SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2012-2.3.1 – Third Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-3IP

PRACE Third Phase Implementation Project

Grant Agreement Number: RI-312763

D7.1.1
Applications Addressing Major Socio-economic Challenges

Final

Version: 1.0
Author(s): Maciej Szpindler, ICM
Marcin Zieliński, SURFsara
Date: 25.05.2013

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-312763**	
	Project Title: PRACE Third Phase Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D7.1.1**>	
	Deliverable Nature: < Report >	
	Deliverable Level: PU *	**Contractual Date of Delivery:** 31 / 05 / 2013
		Actual Date of Delivery: 31 / 05 / 2013
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

Document	**Title: Applications Addressing Major Socio-economic Challenges**	
	ID: D7.1.1	
	Version: <1.07>	**Status: Final**
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D7.1.1.docx	
Authorship	**Written by:**	Maciej Szpindler, ICM Marcin Zieliński, SURFsara

	Contributors:	Lucian Anton, STFC Cevdet Aykanat, BILKENT Eva Casoni, BSC Jacques David, CEA Gunduz V. Demirci, BILKENT John Donners, SURFsara Andrew Emerson, CINECA David Emerson, STFC Hans Eide, UiO Georgios Fanourgakis, CASTORC Massimiliano Guarrasi, CINECA Guillaume Houzeaux, BSC Nevena Ilieva, NCSA Maria Francesca Iozzi,UiO Mohammad Jowkar, BSC Tomáš Karásek, VSB Klaus Klingmueller, CASTORC Soon-Heum Ko, LiU Peter Michielse, SURFsara Charles Moulinec, STFC Martin Plummer, STFC Thomas Ponweiser, JKU Thomas Röblitz, UiO Ole Widar Saastad, UiO Katerina Michalickova, UiO Nikolay Aleksandrov Vazov, UiO Georgios Magklaras, UiO Alan Simpson, EPCC Peter Stadelmeyer, JKU Andrew Sunderland, STFC Ilian Todorov, STFC Ata Turk, BILKENT
	Reviewed by:	Leon Kos, ULFME Dietmar Erwin, PMO
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	29/04/2013	Draft	Initial skeleton version.
0.2	05/05/2013	Draft	All reports added. Formatting. Small missing-texts additions.
0.3	08/05/2013	Draft	Missing section added. Further formatting unification. Contributors list added.
0.4	17/05/2013	Draft	Corrected version after the internal review.
0.5	20/05/2013	Draft	Revised by Mark Bull, EPCC.
0.6	21/05/2013	Draft	Additional text corrections and additions. Formatting fixes.
0.7	24/05/2013	Draft	Further formatting fixes. Last text additions.
1.0	25/05/2013	Final version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Socio-economic applications, Application Support

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] http://www.prace-ri.eu/

[2] The Scientific Case for HPC in Europe 2012 - 2020, http://www.prace-ri.eu/PRACE-The-Scientific-Case-for-HPC

[3] European Exascale Software Initiative (EESI), http://www.eesi-project.eu

[4] Scalable Software Services for Life Sciences (ScalaLife), http://www.scalalife.eu

[5] Big data: The next frontier for innovation, competition, and productivity; MGI Report; http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[6] The Apache Software Foundation, Welcome to Apache Hadoop! Retrieved May 6, 2013 from Apache Hadoop http://hadoop.apache.org

[7] Szpindler M., Zieliński M., "Identification of Applications Associated with Socio-economic Challenges" PRACE whitepaper, (2013) http://www.prace-ri.eu/IMG/pdf/wp65_identification_of_applications_associated_with_socio-economic_challenges.pdf

[8] PRACE-3IP WP7 Task 7.1.C Internal Wiki page https://prace-wiki.fz-juelich.de/bin/view/Prace3IP/WP7/Task71C

[9] Telemac-Mascaret Suite, http://www.opentelemac.org

[10] Dalton Quantum Chemistry Program Suite, http://www.daltonprogram.org

[11] Simen Reine, T. K., Petascaling and Performance Analysis of DALTON on Differen Platforms, University of Oslo, PRACE whitepaper, (2012)

[12] DL_POLY Molecular Dynamics Simulation Software, http://www.ccp5.ac.uk/DL_POLY/

[13] FastQC: a quality control tool for high throughput sequence data http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[14] Cutadapt: a tool that removes adapter sequences from DNA sequencing reads https://code.google.com/p/cutadapt

[15] Martin M., Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet Journal, 17(1): 10-12, (2011)

[16] Burrows-Wheeler Aligner, http://bio-bwa.sourceforge.net

[17] mapDamage: Fast approximate Bayesian estimates of ancient DNA damage parameters http://geogenetics.ku.dk/publications/mapdamage

[18] Ginolhac A., Rasmussen M., Gilbert M.T., Willerslev E., Orlando L., mapDamage: testing for damage patterns in ancient DNA sequences, Bioinformatics, 27(15): 2153-5, (2011)

[19] OpenFOAM, http://www.openfoam.com

[20] Elmer Multiphysical Simulation Software, http://www.csc.fi/english/pages/elmer

[21] Culpo, M., Current Bottlenecks in the Scalability of OpenFOAM on Massively Parallel Clusters, CINECA, PRACE whitepaper, (2012)

[22] Lindi, B., I/O-profiling with Darshan, Norwegian University of Science and Technology (NTNU), PRACE whitepaper, (2012)

[23] Jeffrey Dean, S. G., MapReduce: Simplified data processing on large clusters, Communications of the ACM , 51 (1), 107-113, (2008)

[24] Steve Plimpton, K. D. , MapReduce in MPI for Large-Scale Graph Algorithms, Parallel Computing , 37 (9), 610-632, (2011)

[25] Leskovec, J., Stanford Large Network Dataset Collection. Retrieved May 6, 2013 from http://snap.stanford.edu/data

[26] Modular Earth Submodel System, http://www.messy-interface.org

[27] Sandu, A., KPP – the Kinetic PreProcessor. Retrieved May 7, 2013 from http://people.cs.vt.edu/~asandu/Software/Kpp/

[28] ParaTools Inc., Kppa – ParaTools, Inc. Retrieved May 7, 2013 from http://www.paratools.com/Kppa#Kppa-1

[29] J. C. Linford et. al., IEEE Trans. Parallel and Distributed Systems, 22, 119 (2011)

[30] Delft3d Modeling Suite, http://oss.deltares.nl/web/opendelft3d

[31] SWAN (Simulating Waves Nearshore) Model, http://swanmodel.sourceforge.net

[32] URANIE: The CEA/DEN Uncertainty and Sensitivity platform, Fabrice Gaudier, Sixth International Conference on Sensitivity Analysis of Model Output, Elsevier , Procedia - Social and Behavioral Sciences Volume 2, Issue 6, p. 7660–7661, (2010), http://www.sciencedirect.com/science/article/pii/S1877042810013078

[33] The Uranie Platform, F.Gaudier, 28th ORAP Forum, http://www.irisa.fr/orap/Forums/Forum28/F28_Presentation/F28_Gaudier_UraniePlatform.pdf

[34] Modelling of the Uncertainty of Nuclear Fuel Thermal Behaviour Using the URANIE Framework, Porto, Portugal, September 20-September 25 ISBN: 978-0-7695-3773-3

[35] Uranie salient features, V.Bergeaud & alii, 7th FP framework, NURISP project, final seminar http://www.nurisp.eu/www/nurisp/final_seminar/SP4_Uranie.pdf

[36] A. Dora, L. Bryjko, T. van Mourik, and J. Tennyson , J. Chem. Phys. 146 (2012) 024324

[37] A. Dora, L. Bryjko, T. van Mourik, and J. Tennyson, J. Phys. B 45 (2012) 175203

[38] J. M. Carr, P. G. Galiatsatos, J. D. Gorfinkiel, A. G. Harvey, M. A. Lysaght, D. Madden, Z. Mašín, M. Plummer, J. Tennyson, and H. N. Varambhia, Eur. Phys. J. D 66, 58 (2012)

[39] A. G. Sunderland, C. J. Noble, V. M. Burke and P. G. Burke Comp. Phys. Commun. 145 (2002) 311.

[40] A. G. Sunderland, C. J. Noble and M. Plummer, 'Future Proof Parallelism for Electron-Atom Scattering Codes on the XT4', NAG HECToR dCSE project report (2010) http://www.hector.ac.uk/cse/distributedcse/reports/prmat/prmat.pdf

[41] A. G. Sunderland, M. Plummer, M. A. Lysaght and M. Pettipher, The PFARM Wiki, (PRACE 2013) http://hpcforge.org/plugins/mediawiki/wiki/pfarm/index.php/Main_Page

[42] Alya Red - HPC-based Computational Biomechanics for Supercomputers, http://www.bsc.es/computer-applications/alya-red-ccm

[43] PLUTO: A Modular code for computational astrophysics, http://plutocode.ph.unito.it/

[44] A G Sunderland, C. J. Noble, V. M. Burke and P. G. Burke, CPC 145, 311-340, (2002)

[45] J. M. Carr, P. G. Galiatsatos, J. D. Gorfinkiel, A. G. Harvey, M. A. Lysaght, D. Madden, Z. Mašín, M. Plummer, J. Tennyson, and H. N. Varambhia, Eur. Phys. J. D 66, 58, (2012)

[46] Parallel Flexible Asymptotic R-matrix Package (PFARM), http://hpcforge.org/plugins/mediawiki/wiki/pfarm/index.php/Main_Page

[47] Sequence Alignment/Map Tools, http://samtools.sourceforge.net

[48] MUSCLE Multiple Alignment Program, http://drive5.com/muscle

[49] CLUSTALW Multiple Sequence Alignment Program, http://www.ebi.ac.uk/Tools/msa/clustalw2

[50] RAxML Evolutionary Placement Algorithm, http://www.exelixis-lab.org

[51] The R Project for Statistical Computing, http://www.r-project.org

[52] libMesh – C++ Finite Element Library, http://libmesh.sourceforge.net

[53] Netgen Mesh Generator, http://sourceforge.net/apps/mediawiki/netgen-mesher

[54] Pegasus Parallel Graph Mining Library, http://www.cs.cmu.edu/~pegasus

[55] Mahout Parallel Machine Learning Library, http://mahout.apache.org

[56] NEST (NEural Simulation Tool), http://www.nest-initiative.uni-freiburg.de/

[57] ECHAM Atmospheric Circulation Model, http://www.mpimet.mpg.de/en/science/models/echam.html

[58] P. Jöckel et al., Geosci. Model Dev., 3, 717-752, (2010)

List of Acronyms and Abbreviations

ADMM	Alternating direction method of multipliers
AGBNP2	Analytical Generalised Born plus Non Polar
AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
AMR	Adaptive Mesh Refinement
APSP	All Pairs Shortest Path
APSS	All Pair Similarity Search
BCSGPS	Bi-Conjugate Stabilized Gradient Poisson Solver
BSC	Barcelona Supercomputing Center (Spain)
CaSToRC	Computation-based Science and Technology Research Center (Cyprus)
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CFD	Computational fluid dynamics
CGPS	Conjugate Gradient Poisson Solver
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CPU	Central Processing Unit
CSC	Finnish IT Centre for Science (Finland)
CSM	Computational structural mechanics
CUDA	Compute Unified Device Architecture (NVIDIA)
DECI	Distributed European Computing Initiative is an European single-project HPC access scheme supported by PRACE projects
DFT	Density functional theory
DIRAC	Code that computes molecular properties using relativistic quantum chemical methods
DNA	DeoxyriboNucleic Acid
DP	Double Precision, usually 64-bit floating point numbers
EC	European Community
EESI	European Exascale Software Initiative
EMAC	ECHAM/MESSy Atmospheric Chemistry
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
FMM	Fast multiple method
FSI	Fluid-structure interactions
GB	Giga (= 2^{30} ~ 10^{9}) Bytes (= 8 bits), also GByte
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
GUI	Graphical User Interface
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
IBM	Formerly known as International Business Machines
ICHEC	Irish Centre for High-End Computing (Ireland)
ICM	Interdisciplinary Centre for Mathematical and Computational Modelling (Warsaw, Poland)
IEEE	Institute of Electrical and Electronic Engineers
IGBSPE	Iterative Grid-Based Solver of the Poisson Equation
ITER	International Thermonuclear Experimental Reactor (Cadarache, France)

I/O	Input/Output
JET	Joint European Torus (Culham, United Kingdom)
JKU	Johannes Kepler University (Linz, Austria)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MECCA	Module Efficiently Calculating the Chemistry of the Atmosphere
MESSy	Modular Earth Submodel System
MD	Molecular dynamics
MHD	Magnetohydrodynamics
MPI	Message Passing Interface
MPMD	Multiple Program Multiple Data
NCSA	National Centre for Supercomputing Applications (Sofia, Bulgaria)
OS	Operating System
PDE	Partial Differential Equation
PGI	Portland Group, Inc.
PM	Person Month; unit to measure effort of work done / to be done
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PR	PRoject identifying a socio-economic challenge
QP	Queued Project
SAN	Storage Area Network
SARA	Stichting Academisch Rekencentrum Amsterdam (The Netherlands); SURFsara from January 2013
SSC	PRACE Scientific Steering Committee
SNIC	Swedish National Infrastructure for Computing (Sweden)
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
SURFsara	Dutch national High Performance Computing & e-Science Support Center (The Netherlands)
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
Tier-1	See Tier-0
UiO	University of Oslo (Norway)
VSB	Technical University of Ostrava (Czech Republic)
WP7-3IP	Work Package 7 “Application Enabling and Support” of PRACE-3IP

Executive Summary

The PRACE project[1] provides continuous support for European HPC applications. One of the objectives of the PRACE-3IP Work Package 7 'Application Enabling and Support' is to extend the support with a focus on key community codes that address major socio-economic challenges.

