INSTITUTO DE ESTUDIOS DE POSTGRADO
PROGRAMA DE DOCTORADO EN BIOCIENCIAS Y CIENCIAS AGROALIMENTARIAS

MANY-CORE BIOINFORMATICS PLATFORM: DEVELOPMENT AND OPTIMIZATION

PLATAFORMA BIOINFORMÁTICA MULTINÚCLEO: DESARROLLO Y OPTIMIZACIÓN

FRANCISCO JOSÉ ESTEBAN RISUEÑO

LÍNEA DE INVESTIGACIÓN: BIOTECNOLOGÍA AGROALIMENTARIA

MEMORIA DE TESIS PARA OPTAR AL GRADO DE DOCTOR

DIRECTORES:

DR. D. GABRIEL DORADO PÉREZ
DR. D. JUAN ANTONIO CABALLERO MOLINA
DR. D. SERGIO GÁLVEZ ROJAS

CÓRDOBA, FEBRERO DE 2014

TITULO: *Plataforma Bioinformática Multinúcleo: Desarrollo y Optimización*

AUTOR: *Francisco José Esteban Risueño*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

To Aurora

Acknowledgments

My first words in this acknowledgments section are for my family: My parents, Francisco and María Cristina, always had our education as their highest priority. My wife Aurora and my daughters, Aurora, Cristina and Patricia, have supported this effort for many years; I would have no results without them.

My supervisors, Professors Gabriel Dorado, Juan Antonio Caballero and Sergio Gálvez, have oriented my work patiently, providing their best effort and experience. This has been determinant to the success of this thesis, as my labor situation has prevented a continuous dedication to it.

This has been a collective work: the research group members Pilar Hernandez and David Díaz have been essential components in every work we have undertaken. This is especially true with David, whose talent and dedicated work has allowed to overcome all the difficulties. I expect I can help him in his further projects.

Research funding has been limited in our country, so we have to be grateful to our administrations: Gobierno de España – Ministerio de Economía y Competitividad, Junta de Andalucía – Consejería de Innovación, Ciencia y Empresa and Universidad de Córdoba – Programa de Ayuda a Grupos de Investigación, for supporting our research.

We are grateful to Tilera for providing the hardware and software tools used for these developments.

Finally, my former supervisor, Professor Rafael Medina was an excellent mentor in my initial doctorate studies, and he had the generosity of letting me accept the project that now concludes.

Table of Contents

Introduction

Molecular biology has experienced remarkable advances in last years, especially in genomics and proteomics. These developments have been produced by the technical improvements in the chemistry and instrumentation, as well as the bioinformatics software used by life scientists, including genome and proteome sequencing, assembly and analysis tools. Thus, along with the solely technological improvements, manifested in better equipment, capable of reading the nucleic acids, including the DeoxyriboNucleic Acid (DNA) and the RiboNucleic Acid (RNA), as well as the peptides (eg., proteins) faster, with higher throughput and at a fraction of cost of their predecessors, a methodological refinement has taken place, with better algorithms, databases and programs that take the most out of the state-of-the-art in computing resources.

On the other hand, the evolution in microprocessors has entered a point through past years in which there is little margin for further improvements by increasing just the Central Processing Unit (CPU) clock frequency, so more computer power is achieved by grouping processor elements together, both in loose and tight aggregations. At the loose level, network computing is now a well-established paradigm with different approaches like cluster computing, grid computing or cloud computing. At the tight level, the presence of various processor elements in a single computer has evolved from the first multi-CPU systems to the package of such processors in a single die, both in the CPU, by building multi-core and many-core systems, as well as in auxiliary processor elements, specially the Graphic Processing Unit (GPU).

In this context, our research group has used the first commercially available many-core CPU microprocessors, the Tile64 by Tilera, including 64 tiles (cores) in each microprocessor die. From this point on, the work developed during this thesis has been centered in getting the most from this platform in the field of bioinformatics, by means of parallelizing and optimizing some bioinformatics algorithms to this architecture. This research line gives our thesis its thematic unit, reflected in every generated publication.

The Tile64 microprocessor is shipped both in stand-alone solutions and in expansion cards to be used in conjunction with a Personal Computer (PC). We have used the last configuration, by means of TILExpress-20G cards, which pack Tile64 microprocessors with 8 Gigabytes (GB) of Random Access Memory (RAM) and two Ethernet ports following the 10GBase-CX4 specification, so bringing an aggregate communication capacity of 20 Gigabits per second (Gbps).

The TILExpress-20G cards have the ability to run a separate Linux operating system in each core, in which programs compiled for its Reduced Instruction Set Computer (RISC) architecture can be launched independently of the PC execution, although there are several interaction possibilities with the host PC. The intrinsic parallelism of the Tile64 microprocessor can be exploited by the regular Unix mechanisms and, especially, with the libraries provided by the manufacturer, which includes functions for program spawning or replication, execution coordination and various inter-process

communication techniques, like shared memory, channel communication or message passing.

Two major phases can be viewed in our research. In the first one, we have worked with isolated single-card solutions, so dealing with independent 64-core parallel environments built in regular PC machines. A further phase has involved getting several cards connected through their built-in high-speed Ethernet ports, constructing a computing cluster of any potential number of processors. In our developments, we have taken into account the usual considerations that the evaluation of a new architecture implies, from performance factors to ease of use, taking special care of methodological aspects, both from the biological and the informatics points of view; thus, this is the bioinformatics approach of this thesis.

Hypothesis and goals

The main goal of this work has been to deploy and study a massively parallel bioinformatics platform, based on the Tile64 microprocessor, further connected through a 10 Gbps network. These developments are based on the hypothesis that such a platform could provide the computer power needed by nowadays bioinformatics tasks in a desktop PC, thus effectively allowing algorithm performance gains, easy portability of some existing programs, and therefore providing the life scientists with a powerful tool for new relevant discoveries from a biological point of view.

The publications included in this compendium show the platform construction progress and the obtained results. The first article presented, published in the *Bioinformatics* journal shows a first approach of a well-known pairwise-alignment implemented from scratch in the platform, with a significant speedup against alternative developments. As a complement of this publication, the book chapter included in the *5th International Conference on Practical Applications of Computational Biology & Bioinformatics* shows the web interface developed to provide an intuitive access to the system. The article published in the journal *Parallel Computing* deeply analyzes the first implementation, centering the discussion in the informatics point of view.

The book chapter published in the *IT Revolutions; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST)*, introduces other bioinformatics tasks like multiple alignment, de-novo genome assembly and peptide (protein) folding. The article in the journal *Future Generation Computing Systems (The International Journal of Grid Computing and eScience)* discusses the results generated when porting some well-known algorithms to the Tile64 platform.

As a final stage in the bioinformatics algorithm development, the article in the *PLoS ONE* journal discusses a novel implementation of one of the most used multiple-alignment algorithms: ClustalW, in which both the first and third stages are parallelized, and a novel hybrid computing approach is exploited. The last phase of this work is treated in the two last publications, dealing with the cluster construction and its performance evaluation.

Publications

Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment

Authors: Galvez, S, Díaz, D, Hernández, P, Esteban, F.J., Caballero, J.A. Dorado, G
Type of publication: Journal article
Year: 2010
Journal: Bioinformatics
ISSN: 1367-4803
Volume: 25
Issue: 5
Pages: 683-686
DOI: 10.1093/bioinformatics/btq017
Abstract: Motivation: Bioinformatics algorithms and computing power are the main bottlenecks for analyzing huge amount of data generated by the current technologies, such as the 'next-generation' sequencing methodologies. At the same time, most powerful microprocessors are based on many-core chips, yet most applications cannot exploit such power, requiring parallelized algorithms. As an example of next- generation bioinformatics, we have developed from scratch a new parallelization of the Needleman–Wunsch (NW) sequence alignment algorithm for the 64-core Tile64 microprocessor. The unprecedented performance it offers for a standalone personal computer (PC) is discussed, optimally aligning sequences up to 20 times faster than the non-parallelized version, thus saving valuable time. Availability: This algorithm is available as a free web service for the scientific community at http://www.sicuma.uma.es/multicore. The open source code is also available on such site. Contact: galvez@uma.es Supplementary information: Supplementary data are available at Bioinformatics online.

