3rd BSC International Doctoral Symposium

2016

4th, 5th & 6th May, 2016

Book of abstracts

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Book of Abstracts
3rd BSC International Doctoral Symposium

Editors
Nia Alexandrov
María José García Miraz

Graphic and Cover Design:
Cristian Opi Muro
Laura Bermúdez Guerrero

This book is available at
http://www.bsc.es/doctoral-symposium-2016

published by:
Barcelona Supercomputing Center

supported by:
The "Severo Ochoa Centres of Excellence" programme

3rd Edition, September 2016

ACKNOWLEDGEMENTS

The BSC Education & Training team gratefully acknowledges all the PhD candidates, Postdoc researchers, experts and especially the Keynote Speaker Francisco J. Doblas-Reyes and the tutorial lecturers Vassil Alexandrov and Javier Espinosa, for contributing to this Book of Abstracts and participating in the 3rd BSC International Doctoral Symposium 2016. We also wish to expressly thank the volunteers that supported the organisation of the event: Carles Riera and Felipe Nathan De Oliveira.

BSC Education & Training team
education@bsc.es

CONTENTS

EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 3rd BSC International 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 34 presentations were given, 32 posters were exhibited, a two days training on an Introduction to Scientific Writing was conducted; and we reached more than 90 attendees. Furthermore, we opened the participation to students all over the world and have succeeded enrolling students from different countries

In this third edition of the BSC Doctoral Symposium we have planned a keynote speaker' talk, and two extensive training courses on Algorithms and Techniques for Data Intensive Problems and on Scientific Visualization of Data.

The talks will be held in six different sessions and will tackle the topics of: Postdoc research at BSC; Algorithms, Physics & Data Science Algorithms, Numerical Methods and Data Science; Life Sciences; Simulations and Modelling; and Performance. The posters will be exhibited and presented during four poster sessions that will give the authors the opportunity to explain their research and results.

The keynote speaker prof. Francisco Doblas Reyes will give the lecture: Big Data for the Study of Climate Change and Air Quality. He is Director of the Department of Earth Sciences at BSC and ICREA research professor at the Catalan Institute of Climate (IC3). At BSC he coordinates the largest FP7 project on climate prediction. The Department hosts more than 50 engineers, physicists, mathematicians and other air quality and climate researchers who try to bring the latest developments in supercomputing and Big Data to provide the best information and services. He is author of more than 100 peer-reviewed papers.

The training courses will review the fundamental Algorithms and Techniques for Data and Computationally Intensive Problems and introduce Scientific Visualisation of Data as well as provide some practice regarding these concepts.

This Book of Abstracts follows the order of the programme of the 3rd BSC International Doctoral Symposium, and all the information is arranged according to the order of participation of the students. Information about the participants and accepted extended abstracts are published after a rigorous review process and with permission of the authors.

We hope that this publication is a valuable contribution to the reflection and dissemination of how the usage of high performance computing technologies and resources are key-factors to promote the scientific and social development in many different areas.

BSC Education & Training team
education@bsc.es

WELCOME ADDRESS

I am delighted to welcome all the PhD students, Postdoc researchers, advisors, experts and attendees participating in the 3rd BSC International Doctoral Symposium.

This year edition has consolidated international relevancy and students from different countries and organizations will take part in the symposium. Nevertheless, 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 Francisco Doblas and the invited lecturers Vassil Alexandrov and Javier Espinosa, 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 in your career and I really hope you enjoyed this great opportunity to meet other colleagues and share your experiences.

Dr. Maria Ribera Sancho

Manager of BSC Education & Training

PROGRAM

Day 1 (4th May)

Start time	Activity	Speaker/s
8.30 h Registration		
9.00h	Welcome and opening	Mateo Valero, BSC Director
9.20h	Keynote talk: Big Data for the Study of Climate Change and Air Quality	Francisco Doblas Reyes, Head of Earth Sci Department, BSC
	Abstract: The extraction of a significant message oriented towards the action of a range of users based on the large set of heterogeneous data that the weather, climate and air quality communities produces and has produced is the main challenge of Big Data for these communities. The kind of problems to deal with include the operational nature of many of the activities, which implies sharing the data with very strict schedules, the need to extract information from large heterogeneous datasets by users that in some cases are not aware of the limitations of those data (uncertainty level, covariances, etc), or also the management of large datasets with high levels of documentation, curation and long-term availability. These communities are also special in that they are highly organised and collaborative worldwide and have a strong statistical background, substantial distributed computational power. This presentation will address these characteristics and illustrate them with recent examples.	
10.20h Event Photo		
Coffee break & First Poster Session: Algorithms and Models		
10.40h	Development of a wind energy climate service based on seasonal climate prediction, Veronica Torralba, Earth Science, BSC Assessment of Meteorological Models for Air Pollution Transport: Analysis between Mexico and Puebla Metropolitan Areas, Sergio Natan González Rocha, Earth Science, BSC Enhanced Monte Carlo methods for Sparse Approximate Matrix Inversion, Oscar Esquivel Flores, ITESM & Diego Davila, Computer Science, BSC/ UPC Dynamic Load Balancing for hybrid applications, Marta Garcia Gasulla, Computer Science, BSC Effects of detailed ventricular anatomy on the blood flow, Federica Sacco, DTIC, Universitat Pompeu Fabra	
First Talk Session: Post Doc research		
11.30h	Probabilistic seismic risk assessment using CRISIS2015 & USERISK2015. Application to buildings of Barcelona, Spain.	Armando Aguilar, CASE, BSC
11.50h	DimLightSim: Optical/Electrical Network Simulator for HPC Applications	Hugo Daniel Mayer, Computer Sci., BSC
12:10h	Comparing electoral campaigns by analysing online data	Javier Espinosa, Computer Sci., BSC
12.30h	Integrated approach to assignment, scheduling and routing problems	Laura Hervet- Escobar, ITESM
12.50h Lunch break		
14.00h	Tutorial 1: Algorithms and Techniques for Data and Computationally Intensive Problems	Vassil Alexandrov, Computer Science, BSC
	Abstract: This tutorial will focus on key research methods, algorithms and techniques for Data and Compute intensive problems, ranging from theory creating and theory testing approaches to conceptual-analytical approaches and experimental ones that are able to lead to discovering global properties on data as well as providing efficient ways of parallel computation. These will be mainly deterministic and hybrid (stochastic/deterministic) methods and algorithms including: • Network Science, a graph based approach enabling the discovery of global properties of networks and global properties of data. • Multi-Objective and Multi-Constrained Optimization, a method for efficient classification and optimization. • Monte Carlo and quasi-Monte Carlo methods allowing the design of scalable, fault-tolerant and resilient algorithms for variety of problems that can run efficiently on novel parallel computer architectures.	
16.00h Coffee break		
	Tutorial 1 continues	
18.00h	Adjourn	

Day 2 (5th May)

Start time	Activity	Speaker/s
9.00h	Opening of the second day	
Second Talk Session: Algorithms, Numerical Methods & Data Science		
9.10h	Innovative algorithm for particles transport in a fluid	Edgar Olivares Mañas, CASE, BSC
9.30h	Validating the Reliability of WCET Estimates with MBPTA	Suzana Milutinovic, Computer Science, BSC
9.50h	Efficient and versatile data analytics for deep networks	Jonatan Moreno, Computer Science, BSC
10.10h	Numbering along advection for Gauss-Seidel and Bidiagonal preconditioners	Paula Córdoba Pañella, CASE, BSC
Coffee break & Second Poster Session: Simulations and Modelling		
10.30h	Exploring the protonation properties of photosynthetic phycobiliprotein pigments from molecular modelling and spectral line shapes, Marina Corbella Morató, Computational Biology, UB Clustering the Roman Empire: the use of multivariable analysis to understand cultural dynamics, Maria Coto-Sarmiento, CASE, BSC - UB Assessing drug-protein binding by simulation of stereoselective energy transfer dynamics: electronic interactions between tryptophan and flurbiprofen, Silvana De Souza Pinheiro, Computational Biology, UB Extrapolations of the fusion performance in JET, Dani Gallart Escolà, CASE, BSC Regional Arctic sea ice predictability and prediction on seasonal to inter-annual timescales, Rubén Cruz García, Earth Science, BSC	
Third Talk Session: Life Sciences		
11.20h	Identification of genetic variants associated with risk for a variety of cancers through the re-analysis of publicly available genome-wide data from more than 20,000 individuals	Marta Guindo, Life Sciences, BSC
11.40h	Genomic Instability Promoted by Expression of Human Transposase-Derived Gene	Elias Rodríguez Fos, Life Sciences, BSC
12.00h	SMuFin2: Identification of virus in Cancer genomes using an improved version of SMuFin.	Mercè Planas-Fèlix, Life Sciences, BSC
12.20h	Docking through Democracy Re-ranking protein-protein decoys with a voting system	Didier Barradas-Bautista, Life Sciences, BSC
12.30h Lunch break		
14.00h	Tutorial 2: Scientific Visualisation of Data	Javier Espinosa, Computer Science, BSC
	Abstract: Data visualization has become more important than ever in almost every discipline dealing with data. From creating a visual representation of data points as part of a research experiment, for showcasing progress, or for analysing 3D models, data visualizations are a critical and a valuable tool for gaining insight about data. When it comes to big data, weak tools with basic features get to their limits. In consequence, specific techniques should be developed and applied. This tutorial will address different techniques for visualizing big data collections. It will include an overview of the visualization process as a complex and greedy task and then it will discuss out of the box solutions that can help to analyse and interpret big data in aggregated and analytic ways. The tutorial will have a strong hands-on component for experimenting with the visualization of data collections of different types.	
16.00h Coffee break		
	Tutorial 2 continues	
18.00h	End of the Training and Adjourn	

Day 3 (6th May)

Start time	Activity	Speaker/s
8.50h	Opening of the third day	
Fourth Talk Session: Simulations and Modelling		
9.00h	Block-Based Execution on an Integrated Vector-Scalar In-Order Core	Milan Stanic, Computer Science, BSC
9.20h	Photoprotection and triplet energy transfer in higher plants: the role of electronic and nuclear fluctuations	Lorenzo Cupellini, University of Pisa
9.40h	Modelling the Co-evolution of Trade and Culture	Simon Carrignon, CASE, BSC-UPF
10.00h	Simulating Gravitational Collapse with Arbitrary-Precision Arithmetic	Daniel Santos-Oliván, GWART, Institute of Space Science
10.20h	Crowd Simulation and Visualization	Hugo Perez, Computer Science, BSC
Coffee break & Third Poster Session : Computer Science & Applications		
10.40h	Reproducing crowd turbulence with Verlet integration and agent modeling, Albert Gutierrez-Milla, Dani Gallart Escolà, CASE, BSC Generation of a simulation scenario from medical data: Carto and MRI, Mariña López-Yunta, CASE, BSC How Can We improve Energy Efficiency through User-directed Vectorization and Task-based Parallelization? HelenaCaminal, Computer Science, BSC Using Graph Partitioning to Accelerate Task-Based Parallel Applications, Isaac Sánchez Barrera, Computer Science, BSC Improving Scalability of Task-Based Programs, Iulian Valentin Brumar , Computer Science, BSC	
Fifth Talk Session: Life Sciences		
11.30h	Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses	Morena Pappalardo, Biosciences, University of Kent
11.50h	Enrichment of Virtual Screening results using induced-fit techniques	Jelisa Iglesias, Life Sciences, BSC
12.10h	On the way to real time protein-ligand sampling	Daniel Lecina Casas, Life Sci BSC
12.30h	PMut2: a web-based tool for predicting pathological mutations on proteins	Víctor López Ferrando, Life Sciences, BSC
12.50h Lunch break		
13.50h Visit to MareNostrum III (14.00 in Torre Girona)		
Sixth Talk Session: Performance		
14.40h	Per-Task Energy Metering and Accounting in the Multicore Era	Qixiao Liu, Computer Science, BSC
15.00h	Task Dependences Management Hardware Acceleration for Task-based Dataflow Programming models	Xubin Tan, Computer Science, BSC
15.20h	The OmpSs Reductions Model and How to Deal with Scatter-Updates	Jan Ciesko, Computer Science, BSC
15.40h	Runtime Estimation of Performance–Power in CMPs under QoS Constraints	Rajiv Nishtala, Computer Science, BSC
Coffee break & Fourth Poster Session: Life Science Applications		
16.00h	Conformational landscape of small ligands: A Multilevel strategy to determine the conformational penalty of bioactive ligands, Toni Viayna Gaza, Facultat de Farmàcia i Ciències de l'Alimentació, Santa Coloma de Gramanet, UB Characterization of Protein-Protein Interfaces and Identification of Transient Cavities for its Modulation, Mireia Rosell Oliveras, Life Sciences, BSC Improvement of Protein-Ligand Binding Affinity Prediction using Machine Learning Techniques, Gabriela Hernandez Larios, Life Sciences, BSC - EM-DMKM Towards accurate solvation free energies of large biological systems, Sonia Romero Téllez, Departament de Fisicoquímica, UB	
16.40h	Conclusions	
17.30h	End of the Doctoral Symposium	

KEYNOTE SPEAKER

Francisco Doblas Reyes
Head of Earth Science Department, BSC

Big Data for the Study of Climate Change and Air Quality

The extraction of a significant message oriented towards the action of a range of users based on the large set of heterogeneous data that the weather, climate and air quality communities produces and has produced is the main challenge of Big Data for these communities. The kind of problems to deal with include the operational nature of many of the activities, which implies sharing the data with very strict schedules, the need to extract information from large heterogeneous datasets by users that in some cases are not aware of the limitations of those data (uncertainty level, covariances, etc), or also the management of large datasets with high levels of documentation, curation and long-term availability. These communities are also special in that they are highly organised and collaborative worldwide and have a strong statistical background, substantial distributed computational power. This presentation will address these characteristics and illustrate them with recent examples.

Prof. Francisco J. Doblas-Reyes is an expert in the development of seasonal-to-decadal climate prediction systems. He started working on climate variability at the Universidad Complutense de Madrid (Spain) in 1992, where he did his PhD. He then worked as a postdoc in Météo-France (Toulouse, France), at the Instituto Nacional de Técnica Aerospacial (Torrejón, Spain) and for ten years at the European Centre for Medium-Range Weather Forecasts (Reading, UK). He is now the director of the Department of Earth Sciences of the Barcelona Supercomputing Center (BSC-CNS) and the head of the Climate Forecasting Unit at the Institut Català de Ciències del Clima (IC3), where he leads the largest FP7 project on climate prediction. He is author of more than 100 peer-reviewed papers. He is involved in the development of the EC-Earth ESM since its inception, an IPCC lead author (Fifth Assessment Report), and serves in WCRP and WWRP scientific panels.

TUTORIALS

Vassil Alexandrov

Extreme Computing Group Manager,
Computer Science Department, BSC

Vassil Alexandrov is an ICREA Research Professor in Computational Science at Barcelona Supercomputing Centre and the leader of Extreme Computing Research Group. He is a member of the Editorial Board of the Journal of Computational Science, Guest Editor of Mathematics and Computers in Simulation, Guest Editor of special issue on Scalable Algorithms for Large Scale Problems and Route to Exascale of the Journal of Computational Science. He is one of the founding fathers of the International Academy of Information Technology and Quantitative Management.

His research is in the area of Computational Science encompassing Parallel and High Performance Computing, Scalable Algorithms for advanced Computer Architectures, Monte Carlo methods and algorithms. In particular, scalable Monte Carlo algorithms are developed for Linear Algebra, Computational Finance, Environmental Models, Computational Biology etc. In addition, the research focuses on scalable and fault-tolerant algorithms for petascale architectures and the exascale architecture challenge. He currently leads the Extreme Computing research group at BSC focusing on solving problems with uncertainty on large scale computing systems applying the techniques and methods mentioned above. He has published over 100 papers in renowned refereed journals and international conferences and workshops in the area of his research expertise.

Teaching and Professional Training: Lecturing mainly on topics of Data and Computationally Intensive problems, among which, mathematical methods for discovering global properties on data, advanced algorithms and programing techniques for advanced architectures, HPC and Computational Science research methods for scientific and industrial applications. Professor Alexandrov has been Program Director of the MSc in Network Centred Computing (2000-2011), MSc in Computational Science by Research (2007-2011) and the Erasmus Mundus Joint MSc in Network and e-Business Centred Computing for the period October 2005- June 2011).

Tutorial 1: Algorithms and Techniques for Data and Computationally Intensive Problems

This tutorial will focus on key research methods, algorithms and techniques for Data and Compute intensive problems, ranging from theory creating and theory testing approaches to conceptual-analytical approaches and experimental ones that are able to lead to discovering global properties on data as well as providing efficient ways of parallel computation. These will be mainly deterministic and hybrid (stochastic/deterministic) methods and algorithms including:

- Network Science, a graph based approach enabling the discovery of global properties of networks and global properties of data.
- Multi-Objective and Multi-Constrained Optimization, a method for efficient classification and optimization.
- Monte Carlo and quasi-Monte Carlo methods allowing the design of scalable, fault-tolerant and resilient algorithms for variety of problems that can run efficiently on novel parallel computer architectures.

Javier Espinosa

Computer Science Department, BSC

Javier Espinosa is a postdoctoral research fellow at Barcelona Supercomputing Center (BSC) and member of the French-Mexican Laboratory of Informatics and Automatic Control (LAFMIA). He holds a PhD in Computer Science from University of Grenoble, France.

His research concerns databases and distributed systems. In particular, he is interested on Internet Technologies (e.g., Service-Oriented Architectures, Cloud Computing, Data Services) and NoSQL solutions for modern data management. His objective is to design data management services guided by QoS criteria (e.g., security, reliability, fault tolerance, evolution and dynamic adaptability) and behavioural properties (e.g., transactional execution). He has participated in several national and international projects, where he has been responsible of the execution of working packages and the implementation of prototypes (POLIWEB PEPS CNRS; CASES EU-FP7; S2EUNET FP7-IRSES).

Tutorial 2: Scientific Visualisation of Data

Data visualization has become more important than ever in almost every discipline dealing with data. From creating a visual representation of data points as part of a research experiment, for showcasing progress, or for analysing 3D models, data visualizations are a critical and a valuable tool for gaining insight about data. When it comes to big data, weak tools with basic features get to their limits. In consequence, specific techniques should be developed and applied. This tutorial will address different techniques for visualizing big data collections. It will include an overview of the visualization process as a complex and greedy task and then it will discuss out of the box solutions that can help to analyse and interpret big data in aggregated and analytic ways. The tutorial will have a strong hands-on component for experimenting with the visualization of data collections of different types.

TALK PRESENTERS

Hugo Daniel Meyer

Computer Science, Barcelona Supercomputing Center

Hugo Meyer got his degree in Informatic Engineering in 2010 from the National University of Asunción (Paraguay). He obtained his M.Sc. in High Performance Computing in 2011 and earned his Ph.D. degree in High Performance Computing in 2014, both from the University Autónoma of Barcelona - UAB (Spain). In the Computer Architecture and Operating Systems (CAOS) group of the UAB he worked on Fault Tolerance and Performance Analysis in parallel and distributed systems.

Currently, he is working as a researcher in the Barcelona Supercomputing Center researching on Performance Evaluation, Computer Architecture Design, Optical Networks and HPC-based Simulations.

During his career as a researcher, he obtained twelve publications in well-ranked conferences and journals. His main research interests include the design and development of optimizations techniques for parallel computing applications, metaheuristic optimization techniques, simulation and machine learning algorithms.

Talk: “DimLightSim: Optical/Electrical Network Simulator for HPC Applications”

Armando Aguilar Meléndez

CASE Barcelona Supercomputing Center

He was born in Mexico city. He got his degree in civil engineering at National Autonomous University of Mexico (UNAM). Since then, he has been participating in diverse structural projects of buildings. He got his master degree in engineering (structures) at UNAM.

The thesis of his master degree was about CRISIS99, a program for computing seismic hazard. Since 2004 he has been working as a full time professor at Civil Engineering Faculty of the University of Veracruz. He studied the PhD program of earthquake engineering and structural dynamic at Technical University of Catalonia (UPC). His PhD thesis is called “Probabilistic assessment of the seismic risk of buildings in urban areas”.

He has participated in the development of both codes CRISIS2015 and USERISK2015, which were designed to assess seismic hazard and seismic risk, respectively. Currently, he is participating in a

research related to the assessment of seismic hazard and seismic risk in the CASE department, with the support of a posdoctoral fellowship CONACYT-BSC.

Talk: "Probabilistic seismic risk assessment using CRISIS2015 & USERISK2015. Application to buildings of Barcelona, Spain

Laura Hervert-Escobar

Industrial Engineering and numerical methods, Intituto Tecnologico De Estudios Superiores De Monterrey

Postdoctoral researcher at Instituto Tecnológico de Estudios Superiores de Monterrey, her research include, but are not limited to the development, analysis and implementation of metaheuristic techniques for solving complex real-life combinatorial problems with several objectives.

Currently, the principal emphasis will be the study and development of advanced mathematical methods and algorithms both stochastic and hybrid (stochastic/deterministic) ones for Linear Algebra (Matrix Inversion, Solving Large Systems of Linear Equations, etc.) and multi-objective multi-constrained optimization. Also the modeling of complex systems using stochastic and hybrid approaches as well as Network Science techniques, mainly parallel algorithms and parallel computing.

Her work experience includes the development of different projects including new product launch, design and opening of workshops, business model re-engineering, etc.

Talk: "Integrated Approach to Assignment, Scheduling and Routing Problems"

Edgar Olivares Mañas

CASE, Barcelona Supercomputing Center

Edgar is a PhD student in Barcelona Supercomputer Center (BSC). His research topic is particles transport simulations applied to a High Performance Computing (HPC) scope. Particles transport have several applications in engineering and medicine.

Some of his resarch lines include drug delivery, respiratory system, icing on aeronautics, garbage transport on the ocean or combustion. These types of problems have the particles transport in common, but at the same time, every case has its particular behavior: from the properties of the fluid to the forces affecting particles. As a consequence, literature proposes different solutions. A deep study is needed before building up a code capable of bringing the convergence to different approaches.

Talk: "Innovative algorithm for particles transport in a fluid"

Suzana Milutinovic

Computer sciences, Barcelona Supercomputing Center and UPC

Suzana Milutinovic is a Phd student in Computer Architecture at the Universitat Politècnica de Catalunya (UPC). She joined the CAOS group in Barcelona Supercomputing Center in January 2014. Suzana obtained a Master in Innovation and Research in Informatics from UPC and a Bachelor degree in Electrical Engineering and Computing from University of Belgrade, Serbia.

Talk: "Validating the Reliability of WCET Estimates with MBPTA"

Paula Córdoba Pañella

CASE, Barcelona Supercomputing Center

Born and raised in Barcelona, she lived in Sweden for some months to study in the technical university of Stockholm, Kungliga Tekniska Högskolan, before finishing her Degree in Physics at the Universitat de Barcelona (2012). After the Master's Degree in Advanced Mathematics and Mathematical Engineering at the Universitat Politècnica de Catalunya (2014), she started her PhD at the Computer Applied Sciences and Engineering department in Barcelona Supercomputing Center. Her thesis is based on preconditioning parallel sparse iterative solvers for several multi-physics problems, focusing especially in CFD cases.

Talk: "Numbering along advection for Gauss-Seidel and Bidiagonal preconditioners"

Marta Guindo

Life Sciences, Barcelona Supercomputing Center

She studied both biology and biochemistry at the Universidad de Navarra and obtained her Master's Degree in biomedical research at Universitat Pompeu Fabra in 2013. After been working on the wet lab for several years, she started her PhD thesis in 2014 at the Computational Genomics group led by David Torrents under Josep M. Mercader and David Torrents supervision at Barcelona Supercomputing Center (BSC).

Her project is focused on the genetics behind complex diseases aiming to

identify new associated genetic variants and comorbidities through the analysis of up to 70 available GWAS datasets comprising around 60 different diseases and more than 400.000 individuals.

Talk: "Identification of genetic variants associated with risk for a variety of cancers through the re-analysis of publicly available genome-wide data from more than 20,000 individuals"

Elias Rodriguez Fos

Life Sciences, Barcelona Supercomputing Center

He earned a Bachelor's degree in Biology with the speciality in Genetics and Cell Biology at the Universitat Autònoma de Barcelona (UAB). Then, he studied a Postgraduate degree in Bioinformatics that gave him the chance of starting an internship in the Computational Genomics group at the Barcelona Supercomputing Center. In Dr. David Torrents group under the supervision of Dr. Josep Maria Mercader, he worked on developing a new method to perform the analysis of genomic data (from Genome-wide association studies) of complex diseases (Type II Diabetes mostly). During his internship, his interest in doing a PhD in the field of Bioinformatics and specifically in Computational Genomics grew, so he continued his studies obtaining a Master's degree in Genetics and Genomics at the Universitat de Barcelona (UB). Nowadays he is starting his PhD in the analysis and study of complex chromosomal rearrangements in cancer in Dr. David Torrents group.

Talk: "Genomic Instability Promoted by Expression of Human Transposase-Derived Gene"

Mercè Planas-Fèlix

Life Sciences, Barcelona Supercomputing Center

She obtained a Bachelors degree in Biology at "Univeristat Auònoma de Barcelona - UAB". At the same time she was earning her degree, she did a bachelor's internship and received a grant in the Evolutionary group at Genetics and Microbiology department. During her studies she discovered bioinformatics and decided to do a Bachelors internship at "Catalan Oncology Institute – ICO" on a Bioinformatics framework. Then she completed a Master´s in Bioinformatics for Genomics and drug design in UAB and accomplish the master practices at the Structural Genomics groups led by Marc Marti-Renom at "Centre Nacional d'anàlisi genomic - CNAG". Since September 2013 I am doing my PhD in Biomedicine at Computational genomics group led by David Torrents at the BSC with a La Caixa fellowship.

Talk: "SMuFin2: Identification of virus in Cancer genomes using an improved version of SMuFin"

Didier Barradas Bautista

Life Sciences, Barcelona Supercomputing Center

Didier did his bachelor and McS on Biochemistry at the UNAM in Mexico. Currently, he is a PhD student in Biomedicine at Life Science department. He changed from Wet lab to Computer lab. The reason is that while doing the master he saw that bioinformatics is changing the ways biology is made.

Since his bacherlor´s he is interested in how the interaction inside the cell produces a response. Then during his PhD he started to learn system biology and structural biology. This combination led him to the fascinating interface between molecular biology and biophysics. His thesis topic is related to the use of protein docking tools to find sites in the 3D structure that harbor mutations that alter the interactions resulting in a disease. At his talk he discussed how they have enhanced our in-house Docking program

Talk: "Docking through Democracy Re-ranking protein-protein decoys with a web-based integration model of biophysical scoring functions"

Milan Stanic

Computer Science, Barcelona Supercomputing Center

He finished his bachelor studies in the Department of Computer Science and Engineering, Faculty Of Electrical Engineering, University of Belgrade, Serbia (2008). He later completed a Master´s in Information Technology, Barcelona School of Informatics, Universitat Politécnica de Catalunya (UPC) in 2011. He developed two tools that enable rapid manual vectorization and instruction level characterization of applications, and estimate the execution time in many different vector processor configurations.

He joined Barcelona Supercomputing Center in 2009 where he is now a full time researcher at the Computer Architecture for Parallel Paradigms group. His research activity is oriented towards a doctorate title in computer architecture in conjunction with the Universitat Politécnica de Catalunya (UPC).

His current on-going research is related to low power vector processors which involves adding support for vector processing on an in-order ARM core in gem5 simulator. An integrated vector-scalar design is proposed that combines scalar and vector processing mostly using existing resources of scalar core.

Talk: "Block-Based Execution on an Integrated Vector-Scalar In-Order Core"

Simon Carrignon

CASE Barcelona Supercomputing Center / UPF

After a license in Computer Sciences and Life Sciences at Université Claude Bernard (Lyon1) he did a Master´s Degree in Nartural and Artificial Cognition at Ecole Pratique des Hautes Etudes (Paris 4) and another Master´s Degree in History and Philosophy of Science, at Paris Diderot (Paris 7).

He is currently a PhD Student in Bio-Medicine at the Universitat Pompeu Fabra (Barcelona) and he is working in the ERC project EPNet at the Barcelona Computing Center on Cultural evolution and long term economic dynamics : the case study of Rome.

Talk: “Modelling the Co-evolution of Trade and Culture”

Dario Garcia-Gasulla and Jonatan Moreno Vázquez

Computer Science, Barcelona Supercomputing Center

Dario is a PostDoc at BSC since 2015. Before that, he worked as an assistant researcher for four years at the KEMLg group (UPC), participating in several research projects related to Artificial Intelligence (particularly in Knowledge Representation and Reasoning, Machine Learning and Data Mining).

His PhD thesis tackled the large-scale graph mining problem, bringing together AI and HPC topics. As a PostDoc he is now leading a research project in collaboration with IBM for the integration of Deep Learning and graph mining technologies.

Jonatan Moreno Vázquez is a PhD student since 2015. Before that, he worked as a assistant researcher for three years at the KEMLg group (UPC), participating in research projects related with Artificial Intelligence (particularly in the eHealth field, Robotics, Machine Learning and Data Mining).

His Master´s thesis concerns the analysis of sensor data coming from a robotic walker (i-Walker) with all the problems that sensor data entails (noise, outliers, erratic data, etc.). On top of that, it applies some Machine Learning methods over this data (mainly clustering and pattern recognition algorithms). In his PhD he is working in the interpretation of deep neural network models, providing insight into the nature of the learnt artificial knowledge.

Talk: “Efficient and Versatile Data Analytics for Deep Networks”

Lorenzo Cupellini

Chemistry and Industrial Chemistry, University of Pisa

Bachelor degree in Chemistry from the University of Pisa (2013), Master´s degree in Chemistry from the University of Pisa (2013) with a dissertation on nonelectrostatic interactions in continuum models, under the supervision of Prof. Benedetta Mennucci.Currently, he works as a PhD student at the University of Pisa, in the group of Prof. Benedetta Mennucci, with a project on "The interplay between electronic coupling and vibrational motions in excitation energy transfer". His research interest include the study of excitonic systems, QM/classical models in quantum chemistry, and nonelectrostatic interactions.

Talk: "Photoprotection and triplet energy transfer in higher plants: the role of electronic and nuclear fluctuations"

Daniel Santos-Oliván

Institut de Ciències de l'Espai (CSIC-IEEC)

Daniel Santos-Oliván graduated in Physics at the University of Zaragoza (five year degree) in 2012. Then he moved to Barcelona where he studied a Master´s in Astrophysics, Particle Physics and Cosmology at the University of Barcelona.

Currently he is a PhD student at the Institute of Space Sciences (CSIC-IEEC) where he works in the development of Numerical Relativity codes to study the Gravitational Collapse in Minkowski and Anti-de Sitter spacetimes.

Talk: "Gravitational Collapse with Arbitrary-Precision Arithmetic"

Hugo Perez

Computer Science, Barcelona Supercomputing Centre

Hugo Perez got his B.S. degree in Electronic Engineering from National University in Mexico (UNAM). He got his M.Sc. degree in Computers Architecture, Networks and Systems from Universitat Politècnica de Catalunya BarcelonaTech. Currently he is working in the parallel programming models group at the Barcelona Supercomputing Center as a PhD student.

His research project entitled "Crowd Simulation and Visualization" which aims to represent the most realistic possible scenarios in a city which these kind of systems allow: urban planning, simulating disasters, simulate epidemics, among other applications. The project combines differents areas research such as: big data, AI, Parallel Programming Models, HPC, Computer Graphics, HCI between others.

Talk: "Crowd Simulation and Visualisation"

Qixiao Liu

Computer Science, Barcelona Supercomputing Center

Qixiao Liu is a PhD student at Barcelona Supercomputing Center (BSC) and the Universitat Politecnica de Catalunya (UPC), Spain. He received a BS and MS degrees from Northeastern University (NEU), China. His PhD study focuses on quantifying the energy cost of per task in the multicore systems, based on simulated power models of the processor and memory system. His research interests include studying energy efficient design in multicore system and power secure systems.

Talk: "Per-task Energy Metering and Accounting in the Multicore Era"

Xubin Tan

Computer Science, Barcelona Supercomputing Center

She is a third-year Ph.D student in the RoMoL team, in the department of Computer Science in the Barcelona Supercomputing Center and also the department of Computer Architecture in the Universitat Politècnica de Catalunya.

Her current research can be summarized into two parts: First, hardware accelerator/co-processor for task dependence graph management and scheduling for task-based dataflow programming models like OpenMP, OmpSs among others. Second, data exchanging schemes between GPP and accelerators. She had previously worked as IC logic designer and verification engineer for shared-memory heterogeneous multi-cores, and she received both her Master´s and Bachelor´s degrees in E.E. in Beijing, China.

Talk: "Task Dependences Management Hardware Acceleration for Task-based Dataflow Programming models"

Jan Ciesko

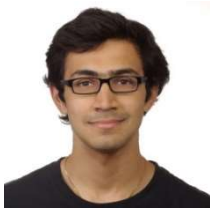
Computer Science, Barcelona Supercomputing Centre

Jan is a PhD student at the UPC and resident student at the Barcelona Supercomputing Center. His research is focused on programming model- and architectural support for irregular array-type reductions (scatter-update). Prior to his research position, Jan studied computer science at the Friedrich-Alexander University in Germany and initiated several mobile app and business ventures including Numerical Monkeys and Realizr.net.

Talk: "Optimizing Scatter-updates Through Inspectors-executors in OmpSs"

Rajiv Nishtala

Computer Science, Barcelona Supercomputing Center



Rajiv Nishtala is a PhD student at Barcelona Supercomputing Center at Universitat Politecnica de Catalunya since June 2013. His research interests include energy-efficient (green) computing and thread scheduling.

Talk: "Runtime Estimation of Performance–Power in CMPs under QoS Constraints"

Biosciences, University of Kent

Morena obtained her Bachelor and Master´s degrees in Pharmaceutical Chemistry at the University of Catania (Italy) in 2013. She has been working on a combined computational and experimental project to investigate Protein-Protein and Protein-Ligand Interactions. Then she she started her PhD at the University of Kent, in Canterbury (United Kingdom) where she is researching Single Nucleotide Variants and where she has been involved in the development of a computational tool called VarMod, which is a freely available algorithm for the prediction and the modeling of the effects of non synonymous variants. She has also been researching sequence conservation in Ebola viruses in order to understand the virus pathogenicity.Her main research interest is in Structural Bioinformatics and precisely in those approaches that help understanding the relationship between genetic variation and disease.

Talk: "Conserved differences in protein sequence determine the human pathogenicity of Ebola viruses"

Jelisa Iglesias

Life Science, Barcelona Supercomputing Center

She obtained her degree in Biotechnology at the Universitat de Barcelona in 2013; during her final project she discovered the bioinformatics world and loved it thus she did an MSc in Bioinformatics at the Universitat Autònoma de Barcelona in 2014. She has been doing her PhD since 2015 in the Barcelona Supercomputing Center and the Universitat Politècnica de Catalunya, working on improving the results of VS using induced-fit techniques.

Talk: "Enrichment of Virtual Screening results using induced-fit techniques"

Daniel Lecina Casas

Life sciences, Barcelona Supercomputing Center

Daniel studied physics in the Universitat de Barcelona. He did a Master´s on computational physics in a joint program between the Universitat de Barcelona and the Universitat Politècnica de Catalunya, with a master thesis on applying quantum annealing to color an Erdos-Reny random graph. Later on, he started working as a software developer in Victor Guallar's group, where he is now pursuing his PhD.

Talk: "On the Way to Real Time Protein-Ligand Simulations"

Víctor López Ferrando

Life Sciences, Barcelona Supercomputing Center

He studied Mathematics and Computer Engineering at the Universitat Politècnica de Catalunya, having an special interest in algorithmics. His degree thesis was titled "Topology and Time Synchronization algorithms in wireless sensor networks" and was the result of a two year work in the WISEBED European project.

After finishing his studies he worked as a software engineer for two years in a spin- off of the Department of Computer Architecture of UPC named Talaia Networks.

In September 2014 he got a La Caixa-Severo Ochoa Fellowship and started his PhD in Bioinformatics in the Barcelona Supercomputing Center.

Talk/Poster's title: "PMut2: a web-based tool for predicting pathological mutations on proteins"

POSTER PRESENTERS

Verónica Torralba Fernández

Earth Sciences, Barcelona Supercomputing Center

She studied physics at Complutense University of Madrid (UCM), where she also obtained a Master´s in Meteorology and Geophysics. She joined the Climate Forecasting Unit (CFU) in IC3 in October 2013, where she was involved in the development and communication of climate services for wind energy.

She is currently working in the Earth Sciences Department at the Barcelona Supercomputing Center where she is a PhD student working on different European Projects. The overall aim of the PhD is the assessment of different statistical post-processing methods which are applied to improve the the quality of the seasonal forecasts produced by the state-of the-art prediction systems. Then based on the post-processed seasonal forecasts, alternative methodologies to compute some tailored variables, as capacity factor of wind power are explored.

Poster: "Development of a wind energy climate service based on seasonal climate prediction"

Sergio Natan González Rocha

Earth Sciences, Barcelona Supercomputing Center

Sergio Natan González Rocha has a degree in Chemical engineering from the University of Veracruz (UV) in Mexico, MsC in computation (UV), MsC in Environmente (UV), and a PhD in Environmental Management for development from the Popular Autonomous University of Veracruz (UPAV) in Mexico; he has worked as academic technician, engineer, consultant and full time professor in environmental, computing and chemical areas; Professor in Master and PhD degree in UV, UNID and UPAV; postdoctoral CONACyT in the BSC-CNS in Earth Sciences Department.

Since November 2014, he has expanded his interest in modeling in high resolution models WRF, CMAQ and CALIOPE for Health impacts assessments in Europe and Mexico.

Poster: "Assessment of Meteorological Models for Air Pollution Transport: Analysis between Mexico and Puebla Metropolitan Areas"

Marta Garcia Gasulla

Computer Science, Barcelona Supercomputing Center

Marta Garcia obtained her degree in Informatics in 2006 from Universitat Politecnica de Catalunya (UPC), Facultat d'Informatica de Barcelona (FIB). That year started a Master´s in Computer Arquitecture and Network Systems at Universitat Politecnica de Catalunya, Departament d'Arquitectura de Computadors (DAC), finishing it in 2008.

In July 2006 she started working as a research student at Barcelona Supercomputing Center. In 2011 she started her PhD. She also worked as associated professor from 2008 to 2013 teaching different courses from Operating systems to parallel programming.

Poster: "Dynamic Load Balancing for hybrid applications"

Federica Sacco

Departament de Tecnologies de la Informació i les Comunicacions (DTIC), Universitat Pompeu Fabra

She is an Italian PhD student working in Barcelona. In 2015, she obtained her Master´s degree in Biomechanics and Biomaterials (a Biomedical Engineering Master program) at Politecnico di Milano.

In September 2015 she moved to Barcelona to start her PhD at Universitat Pompeu Fabra under the supervision of Dr. Constantine Butakoff, in collaboration with the Barcelona Supercomputing Centre and under the guidance of Dr. Jazmin Aguado-Sierra. The project she is working on is about the creation of a detailed full heart model and the study of the influence of trabeculae and papillary muscles on its function from the electro-physiological and hemodynamic points of view. Most of the early electro-mechanic physiological heart models are simplified and consider a smooth endocardial surface. With her work, she is aiming to understand the effect of neglecting these structures on cardiac simulations.

Poster: "Effects of detailed ventricular anatomy on the blood flow"

Marina Corbella Morató

Facultat de Farmàcia, Universitat de Barcelona

After obtaining a degree in Chemistry at the Universitat Autònoma de Barcelona she worked in the R&D department of Fine Chemicals in Applus+ LGAI technological center, after three years she moved to the Faculty of Pharmacy of the Universitat de Barcelona where she completed her master hesis on the synthesis of compounds as potential antitumor agents.

Now, she is undertaking a PhD in Computational Biology and Drug Design group under the supervision of Dr.Carles Curutchet. Her research topic is charge transfer in biological systems such as DNA and phycobiliproteins (antenna proteins involved in light-harvesting processes).

Poster: "Exploring the protonation properties of photosynthetic phycobiliprotein pigments from molecular modelling and spectral line shapes"

Maria Coto-Sarmiento

Case, Barcelona Supercomputing Center

Maria Coto is a PhD student in archaeology between Barcelona Supercomputing Center and University of Barcelona, under the project EPNET (Production and Distribution of Food during the Roman Empire: Economic and Political Dynamics). She studied History at University of Seville. She also had a Mphil in Archaeology and Cultural Heritage at the same university. As an archaeologist, she is interested in culture evolution and social changes from an evolutionary perspective.

Poster: "Clustering the Roman Empire!: the use of multivariable analysis to understand cultural dynamics"

Oscar A. Esquivel

Industrial Engineering and numerical methods, Intituto Tecnologico De Estudios Superiores De Monterrey

Oscar A. Esquivel-Flores received his Bachelor's degree in Applied Mathematics and Computing from Universidad Nacional Autónoma de México, M.S. degree in Computer Sciences from Universidad Autónoma Metropolitana, México and PhD degree in Computer Engineering from UNAM in 2013. His research in parallel algorithms and high performance computing areas started during a posdoctoral position at the Barcelona Supercomputing Center (2014 - 15) as part of an international agreement with National Council of Science and Technology of México. He currently holds a posdoctoral position at Instituto Tecnológico y de Estudios Superiores de Monterrey, México. In 2015 he was elected as candidate of National Researchers System (SNI-CONACYT, México).

Diego Dávila

FIB, Universitat Politècnica de Catalunya

Diego Dávila is a High Performance Computing(HPC) enthusiastic currently studying a Master´s degree in Innovation and Research in Informatics at the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain.

Diego has been working within the HPC environment since 2013 starting as a system administrator at the National Supercomputing Laboratory in México.

Recently he joined the Extreme Computing research group at Barcelona Supercomputing Center (BSC) where he aims to develop highly efficient stochastic and hybrid parallel algorithms. He has been awarded the Severo Ochoa International Scholarship in 2014, the CENEVAL Excellence Award in 2013, the PROMESAN Scholarship in 2010 and the UASLP Excellence Award in 2008.

Poster: "Enhanced Monte Carlo Methods for Sparse Approximate Matrix Inversion"

Albert Gutierrez-Milla

CASE, Barcelona Supercomputing Center

Albert obtained his bachelor in Computer Science at UAB and a MSc in HPC and Information Theory in the same university. He was granted with a PIF grant at the Computer Architecture and Operating Systems department at UAB were he is still developing his PhD. His research focuses on crowd dynamics, agent based modelling, HPC and simulation. Currently he is also employed in the Barcelona Supercomputing Center in the Fusion group inside the CASE department. His work at BSC involves HPC support to the EuroFUSION community and to the Fusion group.

Poster: "Reproducing crowd turbulence with Verlet integration and agent modeling"

Silvana De Souza Pinheiro

Physical Chemistry, University of Barcelona

PhD student in Research, Development and Drug Control - University of Barcelona (UB). Master in Genetics and Molecular Biology - Federal University of Pará (UFPA). Graduated in Full Degree in Natural Sciences with specialization in Chemistry - University of Pará (UEPA). Graduated in Generalist Pharmacy - University Center of Pará (CESUPA). She has experience in education, with emphasis on Teaching of Chemistry and Natural Sciences, Theoretical Chemistry and Computational, with applications in Drug Planning. She operates mainly in the following areas: Bioinformatics, Computational Biology, Molecular Modelling, Hybrid methods QM / MM and energy transfer.

Poster: "Assessing drug-protein binding by simulation of stereoselective energy transfer dynamics: electronic interactions between tryptophan and flurbiprofen"

Dani Gallart Escolà

Case, Barcelona Supercomputing Center

Dani Gallart studied fundamental physics at UB. After obtaining his degree in physics he enrolled the nuclear engineering MSc at UPC-ETSEIB where he obtained one of the grants from Fundació Catalunya – La Pedrera.During his Master thesis he joined the Barcelona Supercomputing Center Fusion group (CASE) under the supervision of ICREA researcher Dr. Mervi Mantsinen. During this period he was awarded one of Fundació "La Caixa" grants in order to start his Phd.

Currently, he is developing his Phd studies under the supervision of Dr. Mervi Mantsinen. His work is mainly focused on fast particle physics and plasma heating. He works closely with the main European fusion facilities like the Joint European Torus (JET) in Oxford, UK, and Asdex Upgrade (AUG) in Garching, Germany, under the EUROfusion umbrella.

Poster: "Extrapolation for high fusion performance in JET"

Rubén Cruz García

Earth Sciences, Barcelona Supercomputing Center

Rubén studied Environmental Sciences at University of Murcia, where he discovered his passion for investigating the climate change. Thereafter, he studied a Master´s degree in Geophysics and Meteorology at University of Granada. He started his PhD project in the Climate Prediction Group (CPG) at the Earth Sciences Department of the Barcelona Supercomputing Center in October 2015. His project focuses on trying to improve 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.

Poster: "Regional Arctic sea ice predictability and prediction on seasonal to interannual timescales"

Mariña López Yunta

CASE, Barcelona Supercomputing Center

After graduating in Mathematics from the University of Santiago de Compostela, she did a Master in Mathematical Modelization in University Pierre et Marie Curie (Paris 6). She is a PhD student at CASE departament at Barcelona Supercomputing Center. She is working on processing medical data and electromechanical cardiac simulations in close colaboration with CNIC (Centro

Nacional de Investigaciones Cardiovasculares, Madrid).

Poster: "Generation of a simulation scenario from medical data: Carto and MRI"

Helena Caminal

Computer Science, Barcelona Supercomputing Center

Helena is a research assistant at Barcelona Supercomputing Center (BSC) and currently enrolled in the High Performance Computing Master of Science degree at the Polytechnical University of Catalonia (UPC). In december 2014 she received a 5-year degree in Industrial Engineering with Electronics specialization by the UPC. She spent 8 months at Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle (IRIDIA-CoDE) research center in Brussels, Belgium, developing her final degree project in image processing for swarm robotics.

Since she started working at BSC, Helena has been working with a representative set of benchmarks and porting their vectorized version to OmpSs programming model in order to analyze how to improve vector architectures and runtime systems. She has submitted a paper that shows the performance and energy reduction of user-directed vectorized codes and its differences between manually vectorized codes.

Poster: "How Can We Improve Energy Efficiency through User-directed Vectorization and Task-based Parallelization?"

Isaac Sánchez Barrera

Computer Science, Barcelona Supercomputing Center

Isaac Sánchez Barrera received his BSc degrees in Mathematics and Computer Science from Universitat Politècnica de Catalunya (UPC) in 2015. In April 2015, he joined the RoMoL team from BSC, and in September 2015 he started the Master's degree in Advanced Mathematics and Mathematical Engineering from UPC. He is part of the Editorial Committee of RSME's Bulletin (the Royal Spanish Mathematical Society) and the Publications Commission of ANEM (the National Association of Mathematics Students in Spain), and is also member of both societies.He is Joint Coordinator for the XVII Encuentro Nacional de Estudiantes de Matemáticas, which will be celebrated in Barcelona in July 2016. His main interests are optimization and approximation algorithms and he is currently studying graph partitioning techniques to minimize data movements in parallel applications.

Poster: "Using Graph Partitioning to Accelerate Task-Based Parallel Applications"

Mireia Rosell Oliveras

Life Sciences, Barcelona Supercomputing Center

Mireia Rosell is a PhD student at the Protein Interactions and Docking Group in the Barcelona Supercomputing Center, lead by Juan Fernandez-Recio. She did a B.Sc. In Biochemistry at Universitat Autònoma 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.

Poster: "Characterization of Protein-Protein Interfaces and Identification of Transient Cavities for its Modulation"

Gabriela Hernández Larios

Life Sciences, Barcelona Supercomputing Center

She graduated from the Autonomous University of Zacatecas with a Bachelor's degree in Electronics and Telecommunications Engineering in 2013. Afterwards, she worked as a VoIP and Unified Communications Engineer at ASLO Information Technologies. In 2014, she enrolled the Erasmus Mundus Master Program in Data Mining and Knowledge Management. She carried out the first year of the Master at the École Polytechnique de l'Université de Nantes with the specialization in Knowledge and Decision, and she studied the third semester at the Universtà degli Studi del Piemonte Orientale with the specialization in Relational Data Mining.

Currently, she is doing her thesis in the Barcelona Supercomputing Center and her research is mainly focused on the improvement of the protein-ligand binding affinity through Statistics and Machine learning techniques. She is interested in Data Mining and Machine Learning Algorithms applied to life science areas, in knowledge extraction and data visualization.

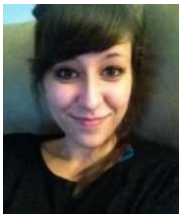
Poster: Improvement of Protein-Ligand Binding Affinity Prediction using Machine Learning Techniques

Sonia Romero Téllez

Facultat de Farmàcia, Universitat de Barcelona (UB)

Theoretical Chemistry Computational Modelling (TCCM) Master Student, doing her TFM (final work) in Carles Curutchet's group, working on the re-parameterization of MST Solvation Model applying

the new ddCOSMO algorithm to obtain more accurate solvation free energies at quantum level in a cheaper and faster way even in biological systems.



Before, she did an experimental Master in the Faculty of Biochemistry. She worked both experimentally and theoretically, finding inhibitors for DXS enzyme from Methyl-erythritol phosphate pathway(MEP) (non-mevalonate pathway). Some bacteria, such as Plasmodium Falciparum use this pathway to create isoprenoids and triggers malaria disease.

Sonia graduated in 2012 in Chemistry at University of Barcelona (UB). Her final presentation was about choosing properly DFT functionals in the unit cell of transition metals.

Poster: "Toward accurate solvation energies of large biological systems"

Iulian Valentin Brumar

Computer Science, Barcelona Supercomputing Center

Iulian Brumar has a degree in Computer Science with Computer Engineering Mention from Universitat Politècnica de Catalunya and is currently undertaking a Master's Degree in Innovation and Research in Informatics at the same university.

He has been collaborating with Barcelona Supercomputing Center for almost three years, focusing on high performance computing, task based programming models and performance analysis.

Poster: "Improving Scalability of Task-Based Programs"

Antonio Viayna Gaza

Campus Torribera – Facultat de Farmàcia, Universitat de Barcelona

I obtained my Bachelor's Degree in Pharmacy from Universitat de Barcelona (UB) in 2012. During my Bachelor's Degree I did two interships, first at the Unitat de Química Farmacèutica of the UB, under the supervision of Dr. Diego Muñoz-Torrero, and a 3-month intership in the Dipartimento di Scienze Pharmaceutiche of the Università di Bologna. After finishing my degree, between 2012 and 2013, I obtained a "Master de Industria Farmaceutica y Parafarmaceutica" in the "Centro de Estudios Superiores de la Industria Farmacéutica (CESIF)".

Later I started a 2-year work experience in the pharmaceutical industry. From December 2012 to September 2014, I was working in the R&D Department of the pharmaceutical holding Grifols, a group of companies devoted to the investigation, development, manufacturing and commercialization of hemoderivates, serums, clinical nutrition products, diagnostic systems and medical-sanitary materials.

In October 2014, I changed my work area to the Computational Chemistry field, starting a PhD in the Computational Biology and Drug Design (CBDD) group under the supervision of Dr. F. Javier Luque, working in conformational studies of ligands and the design of new anti-malarical compounds.

Poster: "Conformational landscape of small ligands: A Multilevel strategy to determine the conformational penalty of bioactive ligands"

EXTENDED ABSTRACTS

Development of a wind energy climate service based on seasonal climate prediction

Verónica Torralba[1], Isadora Jiménez[1], Llorenç Lledó[1], Nube González-Reviriego[1] Albert Soret[1] and Francisco Doblas-Reyes[1,2]
(1) *Earth Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain*
(2) *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
veronica.torralba@bsc.es

Abstract- *Climate predictions tailored to the wind energy sector represent an innovation to better understand the future variability of wind energy resources. In this work an illustration of the downstream impact of the forecasts as a source of climate information, the post-processed seasonal predictions of wind speed and temperature will be used as input in a transfer model that translates climate information into capacity factor. This transfer model is based on multivariate regression that assumes a linear relationship between wind speed and temperature with the capacity factor.*

Key words: wind energy, seasonal forecasts, capacity factor, bias-correction, climate services

A. INTRODUCTION

Operational and economic issues related to wind energy require the modeling and forecasting of the wind power generation processes at a range of temporal and spatial scales (from minutes to decades) [1]. The wind industry has traditionally used forecasts at short (from hours to a few days) time scales due to the strong dependency between the wind energy production and the wind speed synoptic-scale variability [2]. At longer time scales, the need of climate information representative of the next few decades for resource evaluation has raised the interest of the wind energy users in climate projections [3]. Hence, climate predictions whose time scales vary from one month to a decade into the future can cover the existing gap between weather forecasting and climate projections.

At seasonal time scales current energy practices employ a simple approach based on a retrospective observed climatology. Instead, probabilistic seasonal forecasting can better address a long list of challenges to produce climate information that responds to the expectations of the users [4] and that can be used to make specific decisions that affect energy demand and supply, as well as decisions relative to the planning of maintenance work.

The large amount of information that arises from the seasonal forecasts (e.g. uncertainty, skill and reliability assessments, bias-correction techniques, probabilistic approaches) is hard to understand and in most cases the users are not able to incorporate it in a useful manner for their daily activities.

The main goal of this work is to develop tailored climate information that can be afterwards used as a tool to inform wind energy users with greater accuracy than their current approaches.

B. DATA AND METHODS

The probabilistic seasonal forecasts of 10-m wind speed and 2-m temperature from the ECMWF seasonal prediction system, System 4 (S4), in the 1981-2014 period have been used. In addition, the 10- m wind speed and 2-m temperature data from ERA-Interim have been also used as reference dataset.

Figure 1*. ECMWF S4 10-m wind speed seasonal forecast for JJA 2015 initialized the 1st of May. The most likely wind speed category (below-normal, normal or above normal) and its percentage probability to occur is shown. White areas show where the probability is less than 40 % and approximately equal for all three categories. Grey areas show where the climate prediction model doesn't improve the climatology.*

Probabilistic forecasts have been used in order to provide users with information about the forecast uncertainty. An example of the probabilistic seasonal forecasts of 10-m wind speed has been illustrated in Figure 1.

The capacity factor (CF) is the average power generated over a period of time, normalized by the maximum power of a wind-turbine. It is a widely used indicator for the wind energy users which provides information about the extent of use in a power plant. The state-of-the art climate prediction systems don't produce this kind of variables, for that reason it is needed the development of an impact model to generate

CF seasonal predictions from climate variables. However, the capacity factor depends on many factors as for example the operating limitations of a wind farm that have not been considered in this analysis.

The impact model is based on a multivariate linear regression (Equation 1) that relates CF with wind speed (ws) and temperature (T).

$$CF(ws, T) = A\,ws + B\,T + C \quad (1)$$

This equation has been used to relate the climatological information of wind speed and temperature from the ERA Interim reanalysis and the seasonal forecasts of such variables for a target period.

The first step has been the estimation of the capacity factor values from the ERA-Interim reanalysis which has been used as reference by the methodology described in [5]. The wind speed, temperature and the derived capacity factor from ERA-Interim are fitted to the equation (1) in order to find the A, B and C coefficients.

