Abstract: In today’s world, ruled by a great amount of data and mobile devices, cloud-based systems are spreading all-over. Such phenomenon increases the number of connected devices, broadcast bandwidth, and information exchange. These fine-grained interconnected systems, which enable the Internet connectivity for an extremely large number of facilities (far beyond the current number of devices) go by the name of Internet of Things (IoT). In this scenario, mobile devices have an operating time which is proportional to the battery capacity, the number of operations performed per cycle and the amount of exchanged data. Since the transmission of data to a central cloud represents a very energy-hungry operation, new computational paradigms have been implemented. The computation is not completely performed in the cloud, distributing the power load among the nodes of the system, and data is compressed to reduce the transmitted power requirements. In the edge-computing paradigm, part of the computational power is moved toward data collection sources, and only after a first elaboration collected data are sent to the central cloud server. Indeed, the “edge” term refers to the extremities of systems represented by IoT devices. This survey paper presents the hardware architectures of typical IoT devices and sums up many of the low power techniques which make them appealing for a large scale of applications. An overview of the newest research topics is discussed, besides a final example of a complete functioning system, embedding all the introduced features.

Keywords: Edge computing; Internet of Things(IoT); MicroController(MCU); low power; embedded system

1. Introduction

The Internet of Things (IoT), term coined in 1999 by Kevin Ashton, is gaining more and more attention in these years due to the increasing amount of connected devices and consequently to the amount of data. In the big data era, recording data from several environments and users is extremely valuable from a statistic as well as a business and economic point of view. Nowadays, almost every device present in everyday life presents some embedded electronics, which turns it into a potential IoT node. Indeed, IoT nodes are able to sense information and transmit it, thanks to a communication interface. So far the IoT paradigm had a huge impact on both consumers lives and business models, due to the decreasing cost of implementation of these devices and the increasing demand. The trend is expected to rapidly increase, as shown in Figure 1. Gartner [1] (world’s leading company in research and advisory fields) states that 23.14 billion of connected devices have been produced in the past year (2018), up to 30.73 billion are expected for 2020. This represents a great opportunity for investors, producers and companies to collect big data. In fact, companies are expected to spend around 5 trillion dollars in 2021.
to expand the market [2] and introduce new applications, embedded in everyday gadgets.

In a highly dynamic scenario, as depicted above, the opportunities to diversify the possible solutions and applications are many. Besides, there still exists a main common factor: the hardware implementation. Indeed, hardware architectures are quite similar regardless of the final use since their organization relies very often on microcontroller-based platforms. Most part of IoT devices leans on batteries or energy harvesters. Given that their energy budget is limited, even the power that can be consumed will need smart energy management, driving the hardware engineering toward an ultra-low power approach. Limited power represents a huge constraint to many components of the architecture, especially the energy-hungry ones, like wireless transmitters. In such a case, communication and data to be broadcast must be reduced to the essential, which translates to low energy technology, such as Ultra-Wide Band (UWB) [3–8], and transmit only useful features by exploiting state of the art techniques, like Compressing Sensing.

IoT nodes have to sense and collect data with respect to their specific task. The tasks could be many, ranging from smart household appliance, sleep monitoring, physical activity tracking, caretaker condition monitoring etc. Such information, collected over thousands of individuals, must ensure absolute privacy for final users. Indeed, such personal data must be kept safe from outside attacks, avoiding any hacking attempt. Even localization-based services are affected by this problem. The position is part of a privacy policy that must be prevented from threats [9]. In order to protect data, the common procedure is to encrypt the communication so that nobody can steal precious info. Device security, together with encryption techniques are a matter of discussion and research widely spread around the globe.

Many of the aforementioned tasks, including the last one described above, can be performed directly on the platform without accessing the cloud or a remote hosting service. This considerably reduces the power needed to transmit and receive data before and after elaboration, relieving much of the effort from servers. This change of paradigm is called Edge Computing: part of the workload is decentralized and distributed among the IoT nodes, turning them from simple sensors into more powerful and smart embedded systems, capable of several new features.

Thanks to this innovative and effective approach, measured information can be further analyzed directly on the field, allowing for a more responsive application and a faster post-processing operation once data has been transmitted. Edge Computing paradigm is considered, from many, an environmentally friendlier alternative to the Cloud Computing one, due to its ability to restrain the volume of data to be moved, consequently cutting down the energy cost.

As described above, the decentralization of the workload is the focus of many works of research. The

![Figure 1. Expected adoption growth of IoT devices](image-url)
aim is to reduce the latency by offloading some of the tasks on surrounding servers \[10,11\]. In fact in urban environments is possible to rely on such infrastructure to enhance performances. However, in order to correctly manage the workload, there is often the need to split it homogeneously and to synchronize all the different duties. To achieve this goal, something more is required, such as an Operating System (OS). The use of a complete OS instead of a limited embedded firmware becomes a powerful tool when handling complex tasks. In order to manage multiple users and different scenarios, required for IoT, the main OS can run virtualized operating systems (which emulates the entire hardware resources required by an OS) or it can exploit the containerization paradigm, which let the user have multiple instances of the OS running at the same time while sharing the kernel for resources allocation. The containerization of tasks makes the organization and the allocation of workloads more efficient, by exploiting the bandwidth of the system at its maximum, without exceeding constraints \[12\]. For sake of clarity, since this current survey is focusing on hardware aspects, OS and software related technologies are not discussed in the following sections; however, the reader can refer to the provided references for further details.