MC64: A web platform to test bioinformatics algorithms in a many-core architecture

Authors: Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S.
Type of publication: Book chapter
Year: 2011
Book title: 5th International Conference on Practical Applications of Computational Biology & Bioinformatics
ISBN: 978-3-642-19913-4
Editors:
Rocha, M.P.
Rodriguez, J.M.C.
Fdez Riverola, F.
Valencia, A
City: Berlin
Publisher: Springer-Verlag Berlin
Volume: 93
Pages: 9-16
Accession Number: WOS:000291366100002
Abstract: New analytical methodologies, like the so-called “next-generation sequencing” (NGS), allow the sequencing of full genomes with high speed and reduced price. Yet, such technologies generate huge amounts of data that demand large raw computational power. Many-core technologies can be exploited to overcome the involved bioinformatics bottleneck. Indeed, such hardware is currently in active development. We have developed parallel bioinformatics algorithms for many-core microprocessors containing 64 cores each. Thus, the MC64 web platform allows executing high-performance alignments (Needleman- Wunsch, Smith-Waterman and ClustalW) of long sequences. The MC64 platform can be accessed via web browsers, allowing easy resource integration into third- party tools. Furthermore, the results obtained from the MC64 include time- performance statistics that can be compared with other platforms.

Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture

Authors: Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Dorado, G, Gálvez, S.
Type of publication: Journal article
Year: 2011
Journal: Parallel Computing
ISSN: 0167-8191
Volume: 37
Issue: 4-5
Pages: 244-259
DOI: 10.1016/j.parco.2011.03.003
Abstract: Current computer engineering evolves at an accelerated pace, with hardware advancing towards new chip multiprocessors (CMP) architectures and with supporting software gear- ing towards new programming and abstraction paradigms, to obtain the maximum effi- ciency of the hardware at a low cost. In this context, Tilera Corporation has developed a brand new CMP architecture with 64 cores (tiles) called Tile64, and has launched several Peripheral Component Interconnect Express (PCIe) cards to be used and monitored from a host Personal Computer (PC). These cards may execute parallel applications built in C/ C++ and compiled with the Tile-GCC compiler. We have previously demonstrated the use- fulness of the Tile64 architecture for bioinformatics [S. Gálvez, D. Díaz, P. Hernández, F.J. Esteban, J.A. Caballero, G. Dorado, Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment, Bioinformatics, 26 (2010) 683–686]. We have chosen a bioinformatics algorithm to test this many-core Tile64 architecture because of actual bioinformatics challenging needs: data-intensive workloads, space and time-consuming requirements and massive calculation. This algorithm, known as Needleman–Wunsch/Smith–Waterman (NW/SW), obtains an optimal sequence alignment in quadratic time and space cost, yet requires to be optimized to take full advantage of computing parallelization. In this paper we redesign, implement and fine- tune this algorithm, introducing key optimizations and changes that take advantage of spe- cific Tile64 characteristics: RISC architecture, local tile's cache, length of memory word, shared memory usage, RAM file system, tile's intercommunication and job selection from a pool. The resulting algorithm – named MC64-NW/SW for Multicore64 Needleman– Wunsch/Smith–Waterman – achieves a gain of 1000% when compared with the same algorithm on a 86 multi-core architecture. As far as we know, our NW/SW implementa- tion is the fastest ever published for a standalone PC when aligning a pair of sequences larger than 20 kb

Many-Core Processor Bioinformatics and Next-Generation Sequencing

Authors: Esteban, F.J., Díaz, D,, Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.
Type of publication: Book chapter
Year: 2012
Book title: IT Revolutions. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering
ISBN: 978-3-642-32303-4
Editors:
Liñán Reyes, Matías
Flores Arias, José M.
González de la Rosa, Juan J.
Langer, Josef
Bellido Outeiriño, Francisco J.
Moreno-Muñoz, Antonio
City: Berlin
Publisher: Springer Berlin Heidelberg
Volume: 82
Pages: 172-188
DOI: 10.1007/978-3-642-32304-1_15
Abstract: The new massive DNA sequencing methods demand both computer hardware and bioinformatics software capable of handling huge amounts of data. This paper shows how the many-core processors (in which each core can execute a whole operating system) can be exploited to address problems which previously required expensive supercomputers. Thus, the Needleman- Wunsch/Smith-Waterman pairwise alignments will be described using long DNA sequences (>100 kb), including the implications for progressive multiple alignments. Likewise, assembling algorithms used to generate contigs on sequencing projects (therefore, using short sequences) and the future in peptide (protein) folding computing methods will be also described. Our study also integrates the last trends in many-core processors and their applications in the field of bioinformatics.

Direct approaches to exploit many-core architecture in bioinformatics

Authors: Esteban, F.J., Díaz, D., Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.
Type of publication: Journal article
Year: 2013
Journal: Future Generation Computer Systems
ISSN: 0167-739X
Volume: 29
Issue: 1
Pages: 15-26
DOI: 10.1016/j.future.2012.03.018
Abstract: Current trends in computer programming look for solutions in the challenging task of porting and optimizing existing algorithms to many-core architectures with tens of Central Processing Units (CPUs). Yet, the lack of standardized general-purpose parallel programming and porting methodologies represents the main bottleneck on these developments. We have focused on bioinformatics applied to genomics in general and the so-called ‘‘Next-Generation’’ Sequencing (NGS) in particular, in order to study the viability and cost of porting and optimizing well known algorithms to a many-core architecture. Three different methods are tackled in order to implement existing algorithms in Tile64, corresponding to a microprocessor containing 64 CPUs, each of them being capable of executing an independent Linux operating system. Three different approaches have been explored: (i) implementation of the Needleman–Wunsch/Smith–Waterman pairwise aligner from scratch; (ii) direct translation of the Message Passing Interface (MPI) C++ ABySS assembly algorithm with changes on the communication layer; and (iii) migration of the ClustalW tool, parallelizing only the most time-consuming stage. The performance-gain/development-cost tradeoffs indicate that the Tile64 microprocessor has the potential to increase the performance of bioinformatics in an unprecedented way for a standalone Personal Computer (PC). Yet, the effective exploitation of these parallel implementations requires a detailed understanding of the peculiar many-core characteristics when migrating previous non-parallel source codes.

MC64-ClustalWP2, a high parallel strategy to align multiple DNA sequences in many-core architectures

Authors: Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Guevara, A., Dorado, G., Gálvez, S.
Type of publication: Journal article
Year: 2013
Journal: PLoS ONE
ISSN: 1932-6203
Volume: ***Currently in press***
Issue:
Pages:
DOI:
Abstract: We have developed the MC64-ClustalWP2 as a new implementation of the Clustal W algorithm, integrating a novel parallelization strategy and significantly increasing the performance when aligning long sequences in architectures with many cores. The analysis of both the software and hardware features and peculiarities is required to reveal key points to exploit the full potential of parallelism in many-core CPU systems. The new parallelization approach has focused into the most time-consuming stages of this algorithm. In particular, the so-called progressive alignment has drastically improved the performance, due to a fine-grained approach where the forward and backward loops were unrolled and parallelized. We have implemented the new algorithm in a hybrid computing system, integrating an Intel Xeon multi-core CPU and a Tilera Tile64 many-core card. A comparison with other Clustal W implementations reveals the high-performance of the algorithm in many-core CPU architectures, in a scenario where the sequences to align are relatively long (more than 10 kb). The MC64-ClustalWP2 runs multiple alignments 20 times faster than the original implementation and it is publicly available through a web service. These developments have been deployed in cost-effective personal computers and should be useful for life- science researchers, including the identification of identities and differences for mutation/polymorphism analyses, biodiversity and evolutionary studies and for the development of molecular markers for paternity testing, germplasm management and protection, to assist breeding, illegal traffic control, fraud prevention and for the protection of the intellectual property (identification/traceability), including the protected designation of origin, among other applications.

