imec
embracing a better life

WHITE PAPER ARTIFICIAL INTELLIGENCE

AI at the edge - a roadmap

Jan M. Rabaey, Marian Verhelst, Jo De Boeck, C. Enz, Kristiaan De Greve, Adrian M. Ionescu, Myunhee Na, Kathleen Philips (editors)
with F. Conti, F. Corradi, K. De Greve, M. De Ketelaere, B. Dhoedt, M. Hartmann, K. Myny, I. Ocket, W. Philips, W. Verachtert, D. Verkest

IMEC, KU Leuven, Ghent University, VUB, EPFL, ETH Zurich and UC Berkeley

Contents

Introduction

With the explosive growth in the adoption of Artificial Intelligence (AI) to address a large range of problems that were deemed very hard, the information technology and semiconductor communities have rushed to develop computing platforms that provide the desired performance at an acceptable energy cost. So far, the majority of attention and investment has been directed towards "Cloud AI", with "Data" being the largest common denominator in creating value for industries, governments, and individuals' lives. However, the quest for intelligence is fast becoming a prominent and essential feature at the "Edge" as well, where trillions of "things" will combine to generate even more data. Given the severe constraints that govern edge devices in terms of efficiency, footprint, robustness and cost, it is self-evident that bringing true intelligence to the edge will require profound innovation at all levels of the stack from the computational concepts all the way down the implementation technology. This observation spans the full range of applications ranging from automotive, mobile, industrial, immersion, IoT to wearable and implantable. In this white paper, we evaluate a number of plausible options, explore possible paths forward, and present a set of recommendations on how to make it happen.

This White Paper is the direct outcome of a brainstorming workshop that was held in Leuven on September 17, 2019 with the authors and contributors attending.

Why focus on the edge?

There is little doubt that artificial intelligence (AI) and more particularly, Machine Learning (ML) will be an essential if not dominant part of the computing landscape over the next decade. Various forms of AI are already adopted by virtually every branch of industry, government and society, and this may just be the beginning. It hence comes as no surprise that all major industrial nations and regions are investing heavily in the development of the next generation of AI services, products, soft-and hardware platforms. At the latest count, there are 8705 startups and companies listed in Crunchbase who are relying on machine learning for their main and ancillary applications, products and services. Artificial Intelligence-related companies raised $9.3B in 2018, a 72% increase over 2017, according to PwC/CB Insights MoneyTree Report, Q4 [CBI18]. In the second quarter of 2019 alone, $7.4 Billion was invested in AI startups. And while the US was responsible for the lion share of the investment until recently, regions/countries such as China, Japan, Europe and Israel are accelerating very quickly.

"The AI growth factor for the Edge is expected to be even larger than for the Cloud over the next decade."

From a computational perspective, the data-driven and learning-based nature of AI translates into programming models and compute platforms that fundamentally differ from those that have dominated computing for the last 7 decades. While it is true that most AI tasks can be executed on existing compute-servers, they cannot do so efficiently. Traditional processors such as CPUs and conventional GPUs do not natively map the processing kernels and data flow patterns found in most AI algorithms. Complexity, performance, and power considerations combined with the need to process massive data sets, preferably with large degrees of inherent parallelism, have therefore spawned a new generation of computer architectures that are optimized for specific AI tasks. The processing units in these architectures are often denoted as NPUs (neural processing units) or TPUs (tensor processing units). They form an addition to traditional CPUs and GPUs in the form of either specialized accelerators, standalone processors, or modifications of existing processor architectures (such as GPUs with added tensor cores). As such, the market for chip sets focused on AI has exploded. The AI chip set market as a whole is expected to more than double over the next five years [Eet19]. More than 60 companies worldwide are in advanced stages of either building or selling specialized processors to accelerate AI applications [Ark19]. In total, venture capital firms have invested more than $2.6 Billion across 40 startups in this field.

However, some caveats are worth raising. Most of the investment in AI hardware has been focused on cloud intelligence and big data – for very good reasons, as the vast majority of AI applications and services have been cloud-based to date. While this has created opportunities for semiconductor companies such as Nvidia or Intel, the momentum is changing towards a vertical integration model monopolized by cloud companies and specialized consumer companies. Amazon, Google, Facebook, Apple, Tesla, Alibaba, Baidu, etc. are all developing or deploying their own customized hardware that is optimized for their own specific application domains and functions, thereby reducing the market for the traditional semiconductor companies. Having said this, there is no question that there is still plenty of room for creativity, innovation and business opportunity in this space.

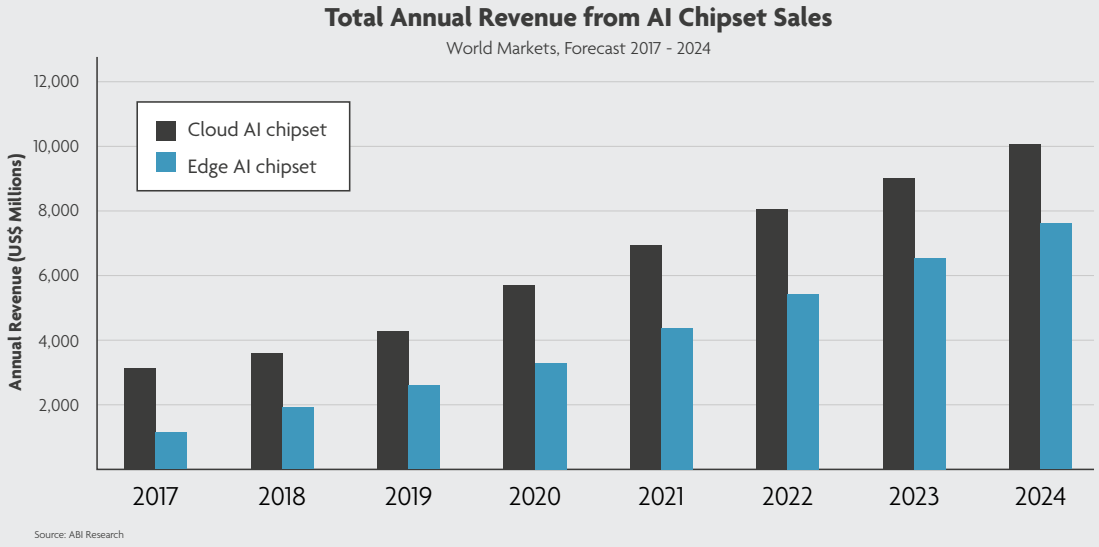


Fig. 1: Total annual revenue from AI chipset sales, 2017 to 2024 (Source: ABI Research and embedded.com) [ABI19]

While the focus of AI hardware has been primarily on the cloud, another development has been happening somewhat under the radar screen, primarily driven by the explosive growth in the mobility, IoT, wearables and Industrial Internet application domains. While early IoT devices were primarily sense-communicate modules hooked to the cloud, the trend nowadays is towards adding local intelligence for a variety of reasons including communication cost, latency, robustness, security, privacy, latency and power efficiency. Hence an increasing interest in Edge AI is being observed. Here, inference and often even learning functions are deployed in close proximity to sensors and actuators within resource-constrained devices, such as smartphones, IoT devices, smart wearables, domotics, and autonomous vehicles. The AI growth factor for the Edge is expected to be even larger than for the Cloud over the next decade [Tra19, Eet19]. Looking at the AI chip market, revenue from Edge AI may grow at the same if not even faster rate than that from Cloud AI, as illustrated in Figure 1 [Eet19, ABI19]. A healthy growth from $1.9B in 2018 to $7B in 2024 is projected, presenting plenty of opportunity for both established semiconductor companies and start-ups alike. However, given the broad application space, that market is likely to be more fragmented.

While one may argue that there is a large common denominator between Cloud and Edge AI, there are also many reasons to believe that their roadmaps will diverge in substantial ways over coming years. Their applications differ and so do their implementation constraints. Most Cloud applications are focusing on processing and interpreting large data sets, and as such data memory bandwidth, and computational performance are paramount. Edge devices, on the other hand, often face extremely tight constraints in terms of size and energy budget. Moving massive amounts of data from the Edge to the Cloud is probably not sustainable with the current (and even the next) generation of communication technology (Fig. 2), and is very inefficient from an energetic point of view. Even more, it poses additional security risks. Another important differentiation is the stringent round-trip latency requirement (often of the order of milliseconds or smaller), in applications that require real-time decision making such as autonomous mobility and healthcare. As a result, realizations at the edge often end up using entirely different computational models and optimization methods, architectures and ultimately also integrated system and circuit technologies.

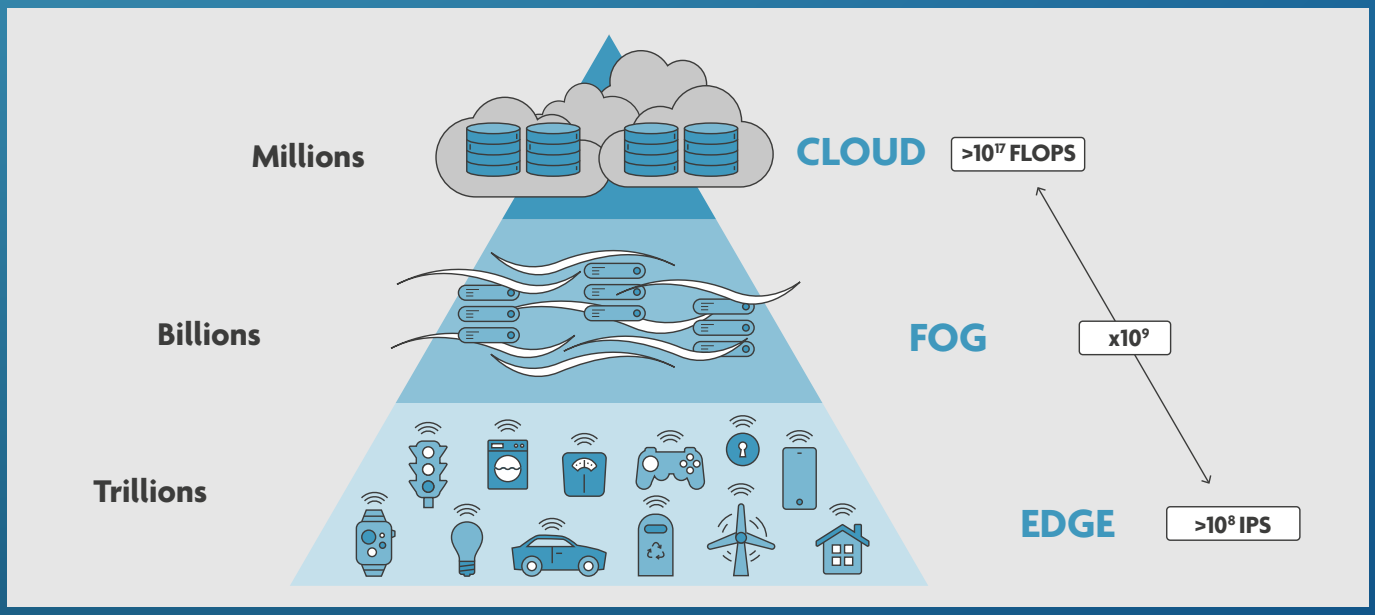


Fig. 2: Edge-to-Cloud information processing model for AI leads to unsustainable and inefficient zettabyte data movement (Source: EPFL View on Future Grand Challenges [AI19]).