This survey is intended to be an overview of the key aspects of IoT hardware platforms, designed for Edge Computing. Several state of the art techniques, suitable for low power applications, are introduced and discussed through real examples. Section 2 presents the motivations that drive the continuous development of new architectures and techniques, with a special focus on the Edge Computing approach, by describing and emphasizing its importance. Section 3 depicts the landscape of a typical IoT node system, comprehensive of all its main features, ranging from the "brain" (Central Processing Unit) to its peripherals. First, Input/Output (IO) systems are presented as the main communication channels of the system (3.1); a detailed analysis of several IO types is listed, providing pros and cons for every possible choice. Then, memories are introduced as fundamental elements for data retention. Even in this case, several solutions are presented, with a special focus on non-volatile memories (3.3) (memory elements in which data are retained even in case of power failure or interruption).

Power management (Section 3.4) is responsible for deciding which parts of the system should be turned off, when they are detected as not useful for the current task, or even for adjusting the local power supply parameters, such as the voltage level. After this unit, near-threshold behavior is described (3.5) in order to better understand what are the consequences of power supply parameters tuning. Since Edge Computing heavily relies on data processing, an entire section is dedicated to this topic (Section 3.6). State of the art algorithms and techniques able to reduce significantly the amount of data to be transmitted are discussed.

Section 3.8 is completely devoted to the Central Processing Unit (CPU), explaining the evolution of single-core architectures and the reason for adopting the multi-core paradigm. Indeed, most part of modern IoT nodes is equipped with powerful processors, able to effectively distribute the computational load among the cores and, consequently, level out the power dissipation. Finally, a real platform, which sums up all the above-listed features, is presented in Section 4, while Section 5 concludes the paper by providing new challenges and future perspectives.

1.1. Methodology and Organization

This section is intended to explain the review methodology, which this paper is based on. Researchers have at their disposal several different approaches \[13\] to collect and summarize the state of art literature on a specific topic. A review paper is the result of five generic steps:

1. topic and objectives definition;
2. primary search;
3. information refinement and secondary search;
4. data retrieving;
5. analyzed data presentation.
Namely, once the main subject of the manuscript is defined, primary search is fundamental in order

to create a pool of articles from which the topic is clearly presented and the reader can figure out

the principal aspects. Then, a refining process is necessary in order to discard loosely related articles

or misleading essays. Secondary search is intended to integrate the current information pool and to

deepen some important points. Once the ensemble of articles has been consolidated, data and topic

are extracted, reworked and presented.

All the aforementioned points seem to be sequential and to be followed in a linear path, but the

review process is strongly iterative, going back and forth through stages many times, before reaching a

satisfactory result [14].

The methodology chosen for this paperwork is the narrative review, that is a very traditional way of

reviewing contemporary literature. It consists of a summary of the found material, which depicts a

quasi-general overview of the topic. Indeed, authors can decide to focus more on a certain aspect than

others. As a matter of fact, this paper deals with IoT and edge computing, with special attention to

hardware features.

In particular, several efforts have been spent to ensure a rigorous approach. The one pursued in the

following sections relies on the one developed by Levy et al. [15], which includes these three steps:

1. literature collection and screening;

2. data extrapolation and examination;

3. composition of the review script.

Thus, this current overview paper presents a narrow and focused spectrum, as allowed by the flexibility

of the narrative review approach.

2. Definitions and Motivations

The number of IoT devices is constantly increasing for several different reasons: low production

cost, pervasive electronics, and technology in daily life, availability of wireless and wired

communication networks, etc..

Since these devices are spreading all over, ethic and practical questions had been posed related to

energy footprint and sustainability. Indeed IoT systems are often battery powered devices able to

acquire and send data via wireless transmission. This means, from an architectural point of view, that

these devices must be energy efficient, thus requiring low power solutions.

In order to meet all these constraints, IoT systems must be composed of low power sensors, such

as MicroElectroMechanical Systems (MEMS), while data are elaborated by means of low power

MicroController Units (MCU) and then transmitted. Generally, data are sent as once-in-a-while

packets, by exploiting low energy radio transmitters/receivers based on new technologies, such as the

UWB communication.

Since transmission is the most power consuming task among the three steps described above, many

techniques have been developed in the past years to make it as low-power as possible. One of the

most promising techniques, compressive sensing (described in section 3.7), represents a very efficient

solution, which allows to directly reduce the amount of data collected by sensors. As a consequence,

the power needed to transfer data to the cloud for post-processing is much smaller than conventional

Nyquist-rate sampling. The energy budget for IoT nodes usually consists of small battery sources or

energy harvesters; thus, the ability to reduce the amount of data to be transmitted (and consequently

the power) plays a key role in today’s systems. Alternative approaches have been proposed in the

literature to reduce the rate of data to be transmitted, by relying on bio-inspired techniques.

Many of modern MCU architectures cannot deal with the power budget imposed by IoT nodes, but

even in this strict scenario, researchers developed a broad spectrum of energy saving approaches,

such as drastic voltage and frequency scaling. Near-threshold operation of embedded transistors

introduces limitations such as performance degradations but still, for certain applications, it represents

a very promising solution. Sometimes, for some particular cases, a dedicated hardware approach
such as Application Specific Integrated Circuits (ASICs) could be preferred in order to avoid the power consumption of the general purpose MCUs. The main drawback is the loss of flexibility of the architectures. This is overcome by the use of semi-specific processors, called Application Specific Integrated Processors (ASIPs), that are designed so that recurrent application-specific operations (for example convolution or bit-wise operations) are accelerated with dedicated hardware. The overall processor is not as specifically tailored on the whole application as in ASICs in order to remain flexible and future-ready.

