1.2 Brain-Inspired Technologies: Towards Chips that Think?

Barbara De Salvo, Chief Scientist and Scientific Director

CEA-Leti, Université Grenoble Alpes, France

Abstract

The advent of the *Internet-of-Things* has introduced a new paradigm that supports a decentralized and hierarchical communication architecture, where a great deal of analytics processing occurs at the *edge* and at the *end-devices* instead of in the *Cloud*. To map the embedded-systems requirements, we present a *holistic research approach* to the development of *low-power architectures inspired by the human brain*, where process development and integration, circuit design, system architecture, and learning algorithms are simultaneously optimized. This paper is organized as follows: We begin with a survey of recent research on the *human brain* and a historical perspective of *cognitive neuroscience*. Then, *artificial intelligence* is introduced, and the challenges of *Deep Learning* systems (in terms of power requirements) are addressed. The key reasons to distribute intelligence over the whole network are discussed. To emphasize the need for low-power solutions, a quantitative *benchmark* of existing *specialized edge platforms* that can execute *machine-learning algorithms* on conventional embedded hardware is presented. The primary focus of this paper will be on the implementation of optimized *neuromorphic hardware* as a highly promising solution for future ultra-low-power cognitive systems. We show that *emerging technologies* (such as advanced CMOS, 3D technologies, emerging resistive memories, and Silicon photonics), coupled with *novel brain-inspired paradigms,* such as spike-coding and spike-time-dependent-plasticity, have extraordinary potential to provide intelligent features in hardware, approaching the way knowledge is created and processed in the human brain. Finally, we conclude with our vision of the *enabled future disruptive applications* and a discussion of the *main challenges* which should be tackled to exploit the full potential of brain-inspired technologies.

1.0 The Brain and Cognitive Science Perspective

More has been learned about the brain in the past decades than in all prior human history. In the quest for understanding the brain, neuroimaging has played a pivotal role, enabling *in-situ*, non-invasive brain mapping [1]. Medically, this has facilitated early diagnosis and treatment of patients with specific neurological or psychiatric diseases. *Magnetic resonance imaging (MRI*) has become the reference technique to investigate the human brain *in-vivo*, making anatomical/functional imaging and cerebral connectivity mapping possible (see **Fig. 1.2.1**). The MRI scanner operating at 11.7Tesla, currently installed at *CEA-Neurospin*, is being used to study the brain on a 100μm scale (addressing volumes corresponding to a few thousand neurons). Recent discoveries suggest that the brain organization has been shaped by a trade-off between a parsimonious principle of minimizing costs and maximizing adaptive values and robustness [2]. The large-scale neuronal networks of the brain are arranged globally as hierarchical modular networks, with dense modules at the local level (cellular circuits, laminar compartments) that are encapsulated in increasingly larger modules (cortical columns, areas and whole lobes), but with very sparse overall connectivity. Such a topology fundamentally enhances the brain's dynamic stability and information-processing abilities. An important research target will be to understand how the three-dimensional organization of brain cells, neurons and glial cells, connected in networks within the layers of the brain cortex, are responsible for the emergence of genetically-determined elementary operations. These operations combined together and interacting with the environment, give rise to higher-order functions, such as language, calculations, and consciousness.

This period of rapid discoveries has also seen the rise of *cognitive science,* a unified science based on interdisciplinary efforts among researchers in various fields (neurosciences, physics, biology, psychology, linguistics, artificial intelligence, robotics, and philosophy) whose aim is to investigate the functional architecture of cognition through computational models. Since its inception in the mid-1950s, cognitive science has moved through a series of different paradigms, which have strongly influenced the evolution of *artificial cognitive* systems. The first shift away from classical *behaviorism* (which asserts the dominant role of environmental factors on mental processes) came with *cognitivism*, under which thinking corresponds to a logical manipulation of symbols representing external phenomena. The rules that specify how symbols are transformed were taken to govern cognitive performance. Since the early 1980s, the important discoveries in the field of brain neurophysiology have led to the emergence of *connectionism. Connectionist* systems rely on parallel processing of non-symbolic distributed activation patterns using statistical properties, rather than logical rules. Their models reflect the concept of "*emergence*" in the brain's organization (cooperative interactions of individual components determining the "emergent" functionalities of the whole entity, given that these functionalities do not exist individually). The most common connectionist models are *neural networks*. By the repeated presentation of a training set and application of the learning rule, networks can learn to produce the correct responses to a set of inputs. In the past decade, connectionist models have strongly evolved and a new class of architectures (such as feedforward, fully-recurrent, simply recurrent) and learning frameworks (such as supervised, unsupervised, reinforcement), with almost no resemblance to biological systems, have been developed in order to implement them in artificial cognitive systems. It is worth mentioning that in recent years, a new approach called *embodied or enactive cognitive science* has emerged [3]. Whereas traditional cognitive science rests on a fixed inside–outside distinction, assuming that the mind is separate from the outside world, the embodied cognition approach views the *mind* as a biological system *rooted in body experience*, and *interacting with the environment and other individuals*. Embodiment refers to both the embedding of cognitive processes in brain circuitry and to the origin of these processes in an organism's sensory–motor experience. Action and perception are no longer interpreted in terms of the classic physical–mental dichotomy, but rather as being closely interlinked. The possible implications of this last paradigm in the design of future cognitive agents that interact with the physical world have yet to be fully explored.