As the prediction of wind speed and temperature is affected by biases, three different techniques for the bias-adjustment of ensemble forecasts are considered: simple bias correction, quantile-quantile mapping and calibration method. These methods have been applied on daily data and in "one-year out" cross-validation and they produce corrected forecasts with improved statistical properties. The impact of these different bias corrections techniques over the forecast quality has been explored in a forecast quality assessment.

The bias corrected seasonal predictions of wind speed and temperature has been used as input in the regression model together with the coefficients estimated previously for the reanalysis, and the seasonal predictions of CF are produced.

C. RESULTS

The seasonal average of the CF computed from ERA-Interim wind speed and temperature for a particular region in Canada has been illustrated in the Figure 2. It shows the relationship between the three variables used in the regression model and evidences that wind speed is the main driver of the CF. However, can be noticed that the highest values of CF are mainly produced for high temperatures.

The impact of the bias-adjustments of wind speed and temperature has also been explored. The forecast quality assessment demonstrates that the three considered methods produce forecasts with improved skill and reliability. This is a critical aspect of the forecasts from the user perspective because a good reliability guarantees the trustworthiness of the predictions.

***Figure 2**. Scatter plot of the mean capacity factor (%) for a region in Canada [49.6°-51.7°N and 246°-248.2° E] in DJF. Marginal histograms of ERA-Interim 10-m wind speed in the x-axis and 2-m temperature in the y-axis for the period of 1981-201.*

The estimated CF from the reanalysis and the bias-corrected forecasts have been used to estimate the predictions of the capacity factor. The output of the regression model is in terms of capacity factor values aggregated in terciles where each percentage indicates the probability of capacity factor being below-normal, normal and above normal. The forecast quality assessment of these predictions reveals that the estimated forecasts for this particular region has positive values which indicates the added value of the seasonal predictions of capacity factor relative to the climatological capacity factor.

D. CONCLUSIONS

This study describes a simple methodology to develop useful information for the wind industry that can be easily integrated in their decision-making processes. Transforming climate variables into a user friendly capacity factor is essential for the wind industry. The quality of these predictions as well as their benefit to the wind energy users should be further explored in different regions around the world.

REFERENCES

1. Fant, C., Adam Schlosser, C. & Strzepek, K. The impact of climate change on wind and solar resources in southern Africa. *Appl. Energy* **161,**

556–564 (2016).

2. Alessandrini, S., Sperati, S. & Pinson, P. A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data. *Appl. Energy* **107,** 271–280 (2013).

3. Reyers, M., Pinto, J. G. & Moemken, J. Statistical-dynamical downscaling for wind energy potentials: Evaluation and applications to decadal hindcasts and climate change projections. *Int. J. Climatol.* **244,** 229–244 (2014).

4. Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P. & Rodrigues, L. R. L. Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdiscip. Rev. Clim. Chang.* **4,** 245–268 (2013).

5. MacLeod, D., M. Davis, F. J. Doblas.-Reyes. Modelling wind energy generation potential on seasonal timescales with impact surfaces. *SPECS Technical Note* No.3, 24 pages (2014).

Assessment of Meteorological Models for Air Pollution Transport: Analysis between Mexico and Puebla Metropolitan Areas

Sergio Natan González-Rocha[1,3], Osiel O. Mendoza-Lara[2], Vicente Fuentes-Gea[2], Jose M. Baldasano[1,4]
[1]Barcelona Supercomputing Centers (BSC – CNS) – Earth Sciences
[2]National Autonomous University of Mexico (UNAM) – Faculty of Engineering
[3]University of Veracruz (UV) – Poza Rica – Tuxpan Campus
[4]Universitat Politecnica de Catalunya (UPC), Barcelona, Spain
sergio.gonzalez@bsc.es

Abstract- *This poster presents the results of research in the metropolitan areas in Mexico and Puebla valleys. The objective is assess and conduct a sensitivity analysis of meteorological conditions that could influence air pollutant transport between both valleys. The simulations were performed with CALMET v6.4 and WRF v.3.5, latter performed in the Mare Nostrum III super computer in the BSC-CNS; six days simulations considered statistically by Spearman correlations were selected in March and May months in 2012 year. It was found that WRF presented better results in domains to 9, 3 and 1 km in contrast to CALMET, considering wind speed and temperature variables.*

I. Introduction

Urban development and air pollution in last 50 years, has brought a lot of problems, these consequences are just beginning to be recognized [1]. In accordance with World Health Organization (WHO), one way of assess air quality is measuring the concentration of criteria air pollutants (PM, O_3, SO_2, NO_2 y CO) [2]. Furthermore, there have been multiple studies using mathematic models, with the aim of to know the transport and influence between urban zones. Currently, air quality models constitute a complementary approach to monitor and characterize air pollution [3].

The Mexico City Metropolitan Area (MCMA) lies in an elevated basin at an altitude of 2240 m.a.s.l. (Meters above sea level) and 780 hPa mean atmospheric pressure (data from Mexico City International Airport). The MCMA's large population, industries, 5 million vehicles, complex topography, and meteorology cause high pollution levels. The mountains together frequent thermal inversions trap pollutants within the basin. The high elevation and intense sunlight also contribute to photochemical processes that create O_3 [4]. Some previous research suggests that air pollution emitted by MCMA is transported to cities near of this as Toluca and Cuernavaca [5][6]. However, it should be mentioned that the MCMA is a receptor of air pollution, e.g. from the Tula Metropolitan Area as a result of the emissions of industrial complexes in that area [7], there is few information related with the transport between MCMA and Puebla Metropolitan Area.

Usually air quality modelling systems (AQMS) require detailed information about topography, meteorological and pollutants emissions. Atmosphere is the place where are carried out the transport phenomena, therefore is need to obtain reliable and validated data to achieve accurate results for meteorological modelling as well as in air quality modelling.

The objectives in this work are threefold. First, to analyse synoptic and meso meteorology in the study area to find meteorological variables that impulse the air pollutants transport between both MCMA and Puebla Metropolitan Area (PMA); second, to assess the sensitivity of meteorological models (WRFv3.5 and CALMETv6.4); and the third is generate 3D weather information for use in an AQMS.

II. Methods

II.1 Study zone and selection of days for modelling

The study area was defined as a rectangular grid projection. The left corner reference coordinates of Southwest are latitude 18.603864°N, longitude -99.104269°W. The grid has 115 per 115 (*x,y*), with 1 km of spacing. The Metropolitan Areas within study area are: Mexico City, Puebla-Tlaxcala and Cuernavaca. There are three important elevations that creates a particularly situation: in the centre the "Popocatepetl" and "Iztaccihuatl" volcanoes, and northeast "La Malinche" volcano with 5500 m, 5220 m and 4420 m.a.s.l. respectively. The days selection, was performed by statistics analysis of measuring air pollutants data between 2001 and 2013 from SIMAT (Sistema de Monitoreo Atmosférico de la ciudad de México, spanish acronym) in MCMA and REMA (Red Estatal de Monitoreo Atmosférico, spanish acronym) in PMA, (2001-2013), this database must fulfil the criteria on the 75 % minimum. The third stage calculated the maximum mixing height layer with Holzworth method [8]. This calculation is aimed at getting to know in which days the heights exceeds volcanoes elevation, and was realized using information from radiosounding from International Airport "Benito Juárez" in Mexico City (BJIA, English acronym). Finally, was performed a Spearman correlation of measuring air pollutants data between MCMA and PMA to select days most correlated.

II.3 CALMET configuration

CALMET as a diagnostic meteorological model uses meteorological measurement, orography and land use data, for CALMET v6.4 configuration was used local meteorological data observations from different institutions, SIMAT, BUAP (Benemérita Universidad Autónoma de Puebla, Spanish acronym) and SMN (Servicio Meteorológico Nacional de

México, Spanish acronym), this study used 16 meteorological stations datasets; the vertical meteorological information was obtained from BJIA radio sounding. The key configuration are z levels: 0., 20., 40., 80., 160., 300., 600., 1000., 1500., 2200., 5000, and maximum mixing height is 5000 m.a.s.l.

II.2 WRF configuration

The Weather Research Forecasting (WRF v3.5) is a mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting, configuration in this work considered the domains: 27 (D0), 9 (D1), 3 (D2) and 1 (D3) km, centered on the coordinates 19.124º N, -98.556º W, with geographical resolutions of 10 m, 5 m, 2 m and 30s; GRIB2 files (6 hourly files) in 1º x 1º resolution were downloaded from the NCEP data base ds083.2 [9]; according to the procedure of preprocessing in WPS, the geogrid, ungrib and metgrid for each domain corresponding files were generated, being D0 the parent domain and D1, D2, and D3 nesting domains. It was considered in the parameterization 50 eta-levels, and mp_physics 4 value (D0, D1) and 5 (D2, D3). The execution was performed in the super computer MNIII, with 128 processors and 7:00 hours for each day simulated.

III.4 Sensitivity assessment

The sensitivity assessment considered only the meteorological variables wind-speed and temperature; the BIAS (1) and MSRE (2) equations were calculated with METAR observations of BJIA (METAR MMMX) in the days of this evaluation, the WRF domains D1, D2, D3 and CALMET domain. Φ_i is forecast value for i cell, Φ_{obs} is observation for i cell and N is the number of analysed values. The selection of these variables was the lack of information in other meteorological variables as relative humidity or wind-direction.

$$BIAS = \sum_{i=1}^{N} \frac{(\phi_i - \phi_{iobs})}{N} \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\phi_i - \phi_{iobs})^2}{N}} \quad (2)$$

III. RESULTS

The database year concentration of air pollutants with better performance than satisfy the 75% sufficiency criteria was 2012. The calculation of the estimate of the maximum mixing layer height (MMLH) used the BJIA radio survey, calculating in March, April and May months, MMLH higher than 3250 m (minimum height between the volcanoes), therefore these months were selected by Spearman correlation, obtaining the six days with higher correlation: 14, 19, 30 and March; 18 and May 27, 2012.

CALMET v6.4 and WRF v 3.5 simulations were calculated in different geopotential heights; a qualitative analysis of the behavior of synoptic and meso-scale meteorology was performed with wind-speed, wind direction, temperature, planetary boundary layer (WRF: PBLH) and mixing layer height (CALMET: MLH) variables; Figure 1 shows an example of these results. In sensitivity assessment were compared hourly values of temperature and wind speed between modelled results and observations from MMMX stations in selected days at 00 - 23 (00 UTC-6). Tables 1 and 2 shown wind speed magnitude at z500 height, pressure at 500 hPa and the geopotential height in D1. It can be observed conditions for possible pollutants transport. The temperature modeled in CALMET v6.4 presented better BIAS results between -0.45 and 2.33, however, the RMSE did not.

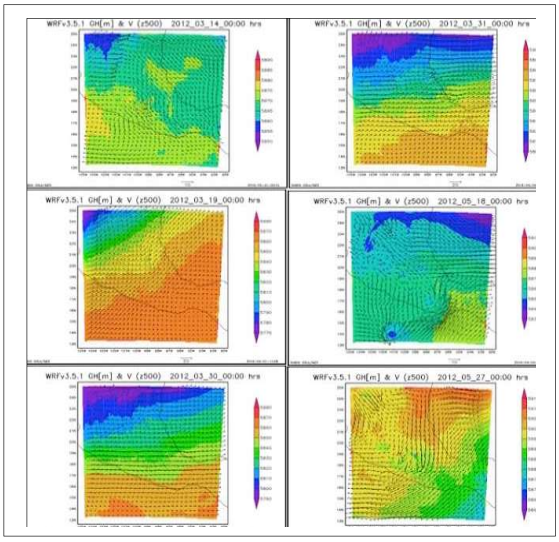


Fig. 1. Wind-speed, wind-direction at z500 (500 hPa) and geopotential height for domain (D1) in selected days for simulation.

Table 1. Sensitivity assessment results in Temperature (ºK).

Temperature [ºK]								
Day	**BIAS**				**RMSE**			
	CM	**WD_1**	**WD_2**	**WD_3**	**CM**	**WD_1**	**WD_2**	**WD_3**
1	0.01	1.42	2.03	-0.87	3.33	2.30	1.42	2.76
2	2.33	1.42	1.30	1.46	3.53	1.58	1.14	1.65
3	-0.06	1.16	1.20	1.32	2.58	2.55	1.10	2.58
4	0.03	1.31	1.46	1.61	2.69	2.13	1.21	2.37
5	-0.45	1.71	1.75	2.07	3.00	2.02	1.38	2.25
6	0.08	1.65	1.66	1.64	3.33	1.85	1.29	1.88

CM: CALMET. **WD1:** WRF dominium 1. **WD1**: WRF dominium 2. . **WD1:** WRF dominium 3.

Table 2.- Sensitivity assessment results in wind speed (m/s).

Wind speed [m/s]								
Day	**BIAS**				**RMSE**			
	CM	**WD_1**	**WD_2**	**WD_3**	**CM**	**WD_1**	**WD_2**	**WD_3**
1	2.05	1.72	1.69	1.71	2.55	2.07	2.09	2.11
2	1.45	1.73	1.29	1.22	2.17	2.39	2.21	2.23
3	3.09	2.02	2.10	2.08	3.82	2.93	3.05	3.14
4	2.81	2.34	2.19	2.07	3.60	3.16	3.17	3.19
5	2.54	1.65	1.60	1.58	3.32	2.08	2.08	2.04
6	2.03	1.83	1.82	1.73	2.87	2.49	2.52	2.44

CM: CALMET. **WD1:** WRF dominium 1. **WD1**: WRF dominium 2. . **WD1:** WRF dominium 3.

Wind speed modeling WRF v3.5 presented better performance in Wind-speed BIAS values of 1.22 (minimum) and 2.34 (maximum) at different domains (D1, D2 and D3)

and RMSE values of 2.04 (minimum) and 3.19 (maximum) corresponding at D3.

IV. Conclusions

According to the sensitivity analysis performed on selected days (14.03, 19.03, 30.03, 18.03 and 27.05, 2012), it concluded that:

Meteorology for selected days was analyzed, concluding that could exist conditions for air pollution transport between these metropolitan areas.

Modelling in WRF v.3.5 reflected better sensitivity to temperature and wind speed, although CALMET v.6.4 model is influenced by the surface meteorological information provided by only 16 stations.

WRF Outputs model generated datasets for their use in an AQMS in selected days in MCMA and PMA; it is necessary to confirm the air transport of pollutants between both metropolitan areas.

Acknowledgment

Thanks to the Barcelona Supercomputing Center – Centro Nacional de Superomputación (BSC-CNS) for the support in this mobility project and the facilities for the stay in the Earth Sciences Department, thanks to CONACyT for the "Beca mixta" support and the Mobility Department in UNAM.

References

[1] Baldasano, J., Valera, E., & Jiménez, P. (2003). Air quality data from large cities. *The Science of the Total Environment, ELSEVIER*, 141-165.

[2] WHO. (2014, Marzo). *World Health Organization. Ambient (outdoor) air quality and healt.* Retrieved 2015, from Fact sheet N°313: http://www.who.int/mediacentre/factsheets/fs313/es/

[3] Valverde. (2016). Characterization of atmospheric pollution dynamics in Spain by means of air quelity modelling,. *Ph.D.Thesis*. Barcelona, Cataluña, España: UPC-BSC (Universidad Politécnica de Cataluña-Barcelona Supercomputing Center-Centro Nacional de Supercomputación).

[4] Molina, L., & Molina, M. (2004). Megacities and Atmospheric Pollution. *ISSN 1047-3289. Air & Waste Management Association. 54*, 644-680.

[5] Chuquer, D. (2014). Transporte de contaminantes atmosféricos entre la Zona Metropolitana del Valle de México y la Zona Metropolitana del Valle de Toluca. Ciudad de México, México: Tesis de maestría, UNAM.

[6] Luna, C. (2015). Estudio de calidad del aire de la Zona Metropolitana de Cuernavaca y la posible aportación de contaminación proveniente de la Zona Metropolitana del Valle de México. Ciudad de México, México: Tesis de maestría, UNAM.

[7] García, J., García, J., Jazcilevich, A., & Ruiz, L. (2014). The influence of the Tula, Hidalgo complex on the air quality of Mexico City Metropolitan Area. *Atmósfera, Vol. 27, N° 2*, 215-225.

[8] Holzworth, G. C. (1972). *Mixing Heights, Wind Speeds, and Potential for Urban Air Pollution Throughout the Contiguous United States. AP-101. U.S. Environmental Protection Agency, Research Triangle Park, NC.*

[9] National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce (2000), NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999, http://dx.doi.org/10.5065/D6M043C6, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, Colo. (Updated daily.) Accessed† 07 03 2016.

Enhanced Monte Carlo Methods for Sparse Approximate Matrix Inversion

Oscar A. Esquivel-Flores1, Diego Dávila2, Vassil Alexandrov3
1ITESM-MTY, Mexico, 2UPC, Spain, 3ICREA-BSC
1oscar.esquivel@bsc.es, 2diego.davila@est.fib.upc.edu, 3vassil.alexandrov@bsc.es

Abstract-*Novel scalable mathematical methods and algorithms for fundamental linear algebra problems such as solving Systems of Linear Algebraic Equations and/or matrix inversion with focus on large scale systems are subject of intensive study. This research presents an enhanced parallel Monte Carlo method and algorithm for computationally intensive problems such as sparse approximate matrix inversion..*

I. INTRODUCTION

Solving systems of linear algebraic equations (SLAE) or inverting a real matrix are well-known Linear Algebra problems. Iterative or direct methods to solve these systems may be a costly approach in some cases and/or for large systems. One option of reducing the effort of solving these systems is to use preconditioners before applying an iterative method. Standard deterministic preconditioners can be computed by using the optimized parallel variant of SPAI- the Modified SParse Approximate Inverse Preconditioner (MSPAI). One approach is to replace MSPAI with a Monte Carlo preconditioner that relies on the use of Markov Chain Monte Carlo (MCMC) methods [1].

Key physical problems imply solving a SLAE arising as a result of discretization of partial differential equations. Iterative solvers are often the method of choice due to their predictability and reliability when considering accuracy and speed. They, however, may be prohibitive for large-scale problems as they can be very time consuming to compute. Iterative Methods are dependent on the size of the matrix and so the computational effort grows with the problem size. The complexity of these methods is $O(kn^2)$for dense matrices [2] in the iterative case and $O(n^3)$ for direct methods with dense matrices while solving SLAE if common elimination are employed [3] On the other hand, Monte Carlo (MC) methods, performing random sampling of a certain variable whose mathematical expectation is the desired solution, depend linearly on the matrix size and might lead to efficient solution for some problems where an estimate is sufficient or even favorable, due to the accuracy of the underlying data. MC methods can quickly yield a rough estimate of the solution. Note that MC methods for matrix inversion(MI) only require $O(NL)$ steps to find a single element or a row of the inverse matrix. Here N is the number of Markov chains and L is an estimate of the chain length in the stochastic process. These computations are independent of the matrix size n and also inherently parallel. Note that in order to find the inverse matrix
or the full solution vector in the serial case, $O(nNL)$steps are required.

II. RELATED WORK

Research efforts in the past have been directed towards optimizing the approach of sparse approximate inverse preconditioners (SPAI) [4]. There have been differing approaches and advances towards a parallelization of the SPAI preconditioner. The method that is used to compute the preconditioner provides the opportunity to be implemented in a parallel fashion. A class of Forbenius norm minimizations that has been used in the original SPAI implementation [5] was modified and is provided in a parallel SPAI software package. One implementation of it, by the original authors of SPAI, is the Modified SParse Approximate Inverse (MSPAI) [6].

The proposed Monte Carlo algorithm has been developed and enhanced in the past decade, and several key advances in serial and parallel Monte Carlo methods for solving such problems have been made. There is an increased research interest in parallel Monte Carlo methods for Linear Algebra in the past year [7], [8], due to their ability to find quickly the approximate solution of the matrix inverse or the SLAE.

III. MONTE CARLO APPROACH

Monte Carlo methods are probabilistic methods, that use random numbers to either simulate a stochastic behaviour or to estimate the solution of a problem. They are good candidates for parallelization because of the fact that many independent samples are used to estimate the solution. These samples can be calculated in parallel, thereby speeding up the solution finding process. We design and develop parallel Monte Carlo methods with the following main generic properties:

- efficient distribution of the compute data
- minimum communication during the computation
- increased precision achieved by adding extra refinement computations
- being naturally resilient and fault-tolerant

Consideration of all these properties naturally leads to scalable algorithms. The key Monte Carlo algorithm has been presented in [9] and allows to extend the Monte Carlo approach for processing both diagonally dominant and non-diagonally dominant matrices.

IV. EXPERIMENTS

We compared matrices from different sets that have been obtained from two collections - The Matrix Market and The University of Florida Sparse Matrix Collection as well as some real life problems from our scientific collaborators. These matrices are used as inputs to both the MSPAI and our Monte Carlo based application to compute the preconditioners. The results from those calculations are two intermediate matrices MSPAI and MC, one for each type of preconditioner. In the last step these preconditioners are used to create an equivalent SLAE solved by the GMRES implementation in Paralution-1.1.0 solver. Numerical

experiments have been carried out on the Marenostrum supercomputer, at Barcelona Supercomputing Center (BSC). It currently consists of 3056 compute nodes that are each equipped with 2 Intel Xeon 8-core processors, 64GB RAM and are connected via an InfiniBand FDR-10 communication network. The experiments have been run multiple times to account for possible external influences on the results. The computation times for both the preconditioner calculated by MSPAI, as well as our Monte Carlo preconditioner are given below. While conducting the experiments, we configured the parameters in both programs to produce preconditioners with similar properties and therefore producing residuals within similar ranges when used as preconditioners for GMRES.

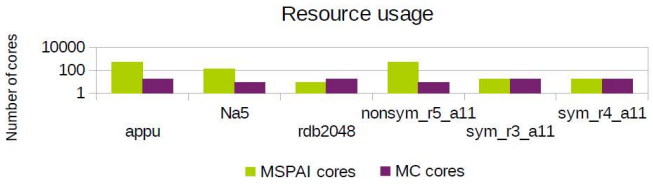


Fig. 1. Top: Run times for MC preconditioner and MSPAI (X – number of cores, Y – time in secs.) Middle: Best time resulting from the addition of the preconditioner time and the solver time (Y- time in secs.). Bottom: Number of cores used for each experiment in the previous middle plot.

A random starting pattern has been chosen in MSPAI for best analogy to the stochastic nature of the Monte Carlo approach. A basic experiment was carried out on various classes of matrices from the matrix market (see Fig 1).
The algorithms run for set or non-diagonally dominant non-symmetric and symmetric matrices as well as diagonally dominant ones. It is evident that in all the cases Monte Carlo preconditioner is obtained faster than MSPAI and in some cases MSPAI is not converging at all. (see Fig 1). In most of the cases the MC preconditioner also leads to similar or faster convergence of GMRES compared to GMRES convergence when MSPAI is used. However, there are problems such as r5_a11 with a non-symmetric matrix where MSPAI+GMRES perform better than MC+GMRES.

V. SOME CONCLUSIONS AND FUTURE WORK

A Monte Carlo based preconditioner for general matrices has been proposed as an alternative to the MSPAI algorithm and its applicability demonstrated. It has been shown that the stochastic Monte Carlo approach is able to produce preconditioners of comparable quality to the deterministic MSPAI algorithm. The proposed approach enabled us to generate preconditioners efficiently (faster), outperforming in many cases the deterministic approach, especially for larger problem sizes.

ACKNOWLEDGMENT

The author would like to thank CONACYT-México for supporting a postdoctoral position in BSC.

REFERENCES

[1] J. Strassburg and V. Alexandrov, "Enhancing Monte Carlo Preconditioning Methods for Matrix Computations", *Procedia Computer Science*, Vol. 29, pages 1580-1589, 2014.
[2] G.H. Golub, C. Loan, "Matrix computations", *Johns Hopkins University Press*, ISBN 9780801854149, 1996 .
[3] M. Ferronato, E.Chow, K.Phoon, "Preconditioning Techniques for Sparse Linear Systems", *Journal of Applied Mathematics*, Hindawi Publishing Corporation, Vol. 2012, article-id:518165, 2012.
[4] G. Allon, M. Benzi, and L. Giraud, "Sparse approximate inverse preconditioning for dense linear systems arising in computational electromagnetics", *Numerical Algorithms*, vol. 16, num. 1, pages 115, 1997.
[5] M. Benzi, C. Meyer, and M. Tuma, "A Sparse Approximate Inverse preconditioner for the Conjugate Gradient Method", *SIAM Journal on Scientific Computing*, Vol.5, pages 1135-1149, 1996.
[6] T. Huckle, A. Kallischko, A. Roy, M. Sedlacek, and T. Weinzierl, "An efficient parallel implementation of the MSPAI preconditioner, Parallel Computing", *Parallel Matrix Algorithms and Applications*, vol.36, num. 56, pages 273-284, 2010.
[7] V. Alexandrov, E. Atanassov, I. Dimov, S. Branford, A. Thandavan and C. Weihrauch, "Parallel Hybrid Monte Carlo Algorithms for Matrix Computations", LNCS, volume. 3516, pp. 752-759. Springer,2005.
[8] V. Alexandrov & A. Karaivanova, "Parallel Monte Carlo algorithms for sparse SLAE using MPI", LNCS, pages 283-290, Springer, 1999.
[9] S. Branford, "Hybrid Monte Carlo Methods for Linear Algebra Problems", PhD thesis, *School of Systems Engineering*, The University of Reading, April 2009.

Dynamic Load Balancing for hybrid applications

Marta Garcia Gasulla, Julita Corbalan and Jesus Labarta
Barcelona Supercomputing Center and Universitat Politecnica de Catalunya
marta.garcia@bsc.es, julita.corbalan@bsc.es, jesus.labarta@bsc.es

*Abstract-**The DLB (Dynamic***

DLB relies on the usage of hybrid programming models and exploits the malleability of the second level of parallelism to redistribute computation power across processes.

I. Introduction

In parallel computing, the loss of efficiency is an issue that concerns both system administrators and parallel programmers. The growth in number of computing units that clusters experienced the last years has helped speeding up applications but has worsened some problems that affect the efficient use of the computational power.

One of the problems that has deteriorated with this growth is load balance. Although it is a concern that has been targeted since the beginning of parallel programming, there is not a universal solution.

In this paper we will talk about the Load Balancing Library, DLB, and a balancing algorithm, LeWI, that can improve the performance of hybrid applications. DLB can load balance an application at runtime without modifying nor analyzing the application.

In a previous work [1] we showed the potential of DLB and LeWI when executed with MPI+OpenMP applications.

In this paper we are showing the results of porting DLB to OmpSs. And how integrating some features of DLB in the runtime the performance can be improved.

II. Dynamic Load Balancing Library (DLB)

The Library

The Dynamic Load Balancing (DLB) is a shared library that helps load balance applications with two levels of parallelism. The current version provides support for:

- MPI+OpenMP
- MPI+OmpSs

The aim of DLB is to balance the MPI level using the malleability of the inner parallel level. One of its main properties is that the load balancing will be done at runtime without analyzing nor modifying the application previously. The algorithm that has showed better performance results is LeWI (Lend When Idle) [1]. And this is the algorithm that we are going to explain in the following section and use for the performance evaluation.

LeWI Algorithm

The philosophy of LeWI is based on the fact that when an MPI process is waiting in an MPI blocking call none of its threads is doing useful work. Therefore, we have one or several CPUs that are not being used. LeWI aims to use these CPUs to speedup other MPI processes running in the same node. The usual behavior of an MPI application is that if a process is blocked in an MPI call it is waiting for one or several other processes to finish. Speeding up processes that are more loaded helps to load balance the application and speedup the whole application.

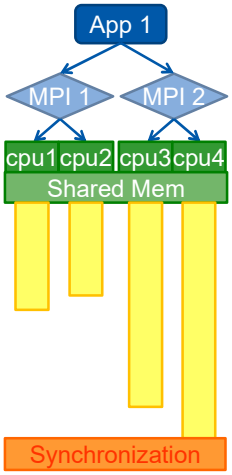


Fig. 1. LeWI Algorithm behavior: Original Application vs. Application load balanced with LeWI.

In Fig. 1 we can see the behavior of the LeWI algorithm when balancing an unbalanced application. On the left shows an unbalanced hybrid application with 2 MPI processes and 2 threads per process. In this example MPI process 2 is more loaded than MPI process 1 and this makes that MPI process 1 must wait in an MPI communication for some time.

At the right we can see the behavior of the same application when executed with the LeWI algorithm. When an MPI process reaches a blocking MPI call it will lend its CPUs to the other MPI processes running in the same node. With the lent CPUs the more loaded MPI processes will be able to finish its computation faster and the MPI process 1 will be less time waiting in the MPI call. The use of the computational resources will be better and the application
will perform better.

The first version of LeWI did not use mapping of threads to cpus, it adjusted the total number of running threads. But with the porting of DLB to OmpSs this offers us the posibility of mapping each thread to a cpu and lending a specific cpu, avoiding a temporal oversuscription.

III. Performance Evaluation

The experiments have been executed on Marenostrum3. Marenostrum3 is based on Intel SandyBridge processors. Its compute nodes are IBM iDataPlex dx360 M4 X servers with two 8-core Intel Xeon processors (E5-2670) per node and 32 GB of shared memory. They also include a hard drive of 500Gb and an MPI network card Mellanox ConnectX-3 Dual Port QDR/FDR10 Mezz Card. For management and GPFS they have two Gigabit Ethernet network cards.

We have executed the BT-MZ a benchmark from the NAS-Multizone benchmark suite. BT-MZ has been executed in one node of Marenostrum (16 cpus) with different configurations of MPI processes and threads.

And Lulesh a mini-app representative of simplified 3D Lagrangian hydrodynamics on an unstructured mesh. Lulesh has a parameter that can be changed to increase or decrease the amoount of imbalance present in the execution. A low value means a good load balance and a high value means more imbalance. Lulesh has been executed in 4 nodes of Marenostrum (64 cpus).

For each execution we can see four different series:

- **Binding**: the original execution of the application without load balancing executed with mapping of threads to cpus.
- **No Binding:** the original execution of the application without binding of threads to cpus.
- **No Binding + LeWI:** Execution with LeWI and without binding of threads.
- **Binding + Mask:** Execution with LeWI and with mapping of threads to cpus.

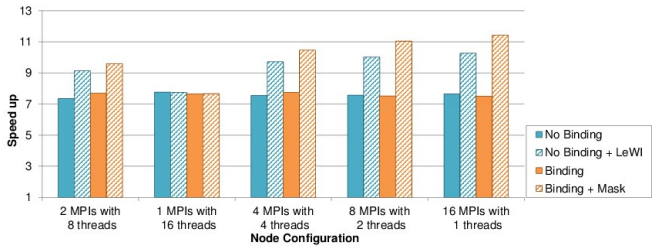


Fig. 2. Speed up obtained by BT-MZ with and without LeWI

In Fig. 2 we can see the speed up obtained by the different executions, when using the load balancing algorithm LeWI we can improve the speed up of the application. But the gain can be higher using a mapping of threads to cpus.

Fig.3. Speed up obtained by Lulesh with and without LeWI

Fig. 3 shows the speed up of Lulesh with a different amount of load imbalance. We can see how the speed up of Lulesh decreases as the amount of load imbalance increases, but when using LeWI the performance is better and maintained independently of the amount of imbalance.

We can se also that the performance when using a mapping of threads to cpus is better when using dynamic load balancing than when not mapping threads to cpus.

IV. Conclusions

In this paper we have presented a load balancing algorithm, LeWI, that has been implemented within a dynamic library, Dynamic Load Balancing (DLB).

The DLB library allows us to balance applications with two levels of parallelism without modifying the application or studying the imbalance it presents. The current version of the library can balance hybrid MPI+OpenMP and MPI+OmpSs applications.

We have shown the relevance of binding of threads to cpus. And how the support from the runtime can help load balance applications.

References

[1]M. Garcia, J.Corbalan, J.Labarta, "LeWI: A Runtime Balancing Algorithm for Nested Parallelism" International Conference on Parallel Processing, ICPP 2009.

[2]I. Karlin, J. Keasler, R. Neely. LULESH 2.0 Updates and Changes. August 2013, pages 1-9

Effects of detailed ventricular anatomy on the blood flow

Federica Sacco[1] federica.sacco@upf.edu , Bruno Paun[1], Mohammad Jowkar[2], Guillaume Houzeaux[2], Mariano Vázquez[2], Jazmin Aguado-Sierra[2] jazmin.aguado@bsc.es , Constantine Butakoff[1] constantine.butakoff@upf.edu
[1] Physense, Universitat Pompeu Fabra, Barcelona, Spain
[2] CASE Barcelona Supercomputing Centre (BSC), Barcelona, Spain

Abstract-*The presented study is a preliminary test and analysis of the role of trabeculae and papillary muscles in the hemodynamics of the left ventricle (LV).*

I. INTRODUCTION

The aim of the present study is to examine the role of trabeculae and papillary muscles in cardiac functionality. Trabeculae and papillary muscles are two tissue structures that project from the inner surface of the ventricular endocardium. The utility of papillary muscles has been related to valve function by pulling the chordae tendinae. However, little has been done to simulate the role of both papillaries and trabeculae in the overall cardiac electro-mechanics and hemodynamics [1, 2]. Most blood flow simulations consider a smooth ventricular surface [2, 3, 4, 5, 6], however are we sure that trabecular and papillary structures don't modify the blood flow pattern?

II. METHODS

A. LV Models

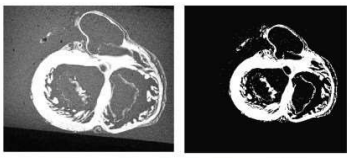
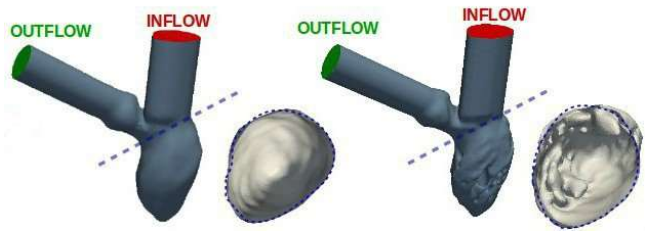


Fig.1: Segmentation of high resolution MRI of ex-vivo human hearts

From MR images of an ex-vivo human heart (Fig. 1) two LV models were created: a smooth-edocardium (Fig. 2, left) and a detailed-endocardium (Fig. 2, right)ventricle *Fig.2:*



Smooth LV and detailed LV models with boundary conditions

comprising trabeculae and papillary muscles. Tubes were attached at mitral and aortic valve levels to extend the inflow and outflow tracts.

B. Meshes and simulations

Iris, an in-house mesh generator was used to generate the two meshes and steady flow simulations were carried out with Alya (code developed at BSC) [7]. For the detailed geometry, a mesh of 1.886 million elements was created. For the smooth geometry two mesh resolutions were tested: a 362.740 and a 19.933 elements mesh. Peak physiological velocity was imposed at the inlet [8], zero pressure at outlet and rigid wall boundary conditions were considered with an approximate Reynolds number of 120.

III. RESULTS

For the smooth geometry case, CFD was solved at two resolutions to verify convergence and identify if the blood flow pattern could be influenced by the mesh resolution. The two smooth models showed that results were visually similar.
By analysing the fluid dynamics in the the smooth and detailed geometries, it can be seen that blood flow has a completely different pattern between them. The trabeculae and papillary muscles disturb the flow creating vortices at the apex (Fig. 3), and mitral valve level (Fig. 4) that are not present in the smooth case.

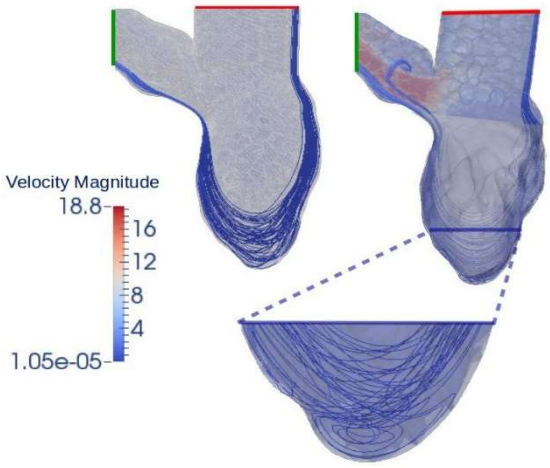


Fig.3: The smooth geometry (left) is characterized by complete laminar flow while in the detailed one (right) vortices can be seen at apex level

IV. LIMITATIONS

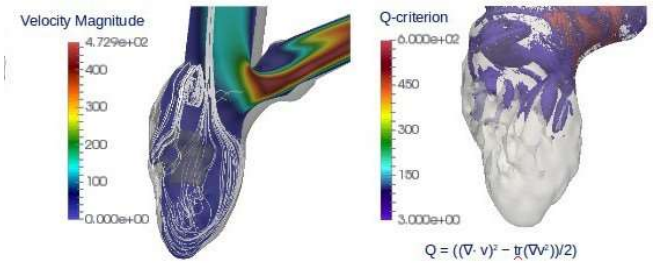


Fig.4: The detailed geometry is also characterized by vortices at the mitral valve level

In this preliminary work, rigid wall boundaries and steady flow conditions were considered, not taking into account physiological pulsatile flow and ventricular contraction during the cardiac cycle. Moreover, the simulations were done without taking into account the mitral and aortic valve.

V. FUTURE DEVELOPMENTS

Ventricular wall motion will be added as boundary conditions and pulsatile flow will be applied at inflow in order to simulate at best the physiological conditions in cardiac contraction. Valves will be attached and simulations will be carried considering valve motion.

ACKNOWLEDGMENT

This research was done with data taken from Visible Heart Lab and simulations were possible by using Alya code thanks to the collaboration with BSC.

REFERENCES

[1] Vedula V. et Al.: *Effect of trabeculae and papillary muscles on the hemodynamics of the left ventricle.* Theoretical and Computational Fluid Dynamics, pp. 1-19 (2015)

[2] Trayanova N. et Al: *Whole-Heart Modeling Application to Cardiac Electrophysiology and Electromechanics. Circulation Research* (2011)

[3] Nguyen V-T et Al.: *A Patient-Specific Computational Fluid Dynamic Model for Hemodynamic Analysis of Left Ventricle Diastolic Dysfunctio*n. Cardiovascular Engineering and Technology (2015)

[4] Choi Y. et Al: *A new MRI-based model of heart function with coupled hemodynamics and application to normal and diseased canine left ventricles*. Frontiers in Bioengineering and Biotechnology (2015)

[5] Lopez-Perez A. et Al: *Three-dimensional cardiac computational modelling: methods, features and applications*. BioMed Central (2015)

[6] Gurev V. et Al.: *Models of cardiac electromechanics based on individual imaging data*. Biomech Model Mechanobiol 10:295-306 (2011)

[7] Houzeaux G., Vázquez M. et Al.: *A massively parallel fractional step solver for incompressible flows*. Journal of Computational Physics 200. 228, 17:6316–6332 (2009) [8] Kulp S. et Al.: *Using High Resolution Cardiac CT Data to Model and Visualize Patient-Specific Interaction between Trabeculae and Blood Flow*. MICCAI, pp. 468 475(2011)

Probabilistic seismic risk assessment using CRISIS2015 & USERISK2015. Application to buildings of Barcelona, Spain.

Armando Aguilar1, Josep de la Puente1, Lluis Pujades2, Alex H. Barbat2, Mario Ordaz3, Nieves Lantada2, Amelia Campos4, Sergio N. González1, Alejandro García5, Alejandro Córdova5
Barcelona Supercomputing Center (BSC) (1); Technical University of Catalonia (UPC) (2); Engineering Institute, UNAM (3); Private Consultant (4); University of Veracruz (5)
aguilar.uv@gmail.com
josep.delapuente@bsc.es

Abstract-*The probabilistic models to assess seismic hazard and seismic risk incorporated into the codes CRISIS2015 & USERISK2015, respectively, are applied to compute the seismic risk of buildings of Barcelona. The main procedures required to assess the seismic risk using these codes are briefly described in the present document. A new version of USERISK, which is being developed in the Barcelona Supercomputing Center was used in the present work. According to the results, the levels of seismic risk of the Eixample District of Barcelona are important due mainly to the high levels of seismic vulnerability of its buildings.*

I. INTRODUCTION

The assessment of the seismic risk in urban zones is an essential task in order to take decisions oriented to increase the resilience levels of the cities [1].

A probabilistic version of the vulnerability index method (VIM_P) to compute seismic risk of buildings in urban areas was proposed by Aguilar et al [2]. This method is a modified version of the vulnerability index method (VIM) that was widely validated in the Risk-UE project [3, 4]. The VIM_P method can be applied by means of two codes: CRISIS2015 [5, 4], & USERISK2015 [6, 4]. A new version of this last code is being developed in the BSC. In the next sections more details about the theoretical background of both codes are included. Results computed in the present work for buildings of Barcelona are mentioned.

II. METHODOLOGY

In the VIM_P method three basic elements are considered to compute seismic risk: a) seismic hazard, b) seismic vulnerability and c) a seismic damage function. The seismic risk is determined when the convolution of the seismic hazard and the seismic vulnerability is computed [3]. For each building, the seismic risk assessed is expressed in terms of the annual frequency with which damage states D_k are exceeded $(\nu[D > D_k])$. For this purpose Eq. 1 is considered.

$$\nu[D > D_k] = \Sigma \ \Sigma P[D > D_k \mid V, I]P[V]\gamma'[I] \qquad (1)$$

where $\gamma'[I]$ is the annual frequency of occurrence of the seismic intensity [7]. $P[V]$ is the probability of occurrence of the seismic vulnerability. $P[D>D_k| I,V]$ is the probability that damage D_k is exceeded given that a seismic intensity I, and a seismic vulnerability V have occurred. In Eq. 1 the total probability theorem is applied and it is considered that the intensity I and the vulnerability V are independent random variables [2].

A. Seismic hazard

In the VIM_P method the seismic hazard must be computed using a probabilistic method, based on the Esteva and Cornell approach [3]. A modified version of this probabilistic approach is incorporated in the CRISIS2015 code which was selected as the standard code to compute probabilistic seismic hazard in the VIM_P method [3]. The seismic hazard results must be expressed in terms of annual frequencies of exceedance of macroseismic intensities.

B. Seismic vulnerability

In the VIM_P method the seismic vulnerability of a building is represented by means of probability density functions (PDFs) beta type [2]. These PDFs describe the variation of a vulnerability index that mainly varies in a range between 0 and 1. Values close to cero mean low seismic vulnerability, and values close to 1 mean high seismic vulnerability [2, 3].

In order to assess the seismic vulnerability of a building is necessary to know basic information about the building. For instance, it is necessary to know data as location, structural typology, construction year, position into the square, etcetera. The seismic vulnerability of buildings can be assessed using USERISK2015.

C. Seismic damage function

In the VIM_P method the seismic damage is assessed by means of a semi-empirical function [4], which allows computing a mean damage grade μ_D (Eq. 2). Additionally, in order to generate a complete distribution of the damage a binomial probability density function is considered. These expressions allow determining probability of occurrence for five damage states. In the damage state five the total destruction of the building occurs [3].

$$\mu_D = 2.5\left[1 + \tanh\left(\frac{I + 6.25V - 13.1}{2.3}\right)\right] \qquad (2)$$

where V is the vulnerability index, mainly with values between 0 and 1; I is the value of the macroseismic intensity EMS-98 [8].

D. Seismic risk

Data of seismic hazard and data of the building are used by USERISK2015 to compute the seismic vulnerability and the seismic risk of the studied buildings. The seismic risk results are expressed in terms of annual frequencies of exceedance of five no null damage states [3].

III. Application and Results

The VIM_P method was used to compute the seismic risk of 69982 dwelling buildings of Barcelona.

A. Seismic hazard

The seismic hazard of Barcelona was computed by means of CRISIS2015. The seismic hazard results are shown in Fig. 1. According to the seismic hazard results the macroseismic intensity equal to 6 has a probability of exceedance of 10% in 50 years. In other words the return period of the macroseismic intensity of 6.0 is equal to 475 years

.

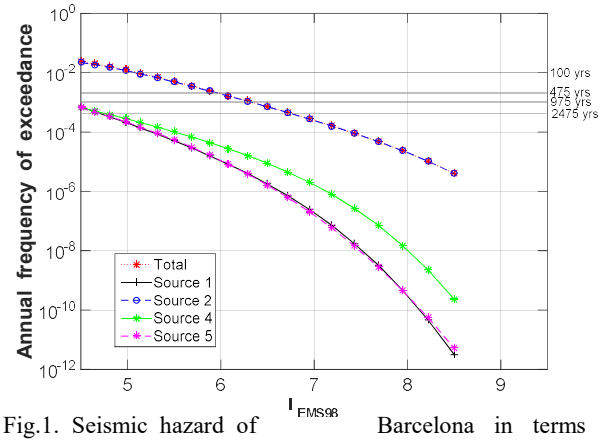


Fig.1. Seismic hazard of Barcelona in terms of annual frequency of exceedance of macroseismic intensities

B. Seismic vulnerability

Three seismic vulnerability curves were computed for each one of 69982 buildings of Barcelona. These curves can be used to obtain representative curves that characterize the seismic vulnerability of a group of buildings. Fig. 2 shows the representative curves of seismic vulnerability of 8432 buildings of the Eixample District of Barcelona. These curves of seismic vulnerability were computed without take into account regional vulnerability modifiers.

C. Seismic risk

USERISK2015 uses seismic hazard results (Fig.1) and seismic vulnerability results (Fig. 2) to compute the seismic risk of the buildings of the Eixample District (Fig.3).

Curves in Fig 3. represent the average seismic risk of 8432 residential buildings of the Eixample District of Barcelona. According to main curve of seismic risk (Fig.3), the buildings of the Eixample District can suffer, in average, a damage state of 1.5 each 475 years. In other words, in

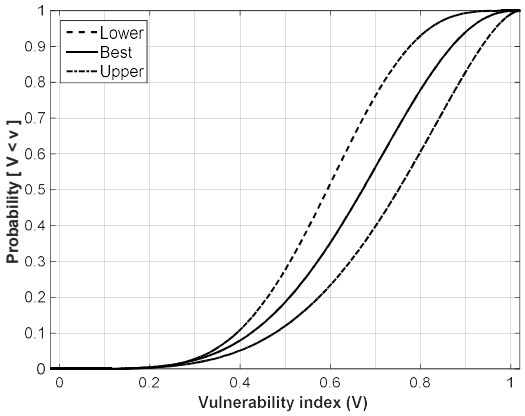


Fig.2 Representative curves that characterize the seismic vulnerability of 8432 residential buildings of the Eixample District

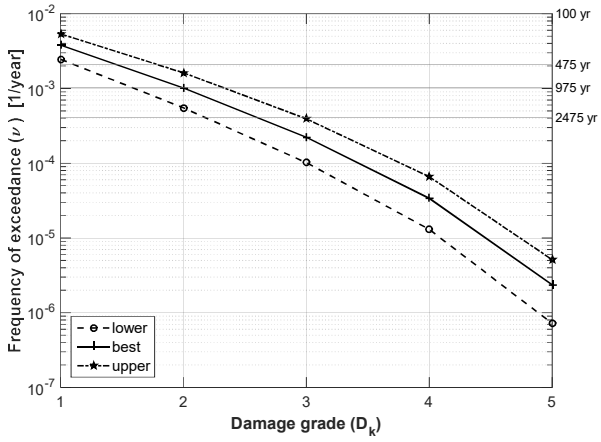


Fig. 3 Average seismic risk of the dwelling buildings of the Eixample District of Barcelona.

average, the damage state equal to 1.5 has a probability of exceedance of 10% in a period of 50 years.

However, if the regional vulnerability modifiers are considered then the damage grade that has a return period of 475 years is equal to 3.0.

.

IV. Conclusions

According to the results, it is possible to affirm that the procedure to assess seismic risk of buildings in urban areas using both codes CRISIS2015 and USERISK2015 is an appropriate procedure, because it is possible to obtain reasonable results in a reasonable time. For this reason, the appropriate use of the procedure described previously with its respective codes, can contribute to increase the resilience of cities of the world. On other hand, the seismic risk of the buildings of the Eixample District is important due mainly to the high vulnerability of many buildings of this district. At the same time, it is convenient verify the seismic vulnerability modifiers that can be used in any region. For the Barcelona case, by the moment, it is recommendable

consider that the minimum seismic risk corresponds to the case where the vulnerability modifiers are not considered, and the maximum seismic risk corresponds to the case where the vulnerability modifiers are considered.

ACKNOWLEDGMENT

A.A. thanks the support of CONACYT-BSC and UV. Thanks to the CONACYT, the Barcelona Supercomputing Center, PRODEP and the University of Veracruz by the support in this project. This research has been partially funded by the Ministry of Economy and Competitiveness (MINECO) of the Spanish Government and by the European Regional Development Fund (FEDER) of the European Union (UE) through projects referenced as: CGL2011-23621 and CGL2015-65913 -P (MINECO / FEDER, UE).

REFERENCES

[1] United Nations. How To Make Cities More Resilient. A Handbook For Local Government Leaders. Geneva, March, 2012. http://www.unisdr.org/files/26462_handbookfinalonlineversion.pdf Last accessed: 2016/01/20

[2] Aguilar, A. Pujades, L., Barbat, A., Lantada. N. 2010. A probabilistic model for the seismic risk of buildings: application to assess the seismic risk of buildings in urban areas. A: US National and Canadian Conference on Earthquake Engineering. "9th US National and 10th Canadian Conference on Earthquake Engineering". Toronto, p. 1-10.

[3] Aguilar, A. Evaluación probabilista del riesgo sísmico de edificios en zonas urbanas. PhD Thesis. Universidad Politécnica de Cataluña, 2011, 297 pp.

[4] Milutinovic, Z. V., Trendafiloski,G. S. 2003. WP4: Vulnerability of current buildings. RISK-UE. An advanced approach to earthquake risk scenarios with applications to different European towns, Contract: EVK4-CT-2000-00014, 109 pp.

[5] Ordaz, M., F. Martinelli, F., Aguilar, A., Arboleda, J., Meletti, C., D'Amico, V. 2015. CRISIS2015. Program for computing seismic hazard. Last accessed: 2016/02/18 https://sites.google.com/site/codecrisis2015/

[6] Aguilar, A., Pujades, L., Barbat, A., Lantada. N. 2015. USERISK2015. Program for computing seismic risk in urban areas. Last accessed: 2016/02/25. https://sites.google.com/site/userisk2015/

[7] McGuire RK. 2004. Seismic hazard and risk analysis. Earthquake Engineering Research Institute. MNO-10. 221 pp.

[8] Grünthal, G. 1998. European Macroseismic Scale 1998. Cahiers du Centre Europ'een de G'eodynamique et de S'eismologie. Luxemburg; 15: 1-99

DimLightSim: Optical/Electrical Network Simulator for HPC Applications

Hugo Meyer, Jose Carlos Sancho
Barcelona Supercomputing Center
{hugo.meyer; jose.sancho}@bsc.es

Abstract *– Optical Packet Switches (OPS) and Optical Circuit Switches (OCS) provide the needed low latency transmissions in today large data centers and HPC systems. These switches can deliver lower latency and higher bandwidth than traditional electrical-based switches. Although light-based transmission has its advantages over electrical-based transmissions, in optical devices packet collisions are possible and this can generate retransmissions. In this work we present an optical network simulator called DimLightSim. DimLightSim models communication events in optical devices at packet level by replaying real application traces. Different experimental evaluations have been made using DimLightSim in order to compare current datacenter networks with the fully optical Architecture-on-Demand (AoD) proposed in the Lightness project. Initial results helped to foresee the impact in HPC applications execution time. In terms of performance improvement, the AoD architecture can outperform Infiniband-based network up to 19%.*

I. IINTRODUCTION

Data centers are growing in size and complexity to accommodate the ever-increasing demand of High Performance Computing (HPC) applications. One of the most challenging issues when scaling out a data center is the network infrastructure. As the size of data centers increases, higher volumes of data have to be transported among thousands of servers very fast. It is predicted that applications will need in the order of several Terabit/s of bandwidth in the near future. In addition, to provide enough network bandwidth there is also need to provide fast access to data, specially to HPC applications, where, for many applications, low latency network is critical to achieve high scalability.

Optical-based network has currently been explored to overcome this bandwidth and latency bottleneck in data centers. The deployment of optical devices leverage on Dense Wavelength Division Multiplexing (DWDM) allows the transmission of more than a hundred of wavelength channels operating at 10, 40, 100 Gb/s and beyond.

Basically, current optical switching architectures are based on Optical Circuit Switching (OCS) and Optical Packet Switching (OPS). OCS is capable to accommodate long-lived high-capacity smooth flows with little latency whereas OPS is ideal for dynamic traffic such as HPC. In order to take advantage of the benefits from both OCS and OPS, a novel architecture-on-demand (AoD) function programmable data center network architecture with the integration of OCS and OPS was recently proposed [1]. AoD is allocating applications to different optical switches depending on the communication characteristic of applications. HPC applications are desirable to be allocated to OPS as their traffic is dynamic. The OCS may not be suitable for HPC traffic in some scenarios, since the mirror reconfiguration time of OCSs is around 25 ms. On the other hand, packet collisions may occur when using OPSs since packets cannot be stored in these switches.

In this work we describe an optical packet-level network simulator named *DimLightSim* that has been designed to model the behavior of fully-optical Datacenter Networks. *DimLightSim* have been designed in order to evaluate the impact of optical network components in HPC applications. *DimLightSim* allows foreseeing how variations in latencies, packet retransmissions, bandwidth, among others, affect execution time of applications.

II. DIMLIGHTSIM: PACKET LEVEL OPTICAL SIMULATION

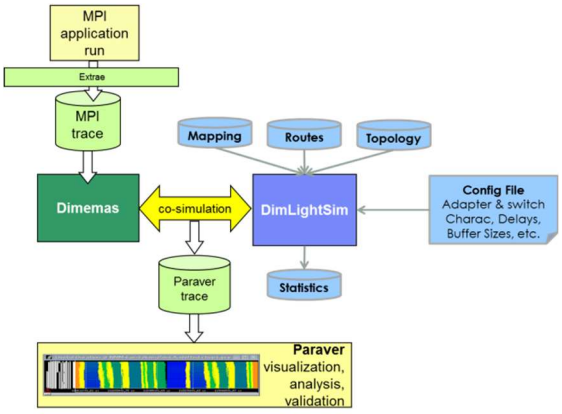


Fig. 1. Optical Network Simulator Framework.

The simulation framework of *DimLightSim* is composed of four open-source tools: a) *Extrae*: extracts information that includes timestamps of events such as message transmissions and other runtime calls; b) *Dimemas*: it reproduces the events from the trace. Communication events are forwarded to *DimLightSim*. c) *DimLightSim*: is the packet level network simulator that models optical devices. It has been developed using the Omnet ++ framework. d) *Paraver*: is a visualization and analysis tool of the computation and communication events.

Fig. 1 depicts the simulation framework and how the different elements interact. *DimLightSim* does the MPI process mapping, setup the preferred network topology, and the routing information as well in order to forward packets from source servers to destination servers. *DimLightSim* allows users to configure the network topology and the routing information in switches.

DimLightSim and *Dimemas* participate in a co-simulation where they take turns during the simulation execution. *Dimemas* starts the simulation execution processing the communication and computation events in the application

trace. *DimLightSim* is waiting for messages from *Dimemas*. When the next event in the application trace is a message transmission then *Dimemas* creates a message and forwards it to *DimlightSim*. *DimLightSim* proceeds with the simulation execution till the transmission is finished or it is interrupted by *Dimemas* because other communication event needs to be scheduled.

