5th BSC Severo Ochoa Doctoral Symposium

24th and 25th April, 2018

Book of Abstracts

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Book of Abstracts
5th BSC Severo Ochoa Doctoral Symposium

Editors
María José García Miraz
Carolina Olmo

Cover
Design based on artwork created by macrovector.com

This book is available at
www.bsc.es/doctoral-symposium-2018

published by
Barcelona Supercomputing Center

supported by
The “Severo Ochoa Centres of Excellence” programme

5th Edition, April 2018

ACKNOWLEDGEMENT

The BSC Education & Training team gratefully acknowledges all the PhD candidates, Postdoc researchers, experts and especially the Keynote Speaker José Ignacio Latorre Sentis and the tutorials lecturers Leonardo Bautista, Gavin Lucas and Antonio Peña for contributing to this Book of Abstracts and participating in the 5th BSC Severo Ochoa Doctoral Symposium 2018. We also wish to expressly thank the volunteers that supported the organisation of the event: Marc Fuster and Alba Gordó.

BSC Education & Training team
education@bsc.es

EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 5th BSC Severo Ochoa Doctoral Symposium.

During more than ten years, the Barcelona Supercomputing Center has been receiving undergraduate, master and PhD students, and providing them training and skills to develop a successful career. Many of those students are now researchers and experts at BSC and in other international research institutions.

In fact, the number of students has never decreased. On the contrary, their number and research areas have grown and we noticed that these highly qualified students, especially the PhD candidates, needed a forum to present their findings and fruitfully exchange ideas. As a result, in 2014, the first BSC Doctoral Symposium was born.

Last year, a total of 26 presentations were given, 32 posters were exhibited, a two days training on Deep Learning was conducted; and we reached more than 70 attendees.

In this 5th edition of the BSC Severo Ochoa Doctoral Symposium we have planned a keynote speaker' talk on *Quantum Disruption* and tutorials on *Key challenges for early-career researchers*, *How to become rich following an academic career* and *Marie Curie Individual Fellowships: Info & Best Practices*.

The talks will be held in four different sessions and will tackle the topics of: Programming Models and Computer Architectures, Simulations and Modelling, Life Sciences and Algorithms and applications. The posters will be exhibited and presented during two poster sessions that will give the authors the opportunity to explain their research and results.

The keynote speaker Dr. José Ignacio Latorre will give a lecture on *Quantum Disruption*. José Ignacio Latorre is Full Professor at the Universitat de Barcelona and is the leader of the Quantic group at Barcelona Supercomputing Center. He got his Ph. D. in Elementary Particle Physics at the Universitat de Barcelona on Elementary Particle Physics, was a Fulbright Fellow at MIT, a postdoc at the Niels Bohr Institute in Denmark, and a Long-term visiting professor at Center for Quantum Technologies in Singapore. He is the founder and director of the Centro de Ciencias de Benasque Pedro Pascual. He has worked as consultant for companies on artificial intelligence. He is a founder of Entanglement Partners. He wrote two popular books: "La Nada" and "Cuántica". He produces wine!

WELCOME ADDRESS

I am delighted to welcome all the PhD students, Postdoc researchers, advisors, experts and attendees participating in the 5th BSC Severo Ochoa Doctoral Symposium.

The goal of the event continues to be providing a framework to share research results of the projects developed by PhD thesis that use High Performance Computing in some degree.

The symposium was conceived in the framework of the Severo Ochoa Program at BSC, following the project aims regarding the talent development and knowledge sharing. Keeping that in mind, the symposium provides an interactive forum for PhD students considering both the ones just beginning their research and others who have developed their research activities during several years.

As a consequence, I highly appreciate the support provided by BSC and the Severo Ochoa Center of Excellence Programme that make possible to celebrate this event.

I must add that I am very grateful to the BSC directors for supporting the symposium, to the group leaders and to the advisors for encouraging the participation of the students in the event. Moreover, I wish to specially thank the keynote speaker José Ignacio Latorre and the invited lecturers Leonardo Bautista, Gavin Lucas and Antonio Peña for their willingness to share with us their knowledge and expertise.

And last but not least, I would like to thank all PhD students and Postdoc researchers for their presentations and effort. I wish you all the best for your career and I really hope you enjoy this great opportunity to meet other colleagues and share your experiences.

Dr. Maria-Ribera Sancho
Manager of BSC Education & Training

KEYNOTE SPEAKER

José Ignacio Latorre

Full Professor at Universitat de Barcelona and leader of the Quantic group at BSC

Quantum Disruption

Quantum Technologies are coming of age. The EU has recently aproved a FET-Flagship on Quantum Technologies, an instrument that will invest 1000 M Euros structured around four pillars: quantum computation, quantum communication, quantum simulation and quantum sensors. In this talk, we shall concentrate in recent progress achieved in quantum computation. The basic idea emerges from the fact that quantum mechanics allows for the manipulation of information in superposition states, called qubits. Furthermore, these superpositions evolved simultaneously following logical gates, providing a genuine parallel computation paradigm. A relevant example of the future use of a quantum computer is illustrated by Shor's algorithm, a quantum circuit that will factor large numbers in polynomial time, and will consequently break all present cryptography. Quantum logic, though, does not correlate in a simple way to classical algorithms. Non-trivial efforts must be devoted to further understand which problems can be addressed efficiently with quantum computation. Finally, it is arguable that quantum computation brings not only a possible dramatic speed up in some computations, but also provides relevant savings in energy. Research teams around the world compete fiercely to get a first demonstration of quantum supremacy over classical computation. Welcome to the quantum race.

José Ignacio Latorre is a Full Professor at the Universitat de Barcelona and is the leader of the Quantic group at Barcelona Supercomputing Center. He got his Ph. D. in Elementary Particle Physics at the Universitat def Barcelona on Elementary Particle Physics, was a Fulbright Fellow at MIT, a postdoc at the Niels Bohr Institute in Denmark, and a Long-term visiting professor at Center for Quantum Technologies in Singapore. He is the founder and director of the Centro de Ciencias de Benasque Pedro Pascual. He has worked as consultant for companies on artificial intelligence. He is a founder of Entanglement Partners. He wrote two popular books: "La Nada" and "Cuántica". He produces wine!

TUTORIALS

Gavin Lucas

Director of ThePaperMill

Troubleshooting Session: Key challenges for early-career researchers

The goal of this short session is help early-career researchers gain new perspectives on some of the key challenges they face, and to acquire practical tools that they can apply in their day-to-day working environment. I will present short modules focused on communication, personal effectiveness, and on sharing and discussing the common challenges faced by all early career researchers, and how they can be addressed.

Topics for this workshop:

- Group awareness - Troubleshooting the challenges of early-career research
- Communication - Understanding my audience and pitching my message
- Project Management – How can I prioritise my tasks?

Gavin Lucas PhD, director of ThePaperMill, is a scientist with 13 years of experience as a biomedical researcher, and 10 years of experience as an academic author's editor, consultant and trainer. In addition to his own solid track-record as a publishing scientist on national, European and international research projects, as an academic author's editor and consultant, he has helped plan, critique, and polish over 300 original scientific articles for dozens of institutes in diverse fields, as well as numerous FP7 and H2020 proposals. He also has extensive experience as a trainer in scientific writing and other transferable skills for researchers, and provides consultancy on training and scientific productivity at numerous academic institutes and agencies.

Leonardo Bautista-Gomez

Research Scientist at BSC

How to become rich following an academic career

The scientific field is very competitive and sometimes it can be even intimidating. This can lead promising young researchers to move to other domains or industries. However, following an academic career also comes with multiple advantages that might be hard to recognize at the early stages. In this talk I will present the perks and benefits of following an academic career.

Dr. Leonardo Bautista-Gomez is a Research Scientist at the Barcelona Supercomputing Center where he leads the European Marie Curie project on Deep-memory Ubiquity, Resilience and Optimization. He was awarded the 2016 IEEE TCSC Award for Excellence in Scalable Computing

(Early Career Researcher). Before moving to BSC he was a Postdoctoral researcher for 3 years at the Argonne National Laboratory, where he investigated data corruption detection techniques and error propagation. Prior to that, he did his PhD. in resilience for supercomputers at the Tokyo Institute of Technology. He developed a scalable multilevel checkpointing library called Fault Tolerance Interface (FTI) to guarantee application resilience at extreme scale. For this work, he was awarded the 2011 ACM/IEEE George Michael Memorial High Performance Computing Ph.D. Fellow at Supercomputing Conference 2011 (SC11), Honorable Mention. Before moving to Tokyo Tech, he graduated in Master for Distributed Systems from the Paris 6 University.

Toni Peña

Research Scientist at BSC

Marie Curie Individual Fellowships: Info & Best Practices

Marie Sklodowska-Curie Individual Fellowships from the European Comission are a great hit in a researcher's career. We will introduce these fellowships, including requirements and benefits, and will advise on how to prepare a successful application.

Dr. Antonio J. Peña is currently a Sr. Researcher at BSC, Computer Sciences Department. He work within the Programming Models group where he leads the "Accelerators and Communications for HPC" team. Antonio is the Manager of the BSC/UPC NVIDIA GPU Center of Excellence and member of the Outreach Working Group. He is also Teaching and Research Staff at Universitat Politècnica de Catalunya. Antonio is a Marie Sklodowska-Curie Fellow, former Juan de la Cierva Fellow, and a recipient of the 2017 IEEE TCHPC Award for Excellence for Early Career Researchers in High Performance Computing. His research interests in the area of runtime systems and programming models for high performance computing include resource heterogeneity and communications. Antonio was formerly at Argonne National Laboratory (U.S.A.) and Universitat Jaume I (Spain).

PROGRAM

DAY 1 (24th of April)

Start time	Activity	Speaker/s	Chair
8.30h Registration			
9.00h	Welcome and opening	Mateo Valero, BSC Director	
9.20h	Keynote talk: Quantum Disruption	Jose Ignacio Latorre Sentis Quantic Group Leader, CASE, BSC	
	Abstract: Quantum Technologies are coming of age. The EU has recently approved a FET-Flagship on Quantum Technologies, an instrument that will invest 1000 M Euros structured around four pillars: quantum computation, quantum communication, quantum simulation and quantum sensors. In this talk, we shall concentrate in recent progress achieved in quantum computation. The basic idea emerges from the fact that quantum mechanics allows for the manipulation of information in superposition states, called qubits. Furthermore, these superpositions evolved simultaneously following logical gates, providing a genuine parallel computation paradigm. A relevant example of the future use of a quantum computer is illustrated by Shor's algorithm, a quantum circuit that will factor large numbers in polynomial time, and will consequently break all present cryptography. Quantum logic, though, does not correlate in a simple way to classical algorithms. Non-trivial efforts must be devoted to further understand which problems can be addressed efficiently with quantum computation. Finally, it is arguable that quantum computation brings not only a possible dramatic speed up in some computations, but also provides relevant savings in energy. Research teams around the world compete fiercely to get a first demonstration of quantum supremacy over classical computation. Welcome to the quantum race.		
10.30h Event Photo			
Coffee break & First Poster Session			
10.40h	1. Application of the edge-based finite element method for fusion plasma simulations, Marc Fuster 2. Skip RNN Learning to Skip State Updates in Recurrent Neural Networks, Víctor Campos 3. Recurrent Semantic Instance Segmentation, Míriam Bellver 4. Improving Time-Randomized Cache Designs, Pedro Benedicte 5. Co-Evolution of Morphology and Behavior in Self-Organized Robotic Swarms, Jessica Meyer		
First Talk Session: **Life Sciences**			
11.40h	FrAG-PELE: Novel Fragment-based Growing Tool for hit-to-lead in Early Drug Discovery	Carles Pérez López	
12.00h	Sampling nterfacial Water Effects over Protein Specificity with PELE	Martí Municoy Terol	
12:20h	Characteritzation of pathological mutations affecting protein-protein interactions for drug discovery	Mireia Rosell Oliveras	
12.40h Lunch Break			
14.30h Tutorial 1			
	Title: Troubleshooting Session: Key challenges for early-career researchers	Gavin Lucas PhD, director of ThePaperMill	
	Goals & Content		

The goal of this short session is help early-career researchers gain new perspectives on some of the key challenges they face, and to acquire practical tools that they can apply in their day-to-day working environment. I will present short modules focused on communication, personal effectiveness, and on sharing and discussing the common challenges faced by all early career researchers, and how they can be addressed.

Topics for this workshop:

- Group awareness - Troubleshooting the challenges of early-career research
- Communication - Understanding my audience and pitching my message
- Project Management – How can I prioritise my tasks?

16.30h Adjourn

DAY 2 (25th of April)

Start time	Activity	Speaker/s	Chair
9.00h	Opening of the second day		
Second Talk Session: Simulations and Modelling			
9.10h	Comparison of seismic ground motions in Mexico City due to damaging earthquakes applying Seismograms Analyzer-e	Armando Aguilar-Melendez	
9.30h	Earthquake simulation by Fiber Bundle Model and Machine Learning techniques	Marisol Monterrubio Velasco	
9.50h	An introduction to FE2 multi-scale methods and why HPC is so crucial.	Guido Giuntoli	
10.10h	Effect of population structure, parameter estimation of complex model, and effect of LITB on TB dynamics	Nura Mohammad Rabiu Ahmad	
10.30h	An assessment of regional sea ice predictability in the Arctic ocean	Ruben Cruz-García	

10.50h Coffee break & Second Poster Session:

1. A Unified Memory approach to GPU acceleration on task based programming models, Aimar Rodríguez
2. A Machine Learning Workflow for Hurricane Prediction, Albert Kahira
3. Evaluation of traffic emission models coupled with a microscopic traffic simulator and on-road measures, Daniel Rey
4. Accelerating binding free energy calculations by combining Monte Carlo simulations, enhanced sampling and Markov State Models, Joan Francesc Gilabert

Third Talk Session: Programming Models and Computer Architectures

12.00h	Model-based Machine Learning for Retrospective Event Detection	Joan Capdevila Pujol
12.20h	Detailed Tuning and Validation of Hardware Simulators through Microbenchmarks	Rommel Sánchez
12.40h	Enabling a Reliable STT-MRAM Main Memory Simulation	Kazi Asifuzzaman
13.00h	A Linux Kernel Scheduler Extension for Multi-Core Systems	Aleix Roca Nonell

13 20h Lunch break

14.30h Tutorial 2

Title: How to become rich following an academic career	Leonardo Bautista-Gomez, Research Scientist at BSC
Content&Goals The scientific field is very competitive and sometimes it can be even intimidating. This can lead promising young researchers to move to other domains or industries. However, following an academic career also comes with multiple advantages that might be hard to recognize at the early stages. In this talk I will present the perks and benefits of following an academic career.	
Title: Marie Curie Individual Fellowships: Info & Best Practices	Toni Peña, Sr. Researcher at BSC
Content&Goals Marie Sklodowska-Curie Individual Fellowships from the European Comission are a great hit in a researcher's career. We will introduce these fellowships, including requirements and benefits, and will advise on how to prepare a successful application.	

15.40h Coffe break

Fourth Talk Session: Algorithms and applications

16.00h	Modelling of Alfvénic instabilities in complex toroidal magnetic geometries for fusion	Allah Rakha
16.20h	Robust point-location method for linear and high order meshes. Application to particle transport	Edgar Olivares
16.40h	On the quest to reach nuclear fusion as a future energy source	Dani Gallart Escolà
17.00h	Fuzzy Finite State Machines in Crowd Simulation	Leonel Antonio Toledo Díaz
17.20h	Top View Human Head and Shoulder Classification Using CNN	Ivan Rivalcoba
17.40h	Conclusions	

17.50 End of the Doctoral Symposium

Talk Abstracts

Comparison of seismic ground motions in Mexico City due to damaging earthquakes applying Seismograms Analyzer-e

Armando Aguilar-Meléndez*†, Josep De la Puente*, Héctor Rodríguez-Lozoya‡ Marisol Monterrubio-Velasco*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universidad Veracruzana, Poza Rica, Mexico
‡Universidad Autónoma de Sinaloa, Culiacán, Mexico
E-mail: {armando.aguilar, josep.delapuente, marisol.monterrubio}@bsc.es

Keywords—Seismic ground motions, acceleration, PGA, earthquakes.

I. Introduction

Earthquakes are one of the natural phenomena that eventually can trigger damage to cities depending on diverse factors as the occurrence site, the size, duration of the earthquake, etcetera. The released energy during an earthquake is partially dissipated by seismic waves. The seismic waves that are triggered by an earthquake are unique. In other words, there are not two earthquakes that have triggered exactly the same seismic waves. The analysis of the seismic records is an important source to obtain valuable information about both the features of the seismic waves on a site and the characteristics of the earthquake that triggered these seismic waves. In the present document, we described some relevant aspects of the analysis of seismic records that we processed and analyzed applying the computer code Seismograms Analyzer-e (see Figure 1).

II. Seismic ground motions in Mexico City

Table I shows basic data about two earthquakes that occurred on September 19 in the years 1985 and 2017. Both earthquakes have significant differences in the magnitude and in the distance from the epicenter to Mexico City. Unfortunately, in both earthquakes, some buildings in Mexico City had a partial or total collapse and as a consequence, some people died.

TABLE I. Main data of two earthquakes that triggered significant damage in buildings of Mexico City [1] [2] [3]

Data	Michoacan Earthquake	Puebla Earthquake
Date	Sept,19,1985	Sept,19,2017
Site of occurrence	Coast of Michoacan	Limits Puebla
Magnitude	8.1	7.1
Depth	27.9 km	57.0 km
Distance from the epicenter to the Mexico City station	419.5 km	116.4 km

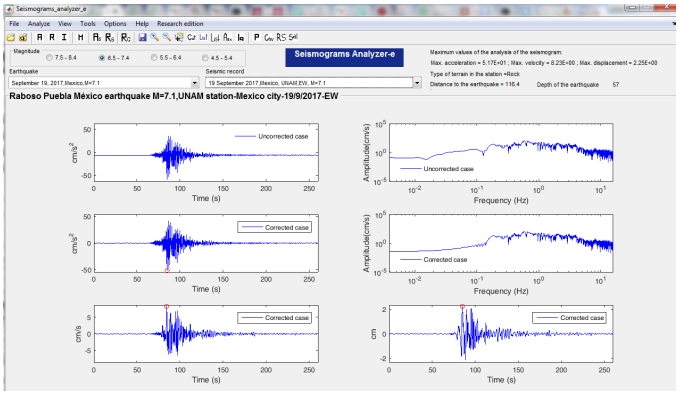


Fig. 1. Main screen of Seismograms Analyzer-e [4] [5].

A. The Michoacán, Mexico earthquake of September 19, 1985 (Ms = 8.1)

The Michoacán earthquake of September 19, 1985, generated seismic waves that in a rock site of Mexico city triggered a peak ground acceleration (PGA) about 31 cm/s^2 in the North-South component, and about 33.8 cm/s^2 in the East-West component (Figure 2). In this earthquake, the distance from the epicenter to the seismic station in Mexico City was of 419.5 km.

B. The Puebla, Mexico earthquake of September 19, 2017 (Mw = 7.1)

The Puebla earthquake of September 19, 2017, generated seismic waves that in a rock site of Mexico city triggered a peak ground acceleration (PGA) about 44.3 cm/s^2 in the North-South component, and about 51.70 cm/s^2 in the East-West component (Figure 3). In this other earthquake, the distance from the epicenter to the seismic station in Mexico City was of 116.4 km.

III. Response Spectrum

A seismic record can be used to determine a response spectrum that gives us information about the effects that the seismic waves recorded could have triggered on the buildings. For instance, a pseudoacceleration response spectrum shows values of pseudoacceleration that the earthquake could have

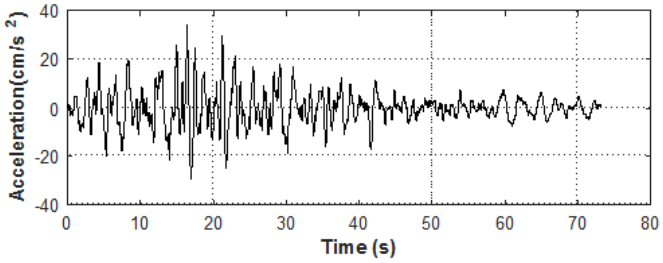


Fig. 2. Accelerogram of the component East-West of the earthquake of September 19, 1985 (Table I) recorded in a station of CU in Mexico City and processed by Seismograms Analyzer-e [4].

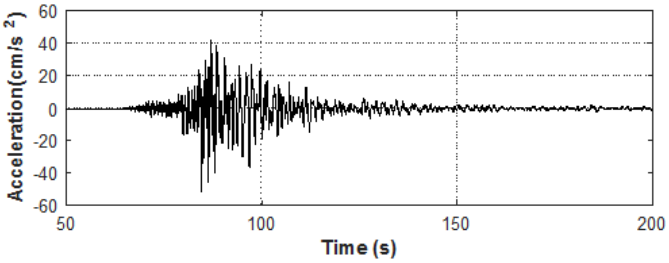


Fig. 3. Accelerogram of the component East-West of the earthquake of September 19, 2017 (Table I) recorded in a CU station of CU in Mexico City and processed by Seismograms Analyzer-e [4].

generated in the roof floor of different buildings depending on their structural period.

Figure 4 and Figure 5 show the response spectra that were determined with the acceleration data obtained in a seismic station located in the National Autonomous University of Mexico in Mexico City during the two earthquakes of Table I. According to the response spectrum of Figure 4, in the roof floor of a building with a structural period of 0.5 s (as a reference some buildings of reinforced concrete about 5 levels have a structural period near to 0.5 s) the maximum value of pseudoacceleration was of 51 cm/s^2 during the Michoacan earthquake, but of 102 cm/s^2 during the Puebla earthquake. Similarly, according to the response spectrum of Figure 5, in the roof floor of a building with a structural period of 0.5 s the maximum value of pseudoacceleration was of 60 cm/s^2 during the Michoacan earthquake but of 161 cm/s^2 during the Puebla earthquake. Therefore, it is possible to identify that the recent earthquake of Puebla generated seismic waves that in some cases triggered higher values of pseudoacceleration that the values that were triggered by the seismic waves due to the Michoacan earthquake.The high values of pseudoacceleration triggered during the Puebla earthquake are a part of the factors that explain why some buildings of Mexico City suffered partial or total collapse [3] .

The analysis that was summarized in the present work is an example of the type of analysis that it is possible to do with the support of Seismograms Analyzer-e [4]. Therefore, we believe that this kind of software must be widely sharing to contribute to that more people can be able to do these types of analysis before taking decisions about existing buildings or new buildings, in order to reduce the seismic risk of buildings.

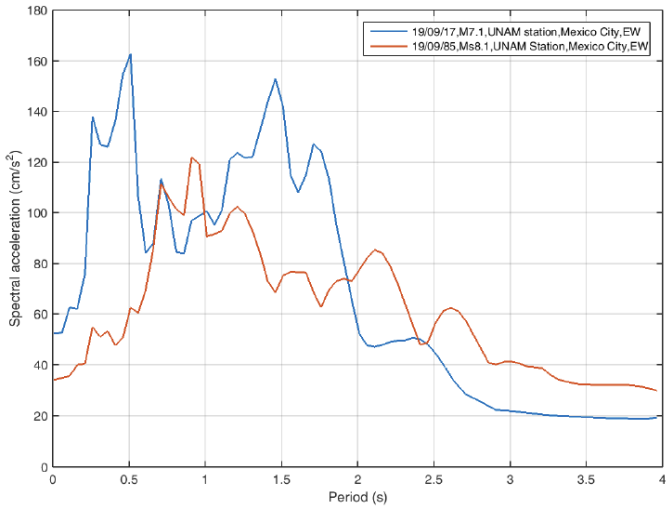


Fig. 4. Response spectra of pseudoaccelerations determined by Seismograms Analyzer-e [4] for the component North-South of the two earthquakes that occurred in the same day (September 19), but in different years (1985 and 2017) (Table I). The seismic records were obtained in a rock site of Mexico City.

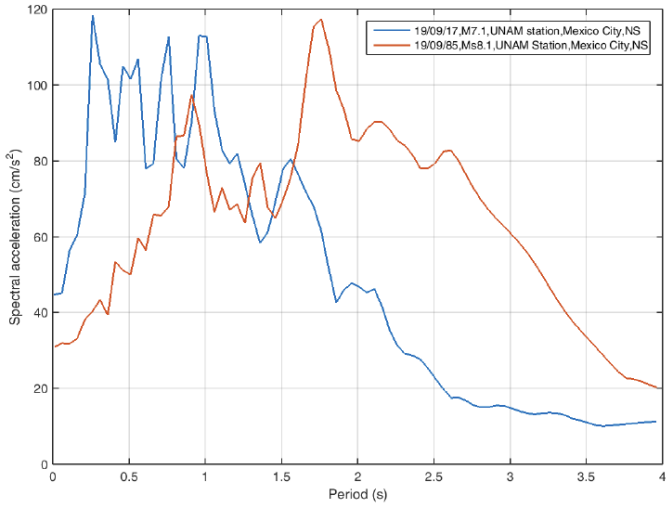


Fig. 5. Response spectra of pseudoaccelerations determined by Seismograms Analyzer-e [4] for the component East-West of the two earthquakes that occurred in the se day (September 19), but in different years (1985 and 2017) (Table I). The seismic records were obtained in a rock site of Mexico City.

IV. Conclusion

According to the results that we showed in the present document, we can to affirm that the significant damage in buildings of Mexico City during the recent earthquake (09/19/2017) was due to the combination of high values of PGA with high values of seismic vulnerability of the buildings that suffered significant damage. For instance, about the values of PGA, it is possible to highlight that the highest value of PGA that was recorded in a rock site of Mexico City during the 2017 earthquake was 53 percent greater than the value of PGA that was recorded in the same site but during the earthquake of 1985.

The analysis of seismic records is a valuable procedure to know features of earthquakes and their effect on buildings. This type of analysis can be appropriately done applying the

software Seismograms Analyzer-e (SA-e) [4].

V. Acknowledgment

The authors would like to thank to CONACYT and BSC.

References

[1] Aguilar-Melendez, A. and De la Puente, J. and Rodriguez-Lozoya, H. E., "Comparación de movimientos del terreno generados por los sismos del 19 de septiembre de 1985 y del 19 de septimbre de 2017 , en la UNAM (Ciudad Universitaria) en la Ciudad de Mexico. Short Technical Note," https://goo.gl/559McM, 2015.

[2] SSN-Mexico, "Sismos Fuertes, Servicio Sismológico Nacional, Mexico," http://www2.ssn.unam.mx:8080/sismos-fuertes/, 2015.

[3] CIRES-México, "El sismo del 19 de Septiembre de 1985, México," http://www.cires.org.mx/, 2015.

[4] A. Aguilar-Melendez *et al.*, "Seismograms Analyzer-e. Program for analysis of seismic records," 2018. [Online]. Available: https://sites.google.com/site/seismogramsanalyzere/home

[5] A. Aguilar-Melendez and L. Pujades, "Seismograms Analyzer-e, un software para analizar registros sismicos," in *Proccedings XXI Congreso Nacional de Ingenieria Sismica. Guadalajara, Jalisco*, 2017.

Armando Aguilar Armando Aguilar is a Posdoctoral Researcher Conacyt-BSC in the CASE Department. He is mainly interested in earthquake engineering, seismic risk, seismic hazard and seismic vulnerability.

Josep De la Puente received his PhD in the Ludwing Maxmillian University in 2007. His main topic research is about computational seismology. Currently, he is the manager of the geosciences application group of the Barcelona Supercomputing Center.

Marisol Monterrubio-Velasco received her PhD from the University Politcnica de Catalunya (Spain) and was a postdoctoral researcher at Geosciences center, UNAM. At present she is postdoc in the Barcelona Supercomputing Center. Her research interests focus on computational physics, statistical analysis and numerical simulation applied to Earth phenomena (fractal structures, earthquakes behavior, statistical analysis)

Héctor Rodríguez-Lozoya received his PhD degree in Seismology from the CICESE, Mexico. Currently he is a full time Professor in the Faculty of Engineering at the Autonomous University of Sinaloa.

Enabling a Reliable STT-MRAM Main Memory Simulation

Kazi Asifuzzaman*†, Rommel Sánchez Verdejo*†, Petar Radojković*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {kazi.asifuzzaman, rommel.sanchez, petar.radojkovic}@bsc.es

Keywords—STT-MRAM, Main memory, High-performance computing.

I. Extended Abstract

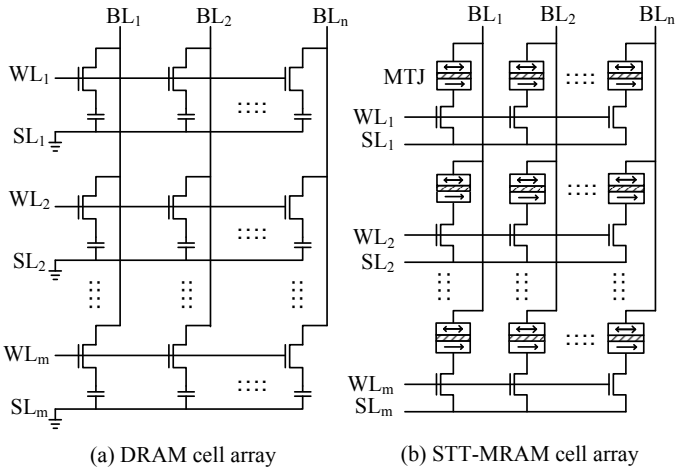
Memory systems are major contributors to the deployment and operational costs of large-scale HPC clusters [1][2], as well as one of the most important design parameters that significantly affect system performance. In addition, scaling of the DRAM technology and expanding the main memory capacity increases the probability of DRAM errors that have already become a common source of system failures in the field. It is questionable whether mature DRAM technology will meet the needs of next-generation main memory systems. So, significant effort is invested in research and development of novel memory technologies. A potential candidate for replacing DRAM is Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM).

The main objective of this work is to understand and publish detailed STT-MRAM main memory timing parameters enabling a reliable system level simulation of the novel memory technology. The approach that we present converged through research cooperation with Everspin technologies Inc., one of the leading MRAM manufacturers, and it provides reliable STT-MRAM timing parameters while releasing no confidential information about any commercial products.

A. STT-MRAM

The storage and programmability of STT-MRAM revolve around a Magnetic Tunneling Junction (MTJ), see Figure 1(b). An MTJ is constituted by a thin tunneling dielectric being sandwiched between two ferro-magnetic layers. One of the layers has a fixed magnetization while the other layer's magnetization can be flipped. If both of the magnetic layers have the same polarity, the MTJ exerts low resistance therefore representing a logical "0"; in case of opposite polarity of the magnetic layers, the MTJ has a high resistance and represents a logical "1". In order to read a value stored in an MTJ, a low current is applied to it. The current senses the MTJ's resistance state in order to determine the data stored in it. Likewise, a new value can be written to the MTJ through flipping the polarity of its free magnetic layer by passing a large amount of current through it [3].

STT-MRAM main memory timing parameters has neither been standardized nor been released by any industry. This is perhaps due to the perpetual evaluation of the STT-MRAM technology that is constantly changing over a short duration



(a) DRAM cell array (b) STT-MRAM cell array

Fig. 1. STT-MRAM cell and cell-array

of time. Memory manufacturers, who are developing STT-MRAM are judiciously not revealing these parameters ahead of time; so, at this point, we have to accept that there is no reliable information on how these timing parameters will change for the upcoming STT-MRAM devices.

Industrial patents [4][5][6] suggest STT-MRAM manufacturers are adopting STT-MRAM technology in to DDRx interface and protocols in order to enable a seamless integration into rest of the system. STT-MRAM memory devices are DDRx compatible, with the same or very similar organization and CPU interface, as the conventional DRAM. Also, both, DRAM and STT-MRAM main memory devices use row buffer as an interface between the cell-arrays and the memory bus. Since the circuitry beyond the row buffer for DRAM and STT-

TABLE I. DRAM and STT-MRAM parameters associated with row operation (DDR3-1600 cycles)

Timing Parameters	Description	DRAM	ST-1.2	ST-1.5	ST-2.0
tRCD	Row to column command delay	11	14	17	22
tRP	Row precharge	11	14	17	22
tFAW	Four row activation window	24	29	36	48
tRRD	Row activation to Row activation delay	5	6	8	10
tRFC	Refresh cycle time	208	1	1	1

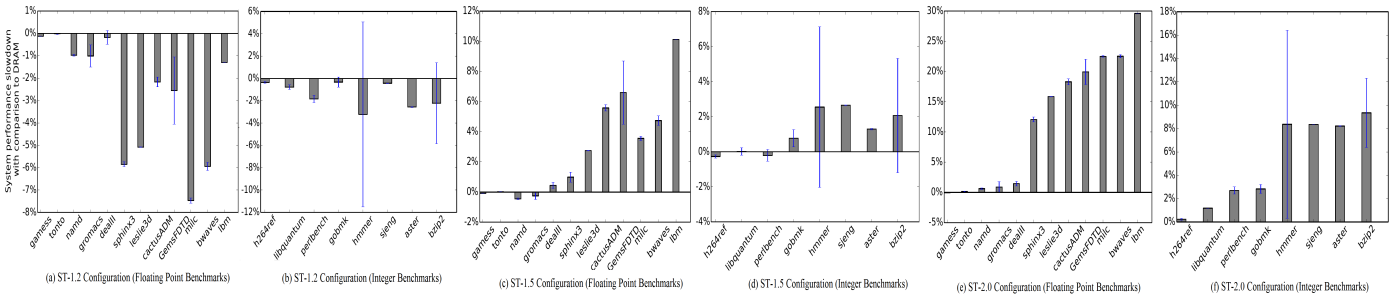


Fig. 2. STT-MRAM slowdown with respect to DRAM main memory

MRAM is essentially the same — once the data is in the row buffer, STT-MRAM timing parameters for the consequent operations are the same as DRAM. Therefore, the values of all the timing parameters that are not associated with row operations do not change from DRAM to STT-MRAM.

The only fundamental difference in STT-MRAM and DRAM main memory is their storage cell technology (see Figure 1), MTJ and capacitor, respectively. Due to the difference in the cell access mechanism of these two memory technologies, the timing parameters associated with STT-MRAM row operations would deviate from DRAM, and there is no reliable information on how these timing parameters will change for the upcoming STT-MRAM devices. Therefore, a sensitivity analysis is performed on timing parameters that would deviate from DRAM. In this study, we selected three set of timings naming ST-1.2, ST-1.5 and ST-2.0 with deviations of 1.2x, 1.5x and 2x from respective DRAM timing parameters as summarized in Table I. All the timing parameters that are *not* listed in this table will be same as DRAM. The presented methodology converged through our research cooperation with Everspin Technologies Inc.

