Energy efficient computing and sensing in the Zettabyte era: from silicon to the cloud

Adrian M. Ionescu

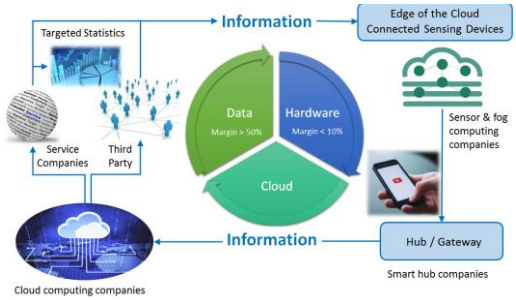
[1]Nanolab, Ecole Polytechnique Fédérale de Lausanne, Switzerland, email: adrian.ionescu@epfl.ch

***Abstract*— In this paper we will present and discuss some of the great research challenges and opportunities related to 21st century energy efficient computing and sensing devices and systems, in the context of the Internet of Things (IoT) revolution. In the future, major innovations will require holistic approaches encompassing silicon and cloud technologies and will be centered on big/abundant data and context. There is still an important role to be played by innovations in energy efficient technologies, devices, and system design, building on the success of silicon CMOS. The predicted future global amounts of stored, computed, communicated, and sensed information will certainly challenge the world capability to process and make sense of zettabytes of data, requiring orders of magnitude improvements in energy efficiency.**

I. INTRODUCTION

The last decade has seen a decline in how much performance improves with each new generation of cutting-edge nanoelectronics microprocessors. With the end of Dennard-scaling already here, nobody can exactly predict when the era of Moore's Law [1] will come to the end. However, in addition to dimensions and processes attaining close to atomic scale limits, the energy efficiency of computing became one of the most important challenges for our community. According to Koomey [2], until 2009, the computations per kilowatt-hour doubled every 1.57 years, but today it takes almost 3 years for peak-output efficiency to double. Certainly, the energy efficiency benchmark has to be re-adapted to the current computation needs (including for instance average and idle energy efficiency metrics) but the slow-down is obvious. Historically, 10 years of scaling boosted efficiency by a factor of 100x but at current rate it would take around 20 years to see such improvement. This is the reason why it was claimed that Koomey's law is replacing today Moore's Law, in terms of importance. It is worth noting that past progress was not only due to technology scaling but also due to progress in power management design, algorithms, and software. On the other hand, today we have a much larger variety of processors (from high performance to mobile computing), which results in a much larger dispersion in the energy efficiency metrics and the need of new metrics for measuring the efficiency, such as a "typical-use efficiency" (computations per kilowatt-hour).

We are living in an era when there is an exponentially increasing pressure on the amounts of information that the society is processing with its information and communication technologies (ICTs) [3]. Only in 2007, humankind was able to store 2.9×10^{20} optimally compressed bytes, communicate almost 2×10^{21} bytes, and carry out 6.4×10^{18} instructions per second on general-purpose computers. The same year, 99.9% of the communication data and 94% of the stored data were in digital formats. In the perspective of the human technological progress, the amount of bits stored by humanity in all of its technological devices today is approaching the roughly 10^{23} bits stored in the DNA of a human adult, but remains still negligible compared with the 10^{90} bits stored of observable universe. However, with the emergence of the Internet-of-Things (IoT) and Cloud technologies, this amount of information is exponentially increasing, creating new challenges for computation and its energy efficiency. The 21st century is a new era for computing, characterized by the need to handle over zettabytes (10^{21} bytes, or ZB) of data. The world's capacities to *sense*, *transmit*, *store*, and *process information* need a technology capable of an improvement of three orders of magnitude, while keeping the consumption of energy at a level similar to the one that we have today. This translates into a big challenge for the new generations of researchers in electrical engineering and computer science to achieve about 1000x improvement in the energy efficiency of ICT (or in the computational performance per Watt).



Year	Storage	Commu-nication	Computing (general purpose)	Computing (Special purpose)	Edge IOT Sensing
1985	21PB	59PB	0.3PIPS	0.44PIPS	NA
2007	277EB	537EB	6.39EIPS	189EIPS	NA
2020	140ZB	272ZB	18ZIPS	2570ZIPS	>50ZB

Figure 1. *Top*: Schematics of the information flow involving edge of the cloud sensors, fog and cloud computing and the new ecosystem of companies. *Bottom* (adapted after [4]): estimated information storage, communication, computation and edge IoT sensing in the Zettabyte era.

IoT is perceived as the semiconductor industry's next big wave, following PCs, networking, and mobile systems with the promise of "everything connected" being certainly cloud dependent. The IoT is not only a technological revolution of smart objects but is also seen as a seamless combination of

embedded intelligence, ubiquitous connectivity and deep analytical insights that creates unique and disruptive value for companies, individuals, and societies. It leverages the emergence of Edge of the Cloud (EoC) sensing devices, which support the creation of a new ecosystem of services, business and innovation that puts altogether the Hardware, the Cloud and the Data. In this new ecosystem, the Big and Deep Data will be highly valued and are estimated to offer more than 50% of the business margin [4]. It is worth noting that Big Data is not wholly about size and has three major characteristics, called the 3Vs: *volume, variety*, and *velocity*. Moreover, data collected by IoT devices can contain very sensitive personal information and must be managed under security (involving standardization and certification as foundational principles) and anonymity to avoid any user privacy violations. The EoC sensing will massively contribute to generate vast amount of data of large variety supporting new service platforms and will push the humanity deeper into the *Zettabyte era*. The real value of IoT is at the intersection of gathering data and leveraging it: it requires an infrastructure in place to analyze it in real time and a true artificial intelligence for supporting and generating the new services for the humanity using cloud-based applications.

Most critically, we cannot escape facing the great challenge of energy efficiency in the Zettabyte era. To better understand how critical is the need of energy efficiency, let us observe that the energy required in the future is increasing exponentially and will be unsustainable in a world where 40ZB of information will be generated per year in 2020, without an 1000x improvement in energy efficiency. Additionally, IoT sensors are expected to put even more strain on big data infrastructure by generating thousands of trillion of bytes of data by 2030.