This deliverable contains a report on major socio-economic challenges that are addressed with a selected set of community codes. These challenges are focused on problems that have substantial impact on our lives and environment or on areas that drive Europe's economic competitiveness. The list of identified socio-economic challenges includes:

- *Safe and Environmental-friendly energy production*.
- *Rational drug design*.
- *Future aircraft transportation*.
- *Sustainable food supply*.
- *'Big data' management and processing*.
- *Multiscale modelling of the human cells and organs.*
- *Understanding of climate change*.
- *Natural environment protection*.

A set of European and other community codes has been carefully selected to address challenges in the identified areas. For each code, this Task will either enable the application to use HPC systems, or improve its performance. The application codes have been chosen with the support of the PRACE Scientific Committee, assuring the independence and quality of the selection. The process used for selection of these applications is described in the report.

For each of the identified socio-economic challenges at least one enabling project has been defined, including its objectives, work scope, current state and the expected improvement of a given application or a set of applications. The objectives of these projects include code porting to selected Tier-0 architectures, tuning specific computational kernels to enable simulations at a desired scale, optimization of an application workflows and multi-application environment couplings.

The enabling support of the selected codes associated with socio-economic challenges has already started. This deliverable presents an initial progress report for each application.

1 Introduction

PRACE provides permanent support for application codes targeting Tier-0 and Tier-1 European HPC systems with a number of tasks and activities. The PRACE-3IP project continues and extends previous efforts in Work Package 7 'Application Enabling and Support' (WP7-3IP for short). Task 7.1 'Scaling and Optimization of Applications Codes' continues supporting projects applying for PRACE Preparatory Access and PRACE DECI. A new, third focus extends the support to community codes that address major socio-economic challenges. These challenges arise in area that has substantial impact on human life and environment, or that drive Europe's economic competitiveness.

The capabilities of HPC systems are constantly growing, and evolution from Petascale to Exascale is expected to be complete by the end of this decade [2]. HPC application codes have proven their usefulness in a wide range of research applications, and clear benefits to the communities have been identified. These scientific and engineering applications enable breakthrough research by simulating problems with increasing scale and precision. The growing performance of HPC systems and applications opens possibilities for tackling important socio-economic challenges by both simulation and data analysis at the desired scale. Addressing socio-economic problems using HPC systems is an important factor in maintaining Europe's competitiveness and innovation. Identification of the appropriate socio-economic challenges and associated applications is a first step in supporting these application codes and bringing the benefit to the user communities.

This deliverable explains the process used to identify key socio-economic challenges and associated community codes, followed by initial enabling of the selected applications. The socio-economic challenges identified include scientific and computational problems in the following areas: *Energy Sources and Management*, *Life Sciences and Medicine*, *Climate Change*, *Big Data*, *Environment Protection*, and *Engineering*. The document is intended for a public audience while focusing on community applications codes and effort supported by the PRACE-3IP project.

The remainder of this document is organised as follows: Section 2 covers the objectives, and the process used to identify the set of key socio-economic challenges and select related representative community application codes. Section 3 contains the initial report on the applications enabling, focusing on the projects addressing selected socio-economic challenges. Section 4 summarizes the report and outlines future work. The Appendix contains the timeline of the process, the selection criteria, and a complete list and details of the submitted proposals.

2 Identification of socio-economic challenges

This section describes identification of the major socio-economic challenges and associated applications to be supported and 'enabled' on European HPC systems. In the first step, a set of socio-economic challenges with their corresponding computational problems was identified. As a second step, associated application codes addressing these problems have been selected. It was also decided, that due to the available manpower, only a limited number of the key community applications will be supported. The timeline of the selection process is given in Appendix 5.1.

2.1 Socio-economic challenges and scientific domains

The *PRACE Scientific Case* [2] was used as the primary source of input and main reference for the selection of scientific and computational challenges. This is a natural choice, since the document contains an up-to-date summary of previous PRACE achievements, findings, future directions and requirements for HPC directed applications. Other input includes PRACE reports and deliverables, and results of other European projects, for example from EESI [3] and ScalaLife [4].

Using these inputs, and taking into consideration the available scientific and technical expertise of the PRACE-3IP partners, a set of socio-economic challenges have been identified, which are described in the following sections. For each challenge, the required applications enabling work is briefly outlined.

2.1.1 *Safe and Environmental-friendly energy production*

Energy production has a crucial impact on the economy and on human living. For existing sources of energy and for nuclear plants in particular, safety is one of the most essential aspects. A growing usage of renewable and environmental friendly sources is also one of the key topics for energy management.

There are a number of existing application codes that provide nuclear reactor safety analysis and that rely heavily on HPC. In order to improve safety and efficiency of the facilities, increased performance and flexibility of these application codes is required.

HPC applications are also widely used in the development of renewable energy sources, for example the placement of wind and marine turbines. For the optimal placement of off-shore energy-generating platforms, large scale and multi-discipline coupled simulations need to be enabled for large HPC systems.

2.1.2 *Rational drug design*

Rational computer-aided drug design is a safe and economic tool for science and industry. It also provides an alternative to the trial-and-error approach. Molecular modelling is a key tool for computer-aided drug design. Community application codes implement a variety of computational methods addressing different simulation scales including molecular dynamics and the more precise (but expensive) first-principles quantum chemistry.

In order to allowing larger experiments with longer timescales and higher precision, improved scalability of current application codes is necessary. This requires algorithmic improvements as well as the identification and removal of performance bottlenecks.

2.1.3 *Future aircraft transportation*

Development of future aircraft, “green” fuel-efficient airplanes requires virtual modelling with precise multidisciplinary interactions, and with reduced time costs. In order to further improve existing designs (especially wing designs in combination with high-lift devices) it is necessary to have high-fidelity simulations available that couple computational fluid dynamics (CFD) and computational structural mechanics (CSM).

2.1.4 *Sustainable food supply*

Food production is one of the key economic challenges. Establishing and monitoring genetic diversity of species, for example fish stocks in relation to fishing pressure and aquaculture, are of high importance for many countries.

Genomic research for studying such problems involves massive amounts of data in application workflows that require large computing resources. Development of workflows for community codes that would be applicable to large HPC systems would address the underlying computational challenge.

2.1.5 *'Big data' management and processing*

The ever increasing amount of data produced and stored by a number of social activities and industrial applications is becoming a real challenge. The processing and analysis of these data sets is no longer possible with standard tools. One of the novel research areas with a wide applicability to social sciences is large datasets analysis ("Big Data") [5]. The methodology used includes not only the data-driven approach but new tools and programming paradigms[6].

HPC provides unique processing capabilities for data analysis. It needs to be investigated whether the "Big Data" approach and programming techniques are applicable for HPC codes.

2.1.6 *Multiscale modelling of the human cells and organs*

Recent achievements in Life Sciences and Computer-aided Medicine provide detailed information on structures and processes at very different scales. The challenge is to integrate all this data into multiscale and multidiscipline models that will describe entire cells and even organs.

These challenges require development of Exascale systems and software to be fully addressed. Most application codes in use today cannot scale up to hundreds of thousands of processors [2]. Enabling of the selected applications for improved scalability and efficient use of Petascale systems is the first step to Exascale.

2.1.7 *Understanding of climate change*

Atmospheric chemistry has a huge socio-economic impact, affecting, for example climate change, human health and agriculture. Simulations help to identify unfavourable trends and assist adaptation to changing environmental conditions.

Atmospheric chemistry processes are one of the most computationally intensive tasks and higher accuracy modelling requires existing computational kernels to be enabled for improved scalability and acceleration.

2.1.8 *Natural environment protection*

One of the important aspects in environment protection for a number of regions in Europe is preservation, development and sustainability of water reservoirs. One of the examples is the design process of a lake and its environment: this needs an interactive approach in which the economic, engineering, recreational, flooding, and ecology aspects from different stakeholders can be combined. To address the challenge a set of applications need to be

coupled and enabled for HPC processing. Existing community application codes should also be optimized for improved runtime.

2.2 Selection process for applications with socio-economic relevance

2.2.1 *Planned structure of work*

Based on the number of the socio-economic challenges identified, the number of scientific aspects that could be addressed within those challenges, and on feedback from contributing partners, it was decided to manage the work on a project basis.

Each of these scientific aspects is a challenging and complex computational objective to solve by means of specialised software application or tool. Therefore the projects will either concentrate on improving the performance of a specific application code, or on bringing additional, HPC expertise to help solve a relevant part of that scientific problem.

2.2.2 *Selection criteria*

The projects and the corresponding specialised software applications and tools, should fulfil certain criteria. Those criteria were carefully discussed and agreed with all the contributing partners and further approved at the end of M4:

- The project's aim should conform to the PRACE Scientific Case [2] document for socio-economic significance.
- The project should introduce novelty with respect to the already on-going PRACE activities and other European projects.
- The project's corresponding software applications, tools or data should be freely available.
- The project should have clearly identified benefits for the community, and for related European software applications or tools.

2.2.3 *Call for projects*

To follow the pre-determined scheme of work, which would efficiently address the highlighted aspects of the socio-economic challenges, an internal call for projects proposals was initiated. Contributing PRACE-3IP partners submitted preliminary proposals for projects within specified guidelines (see Appendix 5.2). An internal presentation of those proposals was made during the face-to-face meeting at the end of M4, which resulted in constructive feedback. The presented proposals were adjusted to reflect the gathered feedback, before the call was closed in mid-M5.

2.2.4 *Proposed projects*

The internal call for proposals resulted in 15 submitted projects. The scientific aspects covered in the proposals were classified to the following socio-economic challenges:

- Safe and environmental-friendly energy production (4 proposals)
- Rational drug design (2)
- Future aircraft transportation (2)
- Sustainable food supply (1)
- Big Data" management and processing (1)
- Multiscale modelling of the human cells and organs (3)

- Understanding of climate change (1)
- Natural environment protection (1)

The full description of project proposals can be found in Appendix 5.3.

2.3 Evaluation and final selection

2.3.1 *Review by the Scientific Steering Committee*

In order to assess the transparency of the process, and the scientific significance and quality of the projects, the PRACE Scientific Steering Committee (SSC) was consulted. A one-day workshop with the WP7 leaders, task leaders and members from the SSC was held, in which the proper identification of socio-economic challenges and the adequacy of the selection criteria were assessed and confirmed. All fifteen projects from contributing PRACE partners were evaluated against the agreed selection criteria, and advice and recommendations on the final project selection were given. A ranking of the submitted projects (listed in Appendix 5.3) was created and is shown in Table 1.

A scale of four ranks was used, taking into account SSC members' assessments on conformance with a given socio-economic challenge, scientific relevance, applications choice, enabling practicability and importance for related community. The ranks give the following recommendations on projects acceptance:

A – overall good quality and conformance to the selection criteria;

B – minor concerns on criteria conformance or enabling practicability;

C – some criteria unmet;

D – not related to the socio-economic challenge.

Moreover for proposals receiving the A rank, the associated risk in enabling feasibility was discussed. A perceived risk is indicated by additional "-" (minus) sign.

Socio-economic challenge	Project Number	Contributor	Rank	Recommendation
Safe and Environmental-friendly energy production	PR1	CINECA	D	Reject
	PR2	STFC	A-	Approve
	PR3	STFC	C	Reject
	PR4	CEA	B	Re-evaluate
Rational drug design	PR5	LiU	A-	Approve
	PR7	NCSA	A	Approve
Sustainable food supply	PR8	UiO	A-	Approve
Future aircraft transportation	PR9	JKU	A	Approve
	PR10	VSB	D	Reject

Socio-economic challenge	Project Number	Contributor	Rank	Recommendation
'Big data' management and processing	PR11	BILKENT	A-	Approve
Multiscale modelling of the human cells and organs	PR6	STFC	B	Re-evaluate
	PR12	BSC	B	Re-evaluate
	PR13	ICM	A-	Approve
Understanding of climate change	PR14	CASTORC	A	Approve
Natural environment protection	PR15	SURFSARA	A	Approve

Table 1: Ranking of the projects addressing socio-economic challenges

2.3.2 *Final selection*

The feedback from the SSC members and the ranking was forwarded to the projects' authors, who were given the opportunity to respond to the evaluations. Since only a limited number of applications could be supported, a compromise between supporting all projects and assuring adequate effort levels has been found. The highest ranked projects have been marked for approval and the remaining positively evaluated projects formed a reserve list. As a result of the selection process a total of eight projects were selected from the original list of fifteen submitted projects. Additionally, a reserve list of four projects was selected. These projects will be re-evaluated after the initial phase of applications enabling (at M11).

Proposers of unsuccessful projects were invited to collaborate with the selected projects. The final list of approved projects is summarized in Table 2, and projects on the reserve list for re-evaluation are shown in Table 3.