B. Experimental Environment

STT-MRAM main memory was evaluated on a set of eight integer and twelve floating point benchmarks from the SPEC CPU 2006 suite [7]. We used ZSim [8] system simulator for the experiments. The simulated hardware platform comprises a detailed model of Sandy Bridge-EP E5-2670 cache hierarchy [9]. This Sandy Bridge E class processor has eight cores, dedicated L1 instruction and data cache of 32 KB each, dedicated L2 cache of 256 KB and a shared L3 cache of 20 MB. Both DRAM and STT-MRAM main memory is simulated with DRAMSim2 [10].

C. Results

Figure 2 shows overall system performance impact of STT-MRAM configurations on SPEC integer benchmark. The vertical bars represent system performance deviation from DRAM for the corresponding benchmarks listed at *X* axis. Both Floating point (Figure 2(a)) and Integer (Figure 2(b)) benchmarks with ST-1.2 configuration experience a speedup with the STT-MRAM main memory. This is due to the operation sequence of STT-MRAM, which is different from DRAM. Unlike DRAM, STT-MRAM has a non-destructive read which does not have to write-back; meaning it can issue precharge command sooner [11]. For ST-1.5 configuration (Figure 2(c)(d)) system performance degradation ranges from -0.2% (h264ref) to 10.1% (lbm). For the most pessimistic configuration ST-2.0 (Figure 2(e)(f)), slowdown ranges between 0.2% (h264ref) and 29.6% (lbm).

D. Conclusion

In this study, we publish reliable timing parameters of STT-MRAM main memory and measure its performance degradation w.r.t DRAM. We believe this study will enable researchers to perform a reliable STT-MRAM main memory simulation.

II. Acknowledgment

This work bas been published in proceedings of the International Symposium on Memory Systems (MEMSYS), 2017 [12].

References

[1] P. Kogge *et al.*, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," DARPA, Sep. 2008.

[2] A. Sodani, "Race to Exascale: Opportunities and Challenges," Keynote Presentation at the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Dec. 2011.

[3] Y. Xie, "Modeling, Architecture, and Applications for Emerging Memory Technologies," *IEEE Design Test of Computers*, 2011.

[4] H. Kim *et al.*, "Magneto-resistive memory device including source line voltage generator," 2013.

[5] H. Oh, "Resistive Memory Device, System Including the Same and Method of Reading Data in the Same," 2014.

[6] C. Kim *et al.*, "Magnetic Random Access Memory," 2013.

[7] J. L. Henning, "SPEC CPU2006 Benchmark Descriptions," *SIGARCH Comput. Archit. News*, 2006.

[8] D. Sanchez and C. Kozyrakis, "Zsim: Fast and accurate microarchitectural simulation of thousand-core systems," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ser. ISCA, 2013.

[9] Intel, "Intel 64 and IA-32 Architectures Optimization Reference Manual," http://www.intel.com/content/www/us/en/architecture-and-technology/64-ia-32-architectures-optimization-manual.html.

[10] P. Rosenfeld *et al.*, "DRAMSim2: A Cycle Accurate Memory System Simulator," *IEEE Computer Architecture Letters*, 2011.

[11] J. Wang *et al.*, "Enabling High-performance LPDDRx-compatible MRAM," ser. ISLPED, 2014.

[12] K. Asifuzzaman *et al.*, "Enabling a reliable stt-mram main memory simulation," in *Proceedings of the International Symposium on Memory Systems*, ser. MEMSYS '17.

Kazi Asifuzzaman received his BSc degree in Computer Engineering from North South University (NSU), Bangladesh in 2008. The following year, he worked at the IT department of Shimizu Densetsu Kogyo Co. Ltd (SEAVAC) in Japan. He completed his MSc degree in Electronic Design from Lund University, Sweden in 2013. Since 2014, he has been with the Memory Systems group of Barcelona Supercomputing Center (BSC) as well as a PhD student at the department of computer architecture of Universitat Politècnica de Catalunya (UPC), Spain.

Model-based ML for Retrospective Event Detection

Joan Capdevila*†, Jesús Cerquides‡, and Jordi Torres*†
* Barcelona Supercomputing Center (BSC), Barcelona, Spain
† Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
‡ IIIA-CSIC, Bellaterra, Spain
E-mail: {joan.capdevila, jordi.torres}@bsc.es, cerquide@iiia.csic.es

Keywords—Model-based Machine Learning, Probabilistic Models, Variational Inference, Social Computing, Twitter.

I. Introduction

The problem of event detection in Twitter has lately attracted the interest of several distinct communities ranging from data miners, spatial statisticians to machine learners. Each community has proposed tailored solutions to a problem which has multiple definitions in the literature. This has led to a myriad of techniques which address similar but slightly different problems [1]. In this paper, we focus on Retrospective Event Detection (RED) from geo-located tweets. RED seeks to identify and characterize groups of tweets which belong to a past unseen event [2]. This problem can be of interest to news agencies, city councils or geo-marketing companies.

One common approach to RED formulates the problem as one of clustering with noise. That is to say that event-related tweets are associated to event clusters while non-event tweets are assigned to a noise or background component. For example, Capdevila et al. presented Tweet-SCAN in [3], a technique for RED that extends the popular clustering technique called DBSCAN [4] to cope with geo-located tweets. However, this type of approach to machine learning constraints the solution to the assumptions of the chosen technique. For example, Tweet-SCAN suffers from the same weakness than DBSCAN which is that all event-related clusters must have the same minimum tweet density. Addressing this shortcoming within this approach would certainly require to replace DBSCAN with OPTICS or another technique. In most cases, this type of changes prevent us to reuse parts of the original solution when trying to incorporate more complex assumptions.

Model-based Machine Learning [5] offers an alternative approach which separates assumptions from computation and tasks. It enables to explicitly specify the assumptions in a compact modeling language, then define an inference algorithm to learn the model from data and finally carry out the tasks as predictions on the trained model. Moreover, Blei [6], inspired by the work of George E.P. Box, proposed to close the loop with criticism, so that assumptions can be revised when the solution does not meet the requirements. In this paper, we present WARBLE [7], a model-based solution that specifies the assumptions through a probabilistic graphical model, defines a variational inference algorithm to learn the model from "La Mercè" data set and performs event detection through a Maximum a Posteriori (MAP) query on the event assignment variables. We concluded in criticism section presenting some results and conclusions of this approach.

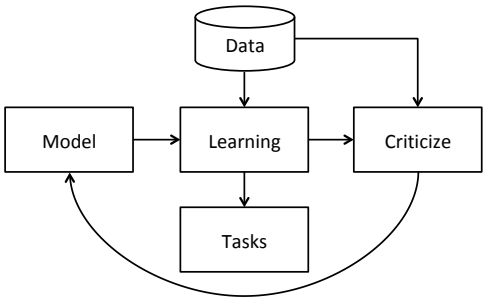


Fig. 1: Box's loop applied to Model-based Machine Learning [6].

II. The Warble solution

A. The Probabilistic Model

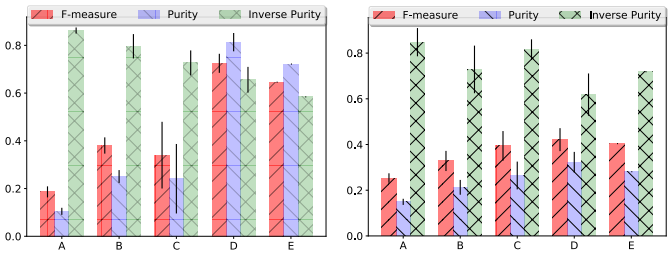
The WARBLE model is a heterogeneous mixture model with two types of mixture components. On the one hand, the event components represent the different events in terms of their spatio-temporal and textual features. On the other hand, the non-event component depicts the non-event tweets in its steady conditions. The generation process for event and non-event tweets can be described through the following generation processes,

Event tweet	Non-event tweet
$l_n \sim \text{Normal}(\mu_k, \Delta_k)$	$l_n \sim \text{Hist}(L_B)$
$t_n \sim \text{Normal}(\tau_k, \lambda_k)$	$t_n \sim \text{Hist}(T_B)$
For $m = 1...M_n$:	For $m = 1...M_n$:
$z_{n,m} \sim \text{Discrete}(\theta_k)$	$z_{n,m} \sim \text{Discrete}(\theta_K)$
$w_{n,m} \sim \text{Discrete}(\phi_{z_{n,m}:})$	$w_{n,m} \sim \text{Discrete}(\phi_{z_{n,m}:})$

where l_n and t_n are the tweet location and time which we assume to be normally distributed for event tweets. Non-event tweets follow a Histogram distribution that models the varying spatio-temporal tweet density in steady conditions, addressing one of the shortcomings of Tweet-SCAN [3] . $w_{n,:}$ correspond to the M_n words in the tweet, which are distributed according to a mixture of topics ϕ for both type of tweets. The proposed model also incorporates some prior distribution over the means, precisions and proportions of the above distributions (see [7] for more details).

B. The Variational Learning Algorithm

The Bayesian approach of learning consists in computing the posterior distribution over all the model unknowns. How-



(a) Mercè 2014 (b) Mercè 2015

Fig. 2: (A) McInerney & Blei model (B) WARBLE w/o simultaneous topic-event learning (C) WARBLE w/o background model (D) Complete WARBLE (E) Tweet-SCAN

ever, this posterior distribution is intractable in the WARBLE model since it involves computing a normalizing constant which has to marginalize all latent variables and parameters. Approximation methods have been developed to bound this normalizing constant and solve a simpler optimization problem. Mean-field variational inference assumes a factorized distribution $q(X|\eta)$ over all model parameters $X = \{c, z, \pi, \tau, \lambda, \mu, \Delta, \theta, \phi\}$ and maximizes the evidence lower bound (ELBO) w.r.t the variational parameters η

$$\text{ELBO}(\eta) = \mathbb{E}_{q(X|\eta)}\left[p(l,t,w,X;\Gamma)\right] - \mathbb{E}_{q(X|\eta)}\left[q(X|\eta)\right] \quad (1)$$

where Γ refers to the set of hyperparameters. Thanks to the mean-field approximation and the local conjugacy, we can build a coordinate descent algorithm with close-form updates for each variational parameter.

C. The Event Detection Task

The event detection task can be seen as MAP query on the event assignment variables in which we assign the most probable mixture components (event or non-event) to each tweet. However, this inference task is also intractable and we need to resort to its variational approximation:

$$c^* = \underset{c}{\text{argmax}}\, p(c|l,t,w;\Gamma) \approx \underset{c}{\text{argmax}}\, q(c;\eta). \quad (2)$$

We note that this approach enables to formulate other tasks as queries to the probabilistic model. For example, we could ask for the most likely location for a tweet without geo-location.

D. "La Mercè" data sets

We have crawled two datasets of tweets geo-located in the city of Barcelona during its local festivities in 2014 and 2105 [1]. Several event-related tweets were manually tagged based on the festivities agenda. Tags were only used for evaluation purposes. Moreover, the histogram distributions for the spatio-temporal profiles for the non-event component were built from tweets generated during the days previous to the period of interest.

III. CRITICISM

In Fig. 2, we compare the performance of WARBLE against other non-model-based techniques like Tweet-SCAN [3] and against other model-based approaches like McInerney & Blei model [8]. We observe that WARBLE outperforms all techniques in both datasets in terms of F-measure. More importantly, we show how the iterative process of Model-based machine learning works. McInerney & Blei model did not distinguish between event and non-event tweets. By replacing their homogeneous mixture with a heterogeneous, we were able to improve accuracy from A to B. McInerney & Blei model did not perform simultaneous learning of topics-events. By doing so, we improve results from A to C. When combining both features into a more complex solution, we come up with WARBLE which has a performance as in D.

As we have seen, this approach to machine learning enables to be explicit about the assumptions, and hence criticize and improve different parts of the model. Moreover, it allows to decouple the model from inference and think of the computational aspects independently. This has pushed the field of *probabilistic programming* to develop software solutions which automatize inference for a given model.

IV. ACKNOWLEDGMENT

This work is partially supported by Obra Social la Caixa. The extended abstract has been already published in [7].

REFERENCES

[1] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.

[2] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 28–36.

[3] J. Capdevila, J. Cerquides, J. Nin, and J. Torres, "Tweet-scan: An event discovery technique for geo-located tweets," *Pattern Recognition Letters*, vol. 93, pp. 58 – 68, 2017, pattern Recognition Techniques in Data Mining.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[5] C. M. Bishop, "Model-based machine learning," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20120222, 2013.

[6] D. M. Blei, "Build, compute, critique, repeat: Data analysis with latent variable models," *Annual Review of Statistics and Its Application*, vol. 1, pp. 203–232, 2014.

[7] J. Capdevila, J. Cerquides, and J. Torres, "Mining urban events from the tweet stream through a probabilistic mixture model," *Data Mining and Knowledge Discovery*, Aug 2017.

[8] J. McInerney and D. M. Blei, "Discovering newsworthy tweets with a geographical topic model," *NewsKDD: Data Science for News Publishing workshop. Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.

[1] https://github.com/jcapde/Twitter-DS/tree/master/MERCE/

Joan Capdevila is a Ph.D. student at the Computer Architecture Department at Universitat Politècnica de Catalunya (UPC) and associated to the Barcelona Supercomputing Center (BSC). He is supported by "la Caixa" Foundation through the UPC - 'la Caixa" program. He has an Eng. degree in Telecommunications and a M.Sc. in Information and Communication Technologies by UPC. His research interest are in probabilistic modeling and approximate inference methods.

An assessment of regional sea ice predictability in the Arctic ocean

Rubén Cruz-García*, Virginie Guemas, Matthieu Chevallier, François Massonnet
*Barcelona Supercomputing Center (BSC-ES), C/ Jordi Girona 29, 08034, Barcelona, Spain.
Email: ruben.cruzgarcia@bsc.es

Abstract*—Arctic sea ice plays a central role in the Earth's climate. Changes in the sea ice on seasonal-to-interannual timescales impact ecosystems, populations and a growing number of stakeholders. A prerequisite for achieving better sea ice predictions is a better understanding of the underlying mechanisms of sea ice predictability. Previous studies have shown that sea ice predictability depends on the predictand (area, extent, volume), region, and the initial and target dates. Here we investigate seasonal-to-interannual sea ice predictability in so-called "perfect model" 3-year-long experiments run with the EC-Earth 2.3 climate model initialized in early July. Consistent with previous studies, robust mechanisms for reemergence are highlighted, i.e. increases in the autocorrelation of sea ice properties after an initial loss. We find that Arctic regions can be classified according to three distinct regimes. The central Arctic drives most of the pan-Arctic sea ice volume persistence. In peripheral seas, we find trivial predictability for the sea ice area in winter but low predictability throughout the rest of the year, due to the particularly unpredictable sea ice edge location. The Labrador Sea stands out among the considered regions, with sea ice predictability extending up to 1.5 years if the oceanic conditions upstream are known.

I. Methodology

We used a 300-year long present day control experiment under perpetual 2005 forcing (*ControlRun*). This single-member experiment provides the initial conditions used to perform a set of idealized climate prediction experiments initialized from July 1st (*IdealPred*). The predictions last 3 years and consist of 8 members, each of them with slightly different perturbations of the initial sea surface temperature (SST; 10^{-4} K magnitude).

For evaluating the predictability we consider the **prognostic potential predictability** (**PPP** hereafter). It compares the ensemble spread with an estimation of the amplitude of the natural variability of the system based on the standard deviation of the control simulation, and addresses the initial value predictability. A PPP value of 1 would mean that we have a perfectly predictable system. Predictability is estimated both in a prognostic (PPP) and diagnostic (lagged *ControlRun* properties) way. The **prognostic approach** suffers from insufficient sampling, in contrast with the long control simulation, that can be used to supply that problem.

Breaking down the **analysis into sectors** is essential since the pan-Arctic sea ice extent (SIE), area (SIA) and volume (SIV) integrate a large variety of regions which are regulated by different physical mechanisms. Thus, predictability was investigated for each Arctic sea (Fig. 1). Lagged SIE autocorrelation for each month against lead time shows the **September to September** correlation reemergence (**leadtime 12**), consistent with Blanchard et al. (AMS, 2011) mechanism from one summer to the next. The **melt-to-freeze** mechanism is present in July (**leadtime 3**). The SIV memory regime is characterized by its vast persistence for all start months.

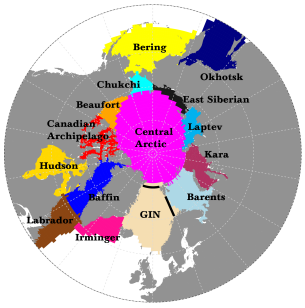


Figure 1: Map of the Arctic seas as defined in this study. The black lines indicate the sections used for the calculation of the Atlantic heat transport into the Arctic (Fram Strait plus Barents Sea Opening). The GIN region is formed by the Greenland, Icelandic and Norwegian seas.

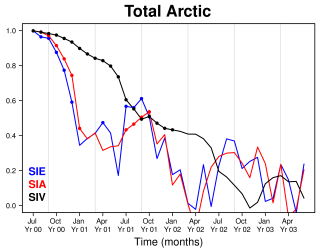


Figure 2: Potential predictability of the total Arctic SIE (blue), SIA (red) and SIV (black) measured with the PPP of *IdealPred* using the natural variability of *ControlRun* as a reference. Dots indicate significant values at the 95 % level, estimated by Fisher's test. September and March are marked by thin gray vertical lines.

II. Pan-Arctic sea ice

The **melt-to-freeze** mode is not only present in the lagged correlations (not shown), but it is also a feature in the predictions initialized in July (Fig. 2).

The **long SIV predictability** agrees with the lagged correlations (not shown). This persistence comes almost entirely from the central Arctic SIV memory, as can be checked when comparing the lagged correlation of central and pan-Arctic SIV (Fig. 2; blue and black lines).

Summer-to-summer memory reemergence has its origin in the summer SIT memory (from the central Arctic). Over three continuous years, the **central SIV and SIE are synchronously correlated in September** (not shown).

III. Regional Arctic sea ice

In the Barents Sea, peaks of reemergence occur the second and third summer (not shown). Synchronous correlation between the SST and the SIE (Fig. 3, red line) reveals that **SST** is a **source of SIE predictability** in December. Correlation between the gridpoint SST in December and SIE from December to February confirms this timeseries. The SST during the previous spring also provides predictability to the December SIE (not shown).

The PPP of the SIE in the interior basins (e.g. the Canadian Archipelago; Fig. 4) saturates in winter because of the **extremely**

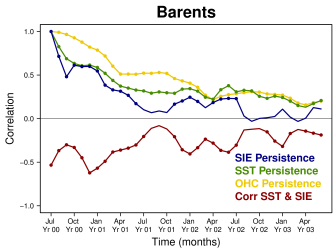


Figure 3: The persistence of the SIE (blue), the SST (green), OHC (0-300 m depth, yellow) for the Barents Sea. In red, the synchronous correlation between the SST and the SIE. Correlations were calculated using the *ControlRun* during the three subsequent years. The dots represent significant values at the 95% level as estimated from a one-sided student-T distribution.

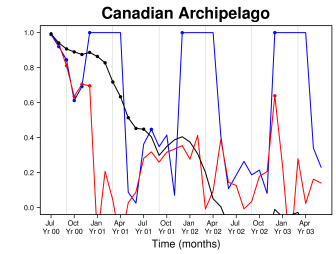


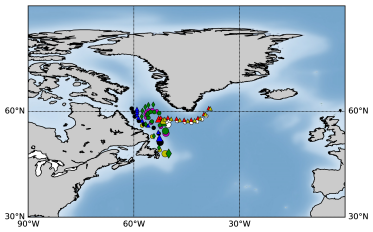
Figure 4: As Fig. 2 for the Canadian Archipelago.

low sea ice variability. SIA PPP differs from SIE signal because its variability is larger in winter, with non-fully covered ice regions. In most of central regions SIV is potentially predictable up to one year before.

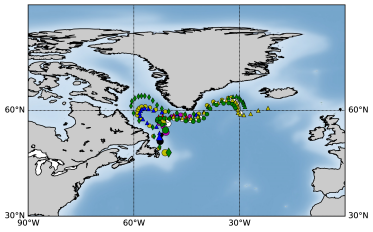
Backward trajectories from the **Labrador Sea** reveal that the **water masses origin is the Irminger Sea**, and the North Atlantic Ocean in a longer term (Fig. 5a-b). The **Irminger Sea SST and ocean heat content (OHC)** at the moment of the initialization and the Labrador Sea SIE are significantly anticorrelated from February to July the two first years, matching exactly the time when the PPP reemergence in the Labrador Sea occurs (Fig. 5c).

IV. Conclusions

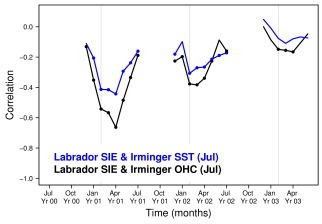
- Pan–Arctic SIE experiences melt-to-freeze reemergence both in the prognostic ensemble potential predictability and in the control run lagged correlations. The SIV shows greater predictability, attributable to the long-lasting persistence of the SIT in the central Arctic SIT.
- The summer-to-summer reemergence of the PPP of pan-Arctic SIE is due to the persistence of SIT anomalies in the central Arctic.
- In the peripheral seas of the Atlantic Sector, significantly high PPP values over 1 year are driven by the persistence of local oceanic thermal anomalies (SST and OHC).
- In the Labrador Sea, which is ice-free in July, the PPP peaks between January and April as result of the advection of ocean temperature anomalies from the Irminger Sea and the Eastern North Atlantic Ocean.
- In the interior Arctic seas, winter SIE potential predictability is trivial due to complete ice coverage. No significant predictability was found for the SIA. In contrast, the SIV has a longer predictability in this set of seas as a result of the long SIT persistence.



(a)



(b)



(c)

Figure 5: Map of the backward trajectories followed by water masses travelling from different locations in the Labrador Sea from **(a)** the first and **(b)** the second February until the first July. Each lead time is marked with a dot, while the initial positions (corresponding February) are marked with bigger dots. **(c)** Correlation between the Labrador Sea SIE and the Irminger SST (in blue) and the Irminger OHC (0-300 m depth; in black) the first July for the *ControlRun* during the three following years. The dots represent the significant values at the 95% level estimated from a one-sided student-T distribution. The vertical grey lines represent the months of February. The SST and OHC were integrated for the corresponding area in Fig. 1.

Rubén Cruz-García. Rubén studied his bachelor in Environmental Sciences at University of Murcia, where he discovered his passion for investigating the climate change. After that, he began a Master degree in Geophysics and Meteorology at University of Granada. He started his PhD project in the Climate Prediction Group at the Earth Sciences Department of the Barcelona Supercomputing Center in October 2015. His work focuses on assessing the predictability and the prediction skill of the Arctic sea ice conditions at the regional scale using two state-of-the-art dynamical models, EC-Earth and CNRM-CM.

Improving trimAl ability to cope with heterogeneous multiple sequence alignments.

Víctor Fernández-Rodríguez[#1],, Toni Gabaldón[#3,4,5], Salvador Capella-Gutierrez[#2]

[#1,2]*Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain.*

[1]victor.fernandez@bsc.es, [2]salvador.capella@bsc.es

[#3]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain*

[#4]*Universitat Pompeu Fabra (UPF), Barcelona, Spain*

[#5]*Institucio Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Spain*

[3]toni.gabaldon@crg.eu

Keywords - Multiple Sequence Alignment, Alignment filtering, Phylogenetic analysis

EXTENDED ABSTRACT

Alignments of biological sequences, called Multiple Sequence Alignments (MSA), are the entrypoint for many biological applications including evolutionary studies. However, the current algorithms used to reconstruct them tend to minimize (or maximize) mathematical functions rather than truly representing biological events. This is especially relevant for highly variable sequences regions where the positional homology is difficult to infer. This often tends to produce MSAs with a high noise-to-signal ratio, which will be eventually amplificated on downstream analyses that rely on them.

Thus, MSAs refinement has become a common practice in many biological domains. However, MSAs refinement algorithms are not except of errors so further investigation is needed making this area a very active research field.

Here we present a revisited version of trimAl, a popular resource aiming to improve MSAs using manual and/or automated methods. We will explain why is important to refactor trimAl's source code including issues found and solutions applied. Finally we will introduce a set of new functionality only achieved after improving the existing source code.

A. *Introduction*

trimAl was born as a internal laboratory script, that grew fast in functionality and code length. The original code was written in C, and later moved to C++ to exploit the Object Oriented Programming Paradigm (OOP).

The fast growth of the code, due to addition of new functionality and lack of a project pre-production phase led to a fully functional and almost bug-free but coupled code with some evident issues.

For this reason, in the present document we explain some relevant aspects of the refactoring step performed and the results obtained through the process.

B. *Format Machine State*

We have implemented a machine state to load and save MSAs in different formats, which allows to isolate the format handling code from the rest of the program.

This new paradigm allows to remove and/or implement new format handlers with ease, and also, allows the community to provide their own format handlers.

C. *Memory Improvement*

The original implementation loaded into memory a copy of the complete MSA each time any operation was applied. This leads to have a high degree of redundancies among loaded copies e.g. sequences names, metadata, etc. Indeed, up to three MSAs containing subsets of the same information are allocated at the same time in memory: original alignment, so we can compare the result obtained with the original; current alignment, the one that is being processed at the current step, and the resulting alignment.

In the new implementation, we followed a different memory management strategy, at a potential cost of performance.

We have the data on memory once, and all copies would point to that information and contain a pair of vectors indicating if we would reject or keep specific columns and/or sequences

D. *Speed Improvement*

One of the effects that highly coupled code had on the original implementation was that some statistics were computed more than once, increasing the time needed to perform an analysis.

The new approach allowed us to detect and avoid repeating calculations, and thus, reduce the time needed to perform the same analysis, with the same results.

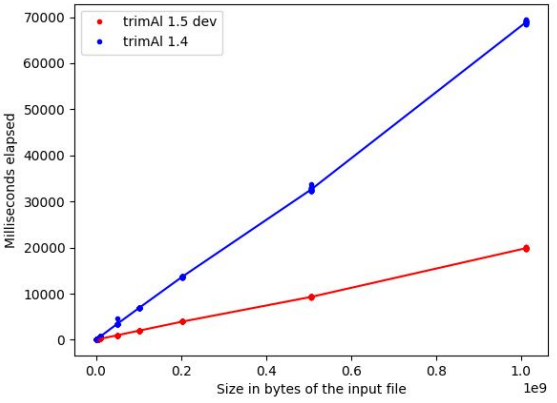


Fig.1. Time needed by the original and new implementations of trimAl using the strict algorithm (the most consuming of the program).

E. *Reporting Improvement*

Reporting has been improved in several ways: Statistics

report has been eased visually, using the correct tabulation and adding a header specifying the statistic being reported and the original filename of the alignment which it is extracted from.

More relevant is the new format for trimming reporting: The original implementation outputted an HTML file with a graphic visualization of the results of the trimming steps and the statistics used to perform these steps.

This allows the user to have an insight of what was removed and why.

The new format, SVG, allows faster load, and better representation of the statistics, using a graph approximation, where the original used categories.

This lead to a more informative reporting, and also, more useful, as the report can be treated as a vectorial image, allowing to cut and scale it as much as needed.

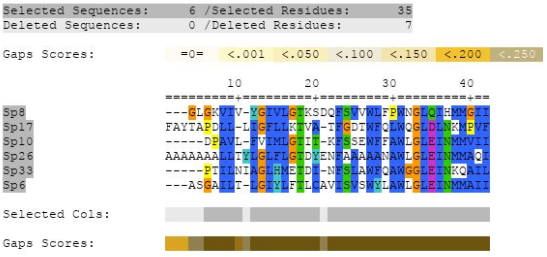


Fig.2. HTML version of the trimming report.

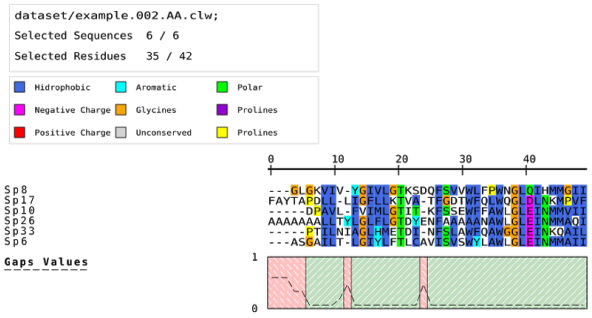


Fig.3. SVG version of the trimming report.

F. *Error Reporter*

Centralization and standardization of the errors and warnings that we provide to the users was required, as some problems had arisen from the lack of them.

These problems include reporting the same error with different messages, which would lead to confusion to the final user or having to do code scraping to find all the calls to an error message that we would like to change.

An Error Reporter has been created, isolating the code of error reporting from the rest of the code.

This allowed us to create a numbered list of errors, that allows to a better understanding of the situations that may arise from the use of the program.

It also allowed us to add a verbose option, allowing for better reporting control to the end user.

G. *Time Tracker*

To have a better understanding of the flow performed by the program, an auxiliary class has been implemented: the Time Tracker.

This class tracks the calls to most of the methods in the program, and outputs a tree where we can easily understand which methods calls to others in a specific execution, and calculate the time each method lasts, including and excluding calls to other tracked methods.

This allows us to see if the program behaves exactly as expected, and to pinpoint which methods are candidates to optimization.

This functionality has been enhanced by adding the ability to track memory before and after each method call. This allows to have a better understanding of the memory management on each method and globally.

H. *Conclusion and Future Enhancement*

Short-term future foresight includes containerization of the binaries, a complete revamp of the suite website and the extension to support Next Generation Sequencing (NGS) data. In the long run, we will deeply analyse the existing trimming algorithms to propose new ones, which can cope with MSAs made up to ten of thousands of sequences. Any newly developed method will be extensively benchmark to ensure their scientific and technical relevance for current and future end-users.

References

[1] Capella-Gutiérrez, S. & Gabaldón, T. Measuring guide-tree dependency of inferred gaps in progressive aligners. Bioinformatics 29, 1011–1017 (2013).

[2] Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma. Appl. NOTE 25, 1972–197310 (2009).

[3] Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37 (2010).

Author biography

Víctor Fernández was born in Valencia, Spain, in 1992. He studied Biology in Universitat de València, Burjassot, where he developed an strong interest in genetics. After his degree, he worked on a startup developing video games, with the intention of improving his programming abilities to be able to obtain the maximum possible of his master's degree in bioinformatics.

Since November 2016 he has been enrolled in the Bioinformatics and Biological Computation Master degree at the Escuela Nacional de Sanidad (Madrid), and since July, he is part of the team developing the new version of trimAl at the BSC.

Improving trimAl ability to cope with heterogeneous multiple sequence alignments.

Víctor Fernández-Rodríguez[#1],, Toni Gabaldón[#3,4,5], Salvador Capella-Gutierrez[#2]

[#1,2]*Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain.*

[1]victor.fernandez@bsc.es, [2]salvador.capella@bsc.es

[#3]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain*

[#4]*Universitat Pompeu Fabra (UPF), Barcelona, Spain*

[#5]*Institucio Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Spain*

[3]toni.gabaldon@crg.eu

Keywords - Multiple Sequence Alignment, Alignment filtering, Phylogenetic analysis

EXTENDED ABSTRACT

Alignments of biological sequences, called Multiple Sequence Alignments (MSA), are the entrypoint for many biological applications including evolutionary studies. However, the current algorithms used to reconstruct them tend to minimize (or maximize) mathematical functions rather than truly representing biological events. This is especially relevant for highly variable sequences regions where the positional homology is difficult to infer. This often tends to produce MSAs with a high noise-to-signal ratio, which will be eventually amplificated on downstream analyses that rely on them.

Thus, MSAs refinement has become a common practice in many biological domains. However, MSAs refinement algorithms are not except of errors so further investigation is needed making this area a very active research field.

Here we present a revisited version of trimAl, a popular resource aiming to improve MSAs using manual and/or automated methods. We will explain why is important to refactor trimAl's source code including issues found and solutions applied. Finally we will introduce a set of new functionality only achieved after improving the existing source code.

A. *Introduction*

trimAl was born as a internal laboratory script, that grew fast in functionality and code length. The original code was written in C, and later moved to C++ to exploit the Object Oriented Programming Paradigm (OOP).

The fast growth of the code, due to addition of new functionality and lack of a project pre-production phase led to a fully functional and almost bug-free but coupled code with some evident issues.

For this reason, in the present document we explain some relevant aspects of the refactoring step performed and the results obtained through the process.

B. *Format Machine State*

We have implemented a machine state to load and save MSAs in different formats, which allows to isolate the format handling code from the rest of the program.

This new paradigm allows to remove and/or implement new format handlers with ease, and also, allows the community to provide their own format handlers.

C. *Memory Improvement*

The original implementation loaded into memory a copy of the complete MSA each time any operation was applied. This leads to have a high degree of redundancies among loaded copies e.g. sequences names, metadata, etc. Indeed, up to three MSAs containing subsets of the same information are allocated at the same time in memory: original alignment, so we can compare the result obtained with the original; current alignment, the one that is being processed at the current step, and the resulting alignment.

In the new implementation, we followed a different memory management strategy, at a potential cost of performance.

We have the data on memory once, and all copies would point to that information and contain a pair of vectors indicating if we would reject or keep specific columns and/or sequences

D. *Speed Improvement*

One of the effects that highly coupled code had on the original implementation was that some statistics were computed more than once, increasing the time needed to perform an analysis.

The new approach allowed us to detect and avoid repeating calculations, and thus, reduce the time needed to perform the same analysis, with the same results.

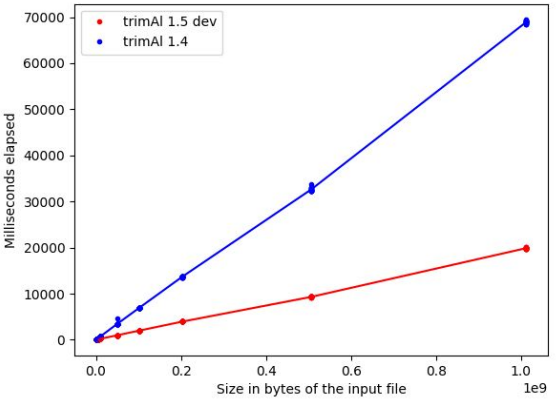


Fig.1. Time needed by the original and new implementations of trimAl using the strict algorithm (the most consuming of the program).

