scene/object wavelength analysis is another modality enabled by optical sensors. The latest ambient light sensors (ALS) feature multi-wavelength reporting and this can be extended further to multi- and hyper-spectral sensing, enabling spectrometer like functionality in the visible/NIR/SWIR range [16]. Multi-wavelength ALS provide data for ideal display configuration as well as light and colour information for optimal image sensor set-up. Multi-spectral sensors enable a wide range of applications including gas and chemical sensing (allowing features such as food freshness analysis), cosmetics, medical, automotive, and agricultural imaging. The wavelength response range of such devices has been shown to be enhanced by above- IC technologies such as quantum dots [17]. The addition of AI to such devices makes possible the extraction of key scene spectral signatures, which are learnt and acted upon. An example of this can be found in [18] where food products are imaged and their spectral signature analysed for chemical composition.

3.2 Motion MEMS sensors

In the year 2000 a paradigm shift in the man-machine interface happened with MEMS technology finding applications in automotive (accelerometer in car airbags), in personal electronic devices (smartphone orientation), and gaming (handheld game controller). Since then, MEMS have made considerable progress: in the past 10 years, we have seen 91% power consumption reduction, 70% size reduction, and over 70% MEMS noise reduction (Figure 1.2.6).

The major technology drivers for MEMS sensors today are Augmented Reality, 5G, connectivity technology evolution, and the proliferation of advanced AI-enabled use cases. For all of these, increasing data processing capability complemented by AI within

the sensor is key (Figure 1.2.7). In this respect, as mentioned before, the TinyML community targets also embedded AI into very low processing power devices such as sensor systems, leading to μ W levels of power consumption, ultra-low latencies, minimized silicon area, greater accuracy, and enhanced privacy. Embedded AI research on inertial sensing currently addresses sensor design and selection, calibration, and error modeling, navigation, and motion-sensing algorithms, multi-sensor information fusion, system evaluation, and practical application [19].

A clear challenge for sensor makers is to add AI processing capabilities without disrupting their proven solution for low noise, low power, and high accuracy, and also without changing the package form factor. This last point would offer system designers flexibility on sensor selection without changing board design, as well as offering advantages for manufacturing.

Different solutions are being proposed to meet these challenges.

Sensor packaging approaches face the duality where the possibility to integrate on same SiP (System in Package) different sensing technology, data processing or transmission capability go in contrast with ease of testing and calibration (Figure 1.2.8). The possibility to build fully functional sensors in tiny Chip Scale Packages (CSP) [20], [21], that can be integrated as bricks in a bigger SiP, will be a game changer in the coming years (Figure 1.2.9).

3.3 Audio Sensors

MEMS microphones can be considered as smaller versions of conventional microphones. These products are characterized by their signal type (digital as PDM, I2S, and analog), and in applications by the band of interest (voice, audio, ultrasound). Audio sensors are not only microphones, since sound can also be captured through the vibrations propagated in a solid material. Three significant microphone evolutions are emerging with different technology requirements.

First, microphones with wake-on-sound function, where the sensor uses sound energy to wake up a system from full power-down. The key role of the sensor technology is to determine the speed of wake up. For example, piezoelectric technology allows the microphone in power down state (current consumption less than 10uA) to detect a sound above a certain predefined acoustic threshold; this threshold crossing generates an interrupt event triggering the application processor to wake-up the system.

Second, microphones for vibration detection. For environmental monitoring wide bandwidth is key, while in wearable applications the integration of audio sound with other sensors is the focus. This integration is essential for True Wireless Stereo (TWS), a unique Bluetooth feature that enables true stereo sound quality without the use of wires. Here, a specific wide bandwidth sensor captures the vibration of the voice that propagates through the jawbone. This bone detection generates a signal improving the audio quality of the microphones. Today's discrete solutions are based on the combination of inertial sensors (for user interface or 3D audio), wide-bandwidth sensors and MEMS microphones. Next-generation designs target performance improvement, size, and power reduction delivered in new intelligent and highly integrated sensors. Ultimately, a single sensor could provide head tracking (for 3D audio), bone conduction, user interface, and in-ear detection (today performed with a dedicated infra-red sensor) (Figure 1.2.10).

Third, the integration of arrays of audio sensors in size-optimized and cost-optimized embedded systems enables advanced speech and acoustic processing features, such as always-on listening for specific keywords in noisy environments. This brings challenges for system design, particularly in terms of processing speed, compact packaging, and low power consumption. Voice processing and context awareness capabilities contribute to an enhanced and more personalized user experience. Edge processing combined with cloud connectivity brings flexibility to system design and enables optimal balance between offline and online capabilities. Consider for example the in-ear translators and in-ear-piece translator devices [22], which require a perfect fusion of these technologies, as an example of a system where improvements need to be made in the coming years.

