9th BSC Doctoral Symposium

10th, 11th and 12th May, 2022

Book of Abstracts

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Book of Abstracts
9th BSC Doctoral Symposium

Editor
Carolina Olmo

Cover
Design based on artwork created by macrovector.com

This book is available at
https://www.bsc.es/education/predoctoral-phd/doctoral-symposium

published by
Barcelona Supercomputing Center

9th Edition, May 2022

ACKNOWLEDGEMENTS

EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 9th BSC Doctoral Symposium.

During more than fifteen 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.

In this 9th edition of the BSC Doctoral Symposium we are offering a keynote talk titled "Competences as the key to success for strong technological developments" by Dr.-Ing. Bastian Koller and a tutorial on Effective Team Work and External Collaborations.

The talks will be held in six different sessions and will tackle the topics of: Life Sciences, Performance, Simulations and Modelling and Machine Learning and its applications and there will be three poster sessions.

This 9th edition of the BSC Doctoral Symposium recovers the face to face scenario giving visibility to our Phd student's research. This Book of Abstracts is the result of their contributions.

WELCOME ADDRESS

I am delighted to welcome all the PhD students, Postdoc researchers, advisors and experts to the 9th BSC Doctoral Symposium.

This 9th edition of the BSC Doctoral Symposium recovers face to face format, which we have all very much longed for. The goal of the occasion is to provide a framework to share research results of the projects developed by PhD thesis that use High Performance Computing in some degree. The symposium provides a forum for PhD students considering both the ones just beginning their research and others who have developed their research activities during several years.

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.

I would also like to thank all PhD students and Postdoc researchers for their papers and presentations. I wish you all the best for your career and I really hope you enjoy this great opportunity to meet other colleagues in some cases for the first time and share your experiences.

Last but not least, I wish to thank the Education and Training Team who put great effort and enthusiasm on the event.

Dr. Maria Ribera Sancho
Manager of BSC Education & Training

KEYNOTE SPEAKER

Bastian Koller
Managing Director of HLRS

Competences as the key to success for strong technological developments

Bastian Koller will present the idea behind the activity of building National Centres of Excellence in Europe and how it is being implemented. First successes will be presented and the embedding of the network of competence centres in the European HPC ecosystem will be shown. It will also become clear how important competences are in the European HPC strategy and how to interact with the (constantly growing) network at national and European level.

Dr.-Ing. Bastian Koller joined the High Performance Computing Center Stuttgart (HLRS) in 2004 after his diploma degree in computer science and received his PhD in engineering in 2011. Since 2014 he is Managing Director of HLRS. Dr. Koller was and is involved in many national and international research activities, covering for example HPC, HPDA, Cloud or AI topics. Currently, he is coordinating the EuroCC and CASTIEL EuroHPC actions, establishing a network of National Competence Centres in 33 nations.

TUTORIAL

"Effective Team Work and External Collaborations"

The purpose of the workshop is to deliver a tailor made training experience with content and strategies focused on the following:

- how to deal with a director and teammates,
- how to work as a part of a team and independently,
- how to work with people from other institutions,
- how to network and use the scientific community to support your own growth,
- how to deal with conflict.

 Moisès Martínez is the Biopolis World Founder. Creator of the Biopolis Methodology. Lifelong learner, his passion is to enjoy helping people, teams and organizations to grow in all their areas. He engages PhD's, Researchers, Entrepreneurs, Managers, Psychologists and Coaches trough Mentoring, Training, Supervisión, Coaching and Consulting processes on three continents..

PROGRAM

DAY 1 (May 10th)

Start time	Activity	Speaker/s	Chair
8.30h Registration			
9.00h	Welcome and opening	**Josep Mª Martorell,** BSC Associate Director	**Maria Ribera Sancho**
9.20h	Keynote talk: Competences as the key to success for strong technological developments	**Bastian Koller,** Mananging Director, HLRS	
	Abstract: I will present the idea behind the activity of building National Centres of Excellence in Europe and how it is being implemented. First successes will be presented and the embedding of the network of competence centres in the European HPC ecosystem will be shown. It will also become clear how important competences are in the European HPC strategy and how to interact with the (constantly growing) network at national and European level.		
10.30h Event Photo			
10.40h Coffee break & First Poster Session			
	A platform for antibody design, **Anna Maria Díaz-Rovira** (LS - Electronic and Atomic Protein Modelling)		
	Designing a Cloud and HPC Based M&S platform to Investigate the IVD Diseases Mechanism, **Maria Paola Ferri** (LS - INB)		
	Low Input Promoter Capture Hi-C Method Enables to Decipher the Molecular Mechanisms Underlying Genetically Complex Diseases, **Paula López-Martí** (LS - Computational Biology Life Sciences)		
	FAIRsoft - A practical implementation of FAIR principles for research software, **Eva Martín del Pico** (LS - INB)		
	CSDownscale: an R Package for Statistical Downscaling, **Jaume Ramon** (ES - Earth System Services)		
	JLOH: Extracting Loss of Heterozygosity Blocks from Short-Read Sequencing Data, **Matteo Schiavinato** (LS - Comparative Genomics)		
11.40h	**First Talk Session: Performance**		
11.40h	TACL: Interoperating asynchronous device APIs with task-based programming models, **David Álvarez** (CS - Programming Models)		
12.00h	PredIG: a predictor of T-cell immunogenicity, **Roc Farriol-Duran** (LS - Electronic and Atomic Protein Modelling)		**Víctor Guallar Tasies**
12.20h	Exploring Task Scheduling Sensitivity to L1 Cache configuration on a 32-core RISC-V Processor, **Lucas Morais** (CS - Computer Architecture for Parallel Paradigms)		
12.40h	Investigating Memory Prefetcher Performance over Parallel Applications: From Real to Simulated, **Valéria Soldera Girelli** (CS - Memory technologies)		
13.00h Lunch Break			
14.00h Tutorial 1st part			
	"Effective Team Work and External Collaborations"	**Moisès Martínez**	

The purpose of the workshop is to deliver a tailor made training experience with content and strategies focused on the following:

how to deal with a director and teammates,
how to work as a part of a team and independently,
how to work with people from other institutions,
how to network and use the scientific community to support your own growth,
how to deal with conflict.

17.00h Adjourn

DAY 2 (May 11th)

Start time	Activity	Speaker/s	Chair
9.00h	Opening of the second day		
9.10h	**Second Talk Session: Simulations and Modelling 1**		
9.10h	The development of high-order methods for finite elements to study turbulent reacting flow, **Antonio Blanco Casares** (CASE - Propulsion Technologies)		
9.30h	Implementation of cellular transport models within a multiscale simulation software, **Othmane Hayoun Mya** (LS - Computational Biology)		**Eduardo Pérez**
9.50h	Numerical Investigation of Unsteady Processes in Combustion using Flame Transfer Function, **Phivos Andreou Koumides** (CASE - Propulsion Technologies Group)		
10.10h	Non-Reacting Large Eddy Simulation of a Low Swirl Burner, **Fredherico Rodrigues** (CASE - Propulsion Technologies)		
10.30h	**Coffee break & Second Poster Session**		
	Distributed Tabulation of Flamelet Lookup Tables, **Nicholas Abel** (CASE - Propulsion Technologies)		
	Discovering the Ethereum2 P2P Network, **Mikel Cortes Goicoechea** (CS - Resilience in Distributed Systems)		
	Multi-model Forecast Quality Assessment of CMIP6 Decadal Predictions, **Carlos Delgdo-Torres** (ES - Earth System Services)		
	Ground-Level Ozone Simulations Improved by Updating Land Cover Databases, **Franco López** (ES - Atmospheric Composition)		
	Rational Enzyme Engineering of Different Active Sites on a Xylanase, **Rubén Muñoz Tafalla** (LS - Electronic and Atomic Protein Modelling)		
	Development of corrections for the absolute free binding energy prediction, **Ignasi Puch-Giner** (LS - Electronic and Atomic Protein Modelling)		
11.40h	**Third Talk Session: Performance 2**		
11.40h	Improving resource usage in large FPGA accelerators, **Antonio Filgueras** (CS - Programming Models)		
12h	CAMP First GPU Solver: A Solution to Accelerate Chemistry in Atmospheric Models, **Christian Guzman Ruiz** (ES - Computational Earth Sciences)		**Eduard**

12.20	b8c: an FPGA-Friendly Sparse Matrix Representation Suitable for the SpMV kernel, **José Oliver Segura** (CS - Programming Models)	**Ayguadé**
12.40	HBM, Present and Future of HPC based on FPGAs, **Elias Augusto Perdomo Hourné** (CS - European Exascale Accelerator)	

13.00h **Lunch Break**

14.00h **Tutorial session 2**

"Effective Team Work and External Collaborations"	**Moisès Martínez**

The purpose of the workshop is to deliver a tailor made training experience with content and strategies focused on the following:

how to deal with a director and teammates,
how to work as a part of a team and independently,
how to work with people from other institutions,
how to network and use the scientific community to support your own growth,
how to deal with conflict.

17.00h **Adjourn**

DAY 3 (May 12th)

Start time	Activity	Speaker/s	Chair
9.00h	Opening of the third day		

9.10h **Fourth Talk Session: Simulations and Modelling 2**

9.10h	Numerical Simulations for the Atomic Beam Probe, **Mátyás Aradi** (CASE - Fusion)	
9.30h	Nuclear Fusion Reactor Materials: Modelling Atomic-Scale Irradiation Damage in Metal, **Mary Kathleen Chessey** (CASE - Fusion)	
9.50h	Comparison of Five Strategies for Seasonal Prediction of Bioclimatic Indicators in the Olive Sector, **Chihchung Chou** (ES - Earth System Services Group)	**Dani Guallart**
10.10h	Adaptive points to estimate the Lebesgue constant on the simplex, **Albert Jiménez Ramos** (CASE - Geometry and Meshing for simulations)	
10.30h	Preparing plasma heating in ITER using integrated modelling, **Tomás Bensadon** (CASE - Fusion)	

10.50h **Coffee break & Third Poster Session**

Foreign and domestic contributions to surface ozone on Spain, **Roger Garatachea** (ES - Atmospheric Composition)	
Respiratory tract microbiome modifications after lung transplantation and its impact in CLAD, **Olfat Khannous** (LS - Comparative Genomics)	

Organic Particle Export, Remineralization and Advection in the North Atlantic mesopelagic layer, **M. Andrea Orihuela-García** (LS - Climate Prediction)

Characterisation of Ozone levels and associated NMVOC emissions in Spain: a preliminary assessment, **Kevin Oliveira** (ES - Atmospheric Composition)

Experimental Analysis on the NXP's T2080 Cache Coherence: A Step Towards MPSoCs in Critical Systems, **Roger Pujol Torramorell** (CS - Operating System / Computer ArchitectureInterface)

Space Shuttle: A Test Vehicle for the Reliability of the SkyWater 130nm PDK using OpenLane and the Google/E-fabless Shuttle Run, **Ivan Rodriguez-Ferrandez** (CS - Operating System / Computer Architecture Interface)

11.50h   Fifth Talk Session: Life Sciences

11.50h	Characterisation of inter-individual genomic variability across populations of different genetic ancestries, **Rubén Chazarra Gil** (LS - Transcriptomics and Functional Genomics Lab)	
12.10h	Transcriptome Analysis of Differential Gene Expression in Disease, **José Miguel Ramírez** (LS - Transcriptomics and Functional Genomics Lab)	Raquel García
12.30h	RosneT: A Block Tensor Algebra Library for Out-of-Core Quantum Computing Simulation, **Sergio Sánchez Ramírez** (CASE - Quantic)	
12.50h	Exploring the recent evolution of yeast pathogens using the CandidaMine database, **Miquel Àngel Schikora Tamarit** (LS - Comparative Genomics)	

13.10 Lunch

14.00h   Sixth Talk Session: Machine Learning and its applications

14.00h	Focus! Rating XAI Methods and Finding Biases, **Anna Arias Duart** (CS - High Performance Artificial Intelligence)	
14.20h	Reaching a modular, generic and containerized development in Biomedical Natural Language Processing systems, **Javier Corvi** (LS - INB)	Daniele Lezzi
14.40h	QUARQ QUick Approximate and Relaxed Querying, **Victor-Alejandro Ortiz** (CASE - Smart Cities)	
15.00h	Hyperparameter Optimization using Agents for Large Scale Machine Learning, **Pere Vergés Boncompte** (CS - Workflows and Distributed Computing)	

15.20h   Final remarks and End of Doctoral Symposium

Abstracts

Distributed Tabulation of Flamelet Lookup Tables

Nicholas Abel[*][†], Ricard Borrell Pol[*], Daniel Mira Martinez[*]

[*]Barcelona Supercomputing Center, Barcelona, Spain

[†]Universitat Politècnica de Catalunya, Barcelona, Spain

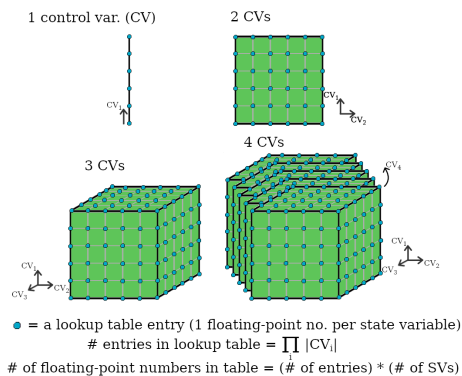E-mail: {nicholas.abel, ricard.borrell, daniel.mira}@bsc.es

Fig. 1. The discretization of the flamelet lookup table can be visualized on the p-dimensional unit hypercube, where p is the number of controlling variables. If there are N discretization points of each controlling variable, the total number of floating point numbers stored in the lookup table is SN^p where S is the number of state variables, or arguments to the function being retrieved.

Keywords—combustion, tabulated chemistry, mpi3-rma, High-Performance computing.

I. Extended Abstract

One of the fundamental questions in combustion simulation is how to account for detailed chemistry effects, while controlling both the error of the chemical scheme and the computational cost. Combustion chemistry is important for resolving processes such as flame propagation and pollutant formation, which are non-linear processes that can be computationally expensive. The direct solution of the governing equations of turbulent reacting flows can be prohibitively expensive as the chemical integration is often stiff. Tabulated chemistry methods with flamelet modelling emerge as an alternative to perform direct integration of the chemical source terms and has been extended to a wide range of conditions [1].

In flamelet methods, the chemical time scale is assumed smaller than the time scales of the turbulence, so the flame structure is not affected by the turbulence. In flamelet methods, the thermochemical states of the flame are computed in a pre-processing step, and these values are retrieved from a lookup table loaded into memory at the beginning of the simulation. The flame structure can be recovered through the use of controlling variables, which represent dimensions along the multidimensional space of the flame manifold.

A restriction of the tabulated chemistry method is that it is bound by memory. Increasing the number of dimensions, the number of tabulated values per dimension, and the number of discretization points on the flamelet manifold as desired
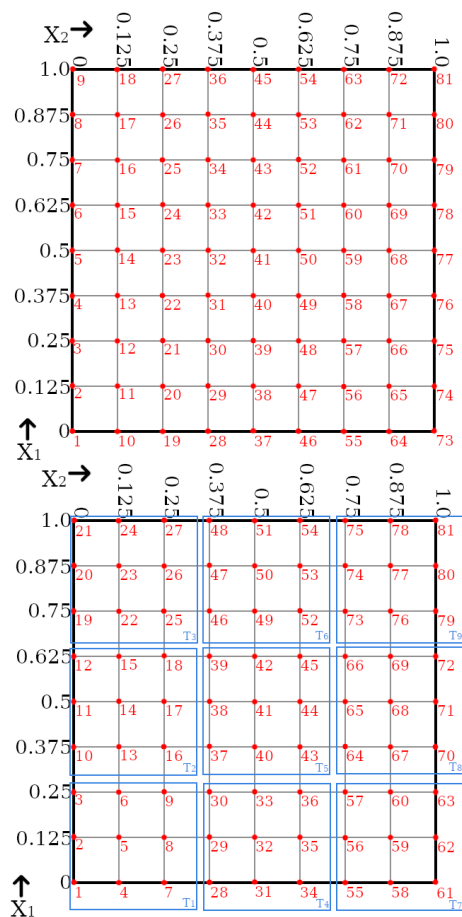

Fig. 2. An example of block structuring applied to a small 2D lookup table, with controlling variables labelled X_1 and X_2. Numbers 1-81 give the ordering of entries as they appear in memory before and after applying the block structure. Here, the table size is 9×9 and the block size is 3×3. The blocks are labelled T_1 to T_9.

are not possible given the way that tabulated chemistry is currently implemented in combustion codes [2]. In particular, the entirety of the lookup table is stored on each core on distributed-memory machines. Modern supercomputers typically have about 2GB of memory on each core, and tables quickly outgrow this limit as they increase in dimension (see Figure 1).

Since combustion simulations are typically run in parallel across many compute cores, the memory-boundedness of the current implementation may be addressed by storing the lookup table in a distributed manner, as opposed to keeping a

D_T ($D_R = 50$)	1	2	5	10	20	25	50	100	250	500
Wall Time (s)	1.3858	1.2581	1.2532	1.2344	1.1940	1.2697	1.2209	1.1985	1.2733	1.2264
Speedup	1.0000	1.1014	1.1057	1.1227	1.1606	1.0914	1.1350	1.1563	1.0884	1.1299
D_T ($D_R = 100$)	1	2	5	10	20	25	50	100	250	500
Wall Time (s)	1.8338	1.6665	1.4573	1.9551	1.6779	1.4743	1.6022	1.4384	1.4908	1.6966
Speedup	1.0000	1.1004	1.2584	0.9380	1.0929	1.2439	1.1445	1.2749	1.2301	1.0808

TABLE I.   WALL TIME (IN SECONDS) AND SPEEDUP FOR 10^8 RANDOM ACCESSES TO THE TABLE, FOR ACCESS RANGES $D_R = 50, 100$ AS BLOCK SIZE, GIVEN BY DT, VARIES.

copy of the entire table in memory at each core. We refer to this method as **distributed tabulation**. The goal of our project is to develop, implement, and test distributed tabulation in the multiphysics code Alya [3].

A. Methodology

The most naive approach to distributed tabulation is as follows. Consider a lookup table that contains N_E total entries, used in a simulation run on N_R total compute cores. Then, at the beginning of runtime, we load N_E/N_R table entries into memory onto the each core (assume for simplicity N_E is divisible by N_R). Then, using MPI 3's remote memory access capabilities [4], we open a memory window on each rank's table entries to permit one-sided access during the solve phase of the simulation. If a sub-domain needs an entry which is stored on another sub-domain, it will retrieve that entry using a MPI Get call. This technique alone will solve the problem of memory-boundedness given enough parallel resources, but in the worst case it will incur a significant cost in overhead due to communication.

To reduce communication, we exploit a pattern of locality which tends to occur in flamelet table lookups. When a sub-domain requires a lookup table entry, successive entries requested by that sub-domain will tend to be spatially near previous entries in the flamelet manifold. To exploit this, we organize the table in a block structure and keep recent remotely-retrieved entries in a small local pile.

The block structure is a partitioning over the unit hypercube. Table entries which are spatially collocated on the unit hypercube are rearranged to be contiguous in memory. An example of block-structure reorganization on a table with 2 controlling variables is illustrated in Figure 2.

Then, when a sub-domain requests a non-local entry, it will use MPI remote memory access to retrieve all entries on its block and store the block locally on a pile. When a sub-domain needs a non-local entry, it looks first in the local pile to see if the entry is available there. The local pile is restrained by a maximum number of stored blocks. When a block is added to the pile or used locally, it is moved to the front. As a result, blocks that are not used for some time are removed from local memory.

B. Test on locality optimization

Distributed tabulation reorganizes lookup tables such that entries of the tables which are spatially co-located on small subsets of the unit hypercube, called blocks, are stored contiguously in memory. Combustion simulations tend to successively request lookup table values which are close to one another in

space, but current lookup table strategies do not necessarily store spatially close values near one another in memory. While block structuring of lookup tables was devised with message passing in mind, it may also increase CPU cache efficiency and decrease the total time needed to retrieve local lookup table values. To test this effect, we run the following test on a single MPI rank.

Using a three-dimensional lookup table of size 1000×1000×1000, we choose a range R on the table with dimensions $D_R \times D_R \times D_R$, for $D_R = \{50, 100\}$, in which we access 10^8 randomly chosen elements. For each size of D_R, the size of the blocks T on the cube are chosen as $D_T \times D_T \times D_T$, for $D_T = \{1, 2, 5, 10, 20, 25, 50, 100, 250, 500\}$. Table I shows the wall-time (in seconds) and speedup comparing 10^8 accesses on R for $D_R = 50$ and $D_R = 100$, as the block size T defined by D_T varies. This demonstrates that the block structure can yield modest sequential speedup for repeated access of spatially close entries on the table, a side-effect of a functionality intended for point-to-point communication.

REFERENCES

[1] J. Benajes, J. García-Oliver, J. Pastor, I. Olmeda, A. Both, and D. Mira, "Analysis of local extinction of a n-heptane spray flame using large-eddy simulation with tabulated chemistry," *Combustion and Flame*, vol. 235, p. 111730, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010218021004739

[2] E. Illana, D. Mira, and A. Mura, "An extended flame index partitioning for partially premixed combustion," *Combustion Theory and Modelling*, vol. 25, no. 1, pp. 121–157, 2021. [Online]. Available: https://doi.org/10.1080/13647830.2020.1841912

[3] M. Vázquez, G. Houzeaux, S. Koric, A. Artigues, J. Aguado-Sierra, R. Arís, D. Mira, H. Calmet, F. Cucchietti, H. Owen, A. Taha, E. D. Burness, J. M. Cela, and M. Valero, "Alya: Multiphysics engineering simulation toward exascale," *Journal of Computational Science*, vol. 14, pp. 15–27, 2016, the Route to Exascale: Novel Mathematical Methods, Scalable Algorithms and Computational Science Skills. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877750315300521

[4] T. Hoefler, J. Dinan, R. Thakur, B. Barrett, P. Balaji, W. Gropp, and K. Underwood, "Remote memory access programming in mpi-3," *ACM Trans. Parallel Comput.*, vol. 2, no. 2, jun 2015. [Online]. Available: https://doi.org/10.1145/2780584

Nicholas Abel received his BSc degree in Mathematics from The University of New Mexico (UNM) in 2017, and completed his MSc degree in Mathematics from UNM in 2020. He currently is a member of the Computer Applications in Science and Engineering (CASE) group at the Barcelona Supercomputing Center (BSC), as well as a PhD student in the School of Mathematics and Statistics at Universitat Politècnica de Catalunya (UPC).

TACL: Interoperating asynchronous device APIs with task-based programming models

David Álvarez*†, Kevin Sala*†, Vicenç Beltran*

*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {david.alvarez,kevin.sala,vbeltran}@bsc.es

Keywords—Task-based Programming Models, OmpSS-2, Heterogeneous Computing, High-Performance Computing

I. EXTENDED ABSTRACT

Heterogeneous architectures have become commonplace in modern HPC systems. Eight of the world's top ten supercomputers have accelerators, and the up-and-coming MareNostrum 5 will feature accelerated partitions. However, programming these heterogeneous systems is difficult, as users have to insert data transfer operations, kernel launches and synchronizations manually from the host system to its accelerators. This is even more challenging in distributed heterogeneous systems, as programmers have to coordinate the previous activities with inter-node communications between hosts. This work presents the Task-Aware Ascend Computing Language (TACL), which interoperates with the OmpSs-2 programming model and greatly simplifies kernel execution, data transfers and synchronizations between host and accelerators by naturally leveraging the data-flow execution model of OmpSs-2.

A. OmpSs-2

OmpSs-2 [1] is a task-based parallel programming model developed by the System Tools and Advanced Runtimes (STAR) research group at BSC. OmpSs-2 programs are composed of C, C++ or Fortran code annotated with compiler directives, which can express complex computational patterns using fine-grained data dependencies between tasks.

This programming model has limited support for heterogeneous systems, as implementing device tasks requires extensive changes in the OmpSs-2 compiler and runtime. The latest release of OmpSs-2 has support for CUDA tasks only when used in conjunction with Unified Memory.

B. AscendCL

The Ascend Computing Language [2] encompasses a collection of APIs for users to leverage and operate Huawei's Ascend AI Processors [3]. The Ascend AI lineup is based on Huawei's proprietary Da Vinci architecture, intended initially to accelerate neural network training and inference workloads. However, it is possible to write custom kernels to execute directly on the accelerator hardware and thus repurpose the architecture for HPC workloads. Moreover, the AscendCL library exposes a small subset of BLAS subroutines, such as `gemm` and `gemv`, for matrix multiplication.

C. TACL

TACL is a wrapper around the AscendCL library that implements two key features to allow the interoperability of tasks and the accelerator APIs: (1) TACL transparently handles a pool of ACL streams for tasks to re-use intelligently and (2) TACL can bind the release of a task's dependencies to the completion of all operations in an ACL stream in a non-blocking manner. With these two capabilities, programmers can call ACL functions from their tasks without blocking host CPUs in synchronization functions or risking deadlocks from inter-device communication functions. The idea to deliver non-blocking synchronization by delaying the tasks' dependency release was initially pioneered by TAMPI [4], [5] for MPI communications. In this work, we explore its applicability in heterogeneous computing.

We will use the example in Listing 1 to illustrate the use of TACL in task-based applications. In this example, a producer task runs in the host, followed by an offloading step, which will copy the relevant data to the accelerator, execute a kernel, and then copy the data back to the host. This is a standard workflow for most heterogeneous HPC programs.

However, this example features two problems. First, a new device stream is created and destroyed for every offloading task. This is done because re-using the same stream for all tasks would serialize the execution, but creating streams is generally an expensive operation, which should be avoided. Moreover, in line 13, the execution of this task is suspended until all the operations pending on `stream` finish. However, in this case, the thread executing the offloading task will be blocked in this operation, while it could be busy executing other tasks. Therefore, we are wasting resources on blocking synchronizations.

We can modify the code as showcased on Listing 2 to solve the problems mentioned above. We substituted the highlighted functions with calls to the TACL library. Substitution is one-to-one and existing code can be easily adapted. After these changes, the two previous problems no longer exist. Firstly, TACL manages a pool of re-usable streams, amortizing stream creation costs. As such, `taclrtGetStream` is a much faster operation than creating a new stream from scratch. Secondly, the call to `taclrtSynchronizeStreamAsync` is a non-blocking version of the blocking `aclrtSynchronizeStream` call. The TACL call returns immediately without blocking, and the task can proceed and return. However, the dependencies of

```
1 #pragma oss task out (A[0;size]) label("producer")
2 produceDataHost(A, size);
3
4 #pragma oss task inout(A[0;size]) label("ascend offload")
5 {
6   aclrtStream stream;
7   aclrtCreateStream(&stream);
8
9   aclrtMemcpyAsync(devA, A, ..., ACL_MEMCPY_HOST_TO_DEVICE,
      stream);
10  aclopExecute(devA, ..., stream);
11  aclrtMemcpyAsync(A, devA, ..., ACL_MEMCPY_DEVICE_TO_HOST,
      stream);
12
13  aclrtSynchronizeStream(stream);
14  aclrtDestroyStream(stream);
15 }
16
17 #pragma oss task in(A[0;size]) label("consumer")
18 consumeDataHost(A, size);
```

Listing 1. Pseudocode of an application offloading computation to an Ascend accelerator

```
1 #pragma oss task out (A[0;size]) label("producer")
2 produceDataHost(A, size);
3
4 #pragma oss task inout(A[0;size]) label("ascend offload")
5 {
6   aclrtStream stream;
7   taclrtGetStream(&stream);
8
9   aclrtMemcpyAsync(devA, A, ..., ACL_MEMCPY_HOST_TO_DEVICE,
      stream);
10  aclopExecute(devA, ..., stream);
11  aclrtMemcpyAsync(A, devA, ..., ACL_MEMCPY_DEVICE_TO_HOST,
      stream);
12
13  taclrtSynchronizeStreamAsync(stream);
14  taclrtReturnStream(stream);
15 }
16
17 #pragma oss task in(A[0;size]) label("consumer")
18 consumeDataHost(A, size);
```

Listing 2. Code from Listing 1 adapted with TACL

the offloaded task (the `inout(A[0;size])`) will not be released until all pending operations in the stream finish. This allows the OmpSs-2 runtime to re-use the thread to execute other tasks instead of wasting resources waiting for blocking operations.

D. TACL Architecture

The architecture of the TACL library has two main components: the stream pool, managing device streams, and the TACL polling service, as illustrated in Figure 1.

When a program using TACL asks for a stream, one is returned from the pool. Then, upon calling to `taclrtSynchronizeStreamAsync`, the library will use the OmpSs-2 external event API to block task dependency release until all outstanding operations are completed. An event is registered on the device stream to check for the completion of enqueued operations. This event will be completed once all previous stream operations have finished, and thus it enables TACL to check their status. Both the device event and the OmpSs-2 task handle are stored inside the library.

The polling service component periodically checks the completion of events. In OmpSs-2, it is possible to have a recurrent task that has a configurable deadline. Each time the deadline passes, the task will be scheduled at the next possible opportunity. This is how the polling service is implemented
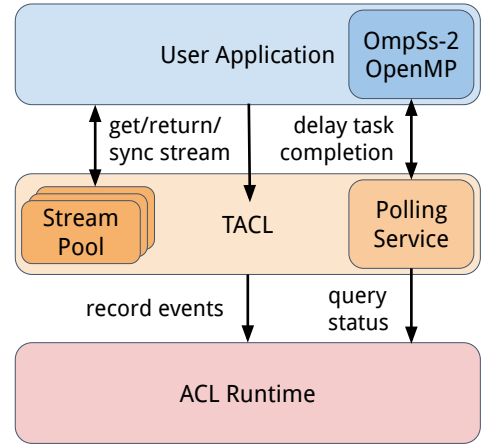

Fig. 1. Architectural diagram of TACL

in TACL. Internally, this polling service queries the status of every outstanding device event. Once an event is completed, TACL uses the OmpSs-2 external event API to notify that the task that initially requested synchronization should no longer wait for its offloaded operations to release its dependencies.

E. Conclusions and Future Work

We have shown how TACL can be leveraged to prevent common problems when writing heterogeneous programs using task-based programming models. Future work is centered around porting existing accelerated applications such as HPL-AI to TACL and evaluating their performance.

II. ACKNOWLEDGMENT

REFERENCES

[1] BSC STAR Research Group, "Ompss-2 specification." [Online]. Available: https://pm.bsc.es/ftp/ompss-2/doc/spec

[2] Huawei, "AscendCL Overview." [Online]. Available: https://support.huawei.com/enterprise/en/doc/EDOC1100155021/d63e3d89/ascendcl-overview

[3] ——, "Ascend 910 AI Processor." [Online]. Available: https://e.huawei.com/en/products/cloud-computing-dc/atlas/ascend-910

[4] K. Sala *et al.*, "Improving the interoperability between mpi and task-based programming models," in *Proceedings of the 25th European MPI Users' Group Meeting*, ser. EuroMPI'18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3236367.3236382

[5] K. Sala, S. Macià, and V. Beltran, "Combining one-sided communications with task-based programming models," in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, 2021, pp. 528–541.

David Álvarez received his BSc degree in Informatics Engineering from UPC in 2019 and his MSc degree in Research and Innovation in Informatics at UPC in 2021. Since 2019, he has been with the System Tools and Advanced Runtimes (STAR) group of Barcelona Supercomputing Center (BSC). He is currently a PhD Student and an Associate Part-time Professor in the Computer Architecture Department at UPC.

Numerical Simulations for the Atomic Beam Probe

Mátyás Aradi*, Mervi Mantsinen*†, Shimpei Futatani‡

*Barcelona Supercomputing Center, Barcelona, Spain

matyas.aradi@bsc.es, mervi.mantsinen@bsc.es

†ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

‡Universitat Politècnica de Catalunya, Barcelona, Spain

shimpei.futatani@upc.edu

Keywords—fusion, plasma, tokamak, COMPASS, plasma diagnostic, Atomic Beam Probe, CUDA.

EXTENDED ABSTRACT

Fusion plasmas are complex systems with several physical parameters which need continuous control and adjustment. Most plasma diagnostics have limitations of applications regarding the spatial, temporal resolution with typically high relative error. A typical physical parameter is combined from the output of several diagnostics to reduce these limitations.

The *Atomic Beam Probe (ABP)* is a novel diagnostic technique, and in our recent work, we are working on supportive numerical modelling procedures. Our tool has already been supporting diagnostic and scenario design, and we would like to give a hint about possible publication opportunities.

A. Atomic Beam Probe

The Atomic Beam Probe is a novel fusion plasma diagnostic tool [1], [2] installed on COMPASS tokamak [3], [4].
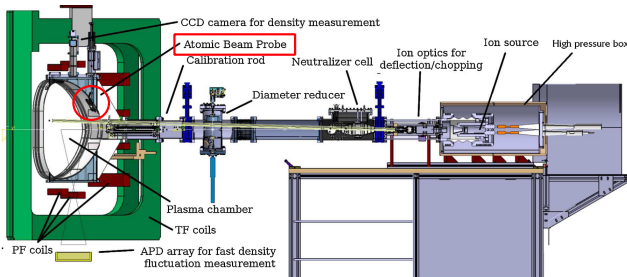

Fig. 1. Atomic Beam Probe conceptual design [2]

A light neutral alkaline atomic beam, typically a 50–100 keV lithium [4] or sodium [5] beam, is injected into the plasma through the middle plane port [6]. The beam particles are ionised in the plasma [2], close to the last close flux surface. The ions interact with the magnetic field, and their trajectories are diverted from the injection axis. The alkaline ions are collected by a Faraday cup detector matrix [3], [4], [5] installed on the upper port.

The detector is adjustable, both the angle and the toroidal position. In the most optimal scenarios, the detector coordinates and the resolution of the segments are defined by the result of the simulations.

The detector receives a high temporal resolution signal of the alkaline ion current, which maps the fast magnetic field changes in the plasma.

The measurements of the high temporal resolution magnetic field are limited in fusion plasmas. However, they are crucial in the understanding of some critical phenomena, such as edge localised modes (ELMs).

B. GPGPU Trajectory simulation

The feasibility studies, the detector design, the deployment and the application required a trajectory simulator. In the early stage of the study, a sequential Matlab code (*ABPIons*) [1], [2] supported the development. Nevertheless, the limitation of the previous solution was a severe lack of performance. Thereby, as a greenfield project, a new GPU-based simulator was written in CUDA C: the *Trajectory simulator of ABP Ions with GPU Acceleration (TAIGA)* [3].

The focus was on the application ability, even on older graphics cards, in the architecture design. Consequently, we are using SDK 2.1 which supports all cards from NVidia Fermi architecture (GeForce 450+). The numerical solver is atomistic. Each ion is modelled as a non-interactive particle, and their trajectories are calculated, in parallel. Thereby we can model the trajectories of millions of particles in less than a minute, even on older cards [3].
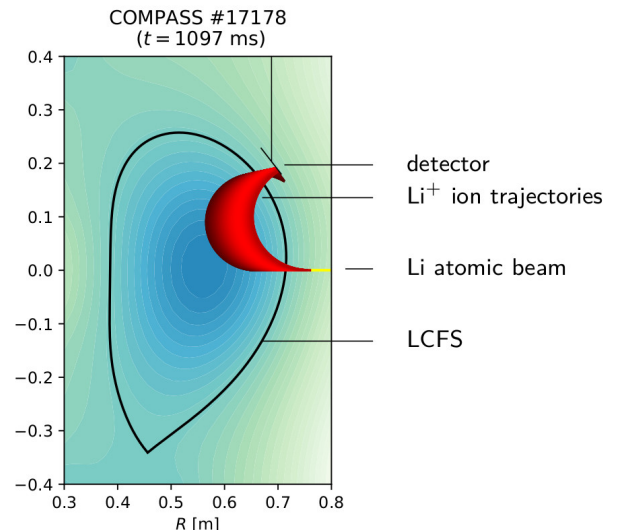

Fig. 2. GPGPU trajectory simulation for beam particles

With the new trajectory simulator, we can map a higher range of parameters, such as the influence of local magnetic perturbations.

We implemented several major improvements since the latest publication [3] such as numerical solvers, electric field,

perturbative magnetic field integration. The improvements also include modern code development solutions: continuous integration and a strict testing environment. We deliver the new functionalities in a test-driven development framework. For scientific accuracy, both CPU, GPU or transfer modules are tested directly or mocked, in multilayer automatised testing.

C. Synthetic diagnostic

Synthetic diagnostics are popular tools in fusion plasma physics lately and bring us closer to the fusion flight simulator.

A synthetic diagnostic is a framework that models and combines physical models which influence the measurement. Additionally, the framework mimics the physics of the diagnostic tool and provides an equivalent synthetic detector signal [7]. Thus the measurement and the synthetic data are comparable. If we had all physical processes taken into account, the signals would be equal, and we could predict future scenarios.
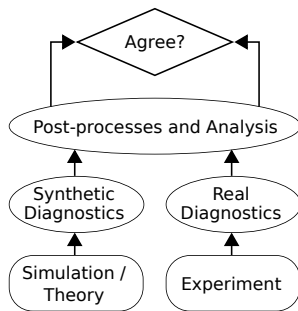
Fig. 3. Definition of synthetic diagnostic [7]

For the Atomic Beam Probe, we are implementing a synthetic diagnostic, combining the trajectory simulator, atomic physics libraries (such as RENATE-OD [8]) and other codes.

Explaining the details of the trajectory solver and the synthetic diagnostic exceeds the limits of this paper. We are planning to publish them in a peer-reviewed journal in due course.

D. Conclusion

A new plasma diagnostic, the Atomic Beam Probe was installed on the COMPASS tokamak. We mentioned the GPGPU trajectory simulator, which supports the measurements. We presented the abilities of the simulator and highlighted the direction of the development. We highlighted the subject of publication opportunities: the trajectory solver and the synthetic diagnostic.

E. Acknowledgement

REFERENCES

[1] M. Berta, A. Bencze, G. Anda, M. Aradi, D. Dunai, G. Veres, and S. Zoletnik, "Concept of an atomic beam probe diagnostic on compass tokamak," in *36th EPS Conference on Plasma Physics*, vol. 33E, 2009, P5.139.

[2] M. Berta, G. Anda, M. Aradi, A. Bencze, C. Buday, I. Kiss, S. Tulipán, G. Veres, S. Zoletnik, J. Havlícek, and P. Háček, "Development of atomic beam probe for tokamaks," *Fusion Engineering and Design*, vol. 88, 11, 2875–2880, 2013, doi:10.1016/j.fusengdes.2013.05.064.

[3] P. Hacek, M. Berta, G. Anda, M. Aradi, A. Bencze, D. Dunai, J. Krbec, R. Panek, D. I. Refy, J. Stockel, V. Weinzettl, and S. Zoletnik, "Development of an ion beam detector for the atomic beam probe diagnostic," *Review of Scientific Instruments*, vol. 89, 11, 113506, 2018, doi:10.1063/1.5044529.

[4] D. I. Réfy, S. Zoletnik, D. Dunai, G. Anda, M. Lampert, S. Hegedűs, D. Nagy, M. Palánkai, J. Kádi, B. Leskó, M. Aradi, P. Hacek, and V. Weinzettl, "Micro-faraday cup matrix detector for ion beam measurements in fusion plasmas," *Review of Scientific Instruments*, vol. 90, 3, 033501, 2019, doi:10.1063/1.5084219.

[5] D. Réfy, P. Hacek, S. Zoletnik, M. Aradi, D. Dunai, G. Anda, J. Krbec, E. Skáre, E. Glocker, the COMPASS Team, and the EUROfusion MST1 team, "Plasma edge current fluctuation measurements with the atomic beam probe diagnostic at compass," in *47th EPS Conference on Plasma Physics*, vol. 45A, 2021, P5.1008.

[6] S. Zoletnik, G. Anda, M. Aradi, O. Asztalos, S. Bató, A. Bencze, M. Berta, G. Demeter, D. Dunai, P. Hacek, S. Hegedűs, G. H. Hu, T. Krizsanóczi, M. Lampert, D. Nagy, J. Németh, M. Otte, G. Petravich, G. I. Pokol, D. Réfy, B. Tál, and M. Vécsei, "Advanced neutral alkali beam diagnostics for applications in fusion research (invited)," *Review of Scientific Instruments*, vol. 89, 10, 10D107, 2018, doi:10.1063/1.5039309.

[7] L. Shi, E. J. Valeo, B. J. Tobias, G. J. Kramer, L. Hausammann, W. M. Tang, and M. Chen, "Synthetic diagnostics platform for fusion plasmas (invited)," *Review of Scientific Instruments*, vol. 87, 11, 11D303, 2016, doi:10.1063/1.4961553.

[8] O. Asztalos, B. Szondy, K. Tőkési, and G. Pokol, "The modeling of atom - neutral collisions for beam emission spectroscopy applications," *The European Physical Journal D*, vol. 73, 6, 116, 2019, doi:10.1140/epjd/e2019-90690-2.

Mátyás Aradi is a second-year PhD student at the Barcelona Supercomputing Center and the Universitat Politècnica de Catalunya.

He joined the fusion community as a secondary school student in 2006 and participated in the Atomic Beam Probe modelling.

He graduated as a physicist, received his BSc degree at the Budapest University of Technology and Economics with a nuclear physics specialisation.

He received his MSc degree with computational physics specialisation at Eötvös University, Budapest, Hungary. He wrote his theses related to the Atomic Beam Probe.

Thereafter, he worked for 1.5 years at the Budapest University of Technology and Economics. He implemented and integrated a runaway electron fluid module into the European Integrated Modelling framework.

As a next project, he worked as a university assistant at Graz University of Technology, Austria. He was teaching numerical methods and Matlab practical courses. In his research, he focused on neoclassical modelling.

He spent a year as a software developer at Sigma Technology in an Ericsson telecommunication project. His team implemented 5G mobile positioning into the market-leading software. He gained both modern software development and project management experiences.

He joined the BSC Fusion Group in September 2020. He has been working on the Atomic Beam Probe synthetic diagnostic in an international cooperation.

Focus! Rating XAI Methods and Finding Biases

Anna Arias Duart*†, Férran Parés*, Dario Garcia-Gasulla*

*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {anna.ariasduart, ferran.pares, dario.garcia}@bsc.es

I. EXTENDED ABSTRACT

Explainability has become a major topic of research in Artificial Intelligence (AI), aimed at increasing trust in models such as Deep Learning (DL) networks. However, trustworthy models cannot be achieved with explainable AI (XAI) methods unless the XAI methods themselves can be trusted.

To evaluate XAI methods one may assess interpretability, a *qualitative* measure of how understandable an explanation is to humans [1]. While this is important to guarantee the proper interaction between humans and the model, interpretability generally involves end-users in the process [2], inducing strong biases. In fact, a qualitative evaluation alone cannot guarantee coherency to reality (*i.e.,* model behavior), as false explanations can be more interpretable than accurate ones. To enable trust on XAI methods, we also need *quantitative* and objective evaluation metrics, which validate the relation between the explanations produced by the XAI method and the behavior of the trained model under assessment.

In this work we propose a novel evaluation score for *feature attribution* methods, described in §I-A. Our input alteration approach induces in-distribution noise into samples, that is, alterations on the input which correspond to visual patterns found within the original data distribution. To do so we modify the context of the sample instead of the content, leaving the original pixels values untouched. In practice, we create a new sample, composed of samples of different classes, which we call a *mosaic image* (see examples in Figure 2). Using *mosaics* as input has a major benefit: each input quadrant is an image from the original distribution, producing blobs of activations in each quadrant which are consequently coherent. Only the pixels forming the borders between images, and the few corresponding activations, may be considered out of distribution.

By inducing in-distribution noise, *mosaic images* introduce a problem in which XAI methods may objectively err (focus on something it should not be focusing on). On those composed mosaics we ask a XAI method to provide explanation for just one of the contained classes, and follow its response. Then, we measure how much of the explanation generated by the XAI is located on the areas corresponding to the target class, quantifying it through the *Focus* score. This score allows us to compare methods in terms of explanation precision, evaluating the capability of XAI methods to provide explanations related to the requested class. Using *mosaics* has another benefit. Since the noise introduced is in-distribution, the explanation errors identify and exemplify biases of the model. This facilitates the elimination of biases in models and datasets, potentially resulting in more reliable solutions. We illustrate how to do so in §I-C.

A. The Focus metric

When a *feature attribution* method is applied to an image to explain the model's prediction regarding a chosen class, it typically produces a map from pixels to real values, referred to as relevance. To formalize mosaics, and later *Focus*, let us define a dataset D composed by a set of images $I = \{img_1, img_2, ..., img_N\}$ and a set of classes $C = \{c_1, c_2, ..., c_K\}$, where N is the number of total images and K is the number of total classes. Every image in I has assigned a unique class from C: $c(img)$. From here we build a set of mosaics $M = \{m_1, m_2, ..., m_J\}$ where J is the total number of mosaics in M. A mosaic m is composed by four images $m = \{img_1, img_2, img_3, img_4\}$ and characterized by a target class $tc = c(m)$, the specific class the XAI method is expected to explain. While two images of the mosaic belong to the target class $c(img_1) = c(img_2) = c(m)$, the other two are randomly chosen among the rest of classes $c(img_3) \neq c(m); c(img_4) \neq c(m)$. Mosaics are implemented as two by two, non-overlapping grid, with the position of each image being random.

The *Focus* metric estimates the reliability of XAI method's output as the probability of the sampled pixels lying on an image of the target class of the mosaic $c(m)$. This is equivalent to the proportion of positive relevance lying on those images:

$$F_{\mathcal{A},\theta}(m) = \frac{R_{c(m)}(img_1) + R_{c(m)}(img_2)}{R_{c(m)}(m)} \quad (1)$$

where $R_c(r)$ is the sum of positive relevance toward class c on the region of the mosaic r. This probability can be interpreted as a precision of the relevance. In an sort of eye-tracking game, the *Focus* metric asks to the XAI method "*Why does mosaic m belong to class $c(m)$?*" on a mosaic m which contains both samples belonging and not belonging to the target class $c(m)$. Given the previous question and a good underlying model, a reliable *feature attribution* method should be able to concentrate most of its explanation relevance on the two appropriated images of the mosaic (img_1 and img_2).

B. Evaluation of XAI methods

We evaluate GradCAM, LRP, SmoothGrad, LIME, Grad-CAM++ and IG, using three architectures (AlexNet, VGG16 and ResNet-18) and four target datasets (Dogs vs. Cats, MAMe, MIT67 and ImageNet). Figure 1 shows an example of the *Focus* distribution obtained for the MAMe dataset experiments.
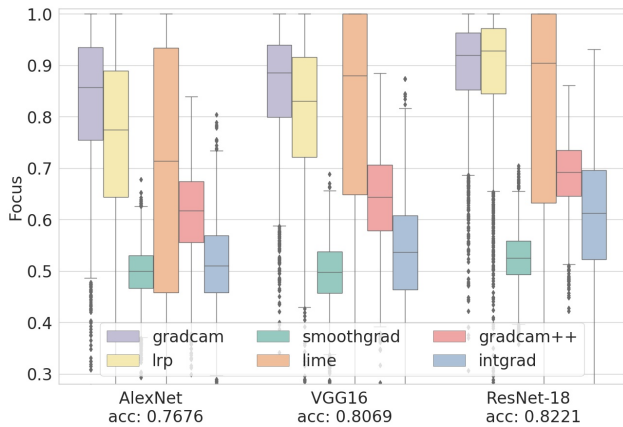
Fig. 1. *Focus* distribution boxplot for different XAI methods applied to models trained for the MAMe dataset.
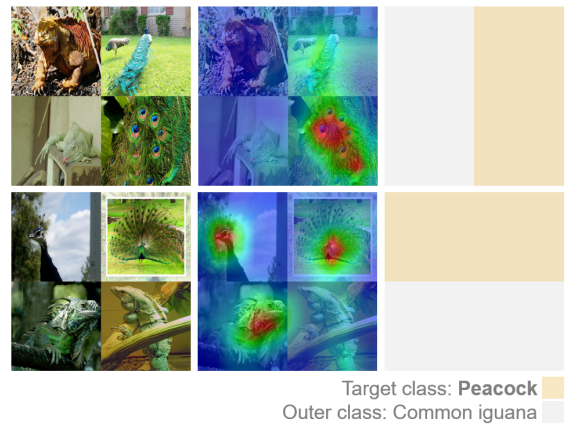

Fig. 2. GradCAM explanations obtained on the ResNet-18 trained with ImageNet. Two examples of mosaics are shown in the first column. The second column shows the corresponding GradCAM explanations for the target class. The third column specifies the positions of the classes within the mosaic. The target class is the *Peacock* class and the outer class is the *Common iguana* class. The example above obtains a high *Focus* score (0.8176) and the one below a lower one (0.4940).