Socio-economic challenge	Project Number	Contributors	Associated applications
Safe and Environmental-friendly energy production	PR2	STFC	Tomawac, Telemac-3D, Sisyphe (coupled)
Rational drug design	PR5	LiU	LSDALTON
	PR7	NCSA	DL_POLY_4
Sustainable food supply	PR8	UiO	FastQC, cutadapt, BWA, mapDamage (workflow)
Future aircraft transportation	PR9	JKU VSB	OpenFOAM, Elmer
'Big data' management and processing	PR11	BILKENT	Pegasus, Mahout
Understanding of climate change	PR14	CASTORC	EMAC/ ECHAM, MESSy (coupled)
Natural environment protection	PR15	SURFSARA CINECA	Delft3D, Delft3D-DELWAQ, Delft3D-SWAN (coupled)

Table 2: Final list of approved projects

Socio-economic challenge	Project Number	Contributors	Associated applications
Safe and Environmental-friendly energy production	PR4	CEA	URANIE
Multiscale modelling of the human cells and organs	PR6	STFC	PFARM
	PR12	BSC	Alya Red
	PR13	ICM	NEST

Table 3: Reserve list

The selection of the 12 proposals met the formal Milestone MS7.1 at Month 6. In addition, a White Paper "Identification of Applications Associated with Socio-economic Challenges"[7] was produced and published on the PRACE web site. Based on the SSC recommendations and the White Paper, the PRACE Management Board in its meeting on 5th of February 2013 approved the selection and decided that the four projects in the reserve list should be worked on without further re-evaluation.

2.4 PRACE Tier-0 resources committed

Following the guidelines for applying for access to the PRACE Tier-0 systems for the PRACE tasks, a Preparatory Access Type B (PA/B) proposal was submitted in the end of M5. The proposals followed a simplified reviewing process and were approved by mid-M7. Since it has been foreseen that the first stage of the enabling (M7-M11) will not make heavy use of the granted computing time per challenge, the application has been submitted in a form of a cumulative proposal, including all the projects from the final list.

Due to exhausted resources for the hybrid partition of CURIE and for HERMIT for the year 2012, the task has been granted separate budgets for the following Tier-0 systems:

- 250,000 core hours on FERMI (BlueGene/Q) at CINECA, Italy
- 250,000 core hours on JUQUEEN (BlueGene/Q) at Jülich, Germany
- 200,000 core hours on CURIE (x86-64) at CEA, France

The lack of a hybrid partition of CURIE for this first stage of the enabling period was unfortunate since one of the projects had to be stalled and another route undertaken. For those reasons, a separate Preparatory Access Type-B proposal requesting access to a hybrid partition of CURIE for the June cut-off has been submitted in M9.

A general guideline for partners on how to obtain access to the granted systems was circulated at the end of M7 and in M8 and put on the internal PRACE Wiki [8]. The resulting popularity of the systems in terms of signups was spread fairly evenly, with a slight preference for the JUQUEEN system.

The popularity of the JUQUEEN system is also very visible in terms of the used budget per system. The state of the budgets allocation and exploitation of each of the systems for M11 is as follows:

- CURIE: ~600 core hours used (0.3%)
- FERMI: ~30000 core hours used (12%)
- JUQUEEN: ~450000 core hours used (180%)

The initial enabling reports and the future plans and goals of the running projects, described in Section 3 suggest that the remaining computing time on the two systems will be exhausted in

the period of M12-M14. This will adequately overlap with the period of the subsequent PA/B proposal.

The overuse of 80% above the initial computing time on the JUQUEEN system is due to a project on Big Data. It is also foreseen that this project will start simulations and tests on CURIE as well. Hence, at this stage of the task, it will be determined whether a separate PA/B proposal for that challenge only should be put forward, to eliminate any future clashes. Thus, the next PA/B proposal will, most likely, cover only the other six projects.

3 Initial report on the applications enabling

This section contains initial reports on the enabling work performed during the period of M7-M11. Project progress was monitored during monthly teleconferences and evaluated during a dedicated face-to-face session in M10.

Enabling of the selected applications includes tasks such as code porting, tuning specific computational kernels, optimization of application runtime, scalability testing and improvement, testing and validating the results and multi-applications coupling. The aim of the enabling work is either to allow a particular application to run more effectively on the selected HPC system, or to allow solving larger problems, or to reduce resource consumption.

Contributors were asked to follow a set of reporting guidelines. The scope of the initial reports (Sections 3.1 to 0) includes: details on the supported applications, enabling objectives, actual work description and goals for the final enabling stage (for the period of M12-M24). The work description includes details of:

- identified issues on applications performance and scaling bottlenecks,
- preliminary measures for the improvement in performance and scaling,
- identified risks and potential problems,
- a description of work already done and achieved results in terms of a given measures.

For the projects in Sections 3.9 to 3.12 that have been started recently (M11) only the project abstract is provided and the applications enabling objectives defined.

3.1 PR2: Impact and optimum placement of off-shore energy generating platforms

The hydrodynamic suite used for this project is the TELEMAC-MASCARET system [9]. All TELEMAC freeware is distributed under the GNU General Public License (GPL), except for the BIEF library (Finite Element library) which is provided under the Lesser GNU General Public License (LGPL). The suite was initially developed by EDF R&D from 1987 on, but a consortium of 6 institutions (including EDF R&D and STFC Daresbury Laboratory) is now managing it since 2010. The modules used in this project are TOMAWAC to compute wave propagation, TELEMAC-3D for the hydrodynamics (solving the Navier-Stokes equations for free-surface flows) and SISYPHE for the sediment transport.

3.1.1 *Application enabling principles*

The socio-economic impact is twofold, first energy-related and, then ecology- and environment-related.

On the energy side the impact and optimum placement of off-shore energy generating platforms will be investigated, by simulating the placement of marine turbines in coastal waters and testing the fully-coupled performance of TELEMAC modules (TOMAWAC,

TELEMAC-3D, SISYPHE) at very large-scale. This has not been done before but is necessary to understand local (1-10 m scale) and distant (10-100 km scale) impact.

On the ecological and environmental side, it will consist of performing large-scale calculations to assess the near and long-term impact of marine turbine farms on local and distant coastal waters e.g. sediment morphology.

3.1.2 *Initial report on enabling*

The TELEMAC suite has successfully been ported on the STFC Blue Gene/Q (Blue Joule) and on the UK National Facility Cray XE6 (HECToR).

The whole coupling "TOMAWAC, TELEMAC-3D and SISYPHE" has never been tested in parallel at intended scale. At the start of the project, TOMAWAC was not running in parallel, but TELEMAC-3D and SISYPHE were. The version V6P2 of TOMAWAC is now running in parallel and scalability of TOMAWAC and TELEMAC-3D are presented in Figure 1. The test case has been obtained from University of Liverpool and concerns the flow in Liverpool Bay.

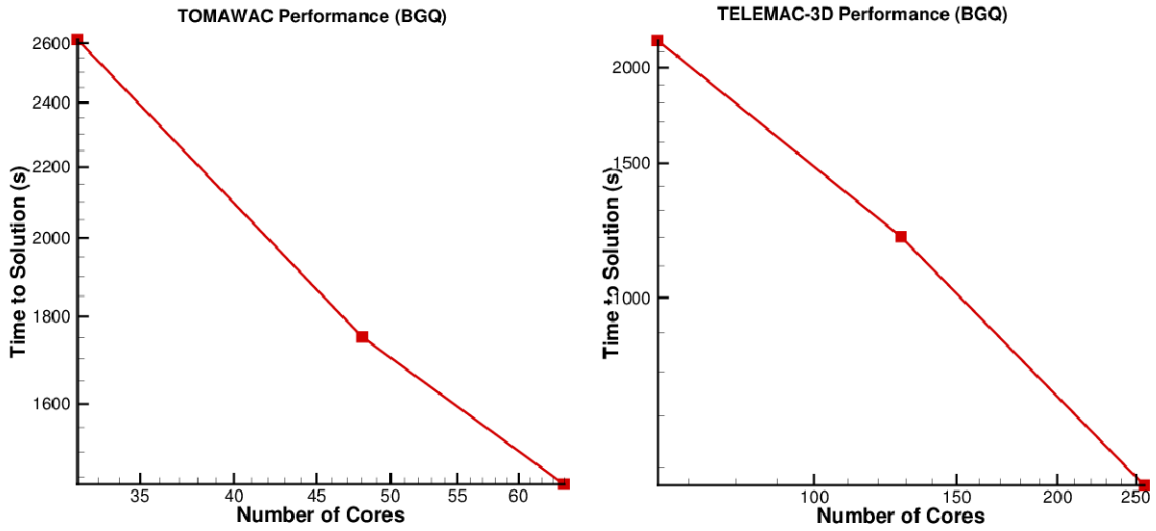

Figure 1: Performance on STFC Blue Gene/Q (BlueJoule). Left: Performance of TOMAWAC (0.6M element mesh). Right: Performance of TELEMAC-3D (0.6M element mesh).

Some memory issues are currently observed on a POWER7 cluster (XL compiler) and dealt with. This cluster is used to complete the debugging of the coupling between TOMAWAC and TELEMAC-3D, because the compiler is 'similar' to the compiler used on Blue Gene/Qs.

3.1.3 *Plans and goals for a final enabling stage*

The plans for the remaining part of the project are threefold:

- After completing the TOMAWAC - TELEMAC-3D coupling debugging, add the coupling with SISYPHE (sediment transport) to the TOMAWAC - TELEMAC-3D coupling for the current problem (Liverpool Bay).
- Generate a larger dataset (finer mesh), in order to have about 10k elements per MPI task.
- Test the full coupling TOMAWAC - TELEMAC-3D - SISYPHE on this larger dataset. This large case will be run on two Tier-0 machines: Fermi and JUQUEEN.

3.2 PR 5: Computer-aided drug designing

The advent of Tier-0 systems in Europe nowadays facilitates the ab-initio based evaluation of the protein-drug molecule interaction energy for drug design. The DALTON family of codes has been renowned as an accurate tool for the electronic-structure calculations, which results in a wide use in life-science community such as the ScalaLife Project [4] under FP7 program. The DALTON code [10] has been developed since 1983 and more than 60 developers (mainly Scandinavian research community members) contribute to improvements and new functionality. The code is distributed freely, under the personal or site license agreement.

Two major branches exist under DALTON code now: traditional DALTON and LSDALTON (Linear Scaling DALTON). This project uses the LSDALTON code, which is implemented under AO (Atomic Orbital) basis. This characteristic leads LSDALTON to provide better efficiency over the traditional DALTON code (which uses molecular orbital basis) on large-scale simulations. MPI-parallelism has been introduced to LSDALTON version through prior PRACE work[11], which enables LSDALTON to have good scalability for Density Functional Theory (DFT) calculations at Tier-0 scale.

3.2.1 *Application enabling principles*

The following issues will be resolved in the initial enabling phase:

- choice of LSDALTON or DALTON
- determination of technical targets for the final goal
- communication/partnership with DALTON developers and users

As has been discussed, the LSDALTON code was chosen for this project after intensive conversations with the DALTON developers (Dr. Simen Reine at University of Oslo; Dr. Dan Jonsson at University of Tromsø; Dr. Stefan Knecht at University of Southern Denmark). It is assumed that LSDALTON is better suited for large-scale ab-initio computation on Tier-0 systems, but the current MPI version still has a large potential for improvement (for example, reducing additional message overhead in the exact exchange approach; greatly reducing computational overheads in the integral-driver interface; scalability improvement for certain important approximations to DFT; fully memory-distributed integral-component constructions, etc.). The performance improvement of this MPI-parallelized LSDALTON will benefit many DALTON users in the life-science community. The improvements are being done in collaboration with the main DALTON developers Dr. Simen Reine and Dr. Dan Jonsson.

3.2.2 *Initial report on enabling*

Scalability runs were performed for two different molecular structures: Valinomycin (168 atoms) and Insulin (787 atoms). DFT calculations (BLYP/cc-pVDZ) with/without density fitting have been conducted. The resource used is a local x86-64 cluster, in which each node has a 16-core Sandy Bridge CPUs and 32 GB memory. The scalability tests have been performed up to 1024 cores. Since the architecture and the software stack of the current system are very similar to CURIE Tier-0 thin nodes, the same characteristics are expected to be found on the CURIE system.

The DFT approach with/without density fitting was applied for Valinomycin simulation and both cases converge successfully. In view of performance, the gain by density fitting is stronger when a smaller number of cores is used (583 seconds with density fitting, 8514

seconds without density fitting in case of 16 cores). This gap reduces as the number of processors increases (350 seconds vs. 889 seconds at 512 cores). The simulation time increases as the number of cores increases to 1024 (698 seconds vs. 1020 seconds). This result reveals that computational overheads become an issue at large number of cores.

In case of the insulin simulations, the large memory consumption becomes an important limitation. It was observed that a huge amount of memory is required in storing the `lstensor` 5-dimensional array: more than 3 GB per core for the regular calculation and even higher with density fitting. Thus, the scalability was measured through regular DFT simulation by using half of the cores per node. Since the convergence is difficult for the BLYP calculations on insulin, per-iteration performance was measured. It is expected that the better convergence can be achieved through the use of hybrid DFT method. Good scalability was observed up to 512 cores. It takes 14592 seconds with 16 cores, which reduces to 704 seconds at 512 cores. The scalability becomes poor at 1024 cores (543 seconds), which is the same phenomenon as in the Valinomycin simulation.

In summary, the scalability becomes poor as the number of processors approaches 1000. Also, the density-fitting approximation improves performance, but requires additional memory space. It was observed that the initial procedure of the density-fitting approximation is sequential. Distribution of this task will help to reduce the unbalanced memory requirement, so that the density-fitting approximation could be applied efficiently on far larger molecular system than with the current implementation.

3.2.3 *Plans and goals for a final enabling stage*

Two important issues for the simulation of large molecular systems of biological interest are good convergence of the solver and reducing memory requirements. As has been noted, one solution to this problem is to use the so-called range-separated functionals. The recently implemented ADMM exchange approximation in LSDALTON allows the use of range-separated functionals for large molecular systems at a cost similar to the traditionally much faster pure DFT functionals. This project will work on the better scalability of DFT for large molecules. The first and most important step will be to replace the `lstensor` structure with regular Fortran arrays, as this will lead to both a reduction in memory consumption and improved scalability with the number of nodes. The second step is to enable a fully parallelized and memory-distributed application of the density-fitting approximation. The third and final step involves the improved scalability of the exchange contribution, in particular applying the ADMM exchange.