III. Experimental Evaluation

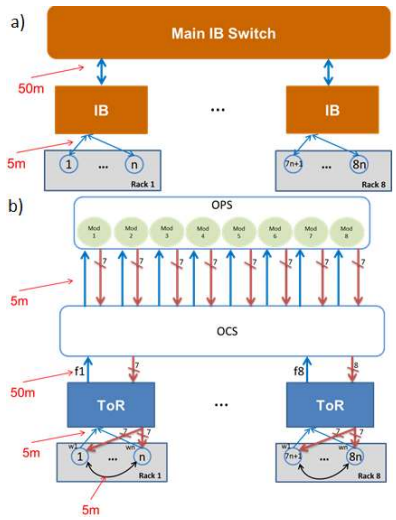


Fig. 2. Experimental Networks. a) IB network with 8 racks, 8 Top-of-the-Rack (ToR) switches and a main IB Switch. b) Optical AoD network with 8 racks, 8 ToRs connected to an OCS and one OPS.

DimLightSim has been validated in [2] by comparing real transmission times with simulated times. It is also very important to highlight that all the parameters of the devices used in *DimLightSim* correspond to real measures taken from the optical devices [3].

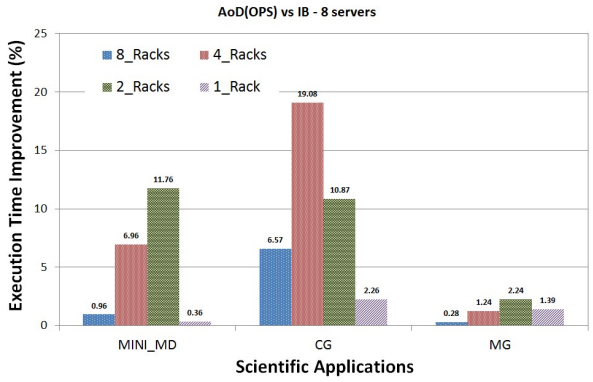


Fig. 3. Performance Comparison of OPS and Infiniband Switching using DimLightSim.

Fig. 2 shows the experimental configurations used to compare an IB-based network with an optical AoD network. Considering the experiments made using AoD infrastructure, NICs are able to communicate directly between each other when residing in the same rack and for communication between racks, all traffic goes through the OPS switch. Optical Top-of-the-Rack (ToR) switches are in charge of multiplexing/demultiplexing optical wavelengths to fibers in a negligible time. The next configuration parameters were set: NIC Delay=300ns; OPS Delay=25ns; IB switch Delay=200ns; Bandwidth = 8 Gbps. The cables lengths are depicted in Fig. 2.

Fig. 3 depicts results obtained using *DimLightSim* when comparing Infiniband (IB) switching with OPS assuming the usage of the AoD network infrastructure. According to the obtained results, OPS-based networks can outperform IB-based networks in up to 19% in terms of execution time. The showed results consider also the penalty of packet retransmission in the execution time when using OPS.

Further results and proposals that base their research in *DimLightSim* can be found in [3, 4].

IV. Conclusions

Foreseeing the impact of new network technologies in HPC applications is highly important in order adapt or configure properly these new improvements. In this work we presented *DimLightSim,* which is able to mimic the behavior of optical networks at a packet level and enable us to analyze how the technology improvements and its limitations affect execution of applications and resource usage. *DimLightSim* has been used in the Lightness project [5] and other works to drive design decisions and foresee performance impact in applications. Obtained results helped to determine the benefits of optical networks when comparing to electrical-based networks. In particular, OPS can outperform in up to 19% Infiniband-based networks.

Acknowledgment

This work has been supported by the FP7 European Project LIGHTNESS (FP7-318606).

References

[1] S. Peng, D. Simeonidou, G. Zervas, R. Nejabati, Y. Yan, Y. Shu, S. Spadaro, J. Perello, F. Agraz, D. Careglio, H. Dorren, W. Miao, N. Calabretta, G. Bernini, N. Ciulli, J. C. Sancho, S. Iordache, Y. Becerra, M. Farreras, M. Biancani, A. Predieri, R. Proietti, Z. Cao, L. Liu, S. J. B. Yoo, *A novel sdn enabled hybrid optical packet/circuit switched data centre network: The lightness approach*, in*: Networks and Communications* (EuCNC), 2014 *European Conference* on, 2014, pp. 1{5. doi:10.1109/EuCNC.2014.6882622.

[2] Hugo Meyer, Jose Carlos Sancho, Wang Miao, Harm Dorren, Nicola Calabretta, and Montse Farreras. *Performance Evaluation of Optical Packet Switches on High Performance Applications*. In Waleed W. Smari, editor, *Proceedings of the 2015 International Conference on High Performance Computing & Simulation* (HPCS 2015) Amsterdam, the Netherlands, pages 356–363. IEEE Computer Society, 2015.

[3] Wang Miao, Jun Luo, Stefano Di Lucente, Harm Dorren, and Nicola Calabretta, "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," *Opt. Express* 22, 2465-2472 (2014).

[4] Hugo Meyer, Jose Carlos Sancho, Milica Mrdakovic, Shuping Peng, Dimitra Simeonidou, Wang Miao, and Nicola Calabretta. *Scaling architecture-on-demand based optical networks.* In *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ICDCN '16, pages 10:1–10:10, New York, NY, USA, 2016. ACM.

[5] *Low Latency and High Throughput Dynamic Network Infrastructure for High Performance Datecentre Interconnects* (Lightness) European Project, 2012 [online]. Available: www.ict-lightness.eu.

Comparing electoral campaigns by analysing online data

Javier A. Espinosa-Oviedo[156], Genoveva Vargas-Solar[256], Vassil Alexandrov[134], Géraldine Castel[7]
[1] *BSC, Barcelona Supercomputing Centre, Spain*
[2] *CNRS, French Council of Scientific Research, France*
[3] *ICREA, Catalan Institution for Research and Advanced Studies, Spain*
[4] *ITESM, Tecnológico de Monterrey, Mexico*
[5] *LAFMIA, French-Mexican Laboratory of Informatics and Automatic Control, France*
[6] *LIG, Laboratory of Informatics of Grenoble, France*
[7] *Université Stendhal, Grenoble 3, France*
{espinosa, gvargas}@imag.fr, vassil.alexandrov@bsc.es, geraldine.castel@u-grenoble3.fr

Abstract- *The use of information and communication technologies (ICT) in the political sphere is nowadays a key aspect for running electoral campaigns. Thus, our work addresses the influence of ICT and candidate practices during electoral campaigns. Our approach is based in the collection of data produced by political candidates so that experts can analyse them through an analytics processes. Accordingly, this paper presents results concerning three of the data collections life cycle phases: collection, cleaning, and storage. The result is a data collection ready to be analysed for different purposes. The paper also describes our experimental validation for comparing political campaigns behaviour in France and the United Kingdom during the European elections in 2014.*

I. INTRODUCTION

The use of ICTs for political purposes is a relatively new research field as the first publications date from 1980s [1], [2]. This continuous evolution of tools has been paralleled by a shift in attention from sites to forums [3], [4], blogs [5], [6], or social networks [7], [8]. The problem of integrating different data sources for supporting the analytics processes is not new in the database domain [9]. Most proposals assume that data providers (heterogeneous or not) are known in advance [10] and thus integration is based on knowledge about the data structure [11], content, semantics [12] and constraints. However, the emergence of new kinds of data providers like services (e.g. Twitter, Facebook), where there are no schemas, introduced new challenges [11]. Data also started to acquire "new" properties (more volume, velocity, variety) and with them emerged the need of building huge curated data collections [13], [14]. The challenge is thus to collect political data continuously [15] in order to analyze the influence of ICT during electoral campaigns.

II. OVERVIEW OF THE APPROACH

Figure 1 shows the overview of our approach. This process is recurrently executed since new data is produced.

1.Data collection. Data is collected according to different modes (push, pull) and at different rates when data are produced continuously. In the case of Web pages and blogs we collect their content using crawling tools and Web scrapping techniques.

2.Data analytics. For each attribute of a given data structure we identify the distribution of the values within the collected data. We consider some level of uncertainty so we identify missing values and infers some proposals based on computed values distributions (e.g., using extrapolation), as well as discovering possible relations among data attributes (e.g., equivalence, functional dependency, temporal or casual correlations). This phase generates views that provide an abstract representation of the data collection contents.

3.View storage. Depending on the characteristics of the data, views can be materialized and stored together with raw data. These decisions consider the probability of data to be accessed and processed together based on their possible dependencies. Data organization can ensure performance and reduction of memory and communication resources consumption during the data analytics processes.

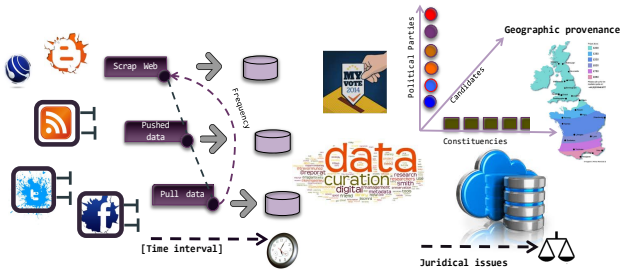


Figure 1 Data collection and curation overview

III. BUILDING AND MAINTAINING DATA VIEWS

The main idea of our approach is not to transform collected data but to generate an abstract aggregated view and then tag it with information that can be used for further data processing tasks. In this sense views can be seen as a kind of schema in the relational world, but extracted *a posteriori* after having created a database. As shown in the class diagram of figure 2, a View characterizes the content provided by a given dataProvider as a document composed of set of attributes, where an Attribute provides a snapshot of a given attribute's values domain for a given dataset. An attribute within the dataset has maximum and minimum values, a standard deviation of the values assigned to the attribute in the different documents collected in the dataset, and the variation of values across the dataset elements represented by a histogram. Within a dataset an attribute can

have null and missing values that must be inferred in order to characterize its domain type as precisely as possible. Indeed, many data collections represent missing values by dummy values and therefore we want to represent those cases.

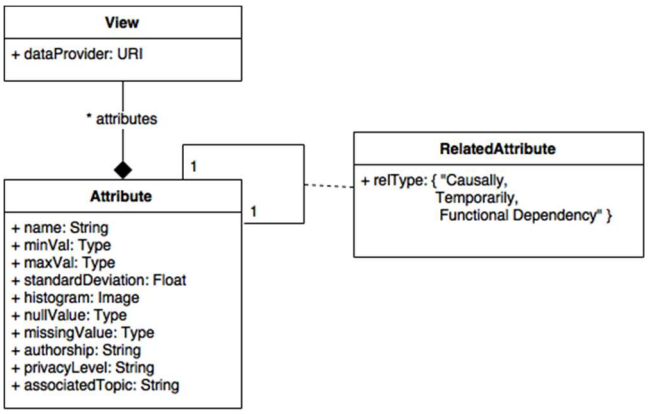


Figure 2 UML class diagram of the concept of View

IV. Validation

We built a system to analyse and compare campaigns in UK and France of the European elections in 2014 based on our approach (cf. We collected 30Gb of data comprising 12 parties and 100 candidates in France and UK, and they concern only online activities reported in Twitter, Facebook and official sites, pages and blogs. We used JSON as data model and we then implemented document processing tasks to characterize the content of collected data.

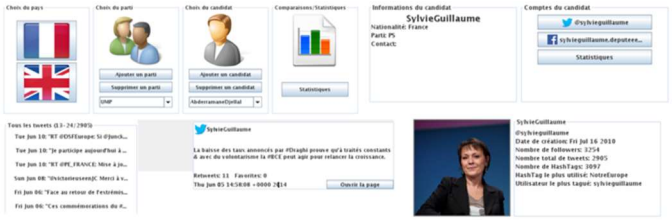


Figure 3 Profiling a candidate's campaign on social networks

V. Conclusions

This paper introduced our approach for building and curating political data collections and preparing them for the analytics process. Our first contribution in this paper regards the strategies used for characterizing and inferring data content through the notion of view. Some inference had to deal with uncertainty that we addressed associating accuracy probabilities to inferences so as to guide the data scientist in her further data analytics design.

References

[1] J. B. Abramson, F. C. Arterton, and G. R. Orren, "The Electronic Commonwealth: The Impact of New Media Technologies on Democratic Politics," *Michigan Law Review*, vol. 87, no. 6, p. 1393, May 1989.

[2] J. D. H. Downing, "Computers for Political Change: PeaceNet and Public Data Access," *Journal of Communication*, vol. 39, no. 3, pp. 154–162, Sep. 1989.

[3] S. Wojcik, "Les forums électroniques municipaux,espaces de débat démocratique ?," *Sciences de la Société: Démocratie locale et Internet*, no. 60, pp. 107–125, 2003.

[4] M. Marcoccia, "Les webforums des partis politiques français : quels modèles de discussion politique ?," *Mots Les langages du politique [En ligne], URL: http://motsrevuesorg/512*, pp. 49–60, 2006.

[5] F. Greffet, "Les blogues politiques. Enjeux et difficultés de recherche à partir de l'exemple français," *Communication Information médias théories pratiques*, vol. 25, no. 2, pp. 200–211, 2007.

[6] S. Gadras, "Public Sphere and Political Communication: how Does the Public Sphere Evolves with the Development of ICTs in French Local Politics?," in *The European Public Sphere: From critical thinking to responsible action*, Brussels, Peter Lang, 2012.

[7] N. Jackson and D. Lilleker, "Microblogging, Constituency Service and Impression Management: UK MPs and the Use of Twitter," *The Journal of Legislative Studies*, vol. 17, no. 1, pp. 86–105, Mar. 2011.

[8] M. Margaretten and I. Gaber, "The Crisis in Public Communication and the Pursuit of Authenticity: An Analysis of the Twitter Feeds of Scottish MPs 2008-2010," *Parliamentary Affairs*, vol. 67, no. 2, pp. 328–350, Apr. 2014.

[9] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, vol. 6, no. 11, pp. 1245–1248.

[10] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *Proceedings of the VLDB Endowment*, vol. 3, no. 1–2, pp. 1358–1369, Sep. 2010.

[11] V. Cuevas-Vicenttín, J. L. Zechinelli-Martini, and G. Vargas-Solar, "Andromeda: Building e-Science Data Integration Tools," in *Proc. of the 17th Int. DEXA Conference*, 2006, pp. 44–53.

[12] F. Osborne and E. Motta, "Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks," in *Proc. of the 14th Int. Semantic Web Conference (ISWC 2015)*, 2015, pp. 408–424.

[13] M. Adiba, J. C. Castrejón, J. A. Espinosa-Oviedo, G. Vargas-Solar, and J.-L. Zechinelli-Martini, "Big Data Management: Challenges, Approaches, Tools and their limitations," in *Networking for Big Data*, CRC Press, 2015.

[14] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012.

[15] M. Ma, P. Wang, and C.-H. Chu, "Data Management for Internet of Things: Challenges, Approaches and Opportunities," in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, 2013, pp. 1144–1151.

Integrated approach to assignment, scheduling and routing problems

Laura Hervert-Escobar, Francisco López-Ramos, Oscar A. Esquivel-Flores
Instituto Tecnológico de Estudios Superiores de Monterrey
laura.hervert@itesm.mx

Abstract-*This research considers a real life case study that determines the minimum number of sellers required to attend a set of customers located in a certain region taking into account the weekly schedule plan of the visits, as well as the optimal route. The problem is formulated as a combination of assignment, scheduling and routing problems. In the new formulation, case studies of small size subset of customers of the above type can be solved optimally. However, this subset of customers is not representative within the business plan of the company. To overcome this limitation, the problem is divided into three phases. A greedy algorithm is used in Phase I in order to identify a set of cost-effective feasible clusters of customers assigned to a seller. Phase II and III are then used to solve the problem of a weekly program for visiting the customers as well as to determine the route plan using MILP formulation. Several real life instances of different sizes have been solved demonstrating the efficiency of the proposed approach.*

I. INTRODUCTION

Network models and integer programs are applicable to an enormous known variety of decision problems. In a real life, the cost efficient management decision is defined by a combination of different models.

Consider a set of customers ***C={1,2,...,i,...,j,...N}*** dispersed in a given region where their locations are given by coordinates ***(gx_j, gy_j).*** It is required to design a business plan that includes the minimum number of sellers ***Y={1,2,...,s,...Y}*** to attend these customers, in days ***D={1,2,3,4,5,6}*** denoted by index *t* in the scheduling plan per week, providing the optimal daily routing. The decision should consider the demand ***(Dem)*** and the service time ***(T_i)*** of the customers. Also, the daily capacity *(Cap)* and available time of the sellers ***(T_s)***. Decision variables of the model are as follows:

Y_i^s Binary variable denoting whether customer *i* is assigned to seller *s*

V_{it}^s Binary variable denoting whether seller *s* visits a customer *i* at day *t*

X_{ij}^{st} Binary variable denoting whether customer *i* is visited before customer *j* by seller *s* at day *t*

e_i^{st} Continuous variable denoting the order in which customer *i* is visited in the route plan of seller *s* during day *t.*

The formulation of the matemathical model is as follows:

$$min \sum_{s \in S} \sum_{i \in} 2^{s-1} Y_i^s + \sum_i \sum_j \sum_s \sum_t d_{i,j} \cdot X_{i,j}^{s,t} \qquad (1)$$

Subject to:

$$\sum_{s \in S} Y_i^s = 1, \quad \forall i \qquad (2)$$

$$V_{i,t}^s \le Y_i^s, \quad \forall i, t, s \qquad (3)$$

$$X_{i,j}^{s,t} \le V_{i,t}^s, \quad \forall i, j, t, s, i \ne j \in N \qquad (4)$$

$$X_{i,j}^{s,t} \le V_{j,t}^s, \quad \forall i, j, t, s, i \ne j \in N \qquad (5)$$

$$\sum_i X_{i,j}^{s,t} = \sum_j X_{i,j}^{s,t}, \quad \forall i, j, t, s, i \ne j \in N \qquad (6)$$

$$\sum_i X_{i,j}^{s,t} + \sum_j X_{i,j}^{s,t} = 2V_{i,t}^s, \quad \forall i, j, t, s, i \ne j \in N \qquad (7)$$

$$\sum_i T_i \cdot V_{i,t}^s \le T^s, \quad \forall t, s \qquad (8)$$

$$\sum_t \sum_s V_{i,t}^s = \left\lceil \frac{Dem_i}{6 * Cap} \right\rceil, \quad \forall i \qquad (9)$$

$$V_{i,t}^s + V_{i,t+1}^s \le 1, \quad \forall i, s, t \le 5, \; Freq_i \le 3 \qquad (10)$$

$$e_i^{s,t} - e_j^{s,t} + MX_{i,j}^{s,t} \le M - 1, \quad \forall i, j, t, s, i \ne j \in N \qquad (11)$$

$$e_i^{s,t} \le \sum_j V_{j,t}^s, \quad \forall i, j, t, s, i \ne j \in N \qquad (12)$$

The objectivefunction (1) represents the sum of two goals, the minimization of the number of sellers required to service the customers and the minimization of the traveling distance to visit each customer for each routing plan.

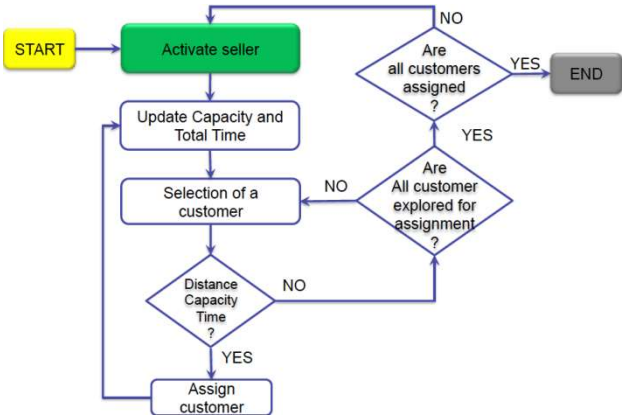
As for constraints, (2) ensures that a customer is attended by only one seller. Equation (3) guaranties that a customer is assigned to the seller that actually visits that customer. Equations (4) and (5) link the scheduling variables to the routing ones. Equations (6) and (7) are used for connectivity purposes. Next equation (8) ensures that the available time of the seller is not compromised during the scheduling of visits to the customers assigned per day. In this way, equation (9) establishes the number of visits to carried out per customer according to the given frequency. Equation (10) avoids consecutive visits to those customers whose frequency is less than 4 visits per week. Finally, equations (11) and (12) allow to assign the proper order of visits to customers during the routing plan to avoid sub-tours.

II. SOLUTION METHOD

The solution method for the problem is divided into three phases. The Phase I find a cluster of customers using the nearest neighbor approach. This objective is known as the tightest cluster of *m* points. This is similar to the one facility version of the max-cover problem [1], for planar models [2],

for one facility [3], and for several facilities [4], where we wish to find the location of several facilities which cover the maximum number of points within a given distance. The procedure is illustrated in Fig. 1.

Fig. 1. Greedy algorithm for phase I, the assignment of customers to sellers.



After defining the clusters of customers, the sequence of visits for each seller (cluster) is determined by solving a scheduling problem (phase II). Finally, the routing is solved per day and seller in phase III.

Caceres et al. [5] present a survey on VRPs apply to real life problems. A classification that applies for this case study is Multi-Period/Periodic VRP with Multiple Visits/Split deliveries. In this classification, the clients are visited several times as vehicles may deliver a fraction of the customer's demand. Moreover, optimization is made over a set of days, considering a different frequency of visits to each client.

III. Results

To test the performance of the proposed models, several instances were tested. The data for each instance correspond to a real life case consisting of a soft-drinks manufacturer. The solving time is an important issue for the company due to the deadline to generate the business plan each week. Therefore, the results are given in terms of both objective functions as well as solving times. The greedy algorithm, which determines the total number of sellers needed to satisfy the customers demand, was implemented in C++ 9.0.21. The scheduling and routing models were implemented using AMPL to call the optimizer CPLEX v.12.6.0. A time limit of 3600 sec is used as a stopping criteria when scheduling and routing are jointly solved. The cover area criteria for the greedy algorithm was set to 10 km.

The results are given in table 1. For each combination of territory and seller, the table provides the total number of customers per territory, the total number of required sellers, the optimal solution provided by the scheduling solution from the three-phase approach (OF(Scheduling)) and the computational times of both approaches.

TABLE I
Results of solution methods

Territory	TypeSeller	#Cusm	Sellers	OF(Scheduling)	T_{CPU} (2-p/3-p)
T1	D	15	5	1207	0.03/0.06
T4	D	17	6	1446	0.04/0.06
T3	AS	26	2	8472	0.19/0.4
T2	AS	33	5	5426	0.1/0.51
T1	B	37	1	13957	0.28/2.46
T4	A	59	7	6467	0.11/0.22
T3	F	112	3	28118	0.49/14.25
T1	C	163	5	26784	0.47/11.59
T4	F	243	6	28985	0.36/16.75
T3	C	405	11	38512	0.57/14.5
T2	E	1645	6	103299	4.26/611.46

The results shows that the total number of required sellers is not related to the size of the instance but to a combination of distance and demand of the customers. On the other hand, the objective function of the scheduling increases with the customers per territory. Concerning the computational time, it should be noted that the three-phase approach is faster than the two-phase one. Moreover, the three-phase approach achieves the same quality of the solution or even better. The computational performance improves with tight time windows and high node geographical density. Due to the use of the greedy algorithm, the critical size of the cluster-based MILP formulation significantly decreases and the hybrid approach becomes much more efficient.

IV. Conclusions

The described approach allows tackling the uncertainties stemming from practical problems such as different sizes of territory and particular features of the demand such as the distance and the service time. Future research lines include the development of a metaheuristic for further improving the solution provided by the three-phase approach, as well as the addition of stochastic data to represent a raise/fall in the clients demand and the appearance/loss of clients. Moreover, this approach is better suited for parallel implementation for larger problems.

ACKNOWLEDGMENT

This research has been submitted for presentation at ICCS 2016.

REFERENCES

[1] Daskin M (1995) Network and discrete location: models, algorithms and applications. Wiley, New York.

[2] Current J, Daskin M, Schilling D Discrete network location models. Ch. 3 In: Drezner Z, Hamacher HW (eds) Location analysis: applications and theory (2002), 81-118.

[3] Drezner Z (1981) On a modified one-center model. Manage Sci 27:848 - 851.

[4] Watson-Gandy CDT (1982) Heuristic procedures for the M-partial cover problem on a plane. EurJ Oper Res 11:149-157.

[5] J. Caceres-Cruz, P. Arias, D. Guimarans, D. Riera, and A. Juan. 2014. Rich Vehicle Routing Problem: Survey. ACM Comput. Surv. 47, 2, Article 32 (December 2014)

Innovative Algorithm for Particles Transport in a Fluid

Edgar Olivares*, Guillaume Houzeaux*
* *Barcelona Supercomputer Center (BSC - CNS)*
edgar.olivares@bsc.es

Abstract-*Particles transport in a fluid simulations have plenty of applications in the medicine or different fields of the engineering; from drug delivery simulation in the respiratory system to the friction of a car's break with its wheels or the icing of water droplets on a wing. But its implementation has also very different possible approaches: depending on the fluid, the size of the particle and the number of particles, literature proposes different solutions. In this paper, we want to show a generalized solution and compare it with proposed algorithms in the literature.*

I. INTRODUCTION

Particles in a fluid are transported because of the action of different forces. Depending on the case, gravity, buoyancy, Coriolis, Brownian motion or other forces may become necessary. Although involved forces may vary in every problem, drag force [1] and lift force [2] become essential when transported by a fluid.

Let F_p, a_p and m_p be the force, acceleration and mass of particle p. Applying the Newton's second law, the total acceleration applied on each particle is given by the summatory of all the forces involved

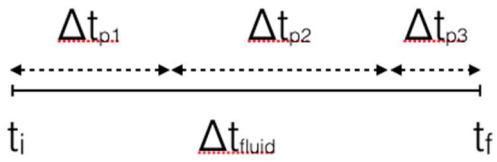
$$\Sigma F_p = m_p a_p \quad (1)$$

The calculation of a_p every step from initial time t_i to final time t_f requires an integration scheme. The time interval Δt of every step will be

$\Delta t = t_f - t_i$.

II. INTEGRATION SCHEME

The proposed integration scheme is a semi-implicit Newmark-β[3]. In this scheme, the actualization of the velocity u_{n+1} and position x_{n+1} of the next time step is given by two equations:



$$u_{n+1} = u_n + [(1-\gamma)a_n + \gamma a_{n+1}]\Delta t \quad (2)$$

$$x_{n+1} = x_n + u_n\Delta t + [(1-2\beta)a_n + 2\beta a_{n+1}]\Delta t^2/2 \quad (3)$$

Where β and γ are constants. If β=1/4 and γ=1/2 the method is implicit unconditionally stable and acceleration within the time interval Δt is presumed to be constant. If, otherwise, a linear variation of the acceleration during the time time interval is assumed, then the values will be β=1/6 and γ=1/2. As far as these values are the most commonly used in our simulations, becoming in both cases in an implicit method, a Newton-Raphson is needed in order to solve the dependence on u_{n+1}.

Let u_{n+1} be the function whose root is desired. The Newton-Raphson is described in this case by:

$$u_{n+1} = u_n - w(u_n) \quad (4)$$

Where,

$$w(u_n) = f(u_n)/f'(u_n) \quad (5)$$

And

$$f(u_n) = u_{n+1} + [(1-\gamma)a_n + \gamma a_{n+1}]\Delta t - u_n \quad (6)$$

$$f'(u_n) = -1 + \Delta t\gamma\, da/du|_{n+1} \quad (7)$$

To ensure the convergence

$$||w(u_n)|| / ||u_n|| < \varepsilon_c \quad (8)$$

is imposed. ε_c means the desired precision in the convergence.

The Newton-Raphson will be compared to an explicit Runge-Kutta 4, which is one of the most widely integration schemes used when particles transport, e.g. [4],[5] or [6]. The behavior of both cases will be discussed in terms of mathematics and High Performance Computing (HPC).

III. ADAPTIVE TIME STEP

When particles transport is solved, firstly, the fluid is solved in the chosen interval Δt_{fluid}. The particles must be solved during the same time interval, but using adaptive time step, smaller independent intervals Δt_p for each particle p can be computed as shown in *figure 1*. A constant variation of the velocity of the fluid is supposed during Δt_p.

Figure 1: *Scheme of adaptive time step. Particles can adopt a smaller time step than the fluid.*

The time step may vary because of three reasons:

A. One element per time step

Particles cannot cross more than one element from one time step to another. Otherwise, time step is automatically decreased. This is why, if particles change subdomain, only

the first neighborhoods list is looped. These elements are called halo elements. In *figure 2* an example is shown.

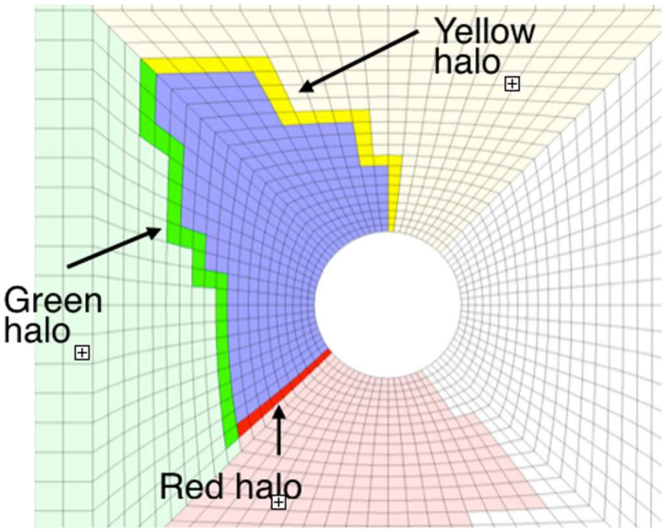


***Figure 2**: Blue subdomain is bordering red, green and yellow subdomains. In the case, a particle in blue subdomain changes subdomain, time step will be decreased if necessary, until it belongs to a a halo element.*

B. Reaching convergence of the integration scheme

Time step is decreased when convergence in Newmark-β is not reached. The convergence factor can be controlled by the user before the simulation starts, choosing the value ε_c defined by (8).

C. Control the error due to discretization of the time

As in the point B, before the simulation starts, the user must define the maximum acceptable error of the discretization ε_{err}. This error is normalized using a characteristic length L.

Some of the proposed characteristic lengths are the diameter of the particle, the length of the element or the instant velocity of the particle multiplied by a characteristic time τ. The discussion about which is the right characteristic length is not always straightforward and may vary depending on the properties of the problem.

In oder to estimate the new Δt_{err}, let x^{exa}_{n+1} be the exact solution applying Taylor series.

$$x^{exa}_{n+1} = x_n + u_n\Delta t + 1/2a_n\Delta t^2 + 1/6(da_n/dt)\Delta t^3 \quad (9)$$

It is necessary to subtract the equation (9) and (3), obtaining as result

$$x^{exa}_{n+1} - x_{n+1} = \beta(a_{n+1} - a_n)\Delta t^2 - 1/6(da_n/dt)\Delta t^3 \quad (10)$$

Now, applying the approximation

$$da_n/dt = (a_{n+1} - a_n)/\Delta t \quad (11)$$

Dividing by characteristic length L, and finally isolating Δt, it is obtained

$$\Delta t_{err} = \{\varepsilon_{err}L/[(\beta-1/6)(a_{n+1} - a_n)]\}^{1/2} \quad (12)$$

Let define ε_{trn} as the truncation error. According to equation (3), we also require the second order term (dependent on the velocity u_n) to be ε_{trn} times smaller than the first order term (dependent on the accelerations a_n , a_{n+1}). This means:

$$[(1-2\beta)a_n + 2\beta a_{n+1}]\Delta t^2/2 = \varepsilon_{trn}\, u_n\Delta t \quad (13)$$

Therefore, the truncation time step Δt_{trn} which satisfies this criteria is

$$\Delta t_{trn} = 2\,\varepsilon_{trn}\,|\,u_n/[(1-2\beta)a_n + 2\beta a_{n+1}]\,| \quad (14)$$

Both time steps will be used to estimate a new time step. To obtain that, we the define the accuracy α as

$$\alpha = \min(\Delta t_{trn}\,,\,\Delta t_{err})/\Delta t \quad (15)$$

The new time interval is only accepted if $\alpha > 0.9$. Otherwise, the process is repeated using $\Delta t_{new} = \Delta t$.

When showing results we will compare the error accuracy with the adaptive time step and without. Inasmuch as its impact on the performance of the code.

REFERENCES

[1] G.H. Ganser (1993) A rational approach to drag prediction of spherical and nonspherical particles. Powder Technol.

[2] A. Li, G. Ahmadi (1992) Dispersion and deposition of spherical particles from point sources in a turbulent channel flow. Aerosol Sciance and Technology.

[3] N.M. Newmark (1959) A method of Computation for Structural Dynamics. Journal of the Engineering Mechanics Division, ASCE, pp. 67-94.

[4] B.Asharian, S.Anjivel (1994) Inertial and Gravitational Deposition of Particles in a Square Cross Section Bifurcating Airway. Aerosol Science and Technology.

[5] ST Jayaraju, M Brouns, S Verbanck, C Lacor (2007) Fluid flow and particle deposition analysis in a realistic extrathoracic airway model using unstructured grids. Journal of Aerosol Science

[6] J.M. Dias, J.F. Lopesa, I. Dekeyserb (2001) Lagrangian transport of particles in Ria de Aveiro lagoon, Portugal. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere

Validating the Reliability of WCET Estimates with MBPTA

Suzana Milutinovic[1,2], Jaume Abella[2], Francisco J. Cazorla[2,3]
[1]*Universitat Politècnica de Catalunya (UPC),*
[2]*Barcelona Supercomputing Center (BSC-CNS),*
[3]*Spanish National Research Council (IIIA-CSIC)*
suzana.milutinovic@bsc.es

Abstract-*Estimating the worst-case execution time (WCET) of tasks in a system is an important step in timing verification of critical real-time embedded systems. Measurement-Based Probabilistic Timing Analysis (MBPTA) is a novel and powerful method to compute WCET estimates based on measurements on the target platform. To provide reliable estimates, MBPTA needs to capture at analysis time the events with high impact on execution time. We propose a method to assess and increase the confidence that MBPTA captures the relevant events during analysis.*

I. INTRODUCTION

The worst-case execution time estimate (WCET) is an important metric in critical real-time systems to prove that each critical task will complete its function in time. The WCET estimates need to be reliable according to the domain-specific safety standard (e.g. ARP4761 in the avionics domain [1]), and as tight as possible to avoid wasting hardware resources during task scheduling. As real-time systems deploy increasingly complex hardware and software, satisfying both requirements for WCET estimation becomes a difficult challenge [2].

Measurement-Based Probabilistic Timing Analysis (MBPTA) [3] is a novel method to derive WCET estimates based on measurements on the target platform. MBPTA deploys Extreme value theory (EVT)[4], a statistical method used to describe tails of distributions. Based on a sample of collected measurements EVT returns the distribution of high execution times of a task with corresponding probabilities of exceeding them, Fig. 1. The pWCET estimate is the execution time

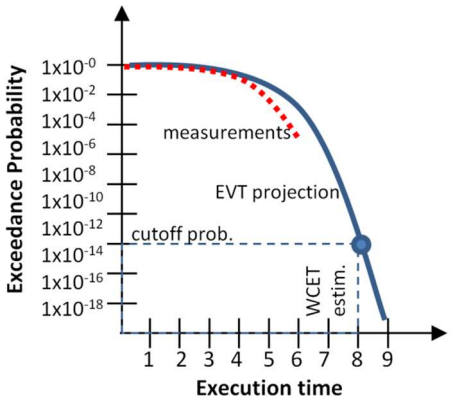


Fig. 1. Example of pWCET distribution.

value of that distribution that can only be exceeded with the cutoff probability selected in line with safety standard requirements.

MBPTA requires that the conditions under which the measurements during analysis are collected are equal or worse to the conditions at the system's deployment [5]. This requirement is ensured by: 1) forcing the sources that cause low variation of execution time to take its worst latency during analysis; and 2) time-randomizing the behavior of the sources that cause high variation of execution time. The main example of a hardware component whose timing behavior is randomized is a cache implementing random replacement and random placement policies.

The strength of MBPTA lies in the fact that it doesn't need to observe the longest execution time at a single measurement to predict that it can occur. If the worst outcome of each time-randomized resource is observed across different measurements, EVT will be able to predict the execution time when these bad scenarios occur together and upper-bound their probability of occurring simultaneously.

Our work focuses on deriving a method to guarantee with the specified level of confidence that the sample of measurements provided to EVT includes the events causing the worst outcomes of each time-randomized resource (we call them *events of interest*). As a consequence, the method guarantees that EVT will return a reliable pWCET estimate.

II. MBPTA: CAPTURING EVENTS OF INTEREST

In the previously studied architectures, MBPTA successfully captures the events of interest for each time-randomized resource, apart from time-randomized caches (TRc). TRc deploy the random placement policy, which maps addresses to randomly chosen sets. The mapping is changed across different runs, but kept constant during a single run. The authors in [6] observe that the execution time increases significantly if the number of addresses mapped to the same set exceeds the cache associativity. Some of these mappings, which we call *cache placements of interest*, may occur with a probability considered relevant by a safety standard, but low enough not to be captured during measurements at analysis time. Failing to provide the measurements capturing the cache placements of interest to the EVT method may cause the method to deliver optimistic pWCET estimates, and therefore return the unreliable results.

The HoG method proposed in [6] solves the problem of capturing cache placements of interest assuming that all addresses have the same impact on execution time. This is the case, for example, for a sequence that accesses all

addresses in a round robin fashion. We extend such solution to the general case with arbitrary access patterns by proposing the Representativeness Validation by Simulation (RVS) method [7].

III. RVS METHOD

The RVS method analyzes the miss count impact f the different groups of addresses if they are placed in the same set by means of simulations and determines analytically the probability of this to occur. The validation is done in a miss count domain, as cache misses highly correlate with the execution times [7]. Then, by testing whether those pairs <miss count, probability> are upper-bounded by the pWCET curve obtained from applying MBPTA in the miss count domain with the default MBPTA number of runs (R), RVS can detect whether R runs are enough or, instead, extra runs are needed. Then RVS precisely identifies the number of runs needed (R') so as to guarantee that all cache placements of interest have been observed

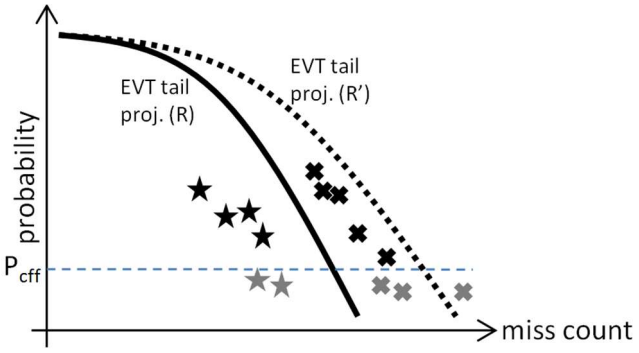


Fig. 2. Illustrative application of RVS.

with sufficient confidence.

This is illustrated in Fig. 2 where we show an example pWCET curve obtained with R runs that only upper-bounds some of the cache placements (marked with stars), but not some others (marked with crosses). Gray marks correspond to those cache placements that occur with negligible probability according to standards. RVS detects those cases leading to optimistic pWCET estimates and requests more runs (R') so that they properly upper-bound all meaningful cache placements.

We show the result of applying the RVS method for *aifirf* benchmark from the EEMBC automotive suite [8], in Fig. 3. This particular benchmark fails to pass the validation step with R runs, but passes it successfully with R' runs, as determined by RVS, once the cache placements of interest are observed so that MBPTA can upper-bound them. In the figure we show the pWCET curve with R and R' runs respectively, and the empirical complementary cumulative distribution function (ECCDF) for a large number of runs (orders of magnitude larger than R and R').

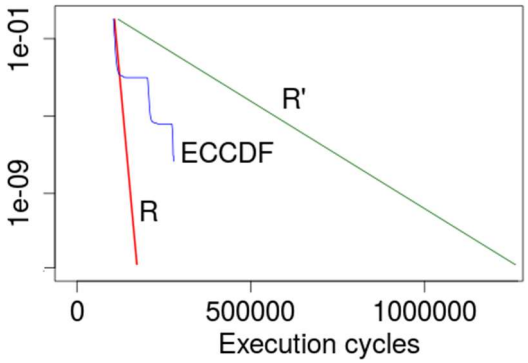


Fig. 3. pWCET for aifirf.

IV. CONCLUSIONS AND FUTURE WORK

MBPTA relies on EVT to estimate the pWCET of tasks, but to be reliable it requires that the execution time measurements used by EVT include all relevant (random) cache placements of interest. We propose the RVS method that determines the minimum number of measurements needed to feed MBPTA to produce reliable WCET estimates [7]. This is achieved by verifying that the pWCET estimate upper-bounds relevant cache placements in the miss domain. Our future work focuses on reducing the computation cost RVS and extending RVS toward multipath programs and cache hierarchies.

ACKNOWLEDGMENT

This work has received funding from the European Community's FP7 programme [FP7/2007-2013] under grant agreement 611085 (PROXIMA, www.proxima-project.eu). Support was also provided by the Ministry of Science and Technology of Spain under contract TIN2015-65316-P and the HiPEAC Network of Excellence. Jaume Abella has been partially supported by the MINECO under Ramon y Cajal postdoctoral fellowship number RYC-2013-14717.

REFERENCES

[1] SAE International. ARP4761: Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment.

[2] J. Abella *et al*., "WCET analysis methods: Pitfalls and challenges on their trustworthiness," *10th IEEE International Symposium on Industrial Embedded Systems (SIES),* Siegen, 2015, pp. 1-10.

[3] L. Cucu-Grosjean *et al*., "Measurement-Based Probabilistic Timing Analysis for Multi-path Programs," *24th Euromicro Conference on Real-Time Systems (ECRTS),* Pisa, 2012, pp. 91-101.

[4] S. Kotz *et al*., "Extreme value distributions: theory and applications, " World Scientific, 2000.

[5] F.J. Cazorla *et al*. "Upper-bounding program execution time with extreme value theory, " In WCET Workshop, 2013.

[6] J. Abella *et al*. "Heart of Gold: Making the improbable happen to extend coverage in probabilistic timing analysis," In ECRTS, 2014.

[7] S. Milutinovic, J. Abella, F.J. Cazorla, "Modelling Probabilistic Cache Representativeness in the Presence of Arbitrary Access Patterns, ", In ISORC, 2016.

[8] Jason Poovey, "Characterization of the EEMBC Benchmark Suite," North Carolina State University, 2007.

Efficient and versatile data analytics for deep networks

D. Garcia-Gasulla[1], J. Moreno-Vázquez[1], J.A. Espinosa-Oviedo[14], J. Conejero[1], G. Vargas-Solar[34], R.M. Badia[1], U. Cortés[2], T. Suzumura[5]

[1] *Barcelona Super Computing Centre, Spain*
[2] *Universitat Politècnica de Catalunya, Spain*
[3] *CNRS, France*
[4] *LIG-LAFMIA labs, France*
[5] *IBM T.J Watson Research Center, USA*

{dario.garcia, jonatan.moreno, javiera.espinosa, rosa.m.badia }@bsc.es, ia@cs.upc.edu genoveva.vargas@imag.fr suzumura@acm.org

Abstract-*Deep networks (DN) perform cognitive tasks related with image and text at human-level. To extract and exploit the knowledge coded within these networks we propose a framework which combines state-of-the-art technology in parallelization, storage and analysis. Our goal, to make DN models available to all data scientists.*[1]

I. INTRODUCTION

Deep learning networks (DN) learn data representations through the millions of features composing them [6]. This provides a trained DN with a rich representation language to, for example, perform object classification at a human-level. In the case of images, each feature learnt by a convolutional neural network provides a significant piece of visual information on the image, even if these features are not optimal for discrimination (only the top layer features are). In a sense, by considering all feature values for a given image, one is in fact looking at everything the network sees in an image. The goal of the Tiramisu environment (see Fig. 1) is extracting and exploiting the knowledge coded within DN. Therefore, using a pre-trained network (e.g., GoogLeNet on ImageNet data [7]), and extracting the features through a deep learning toolkit ((DLT) e.g., Caffe [5]), Tiramisu builds alternative data representations and runs various analytics methods on top of them [4]. This paper presents the Tiramisu architecture, focusing on its unique requirements in terms of parallelism and data management and storage.

II. RELATED WORKS AND MOTIVATION

Data management in many data analytics work- flows is guided by the RUM conjecture (Read, Update, Memory (or storage) overhead) [2]. Several platforms address some aspect of the problem like Big Data stacks [3, 1]; data processing environments (e.g., Hadoop, Spark, CaffeonSpark); data stores dealing with the CAP theorem (e.g., NoSQL's); and distributed file systems (e.g., HDFS). The principle is to define API's (application programming interface) to be used by programs to interact with distributed data management layers that can cope with distributed and parallel architectures. The challenge is to have tools that can change their preferences towards RUM and provide elastic strategies for implementing these operations. Such strategies should evolve as data acquire different structures and semantics as a result of the data processing operations applied on them.

III. PARALLELISM AND STORAGE REQUIREMENTS

Tiramisu depend on the analytics to be executed. The parallel execution of various DLT instances that process various inputs must be handled by a general purpose parallel programming model and execution platform (e.g., PyCOMPSs). Other data analytics processes, such as vector distances or graph clustering, may be either built in programs in Tiramisu, or provided by third parties (e.g., Spark, ScaleGraph). The challenge for Tiramisu is to provide an architecture that (i) connects to each of those components, and (ii) accesses the most appropriate on each specific context.

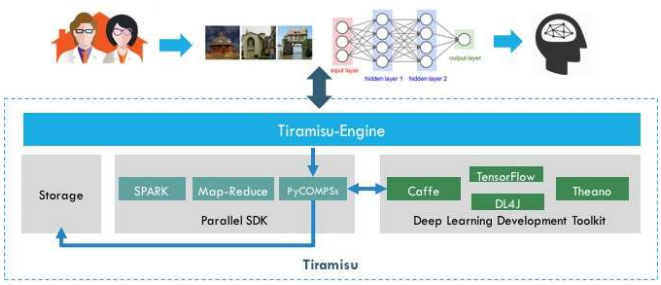


Fig. 1. Tiramisu architecture

IV. MANAGING DATA FOR TIRAMISU

The classification approach in Tiramisu is performed by a data centric workflow. It consists in a sequence of parallel analytics operations that process a data set of images and

[1] Details about the approach and implementation can be found in the proceedings of the 7th International Supercomputing Conference in Mexico (ISUM 16) organized from 11-13th April 2016 in Puebla, Mexico.

generate subsequent data collections with incremental content and semantics. Such operations require prepared input data collections (tagged and indexed), clever data colocation across processing nodes and results storage (cf. Fig. 2). A typical Tiramisu object contains a set of metadata (e.g., source image), and a set of values (e.g., millions of floats). Depending on the specific data representation, different types of persistence may be used, which must be consistent with the set of tools providing parallel analytic services. Indeed, well adapted data look-up, querying and exploration tools must be provided to ensure transparent access to data scientists.

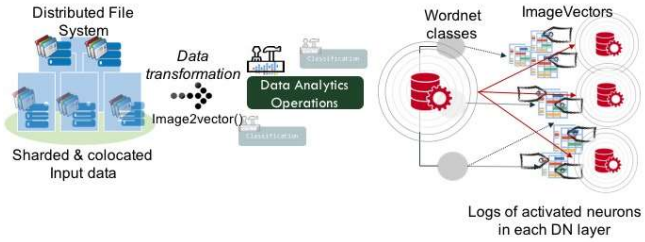


Fig. 2. Tiramisu data flow

ACKNOWLEDGMENT

This work is partially supported by the IBM/BSC Technology Center for Supercomputing (Joint Study Agreement, No. W156463) and the French Council of Scientific Research through its UMI 3175.

REFERENCES

A. Alexandrov, R. Bergmann, S. Ewen, J.C. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, et al. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6):939–964, 2014.

M. Athanassoulis, M. Kester, L. Maas, R. Stoica, S. Idreos, A. Ailamaki, and M. Callaghan. Designing access methods: The rum conjecture. *In International Conference on Extending Database Technology (EDBT)*, 2016.

Matthew Franklin. The berkeley data analytics stack: Present and future. *In Big Data, 2013 IEEE International Conference*, pages 2–3. IEEE, 2013.

D Garcia-Gasulla, J Béjar, U Cortés, E Ayguadé, and J Labarta. A visual embedding for the unsupervised extraction of abstract semantics. arXiv preprint arXiv:1507.08818, 2015.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.

Numbering along advection for Gauss-Seidel and Bidiagonal preconditioners

P. Córdoba*, G. Houzeaux, J.C. Cajas
Barcelona Supercomputing Center (BSC-CNS)
*paula.cordoba@bsc.es

Abstract- *Domain decomposition methods (DDM) are often chosen to precondition sparse linear systems of equations, as they are famous to well-improve the convergence of iterative solvers. But at the same time, they are difficult to implement and can be computationally expensive. In this work a new mesh numbering to adapt preconditioning techniques to the physics of different problems is proposed as an alternative to DDM preoconditioning.*

INTRODUCTION

Complex physical problems for both, applied fields and basic research, such as fluid dynamics, heat transfer problems, solid dynamics or general transport equations, are often represented by partial differential equations which have to be dicretized and solved numerically. This takes the continuum formulations of physics to systems of algebraic equations, and in order to obtain good approximations to the real life solutions of such problems it is necessary to solve systems with a great number of unknowns. The resulting matrices obtained from this discretizations are often very sparse, that is, only a few entries of the matrix differ from zero. Sparse linear systems of equations (SLSE) are usually solved with iterative solvers, as they are cheaper in terms of computer storage and CPU-time, but at the same time they are less robust than direct methods and often converge slowly to the desired solution. To cope with this problem, equivalent preconditioned systems can be solved instead of the original one, this means multiplying the system by a matrix called preconditioner, which has part of the information contained in the original matrix.

Finding a good preconditioner for solving SLSE is not an easy task and several aspects have to be taken into account, on of them is the physics of the problem, as the coefficients of the matrix highly depend on this.

In the present contribution the construction, implementation and results of a closely-related-to-the-physics preconditioners for convection dominated problems is studied. In this case, the information propagates mainly in the direction of advection. Then, independently of the discretization scheme considered (Finite Element, Finite vVolume, Finite Difference, etc,) the main contribution in every row of a certain node of the resultant matrix, apart from the diagonal term, comes from the closest neighboring in the opposite direction direction of the advection field. Thereby, a mesh node numbering along the flow direction (streamwise direction) is proposed in such a way that the main coefficient of each row will is, apart from the diagonal term, the first left off-diagonal term. Knowing this, several numerical examples in two and three dimensions have been tested using both Gauss-Seidel and Bidiagonal preconditioning together with Krylov subspace methods, inparticular the GMRES and BiCGSTAB solvers are used. The examples have been executed in sequential and in parallel and compared between them.

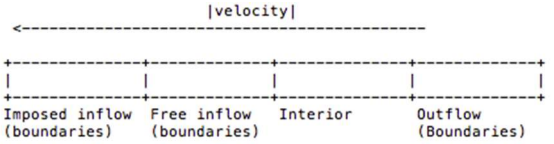


Figure 1. Node ordering by its velocity module.

METHODOLOGY

The numbering algorithm is based on two main ideas. First, the nodes are ordered by its velocity module in an increasing way, starting with the 'imposed' inflow nodes and ending with the nodes of the outflow. This is clearly shown in Figure 1.

After the nodes are put into different groups following what we call the 'minimum angle criterium' achieving like this the final ordering. This is done as it follows:

1. Starting with an inflow node, the forming vector between this node an each of its neighbors is computed.

2. Then different the cosines of the angle that these vectors form with the velocity vector that the inflow node has are computed and compared.

$$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

3. Discarding the nodes which have a negative or zero cosine, the next node in the group will be the one that forms the smallest angle with the velocity vector (maximum cosine), or what is the same, the chosen node will be the one which is closest to the direction of the advection velocity.

4. This procedure will be repeated recursively until no positive values of the cosine are found.

5. When this happens another node of the inflow will be taken and the above process will be repeated until all the nodes in the mesh are numbered.

In the parallel case this procedure will be done in each subdomain except for the interface nodes as the interfaces cut the advection lines, in this case a Jacobi preconditioner will be used instead.

RESULTS

To prove the algorithm several cases have been tested in sequential and in parallel. In this work it is shown an e example of each, solved in both cases with GMRES and BiCGSTAB solvers and preconditioned in each case with a Gauss-Seidel, Bidiagonal and Jacobi preconditioners.

A. In Sequential

Figure 3 shows the results for the example in sequential, this corresponds 2D heat convection with the following rotating advection field centered in (0.5, 0.5), so that v = (-y + 0.5, x-0.5). This has been solved on a 200 element mesh. In this case a mesh of 40430 elements has been used.

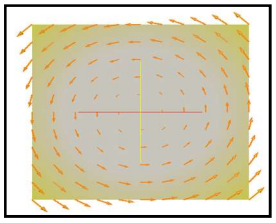


Figure2. Rotating advection in a heat convection problem.

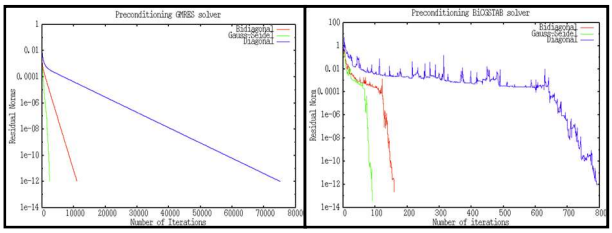


Figure 3.Heat convection problem solved in sequential

B. In Parallel

Figure 4 shows the problem tested in parallel. This is a 2D that simulates the plastic barrier designed for the ocean clean-up problem using Navier-Stokes equations and with an inflow velocity of 50m/s.

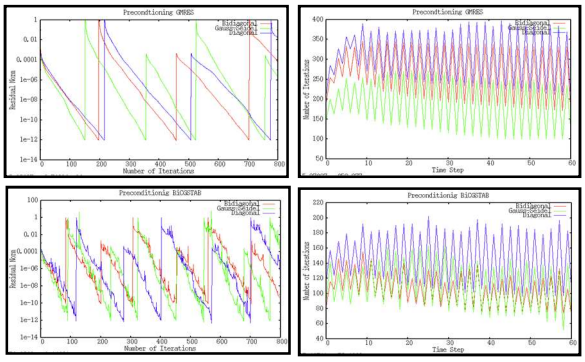


Figure 4. Navier Stokes equations solved in parallel.

CONCLUSIONS AND FUTURE WORK

A new node numbering for convection dominated problems has been developed and tested in different problems with the Gauss-Seidel and Bidiagonal preconditioners. Either in sequential and in parallel it is shown that the convergence of the GMRES and BiCGSTAB solvers is improved if compared with the Jacobi preconditioner. Although in the parallel case it still has to be checked the comparison between Bidiagonal and Gauss-Seidel preconditioning if a BiCGSTAB solver is used. Also in parallel it is expected that the efficiency of the strategy will decrease with the number of subdomains, as the streamlines are cut on subdomain interfaces.