MC64-Cluster: A Many-Core CPU Cluster for Bioinformatics Applications

Authors: Esteban, F.J., Díaz, D., Hernández, P., Caballero, J.A., Dorado, G., Gálvez, S.
Type of publication: Book chapter
Year: 2013
Book title: Advances in Information Systems and Technologies
ISBN: 978-3-642-36980-3
Editors:
Rocha, Álvaro
Correia, Ana Maria
Wilson, Tom
Stroetmann, Karl A.
Publisher: Springer Berlin Heidelberg
Volume: 206
Pages: 819-825
DOI: 10.1007/978-3-642-36981-0_76
Abstract: The current developments in life sciences face a big challenge, with the need of dealing with huge amounts of data and the increasing demand of computational resources, both in hardware and in software, pushing the limits of the available state-of-the-art at an affordable price. This paper introduces a computer cluster whose building blocks are the first commercially available many-core CPU systems: the Tile64 by Tilera Corporation, packed in PCIe cards (TILExpress-20G). We have developed the main software components of the cluster (resource manager and scheduler) and a communication library, in order to offer a high-performance general-purpose platform to speedup bioinformatics applications.

MC64-Cluster: Architecture of a many-core CPU cluster and performance analysis in B-tree searches

Authors: Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S.
Type of publication: Journal article
Year: 2013
Journal: Journal of Parallel and Distributed Computing
ISSN: 0743-7315
Volume: Under Review
Issue:
Pages:
DOI:
Abstract: We present the MC64-Cluster, a computer cluster built with the first commercially available many-core CPU system: Tile64. This microprocessor contains 64 RISC tiles/cores and it is boarded on a PCIe card (TILExpress-20G) that can be used in any standard PC. The architecture of the MC64-Cluster is analyzed in terms of both hardware and software, including the commands available to manage the jobs and the API provided to developers to communicate and synchronize tiles. To analyze the performance of the MC64-Cluster, we have selected the problem of the massive concurrent search of keys in a B-tree. This scenario is of particular relevance to efficiently carry out many bioinformatics tasks, like the comparison of an increasingly amount of available genomes, transcriptomes and proteomes, where both computational power and methodological optimizations can improve yields. Indeed, our results offer remarkable performance improvements when the cluster resources in this sort of architecture are combined with those available in the host machine.

General Discussion

As an initial work, a development environment was set up, consisting of three Dell Precision T5400 personal computers with an Intel Quad Core Xeon 2.0 GHz microprocessor and 8 GB of DDR2 memory. Each of them was equipped with a TILExpress-20G card. As recommended by the card's manufacturer, CentOS 5.3 and Red Hat Enterprise 4.3 Linux distribution were installed on those workstations. The software environment also included the Tilera Multicore Development Environment (MDE), an Eclipse-based platform capable of generating code in the Tilera's binary format, executing it both in the hardware and in a simulator, and optimizing it by means of an integrated debugger and profiler.

Although the programming environment was known and so were the available languages (C and C++), the first mandatory task was to learn the different parallelizing techniques available and their associated functions in the provided Application Program Interface (API), in order to exploit the parallelism in Tile64 processor. Three main techniques where addressed in this work: channel communication, message-passing and shared memory. Once equipped with the development environment and the skills needed to exploit it, the bioinformatics algorithm selection phase began, along with the migration strategy and its subsequent parallelization.

The first algorithm selected was the Needleman-Wunsch (NW) global pairwise alignment, implemented from scratch with different parallelization strategies and subsequently optimized, finally obtaining a significant speedup against other implementations. In order to increase the usability of the above development, and to open the research products to the scientific community, the need to develop a web interface for the Tile64 environment naturally arose, so a web site was set up with the ability to launch batch jobs to a TILExpress-20G card, and to return the results from it and to the user. This system is available at <http://galactus.uma.es/manycore> and currently includes the latest developments.

The results of this work represent the first milestone: an article, published in the *Bioinformatics* journal, which deals with a new NW implementation. As a complement of this publication, the book chapter included in the *5th International Conference on Practical Applications of Computational Biology & Bioinformatics* shows the developed web interface. It provide an intuitive access to the system. Thus, The MC64 web platform allows any life-science researcher to execute basic bioinformatics algorithms in the Tile64 architecture and to test the relative performance when the same algorithms are executed in a usual x86 multi-core architecture.

Given the similarities (from the informatics point of view) between the NW global alignment and the proposed local-variant by Smith and Waterman (SW), the algorithm implementation of this latter was the next achieved task. The article, published in the journal *Parallel Computing*, deeply analyzes these first implementations, centering the discussion on the informatics point of view.

The MC64-NW/SW was developed from scratch and meticulously optimized, taking advantage of the Tile64 architecture characteristics, being able to align very large sequences and exploiting parallelism in order to decrease execution time. To our knowledge, such algorithm in the TilExpress-20G card can align a pair of sequences of up to 1 Mb 15 times faster than any other standalone PC implementation to date. It is important to highlight that a proper parameterization of the algorithm is essential to achieve optimal results. Thus, it must take into account: (i) the dimensions of a sub-process' matrix; (ii) the average number of working tiles; and (iii) the memory required to save Job Fragments. As the Controller scheduling-time and the memory access-time through iMesh are much lower when compared to processing time, the parameterization should maximize the number of processing tiles, yet avoiding an excessive fragmentation of the matrix. Even so, to our knowledge, performance is still greatly superior in a standalone system to any other optimal alignment algorithm to date.

The last extension to this first set of bioinformatics algorithms was the ability of performing multiple alignments. This was carried out with a ClustalW implementation in which the first stage of the algorithm was performed by the previously developed and parallelized pairwise aligners. Thus this work produced a parallel solution to the most computationally-demanding task.

On the other hand, a different approach was considered in the porting strategies, choosing an algorithm already parallelized, in which a more direct solution could be tested. This is the case of ABySS, a de-novo sequence assembler in which a graph was distributed along all the available computing nodes, using the well-known MPI library. The book chapter published in the *IT Revolutions; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST)* introduces these other bioinformatics tasks: multiple alignment, de-novo genome assembly and protein folding.

As a second milestone, the article in the *Future Generation Computing Systems (The International Journal of Grid Computing and eScience)* journal discusses the results obtained when porting well-known bioinformatics algorithms to the Tile64 platform, focusing the discussion in the great differences obtained, depending on the method used to accomplish the migration.

The practical usefulness of the Tile64 parallelization depends on the particular algorithm used. In general, and not surprisingly, it is needed to thoroughly optimize the code to improve the global performance, as with other chips. But, unlike other architectures like CUDA, the algorithms may be rewritten by using a higher level of abstraction, as in a cluster of standalone pico-computers connected by an ultra-fast network and shared resources. Indeed, the Tile64 microprocessor can accomplish very fast response times that were previously unreachable by a standalone PC, for algorithms which can be divided into independent parallel tasks. However, when compared to other x86 multi-core systems, the Tile64 many-core has lower memory-amounts-per-core ratios and more shared resources that may produce contention and latency. Alternatively, the Tile64 microprocessors have more cores (64) than the current x86 alternatives (10).