When applied to SmoothGrad or IG, *Focus* finds these methodologies as quasi-random in its explanations with respect to the model. On the contrary, LRP and GradCAM are both found to be consistently reliable methods. GradCAM performs well on all experiments conducted, even when the underlying model is not particularly well fit to the task. LRP performs very well for high performing models, but it becomes more unreliable on less accurate models. This also seems to be the case of LIME, which suffers from an even larger variance. GradCAM++ performs better than random, but not as well as GradCAM and LRP.

C. Bias detection

We can also use the *Focus* to automate the bias identification in models and datasets. This is possible because mosaics induce in-distribution noise, where *Focus* errors directly correspond to visual biases of the model. The proposed procedure is as follows. First, for a better detection of biases between pairs of classes, we use mosaics with two classes. Therefore, in the mosaics used for this section, samples different from the target class actually belong to the same class: $c(img_3) = c(img_4) \neq c(m)$. We concentrate on the most relevant biases by finding the pairs of classes obtaining the lowest mean *Focus* in their joint mosaics. For each of these pairs we extract the mosaics with highest and lowest *Focus*, and present them to a human evaluator who must review the explanations produced. The role of the evaluator is to interpret the rationale behind the explanations (both correct and incorrect) and its degree of generalization for the task. Based on that assessment, corrective measures can be implemented.

To conduct this experiment we use the GradCAM method and the ResNet-18 architecture, a configuration which obtains a particularly robust *Focus*. An example is shown in Figure 2, divided in two rows: the top one corresponds to a high *Focus* and the bottom one to a low *Focus*. In this example, the model is able to correctly attribute relevance to the *Peacock* images on the upper mosaic, while, for the bottom mosaic, some of the relevance incorrectly fall on the head of the *Common iguana*. The fact that most of the incorrect relevance in the *Common iguana* falls in the subtympanic shield (the characteristic circle in its jowl) seems to be related with its visual similarity with the ocellus of the *Peacock* (the circular spot in the feathers).

Notice the iguana's subtympanic shield is hardly visible in the top mosaic. After the identification of these biases, and an assessment of their impact, one could try to mitigate their relevance for the model. For example more images of the target class without the characteristic pattern found in the outer class could be added to the training set (*Peacocks* images where the ocellus is not visible).

D. Conclusion

With the aim of evaluating XAI methods in a quantitative manner, we introduce a novel metric—the *Focus*—to assess the faithfulness of a XAI method to the underlying model. We show the methodology to be consistent across tasks and architectures, providing strong empirical evidence of their performance. We introduce another application of *Focus*, using it for the identification and characterization of biases found in models. This empowers bias-management tools, in another small step towards trustworthy AI.

REFERENCES

[1] L. H. Gilpin *et al.*, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[2] S. Mohseni *et al.*, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2018.

Anna Arias Duart obtained a Bachelor's degree in Telecommunications Technology and Services Engineering in 2015 from the Universitat Politècnica de València (UPV). In 2018 she completed the Double Diploma awarded by the UPV and Télécom Paris-Tech (Paris). She is currently a student of the Doctoral Program in Artificial Intelligence since 2019 at the Universitat Politècnica de Catalunya within the Industrial Doctorate Program in collaboration with SEAT, S.A.

Preparing plasma heating in ITER using integrated modelling

Tomás Bensadón[*], Mervi Mantsinen[*†], Thomas Jonsson[‡]
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[‡]Royal Institute of Technology (KTH), Stockholm, Sweden
E-mail: tomas.bensadon@bsc.es

Keywords—Plasma, fusion, ICRF heating.

I. EXTENDED ABSTRACT

Nuclear fusion takes place when two light nuclei combine to make a heavier nucleus, releasing energy in the process. Magnetic confinement fusion attempts to achieve fusion and use this energy by confining the fuel in the form of a plasma. A plasma is a fully ionized gas whose behaviour is no longer dominated by short-ranged Coulomb forces, but by long-range electric and magnetic forces. Typically, plasmas are composed by hydrogen (H), helium (He), deuterium (D) or tritium (T) ions, or a combination of these. In this abstract, we tackle magnetic confinement fusion using tokamaks. Tokamaks are toroidal devices that have axial symmetry. They use poloidal and toroidal magnetic fields to create twisted magnetic field lines along which the charged particles travel in helical trajectories. In order for the plasma to reach the necessary temperatures for fusion to take place, auxiliary heating systems are used. In this abstract we focus on heating the plasma with electromagnetic waves in the ion cyclotron range of frequencies (ICRF).

Significant advances have been made in the technological development of these magnetic confinement devices in order to progress towards the main goal of fusion research: to achieve electricity-producing fusion power stations that can provide energy reliably, safely and efficiently. The ratio of fusion power produced to the external power required to maintain the plasma in a steady state is known as the Q-factor. Fusion reactions provide energy to the plasma, which leads to self-heating and eventually to a self-sustained reaction, known as ignition. One of the long-term purposes of fusion research is to achieve this ignition (Q=infinite). However, no device so far has achieved a sustainable Q=1 plasma. The main candidate to do so is ITER ("The Way" in Latin), the largest tokamak nuclear fusion reactor, which is being built in the south of France and which is projected to start operating in 2025. The aim is for ITER to maintain Q≥5 and to reach Q=10 for a duration of 400-600s, demonstrating the feasibility of fusion power and of a ten-fold gain of plasma heating power.

The commissioning of ITER is taking place through a staged approach. The First Plasma will be followed by an upgrade of the capabilities of the tokamak and two Pre-Fusion Power Operation (PFPO I and II) phases. In the PFPO phases, the basic controls and protection systems will be demonstrated, including the auxiliary heating and diagnostics systems, in experimental hydrogen (H) and helium (He) plasmas. The Fusion Power Operation (FPO) will start and a transition to deuterium (D) and deuterium-tritium (D-T) campaigns will be made.

Most of the ICRF modelling that has been carried out so far for ITER has focused on heating scenarios relevant for the D, T and D-T plasmas in the FPO stage. There is a need to improve our understanding on the performance of ICRF in H and He plasmas in the PFPO phase of ITER, and on the heating schemes planned for this phase. In this abstract we use the ICRF heating code PION [1] integrated into the transport modelling workflow European Transport Solver (ETS) [2] to study and predict how the plasma will be heated when ICRF heating is applied to ITER PFPO plasmas. The integration into a transport modelling workflow is relevant because PION calculates the ICRF power deposition but it does not predict on its own how the heating will affect the plasma and how its parameters will evolve. A transport modelling workflow such as ETS, which has been developed inside the ITER Integrated Modelling & Analysis Suite (IMAS) [3], can calculate the evolution of the plasma discharge and provide the capabilities for self-consistent predictive simulations.

A. The Physics of ICRF

The basic mechanism by which the plasma absorbs energy from radio-frequency (RF) waves is a wave-particle resonance. In the case of ICRF waves, this resonance takes place when the Doppler shifted frequency of the wave matches an exact harmonic of the ion cyclotron frequency. This condition is described by $\omega = k_{\parallel}v_{\parallel} + l\omega_c$ [4] where $l = 0, 1, 2$, etc. Here, ω is the frequency of the wave, $k_{\parallel}$ is the wavenumber parallel to the background magnetic field, $v_{\parallel}$ is the parallel velocity, ω_c is the ion cyclotron frequency and l represents the harmonics of the wave. The $l=0$ resonance is known as Landau damping, the $l=1$ resonance corresponds to the fundamental and the $l=2$ to the second harmonic. There are several schemes or approaches to heat the plasma with ICRF; in this abstract we will consider the minority heating scheme. In minority schemes, a small amount (usually up to a few percent) of a minority ion species with a larger charge-to-mass ratio than the main ion species is added to the plasma. The ICRH frequency can then be set to target both fundamental and harmonic resonance frequencies of this minority ion species.

B. Modelling and Results

Synthetic ITER discharge 110005 with ^{4}He plasma and H minority was modelled using the PION+ETS integration. The

TABLE I.   Parameters from ITER PION simulations; the H concentration ([H]), magnetic field (B_0), frequency (f), ICRF power (P), electron density (n_e), ion temperature (T_i) and electron temperature (T_e)

	[H] (%)	B_0 (T)	f (MHz)	P (MW)	n_e x 10^{19} (m^{-3})	T_i (keV)	T_e (keV)
ITER	1,2.5,5,10	2.65	40	10	3.3	10.9	10.2

plasma parameters are summarised in Table I. For this ICRF frequency, magnetic field and plasma composition, there are several competing absorption mechanisms at play: the central resonance of the fundamental minority H heating, the second harmonic ^{4}He heating and direct electron damping. A typical ITER pulse will last for 400s to 600s. This particular ITER synthetic discharge has a pulse duration of 647s. However, here we are not simulating a full ITER discharge, but rather a shorter duration aimed at demonstrating the capabilities of the PION+ETS integration. The plasma was pre-heated to the electron and ion temperatures shown in Table I until t=300s, at which point ICRF heating was applied. The chosen duration for the simulation was 5s. This is longer than the slowing down time of the fast ions, which guarantees that the plasma reaches a steady state.

Table II summarises the results of the simulation. A trend can be seen in the increase of energy content and parallel energy content with the H concentration, excluding ΔW at 2.5%, which is slightly lower. It can be noted that the power absorbed by the H ions does also follow the same trend and increases with increasing H concentration, reaching its maximum at 5%. The collisional power transferred from the resonant ions to the electrons is higher than the fraction transferred to bulk ions and hence electron heating dominates over bulk ion heating for H concentrations below 5%. This is due to the formation of a high-energy tail in the resonating ions distribution function with a critical energy lower than the average energy of the fast H ions, which leads to collisional energy transfer to the background electrons.

TABLE II.   Results from PION simulation with ITER discharge, including total (ΔW) and parallel ($\Delta W_\parallel$) energy content of the main resonant ion, power absorbed by the main resonant ion species ($P_{abs.H}$), fraction of the power transferred from resonant ions to bulk ions (P_{ci}) and electrons (P_{ce}).

[H] (%)	ΔW (MJ)	$\Delta W_\parallel$ (MJ)	$P_{abs,H}$(MW)	P_{ci}(%)	P_{ce}(%)
10	4.36	1.16	7.96	65.83	43.17
5	3.16	0.63	8.02	56.11	43.89
2.5	2.83	0.41	7.52	41.62	58.38
1	2.87	0.22	6.81	21.15	78.85

We also present the RF power density absorption profile of electrons and resonant ions at [H]=5% as a function of the normalized flux surface, s, in Figure 1. Figure 1 shows that most of the RF power is absorbed by the resonant H ions, yet there is some competitive absorption from the ^{4}He ions. An off-axis absorption can also be noted, as the power density peaks at around s=0.3 in the case of H, while the ^{4}He absorption is more central at the steady state. The broadening and flattening of the electron's power absorption profile is also due to the presence of high-energy fast ions.

C. Conclusion

In this study we present results of predictive simulations at different H concentrations of an ITER synthetic discharge using the PION+ETS integration. Fundamental minority H
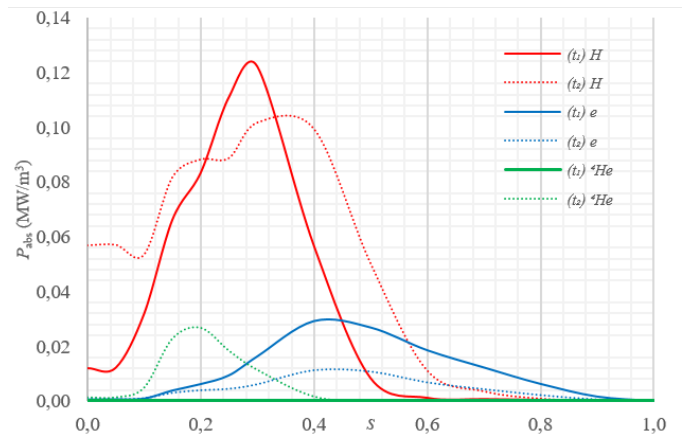

Fig. 1.   RF-power density absorbed in the ITER scenario by the resonant ion species, H and ^{4}He, and the electrons as a function of the normalized flux surface, s, for the first time step t_1=300s (solid) and steady-state t_2=305s (dashed), at [H]=5%

heating was found to be a strong ICRF power absorption mechanism as compared to the competing absorption mechanisms, yielding increasing power absorption with increasing H concentration and reaching a maximum of 80.2% at [H] $\sim$ 5%. Further work is needed to exploit the full capabilities of the workflow integration and study the effect of ICRF heating on transport.

II.   ACKNOWLEDGMENT

REFERENCES

[1]   L. Eriksson *et al.*, "Comparison of time dependent simulations with experiments in ion cyclotron heated plasmas," *Nuclear Fusion*, 1993.

[2]   D. Coster *et al.*, "The European Transport Solver," *IEEE Transactions on Plasma Science*, 2010.

[3]   F. Imbeaux *et al.*, "Design and first applications of the ITER integrated modelling analysis suite," *Nuclear Fusion*, 2015.

[4]   T. Stix, "Waves in plasma," 1992.

Tomás Bensadón graduated in Quantum Physics and Physical Chemistry (2018) from University College London (UCL), UK, where he also obtained an MSc in Particle Physics (2019). He then worked for the United Nations at the Fusion Department in the International Atomic Energy Agency (IAEA), Austria. In 2020 he joined the Fusion Group (CASE) at Barcelona Supercomputing Center (BSC) as a PhD student under the supervision of ICREA professor Mervi Mantsinen.

The development of high-order methods for finite elements to study turbulent reacting flow

Antonio Blanco Casares, Daniel Mira Martínez, Oriol Lehmkuhl

Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {antonio.blanco, daniel.mira, oriol.lehmkuhl}@bsc.es

Keywords—Numerical methods, LES, FEM.

I. EXTENDED ABSTRACT

Numerical methods for solving differential equations are the base of any computational simulation. The finite element method (FEM) has emerged as one of the most powerful numerical methods so far devised, mainly because of the the ease in modeling complex geometries, the consistent treatment of differential-type boundary conditions and the possibility to be programmed. To solve a problem, the FEM subdivides a large system into smaller, simpler parts that are called finite elements. Focusing on Computational Fluid Dynamics (CFD), the time is an important variable in the simulations, the set of equations is based on solving a set of partial differential equations given the initial and boundary conditions and the discretized domain.

It is challenging to find the perfect numerical method, the case of LES simulations requires methods very low numerical dissipation. In the last decades, the advance on computation has allow to solve complex algorithms and great efforts have made on their development. The key point is to have a algorithm with low computational cost that leads an acceptable solution with the lower number of operations. The most recent researches indicate that complex high-order numerical methods can save costs with respect to the low-order and more naive ones, because they achieve faster the convergence even though they require a high number of operations per iteration and mesh point. Nowadays in HPC this issue has been solved by the hardware accelerators, i.e. GPUs, that allow to compute a large amount of numerical operations in a parallel way.

The main objective of this work is to show the importance of the high order methods and introduce a methodology for making a proper comparison of the most popular numerical methods to solve problems of fluid flows with reactions and LES. Section A gives an insight of the turbulence models and remarks the requirements of the numerical model, section B shows the most common stabilization methods for FEM and section C shows the comparison methodology and the results of some stabilization methods.

A. Turbulence models

The Navier–Stokes equations describe the motion of viscous fluid flows. When a simulation resolve these equations for all relevant spatial and temporal scales, it is designed as direct numerical simulation (DNS). For most fluid dynamical engineering applications turbulence effects are important, and because of the big
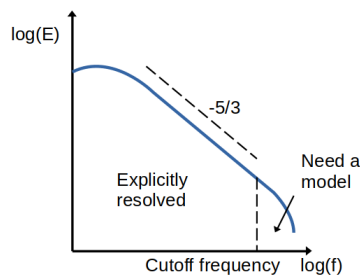

Fig. 1: Energy cascade with LES representation.

amount of scales to solve, the huge computational cost makes the DNS of such turbulent flows impractical. Hence, the Navier–Stokes equations are often coupled with turbulence models. The principal idea behind LES is to reduce the computational cost by ignoring the smallest length scales, which are the most expensive to solve, by applying a filter at an specific frequency, as shown in Figure 1. The effect of frequencies lower than the cutoff is modeled.

For LES simulations, it is vital have a mesh resolution and numerical methods that capture all the energy of the smallest eddies scales without numerical dissipation errors, the quality of a computation is known to be as dependent on the accuracy of the numerical scheme and the computational mesh as on the LES model itself. A poor mesh resolution and the use of a low-order scheme can contribute dramatically to the dissipation of eddies and to the distortion of their form (through numerical dispersion errors) [1]. For complex geometries, when unstructured or structured multi-block grids are required, the simplest and most common way to achieve the required level of accuracy is to use a relatively low-order finite-volume or finite-element scheme with a very fine mesh. Although the results of such computations may be impressive, the approach can be extremely costly. This requirement for accuracy leads us to consider high-order methods, since the total cost of a calculation is likely to be lower if we increase the order of the numerical scheme rather than increasing the number of mesh points [2].

B. Galerkin stabilization methods

Standard finite element approximations are based upon the Galerkin formulation of the method of weighted residuals. The main difficulty is due to the presence of convection operators which are non-symmetric, in practice this leads to spurious oscillations, this has motivated the development of a wide variety
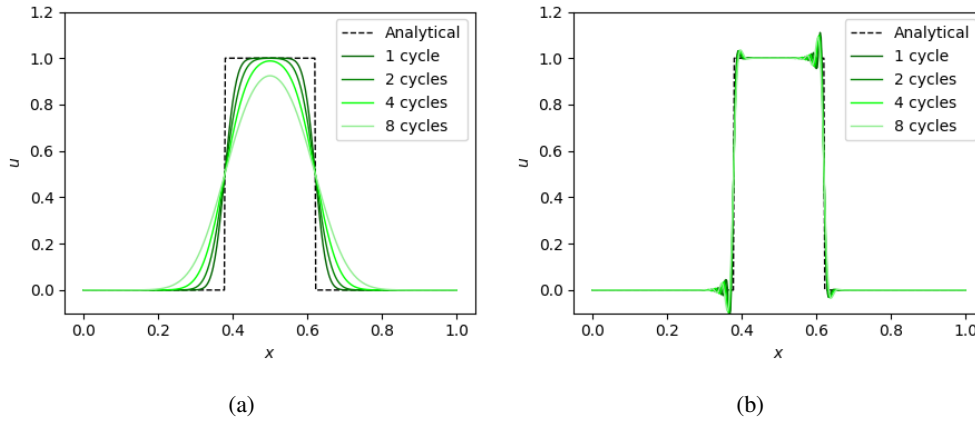
(a)   (b)

Fig. 2: Solution of a pure convection problem with an initial square signal applying FEM with linear elements and using (a) SUPG method and (b) 3rd order Taylor-Galerkin method.

of stabilization methods. Essentially, all these methods consist in the addition of a stabilization term to the original Galerkin formulation of the problem [3].

The first alternative is SUPG, which is a commonly used formulation based on applying an specific amount of viscosity in the streamline direction, that amount is related to the residual of the PDE of the problem.

Taylor-Galerkin (TG) methods are constructed by combining the PDE equation of the problem with a Taylor expansion with the desired order. According to Colin [2], it is found that the TG schemes are too dissipative at intermediate and high frequencies and that these schemes are therefore unsuitable for practical LES applications. As a result, a new class of two-step TG schemes was developed that gives third-order accuracy with less overall dissipation, it is denominated as TTGC.

Entrophy viscosity is a relatively modern and out-standig method. The entropy satisfies a conservation equation only in the regions where the solution is smooth and satisfies an inequality in shocks; this inequality then becomes a selection principle for the physically relevant solution. By making the numerical diffusion to be proportional to the entropy production, we add a large numerical dissipation in the shock regions and almost no dissipation in the regions where the solution remains smooth [4].

C. Results

In order to overview all the methods, a 1D pure convection problem with FEM is proposed, considering periodic boundary conditions and setting a square signal as the initial condition, as shown in Figure 2. If we were to simulate the pure Galerkin method, we would observe spurious oscillations that end up breaking the simulation.

It is clearly observed in Figure 2 that the drawback of the SUPG method is the high numerical diffusion compared to TG, as the signal propagates through the domain, the shape becomes smoother and the maximum value lowers. In the other hand TG3 preserves very well the shape of the wave but overshoots are present, however they do not seem to grow with time. TG family of methods seems to be an appealing choice for solving our turbulent reacting problem.

D. Conclusion

In this study, a methodology for comparing numerical methods has been presented. The TG family of methods seems to be a good choice, to our knowledge the TTGC and Entropy viscosity are the most suitable methods for turbulent reacting flow, future works will be focused on comparing these two methods in 1D and 2D domains.

REFERENCES

[1] C. Canuto *et al.*, "Evolution to complex geometries and applications to fluid dynamics," *Spectral Methods*, vol. 596, 2007.

[2] O. Colin and M. Rudgyard, "Development of high-order taylor–galerkin schemes for les," *Journal of computational physics*, vol. 162, no. 2, pp. 338–371, 2000.

[3] R. Codina, "Comparison of some finite element methods for solving the diffusion-convection-reaction equation," *Computer methods in applied mechanics and engineering*, vol. 156, no. 1-4, pp. 185–210, 1998.

[4] J.-L. Guermond *et al.*, "Entropy viscosity method for nonlinear conservation laws," *Journal of Computational Physics*, vol. 230, no. 11, pp. 4248–4267, 2011.

Antonio Blanco Casares received his Bachelor and Master degree in Engineering from University of Valladolid (UVa), Spain in 2018 and 2020. The following year, he was a visitor Researcher in San Diego State University (SDSU) where we worked on the computational department of Aerospace engineering, on studying supersonic flows over blunt bodies. Since 2021, he has been in the Computer Applications in Science and Engineering (CASE) Department of Barcelona Supercomputing Center (BSC) as well as a PhD student at the Physics department of Universitat Politècnica de Catalunya (UPC), Spain.

Characterizing transcriptome variation in human populations at single-cell resolution

Ruben Chazarra-Gil*, Marta Melé*

ruben.chazarra@bsc.es, marta.mele@bsc.es

*Barcelona Supercomputing Center, Pl. Eusebi Güell, 1-3, Barcelona, Spain

Keywords

Ancestry, single cell RNA-seq, eQTL

EXTENDED ABSTRACT

Phenotypic diversity in human populations is a direct consequence of genetic variation, which acts in conjunction with environmental and behavioral factors to produce phenotypic variation, from eye color and height to disease susceptibility and responses to drugs (1). High population-specific variability in disease's prevalence have been described, including multiple examples where a disease is strongly overrepresented in a single population, for instance sickle cell anemia in Africans, hemochromatosis in Northern-Europeans or familial Gaucher's disease in Ashkenazi Jews (2). In addition, population differences in response to drugs have been documented, for instance, 5-Fluorouracil (cancer chemotherapeutic), Warfarin (anticoagulant for preventing thrombosis and embolism) or nicotine (3).

In this context, there has been a growing interest in profiling the molecular causes underlying infectious disease-related phenotypic differences across individuals from populations of different genetic backgrounds. Studies using RNA-seq data from primary monocytes, as a model of an innate immunity, have shown that human populations differ in their transcriptional responses to immune challenges, which are largely controlled by genetics and have been shaped by natural selection (4). In addition, a similar work focusing on alternative splicing characterisation upon immune activation highlights the contribution of positive selection to diversify the splicing landscape of human populations (5).

Moreover, additional works using single-cell RNA-seq indicate that most of the ancestry effects on the immune response are cell type specific, exceptuating interferon (IFN) response which is strongly correlated with European ancestry after infection with influenza A virus (*Figure 1*) (6). Also in line with previous evidence, it has been seen that eQTLs explains > 50% of population differences in response to infection, stressing the key role played by genetics in shaping population differences in immune responses (6).

This evidence suggests that genetic ancestry is a main driver of inter-individual differences in response to infection. In turn, these findings highlight the importance of studying the effect of human population genetic variation over disease and disease response for developing effective treatments, and in order to lay the foundations for the establishment of personalized medicine. Moreover, the characterisation of the transcriptome differences derived from human genetic variation can provide further insights into the evolution of human populations.
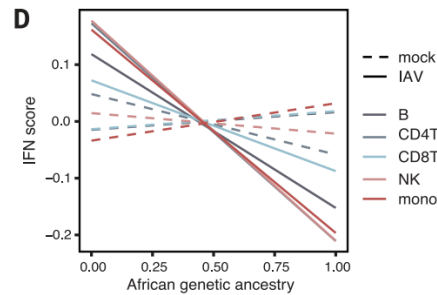
Figure 1. *Correlation between African genetic ancestry proportion and IFN score in mock (dotted lines) and IAV-infected conditions (solid lines) for different peripheral blood mononuclear cell types. A consistent reduction in the IFN response is observed as the estimated african ancestry of individuals increases. INF score represents the average expression of interferon genes, while the proportion of african genetic ancestry is estimated based on polymorphism presence. (Adapted from Randolph et al., (Science, 2021)).*

Data

This project is defined within the eQTLGen Consortium, a large-scale, international collaborative effort with the aim of finding disease-related genetic variants in individual immune cell types. This study will perform **single-cell RNA-seq** and **ATAC-seq** in a collection of peripheral blood mononuclear cells (**PBMCs**) from different African populations, including rural and urban populations, and first generation immigrants. In addition, *ex vivo* immune stimulation will be performed. **Genome variation data** is also expected to be present.

Objectives

The main aim of this project is performing a **detailed characterisation of the transcriptomes of different human populations at single cell resolution**. Some additional objectives are exposed below.

A key step in any scRNA-seq data analysis involves conferring identity to each cell. For this aim, we will use available machine learning methods which are able to project cell type labels from existing annotated references onto our query data. Fortunately, initiatives like the Human Cell Atlas are a source of curated and well annotated scRNA-seq reference datasets across many tissues. In addition, we will manually confirm the assigned cell type identities by checking the expression of established markers for each cell population present.

For the **estimation of genetic ancestry** of each individual we propose the usage of methods (like ADMIXTURE) based on polymorphism datasets.

In order to characterize genetic ancestry effects on gene expression across cell types, we will perform **differential expression analysis** (DEA) to determine population differentially expressed genes (popDEG). In this regard, it will be interesting to confirm the observation of ancestry effects being cell type specific in a non-viral stimulation context (6). Pseudo-bulking (aggregating all the cells from the same cell type and donor into a single profile) is often considered in single-cell DEA to increase statistical power and improve effect sizes. We

will consider both standard single-cell DEA and pseudo-bulking, in case the first strategy is unable to capture population differences. Next, to identify functionally enriched pathways with genetic ancestry, we will perform **gene set enrichment analysis** with the popDEGs. We will consider established sources of terms as GO, KEGG and additional gene signature databases as MSigDB.

In addition, we will assess the contribution of genetic variation to genetic ancestry-associated differences by performing **expression Quantitative Trait Loci** (eQTL mapping). We will search for cis-eQTLs (SNPs within 100kb distance from the gene) and trans-eQTLs (SNPs regulating gene networks over long genomic distances).

We will also identify genetic associations with epigenetic variation (**chromatin accessibility QTLs**) and **response to stimulation QTLs** (reQTLs). It has been reported that cis-eQTLs explain a large fraction of the variance in ancestry-associated expression differences (6). In order to validate this, we aim to quantify the fraction of population differences in gene expression that can be attributed to genetics as done in (6).

Moreover, in order to screen for **polygenic selection**, we will perform GSEA on the intersection of popDEGs and genes with an eQTL (eGenes). In this context, encountering consistent population differences in the expression of genes within the same pathway, can be due to genetic drift, or due to polygenic selection. We will employ different strategies to assess if natural selection, as opposed to genetic drift, has contributed to differences between populations.

Additional analysis we envision to perform include **gene regulatory networks** and **trajectory analysis** comparison across populations.

Author biography

Ruben Chazarra Gil obtained his BSc in Biotechnology from the Polytechnic University of València in 2018. Motivated by his interest in bioinformatics, he performed a 1 year Erasmus internship at the Sanger Institute, (Cambridge, UK). Here he became interested in single-cell transcriptomics, and method development. His project consisted in building up a pipeline for benchmarking single-cell RNA-seq data integration methods which was published in NAR journal. Following his interest in gene expression and method development, he performed a second internship at the European Bioinformatics Institute focusing on cell-type annotation methods in the Array Express infrastructure. Next, he spent one year in the Core Bioinformatics Team at the Cambridge Stem Cell Institute, performing extensive single-cell expression data analysis on disease progression datasets in Liver and Lung. In addition, he developed the team's bulk RNA-seq pipelines. He has further interests in multi-omics data integration and application of machine learning in cell-type classification. In addition he is a fan of sports (cycling, running, swimming) and music, playing guitar often and singing in the University of Barcelona choir.

References

[1] Rahim, N.G., Harismendy, O., Topol, E.J. *et al.* Genetic determinants of phenotypic diversity in humans. *Genome Biol* 9, 215 (2008). https://doi.org/10.1186/gb-2008-9-4-215

[2] Xiao, Q., Lauschke, V.M. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *npj Genom. Med.* **6,** 41 (2021). https://doi.org/10.1038/s41525-021-00203-x

[3] Bachtiar, M., Lee, C.G.L. Genetics of Population Differences in Drug Response. *Curr Genet Med Rep* **1,** 162–170 (2013). https://doi.org/10.1007/s40142-013-0017-3

[4] Quach, Hélène, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* 167.3 (2016): 643-656 https://doi.org/10.1016/j.cell.2016.09.024

[5] Rotival, M., Quach, H. & Quintana-Murci, L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat Commun* **10,** 1671 (2019). https://doi.org/10.1038/s41467-019-09689-7

[6] Randolph, Haley E., et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* 374.6571 (2021): 1127-1133.https://doi.org/10.1126/science.abg0928

Nuclear Fusion Reactor Materials: Modelling Atomic-Scale Irradiation Damage in Metal

Mary K. Chessey[a,1], J. Julio Gutiérrez Moreno[a,2], Marina García Romero[b],
Paolo Settembri[c], Eoin Kearney[d], Mervi Mantsinen[a,e]

[a]*Barcelona Supercomputing Center (BSC), C/Jordi Girona 31, 08034 Barcelona, Spain*
[1]`mary.chessey@bsc.es`, [2]`julio.gutierrez@bsc.es`
[b]*Universitat Autònoma de Barcelona (UAB), Bellaterra, Spain*
[c]*Università degli Studi dell'Aquila, Italy*
[d]*University of Edinburgh, Scotland, UK*
[e]*ICREA, Pg. Lluís Companys 23, 08034 Barcelona, Spain*

EXTENDED ABSTRACT

Achieving nuclear fusion as an energy source on Earth is a practical goal that relies on continuing scientific and engineering innovation. Functional fusion reactors around the world today allow scientists and engineers to plan improvements that will eventually allow for greater energy output than the input required to operate the machine (including heating the plasma and operating the superconducting electromagnets that confine the plasma, among other energy inputs). The fusion reaction between nuclei of hydrogen isotopes is a carbon-free source of massive amounts of energy that could be paramount in a global turn towards greener energy. The fusion fuel needed to provide one person's energy use for 100 years (assuming 20 kWh per day) is contained within roughly one and a half bathtubs of water and 3 laptop batteries. Given the enormous payoff of fusion, continued research and development are of great interest so that current challenges of heating and confining plasma, mitigating plasma disruptions, improving efficiency of magnets, and extending the lifetime of materials subjected to the harsh conditions surrounding the plasma may be overcome. Fusion reactor materials research carried out here at the BSC contributes to this ambitious goal.

The idealistic goal for fusion materials research is to provide predictions about material behavior with the accuracy of quantum mechanical calculations at the scale of a full fusion reactor. Using strategic approximations and working at a small scale, computational fusion materials researchers can accurately reproduce and explain experimentally observed physical phenomena, such as the formation of microstructural defects in metals under neutron-irradiation, and offer the best predictions available for behavior of materials in future fusion reactor environments, where data about what will happen simply do not exist yet.

In the study presented here, we examined the thermal conductivity, or how quickly a material allows heat to flow, of tungsten (W). W has been selected for plasma-facing components in ITER, which is currently under construction. We used LAMMPS atomic modelling of materials software and found that the thermal conductivity of W is significantly decreased in the presence of defects.

A. Introduction

Fusion reactors depend on cutting edge scientific and engineering knowledge. As such, the materials used to build a fusion reactor are carefully designed to fulfill a variety of different purposes [1]. The material that serves as the focal point of this study, tungsten (W), is used in protective plasma-facing components (PFC), which require high melting temperatures and high thermal conductivity among other characteristics. One of the primary purposes of W in the fusion reactor is to carry excess heat away from the plasma quickly without melting and without allowing other nearby components to melt. A PFC called the divertor is expected to undergo extreme heat fluxes of 10-20 MWm^2 during ITER operation. The ability of PFC to dissipate heat is given by their thermal conductivity. Materials like W conduct heat through electrons and phonons (quasiparticles representing vibrations in a crystal lattice). This study considers the phonon thermal conductivity, which is expected to be substantially affected by defects. High energy neutrons (14.1 MeV) are emitted by the plasma due to the nuclear fusion reaction between hydrogen isotopes deuterium and tritium. The neutrons provide energy output by heating special parts of the reactor, but also create damage in the protective materials around the plasma. Typical defects associated with this characteristic fusion-produced neutron-irradiation of W include self-interstitial atoms (SIA, W atoms that crowd the normal empty spaces of the crystalline lattice structure), vacancies (empty points in the lattice where a single W atom is missing), voids, and extended defective microstructures formed by the movement of defects throughout the metal, among others.

Computational methods are needed to make predictions about material behavior in future fusion reactors like ITER where experimental data do not yet exist. Quantum mechanical approaches are highly accurate but computationally expensive, and are typically restricted to ground state calculations of a few hundred atoms. Classical molecular dynamics (MD) is a simplified model of particles moving within an external potential. The interatomic interactions are determined using potentials (fitted to quantum mechanical calculations and sometimes experimental data) to calculate the energy and forces between atoms, which are then used to update momenta and positions. MD codes like LAMMPS [2] allow for simulation of larger systems for longer times at realistic temperatures due to a decreased computational cost. In this study, we used MD to explore how the phonon thermal conductivity of W changes in the presence of defects.

B. Methods

We estimated thermal conductivity of W using LAMMPS and a non-equilibrium molecular dynamics (NEMD) approach based on the classic Fourier heat transfer law:

$$\kappa = \frac{J}{\partial T / \partial x}$$

where κ is the thermal conductivity, J is the heat flux in the x direction, and $\partial T / \partial x$ is the gradient of temperature in the x

direction [3]. We created a simulation cell with 3D periodic boundary conditions containing a few hundred thousand to a few million atoms, corresponding to physical dimensions of $1583 \times 158 \times 158$ Å (one Ångström is one tenth of one nanometer) for the largest system. The system was then equilibrated to 300 K and zero pressure using NPT (isothermal-isobaric) and NVT (canonical ensemble) relaxations in LAMMPS. Next, we applied a local heat source and sink to create regions of hot and cold, resulting in a steady state heat flux J after 400 picoseconds of simulated time (400 cpu hours). Between the heat source and sink regions, we placed defects and measured their effect on κ.
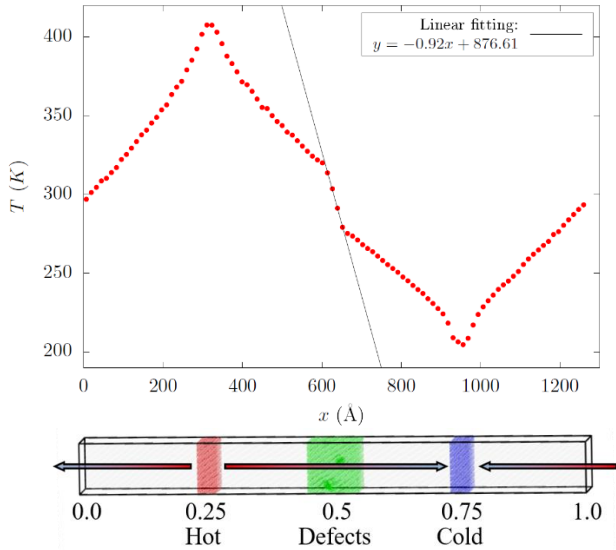
Fig. 1 A sample temperature profile of a W system with a void placed in the center, and a schematic indicating heated and cooled regions and placement of defective material. Arrows indicate the direction of heat flow.

We tested a variety of types of defects including spherical voids in the W lattice and defective microstructures consisting of vacancies and clusters of interstitial atoms. The voids had radii of up to 30 Å, placed in the center of the simulation cell. The effect of a void with radius 20 Å is visible in Fig. 1 as a sharply steeper slope in the temperature profile, signifying a lower thermal conductivity in the damaged region. An initial strategy that we tried for creating the types of defective microstructures associated with high energy neutron-irradiation was a cascade procedure starting with one high energy atom (up to 200 keV). Due to the heavy computational cost of simulating defect cascades, a new simplified method was instead used: Creation Relaxation Algorithm (CRA) [4]. CRA can produce damage levels that are several orders of magnitude higher than defect cascade approaches for similar computational time. This algorithm iteratively inserts Frenkel pairs (one SIA and one vacancy) randomly, then performs an energy minimization by making small adjustments to the positions of all atoms. Building up 250 iterations of inserting 1000 Frenkel pairs then performing the energy minimization required 1200 cpu hours (about 5 cpu hours per minimization) in a 2.5M atom system.

C. Results

For a perfect W system of 320k atoms without structural defects, we found a phonon thermal conductivity of 13.33 W/mK. Thermal conductivity dropped by 74% in the example shown in Fig. 1, where the cross sectional area of the spherical void (circle of radius 20 Å) is about 30% of the cross sectional area of the simulation cell (square of side length 63 Å). A small void where only 11 atoms were removed from a system of half of a million atoms resulted in a 5% decrease in thermal conductivity. In the presence of a highly defective microstructure corresponding to the insertion of one Frenkel pair for every atom within the central damaged region of the simulation cell, W thermal conductivity decreased by 21%.

D. Discussion

One implication of the decreased thermal conductivity of W in the presence of defects is the risk of melting of reactor components due to a build-up of heat on the plasma-facing side. This predicted effect in fusion reactor materials is one way that the lifetime of machine components is limited, reducing the overall efficiency of energy output to machine maintenance in future fusion power plants.

This work can be extended by using alternative methods for estimating thermal conductivity, such as the Green-Kubo equilibrium molecular dynamics approach. A wider range of temperatures could be examined, since operating conditions of W in ITER will include 1200-1600 K. Further, voids could be filled with gases expected inside fusion reactors, such as the helium that results from the nuclear fusion of deuterium and tritium. Another major avenue to extend this work will be to compare results given alternative interatomic potentials.

References

[1] G. Pintsuk *et al.*, Fusion Engineering and Design **174**, 112994 (2022).
[2] A.P. Thompson et al., Comp Phys Comm **271**, 10817 (2022).
[3] P.K. Schelling *et al.*, Phys Rev B **65**, 144306 (2002).
[4] P.M. Derlet and S.L. Dudarev, Phys Rev Materials **4**, 023605 (2020).

Author biography

Mary Kate Chessey completed a BSc in physics at Drexel University (Philadelphia, USA) in 2013, and a MSc in physics and PhD in physics education research in 2018 at the University of California Davis. She was then a Postdoctoral Research Associate at University of Maryland College Park. In 2021 she joined the CASE Department at the BSC, working most closely with Dr. Julio Gutiérrez in the Fusion Group led by ICREA Prof. Mervi Mantsinen.

Comparison of Five Strategies for Seasonal Prediction of Bioclimatic Indicators in the Olive Sector

Chihchung Chou[#1], Raül Marcos-Matamoros[*#2], Nube González-Reviriego[#3]

[#]*Earth Sciences Department, Barcelona Supercomputing Center - CNS, Plaça d'Eusebi Güell, 1-3, Barcelona 08034, Spain*
[1]`chihchung.chou@bsc.es`, [3]`nube.gonzalez@bsc.es`

[*]*Department of Applied Physics, University of Barcelona, Av. Diagonal 647, Barcelona 08028, Spain*
[2]`raul.marcos@bsc.es`

Keywords── **Seasonal Prediction, Blending Approach, Olive Sector, Bioclimatic Indicators, Climate Services**

EXTENDED ABSTRACT

The forecast quality of five seasonal prediction strategies used to obtain tailored bioclimatic indicators in the olive sector has been assessed over the Iberian Peninsula (IP). In total, five indicators have been selected considering their importance in the management of the olive orchard. As time progresses through the indicator target period, the impact of the increasing share of actual observations included in its computation has been evaluated by examining the variabilities of correlation and fair Rank Probability Skill Score (fair RPSS) in each initialization date.

The results show that blending either seasonal predictions or climatology with observations enhanced the capability of forecasting the tercile category for all the indicators when compared with the use of climatology or seasonal predictions alone. In fact, for Spring Temperature Maximum (SPRTX) and Growing Season Temperature (GST) indicators, the combination of observations and SEAS5 prediction could outperform the other methods for most of the start months. As for those threshold-defined indicators, namely Spring Heat Days (SPR32) and Summer Heat Stress Days (SU36 and SU40), the end-users are highly encouraged to use climatology in the first month and combine it with observations as soon as they become available.

A. Introduction

Olive is one of the key staples in many Mediterranean countries. Approximately 70% of the world's olive oil came from the EU in 2020-2021 (IOC, 2018). Additionally, Mediterranean Spain accounts for more than half of the total cultivation areas in the EU (EPRS, 2017). However, a significantly hotter and drier Mediterranean basin has been projected with regional climate models under varying scenarios by the end of this century (Cos et al., 2022). As such, scientists have warned that an advanced adaptation driven by climatic information is essential to tackle these foreseen threats induced by climate change (Ranasinghe et al., 2021).

Seasonal predictions have been applied to a range of sectors but the forecast skill over non-tropical regions are limited (Doblas-Reyes et al., 2013). This could prevent end-users from using them in their decision making. One of the solutions proposed to solve this problem is the inclusion of observation in the calculation of the bioclimatic indicators (Chou et al., under review). This work aims to understand the transferability of the blending approach (already applied to the wine sector) to five indicators of the olive sector by comparing five seasonal prediction strategies.

B. Data and Spatial Domain

The daily temperature predictions of the ECMWF SEAS5 prediction (S5) in the 1993-2016 period has been used (Johnson et al., 2019). As for the observational data, we have used the ECMWF ERA5 reanalysis (Hersbach et al., 2018) as the reference data over the same hindcast period. The IP has been selected as the spatial domain of interest because it holds more than 50% of the total olive cultivation areas in the EU.

C. Bioclimatic Indicators for the Olive Sector

The five indicators evaluated in this work have been co-defined taking into account the olive grower's perspective.
- SPRTX: mean temperature maximum from April-May
- SPR32: accumulated number of days with the maximum temperature exceeding 32ºC from 21st April to 21st June
- SU36 and SU40: accumulated days with the maximum temperature above 36 and 40ºC, respectively, from 21st June to 21st September
- GST: the mean of daily average temperatures from April to October.

D. Five Strategies to Generate Predictions of Indicators

The acronyms of the five strategies are listed below with the relevant information about the skill metrics calculation followed by the approaches of generating the predictions of indicators.
1. E5: ERA5 climatology (used as a benchmark)
2. S5: bias corrected SEAS5 prediction
3. B-S5: SEAS5 prediction blended with the past observations
4. B-E5: ERA5 climatology blended with the past observations
5. P: persistent prediction

The basic concept of the blending approach is to progressively combine the (passed) observed data with the predicted data. This technique has been applied for B-S5 and B-E5 with the only difference being the predicted data used in each case. The former (B-S5) took SEAS5 prediction while an ensemble of observations was created for the latter (B-E5).

As for the P strategy, the assumption is that tomorrow remains the same as today. Thus, the key step is to decide the month(s) used for each initialization date. To uptake as much information of observations as possible, we have used all the days of the closest month(s) being available. Taking GST as an example, the daily data in March (April) has been taken to compute the indicator for the start month of April (May). Thereafter, from the start month June onwards, the months from April to the month before the present one have been used.

E. Evaluation Methods

Correlation coefficient and fair RPSS have been applied for evaluation. To understand the effect of the increasing share of the observations within the indicator on the skill metrics, they have been computed for each initialization date of each

indicator period. It is worth noting that only the highest positive fair RPSS is interesting in the final comparison.

For the P strategy, we assigned 1 to the tercile category in which the prediction belonged due to the lack of ensemble members. As such, the annual fair RPSS would be either one or negative values.

F. Results and Discussions

According to the indicators' features, the results have been grouped into two categories: the threshold-defined indicators including SPR32, SU36 and SU40; and the SPRTX and GST belong to the period-average indicator.
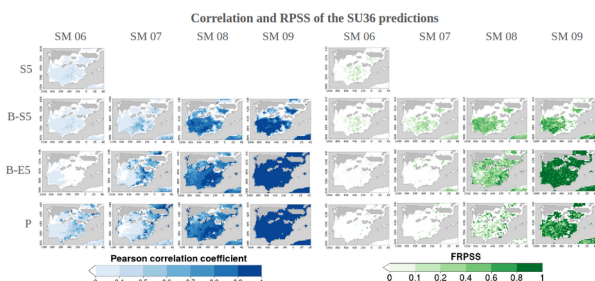
Fig. 1. Correlation and fair RPSS of the SU36 predictions of the four strategies (S5, B-S5, B-E5 and P, top to bottom) for each initialization date (1st June to 1st September, left to right)

Here, the SU36 indicator is used as an example to describe the overall improvement of the verification metrics and the comparison between the different strategies. The increased value in correlation when blending more observations within the season could be seen in most cases as shown in Fig. 1 (left panel). Regarding fair RPSS (the right panel), when comparing B-S5 with B-E5, the former outperformed the latter in July for the southern half of IP (see the SM07 column). After that, the inclusion of July observations substantially enhanced the performance of B-E5 in SM08 as shown in the third column. In SM09, B-E5 performed better than B-S5 throughout the domain when three out of four observations were available. In terms of P, positive fair RPSS started to appear in SM08 when the critical month (July) was uptaken and attained values above 0.6 over a large area in September.

To better understand the spatial and temporal variabilities of the best prediction for each indicator, Fig. 2 depicts the best performer (with the highest fair RPSS) for each pixel and start month for the threshold-defined indicators. To sum up, E5 (Climatology) and B-E5 are recommended for the threshold-defined group (SPR32, SU36 and SU40) while S5 and B-S5 have a better performance for the period-average indicators (SPRTX and GST, not shown here).
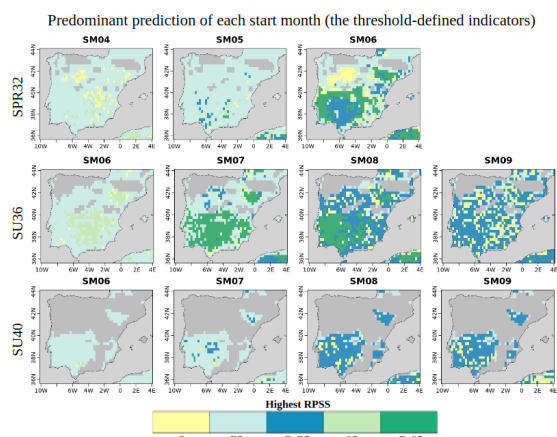

Fig. 2. The best prediction for the threshold-defined indicators (SPR32, SU36 and SU40, top to bottom) in each start month (left to right)

G. Conclusions

It is recommended to uptake the available observations when seasonally predicting bioclimatic indicators for the olive sector. Over the IP, end-users are encouraged to use blending SEAS5 (B-S5) prediction for those period-average indicators while the climatology (E5) and blending past observations with climatology (B-E5) would be a better choice when involving threshold-based indicators.