2.0 Hyperconnectivity and Deep-Learning Power Challenges

In the past few decades, the world has experienced great transformations. Enabled by the convergence of miniaturization, wireless connectivity, increased data-storage capacity, and data analytics, the Internet-of-Things (IoT) has been the epicenter of profound social-, business-, and political-changes. With billions of easy-access and low-cost connected devices, the world has entered the era of hyperconnectivity [4], enabling people and machines to interact in a symbiotic way (anytime, anywhere) with both the physical and cyber worlds. *Artificial intelligence (AI)* has been at the center of this revolution. In recent years, we have seen a boost in the performance and applications of *machine learning (ML)*, driven by several factors: (i) the enormous storehouses of *data* (images, video, audio, and text files strewn across the Internet) which have been essential to the dramatic improvement of *learning/training* approaches and algorithms; (ii) the increased *computational power of modern computers* (the advent of parallel computing for neural network processing having compensated the slowing down of Moore's Law below the 10nm node). Among the many fields of ML, *Deep Learning (DL)* is the most popular. Today, for tasks such as image or speech recognition, ML applications are equaling or even surpassing expert human performance. Other tasks considered as extremely difficult in the past, such as natural language comprehension or complex games, have been successfully tackled. The particular case of the AlphaGo program from Google is remarkable in that it demonstrates how to increase performance by refining the algorithm architecture and combining several techniques of ML (DL techniques with reinforcement learning). In the future, new applications will require more and more analysis, understanding of the environment and intelligence. Self-driving cars will have to be able to recognize and analyze their environment through multiple sensors. Personal digital assistants will require voice and context analysis. For ML algorithms to become pervasive, increased computational resources will be needed. However, for the time being, data are transmitted in hierarchical infrastructures, and applications must deal with many different levels of analysis: *Cloud* computing, the *edge* (networked mobile devices), and the *end-devices* (wireless sensor nodes). Most of the data processing for *DL training,* and even for the *inference* phase, happen in the *Cloud* (data are sent to a data center and then processed there, before pushing operational decisions back to the edge platform). But, AI algorithms are not useful in settings where connectivity is sparse. Moreover, training a DL network in the *Cloud* (with conventional processors or GPU) on extremely large datasets involves intensive computing tasks and can take several weeks [5]. As well, the power limitations of servers used for DL are expected to slow down the pace of performance improvements. This poses a great challenge to computing platform designers.

3.0 Towards Distributed Intelligent Systems

Bringing intelligence to the *edge* or to *end-devices* means doing useful processing of the data as close to the collection point as possible, and allowing systems to make some operational decisions *locally*, possibly semi-autonomously. Distributing the intelligence over the network is important for a number of reasons: *Safety* will require local decision making, *in real time*, without having to rely on a connection that could be interrupted for various reasons. Running *real-time* DL locally is essential for many applications, from landing drones to navigating driverless cars. The delay caused by the round-trip to the Cloud could lead to disastrous or even fatal results. *Privacy* will require that key data not leave the user's device, while transmission of high-level information, generated by local neural-network algorithms, will be authorized. Raw videos generated by millions of cameras will have to be locally analyzed to limit *bandwidth issues* and *communication costs.* For all these reasons, new concepts and technologies that can *bring artificial intelligence closer to the edge* and *end-devices* are in high demand. The primary design goal in distributed applications covering several levels of hierarchy (similar to what happens in the brain), is to find a global optimum between *performance* and *energy consumption.* This imperative requires a holistic research approach, where the technology stack (from device to applications) is redesigned. As shown in **Fig. 1.2.2**, to address *embedded applications*, major industrial players and start-ups have developed specialized edge platforms that can execute ML algorithms (*inference*) on embedded hardware (CPU and GPU), such as Movidius Myriad X, MobilEye Eye Q5, Jetson TX2. Impressive power improvements (down to a few Watts) have been achieved by exploiting Moore's Law (pushing the FinFet technology down to the 7nm node) and by hardware-software co-optimization. Since many mobile applications are "*always-on*" (e.g., voice commands), low power is critical for mobile IoT [6]. In this context, several research groups have focused on hardware designs of C*onvolutional Neural Network (CNN)* accelerators. Precision-Scalable Processors (implemented in 40nm LP CMOS) for deep neural networks have shown power consumption in the range of 70mW [7]. The need for off-chip storage devices, such as DRAMs, significantly increases power consumption. Recently, mobile-oriented applications (keyword spotting and face detection) have been demonstrated with a low-power programmable *DL accelerator* [8] (incorporating on-chip weight storage) which consumed less than 300μW.

It is worth mentioning that the challenges of bringing intelligence into *low-power IoT-connected end-devices* (with applications ranging from habitat monitoring to medical surveillance) are much more demanding than those associated with traditional networked mobile devices at the edge [9]. Most connected end-devices are wireless sensor nodes containing microcontrollers, wireless transceivers, sensors, and actuators. The power requirement for these systems is extremely critical (<100μW for normal workloads), as these devices often operate using energy harvesting sources or a single battery for several years. The unreliable, noisy and complex environments where these systems are deployed create difficulties in modeling and predicting this environment (as in the case of energy harvesters and wireless communications). Fixed or non-intelligent communication protocols may dissipate the energy harvested at the nodes. To address this issue, *adaptive mechanisms* have been proposed which reduce energy requirements at the architectural or circuit level (dynamic and frequency scaling approximate computing). However, finding the best system configuration relies on knowledge of the system state [10]. Learning about the changing environment and configuring the system accordingly, using various techniques, is the key to achieving energy savings [11]. Moreover, fast and accurate *decision making* in IoT end-devices can be achieved using learning techniques. Many applications can be foreseen for ML, such as power and reconfiguration management, non-volatility control, and security-countermeasure activation. An example is provided in [12], where *neural cliques* behaving as an associative memory allow for very fast and accurate decision makings. Furthermore, *reinforcement-learning techniques* (where the learner must discover which actions yield the most rewards by trying them) can also be applied for end-device control with good accuracy at a very low cost [13]. When communication with the Cloud or edge devices is not possible, *live in-node data processing* and *classification* are required, and should be optimized so that they consume minimal energy, and preserve the quality of the information. *Genetic machine learning algorithms* have been explored for this purpose [14, 15], but integration of learning algorithms into low-power devices still remains an issue. Evaluations of computational requirements for embedding deep learning in low-power IoT devices (like, for example, a smart glass performing real-time recognition on the video stream that it captures) have shown a large processing efficiency gap between the capabilities of current computing platforms and the requirements imposed by such distributed applications [5]. Finally, to improve implementation efficiency of ML, various approaches have been explored. Nevertheless, given the energy costs related to the memory system, and the constraints on both parallelism and technology scaling, it might seem like there is not much room for additional energy improvements [16]. Finding new affordable, energy efficient ways to implement inference and learning through *new specialized low power and distributed compute engines* is thus key for future intelligent systems.