II. ENERGY EFFICIENT COMPUTING

A. Limits of energy computing: enough room at the bottom

Thermodynamics and quantum mechanics set fundamental limits for the energy transfer during binary switching. For the relationship between switching energy ΔE and transition time τ_d, the Heisenberg uncertainty principle requires $\Delta E \geq h/\tau_d$. The minimum energy required to preserve a binary state can be estimated from the Boltzmann probability as $E_{bmin} = 3\ k_B T \ln(2) \approx 10^{-21}$J (at T=300K). Irreversible or many-to-one operations such as AND or ERASE require dissipation of at least E_{bmin} for each bit of information lost. In principle, reversible or one-to-one logical operations such as NOT can be performed without dissipation, as shown by Landauer [5]. One drawback of reversible or adiabatic computation is that system switching speed is proportional to the energy dissipation; hence to achieve significant energy savings, prohibitively low speeds may be required. A detailed discussion of the ultimate limits of a computer was proposed by Lloyd [6]: it was suggested that the speed per logical operation is limited by its energy, and the amount of information that can be processed is limited by the number of degrees of freedom of the computing system. Today's advanced CMOS technology operates at energies on the order of 10^4 $k_B T$ per binary switching [7, 8], which gives us four orders of magnitude space down to the fundamental limit, Fig. 2. The computation power consumption can be expressed according to [7]:

$$\text{Power/Data} = N_{tr} x f x E_{tr} x \text{ComUE} + \text{leakage} \qquad (1)$$

$$E_{tr} = E_{factor} x k_B x T \qquad (2)$$

where the E_{factor} is an energy factor related to the state change in the Field Effect Transistors, depending on physics, materials and voltages and T is the temperature at which the transistor operate. For instance, for 14nm CMOS E_{factor} is today of the order of 1'500, which means it has been reduced by more than 5'000x in last 20 years. However, on the existing technology, there are only two practical knobs today for lowering E_{tr}: changing the physics of the transistor, with significant effect on voltage lowering, or operating at much lower temperatures (cryogenic electronics).

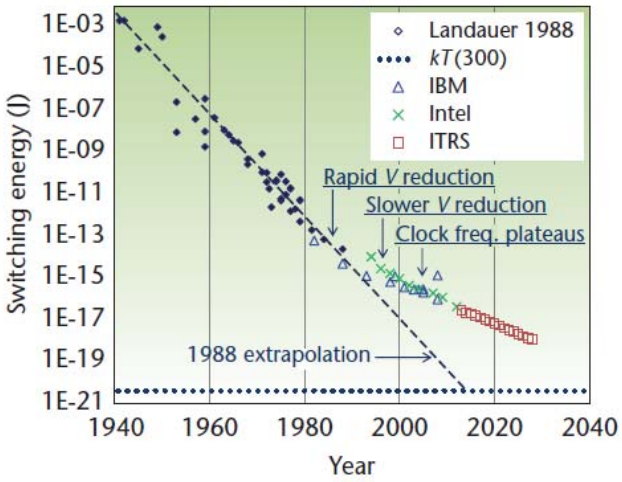


Figure. 2 Minimum switching energy dissipation evolution in logic devices, from. Black diamonds correspond to data from Landauer and dashed line is Landauer's 1988 extrapolation of the trend toward kT (at T=300 K), indicated by dotted line. Triangles and X's are published values from IBM and Intel. Reproduced from Theis and Wong [8].

B. Steep slope devices: new physics & technology-circuit design interactions for energy efficient hybrid cores

Minimizing power consumption in modern integrated circuits (ICs) is related to the voltage supply scaling, V_{dd}. Both dynamic power (proportional to V_{dd}^2), and standby power, (proportional to $I_{off}.V_{dd}$, and comparable or even dominant over the dynamic power in advanced nanometer nodes) depend on V_{dd} scaling. Until early 2000, the industry was able to scale V_{dd} according to Dennard's rule [9] but from 2000 to 2010 this process slowed down but continued thanks to the introduction of new materials. However, after 2010 the voltage scaling has been quasi-saturated at values close to 0.8V. This is due to the fact that conventional CMOS electronics relies on thermal excitation of electrons over a barrier, necessitating an operating voltage many times larger than the thermal voltage, $k_B T/q$, to maintain a good on-off ratio ($>10^5$) for a digital switch. Attempts to further scale down the threshold voltage, V_{th}, would result in an exponential increase in the off (leakage) current, I_{off}, of at least 10x for every 60mV of V_{th} reduction.

To address such fundamental problem, new classes of s *Steep-Slope Switches* have emerged with the main goal to drastically lower the operation voltage and thus the power consumption. Fig. 2 depicts some of the most successful device Steep-Slope Switch concepts that emerged [10]: (i) Tunnel FETs [11], (ii) NEM relays [12], (iii) Negative Capacitance FETs [13], and, (iv) Metal-Insulator-Transition or Phase

Change switches [14, 15]. Each of these categories are 3T devices, with their own performance and integration merits and demerits, which are discussed elsewhere [10] but, in essence, every new steep-slope device family open new challenges for the junction engineering and fundamental characteristics [16].

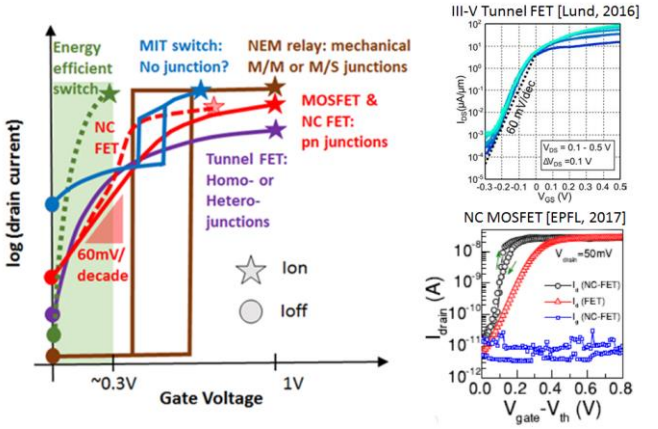


Figure 3. (a) Comparison of most promising steep slope switches in terms of I_{on}, I_{off} and subthreshold slope, S. (b)-(c) Transfer Characteristics of state of the art III-V Tunnel FET with S=48mV/dec [20] and Negative Capacitance MOSFET with S=20mV/dec [21].

Steep-Slope Switches such as Tunnel FETs or devices using multiple steep switching [17], offer an extended design space for energy efficient electronics below 0.3V voltage supply and 0.1V threshold voltage, which cannot be reached by CMOS.