Future work will include checking scalabilities and CPU times of the preconditioners proposed in real cases and a compare them them with some of the existent DDM that are also used as preconditioners.

REFERENCES

[1] F. Magoules, F.X. Roux, G. Houzeaux. *Parallel Scientific Computing*. December 2015, Wiley-ISTE

[2] J. Saad. *Iterative Methods for Sparse Linear Systems*. Siam, 2003

[3] V. G. Korneev, U. Langer. *Domain Decomposition Methods and Preconditioning.* Chapter 22. November 2004, John Wiley & Sons, Ltd.

Exploring the protonation properties of photosynthetic phycobiliprotein pigments from molecular modeling and spectral line shapes

M. Corbella[a], Z. S. D. Toa[b], G. D. Scholes[b], F. J. Luque[a], C. Curutchet[a]
aDepartament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain
bDepartment of Chemistry, Princeton University, Washington Road, Princeton, New Jersey 08544, United States
marinacorbella@ub.edu

Abstract-*In photosynthesis, specialized light harvesting pigment-protein complexes (PPCs) are used to capture incident sunlight and funnel its energy to the reaction center. In Cryptophyte algae these complexes are suspended in the lumen, where the pH ranges between ~5-7, depending on the prolongation of the incident sunlight. However, the pKa of the several kinds of bilin chromophores encountered in these complexes and the effect of its protonation state on the energy transfer process is still unknown. Here, we combine quantum chemical and continuum solvent calculations to estimate the intrinsic aqueous pKas of different bilin pigments. We then use Propka and APBS classical electrostatic calculations to estimate the change in protonation free energies when the bilins are embedded inside five different phycobiliproteins (PE545, PC577, PC612, PC630 and PC645), and critically asses our results by analysis of the changes in the absorption spectral line shapes measured within a pH range from*
4.0 to 9.4. Our results suggest that each individual protein environment strongly impacts the intrinsic pKa of the different chomophores, being the final responsible of their protonation state.

Nature has developed sophisticated and highly efficient molecular architectures to absorb sunlight and convert it into chemical energy, which is finally used by photosynthetic organisms to live and grow. The comprehension of these light-harvesting processes occurring in photosynthetic pigment-protein complexes (PPC) has been an important goal since the first high-resolution structure of the Fenna-Matthews-Olson complex appeared 40 years ago [1]. Cryptomonads are a group of algae which are important primary producers in marine and freshwater environments due to its high quantum yeld. As compared to land plants, the available light that algae can harvest in water environments is significantly reduced. For that reason, cryptomonads' PPC use tunable linear tetrapyrrole chromophores (bilins) covalently bounded to the protein scaffold, which structure and disposition inside the protein have evolved to increase the spatial and spectral cross section (450 – 640 nm) for the absorption of incident light [2]. Unlike in other light-harvesting organisms, in cryptomonads these PPCs are suspended in the lumen, inside the intrathylakoid membrane. Under intense illumination, the reaction centers of photosynthetic organisms are capable of redirecting the excess excitation energy by a change in the thylakoid lumen pH, which triggers a biochemical feedback process in which the absorbed energy is dissipated as heat. Unlike in most photosynthetic organisms, in cryptophyte algae, the increased acidification of the thylakoid lumen directly

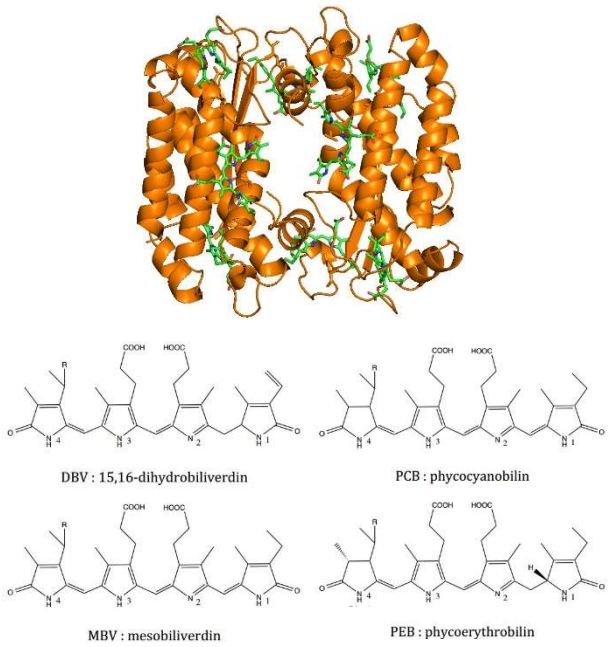


Fig. 1 a) Crystal structure of the pigment-protein complex PC577 and disposition of the containing chromophores (green). b) Structure of the four bilin pigments studied.

affects the local environment of the primary antenna proteins (phycobiliproteins), which are bathed in the lumen. Although many theoretical and experimental studies have been done in order to uncover the basic mechanisms that drive electronic energy transfer in these organisms [3]–[6] and that recently, various PPCs structures have been elucidated (Fig. 1a) [2] (PE545, PC577, PC612, PC630 and PC645), the p*K*a of the chromophores encountered in these complexes and the effect of its protonation state on the energy transfer process is still unknown, and is hard to be determined experimentally because they are covalently bounded to the protein scaffold and they lose their active conformation in solution.

Here, we combine quantum chemical and continuum solvent electrostatic calculations to build a thermodynamic cycle and obtain the change in the Gibbs free energy of the reactions of deprotonation in solution (ΔG_{aq}), governed by the equilibrium constant of the reaction (K_a), so the p*K*a is calculated using (1).

$$pK_a = \frac{\Delta G_{aq}}{RT\ln(10)} \quad (1)$$

We extrapolate the MP2 energy to a complete basis set and we also applied the spin-comonent scaled MP2 correction of the energy proposed by Grimm [7]. We used both MST [8] and SMD solvation methods, giving better results the SMD for the neutral species and the MST for the charged ones, so we performed an average of both sovation free energies to finally obtain the intrinsic p*K*a of each protonable site of each bilin chromophore (PCB, PEB, DBV and MBV) (Fig. 1b, Table I).

Table I. Gibbs free energies in the gas phase and intrinsic p*K*as for each kind of bilin chromophore.

DBV	ΔG_{gas} (kcal mol^{-1})	pKa	PCB	ΔG_{gas} (kcal mol^{-1})	pKa
N4	255.3	17.4	N4	250.2	11.9
N3	234.3	6.7	N3	238.4	7.1
N2	233.1	6.3	N2	238.7	7.4
N1	288.4	29.9	N1	258.1	17.0
MBV	ΔG_{gas} (kcal mol^{-1})	p*K*a	PEB	ΔG_{gas} (kcal mol^{-1})	p*K*a
N4	256.6	16.7	N4	249.4	13.3
N3	236.7	6.6	N3	235.4	6.8
N2	236.6	6.6	N2	234.1	6.5
N1	256.2	16.1	N1	290.4	30.0

Then, we used both Propka server and continuum electrostatic methods (APBS) to estimate the change in the Gibbs free energy of the reaction of deprotonation of the two central pyrrole rings, which are the only ones susceptible to undergo deprotonation, in each particular environment inside each protein (PC577, PC612, PC630, PC645 and PE545). If we observe the crystal structure, we can see that all chromophores instead of MBVs, are coordinated under the two central pyrrole rings by an aspartic or glutamic acid, presumably stabilizing the protonated form of the chromophore with p*K*as ranging between 6 to 7, while the MBVs present p*K*as ranging between 4-5. As we can assume an error of ± 1 or 2 p*K*a units, we finally asses our results by analysis of the changes in the experimental absorption spectral line shapes measured within a pH range from 4.0 to 9.4 (See Fig. 2).

Our results suggest that each individual protein environment strongly impacts the intrinsic p*K*a of the different chomophores, being the final responsible of their protonation state. So, if we observe the experimental spectra, between 5.4 and 8.2 (Fig. 2) there are no apparent changes in the DBVs and PCBs spectral region, while from 6.5 to 7.8 there is a shift of the absorption lineshape within the MBVs region. So, we can assume that we are underestimating the results 1.5 – 2 p*K*a units, and that all chromophores are protonated at a working pH of 7.5, being the MBVs the first ones to deprotonate due to the lack of an stabilizing group coordinating the two central pyrrole rings.

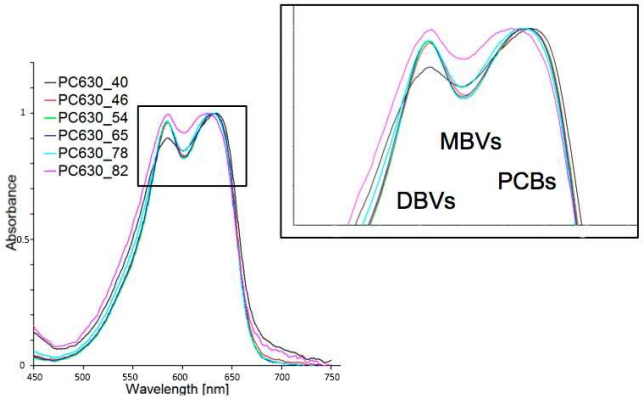


Fig. 2 Absorption spectrum of the antenna protein PC630 at different pHs between 4.0 and 8.2.

ACKNOWLEDGMENT

C.C. and M.C. acknowledges support from the Ministerio de Economía y Competitividad of Spain (grants CTQ2012-36195, RYC2011-08918, EEBB-I-15-09450 and BES-2013-064088), the Generalitat de Catalunya (SGR2014-1189) and computational resources provided by the Consorci de Serveis Universitaris de Catalunya.

REFERENCES

[1] R. E. Fenna and B. W. Matthews, "Chlorophyll Arrangement in a Bacteriochlorophyll Protein from Chlorobium Limicola," Nature, vol. 258, pp. 573–577, 1975.

[2] S. J. Harrop, K. E. Wilk, R. Dinshaw, E. Collini, T. Mirkovic, C. Y.Teng, D. G. Oblinsky, B. R. Green, K. Hoef-Emden, R. G. Hiller, G. D. Scholes, and P. M. G. Curmi, "Single-residue insertion switches the quaternary structure and exciton states of cryptophyte light- harvesting proteins.," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 26, pp. E2666–75, Jul. 2014.

[3] C. Curutchet, J. Kongsted, A. Muñoz-Losa, H. Hossein-Nejad, G. D.Scholes, and B. Mennucci, "Photosynthetic light-harvesting is tuned by the heterogeneous polarizable environment of the protein.," J. Am. Chem. Soc., vol. 133, no. 9, pp. 3078–84, Mar. 2011.

[4] C. Curutchet, V. I. Novoderezhkin, J. Kongsted, A. Muñoz-Losa, R.van Grondelle, G. D. Scholes, and B. Mennucci, "Energy flow in the cryptophyte PE545 antenna is directed by bilin pigment conformation.," J. Phys. Chem. B, vol. 117, no. 16, pp. 4263–73, Apr.2013.

[5] L. Viani, C. Curutchet, and B. Mennucci, "Spatial and Electronic Correlations in the PE545 Light-Harvesting Complex," J. Phys. Chem. Lett., vol. 4, no. 3, pp. 372–377, Feb. 2013.

[6] L. Viani, M. Corbella, C. Curutchet, E. J. O'Reilly, A. Olaya-Castro,and B. Mennucci, "Molecular basis of the exciton-phonon interactions in the PE545 light-harvesting complex," Phys. Chem. Chem. Phys., vol. 16, no. 30, pp. 16302–16311, 2014.

[7] S. Grimme, "Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies," J. Chem. Phys., vol. 118, no. 20, p. 9095, 2003.

[8] C. Curutchet, A. Bidon-Chanal, I. Soteras, M. Orozco, and F. J.Luque, "MST continuum study of the hydration free energies of monovalent ionic species.," J. Phys. Chem. B, vol. 109, pp. 3565–3574,2005

Clustering the Roman Empire: the use of multivariable analysis to understand cultural dynamics

Maria Coto-Sarmiento[1,2], Xavier Rubio-Campillo[1], José Remesal[2]
[1] *Barcelona Supercomputing Center, University of Barcelona*[2]
maria.coto@bsc.es

Abstract- *The aim of this study is to analyze the effects of rates of change of the olive oil amphorae to explore the production dynamics in the Roman Empire. In this case Cultural Evolution theory will be applied to the material culture study because is considered a useful tool to understand the variability of the mechanisms of changes. This analysis can be developed by the fact that we detect differences in the amphorae production through time that they might explain this dynamic of change.*

In this context, it will be presented a research where this methodology has been used to show its capacity to detect the culture trajectories. In particular, our case of study has been focused to understand the dynamics of change of olive oil amphorae production found in Baetica (currently Andalusia) during the Roman Empire (1st-3rd century AD). Specifically, multivariable methods have been applied to distinguish pottery assemblages among different kinds of shapes that it could serve to identify discontinuities in archaeological sequences. Specifically, we want to identify if these changes were produced by cultural reasons as it may be economical, political and social changes.

Finally, the results suggest that different factors as spatial distance can influence the rate of change and that rates will be more or less likely depending on them.

I. INTRODUCTION

Cultural evolution theories [1] provide a set of methods that can be used to account these dynamic of changes, focused on the production of olive oil amphorae during the Roman Empire.

To achieve this goal, multivariable methods were used to evaluate the differences on the pattern production among pottery workshops [2].

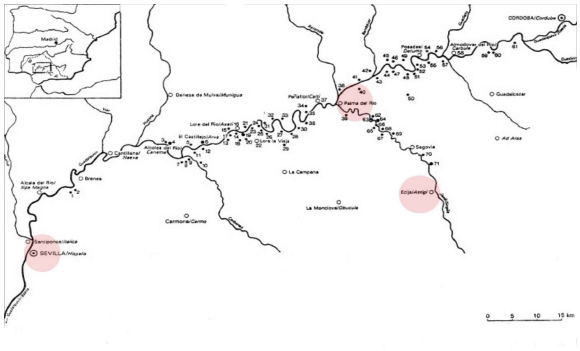


Fig. 1. Distribution map of the four amphorae workshops. The names are Las Delicias (Écija, Seville), Belén and Malpica (Palma del Río, Córdoba) and Parlamento (Seville*)* Specifically we want to identify the origin of these changes and if these changes were produced by cultural reasons depending on the spatial distance and other cultural constraints. As hypothesis, we propose that spatial distribution of pottery workshops is the main influence of the making techniques processes [3]. Four pottery workshops, showed in the map (fig.1) were studied from different spaces in *Baetica*.

II. METHODS

A. *Measurements*

To explore the dynamic of changes we analysed a set of measures among different kinds of amphorae shapes from different workshops. We analysed 413 samples of amphorae from 4 different workshops. These workshops were selected from different spaces of *Baetica* area in order to know if there were differences depending on the space. A database was created using a selection of 80 to 90 samples from each pottery workshops. In each sample of amphora we measured eight measurements among different part of the rim being focused on the rim of the amphora.

B. *Multivariable methods*

Multivariable methods were used to explore these metrical observations [4] with the eight measurements as variables. Principal Component Analysis allowed us to simplify the dataset to see which variable were more relevant. Our results suggested that first and second principal component were more relevant than the rest.

III. RESULTS

Several multivariable methods were used such as Principal Component Analysis and Discriminant Analysis to classify. These methods allowed us to know the differences on the pattern production among workshops. In our case, the first two principal components were taken to see the significal differences among workshops depending on the space. The figure 2 shows the workshops with a minor space such Belén and Malpica share more pottery traits than the rest: Parlamento and Las Delicias.

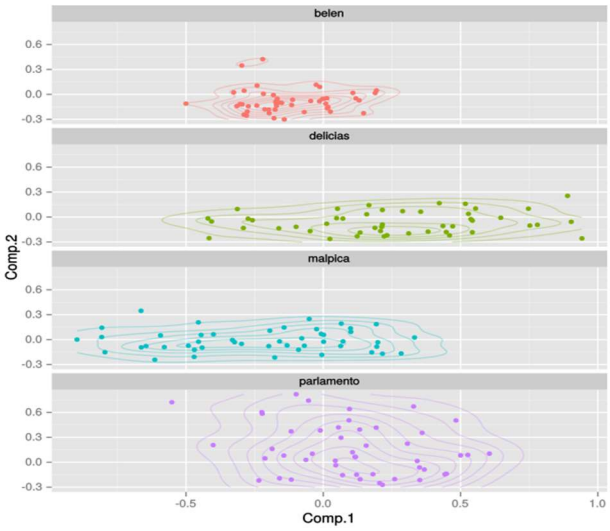


Fig. 2. Plot with the results of the first two principal components given by the PCA.

Once defined the components, we used Discriminant Analysis to find a combination among them to define the groups as well as possible. These results were translated to a confusion matrix which basically means what results were predicted as true or false on the discriminant analysis. As shown in the Figure 3 of confusion, all correct guesses were located in the diagonal of the table. Thus the system had troubles to distinguishing between Belen and Malpica which had a higher number of confusion or number instead of Parlamento with a minor confusion than the rest.

	BELEN	DELICIAS	MALPICA	PARLAMENTO
BELEN	**37**	10	18	8
DELICIAS	4	**32**	8	14
MALPICA	5	3	**18**	6
PARLAMENTO	2	3	4	**20**

Fig. 3. Matrix of confusion. Accuracy: 0.5573 %. P-Value: 0.0006991.

A peer to peer comparison was developed among different workshops. We calculated the geographical distance between each site and the distance among pottery measures, calculated using the previous results. The Figure 4 shows that the pottery distance is correlated with the spatial distance of workshops.

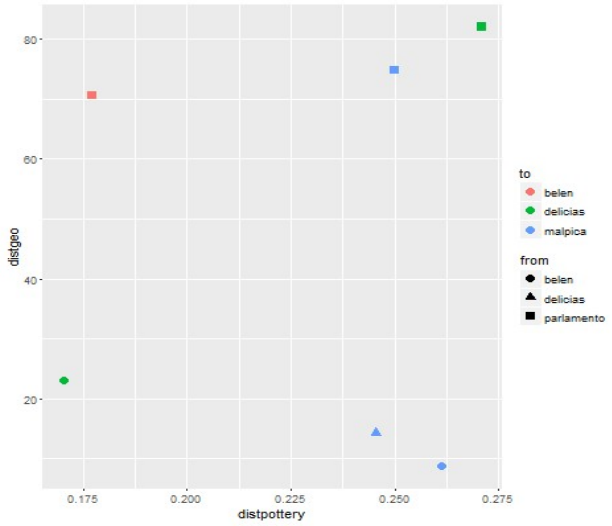


Fig. 4. Distance metrics calculated among different workshops.

CONCLUSION

Differences among pottery workshops were identified using PCA and Discriminant Analysis. As results, Amphorae made in nearby workshops with a minor spacial distance, such as Malpica and Belen, share more traits than amphorae made in pottery workshops farther as Parlamento. It could suggest that the pottery techniques were learned from master to disciple instead of workers with the same level.

ACKNOWLEDGMENT

The Funding for this work was provided by the ERC Advanced Grant EPNet (340828). We thank Museum of Écija and Museum of Palma del Río for their assistance. We also thank to dr. Enrique García Vargas (University of Seville) for his helpful suggestions.

REFERENCES

[1] Mesoudi, A. "Cultural Evolution: A review of Theory, Finding and Controversies", *Evolutionary biology*, 2015, pp. 1-17.

[2] Aguilera, A. "Análisis multivariable: una nueva vía para la caracterización cerámica", *Pyrenae*, 29, 1998, pp. 117-134.

[3] Schillinger, K., Mesoudi, A. & Lycett, S.: "Differences in Manufacturing Traditions and Assemblage-Level Patterns: the Origins of Cultural Differeces in Archaeological Data". *Journal of Archaeological Method Theory*, 2016, pp. 1-19.

[4] Li, Xiuzhen Janice, et al. "Crossbows and imperial craft organisation: the bronze triggers of China's Terracotta Army." *Antiquity,* 88.339, 2014, pp. 126-140.

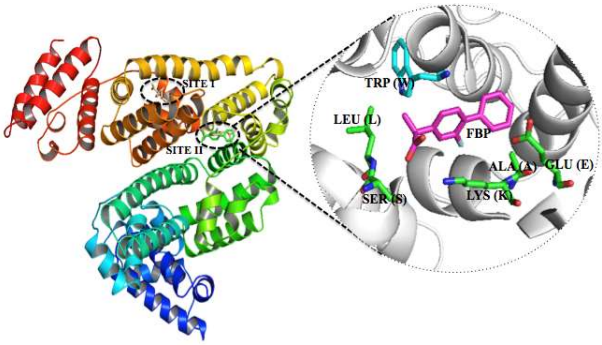
Assessing drug-protein binding by simulation of stereoselective energy transfer dynamics: electronic interactions between tryptophan and flurbiprofen

Silvana Pinheiro & Carles Curutchet
Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Spain
sylsouza_@hotmail.com

Protein fluorescence decays are difficult to interpret and often involving several energy transfer processes among Trp residues or Trp-ligands. In this study, we simulate EET rates by computing MD-averaged electronic couplings V. Fluorescence decays have been observed for the HSA protein bound to the S and R enantiomers of FBP. So far, our results in the HSA-FBP system agree with the experimental hypothesis (stereoselective energy transfer) and strongly support the binding modes proposed for the R and S enantiomers in HSA.

I. INTRODUCTION

The fluorescence of proteins is a complex process often involving several electronic energy transfer (EET) reactions between aromatic amino acids (typically arising from tryptophan), before light emission. In protein-ligand complexes, the ligand can also modify the fluorescence properties by participating in those EET processes, as well as by contributing to electron transfer reactions or the formation of exciplexes. The complex interpretation of optical experiments, however, precludes in a full explotation of the structural information encapsulated in such experiments and related to the drug-binding events observed. The detailed understanding of drug-protein binding is determinant for drug action and drug transport and disposition, which are regulated by various transport proteins such as HSA-*Human Serum Albumin* (*Fig.*1)[1].



Fi gure 1. Drug binding to site 2 in HSA. The detailed binding conformation is show for flurbiprofen.

In this context, we intended to explore the potential of simulation techniques MD/QM-MM in describing EET processes and fluorescence protein-ligand systems in order to determine the binding modes of protein ligands by comparison with fluorescence experiments.[2]

II. COMPUTATIONAL DETAILS

We have simulated how energy transfers involving different flurbiprofen enantiomers modulate the fluorescence properties of model tryptophan-flurbiprofen (TRP-FBP) and flurbiprofen-HSA (*human serum albumin*) complexes *(Fig.2)*, where stereoselective dynamic quenchings have been recently observed.[2] To this aim, we combine classical MD techniques with a polarizable QM/MM methodology that we have recently developed[3] and applied to study the light-harvesting properties of photosynthetic systems.[4,5]

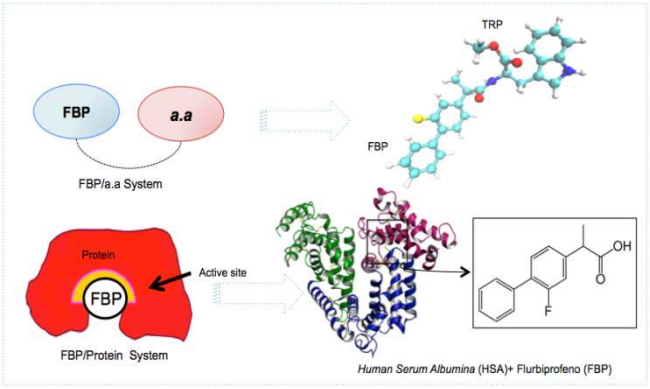


Figure 2. Model system (FBP/TRP) and biological system (HSA/TRP)

The QM/MMpol linear response approach

The QM/MMpol model[3] combines a quantum-chemical description of the pigment's excited states (TD-DFT, CIS or ZINDO) with a polarizable MM description of the surrounding environment, where MM atoms are assigned with a partial charge and an isotropic polarizability:

$$\hat{H}_{eff}\left|\Psi\right\rangle = \left(\hat{H}_{QM} + \hat{H}_{QM/MM} + \hat{H}_{MM}\right)\left|\Psi\right\rangle = E\left|\Psi\right\rangle$$

The electronic coupling (V) between relevant excited states, is obtained perturbatively from the transition

densities computed for the non-interacting donor (D) and acceptor (A):

$$V = V_0 + V_{\text{env}}$$

$$V_0 = \int d\vec{r}' \int d\vec{r} \rho_A^{T*}(\vec{r}') \left[\frac{1}{|\vec{r} - \vec{r}'|} + g_{xc} \right] \rho_D^T(\vec{r}) - \omega_0 \int d\vec{r}' \int d\vec{r} \rho_A^{T*}(\vec{r}') \rho_D^T(\vec{r})$$

Direct interaction between D/A transition densities. Includes Coulomb, exchange-correlation and overlap terms.

$$V_{\text{env}} = -\sum_k \vec{\mu}_k^{ind} \left(\rho_D^T \right) \int d\vec{r} \frac{\rho_A^T(\vec{r}) \left(\vec{r}_k - \vec{r} \right)}{\left| \vec{r}_k - \vec{r} \right|^3}$$

Environment-mediated D/A interaction described in terms of the MM polarization response.

III. Results and Discussion

In order to interpret the fluorescence experiments on the Flurbiprofen complex with HSA protein, we studied the ability of different methods in order to describe the electronic states involved. The table shows the results of the transition energies obtained for the π→π* state of flurbiprofen (FBP) and the states La and Lb of tryptophan (TRP). The transitions energies for the Flurbiprofen lowest p → p* state and the tryptophan La state are in excellent agreement with the experimental values, which indicates the goodness of the semiempirical ZINDO method in order to describe the properties of the system.

	Flurbiprofen (donor)			Tryptophan La state			Tryptophan Lb state	
	exp / eV	ΔE / eV	f	exp / eV	ΔE / eV	f	ΔE / eV	f
Protein *S*	4.29	4.12	0.56	4.16	4.21	0.18	4.01	0.05
Protein *R*		4.11	0.56		4.18	0.19	4.01	0.06
Model *S*		4.26	0.46		4.14	0.17	4.02	0.04
Model *R*		4.25	0.45		4.13	0.17	3.99	0.04

Then we proceeded to estimate the rate of energy transfer (EET) for the two enantiomers of FBP, to investigate whether changes in fluorescence (experimentally observed) are due to processes EET, and to validate the binding mode of each enantiomer predicted theoretically. The results ZINDO level found for HSA-FBP biological system suggest a process EET approximately 30% faster than FBP for TRP in the case of the *S* enantiomer, according to the experimental observation, giving validity to the proposed model.

The results of coupling squared (V^2) calculated ZINDO level are shown in Figure 3. For the FBP-TRP model in solution, found comparable results between *R* and *S* enantiomers. These results are in contrast with the experimental observation where it postulates a EET 2-3 times faster for the *R* enantiomer .

However, this system is very flexible from the conformational point of view. Currently, there are increasing in the number of structures studied to converge predictions, besides, it is necessary to study the validity of the conformational preferences predicted by MD simulations that could be affecting our results significantly.

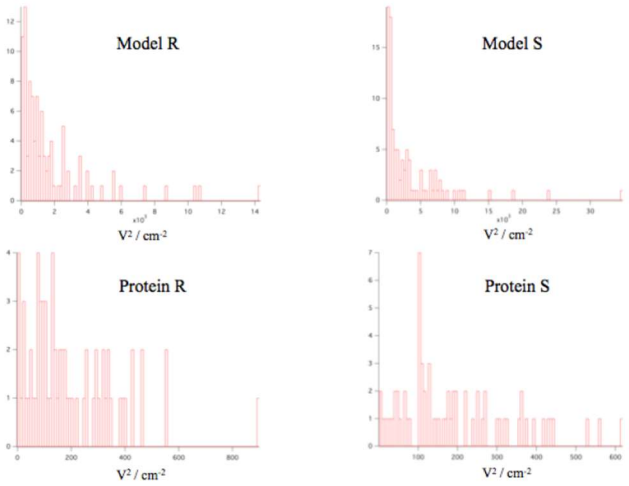


Figure 3: Coupling distribution (V2) for the model (FBP / TRP) and to the biological system (HSA / TRP) calculated over the simulation.

In general, preliminary analysis of MD-QM/MMpol simulations at the ZINDO semiempirical level (100 snapshots) predict a distribution of squared couplings along the simulation comparable for the *S* and *R* enantiomers, opposite to the experiments. In contrast, the results for the protein systems suggest a slightly faster EET from FBP to Trp in the *S* case, in agreement with the experimental observation. Nowadays, we are extending the simulations to more structures in order to properly converge the estimated EET rates. We are also performing *ab inito* CIS and TD-DFT calculations in order to verify the semiempirical results. Overall, the results obtained strongly support the hypothesis that changes in the fluorescence of the HSA-FBP system arise from EET processes.

ACKNOWLEDGMENT

Thank to CNPQ – Ciência sem fronteiras; Universitat de Barcelona; Ministerio de economia y competitividad – Gobierno de España.

REFERENCES

[1] J. Ghuman, Patricia A. Zunszain, Isabelle Petitpas, Ananyo A. Bhattacharya, Masaki Otagiri and Stephen Curry. J. Mol. Biol. 2005, 353, 38-52.

[2] I. Vayá, P. Bonancía, M. C. Jiménez, D. Markovitsi, T. Gustavsson and M. A. Miranda. Phys. Chem. Chem. Phys. 2013, 15, 4727- 4734.
[3] C. Curutchet, A. Muñoz-Losa, S. Monti, J. Kongsted, G. D. Scholes and B. Mennucci. J. Chem. Theory Comput. 2009, 5, 1838- 1848.
[4] C. Curutchet, J. Kongsted, A. Muñoz-Losa, H. Hossein-Nejad, G. D. Scholes and B. Mennucci. J. Am. Chem. Soc. 2011, 133, 3078- 3084.
[5] C. Curutchet, V. I. Novoderezhkin, J. Kongsted, A. Muñoz-Losa, R. V. Grondelle, G. D. Scholes and B. Mennucci. J. Phys. Chem. B 2013, 117, 4263-4273.

Extrapolations of the fusion performance in JET

D. Gallart[1], M.J. Mantsinen[1,2], L. Garzotti[3], C. Challis[3] , J. Garcia[4], A. Gutiérrez[1], X. Sáez[1]
[1]*Barcelona Supercomputing Center, Barcelona Spain*
[2]*ICREA, Barcelona, Spain*
[3]*CCFE, Culham Science Centre, Abingdon, UK*
[4]*CEA, IRFM, Saint-Paul-lez-Durance, France*
daniel.gallart@bsc.es

Abstract- *In preparation of the forthcoming high power campaign with the reactor-relevant deuterium-tritium (DT) fuel mixture in the Joint European Torus (JET), significant efforts are being devoted to DT scenario extrapolation using computer modelling. We report on simulations aimed at optimizing external heating using neutral beam injection (NBI) and radiofrequency waves in the ion cyclotron range of frequencies (ICRF) for high DT fusion yield. Our results show that by increasing external heating power to the maximum power available, the fusion neutron rate can be enhanced by a factor of 4-5 with respect to the recent record values. The comparison of two ICRF schemes using different resonant ion species, i.e. ^{3}He and H minority ions, shows that the ^{3}He minority heating scenario achieves a higher fuel ion temperature but not necessarily a better fusion performance. Finally, we study the dependence of the performance of external heating on key experimental parameters.*

I. INTRODUCTION

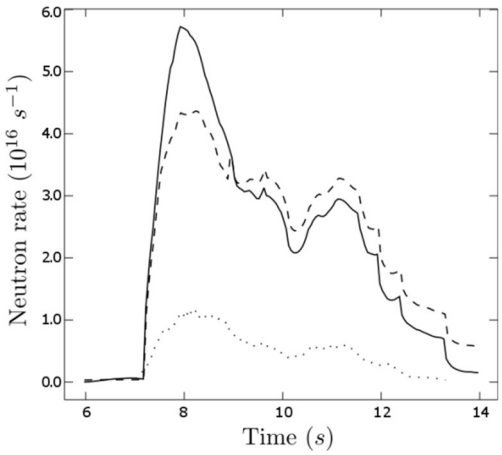
In preparation of the second high power campaign with the deuterium-tritium (DT) fuel mixture in the Joint European Torus (JET) [1], we report on simulations aimed at optimizing external heating using neutral beam injection (NBI) and radiofrequency waves in the ion cyclotron range of frequencies (ICRF) for high fusion yield. As a reference we use a high-performance 2.9 T/2.5 MA hybrid discharge (86614), i.e. high beta plasma with good confinement (H-mode). Here, beta is the ratio of the plasma pressure to the magnetic pressure $\beta = 2\mu_0 n k_B T/B^2$ where μ_0 is the magnetic permeability, n the plasma density, k_B the Boltzmann constant, T the plasma temperature and B the magnetic field. This discharge yielded the record fusion reaction rate (DD) so far in the JET campaigns with the ITER-like Wall (ILW) [1] and has been extensively analyzed prior to this work e.g. in Ref. [2]. In particular, we compare the performance of ^{3}He and H minority heating using the ICRF modelling code PION [3] coupled to the beam deposition code PENCIL [4]. PION + PENCIL modelling takes into account the synergy between ICRF waves and resonant NBI ions. The 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 and DT in the cases studied here. For good RF accessibility and absorption we choose the cyclotron frequency of the minority species that is higher than that of main ion species. Thus, we use the hydrogen minority resonance $\omega = \omega_{cH} = 2\omega_{cD}$ and ^{3}He minority resonance with $\omega = \omega_{c}{}^{3}{}_{He} = 2\omega_{cT}$, where ω is the frequency of the launched wave and the ion cyclotron frequency is defined as $\omega_c = ZeB/(Am_p)$. Here, Ze and Am_p are the ion charge and mass, respectively, and B is the confining magnetic field.

We have carried out our simulations with the coupled PION and PENCIL codes not only for fixed but also for evolving plasma parameters as calculated by the coupling of these codes to the plasma transport code JETTO [5] in order to take into account the plasma response to the applied plasma heating and fueling.

II. FUSION PERFORMANCE OF D PLASMA AT HIGH INPUT POWER

The extrapolation of the JET reference hybrid discharge to high fusion performance with D plasma considered here consists of increasing the external heating power with ICRF waves and NBI, and the toroidal magnetic field and the plasma current to 3.25 T and 2.7 MA, respectively.

Fig.



1.Neutron production rate (DD) for the reference pulse (dotted line) and for the extrapolation to high power using ^{3}He (solid line) and H (dashed line) minority ICRF heating with a minority concentration of 4% in a deuterium plasma.

Figure 1 compares the experimental total fusion reaction rate of the reference discharge which has a total of 27 MW of external heating power with two simulated cases using coupled PION, PENCIL and JETTO modelling. In the simulated cases a higher total power of 40 MW is assumed while keeping the same plasma density as in the reference discharge. The input power consisted of 34 MW of D NBI and 6 MW of ICRF power, which is the maximum power

foreseen to be available for the presently planned future JET experiments. As we can see in Fig 1, our simulations suggest that the peak fusion reaction rate can be increased by a factor of about 2-3 by increasing the total injected power by a factor of 1.48 to its maximum value.

The two simulated scenarios in Fig. 1 have identical fuel mixtures except for the different minority ion species resonant with the launched waves In one of the simulations we have $\omega = \omega_{c^3He} = 33$ MHz while in the other simulation we have $\omega = \omega_{cH} = 2\omega_{cD} = 51.5$ MHz. As we can see in Fig. 1, the $\omega = \omega_{c^3He}$ scenario gives rise to a better fusion reactivity in the high performance phase up to t = 9 s as compared to the $\omega = \omega_{cH} = 2\omega_{cD}$ scenario while the situation is opposite in the lower-performance phase from t = 9 s onwards. In both scenarios, the ion temperature reaches its maximum at a minority concentration of about 4%, which are the cases considered in Fig. 1. However, the ^{3}He scenario results in a higher ion temperature during all the NBI and ICRF phase with a maximum of 16 keV at the high performance phase and 12 keV on average at the low performance phase. Although the H scenario gives rise to a lower temperature (12 keV at the high performance phase and 10 keV at the low performance phase), the synergy between the deuterium NBI and ICRF heating enhances the second deuterium harmonic damping and, thereby, the number of fast deuterons. This in its turn improves the fusion yield of the H minority scheme in the low performance phase as compared to that of the ^{3}He scenario.

III. COMBINED NBI + ICRF HEATING IN JET DT PLASMAS

We have performed our first series of simulations with coupled PENCIL and PION codes to study the dependence of the combined NBI + ICRF heating on key plasma parameters. The analysis presented here is for a 50%-50% DT plasma mixture with 5% of ^{3}He assuming equal ion and electron temperatures. A total external power of 28 MW consists of 22 MW of NBI (11 MW of T NBI and 11 MW of D NBI) and 6 MW of ICRF with a frequency of 33 MHz tuned to a central $\omega = \omega_{c^3He} = 2\omega_{cT}$ resonance. While our reference discharge had a plasma electron density of $6.2\text{x}10^{19}$ m^{-3} and a temperature of 9 keV, in our simulations we have multiplied the density and temperature by constant factors in order to perform a scan in these two quantities. The factor of 0.5 for the density and 1.0 for the temperature correspond to the values of the reference discharge.

Figure 2 shows the power absorption from ICRF waves by resonant ions, i.e. ^{3}He minority ions and tritons. Each point in the surface presents one simulation with the coupled PION and PENCIL codes. The power not absorbed by the resonant ^{3}He minority ions and tritons is absorbed by direct electron damping by electron Landau damping and transit time magnetic pumping. As we can see in Fig. 2, absorption by resonant ions decreases weakly with plasma density and temperature. Nevertheless, it remains dominant, accounting for 65-90% of the total ICRF power, in the whole temperature and density range under consideration.

The resonant ions heat the bulk ions and electrons through collisions. The collisional bulk ion heating fraction, which is a key quantity for high fusion performance, depends on the average resonant ion energy with respect to the critical energy. They both depend on the plasma parameters. According to our simulations the resulting collisional bulk ion heating by resonant ^{3}He ions and tritons increases modestly with plasma density and temperature as shown in Fig. 2.

IV. CONCLUSION

We have extrapolated a reference record JET discharge to high NBI+ICRF power using two different ICRF scenarios, i.e. ^{3}He and H minority heating. While ^{3}He minority heating results in a z better fusion neutron rate in the high performance phase, the H minority scenario performs better in the low performance phase due to second harmonic damping of launched wave on deuterons which increases the fusion yield. We have also presented our first results on the dependence of combined ICRF+NBI heating on the plasma density and temperature for the ^{3}He minority in 50%-50% DT fuel mixtures. Our next step is to compare with the H minority scenario and analyse in detail the fusion reactivity in DT plasmas for both cases.

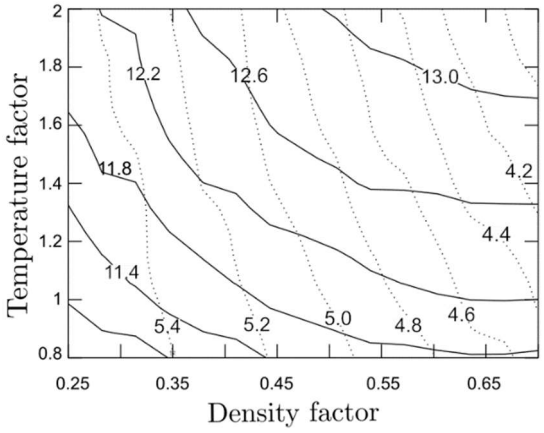


Fig. 2.Contours of constant ^{3}He and T absorption (dotted lines) and collisional bulk ion heating by resonant ^{3}He ions and tritons (solid lines) in MW as a function of plasma density and temperature factor. The total ICRF power is 6 MW and the total T NBI power is 11 MW.

ACKNOWLEDGMENT

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. Dani Gallart would like to express his gratitude to “La Caixa” for support of his PhD studies *and* to acknowledge FuseNet for financial support of some of the courses and events that he has attended.

REFERENCES

[1] F. Romanelli and JET EFDA Contributors 2013, *Nucl. Fusion* **53** 104002.
[2] M.J. Mantsinen, et al., Europhysics Conference Abstracts 2015, vol. 39E, pp. P2.171: 1-4
[3] L.-G. Eriksson, T. Hellsten, and U. Willén, Nuclear Fusion **33**, 1037, 1993
[4] P.M. Stubberfield and M.L. Watkins, *Multiple Pencil Beam*, JET-DPA(06)/87, 1987
[5] M. Romanelli, et al, *JINTRAC: A System of Codes for Integrated Simulation of tokamak Scenarios*, Plasma and Fusion Research **9**, 3403023, 2014.

Regional Arctic sea ice predictability and prediction on seasonal to interannual timescales

Rubén Cruz García, Virginie Guemas, Matthieu Chevallier
Barcelona Supercomputing Center
ruben.cruzgarcia@bsc.es

Abstract- *The fast depletion of the Arctic sea ice extent observed during the last three decades has awakened concerns about the consequences of such changes at hemispheric scales, and opened socio-economic opportunities such as maritime transport. This PhD project aims at investigating the sources of predictability and prediction skill of Arctic sea ice conditions at the regional scale. The first months have been dedicated to the investigation of the mechanisms behind the development of model systematic errors in seasonal regional predictions.*

I. INTRODUCTION

Over the last three decades (since the advent of satellite imagery), the Arctic sea ice extent has experienced a steady depletion by about 3% per decade [1]. Whereas the average September sea ice extent over the period 1979-2000 was estimated to be 7.04 million km^2, a record low of 3.41 million km^2 was reached on 18 September 2012 as reported by the National Snow and Ice Data Center (NSIDC).

Such a rapid sea ice decline is projected to accelerate in the coming decades, with a summer Arctic ice-free expected within the next 50 years [2]. Advanced knowledge about the potential opening of maritime routes such as the Northern Sea route (north of Russia) and the Northwest Passage (through northern Canada) could offer faster and cheaper shipping between the Atlantic and Pacific [3, 4].

Information on the marine accessibility of Arctic seas and the duration of the ice-free season in the marginal ice zone (MIZ) would allow planning of the exploitation of resources, ship supplies and fishing and hunting activities, which are of particular interest for the Inuit populations. The growing polar ecotourism industry could also benefit from sea-ice predictions.

II. OBJECTIVES

This project has mainly two objectives:

1. Investigating the sources and mechanisms of predictability of the Arctic sea ice at regional scale, including in some case studies. This objective could be divided into three different sub-objectives:
   A.Attribution of mechanisms leading to successful predictions of the regional Arctic sea ice conditions, which goes beyond most current studies focused on the global scale.
   B. Diagnosing potential causes for failures in predicting the Arctic sea ice conditions in some regions.
   C. Improving the representation of processes, as the sea ice deformation, by increasing the horizontal resolution for example.
2. Investigating the mechanisms for the development of model biases during the predictions.

This project will rely on two state-of-the-art forecast systems, EC-Earth and CNRM-CM.

III. FIRST RESULTS

Within the second objective, we have obtained initial results.

The PredEC2.3 is a retrospective seasonal prediction experiment which climate predictions have been initialized on 1st May and 1st November every year from 1979 to 2012. 5 members were run with the EC-Earth2.3 model for each startdate. We estimated the drift, i.e. the evolution of the prediction bias as a function of the forecast time, compared to the NSIDC observational data.

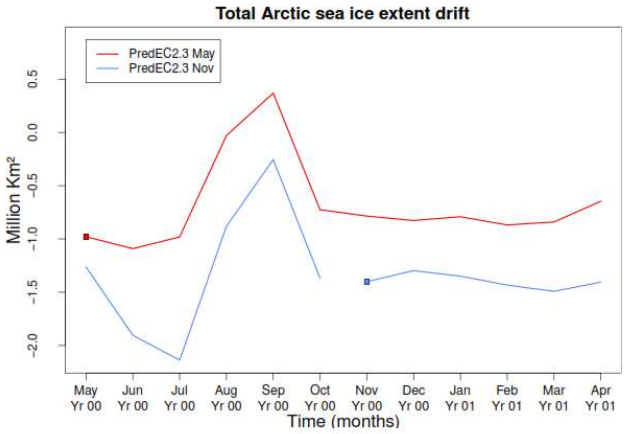


Fig. 1. Total Arctic sea ice extent prediction drift for the PredEC2.3 experiment initialized in May (in red) compared with the PredEC2.3 experiment initialized in November (in blue).

The evolution of the bias from one month to the following throughout the year is similar whether we initialize the forecast in May or November (Fig. 1). Looking at the total

Arctic sea ice extent drift for the November initialized experiment, we can note that there is an underestimation of 1.5 million km from November to May with respect to the observational data, which makes a larger bias by 0.5 million km compared to the May initialized forecast. Note also the higher underestimation during the summer.

ACKNOWLEDGMENT

This work is conducted within the framework of the European SPECS (FP7 GA 308378) and PRIMAVERA (H2020 GA 641727) projects.

REFERENCES

[1] Meier, W. N., Stroeve, J., Barrett, A., and Fetterer, F.2012: A simple approach to providing a more consistent Arctic sea ice extent time series from the 1950s to present, The Cryosphere, 6, 13591368, doi:10.5194/tc-6-1359-2012.

[2] Koenigk, T. and L. Brodeau, 2013: Ocean heat transport into the Arctic in the 20th and 21st century in EC-Earth. Clim Dyn, doi: 10.1007/s00382-013-1821-x

[3] Hassol SJ. 2004. Impacts of a Warming Arctic: Arctic Climate Impact Assessment. Cambridge University Press: Cambridge, UK. http://www.amap.no/arcticclimate-impact-assessment-acia (accessed 6 July 2014).

[4] Smith LC, Stephenson SR. 2013. New Trans-Arctic shipping routes navigable by midcentury. Proc. Natl. Acad. Sci. U.S.A. 110: 4871–4872, doi:10.1073/pnas.1214212110.

Genomic Instability Promoted by Expression of Human Transposase-Derived Gene

Elias Rodríguez-Fos[1], Santi González[1], Montserrat Puiggròs[1], Anton G. Henssen[2], Alex Kentsis [2,4,5] and David Torrents[1,3]
[1]Joint BSC-CRG-IRB Research program in Computational Biology, Barcelona Supercomputing Center (BSC-CNS)
[2]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA)
[4]Weill Cornell Medical College, Cornell University, New York, NY, USA
[5]Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Abstract- *DNA Transposases are enzymes that recognize and catalyze the movement of mobile elements in the human genome known as transposons. There are abundant transposase-derived genes in the human genome that have been conserved through evolution. Some of them, such as PGBD5, maintain their enzymatic activity in human cells. The expression of PGBD5 has been related to mobilization of DNA transposons through a motif specific cut and paste mechanism across the genome. The excision and insertion mechanism of transposable elements can cause genomic rearrangements and have a potential mutagenic activity in specific disease cases such as cancer. In this study, we analyze how the expression of PGBD5 leads to genomic instability*

Docking through Democracy Re-ranking protein-protein decoys with a voting system

Didier Barradas-Bautista[1,4] , Iain H. Moal[2], Juan Fernández-Recio[1,3]
[1]Barcelona Supercomputing Center, [2]EMB-EBI Wellcome Trust Genome Campus, Hinxton, Cambridge,[3]Joint BSC-CRG-IRB Research Program in Computational Biology, [4]Universitat de Barcelona.
didier.barradas@bsc.es

Abstract-*We have develop a machine learning framework to enhance protein-protein docking results, using Schulze voting method applied to several models from Support Vector machines.*

Block-Based Execution on an Integrated Vector-Scalar In-Order Core

Milan Stanic1, Oscar Palomar2
[1]Barcelona Supercomputing Center, [2] University of Manchester
milan.stanic@bcs.es, oscar.palomar@manchester.ac.uk

Abstract-*In the low-end processor mobile market, power, energy and area budgets are significantly lower than in the server/desktop/lap-top/high-end mobile markets. It has been shown that vector processors are a highly energy-efficient way to increase performance but adding support for them incurs area and power overheads that could not be acceptable for low-end mobile processors. In this work, we propose an integrated vector-scalar design that mostly reuses scalar hardware to support the execution of vector instructions. The key element of the design is our proposed block-based model of execution that groups vector instructions to execute them in a coordinated manner.*

I. INTRODUCTION

In the last 15 years, energy consumption and power dissipation have become crucial design concerns for almost all computer systems due to several reasons: for example, technology feature size scaling leads to higher power density and therefore to costly cooling. While power dissipation is critical for high-performance systems such as data centers due to large power usage, battery life is a primary concern for mobile systems.

Driven with this goal, researchers have focused on improving performance in an energy-efficient way. Vector processors [1] are energy efficient architectures that yield high performance whenever there is enough data-level parallelism (DLP) [2]. Besides the long and successful history of vector processors in supercomputers, vector units have been adopted in designs of microprocessors [3, 4, 5]. Also, SIMD multimedia extensions [6, 7] are often included in modern microprocessors. Recent research on vector processors shows that they can be a good match even for applications from domains such as column-store databases [8]. The Xeon Phi is a recent massively parallel x86 microprocessor designed by Intel and is based on the Larrabee [9] GPU, that contains a 512-bit SIMD vector processing unit in each core.

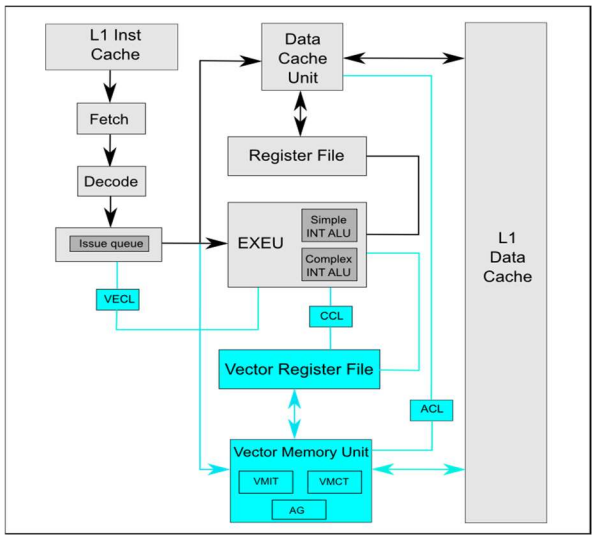
This paper contributes a method to increase the performance of the low-power, low-end embedded systems in an energy-efficient way. The energy efficiency is accomplished by modifying a scalar core to execute vector instructions on the existing scalar infrastructure. In particular, we propose an integrated vector-scalar design that combines scalar and vector processing mostly using existing resources of an energy-efficient processor (in our evaluation environment, it is based on the ARM Cortex A7). In addition to a design that uses a conventional vector execution model, we also contribute a novel block-based model of execution for vector computational instructions.

II. INTEGRATED DESIGN

As a baseline, we use a scalar core based on the highly energy-efficient ARM Cortex-A7. It is an in-order, dual-issue processor that implements the ARM v7 architecture with an 8-stage pipeline (gray blocks in Figure 1).

In our proposed integrated vector-scalar design, we attempt to maximize the reuse of resources already present in the baseline scalar core (white blocks in Figure 1) while adding support for vector instructions. While the front-end of pipeline is the same (fetch and decode stages), in the back-end we added two structures to support the execution of vector instructions on the scalar core: a vector register file, and a vector memory unit (blue blocks in Fig. 1). There is also additional logic that controls the execution of vector instructions: vector execution control logic (VECL), aliasing control logic (ACL) and chaining control logic (CCL). VECL is added in the issue stage to support the execution of computational vector instructions. ACL exchanges information between the vector memory and the data cache unit and forces scalar and vector memory instructions to be executed in-order. CCL is responsible for the execution of chained dependent computational instructions.

Fig. 1. Block diagram of the integrated design.



A. Execution of Vector Computational Instructions

We study two alternatives for executing the vector computational instructions on the existing scalar FUs: 1) the One-By-One model of execution (OBO), in essence the classic vector execution model, in which every instruction is executed to completion, i.e. for all the operations of the vector; and 2) a novel execution model called Block-Based Execution (BBE). In this model, for a block of consecutive vector computational instructions, first all operations on the first element are executed, then the operations of the second element, and so on. Fig. 2 shows an example with a sequence

of vector instructions, illustrating the difference of the two execution models. For this example, we assume that vector instructions operate on floating-point data by using a single floating-point unit and a single data cache port. The first *vecload* instruction is executed in the same way and at the same time on both models, since the models refer only to computational instructions. Regarding computational vector instructions, in OBO (Fig. 2 (a)) all operations of one vector computational instruction (*vecadd*) are executed, and then we move on to the next vector instruction (*vecsub*). In BBE (Fig. 2 (b)), several consecutive vector computational instructions form a block of vector instructions, and we execute one operation from each instruction of the block and repeat this for each operation in the block of vector instructions. In the example, we execute one operation from *vecadd* and then one operation from *vecsub*. The process ends once all operations are computed. The next subsection describes the BBE model in more detail.

A. Block-Based Execution

To support this model of execution, we added a small table that keeps the information of the instructions of the block and simple control logic. In this paper, blocks of vector computational instructions are formed dynamically in a very simple way. Once a computational vector instruction is ready for execution, the control logic examines the next instructions in the issue queue and adds them to the block if they are vector computational instructions, until another instruction type (a scalar or vector memory instruction) is encountered or the block is full.

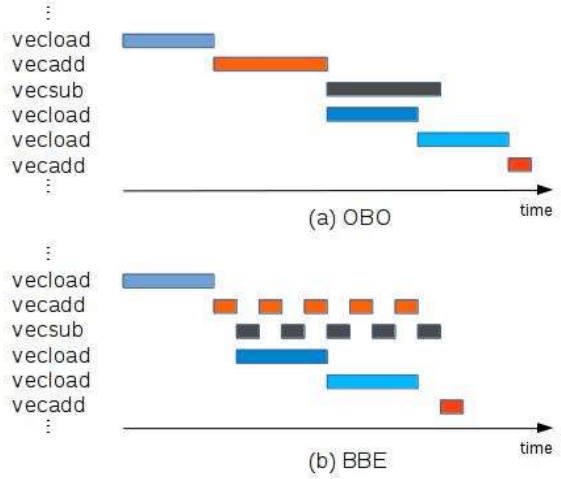


Fig. 2. An example of code with vector instructions executed with one ALU assuming (a) the one-by-one and (b) the block-based execution model.

The number of vector instructions that can be executed in parallel or with chaining using the OBO model is restricted by the number of available ALUs. BBE does not have this limitation, allowing for execution of more vector instructions in parallel. Inherently, more dependent instructions can be chained (scalar bypass logic can be reused) since one vector instruction does not occupy the ALU for all its elements in continuous cycles, and thus it can be interleaved with other instructions using the same ALU. An important advantage of BBE over OBO or a classic vector unit is the following: while a block of vector computational instructions is under execution, BBE allows for the execution of subsequent scalar or vector memory instructions if they are ready for execution and there are free functional units that can execute them. In Fig. 2 (b), the second *vecload* instruction can start execution just after the *vecsub* started with execution of the first operation.

III. CONCLUSION

Using a vector processor is one of the most energy efficient ways of achieving high performance for a wide number of applications that contain a significant degree of DLP. Power dissipation, energy consumption and area are critical concerns in processor design, especially for embedded systems in the low-end market. In this paper, we propose the integrated vector-scalar design that allows for execution of vector computational instructions mostly reusing resources of an in-order core. We also propose block-based execution model to execute vector computational instructions.

REFERENCES

[1] K. Asanovic. *Vector Microprocessors*. PhD thesis, University of California, Berkeley, May, 1998.