3.4 Environmental Sensors

Environmental Sensors are used to sample scalar quantities such as atmospheric pressure, temperature, humidity, and gas concentrations. These measurements are used to monitor ambient conditions in a wide range of applications. New ways of exploiting the sensor data are being deployed; for example, pure pressure measurements have been used to extrapolate indirect measures such as altitude position, individual movements, and other contextual metrics. New rules, standards or legislation can drive evolution, for example, the E911 emergency call feature in the US [23] requires precise altitude and horizontal positioning information (Figure 1.2.11). These additional uses of data enhance the utility of environmental sensors in smart devices, driving their evolution to more power efficiency, higher pressure accuracy, and smaller packaging. Artificial Intelligence also contributes to increasing functionality, especially by way of intelligent aggregation of data from multiple sensors to enable real-time feedback.

4.0 Intelligent Processing

4.1 AI-Augmented Edge Devices

As previously introduced, AI application in distributed systems varies by use case in terms of processing latency, memory footprint, and power consumption while relying on different hardware and software partitioning.

Machine Learning solutions typically used on the Cloud or in high-performance General-Purpose Graphics Processing Units (GP-GPUs) cannot fit on constrained MCUs or Smart Sensors, so ML algorithms have typically to be redesigned and retrained to suit the final device capabilities, widening the project duration. For this reason, the possible exploitation of Automated Machine Learning (AutoML) and Neural Architecture Search (NAS) [24] tools is important. This could greatly simplify the effort required to define an ML algorithm and choose its hyperparameters to trade off performance, cost, and accuracy on a given dataset, thus speeding up development time by finding accurate models. Nevertheless, AutoML and NAS frameworks typically are not automatically deployable on embedded processors for edge AI. For this reason, some NAS tools evolved to limit the research space taking into account the constraints related to the final target [25], [26], [27], [28], [29]. This helps when devising optimized solutions of Neural Networks for constrained devices, thus shortening application development time. Once the ML model is available, its deployment on embedded processors requires code generators and compiler tools that can interoperate with the most widely used AI frameworks (such as TensorFlow/Keras, Pytorch, ONNX, and so on). Most importantly these tools can also automatically optimize the AI solutions for the target platform and facilitate user work when moving from a pure software solution on an MCU to a hardware accelerated solution on a Neural Network Processing Unit (NPU), depending on the selected algorithm and system requirements.

Deployment in embedded systems can be achieved mainly through off-line tools (such as STMicroelectronics STM32Cube.AI for STM32 MCU family; e-AI Development Environment Tools for Renesas processors; NXP® eIQ™ ML software development environment) and interpreters (Google open-source project TensorFlowLiteMicro). Automatic tools to detect the best model for the target platform, as well as for optimized porting on the selected device, are of course important. In supervised learning, data scientists spend at least 70% of their time on data preparation [30], that is, to collect, label, and augment large amounts of data to create datasets suitable for Neural Network training. To decrease the impact of this activity, on-device learning on embedded processors can help, enabling the learning of Neural Network model weights on the same device that run the model inference. On-device training offers better generalization since learning happens in-situ on real-world, just-in-time data acquired by sensors after deployment.

Clearly, on-device learning poses significant challenges for resource constrained devices such as sensors and MCUs, so minimizing the computational complexity and memory requirements of this process is mandatory. Moreover, on-device learning in a well-defined time slot is highly desirable for predictability when deployed in the field. For example, [31] aims to enable on-device training of convolutional neural networks (CNNs) by reducing the computation cost at training time. Applying that to ResNet-110 on CIFAR-10 achieved 68% computation saving while preserving full accuracy and 75% computation saving with a marginal accuracy loss of 1.3%. Aggressive computation saving of 96% is achieved with less than 0.1% accuracy loss when quantization is integrated into the proposed approaches. A more specific example related to TinyML, a system called TinyOL [32], is proposed which enables incremental on-device autoencoder training on streaming data for constrained IoT devices.

4.2 The Advent of Specialized AI Hardware Acceleration

Despite the improvements of TinyML, not all AI solutions can be deployed in pure software due to reasonable processing time requirements of specific applications such as computer vision. Convolutional Neural Networks require deep topologies, millions of parameters, and varying kernel sizes, thus escalating bandwidth, power, and area requirements. Efficiency can be improved with specific hardware accelerators supporting kernel compression and an on-chip reconfigurable data transfer fabric to improve data reuse and reduce on-chip and off-chip memory traffic [33], [34].