Even the widely used MPI shows that an efficient implementation for a platform may produce a very low performance, when directly migrated to another platforms. This is because any high-performance MPI approach has to take into account the internal characteristics of the architecture to be deployed to. Thus, any many-core MPI implementation needs to consider some key features, like limited cache, shared memory usage, inter-core communication methods and intrinsic core limits. Otherwise, some resources may not be optimally exploited (e.g., idle tiles, misused file system in memory, wasted power integer calculus, etc.). Additionally, some resources may be overloaded (e.g., internal network, shared memory, reduced number of tiles, software floating point operations, etc.), so an unbalanced execution may cause a poor overall performance.

This scenario is even worse on GPU/CUDA microprocessors, where any effort to adapt an algorithm (e.g., the ABySS) would have required a greater modification degree, due, for instance, to the lack of inter-processor MPI capabilities. The Tilera architecture is not oriented to execute as many low load threads as possible inside a block code. Instead, it is designed to execute entire programs or even operating systems along a set of tens of cores, providing this set with high-speed communication systems. This is similar to both other message-passing implementations and other approaches like shared memory and interconnection channels. There are many high-throughput developments for CUDA platforms. Albeit, these algorithms have been parallelized exclusively for the GPU approach with a higher development-effort. Besides, many of them have specific limitations, and thus can only work with some data sets.

Therefore, as expected, a rewrite of the algorithms from scratch may be needed to fully exploit and optimize the performance of the Tile64 architecture and similar many-core platforms. Such a strategy may not represent the best performance/effort ratio in all instances. Thus, it may require a significant effort and time to redesign the best possible parallel strategy, taking into account the particular many-core specifications in each case. Yet, if carried out, it should usually produce a more refined and efficient bioinformatics tool, being therefore more useful for life-science researchers.

Later on, a further refinement was made, with a novel implementation of one of the most used multiple-alignment algorithms: ClustalW. In this case, both the first and third stages were parallelized, and a novel hybrid computing approach was taken. The article in the *PLoS ONE* journal discusses this novel ClustalW implementation, in which not only performance gains are described, but also new possibilities from a biological point of view.

The MC64-ClustalWP2 was benchmarked against the original algorithm and other parallel implementations, using different architectures and approaches. At first, stress tests were launched with 10 sequences of different sizes, measuring the speedup of the MC64-ClustalWP2 against other ClustalW implementations. A further test was carried out to analyze families of long sequences, which were aligned in much less time than with other general-global MSA strategies.

The dataset used for these experiments is composed of 10 sets of artificial sequences, whose lengths range from 25 kb to 300 kb. All of these algorithms were run in the same environment, an Intel Xeon Quad Core 2.0 GHz PC with 8 GB of quad-channel DDR2 memory. The MC64-ClustalWP2 used the Tile64 microprocessor on-boarded in a TilExpress-20G card with 8 GB RAM. The results revealed a better performance of the

MC64-ClustalWP2 against every other implementation. This tendency increased with the length of the sequences, because a higher parallelization factor was obtained. Thus, the MC64-ClustalWP2 reached a speedup of more than 7x when compared to the best multi-core implementation to date (ClustalW-MPI). On the other hand, the MC64-ClustalWP2 obtained a speedup factor of more than 18x when compared to the original algorithm.

As a final test, the MC64-ClustalWP2 was used to run an experiment whose execution time is near prohibitive when non-parallel or less-efficient parallel MSA methods are applied. In this stress test, 37 different human herpesviruses were selected to be aligned. The complete genome of the HHV-1 is about 152 kilobase pairs (kbp). The different strains should a priori present several mutation areas, due to the high-mutation rates of this kind of DNA viruses. Therefore, such alignments can be carried out with dynamic-programming algorithms like ClustalW. The MC64-ClustalWP2 aligned the 37 genome sequences in 36,103 seconds (10.03 h; 21,742 s for the first stage, nearly zero for the second one and 14,360 s for the third stage). The final phylogenetic tree revealed small distances, which revealed the proximity of all strains to the reference genome, being modeled from the 17 strain sequencing.

As expected, the alignment allowed to identify the polymorphisms between the virus strains, further confirming the previously reported sequencing results. This alignment was also performed with other global MSA algorithms, and even with other ClustalW implementations, but the quickest among them (ClustalW-MPI) required 274,149 seconds (3.17 days) to complete the alignment, being 7.59 times slower than the approach described in this work. On the other hand, the MSA algorithms do not generate an evolutionary tree of the genomic sequences, as they are oriented towards just finding the similar regions. In fact, they lack accuracy when the aligned sequences have highly polymorphic regions (e.g., with high mutation rates), which is typical for most viruses.

The last effort in this work is related to scalability. Therefore, once the usefulness of this many-core platform had been demonstrated, the possibility to extend its capabilities using the available built-in network interfaces was worth trying. Thus, a review of the methods usually applied to engage similar kinds of systems into collaboration was carried out. As a result, a cluster approach was chosen, further developing its basic elements, and testing general-purpose techniques used in heuristic bioinformatics algorithms.

The last phase of this work is described in the two last publications: the book chapter published in *Advances in Information Systems and Technologies*, and the article in the *Journal of Parallel and Distributed Computing*. Both deal with the cluster construction and its performance evaluation. Two main sets of tests were performed to evaluate the Tile64 system. In the first one, trees with sizes from 1,000 to 10 million of items –with increments of one order of magnitude– were generated. Their keys were uniformly distributed, by setting the distance between two consecutive keys as a random value between 1 and 49. With this search strategy, the range in which the data is expected to be located was fully covered. Both the keys and the queries were equally distributed across all available tiles. Thus, the effective range had a similar size in all tiles, with the exception of the first and the last ones. As a reference, a standard order-50 B-tree implementation with the same keys and search data was also used, being executed in a workstation.

The optimized Tilera implementation showed better performance than the Intel one from the very beginning. Besides, this approach generally scaled well as more tiles were added to the calculus. Only one exception was detected, located around the 110 tile execution with short trees. These anomalies were associated with known memory-contention issues in the card, which tends to have less weight in the overall result, as the number of searches increases. For those larger tree sizes and increasing number of searches, the developed algorithm showed an almost proportional time-to- size behavior. The maximum speedup was obtained with the one-million-key tree, showing a 227.54x gain when all the available tiles in the cluster were used.

In a second set of tests, trees with an increasing number of randomly-generated keys were produced. Thus, they were uniformly spreading across the whole integer range (the size of an integer is four bytes, both in the Tilera and the Intel C language compilers). Afterwards, a set of the same 10 million randomly-generated searches was performed, so observing the performance of a fixed set of queries against different density trees. Once again, this key-generation allowed a similar effective-range size in all participant tiles, being the base of the scalability of this strategy. In order to use the parallelization approach, a previous sort-and-split step is needed in the set of queries, so that a given tile only receives the values inside its coverage area. This previous sorting has an additional advantage: it reduces the cache faults during the search phase, because of the proximity of the successive values to search. As in the prior set of tests, a high improvement could be achieved, although in this case the sort-and-split phase limited the overall speedups. Furthermore, regarding the reference test, it is important to notice that the previous sorting did not actually offer any performance improvements. In fact, when this phase was removed from the reference implementation, the overall speedups obtained by the cluster executions were significantly decreased.

Other possible approaches have demonstrated to be less appropriated in this environment. For instance, searches where every tile reads the corresponding file in its entirety were also performed, with a further selection of the values in the range of interest by each tile. With this approach, the read-task showed a poor performance, as compared to the search itself, even when carried out in an RAM-disk environment. Moving this task to an additional thread (that reads the keys and loads an internal buffer shared with the main thread), initially showed a slight benefit. Yet, when the technique was used with big files and many tiles involved, the performance was strongly penalized, since the same tiles got used for the reading and searching tasks.

The main contributions of this work have been the porting of the ABySS algorithm, the analyses of the different migration strategies and the cluster construction and evaluation. Along with the analyses of the bottlenecks to reduce their impact, the optimization of the hotspots and the supervision of comparisons between heuristic and dynamic-programming algorithms. Thus, this set of tasks constitutes the core of this Thesis.