References

[1] Besnard, G., Khadari, B., Navascués, M., Fernández-Mazuecos, M., El Bakkali, A., Arrigo, N., ... & Savolainen, V. (2013). The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant. Proceedings of the Royal Society B: Biological Sciences, 280(1756), 20122833.

[2] Chou, C., Marcos-Matamoros, R., Garcia, L. P., Pérez-Zanón, N., Teixeira, M., Silva, S., Fontes, N., Graça, A., Dell'Aquila, A., Calmanti, S., González-Reviriego, N., (submitted to Climate Services, under review) Advanced seasonal predictions for vine management based on bioclimatic indicators tailored to the wine sector.

[3] Cos J., Doblas-Reyes F., Martin J., Marcos R., Bretonnière P.A., Samsó M. (2022) 'The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections', Earth System Dynamics, pp. 1–26. doi: 10.5194/esd-2021-65.

[4] Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013: Seasonal climate predictability and forecasting: status and prospects. Wiley Interdiscip. Rev. Clim. Chang., 4, 245–268, https://doi.org/10.1002/WCC.217.

[5] EPRS (European Parliamentary Research Service) Author: Rachele Rossi Members' Research Service PE 608.690 (2017). The EU olive and olive oil sector Main features, challenges and prospects.

[6] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., … & Thépaut, J-N. (2018): ERA5 hourly data on pressure levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on < 17-02-2020 >), 10.24381/cds.bd0915c6

[7] International Olive Council (2018) World Olive Oil Figures. https://www.internationaloliveoil.org/what-we-do/economic-affairs-promotion-unit/#figures. Assessed date: 29 Dec 2021

[8] Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., ... & Monge-Sanz, B. M. (2019). SEAS5: the new ECMWF seasonal forecast system. Geoscientific Model Development, 12(3), 1087-1117.

[9] Ranasinghe, R., A.C. Ruane, R. Vautard, N. Arnell, E. Coppola, F.A. Cruz, … & R. Zaaboul, 2021: Climate Change Information for Regional Impact and for Risk Assessment. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [MassonDelmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, … & B. Zhou (eds.)]. Cambridge University Press. In Press.

Author biography

Chihchung Chou is a postdoctoral researcher of the Climate Services team in the Earth System Services group at BSC. With a background in Environmental Engineering, he studied the land-atmosphere interaction over the South Asia monsoon region in his PhD (University of Melbourne). After that, he participated in preparing a new Catchment Attributes and Meteorological Large Sample dataset for Australia. His research interests include climate services and hydro-meteorology.

Discovering the Ethereum2 P2P Network

Mikel Cortes-Goicoechea*[†], Leonardo Bautista-Gomez*[†]

*Barcelona Supercomputing Center, Barcelona, Spain

E-mails: mikel.cortes@bsc.es, leonardo.bautista@bsc.es

Abstract—**Achieving the equilibrium between scalability, sustainability, and security while keeping decentralization has prevailed as the target solution for decentralized blockchain applications over the last few years. Multiple blockchain teams have proposed several approaches to achieve it, Ethereum being among them. Ethereum is on the path of a major protocol improvement called Ethereum 2.0 (Eth2), implementing Sharding and introducing the Proof-of-Stake (PoS). As the change of consensus mechanism is a delicate matter, this improvement will be achieved through different phases, the first of which is the implementation of the Beacon Chain. As Ethereum1, Eth2 relies on a decentralized peer-to-peer (p2p) network for the message distribution. Up to date, there are around 4.000 nodes in the Eth2 main net geographically distributed. However, the topology of this one still prevails unknown.**

The main objective of this work is to present the results obtained from the analysis we performed on the Eth2 p2p network. Describing the topology of the network as possible hazards that this one implies.

I. INTRODUCTION

Ethereum [1] is a mature decentralized web infrastructure that has proved reliability and has led to the growth of decentralized applications. It features an ecosystem for decentralized applications based on a general-purpose virtual machine [2] and its dedicated programming language [3]. This ecosystem holds certain characteristics, which has generated a large community of developers that have proposed new methods to tackle some of the limitations affecting the Ethereum protocol.

The Ethereum Foundation and the Ethereum community have been working on developing the second generation of Ethereum, Ethereum 2.0 (Eth2). The approach proposed in the Eth2 protocol tackles both: the scalability challenge by introducing Sharding, and the sustainability challenge by changing the consensus protocol to Proof-of-Stake (PoS). This ambitious project is divided into different phases because of its complexity. The first phase (phase 0) is the Beacon Chain [4] launched on December 2020.

Eth2 aims to create a major network of nodes that will provide the infrastructure for Sharding. Over the last two years, the different developer teams have been working to generate Eth2 client implementations, each client targeting a specific type of user. This variety of clients, together with the low computing-power required by the Eth2 protocol, offers to most of the users the possibility to join the network.

In contrast to the Ethereum1 client, the Eth2 one is divided into two main parts, the Beacon Node, and the Validator. The beacon node, or node, performs all kinds of interactions with the Eth2 network. It establishes the connections with other nodes, exchanges the messages, downloads the blocks, and maintains the view of the Beacon Chain as the view of the Shards that the validator is performing at, building among them the entire network. On the other hand, the validator is in charge of the logical part, generating the attestations on the received blocks, as well as proposing the blocks when needed. To perform its duties, the validator needs to follow the current state of the Chain and be in constant communication with a Beacon Node that provides such information.

Given the Eth2 network requirements and the interaction between the clients, it is important to monitor critical points for possible attacks. In this poster, we present a complete analysis of the p2p network of the Eth2 main net. The analysis offers a general overview of the network composition, becoming the first analysis on the recently launched Eth2 main net.

The remainder of this poster is organized as follows. Section II discusses the background and related work. Section III explains the methodology of the study as the results obtained by our study. Finally, Section IV concludes this work and presents some possible future directions.

II. BACKGROUND

With the popularization of blockchain technology over the last decade, new distributed protocols based on blockchain technology appeared, e.g., Bitcoin and Ethereum. The impact of these decentralized protocols has augmented in the industrial and academic fields as shown in [5], triggering remarkable attention on the p2p networks that they rely on.

Previous works [6] [7] provide compelling insights into the actual p2p networks of the most prominent Blockchain protocols, Bitcoin and Ethereum. In these works, the authors show methods to track the number of nodes on the overlay, dig into the peer discovery over the network, and even some light message exchange of information among the peers.

The proposed approach leverages some previous ideas in [8], [9], and [10], such as network crawlers as a tool to exploit possible network weaknesses that might be a hazard for the integrity of the protocol. As explained in those papers, the information about the connections between the nodes that form the network is as important as the information about the nodes them-self. In this approach, we generate a study on the new Eth2 p2p network, including analyzing the discovered topology network and the client distribution.

III. ANALYSIS

The approach proposed in this poster tackles the lack of information about the performance of the p2p network and the peers on the Eth2 main net. The developed tool,

Armiarma [11], generates a custom peer for p2p networks, such as the Eth2 one. It offers a simple yet powerful method that gathers meaningful data about the peers compiling the p2p network in a PostgreSQL database. In this case, the collected data was used to compile the following analysis.

The machine hosting the crawler had 4 Intel Cores at 2.3GHz and 7GB of RAM, 50 GB of storage, and 250 Mbps of network bandwidth. The data used for the analysis corresponds to a data-set snapshot from March 11Th, 2022.
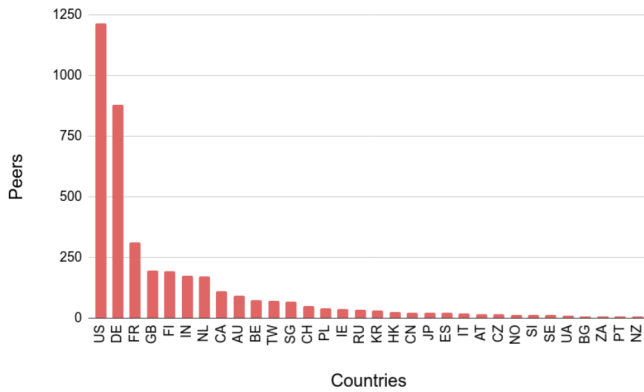

Fig. 1: Observed Geographical distribution.

The tool reported a total network size of around 4054 peers, identifying them in over 65 countries. In Figure 1, we can observe the total peer distribution in each country, where the US and Germany clearly out-stand the rest of the countries with 30% and 21% of total network.

Currently, there are six main Eth2 clients available for participating on the network. With our tool, we were able to detect which client is used by each connected peer, as shown in Figure 2. At the moment, Prysm is the most popular client with 58.8% of the total connected peers, followed by Lighthouse and Teku with 26% and 8.3% respectively.
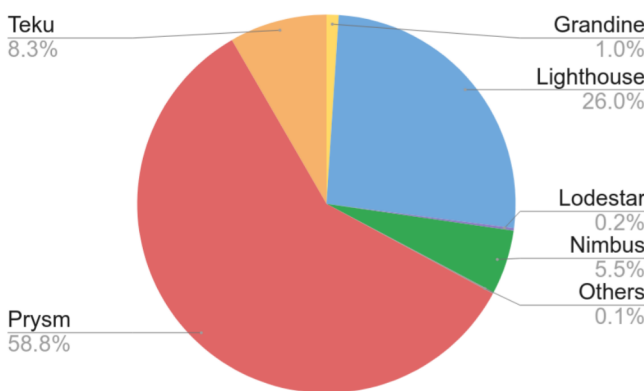

Fig. 2: Observed client distribution in the network.

IV. Conclusion And Future Work

The presented work displays that the network could be facing a critical mono-client dependency. The performed estimations show that about 59% of the network relies on the same client type, which could mean a finalization crisis if a vulnerability or bug is found in this particular client. Back on April 24th, 2021, a bug already impacted that specific client, causing finalization issues during multiple hours [12]. Therefore, This client dependency is one of the most critical points this crawler analysis highlights. The network is still very young, and Eth2 has not been fully deployed yet. Nonetheless, we believe it is crucial to boost the efforts to guarantee a much higher level of client decentralization.

Even though the presented analysis provided meaningful network insights about the Ethereum2 networks' topology, we are aware of the current limitations that the Armiarma tool has. Our future plan is to extend the analysis to include message exchange rates and upgrade the tool towards a distributed crawling strategy to perceive more accurate peer information.

References

[1] "Ethereum whitepaper," https://ethereum.org/en/whitepaper/.

[2] E. Hildenbrandt, M. Saxena, N. Rodrigues, X. Zhu, P. Daian, D. Guth, B. Moore, D. Park, Y. Zhang, A. Stefanescu *et al.*, "Kevm: A complete formal semantics of the ethereum virtual machine," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 204–217.

[3] M. Wohrer and U. Zdun, "Smart contracts: security patterns in the ethereum ecosystem and solidity," in *2018 International Workshop on Blockchain Oriented Software Engineering (IWBOSE)*. IEEE, 2018, pp. 2–8.

[4] "Phase-0," https://notes.ethereum.org/@djrtwo/Bkn3zpwxB.

[5] B. K. Mohanta, D. Jena, S. S. Panda, and S. Sobhanayak, "Blockchain technology: A survey on applications and security privacy challenges," *Internet of Things*, vol. 8, p. 100107, 2019.

[6] V. Deshpande, H. Badis, and L. George, "Btcmap: Mapping bitcoin peer-to-peer network topology," in *2018 IFIP/IEEE International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN)*. IEEE, 2018, pp. 1–6.

[7] T. Wang, C. Zhao, Q. Yang, and S. Zhang, "Ethna: Analyzing the underlying peer-to-peer network of the ethereum blockchain," *arXiv preprint arXiv:2010.01373*, 2020.

[8] S. Ben Mariem, P. Casas, and B. Donnet, "Vivisecting blockchain p2p networks: Unveiling the bitcoin ip network," in *ACM CoNEXT student workshop*, 2018.

[9] S. K. Kim, Z. Ma, S. Murali, J. Mason, A. Miller, and M. Bailey, "Measuring ethereum network peers," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 91–104.

[10] A. Biryukov, D. Khovratovich, and I. Pustogarov, "Deanonymisation of clients in bitcoin p2p network," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 15–29.

[11] "Armiarma repository;," https://github.com/migalabs/armiarma.

[12] "Eth2 mainnet incident retrospective," https://medium.com/prysmatic-labs/eth2-mainnet-incident-retrospective-f0338814340c.

Mikel Cortes-Goicoechea received his BSc degree in Industrial Electronics and Automation (2019) and his MSc degree in Embedded Systems (2021) from The University of the Basque Country (EHU), Spain. Since 2020, he has been with the Resilience in Distributed Systems group of Barcelona Supercomputing Center (BSC) and since 2021, a Ph.D. student at the Department of Telematic Engineering of Universitat Politècnica de Catalunya (UPC), Spain.

Reaching a modular, domain-agnostic and containerized development in Biomedical Natural Language Processing systems.

Javier Corvi*, Josep Lluis Gelpí*†, Salvador Capella-Gutierrez*

*Barcelona Supercomputing Center, Barcelona, Spain

†University of Barcelona, Barcelona, Spain

E-mail: {javier.corvi, josep.gelpi, salvador.capella} @bsc.es

Keywords—Natural Language Processing (NLP), Text mining, Toxicology, Preclinical, Biomaterials.

I. EXTENDED ABSTRACT

The last century saw an exponential increase in scientific publications in the biomedical domain. Despite the potential value of this knowledge; most of this data is only available as unstructured textual literature, which have limited their systematic access, use and exploitation. This limitation can be avoided, or at least mitigated, by relying on text mining techniques to automatically extract relevant data and structure it from textual documents. A significant challenge for scientific software applications, including Natural Language Processing (NLP) systems, consists in providing facilities to share, distribute and run such systems in a simple and convenient way. Software containers can host their own dependencies and auxiliary programs, isolating them from the execution environment. In addition, a workflow manager can be used for the automated orchestration and execution of the text mining pipelines.

Our work is focused in the study and design of new techniques and approaches to construct, develop, validate and deploy NLP components and workflows with sufficient genericity, scalability and interoperability allowing their use and instantiation across different domains. The results and techniques acquired will be applied in two main uses cases: the detection of relevant information from preclinical toxicological reports, under the eTRANSAFE project [1]; and the indexation of biomaterials publications with relevant concepts as part as the DEBBIE project.

A. PretoxTM

The treatment-related findings detected on the test subjects after the administration of the compound are the kind of relevant information included in the toxicology reports that is valuable for the drug development process. Examples of sentences with treatment-related findings are: *"The decrease in food consumption and body weight of the animals from the mid dose onwards is regarded as evidence of general toxicity."* and *"At dose level 3, absolute and relative liver weights were increased in male rats."*.

PretoxTM is a preclinical text mining system that extracts treatment-related findings from toxicological studies and present this information for toxicology experts validation. PretoxTM is being developed as part of eTRANSAFE, a research project funded within the Innovative Medicines Initiative (IMI), which aims at developing integrated databases and computational tools that support the translational safety assessment of new drugs.

In order to develop, train and validate text mining models for the detection of treatment-related findings, a gold standard corpus of preclinical studies annotated by toxicology experts was developed. The formal annotation guideline and the training material are available at the preclinical corpus development materials.

Figure 1.A shows the components of the PretoxTM pipeline. After the extraction of relevant paragraphs from the toxicology reports and the pre-processing steps; the PretoxTM sentence classifier is executed. The classifier was trained with the preclinical toxicology corpus and predicts if a sentence is relevant (contains information on toxicology findings) or not. The model is a maximum entropy classifier, was developed in JAVA and it uses the Stanford CoreNLP framework. The classifier obtains a performance of 0.91 F1-score. Following, the Named Entity Recognition (NER) and Relation Extraction (RE) components are executed to detect concepts and related them to form a toxicological finding. The NER is based on different controlled terminologies such as CDISC SEND (Standard for Exchange of Nonclinical Data). The last step of the pipeline is the mapping of the detected findings to the Study Report Domain (SR-Domain) format, an specific format that is being developed by another partner of the eTRANSAFE project.

The PretoxTM Web App (Figure 1.B) was developed to present the extracted information to the experts in order to provide a user-friendly interface to visualise and validate the findings.

B. DEBBIE

Biomaterials are natural or synthetic materials used for constructing artificial organs, fabricating prostheses, or replacing tissues. The DEBBIE (Database of Biomaterials and their Biological Effect) project aims at integrating biomaterials extracted metadata by storing in its database indexed articles with relevant concepts.

Figure 2 shows the architectural overview of the DEBBIE text mining pipeline. The first step in the pipeline entails the retrieval and standardization of published abstracts from PubMed. Following, is the recognition of relevant abstracts using the DEBBIE SVM model; which performs multiclass classification of abstracts to determine if they are relevant (either clinical or non-clinical studies) or not relevant to the field of biomaterials. The SVM DEBBIE model obtains
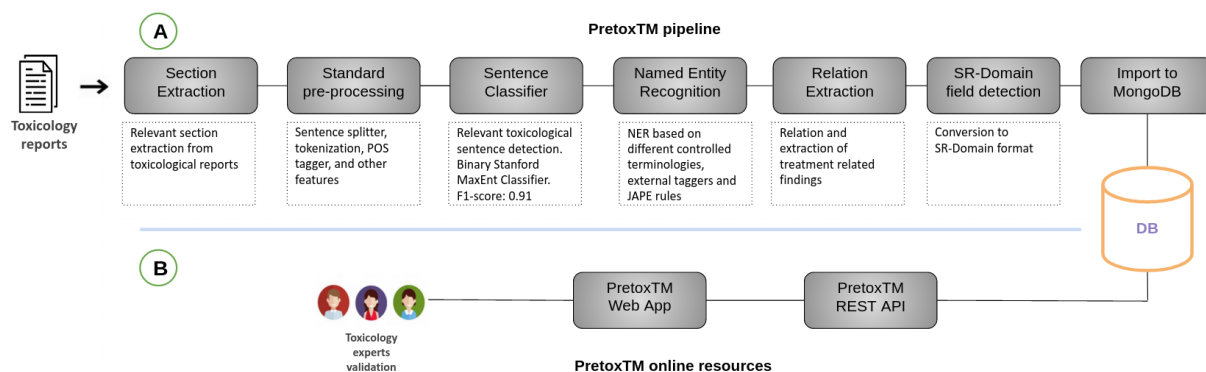
Fig. 1. A) PretoxTM pipeline components for the extraction for treatment-related findings. B) PretoxTM Web App components for expert validation.
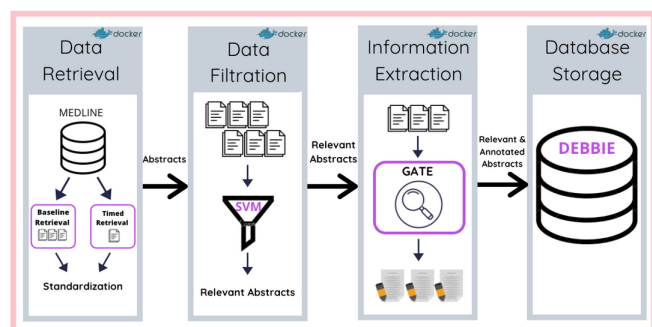

Fig. 2. DEBBIE automated text mining pipeline.

a performance of 0.93 F1-score. For the recognition of relevant biomaterials concepts, DEBBIE relies on the Bio-materials Annotator [2], a Named Entity Recognition (NER) system based on lexical resources specially tailored for the biomaterials literature [3]. The Biomaterials Annotator performs semantic mapping of the annotations by, not only recognizing the category of an entity, but also linking it to the appropriate entry in a well established resource. The Biomaterials Annotator shows a performance of 0.75-F1 score using a strict approach -exact matching of the term- and 0.79 F1-score in the partial-relaxed approach.

In order to access, visualize and retrieve the extracted information, a Web App (https://debbie.bsc.es/search/) was developed to allow the scientific community and general public to search the DEBBIE database using simple keyword queries, and receive results in the form of an intuitive dashboard.

II. ACKNOWLEDGMENT

REFERENCES

[1] F. Pognan, T. Steger-hartmann, C. Díaz, N. Blomberg, F. Bringezu, K. Briggs, G. Callegaro, S. Capella-gutierrez, E. Centeno, J. Corvi, P. Drew, W. C. Drewe, J. M. Fernández, L. I. Furlong, E. Guney, J. A. Kors, M. A. Mayer, M. Pastor, J. Piñero, J. M. Ramírez-anguita, F. Ronzano, P. Rowell, J. Saüch-pitarch, A. Valencia, B. van de Water, J. van der Lei, E. van Mulligen, and F. Sanz, "The etransafe project on translational safety assessment through integrative knowledge management: Achievements and perspectives," *Pharmaceuticals*, 2021.

[2] J. Corvi, C. Fuenteslópez, J. Fernández, J. Gelpi, M.-P. Ginebra, S. Capella-Guitierrez, and O. Hakimi, "The biomaterials annotator: a system for ontology-based concept annotation of biomaterials text," in *Proceedings of the Second Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 36–48. [Online]. Available: https://aclanthology.org/2021.sdp-1.5

[3] O. Hakimi, J. L. Gelpi, M. Krallinger, F. Curi, D. Repchevsky, and M. Ginebra, "The Devices, Experimental Scaffolds, and Biomaterials Ontology (DEB): A Tool for Mapping, Annotation, and Analysis of Biomaterials Data," *Advanced Functional Materials*, vol. 30, no. 16, p. 1909910, apr 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201909910

C. Conclusion and Future Steps

Generic and/or reusable NLP components were developed, encapsulated into software containers and integrated in the text mining workflows of PretoxTM and DEBBIE systems. A first version of the PretoxTM system to extract and visualise treatment-related findings from toxicology reports is already functional. Different state-of-the-art NLP techniques will be explored to enhance the NER and RE components of the PretoxTM pipeline using the preclinical annotated corpus to train the models. As part of the DEB-BIE project, the Biomaterials Annotator was published and different components were developed to build the DEBBIE text mining pipeline and the Web App.

Javier Corvi received the BCS degree from University of La Plata in 2011 and the MSc degree in Bioinformatics from University of Quilmes in 2020. Since 2018 he is a Bioinformatics researcher at the Spanish National Bioinformatics Institute (INB) Coordination Node, Life Sciences Department, Barcelona Supercomputing Center. He is currently doing a PhD in Biomedicine at the University of Barcelona focused on the development of modular, domain-agnostic and containerized Biomedical NLP systems.

Multi-model Forecast Quality Assessment of CMIP6 Decadal Predictions

Carlos Delgado-Torres[#1], Markus G. Donat[#*2], Albert Soret[#3]

[#]*Earth Science Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain*

[*]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

[1]`carlos.delgado@bsc.es`, [2]`markus.donat@bsc.es`, [3]`albert.soret@bsc.es`

EXTENDED ABSTRACT

A. Introduction

Decadal climate predictions are a new source of climate information for inter-annual to decadal time scales (filling the gap between seasonal predictions and climate projections), which is of increasing interest to users. The external forcings (natural and anthropogenic) and the internal climate variability (natural slow variations of the climate system) provide predictability on these time scales. However, due to chaotic characteristics of the climate system, it is not possible to predict its exact evolution. Thus, decadal forecasting provides large ensembles of predictions that, besides predicting the average anomalies based on the ensemble mean, are also used to obtain probabilistic information about the likelihood of certain event types.

Forecast quality assessment is essential to identify windows of opportunity (e.g., variables, regions, and lead times) with skill that can be used to develop a climate service and inform users in specific sectors. Besides, it can help to monitor improvements in current forecast systems. The forecast quality assessment needs to be carried out over a long enough period in the past (when observations are available to compare against) to achieve robust results that can be used as an estimate of how well the forecast system may perform in simulating future climatic anomalies. Thus, retrospective decadal forecasts (also known as hindcasts) are performed with the same forecast systems used to predict future climate variations. For this, the forecast systems are utilized to simulate the evolution of the climate system from our best estimate of the observed initial state, which is referred to as forecast system initialization and the predictions also incorporate information about the external forcings. The hindcasts are also used to apply calibration techniques to partially correct systematic biases of the predictions.

The Decadal Climate Prediction Project (DCPP [1]) of the Coupled Model Intercom-parison Project Phase 6 (CMIP6 [2]) now provides the most comprehensive set of retrospective decadal predictions from multiple forecast systems. The increasing availability of these simulations leads to the question of how to best post-process the raw output from the forecast systems so that the most useful and reliable information is provided to users.

B. Data and Methods

This work evaluates the quality of deterministic and probabilistic forecasts for spatial fields of near-surface air temperature and precipitation generated from all the available decadal predictions contributing to CMIP6/DCPP-A (169 members from 13 forecast systems). The analysis focuses on the predictions for the average of forecast years 1 to 5 during the 1961-2014 period, using the 1981-2010 period as reference for computing the climatology and probabilistic categories. In addition to decadal predictions, the retrospective climate projections (i.e., the uninitialized historical forcing simulations; 195 members from the same forecast systems) are used as benchmark to estimate the impact of model initialisation.

The quality of the deterministic forecasts is estimated with the Anomaly Correlation Coefficient (ACC), while the Ranked Probabilities Skill Score (RPSS) is used to evaluate the probabilistic forecasts (created for tercile categories, i.e., below-normal, normal and above-normal conditions).

C. Results

The DCPP multi-model generally shows high skill in predicting temperature, particularly over land regions (Figure 1). The skill is more limited for precipitation, being significant over some regions like Central Africa and parts of Asia. Also, the comparison against the climatological forecast (defined as equal probabilities for all the categories) shows an added value of using the decadal predictions for temperature.
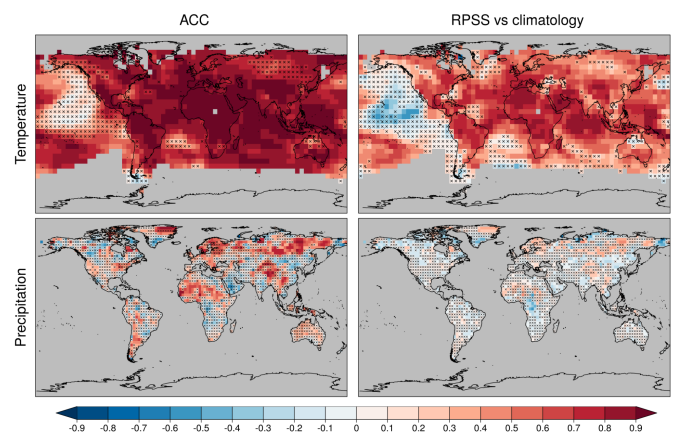

Fig. 1. DCPP multi-model forecast quality for the forecast years 1-5.

In order to assess the benefits and drawbacks of using a multi-model forecast, the multi-model ensemble is compared to the individual forecast systems. To do this, the skill of all forecast systems is first computed. Then, the systems that provide the highest and the median skill are selected for each grid point. The best system usually provides the highest skill (Figure 2). However, the multi-model ensemble is more skilful than at least 50% of the individual systems (not shown). Thus, the multi-model is a reasonable choice for not selecting the best system for each particular variable, forecast period and region, making a potential operational forecast generation more straightforward.
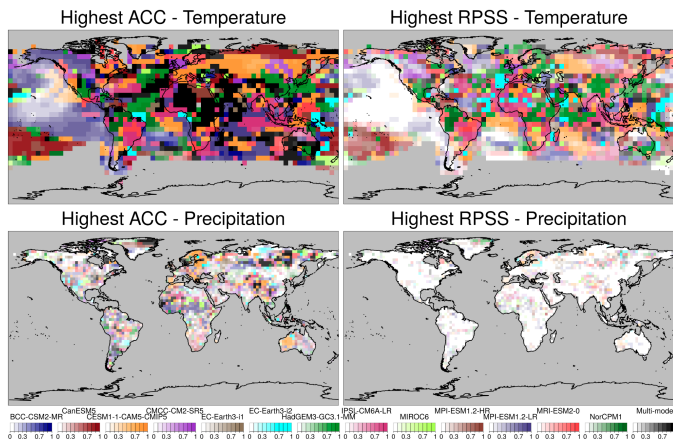
Fig. 2. Highest skill among the individual forecast systems and multi-model for the forecast years 1-5.

The decadal predictions are also compared to the uninitialized historical climate simulations (HIST) to estimate the impact of initialization. For the deterministic forecasts, an added value is found for temperature over several ocean and land regions (Figure 3). For precipitation, the added value is found to be more reduced, although the residual correlation shows significant positive values over several regions meaning that the decadal predictions capture more observed variability than the uninitialised historical simulations. For the probabilistic forecasts, the North Atlantic Ocean is the region that shows the highest added value for temperature.
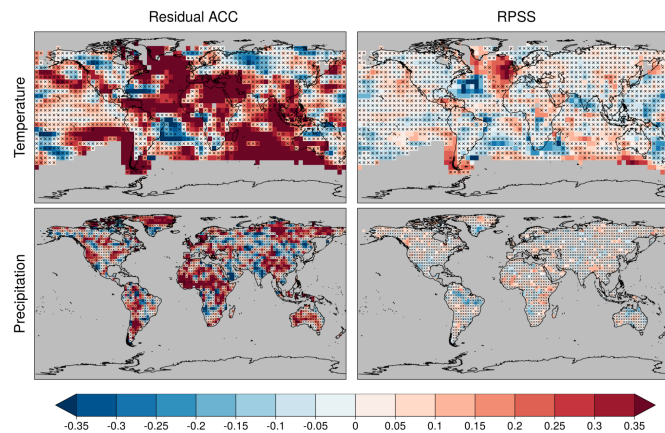

Fig. 3. DCPP multi-model forecast quality in comparison to the HIST multi-model ensemble for the forecast years 1-5.

D. Conclusions

The CMIP6/DCPP multi-model ensemble shows generally high skill in predicting temperature, being lower for precipitation. The comparison between the multi-model ensemble and the individual forecast systems indicates that the best system provides the highest forecast quality for a particular region, variable and forecast period, which would be the best option for a specific climate service. Yet, the multi-model ensemble shows higher skill than, at least, 50% of the systems, which makes it a reasonable choice for operational forecast product generation. The decadal predictions show an added value with respect to the uninitialised historical forcing simulations for temperature and precipitation over several regions.

A paper on this study has been accepted for publication in Journal of Climate [3], in which the impact of calibrating the raw predictions is also assessed, and the comparison of two multi-model ensemble sizes. Besides, the results are also shown for the forecast of the Atlantic Multidecadal Variability (AMV) index and Global Surface Air Temperature (GSAT) anomalies.

References

[1] Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

[2] Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016, 2016.

[3] Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L.-P., Athanasiadis, P. J., Bretonnière, P.-A., Dunstone, N., Ho, A.-C., Pankatz, K., Paxian, A., Pérez-Zanón, N., Samsó Cabré, M., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J.: Multi-model forecast quality assessment of CMIP6 decadal predictions. Accepted in Journal of Climate.

Author biography

Carlos Delgado Torres is a PhD student at the Climate Prediction and Earth System Services Groups of the Earth Science Department. He received a BSc degree in Physics and a MSc degree in Meteorology and Geophysics at the Universidad Complutense de Madrid. His work focuses on the applicability of decadal predictions for climate services.

A platform for antibody design

Anna Maria Díaz-Rovira[#1], Víctor Guallar [#&2]

#Life Science Department, Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell 1-3, Barcelona, Spain
&ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
[1]`anna.diazrovira@bsc.es`, [2]`victor.guallar@bsc.es`

Keywords— **antibody design, protein engineering, protein–protein docking, computer-aided antibody protocol**

EXTENDED ABSTRACT

Antibodies are specialized proteins produced by our adaptive immune system to identify and neutralize specific molecules (antigens) of foreign objects, such as pathogenic microorganisms or infected cells. Due to their chemical specificity and sensitivity, antibodies are widely used in biomedical research and to treat many diseases due to their high efficiency and low risk of adverse events. Nowadays, it is possible to produce a wide variety of artificial antibodies; however, the process of obtaining therapeutic antibodies remains empirical and time-consuming with low success rates. The structural and energetic information of the antibody-antigen complexes, together with computational simulations, complement the experimental results to work towards a more effective and faster "a la carte" antibody design. In this framework, the project aims to develop a general platform that assists in the antibody design and helps the user overcome common challenges such as the lack of structural data of the antibody and the antibody-antigen complex and the flexibility of the antibody complementary determining regions (CDRs).

The platform workflow (Fig. 1) will consist of the following steps. First, if there is no structural data available, DeepAb [1] will be used to generate the 3D structure of the antibody from its amino acid sequence; otherwise, we can jump to the minimization step. In the minimization step, the system will be relaxed to either remove clashes, if it is a predicted structure, or remove crystal artifacts, if we dispose of a crystal structure. Subsequently, we can jump to the point mutation step if we already have information on how the antibody binds the antigen (e.g. crystal of the antibody-antigen complex). Otherwise, we will run the in-house Consensus Docking [2], a pipeline designed to cluster protein-protein poses (PPPs) obtained using several rigid-body protein docking methodologies, which has shown to have a higher success rate than using the docking methodologies individually. This technique extracts 20-30 representative structures from the top 5 most populated clusters that fulfill the consensus among all the docking methodologies used. Then these structures will be refined with the in-house Monte Carlo software, PELE from Protein Energy Landscape Exploration [3], to study if they tend to converge into 2-5 stable binding structures and use these structures to start the affinity maturation of the antibody towards that antigen.

The affinity maturation step will mainly consist on using UEP [4], a fast and accurate in-house predictor of detrimental and beneficial point mutations of the binding energy of protein-protein interactions. For the screening of mutations, we will use the 2-5 binding poses found by PELE or the crystal structure of the complex if it is available. The selection of the most beneficial point mutations identified by UEP will be introduced in the binding structures to perform a local PELE refinement with an exhaustive sampling on the CDRs loops. This refinement will allow the study of the local conformational changes due to the mutation and evaluate the binding energy between antibody-antigen and the system's total energy. Finally, the workflow will output the refined structures with the most beneficial point mutations and a report with the energetic profiles for each mutation.

This protocol will be validated in the future while improving the stability and binding affinity of industry-relevant antibodies.
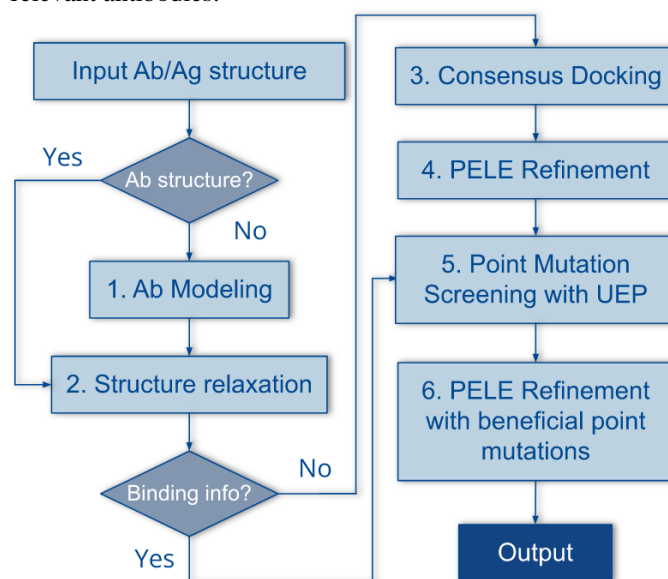
Fig. 1 Platform workflow scheme. Abbreviations: Ab – Antibody, Ag – Antigen, PELE – Protein Energy Landscape Exploration.

References

[1] Ruffolo, J. A.; Sulam, J.; & Gray, J. J. *Patterns* **2022**, 3(2), 100406.
[2] Díaz-Rovira, A. M.; Malo, L. *Consensus Docking*; 2022. https://github.com/laumalo/ConsensusDocking
[3] Borrelli, K. W.; Vitalis, A.; Alcantara, R.; Guallar, V. *J. Chem. Theory Comput.* **2005**, *1* (6), 1304–1311.
[4] Amengual-Rigo, P.; Fernández-Recio, J.; Guallar, V. *Bioinformatics* **2021**, *37* (3), 334–341.

Author biography

Anna Maria Díaz-Rovira was born in Caldes de Montbui, Spain, in 1997. She received a BSc. degree in Physics and a BSc. degree in Chemistry from the Autonomous University of Barcelona, Spain, in 2020. Then she pursued an MSc. in Bioinformatics at the same university, which led her to enroll in Barcelona Supercomputing Center, where she recently started her doctoral studies in antibody modeling.

Since her early years in the degree, she showed great interest in applying scientific knowledge in real case scenarios. That is why, during her time at the university, she collaborated with research institutions such as the Autonomous University of Barcelona's Chemical Department, *Institut de Física d'Altes Energies*, ALBA Synchrotron, Barcelona Supercomputing Center, and Henkel company.

PredIG: a predictor of T-cell immunogenicity

R. Farriol-Duran[#1], M. Vazquez[*2], E. Porta[2], V. Guallar[#3]

[#] *Electronic and Atomic Protein Modelling, Life Sciences Department, Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona (Spain)*

[1]`roc.farriol@bsc.es`, [3]`victor.guallar@bsc.es`

[*]Cancer Immunogenomics, Josep Carreras Leukaemia Research Institute, Josep Carreras Building, Ctra de Can Ruti, Camí de les Escoles, s/n, 08916 Badalona, Barcelona

[2]`eduard.porta@bsc.es`

Keywords— **T-cell immunogenicity, HLA & TCR**

EXTENDED ABSTRACT

The identification of immunogenic epitopes (such as fragments of proteins, in particular peptides, that can trigger an immune response) is a fundamental need for immune-based therapies. A computational tool that could predict such immunogenic epitopes would have vast potential applications in biomedicine ranging, from vaccine design against viruses or bacteria to therapeutic vaccination of cancer patients. While there are several methods that predict whether a peptide will be shown to the immune system via the Human Leukocyte Antigen (HLA) proteins of a patient, most of them cannot predict whether such presentation will indeed trigger an immune response. Additionally, T-cell immunogenicity is determined by multiple cellular processes, some of which are often overlooked by the current state-of-the-art immunogenicity predictors.

The aim of this project is to build PredIG, an immunogenicity predictor that discriminates immunogenic from non-immunogenic T-cell epitopes given the peptide sequence and the HLA typing. After a careful study of the drivers of antigen processing and presentation on HLA class I molecules and an assessment of the physicochemical factors influencing epitope recognition by T-cell receptors (TCRs), we have used a selection of publicly available tools and in-house developed algorithms to identify the most relevant features that determine epitope immunogenicity.

We then used these features to build an immunogenicity predictor (PredIG) modelled by XGBoost against immunogenically validated epitopes by the ImmunoEpitope DataBase (IEDB)(1), the PRIME dataset(2) and the TANTIGEN database(3). Pondering the feature importance in the model, the in-house developed softwares, NOAH for HLA Binding Affinity and NetCleave for Proteasomal Processing were identified as the major contributors to the performance of the model.

Once PredIG was developed, we benchmarked the capacity to predict the immunogenicity of validated T-cell epitopes versus a set of state-of-the-art methods (Fig.1). Relevantly, PredIG showed a greater performance than the Immunogenicity predictors from Prime(2) and IEDB(4). Additionally, our results confirm that predicting T-cell immunogenicity based on data from T-cell assays is more accurate than using HLA Binding assays, the method mostly used in the field. An AUC value of 0.67 and an enrichment factor in the TOP10 epitopes of 90% outperforms the predictive performance of the available methods.

In the context of the immune response against cancers, T-cell immunogenicity of tumoral mutations has been described as a response biomarker for immunotherapies such as immune checkpoint inhibitors. Similarly, the presence of immune infiltrate in a tumor has been related to a better prognosis for many cancer types. What is missing is the link between T-cell immunogenicity of tumoral mutations and the capacity of a tumor to attract immune cells. For this reason, we correlated the PredIG immunogenicity score obtained in a dataset of the The Cancer Genome Atlas (TCGA) against the tumor infiltrate in such tumors demonstrating that rather the total number of mutations a tumor accumulates, the tumor mutation burden (TMB), it is the number of immunogenic mutations what should be accounted for as biomarker of response.
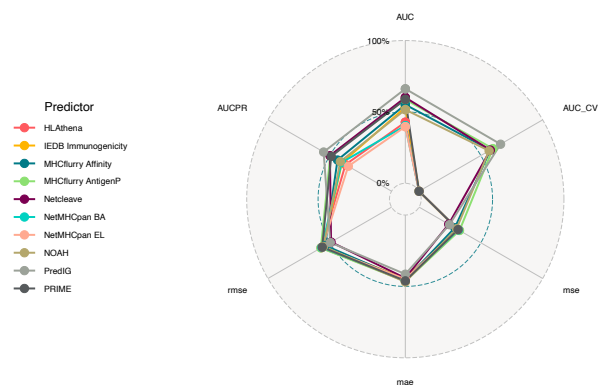

Fig. 1 Radar plot comparing the performance metrics of PredIG versus the benchmarked methods of HLA Binding Affinity (HLAthena(5), MHCflurry(6), NetMHCpan BA & EL(7), NOAH), HLA presentation (MHCflurry Antigen Presentation), Proteasomal Processing (Netcleave)(8) and Immunogenicity (Prime, IEDB immunogenicity). AUC = Area under the ROC Curve. AUC_CV = CrossValidated AUC. RMSE = Root Mean Square Error. MSE = Mean Square Error. MAE = Mean Absoulte Error. AUCPR = Area under the Precision-Recall Curve.

PredIG has shown an increased performance in the prediction of T-cell immunogenicity than the benchmarked state-of-the art methods. Therefore, it can be a useful resource to prioritize T-cell epitopes for their potential capacity to activate an immune response, a functionality that can be applied in vaccine design, cancer immunotherapies or deimmunization of therapeutic proteins.

ACKNOWLEDGEMENTS

References

1.   W. Fleri, *et al.*, The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Front. Immunol.* **8**, 278 (2017).

2.   J. Schmidt, *et al.*, Prediction of neo-epitope immunogenicity reveals TCR recognition

determinants and provides insight into immunoediting. *Cell Reports Med.* **2**, 100194 (2021).

3. L. R. Olsen, *et al.*, TANTIGEN: a comprehensive database of tumor T cell antigens. *Cancer Immunol. Immunother.* **66**, 731–735 (2017).

4. J. J. A. Calis, *et al.*, Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).

5. S. Sarkizova, *et al.*, A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* (2019) https:/doi.org/10.1038/s41587-019-0322-9 (December 23, 2019).

6. T. J. O'Donnell, A. Rubinsteyn, U. Laserson, MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* **11**, 42-48.e7 (2020).

7. V. Jurtz, *et al.*, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

8. P. Amengual-Rigo, V. Guallar, NetCleave: an open-source algorithm for predicting C-terminal antigen processing for MHC-I and MHC-II. *Sci. Reports 2021 111* **11**, 1–8 (2021).

Author biography

Roc Farriol-Duran was born in Sabadell, Catalunya, in 1995. He received the B.E. degree in Biomedical Sciences from the Autonomous University of Barcelona, Barcelona, Catalunya, in 2018, the Msc. degree in Translational Biomedical Research for the Vall d'Hebron Hospital Campus (VH) Barcelona, Catalunya, in 2019 and the Msc. Degree in Omics Data Analysis for the Vic University, Vic, Catalunya.

During 2016-17 he stayed at the Proteomics Group at the Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands. Followingly, he joined the Cellular Immunology Laboratory at the Institute for Biotechnology and Biomedicine, Autonomous University of Barcelona and collaborated closely with the Proteomics Unit at the CSIC-UAB. Then, he joined the lab of Tumor Immunology and Immunotherapy at the Vall d'Hebron Oncology Institute (VHIO). Since May 2020 he has been at the Life Sciences Deparment of the Barcelona Supercomputing Center, first at the Computational Biology group and later at the Electronic and Atomic Protein Modelling group. Currently, he is pursuing a PhD degree in Computational Modelling of T-cell and B-cell immunogenicity.

Designing a Cloud and HPC Based M&S platform to Investigate the IVD Diseases Mechanisms

Maria Paola Ferri[*1#2], Laia Codo[*1], Josep Lluis Gelpi[*1#2]

[#1]*Life Sciences Department, Universitat de Barcelona, Address Including Country Name*
mferrife108@alumnes.ub.edu, gelpi@ub.edu

[*2]*Barcelona Supercomputing Center, Plaza de Usebi Guell 1-3, Barcelona, Spain*
maria.ferri@bsc.es, laia.codo@bsc.es, josep.gelpi@bsc.es

Keywords— **BioBB, Intervertebral Disc Degeneration, Workflow Framework, Portable Platform**

I. Extended Abstract

The main objective of the PhD proposal consists of the creation of a platform for IVD Models & Simulations (M&S) tools and their integration into automated workflows, within the HORIZON MSCA Disc4All project.

Based on the European Open Science Cloud (EOSC) vision, the expected platform will be a Cloud-based one, furnished with a front-end, to guarantee reproducibility, accessibility and easy use for experts and non-experts.

The development of an automated and specialized platform can represent the best hybrid technology with perks on both healthcare data management and computational environments exploitation, given the use of Cloud infrastructures on healthcare software and databases. Though rendering automatic not only the database, but prediction and simulation models in a user-friendly integrated system, may facilitate a difficult diagnosis and forward therapy, especially considering the various forces at play in a multi-omics data analysis of its kind.

A. Introduction

LBP (Low Back Pain) is the largest cause of morbidity worldwide, yet there remains undetected the liable specific cause, leading to poor treatment options and prognosis. As for other highly multi-factorial musculoskeletal (MS) disorders, such as lumbar intervertebral disc (IVD) and Lumbar Disc Degeneration (LDD), the interplay of a wide range of factors is poorly understood. Some genetic factors have been identified as possible biomarkers of specific LDD phenotype, such as Interleukin (IL)-, ECM protein- and protease-related genes, as well as human genetic variants (GDF5, SKT, PARK2 and CHST3)[1]. Beyond molecular data, also nutritional, clinical, social conditions are at stake for such diagnosis: molecular profiling, medical imaging, lifestyle data, sex are crucial factors to determine the predisposition of a single patient to these multi-factorial disorders[2].

Unfortunately, the integration of such data into a holistic and rational map of degenerative processes and risk factors has not been achieved, requiring the creation of professional cross competencies training programs.

The HORIZON MSCA Disc4All project is proposing to provide for an interdisciplinary solution to make up for the lack of diagnosis tools and instruments, considering all the biological, medical and computational contributions. The curation and integration of all the heterogeneous data sets, as well as experimental and theoretical/computational template and algorithms, are going to be embedded to exploit the multidisciplinary and multiscale models and simulations tools, based on image analysis, biophysics and biology[14].

B. Methods

The ultimate purpose of this Disc4All final infrastructure, deployed on the biological, bioinformatics algorithms, image analysis and ML/AI models and tools, is to integrate and interpolate primary patient data and achieve a coherent attribution of a MS phenotype.

To get a homogeneous configuration between all the tools, the first and main aim of the project is to contain all the workflow tools into a similar framework. This would be reached through the use of the BioExcel Building Blocks (BioBBs)[3], a collection of Python wrappers on top of biomolecular/ bioinformatics simulation tools. With their mutual unique syntax, they offer a layer of interoperability between the wrapped tools, to make their interconnection easier in the building process of a biomolecular/bioinformatics workflow.

The BioBBs also allow singular deployment of each tool within a Docker bio-container, but will also be assessed a private Cloud resource and its own Disc4All instantiation.

Sharing this technology, the tools produced for the Disc4all project would be effortlessly combined in the most convenient and comprehensive way, through the CWL (Common Workflow Language) or any other convenient workflow manager language.
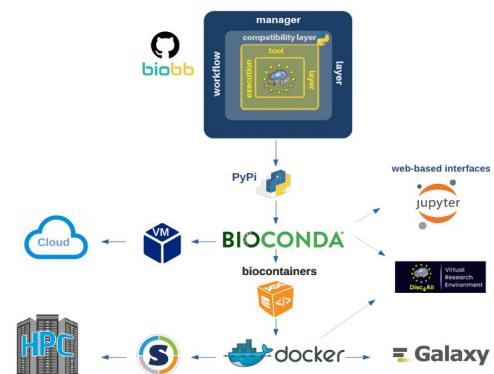

Fig. 1 BioBB's Automated deployment strategy[3]

In the end, thanks to the BioBBs adaptation layer, there are many possibilities for exposing the BioBB based workflow, such as Galaxy, Jupyter Notebooks, etc. One of the strategies for the front-end display is the web-server VRE (Virtual Research Environment)[3]. It will be exploited to construct the front-end of the infrastructure, so to provide for the non-expert users an easy-to-use platform to perform their analysis with the containerized pipeline.