E. *Reporting Improvement*

Reporting has been improved in several ways: Statistics

report has been eased visually, using the correct tabulation and adding a header specifying the statistic being reported and the original filename of the alignment which it is extracted from.

More relevant is the new format for trimming reporting: The original implementation outputted an HTML file with a graphic visualization of the results of the trimming steps and the statistics used to perform these steps.

This allows the user to have an insight of what was removed and why.

The new format, SVG, allows faster load, and better representation of the statistics, using a graph approximation, where the original used categories.

This lead to a more informative reporting, and also, more useful, as the report can be treated as a vectorial image, allowing to cut and scale it as much as needed.

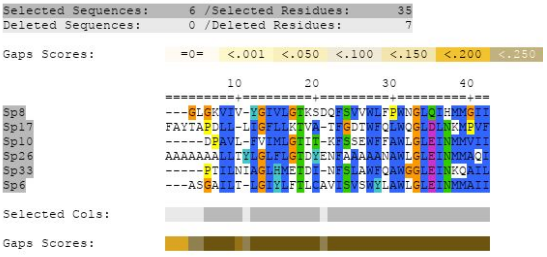


Fig.2. HTML version of the trimming report.

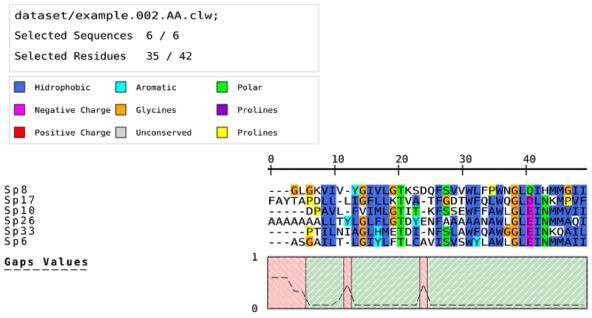


Fig.3. SVG version of the trimming report.

F. *Error Reporter*

Centralization and standardization of the errors and warnings that we provide to the users was required, as some problems had arisen from the lack of them.

These problems include reporting the same error with different messages, which would lead to confusion to the final user or having to do code scraping to find all the calls to an error message that we would like to change.

An Error Reporter has been created, isolating the code of error reporting from the rest of the code.

This allowed us to create a numbered list of errors, that allows to a better understanding of the situations that may arise from the use of the program.

It also allowed us to add a verbose option, allowing for better reporting control to the end user.

G. *Time Tracker*

To have a better understanding of the flow performed by the program, an auxiliary class has been implemented: the Time Tracker.

This class tracks the calls to most of the methods in the program, and outputs a tree where we can easily understand which methods calls to others in a specific execution, and calculate the time each method lasts, including and excluding calls to other tracked methods.

This allows us to see if the program behaves exactly as expected, and to pinpoint which methods are candidates to optimization.

This functionality has been enhanced by adding the ability to track memory before and after each method call. This allows to have a better understanding of the memory management on each method and globally.

H. *Conclusion and Future Enhancement*

Short-term future foresight includes containerization of the binaries, a complete revamp of the suite website and the extension to support Next Generation Sequencing (NGS) data. In the long run, we will deeply analyse the existing trimming algorithms to propose new ones, which can cope with MSAs made up to ten of thousands of sequences. Any newly developed method will be extensively benchmark to ensure their scientific and technical relevance for current and future end-users.

References

[1] Capella-Gutiérrez, S. & Gabaldón, T. Measuring guide-tree dependency of inferred gaps in progressive aligners. Bioinformatics 29, 1011–1017 (2013).

[2] Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma. Appl. NOTE 25, 1972–197310 (2009).

[3] Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37 (2010).

Author biography

Víctor Fernández was born in Valencia, Spain, in 1992. He studied Biology in Universitat de València, Burjassot, where he developed an strong interest in genetics. After his degree, he worked on a startup developing video games, with the intention of improving his programming abilities to be able to obtain the maximum possible of his master's degree in bioinformatics.

Since November 2016 he has been enrolled in the Bioinformatics and Biological Computation Master degree at the Escuela Nacional de Sanidad (Madrid), and since July, he is part of the team developing the new version of trimAl at the BSC.

Improving trimAl ability to cope with heterogeneous multiple sequence alignments.

Víctor Fernández-Rodríguez[#1],, Toni Gabaldón[#3,4,5], Salvador Capella-Gutierrez[#2]

[#1,2]*Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain.*

[1]victor.fernandez@bsc.es, [2]salvador.capella@bsc.es

[#3]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain*

[#4]*Universitat Pompeu Fabra (UPF), Barcelona, Spain*

[#5]*Institucio Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Spain*

[3]toni.gabaldon@crg.eu

Keywords - Multiple Sequence Alignment, Alignment filtering, Phylogenetic analysis

EXTENDED ABSTRACT

Alignments of biological sequences, called Multiple Sequence Alignments (MSA), are the entrypoint for many biological applications including evolutionary studies. However, the current algorithms used to reconstruct them tend to minimize (or maximize) mathematical functions rather than truly representing biological events. This is especially relevant for highly variable sequences regions where the positional homology is difficult to infer. This often tends to produce MSAs with a high noise-to-signal ratio, which will be eventually amplificated on downstream analyses that rely on them.

Thus, MSAs refinement has become a common practice in many biological domains. However, MSAs refinement algorithms are not except of errors so further investigation is needed making this area a very active research field.

Here we present a revisited version of trimAl, a popular resource aiming to improve MSAs using manual and/or automated methods. We will explain why is important to refactor trimAl's source code including issues found and solutions applied. Finally we will introduce a set of new functionality only achieved after improving the existing source code.

A. *Introduction*

trimAl was born as a internal laboratory script, that grew fast in functionality and code length. The original code was written in C, and later moved to C++ to exploit the Object Oriented Programming Paradigm (OOP).

The fast growth of the code, due to addition of new functionality and lack of a project pre-production phase led to a fully functional and almost bug-free but coupled code with some evident issues.

For this reason, in the present document we explain some relevant aspects of the refactoring step performed and the results obtained through the process.

B. *Format Machine State*

We have implemented a machine state to load and save MSAs in different formats, which allows to isolate the format handling code from the rest of the program.

This new paradigm allows to remove and/or implement new format handlers with ease, and also, allows the community to provide their own format handlers.

C. *Memory Improvement*

The original implementation loaded into memory a copy of the complete MSA each time any operation was applied. This leads to have a high degree of redundancies among loaded copies e.g. sequences names, metadata, etc. Indeed, up to three MSAs containing subsets of the same information are allocated at the same time in memory: original alignment, so we can compare the result obtained with the original; current alignment, the one that is being processed at the current step, and the resulting alignment.

In the new implementation, we followed a different memory management strategy, at a potential cost of performance.

We have the data on memory once, and all copies would point to that information and contain a pair of vectors indicating if we would reject or keep specific columns and/or sequences

D. *Speed Improvement*

One of the effects that highly coupled code had on the original implementation was that some statistics were computed more than once, increasing the time needed to perform an analysis.

The new approach allowed us to detect and avoid repeating calculations, and thus, reduce the time needed to perform the same analysis, with the same results.

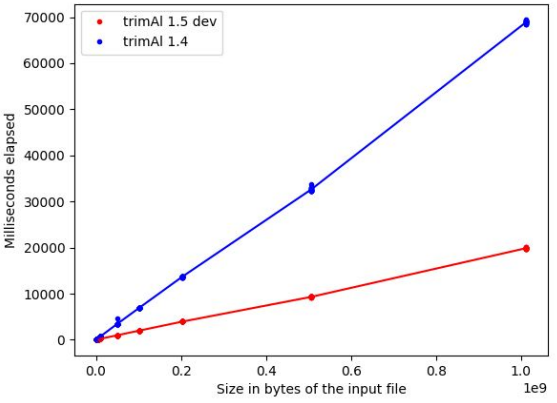


Fig.1. Time needed by the original and new implementations of trimAl using the strict algorithm (the most consuming of the program).

E. *Reporting Improvement*

Reporting has been improved in several ways: Statistics

report has been eased visually, using the correct tabulation and adding a header specifying the statistic being reported and the original filename of the alignment which it is extracted from.

More relevant is the new format for trimming reporting: The original implementation outputted an HTML file with a graphic visualization of the results of the trimming steps and the statistics used to perform these steps.

This allows the user to have an insight of what was removed and why.

The new format, SVG, allows faster load, and better representation of the statistics, using a graph approximation, where the original used categories.

This lead to a more informative reporting, and also, more useful, as the report can be treated as a vectorial image, allowing to cut and scale it as much as needed.

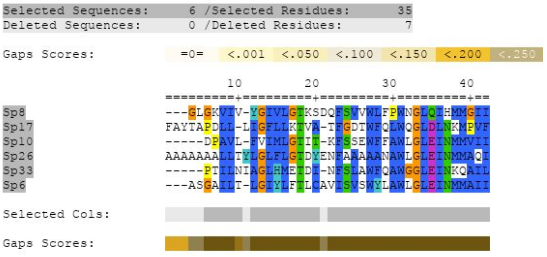


Fig.2. HTML version of the trimming report.

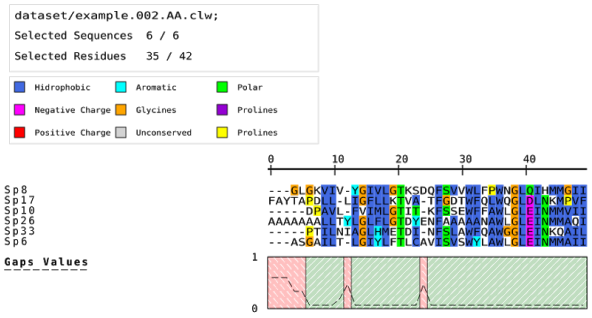


Fig.3. SVG version of the trimming report.

F. *Error Reporter*

Centralization and standardization of the errors and warnings that we provide to the users was required, as some problems had arisen from the lack of them.

These problems include reporting the same error with different messages, which would lead to confusion to the final user or having to do code scraping to find all the calls to an error message that we would like to change.

An Error Reporter has been created, isolating the code of error reporting from the rest of the code.

This allowed us to create a numbered list of errors, that allows to a better understanding of the situations that may arise from the use of the program.

It also allowed us to add a verbose option, allowing for better reporting control to the end user.

G. *Time Tracker*

To have a better understanding of the flow performed by the program, an auxiliary class has been implemented: the Time Tracker.

This class tracks the calls to most of the methods in the program, and outputs a tree where we can easily understand which methods calls to others in a specific execution, and calculate the time each method lasts, including and excluding calls to other tracked methods.

This allows us to see if the program behaves exactly as expected, and to pinpoint which methods are candidates to optimization.

This functionality has been enhanced by adding the ability to track memory before and after each method call. This allows to have a better understanding of the memory management on each method and globally.

H. *Conclusion and Future Enhancement*

Short-term future foresight includes containerization of the binaries, a complete revamp of the suite website and the extension to support Next Generation Sequencing (NGS) data. In the long run, we will deeply analyse the existing trimming algorithms to propose new ones, which can cope with MSAs made up to ten of thousands of sequences. Any newly developed method will be extensively benchmark to ensure their scientific and technical relevance for current and future end-users.

References

[1] Capella-Gutiérrez, S. & Gabaldón, T. Measuring guide-tree dependency of inferred gaps in progressive aligners. Bioinformatics 29, 1011–1017 (2013).

[2] Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma. Appl. NOTE 25, 1972–197310 (2009).

[3] Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37 (2010).

Author biography

Víctor Fernández was born in Valencia, Spain, in 1992. He studied Biology in Universitat de València, Burjassot, where he developed an strong interest in genetics. After his degree, he worked on a startup developing video games, with the intention of improving his programming abilities to be able to obtain the maximum possible of his master's degree in bioinformatics.

Since November 2016 he has been enrolled in the Bioinformatics and Biological Computation Master degree at the Escuela Nacional de Sanidad (Madrid), and since July, he is part of the team developing the new version of trimAl at the BSC.

On the quest to reach nuclear fusion as a future energy source

Dani Gallart*, Mervi Mantsinen*†, JET Contributors‡
*Barcelona Supercomputing Center, Barcelona, Spain
†ICREA, Barcelona, Spain
E-mail: daniel.gallart@bsc.es

Keywords—Plasma, Fusion, ICRF+NBI heating.

I. Introduction

[1] The advantages of nuclear fusion are numerous. It is capable of producing large amounts of energy, its fuel is virtually unlimited as it is extracted from water and it is environmentally friendly as it produces no long-term radioactive waste. However, the goal of achieving controlled nuclear fusion as a future energy source has not yet been reached. Since the early 1950s the scientific community has been working on this field and achieved several milestones such as the design of a toroidal experimental reactor, the so-called tokamak. The main difficulty lies in the complexity of reaching positive energy balance, i.e., $\mathrm{Q} = \mathrm{E_{out}}/\mathrm{E_{in}} > 1$. The plasma, which is a mixture of hot ions and electrons, needs to be confined inside the reactor in order to avoid losses. Stability and confinement are major issues together with radiative losses which are physically inherent to the system. All in all, controlled nuclear fusion is a challenging physical and engineering problem with potential to solve the increasing energy demands of the world's growing population.

Magnetic confinement fusion is one of approaches to develop fusion energy and the subject of this work. It is based upon confining hot plasma using strong magnetic fields by bending the ion and electron trajectories through the Lorentz force. Typically, plasma is composed by hydrogen (H) isotopes such as deuterium (D) or tritium (T) which are heated beyond their ionization energy. Plasma heating is of fundamental importance as it is necessary for fusion reactions to occur and to maximize them. There are two main external methods to heat the plasma, i.e. through the injection of energetic neutral beam particles (NBI) or radiofrequency heating such as heating with electromagnetic waves in the ion cyclotron range of frequencies (ICRF). In some cases the applied heating mechanisms interact with each other. This is the case for example when the frequency of the ICRF wave matches the cyclotron frequency of the beam particles. This effect is known as the ICRF-NBI synergy.

In this work we study the impact of NBI and ICRF heating on the fusion performance of several hybrid discharges at the Joint European Torus (JET), UK. JET is the largest operating tokamak in the world and the only one capable of operating with the reactor relevant D-T fuel mixture. The hybrid scenario is an advanced regime expected to operate in ITER, the fusion reactor being built in France with the main goal to demonstrate the capability of producing Q = 10. Here, a brief summary of the results that have been shown in Refs [1], [2] is presented. These references study the heating performance of the recent hybrid discharges where the performance of ICRF+NBI heating is assessed together with the fusion enhancement through ICRF heating. This analysis is performed with the ICRF code PION [3] and the NBI deposition code PENCIL [4]. Our modelling takes into account the ICRF+NBI synergy by introducing the computed beam source terms from PENCIL as a source term in the velocity distribution function of PION.

[1‡] See the author list of X. Litaudon et al. Nucl. fusion 57 (2017) 102001.

II. Physics of ICRF heating

Heating the plasma with ICRF waves has shown to be a successful mechanism to bring plasmas at high temperatures. There are several ICRF schemes or approaches to heat the plasma. In this work minority heating is considered. Minority heating consists of introducing a small concentration of resonant ion species that is different than that of the principle ion species, i.e., D in the cases studied here. For good ICRF accessibility and absorption we choose the cyclotron frequency of the minority species that is higher than that of main ion species. The cyclotron frequency is defined as $\omega_\mathrm{c} = \mathrm{qZB}/(\mathrm{Am_p})$, where ω_c is the cyclotron frequency, q is the electron charge, Z the atomic number, A the atomic mass and $\mathrm{m_p}$ the proton mass. Therefore, $\omega_\mathrm{cH} = 2\omega_\mathrm{cD}$, i.e., the fundamental resonance of H coincides with the 2nd D harmonic resonance. The ICRF wave launched from an external antenna have a frequency that matches the cyclotron frequency of H. When this condition occurs, the ions become accelerated by the wave field, which effectively damps the wave. In these plasmas, there are three main damping mechanisms competing for the wave energy: the fundamental H resonance, the 2nd D harmonic resonance and the direct electron damping. The ICRF+NBI synergy is taken into account through the D beams.

III. Experimental results

Several hybrid discharges performed to evaluate the impact of ICRH on the fusion performance were modelled. Here, we focus on the role of H concentration and the record neutron rate obtained in one of the best performing discharges.

A. Hydrogen concentration scan

A set of hybrid discharges was carried out at JET to assess the impact of the H concentration on the ICRF heating and

fusion performance. These discharges were prepared in the same way except for the different H concentration, see Table 1. Note that only a few percent of hydrogen is used and there is only a small difference between the hydrogen concentration between the discharges. Nevertheless, it is large enough to have an impact on the way plasma damps the ICRF wave energy and, consequently, on the plasma performance.

TABLE I. HYDROGEN CONCENTRATION FOR THREE HYBRID DISCHARGES

	92321	92322	92323
$n_H/(n_H+n_D)(\%)$	~2.0	~1.5	~3.0

ICRF power was applied with a central $\omega = \omega_{cH} = 2\omega_{cD}$ resonance. In first order, the ratio of H to D damping scales roughly as $n_H/(n_H+n_D)$, as expected. Typically, for this ICRF scenario, the hydrogen minority is the main absorber at low plasma densities and temperatures that take place during the ramp up. Once the deuterium beams are injected and the plasma gets hotter, damping by resonant D ions becomes the main damping mechanism. Channeling the power to D ions is found to be beneficial to this scheme, as the cross section from DD fusion reactions has a maximum at the MeV range, therefore giving a large margin to increase average D energy and consequently the number of DD reactions. This fact is shown in figure 1, where experimental and modelled discharges show this behavior.

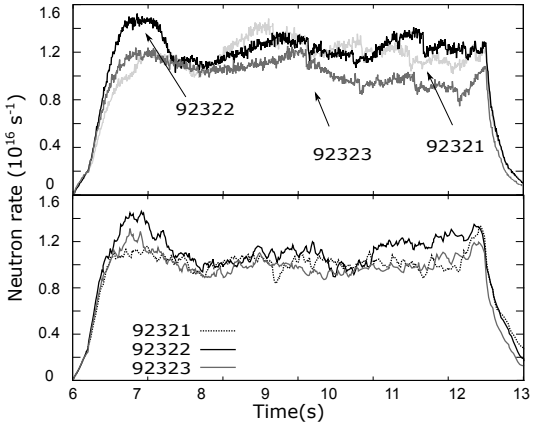


Fig. 1. Comparison of experimental neutron emission rate (top) and modelled neutron emission rate (bottom) of the H scan discharges.

The discharge with the lowest H concentration obtained, in comparison with the highest H concentration discharge, around 10-25% higher number of fusion reactions.

B. High-performing discharge

One of the main goals of the recent experiments with hybrid plasmas was to improve the fusion performance with respect to the previous neutron rate record of $2.3\text{x}10^{16}$ n/s. A total of $2.7\text{x}10^{16}$ n/s was achieved as shown in figure 2.

The modelling of this discharge showed a clear dominance of 2nd D harmonic resonance and low direct electron damping. The ICRF fusion enhancement with respect to NBI was found to be around 15% during the main heating phase.

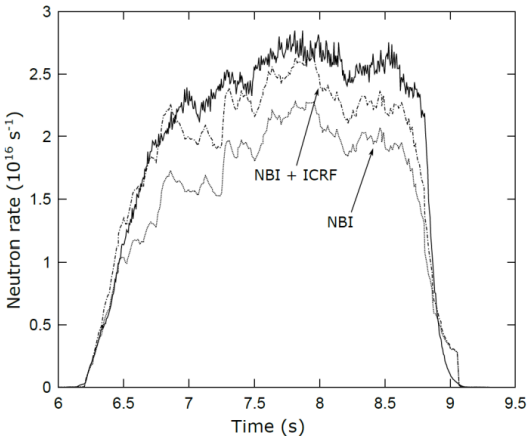


Fig. 2. Neutron production rate modelling of discharge 92398 with ICRF and without ICRF.

C. Conclusions

Many efforts are being devoted to improve the fusion performance of the hybrid and baseline scenarios on JET. Studies of impurity control together with the heating performance as the one presented here are the key to achieve a higher fusion yield. Here, we stated the main idea that channeling the wave power to D ions substantially enhances the fusion performance in this scenario. However this is not necessarily true for the DT scenario, as the cross section peaks at the keV range and, therefore, the way ICRF heats the plasma needs special attention. JET will host in 2020 the second DT campaign of its history, where many of the methods consolidated to improve the DD scenario will be applied to achieve a high performance in DT.

IV. ACKNOWLEDGMENT

This extended abstract is a summary of the work carried out in Refs [1], [2]. The author acknowledges 'la Caixa' for supporting his PhD studies and all the coauthors of Refs [1], [2]. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1] D. Gallart et al., *European Journal of Physics 157 03015*, 2017.
[2] M. Mantsinen et al., *European Journal of Physics 157 03032*, 2017.
[3] L.-G. Eriksson et al., *Nuclear Fusion 33 1037*, 1993.
[4] P. Stubberfield et al., *Multiple Pencil Beam, JET-DPA (06)/87*, 2017.

Dani Gallart studied fundamental physics at Universitat de Barcelona (UB). After obtaining his degree in physics he enrolled in the nuclear engineering MSc at Universitat Politècnica de Catalunya (UPC-ETSEIB) where he obtained one of the grants from Fundació Catalunya-La Pedrera. In 2014, he joined the Fusion group at Barcelona Supercomputing Center (BSC) for his MSc thesis research under the supervision of ICREA Prof. Mervi Mantsinen. In 2015, he was awarded one of the prestigious Fundació 'la Caixa' PhD grants to continue his research at BSC. His research focuses on fast particle physics and plasma heating. He works closely with the main European fusion facilities JET and AUG under EUROfusion.

An introduction to FE2 multi-scale methods and why HPC is so crucial.

Guido Giuntoli*, Mariano Vázquez†, Sergio Oller‡
*†Barcelona Supercomputing Center
‡Universitat Politècnica de Catalunya
*guido.giuntoli@bsc.es, †mariano.vazquez@bsc.es, ‡sergio.oller@upc.edu

Keywords—FE2, HPC, multiscale, composites

I. Methodology

There is no doubt that composite materials are widely used nowadays in almost every engineering areas, mainly because of their excellent properties such as their high resistance, low weight and cost. For this reason it is useful to study their behavior in order to increase their reliability and produce better designs.

There are two main ways of studying composite materials structures. The first and the oldest one is to build a real prototype with the material and to perform experiments, for example, to measure the prototype resistance with a traction test. The second way is to create a computational model applying physical laws and then to predict how it would behave.

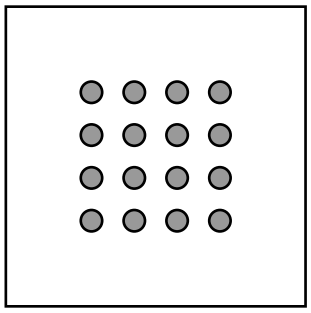


Fig. 1: Representation in two dimensions of a classical composite material structure. In this case the structure is made of two different materials (fibers inside a matrix).

The experimental method can be very accurate because we use directly the real material but it has the counterpart of being expensive because for almost every experiment a new prototype should be manufactured. This is a very slow process specially for those cases where an optimal and reliable design is being searched.

The problem of the cost and speed of the experimental method can be solved with the computational simulations. They involve to develop physical models in order to get accurate solutions, this can be a difficult task depending on the composite material we are dealing with. The most common numerical procedure used in this field is the *finite element method* (FEM) that is combined with constitutive laws to relate the physical variables of the problem. These laws relate the stresses σ and the strains ϵ that are defined at every point in the domain ant they are obtained through experiments.

Composite materials are characterized for being composed of two or more homogeneous materials. In Fig. 1 we represent an example in two dimensions of a typical arrangement of two different materials: a micro structure made of fibers inside a matrix. The distribution and the properties of each material determine the property of the final arrangement. This is a clear example of why computational simulations can allow to find faster an optimal design of a structure because the wide spectra of design options that exist.

In view of what is represented on Fig. 1 it is clear that using the FEM method directly to solve the complete problem can end in a large computational problem that is not feasible to be solved. This is due to the strong difference between the scales that are inside the problem: the structural macroscopic scale and the microscopic scale. A simple explanation of why this happens can be understand by imagining that if at least one finite element is set inside each fiber, then, a large number of finite elements would be needed to discretize the whole problem. Even for a supercomputer the size of this problem can be out of the calculation scale for several orders of magnitude.

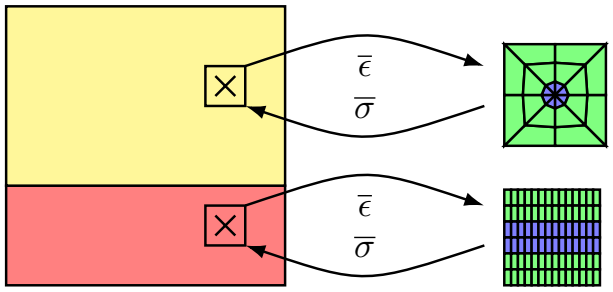


Fig. 2: Representation of the FE2 multi-scale method process. In this example the structure is composed of two different composite materials microstructures (two layers of matrix and fibers with different orientation angles.

To deal with this, the *multi-scale* techniques can be used. The basic idea is to decompose the original problem into two smaller ones: a *macroscopic* and a *microscopic* problem. The macroscopic problem is solved with a coarse FEM mesh and the materials that conforms these elements are homogeneous with constitutive laws obtained using the microscopic model.

For this work we are going to apply the FE2 multi-scale method, here the FEM is applied at both scales which means that in the microscopic scale also the FEM is used to get these properties needed by the macroscopic model. In Fig. 2 we outline this process for a macroscopic structure that is build with two different microscopic structures.

Finally, we should give a briefly idea of the physical models that we are dealing with. For the macroscopic scale we consider the set of equations:

$$\begin{cases} \text{div } \overline{\sigma} = 0 \\ \overline{\sigma} = \langle \sigma \rangle \\ \overline{u} = \overline{u_d} \text{ in } \Gamma_d \\ \overline{\sigma} = \overline{\sigma_n} \text{ in } \Gamma_n \end{cases}$$

where the first is an *equilibrium equation* and the second one is a constitutive law that is obtained using the microscopic model. This is the main difference between the classical single-scale and the multi-scale approach. The others equations are boundary condition equations like in all classical problems.

The macroscopic quantities are calculated with the microscopic model that is defined as:

$$\begin{cases} \text{div}\sigma = 0 \\ \sigma = f(\epsilon) \\ \langle \epsilon \rangle = \overline{\epsilon} \end{cases}$$

here, the first is an equilibrium equation and the second are the constitutive laws of each of the materials that conform the microscopic structure, for this problem these laws are known. The third equation is a supposition and means that the average of the strain field ϵ is equal to the macroscopic strain $\overline{\epsilon}$. This crucial assumption determines which boundary conditions can be imposed in the microscopic model.

The boundary conditions that can be set in the microscopic model are:

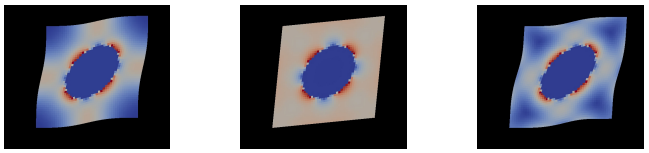
- Periodic
- Uniform strain
- Uniform stress

Each boundary condition produce a different result at the microscopic level, and consequently, at the macroscopic one. Their accuracy of the boundary conditions depends on the problem that is being solved. For example, in the case of aeronautics composite materials, the *periodic* boundary conditions generally give the most accurate results because the microstructure is near to be periodic and the microscopic model is subjected to a similar constrain.

In Fig. 3 we outline the results of a microscopic structure subjected to different boundary conditions.

II. Strategy proposed

The strategy for solving this problem is a distributed memory approach due to the large amount of memory that should be used specially in non-linear problems. In Fig.4 we show the computational scheme that we are going to apply in order to deal with the memory problem. This last consists in applying a domain partition on the macroscopic problem and solve each problem in a different node of a cluster. In this case all of the sub domains are communicated with the others using the MPI protocol, each of them works jointly with a microscopic code that performs the constitutive calculations using the FEM. We plan to add another level of parallelization at the microscopic problem using a *shared memory* memory approach, for example, using *OpenMP* to parallelize the matrix-vector products for solving the linear systems or for performing the assembly of the matrices and RHS.



(a) Periodic. (b) Uniform strain. (c) Uniform stress.

Fig. 3: Results of a microscopic problem subjected to different boundary conditions.

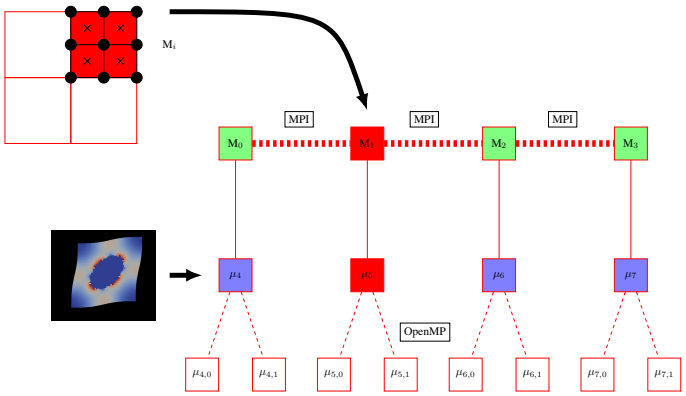


Fig. 4: Distributed strategy that we propose to solve the FE2 problem. In this case the macroscopic problem is divided in four domains and each one is solved in a distributed way among four nodes of a cluster. In each of these nodes an independent microscopic problem is also being solved for retrieving the macroscopic average quantities. The microscopic problem can be parallelized also in a share memory approach considering that each node has more than one CPU.

III. Author biography

Guido Giuntoli: Graduated as nuclear engineer at Balseiro Institute, Argentina in 2015. Working in Argentinian nuclear industry during 2015 and 2016. Currently working as PhD researcher at Barcelona Supercomputing Center and Universitat Politècnica de Catalunya

Effect of population structure, parameter estimation of complex model, and LTBI on TB dynamics

Nura M. R. Ahmad*, Cristina Montañola-Sales†, Daniel López Codinać*
*Department of Physics(BIOCOM), UPC, BarcelonaTech
†Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona
E-mail: {nura.mohammad.rabiu.ahmad, daniel.lopez-codina, cristina.montanola}@upc.edu

Keywords—TB, complex model, parameter estimation, LTBI, age of infection, structure of population, infectious disease.

I. Extended Abstract

The intricate nature of Tuberculosis (TB) infection requires further research to better understand the relationship between the disease mechanisms and the population structure. The influence of the population structure and the role of the infected population on the TB incidence is still not clear. In this study, mathematical modelling techniques are used to elucidate those questions and contribute to understand TB complexity. The work examines the complexity of TB dynamics by using SEI models of different levels of complexity to study the effect of both structure of population and the role of the infected population in TB dynamics. It presents a step by step procedure of how to develop and estimate parameters of complex model for TB transmission. We performed different experiment on more than 20 different countries in order to elucidate if the increase in complexity of the models increases the model accuracy and provides more information about the disease. Our results indicate that parameter estimation could be made easy by the gradual development of simple models. In addition we showed the importance of more complex model over simple model and our result indicate that, unlike simple models, complex model could explain characteristic of the disease such as diagnosis delay time and reinfection. We illustrate how the model without population as a limiting factor dramatic change in behaviour when implemented in high burden incidence. We also demonstrate how the change in age of infection in the latent TB could dramatically alter the dynamics of the disease.

A. Introduction

Tuberculosis (TB) is still a major global health concern and one of the leading causes of death. As reported by WHO, there were an estimated 10.4 million new incident TB cases and 1.7 million deaths worldwide in 2015[1]. Even though most TB cases occur in resource-limited countries, it is still a threat to higher-resource countries. This is due to the nature of the disease's strong interaction with HIV dynamics and also the recently world-wide emergence of drug-resistant TB [2], [3].

The main manifestation and the only infectious form of TB is the pulmonary form, hence worthy of study. Pulmonary TB is an infectious disease caused by *Mycobacterium tuberculosis (Mtb)* and it is transmitted via air borne droplets of the saliva of the sick host. When a sick host sneezes, coughs or talks it can infect susceptible individuals sharing the same environment who inhale the saliva droplets containing the bacterium. The inhalation of the bacilli will usually lead to the initiation of an immune response that can have one of the 3 different clinical outcomes: (1) Complete clearance of the pathogen (2) Latent TB infection (LTBI) or (3) Progression to primary active disease [4], [5]. The objectives of this study are: (i) To understand how the structure of the population that can shape the dynamics of the disease.(i) To show the importance of both complex and simple models depending on each given situation, and to elucidate how parameter estimation can be easily achieved when developing gradually more complex models.(iii) To show the role and importance of the latent infected population and the age of infection in understanding the dynamics of TB. The epidemiological evolution of 20 different countries will be analyzed to exemplify the importance of each type of model to achieve the aims set above.

B. Methodology

TB dynamics presents several characteristics that greatly contribute to its complexity. Compartmental models facilitate obtaining a good understanding of these complexity. Fig. 1 proposed several compartmental models that were used to described the dynamics of TB with different complexity level. In each of the model, starting with the most simple form SEI, population were divided into different compartment, namely Susceptible S, latent infected (exposed) E, Sick (infectious) I, and Reinfection R. We formulated five different models (SEI, SEI2, SE8I, SE8I2, SE8IR), each model was formulated with two different force of infection $\Lambda = \frac{\alpha IS}{N}$ and $\Lambda^* = \alpha I$ to allow the evaluation of effect of population on the dynamics of the disease. Model equation for each model were formulated, and the spectral radius was analyzes.

C. Results

TB epidemiological data from countries all around the world was analyzed and 20 (10 LBC and 10 HBC) different countries were selected to test and validate the models.

Comparison of models performance between low and high incidence burden setting to illustrate the effect of population on Tb dynamics: To address the long over due question of how the population can affect the dynamics of TB, we adhere to our strategy by fitting both set of models with population as limiting factor and without to both LBC and HBC. Fig. 2 illustrate the difference between the two sets of models when implemented in low burden and high burden countries.