Finally, there is one more argument why further divergence is extremely likely. While most early realizations of Edge AI focus on data analysis and classification, or event/anomaly detection (often in collaboration with the cloud), an observed trend is towards increased autonomy and closed-loop feedback. In this, the results of the data analysis serve as inputs to a controller and are translated in concrete and immediate action within a very tight latency window. This is definitely the case in applications such as autonomous mobile devices (cars, drones), robotics, human-machine interfaces (think AR/VR), brain-machine interfaces and wearable medical devices. This is captured in Fig 3, which pictures the evolution of AI towards automated intelligent decisioning systems (ADS). In addition to input (sensors) and output (actuator) modalities, those systems need to combine understanding, reasoning and decisioning. While one can easily imagine that parts of such systems can be run from a centralized location, energy, latency, robustness, security and privacy considerations form powerful arguments for more distributed realizations with a large share of the functionality executed on the Edge device itself (or a cluster/swarm thereof), and eventually penetrating all the way into the sensors and actuators ("in-sensor computing"). In all likelihood, a majority of the systems will combine both centralized and distributed functionality though.

While major progress has been made over the past decade, it is fair to state that **we are far away from the realization of such intelligent decisioning systems in form factors and performance/power budgets imposed by the systems they are embedded in**. Even if we would have access to implementation technologies that would meet those goals and restrictions (and unfortunately, as of yet, we do not ...), the state-of-the-art in artificial intelligence itself is still not up to the needs of the described reasoning and decisioning systems. Indeed, today's AI is not even close to human intelligence - a fair target to pursue - in so many ways that will prove to be crucially important. Beyond reasoning, autonomous entities need to be capable of dealing with novelty, learning from analogy, remember and forget, evolve, and adhere to ethical behavior. To paraphrase Facebook's Yan LeCunn, "we need machines learning to learn, and build their own models of the world as they are encountering it." [LEC19a] These observations will almost without a doubt lead to paradigm-shifting approaches that go beyond the current state-of-the-art, and will increase demands on the implementation platform manifold. A number of research programs are currently under way or have been launched worldwide to explore the potential of alternative approaches to artificial intelligence. An interesting perspective on some of these evolutions is presented in a White Paper published by the UC Berkeley AI group [Ber19], a summary of which is plotted in the chart of Fig. 4. There is no doubt that the outcome of those will continue to inspire innovation in the Edge AI implementation platforms.

DEFINING ARTIFICIAL INTELLIGENCE

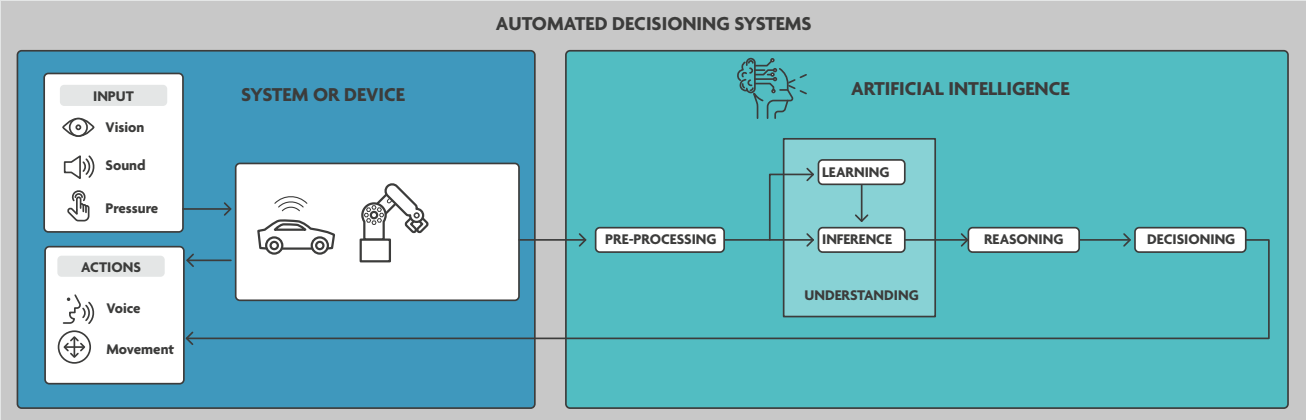


Fig. 3: The evolution to automated decisioning systems [Ket19]

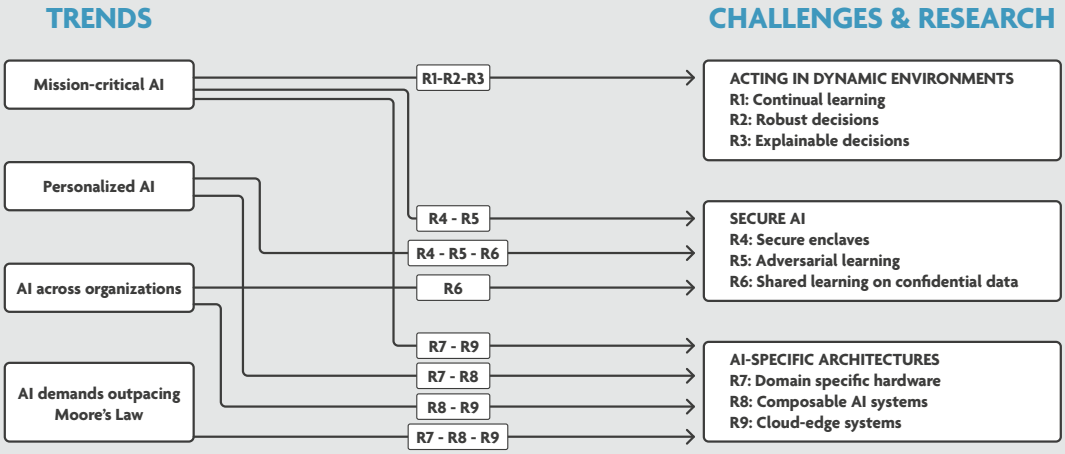


Fig. 4: The Berkeley View on AI. Observe that of all trends, mission-critical, personalized and outpacing Moore's Law all apply to Edge AI. The vast majority of identified challenges also apply to the Edge [Ber19].

In the rest of this document, **we will argue that the effect of profound innovation in hardware realizations combined with the massive deployment of sensor and actuator devices in the physical world will accelerate the creation, the growth, and the evolution of intelligence at the edge**. After all, it was the interplay between rapidly advancing hardware and the availability of huge data sets that led to the breakthroughs in conventional, cloud-based AI [LEC19b]. To help focus the discussion, we first describe the state-of-the-art in Edge AI, discuss metrics to gauge progress, and identify technology opportunities. We then proceed by proposing a challenging set of application drivers ("moonshots") that may help to not only define actionable metrics, but also metricize progress and spur innovative thinking. The paper concludes with a set of recommendations on how to further the field and what concrete actions to take.

The state of the art in (edge) AI and its realizations

While the field of Edge AI is relatively young, progress has been swift. Both within academia and industry, AI building blocks and integrated systems have been realized that have enabled the realization of some AI functionality within mobiles and edge devices. For instance, functions such as facial recognition are now widely available on mobile devices.

Over the past years, tremendous progress has been made in reducing the footprint of AI in terms of power, size and cost. AI accelerators have been designed that allow substantial AI functionality to be moved to the edge. Gains of orders of magnitude in efficiency and complexity were obtained through a number of design choices and optimizations: new network topologies, massive hardware concurrency (MAC arrays), dedicated memory interfaces, customization, exploitation of sparsity, reduction in computational accuracy, in-memory computing, etc. An overview of a number of these techniques was presented in a keynote presentation by HJ Yoo (KAIST) at ISSCC 2019 [Yoo19]. Some of these advances are illustrated in Fig. 5. Many of the underlaying principles were introduced by groups in Europe, including KU Leuven, IMEC and ETHZ [e.g. Moo17, Ban18], as well universities in the United States, China and Korea. Unfortunately, the realization of true Automated Intelligent Decisioning Systems (ADS) and human-like capability will require at least two orders of magnitude of improvements in efficiency and complexity that can only be accomplished through innovation and creativity[1].

In identifying the path forward, it is worth pointing out that the **field of Artificial Intelligence, as it has emerged over the past 4 decades, is broad and actually covers a number of widely diverging approaches, technologies and methodologies**. It is not the intent of this paper to discuss this in great depth (there is a vast literature available doing just that). However, some broad classifications are necessary, as they help rationalize the various computational models and architectural choices.