While general-purpose compute engines (especially GPUs) have spearheaded much of the AI processing, there is an increasing push for more specialized acceleration. This has led to the design of specialized architectures and chips specifically tailored to efficiently execute AI workloads and, more recently, narrowly defined accelerators for target application domains benefiting from more specialization. The increasing demand to include AI capabilities in chips has also incited many IP vendors to create their own lines of hardware accelerators that can be integrated into custom Systems-on-Chip. The approaches here are very diverse. Improvements to existing solutions include DSPs instruction set extensions and customization to accelerate certain operations either with Single Instruction Multiple Data (SIMD) or vector instructions, as well as adding new

data-types processing capabilities to existing platforms such as GPUs. However, there are also many radically different approaches being pursued, explicitly addressing the peculiar data structures required by tensors associated with Neural Network algorithms and their associated degrees of freedom. Neural Networks inference complexity is linked to a high number of operations (as multiply, add, accumulate represent the bulk of the total operations), but the transfer of data and network parameters to/from memory is of equal if not greater relevance especially for embedded systems (Figure 1.2.12). Further advances are emerging from disruptive computing architectures, inspired by biology, in the form of neuromorphic structures. These solutions, such as spiking neurons and convolutional neural networks exploiting innovative memories technologies, have the potential to provide much lower power consumption and higher computing densities. Spiking neural network approaches are explored for their potential of very high-power efficiency, but their training and deployment is not fully mature yet. On the other end, convolutional approaches can already tap into reduced precision computation and advances in semiconductor device technology.

A new class of computing structures for neural network execution moves computation to where the data lives in the memory or close to it. In-Memory Computing (IMC), for instance, can achieve AI inference with up to 100× higher compute densities and power efficiency compared to current state-of-the-art solutions [3]. IMC can be classified as DIMC (Digital In-Memory Computing) and AIMC (Analog In-Memory Computing). In DIMC the computation happens across a group of memory bit cells, that can be accessed in parallel to perform an accumulation operation, thus deterministically generating a precise bit accurate output. Computation within the memory block boundary reduces the data movement and the global I/O power. In AIMC architectures the computation is mapped onto an analog operation such as a current summation or charge transfer, either in a column-wise or a row-wise fashion, with an analog-to-digital conversion (read out) at the end to generate the final output. In AIMC, an especially promising avenue is exploitation of the physical attributes of emerging non-volatile resistive memory devices to perform computations in the memory itself, thus obviating the need to move data from memory to a processing unit [35]. For AI application, amongst the Emerging Memories (EM) technologies, Phase-Change Memory (PCM) is a natural candidate because of its very small cell size, hence allowing a very high computational capacity (TOPS/mm²). Moreover, PCM manufacturability and reliability were proven on high density stand-alone memory products in volume production, showing that PCM is a mature technology which can lead to a much earlier introduction of an industrially proven AIMC solution. IMC-based architectures will provide the freedom to experiment with Deeply Quantized Neural Networks, finding the best trade-offs between accuracy and efficiency, aided in the process by the introduction of tools such as qKeras [36], [37] by Google, and related ecosystems.

While many AI security challenges relate to the overall Machine Learning system, some are more directly linked to hardware-level AI implementation. For example, it is important to protect the confidentiality and integrity of the embedded network configuration to prevent attacks aiming at influencing network outcomes, while protecting the valuable intellectual property that constitutes this configuration. Compromising the confidentiality of the network can be a first step towards adversarial attacks aimed at influencing network outcomes. This is an emerging research topic, for both attack and countermeasures research. It is not yet clear to what extent hardware-level attacks on AI will be as much of a concern as hardware attacks on embedded cryptography. Shifting AI processing can bring privacy benefits as user data may no longer need to be sent to the cloud.

5.0 Intelligent Actuators

Actuators consist of two main parts: a power conversion stage coupled with a motion stage, which translates electric energy into mechanical movement. Industrial robots are an example of an advanced actuator typically used to manipulate parts on an assembly line. The focus today is on reducing size, increasing efficiency, and improving flexibility. This drives three technology trends leading to "Intelligent Actuators".

The first technology trend is power efficiency and energy density. Silicon Carbide (SiC) and Gallium Nitride (GaN) wide bandgap materials allow for smaller adapters through the adoption of planar transformers. This provides a higher efficiency in reduced space in many applications fields, such as wireless power chargers, robotic actuators, and Electric Vehicle charging stations. A GaN transistor has 1/10 the size and weight of a Silicon MOSFET and a 1/40 the capacitance required to switch (Figure 1.2.14).

The second trend is power digitalization. Digital power relies on microcontrollers to run control techniques for power management [38]. This trend goes back to 1985 [39], when actuator drivers and power conversion devices were integrated on silicon with BCD (Bipolar-CMOS-DMOS) technology. In 1995 for the first time, Power MOSFET, Bipolar, CMOS components as well as Non-Volatile Memory cells were implemented on a single chip, integrating a Pulse Wave Modulation (PWM) (up to 2A) current control loop together with an 8-bit microcontroller [40]. In 2018 an internal STMicroelectronics

development featured a 32-bit microcontroller, monolithically integrated with an autonomous numerical torque generator for motor control, a dedicated power stage, and an analog front-end covering smart power conversion functions, from DC/DC conversion to motor drive. The device was created in a 12.7 mm² die in BCD technology. Future advances will rely on these evolving IC technologies as well as associated packaging solutions. These will include super-integration, wafer-to-wafer bonding, and system-in-package, to enable the heterogeneous integration of multiple technologies such as Vertical Intelligent Power, BCD, FD-SOI CMOS, and others.