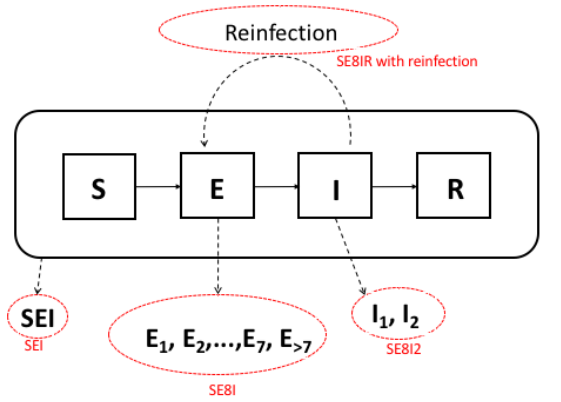


Fig. 1. **Overview of several proposed compartment models with different level of complexity.** SEI is the classical model of three compartments used to understand the epidemiological dynamics of TB in a given population. SEI2 applies to a general context of TB dynamics and has four compartments (Susceptible S; Exposed E; sick and infectious I_1; and sick but not infectious I_2). SE8I refers to a context where the time scale and age of infection in the latently infected population is considered due to its importance in understanding TB dynamics. In SE8I2 two main important features were also added, the diagnosis time delay and the probability of relapse. The sick population are dived into two sub-population, I_1 sick and infectious (thus, spreading the disease) and I_2 sick but under treatment. Finally, in contexts of high incidence, we introduce the concept of people reinfection so $E_i > 1 \rightarrow E_1$

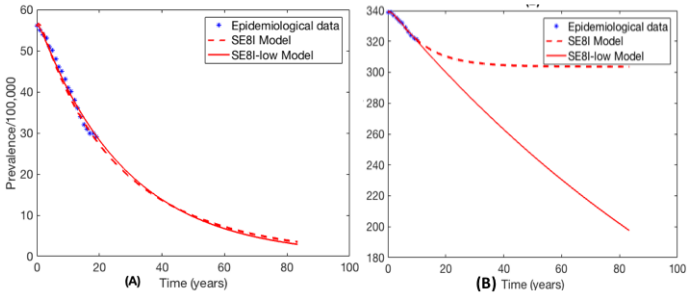


Fig. 2. **The effect of population structure on the dynamics of TB demonstrated in both low and high burden countries.** (A): The model simulation with population as a limiting factor (SE8I model) and model without population as a limiting factor (SE8I low model) in a low burden country. (B): The model with population as a limiting factor (SE8I model) and model without population as a limiting factor (SE8I low model) in a high burden country.

Model complexity and parameter derivation: Developing several models of increasing complexity in a gradual manner allowed the parameters to be derived from the simpler models. Fig 3 shows the simulation results for Argentina with different models of various complexity levels and contrast it with the epidemiological data during 20 years.

D. Conclusion

Our experiment provides a significant evidence that the structure of population plays a vital role in shaping the dynamics of the TB. Although simple models could describe the dynamics of the disease, we show that it is necessary to design more complex model in order to understand some of the more complex structure of the TB dynamics. We also showed that simple models can provide a significant aid in the estimation of

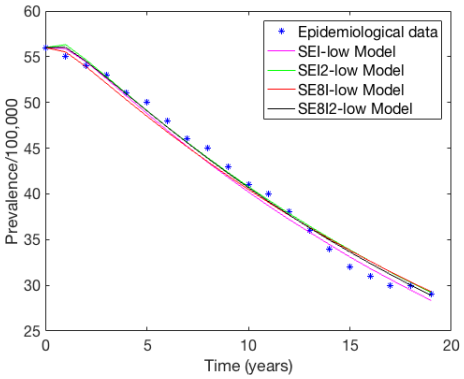


Fig. 3. **Comparison of low incidence models fittings for epidemiological data in Argentina.**

the parameter for more complex models. We finally conclude that the age of the infection and the structure of the latently infected population must to be taken into consideration while designing any TB intervention program

References

[1] W. H. Organization and Others, "Media Centre: Tuberculosis, Fact sheet," 2016.

[2] M. R. Nyendak *et al.*, "Mycobacterium tuberculosis Specific CD8+ T Cells Rapidly Decline with Antituberculosis Treatment," *PLOS ONE*, vol. 8, no. 12, pp. 1–10, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0081564

[3] J.-P. Millet *et al.*, "Predictors of Death among Patients Who Completed Tuberculosis Treatment: A Population-Based Cohort Study," *PLOS ONE*, vol. 6, no. 9, pp. 1–8, 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0025315

[4] J. Davis *et al.*, "Real-Time Visualization of Mycobacterium-Macrophage Interactions Leading to Initiation of Granuloma Formation in Zebrafish Embryos," *Immunity*, vol. 17, no. 6, pp. 693–702, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1074761302004752

[5] K. Bhatt and P. Salgame, "Host Innate Immune Response to Mycobacterium tuberculosis," *Journal of Clinical Immunology*, vol. 27, no. 4, pp. 347–362, jul 2007. [Online]. Available: https://doi.org/10.1007/s10875-007-9084-0

Nura M.R. Ahmad was born in Kano, Nigeria. He receives B Sc. degree in Mathematics from Bayero University in 2009, M Sc in Mathematics in 2015 from the same university. He joined the department of physics Universitat Politècnica de Catalunya (UPC), Spain, as a Ph.D. student in Sep. 2016, and his research interest is Complex system modelling.

Earthquake simulation by Fiber Bundle Model and Machine Learning techniques

Marisol Monterrubio-Velasco*, José Carlos Carrasco-Jiménez*, Armando Aguilar-Meléndez*, Octavio Castillo*, Josep De la Puente*
* Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {marisol.monterrubio, jose.carrasco, armando.aguilar, octavio.castillo, josep.delapuente}@bsc.es

Key words – Seismic rupture processes, Fiber Bundle model, Machine Learning.

I. EXTENDED ABSTRACT

The rupture processes of any heterogeneous material constitute a complex physical problem. Earthquakes are the result of rupture in the Earth's crust. This process is difficult to model deterministically due to the number of parameters and physical conditions, which are largely unknown. Computational physics offers alternative ways to study the rupture process in the Earth's crust by generating synthetic seismic data using physical and statistical models. The Fiber Bundle model (FBM), describes the complex rupture processes in heterogeneous materials in a wide range of phenomena. It has shown the capacity to generate data that depicts the main statistical characteristics of real seismicity. High-performance computing (HPC) combined with Machine Learning (ML) techniques provide a good ground base to perform and improve the simulations, the data management process and data analysis. In this study we show the FBM model versions applied to simulate two stages in the seismic cycle: the rupture stage related with the so-called mainshock and the stress relaxation stage which produces the aftershocks.

A. Introduction

Although we have knowledge about the occurrence of certain major earthquakes, our observational span is still too short to be able to draw strong (predictive) conclusions about when, where and how big the next earthquake will be. Earthquakes can be studied from either a physical or a statistical point of view. The statistical approach considers earthquakes as stochastic point processes [1–3]. Seismic catalogs are the tool that allows study the statistical characteristics of earthquakes. In these catalogs is register the seismic activity occurred in a particular time. It contains at least the information of: the epicentral and hipocentral coordinates, the originate time and the magnitude.

The Fiber Bundle Model (FBM) [4], is an agent-based model that describes the interactions of individual cells, featuring particular transfer load rules and a probability distribution function describing the intrinsic cells properties. Its simplicity offers many advantages and an adaptability to describe a wide range of phenomena. The objective is to study the rupture seismic phenomena via numerical simulations using the FBM. These simulations allow the analysis of natural systems that can not be studied in a laboratory due to their large energetic scale and complexity. Through numerical simulation, we look for example the most important mechanisms related to the seismic migration after a mainshock. In this work our objective is shown the application of the FBM to simulate seismic rupture scenarios, in particular the asperity ruptures and the aftershocks.

The FBM model has been capable to generated seismic synthetic catalogs which reproduces many statistical features of real events [5-7]. At present we developed two different model versions to simulate:

1) the rupture stage (10^0 s-10^2 s) of the faults, due to the excess in the accumulated strain. This stage culminates in the so-called mainshock.

2) The stress relaxation stage (10^1-10^3 days) in the area affected by the mainshock. This stage produces the so-called aftershocks, and it culminates when the background seismic rate is achieved [8-9].

These two stages are those that have a higher risk in our society and can cause major disasters when they occur in highly populated areas.

In our work Machine Learning ML techniques are used to classify and cluster data via training of models of data series which help to find the characteristics that generate patterns, thus making predictions on data [10]. For example, the favorable mechanical properties to produce aftershocks, the geometry pattern and/or number of fractures, among others.

B. Background: FBM applied to earthquake simulation

At the present the development of the FBM versions to simulates seismic scenarios has been divided in two main modules. In chronological order they are:

1) The aftershocks simulator in which the main statistical patterns in time, magnitude and space domains have been studied via parametric and statistical analysis [6]

2) the mainshock simulator developed to study the dynamical rupture and the magnitude's behavior [7]. With this model we use few input parameters coming from observational source inversion, to model dynamic characteristics.

Both versions modelize from simple assumptions and self-evolve in complex patterns similar to the real. Our goal has been the parametric study in order to learn how model behaves and which are the most important variables.

C. Methdology

The two major challenges of this research work are: the generation of a large number of simulations required to gain statistically accurate results and the optimization of simulation parameters. In order to deal with the first major challenge, we use High Performance Computing (HPC) to produce a large number of simulations in a reduced period of time. To deal with HPC we require to adapted the algorithm to be executed in a distributed environment.

Furthermore, to solve the second major challenge, we exploit state-of-the-art ML techniques to analyze and extract patterns from data generated by the simulations to estimate optimal parameters that approximate synthetic events that are similar to real seismic events.

To analyze the synthetic seismic series generated with the FBM we applied supervised learning methods in order to:
1) determine the minimum grid size using machine learning methods such as Random Forest, Flexible Discriminant Analysis, and Support Vector Machines [10].
2) identify the most relevant input parameters of the simulation (i.e. percentage of conserve load, initial organization probability) using statistical analysis and machine learning methods.
Once optimal input parameters are estimated, we validate the model using real seismic sequences and the corresponding spatial distribution of their faults systems.

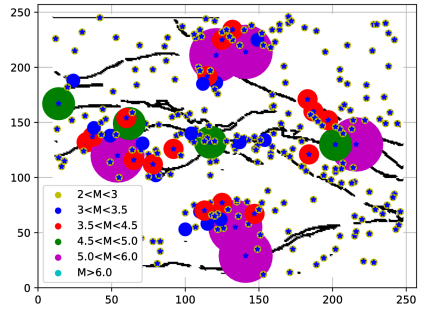


Fig. 1 Example of a simulated aftershocks sequences in space and magnitude domain

D. Conclusion

The novelty of this work lies in the study of the Earth dynamical rupture processes by using an agent-based model which describes the general rupture of heterogeneous materials. This stochastic model requires a large amount of numerical experiments to reduce uncertainties and consequently give robustness to the model. We exploit HPC to increase the number of numerical experiments. Also we introduce ML techniques to better estimates model parameters that better approximates to describes real Earthquakes statistical characteristics.

II. Acknowledgment

Authors thanks to Mexican National Council for Science and Technology (CONACYT) and Barcelona Supercomputing Center (BSC).

References

[1] B. Gutenberg and C.F. Richter, “Earthquake magnitude, intensity, energy and acceleration”. *Bull. Seis. Soc. Am.*, vol.46(2), pp. 105–145,1956.
[2] T. Utsu, Y. Ogata and R.S. Matsu’ura, “The centenary of the Omori formula for a decay law of aftershock activity”. *J. Phys. Earth*, vol. 43(1), pp. 1–33, 1995.
[3] G. King, S. Ross and J. Lin. “Static stress changes and the triggering of earthquakes”. *Bull. Seis. Soc. Am*., vol.84(3), pp. 935–953, 1994.
[4] F. Peirce, “The Weakest Link. Theorems on the Strength of Long and of Composite Specimens”. *J. Textile Inst. Trans*, vol.17(7), pp. T355.
[5] M. Monterrubio-Velasco, X. Lana and M.D. Martínez, “Aftershock sequences of three seismic crises at Southern California simulated by a Cellular automata model based on self-organized criticality”. *Geosciences Journal*, vol.19, pp. 81 – 95, 2014.
[6] M. Monterrubio-Velasco, F.R. Zúñiga, V.H. Márquez-Ramírez and A. Figueroa-Soto, “Simulation of spatial and temporal properties of aftershocks by means of the fiber bundle model”. *Journal of Seismology*, vol.21, pp. 1623 – 1639, 2017.
[7] M. Monterrubio-Velasco, Q. Rodríguiez-Pérez and F.R. Zúñiga, “Simulates Asperities by means of the Fiber Bundle model”. (In review process *Geophysical Journal International* GJI-18-0140).
[8] R. Shcherbakov, D.L. Turcotte and J.B. Rundle, “Aftershock statistics”. *Pure and Applied Geophysics.* vol.162(6), pp. 1051–1076, 2005
[9] A. Fereidoni and G.M. Atkinson, “Aftershock statistics for earthquakes in the St. Lawrence Valley”. *Seismological Research Letters*, vol. 85(5), pp.1125-1136, 2014.
[10] R. Kohavi and F. Provost, “Guest editors' introduction: On applied research in machine learning”. *Machine Learning*, vol.30, pp.271-274, 1998.

Marisol Monterrubio-Velasco received her PhD in computational physics by University Politécnica de Catalunya. She was postdoctoral researcher in the Geosciences center at Universidad Nacional Autónoma de México. At present she is postdoctoral researcher in the Barcelona Supercomputing Center thanks to Mexican National Council for Science and Technology. Her research interests focus on computational physics, statistical analysis and numerical simulation applied to the Earthquakes.

Sampling Interfacial Water Effects over Protein Specificity with PELE

Martí Municoy*, Oliver Carrillo*, Victor Guallar*†
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
E-mail: {marti.municoy, oliver.carrillo, victor.guallar}@bsc.es

Keywords—Drug Design, PELE, Protein Engineering, Water Sampling

I. Extended Abstract

Water is one of the main contributors that determine the shape of a protein, thereby defining its function. It can also be found inside protein cavities, helping proteins to unfold specific interactions with other substrates. Inner water molecules are able to modify the binding mode of the ligand by giving rise to new hydrogen bonds, by changing the polarity of their surroundings or by simply filling empty spaces. Thus, it is essential to take into account mediating water molecules when studying protein-ligand bindings.[1]

A protein environment has regions that are essentially hydrophobic. In this way, inner water molecules need to be highly arranged, trying to look for favorable polar interactions. This arrangement confers on them different properties from those water molecules that are in the bulk solvent. Water trapping has thermodynamic consequences. Enthalpy may be favorable as new hydrogen bonds can be established. But the main drawback is the large loose of entropy that occurs when a free water molecule from the solvent ends up trapped inside the protein.[2]

Therefore, systems like neuraminidase proved to contain water molecules that are responsible for increasing the protein-ligand binding by means of hydrogen bonds.[3] Yet, other systems such as HIV-1 protease show bindings that are favorable thanks to the gain of entropy that comes from the displacement of a trapped water to the bulk solvent.[4] Since thermodynamic properties are used to describe the affinity of a drug towards a target, drug design software needs to take into account the effects of water molecules in binding sites.[5]

In this work, we introduce the advantages of performing protein-ligand sampling by including sampling of mediating water molecules as well. Our approach uses the Protein Energy Landscape Exploration (PELE) program which is a tool that does protein-ligand sampling by means of a Monte Carlo method.[6] We aim to add a new routine to PELE to carry out sampling of interfacial waters from the binding site of a protein. Currently, PELE works with an implicit solvent, hence, interfacial water molecules are ignored by default. This tool should perform protein-ligand sampling while interfacial mediating waters are perturbed according to Monte Carlo method.

A. PELE methodology

PELE offers a methodology to perform protein-ligand sampling with a significantly reduced computational cost than that of conventional molecular dynamics simulations. It relies on a sampling procedure made up of three main steps. Firstly, the current state of the protein system is perturbed by translating and rotating the ligand and applying a perturbation on the protein according to the main vectors calculated with the Anisotropic Network Model (ANM). As a second step, PELE attempts to relax the system by applying a side chain prediction and a global minimization. Finally, the last step stands for either accepting or rejecting the perturbation according to the Metropolis criterion.

During the last years, PELE has been applied to perform different studies. From mapping ligand migration pathways to studying the substrate recognition of enzymes. In many cases, PELE has proved to be an outstanding tool to study protein-drug interactions.[7]

B. Water mediation in neuraminidase

An initial test has been conducted to see how the current version of PELE explores water-dependent systems. The system that is chosen is the native influenza virus neuraminidase bound to an inhibitor, a sialic acid. The X-ray structure of the complex was determined at 1.8 resolution (PDB: 1F8B).[3] In this complex, two water molecules seem to play a crucial role in the interaction between the protein and the ligand as they bridge them by means of hydrogen bonds.

Subsequent theoretical studies on the same crystallographic structure could classify the water molecules of the binding site.[8] They concluded that one water molecule (Wat A) seems to unfold very strong electrostatic interactions, while the other one (Wat B) presents an unfavorable binding free energy. This could be the main reason why Wat B exhibits displacement with the entrance of the ligand in some influenza neuraminidase complexes, while Wat A is conserved in all of them.

The role of Wat A on the protein-ligand binding was analyzed with PELE. Two local explorations were performed taking the previous crystallographic structure as starting point. The first exploration included both water molecules, Wat A and Wat B. Then, in the second exploration, Wat A was removed from the binding site. Other eight water molecules from the binding site were included in both explorations. All of them were treated by PELE as part of the protein chain but they

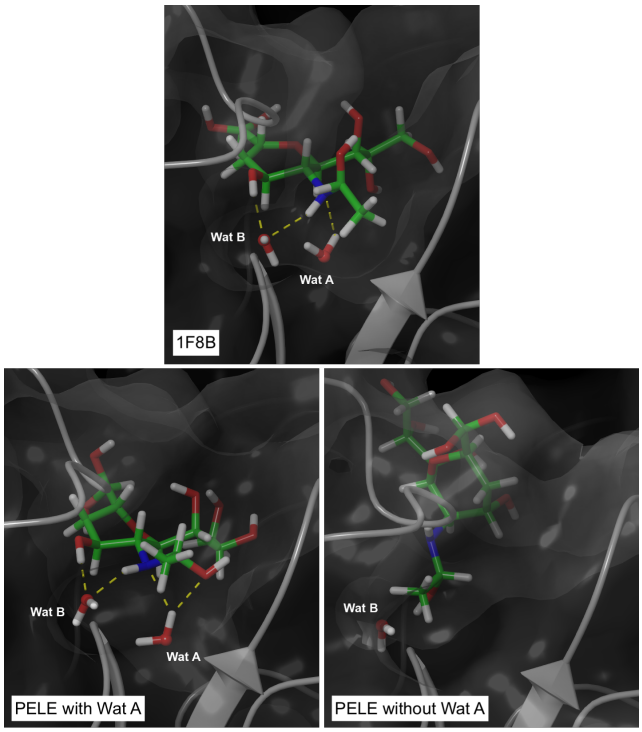


Fig. 1. Structural comparison between the crystallographic structure and the structures with the highest binding affinity from PELE explorations.

TABLE I. ENERGETIC AND STRUCTURAL ANALYSIS OF SIALIC ACID STRUCTURES.

Wat A	Binding Energy	SASA	RMSD
Yes	−74.9072	0.0422	1.6639
No	−50.6205	0.1662	5.9820

were omitted in the ANM perturbation. Figure 1 shows the structures which presented the highest protein-ligand binding affinity for each PELE exploration. Table I compares these two structures with the crystallographic system.

With the system that contains Wat A, PELE was able to predict a structure very close to the reference; the RMSD comparison with reference has a value of 1.6639. However, we appreciate significant structural changes after removing Wat A. The resulting structure is no longer similar to the crystallographic structure. In this case, the RMSD comparison increases up to 5.9820.

The absence of Wat A makes the sialic acid to be more exposed to the solvent as it tries to stabilize its polar chains. Wat A seems to have a key role on stabilizing the amine of the sialic acid. Thus, with the presence of Wat A it does not need to expose itself to the solvent to gain stability.

These different structural arrangements entail a change on the protein-ligand binding energy. As a result of this, the binding affinity of the sialic acid is significantly decreased when Wat A is missing in the binding site.

C. Conclusions

When all mediating water molecules are placed correctly in the binding site, PELE is able to find a binding site for sialic acid which strongly matches with the reference complex. However, when Wat A is missing or slightly shifted from its original place PELE explorations point to other structures that do not match with the reference. This is due to the fact that PELE does not contain a method to sample the waters of the binding site. Then, the role of Wat A is ignored when it is not placed correctly.

We believe that if we include a Monte Carlo method to sample water molecules along a PELE run, we would not need to previously place mediating water molecules to a suitable location. Moreover, PELE could find new poses of the ligand which require water molecules from the binding site to be shifted. For instance, we could deal with systems like HIV-1 protease where there are water molecules which need to be shifted out when the ligand binds to the protein.

II. ACKNOWLEDGMENT

The authors would like to thank the cooperation in the project to all the members of Electronic and Atomic Protein Modeling group in the Barcelona Supercomputing Center and the assistance provided by some members of Nostrum Biodiscovery.

REFERENCES

[1] F. A. Quiocho, D. K. Wilson, and N. K. Vyas, "Substrate specificity and affinity of a protein modulated by bound water molecules," *Nature*, vol. 340, p. 732, aug 1989.

[2] J. E. Ladbury, "Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design," *Chemistry & Biology*, vol. 3, no. 12, pp. 973–980, dec 1996.

[3] B. J. Smith, P. M. Colman, M. V. Itzstein, B. Danylec, and J. N. Varghese, "Analysis of inhibitor binding in influenza virus neuraminidase," *Protein Science*, vol. 10, no. 4, pp. 689–696.

[4] P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, and al. Et, "Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors," *Science*, vol. 263, no. 5145, pp. 380 LP – 384, jan 1994.

[5] D. Bucher, P. Stouten, and N. Triballeau, "Shedding Light on Important Waters for Drug Design: Simulations versus Grid-Based Methods," *Journal of Chemical Information and Modeling*, vol. 0, no. 0, p. null.

[6] K. W. Borrelli, A. Vitalis, R. Alcantara, and V. Guallar, "PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique," *Journal of Chemical Theory and Computation*, vol. 1, no. 6, pp. 1304–1311, 2005.

[7] S. Acebes, E. Fernandez-Fueyo, E. Monza, M. F. Lucas, D. Almendral, F. J. Ruiz-Dueñas, H. Lund, A. T. Martinez, and V. Guallar, "Rational Enzyme Engineering Through Biophysical and Biochemical Modeling," *ACS Catalysis*, vol. 6, no. 3, pp. 1624–1629, 2016.

[8] C. Barillari, J. Taylor, R. Viner, and J. W. Essex, "Classification of Water Molecules in Protein Binding Sites," *Journal of the American Chemical Society*, vol. 129, no. 9, pp. 2577–2587, 2007.

Martí Municoy has a double Bachelor's Degree in Physics and Chemistry in the Autonomous University of Barcelona (UAB). He is currently enrolled at the Master of Modelling for Science and Engineering in the Autonomous University of Barcelona (UAB). He is also an internship student from the Electronic and Atomic Protein Modelling (EAPM) group in Barcelona Supercomputing Center (BSC), where he is developing his Master Thesis.

Inception: We need to go wider

Rajiv Nishtala
Barcelona Supercomputing Center
rajiv.nishtala@bsc.es

Paul Carpenter
Barcelona Supercomputing Center
paul.carpenter@bsc.es

Xavier Martorell
Universitat Politécnica de Catalunya &
Barcelona Supercomputing Center
xavier.martorell@bsc.es

I. Introduction

Modern HPC systems are typically built with multiple racks of several multi-core chips put together as a single system. Each such chip has a local DRAM, and they are collectively called as a node. Each node is connected using a high-speed interconnect. This enables the programmer the benefit of transparently issuing a memory request either to local or remote at relative costs.

However, traditional HPC workloads have been shown to have a large variation in their memory footprint, depending on the application domain, the number of processes and whether it is strong or weak scaling. In such circumstances, the programmer is bound to use either the large memory nodes available on commonly deployed HPC systems or suffer large latency delays in disk accesses. For instance, MareNostrum-4 has 128 large memory nodes with 128 GB, as opposed to the typical 32 GB nodes.

On this end, we aim to (1) develop a model to compute the performance overhead of NUMA access latency and application load balancing; (2) reduce the need for large memory nodes by providing HPC resource manager (for example, SLURM) support for memory capacity sharing among nodes using the UNIMEM architecture.

The rest of the paper is organized as follows: Section II provides a background on the related programming models and the UNIMEM architecture. Section III introduces the methodology for validating Inception. Finally, Section IV validates Inceptionusing two different simulators: TaskSim and ZSim .

II. Background

[OmpSs programming model]: The OmpSs programming model is a task-based programming model that provides an abstraction to the implementation of parallel applications. In OmpSs , the task construct enables the annotation of function declaration with the task directive. Every invocation of this function generates a task that is executed concurrently with other tasks or parallel loops. The OmpSs environment is built on top of the (a) Mercurium compiler and (b) the Nanos++ environment which serves as a runtime platform. Mercurium is a source-to-source compiler to translate OmpSs annotation clauses to source code. The Nanos++ is responsible for the internal creating and execution of tasks.

[UNIMEM]: State-of-the-art HPC infrastructures have computing cores in the order of millions, if not billions, working and communicating in tandem to solve a problem. There are two sources for this problem: (1) Big-data applications require large and fast memory subsystems for computations, else suffering from a high disk access latency. (2) These chips spend a lot of energy communicating amongst each other, rather than the actual computation. A large portion of the energy is spent on transferring the communicated data from the remote node buffer to the local nodes' buffer. This, in large, translates to magnitudes of wasted energy and computational overhead.

In contrast to such architectures, the Unified Memory (*UNIMEM*) architecture, offers the ability to access areas of memory located in the remote nodes at a relatively "low" latency and communication cost. UNIMEM achieves this by communicating using the Remote Direct Memory Access (RDMA) operation through its Global Address Space, which delivers data in-place and avoid receiver-side copying. The "low" latency and communication cost is achieved using the Input/Output Memory Management Unit (IOMMU) and DMA Engine Virtualization, which allows user-level initialization of RDMA operations. This allows the UNIMEM architecture to facilitate sharing of large memory nodes.

III. Methodology

[Benchmarks]: We use four scientific workloads implemented in the OmpSs programming model: Blackscholes, dedup, freqmine and fluidanimate. These benchmarks are regularly executed on hundreds of thousands of processing cores. These benchmarks are available as a part of the PARSECSs benchmark suite [5]. For all PARSECSs benchmarks, we use medium and native input datasets [1].

[Hardware resource]: We perform evaluation on two simulators: ZSim and TaskSim . The parameters used in TaskSim and ZSim are presented in Table I.

Nord-III. We collect the traces for the applications running on ZSim on the Nord III cluster [3]. Nord-III contains 84 compute nodes. Each node contains two Intel SandyBridge-EP E5-2670 sockets that comprise of eight cores operating at 2.6 GHz. Hyperthreading was disabled as in most HPC systems. SandyBridge processors are connected to main memory through four channel and each channel is connected to a single 4 GB DDR3-1600 DIMM. Applications running on Nord III were compiled using gcc version 6.2.0, ompss, nanos-0.15a and mercurium-2.1.0.

ZSim. We simulate the multi-core processor and a DRAM using ZSim [13] and DRAMSim2 [12], respectively. ZSim deploys three techniques to achieve accuracy, speed, and scalability. ZSim ensures the accurate x86 code instrumentation

	TaskSim/MUSA	ZSim
Platform	MareNostrum-4	Nord-III
System Configuration		
Core Frequency	3.0 GHz	3.0 GHz
Number of Cores	8	8
Core Model	4-issue, out-of-order	4-issue, out-of-order
Architecture	Simpled CPU model	Intel® Sandy Bridge
Memory Subsystem		
Cacheline Size	64	64
Private L1 I Cache	32 kB 8-way set associative	32 kB 8-way set associative
Private L1 D Cache	32 kB 8-way set associative	32 kB 8-way set associative
Private L2 Cache	256 kB 8-way set associative	256 kB 8-way set associative
Shared L3 Cache	20 MB 20-way set associative	20 MB 20-way set associative
DRAM		
Simulator	Ramulator	DRAMSim2
Standard	DDR3-1600	DDR3-1600
Capacity	8 GB	8 GB
Organization	2 ranks, 8 banks, DDR3 512 MB	2 ranks, 8 banks, DDR3 512 MB
Instrumentation Tool	DynamoRio	Pin

TABLE I: System parameters for TaskSim and ZSim

through dynamic binary translation tool called pin [10]. It speeds up simulation by categorizing memory requests into two-phases: bound and weave phase. Furthermore, it scales well using a user-level OS virtualization layer. The integration of ZSim and DRAMSim2 enables a cycle-accurate simulation for memory requests by creating precise timing events for the weave phase of the ZSim simulator. The simulated multi-core processor is similar to SandyBridge architecture [14].

MareNostrum-4. We collect the traces for the applications running on TaskSim on the MareNostrum-4 supercomputer [2]. MareNostrum-4 contains 3456 compute nodes. Each node contains two Intel Xeon Platinum 8160 sockets that comprise of 48 cores operating at 2.10 GHz. Hyperthreading was disabled as in most HPC systems. Xeon Platnium processors are connected to main memory through six channel and each channel is connected to a single 48 GB DDR4-2666 DIMM. Applications running on MareNostrum-4 were compiled using gcc version 7.1.0, ompss, nanos-0.11a and mercurium-2.1.0.

TaskSim. We simulate the multi-core processor and a DRAM using TaskSim simulator [11] and Ramulator [9]. The simulated multi-core processor is similar to simple CPU model that issues and commits instructions faster. The TaskSim infrastructure uses Nanos++ scheduling delays that happen at the rate of 1 CPU run, and add delays on thread migration and therefore, new task assignment might behave better/worse in some extreme cases. Additionally, TaskSim does not implement a cache coherence protocol, and thereby does not conflict between CPUs. The instrumentation tool used for TaskSim is DynamoRio [7].

[Why Simulator?]: State-of-the-art architectures like Cavium ThunderX [4] provide a dual-socket configuration with large shared memory and high memory bandwidth. In a shared memory system, all processes share a global memory and each processor accesses memory through a shared bus and have a local cache. There is a fixed latency for a memory requests from either socket. The two main concerns with a shared memory system are: contention and coherence. The performance degradation when multiple processors are trying to access the shared memory simultaneously results in contention for memory bandwidth; whereas, having stale data across different cache might result lead to a coherence problem.

Current architectures do not natively provide the possibility to modify the memory access latency or bandwidth from different sockets - as in UNIMEM architectures - and therefore it is hard to emulate multiple memory islands at different latencies.[1]

We emulate the UNIMEM architecture using the aforementioned simulators. These simulators define the core, caches, DRAM, etc., as modules. Every module is connected to one another using "ports". The simulator parameters are configured in the simulators' configuration file. We instantiate a fixed number of DRAM modules at runtime, which are connected to the memory controller. A read or write request from the memory controller is directed to a specified DRAM, at fixed latency, based on the first touch policy [8], which is the default policy in Linux. In the first touch policy, memory is allocated to the same node as the thread that accesses the memory page - this allows to maximize local accesses over remote accesses. However, this is not guaranteed because the data can be shared by threads on multiple nodes.

IV. Evaluation

The aim of this work is two-fold (a) Cross-validating the results obtained from TaskSim (a trace based simulator) and ZSim (an execution driven simulator) with a real-machine. (b) Introduce a simulated contention to the local DRAM from an application from a remote node.

At the time of writing this paper, we are generating the results required.

V. Author biography

Rajiv Nishtala is a Post-doc at the Barcelona Supercomputing Center. His research interests include dynamic resource allocations, energy efficient computing and thread scheduling. For more details: nishtala.github.io

References

[1] C. Bienia and et al. The PARSEC benchmark suite. In *PACT 2008*.
[2] BSC, MareNostrum IV System Architecture.
[3] BSC, Nord III System Architecture.
[4] Cavium. Cavium ThunderX ARM Processor, 2018.
[5] D. Chasapis and et al. PARSECSs. *ACM TACO*, 2015.
[6] H. David and et al. Memory power management via dynamic voltage/frequency scaling. ICAC '11.
[7] DynamoRio: Dynamic Instrumentation tool Platform.
[8] F. Gaud and et al. Challenges of memory management on modern NUMA systems. *Communications of the ACM*, 2015.
[9] Y. Kim and et al. Ramulator: A fast and extensible dram simulator. *IEEE CAL*, Jan 2016.
[10] S. Naftaly. Pin: A Dynamic Binary Instrumentation Tool.
[11] A. Rico and et al. On the simulation of large-scale architectures using multiple application abstraction levels. *ACM TACO*, 8(4):1–20, 1 2012.
[12] P. Rosenfeld and et al. DRAMSim2: A Cycle Accurate Memory System Simulator. *IEEE CAL*, 2011.
[13] D. Sanchez and et al. ZSim. ACM Press, 2013.
[14] R. S. Verdejo and et al. Microbenchmarks for Detailed Validation and Tuning of Hardware Simulators. In *HPCS 2017*, 2017.

[1] Natively implies without any modifications to the RAS-to-CAS delay and back-to-back CAS delay [6]

Robust Point-Location Method for Linear and High Order Meshes. Application to Particle Transport

E. Olivaresn*, R. Borrell*, G. Houzeauxć*, B. Eguzkitzać*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {edgar.olivares}@bsc.es

Keywords—Point-location problem, Inclusion test, Ray casting, Particle transport, High order elements, Finite elements method, High performance computing

I. Extended Abstract

II. Introduction

In computational geometry the point-location problem is a fundamental topic. In the Finite Elements Method (FEM) context, it is used to find which is the host element of a given point in the computational domain. This process is required in many application such as the measurement of flow properties on specific points (probes) in computational fluid dynamics (CFD), the projection from one mesh to an- other in adaptivity or for Lagrangian particle transport simulations.

III. Point-location algorithm

In this work, an efficient solution to the point-location problem applied to FEM is presented. The robustness of the proposed approach is evaluated in the context of particle transport simulations in the respiratory system airways. Respiratory system simulations involve CFD and millions of transported particles along with tens or hundreds of thousands of time-steps [1]. As a consequence, the location process is one of the critical parts of the simulation. In other words, an efficient and robust inclusion test becomes essential not only for the accuracy of the results, but also to achieve a good computational efficiency.