C. Discussion and Results

The platform will provide the necessary environment to develop workflows using the tools developed in the Disc4All project.

The current objective for the development of the platform is the adaptation of the tools and software, given their computation and data requirements, into BioBBs. Since more of the half of the tools are in development, the best course of action is to start for each of them a single BioBB instantiation, so as to first define the basic pre-processing steps and environment. This introductory BioBBs are in production, one for Genomics data retrieval and the other one for a pre-Learning Active Cycle extracting features from imaging data.

Different modular Cloud containers are deployed for each BioBB, to have independent environments that could be possibly concatenated, following official FAIR workflow standards.

On one of them, the first draft for the final front-end platform is instantiated and exposed, in compliance with the openVRE environment. It is furnished with a simple case-use of the project (which is the aforementioned pre-processing BioBB data retrieval), that is up and working. This preliminary interface then is going to be completed with the rest of the tools, that could be called individually, and with the integral pipeline.

D. Future Enhancement

In the foreseeable future, the work on the adaptation to the BioBB framework will continue with the advancement of the tools, as well as the exploration of federated data access strategies with the furnished data model, for smooth flow of information through the platform, as well as exploring other data models adapters.

For the long term results, should be taken due account the different types of instantiations on Cloud-based and HPC environments, to build the computational infrastructure through Docker & EOSC cloud-based services, such as BSC's MareNostrum or StarLife (for users with granted HPC credentials), and to be able to probe the perks of hybrid workflows.

The front-end is a key point for the users (end-users and developers), to render available the workflow in the appropriate computational environments, so its finalization would be carried out in conformity with the progression of the study.

E. ACKNOWLEDGMENTS

References

[1] Gantenbein B, May RD, Bermudez-Lekerika P, Oswald KAC, Benneker LM, Albers CE (2021) EGR2, IGF1 and IL6 Expression Are Elevated in the Intervertebral Disc of Patients Suffering from Diffuse Idiopathic Skeletal Hyperostosis (DISH) Compared to Degenerative or Trauma Discs. Applied Sciences 11(9): https://doi.org/10.3390/app11094072

[2] Eskola, P. J. et al. Gender difference in genetic association between IL1A variant and early lumbar disc degeneration: A three-year follow-up. Int. J. Mol. Epidemiol. Genet. 3, 195–204 (2012)

[3] BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. Pau Andrio, Adam Hospital, Javier Conejero, Luis Jordá, Marc Del Pino, Laia Codo, Stian Soiland-Reyes, Carole Goble, Daniele Lezzi, Rosa M. Badia, Modesto Orozco & Josep Ll. Gelpi. Nature Scientific Data, 09/2019,Volume 6, Issue 1, p.169, (2019).

[4] Towards FAIR principles for research software, Lamprecht Anna-Lena, Garcia Leyla, Kuzak Mateusz, Martinez Carlos, Arcila Ricardo, Martin Del Pico Eva, Dominguez Del Angel Victoria, van de Sandt Stephanie, Ison Jon, Martinez Paula Andrea, McQuilton Peter, Valencia Alfonso. Harrow Jennifer. Psomopoulos Fotis, Gelpi Josep Ll., Chue Hong Neil, Goble Caroleu, Capella-Gutierrez, Salvador. 10.3233/DS-190026. Data Science, vol. Pre-press, no. Pre-press, pp. 1-23, 2019

[5] Felipe da Veiga Leprevost, Björn A Grüning, Saulo Alves Aflitos, Hannes L Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I Nesvizhskii, Yasset Perez-Riverol, BioContainers: an open-source and community-driven framework for software standardization, *Bioinformatics*, Volume 33, Issue 16, 15 August 2017, Pages 2580–2582, https://doi.org/10.1093/bioinformatics/btx192

Author biography

Maria Paola Ferri was born in Rome, Italy, in 1995. She graduated in 2018 in Biotechnologies, from the University of Rome La Sapienza, and in 2021 she obtained the Internation Master degree in Bioinformatics from the Alma Mater University of Bolumn. Her master thesis was focused on the creation of Cloud architectures designed to support and instantiate Bioinformatics applications, curing the portability and the user-friendly access to them. Right now, she is holding a position in the Life Science's Department at the BSC (Barcelona Supercomputing Center). Her role in the Disc4All project would be to extend the implementation of the IDD identification and prediction software in development into a Cloud and HPC environment, so to create in the end a portable and ready-to-use front-end workflow in the LBP (low back pain) investigation, for user and non-experts.

Improving resource usage in large FPGA accelerators

Antonio Filgueras*†, Carlos Álvarez*†, Daniel Jiménez-González*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {antonio.filgueras, daniel.jimenez, carlos.alvarez}@bsc.es

Keywords—FPGA, High-performance computing, Heterogeneous computing.

I. EXTENDED ABSTRACT

A. Introduction

In modern FPGA devices, place and route has become a difficult task for the underlying FPGA implementation tools. This is caused by an increase of device size and complexity. As devices grow in size and number of resources, their topology also grows in complexity. Larger devices are divided in different regions. While this allows to pack a larger number of resources in a single device, it creates a new set of challenges in order to obtain good quality of results while using as many resources as possible. Devices such as Xilinx's Alveo accelerators are comprised of multiple regions called Super Logic Regions (SLR). Crossing from one region to another adds some delay to signal propagation. This can hurt overall timing if implementation tool decides to scatter a single accelerator among different SLRs. Thus, the design may not reach operating frequencies expected by the user. In a similar fashion as the SLRs, they usually have multiple independent memory banks that interface with DDR modules. This requires memory allocations and interconnection to be manually managed by the user, causing extra burden to users. Otherwise, the design will not be able to take profit of the aggregated available bandwidth.

We propose methods to improve resource and bandwidth usage that allow a user to direct how a design is built and implemented while maintaining device abstraction and minimal development overhead.

B. Design

1) Memory access: In order to address memory access contention, data allocations are scattered across device's different memory banks. This is done by adding interleaver modules between accelerators or PCIe and the memory interconnection. These modules are shown in the right diagram of figure 1 as the highlighted blocks. These modules scatter memory accesses across all available memory modules. This allows memory accesses to a single piece of data to be performed in parallel by different accelerators as different parts of the same data structure can be stored in different banks. This increases the overall available bandwidth.

However, early designs consisted of two level interconnection, as shown in the left diagram of figure 1. While this allowed reducing resource usage due to each module being much simpler, it also prevents accesses from being performed

in parallel due as a single bus is connecting both interconnection modules. In order to solve this issue, interconnection has been reworked so it can be implemented in single stage as shown on right hand side of figure 1. Then high performance configuration options can be enabled to allow parallel accesses to different memory modules.
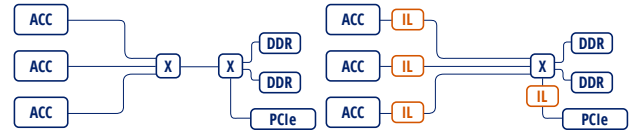

Fig. 1: Interconnection modes: 2 stage vs. single stage with interleave modules.

2) Placement: In order to prevent accelerators to be placed among different SLRs and properly implement region crossing, users are allowed to assign accelerator instances to SLR regions. Implementation tool is then instructed not to allow any resource belonging to an accelerator to be placed outside the region assigned by the user. Along with this, register slices are then properly inserted between accelerator interfaces and interconnection infrastructure when needed. This improves timing as region crossings are done in a controlled fashion.

Figure 2a shows a diagram of a design containing multiple accelerators. Each of them is restricted to its region, and SLR crossings are done by special register slices.
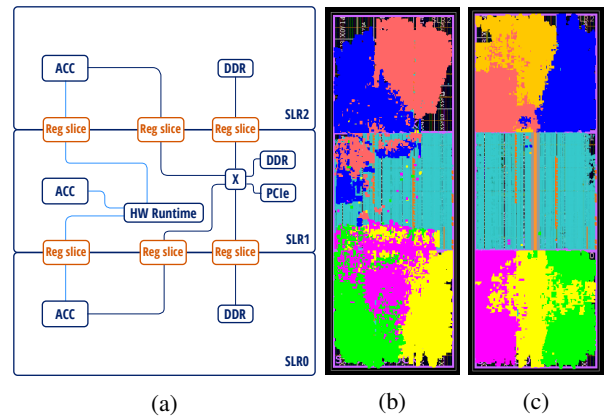

Fig. 2: Placed accelerator diagram and implemented designs.

Figures 2b and 2c show a physical view of the device without (2b) and with (2c) placement options enabled. Highlighted cells correspond to resources belonging to different accelerator instances. Note that the highlighted cells belonging

to accelerators are not scattered among different SLRs in figure 2c. Also note that device resource usage is higher and we can fit one extra accelerator therefore increasing application performance.

C. Evaluation

1) Experimental setup: Benchmarks used for evaluation have been implemented using OmpSs@FPGA and Vivado HLS pragmas. Experiments are run a Xilinx Alveo U200 (XCU200-FSGD2104). This card is attached to a host via PCIe. The host itself consists of dual Intel Xeon CPU X5680 @ 3.33GHz.

2) Benchmarks: To perform the evaluation, we used two different applications that cover a wide range of characteristics of HPC benchmarks: a matrix multiplication (Matmul) and a Cholesky decomposition. All applications have been implemented by dividing computation in smaller tiles so they can be efficiently implemented in an FPGA. All applications use single precision floating point data.

The matrix multiplication benchmark is a well-known embarrassingly parallel application with a regular dependence pattern. The application operates with three square matrices of size $N \times N$, A, B and C, and computes $C = C + A \times B$.

The Cholesky benchmark performs a Cholesky decomposition of a Hermian, positive definite matrix into a lower triangular matrix, which multiplied by its transpose results in the original matrix. I.e., the application generates an output matrix L from an input C, assuring that $C = L \times L^T$ providing that C fulfils the restrictions. The code uses four kernels: *gemm*, *trsm*, *syrk* and *potrf*. In addition, the particularity of this benchmark is that the potrf kernel is hard to accelerate in FPGA due to its memory access pattern. Previous evaluations [1] demonstrate that it is faster to execute it in the host rather than in the FPGA. We use the OpenBLAS implementation of the kernel, which is in fact a Cholesky decomposition of a single block.

3) Results: Figure 3 shows how proposed features affect these applications. It shows performance (in GFLOPS) of the applications in three different cases. First, without using any of the proposed features (Baseline), then using the proposed memory access improvements (Memory) and finally using both improvements in memory access as well as placement (Mem & placement).

In the Cholesky application, bandwidth increase due to interconnection and interleave features have not a dramatic impact. This is due to the application complexity as it involves different kernels with a critical path more demanding than the Matmul application Also one of them (*potrf*) is run in the main CPU. We are also able to fit two more *gemm* accelerators in the improved version compared to baseline or memory improved version due to the improvements regarding accelerator placement. This further increases performance, but as is the case with memory related optimizations, complex application structure prevents performance improvements from being dramatic even though performance increase is still significant (18%).

In Matrix multiply we can see a big performance increase (50%). This is caused, on one hand, due to increased available
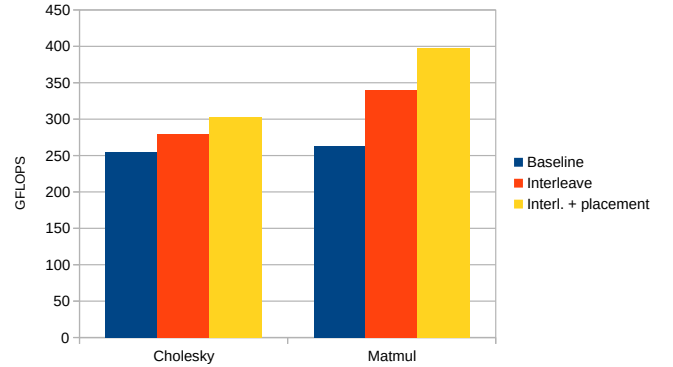

Fig. 3: Application performance.

bandwidth that memory access interleaving and single level interconnection provide. This result is significant not only by itself but because it paves way to even more improvements in future prototypes that use new memory subsystems like HBM. On the other hand, we were able to place two more accelerator instances in the placement improved version, increasing usage of available computational resources, which allows further performance increase over the memory improved version.

D. Conclusion

This paper presents an extension of the OmpSs@FPGA ecosystem in order to more efficiently use available resources. Proposed extensions allow better use of available memory bandwidth through automated memory access interleaving. It also allows improved computational resource usage through accelerator placement improvements. Moreover this is done without causing extra burden to the user and abstracting low level architectural features.

II. ACKNOWLEDGMENT

REFERENCES

[1] J. M. de Haro, J. Bosch, A. Filgueras, M. Vidal, D. Jiménez-González, C. Álvarez, X. Martorell, E. Ayguadé, and J. Labarta, "Ompss@fpga framework for high performance fpga computing," *IEEE Transactions on Computers*, vol. 70, no. 12, pp. 2029–2042, 2021.

Antonio Filgueras earned a degree in Computer Science from Universitat Politècnica de Catalunya (UPC) in 2012. Currently, he's a PhD student in the departemt of Computer Architecture at UPC and working in the OmpSs@FPGA team within the Programming Models group at BSC, focused primarily on programming models for reconfigurable systems in High Performance Computing. He has been a researcher in the AXIOM, Legato, EuroEXA and MEEP European projects among others.

Foreign and domestic contributions to surface ozone in Spain

Roger Garatachea Solé[1*], Oriol Jorba[1], Carlos Pérez García Pando[1,2]

[1]*Barcelona Supercomputing Center, Barcelona, Spain*

[2]*ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain*

*`*roger.garatachea@bsc.es`

EXTENDED ABSTRACT

Tropospheric ozone (O_3) exerts strong adverse impacts on human health, climate, vegetation, biodiversity, agricultural crop yields and thus food security. O_3 is formed in the atmosphere through non-linear photochemical reactions involving volatile organic compounds (VOCs) and nitrogen oxides (NOx) precursors [1]. Furthermore, meteorological stagnation, high solar radiation, high temperatures and low precipitation favor the formation of tropospheric O_3 at surface levels exceeding target regulatory values [2]. Due to the complex and poorly constrained physico-chemical O_3 formation and removal pathways, no straightforward strategies currently exist for reducing O_3. Currently, there are no observational methods that differentiate the origin of O_3. Despite their inherent uncertainties, chemical transport models (CTMs) allow for the apportionment of the contribution of any source to O_3 concentrations. The mass-transfer source apportionment method is an optimal approach to study the contribution of different sources to ozone levels [2].

In this study, we provide a quantitative estimation of the foreign and domestic contributions to surface ozone on Spain, relative to European countries and the contribution of hemispheric background ozone. For that, we use the CMAQ-ISAM within the CALIOPE air quality modelling system to simulate the O_3 dynamics over Europe quantify national contributions for the ozone season from May to October in 2015. We tag both O_3 and its precursors, NOx and VOCs, from the different European countries, all the way through their lifetime, from emission to deposition.

A. Air quality model and simulation set-up

The CALIOPE air quality model [3] was used to simulate the surface ozone concentration over Europe. CALIOPE integrates a meteorological model (WRF-ARWv3.6), an emission model of anthropogenic emissions (HERMESv3) [4] and biogenic emissions (MEGANv2.04) [5], and a chemical transport model (CMAQv5.02) [6]. For the present study, the system is configured with a domain over Europe at a resolution of 18 km x 18 km and 37 vertical levels (top of the atmosphere at 50hPa) and uses meteorological and chemical boundary conditions from the ERA-interim and CAMS analysis, respectively. CMAQ uses the gas phase mechanism CB05 with active chlorine chemistry and toluene mechanism (CB05TUCL) [7] and the sixth generation CMAQ aerosol scheme.

The Integrated Source Apportionment Method (ISAM) within CMAQ allows us to quantify the contribution of O_3 precursors from different countries to the O_3 surface concentration over the European domain [6]. The CMAQ-ISAM tracks both the O_3 and its gas precursor emissions, NOx and VOCs, from the different European countries, boundary conditions and initial conditions. Each tagged species is subject to the physical and chemical processes that it might undergo in the atmosphere. In this study, we tracked the contributions of 35 European countries, the non-European countries, the sea, the chemical boundary conditions (BCON) and the initial conditions.

We focus the study over the period May to October 2015. The summer of 2015 was a hot summer in Europe compared with the reference period 1981-2010; it was especially hot in central Europe and in the Mediterranean. In Spain, there was a heatwave of 26 days during the month of July, making the year 2015 of particular interest to understand O_3 formation.
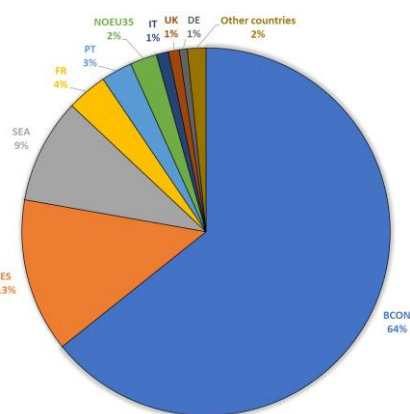

Fig. 1 Foreign and domestic contributions to the MDA8 O3 in 2015 (May-October) in Spain.

B. Results

The maximum daily 8 hours averaged (MDA8) O_3 concentration in Spain for the period of study is 116.10 µg/m^3. In figure 1, we show the relative contribution of the different tagged sources to the MDA8. The imported O_3 is by far the most significant contributor to the ground level O_3 concentration (86 %) and only 14 % is coming from the emitted precursors in Spain. From the foreign contributions, the BCON plays a key role in the O_3 concentrations in Spain. This result shows the importance of the hemispheric and intercontinental transport of O_3 background in the concentrations of Spain. The emissions coming from the sea contribute an overall 10,71 µg/m^3, being the third most important contributor in Spain. Although ship emissions mainly affect coastal regions, their impact on O_3 concentrations is close to the sum of the main country emissions

The contributions of individual European countries to O_3 concentrations in Spain remain below 4%. France and Portugal are the main countries affecting Spain (4 % and 3 %, respectively), and the sum of the other countries contributes a total of 7 %.

Figure 2 shows the MDA8 O_3 surface concentration over Europe for our period of study due to the tagged species (emissions) from Spain. Higher concentrations are found over Spain, but Spain contributes to the O_3 surface concentrations of other European countries, as Portugal and France, clearly affected by proximity. Spain contributes a total of 8.5% and 3.5% of the total ozone concentration in Portugal and France, respectively. Interestingly, the balance between neighbor countries is quite different. Spain is exporting O_3 potential formation to Portugal while importing it from France. In relation to the other European countries, the contribution of Spain is nearly negligible.
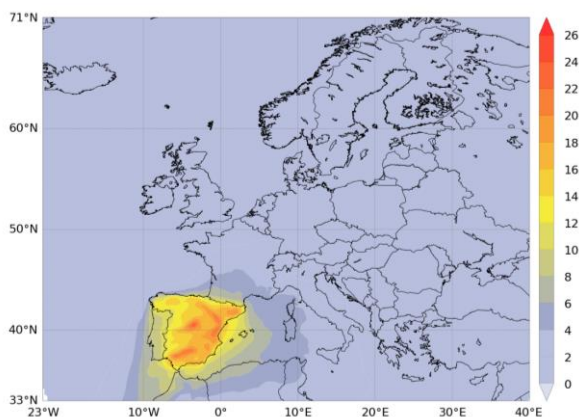

Fig. 2 MDA8 O3 surface concentration over Europe in ug/m3 due to the Spain emissions.

The most polluted regions in Spain are the air basins of Madrid and Barcelona, due to the respective emissions of O_3 precursors in the main urban areas. But the south of Spain also shows important O_3 levels, up to 130 µg/m^3 on MDA8 O_3 average in the studied period. Concentrations there are higher than in Madrid and Barcelona. Note that the south of Spain is not strongly affected by the O_3 contributions from other countries, despite the non-negligible contribution of the North African countries and ships emissions. Our results indicate the importance of understanding the source of the pollutants and how the topography and meteorological conditions play an important role in the O_3 surface concentration. This information is a fundamental step to design effective policies to reduce O_3 in a country and quantify the margin of action that local actions might have.

C. Conclusions and future work

Our study has provided a first estimation of the principal contributors to the O_3 surface concentration in Spain during the ozone season in 2015. The results show the importance of the imported O_3 over Spain and point out the relevance of the long-range transport. To achieve the European air quality standards in Spain and avoid the exceedances of the O_3 target values, it is necessary to decrease not only the domestic emissions of O_3 precursors, but also the emissions from the intercontinental transport, the shipping transport, and the other European countries. So, the regulatory efforts have to be designed as a combination of local and foreign emission reduction actions.

Future work will extend this analysis to other European countries and different years to provide a general picture of the balance between the foreign and domestic contributions to surface ozone among Europe. Furthermore, we plan to combine these results with health information and advance our understanding of the impact of O_3 upon the mortality burden in the continent.

References

[1] Crutzen, P. J.: Photochemical reactions initiated by an influencing ozone in the unpolluted troposphere, Tellus, 26, 47–57, 1975.

[2] Pay, M. T. et al. Ozone source apportionment during peak summer events over southwestern Europe. Atmospheric Chemistry and Physics, 19(8), 5467–5494, 2019.

[3] Baldasano, J. M. et al. An operational air quality forecasting system for the Iberian Peninsula, Balearic Islands and Canary Islands – first annual evaluation and ongoing developments, Adv. Sci. Res., 2, 89–98, 2008.

[4] Guevara, M. et al. HERMESv3, a stand-alone multi-scale atmospheric emission modelling framework – Part 1: global and regional module. Geoscientific Model Development, 12(5), 1885–1907, 2019.

[5] Guevara, M., et al. An improved system for modelling Spanish emissions: HERMESv2.0, Atmos. Environ., 81, 209–221, 2013.

[6] Kwok, R. H. F. et al. Photochemical grid model implementation and application of VOC, NOx , and O3 source apportionment, Geosci. Model Dev., 8, 99–114, 2015.

[7] Sarwar, G. et al. Examining the impact of heterogeneous nitryl chloride production on air quality across the United States, Atmos. Chem. Phys., 12, 6455–6473, 2012.

Author biography

Roger Garatachea Solé received his BSc degree in Chemistry from the University of Barcelona (UB) in 2018. In 2021, he completed his MSc degree in Environmental Engineering at Polytechnic University of Barcelona (UPC), and he did his master's thesis at the Barcelona Supercomputing Center (BSC). In September 2021, he enrolled the Atmospheric Composition Group of BSC as a Junior Researcher Engineer.

CAMP First GPU Solver: A Solution to Accelerate Chemistry in Atmospheric Models

Christian Guzman-Ruiz, Mario Acosta, Oriol Jorba

Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {christian.guzman, mario.acosta, oriol.jorba}@bsc.es

Keywords—Parallelism and concurrency, ODE solver, Chemistry software, Performance

I. Extended Abstract

Atmospheric models are a representation of dynamical, physical, chemical, dynamical, and radiative processes in the atmosphere [1]. The load of these models is often spread across multiple processes in HPC environments.

Most of this load comes from the resolution of chemical processes, which can take up to 90% of the total time execution [2]. Recent studies reported a relevant speedup by translating a chemical module to GPUs [3] [4]. This study is based in some previous works of the authors. These works are tested in the Chemistry Accross Multiple Phases (CAMP) module [5] simulating the conditions of an atmospheric model experiment. In our first approach we present an strategy to efficiently integrate GPU routines without needing to translate the entire chemical module to GPU code [6]. In our second and last work, we integrated a GPU version of the linear solver used in CAMP and evaluated multiple kernel configurations, achieving up to 34x speedup from the base CPU linear solver in a single-thread execution, in addition to a 2.7x for an equivalent MPI implementation with the maximum number of physical cores available on a node (40) [7].

The main objective of this work is to develop a GPU version of the entire CAMP solving algorithm. Our second objective is to evaluate the performance of our work, comparing the results with other state of the art GPU chemical modules.

A. CAMP solving algorithm

We will refer to the system of equations formed by the chemical processes as cells, in reference to the discretization in space grid-cells of the atmospheric model [1].

CAMP uses the CVODE package to solve these cells. CVODE is defined as a solver for stiff and nonstiff ordinary differential equation (ODE) systems (initial value problem) [8]. It includes different algorithms options depending on the user choice. For our case, CAMP uses the Backward Differentiation Formula (BDF) and the KLU linear solver as the default solving options. These options are configured for a single-thread CPU execution. In addition, CAMP includes the Biconjugate gradient (BCG) linear solver for GPU execution as an alternative to the CPU option, thanks to an earlier work of the authors [7].

In this work we use the parallelization strategy with the highest speedup obtained in the linear solver study [7], named as Block-cells (1). With it, we configure the GPU kernel as follows: The number of blocks is equal to the number of cells, and the threads per block are as many as the concentrations of different chemical species in each cell. In this way, each cell is independent of each other, and we avoid any communication between GPU blocks, which is usually very computationally expensive [9]. Similarly, we take advantage of some basic functions developed for BCG, such as array multiplication or array reduction to a single value.

B. Experimental environment

All the tests and executions were performed in the CTE-POWER cluster provided by the Barcelona Supercomputing Center (BSC) [10].

The chemical configuration corresponds to the Carbon Bond 2005 (CB05) mechanism [11] [12] and fixed photolysis reaction rates during time. The initial conditions such as species concentration, temperature and pressure are extracted from an experiment on the Multiscale Online Nonhydrostatic AtmospheRe CHemistry (MONARCH) model [13] [14]. The values are extracted at the first iteration just before calling the chemical module, therefore it corresponds with the initial conditions on the atmosphere after the transformations previous to the chemistry. The source code and configuration files are available in the CAMP repository [15] in the *test_monarch folder*.

C. Results

We obtain up to 36x speedup from the single-thread base version, plus up to 3x speedup against an equivalent MPI implementation with the maximum number of physical cores available on a node (40). We achieve this maximum speedup at 10,000 cells, for greater number of cells the speedup is mostly equal. These speedups are very similar than the BCG speedup (up to 34x and 2.5x respectively) [7]. This indicates than the advantages of translating the BCG algorithm to GPU also applies to the BDF loop. The main advantage is the parallelism obtained, where each thread computes an operation of the species array. Most of the operations in the BDF loop corresponds to this type, either operating over two arrays or reducing an array to a single variable. Another advantage is the shared memory, which allows to fast communication between the threads of the same block.

D. Conclusion

In this study, we present a GPU version of the BDF loop, which achieves up to 36x times faster than the base implementation. This speedup is very similar than the BCG speedup presented in a previous work [7]. Both cases has the similarity of being iterative algorithms and the same chemistry application. Therefore, we can expect a similar speedup for other chemical solving procedures following the GPU approach presented. However, the CAMP module has the aim to facilitate the coupling of new chemical mechanism. Therefore, we expect that the chemistry required by the users can be easily translated to CAMP, allowing to use the already developed GPU version. In a future work we plan to combine this GPU approach with the CPU implementation, dividing the work between CPU and GPU and evaluating load balancing strategies between these two architectures.

II. ACKNOWLEDGMENT

REFERENCES

[1] M. Z. Jacobson, "Fundamentals of Atmospheric Modeling, Second Edition," p. 437.

[2] M. Christou, T. Christoudias, J. Morillo, D. Alvarez, and H. Merx, "Earth system modelling on system-level heterogeneous architectures: EMAC (version 2.42) on the Dynamical Exascale Entry Platform (DEEP)," *Geoscientific Model Development*, vol. 9, no. 9, pp. 3483–3491, Sep. 2016, publisher: Copernicus GmbH. [Online]. Available: https://gmd.copernicus.org/articles/9/3483/2016/

[3] M. Alvanos and T. Christoudias, "GPU-accelerated atmospheric chemical kinetics in the ECHAM/MESSy (EMAC) Earth system model (version 2.52)," *Geoscientific Model Development*, vol. 10, no. 10, pp. 3679–3693, Oct. 2017. [Online]. Available: https://www.geosci-model-dev.net/10/3679/2017/

[4] K. E. Niemeyer and C.-J. Sung, "Accelerating moderately stiff chemical kinetics in reactive-flow simulations using GPUs," *Journal of Computational Physics*, vol. 256, pp. 854–871, Jan. 2014, arXiv: 1309.2710. [Online]. Available: http://arxiv.org/abs/1309.2710

[5] M. L. Dawson, C. Guzman, J. H. Curtis, M. Acosta, S. Zhu, D. Dabdub, A. Conley, M. West, N. Riemer, and O. Jorba, "Chemistry Across Multiple Phases (CAMP) version 1.0: An integrated multi-phase chemistry model," *arXiv:2111.07436 [physics]*, Nov. 2021, arXiv: 2111.07436. [Online]. Available: http://arxiv.org/abs/2111.07436

[6] C. Guzman-Ruiz, M. C. Acosta, M. Dawson, O. Jorba, C. Pérez, and K. Serradell, "Accelerating Chemistry Modules in Atmospheric Models using GPUs," 2021, the 2020 International Conference on High Performance Computing & Simulation (HPCS 2020).

[7] ——, "Studying a new GPU treatment for chemical modules inside CAMP," Sep. 2021. [Online]. Available: https://events.ecmwf.int/event/169/timetable/

[8] "CVODE | Computing." [Online]. Available: https://computing.llnl.gov/projects/sundials/cvode

[9] S. Xiao and W. Feng, "Inter-block GPU communication via fast barrier synchronization," in *2010 IEEE International Symposium on Parallel Distributed Processing (IPDPS)*, Apr. 2010, pp. 1–12, iSSN: 1530-2075.

[10] "Support Knowledge Center @ BSC-CNS." [Online]. Available: https://www.bsc.es/user-support/power.php

[11] G. Yarwood, S. Rao, M. Yocke, and G. Whitten, "Updates to the Carbon Bond Chemical Mechanism: CB05. Final Report to the US EPA, RT-0400675," 2005.

[12] G. Sarwar, H. Simon, P. Bhave, and G. Yarwood, "Examining the impact of heterogeneous nitryl chloride production on air quality across the united states," *Atmospheric Chemistry and Physics*, vol. 12, no. 14, pp. 6455–6473, 2012. [Online]. Available: https://acp.copernicus.org/articles/12/6455/2012/

[13] C. Pérez, K. Haustein, Z. Janjic, O. Jorba, N. Huneeus, J. M. Baldasano, T. Black, S. Basart, S. Nickovic, R. L. Miller, J. P. Perlwitz, M. Schulz, and M. Thomson, "Atmospheric dust modeling from meso to global scales with the online NMMB/BSC-Dust model – Part 1: Model description, annual simulations and evaluation," *Atmospheric Chemistry and Physics*, vol. 11, no. 24, pp. 13 001–13 027, Dec. 2011. [Online]. Available: https://www.atmos-chem-phys.net/11/13001/2011/

[14] O. Jorba, D. D, B.-B. C, C. Pérez García-Pando, Z. Janjic, J. Baldasano, M. Spada, A. Badia, and M. Gonçalves Ageitos, "Potential significance of photoexcited NO2 on global air quality with the NMMB/BSC chemical transport model," *Journal of Geophysical Research D: Atmospheres*, vol. 117, p. D13301, Jul. 2012.

[15] "CAMP," Nov. 2021, original-date: 2021-07-28T14:31:48Z. [Online]. Available: https://github.com/open-atmos/camp
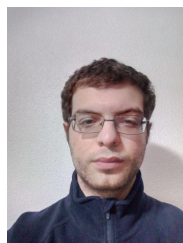
Christian Guzman received his Bachelor's degree in Computer Engineering plus a bachelor's degree in Telecommunication Electronic Engineering, and a Master in Modelling for Science and Engineering by the Autonomous university of Barcelona (UAB). He specialized in techniques for High-Performance Computing (HPC) and is actually developing his pre-doctoral studies on the Barcelona Supercomputing Center (BSC), working on the development of Chemistry Across Multiple Phases (CAMP) module alongside the Multiscale Online Nonhydrostatic Atmosphere Chemistry Model (MONARCH), contributing to the most-computational and logic part from a performance point of view and integrating multitudinal ways of GPU computation in search of speeding-up the system.

Implementation of cellular transport models within a multiscale simulation software

Othmane Hayoun-Mya*†, Miguel Ponce-de-León*, Arnau Montagud*, Alfonso Valencia*‡

*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Universitat Pompeu Fabra, Barcelona, Spain
‡ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Keywords—Agent-Based modeling, Multi-scale modeling, ODE models, Cell transport mechanisms, Systems biology, Biophysics

I. Extended Abstract

Agent-based models (ABM) have been increasingly employed to study dynamics of biological systems. However, these mostly lack transport mechanisms that interface between the agents and their microenvironment. We considered a set of 5 general transport mechanisms (see Fig. 1), written as ODE models and built on top of Fick's Second Law of Diffusion and R-MM kinetics, and their implementation within an ABM software. Unit testing of said models was performed on static, liposome-like agents. We also studied the effect of varying agent and microenvironment-related parameters on the dynamic. We then connected a few of these mechanisms to the agent phenotype, developing a toy example that emulates the experimental decreased tumorigenic growth dynamics of Cytochalasin β.

II. Introduction

Cell transport mechanisms are remarkably relevant for many biological system dynamics. At the same time, agent-based modeling (ABM) software employed for studying tissue and cell-level dynamics usually lacks said mechanisms. In light of this, we set to develop and implement a compendium of general transport systems within an open-source, physics and agent-based multiscale simulation software: PhysiCell [1].

III. Materials & Methods

All simulations have been carried with PhysiCell v1.9.1 (https://github.com/MathCancer/PhysiCell.git) [1]. Note that, as a multiscale simulation software, it works with three different timescales: The phenotype (6 *min*), the mechanics (0.1 *min*) and the diffusion timescale (0.01 *min*). For the development and implementation of the transport models, we will mostly work at the diffusion timescale. To obtain quantitative results of these dynamics, a small extension of *tools4physicell* (https://github.com/migp11/tools4physicell) was developed.

IV. Preliminary results & discussion

Model building. In order to generalize the different transport systems, we two common mechanisms: Diffusion and two-step enzymatic reactions. For this reason, all models will be built through tweaks and combinations of the simple diffusion and facilitated diffusion carriers.
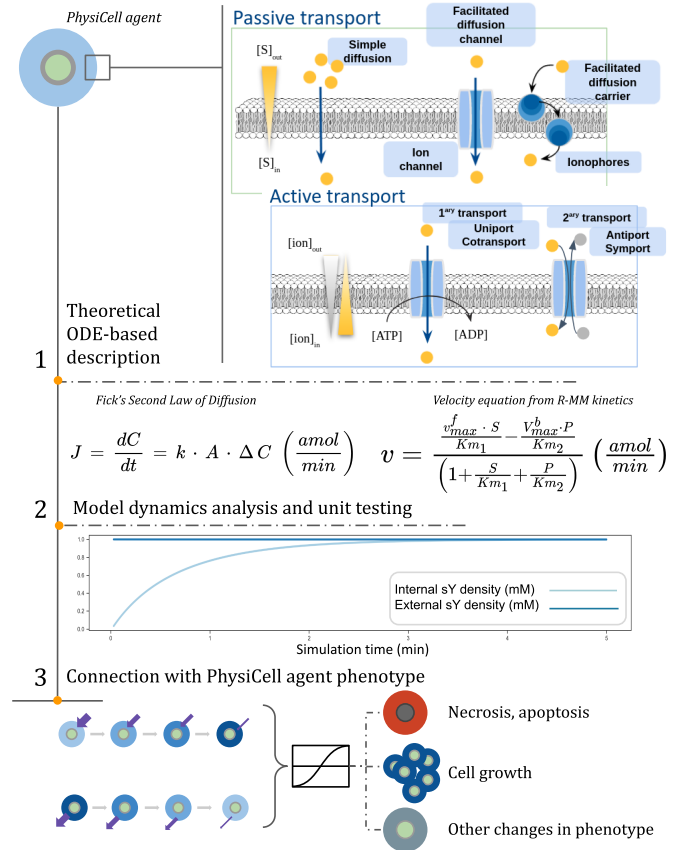

Fig. 1. Graphical abstract describing the employed workflow for the development and implementation of different transport mechanisms within PhysiCell. In point 1, R-MM refers to Reversible Michaelis-Menten kinetics.

Simple diffusion (SD) across the cell membrane is the simplest mechanism. Implementation is based on a small variation of Fick's Second Law of Diffusion, in order to align it with the space (μm) and time (*min*) of PhysiCell (see Point 2 of Fig. 1) by accounting for the permeability coefficient (k ($\frac{\mu}{min}$)), the agent surface (A (μm^2)) and the concentration gradient (ΔC (mM)). Implementation within PhysiCell consisted in integrating the flux at each diffusion timestep (0.01 *min*) with additional biophysical constraints.

The second common mechanism is a facilitated diffusion carrier (FDC). Following a Reversible Michaelis-Menten (R-MM) kinetic, its rate laws can be employed to represent it as a system of four coupled ODEs. From these, and assuming

standard Quasi-Steady State and total substrate and receptor mass balance, we can obtain a relatively simpler velocity equation (See Point 2 of Fig. 1). Implementation consisted in assuming the substrate flux to be equal to the this equation, which was obtained through integration of the ODE system at each diffusion timestep.

Facilitated diffusion channel (FDCh) mechanisms can be understood as the flow of a substrate through a "pore" in the cell membrane that opens or closes under certain conditions. For this reason, we implemented it as a step-function SD model. The channel has two states, Open (O) and Closed (C), and the cause of transitioning between these states is what differentiates the channel types. For Aquaporin-like channels, the condition that determines the state of the channel is just stochastic, obtaining a value from a N(1,1) distribution to compute the amount of Open channels. For Ligand-gated channels, they are considered to be open if there is some Ligand (L) attached to the receptor (R). This amount of RL complex is computed following a Receptor-Ligand kinetic model, implemented as a three coupled-ODE system.

Primary active transport model (PA) was implemented as system of two coupled mechanisms: An independent ATP-ADP exchange system and a carrier enzyme with an FCD-like mechanism, transporting a substrate against gradient each time an ATP is hydrolyzed. Simialrly, the Secondary active (SA) transport model, was built like two coupled FCD-like mechanisms, where one of them moves a substrate along gradient, and the other one moves another substrate against gradient. Here, the coupling is through the equivalence of the pumping rate of a substrate to the passive entry rate of the along-gradient substrate.

***In silico* experiments** Assessment of transport dynamics was carried through two experiments: Experiment A, in which the simulation consisted of agents devoid of substrate, and a microenvironment with certain substrate density (see Point 3 of Fig. 1); And experiment B, where we have the opposite scenario. These experiments were carried with sets of static, liposome-like agents in a closed system, without changes in phenotype, cell cycle nor cell death. This was done in order to avoid seeing dynamics not directly obtained from the transport model.

For the SD model, in both Experiments A and B, both densities inside and outside the agent reached equilibrium following a logarithmic curve while maintaining net balance of the total amount of substrate. This was observed for a wide range of k values (10^{-4} to 1000 $\frac{\mu m}{min}$), indicating that this model allows for simulation of a wide variety of different types of substrate types (see Table 1 from [3]). Identically, for the FDC model, for both Experiments A and B, the same equilibrium-reaching dynamic was seen, and mass balance of both substrate and receptor were maintained.

We explored variations of different parameters of the simulation to assess the effect on the dynamic. For both SD and FDC models, and in both experiments, increasing cell number resulted in a faster equilibrium at a lower concentration, and increasing cell volume resulted om a slower equilibrium, lower concentration. This is explained given that the SD model accounts for the surface, which increases quadratically with the volume.
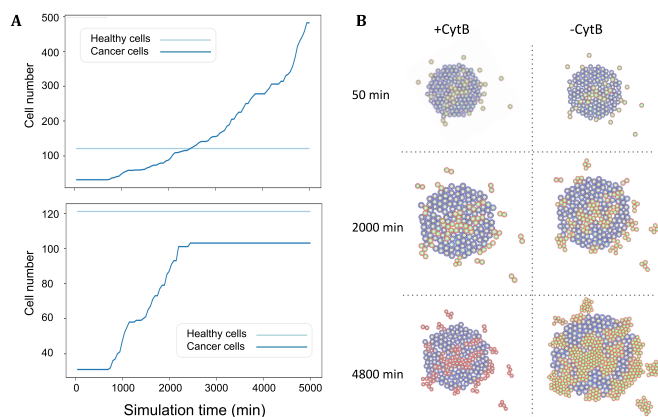

Fig. 2. Blocking of Glucose transport by addition of $Cyt\beta$ at 2500 min. Glucose enters the cells through a FCD mechanism, as a GLUT transporter. $Cyt\beta$ enters agents through an SD model. Both cell types show heterogeneity in the initial amount of receptors per agent, although the cancerous cells contain twice as many receptors [2]. (A) Healthy cell count is constant, as these were set to not divide. (B) PhysiCell output screenshots for both experiments. Red-outlined cells are cancerous cells. Blue-colored cells are healthy cells. Green-coloured cytoplasm indicates levels of Glucose inside the agent.

Next, we set to emulate the dynamics of adding Cytochalasin β ($Cyt\beta$) (Fig. 2). a mycotoxic compound commonly employed to treat cancer by inhibition of GLUT transporters [2], which have an FDC-like mechanism. As expected, cancerous cells divide faster due to the increased amount of GLUT transporters, but addition of $Cyt\beta$ shows a decrease in growth due to the "blocking" of the channels (Fig. 2.A, lower panel), as observed experimentally [4].

Note that the efforts of this work are directed towards fitting said models with specific cell-line pharmacokinetic data in order to represent more realistic behaviours, which is notably valuable in the context of personalized medicine treatment research.

REFERENCES

[1] A. Ghaffarizadeh *et al.*, "Physicell: An open source physics-based cell simulator for 3-d multicellular systems," *PLoS Computational Biology*, vol. 14, 2 2018.

[2] K. Kapoor *et al.*, "Mechanism of inhibition of human glucose transporter glut1 is conserved between cytochalasin b and phenylalanine amides," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 4711–4716, 4 2016.

[3] N. J. Yang and M. J. Hinner, "Getting across the cell membrane: an overview for small molecules, peptides, and proteins," *Methods in molecular biology (Clifton, N.J.)*, vol. 1266, pp. 29–53, 2015.

[4] M. Trendowski, "Anti-cancer agents in medicinal chemistry using cytochalasins to improve current chemotherapeutic approaches," p. 327, 2015.

Othmane Hayoun received his BSc degree in Biotechnology from Universitat de València, Spain in 2020. Now he is carrying his MSc degree in Bioinformatics for Health Sciences (Universitat Pompeu Fabra, Spain), and is doing his MSc Thesis at the Computational Biology group of the Barcelona Supercomputing Center (BSC).

Adaptive points to estimate the Lebesgue constant on the simplex

Albert Jiménez-Ramos, Abel Gargallo-Peiró, Xevi Roca
Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {albert.jimenez, abel.gargallo, xevi.roca}@bsc.es

Keywords—Lebesgue constant, Lipschitzian optimization, adaptivity

I. Extended Abstract

We present a novel adaptive sampling method to estimate the Lebesgue constant of nodal sets in n-dimensional simplices. The main application of this estimation is to assess the interpolation capabilities of a nodal distribution. Given such distribution, the Lebesgue constant corresponds to the maximum of the Lebesgue function, which is non-differentiable. Thus, our method estimates the extremum by only evaluating the function values at a set of sample points that are successively adapted to seek the maximum. Remarkably, our adaptive search does not require storing a mesh to query neighbor points. Furthermore, the search automatically stops by considering specific spatial and Lipschitz-based criteria. The examples, up to four dimensions, show that the method is well-suited to estimate the Lebesgue constant of different nodal distributions.

A. Introduction

The Lebesgue constant is used to assess the interpolation capabilities of a nodal set since it appears in the upper bound of the interpolation error. Given an interpolation set $\mathbf{z}$ in the simplex K^n in $\mathbb{R}^n$, the Lebesgue constant $\Lambda(\mathbf{z})$ is defined as

$$\Lambda(\mathbf{z}) = \max_{\mathbf{x} \in K^n} \Lambda(\mathbf{x}, \mathbf{z}) = \max_{\mathbf{x} \in K^n} \sum_i |l_i(\mathbf{x}; \mathbf{z})|,$$

where $\{l_i(\mathbf{x}; \mathbf{z})\}_i$ are interpolative shape functions. Note that the Lebesgue function $\Lambda(\mathbf{x}, \mathbf{z})$ is non-differentiable, so we can only use function values to estimate the maximum.

One way to estimate the Lebesgue constant is to sample the function in a structured fine grid of points. The maximum value at these sample points determines the estimation of the Lebesgue constant. On the other hand, there exist adaptive sampling methods which refine regions around maxima candidates [1], [2]. In the work presented in [1], a structured triangle subdivision is used to adapt the sample points to the location of the maximum. In [2], the sampling is determined by random point within a box centered at the current maximum which reduces its size in successive iterations. In both works, the method stops once a specific number of iterations is reached. We present a deterministic method based on structured sample points determined by a system of rational barycentric coordinates, and with a stopping criterion that accounts for the function.
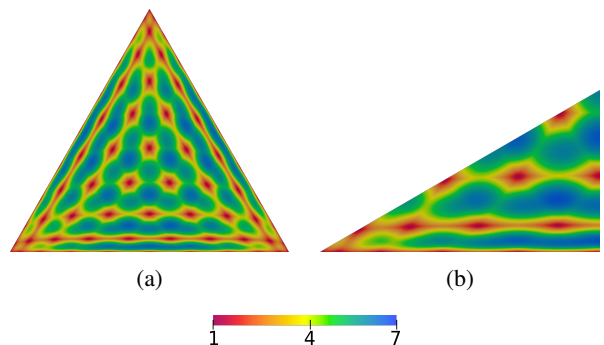

Fig. 1: (a) Lebesgue function for a symmetric nodal distribution of points of polynomial degree 10 in the triangle, (b) and in the symmetric tile, i.e. the sextant.
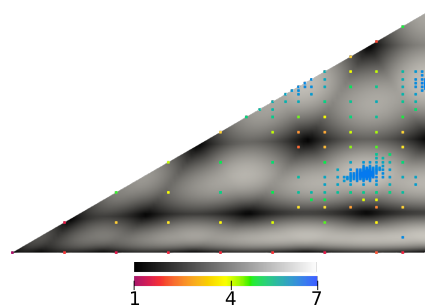

Fig. 2: Sampling for the estimation of the Lebesgue function plotted in Fig. 1.

B. Method

In Fig. 1(a), we plot the Lebesgue function for a symmetric nodal distribution of points of polynomial degree 10 in the triangle [2]. Since the nodal set is symmetric, it is enough to consider the symmetric tile which in 2D corresponds to the sextant of the triangle, Fig. 1(b). The method starts sampling the function at the vertices of the domain. At each iteration, the sampling around the maximum candidate for each resolution is refined, and the maximum is adaptively captured. In Fig. 2, we show the sample points at the last iteration. Note that regions where the Lebesgue function attains higher values correspond to regions where the sampling is finer.

The stopping criterion is based on the planarity of the function. More precisely, we estimate the Lipschitz constant around the maximum sample point, and the method stops

p \ n	1	2	3	4	5	6
6	4.55	8.75	13.66	19.22	25.49	32.63
7	6.93	14.35	23.38	34.08	46.54	61.00
8	10.95	24.01	40.55	60.86	85.24	114.13
9	17.85	40.92	71.15	109.43	156.62	213.76
10	29.90	70.89	126.20	198.08	288.82	400.93

TABLE I: Estimation of the Lebesgue constant for an equi-spaced distribution of points in the simplex for polynomial degree p and dimension n, $p = 6, \ldots, 10$, $n = 1, \ldots, 6$.
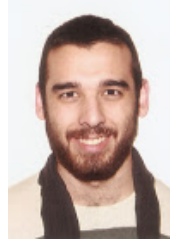
Albert Jiménez-Ramos received his BSc degree in Mathematics from Universitat Politècnica de Catalunya (UPC), Barcelona in 2017. He completed his MSc degree in Advanced Mathematics and Mathematical Engineering from UPC in 2018. Since 2018, he has been a member of the Geometry and Meshing for simulations group of Barcelona Supercomputing Center (BSC) as a PhD student of the Applied Mathematics doctorate program of UPC, Spain.

when this estimate is below a desired threshold. Specifically, if we denote by $\mathbf{x}^\star$ the current maximum point, and by $\mathbf{y}_i$ the neighbor points, $i = 1, \ldots, N$, two conditions should be satisfied. First, we check if the neighborhood of the maximum is sufficiently sampled, that is

$$d\left(\mathbf{x}^\star, \mathbf{y}_i\right) < \delta,$$

for all $i = 1, \ldots, N$. Second, we check if the function is flat at the maximum:

$$\tilde{K}_{\mathbf{x}^\star}\left(\mathbf{y}_i\right) = \frac{\left|\Lambda\left(\mathbf{x}^\star; \mathbf{z}\right) - \Lambda\left(\mathbf{y}_i; \mathbf{z}\right)\right|}{\left\|\mathbf{x}^\star - \mathbf{y}_i\right\|_2} < \varepsilon,$$

for all $i = 1, \ldots, N$.

