Exploration of Ultra Low Power Architectures for Machine Learning at the Edge

Theo Soriano, David Novo, Pascal Benoit

Exploration of Ultra Low Power Architectures for Machine Learning at the Edge

Theo Soriano, David Novo, and Pascal Benoit
LIRMM, University of Montpellier, CNRS, Montpellier, France
E-mail: {firstname}.{lastname}@lirmm.fr

Abstract

In the field of IoT, sensor nodes have received considerable attention from both academia and industry. These small low power devices are designed to embed very simple applications, they can perform some processing, gather sensory data and communicate with other nodes in the network. However, recent advances in machine learning have made it possible to consider the implementation of smart applications in such constrained systems. In this work, we defined a basic parametric model and designed a generic microcontroller architecture to evaluate the energy profile of such applications in low-power sensor nodes.

1. Introduction

Sensor nodes are composed of three main elements: the sensor module to retrieve physical data, the communication module to send the data to the network and the microcontroller that ensures interactions between all modules and data processing. They operate at low power and cost little, which makes them good candidates for large scale deployments. The architecture of these microcontrollers is generally very simple. It includes a core capable of executing sequences of instructions that constitute the application. This core is usually connected to a set of memory arrays to store the application and data. To communicate with other modules, the architecture includes several peripherals such as UART, SPI or I2C.

Microcontrollers are subject to stringent consumption and cost constraints. They operate at low frequencies ($<$100MHz) and are usually equipped with a small internal memory (i.e., less than 1 MB). They are used in applications where little or no data processing is required. They are often used only to collect raw data and transmit it to the communication module. Often, these highly constrained systems do not operate continuously. Instead, they often operate under a normally-off computing paradigm to save as much energy as possible. They are asleep most of the time ($\geq$ 90% of their lifetime) and wake up periodically or after an interruption to collect and send data before going back to sleep.

However, the emergence of machine learning software solutions and their numerous optimisations push sensor nodes to support much more complex applications. In this paper, we present a modular exploration and evaluation tool kit for ultra-low-power architectures with machine learning applications in the normally-off computing paradigm. Section 2 introduces the concept of edge computing and states the problem, Section 3 presents a basic parametric model used to evaluate normally off applications and Section 4 presents ICOBS, a modular FPGA implementable architecture for smart applications evaluation. Finally, Section 5 concludes this paper.

2. Problem statement

In recent years, a lot of research has focused on IoT networks and has led to the emergence of new concepts such as *edge computing*. This concept is based on the decentralisation of storage and processing capacity to the edge of the network. Increasing the processing capacity of the sensor nodes allows for the communication of more meaningful information and thus greatly minimises data exchange. Thus, edge computing reduces the energy impact of network use and management as well as data storage [1] [2]. Of course, the increase in complexity of microcontrollers' missions will lead to an increase in the processing time and memory needs. Furthermore, with the emergence of tiny machine learning algorithms, it is now desirable to implement smart applications in such constrained systems, which further increases the processing phase and power consumption. So, this work is an attempt to address the following question: How can one evaluate smart applications in constrained systems operating in a normally-off computing paradigm?

3. A normally-off parametric model

To evaluate the energy consumption of a system operating in a normally-off computing paradigm, we defined a parametric model including the consumption of each module during each of its operation phases. Thus, for each module, we define a succession of operating phases with a fixed duration (see Fig. 1). The model will therefore al-

low us to build the energy profile of the communication module, the sensor module and the microcontroller.

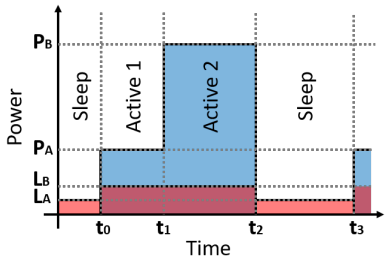

Figure 1. Module energy profile: dynamic power in blue and static power in red

We define the dynamic energy of an active phase of a particular module as:

$$E_D = \sum_{i=1}^{n} (t_i - t_{i-1}) \times P_{i-1}, \quad (1)$$

where P corresponds to the dynamic power consumption of the module during each phase. Similarly, we define the total static energy of this module as:

$$E_S = \sum_{i=1}^{n} (t_i - t_{i-1}) \times L_{i-1}. \quad (2)$$

where L corresponds to the static power consumption of the module during each phase.

We study the share of each module in the total consumption of the system for a wide variety of applications through the use of parameters defining them (e.g. activity and sleep phases duration, wake-up probability, available energy, etc.). For communication peripherals and sensors, the parameters can be the amount of data to collect from the sensor and the amount of data to send to the radio. This model will enable the quick and precise evaluation of the critical modules and phases at the energetic level for each application.

We need as input to our model the duration and consumption of each operating phase of each module. For communication and sensor modules, the parameters can be defined during application design as they depend on the modules' power consumption in each of their states and the amount of data to be exchanged. However, for the microcontroller, the model parameters depend on the execution of the application. It is therefore necessary to fully build the application before using either an instruction set simulator or, as in our case, a tool that emulates the application while capturing its characteristics.

4. Evaluation platform

We developed a typical non-optimized microcontroller architecture based on the RISC-V core Ibex developed by lowRISC. The architecture has enough memory to run a wide range of smart applications and includes peripherals to communicate with other modules (see Fig. 2). The architecture integrates an activity monitor, which will record the evolution of the state of the memories and the core. For this, we defined for each component a set of states (e.g. idle, retention, off, transitions, etc.) that correspond to a consumption and a set of events (e.g. read, write, etc.) that correspond to a quantity of energy. The monitor measures the time spent by each component in each of its states and counts the events.

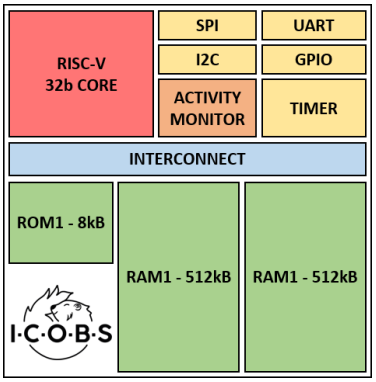

Figure 2. ICOBS architecture

We also defined the entire software development workflow to easily generate applications that can run on the platform. Thus, it is possible to execute applications in real time under realistic conditions thanks to an FPGA implementation combined with communication modules and sensors. The monitor therefore allows us to extract the activity of each part of the architecture for each application and thus make a total evaluation of the microcontroller to feed into our parametric model.

5. Conclusion and future work

The proposed architecture makes it possible to feed our parametric model with the energy profiles of the tested applications. This set of tools will allow us to better evaluate the energy-critical phases of each module at the system level and of each block at the architectural level to better guide our future works.

Acknowledgements

The authors acknowledge the support of the French National Research Agency (ANR), under grant ANR-19-CE24-0017 (NV-APROC project).

References

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the MCC Workshop on Mobile Cloud Computing*, 2012.

[2] M. Peng, S. Yan, K. Zhang, and C. Wang. Fog-computing-based radio access networks: issues and challenges. *IEEE Network*, 2016.