Our algorithm is composed of four main steps: three consecutive filters are applied, followed by the evaluation of the iso-parametric coordinates of the point within the hosting element.

- *Filter 1: Bin/Oct tree.* A list of host element candidates is created using a bin/oct tree strategy [2] in the initial injection. After this, the list of candidates is only formed by direct neighbours of the previous host element.
- *Filter 2: Bounding box of the element.* The list of candidates is looped. The candidate elements which do not contain the target point within its bounding box are discarded.
- *Filter 3: The inclusion test.* An inclusion test method based on ray casting [3] is used to check if the candidate element is the host element between the remaining candidates. This method is based on counting ray intersections with edges or nodes and apply the odd/even parity rule (if the number of intersections is odd, the point belongs to the element). Care must be taken to avoid a ray intersections an edge of the element because invalidates the parity rule.
- *Calculation: Iso-parametric coordinates.* Once the host element is known, the iso-parametric coordinates of the point inside the element can be calculated. To solve the shape function equation which allow to transform global coordinates into local ones, a Newton-Raphson iterator is used.

In figure 1, a flowchart is shown outlining the aforesaid steps.

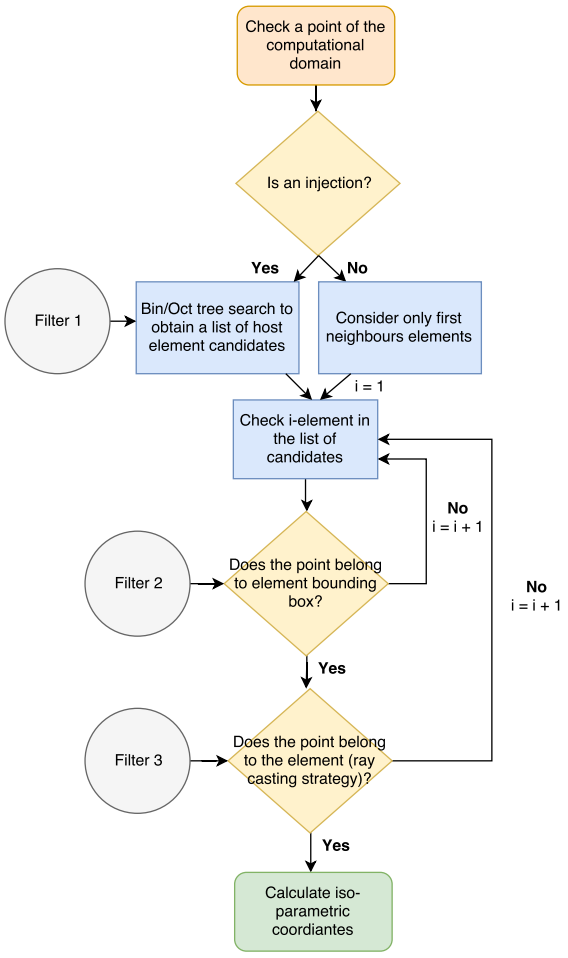


Fig. 1. Flowchart of the point-location process.

IV. HIHG ORDER MESHES

In the particular case of high order meshes, the ray casting strategy has been modified because of the peculiarites of intersection calculus for high order element, which analytic solution may become too complicated or not exist. For this reason, a root finding algorithm comes as the most obvious solution. Thus, a second Newton-Rapshon iterative approach is included.

V. CONCLUSIONS AND PERSPECTIVES

In this work, special emphasis is put on two aspects. First, on the numerical and implementation details that ensure the robustness and efficiency. Second, on the peculiarities of intersection calculus for high order elements, showing a new method based on a modified ray casting strategy which has been proved to properly work with first order meshes.

Both points together allow us to introduce a general solution to the point-location problem, suitable to large and high order meshes.

REFERENCES

[1] G. Houzeaux, M. Garcia-Gasulla, J. Cajas, A. Artiguesa, E. Olivares, J. Labarta, and M. Vázquez, "Dynamic load balance applied to particle transport in fluids," *INTERNATIONAL JOURNAL OF COMPUTATIONAL FLUID DYNAMICS*, vol. 30, no. 6, pp. 408–418, 2016.

[2] G. Houzeaux and R. Codina, "A chimera method based on a dirichlet/neumann (robin) coupling for the navier–stokes equations," *Computer Methods in Applied Mechanics and Engineering*, vol. 192, no. 31, pp. 3343–3377, 2003.

[3] S. D. Ramsey, K. Potter, and C. Hansen, "Ray bilinear patch intersections," *Journal of Graphics Tools*, vol. 9, no. 3, pp. 41–47, 2004.

Edgar Olivares was born in Barcelona in 1985. He received a B.Sc. degree in Physics in UB and a M.Sc. degree in Computational Physics in UPC. He is about to finish his PhD about Lagrangian particle transport simulations in BSC and has recently moved to Fusion group, also in BSC. He has experience in particle transport algorithms, using them in different applications such as Computational Fluid Dynamics (CFD) or plasma simulations. His work has been always focused on a High Performance Computing (HPC) environment, making him developing applied mathematics and computational skills.

FrAG-PELE: Novel Fragment-based Growing Tool for hit-to-lead in Early Drug Discovery

Carles Perez Lopez*, Victor Guallar*†
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Institucio Catalana de Recerca i Estudis Avanats (ICREA), Barcelona, Spain
E-mail: {carles.perez, victor.guallar}@bsc.es

Keywords—Drug Discovery, Fragment-based Growing, Scoring, Hit-to-lead, Free Energy, PK properties, ADMET, Ligand.

I. Extended Abstract

A. Introduction

The pharmaceutical industry has a clear need for improvement in drug design techniques due to the incremental research cost for each new drug delivered to the market. To address this aspect, computer tools play an important role in the reduction of expenses, specially in early drug discovery (EDD).

EED process can be summarized in three main steps: (1) target identification and validation, (2) hit finding and (3) hit-to-lead and lead optimization. In this last step, the main goal is to refine each hit trying to improve their efficacy, selectivity, and adequate their ADMET (Administration, Distribution, Metabolism, Excretion, and Transport) and PK properties[1].

This preclinical research is typically done using repositories of existent molecules. However, in many cases, there is a need for completely novel approaches to cure diseases where current chemical compounds have repeatedly failed. Here is when computational methods come into play.

There are several strategies to construct new molecules. One of those is the growing method, which can be split in two different approaches depending on the size of our ”bricks”: Atom-based and fragment-based.

Fragment-based have become more popular than the atom based counterpart, as indicated by the recently developed programs[2][3]. In fact, the current state of the art for this kind of software focuses on the growing part, while the evaluation of the resultant molecules is lagging behind. This is obviously due to the associated difficulties of computationally predicting accurate protein-ligand binding free energies. For this reason, an improvement of the scoring functions would be highly beneficial for the drug design community.

B. Our approach

In our lab, the Electronic and Atomic Protein Modeling lab at Barcelona Supercomputing Center, a state of the art technique for molecular simulations had been developed. The method, called Protein Energy Landscape Exploration (PELE)[4], combines a perturbation step, trough a random translation and rotation of the ligand, with a relaxation step, via protein structure prediction techniques and energy minimization. The final result of these steps is accepted or rejected according to the Metropolis algorithm. The combination of Monte Carlo sampling with protein structure prediction techniques represents a breakthrough in modeling (sampling) protein-ligand interactions, in fact, it was recently highlighted in the latest CSAR challenge (a blind benchmark for docking and scoring methods) [5] as a remarkable achievement in drug design.

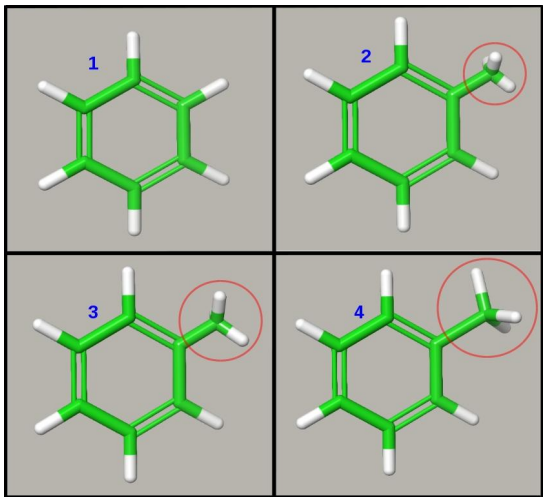


Fig. 1. A simple representation of the slow-growing schema. From a phenyl molecule, we are building a methyl-phenyl after the addition of a methyl in four steps (1-4).

In relation with the growing part, given an initial protein-ligand structure, a fragment in PDB format and the linking site, FrAG (Fragment-based Automatic Growing) is able to set up all necessary files to run PELE and use it to grow.

The whole process follows a slow-growing scheme as shown in the **Figure 1**. Our strategy is based on running successive PELE simulations at the same time that the growth is taking place. Before each simulation, FrAG computes a linear increase of Van Der Waals radius, bond length and charges modification for all atoms of the fragment to avoid major alterations in free energy. Then, once PELE has finished, FrAG analyses the results and choose the best structure that will be used as input for the next simulation.

Through this method, thanks to PELE's protein structure prediction, it is expected to find new spaces to place the molecule which would be difficult to obtain running ordinary simulations.

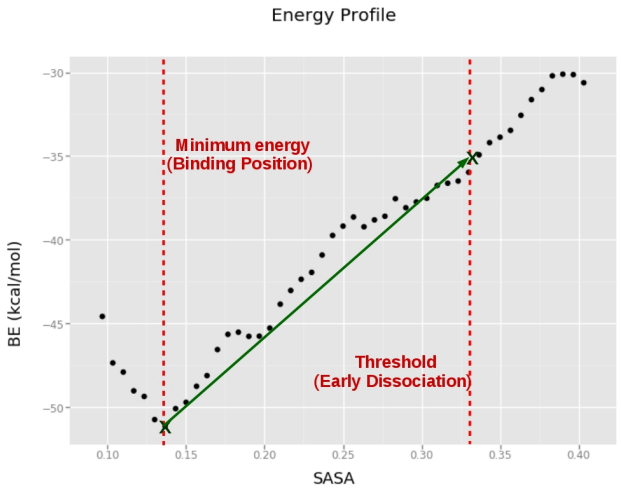


Fig. 2. Graphical representation of the scoring method. The energy profile has been generated from the data obtained during PELE simulation. The first point is identified using the minimum energy value that is considered as the binding position. The second point is obtained after defining a threshold (treated as early dissociation state) of SASA and interpolate the value of BE from the energy profile. Then, the slope between this two points is computed in order to get the score.

When the growing part has finished it is required to score the results. In accordance, it is being developed an unbinding scoring protocol with PELE based on recent work called DUck by Barril's lab[6]. In this method, we will use PELE to lead the ligand outside the binding pocket. During the unbinding, different measurements of Binding Energy (BE) and Solvent Accessible Surface Area (SASA) will be computed, and then, after analyzing the collected data a profile of SASA vs BE can be done. You can see an example in the **Figure 2**. Afterward, our score is the slope computed by the following way:

$$\frac{BEBindingPosition - BEEarlyDissociation}{SASABindingPosition - SASAEarlyDissociation} \quad (1)$$

C. Preliminary results

FrAG-PELE has been tested in a simple case of a T4 lysozyme with a benzene bonded (PDB ID: 181L), where we performed the growing until reaching a phenylethane. This result was compared with the crystallized structure (PDB ID: 1NHB) and we obtained an RMSD between ligands of 0'956.

The scoring method has been proved in a target where high-quality binding data for a series of fragment-sized ligands and the high-quality crystal structure was available[7]. Eleven simulations were performed in DNA Ligase (PDB ID: 4CC5), one for each fragment-sized ligand that we wanted to score. Finally, we set the threshold in a SASA value of 0'4 and we computed the score for the eleven ligands. A regression analysis was performed to compare experimental data (Gibbs free energy calculated) with the score obtained and we got an R-squared of 0'875.

D. Conclusions

Although that the first results are satisfactory, it is needed to perform more tests in order to further evaluate our software. In future updates, we would like to implement a pre or post-filtering method in order to take into account the synthesizability and the ADMET properties of the resultant molecules. Furthermore, we would like to automatize growing process, finding automatically which is the best position to grow the new fragment without user's intervention. When all these would be implemented, we think that FrAG-PELE could be a useful tool for hit-to-lead in EDD.

II. Acknowledgment

The authors would like to thank the cooperation in the project to all the members of Electronic and Atomic Protein Modeling group in the Barcelona Supercomputing Center and the assistance provided by some members of Nostrum Biodiscovery.

References

[1] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *British Journal of Pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.

[2] Y. Yuan, J. Pei, and L. Lai, "LigBuilder 2: A Practical de Novo Drug Design Approach," *Journal of Chemical Information and Modeling*, vol. 51, no. 5, pp. 1083–1091, 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci100350u

[3] N. Chéron, N. Jasty, and E. I. Shakhnovich, "OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4171–4188, 2016.

[4] K. W. Borrelli, A. Vitalis, R. Alcantara, and V. Guallar, "PELE: Protein energy landscape exploration. A novel Monte Carlo based technique," *Journal of Chemical Theory and Computation*, vol. 1, no. 6, pp. 1304–1311, 2005.

[5] H. A. Carlson, R. D. Smith, K. L. Damm-Ganamet, J. A. Stuckey, A. Ahmed, M. A. Convery, D. O. Somers, M. Kranz, P. A. Elkins, G. Cui, C. E. Peishoff, M. H. Lambert, and J. B. Dunbar, "CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma," *Journal of Chemical Information and Modeling*, vol. 56, no. 6, pp. 1063–1077, 2016.

[6] S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N. Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, and X. Barril, "Dynamic undocking and the quasi-bound state as tools for drug discovery," *Nature Chemistry*, no. November, pp. 2–7, 2016. [Online]. Available: http://dx.doi.org/10.1038/nchem.2660

[7] T. B. Steinbrecher, M. Dahlgren, D. Cappel, T. Lin, L. Wang, G. Krilov, R. Abel, R. Friesner, and W. Sherman, "Accurate Binding Free Energy Predictions in Fragment Optimization," *Journal of Chemical Information and Modeling*, vol. 55, no. 11, pp. 2411–2420, 2015.

Carles Perez Lopez has his Bachelor's Degree in Biomedical Sciences in the Autonomous University of Barcelona (UAB). He is currently enrolled in his second year of Master Degree of Bioinformatics for Health Sciences in Pompeu Fabra University (UPF) and at the same time, he is doing his Master Thesis in Electronic and Atomic Protein Modeling (EAPM) group in Barcelona Supercomputing Center (BSC).

Modelling of Alfvénic instabilities in complex toroidal magnetic geometries for fusion

Allah Rakha[1†] M. J. Mantsinen[1, 2] A. López-Fraguas[3] F. Castejón[3]
A.V. Melnikov[4,5] S. E. Sharapov[6] D. A. Spong[7]

[1]Barcelona Supercomputing Center, Spain, [2]ICREA, Barcelona, Spain, [3]Fusion National Laboratory, CIEMAT, 28040, Madrid, Spain, [4]National Research Center 'Kurchatov Institute', 123182, Moscow, Russia, [5]National Research Nuclear University MEPhI, 115409, Moscow, Russia, [6]CCFE, Culham Science Centre, OX14 3DB, UK, [7]Oak Ridge National Laboratory, TN, USA

[†]allah.rakha@bsc.es

***Keywords*— Nuclear fusion energy, MHD instabilities, Alfvén eigenmodes, Modelling and simulation**

I. Introduction

Nuclear fusion is the way to produce enormous amounts of relatively clean energy by fusing light hydrogen isotopes, deuterium (D) and tritium (T), into heavier helium nuclei (He) born at energy 3.5 MeV, and a very energetic neutron (n) of 14 MeV. To achieve the goal of fusion, a mixture of ionized gases i.e. a plasma consisting of D and T ions, is heated to extreme temperatures of T = 10^8 K in specific magnetic fusion devices [1]. The magnetic nuclear fusion devices can have quite complex magnetic field topologies, including nested magnetic surfaces, islands and stochastic domains. To achieve the high temperatures, the fusion plasmas are heated with auxiliary heating systems generating fast ions and/or with fusion produced alpha-particles (He at 3.5 MeV). The fast particles interact with plasma in confining magnetic field and often excite magnetohydrodynamic (MHD) instabilities in the range of Alfvénic frequencies. These MHD instabilities may degrade the confinement of energetic particles (EPs) [2]. Furthermore, these instabilities in complex 3D magnetic geometries may be more dangerous because of the additional coupling between toroidal harmonics which generate specific Alfvén gaps [3] in the Alfvén continuum with some discrete spectrum of weakly-damped Alfvén Eigenmodes (AEs). These AEs are widely investigated in complex toroidal magnetic geometries both in experiments and theory. Numerical modelling and simulations [4] are playing an important role in the explanation of current experimental findings and for predictions in future devices.

In this work we investigate AEs [5] in a 3D toroidal magnetic geometry of the flexible TJ-II heliac shown in Figure 1. TJ-II is a machine with a four period (N_{fp} = 4) magnetic field of B_0 = 0.95 T, with a low magnetic shear, major radius, R_0 = 1.5 m and averaged minor radius, <a> = 0.22 m. Main advantage of the TJ-II flexibility is the ability to provide a platform for investigating various types of Alfvénic modes in quite different magnetic configurations.

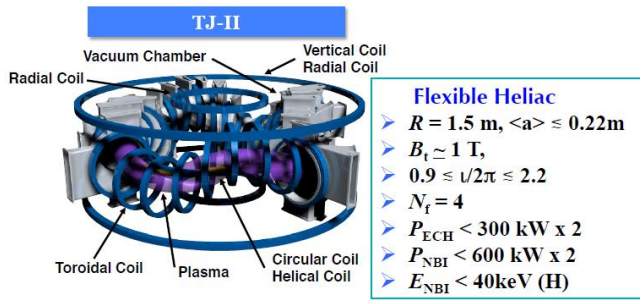


Figure 1: Schematic diagram of TJ-II flexible heliac located at CIEMAT Madrid, Spain.

By varying the magnetic configuration, the TJ-II machine can investigate plasma heating by energetic ions produced with neutral beam injection (NBI), which mimic the mechanism of self-heating with energetic alpha-particles and associate instabilities and demonstrate the fusion self-heating capability of the complex field devices. Here, the comparison of Alfvénic instability modelling results with the experimental findings for TJ-II dynamic discharges (with varying magnetic configuration) is presented. The simulation results for AEs in this complex geometry are in good agreement with the experimental findings. In this paper, the modelling of Alfvén continuum structures in TJ-II plasmas is performed with the STELLGAP [6] code and AEs structures with the spectral code AE3D [7]. Our modelling is focused on investigating the possible gaps in Alfvén continuum structures and AE profiles with their frequencies, combination of prominent mode numbers and radial localization.

II. Physics of Alfvén Eigenmodes

The AEs are coherent MHD waves that exist in toroidal magnetic fusion devices. In a toroidal magnetic geometry, the intersection points of the counter propagating Alfvén waves with equivalent parallel wave vectors $|k_{||}|$ generate the Alfvén gaps in the continuum structures. These gaps are the prominent locations where the AEs can exist and get excited by energetic particles. The lack of axial symmetry and strong shaping associated with complex toroidal 3D magnetic geometries further enhance the coupling between these gaps and generate a more condensed set of prime locations for exciting the AEs. Finding these gap structures and prominent AEs in them with their potential profiles is the main work presented in this paper. The Alfvén continuum solver STELLGAP solves a symmetric generalized matrix eigenvalue continuum equation, by giving the continuum mode structure and eigen frequency. The calculation of discrete AEs is considerably more complicated than calculating the continuum structures. A reduced MHD formulation [8] in the spectral code AE3D is employed to calculate the AEs structures, potential profiles and, radial extents. . The main eigen-value equation which performs these tasks in the AE3D code comes from the vorticity equation and the ideal Ohm's law.

III. Modelling of Alfvénic instabilities in the TJ-II Stellarator

The AEs are modeled in TJ-II by considering discharges in which Alfvén mode activity was experimentally observed. In this paper, we focus on two dynamic discharges to investigate the effect of complex magnetic configurations on AEs.

The dynamic discharges were performed at TJ-II to investigate the chirping behavior of NBI-driven Alfvén modes caused by magnetic configuration variations in the TJ-II stellarator. The experiments found the coexistence of steady and chirping modes [9] as shown in Figure 2. For the modelling of chirping and steady modes, the two similar evolving discharges with dynamically increasing and decreasing iota values, the shots 29834 and 29839 are considered exhibiting the coexistence of chirping and steady modes. These discharges are interesting due to their important features of simultaneous existence of chirping and steady frequency modes.

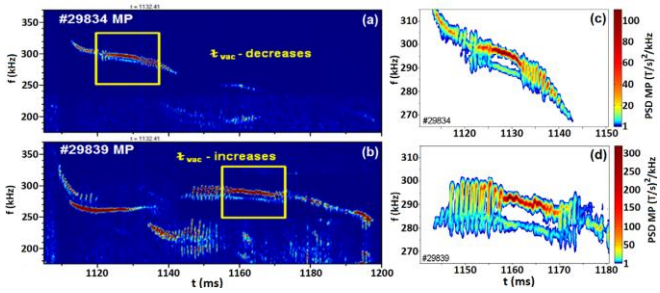


Figure 2: Experimental observations of Alfvén eigenmodes (AEs) modelled using Reduced MHD simulations for TJ-II stellarator discharge 29834 and 29839 [9].

The calculation of the spectra and the radial location of the modes at three different time slices to map the full spectrum of observed modes. For discharge 29834, the simulations are done at t = 1125, 1130 and 1135 ms and similarly for discharge 29839 at t = 1150, 1160, 1170 ms. The simulation results for this section are summarized in Table 1, which are consistent with the experimental findings. The Alfvén continuum gap structures and AE mode structure for one of the modeled cases are presented in Figure 3.

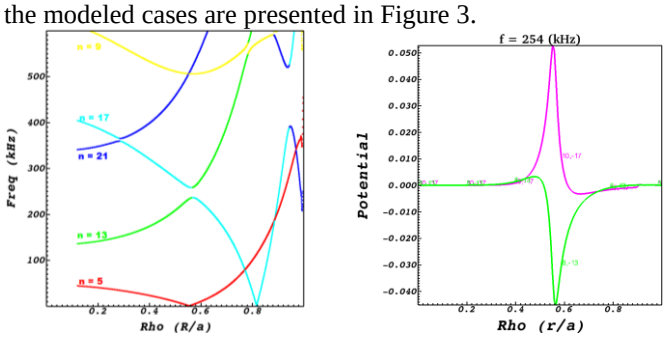


Figure 3: Alfvén continuum gap structures in left and AE structure for steady mode in discharge 29839 at t = 1160 ms. The prominent toroidal mode numbers are shown distinctly with color coding in the graphs

TABLE I
SUMMARY OF MODELLING RESULTS , WHERE, 'S' AND 'C' CORRESPOND TO THE STEADY AND CHIRPING TYPES OF MODES, RESPECTIVELY

Discharges	Modes (m, n)	Frequency (kHz)		Radial location (ρ)
#29834	(11, -19)	276(s)	292(c)	0.65/0.75
	(2, -3)	272(c)	275(s)	0.40/0.45
#29839	(10, -17)	289(c)	254(s)	0.70/0.55
	(8, -13)	251(s)	234(c)	0.55/0.80

IV. SUMMARY AND FUTURE EXTENSIONS

The modelling and simulation analysis of TJ-II dynamic plasmas support the coexistence of chirping and steady AEs. Modelling has revealed that the modes with steady frequencies are relatively localized close to plasma center with lower frequencies, given the higher density at these radial positions. On the other hand the modes with chirping or bursting behavior are localized at larger values of ρ and with relatively lower frequencies.

The extension of this work will lead to model similar discharges with different iota profiles to explore the effect of magnetic configuration on AEs. The fast ions density and pressure effects will also be investigated using non-linear modelling and wave-particle interaction. The resonant interaction of EPs with such modes will be also studied. Furthermore, the comparison of these calculations with the experimental data in TJ-II, and in other 3D devices will be addressed.

V. ACKNOWLEDGEMENTS

This work was supported by the grant from AGAUR-FI, Catalan Government (2016_FI_B_01057) of Spain. The experimental observations with Russian Scientific Foundation, project 14-22-00193 and it was partly supported by the Competitiveness Programme of NRNU MEPhI.

REFERENCES

[1] J. Ongena et al. Nature Physics volume 12, 398 (2016)
[2] N.N. Gorelenkov et al, Nucl. Fusion 54, 125001 (2014).
[3] D. A. Spong. Phys. Plasmas 22, 055602, (2015)
[4] D. A. Spong. Phys. Plasmas 18, 056109, (2011)
[5] R. Jiménez-Gómez et al. , Nucl. Fusion 51,033001 (2011)
[6] D. A. Spong, R. Sanchez and A. Weller. Phys. Plasmas 10, 3217, (2003).
[7] D. A. Spong, D'Azevedo and Y. Todo. Phys. Plasmas 17, 022106, (2010).
[8] S. E. Kruger, C. C. Hegna and J. D. Callen, Phys.Plasmas 5, 4169 (1998).
[9] A.V. Melnikov et al., Nucl. Fusion. 56, 076001 (2016).

Author biography

Allah Rakha, received the M.Phil. degree in Physics from the Pakistan Institute of Engineering & Applied Sciences (PIEAS), Islamabad in 2008 with full fellowship from federal government of Pakistan. He also obtained the M.S. degree in Nuclear Fusion & Engineering Physics from the Ghent University, Belgium, in 2015 with two years Erasmus Mundus fellowship. Since October 2008, he has been working as Lecturer in Physics at Department of Physics & Applied Mathematics (DPAM), PIEAS Islamabad. In 2016, he joined Barcelona Supercomputing Center (BSC) as PhD researcher in Fusion group under the supervision of ICREA Prof. Mervi Mantsinen. He won a prestigious AGAUR FI predoctoral grant from the Catalan government to partially fund his PhD. His current research interests include plasma physics, MHD of fusion plasmas, and Alfvénic instabilities in stellarator devices. He also closely works under EUROfusion education work package (WPEDU).

Top View Human Head and Shoulder Classification Using CNN

Ivan Rivalcoba*†, Isaac Rudomín*, Krely Rodrguez‡
*Barcelona Supercomputing Center, Barcelona, Spain
†Instituto Tecnológico de Gustavo A. Madero, CDMX, México
‡Instituto Tecnológico de Minatitlan, Veracruz, México
E-mail: {ivan.rivalcoba, isaac.rudomin}@bsc.es kyan.itmina@gmail.com

Keywords—Human Detection, Computer Vision, CNN.

I. Extended Abstract

Both industries and scientists consider the human behavior analysis on crowds a powerful source of data for computer applications in the field of security on smart cities and surveillance systems, urban design and planning, video games, autonomous car to mention a few. That is one of the reasons pushing academic and commercial researchers towards a deeper understanding of how humans act, make decisions and plans when they are cruising their environment.

Before even thinking of getting a piece of software capable to analyze a crowd, the first challenge as a priority to tackle is the human detection and tracking. The two main crucial requirements for this topic are high accuracy and real-time speed that means human detectors that are accurate enough to be relied on and fast enough to run on commercial computer hardware, taking as an example of the above mentioned, hardware limited on compute power such as smartphones or tablets. In the present document we focus on the human detection usually the first stage over any approach of human tracking.

Through the years there have been many different types of methods to detect people in video sequences, the first one was the Viola Jones [1] proposed in 2001 who described a machine learning approach for visual object recognition and also introduced the Integral Images", a popular technique to accelerate the calculus of many descriptors. Dalal et al. [2] performed a complete study of Histogram of Oriented Gradients (HOG) applied to the representation of humans. This method offers good results for pedestrian detection by evaluating local histograms of image gradient orientations over a dense normalized overlapping grid, giving a better accuracy in comparison with Viola jones method but with one important drawback and is that it requires a multiscale sliding window causing a bad performance on speed.

All the above methods now are considered as the classical approach to deal with the problem of recognition that was originated by a work published in 2013 by Pierre Sermanet [3], they proposed a multi-scale sliding window algorithm using Convolutional Neural Networks (CNNs). And suddenly after that work, deep learning has become a standard in computer vision tasks [4]. The following section will describe a system developed to classify human head and shoulder as the first stage of a bigger system to model human steering from real humans recorded on video sequences.

Fig. 1. Head and shoulder of humans are Omega shaped.

A. A CNN APPROACH TO CLASSIFY HUMAN HEAD AND SHOULDER REGIONS ON VIDEO SEQUENCES

We present a system capable of classifying head and shoulder sections of a human. We decided to detect the head and shoulders section of the body due to the fact that the human head remains constant over all the scene, unlike other parts of the body. The head and shoulders section are omega (Ω) shaped, see Figure 1. This property makes the head and shoulders the most stable parts of the body to be detected and tracked.

We propose the following CNN architecture designed to prevent the use of huge amount of data to train the neural network (Figure 2).

It is worth to notice that we avoid using aggressive dropout, finding a good balance to avoid overfitting but saving training time. The Dataset used to train the network is the same employed by Li et al. [5], this dataset consist of 3909 images for training and 2143 images for validation, to ensure a better generalization of the omega shape, we added data augmentation to generate more images for our training. The augmentation consisted in the inclusion of transformations including rotation, horizontal and vertical flip. The detailed

TABLE I. Data augmentation allows to get a plausible train with a limited dataset

Transformation	Value
Rotation range	30
Width shift range	0.1
Height Shift Range	0.1
Shear Range	0.2
Zoom Range	0.2
Horizontal flip	TRUE

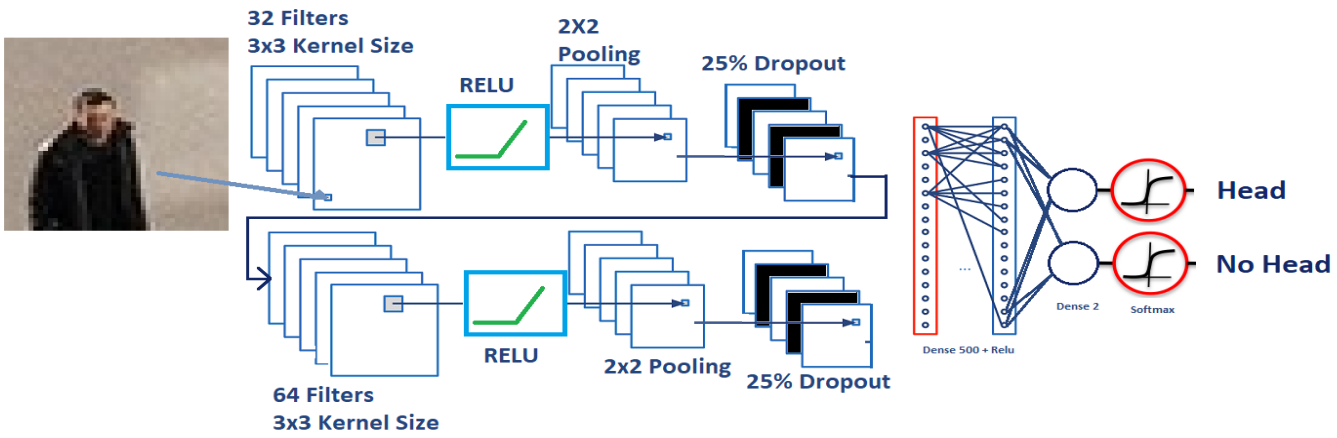


Fig. 2. Architecture of Convolutional Neural Network performing the classification task

transformations are showed in table I.

B. Results

Our method produced 2,068,894 trainable parameters, causing a training time of 15 minutes on a CPU with no GPU enabled. It was trained on 3909 samples and validated on 2143 samples. Producing an accuracy of 91% on validation data and a test score of 18%. We use 10 epochs on the training stage, the necessary to get a good generalization.

C. Conclusion

With the rebirth of neural networks and the end of the well-known winter of AI the classical methods of computer vision have become outdated. The well results of CNNs on computer vision context have proven that the use of CNNs on computer vision task is a trend with empower the computer vision to reach new levels. In this extended abstract we present a CNN architecture to classify head and shoulder from top view images as a first part of a bigger project aimed to model the human behavior in a crowd.

II. Acknowledgment

I would like to thank to SECITI in Mexico for providing the founding for the present research and the TecNM to allowed me to participate in this postdoctoral stay, also I must express my gratitude for all the team that conforms the Barcelona Super Computing Center for facilitate the equipment and the infrastructure for the project.

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–511–I–518, 2001. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=990517

[2] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1467360

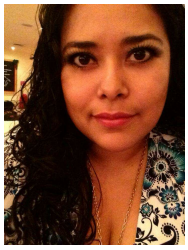
[3] P. Sermanet *et al.*, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," 2013. [Online]. Available: http://arxiv.org/abs/1312.6229

[4] A. Lee, "Comparing Deep Neural Networks and Traditional Vision Algorithms in Mobile Robotics," 2015. [Online]. Available: https://pdfs.semanticscholar.org/1b6f/569b79721037425fca034c7ae47904fb9276.pdf

[5] M. Li *et al.*, "Rapid and robust human detection and tracking based on omega-shape features," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, nov 2009, pp. 2545–2548. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5414008

Ivan Rivalcoba received his Ph.D from Instituto Tecnolgico de Estudios Superiores de Monterrey Campus Estado de Mxico in 2015, He is a full-time professor at Instituto Tecnològico de Gustavo A. Madero in Mèxico City. Currently is working at the Barcelona Supercomputing Centre as a Postdoctoral fellow. Their work focus on using artificial intelligent methods for people detection.

Isaac Rudomin is a senior researcher at the Barcelona Supercomputer Center, which he joined in 2012. His focus is on crowd rendering and simulation including generating, simulating, animating, and rendering large and varied crowds using GPUs in consumer-level machines and in HPC heterogeneous clusters with GPUs. Previously, Isaac was on the faculty at Tecnologico de Monterrey Campus Estado de Mexico (from 1990 to 2012). He finished his Ph.D. at the University of Pennsylvania under Norman Badler on the topic of cloth modeling.



Krely Rodrguez in 2013 she completed her PhD in Mathematics Education at the Escuela Libre de Ciencias in the city of Jalapa, Veracruz. Master of Science in Electronic Engineering from the Instituto Tecnològico de Orizaba. She is currently a professor of Basic Sciences at the Instituto Tecnolgico de Minatitln, she actually collaborates in interinstitutional research projects of Artificial Vision since 2013 with Dr. Jorge Ivn Rivalcoba Rivas.