Figure 2. Comparison between ASIC and ASIP architecture.

This approach leads to processors which usually require higher power than ASICs but are able to adapt to the evolution of standards and technologies without re-designing and producing a new integrated circuit, which cost is a major limitation for the IoT. Ideally, IoT hardware should be at the same time optimized but not too application-specific in order to cut the non-recurrent manufacturing expenses by producing millions of small and cheap integrated circuits. This is a very difficult task, as explained in the following sections. These constraints drive IoT manufacturers towards older silicon technology processes, from 200 nm to 130 nm. These particular nodes are very effective in this field as they have been proven to be mature, cheap, mixed-signal capable, embedded-flash capable, low power and, thus, they are considered a low-risk option. Smaller technological nodes are still needed for high-performance applications, where very fast execution is required.

3. Ultra-Low-Power MCU Architectures

This section introduces the main hardware components that compose a typical IoT node, or more in general, a typical electronic system. Taking into account the power and energy requirements and constraints introduced in Section 2, the following description is oriented to low power solutions. The main component is the Central Processing Unit (CPU), typically on board of an MCU, equipped with some memory. These two component are tightly coupled since the former elaborates data coming from the latter. In order to make things easier and faster, the memory is organized in a hierarchical way, in which smaller and faster components (such as flip-flops) are directly integrated into the CPU, while larger and slower components (such as flash memories) transfer data towards the faster ones. Usually, data to be saved into the memory come from peripherals that are represented by sensors, transceiver modules for communication or, more in general, by connectors.

The steps executed in an embedded application can be roughly represented by fetching data from sensors, elaborate them via the MCU and finally, transmit the results by adopting wireless communications, like Bluetooth, Mobile Networks, Wi-Fi, Zigbee, Z-Wave, LoRaWAN, custom transmission protocols (like IR-UWB) and Light-based systems as. Since this sequence of tasks is performed periodically, energy savings can be obtained by switching off key electronics components when not needed. This process is called duty cycling and consists of waking
up the device only when it has to perform a task, while the rest of the time it operates in a deep sleep mode. During the wake-up phase, generally triggered by an external event, the MCU status is restored, together with the power supply and the clock signal. Once the task is performed, the MCU saves its current state before transitioning again in the low power state.

This behavior is predictable and so the main system can wake up and put to sleep the various actors, including itself, by resorting to well-defined power management techniques.

A general system architecture and organization has been depicted above; now, in order to better understand the details, the following sections will analyze each component listed above. First of all, peripherals will be explained by giving examples of the newest technology in use today. Then the paper will focus on memory types, power management techniques, and data processing, respectively. A final section is devoted to CPU architecture and its evolution through the years.

3.1. IO Architecture

In IoT devices, peripherals are fundamental in order to connect several sensors or external devices. MCUs are equipped with some serial interfaces such as UART (Universal Asynchronous Receiver/Transmitter) [25], SPI (Serial Peripheral Interface) [26] and I2C (Inter-Integrated Circuit) [27]. However, currently, also high bandwidth connections are required, like USB (Universal Serial Bus).

Generally, the CPU has the control of the peripherals, handling associated events and data transfer, but since systems are becoming more and more complex, peripheral subsystems have become smarter, having, sometimes, their own control unit. Peripherals can run even while the CPU is in deep sleep mode and they are able to manage their own power status independently of the CPU, by switching off unused parts. A peripheral can be woken up by events, thus, data transfer can take place without any action from the CPU, thanks to dedicated memory management units.

A not negligible amount of energy is required by off-chip communications, especially for high-speed interfaces, like Double Data Rate (DDR) dynamic RAM memories, wired data transfer (like USB and Ethernet) and video interfaces. IoT nodes tend to be tailored to their application, optimization is indeed the key to create systems able to survive for decades with an extremely low power budget. Nevertheless, there are still applications where high performances are required. This translates in the necessity of low power standard interfaces. The key player in the definition of these standards is the MIPI (Mobile Industry Processor Interface) Alliance [28], founded in 2003 by Samsung, Nokia, Intel, Texas Instruments, STMicroelectronics, and ARM. MIPI interfaces are optimized for low power, high bandwidth and low electromagnetic interferences. The MIPI Alliance works on defining standards for the physical layer (PHY), protocols for multimedia (cameras, displays, audio and touch peripherals), chip-to-chip and inter-processor communications, management of low-speed devices, power management, debugging tools and software integration. As an example, the latest MIPI Display Serial Interface (MIPI-DSI-2) is able to handle very high resolution displays, also thanks to video compression, by reducing the power spent for the transmission of data to the screen and so off-loading the reconstruction of the video stream to the display controller. Another example of MIPI low power, high throughput design is the I3C interface, presented as the successor of the I2C. I3C features a high clock speed and can work at double data rate regime; moreover, it features high power efficiency with respect to its predecessor, as shown in figure 3.