Microcontrollers allow implementation of AI to model complex, non-linear time-varying systems that need to adapt continuously at run time. A recent example [41], shows that a deeply quantized Deep Echo State Network can run the control with only 2.5kB of memory and an inference time of 20μsec, enabling broad deployment in motor control solutions. Another example [42] proposes the improvement of the speed and torque dynamic responses of DC motor by using Neural Network Parameter Tuner (NNPT) for PI-controller; also an example is provided to [43] control the speed of induction motor using a Neural Network estimator and speed controller.

The third trend is safety and security, with galvanic isolation of circuitry preventing electrical shock hazard, while security features support data management.

The Industry 5.0 initiative is continuing the path of Industry 4.0, using digitalization to empower workers in a safer working environment. For example, the EU industrial strategy [44] covers the interaction of the human workforce with robotics using Artificial Intelligence and additive manufacturing. The combined efficiency of manufacturing processes and the related resources depends on the interaction between industrial robots (and the relative ISO 10218-2 Standard) and the humans around it; in fact, the current paradigm of safeguard space must be revised to introduce the new concept of collaborative workspace. A human centric collaborative robotic (cobot) [45][46][47] will quickly require a new generation of intelligent actuators to ensure the safety of workers (Figure 1.2.15). These new actuators will exploit higher efficiency and energy density, as well as new motion control algorithms to deliver increased safety while maintaining security.

6.0 Wireless Connectivity

Connectivity is defined as any secure, reliable link between devices, from Near Field Communication, such as that implemented in our credit cards, up to complex, worldwide mobile 5G networks. Data communication from sensors in the Object Domain node to Cyber Domain and then back to the node actuators is an essential function for AIoT (Artificial Intelligence of Things), IIoT (Industrial Internet of Things), and vehicular system implementations. Multiple connectivity solutions defined by standards or alliances coexist today to address different needs of the systems in term of throughput, coverage, and mobility.

The overall data traffic on the network backbone is constantly growing [47] and, while there is a general trend towards higher throughput and lower latency, with sensors and actuators becoming more and more intelligent, we envision a parallel connectivity trend characterized by lower throughput, shorter transmission time, and lower power. The mix of the two trends will allow more efficient e2e connectivity solutions. In this context, technical innovation will be focused not only on addressing low power, low latency but also secure and reliable connections that are easy to install and capable of operation in dense environments. These trends are reported by all major standardization bodies (IEEE, 3GPP, BLE-SiG) and Alliances (SigFox, LoRA). New emerging technologies include:

- Wake-up Radio (as in IEEE 802.11ba), to extend battery life and keep reduced latency [48];
- Intelligent spectrum partitioning between multiple users implementing sub channelization (as in NB-IoT and Cat-M), and spectrum sharing in frequency or spatial domain (as in Wi-Fi);
- The allocation and use of less congested spectrum portions, positioned in the higher part of the spectrum;
- The constant evolution of algorithms, protocols, and techniques to enhance link security and make system intrusion more difficult.

It is possible to envision a combined use of connectivity solutions (UWB, NFC) with other connectivity links (Wi-Fi, BLE) to enhance security and, at the same time, simplify network installation procedures. Given that increasing reliable area-coverage is in opposition to lowering power, next-generation connectivity solutions will require an extra innovation effort. The endeavor is in the definition of new SoC architectures, both RF and digital, to achieve ultra-low power dissipation in active mode (~10mA when

transmitting and receiving) and in standby mode (~10μA when there is no need to transfer data). The required effort in low power design should address not only the major power-hungry building blocks (oscillators), but also the power generation (low power DC-DC), and management techniques (clock gating, power islands), and the specific use of the selected technology (leveraging low voltage, near threshold). Moreover, low latency connectivity will play a key role also in the future generations of automotive vehicles that will increasingly become part of the network, sharing data collected through their sensors, and making remote vehicle control a reality!

Conclusions

The next Automation Age is being enabled by the proliferation of Cyber Physical Systems composed of interconnected, miniaturized objects and systems able to sense, process, and take actions. The number of such devices is growing exponentially and with increasingly complex functionality, requiring more powerful cloud infrastructure with a consequent energy penalty. For these reasons' new technology trends and their heterogeneous integration are addressing the need to distribute intelligence, such as AI processing, within the CPS. This enables further device miniaturization while bringing analytic capabilities close to or into the sensors, allowing the implementation of sophisticated actions with local smart controls, helping protect user privacy and lowering overall CPS power consumption. The on-going evolution of CPS offers significant opportunities for innovation within the semiconductor industry to enable and enhance the next Automation Age.

Acknowledgment:

I thank the many STMicroelectronics colleagues who have contributed to this paper.