A Linux Kernel Scheduler Extension for Multi-Core Systems

Aleix Roca*, Vicenç Beltran*, Kevin Marquet†
*Barcelona Supercomputing Center
†Univ Lyon, INSA Lyon
E-mail: {arocanon, vbeltran}@bsc.es Kevin.Marquet@insa-lyon.fr

***Abstract*—Current runtime systems take care of getting the most of each system core by distributing work among the multiple CPUs of a machine but they are not aware of when one of their threads (workers) perform blocking calls (*e.g.* I/O operations). When such a blocking call happens, the processing core is stalled, leading to performance loss. In this project, we present two new and independent methods to minimize the effect of I/O operations: The first one is a Linux kernel extension denoted *User-Monitored Threads* (UMT) and the second one is a user-space library named *libsio2aio*. Our Linux kernel extension allows a user-space application to be notified of the blocking and unblocking of its threads, making it possible for a core to execute another worker thread while the other is blocked. The libsio2aio library intercepts the family of read/write system calls, interchanges them by its asynchronous version, and returns control back to the runtime while the I/O operation is being resolved. In both cases we use the Nanos6 runtime to test the new methods.**

Keywords*—Linux Kernel, Process Scheduler, I/O, High-performance computing.

I. Introduction

High performance computing applications usually execute in worker threads that are handled by a userland runtime system, itself executing on top of a general purpose operating system (OS). The main objective of the runtime system is to provide maximum performance by getting the most out of available hardware resources. On a multicore machine, this translates to distributing the work of applications among the machine's available cores and balance each core workload.

Runtime's balancing capabilities are subject to the underlying OS scheduler. When a thread performs a blocking I/O operation against the OS kernel, the core where the thread was running becomes idle until the operation finishes. This problem can lead to huge performance loss as some HPC or high-end server applications perform lots of I/O operations because they heavily deal with file and network requests.

One approach to address this issue is to make the runtime system aware of when blocking and unblocking events happen. In this way, it can chose to execute another worker thread while the first one is blocked. A general approach to detect any blocking operations (such as page faults) requires special kernel support, however, if we narrow the scope of blocking operations to the standard syscalls, a user-space library will suffice. There has been related work on the kernel side [1], [2] but is has been rejected due to its complexity. Instead, both our kernel and user space solutions main advantages are its simplicity.

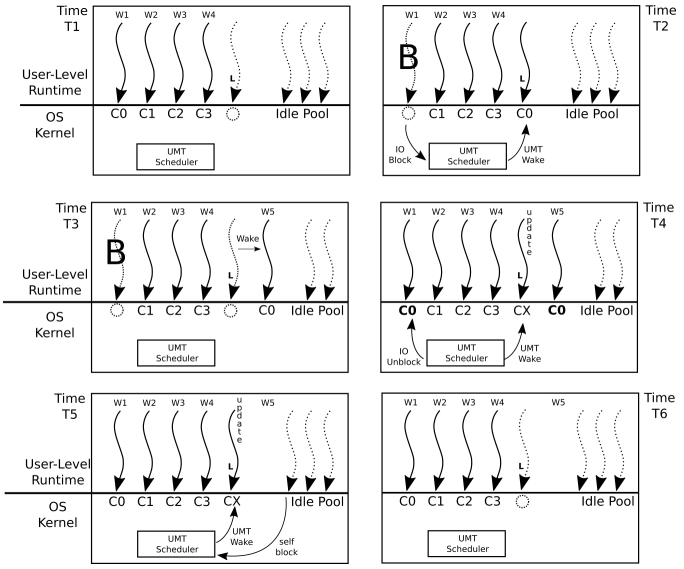


Fig. 1. UMT model overview example

II. UMT Overview

In UMT, the Linux kernel uses a communication channel to notify a user-space application of blocking and unblocking events among their threads. An overview of this functioning is given in Figure 1. The W_i are user-space runtime's worker and L denotes the user-space runtime's *Leader Thread* whose role is to monitor the communication channel. Basically:

- At time T1, four workers W1, W2, W3 and W4 are bound to CPU's C0, C1, C2 and C3 respectively. The Leader Thread is not bound to any CPU and is waiting for UMT events. A pool of idle workers remain blocked until they are needed.
- At time T2, the worker W1 blocks because of an I/O operation and the Leader Thread is notified of the event.
- At time T3, the Leader Thread wakes an idle worker from the pool and waits again for more events. (When W5 wakes, it would also generate an unblock event which is omitted for simplicity). Worker W5 is now running on a CPU; without the proposed mechanism, it would have been idle.
- At time T4, W1 is unblocked after the I/O operation finishes. An unblocking event is generated and the

Leader Thread wakes up. Because there is not any free CPU at the moment, the Leader Thread waits until it momentously preempts another worker. Once it does so, it reads the UMT events and registers that multiple workers (W1 and W5) are running on the same CPU (C0).

- At time T5, after the W5 worker finishes executing tasks, it checks the Leader Thread registers and realizes that there is an oversubscription problem affecting its current CPU. To fix the problem, the worker self surrenders and returns to the pool of idle workers. This generates another event that wakes up the Leader Thread and updates the register of events.
- At time T6, the oversubscription problem has ended and the four workers are running normally.

A. UMT design: Linux kernel-side

Our proposal for the UMT kernel support includes two new system calls to initiate and manage UMT and the infrastructure for the notification channel between kernel- and user-space based on the standard *eventfd*[1] (EFD) file descriptors.

When calling *um_mode_enable* the Linux Kernel initializes an EFD for each CPU on the system and stores them in the context of the calling process. This process' threads start being monitored as soon as each of them allows it by calling the *ctlschedumfd* syscall. The main idea is that each of these EFD keeps a per-CPU count of how many monitored threads are in the ready state. The actual Linux kernel instrumentation of the EFD writing points has been placed into a wrapper around the main context switch entry point called `__schedule()`.

B. UMT design: User-space runtime design

In order to validate the proposal, we have adapted the *Nanos6* runtime[3] of the *OmpSs-2*[4] task-based programming model to work with our kernel extension.

Nanos6 consists of a set of workers threads whose objective is to run tasks and a special management thread called Leader thread. The Leader thread first calls *um_mode_enable* to initialize the UMT kernel structures and then monitors all the per-CPU EFDs using a standard epoll system call. Each worker thread first calls *ctlschedumfd* once to enable monitoring and then start executing tasks. When one of the monitored threads produces an event (it block or unlocks), the Leader Thread wakes up from the epoll sleep and reads the EFD. If the count of ready threads on the CPU that has triggered the event is zero and there are still tasks to execute, the Leader Thread retrieves an idle thread from a pool and gives it a task to execute on the idle CPU.

If the previously blocked worker wakes up while the new worker is running, both threads will have to compete for the CPU. However, this oversubscription problem only prevails for a limited amount of time. Workers have been provided with a oversubscription protection mechanism that consists on checking the counter of ready threads of its CPU after finising executing a task. If the count is greater than one, workers self surrender to allow other workers to run freely on the CPU.

[1] An eventfd is a simplified pipe that was designed as a lightweight inter-process synchronization mechanism. Internally, an eventfd holds a 64 bit counter that can be written to increment its internal value or read to clear and return it.

III. LIBSIO2AIO OVERVIEW

The libsio2aio user-space library defines wrappers for the `pread()`, `pwrite()`, `preadv()` and `pwritev()` syscalls (all Linux Kernel native AIO supported syscalls) which call the asynchronous version of the intercepted syscall. After submitting the request, it checks whether it has immediately completed or not. If it is the case, the wrapper returns immediately as well. Otherwise, it pauses the execution of the current tasks (not the thread) and transfers control to the runtime. The runtime is then able to execute other tasks in the current CPU while the I/O operation is being resolved. The runtime periodically checks whether any AIO request has completed and if it is the case, the task that submitted the AIO request is unblocked. Unblocked tasks execution are later resumed by Workers.

IV. EXPERIMENTATION

We have tested UMT using a synthetic benchmark. The benchmark simply maps a region of memory using *mmap* and creates a set of independent tasks whose purpose is to write and sync random mapped data. As a result we have achieved a speedup of x10. We are currently testing libsio2aio and we have not yet been able to find an appropiate benchmark that benefits from its advantages.

V. CONCLUSION

Finally, we conclude that both UMT and libsio2aio have two main effects: on the one hand, they provide a mechanism to queue more I/O operations which approaches the real I/O rate to the one specified by the manufacturer of the storage device. On the other hand, blocked processes no longer obstruct the core and useful computations can be done while I/O petitions are being served. In the case of UMT, the oversubscription problem limits performance but as results show, it is not always a problem. Future work will focus on finding more I/O intensive applications to test both presented approaches.

REFERENCES

[1] T. E. Anderson, B. N. Bershad, E. D. Lazowska, and H. M. Levy, "Scheduler activations: Effective kernel support for the user-level management of parallelism," *ACM Transactions on Computer Systems (TOCS)*, vol. 10, no. 1, pp. 53–79, 1992.

[2] V. Danjean, R. Namyst, and R. D. Russell, "Linux kernel activations to support multithreading," in *In Proc. 18th IASTED International Conference on Applied Informatics (AI 2000*. Citeseer, 2000.

[3] BSC, "Nanos6 runtime," https://github.com/bsc-pm/nanos6, 2018.

[4] J. M. Perez, V. Beltran, J. Labarta, and E. Ayguadé, "Improving the integration of task nesting and dependencies in openmp," in *Parallel and Distributed Processing Symposium (IPDPS), 2017 IEEE International*. IEEE, 2017, pp. 809–818.

Aleix Roca is a Linux Kernel passionate. He admires the leading Linux developers technical skills and their tremendous effort to manage the world's biggest open source community. Aleix studied computer engineering at UPC-FIB. After obtaining his degree he enrolled the Master in Innovations and Research in Informatics specialized on High Performance Computing at UPC-FIB, where he obtained the *Severo-Ochoa MSc scholarship*. During his master studies he joined the *Barcelona Supercomputing Center* where he developed his final master thesis on the Linux kernel and programming models. Currently he has started a PhD at BSC where he continues his research on the Linux Kernel and HPC.

Characteritzation of pathological mutations affecting protein-protein interactions for drug discovery

Mireia Rosell*, Juan Fernández-Recio*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Institut de Biologia Molecular de Barcelona (IBMB), CSIC, Barcelona, Spain
E-mail: {mireia.rosell, juan.fernandez}@bsc.es

Keywords—Hot-spots, protein-protein docking, structural variants, binding effect.

I. Extended Abstract

Pathogenic single nucleotide variants (SNV) can affect binding affinity or change the specificity of a protein-protein interaction (PPI). It is a known fact that modulating PPIs with small molecules is a long sought strategy in drug discovery. We face three major problems: i) the lack of available structures for the majority of PPIs in human; ii) the absence of natural cavities in protein-protein interfaces that could be used to identify small molecules as in standard enzyme inhibitor discovery; and iii) how a small molecule can compete with a large protein interface. We have developed a strategy for identifying small molecule inhibitors of PPIs when complex structure is not available, based on the integration of molecular dynamics with Amber for the generation of transient cavities, FPocket for the identification of such cavities [1], and computational docking calculations and hot-spot predictions with pyDock to select the best cavities for PPI modulation[2,3].

However, estimating the effects of a given single nucleotide variant on a PPI is extremely challenging. We aim to apply this methodology to known pathological variants from Humsavar and ClinVar databases that affect PPIs. There are very few experimental data for the effect on binding affinity of these SNVs (according to SKEMPI database)[4,5]. Thus, in order to estimate this effect, we initially mapped these structural variants on protein-protein complex structures included in the Protein-Protein Docking Benchmark 5 (formed by complexes with available structure for the complex as well as for the unbound components)[6]. The effect of these SNV on binding affinity is predicted with mCSM[7], pyDock and FoldX[8]. Then, for the protein-protein interactions that are stabilized by pathological variants, we will test on the unbound components of the Docking Benchmark the identification of cavities suitable to find possible small molecule inhibitors.

II. Acknowledgments

This work has been funded by grants BIO2016-79930-R and Severo Ochoa program SEV-2015-0493 from the Spanish Ministry of Economy, Industry and Competitiveness and by Severo Ochoa mobility grant.

References

[1] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10:168.

[2] Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics. 2008;9:447.

[3] Cheng TM, Blundell TL, Fernández-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body proteinprotein docking. Proteins. 2007;68:503515.

[4] Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics. 2012;28:26002607.

[5] David A, Sternberg MJE. The contribution of missense mutations in core and rim residues of protein-protein interfaces to human disease. J Mol Biol. 2015;427:28862898.

[6] Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., et al. Updates to the integrated proteinprotein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. Journal of Molecular Biology. 2015, 427(19), 30313041.

[7] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. L.Bioinformatics. 2014,30(3):335-342.

[8] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Research. 2005,33(Web Server issue): W382W388.

Mireia Rosell is a PhD student at the Protein Interactions and Docking Group in the Barcelona Supercomputing Center, lead by Juan Fernández-Recio. She did a B.Sc. In Biochemistry at Universitat Autónoma de Barcelona and afterwards she obtained an interuniversity M.Sc. in Bioinformatics for Health Sciences by Universitat Pompeu Fabra and Universitat de Barcelona. Her research is focused on the development of a new methodology for the high-throughput structural annotation of sequence variants involved in protein interactions.

Detailed Tuning and Validation of Hardware Simulators through Microbenchmarks

Rommel Sánchez Verdejo*†, Petar Radojković*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {rommel.sanchez, petar.radojkovic}@bsc.es

Keywords—Simulation, Memory simulation, CPU Simulation, DRAM, DRAM Simulation, x86

I. Extended Abstract

Hardware simulators are used by the academia and industry to prototype, explore and evaluate novel microarchitectural features. Because of its importance, it is imperative to pay special attention to their validation. Unfortunately, this process is not standardized. In this work, we describe a set of microbenchmarks for the validation of the CPU execution units and the memory subsystem, including on-chip caches and main memory. Also, we present a case study in which the microbenchmarks are used to validate a simulation infrastructure based on the ZSim [1] and DRAMSim2 [2] vs. an actual Sandy Bridge server. The presented case study shows how the microbenchmarks can be used to isolate the resource behavior of the target architecture and pinpoint the specific differences between the simulator and the target hardware.

A. Microbenchmarks: CPU execution units

All the microbenchmarks are designed using the principle that is presented in Table I. Each benchmark consist of four parts: (1) The register used as a loop iteration counter (`ecx`) is set for 10,000 times of execution; (2) The main section of the benchmark is a sequence of single repetitive instruction of the target ISA; (3) The sequence of target instructions is followed with the decrement of the loop counter register; (4) Finally, the counter value is compared with zero followed by the conditional branch to the beginning of the loop.

B. Microbenchmarks: Caches and main memory

The benchmarks that stress the caches and memory are implemented using the concept of pointer chasing. In the benchmark prologue, we allocate a contiguous section of memory and initialize it to a given array element that contains the address of the next element to fetch. The benchmarks are initialized to (1) Traverse the whole array; (2) Access different cache lines in each memory access; (3) Memory accesses have a random pattern, preventing data prefetchers to bring data to any level of cache.

Table II shows the code for memory latency microbenchmarks, which behavior are explain as follows (1) The register used as a loop iteration counter (`ecx`) is initialized; (2) The initial address of the array is passed to the assembly code as an input parameter; (2) The main part of the benchmark is a sequence of indirect load instructions (`mov(%rax), %rax`)

Line	Source code	Explanation
00001	`mov $10000, %ecx`	Initialize loop counter *ecx* to 10,000
00002	`start_loop:`	beginning of the loop
00003	`ADC %eax, %ebx`	target instruction
00004	`ADC %eax, %ebx`	target instruction
...	`...`	...
10002	`ADC %eax, %ebx`	target instruction
10003	`dec %ecx`	decrement loop counter
10004	`jnz start_loop`	if (counter $\neq$ 0) jump to start_loop

TABLE I. Structure of microbenchmark ASM core.

Line	Source code	Explanation
0001	`register struct line *next asm("rax");`	strcut line owns pointer to the next accesss
0002	`register int i asm("ecx");`	`ecx` is the loop counter
0003	`i = 1000000;`	`C` initialization of the loop counter
0004	`next = ptr->next;`	First memory access in `C` form
0005	`start_loop:`	beginning of the loop
0006	`mov (%rax), %rax`	load instruction (pointer chasing)
0007	`mov (%rax), %rax`	load instruction (pointer chasing)
...	`...`	...
1007	`mov (%rax), %rax`	load instruction (pointer chasing)
1008	`dec %ecx`	decrement loop counter
1009	`jnz start_loop`	if (counter $\neq$ 0) jump to start_loop

TABLE II. Structure of memory latency microbenchmark.

that traverse the memory access pattern; (3) The sequence of target instructions is finalized with the decrement of the loop counter register and an exit condition or jump to the beginning of the iteration. The assembly loop is wrapped-up by the `C` program which reads a previously generated file containing information about the array size and the random access pattern.

C. Case study: ZSim and DRAMSim2 vs. Intel Xeon E5-2670 SandyBridge-EP

We performed a case study in which the microbenchmarks are used to compare the simulation infrastructure integrated with ZSim and DRAMSim2 simulators. ZSim is an execution-driven CPU simulator widely used in the computer architecture research, developed to mimic the Westmere architecture and validated against real hardware using the SPEC CPU 2006 benchmark suite. DRAMSim2 is a cycle accurate model of a DRAM memory controller, DIMMs, and buses by which they communicate. It is validated against DRAM manufacturer's Verilog models.

We validate the simulators vs. a dual-socket platform. Each socket with an Intel Xeon E5-2670 SandyBridge-EP processor [3] operating at 3.0 GHz. The main memory is 16 GB and is connected to the processors using four DDR3-1600 channels. Each processor runs eight cores, the hyper-threading

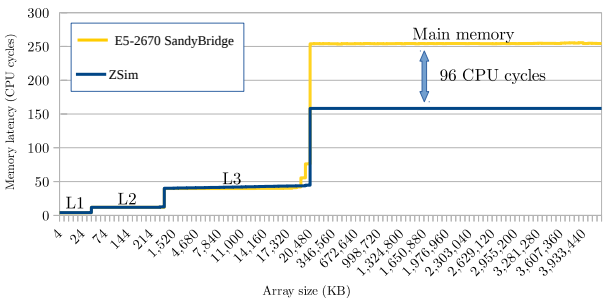


Fig. 1. Memory access latency: L1, L2, L3 cache and main memory.

feature has been disabled like in most HPC systems [4].

D. Case study: CPU Execution units

We automated the creation of 346 microbenchmarks covering a wide range of logic for integer and floating point instructions. Then, we executed the benchmarks in actual hardware and the simulator. Out of 346 tested instructions, the simulator matches around 55% the actual system latency while in 28% the simulation error is moderate or high meaning that exceeds 50% of the CPI. Since each of the microbenchmarks stresses one resource at a time, finding the sources of the simulation error is relativeley simple. The enhanced version of the simulator show much better accuracy 74% of all the instructions correctly match the actual system while less than 9% of them show a simulation error of above 50% of the CPI. Moderate or high simulation error come mainly from the complicated instructions or CPI dependency on the operand values. Detailed analysis of these cases is an ongoing work.

E. Case study: Caches and main memory

To compare the caches and main memory access time between the selected simulators and the actual hardware, we used the same strategy described in Section I-B. Table III summarizes information about array sizes used in this study to stress the memory hierarchy. Results from such comparison are shown in Figure 1. The horizontal axis of the figure represents the size of the traversed array, while the vertical axis displays the memory access latency in CPU cycles. From the figure, we can distinguish four steps of the latency corresponding to the `L1`, `L2`, `L3` cache and main memory. For the cache levels, the lines overlap: ZSim cache contention accurately represents the actual system. However, for the main memory accesses, we detect a significant gap between the simulators and the real system.

These results motivated us to further explore the sources of this error. Because ZSim is a user-level simulator, it does not take into account virtual-to-physical address translation. In the real system to mitigate the address translation overheads, we used huge memory pages (1 GB per page in our study) and contiguous memory space. Simple integration of ZSim and DRAMSim2 may lead to an underestimation of the main memory access latency. ZSim simulates memory access up to the last level of cache, while DRAMSim2 is focused on the detailed timing simulation of the memory device. This implies that a direct merge of ZSim and DRAMSim2 does not consider the delay contributed by all the circuitry between the last level cache and main memory device, including the memory controller and the memory channel.

Memory Level, Size and Scope	Number of Measurements	Array sizes (range, stride)
L1 cache 32 kB, Private	8	4 kB to 32 kB , 4 kB
L2 cache 256 kB, Private	16	46 kB to 256 kB, 14 kB
L3 cache 20 MB, Shared	32	888 kB to 20 MB, 632 kB
Main memory, 16 GB/socket	64	83.69 MB to 4 GB, 63.7 MB

TABLE III. SIZE AND ORGANIZATION OF THE CACHES AND MAIN MEMORY OF THE E5-2670 SANDYBRIDGE-EP USED IN THE STUDY.

The memory latency experiments confirm that the microbenchmarks can be resourceful to detect specific errors in the simulation configurations that might be overlooked. To this date, there exist no guidelines provided by CPU simulator developers that emphasize the importance of the proper integration with the memory simulators while considering latency of the memory controller and the memory channel.

II. CONCLUSIONS

Our study provides first steps in a systematic methodology to validate computer architecture simulators. By comparing the execution of the proposed microbenchmark on both systems, we can check whether a simulator reproduces the system behavior for that particular resource. We presented a case study in which the microbenchmarks are used to validate a simulation infrastructure based on the ZSim and DRAMSim2 simulators vs. a real SandyBridge server. This study opens a discussion about the validation of the state-of-the-art-simulators used in the computer architecture community.

III. ACKNOWLEDGMENT

This work has been published in the International Conference on High Performance Computing & Simulation (HPCS) 2017, and supported by the Collaboration Agreement between Samsung Electronics Co., Ltd. and BSC, Spanish Government through Severo Ochoa programme (SEV-2015-0493), by the Spanish Ministry of Science and Technology through TIN2015-65316-P project, and by the Generalitat de Catalunya (contracts 2014-SGR-1051 and 2014-SGR-1272).

REFERENCES

[1] D. Sanchez and C. Kozyrakis, "ZSim: Fast and Accurate Microarchitectural Simulation of Thousand-Core Systems," in *ISCA*, June 2013.

[2] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "Dramsim2: A cycle accurate memory system simulator," *IEEE CAL*, 2011.

[3] Intel product specification site. [Online]. Available: https://ark.intel.com/

[4] Top 500 supercomputer sites. [Online]. Available: https://www.top500.org/

Rommel Sánchez Verdejo received his M.Eng. in Computer Science from Univ. Nac. Autónoma de México, Mexico (2019). He worked for Intel Corp. in Jalisco, Mexico as a UEFI BIOS Engineer and Software Security Validation Engineer. He is pursuing a Ph.D. at Univeristat Politcnica de Catalunya jointly with the Barcelona Supercomputing Center, Spain (2106).

Fuzzy Finite State Machines in Crowd Simulation.

Leonel Antonio Toledo Diaz*, Isaac Rudomin*,
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {leonel.toledo1, isaac.rudomin}@bsc.es

Fig. 1. An example of Barcelona City. created using Unity's 3d plug -in

Keywords—Crowd Simulation, GPU, Fuzzy Logic, FSM.

I. Extended Abstract

Large crowds of pedestrians are a common phenomenon in big cities, this field of research studies and reproduces this kind of situations in virtual environments. Complex interaction within the agents is desired but it requires modeling their internal state. Through the internal state of the agents we can include reactive behaviors with multiple and changing objectives. The individual (personality) and social (group membership) properties of the agent can also be modified due to external or internal changes, such as environment, communication and mood. The inclusion of this features, and the internal state of the agents, change the behavior of the crowd. One of the main problems when simulating a crowd is to define individual traits for all the agents within the simulation. It is possible to define parameters that individually describe each of the characters involved in the simulation, however this approach turns to be unfeasible as the crowd grows bigger. We propose a method that combines finite state machines with fuzzy logic to represent concepts such as fast or slow which its definition may vary from agent to agent.

A. Fuzzy Finite state machines for crowds in urban environments.

We present a method for simulating crowds in urban environments, to create these scenarios, we use Unitys WRLD plug-in as shown in figure 1. We use the city of Barcelona in our simulation. In a similar fashion as van Essens [1] and Thomnsen [2], where 3D maps are produced using the most relevant features.

Figure 2 shows how we create the urban environment using the previously discussed techniques, we use WRLD3D plug-in which gives us detailed information about geographic locations, in this case we construct the simulation using Barcelona as a reference. Once the environment is created we incorporate the crowd into the simulation, our goal is to make the simulations as complex as possible, to reach that goal we consider two different techniques that we combine; first, we collect real data from GPS traces that describe routes that pedestrians take within the city, this trace includes information about the latitude, longitude, elevation and the time when the sample was taken, and our agents can follow the given routes and be animated accordingly. Second, we consider autonomous characters that can navigate the environment. We include simple behaviors such as patrolling, following, avoiding obstacles or pedestrians just to state a few. This behavior is controlled by finite state machines in which each agent has the freedom to decide how to change states accordingly. Nevertheless, pedestrian behavior cannot be modeled realistically using deterministic models, thats why we incorporate fuzzy logic into the simulation, this way we can create different profiles for each character, and work with concepts such as fast or slow inside the simulation, what is true for an agent might not work in the same way for other. To decide whether a character is moving fast or slow and simulate properly we use a shared library of parameters that all characters inherit from, we can manually tweak each of the variables for any given character or randomly assign values. This allows us to create two different profiles for all the elements in the simulation, the first profile is focused in information such as vision range, maximum speed, weight, turn speed, to state some. The second profile is oriented towards how each character understands fuzzy concepts such as fast or slow, this way even if the members of the crowd have the same physical profile they might behave very different according to their fuzzy parameters. One of the main advantages of this method is that all agents have access to this knowledge and without any changes to the script we can achieve a lot of variety in the crowd behavior.

B. Results

Using this approach, we have been able to simulate both complex urban environments, as well as agent behaviour within the city. In terms of performance, we have achieved a balance between render and simulation, for instance we are capable of simulating one thousand characters in scenes with a little more than 4 million polygons and 3 million vertices. For a scene similar as the one shown in figure 2, we use 1115 draw calls and 39 batched draw calls, 300 megabytes of RAM memory is required and 112 megabytes of video memory. Each frame takes about 28 milliseconds to render, which allow us to have simulation at interactive frame rates (more than 30 fps). On average, each frame has 18,000 objects in viewable space and a total of 43000 total objects in the whole simulation. This outperforms known techniques such as [3] Millans comparison between impostors and point based render, in terms of memory consumption.

Fig. 2. A crowd simulated within the city in Unity

C. Conclusions and Future Work

We present a robust approach for urban crowd simulation that can run at interactive frame-rates. Which is powerful enough to handle large environments in real time without compromising visual quality and the simulation of individual behaviors. At the current stage the system has proven to be successful in achieving meaningful diversity in terms of how characters react in different situation. Nevertheless, this is not enough, significant efforts must be done to further optimize the system, for instance LOD techniques are desirable to further expand the size of the crowd [4]. Visual variety is also considered, at this point, every character in the simulation uses the same mesh.

REFERENCES

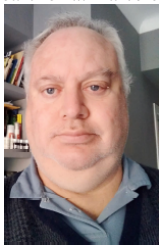
[1] R. van Essen, *Maps Get Real: Digital Maps evolving from mathematical line graphs to virtual reality models*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 3–18.

[2] A. Thomsen, M. Breunig, E. Butwilowski, and B. Broscheit, *Modelling and Managing Topology in 3D Geoinformation Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 229–246.

[3] I. Rudomín, B. Hernández, O. De Gyves, L. Toledo, I. Rivalcoba, and S. Ruiz, "GPU Generation of Large Varied Animated Crowds," *Computación y Sistemas (CyS)*, vol. 17, no. 3 (Special Issue) Supercomputing: Applications and Technologies, pp. 365–380, 2013.

[4] L. Toledo, O. De Gyves, I. Rivalcoba, and I. Rudomn, "Hierarchical Level of Detail for Varied Animated Crowds," 2014.

Leonel Toledo recieved his Ph.D from Instituto Tecnolgico de Estudios Superiores de Monterrey Campus Estado de Mxico in 2014, where he was a full-time professor from 2012 to 2014. He was an assistant professor and researcher and has devoted most of his research work to crowd simulation and visualization optimization. He has worked at the Barcelona Supercomputing Center using general purpose graphics processors for high performance graphics. His thesis work was in Level of detail used to create varied animated crowds. Currently he is a researcher at Barcelona Supercomputer Center.



] Isaac Rudomin is a senior researcher at the Barcelona Supercomputer Center, which he joined in 2012. His focus is on crowd rendering and simulation including generating, simulating, animating, and rendering large and varied crowds using GPUs in consumer-level machines and in HPC heterogeneous clusters with GPUs. Previously, Isaac was on the faculty at Tecnologico de Monterrey Campus Estado de Mexico (from 1990 to 2012). He finished his Ph.D. at the University of Pennsylvania under Norman Badler on the topic of cloth modeling.

Poster Abstracts

Recurrent Semantic Instance Semantic Segmentation

Míriam Bellver*[1], Amaia Salvador*[2], Víctor Campos[1]
Ferran Marqués[2], Xavier Giró-i-Nieto[2], Jordi Torres[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya
E-mail: {miriam.bellver, victor.campos, jordi.torres}@bsc.es, {amaia.salvador, xavier.giro, ferran.marques}@upc.edu

Keywords—Computer Vision, Deep Learning, Image Segmentation.

I. ABSTRACT

***Abstract*—We present a recurrent model for semantic instance segmentation that sequentially generates pairs of masks and their associated class probabilities for every object in an image. Our system is trainable end-to-end, does not require post-processing steps and is conceptually simpler than current methods relying on object proposals. We observe that our model learns to follow a consistent pattern to generate object sequences, which correlates with the activations learned in the encoder part of our network. We achieve competitive results on three different instance segmentation benchmarks (Pascal VOC 2012, Cityscapes and CVPPP Plant Leaf Segmentation).**

Míriam Bellver got her B.S. degree in Telecommunications Engineering in Universitat Politècnica de Catalunya. During the B.S. thesis she started to work in computer vision problems in the Image Processing Group of the university. She also obtained her Master in Telecommunications in the same faculty, and completed the Master Thesis in ETH Zürich. In 2016 she obtained a PhD grant from Obra Social "la Caixa" through La Caixa-Severo Ochoa International Doctoral Fellowship program, to do her PhD in the Barcelona Supercomputing Center about computer vision using deep learning.

∗ First two authors contributed equally.

Improving Time-Randomized Cache Designs

Pedro Benedicte[†,‡], Carles Hernandez[†], Jaume Abella[†], Francisco J. Cazorla[†,⋆]
pbenedic@bsc.es, carles.hernandez@bsc.es, jaume.abella@bsc.es, francisco.cazorla@bsc.es
[†] Barcelona Supercomputing Center (BSC) [‡] Universitat Politècnica de Catalunya (UPC) [⋆] IIIA-CSIC

***Abstract*—Enabling timing analysis for caches has been pursued by the critical real-time embedded systems (CRTES) community for years due to their potential to reduce worst-case execution times (WCET). Measurement-based protabilistic timing analysis (MBPTA) techniques have emerged as a solution to time-analyze complex hardware including caches, as long as they implement some random policies. Existing random placement and replacement policies have been proven efficient to some extent for single-level caches. However, they may lead to some probabilistic pathological eviction scenarios. In this work we propose new random placement and replacement policies specifically tailored for multi-level caches and for avoiding any type of pathological case.**

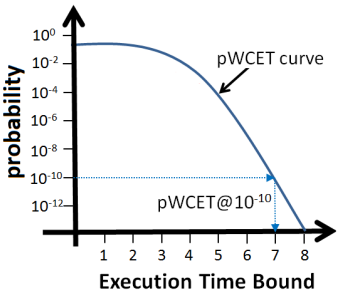


Fig. 1. Example of pWCET distribution.

I. Introduction

WCET estimation for real-time software is needed for the certification of critical systems against safety standards. WCET estimates need to be reliable and as tight as possible. A common misconception is that a WCET estimate overrun necessarily causes a system level failure. However, this is not true since mandatory safety measures are in place to manage sporadic faults. Following the probabilistic approach used to handle random hardware faults [5], MBPTA reasons on WCET as a distribution, aka probabilistic WCET (pWCET) curve (Figure 1), describing the maximum probability with which a WCET estimate can be exceeded.

MBPTA builds on a set of measurements taken during system analysis phase. Those measurements are passed as input to Extreme Value Theory (EVT) [9], a statistical tool to estimate an upper-bound distribution for distribution tails (high execution times in our case). MBPTA imposes how execution time measurements must be collected so that they capture those conditions that lead to execution times matching or upper-bounding those during system operation. EVT, part of MBPTA, requires that the execution times meet several statistical properties related to the degree of independence and identical distribution of the random variable (execution times) modelled, and whether it can be modelled with an exponential tail, which is the most convenient distribution for pWCET estimates of real-time programs [3].

Hardware time randomized caches enable an efficient application of MBPTA. They implement random placement and replacement techniques. Currently, one replacement and two placement MBPTA-compliant policies have been proposed:

Conventional Random Replacement (CRR) [8]: makes random eviction choices so that, in the event of a miss in a given set, for a cache with W ways, the probability of a line in that set to be evicted is $1/W$. CRR builds on a pseudo-random number generator (PRNG) with sufficient quality to allow cache conflicts to be truly random.

hash Random Placement (hRP) [7]: uses a parametric hash function whose input includes the memory address to be accessed and a random seed. It produces the (random) set where the address is placed with that random seed. Thus, whether two addresses are placed or not in the same set is a random event. Any two addresses can be placed in the same set with a probability $1/S$, where S is the number of sets. Upon change of the random seed, addresses are randomly and independently mapped into sets.

Random Modulo placement (RM) [4]: Unlike hRP, RM placement preserves the advantages in terms of spatial locality as modulo placement does. In particular, RM prevents conflicts between cache lines close enough in memory, as modulo placement (MOD) does, but still providing random placement as needed by MBPTA. This is achieved by randomly permuting the location of cache lines within a memory segment using a random seed. Upon a random seed change, addresses in a segment are randomly permuted, thus leading to random placement across segments and no conflicts within segments.