3.3 PR 7: Enabling scalable highly parallelized MD simulations with non-periodic boundary conditions and arbitrary geometry

The DL_POLY project [12] was originally conceived in 1993 by William Smith at the Molecular Simulation Group (now part of the Computational Chemistry Group, MSG) at Daresbury Laboratory, under the auspices of the Engineering and Physical Sciences Research Council (EPSRC) for the EPSRC's Collaborative Computational Project for the Computer Simulation of Condensed Phases (CCP5). The software development has followed the trends of HPC for over 20 years and a number of general-purpose molecular dynamics (MD) packages have been available. The most recent one, DL_POLY_4, authored by Ilian Todorov and William Smith, has a general design, which provides scalable performance from a single processor workstation to a high performance parallel computer. It is supplied in source form under free-to-academia licence and can be compiled as a serial application code, using only a

Fortran90 compiler, or as a parallel application code, provided MPI2 instrumentation is available on the parallel machine. DL_POLY_4 offers fully parallel I/O (with user input to further tuning depending on HPC I/O sub-system) as well as a netCDF alternative (HDF5 library based) to the default ASCII trajectory file. It is also available as a CUDA+OpenMP port, developed in collaboration with the Irish Centre for High-End Computing (ICHEC), to harness the power offered by modern GPU cards.

As general purpose MD software there are over 1600 licenses taken annually (over 14500 since the very first release of the DL_POLY software). Currently, ~15% of the licenses are distributed in China, ~14% in the USA, ~14% in the UK and ~19% in the rest of the European Union. By science domain the distribution is ~35% Chemistry, ~26% Physics, ~17% Materials, ~13% Engineering, 4% ~ Bio-Chemistry, ~3% Mechanics, ~3% Software.

3.3.1 *Application enabling principles*

Two significant issues are being considered and addressed in this implementation phase of PRACE – the following two important inclusions within the framework of DL_POLY_4:

- inclusion of the *Analytical Generalized Born plus Non Polar* (AGBNP2) implicit solvation model, and
- development of an *Iterative Grid-Based Solver of the Poisson Equation* (IGBSPE) for Ewald electrostatics evaluation.

The potential benefits of the successful implementation of the AGBNP2 implicit solvation model will be the huge savings in computation time by excluding the unnecessary degrees of freedom of the explicit solvent and approximating the solvent interaction with the advanced solvation model.

The implementation of an Iterative Grid-Based Solver of the Poisson Equation (IGBSPE) is a challenging alternative for the Ewald electrostatics evaluation within DL_POLY_4. DL_POLY_4 uses the Smoothed Particle Mesh Ewald to evaluate the electrostatics interactions in charged model systems. This method generally scales as O(NlogN) with system size N due to the scalability of the 3D fast Fourier transforms, Daresbury advanced Fourier Transform library (DaFT), used as a central operation of the method. However, this challenge proposes to introduce a purely linear scaling, O(N), alternative of the IGDSPE, by utilizing the already developed machinery within DL_POLY_4. Despite its iterative nature, this solver offers advantages SPME cannot – in particular, an easy extendibility to non-periodic systems that makes it naturally suitable for research in biochemical membrane systems as well as materials interfaces (slabs). *Initial report on enabling*

Because no source implementation of the AGBNP2 algorithm was available as a starting point, the development begun based on the publication of it. First, the requirements for implementation of AGBNP2 within D_POLY_4 framework were investigated and the dependencies upon the internal data structures were identified. These have been accounted for during the implementation of a basic prototype of the central algorithm within AGBNP2 and a couple of helper routines. All this code is developed in a stand-alone module so that it can be easily embedded within DL_POLY_4 later. During preliminary tests it has been found that it is of benefit to keep the 2-, 3- and 4-crossectional volume calculations to the minimum and keep these in arrays. However, due to the irregular patterns of these occurring only within a limited volume of the system (the surface of a protein) the arrays to keep these were too big as well as the search over 2-, 3- and 4- particle groups was too demanding on memory. Because of the preliminary measures the approach was changed to experiment with b-trees using Fortran pointers. Another potential problem is the seamless embedding into the legacy software.

After a trial-and-error study a combination of two grid-based Poisson Solvers – Bi-Conjugate Stabilized Gradient (BCSGPS) and Conjugate Gradient (CGPS) – for inclusion in molecular dynamics calculations were developed. It has been found that the BCSGPS is better suited than CGPS to start with as it does not need a well-defined initial point for convergence and thus starting from a previous state offers a reasonable time-saving. However, the unconditional convergence of BCSGPS leads to a non-stabilized solution (10^{-4}) which lacks the accuracy. Therefore, the application of CGPS may be necessitated to achieve the desired accuracy of the convergence (10^{-6} as needed by MD).
A working stand-alone module for calculation and search of single, 2-, 3- and 4-corssectional volumes has been done. A module demonstrating the BCSGPS and CGPS work together for a grid with one moving charge has been created.
The estimate for the binary tree search over 1 million volumes is 20 microseconds, whereas for the default search that has been developed 20 milliseconds. The target is to do the electrostatic potential evaluation in less than 20 milliseconds for a system of 10,000 charged particles per processor.

3.3.3 *Plans and goals for a final enabling stage*

Further work on AGBNP2 will include:

- Defining a database with Van-der-Waals and ionic radii for a number of chemical species;
- Testing for correctness of the binary tree search, incorporating a user defined option and a switch for the AGBNP2 functionality – the switch will be called within two-body forces, where separate, force, virial and stress components will be accumulated within the routine;

Further work on IGBSPE will include making the prototype domain decomposition compatible. The implemented protocols will be tested. An interface to a multi-grid pre-conditioner will be developed for DL_POLY_4 electrostatics (in collaboration with STFC).

3.4 PR 8: Sustainable food production through fisheries and aquaculture

The data processing of the challenge contains three phases: data acquisition, variant calling and data analysis. The challenge aims to scale the processing of genome data from about 10 genomes (today) to 1000 genomes. This report covers work on the enabling of the first phase *data acquisition*. It is assumed that, for the data acquisition phase, the overall computational complexity scales linearly because each genome can be processed separately. For each genome, the data acquisition runs through a sequence of steps – also called workflow – depicted in Figure 2. The applications used are FastQC, cutadapt, BWA, mapDamage and standard UNIX tools. Table 4: Details of applications of the data acquisition workflow provides details for these applications.

Application name	License, Origin	Methods used
FastQC [13]	GPL v3 Bioinformatics	Detecting problems or biases in sequenced data. FastQC aims to spot problems in the sequencer or in the starting library material.
Cutadapt [14]	MIT Bioinformatics	Removes adapter sequences from sequencing data[15].
BWA [16]	GPL v3, MIT Bioinformatics	Burrow-Wheeler aligner using three algorithms BWA-backtrack, BWA-SW and BWA-MEM.
mapDamage [17]	GPL Bioinformatics	Tracks DNA damage pattern in sequencing reads[18].

Table 4: Details of applications of the data acquisition workflow

3.4.1 *Application enabling principles*

As outlined in the initial plan, enabling the sequence of applications began by considering a pilot use case first. The main difference between the pilot and the full use case is the number of genomes considered, 70 vs. 1000, respectively. For the pilot use case, the workflow of processing steps for sequenced data and all logistics for job submissions, data transfers, etc. is implemented in a generic way such that they can be reused for the full use case. This principle will simplify the exchange of the applications used (see Table 4), and makes reuse of the enabling work for other use cases easier. A second principle being employed is to test the workflow with different input data and parameters. This ensures that potential obstacles are detected as early as possible. Third, the applications are initially used without modifications i.e. considering them as black boxes. If, however, a change is needed, reusing existing tools or libraries helps to limit the amount of self-written application code.

3.4.2 *Initial report on enabling*

All applications used in the data acquisition workflow (see Figure 2) were downloaded to CURIE system (CEA, France) and installed without any problems. A scaling bottleneck is the number of files created in the mapDamage step. For each chromosome in the reference genome, seven files are created resulting in about 200K files for a single test run. On-going enabling work aims to lower that amount by employing a hierarchical data layout, namely HDF5, for which libraries exist. One issue in implementing the overall framework for orchestrating many workflow instances in parallel lies in the available access methods (login) at CURIE. Because it only allows password-based ssh, automating requires to periodically run a script on CURIE and let it fetch new tasks (job submissions, data transfers) from a remote site. Currently, such mechanisms are being implemented.

A single workflow instance runs for around 10 hours on a single core and requires a manual orchestration overhead of approximately one hour. For a single genome there are about 10 independent input files, each being processed in a single workflow instance. Hence, a scientist can analyse a single genome at a time (provided sufficient processing resources are available) and a single genome requires about 20 hours to be analysed. Although, a single workflow instance cannot be speed-up (unless multiple cores may be used), eliminating the manual

orchestration overhead enables huge potential for scaling. Exploiting this potential successfully requires two measures: removing scaling bottlenecks such as the creation of a huge number of files by mapDamage, and automating the logistics between a remote execution site and the home site of a scientist.

Apart from the issues with password-based ssh access on CURIE and the scaling bottleneck of mapDamage no additional problems were identified.

For all steps of the data acquisition workflow, wrapper scripts were implemented and run for both functional testing and observing resource consumption. These tests were performed manually. To gather more information about the resource usage pattern the wrapper scripts were revised to allow for additional user space monitoring. For each step the necessary tools are packed with the actual application into a single MPMD job and executed through the batch system. Job preparation and job submission tasks were decoupled into two different logistics' steps such that a job can be repeated easily. To enable scaling of a large set of genomes, tools for automating the logistics of job creation, job submission and data transfer are being implemented. These tools need further development as well as scalability testing and improvements. At the time of writing, the enabling work is focused on reducing the number of files created by mapDamage. At best, each run will only create a fixed number of files after the modifications.

Initial tests indicate that the overhead of the manual orchestration of a single workflow will be eliminated completely.

3.4.3 *Plans and goals for a final enabling stage*

At the moment, the implementation of the data acquisition phase is being completed. Next, the achieved scalability will be assessed. Subsequently, the other phases for the pilot use case and the full 1000-genome use case will be implemented. To make the use of the implementation as easy as possible, the workflow steps will be made accessible from a portal. Because the authors already use Galaxy as portal solution for other applications, it will be used to implement an interface between the current command-line based implementation and the GUI of the portal.

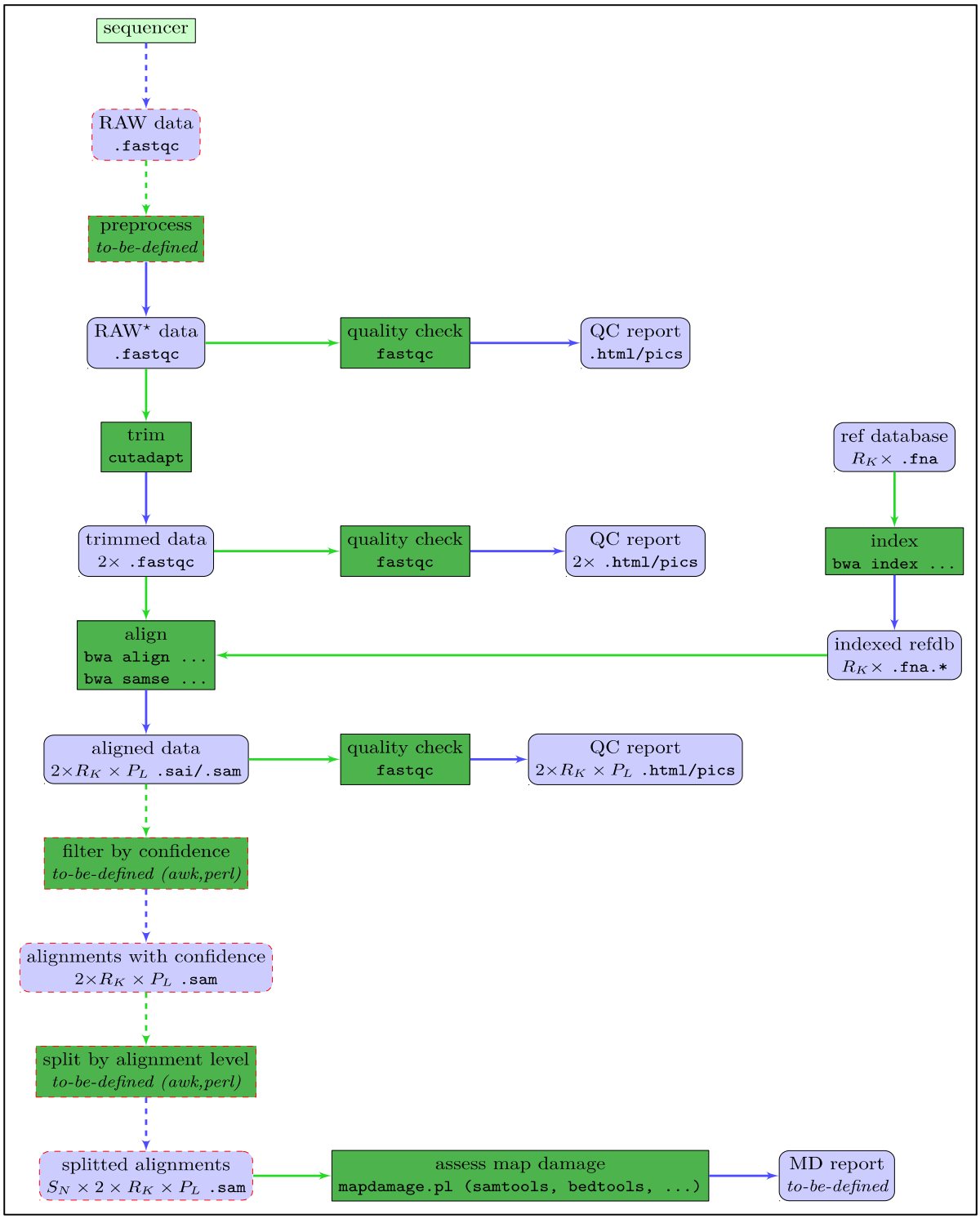

Figure 2: Data acquisition workflow.