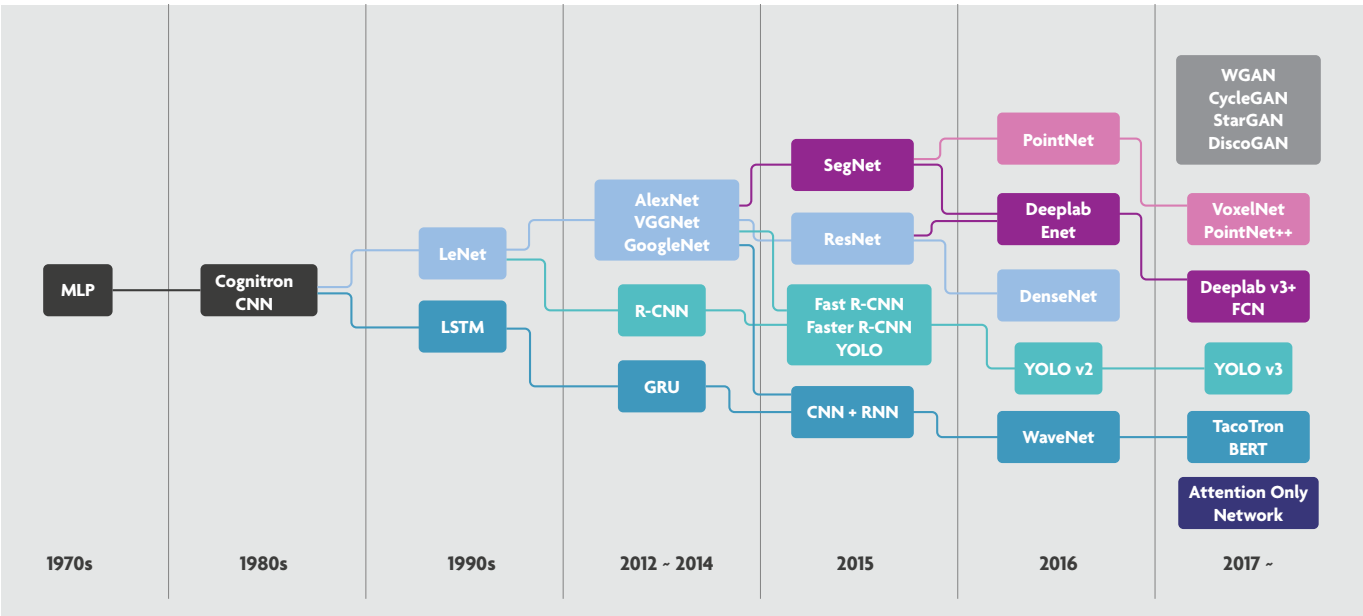


Fig. 5: Evolution of deep neural networks [Yoo19]

- In general, **artificial intelligence** covers **any approach that mimics human intelligence, that is any program that can sense, reason, act, learn and adapt**. As such, rule-based systems fall under this definition, so do smart controllers. Under this general header, a number of current approaches can be identified. It is fair to assume that we are still in the dawn of true Human AI, and that we are bound to see the emergence of other models and approaches in the coming years. Some of these may be inspired by a deeper understanding of the brain and its operational principles, others may arise from advances in technology). To date however, the most commonly-used techniques can be coarsely subdivided in the following classes (see Fig. 6):[2]

- **Machine Learning** (ML) is the scientific study of algorithms and statistical models that computer systems use to **perform a specific task without using explicit instructions**, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task [Wikipedia]. Examples of popular ML approaches include regression and support-vector machines.

  One prominent ML approach is the field of Bayesian Machine Learning. Bayesian statistics is a branch of statistics where quantities of interest are treated as random variables, and one draws conclusions by analyzing the posterior distribution over these quantities given the observed data. While the core ideas are decades or even centuries old, Bayesian ideas have had a big impact in machine learning in the past 20 years or so, because of the flexibility they provide in building structured models of real-world phenomena, their ability to train on small datasets, their capability to bring in expert knowledge and their robustness to missing or faulted observations. *[Definition lifted from https://metacademy.org/roadmaps/rgrosse/bayesian_machine_learning].*

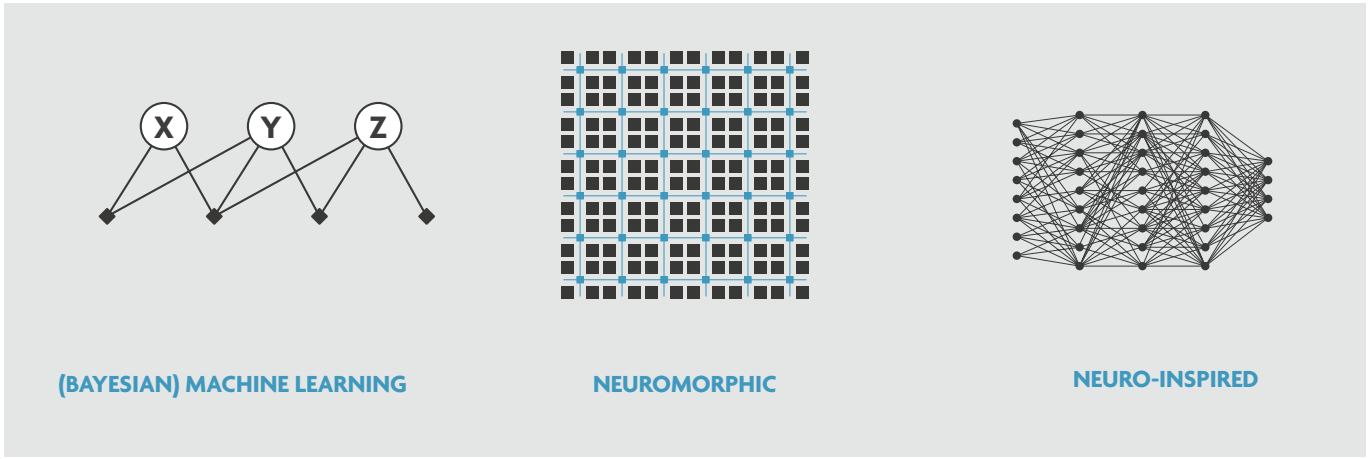


Fig. 6: Artificial intelligence encompasses a plurality of tools and methods.

- **Neuro- or brain-inspired computing** refers to computational **models and methods that are based on abstractions and models of the perceived mechanisms and topologies of the brain**. The goal is to enable the machine to realize various cognitive abilities and coordination mechanisms of human beings in a brain-inspired manner. One can argue that this approach is a subset of machine learning, as programs are constructed from the observation of data. At the same time, the construction and the interpretation of the models is fundamentally different, hence meriting them a sub-class on their own.

  A prominent example of this class of AI is the field of artificial neural nets (ANN), with deep learning networks as its most prominent representative. The latter includes architectures such as deep belief networks, convolutional neural nets and recurrent neural nets. These approaches have been immensely successful over a broad range of specific tasks that range from playing games such as Go over facial recognition to autonomous driving. It builds around an abstract model of a "neuron", an interconnect topology, and overlaying learning and inference mechanisms. It is fair to state the true breakthrough of ANNs was triggered with the availability of huge data sets and fast parallel computing platforms.

  Other examples of neuro-inspired computing approaches include high-dimensional computing (HDC), which uses random patterns in high-dimensional spaces (as inspired by the operations in the cerebellum) to perform a broad range of classification, recognition and reasoning tasks, holographic computing and sparse distributed memory [Kan09, Rah19]. A close relative is the domain of reservoir computing, where a largely untrained part of a neural net acts as a 'reservoir' of dimensionality that makes classification by means of separating hyperplanes easer by virtue of the added dimensions. There is no question that further advances in computational neuroscience will lead to novel schemes, some of which may be very attractive to operation at the edge. Reflect, for instance, about the in-sensor processing happening in our auditory, olfactory, vision, tactile and proprioceptive sensory paths, or the machinery that keeps our human body optimally tuned over a broad range of operational conditions.

imec

- **Neuromorphic computing** is a concept developed by Carver Mead in the late 1980s [Mea90], describing the use of very-large-scale integration (VLSI) systems containing **electronic analog circuits to mimic neuro-biological architectures present in the nervous system**. In recent times, the term neuromorphic has been used to describe analog, digital, mixed-mode analog/digital VLSI, and software systems that implement models of neural systems (for perception, motor control, or multisensory integration) [Wikipedia]. While it technically belongs to the domains of machine learning and neuro-inspired computing, its inspiration is to build physical computing systems that mimic the operation of the brain in a bottom-up fashion. As such, it presents more of a computing architecture than a computational model. Prominent examples of commercial implementations of neuromorphic computers are the IBM TrueNorth processor [Mod14] and the Intel Loihi processor [Loi17]. The Spiking Neural Net (SNN) is one class of neuromorphic NN that has received a lot of attention. Its event-driven executional model makes it particularly attractive for low-energy realizations [Bal18].

Each of the approaches mentioned above comes with pros and cons. For instance, some Bayesian ML approaches work well when an appropriate model can be created. On the other hand, deep neural nets tend to be flexible and rapidly deployable, yet are complex (sometimes requiring 100+ layers and millions of weights) and training (learning) is expensive. Neuromorphic approaches, while being most amenable to advances in implementation technology, are bottom-up and often suffer from a lack of a compelling computational model. Other ML methods overcome these issues by easily integrating expert knowledge, yet, they suffer from reduced task accuracy or require more input regarding model definition. It is our belief that "automated decisioning systems" will combine various flavors of machine learning or hybrid combinations thereof (in combination with some traditional instruction-set processors). The first examples of such are emerging. For instance, the field of Bayesian Deep learning combines ANNs and Bayesian architectures. **The myriad of choices at all levels of the design hierarchy speak to the need for an exploration environment that (i) supports objective comparison the effectiveness and efficiency of the various approaches in light of advanced technology, device and circuit options, and (ii) provides the capability to look forward into the future.** This all hinges on the availability of a number of fair metrics that support an apple-to-apple comparison, as we discuss in the next section.

One thing is quite for certain however – the tradeoffs at the edge will lead to vastly different choices in terms of strategy, architectures and technologies used compared to what we are observing in the Cloud.

The metrics

Building a roadmap requires the identification of a number of figures of merit, and models or projections on how these will improve over time. The technology roadmap for semiconductors (ITRS) was a perfect example of this, as it featured from the very beginning a number of easily trackable metrics, such as: number of transistors per chip, cost per transistor, clock speed, number of operations per second, energy per operation, memory density, etc. When combining these with specific computational (e.g. digital logic) and architectural models (e.g. the instruction set processor), more specific metrics can be defined such as the number of instructions per second or the energy per instruction.