C. Results

In Table I, we show our estimation of the Lebesgue constant for an equispaced distribution of points on the simplex for several dimensions and polynomial degrees. Up to dimension 3, our method reproduces the values reported in the literature. We also compute the Lebesgue constant up to polynomial degree 10 and dimension 6.

D. Conclusion

In this work, we have presented an adaptive deterministic method to estimate the Lebesgue constant. It is based on structured sample points determined by a system of coordinates that circumvents the need for the explicit storage of the neighbor structure and makes the method useful for the n-dimensional simplex. Moreover, the stopping criterion accounts for the planarity of the function which avoids a sub-estimation of the Lebesgue constant. Finally, we find this method convenient when optimizing nodal distributions in n-dimensional simplices targeting good interpolation properties.

II. ACKNOWLEDGMENT

REFERENCES

[1] M. J. Roth, "Nodal configurations and voronoi tessellations for triangular spectral elements," Ph.D. dissertation, 2005.

[2] T. Warburton, "An explicit construction of interpolation nodes on the simplex," *Journal of engineering mathematics*, vol. 56, no. 3, pp. 247–262, 2006.

Respiratory tract microbiome modifications after lung transplantation and its impact in CLAD

Olfat Khannous-Lleiffe[1,2], Toni Gabaldón[1-3*]

[1] Life Sciences Department , Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

[2] Mechanisms of Disease, Institute for Research in Biomedicine (IRB), Barcelona, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[1]olfat.khannous@bsc.es, [3]toni.gabaldon@bsc.es[*]

EXTENDED ABSTRACT

Survival after lung transplantation is limited in large part due to the high incidence of chronic lung allograft dysfunction (CLAD). Infection is a recognized risk factor for the development of CLAD, and both acute infection and chronic lung allograft colonization with microorganisms increase the risk for CLAD. The aim of our study was to analyze respiratory tract microbiome modifications after lung transplantation, with a focus on its relationship with CLAD. Here we used 16S rRNA metabarcoding, observing specific microbiome profiles for both upper and lower respiratory tract and observing lower beta diversity between healthy subjects and patients with NO-CLAD as well as differentially abundant features associated with CLAD.

A. INTRODUCTION

Lung transplantation is an increasing procedure for end-stage lung disorders patients. Most of the mortality within the first year after the transplant is due to infections and development of chronic lung allograft dysfunction (CLAD) [1]. The microbiome is the set of microbial organisms (microbiota) including its genes and metabolites, which inhabit a certain niche. Several studies have recognized its essential role in the regulation of metabolic and physiological processes [2]. Two of the most widely used strategies to study the microbiome are: 16S ribosomal RNA amplicon sequencing, and whole-genome shotgun (WGS). The 16S subunit ribosomal gene has nine hypervariable regions that can be amplified with the use of universal PCR primers [3]. Sequencing of these amplicons and comparison to rRNA sequence databases allows assessing what taxa are present in a sample, and their relative abundances.

In the present work we described the microbiome of lung transplant candidates with different lung disorders and assessed the changes in their microbiome longitudinally after the transplant, and the impact of these changes on the outcome.

B. MATERIALS AND METHODS

This was a longitudinal study in which we analysed the microbiome in both the upper (nasopharyngeal swab (NP) and lower (Lung tissue, Bronchial swab and Bronchoalveolar lavage fluid (BALF)) respiratory tract from 68 lung transplanted (LT) patients the day of the transplant, when patients were discharged and 2-5 months and 12 months after the transplant. We also analysed NP swab samples from 10 healthy subjects. CLAD was assessed after two years of follow-up. DNA was extracted from the samples and the hyper-variable region (V4) of the 16S ribosomal RNA gene was amplified by standard PCR. Amplicons were purified and sequenced in barcoded pools using Ilumina MiSeq technology. Raw sequence reads were demultiplexed by using idemp. The resulting single-end reads were processed using Dada2 [4] pipeline obtaining an amplicon sequence variant table (ASV) to which taxonomy was assigned. Data was normalized by transforming the raw counts to centered log-ratios (clr). Alpha diversity, including both Shannon and Observed indices and beta diversity metrics such as the Aitchison distance, were calculated.

The effect of the clinical variables on the overall composition was evaluated by using the adonis test from the Vegan R package. Differential abundance at the day of the transplant and one year after the transplant was assessed by applying a linear model considering as fixed effect the CLAD variable and as possible source of batch effect the sequencing run. We also performed a differential analysis longitudinally to assess differential trajectories according to CLAD by using the clr of the top25 genus and the function permuspliner of the Splinectome [5] R package.

B. RESULTS AND DISCUSSION

Alpha diversity metrics were calculated for the different types of samples at the day of the transplant observing a significantly increased Observed index, but a decreased Shannon index (although not significantly) in NP swab samples compared to both lung tissue (Lung_Tissue) and bronchial swab samples (B_Swab) of the recipients. Therefore, our analyses suggested more richness but less evennes in upper respiratory tract (NP) as compared to the lower respiratory tract (Lung_Tissue and B_Swab) maybe due to their dynamism and higher influence of external factors.

The type of the sample also influenced the overall composition as suggested by beta diversity such as the Aitchison distance (Adonis test, p-value = 0.001). We detected clustering of the samples according to their source (Fig. 1), observing less distance between lower respiratory tract samples (lung tissue and bronchial swabs) and a composition derived mainly of Actinobacteria at the upper tract as compared to the lower tract in which we observed high presence of Proteobacteria, Firmicutes and Bacteroidetes phyla.
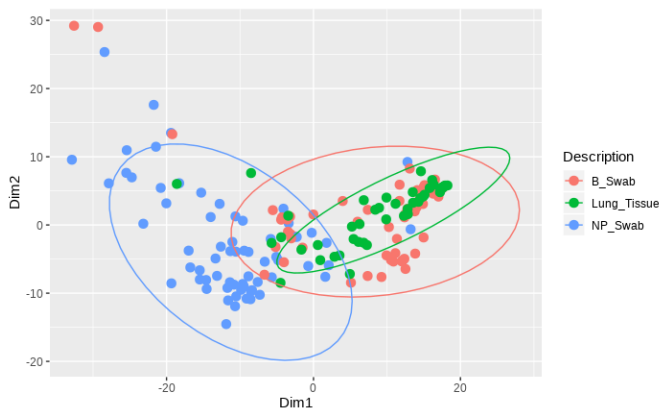
Fig. 1 Overall composition of the samples according to their source. Multidimensional (MDS) plot representing the Aitchison distance. Samples are colored according to the type of the sample: Lung tissue (Lung_Tissue), Bronchial swab (B_Swab) and Nasopharyngeal swab samples (NP_Swab).

One year after transplant a lower beta diversity, as indicated by the Aitchison distance, was observed between healthy subjects and patients with no CLAD. We confirmed a significant effect of the CLAD on the stratification of the samples according to the overall composition (Adonis test, p-value=0.02, R2 (explained variability)=0.049).

Our differential analysis detected 5 and 11 genus as differentially abundant according to CLAD in both the day of the transplant and one year after the transplant, respectively.

We also made a longitudinal follow up of the 25 top genus, detecting significant differences of 2 trajectories between CLAD and Non-CLAD patients: *Veillonella* (p=0.006) and *Neisseria* (p=0.048). Looking more in depth in the particular time points we observed a significant increase in the CLAD group of *Veillonella* 2-5 months after the transplant as compared to Non-CLAD and Healthy subjects (Fig.2).
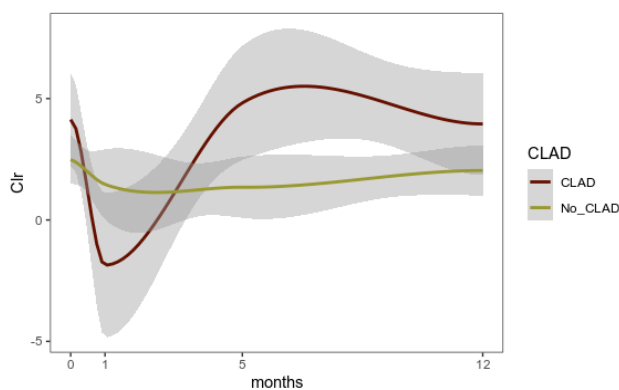

Fig. 2 Veillonella clr over the time distinguishing CLAD and Non-CLAD groups.

We assessed the influence of the donor on the outcome by computing the Aitchison distance between bronchial swab donor samples and distinguishing those distances within donors with patients that ended up with CLAD or not and also between donors of CLAD patients and donors of non-CLAD patients observing more distance within CLAD donors but not clear separation between CLAD and Non-CLAD donors. These suggested no influence of the donor on the outcome of the patient.

REFERENCES

[1] Watzenboeck ML, Gorki AD, Quattrone F, et al. Multi-omics profiling predicts allograft function after lung transplantation. *Eur Respir J*. 2022;59(2):2003292. Published 2022 Feb 3. doi:10.1183/13993003.03292-2020

[2] Brumfield KD, Huq A, Colwell RR, Olds JL, Leddy MB. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One*. 2020;15(2):e0228899. Published 2020 Feb 13. doi:10.1371/journal.pone.0228899

[3] Mas-Lloret J, Obón-Santacana M, Ibáñez-Sanz G, et al. Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample. *Sci Data*. 2020;7(1):92. Published 2020 Mar 16. doi:10.1038/s41597-020-0427-5

[4] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581-583. doi:10.1038/nmeth.3869

[5] Shields-Cutler RR, Al-Ghalith GA, Yassour M, Knights D. SplinectomeR Enables Group Comparisons in Longitudinal Microbiome Studies. *Front Microbiol*. 2018;9:785. Published 2018 Apr 23. doi:10.3389/fmicb.2018.00785

Author biography

Olfat Khannous Lleiffe was born in Vic, Spain, in 1996. She received the BSc in Biotechnology from University of Vic in 2018 and the MSc in Bioinformatics for Health Sciences from University Pompeu Fabra (UPF) in 2020. She is currently doing a PhD in Biomedicine at Toni Gabaldon's Comparative Genomics group with the "Formación de profesorado universitario (FPU)" fellowship from the Spanish ministry of universities. She is also a teaching assistant in both University of Barcelona and ESCI UPF.

Numerical Investigation of Unsteady Processes in Combustion using Flame Transfer Function

Phivos Andreou Koumides[*][†], Daniel Mira[*], Eduardo J. Pérez-Sánchez[*]

[*]Barcelona Supercomputing Center, Barcelona, Spain

[†]Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {phivos.koumides, daniel.mira, eduardo.perez}@bsc.es

Keywords—Flame instabilities, Flame transfer function, Premixed Bunsen flames.

I. Extended Abstract

Hydrogen has been recognized as a clean and sustainable energy carrier to help deliver significant emission reductions and is regarded as an accelerator in achieving a low carbon future. One of the main limitations of the use of hydrogen in combustion systems is the formation of instabilities in lean conditions. Flame instabilities are an undesirable effect, often difficult to predict and eliminate in practical applications. They refer to large amplitude oscillations of one or more characteristic frequencies, arising from the resonant interaction between oscillatory flow and unsteady heat release process [1]. These instabilities can have detrimental effects, such as component vibrations, increased heat transfer rates, flame blow-off or flashback [2].

The control of combustion instabilities is a key aspect to operate in lean conditions and reduce pollutant formation. The key differences of hydrogen as a fuel are its relatively higher adiabatic flame temperature, flame speed and diffusivity, which induce technical challenges to combustor operation. These include increased propensity for flashback and autoignition, as well as higher NOx emissions and different thermoacoustic instability characteristics arising from its higher reactivity and burning rates.

In order to understand the influence of such instabilities on combustion it is necessary to characterize them in simplified configurations for which detailed measurements are available. To pursue this objective, an analysis of the flame in frequency is performed, by obtaining the Flame Transfer Function (FTF) when the flame is excited to velocity perturbations. A multi-slit Bunsen burner experiment investigated by V.N. Kornilov et al. [1] is chosen. The infrastructure developed to compute the FTF when applied to different combustion modelling approaches is validated by comparing numerical results against experiments. The results presented are focused on model assessment and validation, so only the methane case is presented.

A. Methodology

This study is focused on the response of the flame to perturbations at the inlet conditions. Two different numerical approaches are considered to solve the CFD and are compared with experimental data.

1) Experimental configuration: The Kornilov et al. [1] experiment consists of a vessel with a flat perforated disk, of 1mm thickness. The disk contains 8 rectangular 12x2mm slits separated by 3mm. Methane/air mixture is injected below
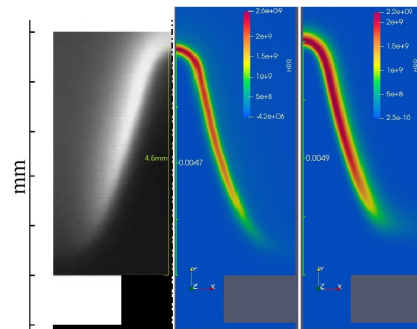

Fig. 1. Flame fields and heat release rate. Left: experimental [1]. Middle: numerical with finite rate model. Right: numerical with FGM method.

the plate at a fixed equivalence ratio $\phi = 0.8$ and mean inlet velocity $\tilde{v} = 0.4 m/s$. The burner plate temperature is measured by an embedded thermocouple. A flow velocity perturbation v' is imposed via a loudspeaker installed upstream the mixture supply tube.

This experimental setup was selected, being a simple 2-dimensional, laminar case, offering well documented results for both steady and perturbed flames.

2) CFD numerical approach: The CFD problem was solved using Alya, a BSC in-house high performance computational mechanics code using the finite element method. It is designed to solve complex coupled multi-physics, multi-scale, and multi-domain problems. One of the advantages of Alya, is its parallelism capability, using Message Passing Interface (MPI) standard, or Open Multi-Processing (OpenMP) interface. Using parallelism, a large number or nodes can operate concurrently, reducing substantially the time to solve.

Two different approaches are considered to evaluate the flame response to flow excitations:

a. Finite rate chemistry (FR)
b. Flamelet Generated Manifolds (FGM)

On one hand, the FR model solves the full set of species mass fractions included in the reaction mechanisms using Cantera in Alya, and integrates the chemistry using an implicit chemical integration method based on backward Euler method.

On the other hand, the flame has been solved with tabulated chemistry based on the FGM method. The FGM model is based on the flamelet concept, a powerful model used for turbulent combustion modelling that describes the turbulent flame as an ensemble of one-dimensional laminar flames called flamelets. In the FGM, all the possible states of the flamelets

are tabulated in a manifold as a function of some controlling variables later transported in the flow. The method features drastic reductions in the computational cost, especially for complex chemical mechanisms [3]. Even though this model is commonly used for turbulent combustion analysis, it is not restricted to this regime and due to its aforementioned notorious reductions in computational cost, it has been used for the simulation of the laminar flames presented in this work. Both combustion models are coupled to a low-Mach solver.

Two different chemical reaction mechanisms were used for the simulations. A detailed mechanism containing 53 species and 325 reactions [4] and a reduced mechanism containing 6 species and 2 reactions [5].

B. Flame Describing Function

To quantify the correlation between acoustics and combustion, a flame transfer function (FTF) was used. In its simplest form, FTF measures the response of the global unsteady reaction rate in the flame, to an inlet velocity perturbation [6].

$$FTF(f) := \frac{q'/\tilde{q}}{v'/\tilde{v}} \qquad (1)$$

where q is the global heat release rate, v the velocity and f the frequency of perturbation. A prime (') denotes the perturbation and a bar () the mean value. FTF(f) is a complex number, where the modulus represents the gain and the phase represents the phase of the FTF.

The FTFs of the CFD simulations were computed using an in-house code developed by the authors. These codes are based on the Fast Fourier Transform (FFT) algorithm written in Python.

C. Results and discussion

Figure 1 shows the steady flame luminosity of the experiment, which is directly related to the heat release [1] (left), the numerical results using the finite rate model (middle) and FGM (right). As it is observed from the figure, the flame height and shape for the simulations is in close agreement with the experiments, especially for the FR model, showing that the flame speed is correctly captured.

Figure 2 gives a comparison of the gain and the phase between the experimental and the simulated results. On one hand, for low frequencies the flame amplifies the response providing gains greater than 1. However, it exponentially drops when increasing the frequency showing that the flame is a low-frequency filter. On the other hand, the phase, which measures the delay between velocity and heat release signals, is almost linear showing that the delay in time between both is almost constant.

This behaviour is well-captured by both models for which an excellent agreement is observed for the gain. In the case of the phase, there is almost a perfect agreement with the FR case, but some differences exist for the FGM model, that even being accentuated when increasing the complexity of the mechanism, are deemed to be small. The origin of such deviations requires further investigation.

The results validated the applied methodology and show that it is able to capture the unsteady phenomena in an accurate way.
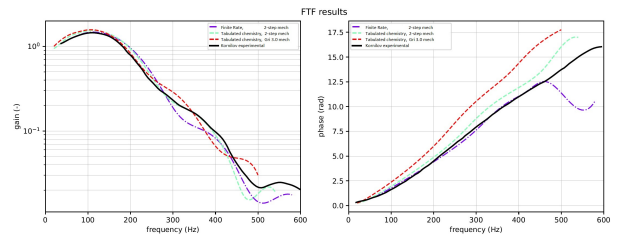

Fig. 2. FTF comparison between experimental and simulated perturbed flames. The experimental results appear in solid lines.

D. Conclusions and perspectives

In this paper we presented the methodology and the tools to numerically study the effect of perturbations on laminar premixed flames. We demonstrated that the FGM method can reproduce the experimental results for the steady as well as the FTF of the perturbed flames. The validation of the methodology resulting from obtaining the FTF by using the FGM as the combustion model is extremely important as it can be used to characterize turbulent flames submitted to noticeable perturbations and therefore, understand their origin as a way to mitigate or suppress them.

The future work is focused on using hydrogen as the fuel at different equivalence ratios, as well as extending this research to turbulent flames.

REFERENCES

[1] V. N. Kornilov, R. Rook, J. H. ten Thije Boonkkamp, and L. P. de Goey, "Experimental and numerical investigation of the acoustic response of multi-slit bunsen burners," *Combustion and Flame*, vol. 156, pp. 1957–1970, 10 2009.

[2] D. Mira, O. Lehmkuhl, P. Stathopoulos, T. Tanneberger, T. G. Reichel, C. O. Paschereit, M. Vázquez, and G. Houzeaux, "Numerical investigation of a lean premixed swirl-stabilized hydrogen combustor and operational conditions close to flashback," *Proceedings of the ASME Turbo Expo*, vol. 4B-2018, 2018.

[3] J. A. van Oijen and E. U. Press), *Flamelet-generated manifolds: development and application to premixed laminar flames*. Technische Universiteit Eindhoven, 2002.

[4] U. of California Berkeley, "Gri-mech 3.0 chemical reaction mechanism," 2022. [Online]. Available: http://combustion.berkeley.edu/gri-mech/index.html

[5] C. Université de Toulouse, "Methane/air combustion reduced chemical reaction mechanism," 2022. [Online]. Available: https://www.cerfacs.fr/cantera/mechanisms/meth.php

[6] L. Crocco, "Research on combustion instability in liquid propellant rockets," *Symposium (International) on Combustion*, vol. 12, pp. 85–99, 1 1969.

Phivos Andreou Koumides received his BScE degree in Mechanical Engineering from the University of New Brunswick (UNB), Canada in 1992. He worked as a process engineer for IBM Canada, as a service engineer for Nissan Motor Co and Volvo Car Corporation, as an operations engineer and as a maintenance manager for the Electricity Authority of Cyprus. He completed his MSc degree in Sustainable Energy Systems at the University of Cyprus in 2020. Since 2021, he has been with the Computer Applications for Science and Engineering department (CASE) of Barcelona Supercomputing Center (BSC), as well as a PhD student in the Department of Physics-Aerospace Division, of Universitat Politécnica de Catalunya (UPC), Spain.

Ground-Level Ozone Simulations Improved by Updating Land Cover Databases

Franco López[#1], Enza Di Tomaso[#], Oriol Jorba[#]

[#] *Barcelona Supercomputing Center (BSC), Barcelona, Spain*
[1] `franco.lopez@bsc.es`

Keywords— **Ozone, air quality modelling, land cover map, satellite observations**

EXTENDED ABSTRACT

A. Introduction

Tropospheric ozone (O_3) is one of the pollutants which raises great concern to the World Health Organization (WHO) because of the health effects associated with its exposure. It is formed in the atmosphere through non-linear photochemical reactions among carbon monoxide (CO), peroxy radicals generated by the photochemical oxidation of volatile organic compounds (VOCs) and nitrogen oxides (NO_X) [1]. High levels of tropospheric ozone cause respiratory malfunction, and a long-term exposure can lead to death [2].

Air Quality Models (AQM) can diagnose and predict air pollutant levels, as well as being a key tool to assess air quality management policies. Nevertheless, modelling ozone surface concentration has several sources of uncertainty, with respect to biogenic and anthropogenic precursor emissions, chemical boundary conditions, chemistry formation, and deposition [3]. Deposition velocity depends on biotic and abiotic parameters that, ultimately, are linked to the land use class (LUC). The use of updated information on land use is critical to improve AQM results.

This study assesses the impact of using an updated land cover map on the skills of an AQM simulating surface ozone concentration over Spain. We used the Multiscale Online Nonhydrostatic AtmospheRe CHemistry (MONARCH) model and enhanced the system with new land use information.

B. Methodology

Two land cover maps were used for ozone modelling: the 1987 United States Geological Survey (USGS) and the 2015 European Space Agency (ESA) ones [4]. The latter has been produced under the ESA Climate Change Initiative (CCI) which aims to realize the full potential of long-term satellite Earth observations. In particular, we have used the medium-resolution land cover CCI product that has been produced at a 300 m resolution using state-of-the-art satellite retrievals. In order to fulfill the modelling requirements, the ESA land cover was re-mapped to a 30s resolution map (using the dominant-class criteria), whereas the LUCs were re-mapped to the USGS 24-category codes following an equivalence matrix. This matrix was made by assigning to each of the ESA 37-category LUCs the USGS code that defines the same, or most similar, physical coverage of the surface.

Ozone atmospheric simulations were carried out with the MONARCH model [5] at a regional scale over Europe at the 0.2° latitude x 0.2° longitude horizontal resolution for the month of July 2019. Simulation outputs have been produced at an hourly temporal resolution.

The model performance has been evaluated by comparing simulated data against the European Environment Agency Air Quality (EEA AQ) e-Reporting observational data. All measurement stations with more than 75% data availability on a monthly basis were selected for comparison.

C. Results and discussion

Over Europe, most of the inland surfaces see a modification in the LUCs classification after updating the land cover map (Fig. 1A). Particularly, over the Iberian Peninsula there is a change in the presence of some classes (Cropland/Woodland Mosaic appears in 1987 USGS map, but it is not present in the 2015 ESA one), as well as a redistribution of classes such as Dryland Cropland and Pasture, Mixed Shrubland/Grassland, and Deciduous Broadleaf Forest. This result is due to both having used land cover information from a more recent year (2015 versus 1987) and the conversion of the ESA 37-category LUC codes into the USGS 24-categories.

The impact of employing an updated land cover map on the ozone surface concentration simulated in July 2019 was not homogeneous over the Iberian Peninsula. Over the northern area there was an increase of the ozone concentration, whereas in the center and southern areas lower surface concentration was simulated, compared to the simulation using the 1987 map (Fig. 1B).

A better fit of the MONARCH simulations to independent EEA AQ e-Reporting observations was obtained when using the 2015 ESA map compared to the 1987 USGS one, as found by calculating monthly statistics of mean bias (0.40 vs. 0.95), root mean square error (4.8 vs. 5.0) and correlation coefficient (0.898 vs. 0.895). Overall, the use of the 2015 ESA land cover map improved the simulation of ground-level ozone over the Iberian Peninsula.

D. Conclusion

The uncertainties of gas-phase pollutant modelling can be reduced by updating the ancillary information used by the model simulation, as suggested by the results shown for ozone surface concentration.
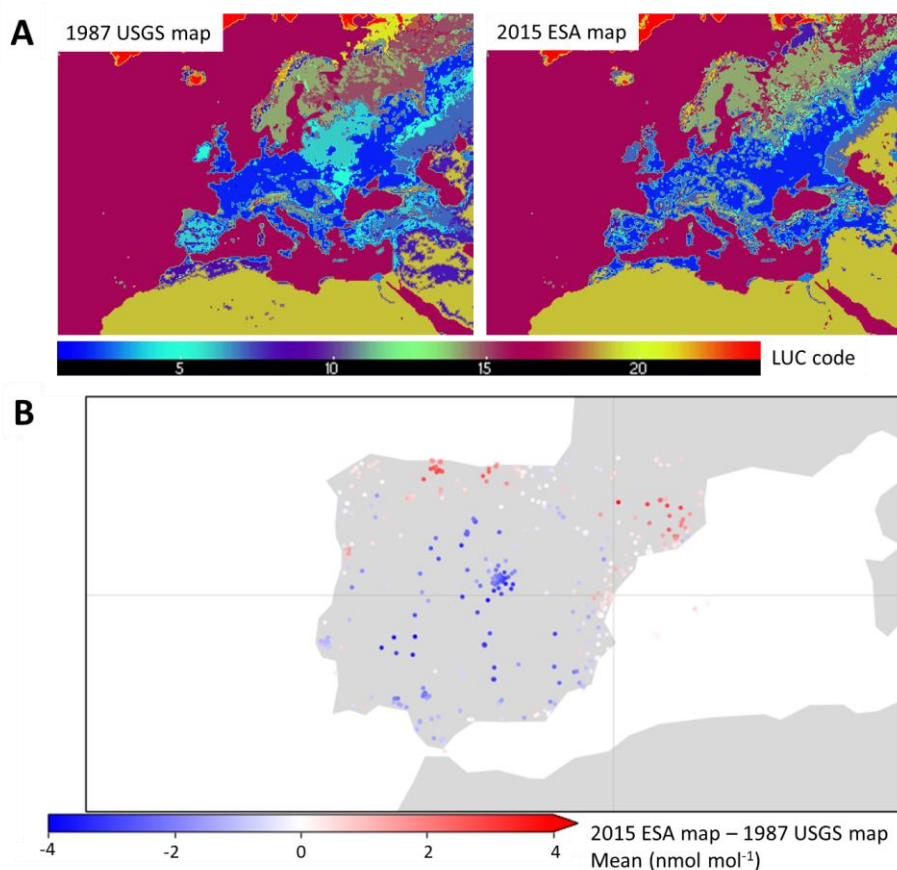
Fig. 1 (A) 1987 USGS and 2015 ESA land cover maps of Europe. USGS 24-category codes were used in the maps. (B) Mean ozone surface concentration difference (nmol mol^{-1}) between MONARCH simulations using the 2015 ESA land cover map and the 1987 USGS land cover map over the Iberian Peninsula EEA AQ e-Reporting sites for July 2019.

We have shown in this work the benefit of using land use information extracted from a satellite-derived product developed under the ESA CCI at a higher spatial resolution compared to what we had previously implemented in our system. Furthermore, the ESA CCI land cover dataset provides a long-term, consistent, global data record at a yearly resolution from 1992 to 2020 that could be further exploited by our long-term simulation runs. In addition, in order to investigate further the uncertainty in ozone modelling, future work should include the effects on the simulations by changes on the deposition parametrization and on the temporal and spatial emissions of the volatile organic compounds and nitrogen dioxide that act as precursors for ozone.

References

[1] P. J. Crutzen, "Photochemical reactions initiated by an influencing ozone in the unpolluted troposphere", Tellus, Vol. 26, Pp. 47-57, 1974.

[2] WHO, "WHO global air quality guidelines. Particulate matter (PM 2.5 and PM 10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.", Geneva: World Health Organization, 2021.

[3] A. Dunker, U. Nopmongcol, and G. Yarwood, "Uncertainty analysis of modeled ozone changes due to anthropogenic emission reductions in Eastern Texas," Atmos. Environ., Vol. 268, Pp. 118798, 2022.

[4] ESA, "Land Cover CCI Product User Guide Version 2", Tech. Rep., Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, 2017.

[5] A. Badia, O. Jorba, A. Voulgarakis, D. Dabdub, C. Pérez García-Pando, A. Hilboll, M. Gonçalves, and Z. Janjic, "Description and evaluation of the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (NMMB-MONARCH) version 1.0: gas-phase chemistry at global scale," Geosci. Model Dev., Vol. 10, Pp. 609-638, 2017.

Author biography

Franco López received the MSc degree in Environmental Engineering from the Universitat Politècnica de Catalunya (UPC) in 2021. He did an internship in the Atmospheric Composition group – Earth Science department at the Barcelona Supercomputing Center (BSC).

Low Input Promoter Capture Hi-C Method Enables to Decipher the Molecular Mechanisms Underlying Genetically Complex Diseases

Paula López-Martí [1,2], Biola M Javierre[2], Alfonso Valencia[1]

[1] Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain

[2] 3D Chromatin Organization, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Spain

E-mail: `paula.lopez@bsc.es`, `bmjavierre@carrerasresearch.org`, `alfonso.valencia@bsc.es`

Keywords— **Capture Hi-C, epigenetics, chromatin organization**

EXTENDED ABSTRACT

A. Introduction

Interactions between promoters and their regulatory elements (e.g., enhancers) are cell-type specific. Moreover, these promoter-enhancer contacts have been shown to be crucial for proper gene expression levels, and their alteration promotes disease [1, 2]. The most widely used experimental methods require tens of millions of cells to obtain the whole high-resolution promoter associations, also called promoter interactome, independently of their activity. This requirement precludes the genome-wide analysis of regulatory promoter interactions in disease-relevant patient samples. This gap of knowledge is particularly problematic for understanding the molecular mechanism underlying inherited risk factors for common human diseases and acquired mutations and epimutations, which all are all highly enriched at regulatory elements.

To provide novel insight into global genomic regulatory mechanisms and gene pathways underlying disease pathologies, Javierre et al. implemented promoter capture Hi-C (PCHi-C) method [1]. It allows systematic genome-wide identification of the genomic regions, including distal regulatory regions, in physical proximity with more than 22,000 promoters independently of the activity status of the interacting regions. In addition, due to PCHi-C method relying just on the sequence capture technology, but not in antibody immunoprecipitation as other methods do, it enables the robust compassion between conditions or cell types and the customization to interrogate any specific interactomes (e.g., enhancer interactome or the interactome in which a collection of variants). Applying this method, it has been possible to associate non-coding disease-associated variants to their distal target promoters, identifying hundreds of potential new disease-candidate genes and/or gene pathways. However, PCHi-C relies on the availability of millions of cells, typically from 20 to 50 million per biological replicate, which prohibits the analysis of rare cell populations such as those commonly obtained in clinical settings.

B. Methodology and results

Here, to circumvent this shortcoming, we present a low input capture Hi-C (liCHi-C) method that allows the generation of high-quality genome-wide promoter interactome maps using very low amounts of starting material. We validate our new method by comparing promoter interactomes for a controlled cell titration of primary human cells (human naïve B cells) against the highest resolution PCHi-C datasets until the date, demonstrating that the interactomes are robust down to 50,000 cells of starting material.

As a proof of its potential for discovering new insights, we used liCHi-C to study normal and malignant human hematopoiesis. We demonstrate that liCHI-C foreshadows developmental trajectories and that the promoter architecture dynamics reshuffle transcriptional decisions along *in vivo* cell commitment. Besides, by using a Bayesian prioritization strategy called Capture Hi-C Omnibus Gene Score (COGS) [1], we prove the applicability of liCHi-C to associate distal regulatory elements and new disease-related non-coding alterations with potential target genes and gene pathways in sparse primary cell types.

Finally, we show its broad applicability by analyzing blasts from acute lymphoblastic leukemias (ALL) from pediatric aspirates, where we demonstrate the power of liCHi-C to improve our understanding of cancer by diagnosing patient specific chromosomal rearrangements and copy number variations (CNVs), as well as identify cancer-specific topological features.

References

[1] Javierre BM, Burren OS, Wilder SP, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-1384.e19.

[2] Choy MK, Javierre BM, Williams SG, et al. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. Nat Commun. 2018;9(1):2526.

Author biography

Paula López was born in Barcelona, Spain in 1997. She received the Bachelor's Degree in Biomedical Sciences from the University of Barcelona in 2019 and the Master's Degree in Bioinformatics for Health Sciences from University Pompeu Fabra in 2021, Barcelona, Spain. She joined Alfonso Valencia's Computational Biology group as a master's student in 2020, and in October 2021 she started her PhD in collaboration with Biola Javierre's 3D Chromatin Organization group in Josep Carreras Leukaemia Research Institute (IJC).

FAIRsoft - A practical implementation of FAIR principles for research software

Eva Martín del Pico*, Salvador Capella-Gutiérrez† Josep Lluís Gelpí Buchaca‡,
*†‡Barcelona Supercomputing Center, Barcelona, Spain
‡Universitat de Barcelona, Barcelona, Spain
Email: *eva.martin@bsc.es, †salvador.capella@bsc.es, ‡gelpi@ub.edu

Keywords—FAIR, Open Science, Research Software

I. Extended Abstract

Computational tools are increasingly becoming constitutive parts of scientific research, from experimentation and data collection to the dissemination and storage of results. Unfortunately, however, research software is not subjected to the same requirements as other methods of scientific research: being peer-reviewed, being reproducible and allowing one to build upon another's work. This situation is detrimental to the integrity and advancement of scientific research, leading to computational methods frequently being impossible to reproduce and/or verify [1]. Moreover, they are often opaque, direcly unavailable or impossible to use by others [2]. One step to address this problem could be formulating a set of principles that research software should meet to ensure its quality and sustainability, resembling the FAIR (Findable, Accessible, Interoperable and Reusable) Data Principles [3]. The FAIR Data Principles were created to solve similar issues affecting scholarly data, namely great difficulty of sharing and accessibility, and are currently widely recognized accross fileds. We present here FAIRsoft, our initial effort to assess the quality of research software using a FAIR-like framework, as a first step towards its implementation in OpenEBench [4], the ELIXIR benchmarking platform.

A. Proposal for a FAIRsoft scoring system

We analysed the FAIR principles for research software [5] to formulate an initial strategy to obtain a quantitative evaluation. The result is a set of measurable indicators generated following a two steps approach. In a first step, we derived a number of requirements that software must fulfil in order to be *Findable, Accessible, Interoperable* and *Re-Usable*, respectively. We call these properties high-level indicators. The second step took us to the desired degree of granularity through low-level indicators. A low-level indicator is one condition that contributes to a software meeting a high-level indicator. To allow for a practical evaluation, each low-level indicator is associated with a well defined evaluation procedure.

B. Measurement of software FAIRness

For each tool, FAIRsoft indicators can be measured using, in most cases, metadata from more than one reference source, which must be accessible and findable by any user in order to be valid. These sources, that include software registries and repositories, e-Infrastructures, software homepages and journal publications, can be extended to increase the indicators coverage for individual entries. We integrated metadata from Bio.tools, Bioconda, Bioconductor, Galaxy ToolShed, SourceForge , and Galaxy Europe as primary sources to discover tools and retrieve an initial collection of metadata. This collection was enriched mining secondary sources: Github, Bitbucket, OpenEBench, PubMed, Europe PMC and Wikidata. The resulting set of metadata was restructured to fit a common data model and then integrated by software instance.

Finally, FAIRsoft scores were computed for each software instance. To this end, low-level indicators were calculated and subsequently combined to generate high-level indicator scores. A weighting scheme was designed and implemeted to reflect the varying importance of individual low-level indicators.

C. Results and Discussion

We obtained a collection of 43,973 unique software instances, 71.4% of which were enriched with metadata from more than one source. The number of instances with available metadata varied greatly among sources, as well as the type of information they provided.

Results for the general four FAIR principles for research software are heterogeneous, as shown in Fig 1. However, patterns can be easily identified among them. Indicator scores for the *Findability* of research software are higher than for other principles, with a remarkable 44.0 % of published software. A lack of structured metadata is the main reason for lower *Findability* scores. The actual usability of the software includes indicators from both *Accessibility* and *Re-Usability* principles. We found that 15.6% of analysed instances score optimally for the main indicator of *Accessibility* (*Existance of a downloadable, buildable or accessible working version of the software*) and 47.6% for the main indicator of *Re-usability* (*Existance of License*), with an impressive predomination of Open Source Licenses (86.0%). Finally, *Interoperability* is the lowest scoring principle, partially due to the absense of agreed standards on how to represent software interoperability, making it difficult to define automatically measurable indicators. Nonetheless, exceptional cases exist, and there is abundant literature on what software interoperability is and how it can be measured both in terms of working with other research software as part of analytical workflows [6], and in terms of interoperating with underlying software components such as software libraries. Indeed, we found information about data types and formats, as well as about dependencies, is only structured when it is required to be machine-readable as in packages repositories (instances enriched with metadata from Bioconda and Bioconductor).
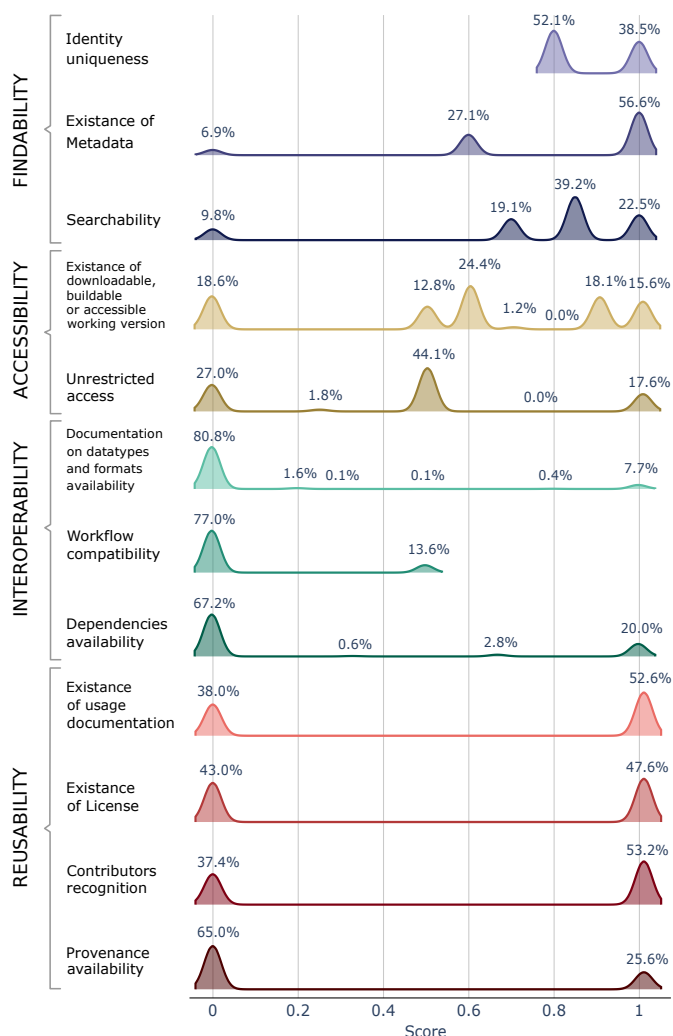
Fig. 1. High-level indicators scores of instances. Each possible score is labelled with the percentage of instances scoring it. Although scores are discrete values, they are shown as density plots for clarity.

D. Final remarks

The proposed FAIRsoft indicators can definitely contribute towards the consolidation of the FAIR principles for research software, driven by community initiatives, e.g. FAIR4RS, a joint work by the Research Data Alliance (RDA), Research Software Alliance (ReSA) and FORCE11. Indeed, this first generation of indicators should serve to improve automated measurement algorithms as well as to reflect the contribution of specific indicators to the four general principles.

This work should be considered an initial effort for having a quantitative overview of the common practices for developing research software in the Life Sciences. The ultimate goal is to contribute to the reproducibility and reliability of scientific outcomes by putting the focus on one of the key elements for success: Research Software. Periodical assessment of research software FAIRness would allow researchers to understand existing and emerging trends and can serve to identify areas of potential improvement for facilitating the generation of high-quality research software.

References

[1] D. B. Allison, A. W. Brown, B. J. George, and K. A. Kaiser, "Reproducibility: A tragedy of errors," *Nature*, vol. 530, no. 7588, pp. 27–29, Feb. 2016. [Online]. Available: http://www.nature.com/articles/530027a

[2] A. Morin, J. Urban, P. D. Adams, I. Foster, A. Sali, D. Baker, and P. Sliz, "Shining Light into Black Boxes," *Science*, vol. 336, no. 6078, pp. 159–160, Apr. 2012, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/science.1218263

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, p. 160018, Dec. 2016. [Online]. Available: http://www.nature.com/articles/sdata201618

[4] S. Capella-Gutierrez, D. d. l. Iglesia, J. Haas, A. Lourenco, J. M. Fernndez, D. Repchevsky, C. Dessimoz, T. Schwede, C. Notredame, J. L. Gelpi, and A. Valencia, "Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking," bioRxiv, Tech. Rep., Aug. 2017, section: New Results Type: article. [Online]. Available: https://www.biorxiv.org/content/10.1101/181677v1

[5] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpi, N. Chue Hong, C. Goble, and S. Capella-Gutierrez, "Towards FAIR principles for research software," *DS*, vol. 3, no. 1, pp. 37–59, Jun. 2020. [Online]. Available: https://content.iospress.com/articles/data-science/ds190026

[6] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M. R. Crusoe, K. Peters, and D. Schober, "FAIR Computational Workflows," *Data Intelligence*, vol. 2, no. 1-2, pp. 108–121, 2020.

Eva Martín del Pico received her BSc degree in Biochemistry from Autonomous University of Madrid (UAM), Spain in 2016. She completed her MSc degree in Bioinformatics for Health Sciences from Pompeu Fabra University, Spain in 2019. Since then, she is a PhD student at the Coordination Node of the Spanish National Bioinformatics Institute (INB) at Barcelona Supercomputing Center (BSC).

Task Scheduling Sensitivity to L1 Cache settings on an area-constrained 32-core RISC-V Processor

Lucas Morais*†, Daniel Jiménez-González*†, Carlos Álvarez*†

*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {lucas.morais, daniel.jimenez, carlos.alvarez}@bsc.es

Keywords—RISC-V, FPGA, Rocket Chip, Task Scheduling, Cache, Design Exploration.

I. Extended Abstract

High-performance applications are highly sensitive to memory performance characteristics. While programs with comparatively low memory-to-computation ratio are less likely to be hampered by limited memory bandwidth, most parallel applications will be severely impacted by the absence of hardware support for low-latency inter-thread synchronization and data sharing.

In this paper, we report a design exploration that sought to identify the cache configuration that maximizes performance of task parallel OpenMP workloads running on a Linux-capable 32-core RISC-V system. We show that, under the constraints of a U200 Alveo FPGA, the best single-level cache configuration consists in 160 KB of coherent, core-private data caches, with a 32/128 split among instruction and program data. With such configuration, we have achieved speedups of up to 28x and 19x for the nbody and cholesky applications, respectively.

A. Background

Rocket Chip [1] is a popular open-source project allowing rapid prototyping of RISC-V systems. While systematic design explorations [2] and tools to that purpose [3] have been proposed in the past, no work had so far focused on optimizing cache configuration for Task Scheduling applications.

Such optimal cache parameters should provide performance gains complementary to those provided by low-latency hardware-accelerated data dependence resolution [4], [5].

B. Rocket Chip's cache parametrization features

Rocket Chip offers a rich set of parameters that might be used to control the organization and characteristics of many of its micro-architectural elements. For example, parameters exist that control, on a core-specific manner, whether a FPU should be integrated or how many miss status handler registers should be included. More to the point of our paper, there are options to control the number of cache sets and ways, as well as the cache line size. It is also possible to select a random, Pseudo-LRU, or a True LRU cache replacement policy.

In our different experiments, we modulated cache capacity by accordingly varying the number of ways of each cache element, as this proved to require much less FPGA resources than changing the number of sets, for example, since other options render the design much less amenable to be mapped to efficient FPGA storage elements.

	LUT	LUTRAM	FF	BRAM
Alveo U-200	1.18M	592K	2364K	2.16K
I-16, D-128	1.07M (90%)	32.6K (6%)	574K (24%)	1.91K (89%)
I-32, D-64	1.04M (88%)	31.9K (5%)	568K (24%)	1.32K (61%)
I-64, D-64	1.08M (91%)	32.6K (6%)	596K (25%)	1.67K (77%)
I-32, D-128	1.08M (92%)	31.9K (5%)	584K (25%)	2.09K (97%)

TABLE I: Resource usage of the different configurations.

Table I shows the resource usage of the different configurations used in the performed study. Each implemented core is an in-order Rocket core, single issue with a floating point ALU. The system runs at 60MHz in FPGA and is Linux-capable.

C. Benchmarks

We evaluate different cache configurations using four benchmarks: Cholesky, N-Body, ramspeed-reading, and ramspeed-writing. Cholesky and N-Body are extensively-used HPC kernels implemented using the Task Parallel paradigm, while the latter two are variants of a Linux tool for evaluating memory performance according to different access patterns.

More specifically, Cholesky is a kernel implementing a matrix decomposition method that is employed, for example, in linear equation solving. In turn, N-Body encapsulates programming logic that is typically found in applications where, given a system of particles, interactions among every pair of particles are evaluated at each simulation step.

Both Cholesky and N-Body make use of OpenMP's automatic data-dependence resolution. However the memory requirements for N-Body are usually lower than for Cholesky. N-Body computation grows with the square of the number of elements. In Cholesky computation grows with the cube of the problem size but data grows square with the problem size usually putting more pressure in the memory system. Both applications have been executed with the optimum task granularity for the system evaluated.

In the case of ramspeed, it evaluates the performance of a series of readings and writings in memory, giving an idea of the scalability of the memory subsytem.

D. Preliminary Results

Figure 1 shows the speedup results of the different benchmarks when executed in a different number of cores against the same application executed sequentially in the same system. Each subfigure shows the performance obtained in each of the different cache configurations analyzed. In addition, Table II shows the absolute performance obtained in each execution.

Our experiments show that the (I-16, D-128) performs poorly in spite of its large data-cache component, which
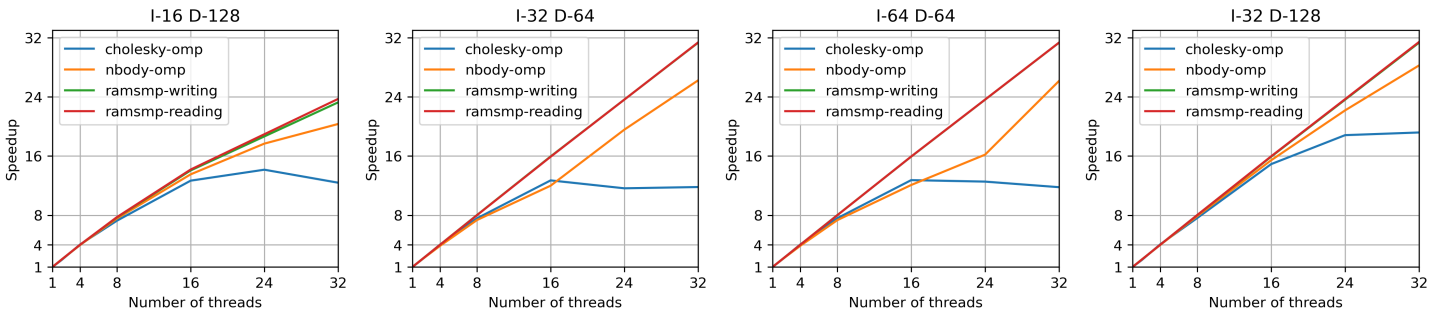
Fig. 1: Performance results across different cache designs.