References:

- [1] A. Djoud, F. Atil, Abdelhak-Djamel, and C. Beddier, "Domain model for cyber-physical systems," vol. 2326, pp. 34–41, 2018.
- [2] A. Andrae, "New perspectives on internet electricity use in 2030," *Eng. Appl. Sci. Lett.*, vol. 3, no. 2, 2020.
- [3] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, no. 7., doi: 10.1038/s41565-020-0655-z, 2020.
- [4] "tinyML Foundation." <https://www.tinymml.org/>.
- [5] "MLCommons." <https://mlcommons.org/en/>.
- [6] Q. Li *et al.*, "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," [Online]. Available: <http://arxiv.org/abs/1907.09693> Jul. 2019.
- [7] D. Cogliati, M. Falchetto, D. Pau, M. Roveri, and G. Viscardi, "Intelligent cyber-physical systems for industry 4.0," 2019, doi: 10.1109/AI4I.2018.8665681.
- [8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.
- [9] K. Eykholt *et al.*, "Robust Physical-World Attacks on Deep Learning Models," [Online]. Available: <http://arxiv.org/abs/1707.08945>.
- [10] "5 Trends Drive the Gartner Hype Cycle for Emerging Technologies" www.gartner.com, 2020.
- [11] H. Tsugawa *et al.*, "Pixel/DRAM/logic 3-layer stacked CMOS image sensor technology," doi: 10.1109/IEDM.2017.8268317, 2018.
- [12] A. Verdant *et al.*, "A 3.0μW@5fps QVGA Self-Controlled Wake-Up Imager with On-Chip Motion Detection, Auto-Exposure and Object Recognition," in *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, vol. 2020-June, doi: 10.1109/VLSI_Circuits18222.2020.9162854, 2020.
- [13] J. L. Robert K. Henderson, Nick Johnston, Sam W. Hutchings, Istvan Gyongy, Tarek Al Abbas, Neale Dutton, Max Tyler, Susan Chan, "A 256x256 40nm/90nm CMOS 3D-Stacked 120dB Dynamic-Range Reconfigurable Time-Resolved SPAD Imager," in *International Solid-State Circuits Conference (ISSCC)*, pp. 106–108, [Online]. Available: 10.1109/ISSCC.2019.8662355, 2019.
- [14] O. Kumagai *et al.*, "A 189x600 Back-Illuminated Stacked SPAD Direct Time-of-Flight Depth Sensor for Automotive LiDAR Systems," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 64, doi: 10.1109/ISSCC42613.2021.9365961, 2021.
- [15] F. Martin *et al.*, "An all-in-one 64-zone SPAD-based Direct-Time-of-Flight ranging sensor with embedded illumination," 2021.
- [16] Z. Yang, T. Albrow-Owen, W. Cai, and T. Hasan, "Miniaturization of optical spectrometers," *Science (80-.)*, vol. 371, no. 6528, doi: 10.1126/science.abe0722, 2021.
- [17] J. S. Steckel and et Alt, "1.62μm Global Shutter Quantum Dot Image Sensor Optimized for Near and Shortwave Infrared," 2021.
- [18] S. N. Esfahani, V. Muthukumar, E. E. Regentova, K. Taghva, and M. Trabia, "Complex Food Recognition using Hyper-Spectral Imagery," doi: 10.1109/CCWC47524.2020.9031258, 2020.
- [19] Y. Li *et al.*, "Inertial Sensing Meets Artificial Intelligence: Opportunity or Challenge?," pp. 1–14, [Online]. Available: <http://arxiv.org/abs/2007.06727>, 2020.