Our future work will include bioinformatics tools that include both dynamic-programming alignment and heuristic methods, which will take benefit from our improvements in massively parallel searches. We will also extend this developments to the new emerging many-core CPU platforms, like the ones by Intel (Xeon Phi) and Adapteva.

Conclusions

Regardless of the results showed in the above publications, the following ones can be considered as the whole-thesis general conclusions:

1. The Tile64 architecture can be effectively deployed on a desktop PC to address bioinformatics challenges, from both the performance and biological points of view [1,4-6].

2. The MultiCore64-NW and MC64-NW/SW algorithms represent the most powerful pairwise alignment solutions in a single PC to date [1-3].

3. Regarding the cluster performance, the MC64-Cluster shows also significant speedup gains with respect to a pure PC implementation [7-8].

4. Our developments can effectively complement the nowadays prevalent heuristic approaches for optimal alignment executions (from a mathematical point of view) of nucleic acids and peptide sequences, as they can be accomplished by means of fully-dynamic programming methods, even with large sequences [6].

5. The optimization of the Tile64-microprocessor performance requires the comprehensive knowledge of both the hardware/software architecture, as well as the fine-tuning of the algorithm programming at their lowest levels [5].

6. Hybrid Computing represents the best approach to fully exploit the parallel potential of many-core CPU architectures [6,8].

References:

1. Galvez, S, Díaz, D, Hernández, P, Esteban, F.J., Caballero, J.A. Dorado, G: **Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment**. *Bioinformatics* vol. 25, issue 5 pp. 683-686 (2010).
2. Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S. MC64: **A web platform to test bioinformatics algorithms in a many-core architecture** in: Rocha, M.P. *5th International Conference on Practical Applications of Computational Biology & Bioinformatics* vol. 93 pp 9-16. Springer-Verlag Berlin (2011).
3. Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Dorado, G, Gálvez, S. **Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture**. *Parallel Computing* vol. 37 issue 4-5 pp. 244-259 (2010).
4. Esteban, F.J., Díaz, D,, Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.: **Many-Core Processor Bioinformatics and Next-Generation Sequencing** in Liñán, M.: *IT Revolutions. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* vol. 82 pp. 172-188 Springer Berlin Heidelberg (2012).
5. Esteban, F.J., Díaz, D., Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.: **Direct approaches to exploit many-core architecture in bioinformatics**. *Future Generation Computer Systems* vol. 29 issue 1 pp. 15-16 (2013).

6. Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Guevara, A., Dorado, G., Gálvez, S.: **MC64-ClustalW, a high parallel strategy to align multiple DNA sequences in many-core architectures**. *PLoS ONE* (in press).
7. Esteban, F.J., Díaz, D., Hernández, P., Caballero, J.A., Dorado, G., Gálvez, S.: **MC64-Cluster: A Many-Core CPU Cluster for Bioinformatics Applications** in Correia, A. *Advances in Information Systems and Technologies* vol. 206 pp 819-825 Springer Berlin Heidelberg (2013).
8. Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S.: **MC64-Cluster: Architecture of a many-core CPU cluster and performance analysis in B-tree searches**. *Journal: Journal of Parallel and Distributed Computing* (under review).

Summary

This thesis shows the building process of a platform for the execution of bioinformatics algorithms in a massively-parallel environment. The Tile64 microprocessor from Tilera has been used, which is the first commercially available general-purpose many-core microprocessor. It has 64 cores, capable of running a whole standard operating system (a customized Linux version) in each core. Processors integrated in PCI-Express cards have been used, which can be inserted in a standard PC, which pack the processor with 8 GB of RAM and two 10 Gigabit Ethernet connectors. In a first step, the following bioinformatics algorithms have been developed in this platform: i) Needleman-Wunsch (global) and Smith-Waterman (local) pairwise aligners, by the development from scratch of a new wave-front parallel version with a master-worker scheme, as well as its later optimization, to get the most of Tilera's characteristics; ii) ABySS "de novo" assembler, by porting its open source code and the later parallelization by the adaptation of the original implementation, written for the MPI library to the message passing library available at Tilera; and iii) ClustalW multiple aligner, using the formerly developed pairwise aligners in the first phase of the algorithm. In a second step, a network between these devices has been built, using the available 10G connectors, so constructing a cluster in which the number of available microprocessors can be arbitrarily extended, keeping a unique point of program execution and administration. To achieve this goal, the usual management elements in this kind of systems have been developed, along with a communication library, in order to extend parallelism to the cluster components. Finally, the performance of this network platform has been evaluated; by developing and executing the standard search techniques typically used in heuristic-based alignment algorithms. As main conclusions, the bioinformatics algorithms performance have been remarkably increased by means of an optimized development to achieve a massive parallelization in this new platform. The best results have been obtained with developments from scratch, along with using hybrid-computing techniques. This strategy allows overcoming the limited resources in the card, effectively contributing extra resources from the host computer. These possibilities open new opportunities in nucleic acid and peptide (like proteins) bioinformatics, since it was not possible to apply optimal alignment methods from a mathematical point of view before these developments, being the most usual algorithms based in heuristic approaches.

Resumen

Esta tesis presenta el proceso de construcción de una plataforma para la ejecución de algoritmos bioinformáticos en un entorno masivamente paralelo. Se ha usado el microprocesador Tile64, del fabricante Tilera, que es el primer microprocesador de propósito general masivamente multi-núcleo disponible comercialmente. Dispone de 64 núcleos capaces de ejecutar un sistema operativo estándar completo (una versión adaptada de Linux) en cada uno de sus núcleos. Se han empleado procesadores integrados en placas PCI-Express, insertables en un PC estándar, que añaden al procesador 8 GB de memoria RAM y dos conectores 10 Gigabit Ethernet. En una primera fase, se han desarrollado sobre esta plataforma los siguientes algoritmos bioinformáticos: i) Alineamientos simples Needleman-Wunsch (global) y Smith-Waterman (local), mediante el desarrollo desde cero de una nueva versión paralelizada, mediante un esquema maestro-trabajadores en frente de onda y su posterior optimización para aprovechar las particularidades de Tilera; ii) Ensamblaje "de novo" ABySS, mediante la migración del código abierto ofrecido por los autores y su paralelización mediante la adaptación de la implementación original, escrita para la biblioteca MPI, a la biblioteca de paso de mensajes disponible en Tilera; y iii) Alineamiento múltiple ClustalW, usando los alineamientos simples desarrollados anteriormente en la primera fase del algoritmo. En una segunda fase, se ha construido una red de estos dispositivos, utilizando los conectores 10G disponibles, según el modelo conocido como "clúster", de modo que el número de microprocesadores disponibles puede incrementarse a voluntad, manteniendo un único punto de ejecución de programas y de administración. Para conseguirlo se han desarrollado los elementos de gestión habituales en este tipo de sistemas y una biblioteca de comunicaciones para extender el paralelismo a los componentes del "clúster". Finalmente, se ha evaluado el rendimiento de esta plataforma en red, mediante el desarrollo y ejecución en la misma de las técnicas de búsqueda estándares típicamente utilizadas en algoritmos de alineamiento basados en heurísticos. Como conclusiones principales, estos desarrollos bioinformáticos sobre la nueva plataforma han permitido incrementar el rendimiento de los algoritmos de forma significativa, mediante la paralelización masiva de los mismos. Los mejores resultados se han obtenido cuando se han llevado a cabo desarrollos desde cero, usando además técnicas de computación híbrida. Esta estrategia permite compensar la limitación de recursos en la tarjeta Tilera, usando recursos extra del ordenador en donde se aloja. Estas posibilidades abren nuevas oportunidades en el estudio bioinformático de los ácidos nucleicos y péptidos (como las proteínas), dado que hasta ahora no era posible aplicar métodos de alineamiento óptimos desde el punto de vista matemático, estando basados los algoritmos más habituales en aproximaciones heurísticas.