	1	2	4	8	16	24	32
I-16 D-128							
cholesky	31.51	61.24	127.31	228.67	399.22	445.92	390.62
nbody	0.43	0.86	1.69	3.22	5.77	7.54	8.67
ram-reading	0.44	0.88	1.75	3.37	6.14	-	10.16
ram-writing	0.22	0.45	0.90	1.74	3.18	-	5.32
I-32 D-64							
cholesky	30.22	59.43	121.75	228.90	384.01	351.68	356.87
nbody	0.43	0.85	1.66	3.16	5.16	8.42	11.28
ram-reading	0.45	0.90	1.81	3.62	7.20	-	14.14
ram-writing	0.23	0.46	0.93	1.86	3.69	-	7.28
I-64 D-64							
cholesky	30.14	59.36	121.86	228.80	384.05	378.02	355.27
nbody	0.43	0.85	1.66	3.15	5.20	6.96	11.25
ram-reading	0.45	0.90	1.81	3.62	7.20	-	14.16
ram-writing	0.23	0.46	0.93	1.85	3.70	-	7.27
I-32 D-128							
cholesky	32.76	63.49	133.30	249.62	488.36	616.61	628.48
nbody	0.44	0.88	1.76	3.48	6.81	9.79	12.47
ram-reading	0.45	0.91	1.81	3.62	7.19	-	14.15
ram-writing	0.23	0.46	0.93	1.86	3.70	-	7.27
	1	**2**	**4**	**8**	**16**	**24**	**32**
				Number of threads			

TABLE II: Performance results. Executions of ramsmp are evaluated in gigabytes per second, while cholesky and nbody performances are measured in megaflops and mega-pairs per second, respectively. Ramsmp only supports power-of-two numbers of threads.

suggests that 16 KB is not enough for ensuring scalability for a system with that much data cache. Moreover, the single-threaded performance of such configuration is substantially inferior to that of all other arrangements.

We also notice that the (I-32, D-64) and (I-64, D-64) designs obtain nearly the same performance for all configurations involving a number of threads different from 24, although the latter features double the data cache of the former. This is evidence that instruction-fetching overhead and related problems are not likely to be reduced by increasing instruction caches beyond 32 KB, at least for the tested workloads. On the other hand, the differences at 24 threads suggest that increasing the instruction cache beyond 32 KB can have both positive or negative effects on load balancing depending on the workload.

E. Conclusions

This work reports how Rocket Chip cache parameters might be selected to obtain maximum performance out of selected Task Parallel applications under the resource constraints of the U200 Alveo FPGA. We show that, among the considered configurations, only the one featuring 160 KB of L1 cache split

in 32/128 instruction/data ratio can ensure scalability beyond 16 threads for a standard task-parallel cholesky implementation. We also indicate that, while serving 32 threads, the (32-I, 128-D) configuration improves performance by up to 54% with respect to (16-I, 128-D) or 76% with respect to (64-I, 64-D), while requiring just 11% or 25% more storage than each of these options, respectively.

As future work, we propose to evaluate how Task Parallel applications might benefit from L2 caches mapped to URAM resources. Also, we would like to evaluate whether our scalability conclusions also hold for runtimes other than OpenMP.

F. Acknowledgement

REFERENCES

[1] K. Asanovic *et al.*, "The rocket chip generator," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016-17*, vol. 4, 2016.

[2] M. Doblas Font *et al.*, "Microarchitectural design-space exploration of an in-order risc-v processor in a 22nm cmos technology," B.S. thesis, Universitat Politècnica de Catalunya, 2020.

[3] S. Bandara *et al.*, "Brisc-v: An open-source architecture design space exploration toolbox," *arXiv preprint arXiv:1908.09992*, 2019.

[4] L. Morais *et al.*, "Adding tightly-integrated task scheduling acceleration to a risc-v multi-core processor," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 861–872.

[5] J. M. de Haro *et al.*, "Towards reconfigurable accelerators in hpc: Designing a multipurpose efpga tile for heterogeneous socs," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022.

Lucas Morais has a Master Degree in Computer Science from the University of São Paulo, where he worked with an international team for adding native Task Parallelism support to RISC-V based multi-core systems. Previously, he received a degree in Computer Engineering from the University of Campinas, Brazil, being accoladed as the best student of his class by the Engineering Council of the São Paulo State. He is currently pursuing a PhD at Universitat Politècnica De Catalunya.

Rational Enzyme Engineering of Different Active Sites on a Xylanase

Rubén Muñoz[#1], Victor Guallar[#2]

#Barcelona Supercomputing Center, Barcelona, Spain
[1]ruben.munoz@bsc.es, [2]victor.guallar@bsc.es

Keywords—— **Protein-engineering, esterase, xylan**

EXTENDED ABSTRACT

In the current times, many processes in industry are switching from using chemical compounds, requiring extreme conditions, to enzymatic technology, which is a more sustainable and greener alternative [1]. In the paper industry, xylanases are enzymes used to facilitate the removal of residual lignin, which gives a dark colour to the paper pulp since these enzymes carry out the depolymerization of the xylan [2].

Xylan polymer consists of a chain of xylose monomers from where several different radicals hang. One of these radical molecules is arabinose, and, in addition, there is sometimes a ferulate bonded to those arabinoses. Usually, to cleave the ferulate from xylan, an enzyme called feruloyl esterase is used, an esterase that gives as a product ferulic acid. This compound is similar to a hydroxycinnamic acid existing in plant materials, mainly used in pharmaceuticals, cosmetics, and food industries owing to its antioxidant, anti-inflammatory, and anti-cancer biological activities [3].

In this work, several mutations are proposed in an endo-xylanase to generate a second catalytic centre, different from the wild-type one, with feruloyl esterase activity. This would allow the enzyme to cut the bond between arabinose and ferulate and thus substantially improve the usage of the enzyme in industry, as only one enzyme should be needed to perform two different reactions [4].

Methods

To design a new active site in the enzyme, we must first explore the protein's surface, searching for unknown binding sites. To accomplish this, we use PELE (Protein Energy Landscape Exploration) software, which is a Monte Carlo technique combined with side-chain prediction [5].

As the xylan polymer is a large structure, the arabinose bonded to a ferulate was used as the ligand in the preliminary PELE global simulations. A bigger ligand was tested once a suitable pocket was found to simulate a more realistic substrate.

Several mutations were introduced and tested with PELE to accommodate the ferulate inside the pocket and incorporate the esterase catalytic residues. Finally, MDs (Molecular Dynamics) were used to simulate the best mutants from PELE in a more realistic environment.

Results

After running the first PELE global exploration, a pocket different -and far away- from the main active site was found to bind the ferulate with slightly better energies than the rest of the protein surface. A PELE local simulation was then performed using the larger ligand. Several mutations were introduced to enlarge the cavity and make it more suitable to bind the ferulic acid hanging from the xylan polymer. After several mutants were tested with PELE, the most promising
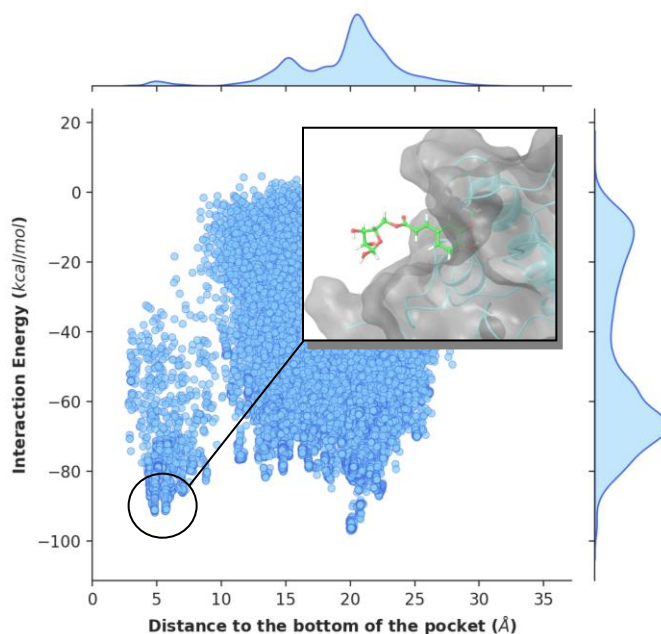

Fig. 1 Plot representing the distance of one substrate's atom to a protein atom at the bottom of the pocket against the interaction energy. The image shows a snapshot of when the ferulic acid enters a protein's pocket.

one showed a significant increase in the binding energy between being inside and outside the pocket.

Several catalytic mutations containing the serine-histidine-aspartic triad were introduced to the mutant with the new pocket residues, and PELE simulations were performed. Once we validated them with MDs simulations, the three most promising mutants were proposed to test experimentally: two different catalytic triads, and one of them with and without the pocket mutations.

References

[1] S. RA. Krüger, C. Schäfers, C. Schröder, and G. Antranikian, 'Towards a sustainable biobased industry – Highlighting the impact of extremophiles', N. Biotechnol., 2018.

[2] P. Bajaj and R. Mahajan, 'Cellulase and xylanase synergism in industrial biotechnology', Appl. Microbiol. Biotechnol., 2019.

[3] S. Liu et al., 'Directed evolution of feruloyl esterase from Lactobacillus acidophilus and its application for ferulic acid production', Bioresour. Technol., 2021.

[4] G. Santiago et al., 'Rational Engineering of Multiple Active Sites in an Ester Hydrolase', Biochemistry, 2018.

[5] Kenneth W. Borrelli, Andreas Vitalis, R. Alcantara, and V. Guallar*, 'PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique', J. Chem. Theory Comput. 2005.

Author biography

Rubén Muñoz is a PhD student at Barcelona Supercomputing Center, in the EAPM group led by Victor Guallar. After graduating in Biochemistry and Molecular Biology at the Rovira I Virgili University (Tarragona), he got his MSc studies in Biophysics at the Autonomous University of Madrid. In BSC, he designed two different pluriZymes, enzymes with more than one active centre. He is currently involved in a European project called FuturEnzymes and a Spanish national project called Furenpol, working with enzymes that synthesize and degrade plastics polymers.

Characterisation of Ozone levels and associated NMVOC emissions in Spain: a preliminary assessment

Kevin Oliveira*, Marc Guevara*,
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {kevin.deoliveira, marc.guevara}@bsc.es

Keywords—Tropospheric ozone, Atmospheric emissions, NMVOC speciation, Emission control strategies.

I. Extended Abstract

In Europe, for the last decades, multiple efforts have been made to regulate and reduce air pollution. Despite the overall reduction trends in emissions, air quality remains poor in many areas, where 99% of the urban population is exposed to tropospheric ozone (referred to as O_3 from here on) concentrations above the World Health Organization Air Quality Guideline value [1].

O_3 is a secondary atmospheric pollutant with complex mechanisms of formation and reaction processes, transport, and deposition. The nonlinear nature of its production is one of the main challenges in controlling ozone levels [2].

Spain shows several areas with O_3 problems. Regarding the target value of O_3 for health protection, out of the 127 areas where it was evaluated in 2019, 34 of them show values above the target value, 81 values between the target value and the long-term target value, and the remaining 12 are below the long-term objective. These results have not substantially improved since 2011 and are constant since 2016 [3].

The Spanish Ministry is currently designing a strategy to tackle the O_3 problem in Spain. In this context, air quality modelling systems become an important complementary tool on which to quantify and evaluate the impact of such air quality plans.

Two major directly emitted precursors drive O_3 formation: nitrogen oxides (NOx) (which includes both nitrogen monoxide, NO, and nitrogen dioxide, NO_2) and non-methane volatile organic compounds (NMVOC), so in order to assess O_3, a correct characterization of its precursors is required.

This work focuses on producing a speciated NMVOC inventory for anthropogenic emissions in Spain to support more effective control strategies. The type of NMVOCs emitted varies widely from one source to another, differing substantially by, e.g. fuel, technology, and others. The significant differences between NMVOC species lead to differences in atmospheric chemical reactivity and result in differences in their influence on the formation of ozone. Therefore, speciated NMVOCs emissions and the estimation of ozone formation potential (OFP) are essential to the reactivity-based control approach.

A. Methodology

In this study, in order to develop a speciated inventory for NMVOC, we select all SNAPs (Selected Nomenclature for reporting of Air Pollutants) related to anthropogenic emissions (i.e. SNAPS 1 to 10). For Spain, for the year 2019, we worked with 158 SNAP sectors. The speciated profiles are crucial to a correct characterization of the inventory. Different profile sources were used, where more recent and/or well-defined profiles were chosen. We used 105 profiles, where the majority come from the SPECIATE database [4], followed by EMEP/EEA guidelines. Specific profiles from state of the art were used when available. Resulting in a total of over 800 species.

The emission of an individual NMVOCs species was calculated as:

$$E_i = \sum_j E_j \times R_{ij} \tag{1}$$

where i is a specific NMVOCs species, j is the emission source, is the total emission of the species i, is the total emission of the source j, and is the ratio of species i to source j.

Then the OFP of an individual NMVOCs species for all sources was calculated as:

$$OFP_i = \sum_j E_{ij} \times MIR_i \tag{2}$$

where OFP_i is the total ozone formation potential of species i, E_{ij} is the emission of the species i for source j, and MIR_i is the maximum increment reactivity of species i. The updated MIR values from [5] were adopted in this work.

B. Results

Figure 1 shows the percentage contribution of each SNAP group for the total emissions and for the total OFP. Despite the slight reduction (<3%) in the contribution of SNAP 6 (solvent use), it's still the major contributor for OFP with around 45%, followed by SNAP 10 (livestock) with 12%, and SNAP 8 (Other mobile sources and machinery) with 11%. The latter has a similar emission contribution as SNAP 2 but gains importance when looking at OFP.

The top 10 major SNAP subgroup sectors contributing to the total of OFP show the majority belonging to the SNAPS
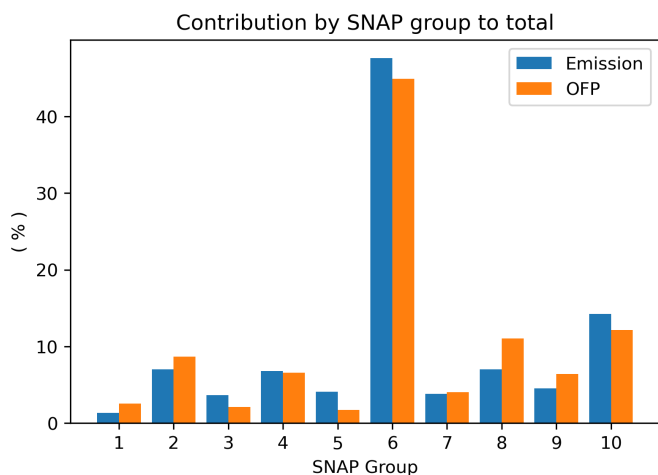
Fig. 1. Contribution, in percentage, of each SNAP sector group to the total of emissions and OFP.

previously identified as major contributors, but other subgroups appear as important as, e.g. residential combustion (0202), burning of agricultural waste (0907), and industry (0405/06).

Figure 2 shows the composition of species emissions for each SNAP where sectors such as the distribution of oil products (SNAP 5) and combustion in the manufacturing industry (SNAP 3) are mainly emitting alkanes (low reactivity). While farm-level agricultural operations (SNAP 10) and production processes (SNAP 4) show a majority of OVOCs.
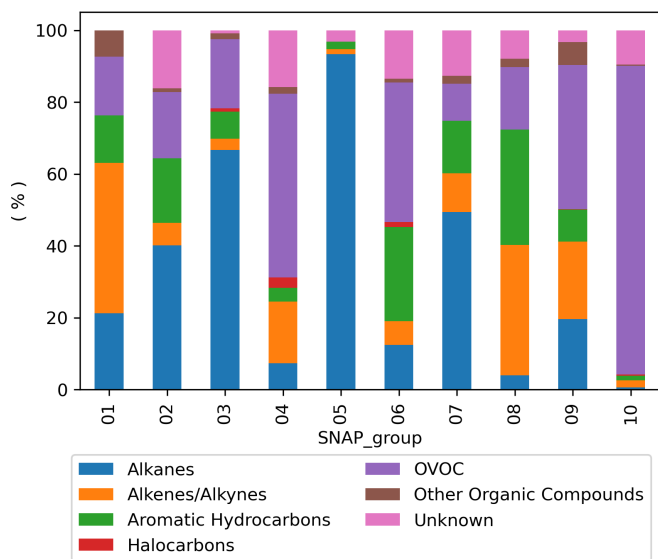

Fig. 2. Composition by grouped species of each SNAP group emissions.

C. Conclusion

When focusing on O_3 problems, this methodology allows prioritising sectors based on their importance to ozone formation, proving to be helpful in terms of control strategies.

The main challenges of adequate characterization of the real world NMVOC emissions are mainly related to the lack of continuous measurements of NMVOCs compounds, limited emission inventories reporting by compound, and lack of recent and more specific speciated profiles for different technologies/fuels. Further complicating this issue is a lack of understanding of the relative importance as to whether total NMVOC emissions or NMVOC speciation is more, less, or equally important, including the relative importance of the changes in speciation and total emissions over time.

Future works will include performing a spatial and temporal analysis of the distribution of NMVOC speciated emissions across Spain. The spatial analysis will allow to determine the importance of acting toward different sectors depending on the area of Spain. While the temporal analysis (i.e. monthly) will deliver information on how to tackle specific sectors based on their activity profile throughout the year. For that, we will make use of the in-house emission model system HERMESv3 [6], which is capable of computing emissions at high spatial (up to 1kmx1km) and temporal resolution (1 hour). To investigate the impact of using more refined NMVOC speciated emissions on the modelling of O_3 levels using the MONARCH atmospheric modelling system [7], which has also been developed at BSC.

II. Acknowledgment

References

[1] EEA, "Air quality in europe—2020 report," 2020.

[2] X. Querol, A. Alastuey, C. Reche, A. Orio, M. Pallares, F. Reina, J. Dieguez, E. Mantilla, M. Escudero, L. Alonso, G. Gangoiti, and M. Millán, "On the origin of the highest ozone episodes in spain," *Science of The Total Environment*, vol. 572, pp. 379–389, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969716316473

[3] MITECO, "Evaluación de la calidad del aire en españa 2019," 2019. [Online]. Available: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/informeevaluacioncalidadaireespana2019$_t$cm30 − 510616.pdf

[4] H. Simon, L. Beck, P. V. Bhave, F. Divita, Y. Hsu, D. Luecken, J. D. Mobley, G. A. Pouliot, A. Reff, G. Sarwar, and M. Strum, "The development and uses of epa's speciate database," *Atmospheric Pollution Research*, vol. 1, no. 4, pp. 196–206, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1309104215305250

[5] M. A. Venecek, W. P. Carter, and M. J. Kleeman, "Updating the saprc maximum incremental reactivity (mir) scale for the united states from 1988 to 2010," *Journal of the Air & Waste Management Association*, vol. 68, no. 12, pp. 1301–1316, 2018.

[6] M. Guevara, C. Tena, M. Porquet, O. Jorba, and C. Pérez García-Pando, "Hermesv3, a stand-alone multi-scale atmospheric emission modelling framework–part 2: The bottom–up module," *Geoscientific Model Development (GMD)*, vol. 13, no. 3, pp. 873–903, 2020.

[7] A. Badia, O. Jorba, A. Voulgarakis, D. Dabdub, C. Pérez García-Pando, A. Hilboll, M. Gonçalves Ageitos, and J. Zavisa, "Description and evaluation of the multiscale online nonhydrostatic atmosphere chemistry model (nmmb-monarch) version 1.0: gas-phase chemistry at global scale," *Geoscientific Model Development*, vol. 10, pp. 609–638, 2017.

Kevin Oliveira received his MSc degree in Environmental Engineering from the University of Aveiro (UA), Portugal in 2018. In the same year, he started working at the research group on emissions, modelling and climate change (GEMAC) from the Department of Environment and Planning of the UA. In 2021, he joined the Atmospheric Composition group of Barcelona Supercomputing Center (BSC) as well as a PhD student of the Universitat Politècnica de Catalunya (UPC), Spain.

b8c: an FPGA-Friendly Sparse Matrix Representation Suitable for the SpMV kernel

José Oliver*†, Eduard Ayguadé*†, Xavier Martorell*† Carlos Álvarez*†

*Barcelona Supercomputing Center, Barcelona, Spain

†Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {jose.oliver, eduard.ayguade, xavier.martorell, carlos.alvarez}@bsc.es

I. EXTENDED ABSTRACT

Sparse Matrix-Vector multiplication (SpMV), computing $y = Ax$ where y and x are dense vectors and A is a sparse matrix, is a key kernel in many HPC applications. SpMV exhibits a kind of memory access that is extremely hard to perform efficiently, due to its random access. In the case of FPGAs, which lack a "default" memory hierarchy including caches to hide the latency of main memory, obtaining good performance when running SpMV is specially challenging. In our work, we have explored different approaches to the SpMV implementation on FPGAs, which have result in the definition of a new sparse matrix encoding format (**b8c**) and its corresponding SpMV implementation using OmpSs@FPGA. In our tests, the **b8c** SpMV implementation shows ≈x25 performance improvement when compared to an HLS-optimized Compressed Sparse-Row (CSR) FPGA version and offers, running at 100MHz, 20% of the performance of a single-core Ryzen 7 running at 3.9GHz and using the same CSR format.

A. Background

Many works build on top of CSR or an alternate representation, ELL(PACK), in order to propose formats suitable for specific instruction sets, like SIMD [1], for vector processors [2] or even for GPUs [3]. The assumption of a cache hierarchy present in the device plus a mechanism for a relatively high number of outstanding memory requests make these formats not suitable for FPGAs targeting high-level synthesis (HLS). There have been different proposals for FPGA-specific matrix encodings and algorithms, even some of them based on CSR or variations of it [4]. Most of these FPGA-specific format proposals, however, target low-level kernel implementations using register transfer level languages (RTL) that are hard to customize and include in high-level synthesis, or require complex architectures that consume a high number of resources, making them difficult to scale in number. Additionally, most of them are limited to process 1 or 2 elements per cycle. By introducing **b8c** we aim to provide a sparse matrix representation that can be used to improve the performance of SpMV operations in FPGA by exposing high levels of parallelism that can be processed without requiring excessive amounts of resources by specialized HLS kernels that can be easily customized and integrated as part of larger projects.
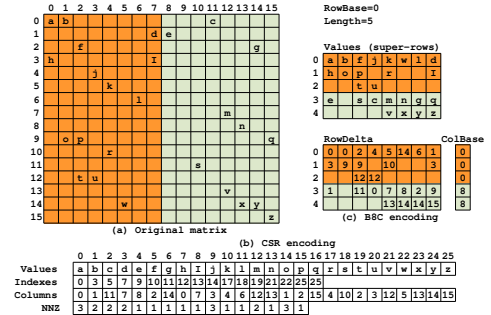

Fig. 1: CSR and b8c encodings for a sparse matrix

B. b8c Matrix Encoding and Algorithm

The **b8c** representation of a sparse matrix seeks to achieve the following goals: a) allow the parallel process of N_elems elements per cycle b) reduce as much as possible non-useful data (padding) c) generate a structure that can be easily partitioned in tasks among different OmpSs@FPGA [5] accelerators and that can leverage the **localmem** feature of OmpSs@FPGA in order to reduce memory pressure. The foundation of the **b8c** representation is built around the concept of what we name **super-rows**. Super-rows are data structures able to enclose matrix values coming from different rows and/or columns in the original matrix. The size (number of elements) of the super-row is chosen to maximize usage of memory bandwidth and to be able to process as many elements in parallel as possible, ideally the same number of elements as the memory is able to serve per cycle. In order to keep the information about the original indexes for each element, each super-row has a corresponding set of metadata associated to it. Since **b8c** is meant to be used in a *blocking* scenario, in which eash task processes all the values corresponding to a submatrix of $N\_rows \times M\_columns$, the current implementation of **b8c** stores the row and column information for each element *relative* to the first row/column of its block. Moreover, as the current implementation forces elements in a super-row to have consecutive column indexes, only the column information for the first element is kept. Figure 1 shows a original sparse matrix (a) represented in both CSR (b) and **b8c** (c). There is an additional constraint that applies to a super-row in the current version of **b8c**: elements in a super-row that do not belong to the same row in the original matrix must belong to rows whose row number modulo N_elems has no collision. N_elems, in this case, is also the number of rows that we can update in parallel for a given-super row. Figure 2 depicts the hardware block diagram for the **b8c** SpMV algorithm implementation.
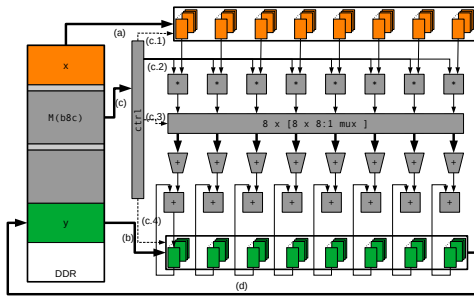
Fig. 2: b8c SpMV hardware diagram

(100 MHz FPGA versions, 1 instance / Ryzen 7 3900MHz CPU version, 1 core)
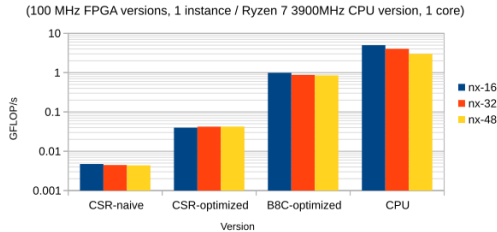
Fig. 3: SpMV FPGA/CPU Performance comparison

As stated previously, **b8c** is meant to be used in a blocking scenario and thus, for each task, the corresponding segments of y and x are automatically loaded into BRAMs using the **localmem** feature of OmpSs@FPGA (steps a and b) prior to processing the elements of the sub-matrix. At that point, the accelerator can start the SpMV computation. The **b8c**-encoded sub-matrix values and the corresponding metadata are kept interleaved in DRAM, and are streamed into the accelerator directly from memory. The use of a bus width of size $N\_elems \times sizeof(elem)$ and a pipelined design of the algorithm allows the processing of 1 super-row (N_elems matrix values) per cycle (except when an interleaved row of metadata is loaded). The metadata of each super-row is used in (c.1) to select the x values (matrix columns) that must be multiplied by the corresponding elements of the matrix that the accelerator is streaming from memory (c.2). After that, partial products are routed into an interconnect that allows the accelerator, using again the meta-data information (c.3) to group values from the products that belong to the same original row, in order to accumulate them using an accumulator tree. Finally, once the accumulation phase is finished, the results are used to update the corresponding positions of y, using the row indexes in the meta-data (c.4). Once the whole sub-matrix has been processed, if no further processing needs to be done in the same segment of y, it gets written back to memory. The fact that all the column indexes for a super-row are consecutive and that the row indexes in the same super-row do not have any collision allows us to partition the BRAMs used to store x and y and to allow, during the computation phase, parallel, 1-cycle access to N_elems values of x and to update, also in the same way, N_elems positions of y. The data structure resulting from this process (the **b8c** representation of the sparse matrix) allows us to split the SpMV computation among different accelerators developed using OmpSs@FPGA and HLS directives, that are able to process each sub-matrix using the process described above.

C. Preliminary Results

In order to evaluate the efectiveness of **b8c**, we have implemented both the transformation and the OmpSs@FPGA SpMV

implementation using it and we have performed experiments using matrices from the HPCG benchmark (with different sizes [$nrows = nx^3$]). Figure 3 shows how **b8c** offers $\approx$x25 performance improvement compared to an HLS-optimized FPGA implementation using CSR and, despite running only at 100MHz, reaches $\approx$25% of the performance achieved by a CPU version using CSR and running on a Ryzen 7 at 3.9GHz.

D. Conclusions

In this work we have introduced **b8c**, an sparse matrix representation and its corresponding SpMV FPGA implementation. Using this representation, our implementation can leverage most of the unique features of FPGAs, like the capacity to perform data-streaming processing, simultaneous 1-cycle memory accesses using on-chip SRAM and simultaneous processing of N_elems per cycle per instance in order to outperform previous HLS optimized FPGA versions and to offer performance on-par with CPU implementations using CSR. We plan to continue working on this line in order to further optimize the matrix representation and the SpMV algorithm using it and to test them against a wider data set using multiple accelerator instances and higher clock frequencies, and also using more advanced memory architectures, such as HBM, that offer higher bandwidth than DDR.

II. ACKNOWLEDGEMENTS

REFERENCES

[1] H. Bian *et al.*, "CSR2: A New Format for SIMD-accelerated SpMV," in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, May 2020, pp. 350–359.

[2] C. Gómez *et al.*, "Efficiently running SpMV on long vector architectures," in *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '21. New York, NY, USA: Association for Computing Machinery, Feb. 2021, pp. 292–303.

[3] S. Filippone *et al.*, "Sparse Matrix-Vector Multiplication on GPGPUs," *ACM Transactions on Mathematical Software*, vol. 43, no. 4, pp. 30:1–30:49, Jan. 2017.

[4] S. Kestur *et al.*, "Towards a universal FPGA matrix-vector multiplication architecture," *Proceedings of the 2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines, FCCM 2012*, pp. 9–16, 2012.

[5] J. M. Deharo *et al.*, "OmpSs@FPGA framework for high performance FPGA computing," *IEEE Transactions on Computers*, pp. 1–1, 2021.

José Oliver is a second year Ph.D student at the Barcelona Supercomputing Center. He received his MSc degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 1998 and his Advanced Studies Diploma from the same university in 2000. From 1998 to 2000 he worked at the European Center for Parallelism at Barcelona. He spent the following 20 years working in different positions in the industry, the last 15 of them as CTO in the Agency for Business Competitiveness. Since 2020 he is with the Programming Models research group of Barcelona Supercomputing Center (BSC) as well as a Ph.D student at the Computer Architecture Department of UPC.

Organic Particle Export, Remineralization and Advection in the North Atlantic mesopelagic layer

M. Andrea Orihuela-García[1], Martí Galí[2], Yohan Ruprich-Robert[3]

[1]Earth Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain, andrea.orihuela@bsc.es,

[2] Department of Marine Biology and Oceanography, Institut de Ciències del Mar, Barcelona, Spain, mgali@icm.csic.es,

[3]Earth Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain, yohan.ruprich@bsc.es

EXTENDED ABSTRACT

The mesopelagic layer of the oceans extends between ~200 and 1000 m depth and plays a fundamental role in global biogeochemical cycles and climate. In addition, it hosts a massive biomass of zooplankton and small fish, what is essential to regulate marine resources. However, scientific understanding and predictive capacity of biogeochemical processes in the mesopelagic zone are still underdeveloped. This lack of information has societal and economic costs because there is uncertainty in estimates of oceanic carbon storage (which inform policies for the reduction of carbon dioxide emissions), and it hampers the management of mesopelagic biological resources. In this way, this thesis would address the study of the carbon cycle and the associated biogeochemical processes. The main objective is to analyse the variability of the transport and transformation of particles that carry organic carbon from the surface to the deep sea in the North Atlantic. For this purpose, this study will focus on the analysis of simulations from dynamical (NEMO4 (Madec & NEMO team, 2008)) and biogeochemical (PISCES-v2 (Amount et al., 2015)) models, together with observations from bio-ARGO floats and satellite data.

Author biography

M. Andrea Orihuela-García was born in Cádiz, Spain, in 1995. She studied the Ocean Sciences degree and the Enronmental Sciences degree at the University of Cádiz, Spain, in 2019. After that she studied a master's degree on oceanography at the Universities of Cádiz and Las Palmas de Gran Canaria, both in Spain. She has been doing different stays abroad in Azores Islands (Portugal) and Valparaíso (Chile) and she has also done an internship in Algarve (Portugal) at the Physics' department of the *Centro de Investigação Marinha e Ambiental*. She has also worked as a scientist observer on board a fishing vessel, in collaboration with *the Instituto Español de Oceanografía*. Since September 2020, she started her Ph.D. at the Climate Prediction group, of the Earth Department in the Barcelona Supercomputing Center, supervised by Martí Galí and Yohan Ruprich-Robert.

References

O. Aumont, C. Ethé, A. Tagliabue, L. Bopp and M. Gehlen (2015). PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies. Geoscientific Model Development Discussions. doi:10.5194/gmdd-8-1375-2015

G. Madec and NEMO System Team (2008). "NEMO ocean engine", Scientific Notes of Climate Modelling Center (27) – ISSN 1288-1619, Institut Pierre-Simon Laplace (IPSL)

QUARQ: QUick Approximate and Relaxed Querying

Victor-Alejandro Ortiz*†, Maria-Cristina Marinescu*, Maria-Ribera Sancho*†

*Barcelona Supercomputing Center (BSC), Barcelona, Spain

†Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

E-mail: {victor.ortiz, maria.marinescu, maria.ribera}@bsc.es

Keywords—Ontologies, Semantic Web, Linked Data, query performance, SPARQL, Online Learning.

I. Extended Abstract

Executing queries over Linked Open Data (LOD) is a complex task. The total number of sources triggered by a single query cannot be known in advance, nor the reasoning complexity applied to each source. In order to avoid this uncertainty, practitioners download full replicas of the open data and build applications on top of the datasets in a controlled environment. With this centralized approach, they lose dynamic data changes, and often they cannot account for the inference capabilities defined in the associated ontologies.

In this work, we explore the feasibility of predicting the performance of Flexible Querying over Linked Open Data [1]. Concretely, we propose QUARQ: QUick Approximate and Relaxed Querying, a tool that using ML provides intelligence to the process of generating alternative queries that run more efficiently than the original ones. With this tool, we propose avoiding the use of replicated Linked Data by seizing the shareable nature of Linked Data and eluding the impracticality of maintaining copies up-to-date or the need to work with outdated data.

A. Overview of the solution

Our solution designs and implements a tool aimed at the Semantic Web practitioners, whose purpose is to query LOD directly through the available public SPARQL endpoints. QUARQ uses ML under an Online Learning (OL) paradigm to predict the performance of a query and, based on parameters provided by the user, generate alternate queries that run more efficiently at the cost of accuracy.

We take the LOD-Cloud diagram as a reference for our project, the most adopted graph of links between multiple datasets. The LOD-cloud provide us with datasets statistics on their published resources. Moreover, a larger part of the Semantic Web community uses this graph to query LOD resources and make available their datasets to the community following Linked Data principles. The diagram currently contains 1301 datasets with 16283 links [1] although not all of them are continuously available [2].

We propose that selecting the flexible querying method with the best performance at a given time to transform an original user query is a task that ML techniques can learn by using OL to predict the performance of alternative queries
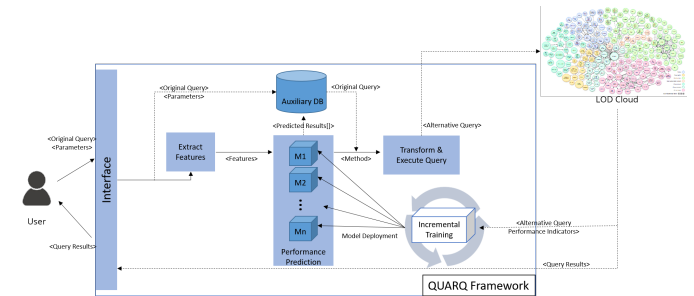

Fig. 1. QUARQ architechture

generated by different Approximate and Relaxed Querying algorithms [3].

QUARQ selects the best possible method to generate an alternate query and executes it directly in public endpoints provided by the datasets published as part of the LOD-cloud.

a) Query Approximation: An Approximate Query (AQ) simplifies the execution process of the query by using data sampling or performing a data synopsis to shorten the response time at the cost of accuracy; it does not rewrite the query [4].

b) Query Relaxation: A Relaxed Query (RQ) rewrites the query to use alternative concepts when the original one cannot return data, results in a query that does not terminate, or can only do so in an amount of time that is not reasonable, ranking the results of a query depending on how closely the original predicate of the query is satisfied [4].

B. Architecture of the solution

Our goal is to build a tool that provides the final user with easy and scalable access to a broad amount of interconnected data, regardless of the user's prior knowledge of data sources, schemas, or technical skills necessary to formulate efficient queries. QUARQ acts as an intermediary between the user and LOD-cloud - through public endpoints, guaranteeing the user will get an approximate valid result and avoiding empty responses.

Figure 1 illustrates QUARQ's architecture, which comprises the following components:

1) Interface: The user interface allows users to input the query they want to execute over the LOD cloud. Moreover, it also provides a parameters settings functionality, where the user inputs levels of runtime, precision and size of the retrieved data from the query. Finally, a query results visualization.

Input and visualization of results are traditional functions on a query interface. Nonetheless, the critical component in

---

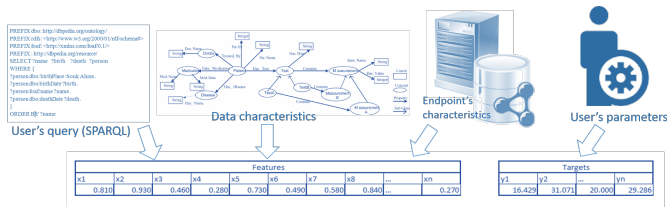[1]https://lod-cloud.net/ consulted on March 28th, 2022

Fig. 2. Feature extraction representation

this section is the parametrization, which is indicative of the user's expectations and used as a target for ML models to predict the queries performance under different alternate query scenarios. The user defines three parameters simultaneously and prioritizes the importance of each one. We can assign a weighted value to each parameter and define an individual assessment with a combined result on the multiple parameter prediction.

2) Feature extraction: Our tool makes predictions based on features derived from an initial input provided by the user - query, endpoint URL and parameters; however, the feature selection process should be thoroughly executed to avoid saturating our model with irrelevant features.

As illustrated in Figure 2, our solution, takes the input provided by the user and extracts different features derived from:

- Original query provided by the user

- Referenced datasets

- Characteristics of endpoints

- Parameters

Based on a previous work presented by [5], we hypothesize that the relevant features are contained in these main sources. However, the objective of this section is to identify and extract from the query only those characteristics that add value to our model, disregarding those irrelevant to the performance's prediction to avoid poor quality or unreliable predictions.

3) Performance prediction: Our solution applies online learning [6], an ML paradigm that does not require all the data upfront. This approach is concerned with improving the models using new incoming data. OL-based models remember previous data as a dynamic training vector without reprocessing it.

The performance prediction component receives the set of features extracted from the original query and runs a regression analysis to predict the expected parameters' values for each of the flexible querying methods implemented in QUARQ.

4) Query Transformation and execution: The query transformation component is in charge of rewriting the original query -given by the user - by applying the flexible querying algorithm determined that is most likely to obtain the targeted performance. This querying method returns either approximate results by applying a Query Approximation (QA) method or relaxes the conditions of a query to avoid empty responses using a Query Relaxation method (QR).

5) Incremental Training: In principle, Incremental learning models - also known as Online learning or data stream learning [7] - allow for both data prediction and retraining requests. The incremental training component is the most critical section of the tool and gives the most added value. Its contribution to our solution is twofold: it allows our predictor to learn in an *empirical* way while granting a level of adaptability to a continuously changing environment.

Every time a user formulates and executes a query, our tool returns results and values for metrics on runtime, accuracy and number of responses. These metrics are labelled as targets to retrain the models that predict the performance for each of the implemented Flexible Querying methods. To avoid delays in the execution of prediction tasks, we propose to perform the retraining process separately on an offline copy. The old model is substituted by a new retrained version of the model, and another new copy stays offline to start the retraining process again.

C. Conclusion

Our tool allows Linked Data practitioners to query the so-called "web of data" in a more efficient way by means of sacrificing some precision but under parameters defined by the user themselves. This enables any user to avoid blank results and always rely on having an answer that approaches their needs regarding runtime, completion or precision.

REFERENCES

[1] R. Hasan, "Predicting SPARQL Query Performance and Explaining Linked Data," pp. 795–805, 2014.

[2] J. A. Nasir and J. P. McCrae, "iLOD: Interplanetary file system based linked open data cloud," *CEUR Workshop Proceedings*, vol. 2821, no. 825182, pp. 27–32, 2020.

[3] R. Frosini *et al.*, "Flexible query processing for SPARQL," *Semantic Web*, vol. 8, no. 4, pp. 533–563, 2017.

[4] S. Chaudhuri *et al.*, "Approximate Query Processing: No silver bullet," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, vol. Part F1277, pp. 511–519, 2017.

[5] R. Hasan and F. Gandon, "A machine learning approach to SPARQL query performance prediction," *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, vol. 1, pp. 266–273, 2014.

[6] J. Liu and E. Zio, "An adaptive online learning approach for Support Vector Regression: Online-SVR-FID," *Mechanical Systems and Signal Processing*, vol. 76-77, pp. 796–809, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2016.02.056

[7] H. M. Gomes *et al.*, "Machine Learning for Streaming Data: State of the Art, Challenges, and Opportunities," *SIGKDD Explor. Newsl.*, vol. 21, no. 2, pp. 6–22, 2019. [Online]. Available: https://doi.org/10.1145/3373464.3373470

Victor-Alejandro Ortiz received his BSc. degree in Informatics from the National Autonomous University of Mexico (UNAM), and the MSc. degree in Big Data and HPC from the University of Liverpool, UK. Currently pursuing a Ph.D. in Computer Science at the Universitat Politècnica de Catalunya (UPC) and the Barcelona Supercomputing Center (BSC), Spain.

HBM, Present and Future of HPC based on FPGAs

Elias A. Perdomo Hourné[*§1], Teresa Cervero[*2], Xavier Martorell[*§3]

[*]*Barcelona Supercomputing Center (BSC), MEEP, Til·lers c/Passeig dels Til·lers, 15. 08034 Barcelona (Spain)*

[§]*Universitat Politécnica de Catalunya (UPC), Computer Architecture Dept., C6, c/Jordi Girona, 1-3. 08034 Barcelona (Spain)*

E-mails: [1]elias.perdomo@bsc.es, [2]teresa.cervero@bsc.es, [3]xavier.martorell@bsc.es

Keywords— **FPGAs, HBM, Parallelism, Performance Optimization and Embedded systems.**

EXTENDED ABSTRACT

In the past decades, advances in the speed of commodity CPUs have far out-paced advances in memory latency. Main-memory access is therefore increasingly a performance bottleneck for many computer applications, including HPC embedded systems. This translates into unprecedented memory performance requirements in critical systems that the commonly used DRAM memories struggle to provide.

High-Bandwidth Memory (HBM) can satisfy these requirements offering high bandwidth, low power and high-integration capacity features. However, it remains unclear whether the predictability and isolation properties of HBM are compatible with the requirements of critical embedded systems. In our research, a deep characterization of the HBM for its use in MEEP applications is performed.

A. Introduction

Custom hardware – from workstations to PCs– has experienced tremendous improvements in the past decades. However, while the speed of commercial microprocessors has increased approximately 70% every year, the speed of commodity DRAM has improved by only around 50% over the whole past decade. As a result, computer systems are experiencing difficulties in achieving high processing efficiency [1].

The traditional approach for boosting performance in systems, particularly those at the edge where huge amounts of data need to be processed locally or regionally; consists on adding more computational capability into a chip and bringing more memory on-chip. But that approach no longer scales since we are now getting to the boundaries of the trifecta of von Neumann architectures, Moore's Law and Dennard scaling. Consequently, engineers have begun focusing on solving the bottleneck between processors and memories by turning out new architectural designs at a rate no one would have anticipated before.

An alternative approach, based on recent technological advances, is moving processing elements closer to, or even into the memory. This solution looks for avoiding the penalty for replicating processing elements, which provides an acceptable tradeoff. When utilizing wide short buses (HBM being the most common example), designers avoid the penalty of going outside the die for accessing to memory and recover some of the performance tradeoff.

This trend is followed by Xilinx, one of the two market giants in the area of FPGAs and the leader in adaptive computing. Xilinx is firmly committed to a transition to HBM memory as a solution to memory bottlenecks, as demonstrated in recent years. In October 2018, Xilinx launched the Alveo U200 with no HBM memory. Only 1-month after, in November 2018 the new Alveo U280 already included 8GB of HBM2 and halved DDR capacity. Their last Alveo Data center card release, the Alveo U55C, doubled HBM capacity and rescinded the DDR memory banks use [2].

B. HBM Features and Standards

HBM stands for High Bandwidth Memory, a type of memory interface used in 3D-stacked DRAM (Dynamic Random Access Memory) in GPUs, servers, high-performance computing and networking and client space.

HBM technology works by vertically stacking memory chips one on top of another in order to shorten how far data has to move, while allowing for smaller form factors. Additionally, with two 128-bit channels per die, HBM memory bus is wider than the one present in other types of DRAM memory. Stacked memory chips are connected through silicon vias (TSVs) and micro-bumps and connected via the interposer, rather than on-chip, to ensure faster accesses [3]. Several versions of the HBM and future standard updates are shown in the Table 1.

TABLE 1
HBM STANDARDS EVOLUTION

HBM Specification	HBM	HBM2 / HBM2E (Current)	HBM3 (Upcoming)
Max Pin Transfer Rate	1 Gb/s	3.2 Gb/s	?
Max Capacity	4 GB	24 GB	64GB
Max Bandwidth	128 GB/s	410 GB/s	512 GB/s

The current HBM2 was adopted as a JEDEC standard in December 2018 and the specifications were updated again in early 2020. It allows a bandwidth of 3.2 GB/s per pin with a max capacity of 24GB per stack (2GB per die across 12 dies per stack) and max bandwidth of 410 GB/s, delivered across a 1,024-bit memory interface separated by 8 unique channels on each stack [4].

It is therefore not surprising that HBM can overcome all the challenges of DRAM to achieve high bandwidth. Even more impressive is that it does so, while its low speed/pin consumption also improves power efficiency, as can be seen in Fig 1.
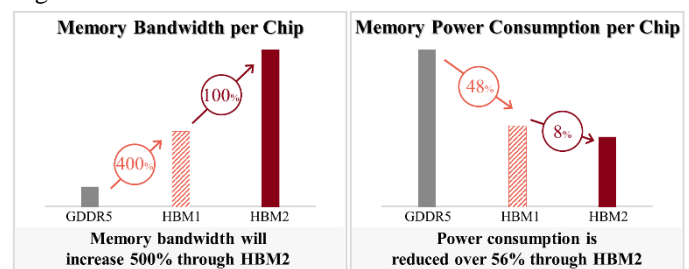

Fig. 1 Bandwidth and power consumption comparison for different memories.

C. HBM in the Alveo U280

Having these precepts for the HBM, we decided to perform a first analysis of HBM from a real-time systems perspective, by exploring how it is built in the Alveo U280 (with a common topology as the Alveo U55C). We focus on the HBM device in order to capture its architectural features and functionalities, which can affect timing predictability such as device access behavior, timing properties and performance metrics.

The HBM controller in the Alveo U280 has 2 stacks of 4GB each, with 8 channels and 16 pseudo-channels and 16 AXI Ports (Fig 2). These ports can access its own memory mapped

or any other memory if the global addressing is activated but having a Lateral AXI Switch Access Throughput Loss in each micro-switch "domain change"[5].
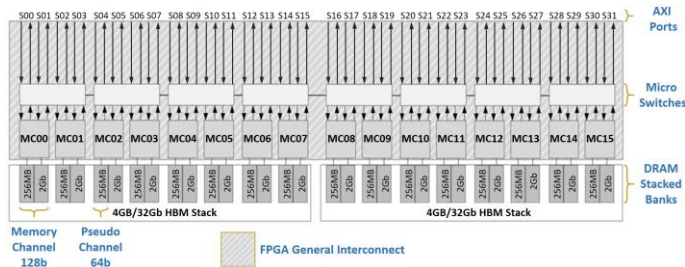

Fig. 2 Alveo U280 HBM topology.