Nanometer CMOS is putting strong quantitative challenges on the figures of merit requested for a new subthermionic switch, which is expected to solve power challenge issues:

- Steepness of the slope: below 10mV/decade over five decades of current to achieve sub-0.3V voltage supply.
- On/Off current ratio higher than 10^5.
- Off current lower than 0.1nA/µm at V_g=0V and V_d=V_{dd}.
- High-enough current density for complementary n- and p-type devices, in the range 0.2-1mA/µm.
- Operation speed in excess of 100's of MHz to 1GHz.
- Energy efficiency metrics: E_{tot_min} or performance/Watt [18], improved by 10-100x versus CMOS for V_{dd} <0.3V.
- Variability comparable with CMOS for similar gate length (technology node) and manufacturability on CMOS.
- Possibility to implement analog and sensing functions with CMOS-limit breaking analog gain at low current/voltage.

To date, the most successful beyond CMOS switch candidate is the tunnel FET, a solid-state semiconductor device designed as a gated p-i-n diode operated in reverse bias and exploiting the quantum-mechanical band-to-band tunneling (BTBT) at one of the junctions. The Tunnel FET configuration benefits from a very low leakage current and, at the same time, the carrier BTBT injection mechanism, potentially allows steep subthermionic subthreshold slope values. Their optimization exploits same additive technology boosters of CMOS [19] and control of defects [20], to limit the trap assisted tunneling, appears to be one main technological challenge. Other concepts for achieving steep slope via sub-unity body factors, such as negative capacitance can be used as performance boosters [21], Fig. 3. Overall, steep slope devices can be envisioned as add-on's on advanced CMOS platforms and large companies have already included some them in their future roadmaps. 1D heterostructure [20] and 2D [22] embodiments of Tunnel FETs appear as most promising and strong research efforts to identify most appropriate device architectures are in still progress.

The complementary figures of merit of MOSFETs (Fig. 3a) and Tunnel FETs generated new ideas about how to innovate and improve performance in hybrid CMOS-TFET design implementations [23] and, especially, extend the design space of CMOS. Of great importance is the possibility to design heterogeneous CMOS/Tunnel FET multicores comprising cores of different device technologies that can enable energy efficient executions. The challenges for heterogeneous cores can be formulated both as power reduction for a given performance (*dark* cores) or as improving performance under power constraints (*dim* cores). Datta [24] showed that on average a heterogeneous multicore has around 20 percent better energy-delay product than the homogeneous multicore and proposed to use hybrid 4CMOS-4Tunnel FET tiles in heterogeneous core design.

The capability of a digital technology to support analog IC is a feature of choice for designers. Due to the steep slope, some key subthreshold analog device metrics are improved by tunnel FETs. One is the ratio between the transconductance and the drain current, g_m/I_d, which reflects the efficiency with which the current (or power) is translated into transconductance; the greater the value of this ratio, the greater the offered transconductance, g_m, at a given current value. In fact, a tunnel FET has the ability to overpass the $40V^{-1}$ fundamental CMOS limit of g_m/I_d, offering significant gain improvement for analog applications operating at very low levels of drain current. Such unique property can be exploited in new design for energy efficient Analog-to Digital Converters (ADC) where the power dissipation per sampling frequency is limited by Signal-to-Noise Ratio (SNR), g_m/I_d and supply voltage, unscalable because of the requirements on SNR.

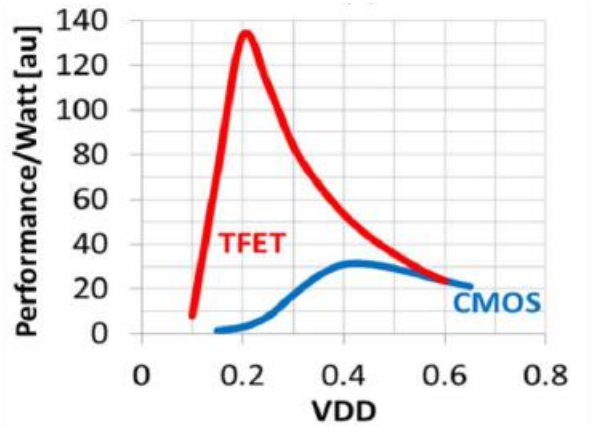


Figure 3. Energy efficiency of the logic operation, including leakage power, showing Tunnel FET higher performance per Watt than near-threshold CMOS, for V_{dd} below 0.3V. Reproduced from U. Avci, D. Morris, Ian A. Young, [18].

Steep slope devices such as Tunnel FETs are expected to notably enrich and extend the energy efficient design capability of nano-CMOS technological platforms by 2030 with two major contributions (i) to offer new possibilities to design more energy efficient heterogeneous digital circuits and dim cores, and, (ii) to supply the increasing energy efficient digital and analog design demands originating in IoT.

C. *3D chips for 1000x better energy efficiency?*

A recent approach that is different from individually focused technology optimizations of performance and energy efficiency in logic and memory functions but still exploits new devices and materials and their fine-grained monolithic 3D integration with ultra-dense connectivity is the Nano-Engineered Computing Systems Technology (N3XT) concept [25]. Fig. 4 depicts one possible embodiment of the concept in which field-effect transistors exploiting atomic-scale nanomaterials (1D or 2D materials) are integrated with huge amounts of dense nonvolatile storage devices (low-voltage resistive RAM) and magnetoresistive memories (spin-transfer torque magnetic RAM) and connected by ultra-dense, fine-grained inter-level vias. The reliable 3D heterogeneous integration of such diverse technologies, each used at its best, is one of the biggest challenges but offers unrivalled system-level benefits for energy efficiency and high performance. Other key architectural innovations of the N3XT chip include computation immersed in memory and thermal management. Such novel architecture will require and end-to-end device-architecture co-design with appropriate software support. For some computational workload benchmarks, it was estimated that the potential gain in energy-delay product of N3XT architecture is up to about 1000x, as compared to a traditional computers when dealing with abundant-data processing applications, specific to future IoT. The key enablers for N3XT are logic and memory devices that have and can be fabricated at low BEOL-compatible temperatures, and in thin layers that allow formation of ultra-dense inter-layer vias.

While in the N3XT chips the combination of logic and memory for intensive computation of abundant data is the main focus, other 3D heterogeneously integrated chips combining CMOS logic, magnetic memory, sensors, ADC and analog/RF circuits and 3D integrated energy harvesters, have been explored by the European Consortia e-BRAINS and e-CUBES to build full IoT sensor nodes with *fog computation* capability [26].