Other scientific contributions

The following contributions to scientific meetings represent complementary activities during this work:

The *3rd International ICST Conference on IT Revolutions 2011*, with a speech about our initial parallelization work, along with a vision on perspectives in bioinformatics.

The *5th International Conference on Practical Applications of Computational Biology and Bioinformatics 2011*, with a speech about the web service to access our bioinformatics algorithms on the Tile64 microprocessor.

The *2013 World Conference on Information Systems and Technologies (WorldCIST'13)*, with a speech about the initial MC64-Cluster implementation.

The *VI Latin American Symposium on High Performance Computing*, with the co-authorship in the contribution: *Many-core Tile64 vs. Multi-core Intel Xeon: Bioinformatics Performance Comparison*. This contribution was later extended and is currently under review to be published as an article in the Latin-American Center for Informatics Studies (CLEI) Electronic Journal (CLEIej).

The *V Jornadas de Divulgación de la Investigación en Biología Molecular, Celular, Genética y Biotecnología*, with the co-authorship in the contribution: *El proyecto de secuenciación (fase de muestreo cromosómico) del cromosoma 4A del trigo*. Two posters where also contributed to this meeting: *Second-generation sequencing and bioinformatics: Many-core processor approaches* and *Second-generation sequencing and bioinformatics: Many-core MC64 web platform*.

The *I Jornadas del Campus de Excelencia Internacional Agroalimentario 2010*, with the posters: *Agrifood Biotechnology: Traceability, Biodiversity, Bioinformatics and Genomics* and *Grupo de Investigación AGR-248*.

Additionally, the following academic and research activities has been performed during the doctorate studies:

A final degree work has been co-supervised, in which a web interface to the MC64-Cluster was developed.

An intervention has been made in two editions of the *Master en Biotecnología Molecular, Celular y Genética*, in which the Tile64 processor was showed to the students, along with a visit to the Computer Center at *Universidad de Córdoba*.

An article was reviewed for the *Expert Systems With Applications* journal, dealing with an alternative use of the Needleman-Wunsch algorithm.

Impact Factors

Source: Journal Citation Report by Thomson Reuters.

Bioinformatics – 2010

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP RETURN TO JOURNAL

2010 JCR Science Edition

Rank in Category: BIOINFORMATICS

Journal Ranking

For **2010**, the journal **BIOINFORMATICS** has an Impact Factor of **4.877**.

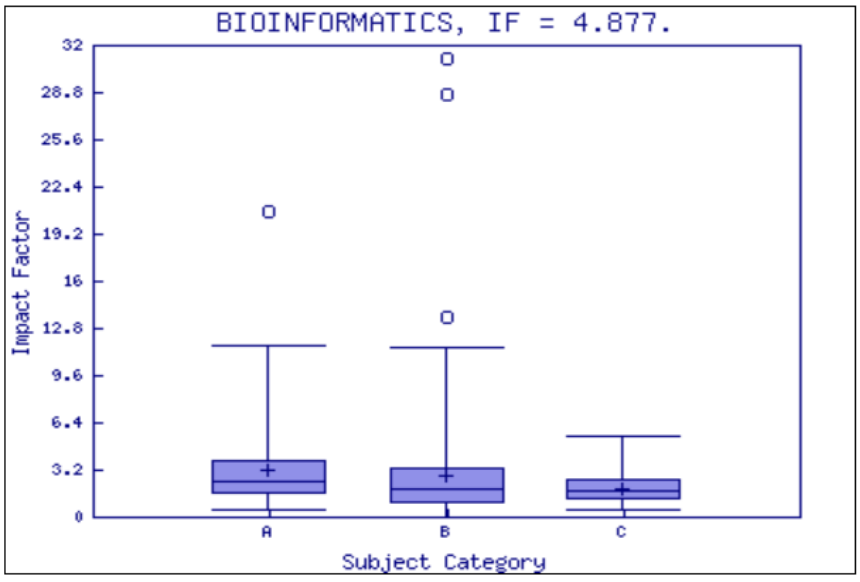
This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
BIOCHEMICAL RESEARCH METHODS	71	12	Q1
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	160	18	Q1
MATHEMATICAL & COMPUTATIONAL BIOLOGY	37	2	Q1

Category Box Plot

For **2010**, the journal **BIOINFORMATICS** has an Impact Factor of **4.877**.

This is a box plot of the subject category or categories to which the journal has been assigned. It provides information about the distribution of journals based on Impact Factor values. It shows median, 25th and 75th percentiles, and the extreme values of the distribution.



Key

A - BIOCHEMICAL RESEARCH METHODS

B - BIOTECHNOLOGY & APPLIED MICROBIOLOGY

C - MATHEMATICAL & COMPUTATIONAL BIOLOGY

Parallel Computing – 2011

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP RETURN TO JOURNAL

2011 JCR Science Edition

Rank in Category: PARALLEL COMPUTING

Journal Ranking

For **2011**, the journal **PARALLEL COMPUTING** has an Impact Factor of **1.311**.

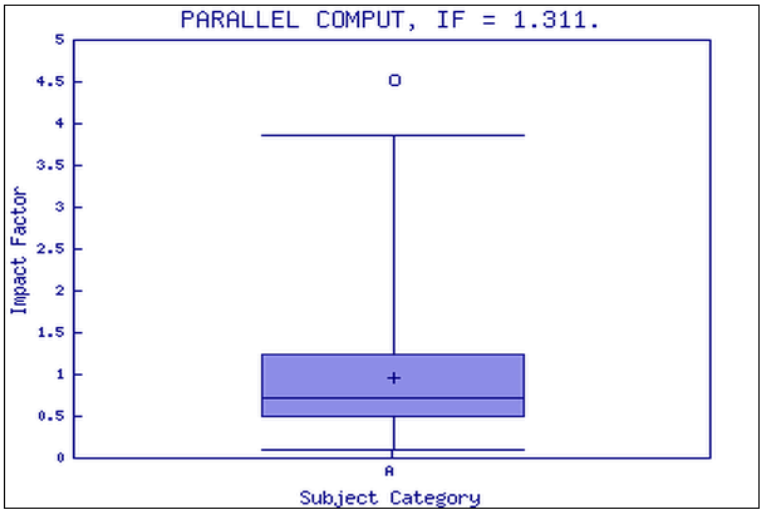
This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
COMPUTER SCIENCE, THEORY & METHODS	99	20	Q1

Category Box Plot

For **2011**, the journal **PARALLEL COMPUTING** has an Impact Factor of **1.311**.

This is a box plot of the subject category or categories to which the journal has been assigned. It provides information about the distribution of journals based on Impact Factor values. It shows median, 25th and 75th percentiles, and the extreme values of the distribution.



Key

A - COMPUTER SCIENCE, THEORY & METHODS

Future Generation Computing Systems – 2012 (2013 ranking is not available by the editing of this work)

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP RETURN TO JOURNAL

2012 JCR Science Edition

Rank in Category: Future Generation Computer Systems-The Internation...

Journal Ranking

For **2012**, the journal **Future Generation Computer Systems-The Internation...** has an Impact Factor of **1.864**.