[2] Y. Lee et al. *Exploring the trade-offs between programmability and efficiency in data-parallel accelerators*. In ISCA 38, pages 129-140, 2011.

[3] C. Kozyrakis and D. Patterson. *Overcoming the limitations of conventional vector processors.* In Proceedings of the 30th ISCA, pages 399-409, 2003.

[4] R. Espasa et al. *Tarantula: a vector extension to the Alpha architecture*. In ISCA 29, pages 281-292, 2002.

[5] C. F. Batten. *Simplified vector-thread architectures for flexible and efficient data-parallel accelerators.* PhD thesis, Cambridge, MA, USA, 2010. AAI0822514.

[6] S. Thakkar and T. Huff. Internet streaming SIMD extensions. Computer, 32:26-34, December 1999.

[7] M. Buxton et al. Intel AVX: *New frontiers in performance improvements and energy efficiency*. White paper, 2008.

[8] T. Hayes et al. *Vector extensions for decision support dbms acceleration*. In MICRO 45, pages 166-176, 2012.

[9] L. Seiler et al. Larrabee: *a many-core x86 architecture for visual computing.* In SIGGRAPH '08, pages 18:1-18:15, 200

Photoprotection and triplet energy transfer in higher plants: the role of electronic and nuclear fluctuations

Lorenzo Cupellini*, Sandro Jurinovich, Ingrid G. Prandi, Stefano Caprasecca and Benedetta Mennucci
Dipartimento di Chimica e Chimica Industriale, University of Pisa
*lorenzo.cupellini@for.unipi.it

Abstract— *The quenching of Chlorophyll triplets by triplet energy transfer (TET) to carotenoids is one of the photoprotection strategies in photosynthetic organisms, and prevents singlet oxygen formation.*
Here we present the study of TET rates in a minor light-harvesting complex (LHC) of higher plants, using a fully atom- istic strategy that combines a molecular dynamic simulation a polarizable quantum/classical calculation.
We find that structural fluctuations of the LHC can largely enhance the TET rates, which are in the sub-nanosecond scale, in agreement with experimental findings.

Photosynthetic organisms employ several photoprotection strategies to avoid damage due to the excess energy in high light conditions. Among these, quenching of triplet clorophylls (Chls) by neighboring carotenoids (Cars) is fundamental in preventing the formation of singlet oxygen.
Singlet excited Chl∗ can decay into triplets (3 Chl∗) which sensitize molecular oxygen to form singlet oxygen, which induces damage in its local environment by destroying lipids and nucleic acids and proteins.[1], [2], [3] Cars are able to accept the triplets from chlorophylls (chls) by triplet energy transfer (TET), and dissipate the excess energy to heat.[4] The efficiency of Chl triplet quenching is 95% in antenna complexes of Photosystem II in higher plants, and the timescale of TET from Chls to Cars has been found to be faster than 500 ps in the major light-harvesting complex of Photosystem II (LHCII).[5]

TET is a spin-allowed process that consists in the transfer of a triplet configuration from a donor to an acceptor molecule. Because TET is based on the Dexter-like mechanism of electron exchange, and it requires an overlap between the molecular orbitals of donor and acceptor, thermal fluctuations are expected to play a relevant role in determining the coupling distribution. Here, we present a fully atomistic strategy, combining classical molecular dynamics (MD) with a hybrid time-dependent density functional theory (TDDFT)/polarizable MM description, to describe TET in the natural environment of the LHC.

In particular, we focused on CP29 (or Lhcb4), a minor light-harvesting complex of the Photosystem II whose crys- tal structure was recently obtained by Pan et al. at high resolution. [6] CP29 contains two strongly couplled Car- Chl clusters, namely those formed by Lutein (Lut) and Violaxanthin (Vio) with the three closest Chls. These two clusters are characterized by a similar arrangement of the Chls around the Car (See Figure 3).

The rate of the TET process can be related to the electronic triplet coupling by Fermi's Golden Rule:

$$kTET = \frac{2\pi}{\hbar^2} |VDA|^2 JDA \qquad (1)$$

Where VDA is the electronic coupling between initial and final states and JDA is the spectral overlap between the Franck-Condon weighted densities of states of donor and acceptor. Here we employ the fragment spin difference (FSD) scheme, a method to compute accurate triplet couplings starting from the eigenstates of the electronic Hamiltonian, namely the adiabatic states. [7] The spectral overlap was obtained from spectroscopic data.
We computed the TET couplings along 100 uncorrelated frames of an 80 ns MD simulation. In Figure 1 we compare the MD rms ($\sqrt{\langle V^2 \rangle}$) couplings with those computed on the crystal structure. In all pairs, except Lut-Chl a610, the MD average coupling is larger than the corresponding one obtained from the crystal. In particular, the coupling between Vio and Chl a603 nearly shows a six-fold increase. This is due to a limited number of favorable configurations with very large coupling values: in fact, excluding the largest 10 couplings result in a 40% drop of the rms coupling, indicating that these configurations account for more than half of the average TET rate.

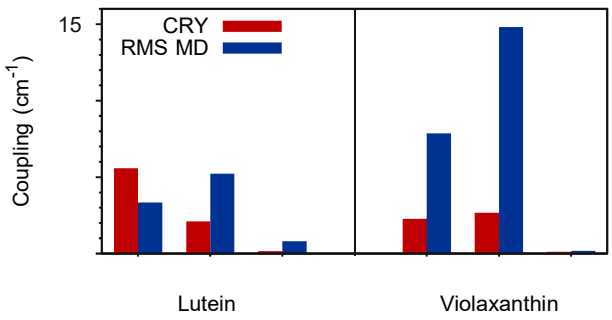


Fig. 1. Comparison between the crystal structure couplings (red) and theMD averaged couplings (blue).

A geometrical analysis of the TET can be performed using the volume of the intersection between the Van der Waals regions of the interacting pigments, defined as a union of interlocking spheres positioned on the atoms of

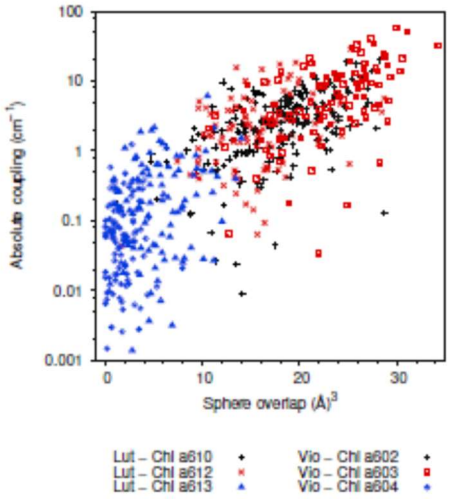


Fig. 2. Scatter plot of absolute coupling values (logarithmic scale) versus geometric overlap. Different Car-Chl pairs are shown in different colours.

the π-backbone with a radius 1.4 times the Van der Waals radius of the atom. To see if the geometric overlap can explain the coupling fluctuations, in Figure 2 we correlate the absolute coupling of all Car-Chl pairs to the geometric overlap. Despite the simplicity of this model, the magnitude of the coupling generally follows the geometric overlap.

The coupling values presented and discussed above are here used to compute the TET rates as obtained from (1), where V^2_{BA} is an average of the squared couplings along the MD. All the results are reported in Table I. In the same Table we also report the TET times obtained from the couplings calculated at the crystal structure. Figure 3 shows the transfer times and relates them to the arrangement of the pigments. In all pairs, except Lut-Chl a610 and a603-a609, the MD average time is shorter than the corresponding one obtained from the crystal structure, by more than one order of magnitude.

These data correlate well with experimental observations. The timescale of TET from Chls to Cars have in fact been found to be faster than 500 ps in the major light harvesting complex of Photosystem II (LHCII).[5] Our results show that the TET quenching mechanism strongly depends on the fluctuations of the surrounding environment. Notably, relying on the crystal structures may result in an underestimate of TET couplings and rates.

TABLE I

TET RATES AND TRANSFER TIMES FOR ALL PAIRS CALCULATED AS AVERAGE ON THE MD SNAPSHOTS. TET TIMES ARE ALSO REPORTED FOR THE SIMULATION ON THE CRYSTAL STRUCTURE (TIME@CRY).

Pair		k_{TET} (s^{-1})	time@MD	time@Cry
Lut	a610	2.13×10^9	470 ps	170 ps
	a612	5.24×10^9	190 ps	1.2 ns
	a613	1.21×10^8	8.2 ns	260 ns
Vio	a602	1.14×10^{10}	140 ps	1.1 ns
	a603	4.05×10^{10}	25 ps	760 ps
	a604	5.20×10^8	190 ns	540 ns

Fig. 3. TET time constants for the pairs investigated in this work. The thickness of the lines connecting the pigments represents the order of magnitude of the transfer time constant.

ACKNOWLEDGMENT

This work has been accepted for publishing [8]. L.C., S.J., S.C. and B.M. acknowledge the European Research Council (ERC) for financial support in the framework of the Starting Grant (EnLight - 277755). I.G.P. acknowledges CNPq - Brazil for PhD scholarship (236693/2012-3).

REFERENCES

[1] M. Ballottari, M. Mozzo, J. Girardon, R. Hienerwadel, and R. Bassi, "Chlorophyll triplet quenching and photoprotection in the higher plant monomeric antenna protein Lhcb5.," J. Phys. Chem. B, vol. 117, pp. 11337–48, Sept. 2013.

[2] R. Croce and H. van Amerongen, "Natural strategies for photosynthetic light harvesting.," Nat. Chem. Biol., vol. 10, pp. 492–501, June 2014.

[3] M. Mozzo, L. Dall'Osto, R. Hienerwadel, R. Bassi, and R. Croce, "Photoprotection in the antenna complexes of photosystem II: role of individual xanthophylls in chlorophyll triplet quenching.," J. Biol. Chem., vol. 283, pp. 6184–92, Mar. 2008.

[4] R. Bittl, E. Schlodder, I. Geisenheimer, W. Lubitz, and R. J. Cogdell, "Transient EPR and Absorption Studies of Carotenoid Triplet Formation in Purple Bacterial Antenna Complexes," J. Phys. Chem. B, vol. 105, pp. 5525–5535, June 2001.

[5] R. Scho¨ del, K. D. Irrgang, J. Voigt, and G. Renger, "Rate of carotenoid triplet formation in solubilized light-harvesting complex II (LHCII) from spinach.," Biophys. J., vol. 75, pp. 3143–53, Dec. 1998.

[6] X. Pan, M. Li, T. Wan, L. Wang, C. Jia, Z. Hou, X. Zhao, J. Zhang, and W. Chang, "Structural insights into energy regulation of light-harvesting complex CP29 from spinach.," Nat. Struct. Mol. Biol., vol. 18, pp. 309–
15, Mar. 2011.

[7] Z.-Q. You and C.-P. Hsu, "The fragment spin difference scheme for triplet-triplet energy transfer coupling.," J. Chem. Phys., vol. 133, p. 074105, Aug. 2010.

[8] L. Cupellini, S. Jurinovich, I. G. Prandi, S. Caprasecca, and B. Mennucci, "Photoprotection and triplet energy transfer in higher plants: the role

Modelling the Co-evolution of Trade and Culture

Simon Carrignon[1,2], Jean-Marc Montanier[1] and Xavier Rubio-Campillo[1]
[1]*Barcelona Supercomputing Center,* [2]*Universitat Pompeu Fabra*
simon.carrignon@bsc.es

Abstract- *We presents a new framework to study the co-evolution of cultural change and trade. The design aims for a trade-off between the flexibility necessary for the implementation of multiple models and the structure necessary for the comparison between the models implemented. To create this framework we propose an Agent-Based Model relying on agents producing, exchanging and associating values to a list of goods. We present the key concepts of the framework and two examples of its implementation which allow us to show the flexibility of our framework. Moreover, we compare the results obtained by the two models, thus validating the structure of the framework. Finally, we validate the implementation of a trading model by studying the price structure it produces.*

I. Introduction

Cultural change comprises processes that modify spread of information by social interaction within a population [1] and numerous social scientists are using an evolutionary framework to model this [2].

Here we use this framework to study economics, a social activity that depends on particular cultural traits: the value attributed to goods used to trade during the economic activity. Multiple cultural processes could influence the way those values evolve through space and time leading to different trade dynamics.

We focus on the way those values are transmitted and vary form individual to individual, and on the bias that affect this transmission. We propose a framework that allow us to implement and test hypotheses and claims made about the nature of such transmission processes and bias and study how those claims and hypotheses affects a given economy.

II. Framework

To explore the co-evolution between trade and cultural change we developed a framework where the different agents produce and trade goods. The model is composed of a population *Pop* of *m* agents. Each agent *i* is defined by 2 vectors Q^i and V^i of size *n*. Q^i store the quantity of each good owned by *i* and V^i represents the price estimated by *i* for each of the *i* good.

Algorithm 1 Model

```
1:  INITIALIZATION:
2:  for i ∈ #Pop do                    ▷ Initialize the agent with no goods and a random value vector
3:      Q^i = (0, ··· , 0)
4:      V^i = (v^i_0, ··· , v^i_n)     ▷ The values of v^i_j are selected randomly
5:  end for
6:  SIMULATION:
7:  loop step ∈ TimeSteps
8:      for i ∈ Pop do
9:          Production(Q^i)
10:     end for
11:     for i ∈ Pop do
12:         for j ∈ Pop do
13:             TradeProcess(V^i, Q^i, V^j, Q^j)
14:         end for
15:     end for
16:     for i ∈ Pop do
17:         ConsumeGoods(Q^i)          ▷ All goods are consumed
18:         if (step mod CulturalStep) = 0 then
19:             CulturalTransmission(V)
20:             Innovation(V^i)
21:         end if
22:     end for
23: end loop
```

Given the prices attributed by the agents for each goods (V^i), trade are done or not (l.13). Given the quantities (Q^i) gathered, a score reflecting the ``economic success'' of each agent is attributed (l.17). Finally, the value attributed to each good V^i is modified (l.19-20).

We propose two different ways to implement this modification:

1. Neutral Model: agent randomly copy a V^i among the population.
2. Trade Model: agent tends to copy more often the V^i of the most successful agents (i.e. with high score).

III. Results

A. Distribution of Cultural Variants

We first compare the impact of different *CulturalTransmission* mechanism on the distribution of frequencies of traits (the belief about the price of each goods).

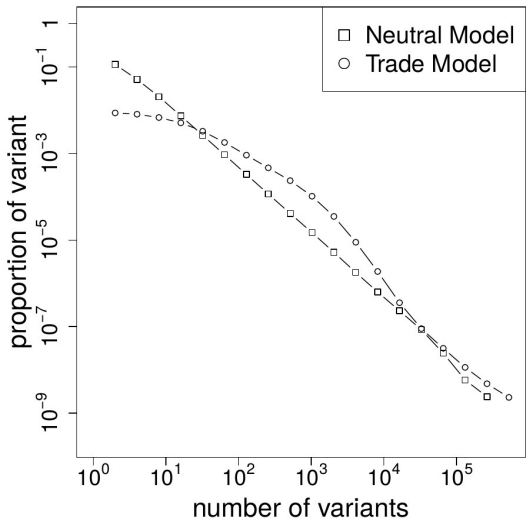


Fig. 1. Comparaison of the distribution of frequencies between the neutral and the trade model.

The figure 1 shows that when *CulturalTransmission* is neutral (agents randomly copy prices) the distribution follow the well know power law [2] but when transmission is not neutral but biased by the economical success of the agents, the power law disappear.

B. Economic Dynamics & Equilibrium

Position figures and tables at the tops and bottoms of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns.

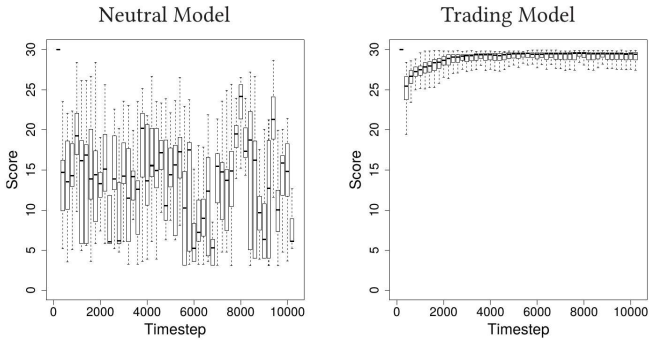


Fig. 2. Evolution of the score within the two different models for two typical run with 500 agents and 3 goods evolving during 10000 timesteps.

As expected when *CulturalTransmission* is random (i.e., agents modify their belief about the prices randomly), the scores evolve randomly (fig 2, left) whereas when a non random copy mechanism is used (i.e. agents tend to copy score of successful agents), scores increased toward the maximum score.

As shown by the figure 3, the raise of the score of the agents comes from the fact that the mechanism of *CulturalTransmission* biased by the economic success of the agents allows them to quickly estimates prices that converge toward their optimal value . Thus it allows them to make more efficient trade and increase their economic success (see also [4]).

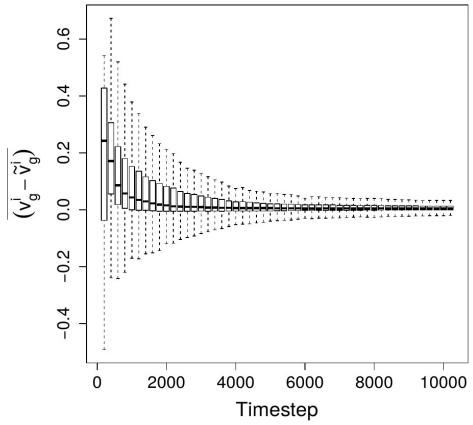


Fig. 3. Evolution of prices toward optimum prices. The figure represents the mean of the difference between a given price for one good g (v^i_g) and the optimal value of this price ($\tilde{v}^i_g$), computed at each timestep for each goods and for 100 runs in a setup with 500 agents where 3 goods are trade.

CONCLUSION

Integrating cultural and economic dynamics into an evolutionary framework is a good candidate to study such systems. It allows one to study precise mechanisms and to easily test and compare different model of such mechanisms.

In future works we hope to fruitfully apply that tool to validate, interpret and propose hypotheses about economics and cultural dynamics at work during the Roman Empire.

ACKNOWLEDGMENT

The Funding for this work was provided by the ERC Advanced Grant EPNet (340828) and the SimulPast Consolider Ingenio project (CSD2010-00034) of the former Ministry for Science and Innovation of the Spanish Government.

REFERENCES

[1] Robert Boyd and Peter J Richerson. *The origin and evolution of cultures*. Oxford University Press, 2005.

[2] Joseph Henrich and Richard McElreath. The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, 12(3):123–135, 2003.

[3] R. Alexander Bentley, Matthew W. Hahn, and Stephen J. Shennan. Random drift and culture change. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1547):1443–1450, 2004.

[4] Herbert Gintis. The emergence of a price system from decentralized bilateral exchange. *Contributions in Theoretical Economics,* 6(1):1–15, 2006.

Simulating Gravitational Collapse with Arbitrary-Precision Arithmetic

Daniel Santos-Oliván, Carlos F. Sopuerta
Institut de Ciències de l'Espai (CSIC-IEEC), c/ de Can Magrans s/n, 08193 Cerdanyola del Vallès, Spain.
santos@ice.cat

Abstract-*The collapse of smooth initial conditions into Black Holes is an important phenomenon to unlock fundamental aspects of the gravitational theory. In this paper we go closer to the formation of the apparent horizon using arbitrary-precision arithmetic (MPFR library) for examining the finer structure that forms during the collapse.*

I. INTRODUCTION

Gravitational collapse is one of the most interesting problems in Einstein's General Relativity. In nature, stars with more than 15-20 solar masses will end their life collapsing into Black Holes (BH) once the nuclear reactions at their cores are not able anymore to compensate the gravitational pull of their own weight. In these astrophysical objects, gravity is extremely strong and they are, therefore, the perfect laboratory for testing our theories of gravitation. In the last years, observations like the gravitational wave emission GW150914 measured by LIGO [1] have shown that Black Holes are not only a mathematical artifact but a true reality that populates all the cosmos.

Studying gravitational collapse is not only interesting because of their astrophysical implications. The appearance of singularities (spacetime points where relevant quantities diverge) from smooth initial conditions represent a key procedure to understand the fundamental features of the theory itself. In order to do this, we need eliminate some realistic aspects that hide some important questions. For example, the fact that astrophysical BHs have a minimum mass is because fermionic matter present quantum effects such as the Exclusion Principle that prevent the object to collapse if the gravitational pull is not strong enough. To avoid this we use bosonic matter that can generate arbitrarily small BH as it was shown by Choptuik [2].

The formation of singularities during the evolution of our system implies that high gradients are going to be present and therefore great accuracy is needed to fully understand the problem. During the development of previous work [3,4], we noticed that some interesting structure appear during the formation of the apparent horizon (AH) of the BH. As we are going to show, this could not be fully tracked with our numerical code so, at the moment, we are developing an improved version using the arbitrary precision library MPFR fully in parallel with OpenMP.

In this paper we are going to present the physical and numerical problem and the current status of the high-precision arithmetic solution that we are developing.

II. PHYSICAL PROBLEM

In order to answer the questions raised in the introduction, we are going to consider the simplest possible scenario: a self-gravitating scalar field in a spherically symmetric flat spacetime. The dynamics of the system is given by Einstein-Klein-Gordon system of couple non-linear partial differential equations:

$$G_{\mu\nu} = 2\left(\phi_{;\mu}\phi_{;\nu} - \frac{1}{2}g_{\mu\nu}\phi_{;\alpha}\phi^{;\alpha}\right), \qquad g^{\mu\nu}\phi_{;\mu\nu} = 0, \quad (1)$$

where $G_{\mu\nu}$ is the Einstein tensor, $g_{\mu\nu}$ is the metric of the spacetime and ϕ is the scalar field. Semicolon indicates covariant derivatives and Greek letters denote the spacetime indices.

Setting a characteristic scheme similar to the one we use in Refs. [3, 4], the problem reduces to the following: we prescribe initial conditions in an ingoing null slice $(r, h(r))$ at a null-timeu_0 where r is the radial coordinate and $h = d(r\,\phi)/dr$ is the field variable. The rest of the information on the slice can compute as direct integrals of $(r(u), h(u,r))$:

$$\bar{h}(u,r) = \frac{1}{r}\int_0^r h\, dr', \quad (2)$$

$$g(u,r) = exp\left(\int_0^r dr' \frac{\left(h-\bar{h}\right)^2}{r}\right), \quad (3)$$

$$\bar{g}(u,r) = g(u,r) - \frac{1}{r}\int_0^r \left(h-\bar{h}\right)^2 g\, dr'. \quad (4)$$

Once this is computed, we can evolve our variables to the next time using the system of ordinary differential equations (ODE):

$$\frac{dr}{du} = -\frac{1}{2}\bar{g}, \qquad \frac{dh}{du} = \frac{\left(h-\bar{h}\right)}{2r}(g-\bar{g}) \quad (5)$$

The first equation if (5) tells us that the points of our grid evolve towards the central region where points will focus in the region where the AH is forming. This feature of our coordinate system is very useful because allows us to have the precision we need without thinking on Adapting Mesh Refinement methods. In Fig.2 we plot the function: $A =$

$\bar{g}/g$ that gives an idea of the curvature induced in the spacetime by the collapsing scalar field and that takes a zero value when the AH is formed.

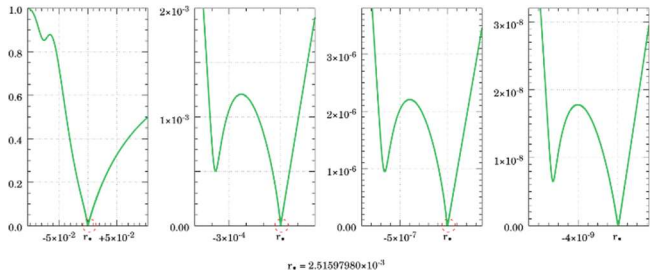


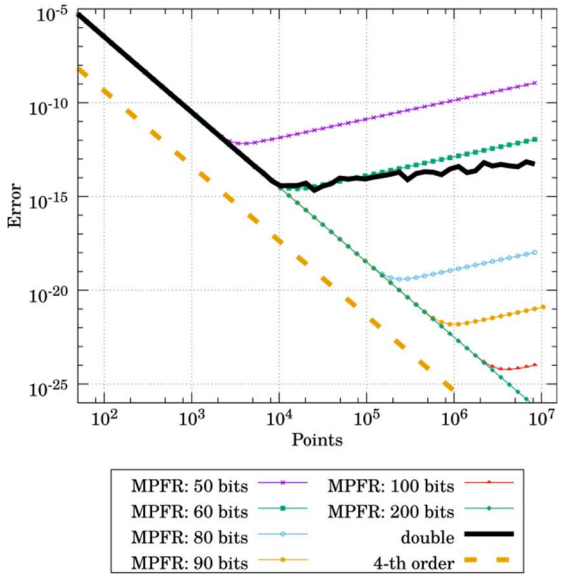
Fig. 1. . Plots of the function A, which monitors the formation of an apparent horizon (AH) for $A \to 0$. The three plots on the left are zoom-in areas of the region indicated by the red circles in the previous one. A repeated structure at different scales appears when we approach the formation of the AH.

When we approach this point, we observed a repeated structure of minima at different scales, each of them closer to AH formation in an exponential way. The last minima observed in the case that we present here is around 10^{-10}. We can guess that new minima appears at a scale below 10^{-12}, but at this scale our numerical noise starts to pile up making impossible to distinguish whether this is the case. At this point we need to go beyond double precision.

III. ARBITRARY-PRECISION ARITHMETIC

The basic numerical ingredients for evolving the system we are interested are the integrals defined in Eqs. (2-4) and (5) that determine the evolution of the system. The main bottleneck in computational time and precision are the integrals. We use arbitrary precision arithmetic with the library MPFR to decrease the numerical error of our simulations using OpenMP for computing the radial integral in blocks. In the MPFR library, one can choose the number of bits of precision. In Fig. 2 we plot the error of a test integral computed with a finite-difference fourth-order scheme for different bit precision. We observe how the error decreases as the expected fourth-order until it reaches round-off error, of course different for the chosen precision.

We can also see that we can improve almost ten orders of magnitude going from double (64-bit) to 100-bit precision.



This is true for this test case and one can expect that for a complete evolution the numerical error is going to be higher but the previous plot gives us the hope that this is the correct path of action. Once the full code is finished we should be able to go much closer to the AH formation and check if the repeated structure is still present at smaller scales.

ACKNOWLEDGMENT

The authors acknowledge high-performance computing resources provided by CSUC and CESGA and contracts ESP2013-47637-P and ESP2015-67234-P (MINECO). DS acknowledges support from a FPI contract BES-2012-057909 from MINECO.

REFERENCES

[1] B.P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), "Observation of Gravitational Waves from a Binary Black Hole Merger,", Phys. Rev. Lett., 116, 061102, 2016.

[2] M.W. Choptuik, "Universality and scaling in gravitational collapse of a massless scalar field", Phys. Rev. Lett., 70, 9, 1992.

[3] D. Santos-Oliván and C.F. Sopuerta, "New Features of Gravitational Collapse in Anti-de Sitter Spacetimes", Phys. Rev. Lett., 116, 041101, 2016.

[4] D. Santos-Oliván and C.F. Sopuerta, "Moving Closer to the Collapse of a Massless Scalar Field in Spherically Symmetric Anti-de Sitter Spacetimes", unpublished., arXiv:1603.03

Crowd Simulation and Visualization

Hugo Perez[1,2], Isaac Rudomin[1], Eduard Ayguade[1,2], Benjamin Hernandez[3],
Javier A. Espinosa-Oviedo[1,4], Genoveva Vargas-Solar[4]
[1]*Barcelona Supercomputing Center, Barcelona, Spain*
[2]*Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain*
[3]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[4]French National Council on Scientific Research (CNRS), LIG-LAFMIA,
BP 72 38402 Saint-Martin d'Hères, France
hugo.perez@bsc.es

Abstract- *This paper presents a methodology to simulate and visualize crowds. Our goal is to represent the most realistic possible scenarios in a city. Due to the high demand of resources a GPU Cluster is used. We use real data from which we identify the behavior of the masses applying statistical and artificial intelligence techniques. In order to take advantge of the processing power of the GPU cluster we use the following programming models during the characters simulation: MPI, OmpSs and CUDA. We developed different visualization schemes: a) In situ, b) Streaming, c) Web. The web scheme is the most flexible, allowing to interact in real time with the simulation through a web browser. For this scheme we use WebGL and Cesium.*

Keywords: Parallel Programming Models, HPC, GPU, AI, Computer Graphics.

I. INTRODUCTION

In this era of big cities one is confronted by emergency situations caused by traffic, natural disasters or special events such as concerts, sports events and protests which require the intervention of qualified personnel in order to generate an orderly and safe urban experience, and save lives. Data management and visualization can support real-time observation, understanding the behavior of this exodus and develop security strategies. We believe that technology and particularly data management and visualization can provide tools that can help to control this complex situation: real-time observation, understanding of the behavior of the people through on-line and post-mortem analytics, real-time decision making and recommendation are some of the activities that can be supported.

Our data processing and simulation are computationally expensive and critical thus we rely on HPC infrastructure with hybrid architecture (CPUs + GPUs) to produce an efficient solution. Heterogeneous clusters provide various levels of parallelism that need to be exploited successfully to harness their computational power totally. Our particular endeavors have been focused on the design, development and analysis of crowd simulations in these systems. Our first efforts [1] combined CUDA and MPI models for in-situ crowd simulation and rendering on GPU clusters and recently [2] we have adopted OmpSs. We proposed a task-based algorithm that allows the full use of all levels of parallelism available in a heterogeneous node.

This paper presents the methodology to simulate and visualize crowds. We address the challenge of visualize on real time and predict the behaviour of individuals and groups moving and evolving within real environments that uses information harvested from different sources.

II. DEVELOPMENT

We use the GeoLife GPS trajectory dataset [3] with data of 182 users, 17,621 trajectories of *ca.* 1.2 million Km. and 48,000+ hours. *These* data is used to compute spatio-temporal people flows in real crowds to provide data driven on-line crowd simulation, enhanced with real places geometric data running on GPU and HPC. Since the data set for a given place and time is sparse, we used agent based microsimulation to complement the actual trajectories in the dataset by using all the trajectories in similar moments that are available in the dataset to derive the most probable trajectories for the simulated vehicles or pedestrians.

Figure 1 illustrates the general overview of our approach using the web visualization scheme (the details of the different schemes are described in [1]) that addresses three main problems: (i) data harvesting, (ii) crowd simulation and analytics and (iii) visualization. We use existing temporal geo located observations concerning individuals' trajectories.

We apply data analytics techniques (temporal and spatial reasoning) for computing trajectories and for identifying crowds, that is people grouped in a sufficient close spatial region that adopt a specific "behaviour" referring to four

well known naïve crowd patterns: (i) casual crowd which is loosely organized and emerges spontaneously, people forming it have very little interaction at first and usually are not familiar with each other; (ii) conventional crowd results from more deliberate planning with norms that are defined and acted upon according to the situation; (iii) expressive crowd forms around an event that has an emotional appeal; and (iv) acting crowd members are actively and enthusiastically involved in doing something that is directly related to their goal.

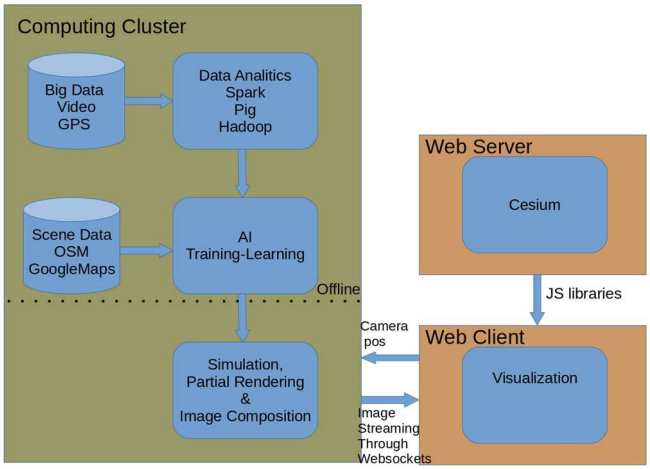


Fig. 1. General overview

Simulation and visualization is based on the work described in [2, 4]. However for clarity sake, we describe the basic charactiristics: Processing these tasks in parallel within a cluster, requires tiling and stencil computations. First, the navigation space (from now on it will be called the World) and information for all the agents is divided into zones. Each zone is assigned to a node which in turn is divided into sub-zones. Then, the tasks performed in each sub-zone can be executed in parallel by either a CPU or GPU. Stencil computations are performed on an inter and intra-node basis. A step by step description of the algorithm is included:

Step 1: Navigation space is discretized using a grid; then the resultant grid is discretized into zones. Each node will compute each zone.

Step 2: Divide each zone into tiles (sub-zones). A CPU or GPU will compute each sub-zone. Each CPU or GPU stores their corresponding tile of the world.

Step 3: Set up the communication topology between zones and sub-zones.

Step 4: Exchange borders (stencils).

Step 4a: (Intra-node) Exchange the occupied cells in the borders between sub-zones

Step 4b: (Internode) Exchange the occupied cells in the borders between zones

Step 5: Update position for each agent in parallel

Step 6: Agents' Information Exchange.

Step 6a: (Intra-node) Exchange agents' information that crossed a border and moved to another sub-zone

Step 6b: (Internode) Exchange agents' information that crossed a border and moved to another zone .

Simulated vehicles and pedestrians will use heatmaps derived from the dataset in order to follow the most popular routes. The details of these modifications are not within the scope of this paper and will be published later.

In general, the principle consists in mapping human perception of the space stemming from cameras and expressed in geographical coordinates (latitude, longitude), for example, into pixels. For instance, as shown in figure 2 "*give me the GPS coordinates of the users evolving in Beijing ordered by time*". Once this query has been evaluated by the appropriate data processing infrastructure (in the work presented here PigLatin [5] execution environment), results are transformed into the appropriate format. Textures and maps are retrieved in order to create the 3D space where individuals' movements will be visualized (simulated) according to the observed information.´

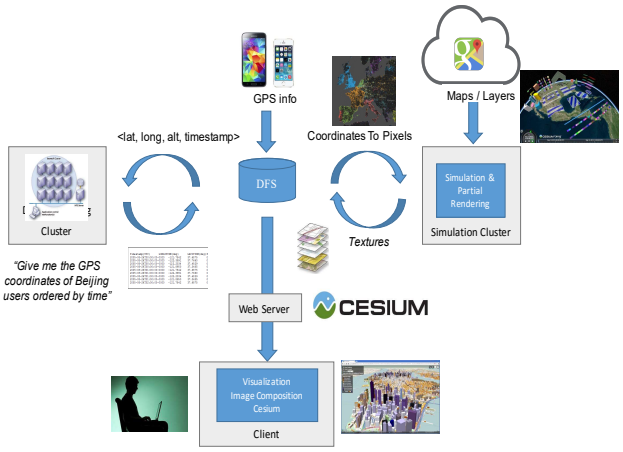


Fig. 2. Visualization process of individuals movement within urban 3D spaces.

III. CONCLUSIONS AND FUTURE WORK

Crowd sourced location data is used to compute spatio-temporal people flows in real crowds. We combine both to provide data driven on-line crowd simulation, enhanced with real places geometric data. This paper presented the general approach for simulating crowd behaviour and thereby supporting individuals' and crowd behaviour in public spaces. The main contribution is combining location based data collections previously harvested together with

online geo-tagged data for visualizing crowds at different levels of precision and detail, according to access control and privacy constraints. Our data processing and simulation process are computationally expensive and critical thus we rely on HPC and GPU infrastructures for producing an efficient solution.

As future work we will use automatic learning techniques (deep learning) so that the system can "react" to events and simulate a synthetic behaviour of the crowd.

ACKNOWLEDGMENT

This research was partially supported by: CONACyT doctoral fellowship 285730, BSC-CNS Severo Ochoa program (SEV-2011-00067), CUDA Center of Excellence at BSC, the Spanish Ministry of Economy and Competitivity under contract TIN2012-34557, and the SGR programme (2014-SGR-1051) of the Catalan Government.

REFERENCES

[1] Hernandez, B., Perez, H., Isaac, R., Ruiz, S., DeGyves, O., Toledo, L.: *Simulating and visualizing real-time crowds on gpu clusters*. *Computacion y Sistemas* 18 (2014) 651–664

[2] Perez, H., Hernandez, B., Rudomin, I., Ayguade, E.: *Task-based crowd simulation for heterogeneous architectures*. In: *Innovative Research and Applications in Next-Generation High Performance Computing*. Qusay F. Hassan (Mansoura University, Egypt). IGI Global 2016.

[3] Zheng, Y., Xie, X., Ma, W.-Y.: GeoLife: *A Collaborative Social Networking Service among User, location and trajectory*. IEEE Data Engineering Bulletin. 33, (2010).

[4] Perez, H., Hernandez, B., Rudomin, I., Ayguade, E.: Scaling Crowd Simulations in a GPU Accelerated Cluster. Springer 2016 (In progress)

[5] Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A.: Pig latin: *a not-so-foreign language for data processing*. In: Proc. of the 2008 ACM SIGMOD Int. Conference on Management of data (SIGMOD'08). ACM Press, Vancouver, Canada (2008).

Reproducing crowd turbulence with Verlet integration and agent modeling

Albert Gutierrez-Milla[1,2] and Remo Suppi[2]
[1]Barcelona Supercomputing Center
[2]Universitat Autònoma de Barcelona
albert.gutierrez@bsc.es, remo.suppi@uab.cat

Abstract *– High density crowds are risk situations that already had turned some events into disasters. There are particular emerging events in these crowds that had led to dangerous situations. One important phenomenon is named "crowd turbulence". It is produced by a propagation of forces among the mass and has already been the cause of several tragedies. We present a new approach to its representation and understanding by a hybrid model composed by two parts: physical interaction among the agents, and psychological factors that produce voluntary interactions. The focus of the present work is contributing with a model able to reproduce such events in a computationally efficient way on SIMD architectures.*

I. INTRODUCTION

A crowd under high density conditions is a potential disaster situation. Understanding and developing models to characterize the crowd helps the security assessment process for building design, event planning, evacuation planning, etc. Previous disasters such as Love Parade (Duisburg), or Hajj (Mina)[1] showed that there is still a lack of knowledge and a there is a need of understanding such situations. There is a particular phenomena reproduced in the case of crowds were the pressure is propagated through the mass potentially causing crashed chests. This effect is named "crowd turbulence". We present a model capable of reproduce crowd turbulence and we implemented a simulator for SIMD architectures using OpenCL.

II. MODEL

In crowd turbulences, interactions among the bodies are described as a wave propagating forces and the mass inertia. We consider that this event can be model by particle simulation and to complete the model we include human behavior in every person. To model the physics of the movement, navigation and inertia we chose Verlet method which integrates Newton's second law of motion. For the psychological part we use Agent Based Modelling (ABM) to model the voluntary actions of the population. Consequently, bodies of the people are modeled as particles and as agents.

Using the second order central method we express the equation of motion as finite differences. Formula 1 shows the equations for coordinates in a bidimensional space. For simplicity we will not consider the mass as variable and the value of the acceleration is a constant defined as a parameter of the model and is applied until the agent reaches its v_{max}, then only the inertia phase is computed. The velocity chosen for the agents follows a normal distribution with a mean of 1.34m/s and a standard deviation of 0.26m/s[2], navigating towards a specific goal *e*. When two agents intersect the collision is solved by a simplified technique to solve inelastic collision losing kinetic energy. This is done by using a factor which depends on the intersection between two agents.

$$x_{n+1} = -x_{n-1} + 2x_n + a\Delta t^2$$
$$y_{n+1} = -y_{n-1} + 2y_n + a\Delta t^2$$

Even though agents are dragged by the mass in case of turbulence, people do not move as particles. They offer resistance to external interactions and also try to gain free space by pushing others expressing their will. Thus, we model intentional and involuntary pushes. Involuntary take place when a person is moved by the crowd and those are modeled by Verlet integration. Voluntary interactions come from psychological factors and are modeled in every agent. Moreover not every person has a tendency in pushing other. Apathy, empathy or neurotic behavior may have a direct impact in the behavior of people during evacuations. To map these non-homogeneous human treats to the model we describe their tendency to push others.

Every agent follows a path declared by a graph were there are some areas named "decisions points" which define the navigation route of the agent. These are nodes, and when the agent is close to it, the goal position *e* is updated to the next node. We assume that agents know the shortest path and there is no uncertainty added. The size of the agent is modeled by a circle with a dimeter of 0.4m and a standard deviation of 0.1m.

III. SIMULATION

To analyze the recreation of crowd turbulences we use the definition of "pressure" (Formulas 2 and 3) proposed by Yu[3] which reflects the irregular/chaotic motion attending the velocity variance and the density. This formula has been used by other studies to validate the reproduction of crowd turbulences[4]. We used as value of R the mean size of the agents.

$$\rho_i = \sum_j \frac{1}{\pi R^2} exp\left(\frac{-\|\vec{r_j} - r_i(t)\|^2}{R^2}\right) \quad (2)$$

$$p = \rho_i Var(\vec{v_i}) \quad (3)$$

We implemented the previously described model in a parallel simulator. The agents are initialized in the GPU and then sent to the GPU. To parallelize the simulation we selected agent division. Every thread is in charge of one agent, executing the same kernel in parallel for all of them during every time step. For the data layout we use a SoA (Structure of Arrays) instead of an AoS (Array of Structures) because the known coalescing issues. Every thread will access the other threads structures but it only will write on its own to avoid any data inconsistency. Transactions between host and GPU occur at each iteration to store the coordinates for postprocesing.

IV. RESULTS

We executed the simulator on the GPU and we also executed a sequential version on a CPU to compare the performance with the parallel version. The experimentation platform is an Intel CPU with 2.1 GHz, 8GB memory, GPU Nvidia 750 GTX, 1.1GHz 2GB memory. The compiler is GCC 4.4.7 with Nvidia libraries for OpenCL 1.2. We used the optimization flag O3 for the compilation process. The scenario we used for the simulation is a T shaped synthetic area.

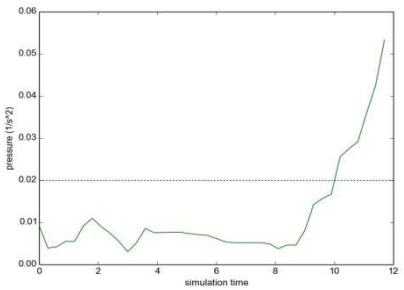


Fig. 1. Plot of the "pressure" in a crowd turbulence reproduced by our model.

Fig. 1 shows the "pressure" with the results of a crowd turbulence reproduced by our simulator. The marked threshold drawn by dots is the value of 0,02 m/s^2 which indicates the start of the turbulence. In the plot we see how the pressure lays behind the threshold from simulation steps 0 to 10. At time 9, pressure starts increasing and at time 10 overpasses the threshold starting the crowd turbulence increasing the local mass pressure and velocity.

GPU performance was compared with the sequential version executed in a CPU. Fig. 2 depicts the performance of both versions doubling the population between 512 and 4096 agent. The time is scaled in logarithmic scale. Increasing the number of agents the performance is comparatively improved in the GPU version making usage of idle SM and making a more efficient usage of the resources. Because of this, the speedup is of only 2 for 512 agents, but 11 for 4096.

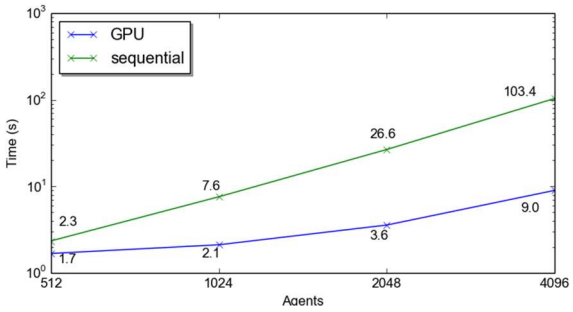


Fig. 2. Performance time in seconds of the sequential and GPU implementations of the simulator. The results are the total simulation time for 1,000 simulation steps.

V. CONCLUSIONS AND REMARKS

We presented a model able to reproduce crowd turbulences with an hybrid approach including a physical model and a psychological model. The simulation was able to reproduce crowd turbulences using the definition of "pressure" to validate the behavior. One of the main features of our model was the simplicity and the suitability for SIMD architecture. The performance of the GPU simulator showed a speedup ranging from 2 to 11 as we increase the number of agents and the work load. Future work may be focused on the improvement of the collision phase and the GPU performance as well as extending the model.

ACKNOWLEDGMENT

This research has been supported by the MINECO (MICINN) Spain under contract TIN2014-53172-P.

REFERENCES

[1] Helbing, Dirk, Anders Johansson, and Habib Zein Al-Abideen. "Dynamics of crowd disasters: An empirical study." Physical review E 75.4 (2007): 046109.

[2] Henderson, L. F. "The statistics of crowd fluids." Nature 229 (1971): 381-383.

[3] Yu, Wenjian, and Anders Johansson. "Modeling crowd turbulence by many-particle simulations." Physical review E 76.4 (2007): 046105.

[4] Golas, Abhinav, Rahul Narain, and Ming C. Lin. "Continuum modeling of crowd turbulence." Physical Review E 90.4 (2014): 042816.

Generation of a simulation scenario from medical data: Carto and MRI

M. López-Yunta*, X. Roca, J. Aguado-Sierra, M. Vázquez
Barcelona Supercomputing Center (BSC-CNS)
*marina.lopez@bsc.es

Abstract- *Multiphysic cardiac models give accurate simulations of normal and pathologic behavior of the heart. It can help to develop new treatments and medical devices. The complexity relies on both mathematical and geometrical models, so that HPC is needed to obtain accurate results using finite element method. From Magnetic Resonance Imaging a complete geometry, including atria and ventricles, is obtained through segmentation. Then the corresponding CAD is generated, where boundary conditions and properties of each heart region are fixed. Finally the volume mesh is built, essential to run simulations.*

A. Introduction

Cardiovascular diseases are the main causes of death in the world. It is important to find new pharmacology treatments and devices. On this field, simulations can help to understand the behaviour of normal and pathologic hearts. Different kind of problems are necessary to face when solving fluid-electromechanical cardiac simulations: geometry and mesh generation, fiber orientation, scar definition, model parameterization and boundary conditions. In this work, we focus on geometry, mesh generation and electrical characterization from the experimental medical data.

When speaking about cardiac infarction is important to differentiate the scar zones. Each zone has different electrical and mechanical properties that have to be included on the final mesh.

The final goal is to obtain a complete volume mesh with all the medical information included and able to run electromechanical cardiac simulations on it.

B. MRI-Based Geometry

Most electrophysiology simulations use simplified geometries based on ellipsoids. This is not enough once we introduce the mechanical problem. A complete heart geometry is needed, that is including atria and ventricles. This geometry is obtained from the Magnetic Resonance Imaging (MRI) through segmentation.

The MRI images are first filtered and then manually segmented using the software *Amira* [3]. Once the geometry contours of atria and ventricles are defined, the corresponding CAD is generated. The CAD is useful to fix boundary conditions and the properties of each heart region and finally create the first volume mesh (Figure 1).

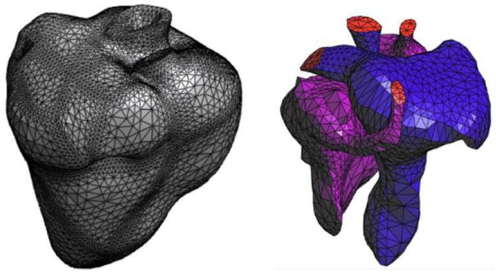


Fig. 1. Complete heart mesh (left) and cavities mesh (right).

When studying myocardial infarction and ventricular tachycardia, it is important to detect and differentiate the heterogeneous and the dense scar zones. The R1 sequences with gladolinium contrast are the best images to detect scar. Considering [4], first the maximum intensity pixel value within ventricles has to be detected. Observing the pixel intensity histogram on ventricles area, the maximum value corresponds to this maximal intensity pixel.

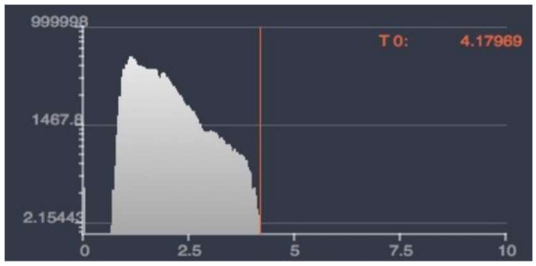


Fig.2: Ventricle pixel intensity histogram. Maximum value: 4.17969

The images are then filtered and the signal intensity normalized to that maximum signal intensity value found. Finally, the regions are classified by thresholds: Dense scar region (0.8 – 1) and heterogeneous region (0.5 - 0.8).

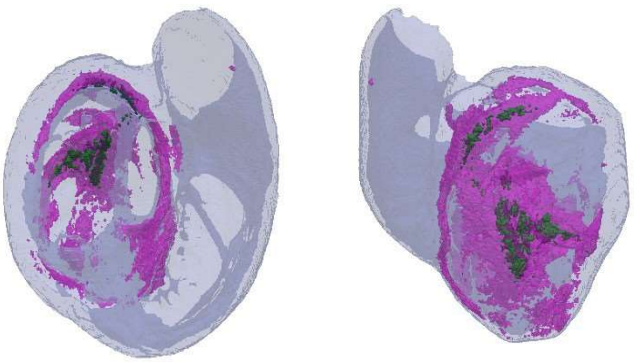


Fig. 3: Reconstruction of ventricles with heterogeneous scar (pink) and dense scar (green).

C. Electrical Properties form Carto

Electromechanical models have to be calibrated with experimental data. Carto data are obtained through a catheter introduced in the heart cavities and pericardia to record the electrical properties on hundreds points of heart surface. This data gives us information about conductivities, activation and restitution times on the different areas of the heart: healthy tissue, heterogenous scar and dense scar. To include this information on the mesh first it has to be compared with the MRI data, to find correspondences between areas.

The scar information coming from MRI data is 3D geometry and the one coming from Carto is surface information. To compare both data, we propose to solve the Eikonal equation on the 3D geometry reconstructed from an MRI segmentation.

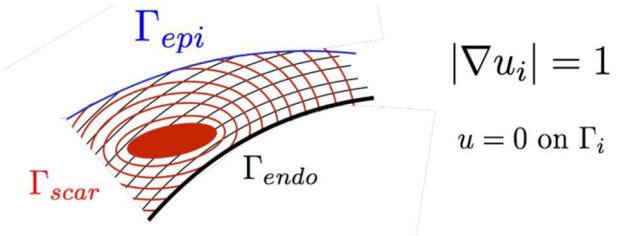


Fig. 4: Scheme of eikonal function applied to ventricles wall and intern scar

I f i = *scar,* the values of the Eikonal function on the walls describe the minimum distance to the scar. If i = *endo (*or i = *epi),* the corresponding values on the epicardium (or endocardium) wall describe the local thickness. Then a normalization of the distance to scar with local wall thickness is applied.

Considering the threshold 0.25 on the normalized surface map, we compare the result scar areas with the Carto data (Figure 4). With this threshold, the 25% of the local ventricle thickness from the epicardium and endocardium are taken into account to measure the distance from the walls to the scar.

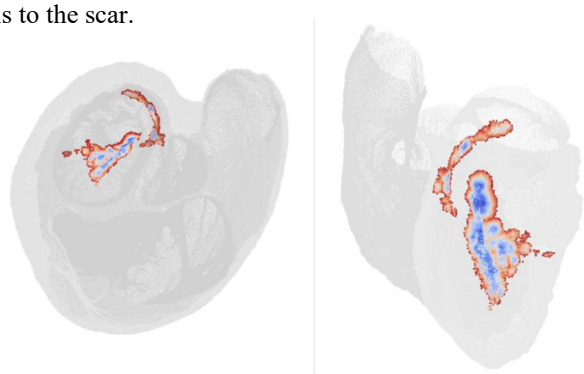


Figure 4: Distance to dense scar surface map normalized by local thickness.

Both methods give similar percentage of dense and heterogeneous scar areas, comparing with the total endocardium and epicardium areas (Figure 5).

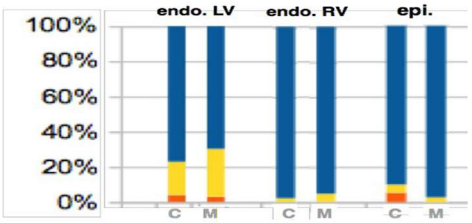


Figure 5: Heart zones areas of Carto data(C) and Mri data (M) in percentages (orange: dense scar, yellow: heter. Scar and blue: healthy zone)

The result relates scar areas taken form different experimental data (MRI and Carto). Now, point-to-point correspondences of data have to be found to get a transformation of Carto surface to MRI geometry. This transformation will be useful to take the electrical data from Carto (conductances and restitution times) of the scar and healthy areas and introduce on the mesh generated from MRI data to parameterized the electromechanical problem.

D. Simulations

The heart mesh obtained form experimental data, including properties of the scar and normal regions, is essential to run our electromechanical model [2]. Cardiac simulations including scar will be run on the multiphysics code Alya, optimized for HPC. The problem combines a mechanical contraction and electrical propagation model that run on the same mesh. Electromechanical simulations are our main objective and the final application of the current work presented.

Acknowledgment

The research leading to these results has received the support of the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government.

References

[1] J. Aguado-Sierra, A. Santiago, M. Rivero, M. López-Yunta, D. Soto-Iglesias, L. Dux-Santoy, O. Camara and M. Vázquez, "Fully-coupled electromechanical simulations of the LV dog anatomy usig HPC: Model testing and verification" *Statistical Atlases and Computational Models of the Heart – Imaging and Modelling Challenges,* 2015.

[2] P. Lafortune, R. Arís, M. Vázquez and G. Houzeaux, "Coupled electromechanical model of the heart: Parallel finite element formulations" *Journal of Numerical Methods in Biomedical Engineering,*, vol. 28, 2012, pp.72-86.

[3] http://www.fei.com/software/amira-3d-for-life-sciences/

[4] Y Tanaka, M. Genet, L. Chuan Lee, A. J. Martin, R. Sievers and E. P.Gerstenfeld, "Utility of high-resolution electroanatomic mapping of left ventricle using a multispline basket catheter in a swine model of chronic myocardial infarctation" *Heart Rhythm Society, 1547-5271.*

How Can We improve Energy Efficiency through User-directed Vectorization and Task-based Parallelization?