With the advent of innovative AI platforms, coming up with insightful metrics may not be that simple[3]. Plots showing trade-off between performance (in GOP/sec) versus power, such as shown in Fig. 7, only make sense when it is possible to define what an equivalent operation means over vastly different implementation platforms. In addition, the type of function being implemented and consequently the definition of system performance may vary substantially over the implementation alternatives. Finally, also the flexibility of the platform as defined by its ability to map a suite of AI algorithms, might differ, yet has a tremendous impact.

All these arguments indicate that a different approach needs to be taken if one wants to effectively measure, compare and project the evolution of various AI implementation strategies.

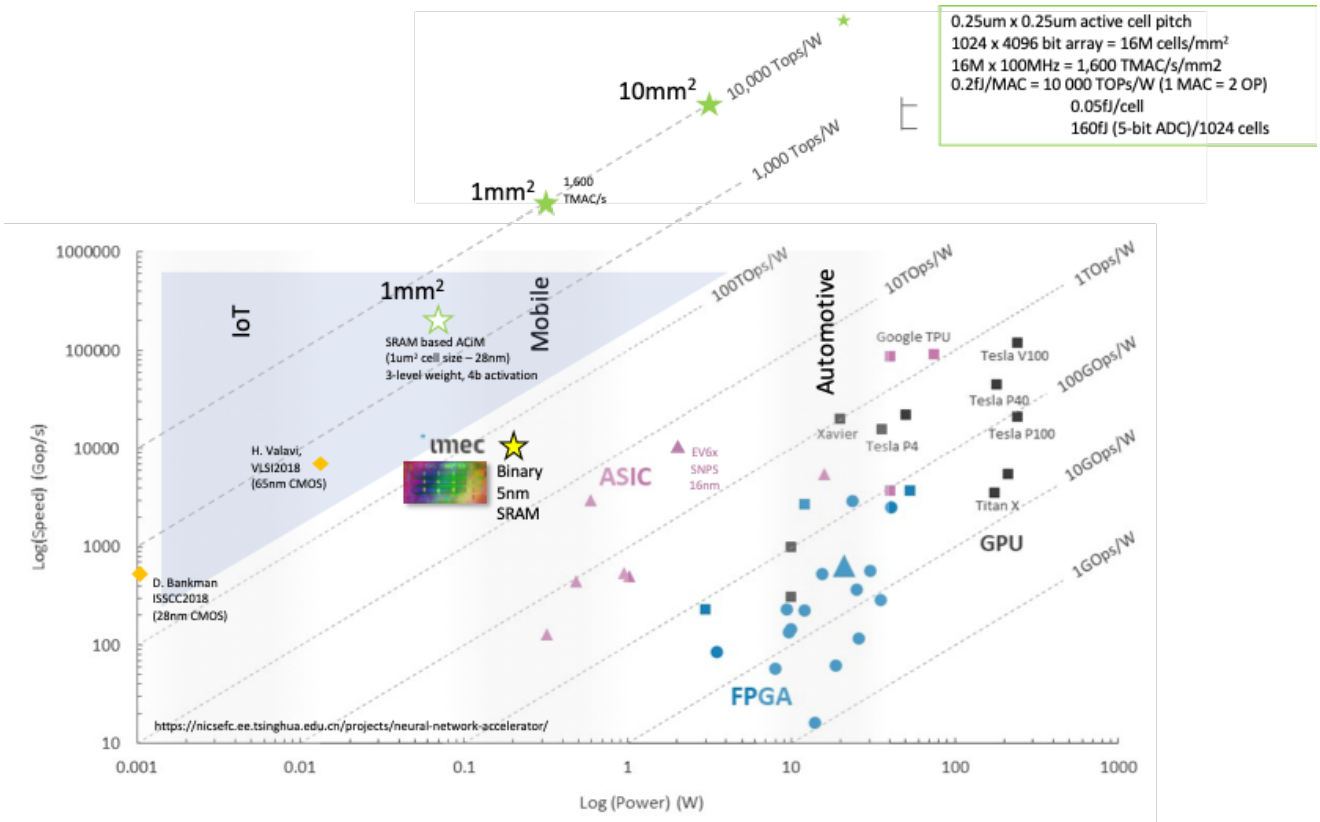


Fig. 7: Energy-performance trade-off for various AI implementation platforms [adopted from Bei18].

The most reasonable alternative is to use a system-level approach, in which an operation is defined as a much larger function or a combination of functions (such as a classification or a recognition task) **and the metrics of interest are functions/sec and energy/function for a given set of specifications and constraints**. Other metrics can be volumetric cost per function (area, size). This approach allows for a fair comparison and enables exploration and forward projection. At the same time, it has some pitfalls, as it may lead to the need of a large and very diverse set of benchmarking functions to be defined, if one wants to cover the whole space of AI. Another danger is the temptation to over-tune an implementation so that it excels on one or more of the benchmark functions (and just those), resulting in a lack of flexibility.

There is no question that this topic needs a lot more reflection and thought. In the meantime, a heat-map identifying clear opportunities and their possible impact rather than a roadmap may be the right target to pursue. Also treating all applications under the same banner may not be the right approach. One possibility is to define a set of "power classes", defined by the amount of energy/power available at an independent node. This influences the functionality that can be implemented for that node, and ultimately will drive the technology choices for its realization. At the edge, power classes could range from the autonomous vehicle (~10Watts) over the mobile (100mWatt-1Watt) to the IoT node (10's of mWatts), the wearable (mWatts) and the implantable (μWatts).

One additional insight is that under certain conditions it is possible to formulate absolute upper or lower bounds for some metrics. For instance, the flow of information is bounded by Shannon's law, independent from the medium, which could be a wire on an integrated circuit or an axon in the brain. Building on this, lower bounds on energy consumption have been defined for simple digital operations and analog functions such as A/D conversion. While currents solutions may be quite afar from those bounds, they present a sense of how efficient a solution is, and how much room for improvement exists (see, for instance, [Mur13]). Another interesting metric for "artificial" solutions is their effectiveness when compared to the solutions provided in nature.

"A heat-map identifying clear opportunities and their possible impact rather than a roadmap may be the right target to pursue."

The technology opportunities

AI functions differ in many important ways from the traditional algorithms inspired by the Von Neuman programming approach. This inevitably translates into different implementation needs, and provides opportunities from an architecture perspective. Just to name a few of the key differences: a learning-based versus a stored program model; the roles of long and short-term memory; the intertwining or separation of logic and memory; the density and the topology of the interconnects; the massive amount of concurrency available; and the statistical nature of the computation. These differences can have a profound impact on how the essential building blocks are realized, and the choice of the underlying circuit and device technologies. This divergence is even more outspoken at the edge where volume, energy and cost of implementation matter the most, and the matching between computational function and architecture is of greater importance. It is fair to state that AI edge applications will become an essential driver for the development of the next-generation hardware implementation technologies and the large-scale manufacturing to support it.

Based on an evaluation of the technology landscape, we believe that the following developments have potentially the largest impact.

3D integration and interconnect

A vast majority of AI functions favor a tight integration between logic and memory providing wide interface lanes and low latency. This immediately invokes a vision of some form of 3D integration with memory stacked on logic supported by dense vertical connectivity[4]. This is part of an ongoing trend where 3D integration allows for increased improvement and addition of functionality, even in a context where horizontal '2D' scaling is slowing down [Itr15]. **In our opinion, this is one of the technology innovations that has the highest potential for major impact on the efficiency and footprint of Edge AI going forward**. It requires a radical rethinking of the topology of both logic and memory and the interconnections between them. Fig. 8 presents a pictorial overview of the broad range of packaging and integration approaches that have become possible and/or available over the past years. They differ widely in terms of scalability, density of interconnect, reliability and cost. Which approach makes sense really depends upon the intended application and the associated cost. Note that 3D integration not only applies to logic-memory integration, but also enables interconnect topologies that were exceedingly hard to accomplish before, or were just extremely inefficient. As an example, consider the very high fan-in requirements underlaying some neuromorphic architectures. Over and beyond, 3D integration also impacts the interface between sensing and computation, which also is crucial for many Edge applications. For example, it allows high-density imagers to be equipped with build-in processing and intelligence.

"A tight integration between logic and memory will have a major impact on the efficiency and footprint of Edge AI."

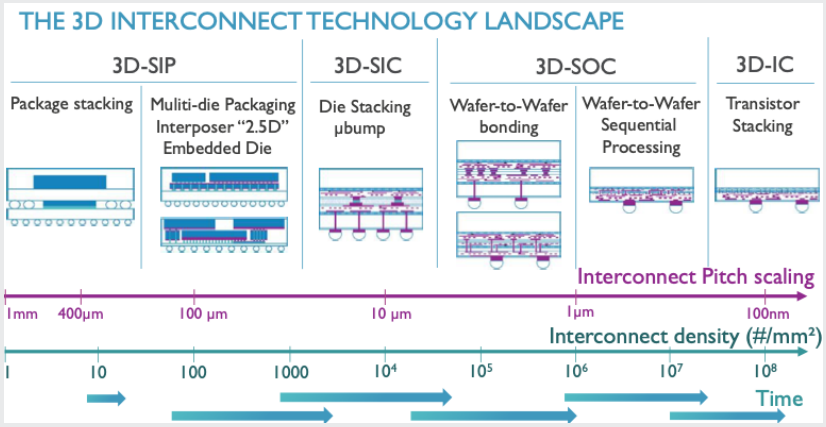


Fig. 8: Evolution in interconnect and packaging strategies [Bey19]

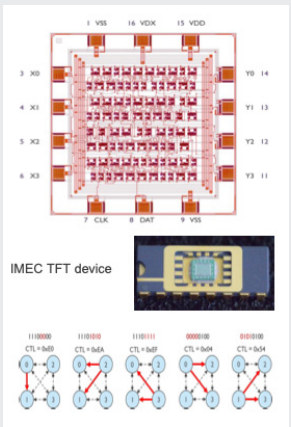


Fig. 9: Thin-film reconfigurable interconnect

Beyond pure integration, these technologies also open the door for options that were extremely hard to accomplish in the traditional CMOS processes. For instance, consider the possibility of dynamically reconfigurable interconnects, in which high-quality switching devices with high R_{off} and low R_{on} are embedded into the interconnect fabric. These switches could be controlled by non-volatile memory placed next or directly above or below. Thin film electronics provide a technology that enables such functionality already today, as was illustrated on a small scale in the IMEC NeurRAM3 prototype [Bal18] (see Fig. 9). Ultimately, one can even think further ahead, and imagine intelligent wiring structures where the resistivity varies dynamically in response to the data flowing through it.