4.0 Advanced Technologies for Brain-Inspired Computing

Inspired by the *human brain*, whose computing performance and efficiency still remain unmatched (see **Fig. 1.2.2**), a radically different approach is being investigated: It consists in implementing bio-inspired architectures in *optimized neuromorphic hardware* to provide direct one-to-one mapping between the hardware and the learning algorithm running on it. This approach, which originated with the pioneering work of Carver Mead [17], has yet to be fully demonstrated and industrialized. Implementation limitations are linked to several elements [18], such as the difficulties to emulate the behavior of neural network elementary components (*neurons, synapse*s) with standard CMOS technologies, and to achieve a 3D brain-like high-density connectivity with 2D-layer technologies. In the following paragraphs, essential brain-inspired operating principles (such as *spike coding* and *STDP*) will be introduced, followed by a detailed discussion on how *emerging technologies* could lead to new neuromorphic hardware, and thereby change the rules of the game.

4.1 Spike Coding and Spike-Timing-Dependent-Plasticity (STDP)

The first brain-inspired operating principle to consider is the way neuron states are encoded in a system. In the past, neuron values were encoded using analog or digital values. However, a recent trend in neuromorphic computing is to encode neuron values as pulses or *spikes* [17, 24, 25]. This parsimonious signal coding was inspired by the way neurons of the central nervous system interact, leading to higher energy-efficiencies. It differs from the traditional *signal rate-coding* (used in today's main industrial neural network applications), which employs the average frequency of spikes in a given time window. The values manipulated in those networks (inputs and outputs of neurons) are numbers representing the "cumulative" effect of spikes over time. However, if input/output signals are represented as *pulses (spikes)*, the multiplication operation between input signals and synaptic weights is reduced to a gating operation at the synapse level. This typically produces a weighted current at the arrival of the pre-synaptic spike that is integrated by the post-synaptic neuron. The higher the frequency of the input spikes, the larger the value integrated by the neuron. Furthermore, if many synapses receive input spikes in parallel, the weighted sum operation is implemented directly at the input node of the post-synaptic neuron following Kirchhoff's current law. Thereby, power consumption can be reduced by implementing this spike or *event-based signal representation* (called *Address-Event Representation, AER)* using asynchronous schemes. Given these features and because this representation is also optimal for transmitting signals across long distances or chip boundaries, most of the recent state-of-the-art neuromorphic computing approaches are using AER. Moreover, spiking neurons offer the additional advantage of being easily interfaced with *low-power spiking sensors* (e.g., image-, audio- , tactile- or chemical sensors [56-60]). The second brain-inspired principle essential to neuromorphic systems is the *learning paradigm* (i.e. the way the synaptic connections among neurons are created, modified and preserved). The computation schemes to define the synaptic weights can be divided into two types: (1) *supervised learning*, where the inference process is based on training examples (this is the case for most neural-inspired machine learning algorithms, which show impressive performance for solving very specific tasks but at the cost of huge power dissipation,); and (2) *un-supervised learning*, which does not use any feedback from an external teacher, but attempts to classify inputs based on the underlying statistics of the data.

Spike-timing-dependent-plasticity (STDP) is a bio-inspired algorithm that enables unsupervised learning. The assumption underlying STDP is that synapses tend to reinforce causal links. That is, when the presynaptic neuron spikes just before the postsynaptic neuron spikes, the synapse between the two becomes stronger. Therefore, if the presynaptic neuron spikes again, the synapse will allow the postsynaptic neuron to spike faster or with a higher occurrence probability.

We will now present the extraordinary potential of *emerging technologies* which could be coupled to the aforementioned *novel brain-inspired paradigms* to provide intelligent features in hardware.

4.1.1 Fully-Depleted Silicon On Insulator (FDSOI)

For the past decade, FDSOI technology has proven to be a viable solution to satisfy Moore's Law requirements for the next CMOS generations [19]. It has been successfully deployed in many applicative fields (including entry-level application processors for smartphones, system-on-chip devices for autonomous driving and the IoT, and mm-wave applications). Thanks to its suitability for low-power design, FDSOI technology is a great candidate for neuromorphic hardware. In the field of *DL architectures*, *high-performance reconfigurable digital processors* based on 28nm FDSOI have shown power consumption in the range of 50mW. This power efficiency has been achieved by introducing optimized data-movement strategy and exploiting FDSOI back-biasing strategies [20, 21]. Recently, *a large-scale multi-core neuromorphic processor (*named *Dynap-SEL),* also based on 28nm FDSOI, was demonstrated (see **Fig. 1.2.3**) [22, 23, 24, 25]. It occupies an area of 7.28mm^2 and comprises four TCAM-based cores and one plastic core. Each TCAM-based core has 256 neurons and 16k TCAM-based programmable synapses, while the plastic core has 64 neurons with 4k plastic synapses, and 4k programmable synapses. In addition it integrates 8.5k × 18-bit SRAMs as Lookup Tables (LUTs), 3-level hierarchical routers, two temperature compensated bias generator circuits for generating 190 on-chip biases, and one input pre-decoder block. Thanks to the scalable architecture and to the on-chip programmable routers, the routing of all neurons on a 16×16 chip array can be easily configured to implement a wide range of connection schemes, without requiring external mapping, memory, or computing support. In order to minimize power consumption, a *mixed-signal design approach* was chosen and analog circuits were used. In this way, the physics of the device was exploited to implement the desired neural network computational primitives. Because these primitives are mainly composed of exponential and logarithmic functions, using sub-threshold analog circuits is the best choice. Indeed, the mixed-signal accelerator demonstrated in [22, 23] consumes 50pJ per spike, approaching the energy efficiency of biological neurons, which is estimated to be a few pJ per spike. Sub-threshold analog circuits reproduce the synapse and neural dynamics expected from theory. They can be used to provide biologically realistic dynamics or fast rectified linear unit transfer functions. They are also fully compatible with spike-based learning algorithms, and can be readily integrated into the next generation of large multi-neuron multicore neuromorphic architectures.