- [20] W. W. Lee, L. T. Nguyen, and G. S. Selvaduray, "Solder joint fatigue models: Review and applicability to chip scale packages," *Microelectronics Reliability*, vol. 40, no. 2, doi: 10.1016/s0026-2714(99)00061-x, 2000.
- [21] J. H. Lau, *Chip Scale Package, CSP*. McGraw-Hill Professional.
- [22] H. Krupakar, K. Rajvel, B. Bharathi, S. Angel Deborah, and V. Krishnamurthy, "A survey of voice translation methodologies-Acoustic dialect decoder," doi: 10.1109/ICICES.2016.7518940, 2016.
- [23] Federal Communications Commission, "Wireless E911 Location Accuracy Requirements Sixth Report and Order and Order on Reconsideration - PS Docket No. 07-114."
- [24] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, 2019.
- [25] J. Lin, W. M. Chen, Y. Lin, J. Cohn, C. Gan, and S. Han, "MCUNet: Tiny deep learning on IoT devices," in *Advances in Neural Information Processing Systems*, vol. 2020-December, 2020.
- [26] Q. Lu, W. Jiang, X. Xu, Y. Shi, and J. Hu, "On Neural Architecture Search for Resource-Constrained Hardware Platforms," [Online]. Available: <http://arxiv.org/abs/1911.00105>, 2019.
- [27] E. Liberis, Ł. Dudziak, and N. D. Lane, "μNAS: Constrained Neural Architecture Search for Microcontrollers," [Online]. Available: <http://arxiv.org/abs/2010.14246>, Oct. 2020.
- [28] N. Nayman, Y. Afalo, A. Noy, and L. Zelnik-Manor, "HardCoRe-NAS: Hard Constrained differentiable Neural Architecture Search," [Online]. Available: <http://arxiv.org/abs/2102.11646>, 2021.
- [29] J. G. López, A. Agudo, and F. Moreno-Noguer, "E-DNAs: Differentiable neural architecture search for embedded systems," doi: 10.1109/ICPR48806.2021.9412130, 2020.
- [30] G. Press, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," *Forbes Tech*, 2016.
- [31] Y. Wu, Z. Wang, Y. Shi, and J. Hu, "Enabling On-Device CNN Training by Self-Supervised Instance Filtering and Error Map Pruning," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, doi: 10.1109/TCAD.2020.3012216, 2020.
- [32] H. Ren, D. Anicic, and T. A. Runkler, "TinyOL: TinyML with Online-Learning on Microcontrollers," doi: 10.1109/ijcnn52387.2021.9533927, 2021.
- [33] G. Desoli *et al.*, "A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 60, doi: 10.1109/ISSCC.2017.7870349, 2017.
- [34] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, doi: 10.1109/JSSC.2016.2616357, 2017.
- [35] J. Hartmann, P. Cappelletti, N. Chawla, F. Arnaud, and A. Cathelin, "Artificial Intelligence: Why moving it to the Edge?," in *ESSDERC-ESSCIRC*, Keynote Speech, 2021.
- [36] C. N. Coelho *et al.*, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nat. Mach. Intell.*, vol. 3, no. 8, doi: 10.1038/s42256-021-00356-5, 2021.
- [37] "QKeras: a quantization deep learning library for Tensorflow Keras." <https://github.com/google/qkeras>.
- [38] R. Attanasio, G. Di Caro, S. Messina, and M. Torrisi, "Digital Controller Eases Design Of Interleaved PFC For Multi-kilowatt Converters," <https://www.how2power.org/article/2017/june/digital-controller-eases-design-of-interleaved-pfc-for-multi-kilowatt-converters.php>, 2017.
- [39] A. Andreini, C. Contiero, and P. Galbiati, "A New Integrated Silicon Gate Technology Combining Bipolar Linear, CMOS Logic, and DMOS Power Parts," *IEEE Trans. Electron Devices*, vol. 33, no. 12, doi: 10.1109/T-ED.1986.22862, 1986.
- [40] D. Rossi, G. Pedrazzini, G. Ricotti, and T. Cavioni, "Fully Integrated Fast Battery Charger with PWM current control and Microprocessor on Board," 1995.
- [41] N. O. Federici, D. Pau, N. Adami, and S. Benini, "Tiny Reservoir Computing for Extreme Learning of Motor Control," *International Joint Conference on Neural Networks (IJCNN)*, July 2021.
- [42] Y. Naung, S. Anatolii, and Y. Htet Lin, "Speed control of DC motor by using neural network parameter tuner for PI-controller," doi: 10.1109/EIConRus.2019.8656911, 2019.
- [43] B. Amarapur, "Neural network based speed control of induction motor," doi: 10.1109/NuICONE.2013.6780161, 2013.
- [44] "Industry 5.0: Towards a sustainable, human-centric and resilient European industry." [Online]. Available: https://ec.europa.eu/info/research-and-innovation/research-area/industrial-research-and-innovation/industry-50_en.
- [45] A. Ghadirzadeh, X. Chen, W. Yin, Z. Yi, M. Bjorkman, and D. Kragic, "Human-Centered Collaborative Robots with Deep Reinforcement Learning," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, doi: 10.1109/LRA.2020.3047730, 2021.
- [46] C. Görür, B. Rosman, F. Sivrikaya, and S. Albayrak, "FABRIC: A Framework for the Design and Evaluation of Collaborative Robots with Extended Human Adaptation." [Online]. Available: <https://arxiv.org/abs/2104.01976>, 2021.
- [47] B. Wilhelm, M. Bender, M. Braun, P. Rally, and O. Scholtz, "Lightweight robots in manual assembly – best to start simply!," [Online]. Available: <http://publica.fraunhofer.de/dokumente/N-415111.html>, 2016.
- [48] D. K. McCormick, *IEEE Technology Report on Wake-Up Radio: An Application, Market, and Technology Impact Analysis of Low-Power/Low-Latency 802.11 Wireless LAN Interfaces*. 2017.



Figure 1.2.1: Key long-term enablers of the "Automation Age".

