MPSoCs for real-time neural signal decoding: A low-power ASIP-based implementation

Paolo Meloni [a,*], Francesca Palumbo [b], Claudio Rubattu [a], Giuseppe Tuveri [a], Danilo Pani [a], Luigi Raffo [a]

[a] Dipartimento Ingegneria Elettrica ed Elettronica, Universitá degli Studi di Cagliari, Cagliari 09123, Italy
[b] PolComIng - Gruppo Ingegneria dell'Informazione, Universitá degli Studi di Sassari, Sassari 07100, Italy

ABSTRACT

In this paper we target the design of a dedicated low-power computing platform for neuroprosthetic applications. The system must be capable of decoding the information encoded in neural signals, to extract the patients' motion intention. To this aim, a highly-portable and reliable integrated processing device is required. However, a commonly acknowledged design methodology, to be used in such kind of design cases, is still not available in literature. In this work, we propose and assess the adoption of the MPSoC paradigm as a prospective solution. We present a design-case of a custom MPSoC integrated solution, implementing an on-line neural signal decoding algorithm. The proposed system executes parallel software tasks onto customized ASIP processing cores. Experimental results, obtained by placement- and activity-aware power evaluations carried out using an industrial 40 nm technology node as a reference, assess that the performance and power-related features of the designed architecture are compliant with the implantability constraints and with the battery lifetime required for real-life use. Moreover, besides the effectiveness of the proposed solution, this paper demonstrates also that custom heterogeneous MPSoCs can successfully challenge ultra-low power bio-medical signal processing problem.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Neuroprosthetics represent a challenging application of bio-engineering research. The key problem is decoding the information contained in physiological neural signals, to extract the motion intentions of the patient encoded in such electrical signals. This would allow an adequate and straightforward control of the robotic prosthesis. An emerging promising solution is represented by neuroprostheses that are controlled according to the sensing of the Peripheral Nervous System (PNS) signals [1]. Electro-Neurographic (ENG) signals are extracted by electrodes directly implanted in the stump of the patient subjected to amputation. Thus, exploiting the natural pathways of motor control, such kind of prostheses can be more easily and finely managed (and then accepted) by the amputees compared to Electromyographic (EMG) controlled ones.

Signal processing techniques have to be exploited to decode the patient's movement intention from the raw neural signal. The information that has to be extracted is encoded in *spikes*, burst of samples corresponding to an electrical oscillation fired by an active motor neuron, i.e. an *action potential*. The neural signal has to be analyzed to detect the spikes, then the activity of the single neurons has to be evaluated through *spike sorting* techniques, that look at the action potential morphology as the fingerprint of different neurons [2]. Unfortunately, spike sorting algorithms are computationally intensive, especially considering the relatively high sampling frequency required by the neural signals [3] and the possible presence of multiple channels [4]. Thus, previous research activities have almost always overlooked the problems related to the development of wearable or even implantable solutions, requiring the decoding algorithm to be executed on a highly portable embedded processing system, with very limited budget in terms of power and energy consumption. Such limitations can strongly influence the performance of the solutions and the possibility of complying with the real-time constraints posed by the application.

This work aims at making a step further in the definition of novel power-efficient methodologies for the implementation of such algorithms. The basic assumption is exploiting the intrinsic parallelism of the application to distribute the execution on a custom Multi-Processor System-on-Chip (MPSoC) and, in turn, achieving improved overall performance. We targeted a state-of-the-art

* Corresponding author. Tel.: +393281559786; fax: +390706755782.
  E-mail addresses: paolo.meloni@diee.unica.it (P. Meloni), fpalumbo@uniss.it (F. Palumbo), Claudio.Rubattu@diee.unica.it (C. Rubattu), Giuseppe.Tuveri@diee.unica.it (G. Tuveri), Danilo.Pani@diee.unica.it (D. Pani), Luigi.Raffo@diee.unica.it (L. Raffo).

PNS signal decoding algorithm [5] that reveals good performance even in case of low Signal-to-Noise (SNR) conditions, both on animals [5] and humans [6]. Custom heterogeneous MPSoCs will allow us to achieve improved computing capabilities, through parallel execution, and efficiency maximization, through macro-architectural and micro-architectural customization.

In this work a complete design case is presented, reporting on all the phases of the design flow, starting from the initial specification of constraints and assumptions until the eventual implementation on a 40 nm standard cell library, aimed at a technology-aware evaluation of the proposed design techniques applied to the target use-case. The performed evaluation allows the assessment of the usability of custom MPSoCs for the construction of implantable prosthesis controllers.

2. Related work

As already mentioned in Section 1, spike sorting techniques aim at recognizing the firing activity of the different neurons on the basis of the information encoded at different levels in the morphology of their action potentials [2]. Several works on neural signal processing have been presented in the years, the largest part of them proposing FPGAs [7,8] as implementation target to guarantee more flexibility than ASICs [9,10] with more parallelism than general-purpose processors.