While the opportunities are seemingly unbounded, one of the main challenges in this road-mapping effort is to identify which approaches provide sufficient gains to offset the increased costs and under what conditions or constraints. This clearly points to the need for a system exploration environment that allows for a trade-off analysis supported by detailed modeling and precision adequate to give meaningful answers.

Memory

There is no question that memory plays a central role in the AI roadmap. Here, its roles and needs vary vastly from what we are familiar with in the domain of traditional Von Neuman-based instruction-set processors. There, memory access is mostly governed by the fetching and storing of instructions and data in mostly irregular patterns. In the deep learning network arena, the number of instructions per function is limited. The main role of memory is to fetch weights and data in vast amounts structured in highly regular patterns, and to synchronize these flows of data. This means that memory controllers need to be rethought. But even more so, a **re-organization of the memory architecture and topology** is necessary. Many AI architectures benefit from or even require distributed – high bandwidth – low latency access. This can, for instance, be enabled by providing wide data-word interfaces, or by changing the modular structure of the memory. Some AI applications prosper from the availability of huge data sets close to the processing elements – think about an autonomous car having to navigate different scenarios in quick succession. Here **density** matters. This could be addressed by a 3D organization as discussed earlier.

"Memory plays a central role in the AI roadmap. Its roles and needs vary vastly from traditional Von Neuman-based instruction-set processors."

Depending upon the chosen AI approach and implementation strategy, some more aggressive memory options may become attractive:

- **Non-volatility** is a desirable feature, especially in the case of quasi-stationary reconfigurable structures for always-on systems at the Edge that are mostly read and only occasionally rewritten. The major attraction is the low amount of leakage for latent memory. While of somewhat lower priority right now, it may become one of the most salient features in the longer term, providing the capability of "long term" reconfiguration at very low energy cost. As shown in Fig. 10, many non-volatile memory options are available with different requirements and constraints in terms of operational voltages, write and read times and resiliency.

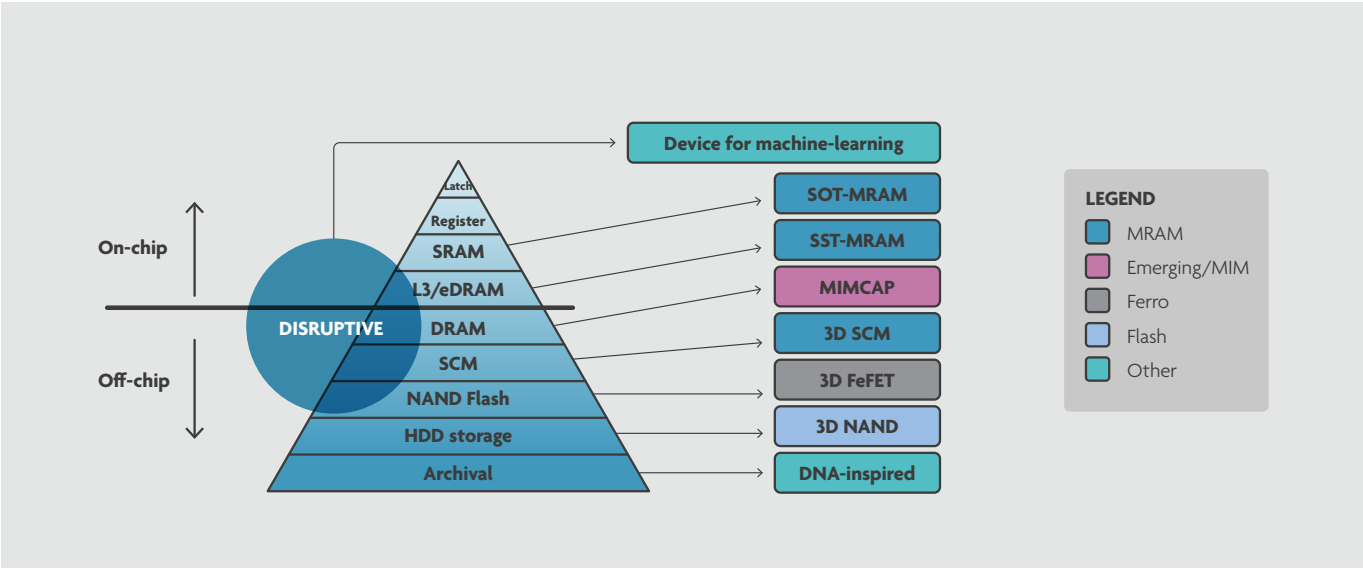


Fig. 10: imec memory and storage landscape [Fur18]

- One of the prominent features of the brain is the intertwining of memory and logic. It therefore makes perfect sense for researchers to explore if this model also translates to physical computing platforms. Called "**in-memory computing**", functions like matrix multiplication, associative search and look-up tables are distributed throughout the memory (which could be SRAM or any type of emerging non-volatile memory). Main advantages are the reduction in data movement (data is local) as well as locality of operation. Realizations vary from the usage of non-volatile memories to store weights to analog distributed computation. The gains of the in-memory approach are potentially huge; however, it comes with challenges with respect to flexibility, programmability, accuracy, variability and scalability that cannot be ignored. Notwithstanding these concerns, research groups and start-up companies have been exploring a broad range of options, some of which may see fruition in the near future. A somewhat less adventurous approach that preserves some of the benefits but requires less change, is the "**near-memory**" approach, in which logic functions are implemented in the periphery, or in-between blocks of memory. This approach connects well with 3D integration and stacking.

- To address the dynamically changing performance requirements (in terms of fault rates, and/or accuracy), a broad range of operating supply voltages may be needed (from 0.3V to 1V)

"One of the prominent features of the brain is the intertwining of memory and logic. Can this model also be translated to physical computing platforms?"

Logic

Most of the hardware accelerators for AI advocated today focus on the efficient realization of "some" abstraction of the neuron function,

$$y = f\left(\sum_{i=1}^{N} w_i \times x_i\right)$$

with N typically being a large number and f a sufficiently non-linear function. A collection of such neurons can be used to represent arbitrary functions (universal approximation theorem of neural nets). Many different approaches towards implementing the neuron function have been proposed, including digital neural processors, large array-vector multipliers, in-memory computing, analog multiplication, memristors, etc.

However, **the** "**neuron**" is not the only logical element that should be pursued in the search of efficient edge AI. **Other components may be equally interesting, including "memristors", "synapse"-like devices, networks of non-linear oscillators (analog and digital), tunable delay lines, and other logical functions**.

In light of the fact that virtually all AI computation is based on statistics, special attention should be devoted to probabilistic and stochastic hardware, and functions/architectures that support statistical behavior [Sha19]. The compelling advantage of this approach is that computation could be performed at a much lower SNR, and hence be a lot more efficient (note: a typical brain synapse operates at an SNR of ~1!). Any such approach, be it based on statistical devices, analog hardware or random data representations such as occur in high-dimensional computing, must support implicit techniques to control the impact of variance on the system performance. These extend well beyond the familiar approaches used in the deterministic computing paradigms that have dominated the traditional semiconductor devices to date.

In this context, materials and devices that have the potential to offer, by their intrinsic physics, efficient realizations of stochastic functions become particularly interesting. Examples include: (i) insulator-metal-transition (IMT) Vanadium Dioxide, in which the inherent physical noise in the dynamics of switching dynamics provides the foundations for building FitzHugh-Nagumo (FHN) neurons with thermal noise along with threshold fluctuations as precursors of bifurcation and ferroelectric doped-high-k materials, and, (ii) doped high-k ferroelectrics exploiting the particular nucleation-limited switching kinetics of the ferroelectrics to emulate Fe-FETs neuron-like integrate-firing activity.

Alternative technologies

Efficient realization of learning-based functions is not restricted to the electronic domain. After all, most of the inspiration in the world of machine learning come from biological systems that exploit an intriguing mixture of chemical and electrical mechanisms. Hence, other technologies could be a part of the solution. For instance, functions such as matrix multiplications and convolutional neural nets can be implemented using light as the prime carrier, inspiring the new field of neuromorphic photonics [Bie19]. Developments in silicon photonics, novel photonic materials and computational imaging could also help advance other computational models such as reservoir computing, discussed earlier. An example of such an all-optical approach is for instance presented by a company called LightOn [Lig19], which uses optical dispersion and speckle patterns as means of mapping data into high-dimensional spaces. Other carriers of data and computation such as magnetics and potentially, in the medium-to-long-term, quantum systems should not be ignored. Finally, the field of synthetic biology is definitely something to keep an eye on. This is especially true now in the age of CRISPR [Cri19], in which active genome editing has become readily available and is enabling profound and controlled change of cellular and organism level functionality. It opens the door for the engineering of biological computational systems, while at the same time offering a deeper understanding of the underlying operational mechanisms.

"Developments in silicon photonics, novel photonic materials and computational imaging could also help advance other computational models such as reservoir computing."

Sensors and actuators

One possible abstraction of edge devices is to consider them as intelligent sensory/actuator systems. This abstraction holds for IoT and AR/VR devices, autonomous mobile entities, as well as wearables and implantables. From that perspective, sensing and its associated processing can be considered as being symbiotic. This sensor-centric perspective is often called "in-sensor computing" [Son15, Oza18]. A prime example thereof in nature can be found in the retina, where the neurons directly connected to the optical sensors perform various forms of feature extraction before sending the information over the optical nerve bundle. Different parts of our peripheral nervous system, such as the olfactory system, show similar hierarchical levels of processing, combining local in- or near-sensor processing of small patterns with global, brain-based processing of the bigger picture. Note, however, that most sensor implementations require technologies that differ substantially from the CMOS technologies used for signal acquisition and computation. As such, tight integration of the two is non-trivial, yet required, and hence is the subject of active ongoing research. A number of options are enticing, such as the realization of both sensing and front-end AI in flexible electronics (e.g. thin film or printable electronics) or the tight integration of the two using 3D packaging.