If no inter-chip communications are required, the only possible power saving technique for peripherals comes from the optimization of the peripheral itself. In particular mixed signal circuits, analog to digital converters (ADC) and digital to analog converters (DAC), are required in almost every embedded application to translate the physical world measure (usually analog) to the digital domain and vice-versa. While digital circuits are inherently robust against continuous time noise, analog circuits suffer from voltage and transistor size scaling. These aspects limit the maximum excursion of the input signal and affect the linearity of active analog parts, like operational amplifiers. Several researchers have struggled to improve the performances, even with the IoT constraints, by reaching a sub-femtoJoule-per-conversion-step Figure of Merit (FOM) [29–34]. Absolute FOM alone is not always
the best indicator for low power conversion systems, as not all input signals require the same precision or speed (this aspect is addressed in Section 3.7). Furthermore, it is important to tailor a converter around the specific signal requirements to achieve the best power performances. As stated by Alioto [35], only a few physical signal types require more than 16 bits and, usually, 8 bits are sufficient for low resolution applications. Also, the converter speed can vary a lot depending on the application: from the low speeds of heart rate and temperature readings to the high speeds required by imaging peripherals. A huge amount of information redundancy (spatial and temporal) is required for classic video stream-based algorithms. Indeed, sensor-level compressive sensing for video capture will be an important step towards low power video applications [36–41].

3.2. Communication and security

This section deals with Internet communication systems and security issues arising from the IoT paradigm. Many standards have emerged in recent years to cope with the low power requirements of IoT nodes. Depending on the application, in particular on the required data rate and data range, it is possible to select the best transmission technology. Sometimes, it is possible to create a local network and to use an aggregator for sending the data to the cloud for further processing or for being stored. However, future IoT nodes will heavily depend on cellular communication, especially with the forthcoming 5G, as the prominent player [42,43]. Indeed, 5G aims to be a revolution for machine-to-machine communication, in particular in situations which were not optimally handled by previous cellular communication systems. In particular, 5G communications can be tailored around the application to improve reliability, reduce latency and energy consumption, and to increase device density. The adoption of 5G in new IoT platforms promises to enable a true pervasive fully-connected era. However, always connected devices through wireless connections can be prone to external hacking attempts, which is a major issue when dealing with sensible data and/or dangerous situations [44–49]. Attacks can involve sensors nodes to collect precious data from users, which would otherwise be unavailable for privacy or secrecy reasons [50]. These data could be used for analysis purposes or to profile users. On the other hand, attacks can involve actuator-based systems, like Autonomous Electric Vehicles [51–54], or Microgrids (small electrical sources able to better distribute the electrical power through the load) [55,56] or healthcare devices (like implantable cardiac devices) or literally every electrical item that will potentially be equipped with a network access. For this reason, a substantial amount of research is devoted to ensure connection security and to avoid eavesdropper stealing precious data. IoT devices will be built with hardware cryptographic accelerators for optimizing power consumption and latency of data transmission. These hardware accelerators can (and should) also be used for anti-tamper protection and to circumvent IP stealing attempts, encrypting the data in inter-chip communications to null reverse engineering efforts.

3.3. Non-volatile Memories

![Figure 3. MIPI I3C bus efficiency and throughput](image-url)
A big problem in edge computing is data retention during the idle mode. Current technologies do not cope well with the low voltages used during power gating and voltage scaling as the memories used are usually volatile, like Dynamic and Static RAM (DRAM and SRAM) and internal Flip-Flops (FFs) or Latch-based registers. These traditional technologies can work at high speed and are relatively easy to manufacture and integrate, but they have the disadvantage that information is lost under a certain supply voltage value. Actually, DRAM cells tend to lose the information even when fully powered due to the leakage through the cell capacitance, thus requiring cell refresh mechanism which is energy and time consuming.

Even though non-volatile integrated memories exist, like Erasable Programmable ROM (EPROM) and Flash memories, these technologies are usually slow, especially during the writing stage. This can be a serious problem, indeed, with volatile memories, part of the computation core cannot be shut down as the information has to be retained. This bottleneck is driving a lot of research towards new non-volatile memories, which have to be easily integrable, fast, small, reliable and cheap. The main new technological approaches are:

- **Resistive RAMs (ReRAMs)** [57–63], which store the information as the variation of resistivity of a thin oxide film. A current is injected in the oxide to change its structure and to modify its resistance value. It is possible to program one cell to high or low resistance and so to assign a logic value to each of these two states. This technology is compatible with the current CMOS process. Moreover, it can achieve switching speeds of up to 10 ns and it features multilevel capability. However, the current required to reset the oxide state is high, usually being difficult to integrate in the circuit.

- **Ferroelectric RAMs (FeRAMs)** [64–66], which work like Dynamic RAMs but store the information in a ferroelectric layer instead of a dielectric one. The technology can be compatible with DRAM process, but it is usually built on old processes (350 to 130 nm). Besides, this type of memory consumes power only to read or write the memory cell, which drastically lowers the consumption with respect to DRAMs. The technology is intrinsically fast, it takes about 1 ns to modify the state of the layer, indeed. Usually, the bottleneck is the electronic control, which is rather complex, like in DRAMs.

- **Phase-Changing RAMs (PCRAMs)** [67–72], in which a chalcogenide glass can change phase from amorphous to crystalline. Moreover, chalcogenide glass can hold also an intermediate state, allowing for multilevel storage. However, the cells are difficult to program, so their use is still limited. These devices are faster than Flash-based memories, in particular for writing operations as PCRAMs are feature the possibility to modify each cell individually. The drawbacks are that the cells are prone to aging (even if they are better than Flash memories) and they are susceptible to temperature variations.

- **Magnetic RAMs (MRAMs)** [73–81], which use electron spin to store information. Currently, MRAMs look a very promising solution and several researchers envision that they might replace both the main memory and the storage memory in future architectures. When reading an MRAM, a current is forced to flow near the magnetic material, and the reading operation is accomplished by sensing the polarization of the magnetic field. When writing, an external current needs to overcome the stored field to impose a new value. As a consequence, writing requires more power consumption than reading. This technology can compete with Static RAM cells speed, while presenting a much lower area utilization.