D. AIT MatMul implementations accessing the Alveo HBM

In MEEP we are enhancing the Accelerator Integration Tool (AIT), a tool in development that automatizes the bitstream generation process custom designs based on several predefined design templates using the Xilinx toolchain. Based on a previous matrix multiplication (*MatMul*) implementation we replaced DDR by HBM and measured performance. To accomplish that we used a *MatMul* implementation which has 4 *matmul_block* accelerators which previously accessed the 2 DDR banks in parallel and now can access 4 HBM *pseudo-channels* in parallel (Fig 3).
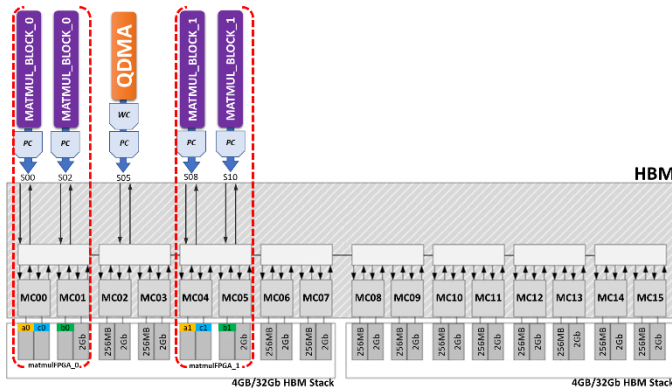

Fig. 3 *MatMul* implementation with 4 *matmul_blocks* and HBM.

Table 2 shows that HBM, even with a port data width half of the DDR one, maintains a comparable performance with the DDR using the double of HBM pseudo-channels than DDR banks. Even whether *MatMul* is a great proof of concept and HBM can use up to 16 times the number of banks; it has its limitations because of the fact of being a CPU bounded application. Thus, we would need a memory bounded application or to stress memory accesses more thoroughly.

TABLE 2
DDR PERFORMANCE METRICS VS HBM COMPARABLE ONES

Implementation	Performance (*GFlops*)
matmul_ait_u280_DDR_PW_512	287.043793
matmul_ait_u280_HBM_PW_256	255.816208

E. Custom Traffic Generators accessing the Alveo HBM

With the purpose of fully characterize the HBM, a Traffic Generator set of test-bench was created. Due to hardware implementations are the only way to stress the HBM memory efficiently, we designed our own Traffic Generators to control memory access pattern generation and to simplifying the debugging task. It is highly recommended to have a common set of tests for the performance, bandwidth and latency comparison between DDR and HBM to start confirming HBM

capabilities. The following table shows the results of a comparison between DDR and HBM with similar implementations (Row-Bank-Column address mapping scheme, Single row accesses and No burst).

TABLE 3
DDR PERFORMANCE METRICS VS HBM COMPARABLE ONES

	Implementations		
	DDR_TG 1 DDR	**DDR_TG 2 DDRs**	**HBM_TG 32 pseudo-channels**
Write Latency	5 clock cycles	5 clock cycles	14 clock cycles
Read Latency	24 clock cycles	24 clock cycles	48 clock cycles
Read Throughput	0.229 GB/s	0.457 GB/s	50 GB/s

F. Conclusion and Future Enhancement

Based on a similar set of test benches HBM proved to have a worse latency than the DDR. Nevertheless, as expected, the throughput performance was more than 25 times better when using all the 32 pseudo-channels in the HBM in parallel than when using the 2 memory banks present in the DDR in the Alveo U280. In order to fully characterize the HBM module, more experiments are being carried out to measure the Lateral AXI Switch Access Throughput Loss, the congestion in a channel or in its memory controller and the impact of the IDs and the reordering option.

References

[1] T. C. Mowry, "Tolerating latency through software-controlled data prefetching," PhD Thesis, Stanford University, 1994.

[2] Xilinx Inc., "ALVEO™ Product Selection Guide Datasheet by Xilinx Inc. XMP451 (v1.7)," *xilinx.com*, 2021. https://www.xilinx.com/support/documentation/selection-guides/alveo-product-selection-guide.pdf (accessed Mar. 15, 2022).

[3] Z. Wang, H. Huang, J. Zhang, and G. Alonso, "Benchmarking high bandwidth memory on FPGAs," *arXiv preprint arXiv:2005.04324*, 2020.

[4] H. Jun *et al.*, "HBM (high bandwidth memory) dram technology and architecture," in *2017 IEEE International Memory Workshop (IMW)*, 2017, pp. 1–4.

[5] Xilinx Inc., "AXI High Bandwidth Memory Controller v1.0 LogiCORE IP Product Guide PG276 (v1.0)," *xilinx.com*, Aug. 06, 2021. https://www.xilinx.com/support/documentation/ip_documentation/hbm/v1_0/pg276-axi-hbm.pdf (accessed Mar. 15, 2022).

Author biography

Elias A. Perdomo Hourné was born in Havana, Cuba, in 1988. He received the B.Eng. degree in Automation Engineering in 2012, and the M.Sc. degree in Digital Systems in 2018 both from the Technological University of Havana, Havana, Cuba. In addition, he received the M.Sc. Advanced Microelectronic Systems Engineering from the University of Bristol, Bristol, United Kingdom in 2019.

Since October 2020, he has been working with the MEEP group of the Barcelona Supercomputing Center (BSC) as Research Engineer as well as a PhD student at the Compuer Architecture Department of the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, under the supervision of Teresa Cervero and Xavier Martorell. His current research interests include Embedded Systems, RTL and RISCV SoC design, FPGAs, Heterogeneous Computing, automatic design generation and memory management for HPC and programming models.

Development of corrections for the absolute free binding energy prediction

Ignasi Puch-Giner*†, Marti Municoy†‡, Victor Guallar*‡

*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
‡Nostrum Biodiscovery, Barcelona, Spain
E-mail: ignasi.puchginer@bsc.es, mmunicoy@nostrumbiodiscovery, victor.guallar@bsc.es

Keywords - Binding energy, forcefield, SBDD, Monte Carlo

I. EXTENDED ABSTRACT

The early stages of drug design rely on hit discovery programs, where initial possible inhibitors' binding affinities are assessed when bound to their biological target. It is an expensive and time-consuming process, requiring multiple iterations of trial and error designs. This sets the perfect ground for computer simulations.

Structure-based drug design has been in the past decade a widely used computational methodology to speed up the drug discovery process for resolved protein-ligand systems[1]. However, providing a fast and reliable answer to the protein-ligand affinity problem can be an arduous task. In this context, the capacity of the software to score the binding affinity of the inhibitors will be crucial to determine possible drug leads that will be later on optimized.

Hence, the main goal of this research is to add physically justified corrections as well as Machine Learning models to the energetic predictions to obtain absolute binding free energies that match the experimental results. To do it we will need to review the physics involved in the forcefields used in the simulations done with the software used in the group: PELE[2].

PELE stands for Protein Energy Landscape Exploration and it is a self-contained Monte Carlo software to model protein-ligand interactions. The reachable conformations by the protein and ligand are explored and energetically assessed with the forcefield. The forcefield is the parameterized functional (eq. 1) that enables a Monte Carlo or a Molecular dynamics simulation to calculate the potential energies involved[3].

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{nonbonded}}$$
$$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}} \quad (1)$$
$$E_{\text{nonbonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}.$$

This functional form does not take into account different energetic contributions that should be addressed. Right now we have considered adding correction terms regarding the strain and the conformational entropy loss of the ligand upon binding, as in eq. 2.

$$\Delta G = \Delta G_{\text{be}} + \Delta H_{\text{strain}} - T\Delta S_{\text{conf}} \quad (2)$$

A. Results

1) PELE's binding energy: First of all, we have addressed the problem of calculating the binding energy of a ligand. This affinity represents the amount of energy needed to separate the ligand from the protein.

In a PELE simulation, thousands of positions are reached each of which has a binding energy assigned with eq. 3.

$$\Delta G_{\text{PELE}} = G_{\text{protein-ligand}} - (G_{\text{protein}} + G_{\text{ligand}}). \quad (3)$$

The binding energy has been calculated for eight different protein-ligand systems with eq. 4

$$\Delta G_{\text{be}} = \frac{\sum_{i=1}^{N} \Delta G_{\text{PELE}}^i e^{-\beta G_{\text{tot}}^i}}{\sum_{i=1}^{N} e^{-\beta G_{\text{tot}}^i}} \quad (4)$$

where $\beta = RT$ which has units (kcal/mol)$^{-1}$ and G_{tot}^i is equal to the total energy of the i's conformation in kcal/mol making the exponent dimensionless.
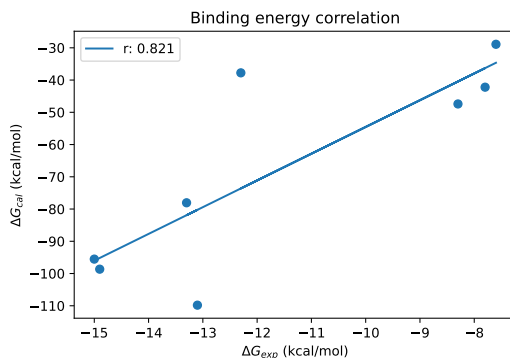

Fig. 1. Correlation between calculated and experimental binding energies. Calculated results are computed with the Boltzmann average of the binding energy of eight different PELE simulations with OPLS. Each simulation corresponds to a different protein-ligand system. (PDB codes: 1CB0[4], 1K27[5], 6QGE[6], 6QGF[6], 1HPV[7], 1HSG[8], 1MSM[9] and 1T3R)[10].

With these results, we have been able to check how well the calculated affinities correlate with the experimental data as shown in Fig. 1.

2) Ligand strain: In the second place, we have considered the strain a ligand undergoes upon its binding in the target's binding site. This energetic term is related to the change in the ligand conformation when solvated in water and its conformation when bound to the protein. In our case, we take it into account by performing two simulations: one with the isolated ligand and the other with the protein-ligand system. From the first simulation, we can obtain the minimum energy conformation of the ligand which will be its solvent conformation energy (H_{sol}). From the second we obtain the ligand energies associated with the conformations adopted inside the binding site of the protein, called bioactive conformations (H_{bc}). With eq. 5 we can associate a ligand strain to a simulation and see how well the results correlate to QM/MM calculated strains, given that ligand strains cannot be measured experimentally.

$$\Delta H_{\mathrm{strain}} = \frac{\sum_{i=1}^{N} H_{\mathrm{bc}}^{i} - H_{\mathrm{sol}}}{N} > 0 \qquad (5)$$

3) Ligand conformational entropy: The last correction implemented for now has been the entropic conformational loss of the ligand upon binding. Since the conformational space available is reduced, there is a toll on the free energy.

To obtain estimations of the entropic loss we have considered the dihedral angles containing the N rotatable bonds of the ligand. With this, we have been able to track the angles reached by the different dihedral angles. Then, we can perform a binning to see the occurrences of each different interval of angles and with that, assign a probability p to each of the m bins. We can associate an entropic term to each rotatable bond (rb) and, consequently, an entropic toll to the protein-ligand simulation and the ligand-in-solvent simulation as shown in eq. 6

$$S = -R \sum_{i=1}^{N} \sum_{j=1}^{m_i} p_j \ln(p_j) \qquad (6)$$
$$-T\Delta S_{\mathrm{conf}} = -T(S_{\mathrm{in}} - S_{\mathrm{s}})$$

where S_{in} and S_{s} are the ligand conformational entropy inside the protein and in the solvent respectively.

4) Further work: First and foremost we need good experimental datasets to be able to validate how good the predictions are. For now, the small amount (and lack) of data impedes good statistics and with that good or definitive results.

Further corrections could involve, for example, taking into account the entropic loss associated with the change in the explorable conformational space of the residues located in the protein's binding site. Another approach could be developing Machine Learning models to fit the calculated results to the experimental data. In the end, we could end up with mixed physically and Machine Learning based corrections to have a good binding free energy estimation.

REFERENCES

[1] Śledź, P., & Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. Current opinion in structural biology, 48, 93-102.

[2] Borrelli, K. W., Vitalis, A., Alcantara, R., & Guallar, V. (2005). PELE: protein energy landscape exploration. A novel Monte Carlo based technique. Journal of chemical theory and computation, 1(6), 1304-1311.

[3] Frenkel, D., Smit, B., & Ratner, M. A. (1996). Understanding molecular simulation: from algorithms to applications (Vol. 2). San Diego: Academic press.

[4] Appleby, T. C., Erion, M. D., & Ealick, S. E. (1999). The structure of human 5-deoxy-5-methylthioadenosine phosphorylase at 1.7 Å resolution provides insights into substrate binding and catalysis. Structure, 7(6), 629-641.

[5] Singh, V., Shi, W., Evans, G. B., Tyler, P. C., Furneaux, R. H., Almo, S. C., & Schramm, V. L. (2004). Picomolar transition state analogue inhibitors of human 5 -methylthioadenosine phosphorylase and X-ray structure with MT-Immucillin-A. Biochemistry, 43(1), 9-18.

[6] Verteramo, M. L., Stenström, O., Ignjatovic, M. M., Caldararu, O., Olsson, M. A., Manzoni, F., ... & Akke, M. (2019). Interplay between conformational entropy and solvation entropy in protein–ligand binding. Journal of the American Chemical Society, 141(5), 2012-2026.

[7] Kim, E. E., Baker, C. T., Dwyer, M. D., Murcko, M. A., Rao, B. G., Tung, R. D., & Navia, M. A. (1995). Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. Journal of the American Chemical Society, 117(3), 1181-1182.

[8] Chen, Z., Li, Y., Chen, E., Hall, D. L., Darke, P. L., Culberson, C., ... & Kuo, L. C. (1994). Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. Journal of Biological Chemistry, 269(42), 26344-26348.

[9] Vega, S., Kang, L. W., Velazquez-Campoy, A., Kiso, Y., Amzel, L. M., & Freire, E. (2004). A structural and thermodynamic escape mechanism from a drug resistant mutation of the HIV-1 protease. Proteins: Structure, Function, and Bioinformatics, 55(3), 594-602.

[10] Surleraux, D. L., Tahri, A., Verschueren, W. G., Pille, G. M., De Kock, H. A., Jonckers, T. H., ... & Wigerinck, P. B. (2005). Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. Journal of medicinal chemistry, 48(6), 1813-1822.

Ignasi Puch-Giner received his BSc degree in Physics at the Universitat de Barcelona (UB). He holds a MSc on Computational Modeling for Physics, Chemistry and Biochemistry from UB and Universitat Politècnica de Catalunya (UPC). Since 2021 he has been in the Electronic and Atomic Protein Modeling group led by Victor Guallar in the Life Sciences department at the Barcelona Supercomputing Center. He is currently enrolled at the UPC's Computational and Applied Physics PhD programme.

Experimental Analysis on the NXP's T2080 Cache Coherence: A Step Towards MPSoCs in Critical Systems

Roger Pujol*†, Mohamed Hassan§, Francisco J. Cazorla†‡,
*Universitat Politècnica de Catalunya, Barcelona, Spain
†Barcelona Supercomputing Center, Barcelona, Spain
‡Maspatechnologies S.L., Barcelona, Spain
§McMaster University, Hamilton, Canada
E-mail: {roger.pujol, francisco.cazorla}@bsc.es, mohamed.hassan@mcmaster.ca

Keywords—Cache Coherence, Critical Systems, MPSoC, Certification.

I. EXTENDED ABSTRACT

The adoption of complex MPSoCs in critical real-time embedded systems [1], [2] mandates a detailed analysis of their architecture to facilitate certification [3]. This analysis is hindered by the lack of a thorough understanding of the MPSoC system due to the unobvious and/or insufficiently documented behavior of some key hardware features [4], [5].

Confidence in those features can only be regained by building specific tests to both, assess whether their behavior matches specifications and unveil their behavior when it is not fully known a priori.

In this line, in this work we develop a thorough understanding of the cache coherence protocol in the avionics-relevant [6] NXP T2080 [1] architecture.

II. SETUP

The NXP T2080 SoC [1] (see Figure 1), assessed for its use in avionics, features four PowerPC e6500 cores [7], each with its private instruction and data cache (IL1 and DL1, respectively) as well as a private Memory Management Unit (MMU). Each core communicates with the shared L2 cache via the cache-core interface (CCI). The L2 cache is shared between all the cores. A "CoreNet" coherence fabric (CCF) provides access to the DDR SDRAM memory controller as well as other interfaces present in the board like the Direct Memory Access (DMA). The L2 has a single port to the CCF whose access is controlled by the Bus Interface Unit (BIU). The L2 cache is the point of coherency in the cluster. DL1 caches contain no modified data as they are write-through. The L2 is inclusive of the DL1 of each core, so if a data line is evicted from the L2 cache, it is invalidated in the corresponding DL1 caches. The T2080 implements a 4-state Modified-Exclusive-Shared-Invalid (MESI) cache coherence protocol. The coherence granularity is a L2 cache line, such that each line has its own coherence state information.

DMA transfers can 'generate' snoop requests to the L2. This can lead to invalidations of data in the DL1 caches and/or the L2 since this data becomes not up to date with respect to the latest value written by the DMA to memory.

A. Observability

Hardware Monitors. Several hardware event monitors in the T2080 [7] provide information about the L2 cache coherence activity, which we show in Table I.

Debugger Support. Using CodeWarrior, the standard IDE for the T2080, we access several flags for each cache line with information about their coherence state: Dirty (i.e. Modified), Valid, Shared, and Exclusive.
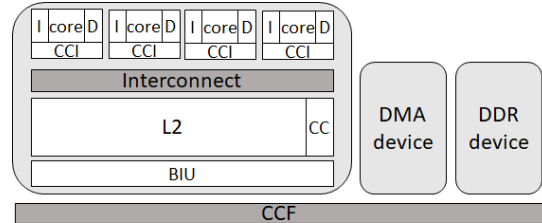

Fig. 1. Simplified block diagram of the T2080

TABLE I. COHERENT RELATED EVENT COUNTERS IN FOR THE L2

Counter	Description
L2DA	Number of L2 data accesses.
L2DM	Number of L2 data misses.
L2SH	Number of L2 cache snoop hits.
L2SP	Number of L2 cache snoop pushes.
ESR	Externally generated snoop requests.
L2SM	L2 snoops causing MINT (Modified INTervention).
L2SS	L2 snoops causing SINT (Shared INTervention).
L2RC	Number of L2 reloads from CoreNet.
BL	BLINK requests from L2 to core (e.g. back invalidates)
CL	CLINK requests from L2 to core(CoreNet data forwarding)

B. Experimental Setup

We seek to observe the behavior of the L2 cache when transitioning from one coherence state to another. To that end, we execute a benchmark during the warm-up phase that sets several lines in the L2 cache to a given coherence state. Afterwards, during the execution phase, another benchmark is executed to force the coherence state of those cache lines to change. Such benchmark accesses a subset of the addresses accessed in the warm-up phase. Moreover, the benchmark in the execution phase can access other addresses not shared with the benchmark executed in the warm-up phase. This is explained on a per experiment basis.

We set a breakpoint right after the warm-up phase (right before the execution phase) to confirm that the state is the one we expect. We also set another breakpoint right after the execution phase to verify that the cache lines have changed to the desired state.

Benchmarks. We use two micro-benchmark types, CPU and DMA, each with read (r) and write (w) variants, resulting in 4 combinations (CPUr, CPUw, DMAr, DMAw). The CPU benchmarks perform 128,000 accesses mapped to a set of either 4,096 or 8,192 different addresses. We refer to these benchmarks as CPUx(4K) and CPUx(8K) respectively, where 'x' is either 'r' or 'w' and the number of accesses is omitted as it is the same in all variants.

The DMAr benchmark reads data from a memory address range including the one used by the CPU benchmarks and writes it to a not overlapping memory address range, while the DMAw benchmark reads from the non-overlapping range and writes to the overlapping one. Both benchmarks, transfer either 2MB or 4MB of data. Hence, they perform 32,768 or

65,536 64B accesses to the overlapping range and the same number to the non-overlapping. We refer to these benchmarks as DMAx(32K) and DMAx(64K) respectively. Note that DMAr and DMAw perform necessarily <u>both</u> read and write operations.

Workloads. A workload comprises the execution of one or several benchmarks during the warm-up phase and one during the execution phase. For instance, $[CPUw(4K), DMAr(32K); DMAw(32K)]$ shows that during the warm-up CPUw(4K) and DMA4r(32K) are executed one after the other; and in the execution phase the DMAw(32K) is executed.

III. ANALYSIS

We analyzed all the Coherence Protocol transitions; here, we only show the two transitions with the most interesting details. The full analysis of all transitions can be found in [8].

TABLE II.   PMCs FOR "M TO S" AND "S TO M" TRANSITIONS.

State		L2DA	L2SH	L2SP	ESR	L2SM	L2SS	L2RC	L2DM	BL	CL
M2S	A	13	3968	3968	65536	3968	0	6432	0	0	6447
	B	12	8064	8064	65536	8064	0	6418	0	0	6418
	C	13	3968	3968	32768	3968	0	3226	0	0	3241
	D	13	8064	8064	32768	8064	0	3226	0	0	3242
S2M	A	127998	0	0	0	0	0	3968	0	3968	0
	B	127998	0	0	0	0	0	8064	0	8064	0

Modified to Shared. During the warm-up phase, we run CPUw that causes the data to be dirty in L2. In the execution phase, we run DMAr to force snoops from the CoreNet addressed to the L2 that has to (i) send the data to the main memory, (ii) clean the dirty flag and (iii) move the cache line to the shared state. We evaluate four scenarios: (A) [CPUw(4K); DMAr(64K)], (B) [CPUw(8K); DMAr(64K)], (C) [CPUw(4K); DMAr(32K)], and (D) [CPUw(8K); DMAr(32K)].

As we start measuring right after the warm-up (i.e. the execution of CPUw), there can be few accesses that are pending to L2, causing the few L2DA, as shown in (M2S) case in Table II.A GetS/GetM message is sent for every address read or written by the DMAr [1]. Hence, the number of expected ESR is 65,536 in (A) and (B) and 32,768 in (C) and (D), matching the observed results in Table II (M2S).The GetS messages hit in the L2 for the memory locations that were accessed previously by the CPUw, which results in 4K[2] L2SH in (A) and (C) and 8K[3] in (B) and (D).

Unexpectedly, the L2 responds to each of these snoop hits in the L2 by requesting other coherent devices to set such lines as shared, but there is no other coherent device in the T2080. This matches L2SP and L2SM values.

Shared to Modified. During the warm-up, we run CPUw that causes the data to be dirty in L2, and then we run DMAr, which sends a request for the data and the L2 sends the dirty data to the CoreNet moving to Shared state. In the execution phase, we run CPUw, which sends GetM requests to the CoreNet to invalidate all the valid copies in any potential device with that data set as shared, despite there is none. As a simultaneous modification of this data could have occurred in another device with a shared copy of the data, the L2 sends a data forward request to CoreNet. This request might bring either no new data or an updated copy if a remote modification occurred between the local modification and its notification

---

[1]Recall that DMAr also performs writes that result in GetM requests.

[2]In reality, we expect $4,096 - 128$ and $8,192 - 128$ respectively, as we perform 31 and 63 iterations respectively of 128 accesses each.

to other coherent devices. We evaluate the following four scenarios: (A) [CPUw(4K),DMAr(64K); CPUw(4K)], and (B) [CPUw(8K),DMAr(64K); CPUw(4K)].

The task analyzed (CPUw) generates accesses from the CPU as captured by the L2DA counter in Table II (S2M). Each of these accesses triggers the L2 to send a GetM message to ask all other sharers to invalidate their own copy of the shared data to the CoreNet, generating around 4K BL in (A) and around 8K in (B). As explained, since other coherent devices could be performing simultaneous modifications of the shared data, the L2 performs 4K L2RC in (A) and around 8K in (B). Those L2RC receive no answer since there is no other coherent device in the platform.

IV. CONCLUSIONS

Our empirical analysis of cache coherence in the T2080 brings some lessons learned. First, we can identify the events triggered by each coherence state transition, providing a clearer understanding of the implemented cache coherence behavior. Second, there are some hardware monitors with ambiguous or incomplete descriptions of the events tracked. And third, we detect unexpected coherence messages for a single L2 coherent cache processor. All these elements help validate the cache coherence protocol itself and allow building other further validation evidence on top of it.

V. ACKNOWLEDGMENT

REFERENCES

[1] Freescale Semiconductor, "QorIQ T2080 Reference Manual," 2016, Also supports T2081. Doc. No.: T2080RM. Rev. 3, 11/2016.

[2] Xilinx, "Zynq UltraScale+ Device Technical Reference Manual," https://www.xilinx.com/support/documentation/user_guides/ug1085-zynq-ultrascale-trm.pdf, 2019, UG1085 (v2.1).

[3] G. Fernandez et al, "Contention in multicore hardware shared resources: Understanding of the state of the art," in *WCET Workshop*, 2014.

[4] J. Barrera et al., "On the reliability of hardware event monitors in mpsocs for critical domains," in *ACM SAC*, 2020.

[5] N. Sensfelder et al., "On how to identify cache coherence: Case of the NXP qoriq T4240," in *ECRTS*, 2020.

[6] D. Radack et al. (Rockwell Collins), "Civil Certification of Multi-core Processing Systems in Commercial Avionics," 2018.

[7] Freescale Semiconductor, "e6500 Core Reference Manual," https://www.nxp.com/docs/en/reference-manual/E6500RM.pdf, 2014, E6500RM. Rev 0. 06/2014.

[8] R. Pujol et al, "Empirical evidence for mpsocs in critical systems: The case of nxp's t2080 cache coherence," in *DATE*, 2021.

Roger Pujol received his BSc degree in Computer Engineering with specialization in Computation from Universitat Politècnica de Catalunya (UPC), Spain, in 2018. That same year, he started working as Research Student at the Computer Architecture - Operating Systems (CAOS) group of Barcelona Supercomputing Center (BSC). He also completed his Master in Innovation and Research in Informatics with specialization in Advanced Computing from UPC, Spain, in 2020. Since then, he remained in BSC's CAOS group and started as a Ph.D. student at the Department of Computer Architecture of UPC, Spain.

Transcriptome Analysis of Differential Gene Expression in Disease

Jose Miguel Ramirez[#1], Raquel García-Pérez[#2], Marta Melé[#3]

#Life Science Department, Barcelona Supercomputing Center (BSC), Eusebi Güell 1-3, Barcelona, Spain

[1]jose.ramirez1@bsc.es, [2]raquel.garcia@bsc.es [3]marta.mele@bsc.es

Keywords— **transcriptomics, disease, demographic traits, diabetes**

EXTENDED ABSTRACT

Many diseases strongly impact the human transcriptome at the gene expression level [1]. However, previous work has focused on accessible tissues [2], and has not incorporated the effect of demographic traits, known risk factors for complex diseases [3]. Here, we leveraged the GTEx dataset to investigate the gene expression changes associated with different diseases. By studying the transcriptomes from an organismal perspective -across tissues and individuals- we can gain deeper insights into disease biology and help preventing complex diseases.

A. The GTEx dataset

The GTEx dataset [4] was originally intended to represent a healthy part of the population, but its latest version (v8) has been able to include diseased samples. This project has collected RNA-seq data from 49 human tissues in more than 838 individuals.

We studied 20 diseases in this dataset and controlled for their different effects across demographic groups, depending on age, ancestry, sex, and body mass index.

B. Gene Expression Changes with Disease

We performed differential gene expression analysis using linear models per tissue accounting for all available diseases and demographic traits. The largest number of differentially expressed genes was found in Hashimoto's thyroiditis in the thyroid, followed by pneumonia in the lung and atherosclerosis in the tibial artery.

Some of these diseases alter the transcriptome in specific directions, for instance, most of the genes affected by type 1 diabetes in the pancreas were downregulated, consistent with the known beta-cell loss in long-standing diabetes [5].

C. Contribution of Disease to Expression Variability

We then asked how much of the expression variability in the whole population can be explained by our diseases compared to demographic traits. In most cases, diseases explain a larger proportion of tissue expression variability than age, ancestry, sex, and body mass index combined, indicating profound alterations in the tissue transcriptome. This means that the thyroid of a healthy young female is more similar to that of a healthy old male than to a young female with Hashimoto's thyroiditis, as we expected.

D. Similar Effects Across Tissues in Diabetes

We benefited from the multi-tissue data to study how diseases affect 49 human tissues for the first time. Although most of the effects are tissue-specific, we find 79 genes dysregulated across tissues with type 1 diabetes and 309 with type 2 diabetes. Some of these genes have previously been associated with the disease, like OXCT1 [6], downregulated in 6 tissues with type 2 diabetes, including both subcutaneous and visceral adipose tissues.

E. Similar Effects Across Diabetes in Nerve

Surprisingly, the most affected tissue in type 1 and type 2 diabetes was the tibial nerve with 491 and 2955 differentially expressed genes, respectively, and neither the pancreas nor the adipose tissue, as we could have expected. This could be explained by diabetic neuropathy, a pathology originating due to continuous high blood sugar levels and known to occur independently with the type of diabetes [7]. We checked if the affected genes were the same in both diabetes and found 327 genes in which this was the case. Not only this number is higher than expected by chance, but all the genes were dysregulated in the same direction.

F. Machine Learning on Nerve Images

To validate the finding in the nerve at the transcriptomic level, we benefitted from the histology images available for the GTEx samples to explore whether the diabetic status could be distinguished at the image level or not.

We extracted Haralick features from the GTEx images and trained a support vector machine. With this simple model, we have already been able to predict diabetic status with an accuracy of 74% and an area under the ROC curve of 73%. This result validates our previous finding, suggesting that type 1 and type 2 diabetes have major impacts on the tibial nerve, a signal that can be detected both at the transcriptome and image level.

G. Additive Effects

Demographic traits often influence disease prevalence and progression [3], so next, we set out to investigate the interplay between demographic traits and diseases.

33.6% of the genes affected by a disease were also affected by a demographic trait, being age and sex the traits overlapping the most with diseases. The most prominent examples were both diabetes in the nerve, in which the overlap with age was significantly higher than expected by chance, but also in particular directionalities. In both cases, the genes upregulated with diabetes were upregulated in older individuals, and the ones downregulated with diabetes were also downregulated in older individuals. This suggests that diabetes impacts the tibial nerve similarly to biological ageing.

H. Conclusion

Some diseases affect more tissues than just the tissue of origin, highlighting the importance of multi-tissue studies to gain insight into disease mechanisms and advance towards prevention and cure.

This work also shows that diseases often alter genes with underlying expression differences in the human population, highlighting the importance of considering additive effects when studying inter-individual gene expression variation in human disease. Understanding the relationship between diseases and demographic traits is crucial to advance towards personalized medicine, as the second ones also tend to influence disease risk and progression.

I. Future Enhancements

In this work, we have explored the effect of disease on gene expression. However, diseases can also affect the transcriptome at the alternative splicing level. The GTEx dataset could also allow this type of analysis that would help us increase the knowledge about diseases and their interplay with demographic traits.

We could also validate the signal observed at the gene expression level with the histology images of other tissues. And, potentially, we could also correlate gene expression matrices with features extracted from the images to pinpoint genes to specific changes in tissue morphology.

J. ACKNOWLEDGEMENTS

References

[1] Casamassimi, A. et al. "Transcriptome Profiling in Human Diseases: New Advances and Perspectives." *International journal of molecular sciences* vol. 18,8 1652 (2017)

[2] Piasecka, B. et al. "Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges." *Proceedings of the National Academy of Sciences of the United States of America* vol. 115,3 E488-E497 (2018)

[3] Huypens P, et al. "Epigenetic germline inheritance of diet-induced obesity and insulin resistance." *Nature Genetics* vol. 48,5 497-9 (2016)

[4] Lonsdale, J., Thomas, J., Salvatore, M. *et al.* "The Genotype-Tissue Expression (GTEx) project." *Nature Genetics* 45 580-585 (2013)

[5] Richardon, Sarah J, and Alberto Pugliese. "100 YEARS OF INSULIN: Pancreas pathology in type 1 diabetes: an evolving story." *Journal of Endocrinology* 252,2 R41-R57 (2022)

[6] MacDonald, MJ. et al. "Decreased levels of metabolic enzymes in pancreatic islets of patients with type 2 diabetes." *Diabetologia* vol. 52,6 1087-91 (2009)

[7] Iqbal, Z. et al. "Diabetic Peripheral Neuropathy: Epidemiology, Diagnosis, and Pharmacotherapy." *Clinical Therapeutics* vol. 40,6 828-849 (2018)

Author biography

Jose Miguel Ramirez is a bioinformatician passionate about data science, machine learning, RNA and the role this molecule plays in disease. He studied a bachelor's degree in Bioinformatics by the UPF, UB, and UPC, and he did his bachelor's thesis on machine learning applied to Oxford Nanopore Technologies. He did a master's in Data Science by the UPC and studied abroad for an exchange at Aalto in Helsinki to focus on high throughput Bioinformatics. He did his master's thesis at the Transcriptomics and Functional Genomics Lab to work with transcriptomics and stayed to work for an extra year. Now, he is starting his PhD at the same lab where he is expanding the work started as a master student.

CSDownscale: an R Package for Statistical Downscaling

Jaume Ramon[#1], Llorenç Lledó[#2], Lluís Palma[#3], Carlos Delgado-Torres[#4], Raül Marcos[*5]

#Barcelona Supercomputing Center, Carrer Jordi Girona 31, Barcelona, Spain

[1]jaume.ramon@bsc.es

*Universitat de Barcelona, Facultat de Física i Química, Carrer Martí i Franquès, 1, Barcelona, Spain

Abstract—**Downscaling is any procedure to infer high-resolution information from low-resolution variables. Many of these techniques have been defined and applied to climate predictions, which suffer from important biases due to the coarse global grids in which they are delivered.**

To help solve this undesirable issue, the R package resulting from this work provides a set of statistical downscaling methods for climate predictions, ready to be applied to refine the output of climate predictions.

A. Introduction

Climate predictions appear helpful in anticipating climate variations and taking timely action to manage their possible effects better. Their applicability in many sectors such as agriculture or renewable energy has increased lately, mirroring their potential value and cost-saving benefits[1]. Climate predictions rely on General Circulation Models (GCMs), which are based on a set of equations representing the dynamics of the atmosphere plus other processes occurring in oceans or the land surface.

These predictions are obtained on a global scale. Therefore, the grids used to compute the model values avoid using fine spatial resolutions in order to reduce computational costs. The downside, however, is that climate values at the fine or local scale are often misrepresented due to biases and drifts produced by the fact that model values represent an average value for a grid cell of thousands of square kilometres[2]. The so-called representativeness errors limit the usage of climate predictions for regional or local applications.

To transfer the coarse-scale information into the fine-scale, downscaling techniques have been introduced into the climate prediction arena. Two types of downscaling techniques do exist. Dynamical downscaling uses a Regional Climate Model, which is coupled to the GCMs, using its output as boundary conditions, to model the fine-scale phenomena that cannot be simulated by the GCMs alone. Although dynamical downscaling may be preferred for some applications, it is known to be computationally expensive[3]. On the other hand, statistical downscaling uses empirical relationships established by the coarse-scale data (predictor) and the fine-scale values (predictand). The statistical downscaling approach (see [4] for a review) is relatively straightforward to implement with climate prediction systems and is computationally cheaper than dynamical downscaling. The present work compiles different statistical downscaling techniques for climate predictions and provides the code using the R programming language.

B. Statistical downscaling methods

Table 1 summarises the six statistical downscaling methodologies available in the R package CSDownscale.

TABLE I
STATISTICAL DOWNSCALING METHODS IN CSDOWNSCALE

Method	Definition
Interpolation	Regrid of a coarse-scale grid into a fine-scale grid, or interpolate model data into a point location. Different interpolation methods, based on different mathematical approaches, can be applied: conservative, nearest neighbour, bilinear or bicubic. Does not rely on any data for training.
Interpolation plus bias adjustment	Interpolate model data into a fine-scale grid or point location. Later, a bias adjustment of the interpolated values is performed. Bias adjustment techniques include simple bias correction, calibration or quantile mapping.
Interpolation plus linear regression	Interpolate model data into a fine-scale grid or point location. Later, a linear-regression with the interpolated values is fitted using high-res observations as predictands, and then applied with model data to correct the interpolated values.
Large-scale predictors and local ECVs	Interpolate model data into a fine-scale grid or point location. Later, a linear-regression with large-scale predictors from the same model (e.g. teleconnection indices) is fitted using high-resolution observations as predictands. Finally, the linear-regression is applied with model data to correct the interpolated values.
Stencil	Interpolate model data into a fine-scale grid or point location. Later, a linear-regression with the four (or nine) nearest neighbours is fitted using high-resolution observations as predictands, and then applied with model data to correct the interpolated values.
Logistic regression	Relate ensemble mean anomalies of the large-scale forecasts directly to probabilities of observing above normal/normal/below normal conditions at the local scale using a sigmoid function. It does not produce an ensemble of forecasts but rather their associated probabilities. It is a statistical method with few parameters to train, and only benefits from local information, but it has shown good performance.

C. Results

The R package CSDownscale has been applied to actual climate data. Seasonal predictions for surface temperatures in February in the Iberian Peninsula have been downscaled from 1° to ~0.3° of horizontal resolution. The SEAS5 seasonal prediction system (Figure 1 (a)) has a spatial resolution of 1

degree, thus being unable to represent localities and particularities of temperatures in highly heterogeneous areas (e.g. mountainous regions or coastlines).

a) Predictions

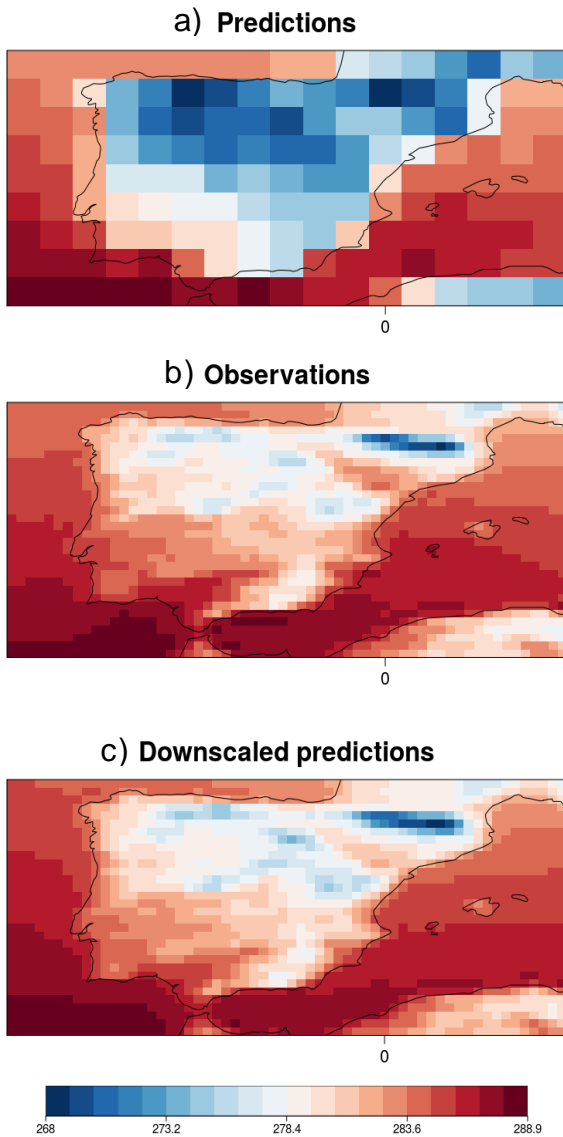### b) Observations

c) Downscaled predictions

Fig. 1 a) SEAS5 seasonal prediction for surface temperature in its original grid, b) ERA5 observations for the same variable in the high-resolution grid and c) downscaled seasonal predictions to the high-resolution grid using the Analogs method. Lead month is zero and the represented month is February 2000. Temperatures are measured in Kelvin (K).

These differences are especially noticeable after comparing the observed temperatures for that particular month (Figure 1 (b)). Predicted values appear systematically underestimated all over the Iberian Peninsula. Moreover, mountainous regions and coastlines features are not properly represented in the prediction field due to representativity issues.

The downscaled seasonal prediction (Figure 1 (c)) appears to reproduce the local singularities: temperatures are systematically lower in mountainous ranges (e.g. Pyrenees) and higher within the valleys (e.g. Ebro river valley). The discontinuity in the temperature values expected along the coastlines is also noticeable.

D. Conclusions

Statistical downscaling is a cheap and efficient tool to generate seasonal predictions at a fine scale. Different methodologies are available in CSDownscale, allowing for rapid comparison of the strengths and weaknesses between methods.

The applicability of CSDownscale is not restricted to the climate field, but other sectors like agriculture or renewable energy may benefit from analyses using CSDownscale. The wind power industry, for example, may benefit from downscaled seasonal predictions at the wind farm scale and advance the amount of renewable generation.

E. ACKNOWLEDGEMENTS

References

[1] A. Troccoli, "Seasonal climate forecasting", Meteorol. Appl., 17, 251–268, https://doi.org/10.1002/met.184, 2010.

[2] A. M. Anders, G. H. Roe, B. Hallet, D. R. Montgomery, N. J. Finnegan, and J. Putkonen, "Spatial patterns of precipitation and topography in the Himalaya", Spec. Pap. Geol. Soc. Am., 398, 39–53, https://doi.org/10.1130/2006.2398(03), 2006.

[3] M. García-Díez, J. Fernández, D. San-Martín, S. Herrera, and J. M. Gutiérrez, "Assessing and improving the local added value of WRF for wind downscaling", J. Appl. Meteorol. Climatol., 54, 1556–1568, https://doi.org/10.1175/JAMC-D-14-0150.1, 2015.

[4] J. M. Gutiérrez, J. Bedia, R. Benestad, and C. Pagé, "Local predictions based on statistical and dynamical downscaling", 1–21, 2013.

Author biography
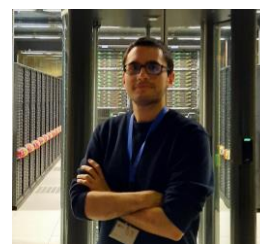
Jaume Ramon holds a PhD in Physics since January 2022, entirely developed at the Earth Sciences department of the Barcelona Supercomputing Center. He is currently a post-doc researcher at the same institution, where he works hard to improve the quality of seasonal predictions. He also does his best at writing clean and efficient code. A lover of gardening and passionate about board games.

Non-Reacting Large Eddy Simulation of a Low Swirl Burner

Fredherico Rodrigues[*][†], José Maria García-Oliver[†] Daniel Mira[*]

[*]Barcelona Supercomputing Center, Barcelona, Spain

[†]CMT - Motores Termicos, Valencia, Spain

E-mail: {fredherico.rodrigues, daniel.mira}@bsc.es, {jgarciao@mot.upv.es}

Keywords—LES, Low Swirl Burner, High-performance computing.

I. Extended Abstract

A. Introduction

Reduction of the emission of pollutants is one of the main concerns in the study of combustion today. Thermal NO_x for example, is generated when high temperatures are achieved in the combustion process. The use of Lean Premixed Prevaporized (LPP) burners addresses this issue by generating flames at lower temperatures near the lean blow-off limit. The resulting flame is, however, less stable and more prone to quenching. For such flames, stabilization can be obtained by swirl, with the introduction of a tangential component in addition to the axial flow.

For High Swirl Burners (HSB), stabilization occurs due to strong recirculation zones, while for a Low Swirl Burner (LSB), this occurs on a low velocity zone in the center region of the flame. The LSB has two main advantages. First, it produces even lower emissions compared to the HSB configuration, and second, the flame is lifted, which is interesting for avoiding overheating of the nozzle.

With the use of High Performance Computing and Large Eddy Simulation, we are able to model and investigate the flow characteristics for the LSB geometry proposed by [1]. In a first moment and presented here, we compute the non-reacting condition, where axial velocity profiles are compared to the experimental results. In the second phase the reacting case is to be solved.

B. Burner design

The chosen LSB [1] has a simple geometry comprised of a plenum with two tangential injections and one central injection. Tangential flow is transported through an annulus and mixes with the axial flow upstream the nozzle exit plane. The geometry is simplified to avoid the use of geometrical arrangements, so that swirl is controlled by the ratio of flow rates.

Due to its design, this burner is characterized by only two parameters. The first is the Reynolds number for the characteristic diameter D, written as

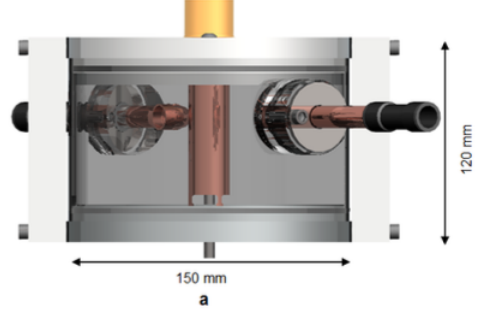$$\mathrm{Re}_D = \frac{\rho u D}{\mu} = \frac{4\dot{m}}{\pi \mu D}, \qquad (1)$$

Fig. 1. General view of the Low Swirl Burner, [1]

where ρ, u, μ and $\dot{m}$ are respectively the density, velocity, dynamic viscosity and mass flow rate.

The second parameter is the swirl number, which is controlled by the ratio of flow rates between axial and tangential injections.

$$S = \frac{\int_{-r}^{+r} \rho \pi u_z \cdot u_\theta r^2 dr}{D \int_{-r}^{+r} \rho \pi u_z^2 r dr}, \qquad (2)$$

with r being the radius and the sub indexes z and θ representing the axial tangential components, respectively.

The basic equations to be solved for the non-reacting incompressible isothermal flows are the conservation of mass (continuity) and conservation of momentum. Following [2], ignoring body forces, the equations are written respectively as

$$\frac{D\rho}{Dt} + \rho \frac{\partial u_i}{\partial x_i} = 0, \qquad (3)$$

$$\rho \frac{\partial (u_i)}{\partial t} + \rho u_i \frac{\partial (u_j)}{\partial x_i} = -\frac{\partial P}{\partial x_j} + \frac{\partial \tau_{ij}}{\partial x_i}, \qquad (4)$$

with τ_{ij} being the viscous stresses for directions i and j.

Considering sufficiently high Reynolds number, flow can be assumed fully turbulent, and Equations 3 and 4 are rewritten in terms of LES filtered quantities such that

$$f = \bar{f} + f', \qquad (5)$$

where $\bar{f}$ represents the filtered part while f' represents the fluctuating part of a generic quantity f. Here, the filtering is
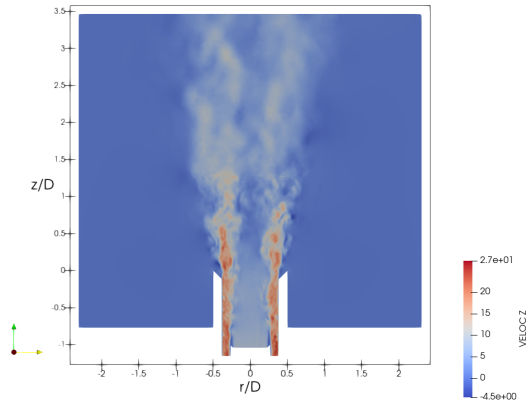
Fig. 2. Snapshot of instantaneous axial velocity at the centerline position of the burner.

Fig. 3. Normalized axial velocity profiles at the nozzle exit plane.

done implicitly in such way that the filter is the mesh itself. Thus we have

$$\frac{D\bar{\rho}}{Dt} + \bar{\rho}\frac{\partial(\bar{u}_i)}{\partial x_i} = 0, \qquad (6)$$

$$\bar{\rho}\frac{\partial(\bar{u}_i)}{\partial t} + \bar{\rho}\bar{u}_i\frac{\partial(\bar{u}_j)}{\partial x_i} = -\frac{\partial\bar{P}}{\partial xj} + \frac{\partial}{\partial x_i}\left[\bar{\tau}_{ij} - \bar{\rho}(\overline{u_i u_j} - \bar{u}_i\bar{u}_j)\right], \qquad (7)$$

with the rightmost term $(\overline{u_i u_j} - \bar{u}_i\bar{u}_j)$ representing the turbulent viscosity, left to be model in the subgrid scale.

C. Modeling approach and preliminary results

The isothermal flow is resolved by the incompressible Navier-Stokes equations in the framework of Large Eddy Simulation (LES) using a low dissipation numerical scheme [3] with the Vreman subgrid model [4]. The numerical domain contains 18 million tetrahedral elements, created in Gmsh [5], and replicates the entire geometry of the burner, plus a combustion chamber opened to the atmosphere (slip wall condition). Air mass flow at standard pressure and $T = 300$ K is supplied to the central and tangential inlets. Additional atmospheric air is supplied as a low speed coflow (5% of the average flow velocity) at the outer rim of the burner to provide some entrainment. Synthetic turbulence is added to the central injection following [6].