"The realization of both sensing and front-end AI in flexible electronics (e.g. thin film or printable electronics) or the tight integration of the two using 3D packaging are enticing options to be considered."

Hardware-enabled security

One of the main challenges to computing in general and Edge AI in particular is to ensure security, safety and privacy – especially since many of the applications can be considered privacy-sensitive, essential or life-threatening if exploited in a malicious way. The constraints of the edge in terms of energy footprint demand solutions that provide inherent hardware-based security mechanisms such as unique and unchangeable identifiers, authentication techniques, location-awareness, etc. These should be tightly intertwined with the machine learning functional blocks, to realize both functions jointly at minimal resource cost (area, power, memory). As always, security appears at the bottom of the list. Yet, in terms of importance, **it should be on par with other constraints such as efficiency and cost**, and hence deserves far more attention in the academic and the industrial world (A common quote in industry is that "security is something that everyone wants and no one wants to pay for"). In fact, the advent of AI everywhere should be considered as an opportunity. For instance, the novel computational paradigms that underlay many AI realizations are based on statistical representations and random projections that can be made to be unique for every single realization (for instance, by using random process variations or exploiting the randomness of nanoscale devices). As another example, some of the most advanced encryption approaches such as **homomorphic encryption** [Hom19] **are using computational models that are similar to machine-learning systems that use high-dimensional representations, and hence may benefit from implementation developments for the latter**.

The application pull - moonshots to drive development

As stated earlier, building a true roadmap for AI technologies requires an application or a system perspective. Given the diversity of the functionality, goals and constraints of "AI at the Edge" no single application will do. One plausible approach would be to compile a list of meaningful and representative functions or sub-systems to serve as benchmarks. While this is definitely commendable and will happen anyhow, benchmarks tend to be backwards looking, create tunnel vision, and miss the overall system settings and constraints. In fact, they often even stifle innovation.

A more effective approach to foster innovation is to identify one or more long-term ambitious goals and visions as "moonshots". Technologies (at any level of the stack) can then be measured in terms of how they advance the state-of-the-art towards reaching the moonshot goals. At the workshop held in Leuven on September 17, 2019, we selected this approach as being the one that would present the best opportunities to advance the field in a forward looking, yet structured and measurable way. In addition, it was judged that a single moonshot or compelling long-term driver would be far from sufficient to cover the full spectrum of Edge AI applications. As such, we have selected a set of three, each of which unique, exciting and audacious.

1. The introspector: towards future Human Avatars for Healthcare

In 2011, Qualcomm launched the Tricorder competition, targeting the development of an automatic non-invasive health diagnostics system able to autonomously diagnose 13 medical conditions (12 diseases and the 'absence of conditions') in a single portable package that weighs no more than 5 pounds [Wikipedia]. As no single team managed to fulfill all goals, a reduced price was awarded in 2017.

Fast forward to 2030 and assume a network of sensors that could be embedded in our daily living environment, or worn on the body, implanted inside, or traveling through the body. We call it the "introspector system". The streams of information would provide a continuous picture of the health and the state of the human body and help to diagnose early-on potential risks or diseases (metabolic, cardiovascular, mental). It may also track everything you have been exposed to over your complete lifespan (your "exposome"). For privacy reasons, most of the data analysis would be done locally – either next to the sensors or on some wearable device. However, the introspector would also build on the immense amount of data and knowledge available in the Cloud. The combined knowledge of all the avatars provides a model of collective learning. In more advanced versions, feedback in the form of actuation/stimulation/drug delivery can be added and used to perform fast response/correction.

In a sense, the system would serve as **preventive/predictive health avatar**, enabling every citizen access to personalized healthcare, healthy lifestyle and disease prevention. The personalization of medicine presents an approach that is proactive, promotes healthy living, prevents disease, and treats disease with precision, while aiming to provide the right treatment at the right time to the right patient. The Human Avatar vision [Heu18] is based on integrative technological and digital data approaches, combined with ethics and behavioral science. Building a Human Avatar involves combining both AI hardware and software, and

"The most advanced omics, smart sensors, nanomedicine, advanced imaging and body-on-chip technologies are needed to realize true experimental physical organ avatars."

data generation and processing technologies (Fig. 11): it will need the most advanced omics, smart sensors (wearables and implantables), nanomedicine, advanced imaging and body-on-chip and technologies to realize true experimental physical organ avatars. These will serve as representative models at the micro/nano-level, and facilitate deep understanding of the function and interaction of organs and the mechanisms underlying their diseases, and they will be used as test models for disease prevention and drug treatments. It will lead to the design and development of the Edge AI components of a specific data infrastructure and subclass of the Internet of Things called the Internet of Healthcare (IoH). In IoH, it will enable Edge AI embedded security, privacy and ethical rules.

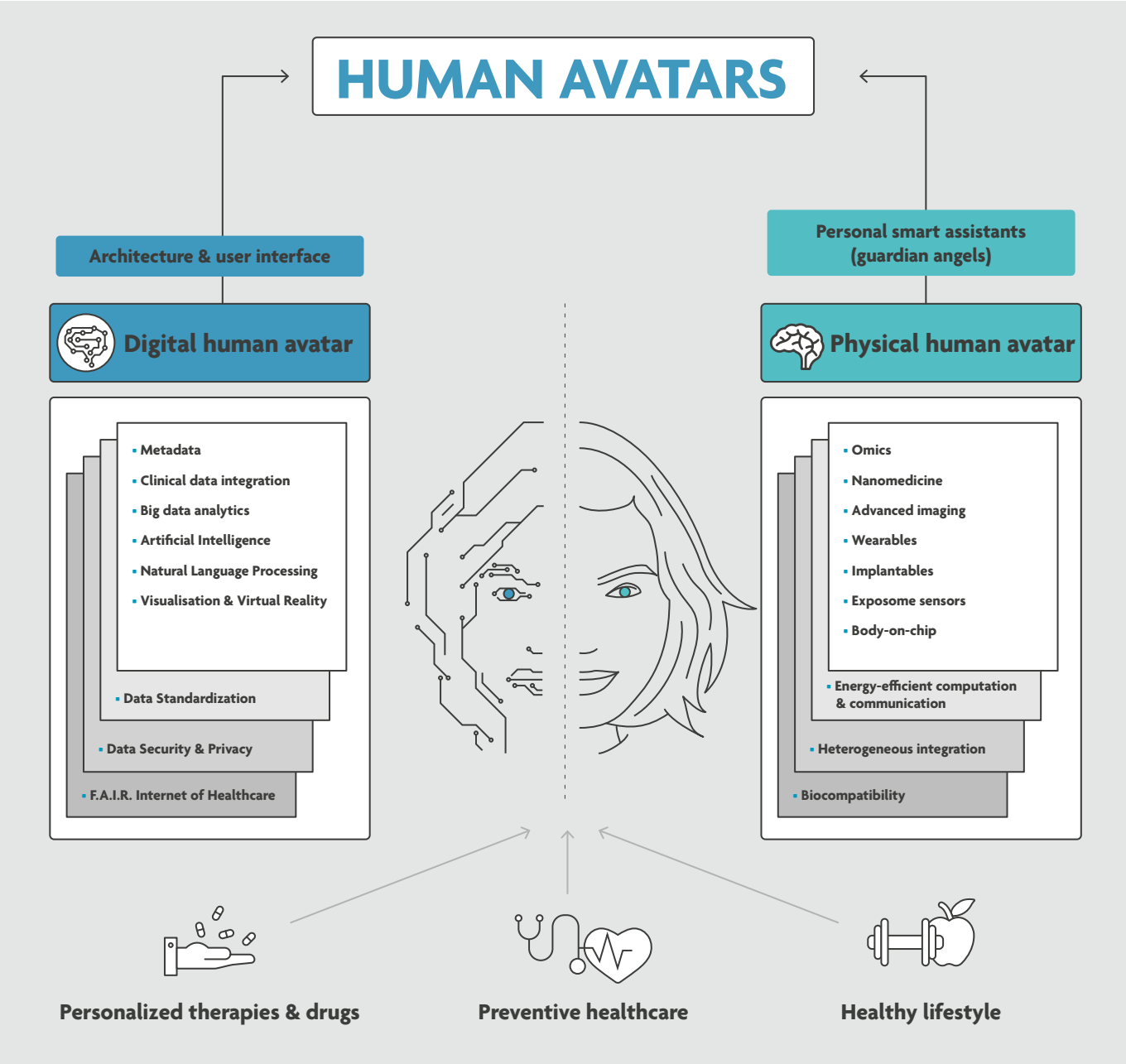


Fig. 11: Human Avatar Platform, as proposed by Health EU, to revolutionize personalized and preventive healthcare, supported by data-generator technology platforms and models based on big data [Source: HEU].