4.1.2 3D Through Silicon Vias (TSVs) and Monolithic 3D

The human brain's intelligence and efficiency is strongly linked to its extremely dense *3D interconnectivity* (roughly 10,000 synapses per neuron, and billions of neurons in the human brain cortex). The hierarchical structure in the cortex follows specific patterns, through vertical arrangements or *µcolumns* (where local data flow on subcortical specialized structures) and *laminar interconnections* (which foster inter-area communications and to build the hierarchy) [29]. Based on these considerations, it is clear that emerging 3D technologies will be a key enabler of efficient neuromorphic hardware. **Figure 1.2.4** shows the evolution (in terms of connection density) and hardware applications of 3D technologies. *Through Silicon Vias* have enabled heterogeneous system integration and are being increasingly used in devices (such as DRAM memory cubes, passive interposers for FPGA or GPU integration, BSI imagers, heterogeneous integration of MEMS and active interposers for High-Performance Computing [26]). Further scaling of 3D interconnects, to achieve pitches in the 1µm range, will be possible using *hybrid bonding* technology [27]. This approach offers a large architectural perspective and a way to overcome the classical limitations of today's imagers. A two-layer 3D partitioned *CNN architecture* is presented in [28]. Each layer comprises a neuronal compute block and the associated memory. This novel circuit uses fine pitch hybrid bonding and presents a substantial 25% improvement in power consumption when compared to a regular 2D version. Today, *3D Sequential Integration* (3DSI), also called monolithic 3D integration, offers new 3D partitioning options at fine granularities thanks to the ultra-small 3D contact pitch (<100nm) [30]. 3DSI consists in stacking active device layers on top of each other in a sequential manner. It differs from 3D packaging, where the tiers are fabricated in parallel, then stacked by bonding. As the top layer's active patterning is defined by the lithographical process-of-reference, the alignment accuracy and feature size of stacked tiers and inter-tier interconnections are dictated only by stepper resolution. This ultra-dense connectivity between memory arrays and computing logic provides much more parallelism capability for high-energy-efficiency computing [31]. Recently, a 3D monolithic integrated nanosystem, based on beyond-Si nanotechnologies, with vertically interleaved layers of computing and data storage, fine-grained and dense connectivity, was demonstrated [32]. Using 3DSI in neuromorphic computing will allow maximum connectivity and reconfigurability between neurons and synapses, a step forward towards cortical µcolumn-like interconnectivity.

4.1.3 Resistive Memories (ReRAM)

Several large-scale neuromorphic systems have been proposed in the last years, taking advantage of the enormous potential of current Silicon technologies. Examples include the Heidelberg's HICANN [35], IBM's TrueNorth [36, 37], and ETH's ROLLS [38] chips. These approaches use standard CMOS technologies to implement both neurons and synapses. The synaptic weights are stored in analog or digital devices such as capacitors or SRAM. Nevertheless, SRAM-based synapses are affected by the problems of area consumption and data volatility. When the network is turned off, the synaptic weights stored in the SRAM are lost, stressing the need for storage in nonvolatile memories (NVMs) during or after the learning process; but NVMs come with additional power and area consumption. Recently, new memory technologies, called ReRAM (such as phase-change memory (PCM), spin-transfer magnetic memory (STT-MRAM), metal-oxide resistive-switching memory (OxRAM), conductive-bridge memory (CBRAM) and Vertical Resistive Memories (VRAM)) have appeared. These memories offer several key features, such as: low voltages (ranging from 1V to 3V), fast programming and reading time (few 10s of ns, even <1ns), long data retention, single-bit alterability, execution in place, good cycling performance (higher than Flash), density and ease of integration in the Back-End-Of-Line of advanced CMOS. ReRAM are currently developed for applications such as microcontrollers [33], servers and high performance computers. Bringing memory close to the processing unit will revolutionize traditional memory hierarchy [34] and facilitate the implementation of in-memory computing architectures. Due to their low power consumption, multi-value properties, and non-volatility, ReRAM memories are also promising for implementing energy-efficient bio-inspired synapses in complex neural network systems [39-41] (see **Fig. 1.2.5)**. In [42], a CNN spike-based architecture for pattern recognition, using HfO_2-OxRAM devices as synapses for convolution kernels has been presented. It was inspired from the mammalian visual cortex organization and consists of two cascaded convolutional layers and a classification module. The CNN was simulated using an in-house special purpose C^{++} event-based simulator (*Xnet*) [43]. Kernels were defined using a backpropagation supervised learning algorithm. The OxRAM based CNN demonstrated high accuracy (recognition rate > 98%) for complex visual pattern recognition applications. This result is in agreement with the state-of-the-art recognition success rate obtained with formal CNN models, implemented with floating–point precision synapses. Thanks to the use of ReRAM synapses to implement the kernel, the convolution operations are performed directly in memory, reducing the latency per image recognition with respect to software implementations on GPU. The use of ReRAM synapses also opens a path towards *online real time unsupervised learning* (through continuous weight updating performed on local synaptic weights) and biological brain *life-long learning* abilities (i.e. once learned, it is almost impossible to train the same algorithm or network on a different task without completely re-learning all parameters). Plasticity will play an important role in achieving these goals. Two main approaches to emulate synaptic conductance modulation have been successfully demonstrated. In the *analog approach*, multiple low-resistance states for emulating long-term potentiation (cumulative increase of conductance, LTP) and multiple high resistance states for long-term depression (cumulative and gradual decrease of conductance, LTD) are used. In the *binary approach,* only two distinct resistive states (LRS and HRS) are used per device, with probabilistic STDP bio-inspired learning rules. This approach is also motivated by biological studies which suggest that STDP learning might be a partially stochastic process in nature. In the case of the binary approach, in order to improve performance, a single synapse could be composed of *n multiple binary cells* in parallel. Several ideas have been proposed to implement STDP with memory devices. A simplified version of STDP is presented in [43], where the analog time dependence of biological STDP is neglected, and only two conditions (increasing or decreasing synaptic weight) are considered. This model requires technologies with multilevel capability. *Phase-change memories* show a strong asymmetry between the SET and RESET process: whereas the SET process is extremely gradual and resembles learning in neural networks, the RESET process is abrupt. In [43, 44] a 2-PCM synapse that recreates artificial symmetry between SET and RESET by employing two devices per synapse has been proposed. This strategy has been shown to achieve unsupervised learning in a fully-connected neural network for automobile tracking. An average detection rate of 92%, and a system power consumption for learning of 112µW have been demonstrated by means of system-level simulations. In [45], an original methodology that uses *conductive-bridge RAM devices* as easy-to-program and low-power binary synapses with stochastic learning rules, is proposed. This learning scheme has been demonstrated on a *fully-connected neural network* able to process asynchronous analog data streams for recognition and extraction of repetitive patterns in a fully-unsupervised way. These demonstrated applications exhibit very good performance (auditory pattern sensitivity >2) and ultra-low synaptic power dissipation (0.55µW) in the learning mode. Low-power neuromorphic computing systems can also be coupled with *Brain-Computer Interfaces* (BCI) to enable the design of *autonomous implantable devices* for rehabilitation purposes, capable of making decisions based on real-time on-line processing of in-vivo recorded biological signals. In [46], a ReRAM-based *two-layer fully-connected neural network* able to identify, learn, recognize, and distinguish between different spike shapes of measured biological signals without any supervision, has been proposed.