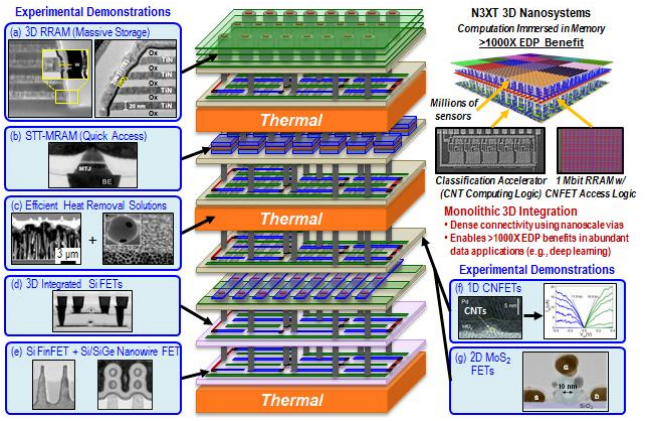


Figure 4. Depiction of one possible embodiment of a 3D nanosystem enabled by Nano-Engineered Computing Systems Technology (N3XT), aiming at energy efficient computation for vast amounts of data, courtesy of H.-S. Philip Wong and Subhasish Mitra, Stanford University. EDP denotes energy × execution time. (a) Left: 3D vertical RRAM, Q. Luo et al. (CAS) [27], Right: 3D vertical RRAM integrated with FinFET, F.K. Hsueh et al. (NNDL) [28]; (b) STT-MRAM integrated with 2x nm CMOS, D. Shum et al. (Globalfoundries) [29]; (c) Efficient heat removal solutions (Stanford) [30]; (d) 3D integrated Si FETs, L. Brunet et al. (CEA LETI) [31]; (e) Left: Si FinFET, Y. Sasaki et al. (IMEC/ASM) [32], Right: Si/SiGe Nanowire FET, H. Mertens et al. (IMEC) [33]; (f) 1D carbon nanotube FET with 5 nm gate, C. Qiu et al. (PKU) [34] ; (g) 2D MoS2 FET with 10 nm gate, C. English et al. (Stanford) [35]; 3D Nanosystems, M.M. Shulaker et al. (Stanford/MIT) [36].

D. *Quantum computing: a convergent technological path with nano-CMOS?*

In strong contrast with a conventional computer, the unit of information in a quantum computer is the quantum bit or qubit. Qubits can simultaneously exist in a superposition of both states |0> and |1> (represented intuitively on the Bloch sphere), providing extraordinary computational power and speed-up. The extraordinary potential of quantum computing (QC) comes from the fact that a small number of particles in superposition states can carry an enormous quantity of information [37]. For instance, 1,000 particles in superposition can represents every number from 1 to $2^{1,000}$ (~10^{300}), and QC can manipulate these numbers in parallel. The big challenge is that at the end of the computation, the rules of quantum mechanics impose the measurement to pick out one of the 10^{300} possibilities, which need smart manipulations. Many computational problems that are not tractable by standard computation algorithms (such as factorization or simulations of quantum physics) are expected to be solved by quantum computers using thousands to millions of qubits. Various qubits implementations: trapped ions, electron/hole spin in semiconductors, single dopants in silicon, nitrogen-vacancies in diamond lattices, etc, have been proposed to date. Practically all qubit technology operate near absolute zero (10's to 100mK), with challenging and/or questionable paths to scalability and possible problems of decoherence at large scale. Existing QC supports only a few to dozens of qubits [38], with machines of large form factors, especially because of the cryogenics required. Even if a quantum computer could be made of general purpose and 1000x faster than a CMOS microprocessor, the economics is still questionable.

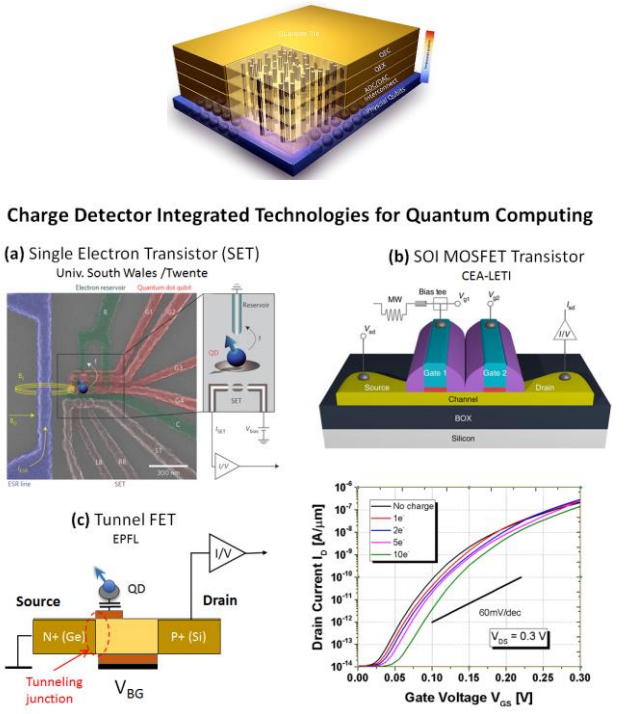


Figure 5. Top: IBM's vision for a future, scalable quantum computer supporting large temperature gradients, with qubits in the bottom layer, at lowest temperature, and various correction layers on the top, reprinted from [39]. (a)-(b)-(c)-(d) Single Electron Transistor (demonstrated) [41], SOI MOSFET (demonstrated) [42] and Tunnel FET (concept and simulations) [43] for integrated highly sensitive qubit charge detection.

Surprisingly, the field of quantum technologies is still witnessing a battle between the never ever-give-up optimists and disbelievers. But something has changed in recent years and CMOS is not anymore the competitor but a potential enabling technology for Qc that require architectures with components operating over large range of temperature (qubits at mK, cryogenic electronics at few K and some other components to interface with the real world at room temperature) [39], Fig. 5. Making quantum computers based on silicon transistors [40] and chips appears more and more possible and realistic. Long coherence times are now possible by the presence of spin-free isotopes of carbon and silicon. Moreover, controlling and reading the quantum state of a single spin in a MOSFET device constitutes today one of the best solid-state qubits, better than other exotic solutions. Figs. 5 (b) and (c) compare Single Electron Transistors (SET) [41] charge detectors with SOI MOSFETs [42]. Emerging steep slope devices using direct bandgap materials having implementations similar with CMOS but with the advantage of an uniquely high charge sensitivity for low currents, no background charge effect and very little dependence of temperature, from deep cryogenic to room temperature are future possible solutions for a scalable energy efficient qubit readouts [43] (Fig. 5c).

With the QC cryogenic consuming large amounts of power and with the claims on a life-time horizon for the realization of large-scale quantum computers operating at low temperatures the thermal management challenges and dissipation in cryogenic components are legitimate questions. It also naturally raises a question about fundamental limitations of energy consumption in scalable quantum computing. Recently, Ikonen et al [44] addressed the *energy efficiency of QC* by exploring the lower bound on the energy consumption of qubits and proposing new protocols capable to reduce by orders of magnitude the power consumption in a QC system.