3.5 PR 9: Multidiscipline Simulations for Aircraft Designs

The overall goal is to demonstrate the opportunities and challenges of high-performance computing in the area of multi-physics simulation and optimization based on an application from aeronautics industry. Integrating high-fidelity models covering multiple disciplines and precisely describing their interactions, especially between fluid and structural dynamics, has become an increasingly important factor in the design processes of aviation industry over the last decade. For example such models are required to compute minimum drag wing designs in order to construct more fuel-efficient, and hence environmentally, friendlier planes.

Work in this task focuses on simulations of fluid-structure interactions (FSI) as they are typically utilized in aircraft design processes. Emphasis will be put on a partitioned formulation of the fluid-structure coupling where two different possibilities of coupling will be explored. First, the coupling is done within one system that is capable of fluid dynamics computations (CFD) and structural mechanics computations (CSM). Second, the coupling is done between two different systems, where one system is exclusively used for CFD and the other one exclusively for CSM. The second possibility is the more relevant industrial use-case today as highly specialized simulation systems can be used for the individual disciplines. OpenFOAM has been chosen for CFD because of its large range of well-established and reliable turbulence models as well as its advanced functionalities for dynamic meshes. CSM will be done with Elmer.

OpenFOAM [19] is a free, open source CFD software package developed by OpenCFD Ltd at ESI Group and distributed by the OpenFOAM Foundation under the GNU General Public Licence version 3 (GPLv3).

Elmer [21] is an open-source multi-physics simulation software mainly developed by CSC - IT Center for Science. Elmer development was started 1995 in collaboration with Finnish Universities, research institutes and industry and is distributed under the GNU General Public License version 2.0 (GPLv2).

3.5.1 *Application enabling principles*

Work for this task is divided into the following steps.

- Preparing a large-scale simulation model consisting of a complete aircraft with a geometrically complex configuration. This model serves as a reference example for a representative industrial use-case and is the basis for profiling the existing systems and the new functionalities described below. Also the work regarding analysing and improving scalability will be driven by the use-case.
- OpenFOAM is extended such that the coupling and both computations (CFD and CSM) are done within OpenFOAM. Compared to a standard CFD simulation the computational complexity of this type of simulation is considerably higher because an additional nested loop has to be added for exchanging data between the fluid and structure models. Runtime behaviour and bottlenecks regarding scalability will be investigated.
- Development respectively integration of a coupling interface for mixed system FSI computations. In addition to the simulation mentioned above explicit data transfer and data synchronization between the two different systems are required. This new functionality adds significant new aspects to the scalability of the whole system. Again the resulting runtime behaviour and bottlenecks regarding scalability will be investigated.

To date no detailed results about scalability for such types of simulations with OpenFOAM in the context of HPC exist. Therefore it is expected to gain new insights about current limitations and valuable information for possible improvements and future development of OpenFOAM and mixed system coupling interfaces in general. The results of this work can also be seen as a reference for the development of other multi-physics / mixed-systems simulations. Furthermore this works contributes to the extension of the application domain of OpenFOAM.

3.5.2 *Initial report on enabling*

Previous profiling work done on OpenFOAM (within PRACE) for standard CFD simulations has shown that I/O operations and metadata handling can become a limiting factor with respect to strong scalability[21]. Since for a given model size FSI simulations have a substantially higher computational complexity than standard CFD simulations it has to be investigated how the I/O behaviour scales in this case. Another key factor for scalability is related to the solution of the systems of equations. As OpenFOAM uses iterative solvers the memory-bound nature of the code has a strong impact on the overall performance[22]. One objective of this work is to analyse in detail how intra-node and inter-node scalability behave for FSI simulations.

Work for this task started in March 2013. By the end of April the preparation of two models (one very well documented model for verifying the results and one complex industrial model) was finished. Also a prototypical implementation for dynamic mesh computations within OpenFOAM has been made available, as it is needed for coupling the fluid and structure computations. Initial small-scale tests have been carried out on CURIE.

3.5.3 *Plans and goals for a final enabling stage*

The work plan for the remaining part of the project includes the following development tasks:

- Implementation of FSI computation within OpenFOAM.
- Implementation of a coupling module for mixed systems using FSI.
- Performance analysis with special emphasis on the topics described above and development of proposals for improvements.

3.6 PR 11: Big Data for Machine Learning

Large-scale machine learning problems exist in a large number of social and industrial applications, such as social network analysis, consumer/voter preference analysis, anomaly detection, credit card fraud control/management systems, and postal automation. This study investigates the scalability of fundamental data mining approaches used for recommendation mining, ranking, clustering, and classification, with the aim of exploring/enabling processing of terabytes of data on PRACE Tier-0 systems. In particular, the applicability of the Map/Reduce parallel programming paradigm is evaluated on supercomputers.

Map/Reduce is a framework originally developed at Google to ease development of parallel and distributed codes[23]. The Map/Reduce paradigm originates from functional programming, where higher order functions map and reduce are applied to a list of elements to return a value. The Map/Reduce framework provides a runtime system that manages mapper and reducer tasks, providing automatic scalability and fault tolerance. With the help of this framework, it is possible to ignore complex parallel programming structures such as message passing and synchronization and the programmer only needs to design a mapper and a reducer function for each distinct map/reduce phase. Along with reducing programming complexity, another important feature of Map/Reduce is that it can operate on massive data sets. That is, Map/Reduce is designed for scalability instead of speedup.

The most widely used open-source Map/Reduce implementation is Apache Hadoop Project[6]. Hadoop is developed in Java and makes use of TCP/IP ports for communication. Since PRACE Tier-0 systems fail to support the requirements of Hadoop, alternative Map/Reduce implementations are considered. Among these, the MapReduce-MPI (MR-MPI)

library [24] developed at Sandia National Labs seemed the most appropriate choice for supercomputing systems.

MR-MPI library is a lightweight Map/Reduce implementation developed in C++. It uses the MPI library for inter-process communication. These properties enable MR-MPI to be used on HPC platforms without an extra overhead. It also has additional functionalities that can be utilized for speed-up. In the original Map/Reduce framework, it is required to submit each Map/Reduce phase as a separate job, which causes a decrease in performance. In contrast, MR-MPI library does not have such requirements, which leads to a performance increase especially in iterative algorithms such as graph algorithms. In the original Map/Reduce framework, initial key-value pairs produced by a map phase are all written to the disk system waiting for the reducer tasks to read their own partitions via remote procedure calls. In the MR-MPI library, whenever a mapper task produces all its key-value pairs, it is not obliged to write all of these key-value pairs to the disk, but instead it is possible to communicate these key-value pairs with reducer tasks while storing them in memory. MR-MPI also provides additional functions to manipulate key-value phases between map tasks and reduce tasks. For example one can reduce some of the key-value pairs and produce new key-values from them. Later it is possible to union old key-value pairs which are not reduced with the new key-value pairs for further reduction operations. With MR-MPI a set of further optimizations can be achieved while designing new efficient Map/Reduce algorithms. The MR-MPI library has the possibility to make an important impact in scientific computing since it eases parallel programming while providing high scalability for HPC platforms.

3.6.1 *Application enabling principles*

Within this project, the MR-MPI library is evaluated with respect to its performance on data mining and machine learning algorithms such as All Pairs Shortest Path (APSP) and All Pair Similarity Search (APSS). APSP is generally used for measuring importance (via metrics such as betweenness centrality) in social networks. APSS is a widely used kernel for many data mining algorithms such as construction of K-Nearest-Neighbour graphs. The M/R-based algorithm for APSS is given in Figure 3.

APSP and APSS algorithms can be thought of as matrix multiplication operations: the only difference is that the APSP algorithm needs $O(\log_2 n)$ consecutive matrix multiplications whereas the APSS algorithm needs only one matrix multiplication of a feature matrix with its transpose. Since both algorithms depend on matrix multiplication operations, a generic matrix multiplication code that uses the MR-MPI library was developed within this challenge. By modifying the binary operators applied to the matrix elements, the generic algorithm can easily be converted to one of the algorithms mentioned above. Figure 3 shows the pseudo code of the two phase matrix multiplication algorithms using MR-MPI library.

3.6.2 *Initial report on enabling*

To test the MR-MPI library, the library was ported to Hermit system (Cray XE6 at HLRS Stuttgart). Compilation of the library caused no problems, and will likely not on most HPC systems since MR-MPI is only dependent on the MPI library. While testing the MR-MPI library, it was observed that to obtain peak performance, MR-MPI must be run on a parallel file system. Hermit provides a parallel file system named Lustre. One should be careful while storing data on this system because different settings to store input files on multiple disks may cause significant performance differences. To efficiently use the parallel file system and the MR-MPI library, it was necessary to perform parallel MPI-I/O in the map and reduce phases of algorithm, during which the data is transferred between the memory and the disk.

A randomly generated graph and the LiveJournal social graph were used for initial testing. The randomly generated graph is recursively generated with power-law degree distributions. These kinds of graphs are commonly used to represent social networks. The LiveJournal social graph is taken from Stanford University Large Network Dataset Collection [25]. LiveJournal is an on-line community in which significant fraction of members are highly active. It allows members to maintain journals and to declare which other members are their friends. Each node in the graph corresponds to a member in network and an edge is added between nodes if two of the members become friends. The properties of these datasets are presented in Table 5.

Table 6 shows the initial results of the tests for APSS on the real data set LiveJournal. The APSS algorithm can be used to find similarities between people in a social graph. If any two members have similar friendship profiles in the network, using the APSS algorithm, it is possible to recommend new friends to these users. As seen in the table, the M/R-based APSS algorithm scales well up to 2K cores. In Table 1Table 7 runtime results of APSP algorithm run on a randomly generated graph are shown. The APSP algorithm is generally used to find central nodes in a social graph. As seen in the table, the code scales reasonably.

Name	Type	Nodes	Edges	Description
Random Graph	Directed	1,024	15,608	Randomly generated network
LiveJournal	Directed	4,847,571	68,993,773	LiveJournal online social network

Table 5: Dataset properties

APSS – LiveJournal Experiments (New)							
K	Map	Collate	Reduce	Map	Collate	Reduce	Total
1024	2.02	4.67	22.21	24.00	35.84	0.49	89.29
2048	1.55	14.69	4.77	3.87	11.72	0.34	36.96
4096	1.71	17.67	4.51	3.65	9.03	0.09	36.68

Table 6: APSS Experiments (results in seconds)

APSP – Random-Graph Experiments	
K	Total (s)
256	**1097.09**
512	**868.08**
1024	**621.93**

Table 7: APSP Experiments

One important observation that should be mentioned is the time spent for the development of the APSS and APSP codes using the MR-MPI library. Once the setting up of the platform on the Hermit system was accomplished, implementing the highly scalable APSS and APSP codes took one and three days, respectively.

3.6.3 *Plans and goals for a final enabling stage*

Currently the MR-MPI library is ported to the systems JUQUEEN and CURIE. In addition, MR-MPI implementations of different machine learning and data mining algorithms will be explored over much larger datasets. This includes collecting information on memory problems, I/O issues, and data gathering issues.

```
Input: MapReduce object M = non-zero coordinates with weights: Key =
(row_i, col_j), Value = (m_ij)
       MapReduce object MT = non-zero coordinates with weights: Key =
(row_i, col_j), Value = (mt_ij)

Map columns using M as input:
      Input: Key = (row_i, col_j), Value = (m_ij)
      Output: Key = (col_j), Value = ('M', row_i, w_ij)
Map rows using MT as input:
      Input: Key = (row_i, col_j), Value = (w_ij)
      Output: Key = (row_i), Value = ('MT', col_j, mt_ij)

Collate columns and rows: Row and column numbers are Key
Reduce:
      Input: Key = (col_i) or (row_i), MultiValue = ('M', row_i, w_ij) or ('MT', col_j, mt_ij)
      ColumnVectorM, rowVectorMT = split MultiValues according to the
tags in the values
      C = columnVectorM × rowVectorN
      For each c_ik ∈ C do
            Output: Key = (row_i, col_k), Value = (c_ij)

Map columns using C as input:
      Input: Key = (row_i, col_k), Value = (c_ik)
      Output: Key = (col_k), Value = ('C', row_i, c_ik)

Convert: collect all partial results in the same column
      Input: Key = (col_k), MultiValue = ('C', row_i, c_ik)
      Output: Key = (col_k), Value = MultiValue

Reduce:
      Input: Key = (col_k), MultiValue = ('C', row_i, c_ik)
      For each c_ik ∈ C_i do
            r_ik += c_ik
      Output: Key = (row_i, col_k), Value = (r_ij)
```

Figure 3: APSS pseudocode.