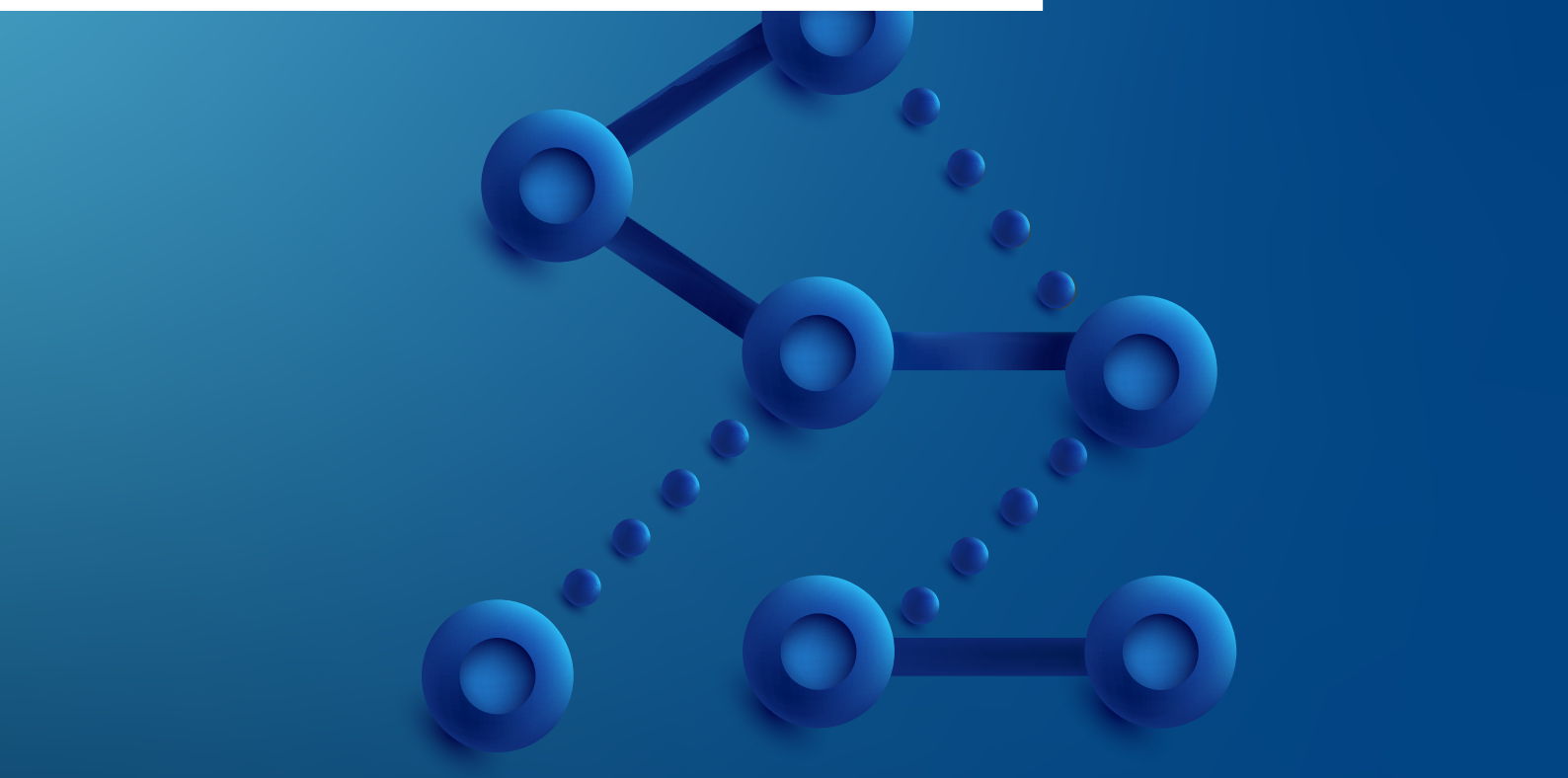
2. The ultimate teacher

From a different outwards-looking perspective, similar networks of sensors and actuators can serve to augment and extend the functions of the human body and fundamentally change how we perceive information from and act on the physical environment around us. Sensory information that extends far beyond what we can conceive today can be made available and transformed so that it can be mapped onto our existing seven senses, couple directly into the brain or the nervous system, or be fed to our digital twin existing in the cyberworld. Similarly, motor output signals can be observed and translated into direct action in the environment (control a car, a drone or an exoskeleton), project imaginary speech, or provide direct stimulation to the body. Prototype systems of this class have been demonstrated or are under exploration today. An example of such are prostheses controlled by a brain-machine interface. AR/VR systems fall in this class as well, but current incarnations still lack the feedback loop.

To leap forward, consider a wearable system that helps you to learn physical or mental tasks in a much faster way by providing instant evaluation of your actions in light of the context you are in, and help you improve on a task by gentle correction. Consider e.g. trying to learn how to play golf. Currently, the feedback loop is very slow – it requires a teacher evaluating your performance and giving verbal feedback. Instead information from sensors in the club, from body-worn motion-tracking sensors and cameras, ball tracking devices, etc. in combination with information extracted from observing similar sensors on expert players, can tell or show you how and where to improve. Even more, a gentle nudge provided using tactile feedback on the arms or hands on the arm can help adjust positioning and swing. In a first instance, the ultimate teacher could be a smart AR/VR device coupling into a set of sensors/devices worn on the body or embedded in the environment. In later stages, more and more actuators and even brain implants can be added. This is just an example of how we could use edge AI to improve on how we as humans deal with novelty. Many other options can be envisioned: helping surgeons during complicated surgeries, learning how to control a really complex machine (or a group of them working in concert), or just learning to play the piano.

In the end, this may have a profound impact on how we learn and acquire skills in the future, and even, at a deeper level, help us understand how we learn in the first place.

"Consider a wearable system that helps you to learn physical or mental tasks in a much faster way by providing instant evaluation of your actions in light of the context you are in. "



3. Swarms on a mission

Imagine a group (swarm) of autonomous drones working actively together to fulfill a given mission or to meet a certain objective. This could be the locating of survivors after a building has collapsed, the mapping of a fast-moving wild fire and the notification of people that may be impacted, the relocation of a number of objects been a selection of source and destinations in a warehouse, etc. While all these tasks are in a sense possible today, all mission planning and task allocation is done in a centralized way, which creates a bottleneck and a single point of failure. An alternative is to have the mobiles work together to realize decentralized decision making and optimize the task distribution based on local observations and sharing of information. This is a perfect example of the ADD functionality with the extra dimension of information sharing between the nodes (addressing questions such as how much information should be shared).

This driver addresses a broad range of interesting AI topics, such as dealing with heterogeneous sensors, adaptation to changing circumstances and environments, trajectory planning based on incomplete information, and informed decision making.

"Have the mobiles work together to realize decentralized decision making and optimize the task distribution based on local observations and sharing of information."

The need for an exploration methodology

As hopefully became apparent over the course of this document, enabling Edge AI to acquire full "Automatic Decisioning System" capabilities, and to potentially come close to human-like functionality, will require progress on many fronts, especially in light of the stringent energy, size, robustness and security constraints. It is also obvious that reaching the stated goals requires innovation and optimization over all the layers of the stack: from the application over the compiler and scheduler, the architecture to the circuit, device and integration levels. More than ever, technology choices will be dominated by system-level considerations and vice versa. The number of options at all levels is large, and experimenting with all of them is terribly expensive in both cost and time. The fact that the answer will most likely be a combination of heterogeneous sub-modules with different flavors does not simplify the design task either.

Hence the need for an environment that supports both early and detailed exploration, and provides insights in what options in the global design space make sense. Back-of-the-envelope modeling is one option that is often used, but that in general provides unreliable and optimistic estimates, as it ignores parasitics, variations and imperfections. Therefore, **an end-to-end framework is required that enables a co-exploration and co-optimization from the system to the device and integration levels with the realism of actual implementation.**

It is based on experimentally-validated realistic models of the various components and integration (packaging) technologies. Using built-in generators and extractors, detailed models of variations and imperfections can be derived. Starting from high-level templates of computational architectures and algorithmic dataflows, the framework would provide tools to map algorithms into architectures and designs. The combination of all these tools in a common framework would allow for estimation, simulation, and parametric and sensitivity analysis of the various options available, hence providing full exploration and what-if functionality. Observe that various attempts were made over the past decades to develop a system-level exploration framework of this nature, but most of them failed for a number of reasons such as an over-constrained system model, lack of the right translational tools or inadequate hardware models. More recently though, the feasibility and prospective benefits of true system-technology co-optimization environments were effectively demonstrated for some constrained architectural platforms such as a single [Kom18] or multicore processor [Aly19] (Fig. 12).

It is clear that no single player on her/his own can provide the full knowledge and skill set needed for the creation of such a framework. It hence requires a consortium of players to come together, including application experts, architects, circuit specialists, technologists and design tool experts. It is up to the community to help create the libraries and models for meaningful building blocks, interconnect approaches, architectural models and block generators, system-level descriptors and compilers. However, if successful the availability of such an exploration environment could help tremendously in expediting innovation and creativity, leading to entirely new computing platforms and a truly cognitive edge.