E. Neuromorphic computing

Neuromorphic computing refers to biology- or brain-inspired computers, devices, and models that contrast the von Neumann computer architecture. The blueprint of a brain-inspired neuromorphic processor is the way in which memory and processing units are organized and interconnected [45, 46].

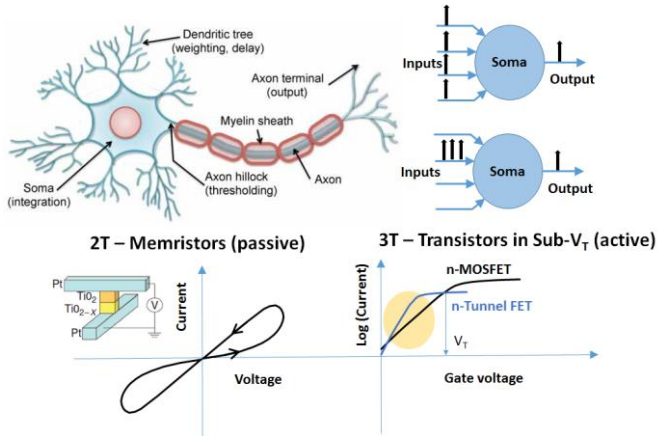


Figure 5. *Top:* Neuro structure and model showing action potentials (spikes in electric potential vs. time, denoted by the vertical arrows) for spatial and temporal summation, and, *Bottom:* depiction of transfer characteristics of memristive and steep slope tunnel FETs, for neuromorphic computing.

The biological processing systems are characterized by co-localized memory and computation exploiting neuron synapse capability to perform at the same time multiple functions such as memory storage and non-linear operations. The energy efficiency of such architecture is outstanding: while conventional supercomputers consume MWatts of power and take long times to carry out complex calculations, a human brain consumes in average 20Watts being capable of much more complex functions in real time. Similarly, complex functions are executed with microWatts power consumption by a fly while performing in real flight control, path planning, food and mate search or even predator avoidance. Neuromorphic computing can greatly benefit from subthreshold CMOS implementations and the new families of low power emerging devices such as 2-terminal memristive devices [47] and 3-terminal steep slope devices, capable to offer unique subthreshold super-exponential characteristics, Fig. 5. Other ongoing neuromorphic computing works consider coupled and scalable relaxation-oscillators utilizing the metal-insulator-metal transition of vanadium-dioxide (VO_2) thin films [48]. Fig. 6 depicts the estimated computational energy efficiency for digital systems, analog signal processing, and potential neuromorphic hardware-based algorithms, suggesting the unique potential of neuromorphic systems, and the enormous room left for improvement (8–9 orders of magnitude better power efficiency for biological systems that the wall of digital computation) [49].

It is worth noting that, among many outstanding recent works, IBM has developed end-to-end technology and ecosystem to create and program energy-efficient, brain-inspired machines that mimic the brain's abilities for perception, action, and cognition. An example is the TrueNorth, a 65 mW real-time neurosynaptic processor that implements a non-von Neumann, low-power, highly-parallel, scalable, and defect-tolerant architecture, having 4096 neurosynaptic cores tiled in a 2-D array, and containing 1 million digital neurons and 256 million synapses [50]; its reported computational energy efficiency is of 400 GSOPS per Watt (GSOPS/W).

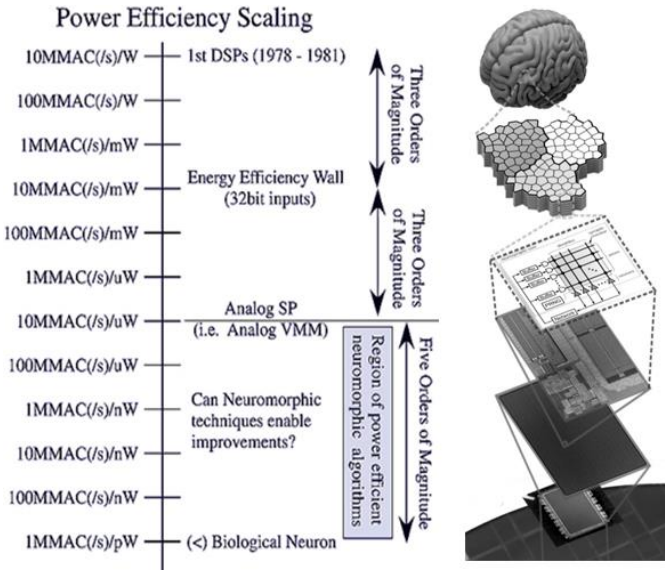


Figure 6. *Left:* Computational efficiency of various technologies: digital technologies, analog signal processing and best estimate of biological neuron computation. Reproduced after [49]. *Right*: TrueNorth architecture, [50].

III. ENERGY EFFICIENTS SENSING: FROM PERSONAL SMART-HUBS TO AUTONOMOUS IOT NODES

Contributions to the quest for "zero-power" technologies and for pushing the scientific and technological limits of energy per processed bit of information (Fig. 7) have been pioneered in 2011 by the FET Flagship project 'Guardian Angels for a Smarter Life' [51]. This initiative was among the first to propose application-driven developments of energy efficient EoC technologies, as intelligent, autonomous electronic personal devices with embedded privacy and security, featuring sensing, computation, and communication beyond human capabilities. This initiative was followed by a large international recognition of the importance of the EoC smart sensors in many scenarios, under the idea of a 'trillion sensors planet or universe'.

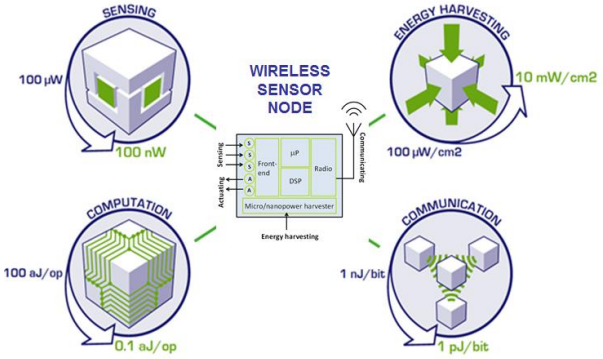


Figure 7. Architecture of a Wireless Sensor Node (WSN) – center – with its key functional components and the future goals in terms of energy efficient processing (energy per useful computed, communicated and sensed bit of information) together with goals for energy harvesting to support the autonomy of WSN (zero power), [51].