Nowadays, when dealing with voltage scaling and power gating, the information contained in a Flip-Flop (FF) is retained thanks to the use of a Non-Volatile Flip-Flop (NVFF) cell [81–95], which

![Non-Volatile Flip-Flop with shadow latch](image)

**Figure 4.** Non-Volatile Flip-Flop with shadow latch
consists on a FF helped by a balloon latch (sometimes called shadow latch) circuit that works with true ground and power supply to retain the logic level inside the FF, as shown in figure 4. Despite this approach works as required, it also increases leakage with respect to true non-volatile memories and it still requires the availability of the power supply. Aside from mass storage applications, the new technologies listed above can be integrated into the balloon part of the NVFF to make it truly non-volatile, as reported by [96].

3.4. Power Management

Power management is a feature of many CPUs and it consists in turning off or switching to a low energy regime parts of the core, peripherals or even sections of the memory hierarchy. There are many reasons to perform such optimization:

- to reduce the power consumption by excluding elements that are not involved in the current task;
- to enhance the lifetime of the battery and consequently of the embedded system;
- to tone down the noise produced by all the components forming the system;
- to reduce the effort and requirements of the cooling apparatus.

Since usually IoT nodes are represented by mobile, wearable battery-based devices, having an onboard unit able to dynamically control the energy consumption is extremely precious. Furthermore, since algorithms and tasks are performed in a sequential fashion, not all the units will be used at the same time. As a consequence, switching off these parts becomes essential for the above-listed reasons. MCUs are already equipped with some low power modes, that consist in turning the entire system in a suspended state, in which peripherals can work independently, while no operation is performed by the core. This situation is usually referred to as deep-sleep status. MCUs can enter this state when no task has to be performed, and, typically, the wake-up signal is produced by an internal timer (deep-sleep for a known period) or by an external interrupt (event wake up).

The aforementioned mode has become very popular in the embedded systems community, despite it is not always effective with some of nowadays benchmarks. Indeed, entering and exiting the deep-sleep mode comes at certain power cost, so, depending on the application, it is not always advantageous.

Many other low power techniques exist, such as clock gating, that consists in stopping the clock from the part of the circuit that is not necessary to the current task. Since the dynamic power is related to the internal activity of combinational circuitry, no transition occurs by disabling the clock. However,
leakage current still persists, but, since in general the dynamic power is greater than the static one, the energy saving will be consistent. Furthermore, clock gating intrinsically retains the state of the circuit, allowing to restore normal operation by just reasserting the clock. The hardware overhead needed to control the clock signal is negligible, and thanks to its fine granularity this solution turns out to be very effective. Besides, leakage power can be reduced by combining clock gating and dynamic voltage scaling, as explained in Section 3.5. However, despite a small overhead is required in the power supply management unit, the time needed to restore the normal state of the circuit is greater as a stable supply voltage is required to have the circuit behaving correctly.

The best approach to reduce power consumption is power gating, in which the supply voltage is disconnected from the circuit. In this case, there is no dynamic power consumption and only a small leakage power is present in current CMOS technology. While the hardware overhead is negligible, as in clock gating, the time to restore the previous status of the circuit could be significant. This aspect must be kept into account since it may affect system performance. In particular, the power gating mode must be entered and exited in a safe way, in order to avoid damaging the circuit and to ensure a correct behavior.

As depicted in figure 7, the hardware overhead of power gating consists of an MOS transistor to be applied between the logic circuit and the supply line. Generally, a header and a footer are applied in order to completely insulate the circuit. MOS transitors behave like switches, namely, when the MOS transistor is open, the voltage is no longer applied to the circuit. As a consequence, the circuit is in a frozen state and no power dissipation occurs.

3.5. Near-Threshold MCU Architectures

Generally, MCUs work with a power supply voltage well above the threshold voltage of transistors. Nevertheless, the power supply voltage can be scaled during deep sleep mode in order to reach sub-threshold condition, as shown in Fig. 6. Indeed, smart voltage scaling to make transistors working in the near-threshold region, leads to a new low-power era. However, making a circuit working in the near-threshold region is a complex task as reliability problems and performance degradation can arise due to several factors, including fabrication process variations.

Indeed, the behavior of a circuit can degrade due to sensitivity to process variation (i.e. channel length, doping concentration, ecc.), voltage and temperature (PVT). PVT compensation requires special circuits to work correctly.

It is worth noting that special circuits are embedded within the main circuit, so they are exposed to the same conditions of the main circuit, i.e. aging, high temperature and current. In order to be effective, these special circuits must probe the current and provide feedback, by adjusting the power supply to prevent unpleasant problems such as meta-stabilities. Common implementations of probing subsystems are canary circuits and razor flip-flops.

Canary circuit is a replica of the critical path that is monitored in order to adjust the power supply. Though being a simple approach, it only provides information about global process variations, while no local information can be evinced as well as local PVT conditions, this is due to the fact that the real critical path is placed somewhere else.

Razor flip-flop approach relies on lowering the supply voltage until a critical point. Working so near to the limit, errors can occur due to time constraints violation. However, no error is propagated thanks to shadow flip-flops that can restore the correct value. Shadow flip-flops are scanned using a delayed clock signal that preserves their integrity and errors are detected by comparing these value to the one in the real critical path. Global and local variations are both considered in this solution since the device works in borderline conditions. In order to correctly apply razor flip-flops feedback, the designer must have access to the low-level circuit and this is not always possible, especially when dealing with externally engineered cores with no additional information than the top level interface protocol.