Figure 1.2.5 shows the topological view of the network architecture: The biological signal is encoded by 32 frequency band-pass filters. The 32 filtered signals are then full-wave rectified and presented to the input layer of 32 neurons where the analog continuous signals are converted into spikes which are then propagated along the synapses to the five output neurons. To solve one of the main challenges of *biological signal treatment in BCI* (the high background-noise level), a synaptic compound using HfO_2-based OxRAM cells, able to implement two different flavors of spike-based synaptic plasticity, the long-term and the short-term learning rules, has been presented [47]. Thanks to long-term plasticity, the system is capable of learning based on an unsupervised paradigm, while the short-term plasticity allows for improved accuracy despite the significant background noise in the input data. Biology teaches us that noise can improve the performance of biological sensory systems. Inspired by this assessment, several studies have been devoted to leveraging *intrinsic device noise* for neuromorphic computing. For example, the stochastic switching behavior of ReRAM under weak programming conditions was used to implement synapses with probabilistic STDP learning rules [45-48], and neuron circuits with stochastic firing [49].

4.1.4 Silicon Photonics

Silicon (Si) photonic technologies are used today in datacenters for high-bandwidth multi-user communication networks. The recent advent of *hybrid platforms* that integrate photonic components on Si wafers in a cost-effective way [50, 51] opens new application fields. Optical interposers to stack and connect computing and memory chiplets together for very fast processing and high energy efficiency have been recently demonstrated [52]. Si photonics has also been explored for application in neuromorphic hardware. Photonic platforms offer an alternative approach to microelectronics, potentially overcoming the fundamental limit of highly-interconnected networks (the bandwidth connection-density tradeoff). The *high speeds, high bandwidth, and low cross-tal*k achievable in photonics seem very well-suited for ultra-fast spike-based information schemes with high interconnection densities. In [53], the use of electro-optic modulators as *photonic neurons* has been proposed. A reconfigurable 49-node Si photonic neural network able to perform emulation tasks has been presented. The results predict a *1960× speed-up over a CPU benchmark*. In [54], photonic hardware is proposed for the implementation of a *Reservoir Computer or Echo-State Networks*. This is a new paradigm in artificial Recurrent Neural Network (RNN) training, where an RNN, the *reservoir,* is generated randomly, and only a readout is trained [55]. Aside from the many potential advantages of photonics in general, it should be noted that photonic neuromorphic computing still remains a very exploratory field and more studies are needed to validate the promises.

5.0 Future Opportunities and Challenges

We are entering a new era where *artificial-intelligence systems* are becoming key players, shaping the future world. With the end of Moore's Law in sight, transformative approaches are needed to address the enduring power efficiency issues of traditional computing architectures. *Brain-inspired hardware,* coupled to new computing paradigms and algorithms, will exploit the full potential of new disruptive technologies and will allow for distributed intelligence over the whole IoT network, all-the-way down to ultra-low power end-devices. This will also open the way to unforeseen new applications. Nevertheless, to make this happen in a way that brings growth to society and benefits to individuals, several challenges still need to be tackled:

a) Despite the tremendous success of connectionist models (such as deep learning) in many important applications, our theoretical *understanding* of these systems is still far from complete. The complexity of the resulting systems makes it difficult to say which of their properties is most responsible for improved performance. Generalization in learning, abstraction, and reasoning abilities remains extremely limited, compared to human general intelligence. *Prediction* remains one of the fundamental problems in neural computation. Recently, neural networks were shown to fail while performing easy tasks where a human would never have failed (e.g., recognizing "*fooling images*", or images changed in a way imperceptible to humans [61]). Indeed, this threat limits market expansion, betrays user confidence, and gives rise to serious *ethical questions.* For these reasons, we believe that more understandable models should be developed, and more efforts should be put into the study of *neural network information and learning theories* [62, 63]. The biological plausibility of artificial systems should not be a burden for engineers' creativity. Nonetheless, we believe that more interactions between AI engineers, neuroscientists, and biologists will be strongly beneficial from a fundamental point of view.

b) The conceptual basis of the *embodied or enactive cognition paradigm* could be highly inspiring when defining new artificial systems suitable for the hyperconnected world. Future artificial cognitive systems will be autonomous physical systems which will need to interact in real time with the environment and individuals everywhere. Physical constraints will shape the dynamics of these interactions: In such systems, as with biological organisms, *the link between the low-level sensory-motor processes, control systems, and cognition will play a key role*. Bio-inspired approaches will force us to think differently. Simpler biological systems, rather than the human brain, will be highly inspirational and instructive. For example, the use of *insects as templates for artificial intelligent systems* [64] highlights the need to think in a systemic way, as *organisms do not decouple sensors and signal treatment*. Future autonomous systems will be required to perform intelligent tasks well beyond the possibilities of current ML systems (designed with a traditional input-output scheme and optimized to address classification tasks). The way they learn autonomously will be essential to define their predictive and interactive capabilities. Moreover, to account for the complexity of the world and the whole spectrum of future demands, it is probable that *a plurality of representational and cognitive architectural approaches* (based on cognitivist connectionist embodied-mind theories) will be needed, leading to *a world of heterogeneous and interconnected mixed-systems solutions*. Each approach will succeed in addressing different classes of empirical behaviors or will be more suitable for specific tasks.