Helena Caminal, Diego Caballero, Juan M. Cebrián, Roger Ferrer, Marc Casas, Miquel Moretó,
Xavier Martorell and Mateo Valero
Barcelona Supercomputing Center

Abstract- *Heterogeneity, parallelization and vectorization are key techniques to improve the performance and energy efficiency of modern computing systems. However, programming and maintaining code for these architectures poses a huge challenge due to the ever-increasing architecture complexity. Task-based environments hide most of this complexity, improving scalability and usage of the available resources. In these environments, while there has been a lot of effort to ease parallelization and improve the usage of heterogeneous resources, vectorization has been considered a secondary objective. Furthermore, there has been a swift and unstoppable burst of vector architectures at all market segments, from embedded to HPC. Vectorization can no longer be ignored, but manual vectorization is tedious, error-prone, and not practical for the average programmer. This work evaluates the feasibility of user-directed vectorization in task-based applications. Our evaluation is based on the OmpSs programming model, extended to support user-directed vectorization for different SIMD architectures (i.e. SSE, AVX2, AVX512, etc). Results show that user-directed codes achieve manually-optimized code performance and energy efficiency with minimal code modifications, favoring portability across different SIMD architectures.*

Keywords SIMD, OmpSs, Performance, Vectorization, Energy Efficiency

I. INTRODUCTION

While transistor shrinking allows to include additional features on the die, the increasing power density prevents the simultaneous usage of all available resources. Instruction level parallelism (ILP) importance subsides, while data level parallelism (DLP) becomes a critical factor to improve the energy efficiency of microprocessors. Among other features, SIMD instructions have been gradually included in microprocessors for various market segments, from mobile to high performance computing (HPC). Each new generation includes more sophisticated, powerful and flexible instructions. The higher investment in SIMD resources per core makes extracting the full computational power of these vector units more important than ever.

From the programmers' point of view, SIMD units can be exploited in several ways, including: a) compiler auto-vectorization, b) low-level intrinsics or assembly code and c) programming models/languages with explicit SIMD support. Auto-vectorization in compilers has strong limitations in the analysis and code transformations phases that prevent an efficient extraction of SIMD parallelism in real applications. Low-level hardware-specific intrinsics enable developers to fine tune their applications by providing direct access to all of the SIMD features of the hardware. However, the use of intrinsics is time-consuming, tedious and error-prone even for advanced programmers. To facilitate the use of SIMD features, some programming models and languages have been extended with a new set of directives that allow programmers to guide the compiler in the vectorization process (e.g., OpenMP 4.0). This approach is high-level, orthogonal to the actual code and portable across different SIMD architectures.

In this abstract, we evaluate the efficiency of an implementation of a user-directed vectorization proposal using a task-based programming model. Our main contributions include:

- Development of a task-based version of a subset of benchmarks from the ParVec benchmark suite [2]. Due to space limitations we only show one of the six benchmarks we have ported.
- We present the code modifications necessary to generate a user-directed code version that achieves similar performance and energy results to those obtained with manual vectorization.
- We discuss our findings and propose improvements for both the manually vectorized versions and the user-directed vectorization module

II. METHODOLOGY

In this document we evaluate three versions of the codes, including: a) two manually-vectorized implementations, one based on pthreads and one based on the OmpSs programming model [4] (labelled pthreads and OmpSs, respectively), and b) a user-directed vectorization (labelled U.D.). Both user-directed and OmpSs versions were developed for this document. The user-directed code is compiled using the Mercurium source-to-source infrastructure. Mercurium's vectorizer recognizes user annotations on the code to produce a SIMD version of the scalar code [1].

The evaluation platform is a dual-socket E5-2603v3 processor running at 1.60GHz, with a total of 12 cores, 30MB of L3 cache and 64GB of DDR3. We use PAPI to measure energy, L1D cache miss-rate and total instruction count. The reported energy numbers account for both sockets. The system runs CentOS 6.5 with Nanox 0.7.12a as runtime for the OmpSs codes.

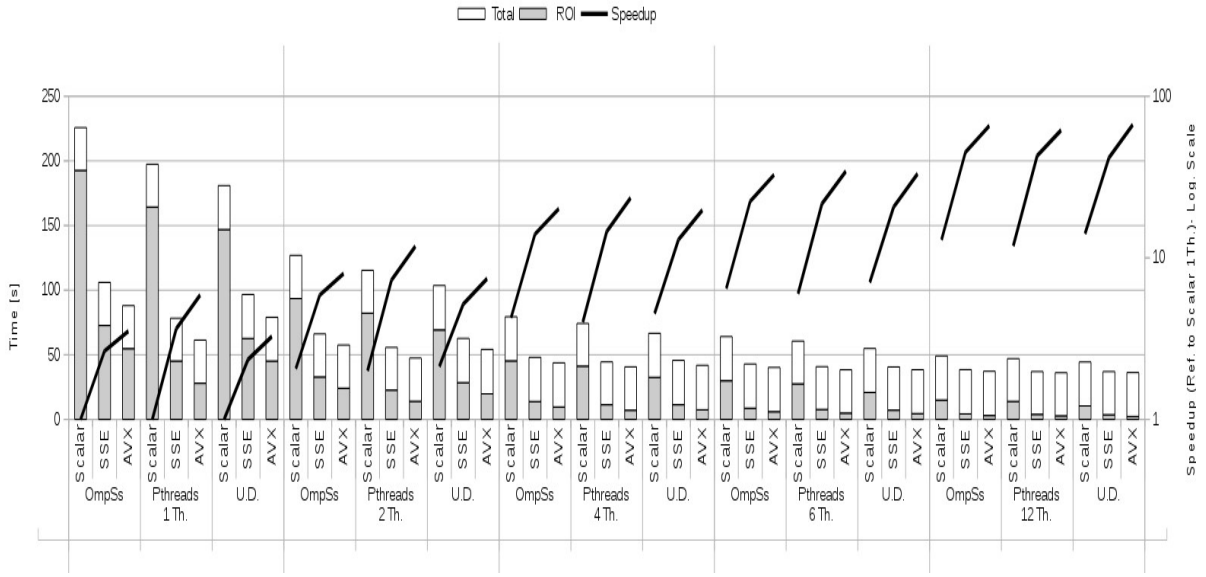


Fig. 1. Blackscholes runtime (Y axis) and speed-up (2nd Y axis

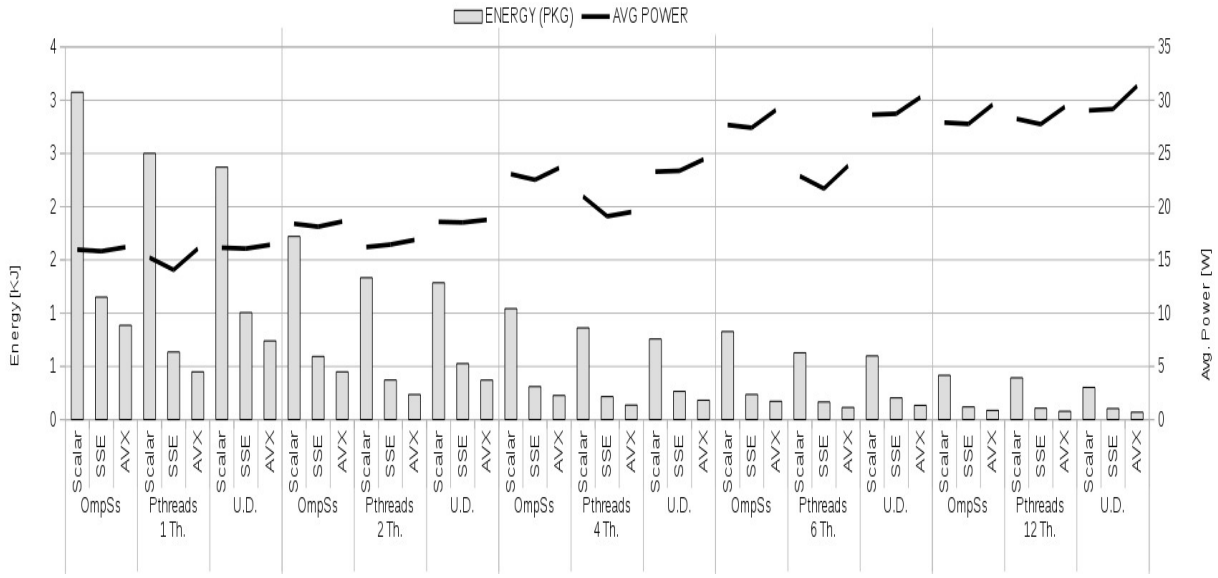


Fig. 2. Blackscholes energy consumption (Y axis) and power (2nd Y axis)

III. EVALUATION

This section shows performance and energy results for only one of the ParVec benchmarks [2] due to of space limitations. Execution times are shown in absolute numbers in order to compare performance between versions. In addition, speed-up is referenced to the scalar sequential
combination of each version to show scalability when varying thread count and vector length.

The blackscholes benchmark shows almost linear scalability with both thread count and vector length (Figure 1). This is mainly because of the high arithmetic intensity of the benchmark (computations per loaded data) and the low L1D cache miss-rate. Instruction count is also reduced linearly with vector length, meaning that we are vectorizing most of the application code.

Pthreads and OmpSs versions have the BlkSchlsEqEuroNoDiv and CNDF functions vectorized manually. In addition, some of the data structures have been aligned. Furthermore, the user-directed version only requires a single directive per function and loop to vectorize all 50 lines of code.

As shown in Fig. 2, power dissipation remains approximately constant in all SIMD versions. Intel platforms share both floating point registers and arithmetic units for scalar and SIMD instructions. While bit-toggling increases power

dissipation due to the extra vector length, the processor spends more time idle, waiting for data dependencies and memory operations, and thus dissipating similar average power independently of running scalar or SIMD code. Finally, it is worth mentioning that Nanos++ has an additional energy overhead when using one and two sockets. As threads spin while searching for work. In the Pthreads version, threads use blocking in the synchronization mechanisms.

IV. CONCLUSIONS

In this abstract, we present an evaluation in terms of performance and energy efficiency of user-directed SIMD implementations using a task-based programming model.

The application shows good performance scalability with vector length. The main reason for that is the reduction of executed instructions and memory accesses with respect to the scalar versions. Power dissipation remains constant when varying vector length. The blackscholes benchmark running with 12 threads can achieve energy improvements up to 35x. User-directed codes achieve similar performance and energy savings to those obtained with hand-vectorized code, while making the code portable between architectures and saving many lines of intrinsics code. As a result, we can confirm that vectorization together with parallelization are key techniques to improve energy efficiency.

REFERENCES

[1] D. L. Caballero de Gea, "PhD Thesis: SIMD@OpenMP: a programming model approach to leverage SIMD features." [Online].Available: http://www.tdx.cat/handle/10803/334171

[2] J. M. Cebrian, M. Jahre, and L. Natvig, "ParVec: Vectorizing the PARSEC Benchmark Suite," Computing, pp. 1077–1100, 2015.

[3] Programming Models, BSC, "The Mercurium C/C++ Source-to-source Compiler Website." [Online]. Available: http://pm.bsc.es/projects/mcxx

[4] A. Duran et al., "OmpSs: A Proposal for Programming Heterogeneous Multi-core Architetcures," Parallel Processing Letters, vol. 21, pp.173–193, Mar. 2011.

Using Graph Partitioning to Accelerate Task-Based Parallel Applications

Isaac Sánchez Barrera, Marc Casas, Miquel Moretó
Eduard Ayguadé, Jesús Labarta, Mateo Valero
Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS), Barcelona, ES
{isaac.sanchez, marc.casas, miquel.moreto, eduard.ayguade, jesus.labarta, mateo.valero}@bsc.es

***Abstract*-*Current high performance computing architectures are composed of large shared memory NUMA nodes, among other components. Such nodes are becoming increasingly complex as they have several NUMA domains with different access latencies depending on the core where the access is issued.

In this work, we propose techniques based on graph partitioning to efficiently mitigate the negative impact of NUMA effects on parallel applications performance, which are able to improve the execution time of OpenMP parallel codes 2.02× times on average when run on architectures with strong NUMA effects.

I. INTRODUCTION

Since the end of Dennard scaling and the subsequent stagnation of the CPU clock frequency, computing infrastructures can only increase their peak performance via augmenting their number of computing units. In the High Performance Computing (HPC) context, this trend has brought an increase in the hardware components count as well as in the heterogeneity among them. As such, shared memory nodes, which are fundamental building blocks of HPC infrastructures, are experimenting an increase in the number of sockets they integrate. Besides the benefits in terms of a unified flat memory address space and large core counts, integrating many sockets into the same node exacerbates its Non-Uniform Memory Access (NUMA) effects, which can become a serious performance bottleneck if they are not properly handled.

To mitigate NUMA effects, techniques consisting in migrating threads, memory pages or both already exist [1]–[3]. These techniques aim to move either computation or data to reduce memory access time. Although these techniques are effective, they do not exploit any kind of application-specific information to predict accesses to remotely allocated data before a particular software component starts displaying this behavior. Oppositely, other approaches transfer the NUMA management responsibility to the programmer exploiting information at the application source code level to carry out NUMA-aware scheduling decisions [4], [5]. However, these approaches require significant code refactoring and programmer effort to be effective.

In this work, we show a novel approach to overcome the limitations of already existing methods for task-based programming models. Our techniques automatically mitigate NUMA effects on multiple NUMA-domain nodes without any kind of specific programmer intervention or application source code change. Our approach leverages runtime system metadata to exploit control and data dependences between the serial parts of parallel workloads and optimally schedule them in the context of a multi-socket NUMA node.

II.PARTITIONING THE TASK DEPENDENCY GRAPH (TDG) TO MITIGATE NUMA EFFECTS

A. *Dependence Easy Placement (DEP)*

Under the Dependence Easy Placement (DEP) policy, tasks are scheduled to the socket where most of their data dependences are allocated. To figure out which specific socket contains a particular block of data, the runtime system keeps a table to map blocks to sockets. The first address of a block is used as its identifier. Tasks that have no inputs, i.e., initialization tasks, are assigned to sockets via a round-robin fashion if most of its output is not allocated yet. In our approach there is a parameter to set the stride of the round-robin approach. When the task to be scheduled is not an initialization task and there is a tie between two or more sockets in terms of the tasks' dependences they contain, the socket is randomly chosen.

B. *Runtime Informed Partitioning (RIP)*

Under the Runtime Informed Partitioning (RIP) policy, task scheduling decisions are based on graph partitioning techniques. The TDG is built at runtime by leveraging information in terms of task dependences. The graph is updated every time new tasks are instantiated and partitioned once the execution goes through a barrier point or a limit in terms of the total number of tasks contained in the graph is reached, which we call the *window size* limit. The graph partitioning algorithm uses the TDG as input, weights its edges depending on the amount of bytes they represent and assigns tasks to

Fig. 1. Speedup results in an SGI Altix UV100 using 2 sockets. For Jacobi using SA, DEP, RIP-DEP and RIP-SP the values are 3.3; for NStream using SA, DEP and RIP-DEP the values are 4.1.

a particular socket taking into account the machine NUMA distances contained in the firmware. Once they are assigned to a socket, they are moved to the corresponding queue. For those tasks that are assigned to a given socket before they are ready to run, they are pushed to the correct queue once their dependences are met, without getting to the temporary queue at all. Once the initial subgraph has been partitioned, we consider three possible options to proceed:

1)RIP with Dependence Easy Placement (RIP-DEP): The RIP-DEP technique consists in propagating the partition obtained from the initial subgraph by taking into account where the tasks' input data resides. As such, if most of the input data of a given task resides in a particular socket, this task is assigned to be run on that socket. This technique is close to the DEP approach, but while DEP applies simple round-robin mechanisms, RIP-DEP partitions the graph.

2)RIP with Socket Propagation (RIP-SP): RIP-SP propagates the partition obtained from the initial subgraph by considering the placement of the predecessors of a particular task and weighting them according to the total amount of data they transfer to the targeted task. As such, the socket where most of the predecessors were executed tends to be chosen by the RIP-SP policy.

3)RIP with Moving Window (RIP-MW): In this case, the graph partitioner is run many times throughout the execution of the program. Once the subgraph contains a particular amount of tasks, the window size, or a barrier point is reached, the partitioning algorithm is run. Once the partitioner finishes its job, the oldest tasks are flushed from the graph and a new subgraph starts getting built, with an intersection between consecutive windows. This intersection is considered to allow the graph partitioner to exploit the already made partitions to generate the new ones, which is an optimization that aims at reducing the overhead.

I. EVALUATION

We evaluate the performance of the proposed mechanisms considering 8 different applications against two schedulers from the Nanos++ runtime:

First-In First-Out (FIFO) task scheduler that is unaware of data location. This is the baseline.

Socket Aware (SA) scheduler, which is driven by annotations at the source code level.

The results for an SGI Altix UV100 machine, with Intel Westmere-EX processors, are shown in Fig. 1. On average, DEP achieves speedups of 1.98× over the FIFO approach, while RIP-DEP, RIP-SP and RIP-MW achieve improvements of 2.02×, 1.28× and 1.09× respectively.

The strong NUMA effects of the Altix system allow the RIP-DEP technique to clearly beat the DEP approach due to the excellent speedups it achieves when dealing with the Gauss-Seidel and the Red-Black applications. The DEP technique is not able to emulate the optimal partition. In contrast, the partition obtained by RIP-DEP is close to the best possible one, which allows the RIP-DEP technique to achieve speedups of 2.01× in Gauss-Seidel and 2.08× in Red-Black, very close to the ones achieved by SA, which is 2.05× faster than FIFO in both cases.

ACKNOWLEDGMENT

This work has been supported by the RoMoL ERC Advanced Grant (GA 321253), by the European HiPEAC Network of Excellence and by the Spanish Ministry of Economy and Competitiveness under contract Computación de Altas Prestaciones VII (TIN2015-65316-P). M. Casas has been partially supported by the Secretary for Universities and Research of the Ministry of Economy and Knowledge of the Government of Catalonia and the Co-fund programme of the Marie Curie Actions of the European Union's 7th FP (contract 2013 BP B 00243). M. Moretó has been partially supported by the Ministry of Economy and Competitiveness under Juan de la Cierva postdoctoral fellowship number JCI-2012-15047.

REFERENCES

[1] M. Dashti, A. Fedorova, J. Funston, F. Gaud, R. Lachaize, B. Lepers, V. Quema, and M. Roth, "Traffic Management: A Holistic Approach to Memory Placement on NUMA Systems," in Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems, 2013, pp. 381–394. doi:10.1145/2451116.2451157

[2] M. Diener, E. H. M. Cruz, P. O. A. Navaux, A. Busse, and H.-U. Heiß, "kMAF: Automatic Kernel-level Management of Thread and Data Affinity," in Proceedings of the 23rd International Conference on Parallel Architectures and Compilation, 2014, pp. 277–288. doi:10.1145/2628071.2628085

[3] M. M. Tikir and J. K. Hollingsworth, "Hardware monitors for dynamic page migration," J. Parallel Distrib. Comput., vol. 68, no. 9, pp. 1186–1200, 2008. doi:10.1016/j.jpdc.2008.05.006

[4] R. Al-Omairy, G. Miranda, H. Ltaief, R. M. Badia, X. Martorell, J. Labarta, and D. Keyes, "Dense Matrix Computations on NUMA Architectures with Distance-Aware Work Stealing," Supercomput. Front. Innov., vol. 2, no. 1, pp. 49–72, Jan. 2015. doi:10.14529/jsfi150103

[5] R. Vidal, M. Casas, M. Moretó, D. Chasapis, R. Ferrer, X. Martorell, E. Ayguadé, J. Labarta, and M. Valero, "Evaluating the impact of OpenMP 4.0 extensions on relevant parallel workloads," in OpenMP: Heterogenous Execution and Data Movements: 11th International Workshop on OpenMP, IWOMP 2015, Aachen, Germany, October 1-2, 2015, Proceedings, vol. 9342, C. Terboven, B. R. de Supinski, P. Reble, B. M. Chapman, and M. S. Müller, Eds. Cham: Springer International Publishing, 2015, pp. 60–72. doi:10.1007/978-3-319-24595-9_5

Improving Scalability of Task-Based Programs

Iulian Brumar, Marc Casas, Miquel Moretó
Barcelona Supercomputing Center
ibrumar@bsc.es, marc.casas@bsc.es, miquel.moreto@bsc.es

Abstract-
In a multi-core era, parallel programming allows further performance improvements, but with an important programmability cost. We envision that the best approach to parallel programming that can exceed the programability, parallelism, power, memory and reliability walls in Computer Architecture is a run-time approach.
Many traditional computer architecture concepts can be revisited and applied at the runtime layer [4][5] in a completely transparent way to the programmer. The goal of this work is taking the computer architecture value prediction and data-prefetching concepts inside a runtime environment like OmpSs.

I.INTRODUCTION

The main objective of this work is researching if *Value Locality* exists in state of the art OmpSs programs and if we can use it in order to obtain better execution times.

Value locality is the property of a static instruction to produce the same output given the same input. If, let us say, a hardware *sum* instruction it is executed twice in a loop, and both times it gets exactly the same inputs, for its second execution we already know that it will generate the same output. However in hardware load instructions, if the input is the same -the address- we are not sure if it will produce the same result. In this case we can only speculate, but even so it has been shown that in many cases, static loads with same input produce the same output [1].

By using this knowledge, we can build a predictor that will skip those instructions that can be well predicted and feed the depending instructions earlier with the predicted output.

In this work we take this concept to a new level for OmpSs tasks and we can distinguish two sub-objectives:

1) Analyze OmpSs benchmarks predictability. We cannot prove that value locality will lead to performance improvement for all possible programs, but we can at least focus on state of the art applications that have been ported to the OmpSs programming model and see how can the value locality concept be extended to our context.

2) Find the ideal speedup using a value locality predictor. This second objective it is a consequence of the previous one. In the cases where value locality exists, what performance improvement can be achieved? We will answer this question using simulation tools.

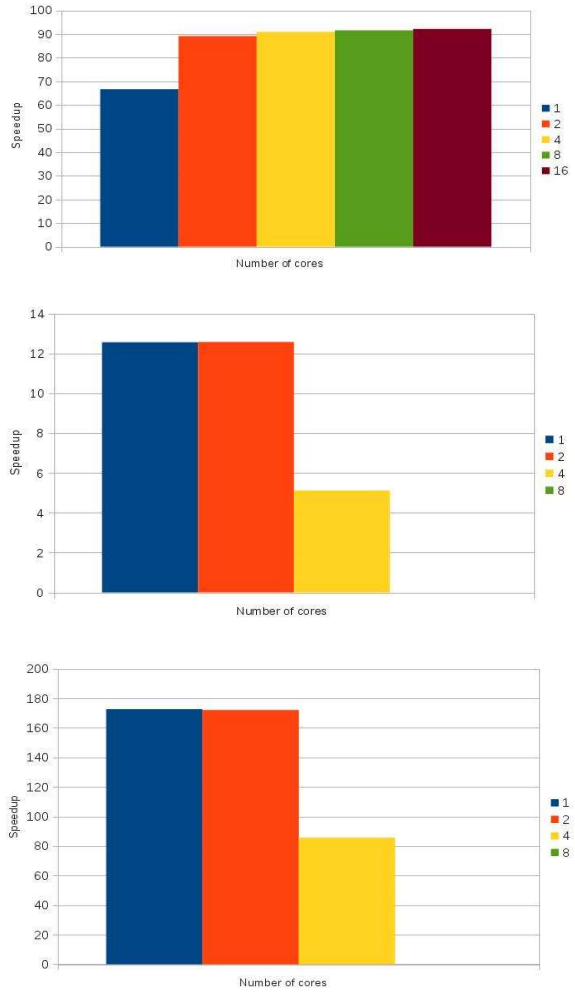
Notice that this is a best case approach in order to discover the limits of the predictability we can have. Also we have to mention that this work was performed with fine grained tasks.

II.RELATED WORK

Since the first moments of computer architecture, it has been seen that the dependencies between instructions were a big wall against *Instruction Level Parallelism* (*ILP)*. A good example of instruction level parallelism is the pipelined processor, which is made of several hardware slots, each one with a specific function. If there are two slots in our processor, namely A and B, an instruction must fulfill both stages in order to complete its execution. We call this an instance of *ILP* because the processor can have two instructions running at the same time. If we hadn't pipelined the processor every instruction would have executed in time time(A) + time(B) but this technique allows us to execute a instruction in time max(time(A), time(B)).

The problem is that the instruction in the first stage (A) might need the result produced by the oldest instruction in (B). In this case the newest instruction will spend one more cycle in stage A and this is a conflict caused by a Read After Write RAW dependency. Even so, back to the 90's, the architects came with a solution [1]. The idea was to continue the execution of the instructions affected by the conflict speculatively. In our example it means that the instruction in stage A can complete the process in this stage speculating the result of the instruction in stage B, and check if the supposition was correct in the next stage.

Now a very good question would be: How can processors predict well the results of hardware instructions? That issue has been explored in the papers of Lipasti [2] and Sazeides [1] which form the motivational base of this work. In the first one the predictor is implemented in two different processor architectures (the out of order PowerPC and the in order Alpha), while the second article gives a more theoretical approach to the issue explaining computational predictors (explained in more detail in [3]) and context based predictors.



III.SOME RESULTS

Figure 1 shows the performance improvement for the Jacobi, Blackscholes and CheckSparseLU benchmarks.

Fig. 1. Performance improvement of Jacobi, Blackscholes and CheckSparseLU.

As we were mentioning in the introduction, those results are obtained using very small task granularities. Additionally, in those three benchmarks, for the same input, the same output is guaranteed to be produced (unlike some programs that don't specify all the data used in their dependencies). For more details on the executions see Table 1. Those speedups are obtained via simulation with TaskSim.

TABLE I

BENCHMARKS CHARACTERISTICS

	Jacobi	Blackscholes	CheckSparseLU
Num. Tasks	64	1024	5000
Bytes/Task	~512	~256	~256
Predicted Tasks	38	899	4800

IV.CONCLUSIONS AND FUTURE WORK

Although huge performance improvements can be achieved using value prediction, we have managed to get these results only at very fine grained levels of parallelism. As part of the same project we have developed a value predictor integrated in the OmpSs runtime together with recovery schemes and data prefetching techniques in case of misspredicition.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Ministry of Science and Innovation under grant TIN2015-65316-P, the HiPEAC Network of Excellence, and by the European Research Council under the European Union's 7th FP, RoMoL ERC Advanaced Grant Agreement number 321253. M. Moreto has been partially supported by the Ministry of Economy and Competitiveness under Juan de la Cierva postdoctoral fellowship number JCI-2012-15047. M. Casas is supported by the Secretary for Universities and Research of the Ministry of Economy and Knowledge of the Government of Catalonia and the Cofund programme of the Marie Curie Actions of the 7th R&D Framework Programme of the European Union (Contract 2013 BP_B 00243).

REFERENCES

[1] SAZEIDES, Yiannakis; SMITH, James E. "The predictability of data values". En *Microarchitecture*, 1997. Proceedings., Thirtieth Annual IEEE/ACM International Symposium on. IEEE, 1997. p. 248-258.

[2] LIPASTI, Mikko H.; WILKERSON, Christopher B.; SHEN, John Paul. "Value locality and load value prediction". ACM SIGOPS *Operating Systems Review*, 1996, vol. 30, no 5, p. 138-147.

[3] LIPASTI, Mikko H.; SHEN, John Paul. *Exceeding the dataflow limit via value prediction*. En *Proceedings of the 29th annual ACM/IEEE international symposium on Microarchitecture. IEEE Computer Society*, 1996. p. 226-237.

[4] Marc Casas, Miquel Moreto, Lluc Alvarez, Emilio Castillo, Dimitrios Chasapis, Timothy Hayes, Luc Jaulmes, Oscar Palomar, Osman Unsal, Adrian Cristal, Eduard Ayguade, Jesus Labarta, and Mateo Valero. *Runtime-aware architectures*. In *Euro-Par*, pages 16–27. 2015.

[5] Mateo Valero, Miquel Moreto, Marc Casas, Eduard Ayguade, and Jesus Labarta. "Runtime-aware architectures: A first approach." *International Journal on Supercomputing Frontiers and Innovations*, 1(1):29–44, June 2014.

Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses

Morena Pappalardo, Miguel Juliá, Mark J. Howard, Jeremy S. Rossman, Martin Michaelis, Mark N. Wass
Centre for Molecular Processing and School of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK.
Email: mp465@kent.ac.uk

Abstract-*This work describes the analysis of 196 Ebolavirus genomes and the identification of specificity determining positions (SDPs) in all nine Ebolavirus proteins that distinguish the non human pathogenic Reston viruses from the four human pathogenic Ebolaviruses. Structural analysis was performed to identify those SDPs that are likely to have a functional effect. This analysis revealed novel functional insights, in particular for Ebolavirus proteins VP40 and VP24. The VP40 SDP P85T interferes with VP40 function by altering octamer formation. The VP40 SDP Q245P affects the structure and hydrophobic core of the protein and consequently protein function. Three VP24 SDPs (T131S, M136L, Q139R) are likely to impair VP24 binding to human karyopherin alpha5 (KPNA5) and therefore inhibition of interferon signaling. Since VP24 is critical for Ebolavirus adaptation to novel hosts, and only a few SDPs distinguish Reston virus VP24 from VP24 of other Ebolaviruses, human pathogenic Reston viruses may emerge.*

I.INTRODUCTION

Four of the five members of the genus Ebolavirus (Ebola viruses, Sudan viruses, Bundibugyo viruses, Taï Forest viruses) cause hemorrhagic fever in humans associated with fatality rates of up to 90% while Reston viruses are non-pathogenic to humans1,2 (see Materials and Methods for the Ebolavirus nomenclature). So far there have been three Reston virus outbreaks in nonhuman primates: 1989-1990 in Reston Virginia, USA, 1992-1993 in Sienna, Italy, and 1996 in a licensed commercial quarantine facility in Texas. All cases were traced back to a single monkey breeding facility in the Philippines.

During these outbreaks five human individuals were tested positive for IgG antibodies directed against Reston virus. Moreover, Reston virus was found in 2008 in domestic pigs in the Philippines. Seroconversion was detected in six human individuals. None of the 11 individuals that were seropositive for Reston virus antibodies reported an Ebola-like disease 3. Our large scale analysis of nearly 200 different Ebolavirus genomes focussed on combining computational methods with detailed structural analysis to identify the genetic causes of the difference in pathogenicity between Reston viruses and the human pathogenic Ebolavirus species. Central to our approach was the identification of Specificity Determining Positions (SDPs), which are positions in the proteome that are conserved within protein subfamilies but differ between them11,12 and thus distinguish between the different functional specificities of proteins from the different Ebolavirus species. SDPs have been demonstrated to be typically associated with functional sites, such as protein-protein interface sites and enzyme active sites12.

The SDPs that we have identified and that distinguish Reston viruses from human pathogenic Ebolaviruses, arguably, contain within them a set of amino acid changes that explain the differences in pathogenicity between Reston viruses and the four human pathogenic species, although a contribution of non-coding RNAs (that may exist but remain to be detected) cannot be excluded6,13. The subsequent structural analysis was performed to identify the SDPs that are most likely to affect Ebolavirus pathogenicity, using an approach that is similar to those used to investigate candidate single nucleotide variants in human genome wide association and sequencing studies by us and others14-17.

II.RESULTS

Specificity Determining Positions (SDP) Analysis. 196 Ebolavirus genomes were obtained from the Virus Pathogen Resource (ViPR18), consisting of 156 Ebola viruses, 7 Bundibugyo viruses, 13 Sudan viruses, 3 Taï Forest viruses, and 17 Reston viruses (online Methods). Phylogenetic analysis of the whole genomes and the individual proteins separated the Ebolavirus species from each other (Supplementary Figure 1).

In accordance with previous studies19-23, we observed high intra-species conservation with greater inter-species variation (Figure 1 and Supplementary Table 1). The surface protein GP exhibited the greatest variation (Figure 1), most likely as a consequence of selective pressure exerted by the host immune response21.

Table 1. SDPs that are likely to alter Reston virus protein structure and function.

Protein	SDP	Interface	Protein Integrity
VP24	T131S	KPNA5 interface	
VP24	M136L	KPNA5 interface	
VP24	Q139R	KPNA5 interface	
VP24	T226A		Loss of Hydrogen bond
VP40	P85T	Octamer interface	
VP40	Q245P		Breaks α helix
VP30	R262A	Dimer interface – loss of Hydrogen bond	
VP35	E269D	Dimer interface	

Using the S3Det algorithm12(Materials and Methods), we identified 189 SDPs that are differentially conserved between Reston viruses and human pathogenic Ebolaviruses (Figure 2, Supplementary Figure 2, Supplementary Tables 2-9). These SDPs represent the most significant changes between the Reston virus and the human pathogenic Ebolaviruses so a subset of these SDPs must explain the difference in pathogenicity. SDPs were present in each of the Ebolavirus proteins representing between 2.4% of residues in sGP to 5.9% of residues in VP30 (Figure 2B). Comparison of the SDPs with previously published mutagenesis studies24 (online Methods) provided no explanation for their functional consequences (Supplementary Table 10).

Structural Analysis. Full-length structures for VP24 and VP40 were available, as well as structures for the globular domains of GP, sGP, NP, VP30, and VP35 (Supplementary Table 11). It was not possible to model the oligeromerization domains of VP30 and VP35 nor the structure of L apart from a short 105 residue segment of the 2239 residue protein, which contained a single SDP. 47 SDPs could be mapped onto Ebolavirus protein structures (or structural models where structures were not available, see online Methods). Most SDPs are located on protein surfaces (Supplementary Figure 3) and are therefore potentially involved in interaction with cellular and viral binding partners and/or immune evasion. Based on our combined computational and structural analysis we find evidence for eight SDPs that are very likely to alter protein structure/function, with six affecting protein-protein interfaces and two that with the potential to influence protein integrity and hence affect stability, flexibility and conformations of the protein (Table 1). Five additional SDPs may alter protein structure/function but the evidence supporting them is weaker (Supplementary Tables 12-18). Two of these weaker SDPs were present in NP (A705R, R105K - all SDPs are referred to using Ebola virus residue numbering and show the human pathogenic Ebolavirus amino acid first and the Reston virus amino acid second). A705R is likely to introduce a salt bridge with E694 and R105K will alter hydrogen bonding (Supplementary Table 12). The three other SDPs with weaker evidence were present in the glycan cap in GP (see below). The eight confident SDPs were present in V24, VP30, VP35, and VP40. The VP40 and VP24 SDPs revealed the most changes that may relate to differences in human pathogenicity (see below).

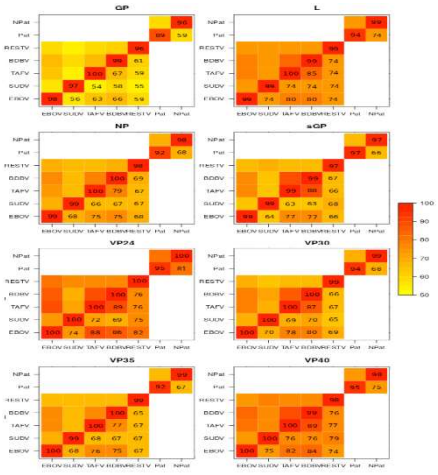
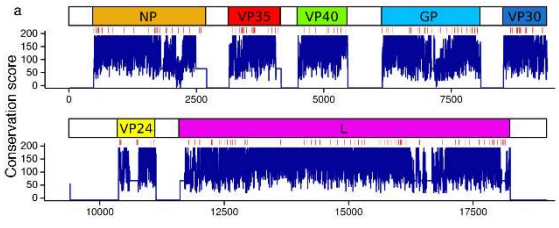


Figure 1: Conservation in Ebolavirus proteins



Protein	length	SDPs	%residues SDPs	SDPs modelled	surface SDPs	probable integrity	probable interface	possible integrity	possible interface
GP	697	30	4.30	8	7				3
sGP	371	9	2.43	8	7				
NP	749	29	3.87	3	3			2	
L	2239	67	2.95	1	1				
VP24	251	9	3.59	8	8	1	3		
VP30	290	17	5.86	6	6		1		
VP35	341	19	5.57	4	4		1		
VP40	331	9	2.72	8	7	1	1		

Figure 2 Ebolavirus SDPs

Multiple SDPs are present in the GP glycan cap. GP is highly glycosylated and mediates Ebolavirus host cell entry. Subunit GP1 binds to the host cell receptor(s). Subunit GP2 is responsible for the fusion of viral and host cell membranes. However, their cellular binding partners remain to be defined1,25-27. Reverse genetics experiments have suggested that GP contributes to human pathogenicity but is insufficient for virulence on its own28. We identified SDPs in both GP1 and GP2 (Supplementary Figure 4 and Supplementary Table 12). Three SDPs (I260L, T269S, S307H) are located in the glycan cap that contacts the host cell membrane (Supplementary Figure 4B-C). These changes (particularly S307H at the top of the glycan cap) alter the electrostatic surface of GP (Supplementary Figure 4D) and may therefore alter GP interactions with cellular proteins, however given the glycosylation of GP, it is unlikely that these residues would physically contact the host cell membrane and none of them are near glycosylation sites. So it is not clear what role they may have. GP binding to the endosomal membrane protein NPC1 is necessary for membrane fusion25. However, residues important for NPC1 binding (identified by mutagenesis studies in25) were conserved in all analyzed Ebolaviruses and the SDPs were not located close to them (Supplementary Figure 5). Thus differences in NPC1 binding do not account for differences in Ebolavirus human pathogenicity. This finding is in concert with very recent data indicating that NPC1 is essential for Ebolavirus replication as NPC1-deficient mice were insusceptible to Ebolavirus infection27. It was not possible to predict the consequences of SDPs in sGP and ssGP (Fig. S23), as there is a lack of functional information available for these proteins3,4. A 17 amino acid peptide derived from Ebola virus or Sudan virus GP exerted immunosuppressive effects on human CD4+ T cells and CD8+ T cells while the respective Reston virus peptide did not29. We identified one SDP in the peptide, which represents the single amino acid change (I604L) previously observed between Reston virus and Ebola virus29, demonstrating that this difference is conserved between Reston viruses and all human pathogenic Ebolaviruses.

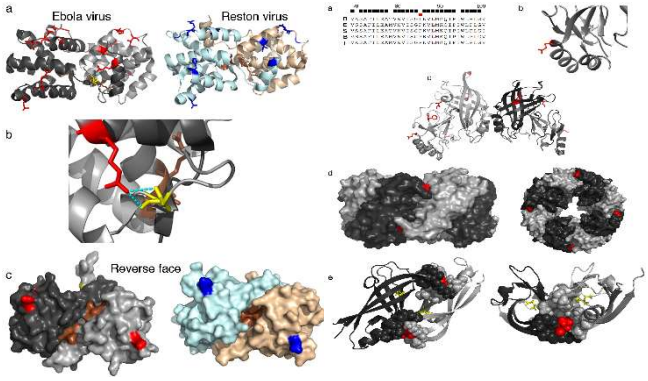


Figure 3-4: SDPs in protein VP30 and in protein VP40

Changes in the VP30 dimer may affect pathogenicity. Analysis of the VP30 SDPs provided novel mechanistic insights into the structural differences previously observed between Reston virus and Ebola virus VP3010 and that may contribute to the differences observed in human pathogenicity between Reston virus and Ebola virus. VP30 is an essential transcriptional co-factor that forms dimers via its C-terminal domain and hexamers via an oligomerization domain (residues 94-112)30.

The VP30 hexamers activate transcription while the dimers do not, and the balance of hexamers and dimers has been suggested to control the balance between transcription and replication31. Crystallization studies have shown that Ebola virus and Reston virus dimers are rotated relative to each other10. We observed two SDPs (T150I, R262A) in the dimer interface that can at least partially explain the structural differences between Ebola virus and Reston virus VP30 dimers. Ebola virus R262 is part of the dimer interface and forms a hydrogen bond with the backbone of residue 141 in the other subunit, whereas Reston A262 does not and is not part of the dimer interface (Figure 3). The removal of the two hydrogen bonds (in the symmetrical dimer) is likely to lead to the different Reston and Ebola virus dimer structures. mCSM predicts this change to be destabilizing with a ΔΔG -0.969 Kcal/mol. T

he Reston virus conformation also buries functional residues A179 and K180 potentially affecting protein function10 (Figure 2). Moreover, our findings show that the Ebola virus confirmation is

conserved in all human-pathogenic Ebolaviruses suggesting that it is relevant for human pathogenicity.

VP35 SDP present in dimer interface. VP35 is a multifunctional protein that antagonizes interferon signaling by binding double stranded RNA (dsRNA). Structural data are available for both the Ebola virus and Reston virus VP35 monomer and an asymmetric dsRNA bound dimer 9,32-34. These structures are highly conserved, however functional studies have demonstrated that Reston virus VP35 is more stable, has a reduced affinity for dsRNA, and exerts weaker effects on interferon signaling32. The increased stability is proposed to be due to a linker between the two subdomains having a short alpha helix in the Reston virus structure32. Our analysis shows that the sequence of this linker region is completely conserved in all of the genomes, however an SDP is located close to the linker (A290V). One SDP (E269D) is present in the dimer interface and the shorter aspartate side chain in Reston virus VP35 results in increased distances with the atoms that this aspartate forms hydrogen bonds with: R312, R322, and W324 (Ebola virus numbering; Supplementary Table 13). mCSM predicts this change to be slightly destabilizing to the complex (ΔΔG -0.11Kcal/mol). This has the potential to alter the stability of the dimer and thus the ability of VP35 to prevent interferon signaling. It has recently been demonstrated that a VP35 peptide binds NP and modulates NP oligomerization and RNA binding to NP35. There are two SDPs (S26T, E48D) in this region. S26T is located on the periphery of the interface. E48D lies outside the solved structure but is within the region required for binding to NP. Both SDPs represent minor changes that maintain the chemical properties of the side chains. Thus, there is no evidence suggesting substantial differences in the binding of this peptide to NP.

VP40 SDPs may alter oligomeric structure. VP40 exists in three known oligomeric forms36. Dimeric VP40 is responsible for VP40 trafficking to the cellular membrane. Hexameric VP40 is essential for budding and forms a filamentous matrix structure. Octameric VP40 regulates viral transcription by binding RNA. Two SDPs (P85T and Q245P) can affect VP40 structure. P85T occurs at the VP40 octamer interface site (Figure 4) in the middle of a run of 14 residues that are completely conserved in all Ebolaviruses (Figure 4B). In the Ebola virus structure, it is located in an S-G-P-K beta-turn, where the proline at position 85 (P85) confers backbone rigidity. The change to threonine (T) at this residue in Reston viruses introduces backbone flexibility and also provides a side chain with a hydrogen bond donor, potentially affecting octamer structure and/or formation. mCSM predicted this change to have a destabilizing effect (ΔΔG -0.626Kcal/mol). The Q245P SDP introduces a proline residue into an alpha helix (Figure 4B), which most likely breaks and shortens helix five, resulting in the destabilization of helices five and six and a change in the hydrophobic core. Interestingly mCSM predicted this change to have little effect on the stability of the protein (predicted ΔΔG 0.059Kcal/mol). Thus, P85T and Q245P may affect VP40 function and human pathogenicity.

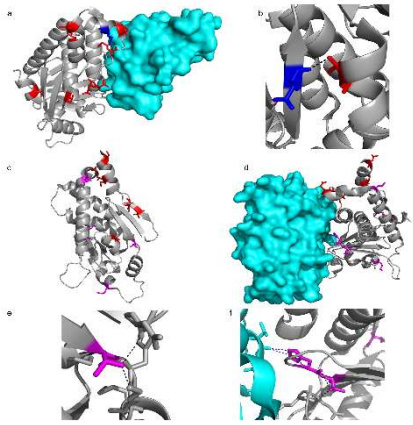


Figure 5: SDPs in protein VP24

VP24 SDPs affect KPNA5 binding. VP24 is involved in the formation of the viral nucleocapsid and the regulation of virus replication1,19,37-39. VP24 also interferes with interferon signaling through binding of the karyopherins α1 (KPNA1), α5, (KPNA5), and α6 (KPNA6) and subsequent inhibition of nuclear accumulation of phosphorylated STAT1 and through direct interaction with STAT124,40-42. Eight VP24 SDPs are in regions with available structural information (Supplementary Tables 17-18). Seven of these are present on the same face of VP24 (Figure 5A) suggesting that they affect VP24 interaction with viral and/or host cell binding partners. The SDPs T131S, M136L, and Q139R are present in the KPNA5 binding site (Figure 5). M136 and Q139 are part of multi-residue mutations in Ebola virus VP24 that removed KPNA5 interactions (Supplementary Table 17)24 and are adjacent to K142 (Figure 5A), mutants of which have shown reduced interferon antagonism43. Therefore, M136L and Q139R can exert significant effects on VP24-KPNA5 binding. Additionally, T226A results in the loss of a hydrogen bond between T226 and D48 in Reston virus VP24 (Figure 5B), with the potential to alter structural integrity and influence protein function. Analysis

using mCSM predits the T226A change to be destabilizing with a ΔΔG -0.935 Kcal/mol. mCSM predicted seven of the eight analysed SDPs to be destabilizing (Supplementary Table 2). VP24-mediated inhibition of interferon signaling may be critical for species-specific pathogenicity24,38,40-42. In this context, VP24 was a critical determinant of pathogenicity in studies in which Ebola viruses were adapted to mice and guinea pigs that are normally insusceptible to Ebola virus disease5,38,44-46. The adaptation-associated VP24 mutations in rodents are located in the KPNA5 binding site with some of them being very close to the VP24 SDPs T131S, M136L, and Q139R that we determined to be in the KPNA5 binding site (Figure 5C-D, Supplementary Table 19). Additionally some of the mutations are similar to the SDPs in that they would remove hydrogen bonds within VP24 (e.g. T187I, T50I, Figure 5E-F, & Supplementary Table 19) or alter hydrogen bonding with KPNA5 (H186Y, Figure 5F & Supplementary Table 19). Thus there is strong evidence suggesting that the VP24 SDPs have a role in rendering the Reston virus non-pathogenic in humans.

III.DISCUSSION

In this study, we have combined the computational identification of residues that distinguish Reston viruses from human pathogenic Ebolavirus species with protein structural analysis to identify determinants of Ebolavirus pathogenicity. The results from this first comprehensive comparison of all available genomic information on Reston viruses and human pathogenic Ebolaviruses detected SDPs in all proteins but only few of them may be responsible for the lack of Reston virus human pathogenicity. Our analysis mapped 47 of the 189 SDPs onto protein structure, so additional SDPs may be relevant but the structural data needed to reliably identify them is missing. Although it is difficult to conclude the extent to which each individual SDP contributes to the differences in human pathogenicity between Reston viruses and the other Ebolaviruses, we can identify certain SDPs that have a particularly high likelihood to be involved. SDPs present in the oligomer interfaces of VP30, VP35, and VP40 may affect viral protein function. VP24 SDPs may interfere with VP24-KPNA5 binding and affect viral inhibition of the host cell interferon response. These findings suggest that changes in protein-protein interactions represent a central cause for the variations in human pathogenicity observed in Ebolaviruses. VP24 and VP40 in particular contain multiple SDPs that are likely to contribute to differences in human pathogenicity. Where possible the SDPs have been considered collectively, such as for VP24, where most of the SDPs are present on a single face of the protein (Figure 5A) and three of them are present in the interface with KPNA5. Beyond this it is difficult to interpret how any combination of SDPs might be responsible for the differences in human pathogenicity. Our data also demonstrate that relevant changes explaining differences in virulence between closely related viruses can be identified by computational analysis of protein sequence and structure. Such computational studies are particularly important for the investigation of Risk Group 4 pathogens like Ebolaviruses whose investigation is limited by the availability of appropriate containment laboratories. The role of VP24 appears to be central given the large number of SDPs we identify as likely to affect function, particularly KPNA5 binding. This is also highlighted by the similarity between these SDPs and the mutations that occur in adaptation experiments in mice and guinea pigs6,33,39-41. Consequently, the mutation of a few VP24 SDPs could result in a human pathogenic Reston virus. Given that Reston viruses circulate in domestic pigs, can be spread by asymptomatically infected pigs, and can be transmitted from pigs to humans (possibly by air)2,47,48, there is a concern that (a potentially airborne) human pathogenic Reston viruses may emerge and pose a significant health risk to humans. Notably, asymptomatic Ebolavirus infections have also been described in dogs2 and Ebola virus shedding was found in an asymptomatic woman49. Thus, there may be further unanticipated routes by which Reston viruses may spread in domestic animals and/or humans enabling them to adapt and cause disease in humans. In summary our combined computational and structural analysis of a large set of Ebolavirus genomes has identified amino acid changes that are likely to have a crucial role in altering Ebolavirus pathogenicity. In particular the differences in VP24 together with the observation that Ebolavirus adaptation to originally non-susceptible rodents results in rodent pathogenic viruses6,33,39-41 suggest that a few mutations could lead to a human pathogenic Reston virus.

ACKNOWLEDGMENT

We would like to thank Antonio Rausell for advise on the use of the S3det algorithm

REFERENCES

1. Feldmann, H. & Geisbert, T. W. Ebola haemorrhagic fever. *Lancet* **377,** 849–862 (2011).

2. Weingartl, H. M., Nfon, C. & Kobinger, G. Review of Ebola virus infections in domestic animals. *Dev Biol (Basel)* **135,** 211–218(2013).

3. Miranda, M. E. G. & Miranda, N. L. J. Reston ebolavirus in humans and animals in the Philippines: a review. *J. Infect. Dis.* **204** Suppl3, S757–60 (2011).

4. Mehedi, M. *et al.* A new Ebola virus nonstructural glycoprotein expressed through RNA editing. *J. Virol.* **85,** 5406–5414 (2011).

5. La Vega, de M.-A., Wong, G., Kobinger, G. P. & Qiu, X. The multiple roles of sGP in Ebola pathogenesis. *Viral Immunol.* **28,** 3–9 (2015).

6. Basler, C. F. Portrait of a killer: genome of the 2014 EBOV outbreak strain. *Cell Host Microbe* **16,** 419–421 (2014).

7. Hoenen, T. *et al.* Soluble Glycoprotein Is Not Required for Ebola Virus Virulence in Guinea Pigs. *J. Infect. Dis.* jiv111, doi: 10.1093/infdis/jiv111 (2015).

8. Zhang, A. P. P. *et al.* The ebola virus interferon antagonist VP24 directly binds STAT1 and has a novel, pyramidal fold. *PLoS Pathog.* **8,** e1002550 (2012).

9. Bale, S. *et al.* Ebolavirus VP35 coats the backbone of double-stranded RNA for interferon antagonism. *J. Virol.* **87,** 10385–10388 (2013).

10. Clifton, M. C. *et al.* Structure of the Reston ebolavirus VP30 C-terminal domain. *Acta Crystallogr F Struct Biol Commun* **70,** 457–460 (2014).

11. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nat Struct Biol* **2,** 171–178 (1995).

12. Rausell, A., Juan, D., Pazos, F. & Valencia, A. Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. USA* **107,** 1995–2000 (2010).

13. Teng, Y. *et al.* Systematic Genome-wide Screening and Prediction of microRNAs in EBOV During the 2014 Ebolavirus Outbreak.*Sci Rep* **5,** 9912 (2015).

14. Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *NatGenet* **43,** 1131–1138 (2011).

15. Chambers, J. C. *et al.* Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* **42,** 373–375 (2010).

16. Chambers, J. C. *et al.* The South Asian genome. *PLos One* **9,** e102645 (2014).

17. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas andcarcinomas. *Nat Genet* **45,** 136–144 (2013). Scientific **Reports** | 6:23743 | DOI: 10.1038/srep23743 10

18. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40,** D593–8 (2012).

19. Morikawa, S., Saijo, M. & Kurane, I. Current knowledge on lower virulence of Reston Ebola virus (in French: Connaissancesactuelles sur la moindre virulence du virus Ebola Reston). *Comparative Immunology, Microbiology and Infectious Diseases* **30,** 391–398 (2007).

20. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345,** 1369–1372 (2014).

21. Liu, S.-Q., Deng, C.-L., Yuan, Z.-M., Rayner, S. & Zhang, B. Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa. *Infect Genet Evol* **32,** 51–59 (2015).

22. Vogel, G. Infectious Diseases. A reassuring snapshot of Ebola. *Science* **347,** 1407–1407 (2015).

23. Hoenen, T. *et al.* Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* **348,** 117–119 (2015).

24. Xu, W., Edwards, M. R., Borek, D. M., Feagins, A. R. & Mittal, A. Ebola Virus VP24 Targets a Unique NLS Binding Site on Karyopherin Alpha 5 to Selectively Compete with Nuclear Import of Phosphorylated STAT1. *Cell Host Microbe* **13,** 187–200 (2014).

25. Miller, E. H. *et al.* Ebola virus entry requires the host-programmed recognition of an intracellular receptor. *EMBO J.* **31,** 1947–1960 (2012).

26. Dahlmann, F. *et al.* Analysis of Ebola Virus Entry Into Macrophages. *J. Infect. Dis.* jiv140, doi: 10.1093/infdis/jiv140 (2015).

27. Herbert, A. S. *et al.* Niemann-pick c1 is essential for ebolavirus replication and pathogenesis *in vivo*. *MBio* **6,** e00565–15 (2015).

28. Groseth, A. *et al.* The Ebola virus glycoprotein contributes to but is not sufficient for virulence *in vivo*. *PLos Pathog.* **8,** e1002847 (2012).

29. Yaddanapudi, K. *et al.* Implication of a retrovirus-like glycoprotein peptide in the immunopathogenesis of Ebola and Marburg viruses. *FASEB J.* **20,** 2519–2530 (2006).

30. Hartlieb, B., Modrof, J., Muhlberger, E., Klenk, H.-D. & Becker, S. Oligomerization of Ebola virus VP30 is essential for viral transcription and can be inhibited by a synthetic peptide. *J. Biol. Chem.* **278,** 41830–41836 (2003).

31. Hartlieb, B., Muziol, T., Weissenhorn, W. & Becker, S. Crystal structure of the C-terminal domain of Ebola virus VP30 reveals a role in transcription and nucleocapsid association. *Proc. Natl. Acad. Sci. USA* **104,** 624–629 (2007).

32. Leung, D. W. *et al.* Structural and functional characterization of Reston Ebola virus VP35 interferon inhibitory domain. *J. Mol. Biol.* **399,** 347–357 (2010).

33. Leung, D. W. *et al.* Structure of the Ebola VP35 interferon inhibitory domain. *Proc. Natl. Acad. Sci. USA* **106,** 411–416 (2009).

34. Kimberlin, C. R. *et al.* Ebolavirus VP35 uses a bimodal strategy to bind dsRNA for innate immune suppression. *Proc. Natl. Acad. Sci. USA* **107,** 314–319 (2010).

35. Leung, D. W. *et al.* An Intrinsically Disordered Peptide from Ebola Virus VP35 Controls Viral RNA Synthesis by Modulating Nucleoprotein-RNA Interactions. *Cell Rep* doi: 10.1016/j.celrep.2015.03.034 (2015).

36. Bornholdt, Z. A. *et al.* Structural rearrangement of ebola virus VP40 begets multiple functions in the virus life cycle. *Cell* **154,** 63–774 (2013).

37. Mateo, M. *et al.* Knockdown of Ebola virus VP24 impairs viral nucleocapsid assembly and prevents virus replication. *J. Infect. Dis.* **204** Suppl 3, S892–6 (2011).

38. Mateo, M. *et al.* VP24 is a molecular determinant of Ebola virus virulence in guinea pigs. *J. Infect. Dis.* **204** Suppl 3, S1011–20 (2011).

39. Watt, A. *et al.* A novel life cycle modeling system for Ebola virus shows a genome length-dependent role of VP24 in virus infectivity. *J. Virol.* **88,** 10511–10524 (2014).

40. Reid, S. P. *et al.* Ebola virus VP24 binds karyopherin alpha1 and blocks STAT1 nuclear accumulation. *J. Virol.* **80,** 5156–5167 (2006).

41. Reid, S. P., Valmas, C., Martinez, O., Sanchez, F. M. & Basler, C. F. Ebola virus VP24 proteins inhibit the interaction of NPI-1 subfamily karyopherin alpha proteins with activated STAT1. *J. Virol.* **81,** 13469–13477 (2007).