F. *Smart energy efficient sensors*

Smart sensors form a fascinating domain for science and technology and they are, in part, responsible for the great success of smart phones that are a new kind of energy efficient computers with multi-sensing capability and wireless connectivity. A sensor is a functional device transforming real world information into electronic information. The EoC sensors support the artificial senses of the IoT, opening capabilities beyond human six senses and enabling disruptive personalized services. They will collect physical, chemical, biochemical, electromagnetic abundant data to enable interpretation and monitoring of any kind of processes, environment, and person's physiological and emotional status, in relation to the environmental and social context.

From the electronic system level perspective, the sensor generates a bit stream in response to changes in some external stimulus. At this level, together with the consumption P, the important metrics are the input full scale (FS), the input bandwidth B (or Nyquist sampling rate), and the effective number of bits (ENOB) N. These quantities are interlinke when considering a sensor to consist of a transducer and an ADC only. For ADCs, Walden has investigated several sources of distortion that may degrade the effective number of bits N, including thermal noise, aperture uncertainty, comparator ambiguity and even the Heisenberg limit. When thermal noise is the main limitation, the relation between power, bandwith and the number of bits stands as:

$$P = 6kT \times 2^{2(N+1)} \times f_s \times (1+NS^{-2}) \qquad (3)$$

Equation (3) tells that if the low power budget of a sensor and ADC is around 100nW, then one should not expect more than 8bit/sample at 1kS/s for a transducer offering a 1% NS. For advanced sensors, the state-of-the-art energy per conversion is as low as tens of pJ in resistive sensors and can be as low as few pJ for capacitive sensors. EoC wearable, environmental and implantable sensors form categories in which the low power consumption is key for quasi-continuous multi-sensing. In advanced energy efficient scenarious, the sensor readout will need to achieve down to the range of 10fJ per conversion.

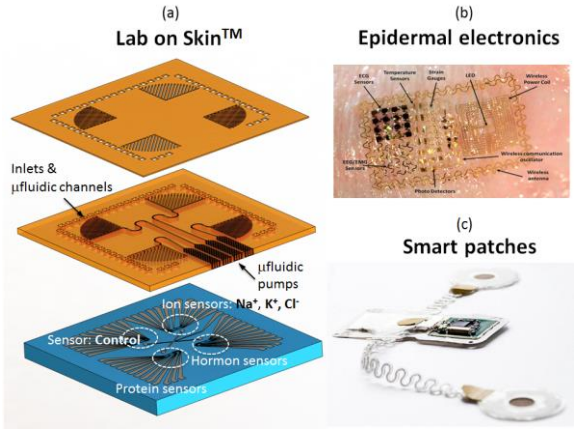


Figure 8. (a) Lab On Skin™ [54] smart multi-sensing systems based on the 3D heterogeneous integration of SU8 micro/nanofludics on UTB FD SOI ISFETs functionalized for multiple analytes, (b) Epidermal electronics tattoo including sensing and wireless communication blocks on ultra-thin biocompatible substrate [52], (c) Smart patch for heart rate and respiration monitoring with motion artefact removal [53].

Fig. 8 depicts three recent smart sensor examples in which energy efficient multi-sensing: (i) the epidermal electronics developed by Rogers [52] (ii) the smart patches for activity and biosignal monitoring developed by IMEC's Holst Center [53] and, (iii) the novel Lab On Skin™ technology by Xsensio and EPFL for monitoring in real time biomarkers in sweat, by combining advanced UTB FD-SOI ISFETsensor arrays with 3D integrated nanofluidics, with the experimental validation for electrolytes sensing in sweat being reported in [54].

G. *Wireless communications for EoC devices*

An essential feature of EoC autonomous smart sensors is their ability to communicate wirelessly using minimal energy per effective information bit. In many state-of-the-art sensor node scenarios, the wireless communication consumes a very significant part of the available energy. The future IoT will require radical progress in the energy efficiency for sensor node communications by factors of 100x to 1000x. The fundamental performance limit for the required transmitted energy in a wireless communication system is derived from the Shannon capacity theorem (assuming a Gaussian transmitted signal using an infinite block code interval). The consequence is that communication is not possible below the received signal-to-noise ratio per bit of -1.6 dB in the wideband limit. Assuming

that only the thermal noise affects a communication device, this capacity theorem sets a fundamental lower bound on the transceiver power. From this fundamental limit of -1.6 dB at the receive side, it can be derived that communication with a transmitted energy of 1pJ/bit is theoretically achievable when transmit-receiver attenuation is limited below 85dB. An energy-proportionality in the function of radio utility is needed. The future needs are to achieve transmitted energy per bit close to 1pJ, with a total system energy per information bit of 10pJ/bit**.** Such energy efficiency metric should include all energy used in the transmitter and receiver, used for transmitted power, in radio front-end operation, and signal processing (PHY, MAC and network layer). To achieve sub-10pJ/bit one needs to combine: (i) RF technology and device innovations (energy efficient RF front ends and RF MEMS) under predefined conditions, and, (ii) innovations in designs of fully self-adaptive/reconfigurable radios, under dynamic conditions.

H. Energy harvesting and storage for WSN

Future energy efficient autonomous systems will need to have embedded ambient energy harvesters and the capability to manage energy generation, storage and efficient power management schemes. Such opportunistic energy interfaces to the real world should be self-adaptive and combine multiple-principles of energy harvesting, based on a large variety of materials requiring heterogeneous integration.Energy harvesting is an emerging field with many challenges and limitations of fundamental, technology and context-dependent nature, including, for instance, the Shockley-Queisser (SQ) limit of 33.7% for single junction solar cells, the Carnot efficiency for thermoelectric harvesters (e.g. 4 % at ΔT=12K around room temperature), proof mass size and internal displacement limits for motion harvesters, and a maximum cell voltage and fixed operating temperature for biofuel cells. New generations of energy harversters will be combined with rechargeable batteries, depending on the energy scenarios and usage. Equally important are energy storage technologies such as supercapacitors using new classes of 1D and 2D materials.

