Non-negative Matrix Factorization on Low-Power Architectures and Accelerators: A Comparative Study ☆

Francisco D. Igual, Carlos García *, Guillermo Botella, Luis Piñuel, Manuel Prieto-Matías, Francisco Tirado

Dpto. Arquitectura de Computadores y Automática, Universidad Complutense de Madrid, 28.040 Madrid, Spain

ARTICLE INFO

ABSTRACT

Power consumption is emerging as one of the main concerns in the High Performance Computing (HPC) field. As a growing number of bioinformatics applications require HPC techniques and parallel architectures to meet performance requirements, power consumption arises as an additional limitation when accelerating them. In this paper, we present a comparative study of optimized implementations of the Non-negative Matrix Factorization (NMF), that is widely used in many fields of bioinformatics, taking into account both performance and power consumption. We target a wide range of state-of-the-art parallel architectures, including general-purpose, low-power processors and specific-purpose accelerators like GPUs, DSPs or the Intel Xeon Phi. From our study, we gain insights in both performance and energy consumption for each one of them under a number of experimental conditions, and conclude that the most appropriate architecture is usually a trade-off between performance and energy consumption for a given experimental setup and dataset.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Recent technological advances in the areas of genetics and molecular biology are the result of interdisciplinary research between diverse fields such as medicine, biology, chemistry and computer science. One of the main challenges these areas will need to face today and in the near future is the ability to process, analyze and interpret massive amounts of data generated in biological experiments.

Different data mining methods have proven to be powerful tools to obtain biological patterns. Two of the most extended techniques in large dataset analysis are *clustering algorithms* and *matrix factorization* techniques. These methods facilitate multidimensional analysis, that allows dimensionality reduction, or discovering certain patterns that dramatically reduce the burden towards biological interpretation. *Dimensionality reduction* is a key step for the effective analysis of high-dimension datasets. In this area, Principal Component Analysis (PCA) [1] or Singular Value Decomposition (SVD) [2] are two of the most popular methods. Non-negative matrix factorization (NMF) is also considered as one of the most efficient methods in the biological disclosure due to its ability to establish relevant relationships between experimental datasets.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. J. Carratero.
* Corresponding author.
E-mail addresses: figual@ucm.es (F.D. Igual), garsanca@ucm.es (C. García), gbotella@ucm.es (G. Botella), lpinuel@ucm.es (L. Piñuel), mpmatias@ucm.es (M. Prieto-Matías), ptirado@ucm.es (F. Tirado).

NMF was proposed by Brunet et al. [3] as a powerful tool for gene expression data. They applied it to cancer micro-array data to extract molecular patterns in leukemia, medulloblastoma and central nervous system tumor datasets based on consensus clustering. NMF is nowadays widely used in computational biology for molecular pattern discovery [4], supervised learning frameworks as class comparison and prediction [5], or text-mining [6]. However, the importance of NMF is not limited to computational biology but also to other areas such as image processing for image classification [7] or face recognition [8], information retrieval to extract semantic recognition or hidden meaning in text [9], web-based logs [10], or hyperspectral image processing [11], among others. However, both NMF primitives and applications based on them pose two big challenges from the computational perspective that are closely related:

Performance requirements. Factorization of large biological datasets makes NMF a very demanding task from the computational point of view that may make this processing unfeasible in practice, unless high-end distributed memory architectures are used. Due to this fact, parallel implementations of NMF have been previously proposed and successfully applied. In particular, previous works have targeted parallel implementations in shared-memory multi-core processors using the OpenMP paradigm [12], distributed-memory clusters based on Message Passing Interface (MPI) [13], and also exploiting modern Graphics Processing Units (GPUs) [14].

Power constraints. Many of the scenarios and applications in which NMF is typically used can be strictly limited by the maximum power that can be supplied to the computing platform (for example, on hand-held devices in biological in-place analysis scenarios), or maximum energy consumed (for example, in on-board hyperspectral image analysis on airborne or spaceborne missions). In such cases, a correct selection of the computing platform and its performance/power ratio is crucial to meet the application requirements in terms of power consumption.

Processor manufacturers are nowadays concerned about concepts such as *Green Computing* and *Power-aware HPC*. Their goal is to develop efficient processors not only in terms of peak performance rates, but also energy consumption and peak power draw. Besides modern and efficient multi-core CPUs, hardware accelerators such as GPUs or the recently introduced Intel Xeon Phi, specific-purpose architectures such as Digital Signal Processors (DSPs), or ultra-low power processors like ARM-based architectures or hybrid architectures using low-power ARM CPUs and programmable GPUs have been recently introduced in the High Performance Computing field in order to meet at the same time performance and power consumption requirements.

The suitability of this type of architectures is primarily dictated by the specific characteristics of the target application, and more precisely, by the efficiency of the underlying software building blocks (e.g. scientific libraries) on top of which those applications are built. As an example, for many biological applications based on NMF, the performance and power efficiency attained on different architectures for this specific operation will dramatically determine the behavior of the overall application. In this paper, we perform a comparative study of several state-of-the-art parallel architectures in terms of performance and power efficiency for a well-known NMF implementation proposed by Brunet et al. [3]. This choice is motivated by the high acceptance of these implementations, but does not exclude the extension to other implementations as Least Squares NMF [15] or Chih-Jen Li approach [16]. To the best of the authors' knowledge, there does not exist either any viability study of NMF under power consumption constraints in the literature or optimized implementations for novel architectures like multi-core DSPs, low-power ARM or Intel Xeon Phi. This paper extends the insights already extracted in [17], with the following new contributions:

- We target novel, state-of-the-art architectures not covered in [17], including ARM Cortex A7 and A15, Nvidia Kepler and Intel Xeon Phi, to adapt the study to new architectures recently emerged in the HPC arena.
- We extend the `pmlib` framework to gather actual power measurements from the most representative architectures among the evaluated ones.
- For these architectures, we extend the power efficiency study already presented in [17] with actual power consumption measurements, describing the measurement environment used in the evaluation process and discussing the suitability of each platform for this specific problem based on realistic considerations.
- We propose new experimental results for single and double precision arithmetic, and discuss the feasibility of each architecture depending on the desired accuracy.

The rest of the paper is organized as follows. Section 2 presents an overview of the selected architectures that are employed to evaluate the performance and power efficiency of our approach. Section 3 introduces the basic concepts of the NMF factorization, together with the specific optimization techniques applied to adapt them to each platform. Section 4 reports detailed performance and power efficiency results and discusses the feasibility of each target architecture for a given experimental condition. Section 5 moves through the description of the power measurement environment used to evaluate the power efficiency of the solutions using realistic power data. Finally, Section 6 closes the paper with some general concluding remarks and future work proposals.

2. Target architectures

In this section, we present the main features of the architectures used in the evaluation of performance and power efficiency of the selected NMF implementations. Four of them (ARM Cortex-A7/A15, CARMA board and DSP) are inherently