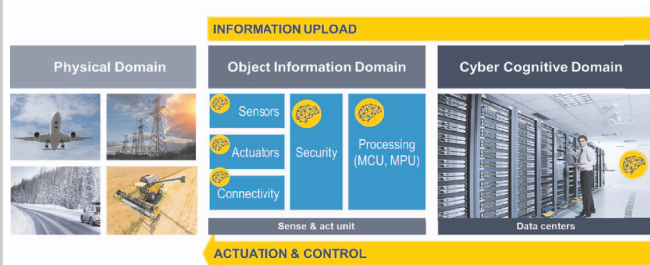


Figure 1.2.2: Distributed intelligence within the Cyber Physical System.

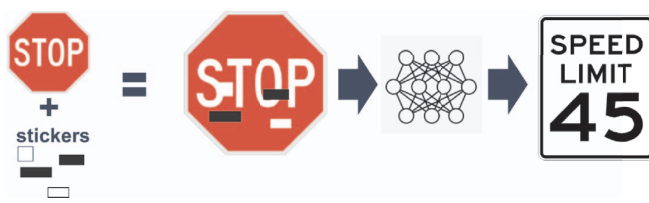


Figure 1.2.3: Adversarial attack on "STOP" sign resulting in misclassification of sign instruction.

Hype Cycle for Emerging Technologies, 2020



Figure 1.2.4: Embedded AI hype cycle (gartner.com/SmarterWithGartner).

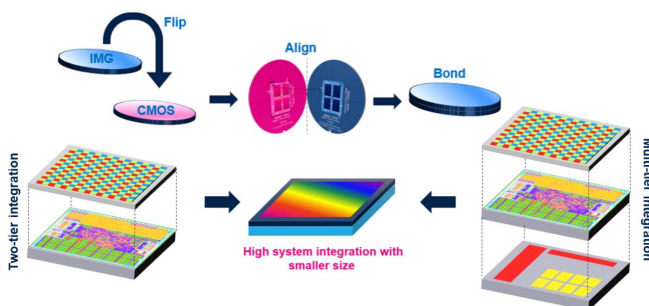


Figure 1.2.5: 2-tier and 3-tier wafer stacking for image sensors.

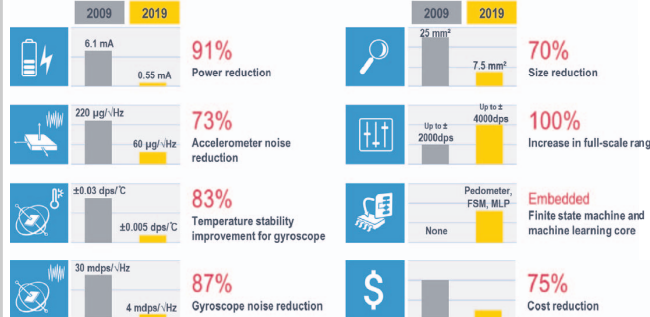


Figure 1.2.6: MEMS performance improvements over the last decade.

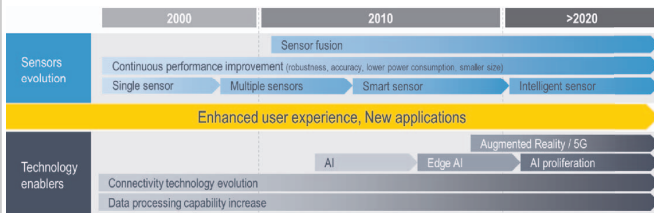


Figure 1.2.7: Key MEMS sensor development milestones.

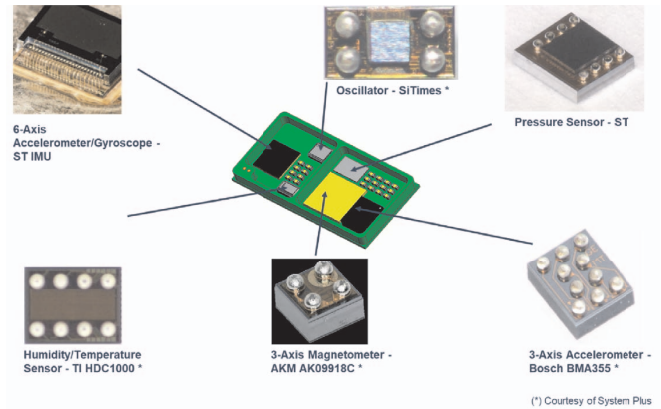


Figure 1.2.8: Integrating different sensing technology, data processing and transmission on same System in Package – an example.

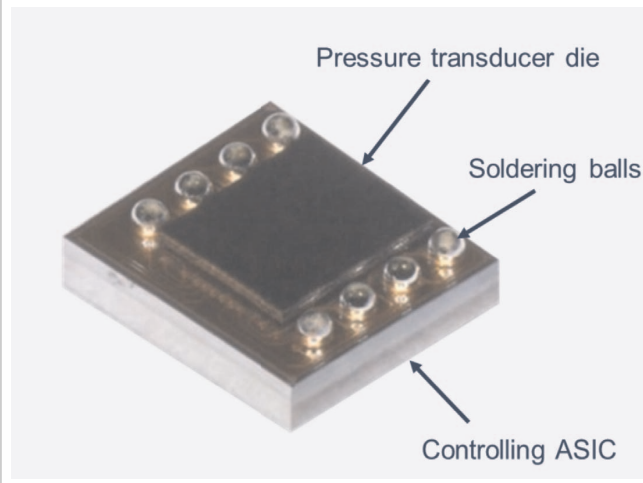


Figure 1.2.9: Tiny Chip Scale Package (CSP) example.