IV. PARADIGM CHANGE FOR PERSONALIZED HEALTH CARE WITH ENERGY EFFICIENT IOT/IOH

In this final section, we discuss the positive impact that energy efficient zettabyte ICT technologies can have on unsustainable field of modern society, such as healthcare. The continuing health of our societies is increasingly threatened by the enormous cost of providing appropriate healthcare (currently more than 4 Billion Euros per day, in Europe) in the rapidly ageing societies. While we cannot avoid medical mistakes, we can avoid most or all of their consequences by making these mistakes in-silico. This change of strategy, exploiting advanced technologies, has been implemented in most areas of our existence, increasing efficiency, saving lives and reducing costs. The healthcare is taking very little of the enormous progress achieved today in ICT. This enormous progress has, as yet, reached neither our health care system nor the way we develop new drugs or can predict the health evolution. In addition to nanotechnology and abundant data analytics, we will need concepts to integrate this information and to predict the effects and side effects of possible therapies (or preventive measures) on every individual. We will show that a truly individualised healthcare and disease prevention system, based on a detailed characterization (clinical, molecular, imaging and sensor based) of the patient/individual and their wellness, health and disease course can be achieved only by federating multidisciplinary research in a large initiative combining *energy efficient IoT* technologies with the idea of building multi-level computer models for future *virtual patients* (avatars), calibrated on omics, imaging and sensor abundant data, Fig. 9.

Such a vision involves a true paradigm change in health care by exploiting the engineering advances in sensing, computing, communication and cloud/fog technologies together with new ways of exploiting big, deep and abundant data, to enable personalized and preventive medicine supported by a specific data infrastructure and a subclass of the Internet of Things called here the Internet-of-Humans (IoH), designed with embedded security, privacy and ethics.

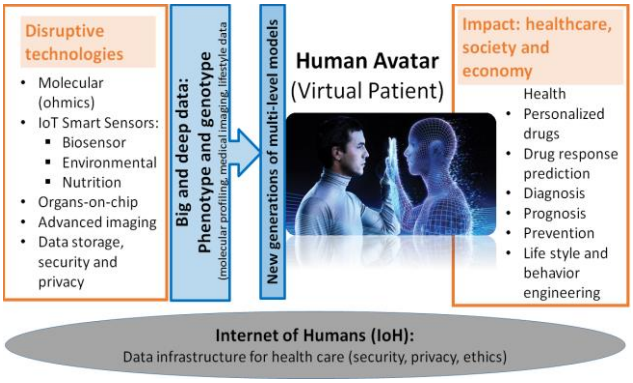


Figure 9. Vision of Future Health FET Flagship new initiative, showing the combination of molecular (omics), IoT sensors, organs-on-chip and advanced imaging technologies used to generate abundant data supporting the build-up of virtual patients (avatars) computer models for future personalized and preventive healthcare.

V. CONCLUSIONS

In the 21st century Zettabyte era, we will certainly see the emergence of new evolutions in electronic and cloud technologies, with more innovation developed in connection with the need and use of abundant data in IoT applications and services that cannot be imagined today. The advanced energy efficient silicon platforms for computation, sensing and communication will support new generations of connected devices, abundant, big and deep data generation and storage but also unique advances in algorithm design. This progress will close more and more the gap between potential and reality in artificial intelligence (AI), beyond simple matching or exceeding the proficiency of humans, by technology-enabled new synergies between human brains and machines, beneficial to a vast number of application domains.

ACKNOWLEDGMENTS

This work has been supported by the ERC Advanced Grant Milli-TEC of the European Research Council and by the H2020 CSA project NEREID (https://www.nereid-h2020.eu/). The author is grateful to prof. Philip Wong and prof. Subhasish Mitra from Stanford University for their careful reading of the paper and very useful feedback.

REFERENCES

[1] G.E. Moore, "Cramming More Components onto Integrated Circuits," Electronics, vol. 38, no. 8, 1965,pp. 1–14..

[2] J.G. Koomey et al., "Implications of Historical Trends in the Electrical Efficiency of Computing," IEEE Annals of the History of Computing, 20011, pp. 46-54.

[3] M. Hilbert, P. López, "The World's Technological Capacity to Store, Communicate, and Compute Information", Science, vol. 332, 2011, pp. 60-65.

[4] Z.-W. Xu, "Cloud-Sea Computing Systems: Towards Thousand-Fold Improvement in Performance per Watt for the Coming Zettabyte Era," Journal of Computer Sci. and Technology, 29 (2), 2014, pp. 177–181.

[5] R. Landauer, "Dissipation and Noise Immunity in Computation and Communication," Nature, vol. 335, 1988, pp. 779–784..

[6] S. Lloyd, "Ultimate physical limits to computation," Nature, vol. 406, 2000, pp. 1047-1054.

[7] J. Summers, "From ZettaBytes to zeptoJoules – will digital demand outstrip the physical limits?," online presentation: https://www.datacentreworld.com/

[8] T. Theis and H.-S. Philip Wong, "The End of Moore's Law: A New Beginning for Information Technology," Computing in Sci. & Engineering, May/June 2017, pp. 41-50.

[9] M. Bohr, ''A 30 Year Retrospective on Dennard's MOSFET Scaling Paper,'' IEEE Solid-State Circuits Newsletter, vol. 12, 2007, pp. 11–13.

[10] A.M. Ionescu, Chapter 5: Beyond-CMOS Low-Power Devices: Steep-Slope Switches for Computation and Sensing, in *Nanoelectronics: Materials, Devices, Applications*, Eds. Marcel Van de Voorde Professor, Robert Puers, Livio Baldi Sebastiaan, E van Nooten, Wiley (2017).

[11] A.M. Ionescu, H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," Nature," 479, 2011, pp. 329-337.

[12] V. Pott et al., "Mechanical Computing Redux: Relays for Integrated Circuit Applications," Proceedings of the IEEE, Vol. 98, 2010, pp. 2076 – 2094.

[13] S. Salahuddin, S. Datta, Nanoletters, "Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices," 8 (2), 2008, pp. 405-410.

[14] Y. Zhou, X. Chen, C. Ko, Z. Yang, C. Mouli, S. Ramanathan, "Voltage-Triggered Ultrafast Phase Transition in Vanadium Dioxide Switches," IEEE Electron Dev. Letts., Vol. 34, 2013, pp. 220-222.

[15] W.A. Vitale et al., "Steep-Slope Metal–Insulator-Transition VO2 Switches With Temperature-Stable High ION," IEEE Electron Device Letters, Vol: 36, 2015, pp. 972 – 974.

[16] A.M. Ionescu, "Nano-devices with advanced junction engineering and improved energy efficiency," 17th International Workshop on Junction Technology (IWJT), 2017, pp. 1- 6.

[17] W.A. Vitale et al., "A Steep-Slope Transistor Combining Phase-Change and Band-to-Band-Tunneling to Achieve a sub-Unity Body Factor," Sci. Reports 7 (1),2017, p. 355.