In contrast to voltage supply control, transistor body biasing represents an alternative to PVT compensation. The main advantage of this solution is that adjusting transistors threshold only modifies
the leakage component, whereas the above approaches impact both on leakage and dynamic power.
Moreover, body biasing is very effective when working near-threshold, as it features a more efficient
and simpler circuit for polarization of p-well and n-well regions than DC-DC regulators for the supply
voltage.

3.6. Data Processing
The term Data Processing is generally used to indicate the collection and then the manipulation of
data in order to extrapolate meaningful information. Sometimes it can also indicate the transformation
of data in an easy to handle format.
When related to hardware systems, data processing involves the CPU, as it must fetch data from
memory or sensors, process them (generally through an application specific algorithm) and finally
store the results in memory.
This data processing technique is important for IoT as it can significantly reduce the amount of data to
store and transmit. IoT nodes can perform a pre-processing in-locio, discarding useless information,
releasing part of the burden from the central unit that is in charge of performing the complete
elaboration.
From a data processing point of view, energy saving can be obtained by increasing the parallelism,
namely the amount of data per cycle that can be handled by the CPU or by optimizing the architecture
with respect to the per-cycle power consumption. MCUs moved from initial 8 bit to 32 bit of today’s
most common devices (like ARM Cortex-M), but still engineers are focusing on reducing the power
per instruction metric since in many application is not required to handle multiple data concurrently.
In MCUs it is common to find optimizations for single cycle multiplication, like Multiply and
Accumulate (MAC) dedicated instructions and arithmetic control support. From the instruction
side, some improvements have been introduced as well, such as defining a set of instructions with
reduced size (such as from 32 to 16 bit), saving energy in the storage and read process from memory.
A very effective way to save power by processing data is through neuromorphic or quasi-digital
approaches. As an example in [97] an Address Event Representation (AER) is adopted. This example
relies on a neuromorphic event-based approach in which the only information to be transmitted is the
source ID when a new event occurs. As a consequence, only when a change in the physical quantity
is sensed, a certain amount of data is produced and sent. Indeed, AER [98,99] allows representing
information very efficiently, by exploiting the neuromorphic approach to reduce the switching activity
and thus the power needed to process and send acquired data is significantly lowered. Stemming from
this concept, [97] proposes an architecture based on the openMSP430 processor that compared with its
regular implementation can save up to 50% of power.

3.7. Compressive Sensing
As briefly introduced in Section 1, not only
architectural optimizations but also algorithmic
optimizations are required to efficiently exploit the
available energy. Usually, the energy bottleneck in
ultra-low power systems is the communication subsystem,
which can be in certain applications a very power-hungry
block, depending on the communication standard and
the hardware architecture used. Many algorithms have
been proposed to manage meshes of a large number of
small interconnected sensor nodes in order to limit the
radio range [100–108] but even in this scenario, it is not
possible to just send the whole measured data to the cloud
for post-processing. This is energetically too expensive and
the burden of data to be processed is (and increasingly will

Figure 8. Compressed data can be directly obtained while acquiring the signal.
be) a major problem in the IoT context. To cope with this problem, a large research effort has been dedicated to the compressive sensing paradigm. Indeed, it is possible, given that one can find a domain where the signal representation is sparse, to undersample the signal, i.e. not work at the Nyquist frequency. By approximating the signal shape it is possible to retain the relevant information and send it instead of the whole ADC reading, or even work directly with signals compressed in the analog domain as in Analog to Information Converters (AIC) [109–117]. In particular, this is possible only when dealing with signals, which waveform can be efficiently approximated resorting to a change of basis and/or by working on the transformed signal (Discrete Fourier Transform, Discrete Sine and Cosine Transforms, Wavelet Transforms and others). If in one domain the signal, or a faithful representation of it, can be compactly encoded, it is possible to send only its representation. For example, a sinusoidal signal in the time domain requires many samples to be represented correctly while in the frequency domain it is fully characterized by a single complex number representation. Even if periodic signals are easy to compress, it is also possible to resort to dictionaries shared by the node and the cloud as proposed in [118]. Indeed, in [118] authors cleverly set up a dictionary with redundancies to form a custom basis set for the signal representation, which adapts its content with respect to the signal shape.

If there is no need to reconstruct the signal but only to extract some features, a different approach can be exploited in order to isolate and send features. As an example in certain bio-inspired applications, it is important to detect only some changes in the amplitude of the signal. This detection can be obtained by a thresholding mechanism, so that when the signal crosses the threshold, the system generates a pulse or an event. In systems where only the event is important, it would be a waste of resources to send the whole signal. In [119] an event-based bio-inspired pseudo-neuromorphic approach is proposed. This system aims to mimic the behavior of neurons, by sending a pulse only when the signal crosses a defined threshold. [120–123] shows how this approach can be effectively coupled with an impulse radio UWB communication system. Indeed, in such a system pulses behave as a spread spectrum signal that can be effectively sent with a very low power consumption and without further processing. This approach works well with low frequency signals, which information can be predicted and where an event is an anomalous deviation from the expected behavior, which is conceptually similar to an asynchronous delta-sigma modulation.

In [124] different types of compressive sensing and reconstruction algorithms are listed, along with a list of successful compressive sensing applications in the fields of imaging, biomedical, communications, pattern recognition, audio and video processing, dimensionality reduction, and very-large-scale integrated (VLSI) systems.