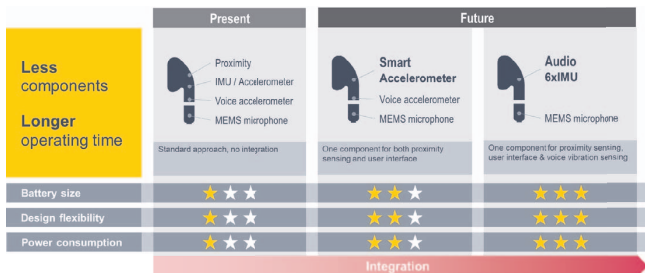


Figure 1.2.10: Microphones for vibration detection towards integration.

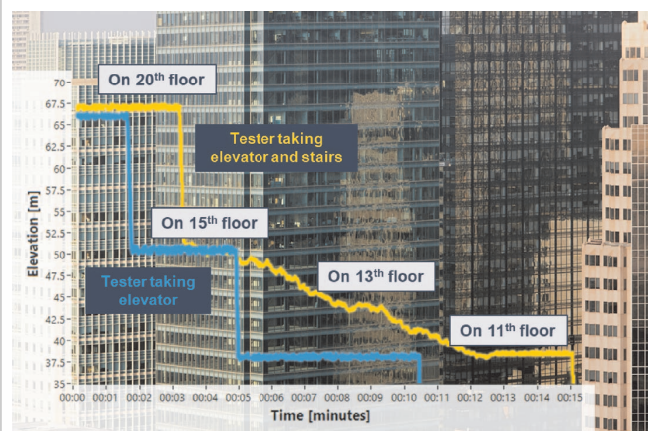


Figure 1.2.11: Indoor vertical positioning using pressure sensor for E911 compliance.

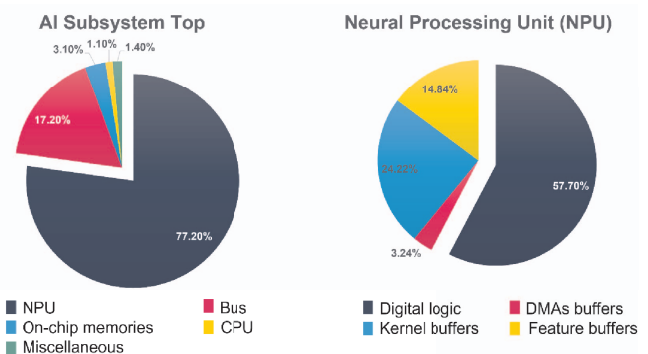


Figure 1.2.12: Relative power consumption breakdown for a typical hardware NPU at system level (left) and NPU core (right).

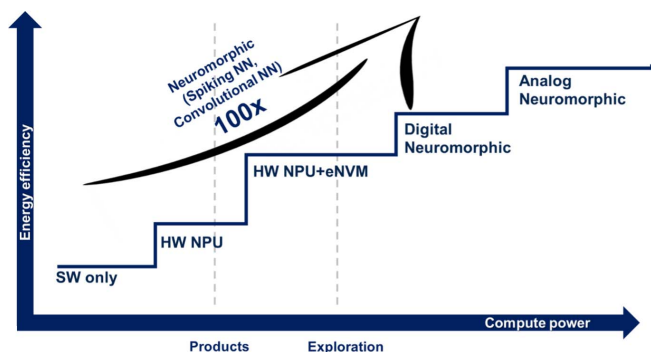


Figure 1.2.13: Toward a higher AI compute density/efficiency.



Figure 1.2.14: Operating frequency vs system power level for different intelligence actuator power technologies.

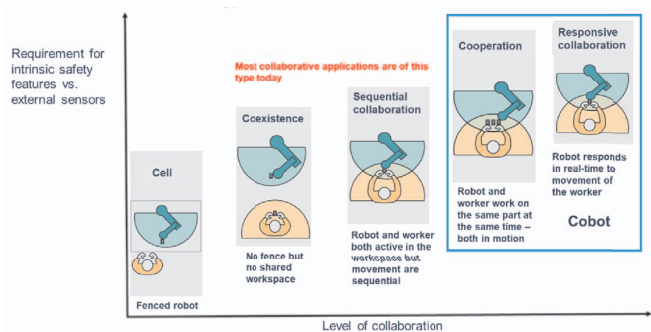


Figure 1.2.15: Interaction levels between human and robot in Industry 5.0 (adapted from [47]).