[18] U. Avci, D. Morris, Ian A. Young, Tunnel Field-Effect Transistors: Prospects and Challenges, IEEE JEDS, Vol. 3, 2015, pp. 88-95.

[19] K. Boucart, A.M. Ionescu, "Double-gate tunnel FET with high-κ gate dielectric," IEEE Trans. Electron Devices, Vol. 54, no. 7, 2007, pp. 1725-1733.

[10] E. Memisevic et al., "Vertical InAs/GaAsSb/GaSb tunneling field-effect transistor on Si with S = 48 mV/decade and Ion = 10 μA/μm for Ioff = 1 nA/μm at Vds = 0.3 V," IEEE IEDM, 2016, p. 19.1.1.

[21] A.Saeidi et al., Negative Capacitance as Performance Booster for Tunnel FETs and MOSFETs: an experimental study, IEEE EDL, vol. 38, pp. 1485-1488, 2017.

[22] D. Jena, "Tunneling Transistors Based on Graphene and 2-D Crystals," Proceedings of the IEEE, Vol. 101, 2013, pp. 1585 – 1602.

[23] D. Morris et al., Novel TFET circuits for high-performance energy-efficient heterogeneous MOSFET/TFET logic, VLSI-TSA 2017.

[24] K. Swaminathan et al., "Steep-Slope Devices: From Dark to Dim Silicon, IEEE Micro, published by Computer Society, Sept./Oct. 2013, pp. 50-59.

[25] M. M. Sabry Aly et al., "Energy-Efficient Abundant-Data Computing: The N3XT 1,000×, Computer, Dec. 2015, pp. 24-33.

[26] http://www.e-brains.org/heterointegration/integration/

[27] Q. Luo et al., "Demonstration of 3D vertical RRAM with ultra low-leakage, high-selectivity and self-compliance memory cells," 2015 IEEE International Electron Devices Meeting (IEDM), 2015, pp. 10.2.1-10.2.4.

[28] F. K. Hsueh et al., "First fully functionalized monolithic 3D+ IoT chip with 0.5 V light-electricity power management, 6.8 GHz wireless-communication VCO, and 4-layer vertical ReRAM," 2016 IEEE International Electron Devices Meeting (IEDM), 2016, pp. 2.3.1-2.3.4.

[29] D. Shum et al., "CMOS-embedded STT-MRAM arrays in 2x nm nodes for GP-MCU applications," Symposium on VLSI Technology, 2017, pp. T208-T209.

[30] M. M. Sabry Aly et al., "Energy-Efficient Abundant-Data Computing: The N3XT 1,000x," in Computer, vol. 48, no. 12, pp. 24-33, Dec. 2015.

[31] L. Brunet et al., "First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers," 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2.

[32] Y. Sasaki et al., "Novel junction design for NMOS Si Bulk-FinFETs with extension doping by PEALD phosphorus doped silicate glass," IEEE International Electron Devices Meeting (IEDM), 2015, pp. 21.8.1-21.8.4.

[33] H. Mertens et al., "Gate-all-around MOSFETs based on vertically stacked horizontal Si nanowires in a replacement metal gate process on bulk Si substrates," IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2.

[34] C. Qiu et al., "Scaling carbon nanotube complementary transistors to 5-nm gate lengths," Science, vol. 355, 2017, pp. 271-276.

[35] C. D. English, K. K. H. Smithe, R. L. Xu and E. Pop, "Approaching ballistic transport in monolayer MoS2 transistors with self-aligned 10 nm top gates," IEEE International Electron Devices Meeting (IEDM), 2016, pp. 5.6.1-5.6.4.

[36] M.M. Shulaker et al., "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," Nature, vol. 547, 2017, pp. 74-78.

[37] S. Aaronson, "The limits of Quantum Computing, Sci. American, 2008, pp. 62-69.

[38] https://www.research.ibm.com/ibm-q/

[39] F. Sebastiano et al., "Cryo-CMOS Electronic Control for Scalable Quantum Computing," DAC '17, June 18-22, 2017, Austin, TX, USA.

[40] M. Fernando Gonzalez-Zalba et al., "Gate-Sensing Coherent Charge Oscillations in a Silicon Field-Effect Transistor," Nano Lett. 16, 2016, pp. 1614−1619.

[41] M. Veldhorst et al., "An addressable quantum dot qubit with fault-tolerant control-fidelity," Nature Nanotech., Vol. 9, 2014, pp. 981-995.

[42] R. Maurand et al., "A CMOS silicon spin qubit" Nature Comms, Nov 2016, pp. 1-6.

[43]. A.M. Ionescu, C. Alper, T. Rosca, "Steep slope charge detector for quantum computing," Patent pending, 2017.

[44] J. Ikonen, J. Salmilehto and M. Möttönen, "Energy-efficient quantum computing", npj Quantum Information, 17, 2017.

[45] C. Mead, "Neuromorphic electronic systems," Proc. of IEEE, vol. 78, 1990, pp. 1629–1636.

[46] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," Proceedings of the IEEE, Vol. 103, Issue: 8, 2015, pp. 1379 – 1397.

[47] T. Chang, Y. Yang and W. Lu, "Building Neuromorphic Circuits with Memristive Devices,"IEEE Circuits and Systems Mag., 2013, pp. 56-73.

[48] S. Datta et al., "Neuro Inspired Computing with Coupled Relaxation Oscillators," Design Automation Conference (DAC), 2014.

[49] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," Front. Neurosci., Vol. 7, Art. 118, 2013, pp. 1-29.

[50] F. Akopyan et al., "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," IEEE Trans. CAD of Integrated Circuits and Systems, Vol. 34, 2015, pp. 1537 – 1557.

[51] A.M. Ionescu and C. Hierold, "Guardian Angels for a Smarter Life: Enabling a Zero-Power Technological Platform for Autonomous Smart Systems, " Procedia Computer Science, Vol. 7, 2011, pp. 43-46.

[52] D. Kim et al., "Epidermal electronics," Science, Vol. 333, 2011, pp. 838-843.

[53] M. Konijnenburg et al., A Battery-Powered Efficient Multi-Sensor Acquisition System with Simultaneous ECG, BIO-Z, GSR, and PPG, ISSCC 2016, pp. 480-482.

[54] F. Bellando et al., "Lab On Skin: 3D monolitically integrated zero-energy micro/nanofludics and FD SOI ionic sensive FETs for Wearable Multi-Sensing Sweat Applications," to appear, IEDM 2017.