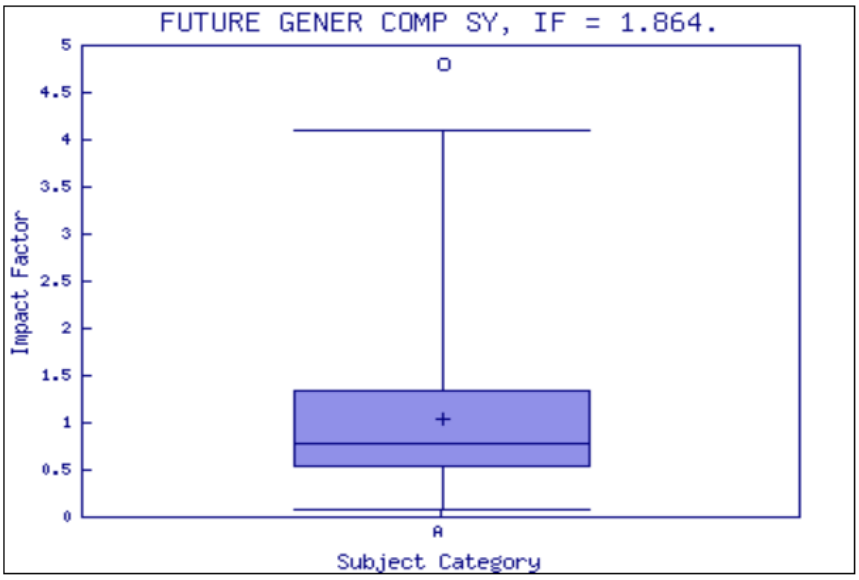
This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
COMPUTER SCIENCE, THEORY & METHODS	100	15	Q1

Category Box Plot

For **2012**, the journal **Future Generation Computer Systems-The Internation...** has an Impact Factor of **1.864**.

This is a box plot of the subject category or categories to which the journal has been assigned. It provides information about the distribution of journals based on Impact Factor values. It shows median, 25th and 75th percentiles, and the extreme values of the distribution.



Key

A - COMPUTER SCIENCE, THEORY & METHODS

Plos ONE – 2012 (2013 ranking is not available by the editing of this work)

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP RETURN TO JOURNAL

2012 JCR Science Edition

Rank in Category: PLoS One

Journal Ranking

For **2012**, the journal **PLoS One** has an Impact Factor of **3.730**.

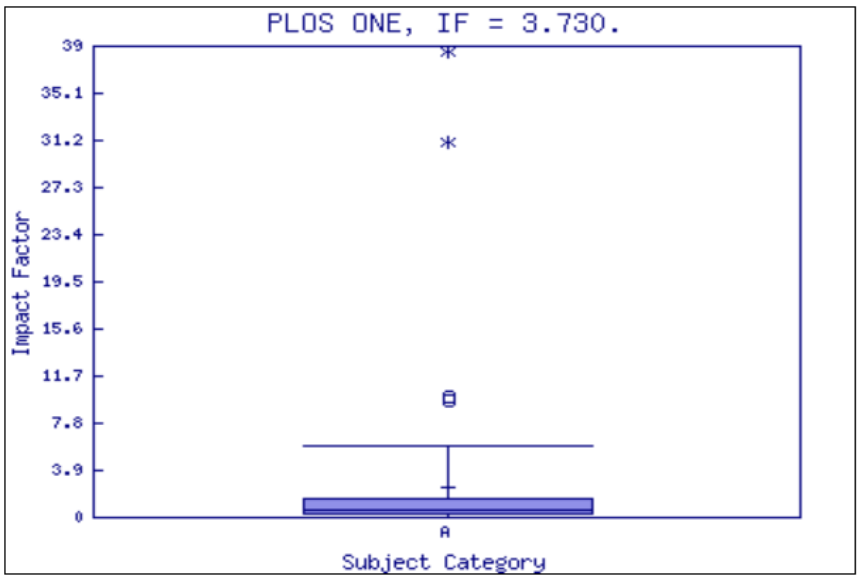
This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
MULTIDISCIPLINARY SCIENCES	56	7	Q1

Category Box Plot

For **2012**, the journal **PLoS One** has an Impact Factor of **3.730**.

This is a box plot of the subject category or categories to which the journal has been assigned. It provides information about the distribution of journals based on Impact Factor values. It shows median, 25th and 75th percentiles, and the extreme values of the distribution.



Key

A - MULTIDISCIPLINARY SCIENCES

Journal of Parallel and Distributed Computing – 2011

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP RETURN TO JOURNAL

2011 JCR Science Edition

Rank in Category: JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING

Journal Ranking

For **2011**, the journal **JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING** has an Impact Factor of **0.859**.

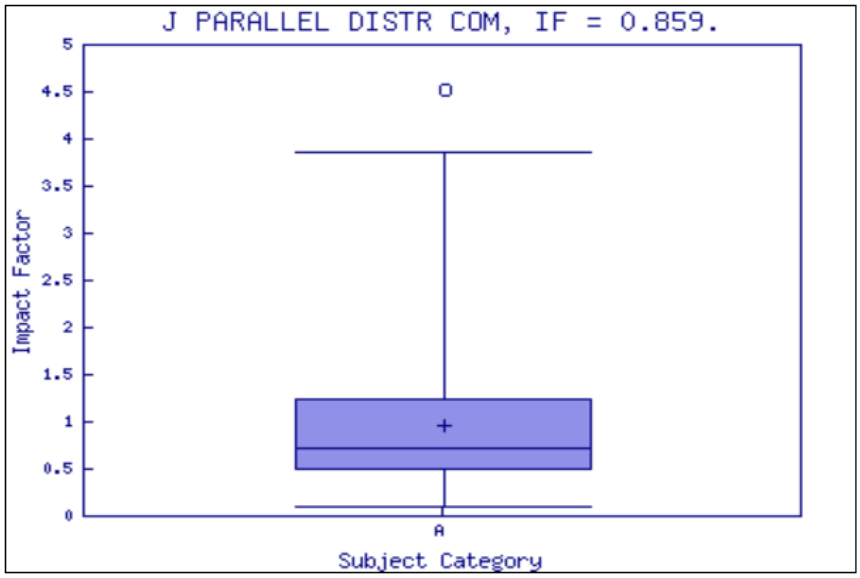
This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
COMPUTER SCIENCE, THEORY & METHODS	99	40	Q2

Category Box Plot

For **2011**, the journal **JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING** has an Impact Factor of **0.859**.

This is a box plot of the subject category or categories to which the journal has been assigned. It provides information about the distribution of journals based on Impact Factor values. It shows median, 25th and 75th percentiles, and the extreme values of the distribution.



Key

A - COMPUTER SCIENCE, THEORY & METHODS

Supervisors' Report

TÍTULO DE LA TESIS:

Plataforma Bioinformática Multinúcleo: Desarrollo y Optimización

DOCTORANDO/A:

Francisco José Esteban Risueño

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

Los doctores D. Gabriel Dorado Pérez (profesor del Departamento de Bioquímica y Biología Molecular de la Universidad de Córdoba), D. Juan Antonio Caballero Molina (profesor del Departamento de Estadística de la Universidad de Córdoba) y D. Sergio Gálvez Rojas (profesor del Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga), informan que el trabajo titulado «Plataforma Bioinformática Multinúcleo: Desarrollo y Optimización» que presenta D. Francisco José Esteban Risueño, Ingeniero de Telecomunicación, para optar al grado de Doctor por la Universidad de Córdoba, ha sido realizado bajo nuestra dirección.

El trabajo referido surge como necesidad de comprobar la efectividad de los recientes progresos en tecnología CPU multinúcleo (del inglés, *many-core*) en el área de la Bioinformática. En este campo, los últimos avances en secuenciación de ácidos nucleicos (ADN y ARN) y péptidos (proteínas) permiten obtener ingentes cantidades de datos, que requieren un tratamiento mediante Computación de Altas Prestaciones para poder obtener resultados en el menor tiempo posible. Los microprocesadores multinúcleo proporcionan una alta capacidad de procesamiento paralelo, cuya adaptabilidad al ámbito de la Bioinformática ha sido el objeto de estudio e investigación en este trabajo.