However, CRR may produce probabilistic pathological cases with relevant probabilities, and hRP and RM do not provide efficient randomization for multi-level caches.

II. Random cache replacement

A. Proposal

Conventional random replacement (CRR) is the most suitable replacement policy for MBPTA due to its probabilistic nature: replacement choices are random and independent. CRR makes pathological replacement patterns probabilistic rather than systematic, though they can still occur. We propose Random Permutations Replacement (RPR) [2], that limits pathological random replacement scenarios by increasing temporal reuse and enforcing random evictions to occur across all cache ways.

- When accessed data fits in a cache set, they will eventually be placed in different cache lines, thus avoiding potentially long mutual evictions by construction.
- When the number of accessed lines exceeds the size of a set, RPR effects are also positive increasing reuse, though the impact of replacement naturally reduces.

To reach its goals, RPR leverages the concept of random permutations [6].

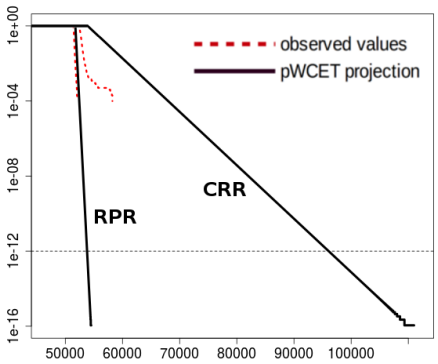


Fig. 2. pWCET curve for `jfdc-tint` Mälardalen benchmark.

B. Results

Figure 2 shows the pWCET distribution when using both CRR and RPR for `jfdctint` Mälardalen Benchmark. Red dotted lines and black straight lines represent the CCDF for the measured data and the pWCET curves respectively. RPR provides increasingly higher gains as the exceedance threshold decreases due to the fact that RPR avoids pathological evictions by construction. Since CRR can produce some such pathological evictions with relevant probability, MBPTA accounts for that by smoothening the shape of the curve and shifting it to the right.

III. Random Cache Placement

A. Proposal

hRP and RM have been proposed for single-level cache systems. For multi-level cache systems, the desired characteristics are the following: (1) **WCET reduction** as the main metric to optimize. (2) **Reduced impact on average performance** due to the importance of this metric for mixed-critical scenarios executing tasks with different criticality levels. (3) **Preserve time composability** by breaking the dependence of WCET estimates on the actual memory addresses used, to favor incremental software development. (4) **MBPTA compliance** to reduce the cost of changing existing timing analysis tools.

Simple solutions consisting of applying just hRP or RM to each cache level do not comply with all characteristics to a sufficient extent. Our proposed solution [1] (shown in Figure 3), uses RM in the L1 and a combination of hRP (at page level) and Modulo (inside each page) in the L2. This design meets all the desired characteristics, while offering a reasonable implementation cost and complexity.

B. Results

We implemented the proposed design in an FPGA and run the EEMBC benchmarks. For all benchmarks we observe (Figure 4) that the pWCET estimate is above the High Water Mark (HWM) (as expected) and for all benchmarks but one the pWCET estimate with hRP+MOD is below HWM+20% (industrial practice for deterministic techniques) obtained for MOD. Hence, hRP+MOD helps reducing WCET estimates w.r.t. current practice while increasing the confidence on estimates w.r.t. just increasing the HWM by a fudge factor of 20%. On average, pWCET estimates are only 8% above the HWM (12 percentage points below HWM+20%).

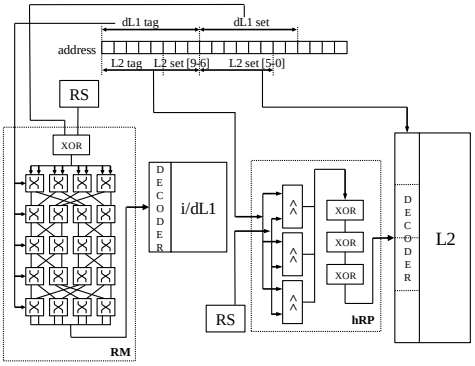


Fig. 3. Multi-level random cache design.

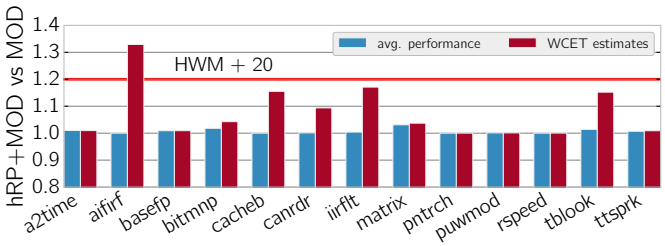


Fig. 4. MOD average results (left) and pWCET hRP+MOD results (right).

IV. Conclusions and Future Work

Existing time-randomized caches help CRTES achieve higher performance while also making them time analyzable. However, they are not sufficiently efficient for multi-level caches and do not avoid pathological cases. Hence, we proposed new MBPTA-compliant placement and replacement policies overcoming those limitations.

Our future work consists of enabling the use of further high-performance computing (HPC) hardware solutions for CRTES in the context of MBPTA. For instance, our next target consists of enabling data prefetching in CRTES.

References

[1] P. Benedicte, C. Hernandez, J. Abella, and F. J. Cazorla. Design and integration of hierarchical-placement multi-level caches for real-time systems. In *DATE*, 2018.
[2] P. Benedicte, C. Hernandez, J. Abella, and F. J. Cazorla. Rpr: A random replacement policy with limited pathological replacements. In *SAC*, 2018.
[3] L. Cucu-Grosjean et al. Measurement-based probabilistic timing analysis for multi-path programs. In *ECRTS*, 2012.
[4] C. Hernandez et al. Random modulo: a new processor cache design for real-time critical systems. In *DAC*, 2016.
[5] IOS. *ISO/DIS 26262. Road Vehicles – Functional Safety*, 2009.
[6] J. Jalle, L. Kosmidis, J. Abella, E. Quinones, and F. Cazorla. Bus designs for time-probabilistic multicore processors. In *DATE*, 2014.
[7] L. Kosmidis et al. A cache design for probabilistically analysable real-time systems. In *DATE*, 2013.
[8] L. Kosmidis et al. Probabilistic timing analysis on conventional cache designs. In *DATE*, 2013.
[9] S. Kotz et al. *Extreme value distributions: theory and applications*. World Scientific, 2000.

Pedro Benedicte is a PhD. Student for the CAOS group at BSC. He obtained his M.S. degree in 2016 and graduated in Informatics Engineering in 2014, both titles obtained from the Universitat Politècnica de Catalunya. His current research focuses on the computer architecture of Real-Time Systems, more specifically, on increasing guaranteed performance in those systems by using techniques commonly found in High Performance Computers as well as Probabilistic Timing Analysis.

Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks

Víctor Campos*, Brendan Jou†, Xavier Giró-i-Nieto‡, Jordi Torres*, Shih-Fu Chang§
*Barcelona Supercomputing Center, †Google Inc, ‡Universitat Politècnica de Catalunya, §Columbia University
{victor.campos, jordi.torres}@bsc.es, bjou@google.com,
xavier.giro@upc.edu, shih.fu.chang@columbia.edu

***Abstract*—Recurrent Neural Networks (RNNs) continue to show outstanding performance in sequence modeling tasks. However, training RNNs on long sequences often face challenges like slow inference, vanishing gradients and difficulty in capturing long term dependencies. In backpropagation through time settings, these issues are tightly coupled with the large, sequential computational graph resulting from unfolding the RNN in time. We introduce the Skip RNN model which extends existing RNN models by learning to skip state updates and shortens the effective size of the computational graph. This model can also be encouraged to perform fewer state updates through a budget constraint. We evaluate the proposed model on various tasks and show how it can reduce the number of required RNN updates while preserving, and sometimes even improving, the performance of the baseline RNN models. Source code is publicly available at https://imatge-upc.github.io/skiprnn-2017-telecombcn/.**

***Keywords*—Deep Learning, Recurrent Neural Networks, Adaptive Computation.**

I. Introduction

Some of the main limitations of Recurrent Neural Networks (RNNs) are their challenging training and deployment when dealing with long sequences, due to their inherently sequential behaviour. These challenges include throughput degradation, slower convergence during training and memory leakage, even for gated architectures [10]. The main contribution of this work is Skip RNN, a novel modification for existing RNN architectures that allows them to skip state updates, decreasing the number of sequential operations to be performed, without requiring any additional supervision signal. The proposed modification is implemented on top of well known RNN architectures, namely LSTM and GRU, and the resulting models show promising results in a series of sequence modeling tasks.

II. Model Description

An RNN takes an input sequence $\mathbf{x} = (x_1, \dots, x_T)$ and generates a state sequence $\mathbf{s} = (s_1, \dots, s_T)$ by iteratively applying a parametric state transition model S from $t = 1$ to T:

$$s_t = S(s_{t-1}, x_t) \tag{1}$$

We augment the network with a binary *state update gate*, $u_t \in \{0, 1\}$, selecting whether the state of the RNN will be updated or copied from the previous time step. At every time step t, the probability $\tilde{u}_{t+1} \in [0, 1]$ of performing a state update at $t + 1$ is emitted. The model formulation implements the observation that the likelihood of requesting a new input increases with the number of consecutively skipped samples:

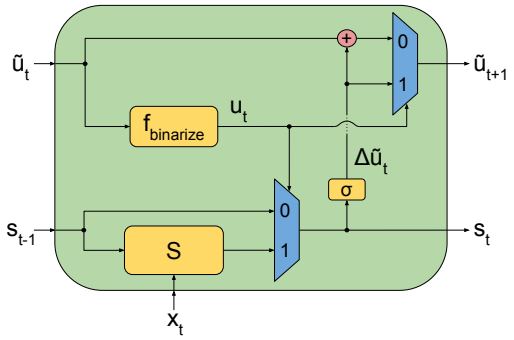


Fig. 1. Model architecture of the proposed Skip RNN, where the computation graph at time step t is conditioned on u_t. In practice, redundant computation is avoided by propagating $\Delta\tilde{u}_t$ between time steps when $u_t = 0$.

$$u_t = f_{binarize}(\tilde{u}_t) \tag{2}$$
$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1} \tag{3}$$
$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p) \tag{4}$$
$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t)) \tag{5}$$

where σ is the sigmoid function and $f_{binarize} : [0, 1] \rightarrow \{0, 1\}$ binarizes the input value. We implement $f_{binarize}$ as a deterministic step function $u_t = \text{round}(\tilde{u}_t)$ and use the straight-through estimator [5] to propagate gradients through it. The number of skipped time steps can be computed ahead of time, enabling more efficient implementations where no computation at all is performed whenever $u_t = 0$

There are several advantages in reducing the number of RNN updates. From the computational standpoint, fewer updates translates into fewer required sequential operations to process an input signal, leading to faster inference and reduced energy consumption. Unlike some other models that aim to reduce the average number of operations per step [10], [6], ours enables skipping steps completely. Replacing RNN updates with copy operations increases the memory of the network and its ability to model long term dependencies even for gated units, since the exponential memory decay observed in LSTM and GRU [10] is alleviated. During training, gradients are propagated through fewer updating time steps, providing faster convergence in some tasks involving long sequences. Moreover, the proposed model is orthogonal to recent advances in RNNs and could be used in conjunction with such techniques,

Model	Accuracy	State updates
LSTM	0.910 ± 0.045	784.00 ± 0.00
LSTM ($p_{skip} = 0.5$)	0.893 ± 0.003	392.03 ± 0.05
Skip LSTM, $\lambda = 10^{-4}$	0.973 ± 0.002	379.38 ± 33.09
GRU	0.968 ± 0.013	784.00 ± 0.00
GRU ($p_{skip} = 0.5$)	0.912 ± 0.004	391.86 ± 0.14
Skip GRU, $\lambda = 10^{-4}$	0.976 ± 0.003	392.62 ± 26.48

TABLE I. ACCURACY AND USED SAMPLES ON THE TEST SET OF MNIST. RESULTS ARE DISPLAYED AS $mean \pm std$ OVER FOUR DIFFERENT RUNS.

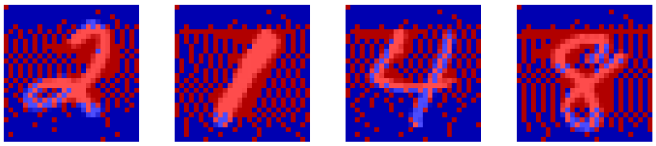


Fig. 2. Sample usage examples for the Skip LSTM with $\lambda = 10^{-4}$ on the test set of MNIST. Red pixels are used, whereas blue ones are skipped.

e.g. normalization [3], [1], regularization [12], [7], variable computation [6], [10] or even external memory [4], [11].

Skip RNN is able to learn when to update or copy the state without explicit information about which samples are useful to solve the task at hand. However, a different operating point on the trade-off between performance and number of processed samples may be required depending on the application, e.g. one may be willing to sacrifice a few accuracy points in order to run faster on machines with low computational power, or to reduce energy impact on portable devices. The proposed model can be encouraged to perform fewer state updates through additional loss terms:

$$L_{budget} = \lambda \cdot \sum_{t=1}^{T} u_t \tag{6}$$

where L_{budget} is the cost associated to a single sequence, λ is the cost per sample and T is the sequence length.

III. EXPERIMENTS: SEQUENTIAL MNIST

The MNIST handwritten digits classification benchmark [9] is traditionally addressed with Convolutional Neural Networks (CNNs) that can efficiently exploit spatial dependencies through weight sharing. By flattening the 28×28 images into 784-d vectors, however, it can be reformulated as a challenging task for RNNs where long term dependencies need to be leveraged [8]. With the goal of studying the effect of skipping state updates on the learning capability of the networks, we introduce a new baseline which skips a state update with probability p_{skip}. We tune the skipping probability to obtain models that perform a similar number of state updates to the Skip RNN models.

Results in Table I show that Skip RNNs solve the task using fewer updates than their counterparts while also showing a lower variation among runs and train faster. We hypothesize that skipping updates make the Skip RNNs work on shorter sub-sequences, simplifying the optimization process and allowing the networks to capture long term dependencies more easily. However, the drop in performance observed in the models where the state updates are skipped randomly suggests that learning which samples to use is a key component in the performance of Skip RNN. Examples such as the ones depicted in Figure 2 show how the model learns to skip pixels that are not discriminative, such as the padding regions in the top and bottom of images, and the attended samples vary depending on the particular input being given to the network.

ACKNOWLEDGMENT

This work has been accepted as a conference paper at ICLR 2018, and we refer the reader to the full publication for extended experiments and results [2]. This research was partially supported by the Spanish Ministry of Economy and Competitivity and the European Regional Development Fund (ERDF) under contracts TEC2016-75976-R and TIN2015-65316-P, by the BSC-CNS Severo Ochoa program SEV-2015-0493, and grant 2014-SGR-1051 by the Catalan Government. Víctor Campos was supported by Obra Social "la Caixa" through La Caixa-Severo Ochoa International Doctoral Fellowship program. We would also like to thank the technical support team at the Barcelona Supercomputing Center.

REFERENCES

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] V. Campos, B. Jou, X. Giró-i Nieto, J. Torres, and S.-F. Chang. Skip rnn: Learning to skip state updates in recurrent neural networks. In *International Conference on Learning Representations*, 2018.

[3] T. Cooijmans, N. Ballas, C. Laurent, Ç. Gülçehre, and A. Courville. Recurrent batch normalization. In *ICLR*, 2017.

[4] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[5] G. Hinton. Neural networks for machine learning. Coursera video lectures, 2012.

[6] Y. Jernite, E. Grave, A. Joulin, and T. Mikolov. Variable computation in recurrent neural networks. In *ICLR*, 2017.

[7] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio, H. Larochelle, A. Courville, et al. Zoneout: Regularizing rnns by randomly preserving hidden activations. In *ICLR*, 2017.

[8] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[10] D. Neil, M. Pfeiffer, and S. Liu. Phased LSTM: accelerating recurrent network training for long or event-based sequences. In *NIPS*, 2016.

[11] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[12] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. In *ICLR*, 2015.

Víctor Campos holds a BsC and a MsC degrees on Electrical Engineering from Universitat Politècnica de Catalunya. He is currently pursuing his PhD on the intersection between Deep Learning and High Performance Computing at the Barcelona Supercomputing Center, supported by Obra Social "la Caixa" through La Caixa-Severo Ochoa International Doctoral Fellowship program. His research interests focus on large scale machine learning.

Application of the edge-based finite element method for fusion plasma simulations

Marc Fuster†, Octavio Castillo†, Shimpei Futatani‡
†Barcelona Supercomputing Center (BSC)
‡Universitat Politècnica de Catalunya (UPC)
E-mail: marc.fuster@bsc.es, shimpei.futatani@upc.edu

Keywords—Nuclear Fusion, Finite Element Method, Nédélec elements, PETGEM.

I. Extended Abstract

Fusion is a clean energy source which shows promise as a future nuclear energy resource. One of the ideas of the nuclear fusion on Earth is that the very high temperature ionized particles forming a plasma can be controlled by a magnetic field, called magnetically confined plasma. This is essential, because no material can be sustained against the high temperature reached in a fusion reactor. In the ideal case, the plasmas in such reactors would remain well confined within the magnetic field in order to allow their core to reach the temperature needed for thermonuclear fusion. Unfortunately such a quiescent confined state is, in general, not observed, and both turbulent small scale motion and collective bulk motion lead to complicated dynamics that cannot be computed analytically and require numerical approaches. The goal of the work is to develop a useful computational tool for fusion applications using the infrastructures given by the PETGEM code [1] which is based on edge elements, a preferable numerical scheme for electromagnetic physics including plasma physics rather than the nodal element approach. The goal of the work is to develop an user-friendly code based on Python which is one of the remarkable growing major programming languages so that the code can be applicable for industrial applications.

The abstract introduces PETGEM code in Section 2. Section 3 shows the achievement of the work such as the mesh generation and the implementation of the initial profiles for the input of PETGEM. The conclusion and the perspectives of the work are summarised in Section 4.

II. PETGEM: based on Edge Finite Element Method (EFEM)

The Parallel Edge-based Tool for Geophysical Electromagnetic Modelling (PETGEM) is a Python HPC scalable tool based on Nédélec Finite Element Method. This code has been developed as open-source (under GPLv3 license) at Computer Applications in Science & Engineering (CASE) of the Barcelona Supercomputing Center (BSC). PETGEM is aimed to solve the marine Controlled-Source Electromagnetic method (CSEM) which is an important technique for reducing ambiguities in data interpretation for hydrocarbon exploration. So as to solve CSEM, one must solve Maxwell's equations:

$$\nabla \times \mathbf{E} = iw\mu_0 \mathbf{H} \qquad \nabla \times \mathbf{H} = \mathbf{J}_s + \sigma \mathbf{E} \tag{1}$$

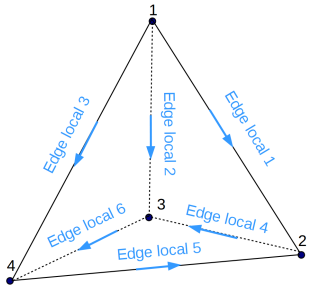


Fig. 1. Tetrahedron discretization.

where ω is the frequency and σ the electric conductivity. Using a perturbation approach ($\mathbf{E} = \mathbf{E}_s + \mathbf{E}_p$ and $\sigma = \sigma_s + \triangle\sigma$) and following the work of [2] the equations to solve become a single equation:

$$\nabla \times \nabla \times \mathbf{E}_s + iw\mu_0\sigma\mathbf{E}_s = -iw\mu_0 \triangle\sigma \, \mathbf{E}_p \tag{2}$$

where $\mathbf{p}$ refers to primary and $\mathbf{s}$ to secondary field. The primary part is the input and the secondary is the output of the PETGEM calculation. In order to obtain $\mathbf{E_s}$ for unstructured meshes, Nédélec Finite Element (a type of edge elements) offers a good balance between accuracy and number of degrees of freedom (DOFs). Nédélec formulation uses vector basis functions defined on the edges of the corresponding elements. As Fig. 1 shows, the discretization method selected is a tetrahedral mesh as these meshes are the easiest to scale-up to very large domains or arbitrary shape. PETGEM is an HPC code due to the fact that Nédélec elements offer a good scalability and it is exploited through the Python Package Petsc4py [3].

III. Results

A. Generation of the finite element mesh

The mesh generation has been performed with GMSH [4] which allows to generate most of kinds of meshes and refine them. The arbitrary function of $\mathbf{E_p}$ is implemented in the generated mesh. Figure 2 shows the generated mesh for finite element method for cylindrical geometry plasma. The initial profile of the electric field $\mathbf{E_p} = -\nabla\phi$ where the electric potential ϕ has been successfully implemented. In the absence of velocity field $\mathbf{u}$ and the simple assumption of the uniform magnetic resistivity η, the profile of the current density $\mathbf{J}$ can be proportional to the electric field through Ohm's law $\mathbf{E} = \eta\mathbf{J} - \mathbf{u} \times \mathbf{B}$ where $\mathbf{B}$ indicates the magnetic field. The cylindrical geometry refers to the mirror plasma confinement

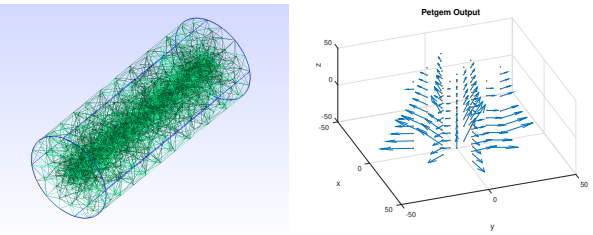


Fig. 2. (Left panel) The generated mesh of the inhomogeneous mesh density for cylindrical geometry. (Right panel) The implemented initial profile of the electric field $E_p = \nabla\phi$ where the electric potential $\phi = e^{-0.02z} \cdot e^{-(x/30)^2-(y/30)^2}$.

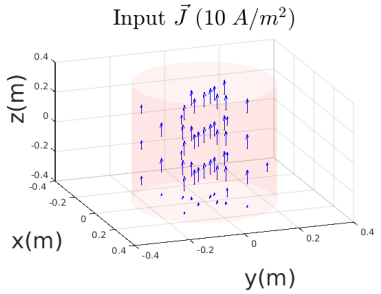


Fig. 3. Initial profile of the antenna current density to be solved by the full wave equation.

devices which are similar with SLPM [5], PANTA [6], etc. The plasma confined in the cylindrical geometry is a useful approach to study the plasma instabilities. The density of the simulation grid can be arbitrarily chosen in the simulation domain. The location of the mesh concentration can be chosen depending on the physics problem to be aimed to study.

B. Steady state plasma

The full wave equation in a magnetically confined plasma with an antenna current as a boundary condition is given by Eq. 3:

$$\begin{aligned} \nabla \times \nabla \times \mathbf{E} - \frac{w^2}{c^2}\mathbf{E} &= \frac{4\pi w i}{c^2}\left(\mathbf{J^p} + \mathbf{J^a}\right) \\ \mathbf{J^p}(\mathbf{r}) &= \int d\mathbf{r}'\sigma(\mathbf{r},\mathbf{r}')\mathbf{E}(\mathbf{r}') \end{aligned} \tag{3}$$

where $\mathbf{J^a}$ is the current of the antenna and σ is the conductivity tensor. This equation models the wave equation with a boundary condition corresponding to the current density introduced by the antenna while Eq. 2 is the perturbation solution of a wave propagating in the earth. Figure 3 shows the initial profile of the current density $\mathbf{J} = \mathbf{J^p} + \mathbf{J^a} = \hat{\mathbf{z}}/f(r)$ where $f(r) = (1 + (r/a)^{2\Lambda})^{1+1/\Lambda}$ and $a = 0.25$m is the radius of the cylinder and $\Lambda = 4$ [7]. The implemented current profile is reasonable approach to compute the initial profile of the current density and the electric field for fusion plasmas [8]. The time variation of the current density will be implemented in the future in order to investigate the interactions between the effect of the antenna and the plasma response.

IV. Conclusions and Perspectives

The work demonstrates the generation of the simulation mesh for finite element method for fusion plasma in the cylindrical geometry. The location and the density of the mesh concentration can be arbitrary adopted according to the physics problem which is aimed to investigate. The implementation of the initial field of any quantities such as electric field and current density, and the application of the reasonable current profile which is specifically aimed for the fusion plasma research have been carried out.

The future work is to solve the full wave equation in the cylindrical geometry which can be solved by current version of PETGEM in order to analyze the precise profile of the electric field and the current density in the steady state plasma. The long-term objective of the work is to go beyond the calculation of the steady state plasma, i.e. implementation of the time integration of the plasma dynamics, for example, magnetohydrodynamic (MHD) which is the combination of the electromagnetism system i.e. Maxwell's equations and the fluid system i.e. Navier-Stokes equation. The primitive approach is to develop the fluid modelling part considering an incompressible ($\nabla \cdot \mathbf{u} = 0$), diffusive model: $\frac{\partial \mathbf{u}}{\partial t} = -\nu \ \nabla \times \nabla \times \mathbf{u}$. is ongoing to be implemented in PETGEM.

References

[1] Octavio Castillo Reyes. Parallel edge-based tool for geophysical electromagnetic modelling (petgem). http://petgem.bsc.es/.

[2] Gregory A. Newman et al. Three-dimensional induction logging problems, part 2: A finite-difference solution. *GEOPHYSICS*, 67.

[3] Lisandro D. Dalcin et. al. Parallel distributed computing using python. *Advances in Water Resources*, 34(9):1124 – 1139, 2011. New Computational Methods and Software Tools.

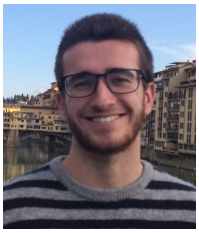
[4] Christophe Geuzaine and Jean-François Remacle. Gmsh. http://gmsh.info/.

[5] F. Castellanos et al. Parallel flows and turbulence in a linear plasma machine. *Plasma Physics and Controlled Fusion*, 47(11):2067, 2005.

[6] Inagaki et al. A concept of cross-ferroic plasma turbulence. 6:22189, 02 2016.

[7] X Shan and D Montgomery. On the role of the hartmann number in magnetohydrodynamic activity. *Plasma Physics and Controlled Fusion*, 35(5):619, 1993.

[8] S. Futatani J. Morales and W.J.T. Bos. Dynamic equilibria and magnetohydrodynamic instabilities in toroidal plasmas with non-uniform transport coefficients. *Physics of Plasmas*, 22:052503.



Marc Fuster Rullan is a Physics bachelor student at UAB. He is at his fourth year and is currently carrying out his bachelor thesis at the Fusion group at the Barcelona Supercomputing Center (BSC). He has won a gold medal at the international physics contest UPHYSICS. His actual work focuses on the computational modelling for fusion. More specifically, he is applying the Edge Finite Element Method code PETGEM to plasma physics.

Accelerating binding free energy calculations by combining Monte Carlo simulations, enhanced sampling and Markov State Models

Joan F. Gilabert*, Victor Guallar*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
E-mail: {joan.gilabert, victor.guallar}@bsc.es

Keywords—AdaptivePELE, PELE, Markov State Models, free energy calculations.

I. Extended Abstract

Computational approaches to the estimation of binding affinities have received a great deal of attention for their potential impact in the drug discovery process. Despite notorious advances and positive results, the current methods still present essential limitations, mainly the difficulty of obtaining thermodynamic sampling in highly-dimensional systems such as protein-ligand energy landscapes.

Biomolecular simulations of atomistic resolution are typically separated into two groups, molecular dynamics (MD) and Monte Carlo (MC) algorithms. Both techniques rely on a force field, a molecular mechanics model of the biomolecule that takes into account many possible interactions, such as electrostatic, bonded or Van Der Waals forces. MD integrates Newton's equations of motion for each atom to obtain the time evolution of the system, while MC techniques apply random movements to sample the energy landscape. Theoretically, MC is expected to generate a higher variety in the obtained conformations, however, the difficulty in generating uncorrelated structures of proteins makes this theoretical expected advantage vanish.

In our group, we developed the Protein Energy Landscape Exploration (PELE)[1], a method that combines Monte Carlo sampling with protein structure prediction techniques to attempt to retain some sampling advantage. Nevertheless, such trajectories exhibit metastability, due to the ruggedness of the energy landscape. To overcome this limitation we have developed an enhanced sampling method, called AdaptivePELE[2]. This method performs several rounds of simulations, combined with clustering and reinforcement learning techniques.

The use of AdaptivePELE drastically improves the efficiency of our simulations, the same reason for this speed-up becomes a hindrance when estimating thermodynamic properties. Enhanced sampling methods introduce a bias and distort the sampled landscape, thus special care has to be taken when performing thermodynamic calculation. Several works have been introduced in this direction[3][4][5], but despite showing great promise, the routine application of such methods is still far from our reach.

Due to the elevated dimensionality of the systems under consideration, many analysis techniques used for molecular simulations are based on some kind of dimensionality reduction, among such methods the use of Markov State Models[6] (MSM) has seen a swift increase in recent years, due to important advances in both its theoretical formulation and its usability. MSMs allow for the extraction of equilibrium properties from simulations of moderate lengths, being particularly suited for the analysis of simulations run in parallel computing setups.

Our work is based on the combination of the three techniques presented here: AdaptivePELE, PELE and MSM. We start with an AdaptivePELE simulation, to quickly map the system landscape, we then cluster the results of the simulation to obtain a few representative structures (approximately 40) to run a longer standard PELE simulation, which will provide a more through sampling. From the PELE simulation we build an MSM, which will give us an estimation of the probability of each state. This probability is used to build a potential of mean force (PMF), g, according to

$$g_i = -k_b T \ln\left(\frac{\pi_i}{V_i}\right) \tag{1}$$

where k_b is the Boltzmann constant, T the temperature, π_i is the probability of state i and V_i the volume of said state. From the PMF, the free energy is calculated as

$$\Delta G^o_{bind} = \Delta W - k_b T \ln\left(\frac{V_b}{V_o}\right) \tag{2}$$

where ΔG^o_{bind} is the binding free energy, ΔW is the depth of the PMF, V_b is the binding volume and V_o is the standard volume, defined in equations 3 and 4 respectively.

$$V_b = \sum_i V_i \exp^{-\frac{g_i}{k_b T}} \tag{3}$$

$$V_o = 1661 \text{ Å}^3 \tag{4}$$

The combination of the three methodologies allows us to obtain a quick estimation of absolute binding free energies for

different scaffolds, which will hopefully help accelerate the drug discovery process.

REFERENCES

[1] K. W. Borrelli, A. Vitalis, R. Alcantara, and V. Guallar, "PELE: Protein energy landscape exploration. A novel Monte Carlo based technique," *Journal of Chemical Theory and Computation*, vol. 1, no. 6, pp. 1304–1311, 2005.

[2] D. Lecina, J. F. Gilabert, and V. Guallar, "Adaptive simulations, towards interactive protein-ligand modeling," *Scientific Reports*, vol. 7, no. 1, p. 8466, 2017. [Online]. Available: http://www.nature.com/articles/s41598-017-08445-5

[3] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics," *Proceedings of the National Academy of Sciences*, vol. 113, no. 23, pp. E3221–E3230, 2016. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1525092113

[4] N. V. Buchete and G. Hummer, "Peptide folding kinetics from replica exchange molecular dynamics," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 77, no. 3, pp. 1–4, 2008.

[5] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, "Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations," *Journal of Chemical Theory and Computation*, p. acs.jctc.7b00373, 2017. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00373

[6] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics," *Current Opinion in Structural Biology*, vol. 25, pp. 135–144, 2014.

Joan F. Gilabert received his BSc degree in Physics Engineering from the Universitat Politècnia de Catalunya (UPC), Barcelona in 2015. He completed his MSc degree in Bionformatics for Health Sciences from Universitat Pompeu Fabra (UPF), Barcelona in 2017. Currently, he is a PhD student of the Computational and Applied Physics of Universitat Politècnica de Catalunya with the Electronic and Atomic Protein Modelling group of Barcelona Supercomputing Center (BSC).

A Machine Learning Workflow for Hurricane Prediction

Albert Kahira*†, Leonardo Bautista Gomez*, Rosa M Badia*,†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {albert.kahira, leonardo.bautista, rosa.m.badia}@bsc.es

Keywords—Machine Learning, Deep Neural Networks, Hurricane

I. Extended Abstract

The Atlantic hurricane season runs from June 1st to November causing massive destruction and loss of life. In 2017, 17 named storms hit the Atlantic causing destruction worth an estimated $316 million and at least 464 fatalities. Meteorologists, by studying previous weather data, predict the expected number of hurricanes in the season. These predictions help authorities prepare for disasters and over the years, better predictions have minimized loss of life and property. However, these predictions rely on human expertise and are often extremely complex due to the thousands of parameters involved and the chaotic nature of weather.

We propose and implement a machine learning model based on deep neural networks to predict the number of hurricanes in the hurricane season. We train the model with more than 100 years of climate data and test it with 5 years. Early results achieve an accuracy of 73% in predicting the number of hurricanes.

A. Background and Motivation

Machine learning has the ability to understand complex models and relationships in data. Recent developments in deep learning models such as Deep Neural Networks (DNN) have led to significant achievements in accuracy [1]. We introduce machine learning model to hurricane prediction to explore the complex relationship between multiple factors such as sea surface temperature, sea level pressure, sea ice cover and wind patterns. We aim to apply a deep learning model to understand the effect of these parameters in the hurricane season and the number of hurricanes. Such insights could significantly improve disaster preparedness and give authorities a better picture of what to expect in the hurricane season.

Previous attempts to use DNN in climate study have been very promising. Liu et al [2] used deep neural networks to detect extreme climate in weather datasets. Zhang et al [3] also used Long Term Short Term memory(LSTM) networks to predict sea surface temperatures. Other studies such as [4] and [5] have also implemented Machine Learning for climate study. However, there has been no studies to predict the number of hurricanes in the hurricane season using Machine Learning.