3.8. From Single Core to Multi Core

Energy per instruction combined with increasing frequency are pushing designers toward the multi-core domain. Even though in many applications single core MCUs would be enough, the IoT trend is demanding more and more computational power. By splitting the workload on more low-frequency simple cores, it is possible to increase the parallel computation maintaining a low power budget. Since multi-core solutions are more complex than traditional MCUs, they introduce several new problems regarding data and connection management. Cache coherency is fundamental in such systems: when a core produces a new write operation, all the caches that contain this variable must be invalidated. Such coherency is maintained by a software locking mechanism that synchronizes data, preventing the invalidation of caches. In fact, every time a cache must be trashed, a lot of energy is wasted in order to flush the pipe of the core, restore cache and recompute previous instructions. Data transfer between cores is power demanding and requires a very complex connection infrastructure. A complex architecture, such as the multi-core one, translates to expensive silicon and power resources; however, many applications cannot be supported nowadays by a normal MCU. That is the reason why
the trend is moving from single core to multi-core, still maintaining low power budget by limiting the memory hierarchy.

Common solutions present heterogeneous cores, in which one core is in charge of the heavy computation, while another core, usually smaller than the previous one, handles the peripheral requests. Since the two cores handle different tasks, they are decoupled, allowing for better energy management since each of them can be turned off completely.

Figure 9 shows the trade-off between single core and multi-core architectures from a power-per-operation perspective. For high workload, a multi-core approach is more efficient. On the other hand, in near-threshold or deep-sleep cases, power leakage is dominant and multi-core, being composed roughly by twice the number of transistors (in case of dual-core), is less energy efficient. However, a multi-core approach is convenient to reduce the active period by a factor equal to the number of cores, when working with a duty cycling behavior in which the MCU is active for a certain period and in deep-sleep for the rest of the time. In this way, dynamic power is reduced by increasing the deep-sleep period.

Introducing a more efficient power management system it is possible to reduce both dynamic and static power by handling each core separately. Dynamic Voltage and Frequency Scaling (DVFS) can be performed on each core, enhancing energy savings and decoupling their working points. Being able to operate each core independently, as well as shutting down them selectively, represents the best low power configuration. However, this needs a hardware overhead and an increased complexity to handle data that cross different frequency and/or voltage domains, requiring handshaking operations. Such cost is not always affordable so, sometimes, simpler solutions could be adopted like power gating. Cores would be subjected to the same voltage and frequency domain, but still capable to be switched off independently. This solution requires a ring of p-type MOS transistor all around the core, and being p-type MOS transistors larger than n-type MOS ones, the area overhead is not negligible. Furthermore, this ring introduces some performance degradation, affecting pull up resistance and consequently the current drain. However, power leakage is strongly reduced, still considering that the core requires a non-negligible time to be restored.
3.9. Memory Hierarchy for Multi-Core Domain

Multi-core systems introduced a new problem: parallel memory hierarchy and management. Since CPU computational power started increasing, memory always represented a bottleneck for both speed and energy consumption. Indeed, CPUs can process data at a higher rate than memories. As aforementioned, multi-core domain requires smart memory management in order to avoid cache invalidation. A schematic view of a typical multi-core system is shown in Figure 10. This type of coherency represents a limit to the number of cores.

IoT is a heterogeneous environment, so it is not possible to refer to a specific architecture. The device architecture is based on the type of application and data dependency. So the memory hierarchy must be designed taking into account the final application and dataflow.

The common multi-core memory hierarchy is composed of two levels, the first one is private and is placed inside the single core, while the second layer is common to all the cores. The first one is smaller and is very energy efficiency, leading to a lower hit latency, but it presents coherency problems. Moreover, when the cores are working on the same dataset, there is the need for data replication that translates to energy wasting. The second level is bigger and stores data in a fixed position in order to maintain integrity. Due to this characteristic, the hitting process is slower, but no data replication is required.

In order to remove coherency problems due to local storage, different types of L2 caches have been developed. Loi et al. [125] present a model of L2 cache shared among different cores that behaves as a final memory stage. The final implementation of the innovative architecture on a 28 nm technology showed a very high bandwidth and low power consumption.

Researchers conducted many studies on memory hierarchy performance for multi-core platforms, presenting some innovative solutions. Farimah R. Poursafae et al. [126] proposed a different approach in which Non-Volatile Memories (NVM) are used instead of conventional static or dynamic RAM for producing a low power hierarchy storage. They exploited the NVM property of consuming less static power while presenting a higher density. Based on the fact that NVM presents a limited lifetime, they proposed a memory management-aware method able to allocate data based on the access patterns defined at compilation time. This approach reduces power consumption but presents low adaptability to the final application and a lower throughput with respect to RAM technology.

Johannes Ax et al. [127] compared different architectures in their work, presenting a new tightly coupled shared data storage with respect to each core cluster. The result shows an improvement of roughly 20% respect to other solutions, tested on 10 different applications.

![Figure 10. Memory management of single-core and multi-core systems](image-url)
4. Example of a Many-Core Low-Power processor: PULP

In this section, a practical ultra-low power many-core architecture is presented as an example of all the topics discussed above. The device under examination is the so-called Parallel-processing Ultra-Low Power Platform (PULP)[128] developed by the Department of Electrical, Electronic and Information Engineering of University of Bologna and the Integrated System Laboratory of ETH Zurich. Figure 11 shows the internal structure of the system, including peripherals, several bus types, memory hierarchy and computational logic. The structure is defined as a System on Chip (SoC) composed by clusters of cores featuring a lot of different functionalities. Each core represents a single Processing Element (PE) of the cluster.