c) Finally, *bio-hybrid interfaces between biological systems and VLSI neuromorphic systems* of varying complexity will play an important role in the future. Primarily intended as a computational tool for investigating fundamental questions related to neural dynamics, the sophistication of current neuromorphic systems makes direct interfacing with large neuronal networks and circuits possible, giving rise to interesting *clinical applications for neuroengineering systems, neuroprosthetics, and neurorehabilitation* [65, 66]. Leti's biomedical research center (Clinatec), dedicated to preclinical and clinical trials [4], is equipped with a cutting edge surgical operating room and medical facilities that are specifically and exclusively used for the qualification of advanced therapies and new prototypes based on micro-technologies. Here*, physicians, biologists, and engineers* work together to provide efficient and rapid validation of diagnostic and therapeutic tools using regulation-based evaluation processes. A high level of miniaturization and real-time data analysis were necessary to develop a *BCI* used in patients' rehabilitation [67] (see **Fig. 1.2.6**).

In the future, we will also witness the introduction of stimulation strategies based on real closed-loop systems, with signals emerging from wearable sensors (for example, sensing gloves). New materials to interface devices with living cells and tissues, new design architectures for lowering power consumption, data extraction and management at the system level, and secured communications are the next domains that will experience intense development. Brain-inspired implantable microdevices, acting as *intelligent neuroprostheses,* and *bio-hybrid systems* represent the new era of cross-disciplinary *brain-repair strategies*, where biological and engineered solutions will complement each-other, probably mediated by artificial intelligence [68, 69].

Acknowledgments - E. Beigne, S. Catrou, M. Belleville, A. Valantian, C. Reita, D. Dutoit, A. Hihi, E. Vianello, T. Dalgaty, S. La Barbera, P. Vivet, M. Causo, D. Morche, G. Sicard, A. Molnos, F. Heitzmann, L. Poupinet, S. Bonnetier from CEA-Leti, M. Duranton, C. Gamrat, A. Dupret, O. Bichler from CEA-List, A. Jerraya from CEA-DRT, B. Yvert from INSERM, Prof. G. Indiveri from ETH Zurich, Prof. J. Casas from University of Tours, Prof. S. Mitra from Stanford University.

References:

[1] "Human Brain MRI at 500MHz, Scientific Perspectives and Technological Challenges", Denis Le Bihan et al., Supercond. Sci. Technol., 30, 2017.

[2] "The Economy of Brain Network Organization", Ed Bullmore and at., Nature, 13, 2012.

[3] "The Embodied Mind", F.Varela et al, MIT Press, 1991.

[4] "Symbiotic Low-Power, Smart and Secure Technologies In the age of Hyperconnectivity", M.N. Semeria, IEDM 2016.

[5] "Efficient Embedded Learning for the IoT Devices", S. Venkataramani, IEEE 2016.

[6] "Can Deep Learning Revolutionize Mobile Sensing?", N.D. Lane et al., ACM International Workshop on Mobile Computing Systems and Applications, 2015.

[7] "A 0.3-2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets", B.Moons et al., VLSI 2016.

[8] "A 288μW Programmable Deep-Learning Processor with 270KB On-Chip Weight Storage Using Non-Uniform Memory Hierarchy for Mobile Intelligence," S. Bang et al., ISSCC 2017.

[9] "Ultra-Low-Power Networked Systems", A. Chandrakasan, Nano Tera Workshop 2015.

[10] "Spendthrift: Machine Learning Based Resource and Frequency Scaling for Ambient Energy Harvesting Nonvolatile Processors," K. Ma et al., 22nd Asia and South Pacific Design Automation Conf., p.678, 2017.

[11] "A Learning Theoretic Approach to Energy Harvesting Communication System Optimization", P. Blasco et al., IEEE Trans. on Wireless Comm., 12, 4, p. 1872, 2013.

[12] "Twin Neurons for Efficient Real-World Data Distribution in Networks of Neural Cliques: Applications in Power Management in Electronic Circuits", B. Boguslawski et al., IEEE Trans. on Neural Networks and Learning Systems, 27, 2, p.375, 2016.

[13] "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots", A. Amravati et al., ISSCC 2018.

[14] "An Approach to Implement Data Fusion Techniques in Wireless Sensor Networks Using Genetic Machine Learning aAlgorithms", A.R. Pinto et al., Information Fusion, 15, p.90, 2014.

[15] "Machine Learning Methods in Data Fusion Systems", R. Nowak et al., 13th International Radar Symposium, p.400, 2012.

[16] "Computing's Energy Problem (and what we can do about it)", M. Horowitz, pp. 10-14, ISSCC 2014.

[17] "Analog VLSI and Neural Systems", C. Mead, Addison-Wesley VLSI Systems Series, 1989.

[18] "Advanced Technologies for Brain-Inspired Computing", F. Clermidy et al., IEEE 2014.

[19] "Planar Fully-Depleted-Silicon-On-Insulator Technologies: Past Research, Current Status and Future Directions", B.Doris et al., Solid State Electronics, 117, p.37, 2016.

[20] "ENVISION: A 0.26-to-10TOPS/W Subword-Parallel Dynamic-Voltage-Accuracy-Frequency-Scalable Convolutional Neural Network Processor in 28nm FDSOI", B.Moons, et al, p. 246, ISSCC 2017.

[21] "A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems", G.Desoli et al., p. 238, ISSCC 2017.

[22] "Scaling Mixed-Signal Neuromorphic Processors to 28 nm FD-SOI Technologies", N. Qiao et al., IEEE Biomedical Circuits and Systems Conf., 2016.

[23] "Analog Circuits for Mixed-Signal Neuromorphic Computing Architectures in 28 nm FD-SOI Technology", N. Qiao et al., IEEE S3S, 2017.

[24] "Neuromorphic Architectures for Spiking Deep Neural Networks", G.Indiveri et al., IEDM 2015.

[25] "Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain", R. Brette, Frontiers in Systems Neuroscience, 9:151, 2015.