42. Zhang, A. P. P. *et al.* The ebolavirus VP24 interferon antagonist: know your enemy. *Virulence* **3,** 440–445 (2012).

43. Ilinykh, P. A. *et al.* Different temporal effects of Ebola virus VP35 and VP24 proteins on the global gene expression in human dendritic cells. *J. Virol.* JVI. 00924–15, doi: 10.1128/JVI.00924-15 (2015).

44. Volchkov, V. E., Chepurnov, A. A., Volchkova, V. A., Ternovoj, V. A. & Klenk, H. D. Molecular characterization of guinea pigadapted variants of Ebola virus. *Virology* **277,** 147–155 (2000).

45. Ebihara, H. *et al.* Molecular determinants of Ebola virus virulence in mice. *PLos Pathog.* **2,** e73 (2006).

46. Dowall, S. D. *et al.* Elucidating variations in the nucleotide sequence of Ebola virus associated with increasing pathogenicity. *Genome Biol.* **15,** 540 (2014).

47. Barrette, R. W. *et al.* Discovery of swine as a host for the Reston ebolavirus. *Science* **325,** 204–206 (2009).

48. Marsh, G. A. *et al.* Ebola Reston virus infection of pigs: clinical significance and transmission potential. *J. Infect. Dis.* **204** Suppl 3,S804–9 (2011).

49. Akerlund, E., Prescott, J. & Tampellini, L. Shedding of Ebola Virus in an Asymptomatic Pregnant Woman. *N. Engl. J. Med.* **372,**2467–2469 (2015).

50. Kuhn, J. H. *et al.* Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virusabbreviations. *Archives of Virology* **155,** 2083–2103 (2010).51. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).

52. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.***7,** 539 (2011).
53. Mistry, J. *et al.* Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions *Nucleic Acids Res.* **41,**e121 (2013).

54. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191–8 (2014).
55. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).

56. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,**1312–1313 (2014).

57. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).

58. Rose, P. W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *NucleicAcids Res.* **43,** D345–56 (2015).

59. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction andanalysis. *Nat Protoc* **10,** 845–858 (2015).

60. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39,** D411–9 (2011).

61. Smith, N. *et al.* DelPhi web server v2: incorporating atomic-style geometrical figures into the computational protocol. *Bioinformatics***28,** 1655–1657 (2012).

62. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* **25,** 1605–1612(2004).

63. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30, 335–342 (2014)

Enrichment of Virtual Screening results using induced-fit techniques

Jelisa Iglesias[a, b], Suwipa Sae-oon[a], Gabriela Hernandez[a, d], Jorge Estrada[a],
Ricard Gavalda[b] and Victor Guallar[a, d].
[a] Barcelona Supercomputing Center
[b] Universitat Politècnica de Catalunya
[c] Université Lumière Lyon 2 – EM-DMKM
[d] Institucio Catalana de Recerca i Estudis Avançats (ICREA)
jelisa.iglesias@bsc.es

Abstract: *This project aims to improve the results of virtual screening and docking techniques used for drug design, using induced-fit techniques and a consensus scoring approach.*

Keywords: Virtual Screening; Drug design; Consensus scoring function.

I. INTRODUCTION.

The drug discovery process aims at discovering new chemical compounds (ligands) that bind to a given target (usually a protein) causing a disease. The ligand is expected to modify the target activity to cure or alleviate the effects of the disease. This process usually requires from 10 to 17 years [1] approximately, costs billions of dollars and has a low success rate.

Virtual screening (VS) procedures have been developed due to the high cost associated with experimentally testing (millions of) chemical compounds. VS is a broad term that includes all the computational methods developed to aid in the drug discovery process complementing the experimental ones. This term includes managing the compounds data, filter them according to their physic-chemical properties and the docking and hit identification processes.

The VS can be divided in two categories depending on the approach used: ligand-based VS and structure-based VS [2]. The docking methods belong to the second category and are used to screen ligands that may bind the target (binders) by ranking them with a prediction of their binding affinity. This process involves two main steps: a) identify the binding pose and b) estimate the binding affinity. In order to achieve a) in a short time they assume the protein to be rigid, which introduces error since the proteins are flexible entities. To do b) they use scoring functions, which use fast and approximate algorithms often designed to discriminate between binders and non-binders.

The overall performance of docking methods is compromised by these two aspects: lack of flexibility and accuracy of the scoring functions.

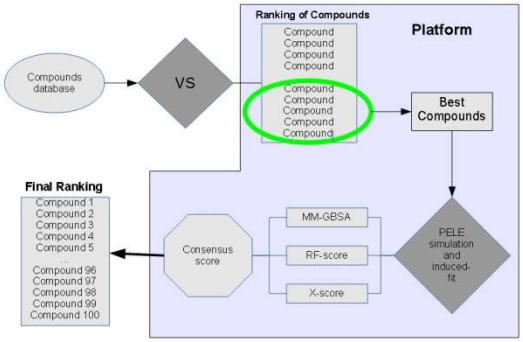


Fig. 1 Scheme of the platform workflow

II. PROJECT DESCRIPTION

This PhD project aims at developing a platform to improve the results from the current VS procedures. This platform will use the output of a VS procedure and will perform induce-fit techniques, to allow the protein to adapt its structure to the ligand, and then will compute a consensus scoring function to score the new structure, giving a better estimation of their binding affinities.

The platform will be able to do all the files conversions necessary, launch the simulations needed and compute all the scoring functions and descriptors used by the consensus scoring function. Protein preparation, for example, is a crucial step where automatic procedures have to be developed with care. The scoring function will be trained and tested on dataset of complexes formed by a protein and a ligand with known binding affinity, and the overall performance of this new platform will be tested on a common dataset for VS and a real case of VS in collaboration with AstraZeneca.

Hypothesis: By adding induced fit techniques to the best (top) thousands virtual screening results, together with the use of multiple scoring functions and machine learning techniques, we will enrich the number of true positive results (binders) in the VS process.

III. WHAT WE HAVE SO FAR.

The present work has been focused in optimizing the protocol to be applied to the top VS poses. For this, we have focused first in improving the affinity prediction by preparing a test set, selecting scoring functions, and initial steps towards adding flexibility with induced fit techniques.

A. Datasets

The training set and the test set for affinity prediction have been compiled and manually prepared using the Protein Preparation Wizard from Schrödinger [3]. The training set consists in a subset of the pdbBind refined core dataset compiled in Ref [4] and is formed by 191 structures from 64 different families. The test set is a subset of the pdbBind refined core set in Ref [5], formed by 64 structures from 64 different families. Both subsets are composed by protein-ligand complexes with known dissociation or inhibition constants (K_d and K_I, respectively)

All the structures have undergone a careful preparation process with the hydrogen added according with the protonation states at experimental pH, the missing loops and chains reconstructed using Prime [6], [7].

Furthermore both test set have been minimized using the Protein Energy Landscape Exploration (PELE) software with and without waters.

B. Scoring function selection

The criteria used to select which scoring functions to use were: their performance [4], [8], their availability (free or commercial), and the type of scoring function. The most commonly used scoring functions were selected, both commercial and free, in a way that there are at least two scoring functions for each type of scoring function according to their method to estimate the binding affinity [9].

The scoring functions being tried are MM-GBSA from Prime [6], [7], AutodockVina [10], PELE interaction Energy, Glide SP [11] and XP [12] scoring functions, XScore [13], DSX [14], RF-Score [15], NN 2.0 score [16] and Rdock [17]. They have been computed for all the structures in both the test and training sets minimized with and without water molecules. Results shown in Table 1

C. Consensus scoring function

The consensus scoring function is under development trying different machine-learning methods, such as random forest and neural networks, using 9 different scoring functions already developed and widely used.

D. Induced fit techniques

The induced-fit techniques aim to reproduce the binding pose of a ligand in a given protein taking into account the conformational changes induced in the protein by the ligand.

To obtain this kind of structure PELE simulations are being performed with some parts of the protein the backbone slightly constrained but leaving the side chains free to move, the objective is to maintain the overall structure of the protein while allowing for the small structural changes induced by the ligand when binding.

FUTURE WORK.

There is still a lot of work to do: we plan to extend the training and test set, to include in this platform a VS method able to deal with thousands of different ligands, to add this platform to the PELE GUI, etc.

TABLE 3; Pearson correlation coefficient for all the scoring functions in the training set

Scoring Function	Correlation	Scoring Function	Correlation
PELE	0.409	Xscore	0.612
MM-GBSA	0.494	DSX	0.533
Autodock Vina	0.532	RF-Score	0.677
Glide SP	0.399	NN 2.0	0.741
Glide XP	0.357	Rdock	0.214

REFERENCES

[1] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs.," *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–83, Aug. 2004.

[2] L. A. and G. C. Di, "Virtual screening strategies in drug discovery: A critical review," *Curr. Med. Chem.*, vol. 20, no. 23, pp. 2839–2860, 2013.

[3] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments.," *J. Comput. Aided. Mol. Des.*, vol. 27, no. 3, pp. 221–34, Mar. 2013.

[4] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang, "Comparative assessment of scoring functions on a diverse test set.," *J. Chem. Inf. Model.*, vol. 49, no. 4, pp. 1079–93, Apr. 2009.

[5] Y. Li, L. Han, Z. Liu, and R. Wang, "Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results.," *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1717–36, Jun. 2014.

[6] M. P. Jacobson, R. a. Friesner, Z. Xiang, and B. Honig, "On the role of the crystal environment in determining protein side-chain conformations," *J. Mol. Biol.*, vol. 320, no. 02, pp. 597–608, Jul. 2002.

[7] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner, "A hierarchical approach to all-atom protein loop prediction.," *Proteins*, vol. 55, no. 2, pp. 351–67, May 2004.

[8] Y. Li, Z. Liu, J. Li, L. Han, J. Liu, Z. Zhao, and R. Wang, "Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set.," *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1700–16, Jun. 2014.

[9] J. Liu and R. Wang, "Classification of current scoring functions," *J. Chem. Inf. Model.*, vol. 55, no. 3, pp. 475–482, 2015.

[10] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–61, Jan. 2010.

[11] R. a. Friesner, J. L. Banks, R. B. Murphy, T. a. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M.

Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin, “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy,” *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, 2004.

[12] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz, “Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes.,” *J. Med. Chem.*, vol. 49, no. 21, pp. 6177–96, Oct. 2006.

[13] R. Wang, L. Lai, and S. Wang, “Further development and validation of empirical scoring functions for structure-based binding affinity prediction,” *J Comput Aided Mol Des*, vol. 16, no. 1, pp. 11–26, 2002.

[14] G. Neudert and G. Klebe, “DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes.,” *J. Chem. Inf. Model.*, vol. 51, no. 10, pp. 2731–45, Oct. 2011.

[15] P. J. Ballester and J. B. O. Mitchell, “A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking,” *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, 2010.

[16] J. D. Durrant and J. A. McCammon, “NNScore 2.0: A neural-network receptor-ligand scoring function,” *J. Chem. Inf. Model.*, vol. 51, no. 11, pp. 2897–2903, 2011.

[17] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, and S. D. Morley, “rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids,” *PLoS Comput. Biol.*, vol. 10, no. 4, pp. 1–7, 2014

On the way to real time protein-ligand sampling

Daniel Lecina-Casas[†], Ryoji Takahashi[†], Victor Guallar[†‡]
[†]Joint BSC-IRB Research Program in Computational Biology, Barcelona Supercomputing Center,
[‡]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
daniel.lecina@bsc.es

Abstract *- Protein-ligand binding free energy is one of the keystones of drug design, and developing a fast method to calculate it would have great impact in personalized medicine. However, it is a daunting task for computational methods, since the conformational space is rugged, having a lot of metastable states that hinder the exploration. Using PELE and an adaptive sampling scheme, one can quickly get thermodynamic properties by traversing the conformational space on a simulation time scale (24h). We show the performance on a new benchmark of a series of different families of proteins and ligands with a large range of binding free energy differences (about 8 kcal/mol).*

I. INTRODUCTION

Protein-ligand binding free energy is one of the keystones of drug design, since it is related to the binding affinity, and developing a fast method to calculate it would have great impact in personalized medicine. Experimental approaches are a standard way to measure thermodynamic properties, and computational methods complement well with them, since they are cheaper, easier to prepare, and give the experimenter an atomistic detail of the process of interest.

Two of the most popular computational methods are Molecular Dynamics (MD) and Monte Carlo (MC). In the first one, we explore the conformational space by numerically integrating Newton equations of motion, whereas in the second by making proposals that are accepted or rejected according to the Metropolis criterion. The protein energy landscape exploration software (PELE)[1][2], our own MC sampling technique, uses small random moves mixed with protein structure prediction algorithms to make proposals, and has been proven to accurately describe protein-ligand interactions and thermodynamics[3][4].

Typically, good estimates of the binding free energy rely on a good sampling of the relevant states. However, both MD and MC often face the problem of getting trapped in metastable minima: in biomolecule simulations there are a lot of competing interactions, which yield a rugged energy landscape that hinders the conformational exploration, oversampling some metastable states whereas undersampling others. To overcome this problem, we developed an adaptive sampling scheme that enhances the traversal of the conformational space.

In this work we will first show the results of applying standard long PELE trajectories on a set of a different proteins and ligands with a large range of binding free energy differences (about 8 kcal/mol). In the second place, we will see the sampling improvements when adaptive sampling is used. This new methodology opens a way in adding PELE and adaptive sampling in pharmaceutical drug design.

II. METHODS

1)System setup

We initialize our system with the closest protein – ligand distance being greater than 15Å, ensuring a sufficient solvent exploration. Instead of exploring the whole protein surface, the ligand is constrained to a sphere of radius ~20Å, that contains the main pocket. We use OPLS2005 with derived QM/MM charges for the ligand and implicit solvent OBC. All explicit waters are removed.

2)Simulations

Unbiased simulations are run with PELE. It combines a stochastic approach, usually called perturbation, with protein prediction algorithms, usually called relaxation. In the perturbation, the ligand is randomly moved, followed by protein backbone displacements using Cartesian coordinate anisotropic network model proposals. PELE will soon incorporate an internal coordinates normal mode analysis, which will allow smoother backbone moves. The resulting structure is relaxed by means of a side chain prediction and a global minimization with constraints on alpha carbons and the center of mass of the ligand. At the end of each iteration, the step is accepted or rejected according to the Metropolis criterion. Simulations are performed for 24h on 128-512 processors, depending on the system.

3)Analysis

In order to analyze the results, we build Markov State Models (MSM) with the ligand center of mass using EMMA[5]. MSM[6] are a methodology based on discrete master equations that allows us to calculate ΔG and obtain a coarse-grain description of the simulations, facilitating the understanding of the molecular mechanisms. We compute the binding free energy, ΔG, using:

$$\Delta G = -k_b T \ln(V_b/V_o) + \Delta W,$$

where ΔW stands for the difference of population in bulk and binding site, V_b is the binding volume and $V_o = 1661Å^3$.

4)Adaptive sampling

Adaptive sampling[7] is an iterative procedure that aims to balance the sampling of the different metastable states. We perform rounds of short simulations (e.g. 15 minutes), and clusterize all visited conformations according to RMSD. This allows us to redistribute simulations taking into account the exploration time and interest of each cluster (exploration-exploitation problem).

III.RESULTS

These are preliminary results since it is work in progress, so they may vary in the final publication.

First, we show the performance on a benchmark of a series of different families of protein and ligands with a large range of binding free energy differences (about 8 kcal/mol). As we can see in Figure 1, we can predict the tendency of binding free energies.

System (PDB id)	ΔG_{comp} (kcal/mol)	ΔG_{exp} (kcal/mol)
3ptb	-7.7	-6.7
1ecv	-9.3	-6.6
1vfn	-8.6	-7.74
1q5k	-9.6	-10.1
1b80	-12.4	-14.7

Table 1: Experimental and computational results for a benchmark of different ligands. Errors were not computed, but are estimated to be in the order of 1kcal/mol.

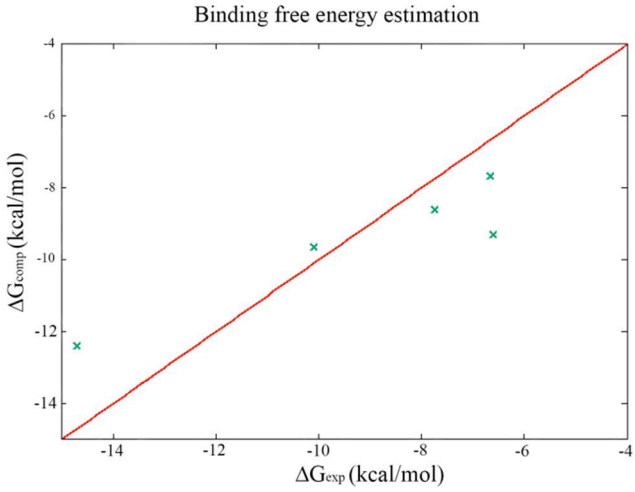


Figure 1: Ligand binding free energy computation with experimental results. Red line serves as guide to compare with exact results.

Estrogen receptors are interesting systems to study the efficiency of adaptive sampling, since we need to reproduce conformational changes in the protein in order to simulate the ligand binding. We will serve of the non-adaptive scheme as reference, and in both scenarios we will run 400 simulations. As we can see in Figure 2 panel A, in the non-adaptive scheme we sample only two binding events, the first being produced at step ~1300, while in the adaptive sampling, the first one is being produced at around step 160. The region of SASA<0.2 is much better sampled in the latter at a fraction of the cost.

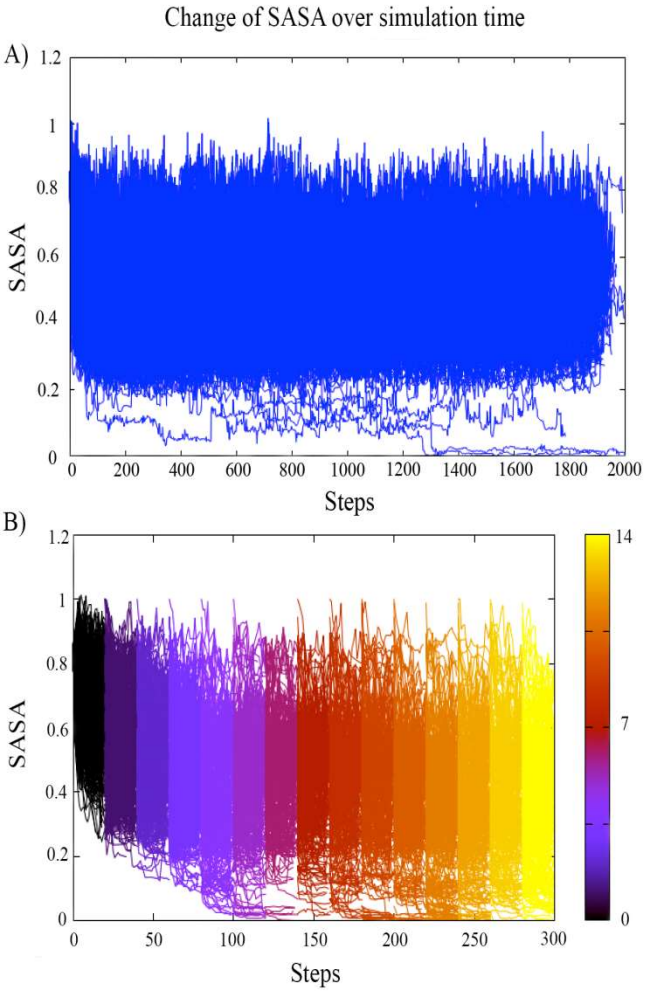


Figure 2: A) Evolution of SASA for the ligand, where 0 means totally buried and 1 means completely solvent exposed, for 400 simulations in 24h of simulation time (around 2000 steps). B) Evolution of SASA for the ligand over different 20-step epochs (color code at the right). A total amount of 15 different epochs where produced. Notice the different scales in the x-axis.

IV.CONCLUSIONS

A benchmark of different systems has been presented: from more rigid to more flexible proteins, and from small to medium sized ligands;

where PELE is able to reproduce a correct ΔG estimation using long trajectories.

Also, we showed preliminary results of applying adaptive sampling to a difficult case, hormone receptors. In this case, the sampling of the binding site improves considerably, showing binding events in an order of magnitude less of computational time.

We still need to test the adaptive sampling scheme in the ΔG benchmark, even though the first results (not shown) seem in good agreement.

The new procedure, mixing PELE sampling with adaptive sampling, seems a promising alternative to be used in a near future in pharmaceutical drug design.

ACKNOWLEDGMENT

D.L-C thanks SEV-2011-00067, awarded by the Spanish government.

REFERENCES

[1] K. W. Borrelli, A.Vitalis, R. Alcantara & V. Guallar , "PELE: Protein Energy Landscape Exploration. A novel Monte Carlo Based Technique" J. Chem. Theory Comput. 1, 1304-1311 (2005).

[2] A. Maddadkar-Sobhani & V. Guallar. "PELE web server: atomistic study of biom." Nucl. Acids Res., 41, W322-W328 (2013).

[3] R. Takahashi, V. A. Gil & V. Guallar. "Monte Carlo free ligand diffusion with Markov State Model analysis and absolute binding free energy calculations." J. Chem. Theory Comput. 10, 282-288 (2014)

[4] K. Edman, A. Hosseini, M. K Bjursell, A. Aagaard, L. Wissler, A. Gunnarsson, T. Kaminski, C. Köhler, S. Bäckström, T. J Jensen, A. Cavallin, U. Karlsson, E. Nilsson, D. Lecina, R. Takahashi, C. Grebner, S. Geschwindner, M. Lepistö, A. C Hogner, V. Guallar. "Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints" Structure, 23, 2280-2290 (2015)

[5] M. Senne, B. Trendelkamp-Schroer, A. S.J.S. Mey, C. Schütte, and F. Noé. "EMMA: A Software Package for Markov Model Building and Analysis." J. Chem. Theory and Comput. 8, 2223−2238 (2012).

[6] G. R. Bowman, V. S. Pande, F. Noé, Eds. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Advances in Experimental Medicine and Biology, Springer: Heidelberg, Germany, 2014; Vol. 797.

[7] N. Hinrichs, V. Pande. Calculation of the Distribution of Eigenvalues and Eigenvectors in Markovian State Models for Molecular Dynamics. J. Chem. Phys. 126, 244101 (2007)

PMut2: a web-based tool for predicting pathological mutations on proteins

V. López-Ferrando[1], X. de la Cruz[2], M. Orozco[1, 3], J.L. Gelpí[1, 3]
1. Joint IRB-BSC-CRG programme for Computational Biology. Life Sciences Dept. Barcelona Supercomputing Center.
2. Vall d'Hebron Research Institute.
3. Dept. Biochemistry and Molecular Biology. Universitat de Barcelona.

Abstract- *Amino acid substitutions in proteins can result in an altered phenotype which might lead to a disease. PMut2 is a method that can predict whether a mutation has a pathological effect on the protein function. It uses current machine learning algorithms based on protein sequence derived information. The accuracy of PMut2 is as high as 82%, with a Matthews correlation coefficient of 0,62. PMut2 predictions can be obtained through a modern website which also allows to apply the same machine learning methodology that is used to train PMut2 to custom training sets, allowing users to build their own tailor-made predictors.*

I. INTRODUCTION

Assessing the impact of amino acid mutations in human health is an important challenge in biomedical research. As sequencing technologies are more available, and more individual genomes become accessible, the number of identified variants has dramatically increased. PMut, released back in 2005 [1], has been one of the popular predictors in this field. PMut was a neural-network-based classifier using sequence data to provide a pathology score for point mutations in proteins.

PMut2 is a new, revised, and much more powerful version of the predictor. It introduces the use of state-of-the-art machine learning algorithms and an updated training set based on SwissVar [2]. It achieves an accuracy of 82% and a Matthews correlation coefficient (MCC) of 0.62. PMut2 includes a fully featured training and validation engine that can be optimized to generate predictors adapted to user specific training sets. The engine is implemented in Python using MongoDB engine for data management. It has been adapted to run at the HPC level to cover large scale annotation projects.

II. METHODS

The process of training PMut2 is based on common machine learning methods. First of all, a training set of mutations annotated as either neutral or pathological must be established. For each of these variants a set of numerical features are computed to best describe them. Finally, a model is selected and trained using the training set and the computed features.

Training set

PMut2 is trained using the manually curated variation database SwissVar (as of December 2015), which contains ~28,000 disease and ~38,000 neutral mutations on ~12,500 proteins.

Features computation

Over 150 numerical features are computed for each mutation. They account for 1) physical property differences between wild type and mutated amino acids, 2) protein interactome information and 3) amino acid conservation. The conservation features are derived from local searches over UniRef100 and UniRef90 clusters [3] using PSI-BLAST [4] and multiple sequence alignments using Kalign2 [5].

Model selection

In order to choose a model for the predictor, different classifiers were tested with different parameter configurations. Random Forest is chosen as it presents the best predictive power and good computing speed.

From the 150 features computed, only 12 of them were selected via an iterative algorithm to be part of the final predictor, as seen in Figure .

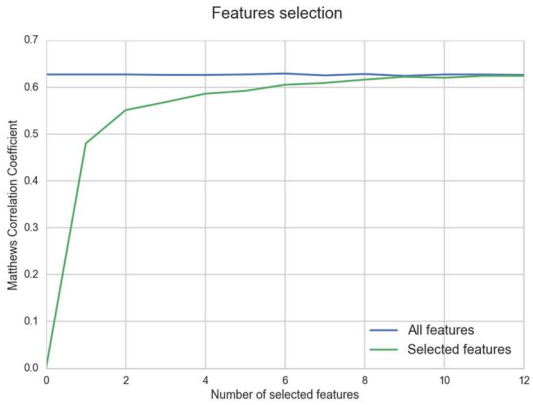


Figure 1. Iterative features selection. With 12 features the model has the same precision as with all 150 features.

III. RESULTS

Predictor evaluation

The predictor was evaluated using 10-fold cross validation with 50% protein identity exclusion in the different folds. This method allows us to estimate the predictor performance when faced with a protein different to the ones in its training set. The predictor obtained a Matthews correlation coefficient of 0,620 ± 0,02 and an Area under the ROC curve of 0,808 ± 0,01.

Predictor comparison

The predictor performance is compared to other popular predictors in the field using new mutations that were added to SwissVar after the model

training (January – March 2016). Table 1 holds the results of this comparison.

The meta-predictor PROVEAN presents the best performance, with an MCC of 0,479; PMut2 follows with an MCC of 0,469. This comparison also outlined how Classic PMut performs poorly.

Table 1. Comparison of different predictors by predicting 573 mutations added to SwissVar in January – March 2016.

Predictor	Coverage	Accuracy	Sensitivity	Specificity	AUC	MCC
PROVEAN	98,3	0,762	0,814	0,819	0,740	**0,479**
SIFT	97,6	0,835	0,835	0,774	0,689	0,395
Polyphen	98,6	0,763	0,884	0,777	0,714	0,463
Condel	95,3	0,758	0,843	0,794	0,724	0,462
Classic PMut	52,9	0,551	0,507	0,791	0,585	**0,154**
PMut2	100	0,743	0,746	0,839	0,742	**0,469**

Web application

A web application allows use of PMut2 predictor and provides other useful functionality (Figure 2). A comprehensive repository of precomputed predictions yields instant access to all possible mutations in all human proteins in UniProt (a total of 803,743,460 variants on 109,106 proteins). Fig. 4 shows one of such proteins in the repository, Fig. 5 is the results page of an analysis of a list of mutations and Fig. 6 is the page of a custom predictor that has been trained using a given training set.

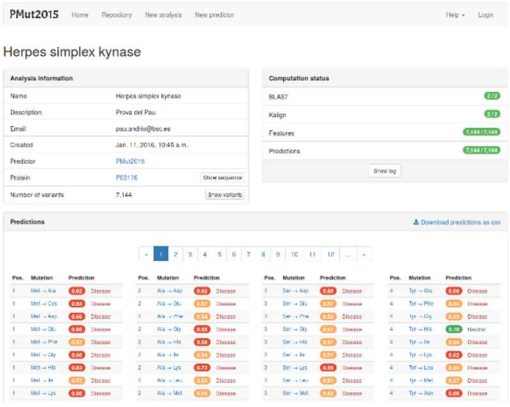


Figure 2: Web application Home page summarizing the three use cases:1) Browser or search the repository, 2) Get predictions of given mutations and 3) Train a tailor-made predictor.

are predicted and their predictions can be observed in the 3D structure

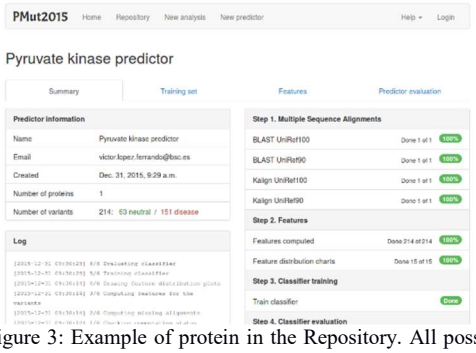


Figure 3: Example of protein in the Repository. All possible mutations

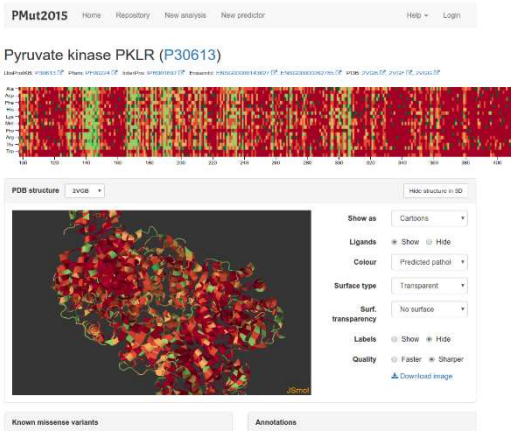


Figure 4: Example of analysis result. After all computations complete,each variant has a corresponding predicted pathology score.

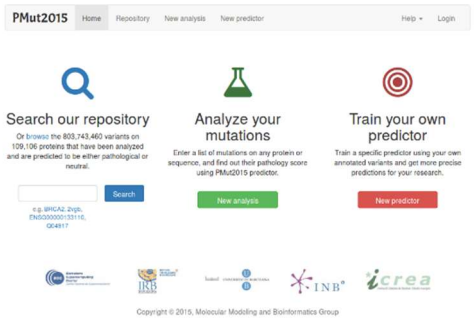


Figure 5: Example of custom predictor for protein Pyruvate Kinase. A training set consisting of a list of variants annotated as disease or neutral was submitted.

IV. CONCLUSION

Using state-of-the-art machine learning algorithms we were able to train a predictor that matches the predictive power of the most popular predictors in the field. Furthermore, the methods used have been automated and offered to the research community via a user-friendly website.

REFERENCES

[1] Ferrer-Costa, et al. «PMUT: A Web-Based Tool for the Annotation of Pathological Mutations on Proteins». Bioinformatics 21, num. 14 (15 July 2005): 3176-78.

[2] Mottaz, et al. «Easy Retrieval of Single Amino-Acid Polymorphisms and Phenotype Information Using SwissVar». Bioinformatics 26, num. 6 (15 March 2010): 851-52.

[3] Suzek, et al. «UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters». Bioinformatics 23, num. 10 (15 May 2007): 1282-88.

[4] Altschul, et al. «Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs». Nucleic Acids Research 25, num. 17 (9 January 1997): 3389-3402.

[5] Timo Lassmann, Oliver Frings, and Erik L. L. Sonnhammer, "Kalign2: High-Performance Multiple Alignment of Protein and Nucleotide Sequences Allowing External Features," Nucleic Acids Research 37, no. 3 (January 2, 2009): 858–65, doi:10.1093/nar/gkn100

Per-Task Energy Metering and Accounting in the Multicore Era

Qixiao Liu, Miquel Moreto, Jaume Abell, Francisco J. Cazorla and Mateo Valero
Barcelona Supercomputing Center, Universitat Politecnica de Catalunya
{Qixiao.liu,Miquel.moreto,Jaume.abella,Francisco.cazorla,Mateo.valero}@bsc.es

Abstract*-Energy has become arguably the most expensive resource in a computing system. As multi-core processors are the preferred processing platform across different computing domains, measuring the energy usage draws vast attention.*

In this thesis, for the first time, we formalize the need for per-task energy measurement in multicore by establishing a two-fold concept: per-task energy metering and sensible energy accounting. The former, for a task running in a multi-core system, provides estimates on the actual energy consumption corresponding to its resource usage. The latter provides estimates on the energy the task would have consumed running in isolation with a given fraction of the shared resources. We have shown how these two concepts can be applied to the main components of a computing system: the processor and the memory system.

I. INTRODUCTION

Chip multi-core processors (CMPs) are the preferred processing platform across different domains such as data centers, real-time systems and mobile devices. In all those domains, energy is arguably the most expensive resource in a computing system, in particular with fastest growth. Therefore, measuring the energy usage draws vast attention. Current studies mostly focus on obtaining finer-granularity energy measurement, such as measuring power in smaller time intervals, distributing energy to hardware components or software components. Such studies focus on scenarios where system energy is measured, and under the assumption that only one program is running in the system. So far, there is no hardware-level mechanism proposed to distribute the system energy to multiple running programs in a resource sharing multi-core system in an exact way.

To elaborate on the need of accurate per-task energy measurement, Figure 1 shows the average power dissipation when executing all the SPEC CPU2006 benchmarks on a POWER7-based system. As shown, different tasks incur different average power dissipation, with the maximum variation being 16%, between *453.povray* and *410.bwaves*. Hence, if a povray-like and a bwaves-like program execute undisturbed in a computing system for a period of time, they will incur significantly different energy consumptions. However, the same amount of energy would be attributed to each, which sum up to the total energy consumption of the system.

In this work, we formalize the need for per-task energy measurement in multicore by establishing a two-fold concept: ***Per-Task Energy Metering (PTEM)*** and ***Sensible Energy Accounting (SEA).***

Given a workload composed by *n* tasks T_1, T_2, ..., T_n running in a processor with n cores, we define per-task

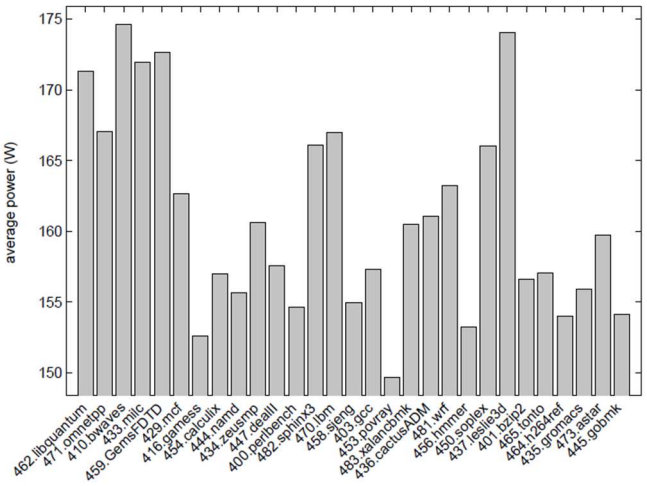


Figure 1: Power consumption of SPECCPU 2006 benchmarks on a PS701 system with an IBM POWER7 processor

energy metering and accounting as follows. ***PTEM*** consists in tracking the energy that a given task, T_i, consumes during a given period of time. ***SEA*** consists in deriving for a given task T_i, the energy that T_i would have consumed if it had run in isolation with a *fraction* of the hardware resources. Both, per-task energy metering and accounting, are complex to derive with the increasing number of hardware shared resources that can serve requests from different tasks concurrently using different resources and/or with different latencies[i].

It is our position that accurately measuring the energy consumed by each task in a computer, instead of considering only the whole energy consumed by the computer, will have plenty of important applications.

Billing. When a customer requests the same computing power to run the same task using the same input, the same energy cost should be accounted[ii].

Energy/Performance optimization. Metering and accounting the energy consumed per task would allow finding the processor setup (e.g. number of cores) and software setup (e.g. mapping) that leads to the lowest system energy consumption.

Selection of appropriate co-runners. Task interaction in hardware shared resources may negatively affect tasks hurting performance and increasing energy requirements. Metering per-task energy can help the OS/runtime scheduler to decide which task to run and when, reducing systems' energy profile.

PROPOSALS

We break energy into its main three components. Dynamic energy corresponds to the energy spent to perform those useful activities that circuits are intended to do. Maintenance energy corresponds to the energy consumed due to useless activity not triggered by the program(s) being run. Leakage corresponds to the energy wasted due to imperfections of the technology used to implement the circuit.

We focus on the case of a shared cache as it is a good representative of the challenges to address when carrying out per-task energy metering and accounting. We assume a multicore architecture where each core has private data and instruction first level caches plus a shared on-chip Last-Level Cache (LLC).

PTEM

Dynamic energy: In order to split dynamic energy across running tasks we consider the number and type of accesses that each task performs to each resource. This requires per-task access counters and the energy consumption per access to be provided by the chip vendor.

Maintenance energy: Maintenance energy is consumed by useless activities in idle resources. It must be attributed to tasks depending on the amount of space they occupy if those resources are stateful (such as caches).

Leakage energy: Splitting leakage energy among running tasks is also proportional to both time and space.

SEA

Dynamic energy: To account the dynamic energy for a task we estimate the number of accesses that each task would perform when it runs with a fraction of LLC space alone.

Maintenance energy: Accounting the maintenance energy based on the execution time and activities estimation when a task runs with a fraction of resources alone.

Leakage energy: Accounting leakage energy relies on the execution time estimation.

Note that we devise hardware mechanisms for both PTEM and SEA in the processor and the memory system[iiiivv]. Through which, the runtime estimation is done for all tasks running in workloads in a multicore processor system.

Experiments

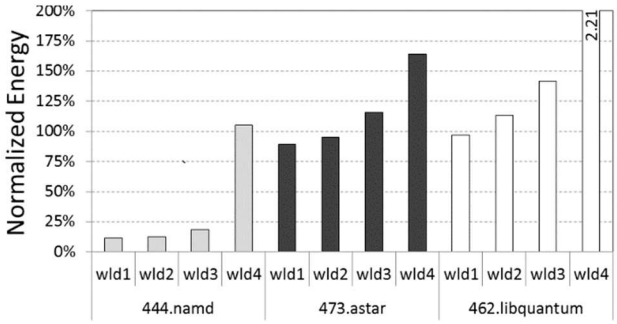


Figure 2: Energy usage of namd, astar, and libquantum in different workloads w.r.t their energy usage when executed in isolation with a fair share of resources.

Our proposed mechanisms have achieved high estimation accuracy in all components, with affordable overheads.

We show in Figure 2, the actual energy metered for three benchmarks when they run in different workloads. The energy consumed by a program will largely depend on its co-runners, due to the fact that they have been sharing resources in the multicore processor system.

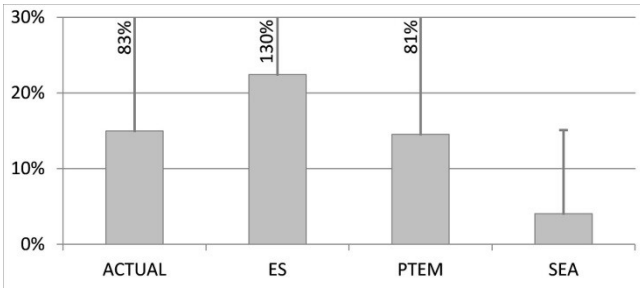


Figure 3: The deviation of mispredicted energy account to tasks running in 8-task workloads under 4-core SMT setup and 16-way LLC

In Figure 3, we show that our proposal significantly improves the estimation of the energy a task would have consumed with a given fraction of resources (SMT core and LLC), compare to other model that measures the energy

.

EXPERIMENTS

In this thesis, we have advanced the field of quantifying per-task energy cost in the multicore systems by proposing a two-fold concept: PTEM and SEA. PTEM derive an estimation of the actual energy a task consumes in a real workload based on resource utilization, and SEA gives an estimation of energy a task would have consumed when it has been given a fraction of resources. These works have already been published in international conferences and journals.

REFERENCES

[1] Q. Liu, et.al. Per-task Energy Accounting in Computing Systems. In IEEE Computer Architecture Letters, 2014.

[1] V. Jimenez, et.al. *Energy-Aware Accounting and Billing in Large-Scale Computing Facilities*. In IEEE Micro, 2011.

[1] Q. Liu, et.al. *Hardware Support for Accurate Per-Task Energy Metering in Multicore Systems.* In ACM Transactions on Architecture and Code Optimization (TACO), 2013.

[1] Q. Liu, et.al. *DReAM: Per-Task DRAM Energy Metering in Multicore Systems*. Euro-Par 2014 Parallel Processing international conference, Porto, August, 2014.

[1] Q. Liu, et.al. *Sensible Energy Accounting with Abstract Metering for Multicore Systems.* In ACM Transactions on Architecture and Code Optimization (TACO), 2015.

Task Dependences Management Hardware Acceleration for Task-based Dataflow Programming models

Xubin Tan, Carlos Álvarez-Martínez, Daniel Jiménez-González, Eduard Ayguadé, Mateo Valero
Universitat Politècnica de Catalunya, Barcelona Supercomputing Center, Barcelona, Spain
{xubin.tan, eduard, maeto.valero}@bsc.es, {djimenez, calvarez}@ac.upc.edu

Abstract- *Task-based programming models have gained a lot of attention for being able to explore high parallelism over multicore and manycore, while hiding the difficulties of parallel programming. For applications with moderate size tasks, performance gains are assured by using these programming models. While for more parallelism by using smaller and more tasks, the performance degrades as a result of runtime overheads. To speed up the runtime, we present a hardware accelerator, Picos Hardware to accelerate task dependence management and scheduling. In this work, we show the performance of the first Picos Hardware prototype realized in a Zynq 7000 All-Programmable SoC by using real benchmarks. Results show that our hardware support greatly outperforms the software-only implementation currentlyavailable in the runtime system for fine-grained tasks.*

I. INTRODUCTION

Parallel computing offers the possibility to scale up the performance over the number of processors. At the same time, it exposes significant challenges for programmers to adapt themselves from sequential to parallel programming. Task-based programming models are quickly developed to target these challenges. For example, Google's MapReduce, Intel's TBB, Open MP 4.0, StarSs and OmpSs programming model [1]. In OmpSs, programmers can gain performance by simply annotating tasks in the source code with directives (input, output, inout) to hint their data dependences. And the remaining actions as task creation, dependence management/dependence graph management and task scheduling are managed by the Nanos++ Runtime system (RTS).

OmpSs is able to expose high parallelism from applications of varied domains with a number of moderate tasks, with both regular and irregular dependence patterns and is fairly easy to use. However, for fine-grained tasks, the runtime overheads (especially dependence management and task scheduling) are too high to scale.

To speedup the runtime and extend the usage of OmpSs to fine-grained tasks, we present a hardware accelerator, Picos Hardware. It accepts general information from master threads as task identification, number of dependences, memory addresses and directions of dependences, and schedules ready tasks to worker threads. In this work, we show a brief description of Picos and its hardware costs, and discuss some challenges during the development, and finally results of the first Picos prototype [2] realized in a Zynq 7000 All-Programmable SoC [3] by using real benchmarks.

II. METHODOLOGY AND CHALLENGES

A. Experimental Setup

First, OmpSs applications are executed in sequential and parallel up to 24 threads in a shared memory machine. It has 2 sockets, each socket is a Xeon E5-2630L with 6 cores with dynamic frequency up to 2.0GHz.

Second, execution time of the same applications of Picos full system are obtained in Zedboard (Zynq 7000 SoC) by using traces. The traces are obtained through instrumenting the sequential execution of OmpSs applications. It includes two main parts: the first part includes task creation/execution time in cycles required to simulate the task creation and task execution processes in ARM processor; the second part includes task and dependence information necessary for dependence management and task scheduling.

Finally, the speedups of OmpSs applications shown in this work are obtained against the sequential execution time.

B. Picos Full System

Fig. 1. shows the organization of the current embedded system integrated with the first Picos prototype. The Programmable Logic part uses a 80MHZ global clock, and a 64bits AXI Timer synchronized with the same clock as the global timer. The ARM processor runs a bare-metal Operating System, and the workers inside are simulating threads.

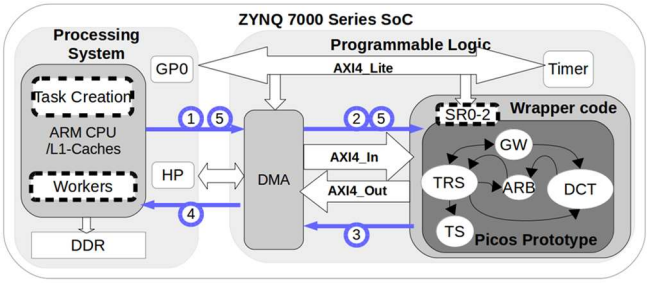


Fig. 1. The Picos Full System employs a close-loop process. Each task is created and sent to Picos for dependence analysis (1, 2). Each ready task is retrieved from Picos to the ARM core for execution in the workers (3, 4). Finally each finished task is sent back to notify Picos (5) to carry on the process until the last task.

Each message between Picos and ARM core carries one task at once and the communication latency for sending or retrieving each task via DMA takes around 200 to 300 cycles for each message.

C. Picos Prototype

Picos prototype has five functional units: Gateway (GW), Task Reserve Station (TRS), Dependence Chain Tracker (DCT), Task Scheduling (TS) and Arbiter (ARB).

GW reads new/finish task information from workers to Picos prototype.

TRS is the major task management unit. It stores in-flight tasks, tracks the readiness of new tasks and manages the deletion of finished tasks.

DCT is the major dependence management unit. It performs address matching of new dependence against the addresses of those arrived earlier, to track data dependences, and also save and control all its live versions.

TS schedules ready tasks notified by TRS to idle workers.

ARB manages communications between TRS and DCT.

The first prototype uses about 5.8% Look-Up Tables, 1.2% Flip-Flops and 17% BRAMs in XC7Z020 [3].

D. Challenges

We encounter several big challenges during the development of the first Picos prototype. Firstly, the balance between speed and hardware cost of TRS and DCT. Since each task can have multiple dependences, this stresses the dependence management unit multiple times more than the task dependence unit. Secondly, the system stalls if new dependences cannot be processed due to the memory capacity and entry conflicts. Thirdly, the communication latency between Picos prototype and the ARM processor.

III. RESULT EVALUATION

We show the speedup (y-axis) of Cholesky, SparseLu (four different block sizes) [4] obtained by Picos Full-system, Perfect Simulator and Nanos++ RTS, with up to 24 threads in Fig. 2.. Results of Perfect Simulator shows the critical-path roofline speedup.

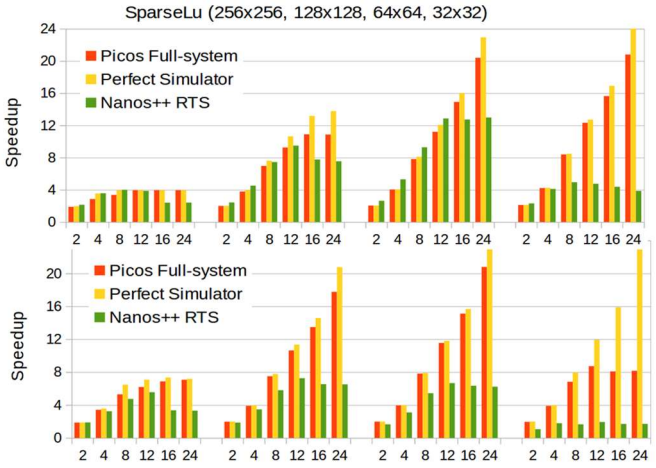


Fig. 2. Speedup of OmpSs applications with 2 to 24 threads

As can be seen, firstly, for each benchmark with a fixed block size, Nanos++ RTS scales up to 12 workers maximum while the Picos prototype continues to scale to 24 workers. For example, for SparseLu and Cholesky in with block size 64, the Picos prototype achieves over 20x with 24 workers while Nanos++ RTS achieves 13x and 7x respectively.

Secondly, for both benchmarks, Nanos++ RTS starts to degrade rapidly after some point s while the Picos prototype keeps on advancing or remains stable as the block size decreases. For Cholesky, although the performance of both Picos prototype and Nanos++ RTS degrade for block size 32, the latter one has a much worse degradation. The reason for the Picos prototype degradation here is that it only uses one TRS and DCT, which is unable to unfold such a high parallelism from Cholesky with block size 32. However, with more module instances Picos Hardware should be able to obtain higher speedup and fill this gap [5].

IV. CONCLUSIONS

In this paper we show a brief description of Picos, as a RTS hardware support to speedup the task and dependence management for task-based dataflow programming models like Open MP 4.0 and OmpSs. The presented implementation has been in a Zynq 7000 All-programmable SoC

Platform. Results of real benchmarks show that the prototype greatly outperforms the existing OmpSs software-only implementation (Nanos++) and as the task granularity decreases, the prototype continues to scale after Nanos++ RTS starts to degrade. More importantly, with a larger design with multiple task and dependence management units upcoming, Picos Hardware could be able to exploit a larger magnitude of parallelism in the applications with very fine granularity, that software alternatives cannot achieve.

ACKNOWLEDGMENT

This work is supported by the Programa Severo Ochoa (SEV-2015-0493) through the TIN2015-65316-P project, the contracts 2014-SGR-1051 and 2014-SGR-1272, and the European Research Council RoMoL Grant Agreement number 321253. We also thank the Xilinx University Program. This work has one accepted paper [2].

REFERENCES

[1] A. Duran, E. Ayguade, R. M. Badia, J. Labarta, L. Martinell, X. Mar-torell, and J. Planas, “Ompss: A proposal for programming heteroge- neous multi-core architectures,” Parallel Processing Letters, 2011.

[2] X. Tan, J. Bosch, D. Jiminez-Gonzalez , C. Alvarez-Martinez, E. Ayguade and M. Valero. ”Performance Analysis of a Hardware Accelerator of Dependence Management for Task-based Dataflow Programming models”. Accepted to the 2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS).

[3] XILINX, “Zynq-7000, etc..” [online], 2015. http://www.xilinx.com/support/documentation/user guides/ug585-Zynq-7000-TRM.pdf.

[4] B. S. Center, “Bsc application repository(bar).” [online], 2014. https//pm.bsc.es/projects/bar/wiki/Applications.

[5] F. Yazdanpanah, C. Alvarez, D. Jimenez-Gonzalez, R. M. Badia, M. Valero, “Picos: A hardware runtime architecture support for ompss,” Future Generation Computer Systems(FGCS), 2015.

The OmpSs Reductions Model and how to deal with Scatter-Updates

Jan Ciesko[1], Sergi Mateo[1], Xavier Teruel[1,2], Vicenç Beltran[1],
Xavier Martorell[1,2], Rosa M. Badia[1,2] and Jesús Labarta[1,2]
[1]*Barcelona Supercomputing Center*
[2]*Universitat Politècnica de Catalunya*
{jan.ciesko, sergi.mateo, xavier.teruel, vicenc.beltran, xavier.martorell, rosa.m.badia, jesus.labarta}@bsc.es

Abstract *– Scatter-updates represent a reoccurring algorithmic pattern in many scientific applications. Their scalable execution on modern systems is difficult due to performance limitations introduced by their irregular memory access pattern that prohibits an efficient use of the memory subsystem. Further performance degradation is caused by techniques that are required in order to eliminate potential data races and come at the cost of overhead. Taking a closer look at algorithmic properties, access patterns and common support techniques reveals that a one-size-fits-all solution does not exist and solutions are needed that can adapt to individual properties of the algorithm while maintaining programming transparency. In this work we propose a solution framework that supports a broad set of techniques, provides the required access pattern analytics to allow dynamic decision making and shows what language extensions are needed to maintain programming transparency. A reference implementation in OmpSs, a task-based parallel programming model, shows programmability and scalability of this solution.*

I. Introduction

The widening gap between processor and memory speeds periodically brings up the discussion on how to improve scalability of algorithms that hit the memory wall exceptionally fast due to their scattered memory updates. At the core of the problem are high memory access latencies that become dominant as a result from the caching and bandwidth inefficiencies of these algorithms and the overheads introduced by techniques that ensure correctness by eliminating the possibility of data races. Among these techniques, only a single generally applicable solution exists, namely access synchronization. Synchronization uses software and hardware assisted techniques to implement atomicity of the update operation (read-modify-write) with overheads that differ between processor architectures. Synchronization constructs are typically members of either the language or runtime specification of a programming model and therefore easy to use but unfortunately do not address the issue of poor locality of these algorithms.

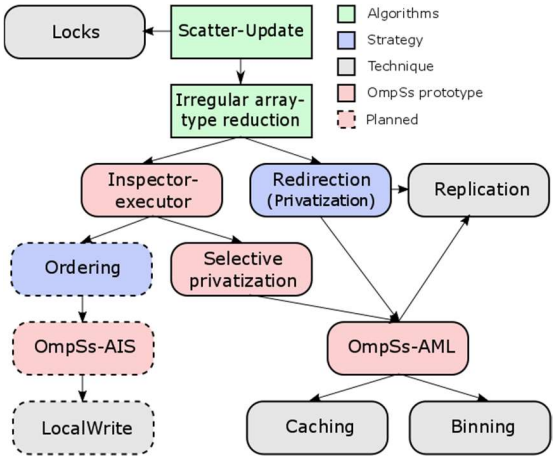


Figure 1 Landscape of algorithms, strategies and techniques

```
1 while(simulation_runs()){
2  #pragma omp loopstep
3  #pragma omp task reduction (+:v:SPB) invariant (v)
4   for(int i = 0; i < iters; i++) {
5     j = f(i);
6     v[j]++;
7   }
8 }
```

Figure 2 Reduction kernel with proposed clauses to support alternative memory layouts with inspectors and executors

A special case occurs when the iterative scatter-update implements a function that is associative, communicative and has no control dependency between iteration loops (algebraic monoid). These algorithms are called reductions and allow a whole set of additional techniques to improve performance and scalability. Main implications of these properties are two-fold: firstly, the order of memory accesses does not matter anymore which allows concurrent executions without maintaining a constant execution order (of tasks, loop iterations or particular instructions) and the existence of the neutral element allows the use of scratch data to temporarily store intermediate results. This led to the development of different support techniques [1] that fall into two strategies. Access redirection is a strategy where accesses are redirected to a scratch

storage while leaving the iteration space untouched. The scratch memory is typically a thread-private copy of the original data (replication) or any data structure that fulfills a similar goal. Ordering is another strategy that avoids redirection and reorders iterations by specific criteria instead. Which of these is used and how they are configured depends on algorithmic properties that require both compiler support and runtime analysis. As of today, none of these techniques other than replication made its way into popular parallel programming models.

In this work we present the OmpSs Reductions Model (OmpSs-RM) which implements a framework to support redirection techniques with alternative memory layouts (OmpSs-AML) such as binning or software caching as well as ordering techniques that require alternative iteration spaces (OmpSs-AIS) such as LocalWrite in near future. In particular we show what new language constructs are needed, how the inspector-executor model can be integrated into the runtime as well as how scientific applications can benefit from these techniques. Figure 1 shows a landscape of algorithms, strategies, techniques and their support in OmpSs.

II. Support in OmpSs

The OmpSs-AML implementation builds on top of the existing functionality of reduction scope definition, pre-allocation, allocation on demand and lazy initialization. In order to support AMLs, we require three additional information from the developer.