El trabajo arranca con un primer estudio sobre los principales problemas a los que se enfrenta la secuenciación de ácidos nucleicos, y cuáles de ellos podrían solucionarse de manera eficiente mediante microprocesadores con muchos núcleos. En este sentido, el inicio del estudio establece como foco de actuación los algoritmos de alineamiento de secuencias (por pares y múltiples), tanto para el descubrimiento de identidades y divergencias en las mismas, incluyendo, por ejemplo, genes comunes y polimorfismos de nucleótido sencillo (del inglés, *Single Nucleotide Polymorphism*; SNP), así como para la generación de la secuencia de genes, cromosomas y genomas mediante el ensamblaje de las lecturas generadas por las plataformas de secuenciación de primera a tercera generación, y la correspondiente obtención de regiones contiguas (del inglés, *contigs*) y andamiajes (del inglés, *scaffolds*) como elementos intermedios. El estudio abunda en la optimización del uso de los recursos proporcionados por los microprocesadores multinúcleo, entre los que se encuentran la memoria compartida y local, las memorias caché internas, el soporte de disco de

estado sólido (del inglés, *Solid State Disk*; SSD), la velocidad de las comunicaciones entre núcleos y el número de instrucciones ejecutadas por ciclo de reloj.

De los resultados de esta investigación se desprende que no resulta eficiente, en la mayoría de los casos, el aprovechar programas paralelos ya existentes para ejecutarlos directamente, o con mínimas modificaciones, en un microprocesador multinúcleo. Ello implica la necesidad de adaptar cada algoritmo, e incluso desarrollarlo desde cero, para sacar el máximo partido de los recursos de cómputo existentes. En tales casos, este trabajo demuestra que el aumento del rendimiento puede ser espectacular y que algoritmos de alineamiento como Needleman-Wunsch, Smith-Waterman o ClustalW pueden ver incrementada la velocidad de ejecución hasta 20 veces, dependiendo de la aproximación paralela que se emplee.

El trabajo propone, además, la utilización de varios microprocesadores multinúcleo, mediante su disposición en grupo (del inglés, *cluster*). Para ello se ha creado la arquitectura necesaria y se han diseñado y programado tanto una interfaz de programación de aplicación (del inglés, *Application-Programming Interface*; API) como una consola de comandos, que permite la gestión y monitorización de los recursos del mismo. La creación de esta arquitectura ha requerido conocer en profundidad el estado del arte actual en materia de *clústeres*: gestores de recursos, gestores de tareas, planificadores de ejecución y sistemas de seguridad y permisos. Con un *clúster* formado por tres microprocesadores de 64 núcleos cada uno se han obtenido excelentes resultados en algoritmos de base que se utilizan en el alineamiento por pares masivo, como es el caso de la herramienta de búsqueda de alineamiento local básico (del inglés, *Basic Local Alignment Search Tool*; BLAST).

Para facilitar el acceso de la comunidad científica a estos resultados, permitir su replicación/comprobación y compartir el código fuente de los programas, se ha creado una plataforma web públicamente accesible. De esta manera, se pone de manifiesto el deseo de realizar una transferencia y divulgación de los resultados de investigación, tal y como queda concretado en el contexto de los proyectos de investigación a cuyo amparo se ha desarrollado esta investigación.

Los estudios, desarrollos, pruebas y análisis realizados en este trabajo han dado lugar a una serie de publicaciones que se indican a continuación, en orden cronológico inverso:

Revistas:

Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Guevara, A., Dorado, G., Gálvez, S.: **MC64-ClustalW, a high parallel strategy to align multiple DNA sequences in many-core architectures**. *PLoS ONE* (en prensa).
Índice de impacto: 3,730; ranking: Multidisciplinary Sciences - 7/56 (Q1).

Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S.: **MC64-Cluster: Architecture of a many-core CPU cluster and performance analysis in B-tree searches**. *Journal: Journal of Parallel and Distributed Computing* (en revisión)
Índice de impacto: 0,859; ranking: Computer Science, Theroy & Methods - 44/97 (Q2).

Esteban, F.J., Díaz, D,, Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.: **Direct approaches to exploit many-core architecture in bioinformatics**. *Future Generation Computer Systems* vol. 29 issue 1 pp. 15-16 (2013).
Índice de impacto: 1,864; ranking: Computer Science, Theroy & Methods - 5/100 (Q1).

Díaz, D., Esteban, F.J., Hernández, P. Caballero, J.A., Dorado, G, Gálvez, S. **Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture**. *Parallel Computing* vol. 37 issue 4-5 pp. 244-259 (2011).
Índice de impacto: 1,311; ranking: Computer Science, Theroy & Methods - 20/99 (Q1).

Galvez, S, Díaz, D, Hernández, P, Esteban, F.J., Caballero, J.A. Dorado, G: **Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment**. *Bioinformatics* vol. 25, issue 5 pp. 683-686 (2010).
Índice de impacto: 4,877; ranking: Mathematical & Computational Biology - 2/37 (Q1).

Capítulos de libros:

Esteban, F.J., Díaz, D., Hernández, P., Caballero, J.A., Dorado, G., Gálvez, S.: **MC64-Cluster: A Many-Core CPU Cluster for Bioinformatics Applications** in Correia, A. *Advances in Information Systems and Technologies* vol. 206 pp 819-825 Springer Berlin Heidelberg (2013).
Indicios de calidad: revisión por pares y editorial de reconocido prestigio internacional.

- Indexado en SCOPUS

Esteban, F.J., Díaz, D,, Hernández, P., Caballero, J.A., Dorado, G., Gálvez, G.: **Many-Core Processor Bioinformatics and Next-Generation Sequencing** in Liñán, M.: *IT Revolutions. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* vol. 82 pp. 172-188 Springer Berlin Heidelberg (2012).
Indicios de calidad: revisión por pares y editorial de reconocido prestigio internacional.

- Indexado en SCOPUS

Esteban, F.J., Díaz, D. Hernández, P. Caballero, J.A. Dorado, G, Gálvez, S. MC64: **A web platform to test bioinformatics algorithms in a many-core architecture** in: Rocha, M.P. *5th International Conference on Practical Applications of Computational Biology & Bioinformatics* vol. 93 pp 9-16. Springer-Verlag Berlin (2011).
Indicios de calidad: revisión por pares y editorial de reconocido prestigio internacional.

- Indexado en Web of Science: Conference Proceedings Citation Index - Science (CPCI-S)
- Indexado en SCOPUS con las métricas:
  - SJR (SCImago Journal Rankings) 2011: 0,136
  - SNIP (Source Normalized Impact per Paper) 2011: 0,202

Comunicaciones en congresos, simposios, jornadas y otras reuniones científicas:

VI Latin American Symposium on High Performance Computing: coautoría en la contribución: *Many-core Tile64 vs. Multi-core Intel Xeon: Bioinformatics Performance Comparison*.

2013 World Conference on Information Systems and Technologies (WorldCIST'13): ponencia sobre la implementación inicial de MC64-Cluster.

5th International Conference on Practical Applications of Computational Biology and Bioinformatics 2011: ponencia sobre el servicio web para acceder nuestros algoritmos bioinformáticos para el microprocesador Tile64.

V Jornadas de Divulgación de la Investigación en Biología Molecular, Celular, Genética y Biotecnología: coautoría en la contribución: *El proyecto de secuenciación (fase de muestreo cromosómico) del cromosoma 4A del trigo*. Asimismo, se enviaron dos pósteres al encuentro, titulados: *Second-generation sequencing and bioinformatics:*

Many-core processor approaches and *Second-generation sequencing and bioinformatics: Many-core MC64 web platform*.

I Jornadas del Campus de Excelencia Internacional en Agroalimentación (ceiA3) 2010: pósteres: *Agrifood Biotechnology: Traceability, Biodiversity, Bioinformatics and Genomics*. *Grupo de Investigación AGR-248 del Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI)*.

Por todo ello, se autoriza la presentación de la citada tesis doctoral.

En Córdoba, a 8 de febrero de 2014

Firma de los directores

Fdo.: Gabriel Dorado Pérez Fdo.: Sergio Gálvez Rojas Fdo.: Juan Antonio Caballero Molina