3.7 PR 14: High Resolution Climate and Atmospheric Chemistry Modelling

The ECHAM/MESSy Atmospheric Chemistry (EMAC) model is a numerical chemistry and climate simulation system that includes sub-models describing tropospheric and middle atmosphere processes. Among them, the MECCA (Module Efficiently Calculating the Chemistry of the Atmosphere) is used to calculate the tropospheric and stratospheric chemistry. Further details can be found on[26]. The source code of MESSy is available under the GNU General Public License (GPL). ECHAM is freely available to the scientific community.

The KPP (kinetic pre-processor) is a software tool that assists the computer simulation of chemical kinetic systems. The concentrations of a chemical system evolve in time according

to the differential law of mass action kinetics. KPP translates a specification of the chemical mechanism into Fortran90 code while it provides a comprehensive suite of stiff numerical integrators. The produced computer code is incorporated in turn (using the appropriate post-processors) into the MECCA sub-model of EMAC. KPP [27] is free software under the GPL License.

The KPPA (kinetic pre-processor accelerated) produces Fortran90 code that can use the GPU or multi-core CPU to deliver improved performance with respect to the KPP. The software [28] is not free and offers a variety of licensing options, including a free 60-day demo license on request.

3.7.1 *Application enabling principles*

Climate change has a series of important socio-economic effects including agricultural production and human health. The EMAC model is used in a various studies by several research and industrial groups. Chemistry kinetics is a bottleneck for the performance of accurate, high-resolution simulations, as it is computationally the most intensive part of a climate simulation requiring sometimes up to 90% of total computing time. Bringing these calculations to the GPUs will significantly reduce the time-to-solution.

The enabling tasks are defined below:

- Getting access to the CURIE supercomputer and other hybrid systems with GPGPUs.
- Installing the existing version of the EMAC model with the PGI compilers, setting up several case studies (includes the transfer of several data bases) and performing reference runs the results of which will be used later on, in order to examine the correctness and the efficiency of our implementation in GPUs.
- Getting and installing a version of KPPA and clarifying license issues regarding its use.
- Running test cases and becoming familiar with the produced CUDA Fortran code that will be incorporated into EMAC later on.

3.7.2 *Initial report on enabling*

Chemical kinetics models may be responsible for over 90% of the computational time. The models are embarrassingly parallel for a given fixed grid employed on a climate simulation since the evolution of the concentrations of the chemical species at a grid point depends only on the initial concentration of these species and meteorology at this particular point. Given the fact that the Fortran code produced by the KPP pre-processor is highly optimized, the only possible way to reduce the time-to-solution is by porting the produced code to the GPUs.

Evaluation of the improvements in performance and scaling of the EMAC model requires moving the calculation of the chemical kinetics to the GPUs. However, based on the results reported from the KPPA developers for box models [29] a 10x speedup or more (for double precision accuracy) for the chemical solver is expected.

During the first period of the project access to the hybrid partition of the CURIE system was not available. For this reason access to a small local system with GPUs was granted. Installation of some packages (e.g. netCDF with PGI compilers) was necessary.

While porting the EMAC code to a GPU cluster using PGI compilers, several problems were encountered, most of them related to the PGI compiler itself. PGI support was contacted to discuss possible solutions of the issues. In several cases parts of the EMAC code were

modified to overcome these problems. In the final version of the code it is planned to use inline directives that will appropriately modify parts of the code when PGI compilers are used.

A free demo license was received for the KPPA package, which was successfully build too. It should be noticed that using the free demo license of KPPA it was possible to produce the necessary computer code for several chemical mechanisms that will be used as case studies during the proposed developments. Therefore, no further use of KPPA is expected to be required during this work.

Currently short simulations using EMAC + KPP are being performed. The results of these simulations will be used throughout this project in order to evaluate the correctness and efficiency of the new developments. Preliminary work has been performed for identifying the part of the EMAC model into which the KPPA code will be incorporated.

3.7.3 *Plans and goals for a final enabling stage*

The plans for the remaining part of the project are:

- Development of the appropriate post-processors that will use the CUDA-Fortran KPPA generated code and will transform it to the appropriate form in order to incorporate it into the EMAC code. The same technique has been used in the latest versions of EMAC in which KPP was employed for the generation of the chemical mechanism Fortran code. It should be emphasized that this step is crucial for the efficient performance of the code. It was found that by proper modification in the original code of the loop structure and (optionally) vector blocking, improved run-time performance was achieved on vector and scalar architectures. In addition, for enhanced performance, the operations in the nested loops contain integer arrays for the indirect addressing of arrays (e.g. in the subroutine that performs the LU decomposition) that prevent the compiler from efficient optimization, and therefore (since the indices do not change) will be replaced by the appropriate sequence of operations without the need of index arrays.
- KPPA produces the code for the chemical mechanism at a specific box (grid point). The possibility to develop a second level of parallelization in which several boxes belonging to the same atmospheric column will be processed in parallel from the same GPU will be investigated. The efficiency of the two, parallelization schemes will be compared. The post-processors developed in the previous step will be used in this case as well.
- A version of KPP with OpenACC directives will be developed along with the appropriate post-processors for the incorporation of the produced Fortran code into EMAC. In this way we can overcome difficulties related to the KPPA licensing policy.
- All previous approaches for bringing the chemical solvers into GPUs will be evaluated by performing a series of simulations with EMAC. During this process, several chemical mechanisms and simulations setups usually employed by end-users in production runs will be selected, while different computer platforms (including CURIE and our local hybrid system) will be used.

3.8 PR 15: Multidisciplinary modelling for interactive design of lakes

Delft3D is a flexible integrated modelling suite, which simulates two-dimensional (in either the horizontal or a vertical plane) and three-dimensional flow, sediment transport and morphology, waves, water quality and ecology and is capable of handling the interactions between these processes. The suite is designed for use by domain experts and non-experts alike, which may range from consultants and engineers or contractors, to regulators and government officials, all of whom are active in one or more of the stages of the design, implementation and management cycle.

The challenge is to enable the design process of a lake and its environment that asks for an interactive approach in which different aspects (economical, engineering, recreational, safety for flooding, ecology) from different stakeholders can be combined. For this purpose, for Lake Marken in the Netherlands, a multidisciplinary coupled model exists. However, the initial runtime for a scenario with the model is four days, so interactive sessions with stakeholders that combine drawing measures with the calculation of its effects with the model are not feasible yet.

Lake Marken is a large lake in the Netherlands that is delineated by the provinces of North-Holland and Flevoland and the dyke called the Houtribdijk (see Figure 4). The lake is surrounded by cities and agricultural regions and is also used for recreational reasons. The development of the lake therefore includes many different stakeholders, each with their own focal points. The complete, numerical model therefore consists of several components that simulate different aspects of the lake:

- Delft3D-FLOW is used for the shallow water flow and Delft3D-WAQ for sediment transport, re-suspension, and sedimentation, water quality, vegetation, algae. Delft3D-FLOW is a multi-dimensional (2D or 3D) hydrodynamic (and transport) simulation program using the finite difference method which calculates non-steady flow and transport phenomena that result from tidal and meteorological forcing on a rectilinear or a curvilinear, boundary fitted grid. In 3D simulations, the vertical grid is defined following the sigma co-ordinate approach. Its parallelization approach is MPI using one-dimensional domain decomposition.
- The SWAN model is used for waves. SWAN is a third-generation wave model that computes random, short-crested wind-generated waves in coastal regions and inland waters. SWAN computations can be made on a regular, a curvilinear grid and a triangular mesh in a Cartesian or spherical co-ordinate system. Two parallelization strategies are available: MPI and OpenMP.

All components are available as open source under the GPL license. Delft3D-FLOW and Delft3D-WAQ are available at [30] and SWAN is available at [31].

3.8.1 *Application enabling principles*

The overall goal is to reduce the runtime from 4 days to 1 hour, to enable running an interactive lake design sessions. The detailed step-by-step plan is explained below.

- (Step 1) Analyse the parallel performance of Delft3D-FLOW applied to shallow lakes on typical PRACE hardware. In 2012 the parallel performance of Delft3D-FLOW was analysed as part of the Work Package 9 “Industrial Applications Support” of PRACE-2IP on the CURIE system. For river applications (schematic Waal model, Zeedelta model) this code turned out to be well-suited to PRACE hardware, especially at the Tier-1 level. However, parallel implementation is by partitioning in one direction only:

for square shaped shallow lakes like Lake IJssel and Lake Marken partitions may be become very thin which may affect parallel performance. For this purpose we consider an existing detailed hydrodynamic model with intrusion from sea, mixing and transport of chloride/salt is considered for Lake IJssel that is used for studies on fish migration and quality of drinking water.

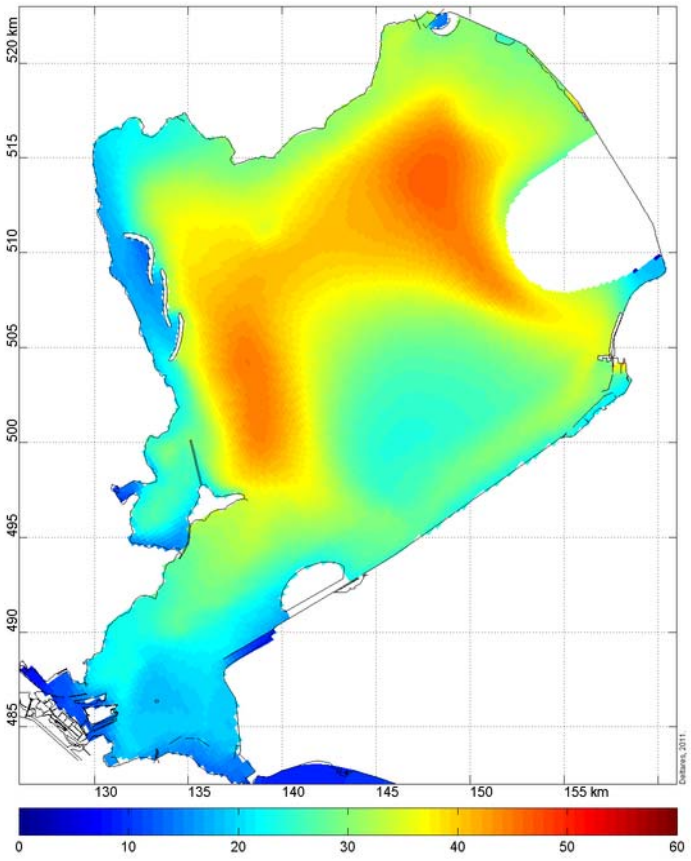

Figure 4: Computational domain for Lake Marken. The figure includes a design option for lake development.

- (Step 2) Demonstration of interactive design for shallow lake with fetch length approach on Collaboratorium at SARA. One of the important driving processes in Lake Marken is re-suspension of sediment due to wind driven waves. For wind driven waves, in the multidisciplinary coupled model for Lake Marken two modelling approaches exist: a fetch length approach which runs relatively fast and a more detailed approach with SWAN for computing waves. The same fetch length approach was used for the Loosdrecht lakes as a rapid assessment tool. This case is used as a demonstrator for interactive design to focus on interaction and visualization aspects first. After this initial demonstration, about half a year later an intermediate design session will follow with uses improved knowledge and techniques (visualization, models, software etcetera) and the project ends with a final session.
- (Step 3) Further development of tools for interactive design. At the moment at Deltares several tools are available (e.g. brownfield navigator, map-/touchtable, Matlab scripts): a selection of (parts of) them will be used and further developed with in mind that they can be reused for other application areas. Also it will be investigated whether components of a map-table approach with both WAQUA (comparable to Delft3D-FLOW in some sense) and SWAN for Lake Ijssel, which has been developed earlier can be reused in the current application.

- (Step 4) Analysis of parallel performance of Delft3D-WAQ on typical Tier-1 hardware. After computing flow (via Delft3D-FLOW) and waves (via fetch length approach or SWAN) the multi-disciplinary coupled model for Lake Marken starts Delft3D-WAQ for computing re-suspension, sedimentation, and transport of silt by means of reaction-advection-diffusion equations. Although, computational times are relatively small, an open question is how OpenMP and MPI (at the moment only for explicit time integration schemes) versions of Delft3D-WAQ perform on typical PRACE hardware and which possible adaptations may improve Delft3D-WAQ.
- (Step 5) Investigation of possibilities to run an MPI version of Delft3D-FLOW coupled to OpenMP version of SWAN on Tier-0/Tier-1 hardware. At the moment there exists a parallel OpenMP version of SWAN for computing waves that scales almost linearly. As for the multi-disciplinary coupled model for Lake Marken, SWAN is computationally most intensive, so the first question is what type of nodes (with 40 cores, for instance at LRZ or with 128 cores, for instance at CURIE) are suitable. Whether the coupling between Delft3D-FLOW and SWAN can be used for respectively MPI version and OpenMP versions at the specific Tier-0/Tier-1 hardware will then be investigated. Depending on step 5 choice will be made between the fetch length approach and SWAN for wave modelling.
- (Step 6) Combining steps 1, 2, 3, 4, and possibly 5 (depending on the outcome of step 5) such that interactive design sessions can be held at the Collaboratorium at SURFsara with the multi-disciplinary model for Lake Marken. The results will be used as a demonstration case to illustrate the new possibilities of this approach.

3.8.2 *Initial report on enabling*

Delft3D-FLOW was already ported as part of the Work Package 9 Industrial applications support of PRACE-2IP to the CURIE system. An initial configuration for Lake Marken that can be used for Delft3D-FLOW for the hydrodynamic simulation is already available. This configuration has been used to create several profiling runs on CURIE. The scalability is quite poor and no more than twelve MPI processes can be used. The poor scalability can be mainly attributed to the small number of grid-points and that this simulation does not yet include all the components. At higher numbers of MPI processes the solution does not converge anymore and the time step needs to be reduced, which negates the increased compute power. The total simulation runs for about eight hours.