In order to avoid cache coherency problems, no local data storage is included in the single PE, instead, a tightly coupled data memory (TCDM) is included in the cluster and connected to all cores by means of multiple parallel ports. The TDCM is split in banks, providing concurrent access through as many ports as the number of banks. A DMA allows to move data from L2 memory to TDCM with minimal energy cost and in the meantime, it is on charge of bridging with other clusters or peripherals.

Outside of cluster domains, an L2 memory provides processing data to clusters and peripherals allowing the SoC to interface with the rest of the world. As exposed by Conti et al. in [128], PULP can operate in two different modes: stand-alone or slave. In the former mode, a flash memory can be connected on the SPI interface allowing the SoC to draw data from it, or from the L2 storage otherwise. In the second mode, instead, PULP is seen as an accelerator that must be coupled with an external processor that is on charge of loading data in the L2 by means of the SPI and synchronize the elaboration by means of dedicated signals directly mapped in the memory. As highlighted in figure 11, in order to improve the power management, SoC domain and cluster domain are exposed to different clock and voltage signals. Each cluster and the SoC itself are equipped with clock dividers, so it is possible to fine tune the frequency of each different part. Moreover, each core can be clock-gated to further reduce the power consumption. This allows to better allocate hardware resources to different workloads.

Since different workloads can require different computational power, a multiplexer denominated BBMUX allows to choose the back bias for each cluster, thus implementing a dynamic body bias. Transitioning from a back bias value to another is possible to pass from a normal mode to a boost one, improving the speed elaboration capability. In order to manage those transitions and make them transparent to the final user, a Power Management Unit (PMU) is introduced. This is in charge of producing all the control signals needed to control the clock gating and the BBMUX selection.

![Figure 11. PULP SoC architecture](image-url)
PULP is composed of cores based on OpenRISC and RISC-V ISA with a parallelism that can vary from 32 to 64 bits. Being synthesized exploiting a 28 nm UTBB FD-SOI technology provided by STMicroelectronics, PULP can reach 211 GOPS/W that makes him one of the most interesting open source ultra-low power core suitable for several tasks, included convolutional networks and image processing.

5. Conclusion and future perspective

This paper proposes an overview of the main techniques to design hardware platforms able to cope with IoT requirements, by exploiting the Edge Computing paradigm. As it can be observed, many approaches have a clear focus on lowering the power consumption. This is, by now, a basic requirement, since the increasing computational effort affects battery lifetime of many mobile devices. From section to section, all the fundamental components of an IoT electronic system have been analyzed. Despite the general architecture is common to different devices, the input/output part, namely peripherals, is the most variable part. In fact, sensors depend on the type of applications and sometimes the design can be very complex.

The main point addressed thought the different sections is that power consumption comes from three main sources: sensor sampling, read and store from and to the memory, transmitting and receiving information. The former is strictly related to the frequency at which data is acquired, and since the power consumption directly depends on the frequency, the faster the system, the larger the energy required. As for memory usage, this represents a huge bottleneck from the power consumption point of view, especially in systems where many accesses are required. Finally, the transmission or the reception of data is a power-hungry task as well, since the intervention of the antenna comes at a non-negligible cost.

The Internet of Things era is rapidly evolving toward a scenario where everything is connected and this requires the adoption of new design rules to adapt available technologies to the new challenges. The actual trend suggests that in the next future IoT will become more and more pervasive, incorporating all the devices around us. However, since it will be impossible to upgrade so many devices in a second moment of their lifetime, one of IoT paradigm requests is to develop now solutions able to withstand these issues. This is a difficult task also from an economic standpoint, as the processes involved to produce such systems could require to sell millions of units in order to cut down the cost for the final user and to make the IoT truly pervasive and accessible. Moreover, Big Data experts must be ready to handle an enormous amount of information never seen before, making Data Mining a fundamental tool.

From a hardware point of view, new devices must be able to accommodate the 5G technology, that will invest the entire technological community, bringing the communication to a totally different level from today, thus further widening the range of IoT applications.

Hardware must also be aware of another important and very pervasive tool, strictly related to the above mentioned Data Mining, that is Machine Learning. Models such as computer vision, speech recognition and many others are becoming more and more popular, and it is just a matter of time before they will change our everyday life. However, these kind of algorithms require quite high computing power and, consequently, a substantial amount of energy. This aspect is strictly related to two interesting research fields, on one hand algorithms must be optimized to meet the hardware and energy constraints dictated by the IoT platforms. On the other hand, researchers need to find new solutions to cope with power requirements of modern devices by working on new battery technologies and more effective energy harvesters.

To conclude, the IoT market is very dynamic and constantly evolving, resulting in an extremely appealing field both for manufacturers and inventors. Since it gathers very open-minded people, it represents a good opportunity for emerging technologies to show off their qualities, especially regarding the hardware world.
Even though this paper presents many of the main aspects characterizing the IoT world, many other points are untouched. In fact, if on one hand all the key points related to the hardware design are analyzed in detail, on the other hand, the software-based literature is still to be investigated in depth. This represents an interesting opportunity for future work.


43. Al Ridhawi, I.; Aloqaily, M.; Kotb, Y.; Al Ridhawi, Y.; Jararweh, Y. A collaborative mobile edge computing and user solution for service composition in 5G systems. Transactions on Emerging Telecommunications


© 2019 by the authors. Submitted to Future Internet for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).