[26] "ITAC: A Complete 3D Integration Test Platform", D. Lattard et al., 3DIC 2016.

[27] "New Perspectives for Multicore Architectures Using Advanced Technologies", F. Clermidy et al., IEEE IEDM 2016.

[28] B. Belhadj et al. CASSES'2014.

[29] "The Neocortical Circuit: Themes and Variations", Harris KD et al., Nat Neurosci., 2, p.170, 2015.

[30] "3D Sequential Integration: Application-Driven Technological Achievements and Guidelines", P. Batude et al., IEDM 2017.

[31] "Energy-Efficient Abundant-Data Computing: The N3XT 1,000x", M. M. Sabry Aly et al., IEEE Computer, 48, 12 2015.

[32] "Three-Dimensional Integration of Nanotechnologies for Computing and Data Storage on a Single Chip", M. M. Shulaker et al., Nature, 547, p.74, 2017.

[33] "Universal Signatures from Non-Universal Memories: Clues for the Future...", L; Perniola, IEEE IMW, 2016.

[34] "Non-Volatile Memory Evolution and Revolution", P. Cappelletti, IEDM 2016.

[35] "A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling", J. Schemmel et al., IEEE Int. Symp. on Circuits and Systems, 2010.

[36] "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface", P. A. Merolla et al., Science, 345, 6197, p.668, 2014.

[37] "TrueNorth Ecosystem for Brain-Inspired Computing: Scalable Systems, Software, and Applications", J. Sawada et al., ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, 2016.

[38] "A Reconfigurable On-Line Learning Spiking Neuromorphic Processor Comprising 256 Neurons and 128K Synapses", N. Qiao et al., Frontiers Neuroscience, 9, p.141, 2015.

[39] "From Memory in our Brain to Emerging Resistive Memories in Neuromorphic Systems", B. DeSalvo, IEEE IMW, 2015.

[40] "Oxide Based Nanoscale Analog Synapse Device for Neural Signal Recognition System", D. Lee et al., IEEE IEDM 2015.

[41] "Large-Scale Neural Networks Implemented with Non-Volatile Memory as the Synaptic Weight Element: Comparative Performance Analysis (Accuracy, Speed, and Power)", G.W. Burr et al., IEEE IEDM 2015.

[42] "HfO2-based OxRAM Devices as Synapses for Convolutional Neural Networks", D. Garbin et al., IEEE Tr. On El. Devices, 2015. [43] "Synapses Made by Two Phase-Change Memory Devices for Efficient Spiking Neural Networks", O. Bichler et al., IEEE Tr. On El. Devices, 2012.

[44] "Phase Change Memory as Synapse for Ultra-Dense Neuromorphic Systems: Application to Complex Visual Pattern Extraction", M.Suri et al., IEDM 2011.

[45] "Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses", M. Suri et al., IEEE Trans. on Electron Device, 60, 7, 2402, 2013.

[46] "Spiking Neural Networks Based on OxRAM Synapses for Real-Time Unsupervised Spike Sorting", T. Werner et al. Frontiers in Neuroscience, 10, 474, 2016.

[47] "Resistive Memories for Spike-Based Neuromorphic Circuits", E. Vianello et al., IMW 2017.

[48] "Spintronic Devices as Key Elements for Energy-Efficient Neuroinspired Architectures", N. Locatelli et al, Design, Automation & Test in Europe Conference & Exhibition, p. 994, 2015.

[49] "Stochastic Neuron Design Using Conductive Bridge RAM", G. Palma et al., IEEE/ACM International Symposium on Nanoscale Architectures, 2013.

[50] "Light Is the Ultimate Medium for High-Speed Communications", C. Kopp et al., EuroPhotonics, 2017.

[51] "Low-Temperature Crack-Free Si3N4 Nonlinear Photonic Circuits for CMOS-Compatible Optoelectronic Cointegration", M. Casale et al., SPIE Photonics West, OE109, 2017.

[52] "10 Gbps, 560 fJ/b TIA and Modulator Driver for Optical Networkson-Chip in CMOS 65nm", J.L. Gonzalez et al., 14th IEEE Intern. NEWCAS Conference, 2016.

[53] "Neuromorphic Photonic Networks Using Silicon Photonic Weight Banks", A.N. Tait et al., Scientific Reports 7, 7430, 2017. [54] "High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification", L. Larger et al., Phys. Rev. X 7, 2017.

[55] "Reservoir Computing Approaches to Recurrent Neural Network Training", M. Lukoševicius et al., Computer Science Review, 3, 3, p.127, 2009.

[56] "A 128 x128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Change", P. Lichtsteiner et al., IEEE ISSCC 2006.

[57] "A QVGA 143dB Dynamic Range Asynchronous Address-Event PWM Dynamic Image Sensor with Lossless Pixel-Level Video Compression", C. Posch et al., IEEE ISSCC 2010.

[58] "Low-Power Spiking Chemical Pixel Sensor", P. Georgiou et al., Electronics Letters 42, 23, p.1331, 2006.
[59] "A 0.5V 55μW 64×2-Channel Binaural Silicon Cochlea for Event-Driven Stereo-Audio Sensing", M. Yang et al., IEEE ISSCC, p. 388, 2016.
[60] "Spike-Based Readout of POSFET Tactile Sensors", S. Caviglia et al., IEEE Trans. on Circuits and Systems, 64, 6, p.1421, 2017.
[61] "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", A. Nguyen et al., IEEE Computer Vision and Pattern Recognition (CVPR), 2015.
[62] "Theory of Deep Learning I, II, III", C. Zhang et al., Tech.
Rep., MIT Center for Brains, Minds and Machines, 2017.
[63] "Deep Learning", Y. LeCun et al., Nature, p.436, 2015.
[64] "Biomimetic Flow Sensors", J.Casas et al., Encyclopedia of Nanotechnology, Springer Verlag, 264, 2012.
[65] "Generation of Locomotor-Like Activity in the Isolated Rat Spinal Cord Using Intraspinal Electrical Microstimulation Driven by a Digital Neuromorphic CPG", S. Joucla et al., Frontiers in Neuroscience, 10, 67, 2016.
[66] "Real-Time Control of An Articulatory-Based Speech Synthesizer for Brain-Computer-Interfaces", F. Bocquelet et al., PLoS Comput. Biol., 12, 11, 2016.
[67] "WIMAGINE®: Wireless 64-Channel ECoG Recording Implant for Long Term Clinical Applications", C. Mestais et al., IEEE Trans. on Neural Systems and Rehabilitation Engineering, 23, 1, 2015.
[68] "Intelligent Biohybrid Systems for Functional Brain Repair", G. Panuccio et al., New Horizons in Translational Medicine, 3, p.162, 2016.
[69] "Trends and Challenges in Neuroengineering: Toward "Intelligent" Neuroprostheses Through Brain-"Brain Inspired Systems", S. Vassanelli et al., Communication. Front. Neurosci. 10, 438, 2016.