In [7], a fully implantable programmable neuroprocessor mapped on a low-power nano-FPGA is presented. It manages data acquisition and reduction by particular compression techniques in order to minimize the output bitrate exploiting the sparse representation of the neural signals. This way, it is possible to overcome the limitation of the wireless telemetry bandwidth by transmitting only the samples associated to the detected spikes to an external device for cortically-controlled Brain-Machine Interfaces. The device has been tested on raw extracellular signals recorded through microelectrode arrays chronically implanted in the brain of sedated rats. The feasibility of this approach in terms of power has been investigated on standard CMOS VLSI [11]. In this approach, the computational complexity is shifted at downstream of the implantable device in order to perform the decoding which can be performed on many-core platforms [12] or FPGA-accelerated solutions [13].

Other energy-efficient implementations for multi-channel spike sorting have been published so far [14], most of them concerning the processing of signals coming from the Central Nervous System (CNS). Some of them focus on the analysis, in terms of necessary hardware resources and accuracy, of some typical processing steps of spike sorting algorithms [15,16]. In these cases, massive parallelization is in contrast with the low-power requirements. To this aim, also coarse-grained reconfigurable approaches have been presented in order to accelerate some computational intensive kernels using the smaller hardware resources set [17]. In this case, a trade-off between hardware reuse maximization and latency minimization must be carefully considered in order to fulfill the relative strict timing constraints.

Although the approach to neural signal decoding based on PNS seems to be the most attractive for the time being [18], there is a lack of studies in terms of architectures able to cope with the application constraints. In [19], the same algorithm taken as starting point for this work [5] has been partially ported on a complex VLIW floating-point processor by Texas Instruments. The claimed real-time results have been obtained on a 300 MHz processor: such an architecture, and the operating frequency, determines an excessive contribution in terms of dynamic power consumption that is not allowable in case of implantable solutions.

The methodological aspects related to power-efficient and effective multiprocessor architectures aimed at implementing in real-time state-of-the-art neural signal decoding algorithms lack in the

scientific literature. In [20], a homogeneous MPSoC architecture has been used to perform a preliminary test of the porting of a neural signal decoding algorithm on parallel processing platforms. Results have shown that real-constraints can be satisfied clocking the system at a reasonable frequency and taking profit from the parallelism to reduce power consumption using a clock-gating programmable manager. The application code has been parallelized effectively using an approach based on software pipeline.

This work extends [20] making one step further to the fine tuning of the system, introducing micro-architecture customization with the use of Application Specific Instruction-set Processors (ASIPs) as building blocks of the system macro-architecture. Such an approach is expected to outperform ASIC-based and FPGA-based solutions either with respect to flexibility or with respect to power-related features.

- The ASIC-based approach does not allow software-based programming of the system, thus the system is not as flexible as when composed by ASIPs.
- FPGA-based approaches, due to obvious technology reasons, have a power consumption significantly higher than the equivalent approach implemented on VLSI technologies.

In [21] a system, similar to the device presented in this paper, is implemented on a low-power FPGA. Reported power figures are almost two orders of magnitude higher than those that will be presented here, thus they cannot comply with the implantability requirements of a real-life use. Some instruments such as micro-FPGA devices [22] can be used to implement ultra-low power systems, but they can accommodate only limited size systems.

Finally, this work presents a detailed technology- and implementation-aware power evaluation, considering a state-of-the-art standard cell library and an industrial-strength RTL-to-routing flow. Such an evaluation assesses the feasibility of implantable MPSoCs for on-line neural-signal decoding and is presented in detail, paving the way for the identification of future potential improvements.

3. Target application and constraints

The decoding algorithm chosen in this work has been described in detail in [5] and, as already said, ported in real-time onto an off-the-shelf floating point DSP processor in [19]. We have decided to opt for it in our studies since, recently, it has also been successfully evaluated in a real scenario on human amputees. Fig. 1 provides an overview of the overall application. The proposed architecture has to be capable of reading and analyzing the neural signal samples, acquired by an adequate analog front-end. Therefore, several steps of the decoding chain have to be performed over the same substrate on a low-power miniaturized embedded system, prospectively implantable.

The chosen algorithm involves three sequential processing steps: preprocessing by *Wavelet Denoising* (see Section 3.1), *Spike Detection* (see Section 3.2) and *Spike Sorting* (see Section 3.3). The inputs to be processed are the PNS signal channels, extracted by the electrodes, whereas the outputs are the indicators, for each incoming spike, of the pertinence with a specific active neuron. Each motor neuron is associated to an average spike waveform, called *template*. Therefore, for each spike, the algorithm calculates a metric based on cross-correlation, which represents the morphological similarity with all the elements in a set of known reference templates. Finding the spike-template pair with the highest similarity metric means identifying which motor neuron has generated the detected spike. This process is actually the so called spike sorting. As soon as the spike sorting is accomplished, pertinence indicators are sent in output to the mechatronic device (see Fig. 1) to train