Furthermore, the Delft3D-FLOW model is being ported to the Fermi system at CINECA, which is based on the IBM BlueGene/Q architecture. This port still crashes with an unknown I/O error. Finally, a preliminary preparation for an interactive lake design session took place.

3.8.3 *Plans and goals for a final enabling stage*

It is expected that the total number of MPI processes can be increased to about 100 when more components are included, but not significantly higher. Ensemble simulations may increase this further. One of the improvements is the combination of MPI and OpenMP for the SWAN wave application. A risk is that the different components cannot be easily integrated to run as one parallel application.

The project follows the previously described steps without any delays. Step 1 and part of the Step 2 have been successfully completed.

3.9 PR4: Uncertainty analysis from numerical experiments, with applications to safety of nuclear power plants

In order to assess safety parameters such as nuclear fuel temperature, numerical simulations are done with comparison to experiments. The URANIE[32][33][34][35] tool uses propagation methods to assess uncertainties in simulation output parameters in order to better evaluate confidence intervals (e.g., of temperature, pressure). This is used for Verification, Validation and Uncertainties Quantification (VVUQ) process used for safety analysis.

While URANIE is well suited for launching many instances of serial codes, it suffers from a lack of scalability and portability when used for coupled simulations and/or parallel codes. The aim of the project is therefore to enhance this launching mechanism to support a wider variety of applications, leveraging on work previously done on analysis of best pathways to take on HPC capabilities, with goal to capitalize on those larger computing resource so to achieve a significant new level of statistical assessment for models.

Uncertainty quantification and data assimilation are key to industrial acceptance of predictive simulations. URANIE has been developed as the working platform of choice in VVUQ for CEA, AREVA, EDF and other partners, leveraging open source tools such as CERN's ROOT toolkit and EDF/CEA's integration platform SALOME, and included as a sub-part in open European projects of NUclear REactor SIMulation in 6th, 7th, and now 8th framework (NURESIM, NURISP, NURESAFE). Others project using URANIE include multi-criteria optimization for laser simulation (CELIA), optimization design for ESS (European Spallation Source – Lund, Sweden), and uses in ALLIANCES platform (ANDRA, French national waste storage agency).

3.9.1 *Enabling principles*

In order to efficiently address HPC usage, some bottlenecks must be addressed, in order to obtain scalability and performance on large scale (>1000s cores) computers. As of now, URANIE is capable of running inside the SALOME framework, which has been ported successfully onto small and medium clusters, and must now be tailored to much larger HPC supercomputer capabilities. The approaches to be applied will build on the ongoing PRACE-2IP "Emerging applications" PA type C project – they will include enhancing MPI parallelization on either single code or multiple (à la "ensemble") simulations.

3.10 PR6: Electron-molecule resonance data for DNA radiation damage studies

Energy deposited in cells by ionizing radiation is channelled into production of free secondary electrons. The reactions of such electrons induce substantial yields of strand breaks in DNA. The strand breaks are associated with electrons trapped in quasi-bound 'resonances on the basic components of the DNA. This project will provide enhanced computational resources for collaborating universities (UCL, Open University, University of St Andrew's and more) to study resonance formation in much more detail than current initial (though impressive) calculations [36] (see also [37]).

The widely-used UKRmol general purpose electron-molecule R-matrix collisions suite[38] will be adapted: the relevant parts of the STFC code PFARM[39][40][41] will be incorporated appropriately into the UKRmol resonance finding and fitting modules RESON and/or TIMEDEL, replacing a serial propagator (coupled PDE solver) module RSOLVE. The UKRmol routines are serial.

PFARM scales to 10000+ cores (thanks in part to completed PRACE-2IP work). Proof of concept exists, as PFARM has already been successfully adapted to run with UKRmol, replacing RSOLVE in an initial scattering matrix calculation over a (coarse or) fine grid of energies. PFARM optimizes load balancing by pipelining energy calculations across cores, effectively overlapping what would be step-by-step operations.

3.10.1 *Enabling principles*

The resonance modules currently use the initial grid data to locate possible resonances and then call RSOLVE, often recursively, for more detailed study and at the computed fitting points. The initial goal of the project is to parallelize the resonance modules such that the search procedure is partitioned, then the additional, currently ad-hoc calls to RSOLVE will be replaced by a (single, ideally) pooled call to PFARM. The subsequent calculation of widths and other fitting parameters will be partitioned or, if time allows, incorporated into the pipelining procedure controlled by PFARM's 'manager' tasks.

To facilitate progress in the time available, the PRACE WP7 developments will be tested on a relatively simple case (such as e+ N2 resonances) in conjunction with scientists at UCL, with final verification tests on a DNA base (adenine or guanine) to be set up at the end of the project.

The UCL scientists and their collaborators will then use the parallelized code to enable new more realistic and highly detailed DNA base resonance studies planned but not currently feasible.

3.11 PR12: Multidiscipline coupling in Cardiac computational mechanics

This project aims at enhancing a high performance computational biomechanics simulation tool, named Alya Red [42], to simulate cardiac mechanics. Although large scale supercomputers are worldwide available to researchers, almost no high performance computational biomechanics simulations tool can profit from these resources, not even at a modest performance level. Alya Red is the application of Alya system to biomechanics problems and, in particular, to cardiac mechanics. Alya is a high performance computational mechanics code developed at BSC-CNS (Spain). It is specially designed for running with the highest efficiency standards in large-scale supercomputing facilities, and capable of solving different physics in a coupled way: fluids, solids, electrical activity, species concentration, etc. Both main features are intimately related, meaning that all complex coupled problems solved by Alya must retain the parallel efficiency.

The socio-economic challenge of Alya Red Project is to perform simulations of biological systems at organ level, creating a computational infrastructure to achieve the following objectives: first to help medical researchers to better understand cardiovascular mechanisms and improve their physiological models, second to reduce the “from the lab to the patient” time in pharmacological and clinical research in both the private sector and academia, reducing production costs and animal essays, and third, to help medical doctors to better understand what causes illness, improve diagnose and design treatments and new healing strategies.

To attain these objectives, it is proposed to develop the fluid-electro-mechanical coupling in Alya Red. The electrical propagation and mechanical equations are solved in the same computational domain (the heart muscle) while the fluid equations solve for the blood motion, inside the heart cavity and down to the vessels. The technical objective is to enhance the volumetric coupling in the first zone, involving the electrical and mechanics problems, as well

as the interface coupling between this first zone and the second one, involving the fluid problem. The main idea is to asynchronize the computation between the zones to avoid idle CPU's.

3.11.1 *Enabling principles*

The cardiac computational model is based on a fluid-electro-mechanical coupled problem. The numerical model for the space discretization is the finite element method, with stabilization for the fluid problem and total Lagrangian formulation in the solid domain. Time discretization is based on finite difference Newmark scheme. The code is parallelized with a Master-Slave strategy, using MPI tasks. For the solid mechanics module of Alya, a hybrid OpenMP/MPI parallelization has been introduced in order to enable the code for massively parallel architectures. The hybridization of the rest of the modules is part of the future developments in Alya. The code has been proved to run for several hundred millions of elements, showing a near-linear scalability.

As far as the computational aspect of this project is concerned, the main bottleneck to address is to develop an efficient fluid-mechanics coupling by overlapping fluid and mechanics computations, with the objective to improve the baseline speedup of 50%.

3.12 PR13: Large-scale simulations of neural networks

With respect to the limited availability of experts that could provide substantial scientific background, support for this project has been postponed. Moreover, a risk related with missing contributors for application enabling has been identified.

4 Summary and future work

This document has summarized the work done in the period between M1 and M11 of PRACE-3IP, within the WP7 task on application scaling and support that address major socio-economic challenges, and presented its current status. The first part of the document (Sections 0 and 2), outlined the process that lead to the identification of a set of key socio-economic challenges and associated scientific aspects. The structure and organization of the work was laid down in detail and the follow up procedure for the submission of the project proposals was explained. A resulting list of proposed and selected projects with the corresponding application codes, covering various scientific aspects that address the selected, key socio-economic challenges, was included. Finally, the Milestone MS7.1 on M6, summarizing this period, was successfully meet by publicising the associated White Paper[7]. The work on the selected application codes within the run projects has started as from M7.

The TOMAWAC Wave propagation module, part of the intended, coupled application suite, has been successfully parallelized. Furthermore, the scalability of the coupled TOMAWAC and TELEMAC-3Ds modules was tested on a real case scenario of the Liverpool Bay, yielding very promising results. Plans for the final enabling stage of the project includes full application coupling with sediment transport module and significant improvement of the modelling resolution to be enabled on the selected Tier-0 systems.

Scalability of the LSDALTON code addressing drug design was tested with two real molecular structures Valinomycin and Insulin. Limiting factors for application scalability and memory consumption are identified and steps allowing removal of these bottlenecks defined. Remaining enabling steps will address these limitations for improved scalability of the DFT

method for large molecules. Moreover, a fully parallelized and memory-distributed density-fitting approximation and scalable exchange contribution will be applied.

Improved computational kernel for DL_POLY code is in the implementation phase with promising perspective for overall time reduction of the molecular dynamics computation proven by preliminary studies with simplified model. The final enabling will include full integration with the application code and consistency tests of the implemented method.

Successful approach to genome sequencing application workflow with optimized data acquisition and full automation has been proposed. Multiple program – multiple data (MPMD) execution model has been implemented resulting in improved logistics of a single workflow and allowing the overhead of manual orchestration to be completely eliminated. As a complement to already enabled data acquisition, scalable workflow for a pilot use case and the full 1000-genome use case will be implemented.

The careful preparations of the two models, concerning the verification of the results and a complex industrial model were finished successfully by the end of M10. Moreover, prototypical implementations within OpenFOAM started off that are essential for the complex coupling of the fluid and structure computations. The final implementation stages of FSI and coupling module for the mixed systems inside OpenFOAM and their extensive tests and analyses are scheduled for the remaining period between M12 and M23.

The porting of the MR-MPI code went through perfectly without any interruptions. The testing and initial runs based on the smaller data sets were successfully executed and resulted with a very promising scalability. Porting of the highly scalable implementations of both APSS and APSP codes are undergoing and their performance and scalability will be soon measured on most of the available PRACE Tier-0 systems. Those scalability and performance tests planned for coming months, will be executed on significantly larger datasets.

The EMAC code was ported to a GPU cluster that uses the PGI compilers, without any major issues. The KPPA GPU code was used as a basis to produce a necessary computer code for several chemical mechanisms. Such version of EMAC, utilizing the KPP code, was initially checked on short, test simulations and in the end performed very well, giving very promising, future perspectives. The remaining stages of work include, among others, a development of a post-processor that will use the CUDA-Fortran, KPAA-generated code and an OpenACC version of KPP code that could overcome the licensing problems of KPPA.

The scalability bottlenecks of the Delft3D-FLOW module were identified after a series of small and medium-scale simulations while, in the meantime, the code was also in the process of porting to the BlueGene/Q system, FERMI. Finally, several preparatory sessions took place that have defined a workflow for the future interactive lake design sessions with several stakeholders. I the remaining year period of the project, the performance of SWAN and FLOW will be further investigated and improved. Those steps are in parallel with a development of a special workflow for the interactive lake design session between various stakeholders.

More extensive reports from the partners documenting their work on the run projects and their current state on M11 are included under the section 3. Furthermore, during the final 12 months of the project the work on the initial eight applications that started at M7 and the four from the reserve list that started at M11 will be completed according to the outlined, detailed plans.

5 Appendix

5.1 Time-line of the process

Key steps undertaken for an identification and selection process of the applications addressing socio-economic challenges and related projects, projects management and applications enabling are summarized below (covering the period from M1 to M11 of the PRACE-3IP Project).

- **Start - PRACE-3IP Kick-Off Meeting, September 2012**
  - Presentation of the activity goals and objectives
- **Contributors gathering, defining sources of input, M3-M4**
  - Intensive internal communication on the activity scope and organization
  - Definition of the work principles
- **Defining partners contributions and expertise, M3-M4**
  - Exchange of ideas and overall experts availability
- **Selection process definition, M3-M4**
  - Internal discussion on the selection principles and criteria
- **Internal call for proposals on socio-economic challenges, M4-M5**
  - Internal process collecting projects proposals on the applications enabling for selected key socio-economic challenges
  - Proposals quality evaluation with face-to-face session
- **Material for the meeting with SSC members gathering, M4-M5**
  - Internal review of the proposals for external presentation (SSC)
- **WP7 meeting with SSC members, November 21st, 2012**
- **Reporting SSC members feedback and final selection of challenges, end of M5**
  - Discussion and agreement on challenges selection process results
  - Decision on the final list of applications addressing selected socio-economic challenges
- **White Paper report on selection results preparation, M4-M6**
  - General description accompanying M6 milestone requirements
- **Preparatory Access application submission, end of M6**
- **Final list of challenges and initial version of accompanying white paper, M6-M8**
  - White Paper accompanying the M6 milestone
  - White Paper publication on the PRACE web site
- **Applications enabling, M7-M11**
  - Preparatory Access (Type B) application processing
  - Initial period of the enabling for applications addressing selected socio-economic challenges
- **Management Board Approval, February 5th, 2013**
  - Management Board approval on the selection process and final list of projects
  - Management Board decision on the four projects in the reserve list receiving support without re-evaluation

- **Projects progress evaluation, M10**
  - Project progress reporting on the face-to-face session
  - Start of the additional four queuing projects on applications addressing socio-economic challenges

5.2 Guidelines for the project proposals

Guidelines used to instruct PRACE partners on the expected contents of the enabling projects proposals are described below. A proposal template has been circulated with the following guidance:

- **Proposal Title**
  - Project should refer to an explicit socio-economic challenge
  - Proposal title is expected to emphasise the context of the challenge it is addressing
- **Contributors**
  - Contributing PRACE partner should be defined
  - Additional collaborators may be listed, e.g. members of the related scientific community
- **Key points of the proposal**
  - Proposal should describe how it addresses selected socio-economic challenge
  - Proposal abstract should focus on well-defined computational techniques for application enabling
  - Proposal should define current state and expected enabling results in terms of application performance or scalability and how it refers to the underlying socio-economic problem to be supported
- **Applications list**
  - Proposal should focus on application code or codes that have potential impact on the solution of the defined problem and benefit for the community
  - References for proposed applications are required
  - Proposal should explain the reasons for application selection
- **Enabling principles**
  - Proposal should include identification of the potential performance bottlenecks and other issues to be address with applications enabling
  - Proposal should state what enabling approaches and techniques are to be applied

5.3 Complete list of the proposals

PR1. Understanding plasma fusion phenomena

Socio-economic challenge: Safe and Environmental-friendly energy production

Main contributor: *CINECA, Italy*

In many applications involving convection and diffusion (parabolic) physical processes, explicit time stepping can become very inefficient. The project's aim is to implement hybrid implicit-explicit scheme time stepping algorithm with AMR capabilities in one of the most important and widely used multi-purpose code for plasma physics and astrophysics: the PLUTO [43] code. The code is well suited for supersonic and super-fast magneto-sonic flows in multiple spatial dimensions and provides a modular structure whereby different integration schemes can be combined together to treat diverse physical regimes.