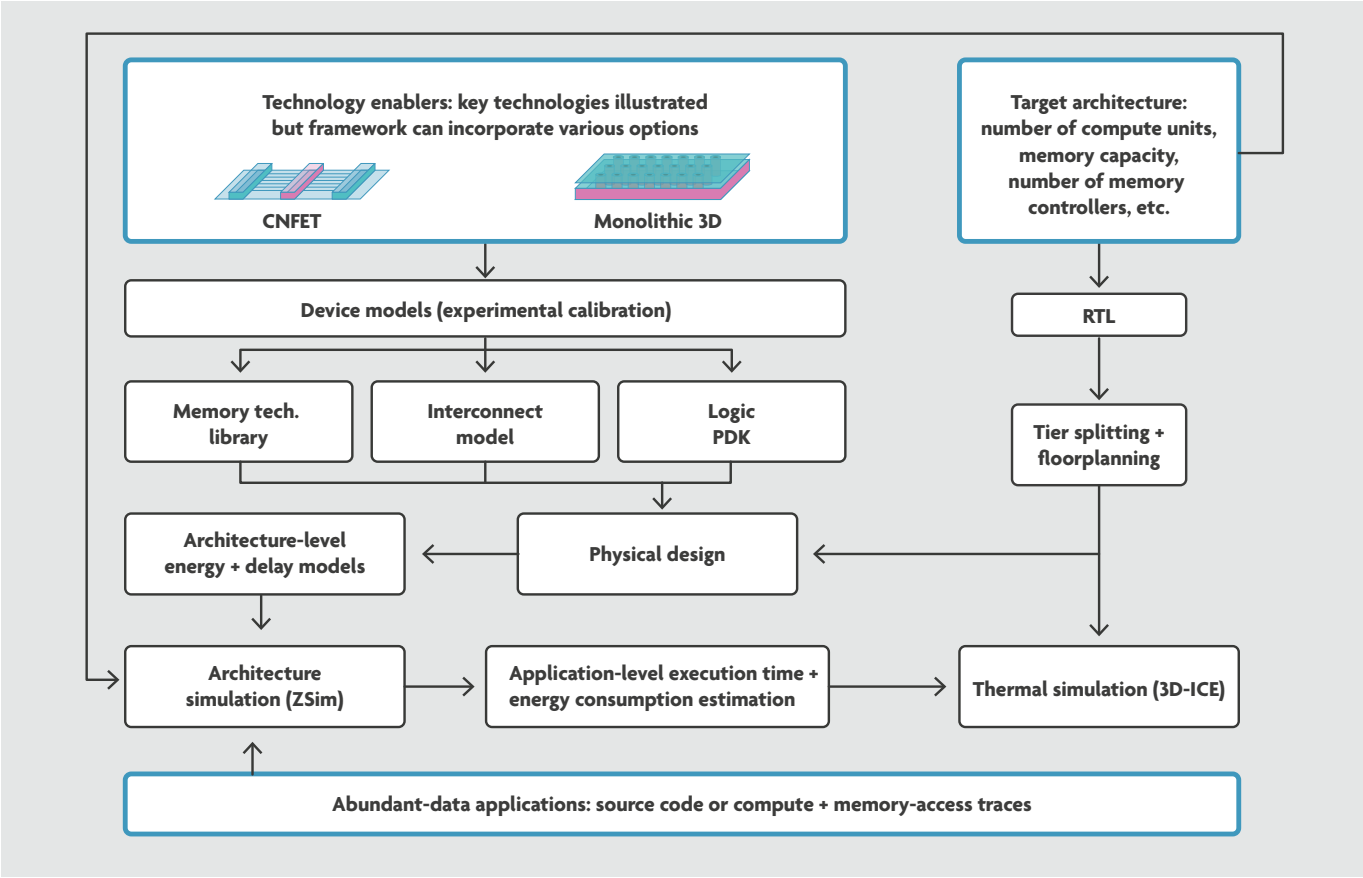


Fig 12: N3XT system-technology exploration environment for multicore processor systems [Aly19]

Recommendations

Over the course of this paper, we have enumerated a set of alternatives and options that can lead to realization of true Edge AI over the coming decade(s) – with the aim of mimicking, augmenting, and potentially even rivaling the capabilities and the prowess of human intelligence. Making this vision come to fruition will require a set of concrete actions and concerted efforts from a broad team of players, representing all parts of the equation. Given the outlandish size of the worldwide competition, the **exploitation of unique strengths** – such as the ready access to a broad range of beyond-state-of-the-art technologies – is of essence. There is no need to try to compete in domains where other players already have a sizable lead. This will however require the taking of **substantial risks** and the **making of informed bets** – just chasing all possible options is an expensive and non-rewarding exercise.

To that effect, we recommend that the following actions be undertaken most specifically by the players that contributed to this white paper, but with outreach to the more global community if needed. A set of concrete milestones needs be defined for each of these.

- Plan for various partners to engage in the **moonshots**. This entails the concrete definition of the overall functionality to be achieved, the identification of the components and the possible solution space, the selection and/or collection of meaningful data sets, and the **outlining of possible prototypes to be built**.

- Definition of a meaningful set of **metrics of relevance to Edge AI (with major focus on energy efficiency and sustainability)**. This will require the selection of a set of representative functions/benchmarks, some of which can be collected from the public domain, others from interested partners. Most preferably, though, they should emerge from or be part of the identified moonshots.

- Develop plan and roadmap for **at-scale availability of the technologies** identified in this white paper. This would include the development of detailed models as well as parameterizable generators to be used in the exploration exercise.

- Support a **continuous evaluation of emerging AI and ML** approaches and concepts (and alternative approaches as may arise over time). This is especially of importance for Edge AI, where efficiency and compactness if of essence. We are still some sizable distance away from what biology can do. This effort would likely be performed in collaboration with universities.

- Develop **exploration methodology** to evaluate joint impact of systems and technologies. This is a substantial effort that will engage a broad selection of partners.

- Develop a strategy to provide **security, safety and privacy within Edge AI**, and how technology innovation can play a role in this. Similarly, packaging can play a role in the creation of secure hardware and sovereignty.

Recommended background reading

1. Ion Stoica et al, A Berkeley View of Systems Challenges for AI, https://arxiv.org/pdf/1712.05855.pdf (2019)
2. Y. Gil and B. Selman, A 20-Year Community Roadmap for Artificial Intelligence Research in the US, Computing Community Consortium (CCC), https://arxiv.org/abs/1908.02624 (August 2019)
3. Beijing Innovation Center for Future Chips, White Paper on AI Chip Technologies, 2018
4. Many Authors, Artificial Intelligence Research Flanders, 2019
5. Y. LeCun, Deep Learning Hardware: Past, Present and Future, Keynote ISSCC 2019.
6. HJ Yoo, Intelligence on Silicon: From DNN Accelerators to Brain Mimicking AI-SOCs, Keynote ISSCC 2019.

End notes

1 This number is an estimate. It is based on estimations of computational equivalence of brain functions that are relatively well understood (Kur05]), and their energy footprint. Hence it should be interpreted cautiously. The reality may actually be worse.

2 This overview is by no means exhaustive, and covers only the most prominent techniques currently in use.

3 This is not a new observation. Creating roadmaps has already become a lot more complicated with the diversification of the application domains addressed by semiconductor technology. The times that everything could be measured by instruction processors and their memory hierarchy have been over for at least 15 years. AI just adds another trace.

4 After all, the brain with its dense interconnect networks and its merged memory/logic structure has long chosen a 3D integration approach.

References

- [ABI19] ABI Research
- [AI19] A.M. Ionescu, Energy efficient computing and sensing in the Zettabyte era: From silicon to the cloud, 2017 IEEE International Electron Devices Meeting (IEDM), 1.2. 1-1.2. 8, 2017.
- [Aly19] S. Aly et al, The N3XT Approach to Energy-Efficient Abundant-Data Computing, IEEE Proceedings, Jan 2019.
- [Ark19] (https://ark-invest.com/research/the-ai-chip-landscape-in-2019
- [Bal18] Balaji, Adarsha; Corradi, Federico; Das, Anup; Pande, Sandeep; Schaafsma, Siebren; Catthoor, Francky, "Power-Accuracy Trade-Offs for Heartbeat Classification on Neural Networks Hardware," Journal of Low Power Electronics, Volume 14, Number 4, December 2018, pp. 508-519(12)
- [Ban18] D Bankman, L Yang, B Moons, M Verhelst, B Murmann, " An Always-On 3.8 J/86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in 28-nm CMOS," IEEE Journal of Solid-State Circuits 54 (1), 158-172.
- [Bei18] Beijing Innovation Center for Future Chips, White Paper on AI Chip Technologies, 2018
- [Ber19] Ion Stoica et al, A Berkeley View of Systems Challenges for AI, https://arxiv.org/pdf/1712.05855.pdf (2019)
- [Bey19] E. Beyne, Evolution in interconnect and packaging strategies, IMEC ITF.
- [Bie19] P. Bienstman et al, Neuromorphic information processing using silicon photonics, Proc. SPIE 11081, Active Photonic Platforms XI, 110811I (5 September 2019).
- [CBI18] https://www.cbinsights.com/research/report/venture-capital-q4-2018/
- [Cri19] CRISPR, https://en.wikipedia.org/wiki/CRISPR.
- [Eet19] EETimes, "AI Chip Market to More than Double in 5 Years", https://www.eetimes.com/document.asp?doc_id=1335096, 2019.
- [Fur18] A. Furnemont, Imec memory and storage landscape, ITF 2019.
- [Heu18] Health EU – Human avatars to prevent and cure diseases, https://www.health-eu.eu/
- [Hom19] Homomorphic encryption, https://en.wikipedia.org/wiki/Homomorphic_encryption.
- [Ins18] CBI Insights, https://www.cbinsights.com/research/report/venture-capital-q4-2018/
- [Itr15] 2015 ITRS Roadmap, http://www.itrs2.net/itrs-reports.html
- [Kan09] P. Kanerva, Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, Cogn. Comput. 1: 139 (2009)
- [Ket19] M. De Ketelaere, TinyAI Program Plan, Internal Document IMEC 2019
- [Kom18] M. Komalan et al, Main memory organization trade-offs with DRAM and STT-MRAM options based on gem5-NVMain simulation frameworks, DATE, March 2018.
- [Kur05] R. Kutzweil, "The singularity is near," Viking Books, 2005.
- [Lec19a] Y. LeCun, The power and limits of Deep Learning, YouTube, https://www.youtube.com/watch?v=zikdDOzOpxY, 2019
- [Lec19b] Y. LeCun, Deep Learning Hardware: Past, Present and Future, ISSCC 2019.
- [Lig19] Lighton.ai
- [Loi17] "Intel unveils Loihi neuromorphic chip, chases IBM in artificial brains". October 17, 2017. AITrends.com
- [Mea90] Mead, Carver (1990). "Neuromorphic electronic systems". Proceedings of the IEEE. 78 (10): 1629–1636. doi:10.1109/5.58356.
- [Mod14] Dharmendra Modha (interview), "A computer that thinks", New Scientist 8 November 2014, Pages 28-29.
- [Moo17] B Moons, R Uytterhoeven, W Dehaene, M Verhelst, "Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," Solid-State Circuits Conference (ISSCC), 2017 IEEE International, 246-247.
- [Mur13] B. Murmann, "Energy Limits in A/D Converters, ", 2013 IEEE Faible Tension Consommation Conference.
- [Oza18] M. Ozatay, L. Aygun, H. Jia, P. Kumar, Y. Mehlman, C. Wu, S. Wagner, J. C. Sturm, and N. Verma, "Artificial Intelligence Meets Large-Scale Sensing "
- [Rah19] A. Rahimi et al, Efficient Biosignal Processing Using Hyperdimensional Computing: Network Templates for Combined Learning and Classification of ExG Signals, IEEE Proceedings, Jan 2019.
- [Sha19] NR Shanbhag, N Verma, Y Kim, AD Patil, LR Varshney, Shannon-inspired statistical computing for the nanoscale era," Proceedings of the IEEE 107 (1), 90-10.
- [Son15] J. Rabaey et al, Sonics Center Overview, September 2015.
- [Tra19] Tractica – tractica.com
- [Yoo19] H.J. Yoo, Intelligence on Silicon: From DNN Accelerators to Brain Mimicking AI-SoCs, Keynote Presentation, ISSCC 2019.