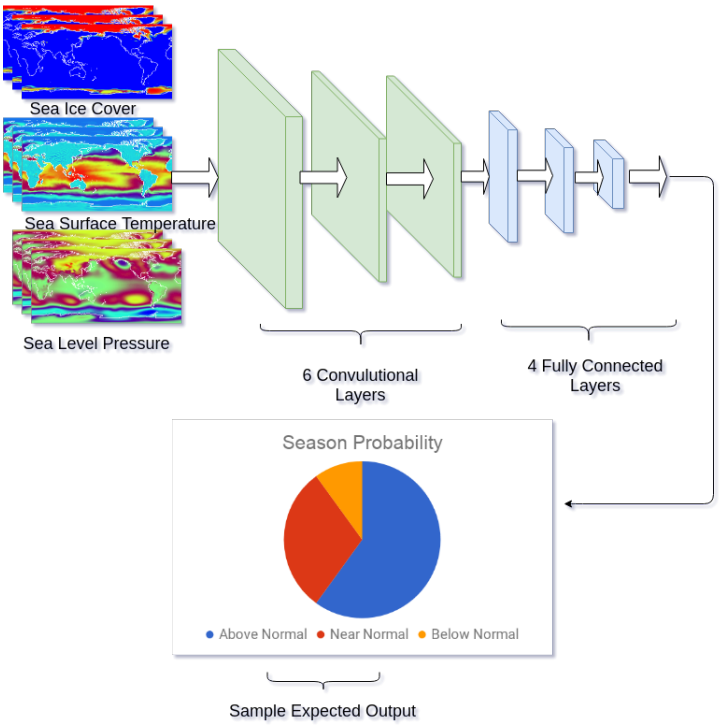


Fig. 1. Work flow

B. Objective

The aim of this study is to introduce Machine Learning to hurricane prediction. Recent scientific advancements have seen geostationary satellites capable of collecting tens of Terabytes of daily data of the weather. On the other side, machine learning models propose efficient techniques to analyse such large data and extract meaningful information. With this large amount of data and the power of high performance computing, machine learning could be an alternative tool for climate study. Furthermore this research aims at introducing approximate computing computing methods to reduce the computational infrastructure generally required for huge amounts of data.

C. Methodology

Monthly averages of 6 weather variables(sea surface temperature, mean sea level pressure, sea ice cover, 2 metre pressure, U wind speed and V wind speed) from 1901 to 2010 are provided by the earth science department of Barcelona Supercomputing Center. Domain expertise shows that these are

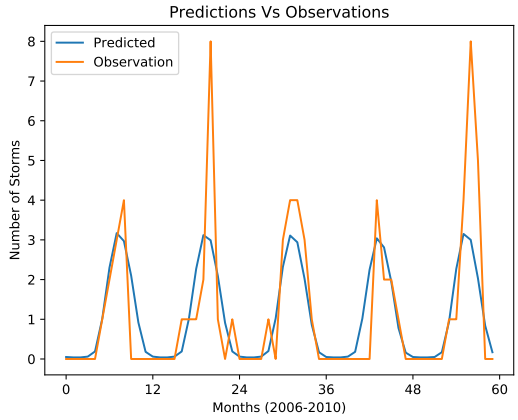


Fig. 2. Predictions for 5 years

the main determinants of the nature and intensity of hurricane season. The total number of named storms for each month in the years 1901 to 210 is also provided which is used as the label for our regression model.

We design a deep learning model with 6 convolutional layers and 4 fully connected layers. Convolutional Neural Networks (CNN) are chosen because of the grid nature of the data. The grid is has the shape (160,320) and therefore the input the layer has the shape (160,320,6). There is a Max Pooling after every 2 convolutional layers and a Dropout layer after every fully connected layer. The Convolutional layers have 6 channels, which are the 6 weather variables. The neural network architecture is summarised in figure 1.

We train our model on single node in MareNostrum4 (48 cores) with 1320 training samples split into training, validation and testing. Initial experiments were aimed at finding the correlation between different weather variables and the number of hurricanes. As such, the initial model contained 1 channel. We trained the model with each channel separately and compared the results. Further experiments were aimed at finding the exact part of grid with most effect on the hurricane season, hence, we crop out some parts of the grid and compare the the results to those of the original grid.

D. Early Results

Using historical data, we train a DNN to classify the hurricane season based on the number of hurricanes likely to occur. Our initial experiments are aimed at finding the most significant factors in storm formation. Preliminary results show that Sea surface temperature has the highest impact on the prediction of the number of storms. Furthermore, given the average sea surface temperatures in a month, our DNN model is able to predict the number of storms with about 60% accuracy. Figure 2 shows the predictions made by our model for a 5 year period.

Our future goal is to develop a complete end to end work flow to continuously learn weather patterns that affect the hurricane season and accordingly, to make predictions.

E. Conclusion

Early results showed a strong relationship between sea surface temperature and the number of storms. Furthermore, cropping the data grid to eliminate land masses and clean up the data have shown improvements with improvements in accuracy to 73%.

F. Future Work

We plan to implement distributed learning using PyCOMPSs (a programming model and runtime which aims to ease the development of parallel applications for distributed infrastructures, such as Clusters and Clouds) to reduce or eliminate the need to expensive computational infrastructure in climate science.

II. Acknowledgment

The authors would like to thank Dr. Alicia Sanchez Lorente and Dr. Louis Philippe Caron from the Earth Science department of Barcelona Supercomputing Center for providing us with the datasets and their invaluable support in this project.

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 713673.

Albert Kahira has received financial support through the la Caixa INPhINIT Fellowship Grant for Doctoral studies at Spanish Research Centres of Excellence, la Caixa Banking Foundation, Barcelona, Spain.

T

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[2] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins *et al.*, "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," *arXiv preprint arXiv:1605.01156*, 2016.

[3] W. Zhang, L. Han, J. Sun, H. Guo, and J. Dai, "Application of multi-channel 3d-cube successive convolution network for convective storm nowcasting," *arXiv preprint arXiv:1702.04517*, 2017.

[4] M. B. Richman, L. M. Leslie, H. A. Ramsay, and P. J. Klotzbach, "Reducing tropical cyclone prediction errors using machine learning approaches," *Procedia Computer Science*, vol. 114, pp. 314–323, 2017.

[5] M. Zhao, I. M. Held, and G. A. Vecchi, "Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of sst anomalies," *Monthly Weather Review*, vol. 138, no. 10, pp. 3858–3868, 2010.

Albert Kahira was born in Lamu, Kenya. He received his BSc degree in Computer Engineering in what is now called Erciyes University and his MSc in Computer Engineering from Abdullah Gul University in Turkey. In 2017 he was awarded the la Caixa INPhINIT Fellowship Grant for Doctoral studies at Spanish Research Centres of Excellence. He is now a research student at Barcelona Supercomputing Center where his research focuses on resilience of machine learning work flows and intelligent distributed systems.

Co-Evolution of Morphology and Behavior in Self-Organized Robotic Swarms

Jessica Meyer* and Joachim Hertzberg†
University of Osnabrück, Osnabrück, Germany
E-mail: *jessy.meyer@gmail.com †joachim.hertzberg@uos.de

Keywords—Swarm Robotics, Co-Evolutionary Robotics, Morphological Computation, Artificial Intelligence, Swarm Intelligence, Bio-Inspired Computing, Evolutionary Computation, Genetic Algorithms, Co-Evolution, Self-Reconfiguring Robotics, Self-Assembling Robotics, Modular Robotics.

I. Extended Abstract

THIS research relies on the premise that robots should be able to improve their swarm by co-evolving their bodies and their minds. Small autonomous robots should work as a swarm and, if and when needed, they should be able to physically cooperate in order to better perform the given task - Figure 1; this might be by physically interacting with each other to temporarily form a larger organism [1].

By co-evolving both morphology and controller, the evolutionary process is expected to converge faster than by doing each evolution at separate times, as seen in [2], which came to these conclusions using a single robot in a simulated environment. There is very little work on any kind of co-evolution in the field of swarm robotics, as it will be discussed in the Related Work; therefore, by providing exclusive data, this research will help solidify the scarce knowledge in the area or even contest it, as the field is still not well established and the state of the art could be restrained to specific scenarios. For example, no research could be found about the impact of evolving the robot's body in a swarm, and consequently, specially at the same time as evolving their controllers.

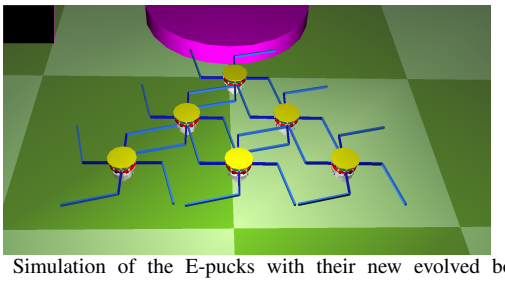


Fig. 1. Simulation of the E-pucks with their new evolved body parts (represented in blue) cooperating to push an object to a determined location.

The controller originally starts with a simple behavior of moving the robots forward, while the interaction between them would be solely due to their morphology. This way, part of the computation that the controller would have to perform is being attributed to the robot's body instead, taking advantage of the morphological computation phenomenon [3], where the controllers tend to be simpler, relying on their physical shapes [4] to compute more sophisticated interactions.

The main two questions that this research raises are whether a trade-off of complexity between morphology and behavior in a swarm of robots exits and if the co-evolution of the morphology and behavior is positive for the swarm.

A. Methodology

The research is being approached using the experimental method. A Genetic Algorithm was developed, where the robots form a population that evolves over time, accordingly to pre-established rules. The algorithm was ran on simulation using Webots, a well-known simulation software. The simulated robots are a direct representation of the E-pucks.

In order to better approach the research hypothesis, the simulation was divided in three stages: the morphological evolution, the behavioral evolution and the co-evolution of both morphology and behavior. In the morphological evolution, arm-like structures are being evolved to improve the robots' bodies. In the behavioral evolution, the conditions that determine the state transitions of a Finite State Machine - FSM - are being evolved to optimize the robots' controllers. In the co-evolution, both arms and transitional conditions are being evolved in order to create the best swarm adapted for the task at hand.

An appropriate task was chosen to test the performance of the robots in a simulated environment specially designed for the situation. As a promising scenario could be search and rescue for example, to start evaluating the robots, a task as simple as pushing a large object forward - Figure 1 - gives an adequate fitness feedback for the evolution to happen.

Following successful co-evolution of the hardware and software of a robotic swarm, the evolutionary process is stopped and the latest robotic generation will be the desired final population of evolved robots, forming an optimal swarm.

In the future, as the robots' controllers and shape specifications can be transferred to their physical bodies - Figure 2, the experiments could be run in a real environment, making use of a 3D printer. Some steps to show its viability were taken and will be presented in the Future Work chapter. With High-Performance Computing (HPC) more available, the processing of the evolution could be run in parallel in simulation, being inspired by Surrogate Models [5], in a continual adaptive process. Controller, morphology and simulation could be co-evolved to address a reality gap between the real world and the simulator [6]. This way, the real robots would be able to adapt to the unforeseen scenarios almost immediately, making them susceptible to evolution in an accelerated pace. It would give the robots a way to predict what could happen and therefore better prepare themselves. Temporal verification techniques for the swarm [7] could also be applied. A combined technique

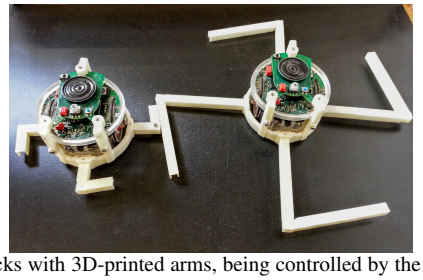


Fig. 2. E-pucks with 3D-printed arms, being controlled by the FSM, showing it is possible to directly transfer the experiment to the real world.

setup like this would greatly benefit and maybe only be possible by having the access to HPC; it could give the robots a sense of ethics as they could simulate the outcome before acting [8], being able to know if their actions would be safe for humans, acquiring what could be called the first steps towards a conscience. The combination of my research with HPC and these new techniques would make the robots as real as they could get while not being biologically alive. They would be able to predict, reflect, adapt, evolve their entire selves, not only based on nature, but faster.

B. Contribution

The direct contribution of this research would be, firstly, new methods and algorithms for hardware/software co-evolution and, secondly, the new types of evolved swarms and their capabilities.

Most importantly, the experiments proposed will gather valuable data in the areas of co-evolutionary robotics and morphological computation, answering some pending questions. Both of these fields are gaining more ground recently and further research is needed to better establish them in the scientific community. Besides the crucial fact that they are both uncharted territory for swarm robotics.

The longer-term impact of the proposed research will be to open up the possibility of robots able to physically evolve and adapt themselves to be able to collectively operate in an unknown or changing environment without human intervention, like for instance in disaster scenarios or planetary exploration.

My research aligned with HPC and 3D printing would enable a real-life robotic evolutionary system that could revolutionize the field. For example, search and rescue robots would be able to be optimized on the go, both physically and behaviorally, giving the victims the best survival chances.

C. Conclusion

It was expected that the robot's morphology would impact on the performance of the swarm, and that the morphology could then be improved through evolution. Given the obtained results, it is shown for the first time that the evolution of the robot's shape can improve the swarm performance for the task of group transport. The robots in their original shape performed worse than the robots with an evolved morphology, independently of the controller's complexity.

The experiments show that the complexity of the controller can be decreased and still achieve good results if there is morphological evolution, thus exploiting the morphological computation phenomena in the transport of objects by multi-robot systems. A more complex controller with a simple morphology does not perform as well as a simpler controller with an evolvable morphology.

It was observed that the evolved robots act as a single organism, they connect with each other almost instantly and combine their forces in the most efficient way, i.e. without creating opposing forces within the swarm. The arms facilitate the connection with the object, making one of the robots touch it, while the other robots connect with one another pushing in unison in the same direction, thus increasing their performance. All of these improvements were solely due to the morphological evolution, with the robots performing better due to the morphological computation. The hypothesized computational savings in the controllers open up possibilities for new improvements in the robot's minds that would not be possible otherwise.

The results from the controller and co-evolution are still being gathered. The swarms that went through the co-evolution seem to be the most successful ones. Since a good controller for a specific shape is not always good for another shape, and vice-versa, evolving both shape and controller concomitantly is not only enabling the best of both to emerge together in a single swarm but, more importantly, they are being tailored to be the best as a whole.

References

[1] P. Levi and S. Kernbach (eds), *Symbiotic Multi-Robot Organisms: Reliability, Adaptability, Evolution.* Springer, 2010.

[2] J. Bongard, "The Impact of Jointly Evolving Robot Morphology and Control on Adaptation Rate." *GECCO*, 2009.

[3] R. Pfeifer and J. Bongard, *How the Body Shapes the Way We Think.* MIT Press, 2007.

[4] J. Bongard, "Taking a biologically inspired approach to the design of autonomous, adaptive machines." *Communications of the ACM*, vol. 59, no. 8, Jun. 2013.

[5] Y. Jin, "Surrogate-assisted evolutionary computation: Recent advances and future challenges." *Swarm and Evolutionary Computation*, vol. 1, pp. 61–70, 2006.

[6] P. O'Dowd, A. Winfield, and M. Studley, "The Distributed Co-Evolution of an Embodied Simulator and Controller for Adaptive Swarm Behaviours." *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011.

[7] C. Dixon, A. Winfield, M. Fisher, and C. Zeng, "Towards Temporal Verification of Swarm Robotic Systems," *Robotics and Autonomous Systems*, no. 60, pp. 1429–1441, Mar. 2012.

[8] A. Winfield, C. Blum, and W. Liu, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection," in *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 85–96.

Jessica Meyer has two BSc's degrees, both obtained in Brazil: one in Computer Science at the Federal University of Bahia and another in System Analysis at the State University of Bahia. Her bachelor theses are about Fuzzy Controllers for a 2D robotic goal-keeper, summa cum laude. She has a MSc's degree in Cognitive Science at the University of Osnabrück, Germany, with focus areas in Artificial Intelligence and Robotics. Her master thesis is about a Mixed Reality Robotics Soccer Team based on Swarms, summa cum laude. Jessica has been involved with RoboCup from 2006 to 2011, in the Soccer Simulation League (2D and Mixed Reality). In 2013 she was involved with the SYMBRION project. She is finishing her PhD in Swarm Robotics at Uni Osnabrück, which first 3 years were at the Bristol Robotics Laboratory, England. Her interests are: Swarm Robotics, Evolutionary Computation, Modular Robotics, Self-Reconfiguring Robotics, Self-Assembling Robotics, Swarm Intelligence, Bio-Inspired Computing.

Evaluation of traffic emission models coupled with a microscopic traffic simulator and on-road measures

Daniel R. Rey*†, Albert Soret*†, Marc Guevara*†, Mari Paz Linares*
*Barcelona Supercomputing Center, Barcelona, Spain
†inLab FIB Universitat Politècnica de Catalunya, Facultad d'informàtica de Barcelona, Barcelona, Spain
E-mail: {daniel.rodriguez, albert.soret, marc.guevara}@bsc.es, mari.paz.linares@upc.edu

***Abstract*—This study aims to compare and contrast the emission results of two instantaneous traffic emission models coupled with a microscopic traffic simulator and COPERT. These will be evaluated with the observed results of on-road measurements done by RSD (Remote Sensing Device) on real driving vehicles in Barcelona. This is done by the comparison of the traffic emission model Panis 2006 which is already integrated into the traffic simulator AIMSUN, and the coupling of AIMSUN to an up to date vehicle emission model, PHEMlight. The studys goal is to assess the divergence between the observed results of the RSD measurements performed at street level, and the modelled emission results in order to evaluate the representativeness of the emission factors applied.**

***Keywords*—Air quality, vehicle emissions, modeling.**

I. Extended Abstract

Air pollution is an important issue for public health, economy and environment. Barcelona is one of the most polluted cities in Europe, and this is directly related with the urban traffic. According to that, air quality measures are everyday more connected with mobility measures (e.g. vehicle restriction, car lanes reduction, increment of parking fees) that aim for the reduction of moving vehicles within the city and the pollution associated to them. To further evaluate and assess the utility of these measures, the coupling of vehicle emission models with traffic simulators have proved to increase the accuracy of emissions [1] but the emission factors used must be as realistic as possible and calibrated to the city and conditions where they are applied. Considering this, the present study compares the NOx (NO2 + NO) emission results of: (I) Panis 2006 [2], (II) PHEMlight [3], (III) the standardized average speed model COPERT IV v10.0 [4] and (IV) COPERT V [5]. Additionally, they were evaluated with observed results of on-road measurements performed with RSD on real driving vehicles in Barcelona [6].

A. Objective

This works goal is to do a primary assessment of the performance of different vehicle emission models by its comparison with observed emission results in Barcelona to be then applied into the air quality integrated system developed during the Ph.D thesis of the author.

B. Methodology

The traffic simulator AIMSUN [7] was used to obtain the representative driving cycle (speed-time data) of a passenger car (PC). The different origin-destination demand matrices needed by AIMSUN as well as the Barcelona network were provided by inLab FIB research centre. The driving cycle of the vehicle studied was introduced into PHEMlight vehicle emission model to obtain its NOx emissions during the whole cycle. These are represented for every time-step of 1.5 seconds in g/h and initially compared to the already coupled but outdated emission model within AIMSUN: Panis 06, and the COPERT IV and COPERT V average speed emission model. Since COPERT works with the cycle average speed, this was set to 28 km/h according to the average speed of the RSD campaign, which was of 28.6 km/h, and to the drive cycle used, whose average speed was of 26.7 km/h. In addition, vehicle degradation factors were applied to COPERT V petrol emission factors according to the vehicle age as stated by EMEP/EEA.

C. Results

The first to notice when looking at the results is the different approach between the instantaneous models (PHEMlight and Panis 06) and COPERT (see Figure 1). While the last calculates the average emissions for the whole cycle, PHEMlight and Panis 06 represent the emission peaks occurred during acceleration periods.

Regarding the average emissions for the whole cycle compared with the RSD observations, there are large differences between petrol and diesel simulated results (see Figure 2). In general, results for diesel are closer to observations for all models, with the largest discrepancies observed in PHEMlight for Euro 5, with an overestimation of 33%, or Panis 06 for Euro 6 vehicles, with an overestimation of 52%. COPERT V agrees well with observations, worth to notice here the difference with

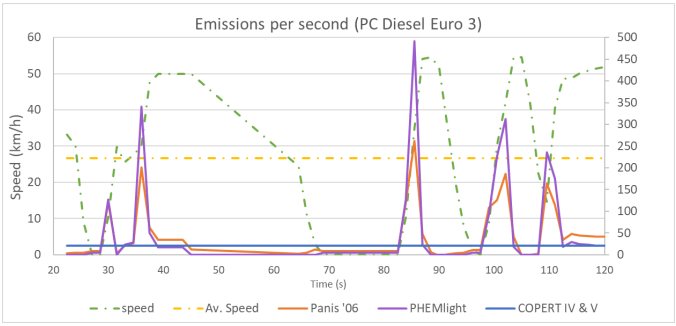


Fig. 1. Speed (km/h) and NOx emissions (g/h) of a passenger car (PC) along time with Panis 06, PHEMlight and COPERT IV and V.

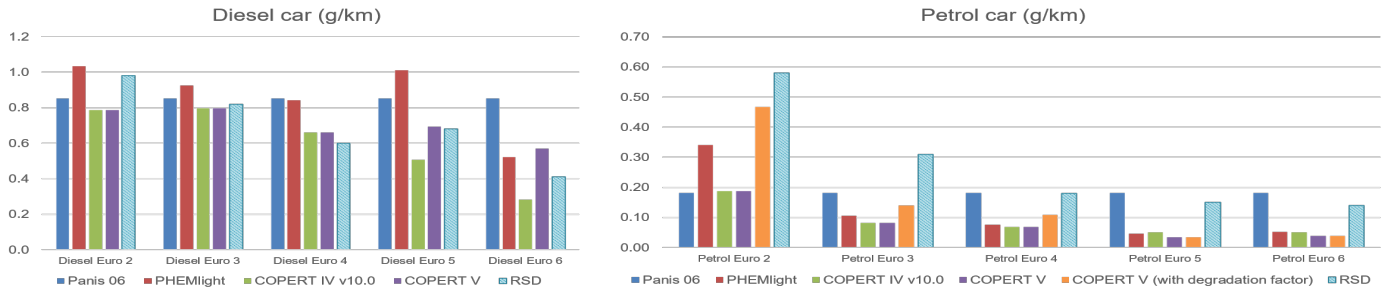


Fig. 2. NOx emissions (g/km) from petrol (left) and diesel (right) estimated by Panis 06, PHEMlight, COPERT IV and V models and the observed average of the RSD study for Euro categories from Euro 2 to Euro 6. Euro vehicles.

COPERT IV for Euro 5 and Euro 6, which underestimates by 34% and 44% respectively.

On the other hand, results for petrol differ more from observations, with large underestimations from all models. In this case the degradation factor applied in COPERT V corrects its results. Applying it, COPERT V underestimates by a factor of 1.2, 2.2 and 1.7 for Euro 2, 3 and 4 respectively. This increases to 3.1, 3.8 and 2.7 without them. For Euro 5 and 6 only Panis 06 agrees with observations, while the rest underestimates by factors of around 4.4 and 3.6.

D. Discussion

Firstly it is noticeable the difference in acceleration peaks that average speed models like COPERT cannot catch. Since for this particular comparison observations were based also on an average speed value, COPERT emission estimations agree reasonably well. Regarding the emissions simulated, diesel values of all models studied agree well with observations, worth to notice the improvement from COPERT IV to COPERT V for Euro 5 and 6 models. However, for petrol there is a large underestimation from all models, with the exception of Panis 06 on Euro 4, 5 and 6. It is outstanding that Panis 06, being such an old model and with its emission factors considering only until Euro 3, agrees so well with the newer petrol vehicle categories, while it largely underestimated the previous ones. It is also worth to stand out the improvement in emission results of the application of vehicle degradation factors to COPERT V.

However, a further study should be made considering observations of instantaneous speed emissions, and not average results, once the data will be available. It is expected then for instantaneous emission models to obtain more accurate results than COPERT V.

II. Acknowledgment

Daniel R. Rey acknowledges the Ministerio de Economa, Industria y Competitividad of Spain for the FPI research grant BES-2016-078116 and RACC institution for the data provided.

References

[1] C. Quaassdorff *et al.*, "Microscale traffic simulation and emission estimation in a heavily trafficked roundabout in madrid (spain)," *Science of the Total Environment*, vol. 566, pp. 416–427, 2016.

[2] L. I. Panis *et al.*, "Modelling instantaneous traffic emission and the influence of traffic speed limits," *Science of the total environment*, vol. 371, no. 1-3, pp. 270–285, 2006.

[3] Technische Universitt Graz, "PHEMlight. User Guide for Version 1," 2017.

[4] EMEP/EEA, "Emission inventory guidebook 2009, updated May 2012," https://www.eea.europa.eu//publications/emep-eea-emission-inventory-guidebook-2009, 2012.

[5] EMEP/EEA, "Air pollutant emission inventory guidebook 2016 Last Update June 2017," https://www.eea.europa.eu/publications/emep-eea-guidebook-2016, 2017.

[6] Area Metropolitana de Barcelona, "Caracterització dels vehicles i les seves emissions a Barcelona i lAMB," 2017.

[7] Transport Simulation Systems, "AIMSUN 8 Dynamic Simulator Users Manual," 2014.

Daniel Rodrguez holds a MSc in Air Pollution Management and Control by the University of Birmingham, and a BSc degree in Chemical Engineering by the UPC. Currently he is doing a Ph.D in Environmental Engineering with Marc Guevara, Albert Soret (from BSC-ES) and Mł Paz Linares (from inLab) in the evaluation of the impact of mobility policies in Barcelonas air quality by the development of an integrated model.

A Unified Memory approach to GPU acceleration on task based programming models

Aimar Rodriguez*, Vicenç Beltran*,
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {aimar.rodriguez, vbeltran}@bsc.es

Keywords—High-performance computing, Programming Models, CUDA, GPGPU.

I. Extended Abstract

Heterogeneous computing has become prevalent as part of High Performance Computing in the last decade, with asynchronous devices such as Graphics Processing Units rapidly advancing. As HPC becomes more specialised and heterogeneous devices improve and develop new features programming models and tools need to adapt in order to keep a competitive performance. In this context, a new version of the OmpSs task based programming model is being developed, which provides an opportunity to introduce the nuances of modern accelerators into the model. In this project, we introduce the implementation of the CUDA GPU programming framework into the OmpSs programming model [1]. The model makes use of the updated *Unified Memory* mechanisms on modern Nvidia GPUs in order to minimise runtime overhead and memory transfer times. This project is developed as part of the Nanos6 runtime project, used for the a new version of the OmpSs programming model.

A. OmpSs-2@CUDA

Graphics Processing Units (GPUs) are used in High Performance Computing (HPC) environments to accelerate the execution of highly parallel workloads. For this, specialised programs called *kernels* are developed, which are then launched using the CPU. A series of characteristics of this form of computation are relevant for the development of a heterogeneous runtime:

- The bulk of the execution is done by the accelerator, while the host CPU is usually only required for setup and launching of the accelerator tasks, leaving the CPU idle most of the time.
- Operations on the GPU are asynchronous, thus, there is no need for the host to be blocked while the accelerator code is running.

Considering this, it is possible to execute accelerator tasks in conjunction with SMP tasks using task based programming models. In order to do this, the runtime can perform a fast operation to launch the asynchronous tasks and use the time in which the GPU is busy to run synchronous CPU tasks.

An implementation of this functionality already exists for the first version of OmpSs [2], however, it present performance issues due to its core design and does not utilise the capabilities introduced in more modern versions of existing GPU programming languages. To solve this, the CUDA support for the new OmpSs runtime, Nanos6 [3], has been redesigned from the ground up, based on the *CUDA Unified Memory* mechanism present on the latest Nvidia GPU architectures.

Launching a CUDA task requires a number of steps:

- Allocating the CUDA memory
- Transfering the data from the host to the CUDA device
- Launching the kernel
- Waiting for the kernel completion
- Copying the data back to the GPU

1) CUDA Streams: Most of these operation are performed on a *CUDA stream*, which is a queue mechanism used to synchronise multiple operations on the CPU. Functions sent to the same stream will be run in the order in which they are invoked, however, there is no guarantee for operations in different streams. Due to this, a usual programming scheme is to launch the copy operations needed for the execution of a kernel before the corresponding kernel in a single stream; this way, it is guaranteed that the kernel will only be run once the transfer operations are completed. This allows to run multiple transfers and kernels in batches without the need for the CPU to intervene beyond an initial launch. This mechanism is used in Nanos6, to launch multiple ready tasks in a batches to different streams.

2) Unified Memory: Three of the five steps involved in launching a CUDA tasks are related to memory management; memory allocation, input transfers and output transfers. In order to leverage this functionality from the runtime and reduce the overhead required to track memory, the *Unified Memory* functionality provided since the CUDA version 6 is used. The UM is a mechanism that provides automatic memory transferences between the host and the GPUs. In the latest Nvidia device architectures this is done with a page fault mechanism, which moves data between CPU and GPU when a page fault is triggered. With this, there is no need to implement memory management in the Nanos6 runtime, and thus the overhead it carries to the execution is reduced. The cost for this is that the memory transferences will be slower due to the page faulting mechanism and that the system will only provide CUDA support for Pascal or more modern architectures.

3) Task synchronisation: In order to detect the finalisation of asynchronous tasks an event polling method is used. To

check for task completions, a CUDA event is recorded in the stream used for the execution of the task immediately after the kernel. This way, it is possible to know that the kernel has finished execution once the status of the event changes. Similar to tasks, events are polled by the worker threads before running their CPU tasks, since it requires a small amount of time and, thus, will not add much overhead to the execution of SMP tasks.

While there are other possible synchronisation methods, polling is chosen since it allows to keep the asynchronous nature of accelerator task execution without adding much overhead. Waiting mechanisms are discarded since they require blocking a CPU to wait for the CUDA tasks, and *CUDA callbacks* are not used due to delays from kernel completion to callback execution.

4) Execution Model: The execution model of the CUDA support for Nanos6 uses all the available worker threads of the runtime to launch asynchronous tasks. Before running a CPU tasks, each worker thread will check if there are any available GPU tasks; if any are available, it will proceed to launch as many as possible in batch before continuing with the SMP task execution. In the same manner, before launching any CUDA task, it will check if any of the currently executing tasks have finished their execution, in order to mark their finalisation and release their dependencies on the runtime.

In addition an additional helper thread exists whose sole task is to launch and check for CUDA task finalisation. This thread runs a loop at a low frequency only performing asynchronous tasks in order to avoid using high CPU time. The reason for the existence of this thread is to avoid starvation of CUDA tasks when all the worker threads are busy executing long running SMP tasks, thus, a minimal service is provided for the GPU tasks as a fallback.

B. Experimentation

The performance of the runtime has been tested using various benchmarks on a single node of CTE-POWER cluster, whose characteristics are the following:

- 2x IBM PowerNV 8335-GTB @ 4.00GHz (10 cores and 8 threads/core, total 160 threads per node)
- 2x nVidia Pascal P100 GPU with 16GB of memory.

A series of benchmarks obtained from the BSC application repository and the Rodinia benchmark suite have been run using both OmpSs and OmpSs-2. A summary of the results can be seen on figure 1

The results of the execution show varying results on different application; both performance losses and improvements have been observed, as well as applications which are not largely affected by the runtime used. Further analysis has shown that the needs for computation or memory transfers on the application determine this.

Compute intensive applications which do not make high usage of memory see speedups when using the new CUDA support at the Nanos6 runtime. This is due to the reduced runtime overhead present on the runtime due to the choice of using Unified Memory and eliminating the need to manually track memory. On the other hand memory intensive applications see their performance lowered, since the Unified Memory mechanism is slower than regular memory transferences when moving data between devices.

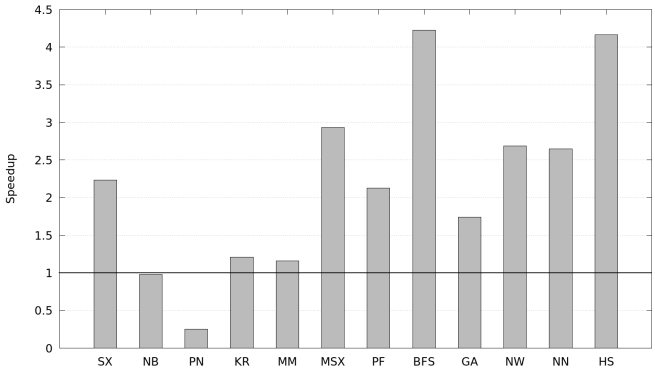


Fig. 1. Speedup evaluation of Nanos6 compared to the previous version

C. Conclusion

Finally, we conclude that the new CUDA support system offers opportunities for performance improvements over the old version. While the Unified Memory mechanism does have drawbacks regarding memory intensive applications and running on older systems, it shows speedups on certain application types. Additional optimizations can be applied to the system, such as improved scheduling and usage of CUDA hints to improve memory times, however, it is also possible to explore the performance of other design choices, such as the usage of streams in way that allows for batch asynchronous task execution.

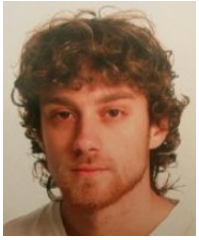
Future work will be focused on implementing a system with manual memory management which includes the design choices on the existing system, as well as finding additional optimizations and exploring different application types and how their performance is affected by the choice of runtime.

REFERENCES

[1] E. Ayguadé, R. M. Badia, P. Bellens, D. Cabrera, A. Duran, R. Ferrer, M. González, F. Igual, D. Jiménez-González, J. Labarta *et al.*, "Extending openmp to survive the heterogeneous multi-core era," *International Journal of Parallel Programming*, vol. 38, no. 5-6, pp. 440–459, 2010.

[2] J. Planas Carbonell, "Programming models and scheduling techniques for heterogeneous architectures," Ph.D. dissertation, Universitat Politcnica de Catalunya, 2015.

[3] BSC, "Nanos6 runtime," https://github.com/bsc-pm/nanos6, 2018.



Aimar Rodriguez is a PhD student in Programming Models for Heterogeneous Systems. Aimar studied computer science engineering at the University of Deusto in Bilbao, Spain, obtaining his degree on 2014. After this, he enrolled on the Master in Innovation and Research in Informatics (MIRI) on the *Universitat Politcnica de Catalunya* (UPC) at Barcelona. During his studies, he joined the *Barcelona Supercomputing Center*, where he developed his final master thesis and is currently working on a PhD on the development of Heterogeneous Architecture support for Programming Models.

NOTES:

NOTES:

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Barcelona Supercomputing Center
Jordi Girona, 31 - Torre Girona
08034 Barcelona (Spain)

education@bsc.es
www.bsc.es

/BSCCNS
@BSC_CNS
/BSCCNS
bsc.es/linkedin