All the improvement that will be made on the PLUTO code (and the explicit/implicit algorithm) can be extended quite easily to other similar Eulerian code for plasma physics. Thus this work would be also a test case for the explicit/implicit algorithm implementation on a multi-purpose MHD code.

PR2. Impact and optimum placement of off-shore energy generating platforms

Socio-economic challenge: Safe and Environmental-friendly energy production

Main contributor: *STFC Daresbury Laboratory, UK*

The project proposed by STFC Daresbury Laboratory in UK deals with the simulation of the placement of marine and wind turbines in coastal waters. Large-scale calculations will be performed to assess the near and long-term impact of marine and wind turbine farms on local and distant coastal waters, with special focus on sediment morphology. The performance of the three fully-coupled TELEMAC [9] modules – TOMAWAC, TELEMAC-3D and SISYPHE – will be tested at very large-scale. This has not been done before but is necessary to understand local (1-10m scale) and distant (10-100km) impact.

PR3. Electron-impact atomic data generation for nuclear fusion reactors

Socio-economic challenge: Safe and Environmental-friendly energy production

Main contributor: *STFC Daresbury Laboratory, UK*

The project's focus is on diagnostics in nuclear fusion tokamaks (e.g. JET, ITER). Related computational problems include: electron-atom electron-ion scattering calculations using R-matrix methods and solving of large-scale closely coupled differential equations based on parallel propagation methods (Hamiltonian systems, partitioning of configuration space, dense linear algebra). Calculations involve many thousands of closely coupled channels, such as those associated with Tungsten ionization (used as lining for tokamaks in ITER and JET).

The associated code suite PRMAT [44] provides highly scalable software to produce independent relativistically corrected collision data for joint data verification with the DIRAC code.

PR4. Uncertainty analysis from numerical experiments, with applications to safety of nuclear power plants

Socio-economic challenge: Safe and Environmental-friendly energy production

Main contributor: *CEA, France*

In order to assess safety parameter such as nuclear fuel temperature, numerical simulations are done with comparison to experience. The URANIE [32][33][34][35] tool uses propagation methods to assess uncertainties in simulation output parameters in order to better evaluate confidence intervals. This is used for the Verification and Validation and Uncertainties Quantification process used for safety analysis.

URANIE is well suited for launching many instances of serial codes but it lacks scalability and portability when used for coupled simulations and/or parallel codes. The aim of the project is therefore to enhance this launching mechanism to support a wider variety of applications.

PR5. Computer-Aided Drug Designing

Socio-economic challenge: Rational drug design

Main contributor: *SNIC-LiU, Sweden*

In molecular modelling, docking is a method to predict the preferred orientation between a large protein molecule and a smaller potential drug molecule. Frequently, a so-called scoring function is used to model the strength of the molecular interaction between protein and molecule. However, even the most state-of-the-art scoring functions still determine the electrostatic part of the interaction from classical point-charge models, due to substantial computational cost. Now, with the advent of cutting edge Tier-0 systems in Europe, it becomes possible to conduct a more ab-initio based evaluation of the protein-drug molecule interaction energy, for example by using electrostatic potentials from ab-initio methods. The ultimate goal of the project coordinated by Linköping University (SNIC-LiU, Sweden) is to enable and improve the LSDALTON/DALTON [10] application code for large-scale, ab-initio docking simulations. The project will investigate the main bottleneck for ab-initio electronic structure calculations on Tier-0 level systems and enable/improve the DALTON/LSDALTON code for better scaling. The project will be performed in close contact with DALTON developers and the ScalaLife Project [4], which uses the DALTON code as one of the main applications.

PR6. Electron-molecule resonance data for DNA radiation damage studies

Socio-economic challenge: Multiscale modelling of the human cells and organs

Main contributor: *STFC Daresbury Laboratory, UK*

The project's focus is on the energy that is deposited in cells by ionizing radiation, which is later on channelled into production of free secondary electrons. Reactions of such electrons induce substantial yields of strand breaks in DNA. The strand breaks are associated with electrons trapped in quasi-bound 'resonances' on the basic components of the DNA. HPC usage will enable the study resonance formation in much more detail than in current initial calculations.

The associated application is UKRmol [45], a widely used, general-purpose electron-molecule collision package and the enabling aim is to replace a serial propagator (coupled PDE solver) with a parallel equivalent module (PFARM [46]) already partially studied in PRACE-2IP.

PR7. Enabling scalable highly parallelized MD simulations with non-periodic boundary conditions and arbitrary geometry

Socio-economic challenge: Rational drug design

Main contributor: *NCSA, Bulgaria*

Molecular dynamics (MD) is widely used method in computational sciences for a broad range of applications. This project, coordinated by NCSA from Bulgaria, is focused on an adequate account for long-range electrostatics in systems with non-periodic boundary conditions and arbitrary geometry (bio-molecules, membranes, surface chemistry, epitaxial slab builds, zeolites). The aim is an essential reduction of computing time for large-scale molecular-dynamic (MD) simulations applying the fast multiple method (FMM). The enabling task is to implement a solver using the FMM method in the DL_POLY_4 [12] application code to achieve significant scaling improvement on very large complex systems.

PR8. Sustainable food production through fisheries and aquaculture

Socio-economic challenge: Sustainable food supply

Main contributor: *UiO, Norway*

Establishing and monitoring genetic diversity of cod and salmon stocks in relation to fishing pressure, detection of fisheries induced selection and enabling sustainable aquaculture by genetically assisted breeding are of high importance for many countries.

Associated methods are applicable to other species and therefore have significant socio-economic impact. Focus of the project coordinated by University of Oslo from Norway is to design and implement workflows that will scale from 1 genome to 1000 genomes and be applicable to computing genomic variety based on short sequences, statistical analysis using genome-wide data across many individuals, integration of phenotypic measurements, population dynamics and climate data and clustering genome-wide information. Workflows include a number of the following application codes: BWA[16], SamTools [47] for scalable short sequence alignment for variant detection and MUSCLE [48], ClustalW [49], RAxML [50] for multiple sequence alignment for clustering and phylogenetics. For multidimensional analysis the R statistics package [51] is considered.

PR9. Multidiscipline Simulations for Aircraft Designs

Socio-economic challenge: Future aircraft transportation

Main contributor: *JKU, Austria*

The design process within aeronautical industry very often suffers from a gap between aerodynamic design and structural design. In order to further improve existing designs, especially wing designs in combination with high-lift devices; it is inevitable to have fluid-structure coupled high-fidelity simulations available. The goal of this task is to demonstrate how HPC is a substantial contribution in accomplishing such types of simulations. Work within this task will concentrate on demonstrating a high-fidelity simulation with fluid-structure coupling based on an integrated simulation model that allows covering multi-scale analysis, ranging from detailed phenomena (boundary layer) to global properties (aero elastic deformation).

Focus of the project coordinated by JKU from Austria is the development of an efficient and robust coupling interface between the computational fluid dynamics (CFD) code and the computational structural mechanics (CSM) code. Identified associated application codes are OpenFOAM [19] for the CFD computations and Elmer [20] package. Both applications are state-of-the-art simulation systems and are available under public licenses.

PR10. Parallel Grid Generation

Socio-economic challenge: Future aircraft transportation

Main contributor: *VSB, Czech Republic*

Parallel grid generation and adaptive refining is an underlying computational step for a broad range of engineering problems such as: aeronautics, combustion, turbulence modelling, aero acoustics and biomedical flow. In particular the process is crucial for application dealing with huge unstructured grids (i.e. more than one billion cells). Any progress in automated parallel generation of such complex structures would be beneficial for a number of applications and communities. The project aims to identify tools and possible methods for extending their capabilities. Tools in scope are LibMesh [52] and Netgen [53] community libraries. The work will be done as a supplementary activity for previous PRACE achievements on engineering applications.

PR11. Big Data for Machine Learning

Socio-economic challenge: ‘Big data’ management and processing

Main contributor: *Bilkent University, Turkey*

One of the novel research areas with a wide applicability to social sciences is large datasets analysis (“Big Data”). Methodology used includes not only the data-driven approach but new tools and programming paradigms. The project explores possibilities related to Big Data and the HPC ecosystem. Aim of the project is to test the scalability of fundamental data mining

approaches used for a variety of problems and to explore/enable processing of terabytes of data on hundreds of computing nodes.

Applications identified in this area are the Pegasus [54] library for parallel graph mining and the Mahout [55] library for parallel machine learning. Both codes are open-source and use the Apache Hadoop implementation of the Map/Reduce paradigm. Project will focus on applicability of Map/Reduce on supercomputers for analysing freely available large data collections.

PR12. Multidiscipline coupling in Cardiac computational mechanics

Socio-economic challenge: Multiscale modelling of the human cells and organs

Main contributor: *BSC, Spain*

The project's aim is to improve scalability of the Alya Red [42] cardiac computational model.The underlying challenge is to understand the human heart with first electro-mechanics-fluid model scalable on thousands of CPU's. The simulation size is 220 million mesh elements and around 100,000 time steps and includes strongly coupled physics in different computational domains.

Enabling focuses on the Alya Red application performance improvement with efficient fluid-mechanics coupling. For this purpose parallel-coupling strategies to overlap fluid and mechanics computations will be developed.

PR13. Large-scale simulations of neural networks

Socio-economic challenge: Multiscale modelling of the human cells and organs

Main contributor: *ICM, Poland*

Understanding the way, how the human brain works remains one of the greatest challenges of the century. Numerical models are fundamental tools for studying neural networks and various aspects of processes in a brain. One of the significant initiatives addressing these challenges is the Human Brain Project, a European Flagship Project.

While the Human Brain Project will focus on using widely recognized applications for in-depth neural cells modelling this project is addressing usability of the available applications for simulating neural network in a scale of a human brain with PRACE HPC systems. Simplified descriptions of neural cells will be used which enables to use the desired simulation scale. The NEST [56]**Fehler! Verweisquelle konnte nicht gefunden werden.** application is in scope for enabling on PRACE systems and scalability study.

PR14. High Resolution Climate and Atmospheric Chemistry Modelling

Socio-economic challenge: Understanding of climate change

Main contributor: *CaSToRC, Cyprus*

Atmospheric chemistry has a huge socio-economic impact, affecting, e.g., climate change, human health and agriculture. Simulations help to prevent unfavourable trends and to adapt to changing environmental conditions. Identified application suite addressing these problems is the ECHAM/MESSy Atmospheric Chemistry model (EMAC) that couples the global climate model ECHAM [57] with the Modular Earth Submodel System (MESSy) [58]. The latter integrates into one software environment a large number of models describing physical and chemical atmospheric processes. Among them, the MECCA sub-model (Module Efficiently Calculating the Chemistry of the Atmosphere) is used to calculate the tropospheric and stratospheric chemistry.

Atmospheric chemistry processes are one of the most computationally intensive tasks and higher resolutions and accuracy modelling needs higher numbers of cores or acceleration. The main goals of the project, proposed by CaSToRC from Cyprus, are to take advantage of recent and forthcoming petascale machines to significantly increase the model resolution. More specifically, the objective is to exploit GPGPU accelerators on hybrid HPC systems.

PR15. Multidisciplinary modelling for interactive design of lakes

Socio-economic challenge: Natural environment protection

Main contributor: *SARA (SURFsara from January 2013), The Netherlands*

The design process of a lake and its environment needs an interactive approach in which different aspects (economical, engineering, recreational, safety for flooding, ecology) from different stakeholders can be combined. For this purpose, for Lake Marken in the Netherlands, a multidisciplinary-coupled model exists. Currently, the simulation runtime for a scenario with the model is four days and interactive sessions (that combines drawing measures with calculations effects with the model with stakeholders) are not feasible yet.

Aim of the project coordinated by SURFsara from The Netherlands is to enable interactive lake design sessions by reduction of runtime from four days to one hour. This includes parallelization improvements of the coupled models Delft3D-FLOW and Delft3D-WAQ and the coupling between the Delft3D [30] sub-models.