Figure 1.2.1: Atlas of brain connectivity, showing the long white matter fiber bundles (connections are visualized using diffusion MRI). A specific color is attributed to each fiber bundle [1].

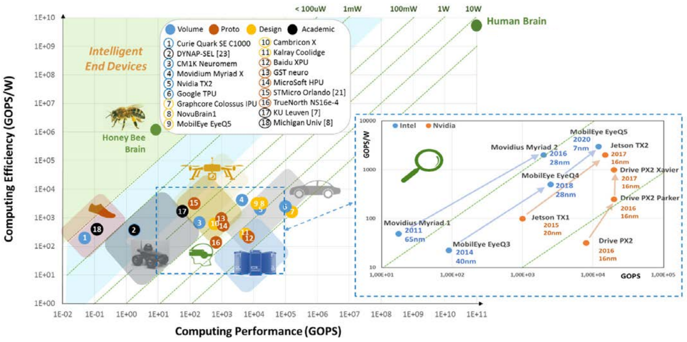

Figure 1.2.2: Comparison of computing efficiency (GOPS/W) during the inference phase versus computing performance (GOPS) of several intelligent chips from literature and the web, showing the gap between the intelligent end-device requirements and existing solutions. *Note that we took the very coarse approximation of a 1:1 correspondence between OPS, FLOPS, IPS, SOPS (SOPS = firing rate × average active synapses).*

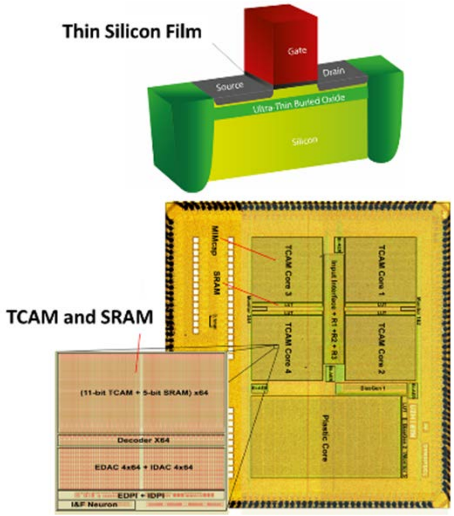

Figure 1.2.3: Dynap-SEL neuromorphic chip, based on a 28nm FDSOI process [22, 23].

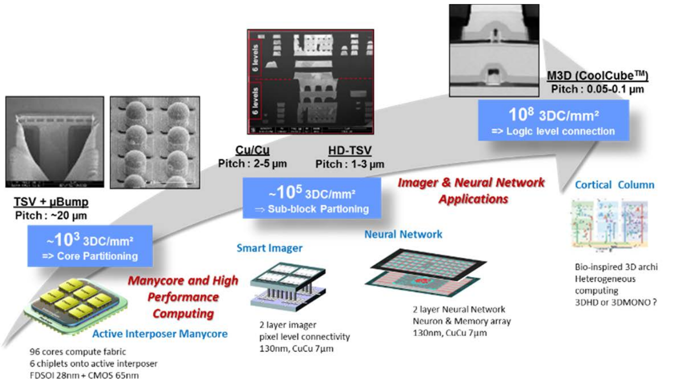

Figure 1.2.4: CEA-Leti roadmap of 3D technologies, showing the connection-density evolution and corresponding hardware applications.

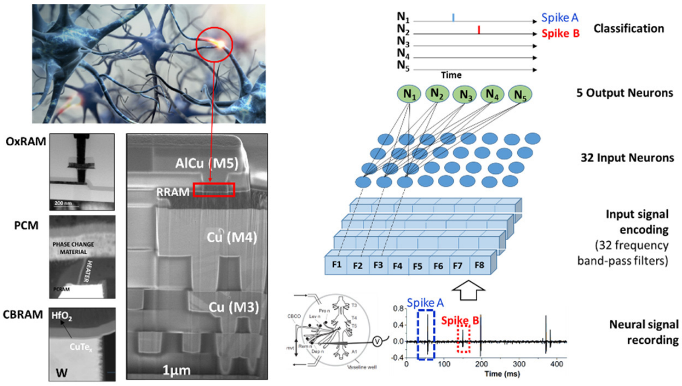

Figure 1.2.5: Left: Illustration of a biological synapse and the concept of using ReRAM as synapses. Right: Functional schematic of a spiking neural network for real-time unsupervised spike sorting [46].

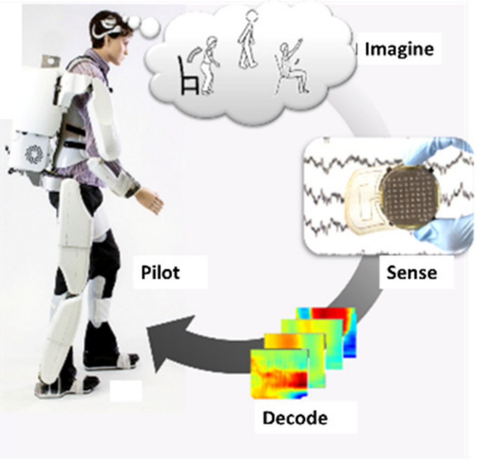

Figure 1.2.6: Illustration of the BCI project, showing functional substitution for tetraplegic subjects via a driven 4-limb exoskeleton [67]. In the future, intelligent neuroprostheses and biohybrid systems for therapeutic purposes are foreseen [68, 69].