The overall geometry and the instantaneous axial velocity values at the centerline of the burner's geometry are shown in Figure 2, while Figure 3 presents the normalized velocity profiles for the axial flow at the nozzle exit. A correction factor of 10% is applied to the numerical profiles to compensate for experimental uncertainties. The black line (——) represents the experimental results from [1], while the blue solid line (——) represents a baseline simulation where the Reynolds number ($Re = 12000$) and ratio of tangential / axial flow at the inlets are imposed as presented in the low swirl condition of the original experiment (a ratio of 3.64 tangential / axial mass flow). Additional profiles varying the air split ratios are included.

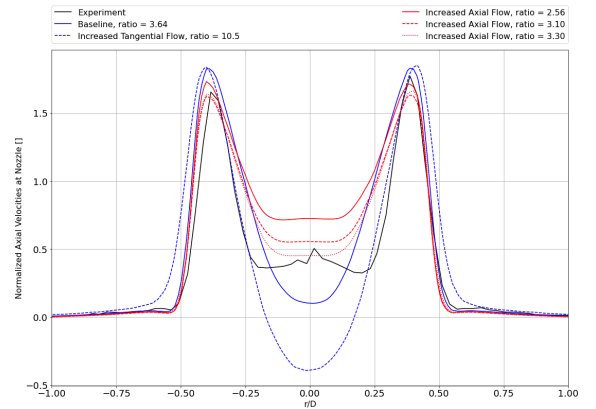For future reacting flows, the constant density assumption can no longer be applied due to thermal expansion and additional equations have to be solved for the transport of each chemical species (Y_K) and enthalpy (h). Using tabulated chemistry it is possible to define a progress variable, Y_c, based on a selection of chemical species, and a mixture fraction, Z, in order to substitute the large number of equations, and solve only for the transport of Z and Y_c and their variances Z''^2 and $Y_c''^2$. This deduction can be find in different books, such as [2]. Further simulations are currently being carried out for both the high swirl non-reacting condition and the reacting cases. The latter are to be performed for the propane and air mixture with a combination of tables constructed with premixed flamelets with varying equivalence ratios and PDF integrated properties.

II. ACKNOWLEDGMENT

REFERENCES

[1] M. Legrand, "ESTUDIO Y CARACTERIZACIÓN DE UN QUE-MADOR ESTABILIZADO POR GIRO," Ph.D. dissertation, Universidad Carlos III de Madrid, 2008.

[2] T. Poinsot and D. Veynante, *Theoretical and Numerical Combustion*. Edwards, 2005.

[3] O. Lehmkuhl *et al.*, "A low-dissipation finite element scheme for scale resolving simulations of turbulent flows," *Journal of Computational Physics*, 2019.

[4] A. W. Vreman, "An eddy-viscosity subgrid-scale model for turbulent shear flow: Algebraic theory and applications," *Physics of Fluids*, 2004.

[5] C. Geuzaine and J. F. Remacle, "Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities," *International Journal for Numerical Methods in Engineering*, 2009.

[6] A. Kempf *et al.*, "Efficient Generation of Initial- and Inflow-Conditions for Transient Turbulent Flows in Arbitrary Geometries," *Flow, Turbulence and Combustion*, 2005.
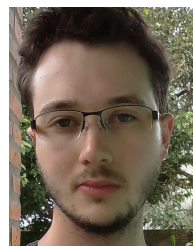
Fredherico Rodrigues is a mechanical engineer with a Bachelor's degree from the Federal University of Rio Grande do Sul, Brazil, graduated in 2017. He obtained his Masters' at the same University in 2020, working with the numerical simulation of combustion phenomena. Currently he is enrolled in a Doctorate program at the Valencia Polytechnic University and working as a researcher in the Propulsion Technologies Group at the Barcelona Supercomputing Center in Spain.

Space Shuttle: A Test Vehicle for the Reliability of the SkyWater 130nm PDK using OpenLane and the Google/E-fabless Shuttle Run for Future Space Systems

Ivan Rodriguez-Ferrandez†*, Alvaro Jover-Alvarez†*, Leonidas Kosmidis†* David Steenari ‡

*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
‡European Space Agency, ESTEC, Netherlands
E-mail: {ivan.rodriguez, alvaro.jover leonidas.kosmidis}@bsc.es
david.steenari@esa.int

Keywords—ASIC, Memory, Space ,Radiation.

I. Extended Abstract

The ASIC industry is experiencing a massive change in the recent years with more and more small and medium business entering the custom ASIC development. This trend is fueled by the recent open hardware movement and relevant government and privately funded initiatives. These new developments can open new opportunities in the space sector, which is traditionally characterised by very low volumes and very high non-recurrent (NRE) costs, if we can show that the produced chips have favourable radiation properties. In this ACM SRC entry, we describe the design and tape-out of Space Shuttle, the first test chip for the evaluation of the suitability of the SkyWater 130nm PDK and the OpenLane EDA toolchain using the Google/E-fabless shuttle run for future space processors.

A. Problem and Motivation

Until very recently, custom ASICs were very rare for small and medium companies due to the overall prohibitive cost of their production, which could only be afforded by few large companies.

This high cost comes mainly from the very expensive commercial EDA tools used for chip designs. However, recently there have been government initiatives such as DARPA's OpenLane/OpenROAD which supported the production of a fully open source EDA toolchain from hardware description languages down to GDSII.

This has been followed by private initiatives such as Google's support for an open source PDK in collaboration with SkyWater for their older 130nm technology, as well as with E-fabless for the sponsored production of multi-project wafers (MPW) of open source designs to test and mature the status of these tools.

The impact of this completely open source option for hardware designs can reduce the hardware production cost many orders of magnitude, lowering the barriers to entry and enabling the production of innovative designs and prototypes, which was not possible before.

One of the sectors that can benefit the most is space due to its different needs than other domains. Space frequently relies on older technology nodes for fabrication of its processors which are normally reused for decades. When the performance of these processors is not sufficient for a mission requirements, the system is normally supplemented with FPGAs. This drives up the cost of missions due to the inability to apply economy of scales in this sector.

However, this low volume production and very specific requirements create a perfect match for new revolution in the space silicon. Therefore, the objective of our project is to validate for the first time this process and tool-chain for space use.

II. Background and Related Work

Space electronics are susceptible to radiation effects such as single event upsets (SEUs) and single event latch-ups (SELs). These radiation effects are well known as well as solutions about how to mitigate them. An example of such mitigation techniques for radiation hardening are for example the ones applied to the widely used in space LEON3 processor[1] over an FPGA. Such mitigations are focused around single error bit flips. Other techniques[2] focus in time delays in redundant structures in order to prevent the same error to impact them in an identical way. Moreover, these methods use checkpoints in order to restore the correct state when these transient errors are detected but not corrected.

III. Approach and Uniqueness

The main goal of our project is to assess the reliability of the SkyWater 130nm manufacturing process and Open Source PDK using the open source OpenROAD EDA toolchain and the Google sponsored shuttle run through e-fabless. Moreover, we aim to evaluate different reliability mitigation techniques.

In order to do so, we have designed a prototype chip called Space Shuttle which has been submitted for tape-out during the second MPW run. In order to assess the reliability of the chip we require methods to be able to observe in a detailed way its internal state, so that not only we can identify whether errors

(e.g. bitflips) have occurred, but also to know which part of the design has been affected. For this reason, the initial reliability assessment has to be performed in a design with relatively low complexity, high level of control and with appropriate facilities to provide the required information.

The most vulnerable parts of a hardware design are its memory structures because they retain their previous values. Therefore, in our chip we have decided to focus on error detection and error correction specifically on flip-flops and registers which are the primary memory storage primitives. Therefore, our design is centered around a register file with 32, 32-bit registers implemented with flip-flops. This is because currently, the OpenRAM compiler is considered experimental and because we want to have more control over the memory structures to implement our detection and correction methods.

We have implemented multiple reliability solutions such as different degrees of replication and ECC around this register file, which can be selectively configured and combined (with some limitations), in order to reliably detect whether errors occur, and in case they do, whether they can be corrected. Moreover, we include a redundant Monitoring Unit consisting of detailed event counters for each of the registers. This way, we can also assess the protection level offered by these different reliability methods and study their trade-offs in order to guide future developments.

In particular, our design has implemented the following characteristics:

- 32, 32-bit register file implemented with flip-flops, organised in 8 banks which can be used in parallel. Each register value can be individually set or inspected. The register file can be either clocked with a user controled signal, offering full control, or with the default chip clock.

- 4 different protection mechanisms, each of which can be enabled selectively and in combination with others:
  - Error Correction Code (ECC): ECC with 1 bit correction and 2 bit correction.
  - Triple Redundancy: The input value is triplicated in the register file.
  - Shadow Register: The input value has a copy in the register file.[3]
  - ECC Shadow Register: The input value has a copy in the register file with ECC protection of 1 bit correction and 2 bit correction.[4]

- Duplicated Monitoring Unit: Individual 32-bit counters per register, reporting the write and read operations performed, as well as the number of detected and corrected errors. Both the monitoring registers and their combinational counter circuit is replicated, ensuring reliable information.

The design has been verified with gate level simulations and it comes with several tests focusing in the verification of different parts of the circuit. A total of 11 tests focus in the verification of the design using the internal I/O pins of the Caravel SoC used for interfacing in all MPW chip designs. 7 additional tests verify the Wishbone interface which allow accessing the performance counters.

IV. RESULTS AND CONTRIBUTIONS

The Space Shuttle design is fully open-source [5] and can be modified and improved by the community. At the final synthesis the chip uses 2.22 mm^2 with a total of 75841 SkyWater cells targetting a frequency of 100 Mhz.

As a result of this work, we have successfully send the design to production inside the second e-fables Google-sponsored shuttle [6], that will be delivered in December 2021.
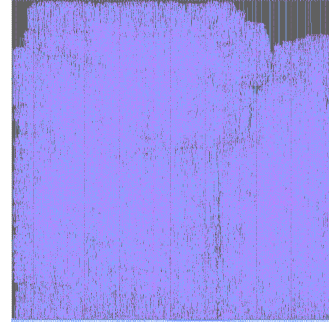

Fig. 1. Floorplan of the resulting chip

When it arrives will be initial tested in normal conditions in order to verify that design behaves and responds to the test that we develop for the simulation. Once it is verified under normal conditions and passes the tests, a set of samples will be send to do a proton base radiation testing in order to characterise its behaviour under radiation.

To our knowledge this is the first attempt to assess the reliability of this new combination of open source technologies, as well as among the first academic tape-outs performed within this program.

REFERENCES

[1] S. Kasap *et al.*, "Survey of Soft Error Mitigation Techniques Applied to LEON3 Soft Processors on SRAM-based FPGAs," *IEEE Access*, vol. 8, pp. 28 646–28 658, 2020.

[2] M. Nicolaidis, "Design for Soft Error Mitigation," *IEEE Transactions on Device and Materials Reliability*, vol. 5, no. 3, pp. 405–418, 2005.

[3] ——, "Time Redundancy based Soft-error Tolerance to Rescue Nanometer Technologies," in *Proceedings 17th IEEE VLSI Test Symposium (Cat. No. PR00146)*. IEEE, 1999, pp. 86–94.

[4] A. Bouajila *et al.*, "An Architecture and an FPGA Prototype of a Reliable Processor Pipeline Towards Multiple Soft-and Timing Errors," in *14th IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems*. IEEE, 2011, pp. 225–230.

[5] I. Rodriguez, "Space Shuttle Repository," 2021. [Online]. Available: https://github.com/jaquerinte/caravel_radiation_harden

[6] "Space Shuttle E-fables MPW-2 Project Entry," 2021. [Online]. Available: https://efabless.com/projects/214

Ivan Rodriguez-Ferrandez received his BSc degree in Computer Engineering from University of Alicante. The following year, enter in Barcelona Supercomputing Center (BSC) at Computer Architecture and Operating Systems (CAOS). He completed his MSc degree in Innovation and Research in Computer Science from Universitat Politècnica de Catalunya (UPC), Spain. In 2021 he started the PhD at UPC that is con-funded by the European Space Agency (ESA).

RosneT: A Block Tensor Algebra Library for Out-of-Core Quantum Computing Simulation

Sergio Sanchez-Ramirez* (ORCID), Javier Conejero† (ORCID), Artur Garcia-Saez*‡ (ORCID)

*QUANTIC, Barcelona Supercomputing Center, Barcelona, Spain

†Workflows and Distributed Computing, Barcelona Supercomputing Center, Barcelona, Spain

‡Qilimanjaro, Barcelona, Spain

E-mail: {sergio.sanchez.ramirez, javier.conejero, artur.garcia}@bsc.es

Keywords—tensor network, quantum computing, simulation, out-of-core, task-based programming, COMPSs, distributed computing, HPC.

I. INTRODUCTION

Abstract—**With recent Quantum Devices showing increasing capabilities to perform controlled operations, further development on Quantum Algorithms may benefit from Quantum Simulations on classical hardware. Among important applications one finds verification and debugging of Quantum Algorithms, and elucidating the frontier for real Quantum Advantage of new devices [1]. Tensor Networks are regarded as an efficient numerical representation of a Quantum Circuit, but exponential growth forces tensors to be distributed among computing nodes. A number of methods and libraries have appeared recently to implement Quantum Simulators with Tensor Networks [2], [3] intended for HPC clusters. In this work we develop a Python library called RosneT using a task-based programming model able to extend all tensor operations into distributed systems, and prepared for existing and upcoming Exascale supercomputers. It is compatible with the Python ecosystem, and offers a simple programming interface for developers.**

Tensors as Quantum States. The Quantum State $\Psi_{\mathbf{i}}$ describing a Quantum computation is represented by a n-order unit tensor on $\bigotimes_{i=1}^{n} \mathbb{C}^2$ where n is the number of qubits and $\mathbf{i} = i_1 \ldots i_n$ are the indices of the qubits. This tensor can be further decomposed in a Tensor Network of tensors A_k. Using a general tensor contraction operation Φ we define the state as $|\Psi_{\mathbf{i}}\rangle = \sum_{i_1 \ldots i_n} \Phi(A_k)|i_1 \ldots i_n\rangle$.

Quantum Gates acting over Quantum States are also represented as tensors of the network, such that any physical observable is obtained solely from operations over the tensors. When contracting the tensor network, intermediate tensors will be created whose order is greater. Generally, as the depth of the networks increases, so does the order of the greatest intermediate tensor. This imposes limits on the size of the tensor networks that we can contract, as the size of tensor grows exponentially fast with its order.

A common method for handling large tensors is tensor cutting or slicing [4], [5], [6]. Slicing is based on the idea of Feynman's Path Integral formulation of quantum mechanics, that all the paths in a tensor network equally contribute to the final quantum state. An index (or edge) in the tensor network represents a vector space of dimensionality D between two tensors. If we select one and project it into one of its basis, then we are filtering $\frac{1}{D}$ of the paths and the intermediate tensors that contain the index will shrink to occupy a $\frac{1}{D}$ of its original space at the cost of $\frac{1}{D}$ fidelity in the final result. Furthermore,
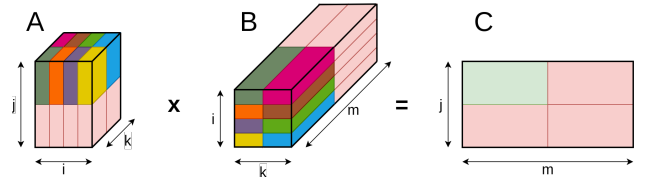

Fig. 1. Example of a block tensor contraction between two 3-order tensors into a 2-order tensor. To compute a block of C (light green), the needed pairwise A, B blocks are colored in a range of colors by matching couples.

if we make copies of the tensor network, each projected to a different basis of the selected index, and repeat this process, we end up with a massively parallel problem and can calculate up to the desired fidelity. However, contraction of non-sliced tensors become redundant and may add a significant overhead to computation. We notice that by viewing tensor projections as block divisions of the original tensors, this compute overhead disappears and becomes an indicator of block reuse while still can fit in a distributed system.

Block Tensor Contraction. Based on the Dimension Extents layout [2], [3], we slice tensors into uniform grids of sub-tensors or blocks. Employing the example in Figure 1, block tensor contraction between tensors A_{ijk} and B_{ikm} is performed following these steps:

1) For each output block in tensor C, we select the blocks from A and B needed to compute the resulting block. These blocks are chosen by matching coordinates from non-contracting indexes j, m.
2) Once the input blocks of A and B have been selected, we group them in pairs by matching the coordinates of contracting indexes i, k.
3) Each of the pairwise coupled blocks, when contracted locally, will compute a partial contribution to the output block.
4) Partial result blocks are summed into the block of C.

II. IMPLEMENTATION

COMP Superscalar (COMPSs) [7] is a task-based programming model for distributed computing. It is supported by a runtime that manages several aspects of the applications' execution, such as task dependency analysis, data synchronization and resource management. In the COMPSs programming model, the developer has to explicitly mark a function as task. On call, COMPSs will execute it in a free distributed

worker selected by the scheduler. The type and direction of the task dependencies can be annotated. COMPSs has bindings for C/C++, Java and Python languages [8]. We opted for Python due to its popularity on the scientific community having already a rich ecosystem above which we can leverage our work.

Alternative distributed runtimes available in Python are Daskand Ray. The main features that differentiate COMPSs are:

- Data is located out-of-core. Our benefits here are two-fold: (1) Fault-tolerance is straightforward and ensured for each task and (2) with a higher secondary storage, the size of the quantum states we can represent grows.

- Detailed tracing and profiling support with Extrae and Paraver.

- Dynamic Heterogeneous Computing. The user may provide several implementations for the same routine and COMPSs will select the implementation on run-time based on the available resources.

III. EVALUATION

Block size has a non-trivial effect on the execution time of simulations. On Figure 2, the execution time of a Random Quantum Circuit simulation for different maximum block sizes is shown. Note that if the maximum block size is too large, the grain of parallelism is too coarse and it does not fully exploit the resources. Moreover, if the maximum block size is too thin, the (de)serialization overhead surpasses the effective execution time.
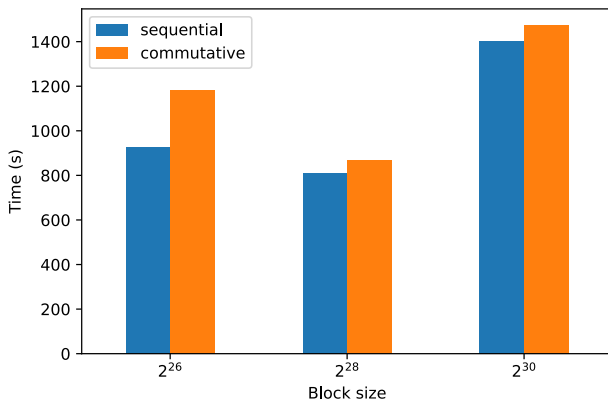

Fig. 2. Execution times for the contraction of a Random Quantum Circuit with 53 qubits and depth 12, varying the maximum block size.

IV. CONCLUSIONS

Using COMPSs runtime our library features out-of-core data location, tracing and profiling support, and potential for heterogeneous computing. As an example thanks to *numpy*'s dispatching mechanism, we can leverage over *numpy* derivatives like `cupy` adding GPU acceleration to `RosneT`. Our library is a novel contribution bringing flexible tools to perform large-scale simulations of Quantum systems to a community of Quantum developers.

REFERENCES

[1] B. Villalonga *et al.*, "Establishing the quantum supremacy frontier with a 281 pflop/s simulation," *Quantum Science and Technology*, vol. 5, no. 3, p. 034003, apr 2020. [Online]. Available: https://doi.org/10.1088/2058-9565/ab7eeb

[2] E. Solomonik *et al.*, "Cyclops Tensor Framework: reducing communication and eliminating load imbalance in massively parallel contractions," p. 13.

[3] D. I. Lyakh, "Domain-specific virtual processors as a portable programming and execution model for parallel computational workloads on modern heterogeneous high-performance computing architectures," *International Journal of Quantum Chemistry*, vol. 119, no. 12, p. e25926, Jun. 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.25926

[4] B. Villalonga *et al.*, "A flexible high-performance simulator for verifying and benchmarking quantum circuits implemented on real hardware," *npj Quantum Information*, vol. 5, no. 1, p. 86, Dec. 2019, arXiv: 1811.09599. [Online]. Available: http://arxiv.org/abs/1811.09599

[5] I. L. Markov *et al.*, "Quantum Supremacy Is Both Closer and Farther than It Appears," *arXiv:1807.10749 [quant-ph]*, Sep. 2018, arXiv: 1807.10749. [Online]. Available: http://arxiv.org/abs/1807.10749

[6] J. Chen *et al.*, "Classical Simulation of Intermediate-Size Quantum Circuits," *arXiv:1805.01450 [quant-ph]*, May 2018, arXiv: 1805.01450. [Online]. Available: http://arxiv.org/abs/1805.01450

[7] R. M. Badia *et al.*, "COMP Superscalar, an interoperable programming framework," *SoftwareX*, vol. 3-4, pp. 32–36, Dec. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352711015000151

[8] E. Tejedor *et al.*, "PyCOMPSs: Parallel computational workflows in Python," *The International Journal of High Performance Computing Applications*, vol. 31, no. 1, pp. 66–82, Jan. 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1094342015594678

[9] S. Sánchez-Ramírez *et al.*, "Rosnet: A block tensor algebra library for out-of-core quantum computing simulation," in *2021 IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)*, 2021, pp. 1–8.

Sergio Sánchez-Ramírez received his BSc degree in Sound and Image Engineering from Universidad Politécnica de Madrid (UPM), Spain in 2019. The following year, he worked as a Researcher at the CITSEM-UPM research center on the NEMESIS-3D-CM project. He completed his MSc degree in Innovation and Research in Informatics, specialization in High-Performance Computing in 2021. Currently, he is a first year PhD candidate at QUANTIC - BSC, working in large-scale quantum computing simulation using tensor networks.

JLOH: Extracting Loss of Heterozygosity Blocks from Short-Read Sequencing Data

Matteo Schiavinato[#1], Valentina del Olmo[#2], Toni Gabaldón[#3]

[#]*Department of Life Sciences, Barcelona Supercomputing Center (BSC-CNS), Plaça Eusebi Güell, 1-3, Barcelona, Spain*
[1]`matteo.schiavinato@bsc.es`, [2]`valentina.delolmo@bsc.es`, [3]`toni.gabaldon@bsc.es`

Keywords— **software, LOH, hybrids**

EXTENDED ABSTRACT

Introduction

Loss of heterozygosity (LOH) happens when a heterozygous genome loses one of the two alleles at a locus. This may have an evolutionary advantage in highly unstable genomes such as those of hybrids (Smukowski Heil et al., 2017). By extracting LOH from a hybrid we understand which alleles were beneficial in its evolution. This is important in wild, clinical, and industrial settings (Gabaldón, 2020). The genomic properties of hybrids are still, however, poorly understood. LOH are studied with reliable short-read sequencing data (Mixão et al., 2019), but the downstream analysis is often done with custom scripts that reduce reproducibility and do not to leverage the power of a computing cluster. Here we present a program called "JLOH" that streamlines LOH extraction from sequencing data maximizing parallel computing.

Materials and Methods

A series of divergent genomes (in short: *Sd*) were simulated from the *S. cerevisiae* genome (*Sc*) using JLOH's *sim* module, with increasing divergence (5%, 10%) and LOH (10%, 20%, 30%, 40%, 50%). Reads were simulated from *Sc* and *Sd* using wgsim (https://github.com/lh3/wgsim), joined together to simulate hybrid reads, and mapped against *Sc* and each *Sd* independently using hisat2 (Kim et al., 2019) with relaxed criteria (--score-min L,0.0,-1.0). The mapping results were filtered and used to call snps with bcftools (Danecek et al., 2021, --multiallelic-caller). The SNPs called were used as input for JLOH's *extract* module. Regions depleted of heterozygous SNPs are good candidates to be LOH blocks and are assessed by JLOH in terms of length, coverage in the up/downstream region, and homozygous SNP density. Called blocks were then compared against the LOH blocks originally introduced in the *Sd* genome to calculate precision and recall.

Results

True positives (TP), false positives (FP), false negatives (FN), precision and recall at each level of divergence and LOH are marked in Table 1.

TABLE I

Div.	LOH	TP	FP	FN	Precision	Recall
5%	10%	110	68	3	0.618	0.973
5%	20%	210	72	5	0.745	0.977
5%	30%	342	67	7	0.836	0.980
5%	40%	526	56	8	0.904	0.985
5%	50%	783	41	15	0.950	0.981
10%	10%	120	109	5	0.524	0.960
10%	20%	232	87	7	0.727	0.971
10%	30%	462	58	7	0.888	0.985
10%	40%	551	62	16	0.899	0.972
10%	50%	838	49	14	0.945	0.984

Conclusions

We conclude that JLOH successfully finds the LOH blocks that were artificially introduced in the simulated genomes using simple input files in common formats. We also conclude that the extent of LOH experienced from a genome may be a limit in JLOH's precision. While all introduced blocks are consistently found at any divergence or LOH rate, the number of false positives is anti-correlated with the level of LOH. Further development of the tool will hopefully reduce this effect.

References

[1] Smukowski Heil CS, DeSevo CG, Pai DA, Tucker CM, Hoang ML, Dunham MJ. Loss of heterozygosity drives adaptation in hybrid yeast. Molecular biology and evolution. 2017 Jul 1;34(7):1596-612.

[2] Gabaldón T. Hybridization and the origin of new yeast lineages. FEMS yeast research. 2020 Aug;20(5):foaa040.

[3] Mixão V, Hansen AP, Saus E, Boekhout T, Lass-Florl C, Gabaldón T. Whole-genome sequencing of the opportunistic yeast pathogen Candida inconspicua uncovers its hybrid origin. Frontiers in genetics. 2019 Apr 25;10:383.

[4] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature biotechnology. 2019 Aug;37(8):907-15.

[5] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb;10(2):giab008.

Author biography

Matteo Schiavinato was born in Italy in 1990. He received both the bachelor and the master degree in Molecular Biology at the University of Padova (Veneto, Italy), completing his studies in 2015. From January 2016 to February 2020 he has been a PhD student in bioinformatics at the University of Natural Resources and Life Sciences (BOKU) in Vienna, Austria. After the PhD, he briefly stayed as a postdoctoral researcher and then moved to the Core Facility Bioinformatics of the same university (BOKU), where he took the role of Operative Head.

Since October 2021 he has joined the research group of prof. Toni Gabaldón at the Barcelona Supercomputing Center (BSC-CNS) as a postdoctoral researcher. His research involves the investigation of the genomic changes underlying hybrid genome evolution.

Exploring the recent evolution of yeast pathogens using the CandidaMine database

Miquel Àngel Schikora-Tamarit[#1], Toni Gabaldón[#]

[#]*Comparative genomics lab, Life Sciences Department, BSC*
[1]`miquel.schikora@bsc.es`

Keywords: Candida, population genomics, drug resistance

EXTENDED ABSTRACT

Background

Advances in medicine (such as chemotherapy or transplants) have extended the life expectancy of patients at the cost of impairing their immune system. This has generated an increasing population of patients highly susceptible to infections. Among them, fungal infections caused by *Candida* species have become a major life-threatening issue, with insufficient diagnostic and therapeutic options. Recent studies have used population genomics in clinical *Candida* isolates to understand their recent evolution, which may also clarify the emergence of phenotypes like drug resistance or virulence. However, there are open questions that hinder our understanding about such evolutionary processes.

Results

We tried to clarify some of these questions by generating variant calling data for ~2,000 isolates from seven major *Candida* pathogens, available at the CandidaMine database. We focused on understanding 1) the role of structural variants (SVs), 2) the genomic determinants of antifungal drug resistance, 3) the role of recombination, 4) the population structure and 5) the similarities among these processes in different Candida species. Importantly, our collection represents a higher sample size as compared to previous studies, suggesting that we have unprecedented power.

Interestingly, we found that SVs drive a significant amount of genetic variation, suggesting that their role should not be overlooked. In addition, we found novel clades and recombination events in several species. Finally, our inter-specific comparison analysis revealed important differences between species, which may be relevant to develop species-tailored diagnostics or therapies.

Conclusions

In summary, our work improves the understanding of the recent evolution of the major *Candida* pathogens.

Author biography

Miquel Àngel Schikora-Tamarit was born in Lleida in 1995. In 2017 he obtains a Bachelor Degree in Human Biology by the Universitat Pompeu Fabra (UPF) of Barcelona. He works on several projects focused on understanding single-cell behavior under the supervision of Dr. Lucas Carey in the Department of Experimental and Health Sciences of the UPF between 2014 and 2018. His research, involving experimental molecular biology techniques and computational analysis, yields four publications in the journals Integrative Biology, Transcription, Genome Research and Cell Reports. He pursues a Master Degree in Bioinformatics at the UPF between 2017 and 2019. He develops the thesis project in the lab of Dr. Toni Gabaldón of the Center for Genomic Regulation (CRG) understanding the evolution of Complex I from fungal genomes (published in Open Biology). He is currently conducting his PhD thesis in Biomedicine at the University of Barcelona (UB), working on the evolution of drug resistance in fungal pathogens under the supervision of Dr. Toni Gabaldón at the Barcelona Supercomputing Center.

Investigating Memory Prefetcher Performance over Parallel Applications: From Real to Simulated

Valéria S. Girelli*[1], Francis B. Moreira[†], Matheus S. Serpa[‡], Danilo C. Santos[§1], Philippe O. A. Navaux[‡]

*Barcelona Supercomputing Center (BSC) - Barcelona, Spain
[†]Department of Informatics, Federal University of Paraná (UFPR) - Curitiba, Brazil
[§]Grenoble Informatics Laboratory, University Grenoble-Alpes (UGA) - Grenoble, France
[‡]Informatics Institute, Federal University of Rio Grande do Sul (UFRGS) - Porto Alegre, Brazil
E-mail: vsoldera@bsc.es, fbmoreira@inf.ufpr.br, danilo.carastan-dos-santos@inria.fr, {msserpa, navaux}@inf.ufrgs.br,

Keywords—*Prefetching, memory system, High-performance computing.*

I. EXTENDED ABSTRACT

In recent years, there have been significant advances in the performance of processors, exemplified by the reduction of transistor size and the increase in the number of cores in a processor. Conversely, the memory subsystem did not advance as significantly as processors, not being able to deliver data at the required rate, and creating what is known as the memory wall [1]. An example of a technology used to mitigate the memory latency is the prefetcher, a technique that identifies access patterns from each core, creates speculative memory requests, and fetches data that can be potentially useful to the cache beforehand.

In High-Performance Computing (HPC) systems, many other problems arise with parallelism. Since HPC applications are highly parallel, with many threads communicating with one another mainly through shared memory, it becomes necessary to keep data coherence in the several cache levels. Moreover, the memory interactions among different threads may also unpredictably change the data path through the memory hierarchy. When considering the memory hierarchy complexity along with prefetcher action, the behavior of the processor's memory subsystem reaches a new level of complexity.

In this work, we seek to shed light on how the prefetcher affects the processing performance of parallel HPC applications, and how accurately state-of-the-art multicore architecture simulators are simulating the execution of such applications, with and without prefetcher. We identify that an L2 cache prefetcher is more efficient in comparison with an L1 prefetcher, since avoiding excessive L3 cache accesses better contributes to performance, when comparing to accessing the L2 cache. Moreover, we show evidence that the prefetchers' contribution to performance is limited by the memory contention that emerges when the level of parallelism increases.

A. Methodology and Experimental Environment

We developed a careful experimental campaign, executing the Numerical Aerodynamic Simulation Parallel Benchmark (NPB) [2] using different levels of parallelism. Hardware counter information was collected using the command line
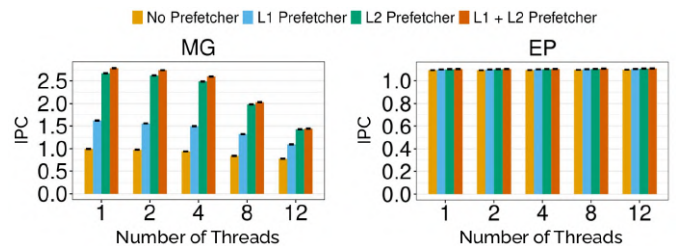

Fig. 1: IPC results for the real execution of the NPB applications with input class A.

tool PAPI [3], on an environment composed of an Intel Xeon Silver 4116 CPU, the Skylake microarchitecture [4]. The state-of-the-art simulators used were ZSim [5] and Sniper [6], configured on an approximation of the real machine. The conducted experiments were based on combinations of all the available prefetchers on both the real hardware and Sniper. A reproducible and open methodology was applied in our investigation, and all the materials are publicly available[1].

B. Investigating Current Architecture Prefetchers

Figure 1 shows the instructions per cycle (IPC) for two NPB applications, MG and EP, running on the real hardware, considering different numbers of threads and prefetchers, and with prefetcher disabled as well. The behavior observed for MG is representative of the general behavior observed for the remaining applications, with the exception of the EP, which is known to not make much use of memory [2], therefore processor stalls due to memory access latency rarely occur during execution.

In general, we observe that the difference in performance from the configuration with only the L2 prefetchers to the configuration with all prefetchers enabled is not much large, with an average of 3.3% of IPC improvement. On Skylake, the L2 cache access latency is of 14 cycles, only 10 cycles higher than the L1 cache, while the L3 latency is measured to be approximately between 60 and 80 processor cycles, presenting a much higher overhead and a higher probability of stalling the processor execution [7]. Therefore, the main performance gains are obtained by the L2 prefetcher and the associated access streams it detects, avoiding long latency accesses to the L3 cache.

---

[1]Work done while being at the Informatics Institute, Federal University of Rio Grande do Sul (UFRGS) - Porto Alegre, Brazil

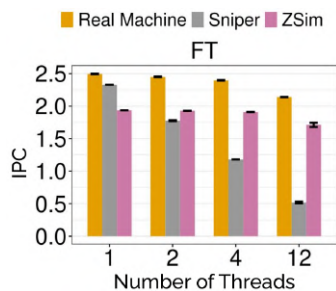[1]https://gitlab.com/msserpa/prefetcher-ccpe

Fig. 2: IPCs with prefetcher disabled on ZSim and Sniper simulations and on the real execution.
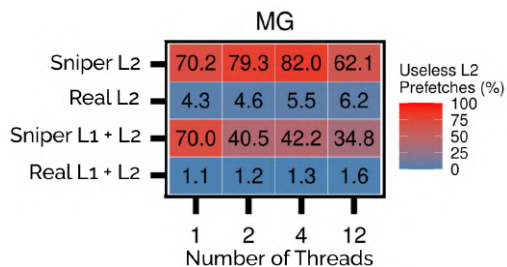

Fig. 3: Useless prefeteches performed by the simulation and by the real hardware, in ratio of the total number of prefetches.

The MG example in Figure 1 also shows a decrease in the IPC when we increase the number of threads. Parallel applications usually perform inter-thread communication that naturally increase in function of the total number of threads. With a large number of threads, the application memory accesses and the inter-thread communication increase, generating contention that hinders the prefetcher's impact over the performance. Therefore, in cases where the number of concurrent threads executed in a processor is very high, the performance benefit of memory prefetchers for applications that use intense memory and inter-thread communications would be negligible.

C. Investigating Prefetchers on Simulation

Figure 2 shows an example of the obtained IPCs when no prefetcher is used for the simulations and the real execution. For applications where communication and contention are more predominant, ZSim simulations can reach a 30% lower IPC when compared to the real execution, while Sniper can provide simulations with an IPC 426% lower. As a general trend, for NPB applications with communication and contention, ZSim tends to underestimate the contention effects, while Sniper tends to overestimate the contention effects as the number of threads in the simulation increases.

When comparing Sniper simulation with prefetchers enabled to the real execution with prefetchers as well, even though the number of prefetch requests issued is quite similar, it does not translate in similar estimations of performance. This may be due to the fact that the prefetches performed by Sniper are different in terms of usefullness, as demonstrated in Figure 3, where we show the mean percentage of prefetches that were not useful during the executions in the real machine and in the Sniper simulation.

D. Conclusions

The obscurity and confidentiality around the real implementation makes accurate prefetching models and algorithms impossible to be reproduced in simulators. This is exemplified by Skylake's non-inclusive L3 cache, which increases the memory system complexity and hinders the accurate simulation of NPB behavior with the Sniper's prefetchers.

On the real machine, we could observe that the use of both L1 and L2 prefetchers does not necessarily warrant significant performance gains, which is not intuitive. With only the L2 prefetcher, we obtained performance gains similar to when using both prefetchers, with the advantages of having more control over the experiments, faster simulation time, and less energy consumption due to the smaller number of prefetch requests being performed. Furthermore, as the parallelism and the inter-thread communication increase, so does the memory contention, becoming a large constraint to the performance.

II. ACKNOWLEDGMENT

REFERENCES

[1] W. Wulf and S. McKee, "Hitting the memory wall: Implications of the obvious," *Computer Architecture News*, vol. 23, 01 1996.

[2] H. Jin *et al.*, "The openmp implementation of nas parallel benchmarks and its performance," NAS System Division, NASA Ames Research Center, Tech. Rep., 1999.

[3] D. Terpstra *et al.*, "Collecting performance data with papi-c." Tools for High Performance Computing. Springer, 2010, pp. 157–173.

[4] J. Doweck *et al.*, "Inside 6th-generation intel core: New microarchitecture code-named skylake," *IEEE Micro*, vol. 37, no. 2, p. 52–62, Mar. 2017. [Online]. Available: https://doi.org/10.1109/MM.2017.38

[5] D. Sanchez and C. Kozyrakis, "Zsim: Fast and accurate microarchitectural simulation of thousand-core systems," *SIGARCH Comput. Archit. News*, vol. 41, no. 3, p. 475–486, Jun. 2013. [Online]. Available: https://doi.org/10.1145/2508148.2485963

[6] T. E. Carlson *et al.*, "An evaluation of high-level mechanistic core models," *ACM Transactions on Architecture and Code Optimization (TACO)*, 2014.

[7] M. A. Z. Alves *et al.*, "Sinuca: A validated micro-architecture simulator," in *2015 IEEE 17th International Conference on High Performance Computing and Communications*, M. Qiu, Ed. IEEE, 2015, pp. 605–610.

[8] V. S. Girelli *et al.*, "Investigating memory prefetcher performance over parallel applications: From real to simulated," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 18, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6207

Valéria S. Girelli is a Research Engineer at the Memory Systems group of the Barcelona Supercomputing Center (BSC), Spain. She received her BSc degree in Computer Science from the Federal University of Rio Grande do Sul (UFRGS), Brazil in 2021. Previously, she was a Computer Architecure Undergraduate Researcher at UFRGS, and a Researcher in Natural Language Processing at Universidade do Vale dos Sinos and Dell Brasil.

[2]http://gppd-hpc.inf.ufrgs.br/

Hyperparameter Optimization using Agents for Large Scale Machine Learning

Pere Vergés Boncompte*†‡ Vladimir Vlassov‡, Rosa M. Badia*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
‡KTH Royal Institute of Technology, Stockholm, Sweden
E-mail: {pere.verges, rosa.m.badia}@bsc.es, vladv@kth.es

Keywords—Machine Learning, Scalable Hyperparameter Search, Distributed Systems, High-performance computing, Task-based Workflow.

I. EXTENDED ABSTRACT

Machine learning (ML) has become an essential tool for humans to get rational predictions in different aspects of their lives. Hyperparameter algorithms are a tool for creating better ML models. The hyperparameter algorithms are an iterative execution of trial sets. Usually, the trials tend to have a different execution time.

In this paper we are optimizing the grid and random search with cross-validation from the Dislib [1] an ML library for distributed computing built on top of PyCOMPSs[2] programming model, inspired by the Maggy [3], an open-source framework based on Spark. This optimization will use agents and avoid the trials to wait for each other, achieving a speed-up of over x2.5 compared to the previous implementation.

A. Background

1) COMPSs and PyCOMPSs: The COMP Superscalar (COMPSs) is a task-based programming model that provides a programming interface for developing applications and a runtime system that exploits the inherent parallelism of applications at execution time. The main features of COMPSs are: sequential programming, programmers do not need to deal with parallel and distributed paradigms, agnostic of the computing infrastructure, offers single memory and storage space. COMPSs is a programming environment for developing complex workflows for parallelization. At runtime generates a task-dependency graph that encodes the existing parallelism of the application workflow.

2) Dislib: The Distributed Computing Library (dislib) is a machine learning library built on top of PyCOMPSs. This library has a very similar interface to scikit-learn. Not only, it can scale to large data, but also it is easy to use in high-performance computing clusters. These features make it easy to use for non-experts.

The main data structure of Dislib is the distributed arrays (ds-array). The ds-array is a matrix divided into blocks that will be stored remotely. All operations performed on the ds-arrays are parallelized using PyCOMPSs. The degree of parallelization is controlled by the size of the blocks in which the ds-array is being split.

It has implemented several ML algorithms, that range from Classification, Clustering, Regression, Decomposition, Pre-Processing, Neighbouring, and Model Selection, including the Grid and Random Search.

3) Agents and Nesting: The agent implementation in COMPSs [4] aims at enabling the offload the execution of functions of the embedded host to other nodes on Cloud-Edge Continuum. The application will request the execution of functions to the runtime. To do so, will show the execution logic, dependencies, resource requirements, data locations. After having received the request, it will invoke the runtime system. This runtime will handle the task execution asynchronously from the resource pool.

Having the agent implementation allows us to have nested tasks. Therefore, we can make a task invoke new tasks. This means that we have a parent task that invokes children. The parent will have to wait for the child to finish before finishing itself.

4) Hyperparameter Algorithms: Several algorithms are being used for hyperparameter [5] tuning, and also there are different approaches for doing parallel hyperparameter optimization that one can take [6]. However, we will focus on the ones implemented in the Dislib ML library. The two algorithms that we have are: Grid Search with Cross-Validation (CV), this algorithm evaluates exhaustively all the combinations of all parameters that have been given to it, and Random Search CV, in this case only a random subset of all the possible combinations will be executed. The core code of both algorithms is shared. We will refer to it as Base Search CV. The Base Search CV is in charge of the trial execution, which consists of a loop that will launch all the trials that have to be evaluated.

TABLE I. EXECUTION TIME OF THE GRID SEARCH CV WITH A DIFFERENT NUMBER OF TRIALS COMPARING THE INITIAL VERSION WITH THE NEW VERSION.

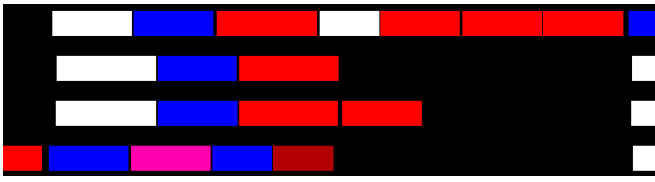| Number of trials	Initial Version time(s)	New Version time(s)
4	23.12	34.66
9	40.90	39.80
16	76.53	40.51
25	192.65	72.02
36	244.69	126.34

Fig. 1. Task view trace showing one iteration of the Grid Search CV algorithm, using the initial implementation and four threads.
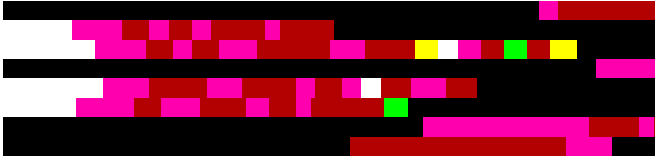

Fig. 2. Task view trace showing one iteration of the Grid Search CV algorithm, using the new implementation and four threads

B. Implementation

1) Initial Version: The initial version had a problem which was a synchronization at the end of the trails. The threads had to wait for each other, hence making them idle, which wasted time and resources. In Figure 1, we see one iteration of the Grid Search algorithm. We can observe four threads that are executing the trials. We see how the three last threads are waiting for the first one to finish its trial to start the following iteration.

2) New Version: The new version overcomes this limitation by using nested tasks. In this version, we are creating a task for each trial. Each trial will be able to invoke a new task with the fit and score tasks of the model being evaluated. This implementation will avoid the synchronization between iterations. Hence, we will optimize the resource and time utilization. In Figure 2, we have the first iteration of the Grid Search CV execution using the new version. It is a bit hard to visualize since we have that every new task will be scheduled in a new virtual thread (but at any point in time, we only have four). However, we can perfectly see that there are no threads idle waiting for other threads to finish.

C. Execution Environment

The evaluation of this implementation is done in the MareNostrum 4 Supercomputer. The configuration selected has been two nodes, where one acts as a master and the other as a worker. Each worker uses 48 CPUs.

D. Results

We have executed the Grid Search CV algorithm for the Cascade Support Vector Machines model. We have done several runs with a different number of trials to execute by the algorithm. In Figure 3 we see that for a small number of trials the new version does not improve. However, once we start to have a bigger number of trials the new versions start performing better, reaching a speedup of over x2.5. In Table I, we reflect the exact times for both executions.

E. Conclusion and Future Work

In this study, we have been able to demonstrate that the use of nesting has been successful at solving the problem
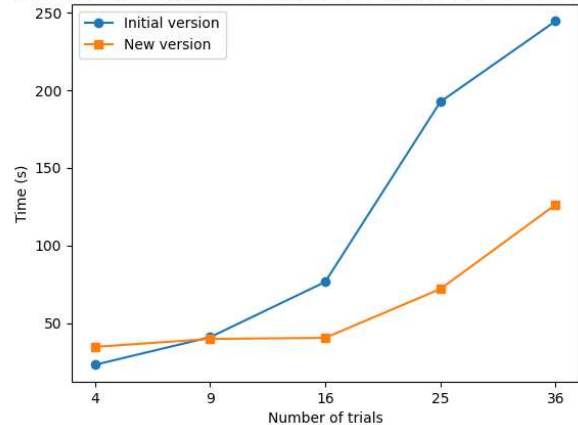

Fig. 3. Execution time of the Grid Search CV with different number of trials comparing the Initial Version against the new Version.

of parallel trial execution on the grid and random search algorithms. The solution has been able to achieve a speedup over x2.5. This speedup is due to avoiding the synchronization that was present in the initial version. This allowed us to take advantage of the time and resources previously being wasted.

As for future work, we have also started and implementation of early stopping criteria, using COMPSsExceptions [7] to emulate a wait for any call, which will also reduce the time execution and resource utilization even further.

REFERENCES

[1] J. Álvarez Cid-Fuentes *et al.*, "dislib: Large Scale High Performance Machine Learning in Python," in *Proceedings of the 15th International Conference on eScience*, 2019, pp. 96–105.

[2] e. A. Lordan, F., "ServiceSs: an interoperable programming framework for the Cloud," in *Journal of Grid Computing*, 2014, p. 67–91.

[3] M. Meister, "Maggy: open-source asynchronous distributed hyperparameter optimization based on apache spark," in *FOSDEM 20*, 2019.

[4] D. L. F. Lordan and R. M. Badia, "Colony: Parallel Functions as a Service on the Cloud-Edge Continuum," in *Euro-Par 2021*, 2021.

[5] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *CoRR*, vol. abs/2003.05689, 2020. [Online]. Available: https://arxiv.org/abs/2003.05689

[6] L. Li *et al.*, "Massively parallel hyperparameter tuning," *CoRR*, vol. abs/1810.05934, 2018. [Online]. Available: http://arxiv.org/abs/1810.05934

[7] e. A. Ejarque J, "Managing Failures in Task-Based Parallel Workflows in Distributed Computing Environments," in *Euro-Par 2020: Parallel Processing*, 2020.

Pere Vergés Boncompte received his BSc degree in Computer Engineering from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain in 2020, coursing one of the semesters at the University of Edinburgh. After that, he started his MSc degree in Innovation and Research in Informatics in Advanced Computing, at UPC and also started working at Workflows and Distributed System group of Barcelona Supercomputing Center (BSC), where also coursed one semester at KTH Royal Institute of Technology, Stockholm, Sweden and now is finishing his Master Thesis at University of California Irvine (UCI).