Firstly, the developer is required to express the intention to use an AML. This step is necessary in order to preserve consistency as with AMLs, the scratch memory is not necessarily a replica of the original data anymore. For this purpose, we prose the extension of the *reduction clause* by the additional parameter *MODE*, where mode is an identifier of a vendor provided privatization technique.

Further, in order to support AMLs that require an inspector-executor, we propose the addition of the *invariant (target)* clause. The invariant clause defined over a *target* specifies that the access pattern of the target as well as the calling order within the scope of a reduction are invariant. This step is important to guarantee that the inspector-executor is always applied to the matching function and that optimization results obtained during the inspection phase are still valid for subsequent function calls or task instances.

```
1 while(...){
2   ID = instance_invariant_identifier;
3   frameID = handle_optimization_frame(ID)
4   task = new reduction_task(frameID, taskcode,...)
5   task.run();
6 }
7 ...
8 taskcode(...){
9   analytics * a; AML * aml;
10  t frameInstanceID = get_frameInstanceID();
11  aml = get_thread_storage(v);
12  analytics = get_analytics(frameInstanceID);
13  if(!analytics.ready)
14    inspect(&v[j],analytics,frameInstanceID);
15  for(...) {
16    j = f(j);
17    (*SPB_get(&v[j], analytics, aml))++;
18  }
19 }
```

Figure 3 Intermediate code prototype showing the main loop body, task code and runtime APIs

Lastly we propose the addition of the *loopstep* pragma. This pragma defines the scope of an optimization frame and is used to differentiate between inspection and execution phases. Figure 2 shows high-level code that uses selective privatization and an AML to implement an array-type reduction.

Figure 3 shows an intermediate code prototype where an instance-invariant identifier (*frameID*) is created to identify an optimization frame and that is subsequently passed to all participating task. By doing so, all tasks sharing one identifier are associated to one optimization frame. Once a task instance is created, the frame identifier is used to generate a new unique identifier for that particular task instance (*frameInstanceID*). In OmpSs and for the context of reductions, the frame identifier is computed as XOR between the reduction target address and value of the reducer function pointer.

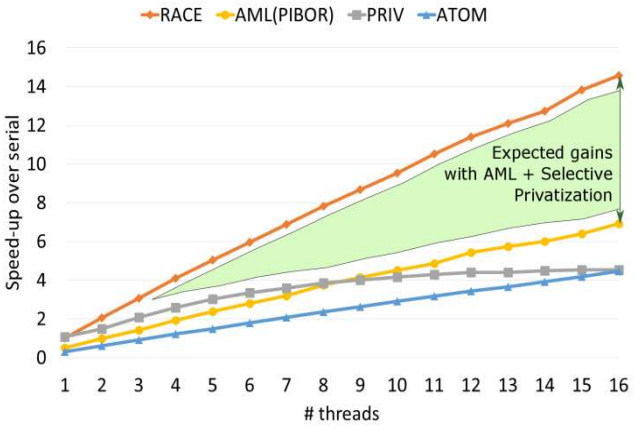


Figure 4 Lulesh reduction kernel scalability on the Xeon E5 processor with different support techniques showing the expected performance when properly exploiting access locality of tasks with AMLs and selective privatization

The instance frame identifier for each particular task is created again as an XOR between frame identifier and a task creation counter. In case of

nesting, new identifiers are created for each nest. Currently, optimization frames across nesting levels are not supported.

III. Case study

Our work on OmpSs AML and the inspector-executor model was largely motivated by Lulesh, a seismic simulation code that contains irregular array-type reductions. Inspecting its memory access pattern revealed a linear access pattern with very small overlaps for boundary iterations. The inspector used in this case records histograms over addresses, over distances between memory accesses and over the rate of distance changes. This information is evaluated once the optimization frame is completed. For the case of Lulesh and AML with selective privatization, the evaluation produces an ownership table that is used in the executor phase to determine whether an update operation accesses task local data or not. Since most accesses in Lulesh are local, the original data can be updated without the need of synchronization nor redirection. Figure 4 shows scalability of a Lulesh reduction kernel implemented with different techniques. We expect that inspector-executor enabled AMLs will be close to a version that is free of any additional overheads but contains data races (RACE). Further evaluation is pending.

IV. Conclusion and Future work

We are currently evaluating OmpSs AMLs and inspectors-executors to derive further knowledge about overheads of the inspection phase, its usability in other applications and architectures as well as the integration of alternative iteration spaces (AIS) into OmpSs. This work aims to influence the OpenMP specification to support this type of algorithms in the future.

ACKNOWLEDGMENT

I would like to thank all my coauthors for their invaluable insights and their patience when exposed to my ideas during countless meetings.

REFERENCES

[1] H. Yu and L. Rauchwerger, Adaptive Reduction Parallelization,14th ACM Intl. Conf. on Supercomputing, 2000

Runtime Estimation of Performance–Power in CMPs under QoS constraints

Rajiv Nishtala, Xavier Martorell, Paul Carpenter
Barcelona Supercomputing Center
rajiv.nishtala@bsc.es

Abstract*-One of the main challenges in data center systems is operating under certain Quality of Service (QoS) while minimizing power consumption. Increasingly, data centers are exploring and adopting heterogeneous server architectures with different power and performance trade-offs. This not only requires careful understanding of the application behavior across multiple architectures at runtime so as to enable meeting power and performance requirements but also an understanding of individual and aggregated behaviour of application and server level performance and power metrics.*

I. INTRODUCTION

Modern data centers increasingly demand improved performance (QoS, quality-of-service) with minimal power consumption. Managing the power and performance requirements of the applications is challenging because these data centers, incidentally or intentionally, have to deal with server architecture heterogeneity [1], [2]. One critical challenge that data centers have to face is how to manage system power and performance given the different application behavior across multiple different architectures.

The objective of this study is to understand individual and aggregated behaviour of thread/server level performance and power trade-offs to solve the online optimization problem. This work is presented in two main parts:

1) *Runtime Estimation of Performance–Power, REPP*, is a scheme for runtime estimation of power and performance at thread or server level parametrized by numerous P-States and Cl-States, leveraging hardware performance counters available on all major server architectures. The model is accurate enough to capture the real behavior, is driven by existing performance counters, and, since the computational complexity at runtime is low, it can be used for fine-grain power management.

2) *Vinson* is a QoS aware thread mapping schema for latency critical (LC) workloads. Vinson aims to accurately meet the required latency for LC workloads and maximizes throughput for batch workloads given a power constraint by adjusting the number of cores allocated and P-States (DVFS, Dynamic Voltage/Frequency Scaling)

II. RELATED WORK

Recently machine learning techniques to predict performance-power and co-allocating LC and batch workloads have garnered significant attention from both academic and industry. Below we summarize the strongly related recent research conducted in the aforementioned areas.

REPP: Prior research works have focused on mapping applications to resources (mainly to CPUs/cores) to improve performance while saving power. In particular, Bellosa [3] used performance monitoring counters (PMCs) at run time to build a power-aware policy at OS level. Isci first showed that using PMCs it is possible to detect fine-grained application phases [4] and then show breakdown of power per component using multilinear models [5] . B. Rountree et al. [6] estimate performance (IPC, Instructions Per Cycle) across P-States by monitoring the number of leading load cycles. Miftakhutdinov et al. [7] predict performance on simulated architectures based on prefetch and variable memory access latencies. Bo Su et al. [8] take advantage of the PMCs available on AMD for estimating the leading loads metric, and predict performance (IPC) across P-States.

Vinson: Most real-time schedulers are based on feedback controllers to quickly adapt to applications' demand. For example, Octopus-Man[9] uses a feedback controller to adjust the number of cores every few seconds in response to changes in measured latency, but not the frequency of the cores. Despite using a heterogeneous architecture they do not leverage using both big and small cores at the same time. Heracles[10] also uses a feedback controller that enables safe colocation of both LC and batch workloads while individually considering CPU, memory and network isolation. However, this paper banks on the cache allocation technology (CAT) and DRAM bandwidth monitor not available on most non-modern Intel architectures and other architectures. Pegasus[11] uses a feedback controller to adjust P-States every few seconds using RAPL in response to changes in measured latency, but not the number of cores and fails for short-term, sub-millisecond variations of applications. In response, Rubik [12] implements a feedback controller to

adjust DVFS at a very small intervals for short-term, sub-millisecond variabilities to cope with diurnal variations similar to Pegasus. However, both, Rubik and Pegasus do not consider varying the number of cores allocated to LC workloads

III. Results To-Date

The results for REPP are validated on AMD Phenom II X4 B97, Intel Corei7-2760QM and ARM Juno R0 – 64bit. As Vinson is work-in-progress, we only show for ARM.

Fig. 1: Runtime power and performance prediction over time (in seconds) for multiprogrammed workload consisting of milc, milc, xalancbmk and blackscholes.

REPP: Figure 1 shows an example of the power and performance prediction in runtime implemented on the Intel architecture for the first 20 seconds of execution the workload (the technique to select workloads is described in [13]). From top-to-bottom, the first (and second) graph represents the power (and performance) as measured using RAPL (and PMC) and the prediction made using REPP. The third and fourth graphs show the random combination of P-States and Cl-States generated for individual cores, respectively, for the first 20 seconds. We highlight two results. First, REPP does show the capability to adapt to workloads consisting of multiple thread phases. For instance, observe at second 12, REPP makes a 11 mW error in predicting power, this is because of the huge changes in P-States and Cl-States. In this scenario, the P-States for core 0, 1, 2, 3 change from 0.8 to 2.4, 0.8 to 2.2, 0.8 to 2.2 and 1.2 to 0.8 respectively and the Cl-States change from 10 to 23, 1 to 31, 41 to 48 and 3 to 35. Observe that these errors only occur with huge changes in P-States and Cl-States in rapid intervals (For example, second 4). Ozlem et al [14] on the other hand, show that rapid changes in power or performance are seldom required in data center environments. Second, REPP can predict power and performance per thread, which can not be accomplished using the in-built RAPL register. In this particular workload, we make an error of 9.4% (384 mJ) and 15.2% (1500 MIPS) when predicting power and performance over 300 seconds, respectively.

Vinson: We present a proof of concept to show that using the number of cores and frequency can help meet the QoS requirements while reducing energy consumption. We simulate memcached from Cloudsuite 3.0 to receive a fixed number of requests per second (RPS) on an ARM platform at all possible core and frequency configurations for a fixed quantum. We sample the latency as the QoS at the 95th percentile (QoS95) and energy consumption. We select those configurations which satisfy QoS with the least energy consumption. On the other hand, for Octopus-Man we select configurations when running on big or small cores exclusively at the highest frequency. Figure 2 shows the average QoS95 and the energy consumed when using Vinson and Octopus-Man for memcached. Vinson leverages the big.LITTLE cores available on the ARM platform truly and reduces energy consumption by 27.74% (on average) over Octopus-Man. Observe that both Octopus-Man and Vinson give same results for very high RPS (greater than 33000) and very low RPS (lower than 22000) because Vinson also runs exclusively on the Big or Small cores. Similar results were observed also for Websearch and multithreaded Parsec 3.0 Benchmarks.

Fig. 2: QoS at the 95th percentile (in ms) and the energy consumed when using Vinson and Octopus-Man on ARM platform for memcached.

PUBLICATIONS: R. Nishtala, M. G. Tallada, and X. Martorell, “A methodology to build models and predict performance-power in cmps,” in Proc. of 44th ICPPW, Sept 2015

REFERENCES

[1] J. Mars, et. al, “Heterogeneity in homogeneous; warehouse-scale computers: A performance opportunity,” CAL 2011.

[2] R. Nathuji, et. al, “Exploiting platform heterogeneity for power efficient data centers,” in Proc. of ICAC ’07.

[3] F. Bellosa, “The Benefits of Event: Driven Energy Accounting in Power-sensitive Systems,” in Proc. of ACM SIGOPS EW 9.

[4] C. Isci and M. Martonosi, “Phase characterization for power: evaluating control-flow-based and event-counter-based techniques,” in HPCA ’06.

[5] C. Isci and M Martonosi, “Runtime Power Monitoring in High-End Processors: Methodology and Empirical Data,” in Proc. of MICRO 36
[6] B. Rountree, et. al, “Practical performance prediction under Dynamic Voltage Frequency Scaling,” in Proc. of IGCC 2011.
[7] R. Miftakhutdinov, et. al, “Predicting Performance Impact of DVFS for Realistic Memory Systems,” in Proc. of MICRO-45.
[8] B. Su, et. al, “Implementing a Leading Loads Performance Predictor on Commodity Processors,” in Proc. of USENIX ATC’14.
[9] V. Petrucci, et. al. “Octopus-man: Qos-driven task management for heterogeneous multicores in warehouse-scale computers,” in Proc. of HPCA 2015
[10] D. Lo, et. al, “Heracles: Improving resource efficiency at scale,” in Proc. of ISCA 2015.
[11] D. Lo, et. al, “Towards energy proportionality for large-scale latency-critical workloads,” in Proc. of ISCA 2014.
[12] H. Kasture, et. al, “Rubik: Fast analytical power management for latency-critical systems,” in Proc. of MICRO-48.
[13] D. Sanchez and C. Kozyrakis, “Vantage: Scalable and Efficient Fine-grain Cache Partitioning,” SIGARCH Comput. Archit. News
[14] O. Bilgir, et .al, “Exploring the Potential of CMP Core Count Management on Data Center Energy Savings,” in Proc. of WEED 2011.

Conformational landscape of small ligands: A Multilevel strategy to determine the conformational penalty of bioactive ligands

Antonio Viayna, Jordi Juárez-Jiménez, Xavier Barril, F. Javier Luque
Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de l'Alimentació,
Avgda. Prat de la Riba, 171, 08921, Santa Coloma de Gramanet
toni.viayna.89@gmail.com

Abstract- *Determining the conformational penalty required for adopting the bioactive conformation is still a challenging question in drug design, because a small uncertainty in this free energy component can lead to significant errors in the predicted activities. Herein, we use the Multilevel strategy, a methodology recently developed by our group, to explore the conformational preferences of ligands in solution, and to estimate the conformational cost of selecting the bioactive conformation.*

I. INTRODUCTION AND OBJECTIVES

In order to enhance the binding affinity, complementarity between the functional groups present in the ligand and the residues of the receptor's binding pocket is essential [1]. To achieve it, some conformational changes are required both in the ligand and the receptor. In the receptor case, those changes typically involve the rearrangement of side chains in the binding cavity and structural modifications in secondary structural elements. With regard to the ligand, conformational changes are associated with the adoption or selection of the bioactive conformation in the bound state. These conformational changes contribute to the binding free energy (ΔG_{bind}), which can be expressed as the addition of the free energy contribution due to the recognition between ligand and receptor in the bioactive conformation (ΔG_{int}) and the cost associated with the structural reorganization of both the ligand and receptor in solution (ΔG^{L}_{conf} and ΔG^{R}_{conf}). (Fig.1)

$$\Delta G_{bind} = \Delta G_{int} + \Delta G^{R}_{conf} + \Delta G^{L}_{conf}$$

Fig. 1. Sum of different contributions to the binding free energy during the ligand-receptor interaction

Focusing on the ligand, the bioactive conformation is just one of the many possible conformations in the physiological media. Intituively, one can expect that a good binder will be recognized by the receptor in a low energy conformation, but this is not always the case. Many times, the bound conformation might not correspond to the global minimum of the free ligand, and then a conformational penalty must be paid to adopt the bioactive conformation. If we consider that biological activity and binding free energy are directly related, then the existence of a high conformational penalty may lead to a significant error in the binding affinity and consequently in the predicted activity [2].

Different research groups have attempted to find computational strategies well suited to estimate the conformational cost needed for the selection of the bioactive conformation of ligands. [3] [4] [5]

Recently, our research group has developed the Multilevel strategy in order to explore the conformational preferences of drug-like compounds in solution and estimate the relative stability of the most populated conformations. [6][7]. The Multilevel strategy relies in two main approximations. First, it relies on the "predominant state approximation" [8], which states that the conformational space can be divided into different wells and the total configurational integral is equal to the sum of the configurational integrals of all the wells. Therefore, the free energy of a flexible molecule can be expressed as the addition of the contributions of the separate conformational wells. Second, the "Multilevel approach" assumes the combination of Low-Level methods to carry out the conformational sampling of flexible molecules to find the conformational minima, and then, High-Level methods are utilized to refine the relative stability of the wells (Fig. 2).

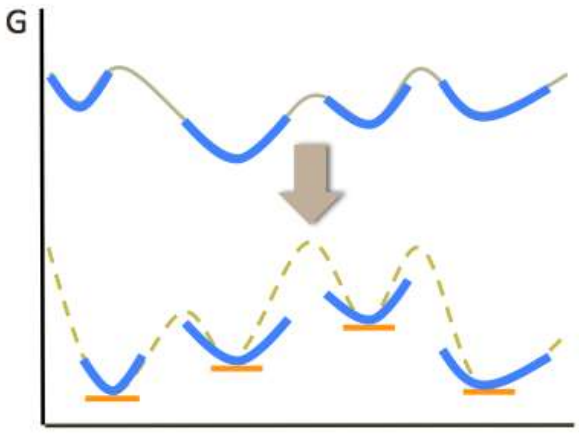


Fig. 2. Schematic representation of the Multilevel strategy. (Upper curve) The free energy surface is first explored at a Low-Level of theory, which permits to identify the major conformational wells. (Bottom curve) The relative stability of the minima is refined at a High-Level of theory, while including the local entropy of the conformaitional wells.

In this work, our interest is to use the Multilevel method to explore the conformational preferences of a diverse set of bioactive ligands in solution and to predict the conformational penalty of selecting the bioactive species, in order to validate this methodology.

II. METHODS

For each ligand, we perform the Low Level part, with classical Molecular Dynamics, using AMBER14 program and gaff force field. Prior to the production runs, we minimized the system, and then the system was equilibrated by rising the temperature from 50 to 298 K at constant volume. Finally, the density of the system was equilibrated at constant pressure, and finally production runs were performed at constant volume. The conformations sampled by the ligand from the trajectories were clusterized by considering the set of active torsions of the ligand, and finally to obtain the different wells.

The High Level refinement was developed taking the representative structure chosen as the minimum energy conformer of the different wells, and submitting it to a IEF-MST/B3LYP/6-31G(d) geometry optimization. The energy of the optimized structure was refined at the MP2/aug-cc-pVDZ level, including the zero-point energy correction, the solvation free energy, and the local conformational entropy of the well.

III. RESULTS AND DISCUSSION

Preliminary results of 12 compounds of the set of drug-like ligands show that the conformational penalty is generally low, corresponding to around 80% of the set. (Fig. 3) This general tendency is not surprising, because most of the molecules are drug-like compounds, and we expect that the conformational cost has been minimized during their design.

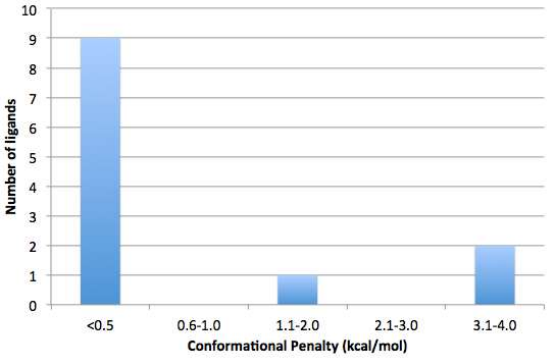


Fig. 3. Histogram confronting the number of ligands of the set and the different intervals of conformational penalty

Nevertheless, we have also detected that some ligands have a high conformational cost (> 2kcal/mol). In all cases, we were able to explain this cost, which was due to either steric hindrance, or breaking of intramolecular interactions.

As an illustrative example, the case of IQP ligand (PDB code 1YDR) (penalty = 3.8 kcal/mol) is a good example. This case is explained in terms of steric hindrance of the methylpiperazine moiety promoted upon filling the protein cavity (Fig. 4).

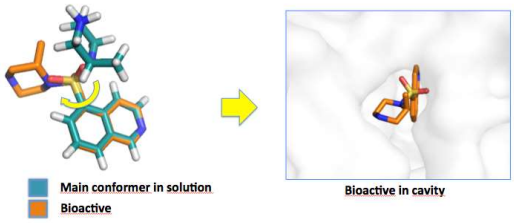


Fig. 4. Left: Superposition of the main conformer in solution according to Multilevel strategy (blue) and bioactive structure (orange) of the IQP ligand. Right: Bioactive structure inside the protein cavity.

Another illustrative example is BMU ligand (PDB code 1KV1) (penalty = 3.3 kcal/mol), which is explained in terms of the forced twisting of the bond between the urea and pyrazol groups. (Fig. 5)

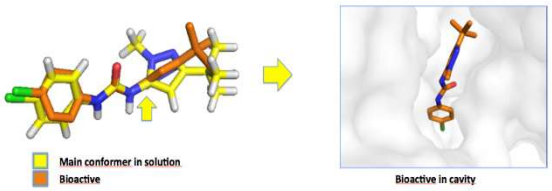


Fig. 5. Left: Superposition of the main conformer in solution according to Multilevel strategy (yellow) and bioactive structure (orange) of the BMU ligand. Right: Bioactive structure inside the protein cavity.

IV. CONCLUSIONS AND FURTHER WORK

As indicated in the results part, the drug-like compounds tend to have low conformational penalties. Only in few cases the cost is larger than 2 kcal/mol, which reflects the curated procedure required for the development of drugs.

Future work will be focused in completing the analysis of the whole set of compounds, to identify the factors that lead increase the conformational stress upon ligand binding, and to assess the possibility of introducing improvements in the Low-Level sampling methods.

ACNOWLEDGEMENTS

We thank Universitat de Barcelona (UB) for the financial support (APIF Fellowship). We thank the Centre de Serveis Científics i Académics de Cataluntya (CESCA) and the Barcelona Supercomputing Center (BSC) for computational facilities.

REFERENCES

[1] C. Bissantz, B. Kuhn and M. Stahl. "A Medicinal Chemsit's Guide to Molecular Interactions" J. Med. Chem. 2010, 53, 5061-5084

[2] S. Chung, J.B. Parker, M. Bianchet, L.M. Amzel and J.T. Stivers. "Impact of linker strain and flexibility in the design of a fragment-based inhibitor" Nat. Chem. Biol. 2009, 5, 407-413

[3] E. Perola and P.S. Charifson. "Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganitzation upon Binding" J. Med. Chem. 2004, 47, 2499-2510

[4] J. Tirado-Rives and W.L. Jorgensen. "Contribution of Conformer Focusing to the Uncertainity in Predicting Free Energies for Protein-Ligand Binding". J. Med. Chem. 2006, 49, 5880-5884

[5] K.T. Butler, F.J. Luque and X. Barril. "Toward Accurate Relative Energy Predictions of the Bioactive Conformation of Drugs" J. Comput. Chem. 2008. 30, 601-610

[6] F. Forti, C. Cavasotto, M. Orozco, X. Barril and F.J. Luque. "A Multilevel Strategy for the Exploration of the Conformational Flexibility of Small Molecules" J. Chem. Theory Comput. 2012, 8, 1808-1819

[7] J. Juárez-Jiménez, X. Barril, M. Orozco, R. Pouplana and F.J. Luque. "Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands" J. Phys. Chem. B. 2015, 119, 1164-1172

[8] M.S. Head, J.A. Given and M.K. Gilson. "Mining Minima: Direct Computation of Conformational Free Energy" J. Phys. Chem. A. 1997, 101, 1609-1618

Characteritzation of Protein-Protein Interfaces and Identification of Transient Cavities for its Modulation.

Mireia Rosell and Juan Fernandez-Recio
Barcelona Supercomputing Center, Barcelona, Spain
mireia.rosell@bsc.es

Abstract- *Protein-protein interactions (PPIs) play an essential role in many biological processes, including disease conditions. Strategies to modulate PPIs with small molecules have therefore attracted increasing interest over the last few years, where successful PPI inhibitors have been reported into transient cavities from previously flat PPIfs.*

Recent studies emphasize on hot-spots (those residues contribute for most of the energy of binding) as promising targets for the modulation of PPI. PyDock is the only computational method that uses docking to predict PPIfs and hot-spots (HS) residues. Using Normalized Interface Propensity (NIP) values derived from rigid-body protein docking simulation, we are able to predict the PPIfs and HS residues without any prior structural knowledge of the complex.

We benchmarked the protocol in a small set of protein-protein complexes for which both structural data and PPI inhibitors are known. We present an approach aimed at identifying HS and transient pockets from predicted PPIfs in order to find potential small molecules capable of modulating PPIs. The method uses pyDock to identify PPIfs and HS and molecular dynamics (MD) techniques to describe the possible fluctuations of the interacting proteins in order to suggest transient pockets. Afterwards, we evaluated the validity of predicted HS and pockets for in silico drug design by using ligand docking.

We present a strategy based on MD and NIP which allows to identify cavities as potentially good targets to bind inhibitors when there is no information at all about the protein-protein complex structure.

I. INTRODUCTION

Protein-protein interactions (PPI) play an essential role in regulating biological processes, such as signaling pathways in cells, and are involved in the majority of diseases, highlighting the interest in protein-protein interfaces (PPIfs) as an attractive target for therapeutic intervention. A detailed structural knowledge of PPIs is needed to understand disease at molecular level, to identify new targets for therapeutic intervention and also to find small molecules capable of inhibiting PPIs [1].

It has been reported that only a few amino acids (so-called "hot-spot" residues) usually contribute to the majority of the free energy of binding. Experimental approaches typically define hot-spots (HS) as those residues that decrease binding energy in more than 1 or 2 kcal/mol upon mutation to alanine [2]. These HS residues are important in the context of drug discovery targeting PPIs because blocking them seems the only way for a small-molecule to compete with a protein-protein interaction. The reason is that PPIfs are usually large and involve higher number of atomic interactions, and hence have higher affinity as compared to protein-ligand interfaces. Other difficulties are that PPIfs do not have clear binding pockets for drug binding, and that very often, both the location of the interface and the binding mode of the PPI are not known. Successful PPI inhibitors have been reported into transient cavities from previously flat PPIfs [3].

Computational approaches such as protein-protein docking and molecular dynamics (MD) are becoming increasingly important tools in drug discovery in order to help solving the difficulties mentioned above. PyDock algorithm (a tool developed in our lab to perform protein-protein docking) is the only computational method that uses docking to predict PPIfs and HS residues when there is no structural information available of the protein-protein complex [4,5,6,7]. The method applies the fast Fourier transform algorithm to the unbound proteins of the complexes, followed by the energy-based scoring from pyDock to calculate the Normalized Interface Propensity (NIP). Using pyDock and MD techniques to suggest putative transient cavities, we present an approach addressed to targeting PPIs and to find potential small molecules capable of modulating PPIs [8].

II. MATERIALS AND METHODS

A. Generation of small benchmark.

From the 2P2I database, we benchmarked the protocol in a small set of protein-protein complexes for which both structural data and PPI inhibitors are known.

B. Hot-spot prediction from protein-protein docking.

We used ZDOCK 2.1 [9] to generate 2000 rigid-body docking poses. We used the top 100 lowest-energy solutions proposed by pyDock algorithm to

calculate the Average Buried Surface (ABS) and Normalized Interface Propensity (NIP) [7]. We applied a cutoff of NIP $\geq$ 0.2 to predict HS residues.

C. *Molecular dynamics and transient cavities detection.*

In the correct predictions of HS in PPIfs, we used AMBER10 to detect transient pockets on the unbound proteins, which were selected based on the predicted HS. For each case, using Fpocket [10], we analyzed 1000 out of 10000 snapshots resulting from 10ns of simulation.

D. Ligand docking.

In those selected snapshots with a putative transient cavity, we used MAESTRO to prepare the structures for docking, as well as the inhibitors. We generated 1000 docking poses from RDOCK (flexible ligand docking).

Results

Assuming the knowledge of the PPI, 6 out of 10 cases are successful. We have focused on these cases to continue the analysis to identify transient cavities. HS and PPIfs predictions are shown in Table I (Figure 1).

[a] Number of predicted hot-spots (NIP$\geq$ 0.2). [b] Number of predicted hot-spots that are located at the PPIfs. [c] Number of predicted hot-spots that are located at the protein-inhibitor interface (PII).

* Correct predictions (these were selected for a more thorough analysis).

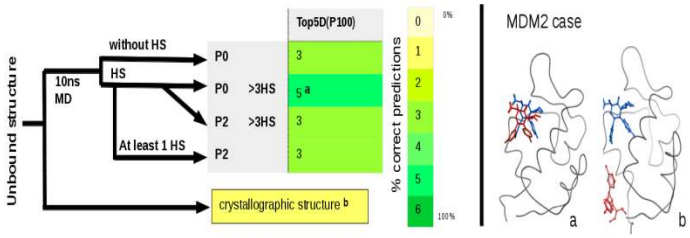
TABLE I

Complex	***HSpred*** [a]	***HSpred at PPIfs*** [b]	***HSpred at PII*** [c]	
Bcl-X_L/BAK	9	2	2	
Xiap_BIR3/Caspase	12	0	2	
HPV_E2/E1	21	12	7	
IL2/IL2R	4	4	4	
Nos/iNos	0	0	0	
Integrase/LEDGF	16	0	1	
MDM2/p53	7	4	4	
Xiap_BIR3/Smac	19	6	7	
TNFR1A/TNFB	14	1	0	
ZipA/FtsZ	0	0	0	

We proposed two strategies to analyze the pockets and identify the putative transient cavity from MD. One is using the top ranked pockets predicted by Fpocket and the second is using the pockets that have at least 2 most frequent residues from all those pockets located at a concrete place (defined using HS) during the simulation (Figure 2.).

In order to evaluate the weight of HS in the role of selection of possible candidates with an interesting transient cavity, we also evaluated both strategies without HS. We compared both strategies with unbound cases. From all strategies, we selected the snapshots with a putative transient cavity in different ways: the top scored by fpocket, the top druggable defined by fpocket and the top druggable from the top 100 scored by fpocket (topD(P100)). We propose the top 5 candidates (from topD(P100)) for further analysis using ligand docking. If we compare unbound structures with respect to the selected cases, we obtain better results in these selections (Figure 3).

Fig. 3. At left, different strategies of selection of candidates with the best transient cavity. In unbound structures selected according to the results obtained from PPIfs selection, we analyzed the pocket and the transient pockets resulting from the simulation. P0 means the top ranked pockets strategy and P2 means the strategy applying most frequent residues. Transient pockets were analyzed: Using HS (at least 3HS) and without HS. Correct predictions were considered with a PPV&COV$\geq$40%. At right, results of ligand docking in MDM2 applying P0 with at least 3HS strategy (a) and directly with the unbound structure (b) of selection.



III. CONCLUSIONS

The characterization of druggable cavities in PPIfs is still unknown where predicting PPIfs from a three dimensional structure is a key task for the modulation of PPIs. The use of the NIP-based HS prediction method improves the identification of transient cavities from MD simulation when compared to known binding cavities. We propose a new tool to predict and characterize PPIs, PPIfs and HS residues. We present a strategy based on MD and NIP which allows to identify cavities as potentially good targets to bind inhibitors. This approach can be extremely useful in a realistic scenario of drug discovery targeting PPIfs, when there is no information at all about the protein-protein complex structure.

REFERENCES

1. Wells J.A., McClendon CL., (2007). Nature, 450,1001-1009.
2. Bogan A.A. and Thorn K.S., (1998). J. Mol. Biol, 280, 1-9.
3. Basse MJ., Betzi S., Bourgeas R., et al. (2013). Nucleic Acid Research, 41, 824-827
4. Teng S., Madej T., Panchenko A., Alexov E., (2009). Biophys J. 96, 2178–2188.
5. Cheng T.M-K.,Blundell T.L.,and Fernandez-Recio J., (2007).Proteins, 68, 503_515.
6. Fernandez-Recio J., Totrov M., and Abagyan R.,(2004). J Mol Biol, 335, 843-865..
7. Grosdidier S., and Fernandez-Recio J., (2008). BMC Bioinformatics, 9:447.
8. Hubert Li, Vinod Kasam,. (2014) J Chem. Inf. Model 54, 1319-1400.
9. Chen R., Weng Z. (2003). Proteins 51, 397-408.
10. Le Guilloux V., Schmidtke P. and Tuffery P., (2009). BMC Bioinformatics, 10-168

Improvement of Protein-Ligand Binding Affinity Prediction using Machine Learning Techniques

Gabriela Hernández[1,2], Jelisa Iglesias[1,3], Suwipa Saen-oon[1]
Supervisors:
Jorge Estrada[1], Ricard Gavaldà[3], Víctor Guallar[1,4]
[1]*Barcelona Supercomputing Center (BSC-CNS);* [2]*Université Lumière Lyon 2 – EM-DMKM;* [3]*Universitat Politècnica de Catalunya (UPC);* [4]*Institució Catalana de Recerca i Estudis Avançats (ICREA)*
ghernand@bsc.es

Abstract- *Predicting protein-ligand binding affinities constitutes a key computational method in the early stages of the drug discovery process. Molecular docking programs attempt to predict them by using mathematical approximations, namely, scoring functions. In the last years, several scoring functions have been developed, encompassing different terms, from electrostatic forces to protein-ligand interaction fingerprints and beyond. However, it has been noticed that usually each individual scoring function cannot be generalized and its predictive power is arguable. The aim of this study is to improve the binding affinity prediction by finding potential models to combine ten different scoring functions, exploiting machine learning techniques.*

Keywords: Protein-ligand binding, Scoring Functions, Drug discovery, Machine learning.

I. INTRODUCTION

The amount of proteins and molecules with publicly-accessible 3D structures is rapidly growing [1]. As a consequence, structure-based drug design (SBDD) is becoming increasingly popular to discover new potential drugs. In this process, the protein-ligand binding plays a fundamental role. For a protein of interest, putative ligand drug candidates are discovered or designed in order to bind the target protein and modulate its activity. The strength of these docked molecules is referred as binding affinity [2]. *In vitro* determination of binding affinity is highly expensive and time consuming. In order to address this issue, *in silico* molecular docking techniques have emerged, using scoring functions (SFs) to estimate the binding affinity of each protein-ligand complex [3]. In general, the SFs can be broadly classified into four categories: 1) force-field based, 2) knowledge-based, 3) descriptor-based and 4) empirical scoring functions [4].

Despite the efforts in develop SFs, underlying different principles, to accurately predict the binding free energy, it has been shown in different studies [5, 6, 7] their limitations and lack of generalization. Nevertheless, it also has been noticed that it is unlikely that a set of SFs will be in error at the same time for a protein-ligand system. Based on this idea, exhaustive studies have been realized to create a most robust scoring function by using the best combination of a set of individual SFs in different fashions. Some attempts were performed in previous works [4, 7], to create consensus SFs based on conventional approaches such as rank-based, percent-based, range-based and vote-based strategies. However, their results are based on a strong assumption which entails that all the individual SFs contribute equally [6]. In other studies, the authors proposed protocols to rescue poor docking results from different SFs by combining conventional approaches such as rank-based with a classifier in order to only discriminate good and bad binders for some target proteins with a set of ligands, without predicting the binding free energy [8, 9]. To the best of our knowledge, no study has fully investigated and assessed the combination of different SFs by using machine learning approaches to better predict the protein-ligand binding affinity, leaving room for improvements.

The purpose of this study is to explore and assess the combination of ten different SFs belonging to the four categories: force-field based (PELE, MM-Gbsa, rDock), knowledge-based (XScore-HMScore, DSX, Autodock VINA), descriptor-based (NNScore and RFScore) and empirical (Glide XP, Glide SP, X-Score) by employing several statistical and machine learning techniques from the perspective of description, regression and intelligibility. To this end, we look forward to discover sets of SFs and models that might be relevant for improving the protein-ligand binding affinity prediction.

II. DATA AND METHODS

A. Protein-ligand complex dataset

In the work by Cheng et al. [10], they built a core set based on the 2007 PDBbind benchmark that circumscribes a diverse set of high-quality protein families. From this core set, we used 64 different

proteins, each of which binds to three different ligands to form a set of 191 unique protein-ligand complexes. By using stratified sampling, we created two disjointed sets for training, with 70% of the complexes, and validation, with the remainder 30%. For both sets, we calculated ten different SFs for each protein-ligand complex, so that each system was described by a 10-dimensional vector. We evaluated the performance of the SFs in both sets through the Pearson Correlation metric, obtaining similar results. Fig. 1 shows the evaluation in the validation set.

Fig. 1. Pearson Correlation of the 10 SFs in the validation set.

B. Machine Learning Techniques

Combining SFs can result in a highly correlated dataset. To tackle this aspect, we attempted to discover the set of most significant SFs to predict the free binding energy by applying four feature selection techniques: correlation analysis to remove highly correlated variables; generalized linear models with convex penalty functions as LASSO and Elastic Net, which perform embedded feature selection; and Recursive Feature Elimination (RFE) with resampling. Table I shows the correspondent SFs selected by each method.

TABLE I

SFs SELECTED APPLYING DIFFERENT FEATURE SELECTION METHODS

METHOD	SCORING FUNCTIONS SELECTED
None	All
Uncorrelated Variables	Autodock VINA, RFScore, NNScore, DSX, PELE, MM-Gbsa, rDock
LASSO	RFScore, PELE, NNScore
Elastic Net	RFScore, PELE, NNScore, XScore-HMScore
RFE with Resampling	RFScore, NNScore, PELE, DSX, rDock

With the resultant sets, we exploited the rationale that each SF brings something distinctive for each protein-ligand complex, in order to develop models based on the ensemble methodology such as AdaBoost, Gradient Boosting and Extra tree regressors. The main idea behind this methodology is to weight several individual models and combine them to obtain a new model that outperforms every one of them. We also performed other well-known machine learning techniques such as Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) regressors for comparison purposes. From the intelligibility perspective, we made an effort to obtain models easy to interpret by using a Generalized Additive Model (GAM) fitted with splines. An important aspect of this model is that it permits to visualize the relationship between the univariate terms of the GAM and the dependent variable, allowing to better understand the behavior of different scoring functions with respect to experimental binding affinity.

III. RESULTS AND DISCUSSION

In the context of regression and prediction, the performance of each model implemented with different selection methods is shown in Fig.2.

Fig. 2. Performance evaluation of the regressor models with different SFs selected according to the method used (see Table I). The performance metric is the Pearson Correlation.

The combination of different SFs has a substantial impact on the performance of the regressor models implemented and is an important step in order to improve the overall binding affinity. In the best scenario, all the models outperform the results of the individual SFs, from which K-NN and GAM stood out, obtaining a notable 0.84 and 0.82 Pearson correlation coefficients, respectively.

From the interpretability aspect, the smooth splines elements of the GAM with the SFs selected by the Elastic Net method are presented in Fig. 3.

Fig. 3. GAM predicted smooth splines of the Experimental binding affinity as a function of the scoring functions: XScore-HMScore, RFScore, PELE, NNScore. The degrees of freedom are in the parenthesis on the y-axis. The gray areas represent the 95% confidence intervals of the smooth splines. The thick marks in the x-axis indicate the distribution of the observations.

IV. Conclusions and Future Work

Heretofore, we have not only achieved promising results in the prediction of the binding free energy, but also we have obtained a clearer understanding on the behavior of the different SFs in individual and embedded manners. To further assess the predictive power and generalization of the developed models, we will test them using a core set based on the 2013 PDBbind benchmark. Furthermore, we attempt to add protein-ligand descriptors for uncovering additional patterns that might be crucial for the improvement of the protein-ligand binding affinity.

REFERENCES

[1] Gohlke, H. and Klebe, G., *Current opinion in structural biology*, *11*(2), pp.231-235, 2001.

[2] Ashtawy, Hossam M., and Nihar R. Mahapatra, *BMC bioinformatics* 16, no. Suppl 4, 2015.

[3] Arciniega, M. and Lange, O.F., *Journal of chemical information and modeling*, *54*(5), pp.1401-1411, 2014.

[4] J. Liu and R. Wang, Journal of chemical information and modeling, vol.55, no.3, pp. 475-482, 2015.

[5] R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake, and J. B. Matthew, Journal of Molecular Graphics and Modelling, vol. 20, no. 4, pp. 281–295, 2002.

[6] Chen, Y.C., Trends in pharmacological sciences. 78-95.

[7] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, and S. Hirono, Journal of chemical information and modeling, vol. 46, no. 1, pp. 380–391, 2006.

[8] A. E. Klon, M. Glick, and J. W. Davies,Journal of medicinal chemistry, vol. 47, pp. 4356–4359, 2004.

[9] M. Jacobsson, P. Lidén, E. Stjernschantz, H. Boström, and U. Norinder, Journal of medicinal chemistry, vol. 46, no. 26, pp. 5781–5789, 2003.

[10] Cheng, T., Li, X., Li, Y., Liu, Z. and Wang, R.c *Journal of chemical information and modeling*, *49*(4), pp.1079-1093, 2009.

Towards accurate solvation free energies of large biological systems

S. Romero,[a] F. J. Luque,[a] X. Barril,[a] F. Lipparini,[b] B. Mennucci,[c] C. Curutchet,[a]
[a]Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Spain
[b]Laboratoire Jacques-Louis Lions - Laboratoire de Chimie Théorique - Institut du Calcul et de la Simulation, Sorbonne Universiteś, Paris, France
[c]Dipartamento de Chimica e Chimica Industriale, Università di Pisa, Italy.
E-mail: sonia.romero@ub.edu

Abstract- *Continuum solvation models like PCM or COSMO are the standard tool to calculate solvation free energies in a quantum level, but have been typically limited to small biological molecules due to its large computational cost. Recently, a new implementation of COSMO based on a domain decomposition strategy (ddCOSMO) [1] has been presented, which speeds up calculations by several orders of magnitude, thus paving the way for its application to very large systems. Here, we report the parameterization of ddCOSMO to the prediction of hydration free energies based on the MST solvation model developed in Barcelona, [2][3]. The parameterization is based on the PM6 semi-empirical Hamiltonian, on a set of over 200 experimental hydration free energies. The new model opens the way to the accurate prediction of hydration free energies of very large biomolecules, thus going beyond the usual classical MM-PBSA or MM-GBSA approaches.*

Keywords: Implicit solvation models, MST Solvation Model, ddCOSMO.

I. INTRODUCTION

In continuum solvation models, the solute is treated at a QM or MM level, and the solvent is described as a continuum dielectric medium.

The solvation free energy is then computed as a sum of electrostatic , cavitation and Van der Waals free energies, where the last two are defined as non-electrostatic term.

$$\Delta G_{solv} = \Delta G_{ele} + \Delta G_{noele} \qquad (1)$$

Where the electrostatic term is defined as:

$$\Delta G_{ele} = \langle \Psi^{solv} | H^0 + {}^1/_2 V^{solv} | \Psi^{solv} \rangle - \langle \Psi^0 | H^0 | \Psi^0 \rangle \qquad (2)$$

There are different strategies to solve the electrostatic problem such as Generalized Born and Poisson-Boltzmann methods, which are used mostly in Molecular Mechanics implementations. In quantum mechanics implementations, Apparent surface charge methods such as Polarizable continuum model (PCM) and Conductor-like screening models (c-PCM or COSMO) are chosen. Nevertheless, its computational cost only allows using these models in small systems.

ddCOSMO is a recently proposed algorithm to solve the polarization equation for the Conductor-like Screening Model (COSMO, where the electrostatic solute-solvent interaction energy is obtained as:

$$E_{ele} = {}^1/_2 \, f(\varepsilon) \int_\Omega \rho(\mathbf{r}) \, W(\mathbf{r}) d\mathbf{r} \qquad (\mathbf{3})$$

Where f(ε) is an empirical scaling introduced to account for the non-conductor nature of the solvent and ε is its dielectric constant, ρ is the charge density of the solute and W is the polarization potential W of the conductor, usually referred to as the reaction field.

The ddCOSMO model[1] solves the COSMO equations based on Schwarz's domain decomposition method, and has been proven to be both smooth and fast; furthermore, linear scaling in both computational cost and memory requirements with respect to the system's size is implicit in the procedure without needing to resort to fast summation techniques. With respect to existing linear-scaling implementations, ddCOSMO can be two to three orders of magnitude faster, allowing computing the solvation energy for very large systems with a reduced computational cost.

In this project, we are re-parameterizing MST solvation model using ddCOSMO algorithm, at B3LYP and more recent PM6 semi-empirical level on a set of over 200 neutral molecules. The aim is to obtain free solvation energies at a quantum level even of large biological systems, in a cheaper and faster way.

II. COMPUTATIONAL DETAILS

The training set is of 238 neutral molecules with known experimental solvation free energies from Cramer and Truhlar data set [3]. All molecules have been optimized in gas phase and solution, (parameterizations are performed for both sets of geometries). The electrostatic free energy is easily computed using MST cavity settings and ddCOSMO method. The non-electrostatic term is isolated in equation (1) and computed using experimental solvation free energies. Then fitted using a multiple linear regression method.

$$\Delta G_{noele} = \sum_{i=1}^{N} \xi_i S_i \qquad (4)$$

The non-electrostatic free energy for each molecule will be obtained multiplying the atomic surface tensors (ξ_i) by the surface of each element/hybridization atom type.

III. RESULTS AND DISCUSSION

Our calculated solvation free energies are compared with the experimental ones, obtaining good results for both PM6 and B3LYP theory levels, being B3LYP slightly better. These results were independently of which geometry optimization phase we used.

Fig. 1. Comparison between experimental and calculated ΔG_{solv}. Parameterization of **a:** PM6 level, molecules optimized in gas phase. **b**: B3LYP level, molecules optimized in gas phase. **c:** PM6 level, optimization in solution. d: B3LYP level, optimization in solution

In the following table, are described the Mean Signed Error (MSE), Mean Unsigned Error (MUE) and Root Mean Squared Deviation (RMSD) of each parameterization.

	MSE	MUE	RMSD
PM6 gas	0,05	0,74	0,97
B3LYP gas	0,03	0,83	1,08
PM6 solv	0,04	0,86	1,14
B3LYP solv	0,03	0,83	1,07

Two different atom type sets were used in the parameterization: i) Element atom type (9), used in MST Model, define an atom type for each element (H, C, O, N,S , F, Cl, Br, P) and ii) Hybridization type(15), (H, Hp, Csp, Csp2, Csp3, Osp2, Osp3, Nsp, Nsp2, Nsp3,S , F, Cl, Br, P). This second atom type definition was definitely better than the element type.

IV. CONCLUSIONS AND FUTURE PERSPECTIVES

Both PM6 and B3LYP parameterizations are able to accurately describe the experimental hydration free energies of neutral molecules with errors below 1 kcal/mol.

Future work will extend the parameterization to charged molecules, based on an automatic rescaling of the cavity size in charged regions, following previous work in the context of the MST model.

ACKNOWLEDGMENT

C.C. and S.R. acknowledges support from the Ministerio de Economía y Competitividad of Spain (grants CTQ2012-6195 & RYC2011 – 08918)

REFERENCES

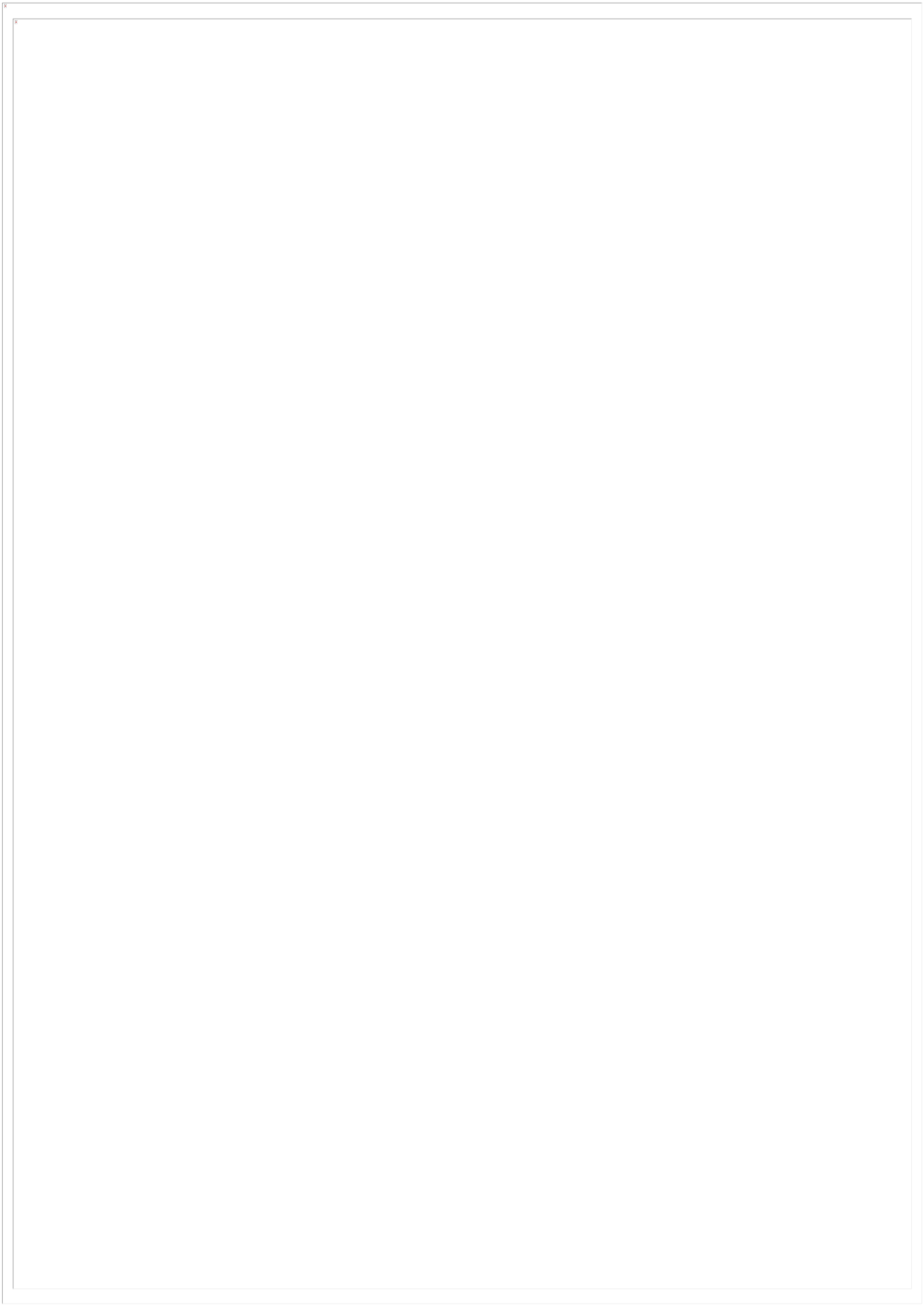
[1] M.Orozco,F.J.Luque, "Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems", Chem.Rev.100(2000), 4187.

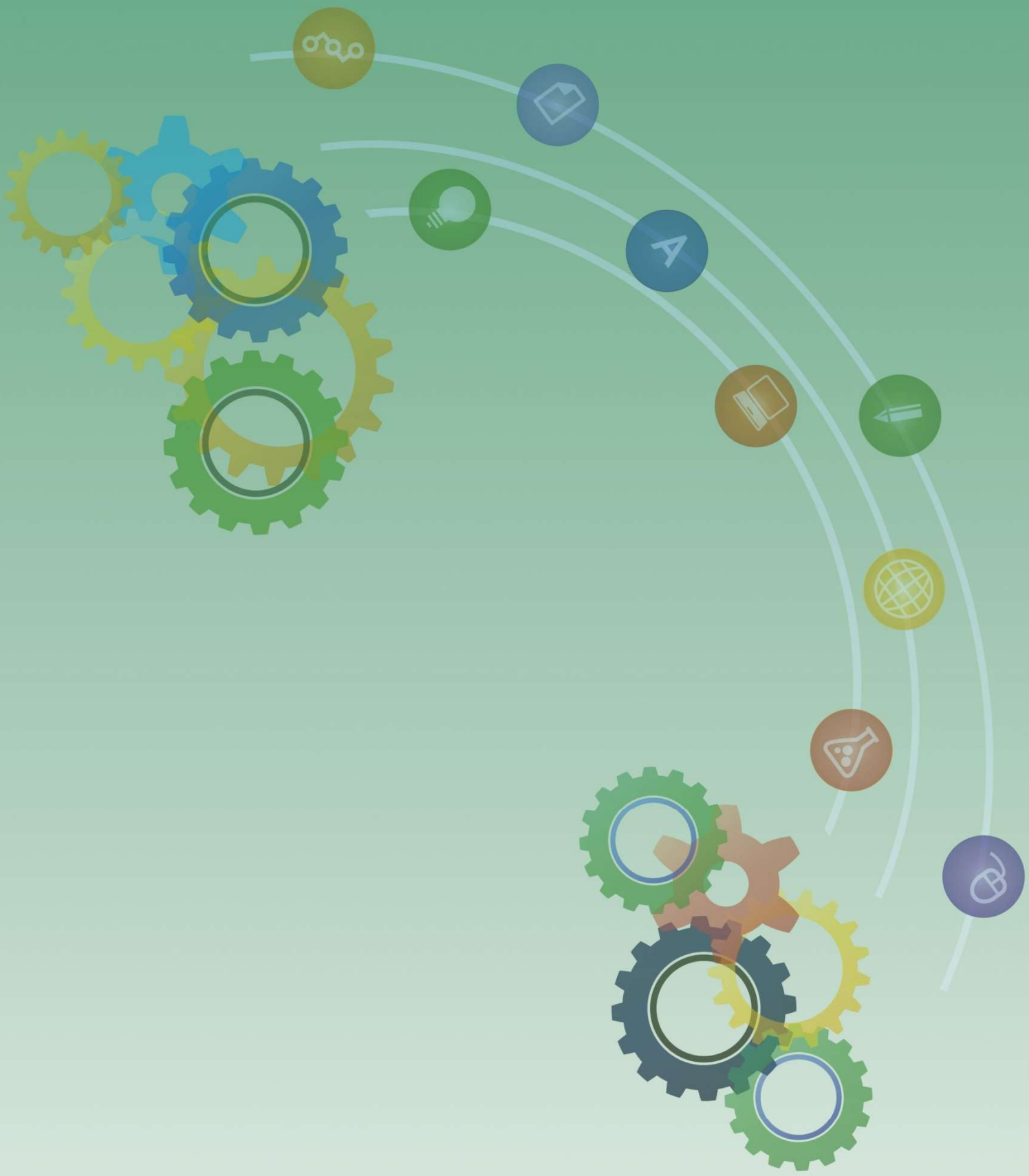
[2]F.Lipparini,L.Lagardère,G.Scalmani,B.Stamm,E.Cancès,Y.Maday,J.-P.Piquemal,M.J.Frisch,B.Mennucci, "Quantum Calculations in Solution for Large to Very Large Molecules: A New Linear Scaling QM/Continuum Approach" J.Phys.Chem.Lett.(2014)953.

[3]C.Curutchet,M.Orozco,F.J.Luque,J.Comput.Chem.22(2001), 1180;C.Curutchet,A.BidonChanal,I.Soteras,M.Orozco,F.J. Luque,J.Phys.Chem.B109(2005),3565.[4]A.V.Marenich,C. J.Cramer,D.G.Truhlar,J.Phys.Chem.B113(2009),6378

[4]A.V.Marenich,C.J.Cramer,D.G.Truhlar,J.Phys.Chem.B113(2009),6378.

POSTERS





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA