European Cooperation in the field of Scientific and Technical Research
- COST -

Brussels, 14 November 2014

COST 103/14

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1406: High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)

Delegations will find attached the Memorandum of Understanding for COST Action IC1406 as approved by the COST Committee of Senior Officials (CSO) at its 191th meeting on 12-13 November 2014.

MEMORANDUM OF UNDERSTANDING

For the implementation of a European Concerted Research Action designated as

COST Action IC1406
HIGH-PERFORMANCE MODELLING AND SIMULATION FOR BIG DATA APPLICATIONS (cHiPSet)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4114/13 "COST Action Management" and document 4112/13 "Rules for Participation in and Implementation of COST Activities", or in any new document amending or replacing them, the contents of which the Parties are fully aware of.

2. The main objective of the Action is to create a long-lasting, sustainable, reference network of research links amongst the High Performance Computing (HPC) and the multiple Modelling and Simulation (MS) research communities addressing Big Data (BD) problems.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 48 million in 2014 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of Section *2. Changes to a COST Action* in the document COST 4114/13.

A. ABSTRACT AND KEYWORDS

The Big Data era poses a critically difficult challenge and striking development opportunities in High-Performance Computing (HPC): how to efficiently turn massively large data into valuable information and meaningful knowledge. Computationally effective HPC is required in a rapidly-increasing number of data-intensive domains, such as Life and Physical Sciences, and Socio-economic Systems.

Modelling and Simulation (MS) offer suitable abstractions to manage the complexity of analysing Big Data in various scientific and engineering domains. Unfortunately, Big Data problems are not always easily amenable to efficient MS over HPC. Also, MS communities may lack the detailed expertise required to exploit the full potential of HPC solutions, and HPC architects may not be fully aware of specific MS requirements.

Therefore, there is an urgent need for European co-ordination to facilitate interactions among data-intensive MS and HPC experts, ensuring that the field, which is strategic and of long-standing interest in Europe, develops efficiently - from academic research to industrial practice. This Action will provide the integration to foster a novel, coordinated Big Data endeavour supported by HPC. It will strongly support information exchange, synergy and coordination of activities among leading European research groups and top global partner institutions, and will promote European software industry competitiveness

Keywords: High Performance Computing, Modelling and Simulation, Big Data, Dynamic Systems, Data Intensive Computing

B. BACKGROUND

B.1 General background

High Performance Computing (HPC) and high throughput computing underpin the large scale processing of grand challenge problems with data-intensive requirements in order to enable complex applications in distinct scientific and technical fields such as high energy physics, genomics, systems and synthetic biology, industrial automation, social and economic data analytics, and medical informatics. This has led to a substantial improvement in the understanding of diverse domains ranging from the evolution of the physical world to human societies. Application performance in HPC systems is nowadays largely dominated by remote and local data movement

overhead (network messages, memory and storage accesses). This poses new challenges to HPC modelling and programming languages, which should enhance data locality and fast data transition. When investigating the behaviour and complexity of those abstractions for large-scale Big Data systems, one employs a series of technologies that have their roots in well-funded large compute cluster environments. With the advent of hardware accelerators (GPU, FPGA), pay-by-use cloud services, and the increased performance of common processors, HPC has become an option for many scientific disciplines. This COST Action shall nurture cross-pollination between the HPC community (developers and users) and those disciplines for which e.g. the continual use of high-throughput data is a novel phenomenon. Data-intensive domains make the issue of efficiency particularly relevant for problems such as multi-dimensional and multi-level integration and model state explosion. Furthermore, these complex systems do not easily lend themselves to straightforward modular decompositions that support parallelisation and hence optimal HPC execution support. These often require a significant amount of computational resources with data sets scattered across multiple sources and different geographical locations.

Modelling and Simulation (MS) are widely considered essential tools in science and engineering to substantiate the prediction and analysis of complex systems and natural phenomena. Modelling has traditionally addressed complexity by raising the level of abstraction and aiming at an essential representation of the domain at hand, leading to a complicated trade-off between accuracy and efficiency. That is to say, the properties of a system can be studied by reproducing "simulating" its behaviour through its abstract representation. Arguably, the context of the application level should be reconsidered. For instance, Monte Carlo simulations must be fed with input data, store intermediate results, and filter and merge output data in an adjusted and reliably robust manner. Thus, MS approaches are particularly affected by the data deluge phenomenon, since they need to use large data sets to enhance resolution and scale, and distribute and analyse data in the different phases of the simulation-analysis pipeline.

Both HPC and MS are well established research areas. However, a better integration of the two, aimed at applications in various domains, will bring substantial progress in addressing Big Data problems. As the focus is on the integration, networking, and coordination of research, a COST Action is the preferred mechanism to establish a common network amongst several established research groups and scientific communities. This Action will systematically foster interconnected research on Big Data through the organisation of meetings, workshops, visits, and schools with the participation of HPC and MS researchers from Europe and overseas.

On the one hand, domain experts need HPC for simulation, modelling, and data analysis, but are often unaware of performance and parallelism exploitation pitfalls in their designs. On the other

hand, designers of HPC development tools and systems primarily focus on absolute performance measures, by definition the raison d'etre for HPC. However, MIPS, FLOPS, and speedups need not be the only measure, and domain-specific considerations ought to lead to consideration of other factors--such as usability, productivity, economic costs, and time to solution--equally, if not more importantly. By further improving collaboration with domain experts, HPC architects ought to be able to develop programming models and architectures better tailored to specific problems, and enhance analysis and validation tools for a sharpened understanding of HPC systems, which are in turn Big Data systems themselves.

It has been observed that HPC is mainly dominated by data-parallel computation, which does not necessarily demand the tight coupling of publicly-funded HPC hardware. Therefore, this Action will be actively promoted as an effective platform for recommendation and support for the best possible use of publicly-funded HPC infrastructures, commercially-provided or community-run clouds, and volunteer computing systems.

The research groups involved in this Action already represent a well-balanced set of competencies, ranging from HPC architectures through high-level parallel programming models, to a number of specific data-intensive MS domains. The proposing consortium consists of research communities and key industrial players. All will significantly benefit from the research networking and training activities supported by the Action. The flexibility of the COST framework will also facilitate the outreach toward larger research audiences, particularly through industrial partners, thus strengthening the links between academia and industry in various and diverse fields.

B.2 Current state of knowledge

HPC is currently undergoing a major change, with "exascale systems" being developed by 2020, which will be very different from today’s systems and will pose technological challenges, e.g. energy consumption and development of adequate programming models for millions of computing elements. Several current exascale research programmes therefore span a 10 to 20-year period. Major experiments depend on HPC for the analysis and interpretation of data and the simulation of models. MS have traditionally been used where the complexity of the problem makes more direct, analytical approaches unsuitable to match observation to theory. This is particularly true for Big Data problems, where the support of HPC infrastructures and programming models is essential, and includes the analysis of HPC systems themselves. The design and optimisation of Big Data HPC-enabled experiments and large scale HPC systems require the realistic description and modelling of the data access patterns, the data flow across the local and wide area networks, and the scheduling

and workload presented by hundreds of jobs running concurrently and exchanging very large amounts of data. Data-intensive (‘Big Data’) HPC is arguably fundamental to address ‘grand challenge’ MS problems.

In fact, several MS approaches are based on discrete-event frameworks due to their efficiency and scalability. MS have addressed problems such as scheduling in distributed, heterogeneous environments, economy-driven resource allocation, Big Data access in distributed environments and more generic HPC concurrent, distributed and cloud architecture. As described in the CERN Big Data HPC infrastructure, stochastic data traffic, management of virtual machines, and job allocation in data centres represent ‘grand-challenge’ HPC-related problems, which require extensive use of MS and HPC itself. Attempts to describe and analyse hardware, middleware and application co-design, an important development direction for HPC, have been made but they currently appear too costly. The complexity can be reduced by means of coarse-grained models--which need precise measures of uncertainty and associated errorsand statistical inference. Simulations have been run in this context for systems with one million cores. Recent trends aim to empower programmers to easily control the hardware performance. Examples include the embedding of HPC facilities in standard OS distributions.

From an application perspective, HPC-enabled MS has started to play a crucial role in a number of diverse knowledge domains. Preliminary proposals with direct porting of existing techniques in HPC, e.g. in climate modelling, and further developments are being sought. In computational electromagnetics, modelling problems with up to 1 billion variables have been addressed with memory and CPU intensive algorithms, sometimes solving major longstanding problems. More structured approaches based on pattern-based parallel programming effectively cater for the design and development of parallel pipelines for MS in Systems Biology, providing developers with portability across a variety of HPC platforms (multi-core, clusters, cloud) and support for massive data sets.

However, HPC-enabled MS has still not reached a fully satisfactory maturity, facing relevant problems in terms of computational efficiency and lack of generality and expressiveness when addressing data-intensive scenarios. The development of new complex HPC-enabled MS applications requires collaborative efforts from researchers with different domain knowledge and expertise. Since most of these applications belong to domains within the life, social, or physical sciences, their mainstream approaches are rooted in non-computational abstractions and are typically not HPC-enabled.

Recent surveys of the use of HPC in life sciences illustrate possible new scenarios for knowledge

extraction and the management of large scale and heterogeneous data collections with application to Medical Informatics. Valuable Big Data diagnostic applications are being developed, with the aim of improving diagnosis by integrating images and large multi-source data sets. These come at the high price of the HPC infrastructure and suffer from the lack of standard protocols for Big Data representation and processing. Once computational results are obtained, large amounts of information need domain-specific validation. For instance, in biomedical studies, wet-lab validation typically involves additional resource-intensive work that has to be geared towards a statistically-significant distilled fragment of the computational results, suitable to confirm the biomedical hypotheses, and compatible with the available resources.

Big Data is an emerging paradigm whose size and features are beyond the ability of the current MS tools. Datasets are heterogeneous, i.e. they are produced by different sources and are of a large size with high and fast in/out rates. Big Data accessibility and the capability to efficiently bring and combine data together will be extremely valuable. Currently, many HPC-enabled MS efforts have been proposed in several Big Data contexts, as diverse as performance evaluation and the management of HPC frameworks, research on blood anti-coagulants, the numerical evaluation of quantum dynamics, computational social network analysis (e.g., the relationship between Internet use and specific emotions, human obesity, or happiness) and genomic sequence discovery. Some approaches have been successful, e.g. leading to potential industrial impact, supporting experiments that generate petabytes of data, like those performed at CERN, and showing interesting flexibility. The Action will offer a unified framework for the systematic advancement of MS and Big Data endeavours supported by leading HPC-enabled models and tools. The state of the art in Data-Intensive MS in distinct domains confirms that significant improvements are still required to have a coordinated effort of HPC and MS experts. On-going research shows both palpable results and challenges. Challenges are due to the inherent complexity of the problems; in part these are intrinsic characteristics of an emerging research trend across different disciplines. The Action will therefore focus on building the missing structured connections between the HPC community and various MS groups from different domains, with the inclusion of strategic industrial partners, aiming at an architectural design that includes human, multidisciplinary competencies in the loop.

B.3 Reasons for the Action

Isolated success stories and still-to-be, potential developments coexist within the current HPC-enabled MS landscape. The capability to turn Big Data into usable knowledge has not yet developed into a fully systematic technological framework supported by a comprehensive theory. A significant

part of software used in the Big Data context is still eminently sequential, or parallelised in a naïve way. A more integrated approach has been attempted with the external help of HPC experts. Nonetheless, after several architectural difficulties, modellers had to revert to an improved version of the script-based approach. HPC architects, who might well contribute to deliver efficient MS solutions, do not always have either the incentive or the resources to develop full-scale application software. Clearly, making MS experts become fully skilled HPC architects is not a viable approach, but a stronger interaction amongst the two fields and the coordinated development of theories and tools to support the work of MS experts on efficient HPC frameworks is fundamental.
The EU itself has clearly pointed out the need and priority to adapt "existing modelling and simulation techniques, and to develop new ones, so that they scale to massive degrees of parallelism", stressing the industrial relevance of a closer collaboration between MS and HPC. That report recognises the need to make it easier and less risky for companies to invest in long term R&D into new modelling and simulation techniques, expanding their user base into new application areas. Moreover, this appears to be a global challenge: “In this time of crisis, the U.S. has the technological tools to maintain our competitive edge and global leadership in manufacturing, but we risk our manufacturing leadership position if we fail to utilize the game-changing tool of high performance computing (HPC) for modelling, simulation, and analysis”, suggesting that Europe has to play a strategic role: the race for leadership in HPC systems is already driven by the need to address scientific grand challenges more effectively and societal grand challenges are expected to become driving forces too. High-end computing has an important role in making Europe more competitive. Particularly for SMEs, access to HPC, modelling, simulation, and product prototyping services is one of the most important drivers towards true competitiveness (European Commission). The Action will arguably constitute a preeminent forum for the creation of future stable synergies amongst MS and HPC experts, from both academia and industry, in this strategic sector, allowing them to exchange ideas on HPC enabled MS for Big Data problems and to engage in projects aimed at devising unified approaches, solutions, methodologies and tools.

B.4 Complementarity with other research programmes

The Action explicitly covers the topic of integrating High-Performance Computing and Modelling and Simulation in the Big Data context. The Action participants have been actively involved in a number of relevant, complementary research projects which provide substantial background to this Action.
The recently approved Action IC1305 - Network for Sustainable Ultrascale Computing (NESUS) is

also motivated by recognising the potentially huge development opportunities that enhanced HPC architectures and programming models will provide in the Big Data scenario. However, NESUS sits in mainstream HPC rather than aiming at HPC-enabled Modelling and Simulation. Consequently, NESUS does not attempt to bring together the two communities. Fruitful interaction amongst the two Actions can be easily envisaged, from the initial sharing of information on the existing common grounds, to comparing proposed solutions and providing feedback and validation from one to the other. In order to establish such collaboration, NESUS key participants will be invited to the kick-off meeting of the Action, should it be approved. Additionally, Action IC0804 dealt with energy efficiency in large-scale distributed systems but had no real objective to bring forward a MS agenda.

Some projects have already addressed different aspects of HPC and MS, although this is on specific topics and typically lacks the required general and multifaceted approach that the Action is fostering. The topics of interest for the Action, taken separately, have been addressed by a number of EU initiatives, especially in cooperative projects of FP5, FP6, FP7, and other major nationally funded projects.

On the one hand, almost all the projects addressing HPC focus on infrastructures (PRACE DATASIM, DASH, EXCESS), middleware (SIMBIO), service-oriented systems (SIMPOSIUM), Big Data (BIG), programming models and tools (Paraphrase, REPARA, HiPEAC, 2PARMA). It is also duly noted that this Action will take advantage of the European Data Infrastructure (EUDAT) for research data access and preservation as a number of Action participants are also EUDAT partners.

On the other hand, MS research projects have exhibited a narrow scientific domain, e.g. MOSAIC, and this similarly holds for COST Actions, e.g. VISTA, where Modelling is instrumental to the main topic of the Action, with the exception of NESUS as discussed above. FP7 MMM@HPC addresses the analysis of materials by means of industry-ready MS applications, e.g. FP7 SINOXYGEN, PLEXMATH, and GREENLION use a simulation of materials and nano-materials; EGEE, SEE-GRID, FP7-CRISP that aim at efficiently supporting numerical studies for LHC experiments at CERN; the flagship Human Brain Project targets simulating the behaviour of a human brain, similarly to FP7 PD-HUMMODEL and TRANSFORM. Other life sciences national projects have employed MS to improve the understanding of distinct human physiological pathways as well as medical informatics. Social science MS including collective behaviour, social interaction, market intelligence, and traffic control have been addressed in ASSISIbf, SIMORG, EEII, CELTIC, and PLASTIC.

C. OBJECTIVES AND BENEFITS

C.1 Aim

The main objective of the Action is to create a long-lasting, sustainable, reference network of research links amongst the High Performance Computing (HPC) and the multiple Modelling and Simulation (MS) research communities addressing Big Data (BD) problems. Such links will enable a novel permanent collaboration framework across HPC and MS, covering both academia and industry in Europe and beyond. Such collaboration does not currently exist in a mature and extensive form, but is nowadays crucial to successfully address problems in the cross-discipline Big Data scenario: huge availability of raw data needs to be transformed into useful knowledge.
On the other hand, there are a growing number of new implementations of memory-demanding applications that have not yet been adapted for HPC environments, mainly because of limited communication between field experts and those with suitable skills for the parallel implementation of data intensive applications. Therefore, another natural objective of this Action is to intelligently transfer the heterogeneous workflows in MS to HPC, which will boost those scientific fields that are essential for both MS and HPC societies. Benefits will be reciprocal. MS experts will be supported in their investigations by properly-enabled HPC frameworks, currently sought but missing. HPC architects in turn will have a wealth of application domains by means of which they will better understand the specific requirements of HPC in the Big Data era. In particular, this will lead to the design of improved data-centred programming models and frameworks for HPC-enabled MS. As well as being timely and necessary, the Action is also strategic, MS being a key asset of the European ICT industry, and HPC-enabled MS a challenge in global competition.

C.2 Objectives

The Action will pursue its main objective:
To **structure** and **co-ordinate research** activity on HPC-enabled Modelling and Simulation for Big Data problems across Europe.
The Action will initially focus on the following core **themes**:

1. Enabling Infrastructures and Middleware for Big-Data Modelling and Simulation

2. Parallel Programming Models for Big-Data Modelling and Simulation

3. HPC-enabled Modelling and Simulation for Life Sciences

4. HPC-enabled Modelling and Simulation for Socio-economical and Physical Sciences

Research on the above themes will be organised in **Working Groups** (WGs). The Action will manage the WGs and their overall coordination, fostering cross-fertilisation and inclusion of early-stage researchers. The Action will strive to include new themes, whenever appropriate, and to react effectively to expected future developments in the area.

The Action will pursue other more specific objectives and will (non-exclusively):

- build an effective, durable and active working community of European researchers in the area, with a span of about 60 research institutions and companies, more than 15 COST and 5 International Partner Countries

- constantly strive to **expand the Action**'s activities to other participants acrossEurope and overseas, increasing both the number of research institutions and companies, and that of COST and International Partner Countries;

- foster the formation of **new multi-disciplinary expertise** of competent researchers exploiting HPC-enabled MS, and contribute, in particular, to the formation of the new generations of such researchers;

- disseminate obtained research results, identified best practices and developed prototypes and supporting tools;

- strengthen the **collaboration with European companies** in order to establish efficient technological transfer of the latest HPC-enabled MS techniques, methods, and tools, encouraging their industrial adoption; and,

- establish the Action itself as a **reference point of competence** that can provide advisory support to industries and can offer informed expertise to policy makers in the strategic field of HPC-enabled MS.

C.3 How networking within the Action will yield the objectives?

The Action will stimulate close collaboration among participants as the main means to achieve its objectives through its structures and activities, including:

- The Management Committee (MC) will pursue the Action **management** according to the COST Rules and Procedures. The MC will be responsible for the progress of the Action, which will be measured against given, clearly measurable, targets (including the ones described in the following).

- The **MC structure** will include a Chair, Vice-Chair, a Steering Group (SG), WG leaders, a Scientific, Dissemination and Training coordinators (Sc, Dc, Tc), as well as National Representatives. Furthermore, the MC will appoint an independent Advisory Board (AB). The MC will meet at least at the plenary meetings (4). The SG will meet at least twice per year, allowing a timely supervision of activities (See Section E for details of the management process).

- The Action will be **fully inclusive** of new partners throughout the whole project lifetime, aiming targeting to expand to more than 15 COST countries and include more than 10 industrial partners, which are active in the deployment of HPC and MS technologies. Dissemination, outreach and participants' connections will be exploited to recruit new partners.

- The Action will hold **plenary meetings**, including three annual meetings and the Final Symposium. These will be typically co-located with main conferences and include peer-reviewed workshops.

- WGs will be responsible for **research activities.** WG **coordination** sessions will be organised at the plenary meetings. WGs will also hold annual meetings in conjunction with specialised workshops. WG leaders will invite selected members of the other WGs and other researchers not in the Action to attend WG meetings for cross-fertilisation and in response to issues and opportunities.

- WGs will publish a state of the art report in the first year and then a **scientific report**, yearly. In order to foster research collaboration, the Action aims to identify key inhibitors and drivers for HPC-enabled MS in sample application areas within the first

year, publish an evaluation of current practices and requirements for future systems within the second year, and a report on best practices, concepts and system architectures geared to the application areas, within the third year.

- WGs will contribute to the development of their area of interest by annually organising **special sessions**, tutorial or discussion meetings at international conferences, and **specialised workshops** to stimulate research collaboration on emerging hot topics, when appropriate.

- Scientific collaboration and management, through SG and WG leaders in particular, will support **transnational project proposals** on topics related to the Action, with a target of 10 submissions in the last phases of the Action and within 2 years of its end.

- Scientific results spawned from the Action coordination will be published in major international peer reviewed journals and conferences, with an expected target in the region of 100 **publications**.

- The organisation of **3 training schools** will contribute to the formation and career development of cross-disciplinary researchers, with particular emphasis on PhD students, PostDocs and Early Stage Researchers (**ESRs**) in general. School organisation will also consolidate developed materials and lead to publication of reference lecture notes.

- The Action can be fruitfully paired with a **Marie Sklodowska-Curie ITN Action** to provide further training support, strengthen networking and suitably fit in the scientific plan of the Action. Involved early-stage researchers, in turn, should benefit from a well-settled, industry-academia pan-European context, a challenging cutting-edge project, and additional career opportunities. The possibility of applying to the 2015-2017 calls is under exploration by some interested Action participants.

- The Action has a target of **10 Short-Term Scientific Missions** (STSMs) **per year**, from the second year (for a total of 30 STSM). These will enable PhD students and ESRs to develop and share their expertise, ultimately helping them create new connections among established researchers.

- **Dissemination** of scientific results shall be mainly carried out via high impact peer-reviewed publications. Tutorials, surveys, and teaching material will also be published and made publically available by the Action. Other forms of dissemination and outreach will be carried out through the organisation of training and scientific events, mobility via STSMs, participation in conferences, industrial collaborations (starting from the participant companies and the tools and solutions to be developed), and, where possible, through outreach towards higher education.

- The Action will develop a **web-site**, and other collaborative and appropriate social media channels to present its overall activity and support the Action's interaction and management.

- Publication of the main achievements will be associated to the final **symposium,** which will include evaluation and forward-looking considerations.

C.4 Potential impact of the Action

This Action will strongly **support collaboration and knowledge exchange between the MS and HPC** communities and research centres across Europe. Putting together competencies, efforts and results from the various national research programmes on these topics will consolidate the internationally leading position of the European research in MS and HPC.
Specific Action benefits will be:

- To develop a conceptual framework that brings **coherence** to the whole field of **HPC-enabled MS**. This will happen by way of a bi-directional transfer of methodologies, techniques and best practices. In one direction, state-of-the-art HPC programming languages and tools from HPC experts will enhance MS methodologies and their capability to address Big Data problems. In the other direction, canonical problems in MS will influence the development of novel HPC high-level programming models, infrastructures and tools to provide the much-needed computational support to the MS domain.

- To influence standardisation bodies with the **best practices** identified through the Action regarding: problem specification languages and data exchange models for MS,

particularly in Life Sciences and Socio-economical and Physical Sciences; concurrent features of programming languages (e.g. C++, Java, R); deployment models of applications in large scale data centres; specification formalisms for distributed Big Data sets and federated architectures and systems with unified standards at the general organizational level for data intensive analytics (e.g. federated clouds based on the OpenStack standards - such technology is successfully implemented in the CERN cluster with the cooperation of Rackspace UK). **Standardisation** results are likely to enhance, in the long term, **the market value** of prototypes developed by industrial partners and commercial software in general.

- To enhance the scale, performance, precision, quality, cost and time-to-market of **MS frameworks** for **applications** in contexts such as personalised medicine within Life Sciences and smart cities within Social Sciences.

- To enhance HPC methodologies for Big Data problems will pave the way, in a broader context, towards **Software Engineering at the exascale**, considered an **enabling technology in H2020** for its scientific, societal, and industrial impacts.

- To bring together the experts in HPC programming and data intensive tools developers for mutual training and development with the public outreach.

- To **transform** the **approach to the HPC-enabled MS** by the **wider community** of researchers, developers, practitioners and users who rely on, or simply exploit, it in their day-by-day activities. This will also happen via the dissemination of results in the described forms of publications, lecture notes, reports that will be made available to such wider community, editorial activities and, where possible, by outreach activities, as outlined in the dissemination plan (see Section H).

- To **train** young researchers as **a new generation of interdisciplinary experts** in future HPC and data-intensive technologies, methods and theories, capable **of leading** multi-disciplinary teams in HPC-enabled MS.

- To allow the best-possible and wider utilisation of the **publicly funded HPC infrastructure**, commercially provided or community-run clouds and volunteer computing platforms.

C.5 Target groups/end users

The benefits of the Action will be enjoyed by the following groups and end users:

- Scientific, medical, engineering and complex systems modelling researchers and academics;
- Big Data analytics, software and visualisation companies;
- Wider European academic, industrial, and practitioner communities, with the provision of experts trained on forefront, strategic competencies, and the opportunity of attracting some of them from overseas; and
- Citizens, governments, governmental organisations and the general public that will benefit from better-informed decisions enabled by Big Data analyses and sophisticated HPC-enabled computational models.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

This Action is based on the idea that key aspects of HPC-enabled MS must be jointly addressed by considering the needs and issues posed by the two communities. When multidimensional, heterogeneous, massively large data sets need to be analysed in a specific Big Data application domain, the methods required to suitably process the data are necessarily determined by the kind of data and analysis to be performed.
Consequently, the features of a programming language, library or execution machinery supporting the efficient implementation of the analysis should not be thought of as independent of the specific data and analysis themselves. Analogously, data characteristics must drive the design and implementation of data storage systems enabling efficient storage, access, and manipulation. Within this vision, the Action will address the specific challenges of both MS and HPC in a combined way. A number of open problems that can be fruitfully investigated by the joint efforts of HPC and MS researchers have been identified according to the expertise and research interest of the Action's

participants. Given the existing links and synergies amongst participants, these topics can be readily investigated, providing an effective bootstrapping of the Action's research and, at the same time, a solid mid-term agenda. They typically sit across the Action's four themes, enforcing cooperation amongst the corresponding WGs and aggregating participant interest and work, for instance around cross-WG publications. The MC will support integration through the Action mechanisms, such as technical and managerial meetings, dissemination, STSM, training schools (see Section E). The following initial list will be extended as the Action develops.
The following initial list will be extended as the Action develops.

1. Innovative hardware and software technologies for the computing and storage resources in data centres and virtualised environments, simultaneously supporting high access performance, availability, and confidentiality of stored data. These are typically sensitive issues for the distributed storage of Big Data sets.

2. Development of novel data-aware programming models for parallel computing with a specific focus on simulation workloads. A key goal of the activity is to support both code and performance portability across different HPC platforms also for problems requiring the management of massive data.

3. Development of novel methodologies supporting simulation precision, typically depending on the quantity of raw data to be produced, filtered and mined. Examples of application domains that will be considered are:

    - Weather and climatology, with a focus on climate change and its multi-dimensional variables and “whole Earth” realistic models, which require high-resolution, data-intensive HPC simulation frameworks to accurately estimate the likely impacts on nature and society;

    - Transfer and abstraction from the experiences of major physical research centres such as CERN, FAIR and ITER in modelling and experiments for challenging astrophysics and plasma physics problems for better reuse of the developed models and technologies for solving general data-intensive scientific problems;

- Modelling of continuous and stochastic dynamics of bio-chemical phenomena, accounting for computationally challenging phenomena like noise, macromolecular crowding and spatial issues (e.g. tissue formation and tissue dynamics as in autoimmunity). Related statistical approaches to stochastic analysis also rely on the replication of experiments, naturally calling for highly-effective parallel simulation support.

4. Improvement of the expressiveness of languages for model specification. Language expressiveness needs to consider aspects of efficient data-intensive management on HPC platforms. Examples of application domains that will be considered are:

5.

- Models capable of describing the interplay between the dynamics of cell populations, e.g. proliferation of stem cells, and intracellular dynamics. While the description of multi-level systems is currently a challenging problem per se, computational efficiency also strongly limits the scaling of models, and multi-level MS;

- Modelling of physical phenomena in engineering and industrial applications would benefit from optimized and uniform modelling specification languages, tunable to specific architectures. The effort to better understand phenomena like gas turbines for aero-engines or power generation will lead to key innovations with widespread scientific and societal impacts.

6. Improvement in the integration of data-intensive processing stages, data-intensive workflows and pipelines. Application domains that will be considered include:

- Omics initiatives which are producing terabyte data sets for single experiments. One example is the Human Microbiome Project 2012: 80 research institutions cataloguing the collective genomes of the microorganisms that live in symbiosis with the human body. The analysis of these data will be used to understand collective behaviour and interactions with the human body;

- Medical records data management. These structured, semi- and unstructured Big Data are processed in data centres and their treatment and mining poses serious risks and offers a valuable opportunity. The definition of an integrated HPC analysis framework in this domain will be addressed.

D.2 Scientific work plan methods and means

Research is organised around the four main objectives introduced in Section C2, which will be addressed by 4 WGs. WGs, with the support of and under the supervision of the MC, will coordinate the scientific activities of the Action. All WGs embed HPC and MS aspects within the context of Big Data problems. WG1 and WG2 focus on HPC infrastructures and programming models for MS, respectively. WG3 and WG4 are thematic umbrellas for data intensive MS in Life Sciences and for Socio-economical and Physical Sciences, respectively. It is envisaged that along with the progression of the Action, new themes and WGs can be added or evolve from the existing ones, also according to the interests of existing and newly added participants and under the supervision of the MC and within the limits of COST Actions. Details of each WG are presented below, followed by a high-level description of their expected collaboration. Details of the WGs' coordination and organisation are in Section E2.

WG1: *Enabling Infrastructures and Middleware for Big-Data Modelling and Simulation*

From the inception of the Internet, one has witnessed an explosive growth in the volume, speed, and variety of electronic data created on a daily basis. Raw data currently originates from numerous sources including mobile devices, sensors, instruments (e.g. CERN LHC, MR scanners, etc.), computer files, Internet of Things, governmental/open data archives, system software logs, social networks, commercial datasets, etc. The so-called 'Big Data' problem requires the continuous improvement of servers, storage, and the whole network infrastructure in order to enable the efficient analysis and interpretation of data through on-hand data management applications (e.g. agent-based solutions in Agent component in Oracle Data Integrator (ODI)). The main challenge in Big Data Modelling and Simulation is to define a complete framework which includes intelligent coordination and communication, data fusion, mapping algorithms, and protocols. The programming abstractions and data manipulation techniques must therefore be designed for (a) the seamless implementation of application solutions with efficient levels of virtualisation of

computational resources (communications, storage, and servers); and (b) the effective normalisation and merging of data with dissimilar types into a consistent format (wide class of data services). WG1 activities will foster research on:

- Comprehensive survey and taxonomy of existing Big Data system architectures and middleware, including Artemis Platform for Neonatal IC, systems using RFID-based technologies, smart meters, smart passports, etc., and with a particular focus on the modelling frameworks of interest for WG3 and WG4,

- Analysis and design of Big Data system components based on the practical use cases and user requirements, particularly those defined in WG3 and WG4.

- Development of novel heterogeneous models, algorithms and techniques for advanced Big Data exploitation, based on the current and emerging multicore system architectures, virtualised servers and data centres, mobile cloud and multi-cloud systems. This objective is tightly coupled with WG2 objectives.

- Definition of a test bed (benchmark suite) and a standardised library for heterogeneous parallel processing of Big Data for life, physical, and social science applications.

- Analysis and development of the new trends in evolution of the Big Data middleware and architectures.

WG1 activities will be structured according to the following overall work plan:

- *Months 1-12:* Survey of state-of-the-art of Big Data systems and middleware, development of a comprehensive full taxonomy of such systems for life, physical, and socio-economical applications - cooperation with WG3 and WG4

- *Months 13-18:* Improvement of existing Big Data systems and middleware, and development of new components based on the analysis of use cases and user requirements defined by WG3 and WG4 - cooperation with WG2

- *Months 19-24:* Development of guidelines for Big Data provisioning and management for MS in cooperation with WG2

- *Months 25-36:* Definition of the test beds for life, social, and physical and science applications - in cooperation with WG3, and WG4

- *Months 36-47:* Development of a comprehensive compendium of the evolved state of the art and novel trends in Big Data systems software, middleware, and infrastructures for MS with special emphasis in Life, Social and Physical sciences. Industrial outreach and potential path-to-market exploration, where applicable.

- *Month 48:* Presentation of the summary of WG1 activities and research results at final Action meeting, discussion of new trends in evolution of the Big Data middleware and systems.

WG2: *Parallel Programming Models for Big-Data Modelling and Simulation*

A core challenge in Modelling and Simulation is the need to combine software expertise and domain expertise. Even starting from well-defined mathematical models, they still have to be manually coded. When parallel or distributed computation is required, the coding becomes much harder. This may impair time-to-solution, performance, and performance portability across different platforms. These problems have been traditionally addressed by trying to lift software design and development to a higher level of abstraction.
In the Domain Specific Languages (DSL) approach, abstractions aim to provide domain experts with programming primitives that match specific concepts in their domain, whereas performance and portability issues are ideally moved (with various degrees of effectiveness) to development tools. Examples include Verilog and VHDL hardware description languages, MATLAB and GNU Octave for matrix programming, Mathematica and Maxima for symbolic mathematics, etc.
In the general-purpose approaches, such as Model-Driven Engineering (MDE), general-purpose programming concepts are abstracted into high-level constructs enforcing extra-functional features by design, e.g. compositionality, portability, parallelizability. In this regard, the number and the quality of programming models enabling the high-level management of parallelism have steadily increased and, in some cases, these approaches have become mainstream for a range of HPC, data-intensive and Big Data workloads: streaming (e.g. Storm, S4, Infosphere stream, FastFlow), structured parallel programming and MapReduce (e.g. Hadoop, Intel TBB, OpenMP, MPI), SIMD (e.g. OpenACC, SkePU).

WG2 activities will be structured according to the following work plan:

- *Months 1-12:* Review of currently used programming models in the development of MS software, and the state-of-the-art parallel programming techniques in both DSL and MDE approaches (in cooperation with WG3, and WG4).

- *Months 6-24:* Quantification of data involved in selected MS applications, their access patterns, computation demand and typical workloads of MS pipelines (in cooperation with WG3 and WG4).

- *Months 12-36:* Study of usage requirements: portability, reactivity, robustness, time-to-market, maintenance, and possible end-users (in cooperation with WG1).

- *Months 24-48:* Identification of evolution of features of current programming models better matching HPC-enabled MS requirements (in cooperation with WG3, and WG4). Industrial outreach and potential path-to-market exploration, where applicable.

- *Month 48:* Presentation of the summary of WG2 activities and research results at final Action meeting, discussion of the new trends in evolution of parallel programming models and languages for HPC-enabled simulation tools.

WG3: *HPC-enabled Modelling for Life Sciences*

Life Sciences typically deal with and generate large amounts of data, e.g. the flux of terabytes about genes and their expression produced by state of the art sequencing and microarray equipment, or data relating to the dynamics of cell biochemistry or organ functionality. Some Modelling and Simulation techniques require the investigation of large numbers of different (virtual) experiments, e.g. those addressing probabilistic and noise aspects or based on statistical approaches. Curation and mining of large, typically multimedia, medical datasets for therapeutic and analytics purposes, are computationally expensive. Recent and future developments, such as personalised medicine need to integrate a mix of genomics, Systems and Synthetic Biology and medical information in a systemic description of a single individual. A surge of large-scale computational needs in these areas spans from the BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) and the flagship effort Human Brain Project, which targets simulating the behaviour of a human brain, to

FP7 projects like PD-HUMMODEL, TRANSFORM. In fact this Action integrates well with the goals of the ESFRI Roadmap, promoted by the EC. Requirements go from pure computational efficiency, to large data file management and storage capabilities and vast memory-bound computational power.
WG3 will foster the much-needed integration of HPC architects and Life Sciences modellers, with the goal of letting them develop and diffuse a coordinated, mature and productive use of HPC facilities. WG3 activities will:

- Identify general classes of MS problems in Life Sciences, e.g. omics and medical data mining and simulation in Systems and Synthetic Biology, stimulate integrated development of HPC-enabled solutions to MS problems, and evaluate the proposed frameworks.

- Encourage and monitor the publication and diffusion of results in relevant venues of both communities (novel interdisciplinary editorial proposals can be attempted, if appropriate).

- Foster collaboration pools within Europe, including academia, industry and overseas partners, for the development of integrated prototype and demonstrative tools, hence facilitating the path to market for innovative technologies and newly developed ideas.

- Coordinate collaboration for grant applications, which will support the follow-up of the activities of the WGs and will contribute to diffuse best practices in the areas;

- Make an outreach attempt towards bio and medical communities, tightening the links between computational, theoretical and experimental approaches in Life Sciences.

WG3 activities will be structured according to the following work plan:

- *Months 1-6:* Analysis of state-of-the-art and identification of relevant classes of MS problems in (selected domains of) Life Sciences. Survey of existing, cutting-edge MS execution environments - preliminary integration with WG1 and WG2.

- *Months 6-12:* Definition of selected major MS open problems and use case for data intensive applications in Life Sciences - coordination with WG1 and WG2, and WG4 in

part.

- *Months 13-24:* Analysis of the prototypical use cases, user requirements and existing models and simulation platforms - coordination with WG1 and WG2. Development of innovative MS approaches. Cooperation with WG4 on shared topics including the modelling of relevant social and economic issues in Life Sciences, like the social and epidemiological aspects of diseases.

- *Months 25-36:* Evaluation, refinement and further development of the HPC enabling solutions for MS under development in WG1 and WG2. Further development of innovative data intensive MS approaches in Life Sciences.

- *Months 37-48:* Delivery of innovative solutions for HPC-enabled MS in Life Sciences, as a result of the cooperation with WG1 and WG2, and, for specific aspects WG4. Industrial outreach and potential path-to-market exploration, where applicable. Identification of the new trends in the HPC-enabled MS in Life Sciences, and final summary/reporting of WG3 activities.

WG4*: *HPC-enabled Modelling for Socio-Economical and Physical Sciences

Many types of decisions in society are supported by modelling and simulation. Some examples are political decisions based on predictive simulations of future climate changes, evacuation planning based on faster-than-real-time simulation of tsunamis, and financial market decisions based on mathematical models emulating current market conditions. In all of these situations, large amounts of data such as global geographical information, measurements of the current physical or financial state, and historical data are used both in the model building and model calibration processes. However, also in the predictive phase, there are many applications that not only benefit from, but require HPC due to the complexity of the models, the computational volume, and the amount of data that is being generated in the simulations.
Some particularly challenging problem features are high-dimensionality (e.g. in finance or quantum physics) where the computational costs grow exponentially with the dimension, multi-scale physics (e.g. in climate and tsunami simulations) where scales that differ in orders of magnitude need to be resolved to capture the relevant physical processes, and computations under uncertainty, where the

impact of uncertain measurements, parameters and models is quantified through multiple evaluations or extended models leading to an increased computational cost (e.g. in safety critical decision problems). Especially in physics, HPC has been successfully employed for a long time. However, existing codes and algorithms are not optimized for modern computer architectures and cannot efficiently exploit massively parallel systems. Furthermore, the increase in available computer power allows for expansion of the horizon of what one can simulate, but the complexity of the systems hampers productivity and progress.

In socio-economical sciences the vast amounts of data expected from the fast growth of the internet of things will provide new challenges for the extraction of knowledge. HPC in a distributed model is going to play a major role in such activities and this Action tackles that approach.

WG4 activities will be structured according to the following work plan:

- *Months 1-12:* Identification of the general problems and analysis of state-of-the-art in physical, chemical, and socio-economical sciences MS. Survey and taxonomy of the simulation environments for solving those problems (analysis of the main criteria and usage conditions). Identification of suitable companies for cooperation and as potential new partners, where appropriate (e.g. Eurocontrol, Indra, Movistar, AMPER, FENOSA, Vodafone, local councils, commodity and transport companies, etc.) - coordination with WG1 and WG2.

- *Months 13-24:* Analysis of user and performance requirements for HPC-enabled MS - coordination with WG1 and WG2, and WG3 in part.

- *Months 25-36:* Evaluation, refinement and further development of HPC-enabled solutions for MS in Socio-Economical and Physical Sciences, building on top of the developments in WG1 and WG2. Further analysis of specific Big Data problems of interest that can be addressed by the improved HPC framework and used for their validation - mutual work with WG1 and WG2 on test-bed scenarios. Acquisition of the realistic test data acquisition from industrial partners and other interested companies.

- *Months 37-*48: Final delivery and dissemination of results, both in terms of architectural solutions, HPC-enabled MS, and successfully analysed scenarios. Industrial outreach and

potential path-to-market exploration, where applicable. Identification of the new trends in the HPC-enabled MS in for Socio-Economical and Physical Sciences Life Sciences, and final summary/reporting of WG4 activities.

Overview of all WGs' activities

- *Months 1-6:* Initial detailed planning for the Action at Kick-off and survey of the state-of-the-art and existing industrial solutions.
- *Months 7-12:* WGs cooperate on the definition of typical use cases and taxonomy of models and toolkits. WGs contribute to inviting potential new Action participants.
- *Months 13-18:* WGs build on top of exchanged information about selected models, infrastructures and technologies for Big Data management and HPC Modelling and Simulation. WGs work on the shared use cases and usage requirements (from WG3 and WG4)
- *Months 19-24:* WGs elaborate on of the prototypal models, system components and technologies, for the adopted use cases. Each WG focuses on its area of interest. WGs further explore test beds and application scenarios of interests. Check point on industrial collaborations.
- *Months 25-30:* Mid-term check-point on results. WGs work on implementation methodologies, system customization and newly developed MS frameworks for Big Data, and frameworks for the provision and management of Big Data, in Life Science and Socio-economical and Physical Science scenarios.
- *Months 31-36:* WGs further develop programming models and frameworks. Check point on the evaluation and feedback on contributed results, both in the HPC context (WG1 and WG2) and in the application context (WG3 and WG4).
- *Months 37-42:* Finalisation of results in mature and coherent frameworks.
- *Months 43-48:* Evaluation of achievements and new trends in HPC-enabled MS for Big Data. Evaluation of maturity and impact of the interdisciplinary trend created. Forward

look for future Actions and activities. Final summary and editing of WGs' final contributions and reports.

E. ORGANISATION

E.1 Coordination and organisation

This Action will be coordinated by the **Management Committee (MC)**, which will include **National Representatives (NRs)** (as standard, see Appendix A). The MC will have a **Chair** and **Vice-Chair**, who will be responsible, in particular, for the proper integration of the two HP and MS communities, enforcing the main aim of the Action starting from the leadership.
The scientific work of the Action will be organised in WGs, as described in Sections D.1 and D.2. (see Section E2 for details on their coordination). The Action will elect a **Scientific Coordinator (Sc)**, who will coordinate the WG work, interactions and knowledge transfer among all groups and will be a direct contact person for **WG Leaders** (WG leaders are also the members of MC). **WG leaders** will supervise WG activities and prepare periodic reports on the WGs' work progress. Additionally, the MC will elect a **Dissemination Coordinator (Dc)**, who will be responsible for outreach and dissemination of the Action's results, including internal publishing, web management, contacts with media and industry; and a **Training Coordinator (Tc)** responsible for the organization and coordination of STSMs and of the training schools. WG leaders and Dc and Tc may organize a sub-committee to provide support for their activities. Overall the MC will be responsible for the proper development of the Action in accordance to the COST Rules.
When possible, nominated roles within the MC will be chosen amongst the NRs, in order to keep the MC to a reasonable size and reduce impact on meeting costs. Furthermore, the Chair, Vice-Chair, WG leaders, Sc, Dc and Tc compose the **Steering Group**, which, in agreement with the MC acts as an agile management group that can promptly intervene in the day-by-day supervision of the Action. The MC will elect the Chair, Vice-Chair, Sc, Dc, Tc, and WG leaders, and will supervise the planned activities of the Action, such as meetings, STSMs, workshops and training schools. Finally, the MC will set up an independent **Advisory Board (AB)**, which is in charge of advising on the progress of the Action and may support the MC regarding corrective actions that might be needed to cope with weaknesses and threats that may emerge during the lifetime of the Action. The AB will consist of 3-5 members, including recognised experts in the covered relevant research areas, a young researcher and an industry representative.

Annual Plenary Meetings: 4 annual plenary meetings and a Final Symposium will be organised,

possibly co-located with main conferences in the area (for scientific interest and cost reduction). MC, SG and WG meetings will be associated to these plenary meetings too. All participants can exchange ideas, report on progress and plan further collaboration. The plenary meetings will also be open to participation by interested researchers and invited experts from academia and industry.

Working Groups Meetings: The work of each WG will be coordinated by the WG Leader who will participate to MC meetings, plenary meetings and SG meetings, reporting on progress. WG members will participate in a continuous process of discussion in a forum on the Action's website, by email, and by Skype audio/video calls. They will also meet individually at conferences, workshops and symposia at least once a year. See Section E2 for the specific organisation of WG meetings.

Training Schools: The Action will organise 3 training schools, around months 16, 25 and 37, conditioned to the budget availability for such initiatives. These will be aimed mainly at PhD students and ESR/young researchers. Teaching will be performed mostly but not exclusively by participants in the Action. Jointly organised training will support the team building between of the application experts who authored the software and their WG collaborators who bring in the insights for further parallelisation. Schools will be co-located with selected external events for increased visibility and to reduce travel.

Short-term Scientific Missions (STSMs): STSMs are primarily intended for PhD students and ESR/young researchers to visit other research groups in order to acquire new expertise, to contribute their expertise to particular projects, and to exchange ideas within participating institutions. STSMs will help PhD students and early-stage researchers to develop contacts and collaborators. STMSs will be organized as short visits, typically of few weeks, but exceptionally of longer duration, between participating sites. The Action will organize about 30 STSMs overall, in years 2 to 4 (about 10 STSMs per year), subject to budget availability.

All the training materials, tutorials and the detailed schedule and scientific plans for both schools and STSMs will be prepared in the first year of the Action. In this period the final Action consortium will be also created (the Action is open for new partners during the whole first year).

Webpage: The website of the Action will be set up in months 1-4 and will be used extensively for discussion in a collection of forum areas (including wiki, blog sections, and collaborative and social media, as appropriate), including dedicated space for MC and each WG. The website will also be essential for dissemination of the results of the Action, see H, and will be a repository for its the overall scientific output.

Other social communication means will be considered for supporting rapid communication and information dissemination, such as an Action Twitter account.

Deliverables and Milestones: The Action will generate the following deliverables (beyond required administrative paperwork):

- D1.i-iii (once per training school): Lecture notes from training schools
- D2.i-iv (once per year) Outreach publications
- D3.i-xvi (once per year per WG) Scientific WGs reports
- D4.i-iv (once per year) Annual reports
- D5 (month 48) Final report
- D6 (month48) Symposium scientific proceedings

The Action will progress through the following Milestones:

- M1 (month 1) Nomination: MC, SG, WG Leaders, Vice-Chair, Sc Dc, Tc
- M2 (month 1) Kick-off meeting
- M3 (months 1-4) Website and social media set up
- M4 (months 13, 25, 37, 48) 4 Plenary meetings
- M5 (months 13-24, 25-36, 37-47) STMSs (10 per year)
- M6 (month 16, 25, 37) Organisation of 3 training schools
- M7 (months 1, 13, 25, 37, 48) MC meetings
- M8 (months 7-9, 19-21, 31-33, 43-45) SG meetings at the conferences and remotely
- M9 (months 1, 13, 25, 37, 48 + once a year at the conferences) WGs meetings
- M10 (month 48) Final meeting and symposium (with proceedings)

These milestones will enable the MC, SG and AB to monitor the progress of the Action towards the achievement of its objectives, and to plan changes in the work programme if necessary. It means

that following the Annual Report or at the prompting of WG leaders, the Action Chair and Vice-Chair, AB, COST, SC may propose to the MC corrections in the Action schedule, management and implementation (and call for MC meetings, under the approval of the SG). Corrections may include among others: new appointees (including chairs, leaders), specific support actions, e.g. more/specific STSM, budget revision.

E.2 Working Groups

The Action research is then organised into the following WGs, reflecting aims, organisation and proposed scientific programme of the Action (see Sections C and D):

- WG1: Enabling Infrastructures and Middleware for Big-Data Modelling and Simulation
- WG2: Parallel Programming Models for Big-Data Modelling and Simulation
- WG3: HPC-enabled Modelling and Simulation for Life Sciences
- WG4: HPC-enabled Modelling and Simulation for Socio-economical and Physical Sciences

WG membership will be determined by the specific interests and expertise of the Action participants. WG1 and WG2 set the HPC foundations, in terms of base platform and coordination, and application programming interfaces and programming models, respectively. WG3 and WG4 provide specific applications in Life and Social and Physical Sciences, being supported by the Big Data infrastructure and associated parallel programming models set by the other two WGs. There is clearly feedback and inter-dependencies from and to the MS WGs and the HPC WGs. Scientific dependencies have been further highlighted in the detailed scientific work plan in D Section.
WG coordination will be addressed by means of the Action organisation. WG leaders, supported by the SG and the MC, are in charge of the scientific coordination of the WGs. There are natural checkpoints for WG coordination about every 6 months, considering the 5 annual plenary meetings and 2-4 other possible plenary meetings. WG meetings will be associated to events of interest that can gather participation. At plenary meetings, each WG meeting will be attended by all the WG leaders and (designated) members from all the other WGs, who will be in charge of disseminating results (especially when prescribed by the scientific work plan, as mentioned). Participants may likely belong to more than one WG, further assuring cross fertilisation in the work carried out by

distinct WGs. Research collaborations and projects are expected to span across WGs, further facilitating WG integration. Finally, media support, e.g. web, will be extensively used to circulate results, plans, issues and decision across the WGs, including WG reports.

E.3 Liaison and interaction with other research programmes

Several research programmes are in the scope of the Action's networking aims and the Action will play the role of a common contact point. Liaisons will be facilitated by Action participants who are directly involved in some of these projects, other contacts will be sought either through interested colleagues or by invitation of key people to the Action's meetings. Some EU and non-EU recent projects to be considered initially are:

- NESUS as key participants will be invited to the kick-off of this Action;
- European Data Infrastructure (EUDAT) for the access of research data access and preservation as a number of Action participants are also EUDAT partners;
- PRACE as a leading endeavour for HPC infrastructures;
- HiPEAC as a significant forum for parallel programming abstractions and tools;
- DATASIM, DASH, EXCESS, SIMBIO, SIMPOSIUM, Big Data BIG, Paraphrase, REPARA, and 2PARMA for background knowledge on HPC infrastructures, middleware, SOA, and parallel programming abstractions;
- MOSAIC and VISTA for general background on MS;
- MMM@HPC, SINOXYGEN, PLEXMATH, EGEE, SEE-GRID, FP7-CRISP, and GREENLION for MS in the physical sciences;
- ASSISIbf, SIMORG, EEII, CELTIC, and PLASTIC for MS in the social sciences; and,
- Human Brain Project, PD-HUMMODEL, and TRANSFORM for MS in the life sciences.

Many of these projects have members participating in the Actions. WGs will be responsible for identifying suitable projects and liaising with them, as well as attracting representatives of

companies with interest in the Action's research or research products.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas.

HPC is a research area with a notoriously low participation of women. In MS, the situation is a bit better, but it is still far from being balanced. This situation is addressed by different initiatives, e.g. the FP7 FESTA-Female Empowerment in Science and Technology Academia, which aims at creating a working environment that is conducive to participation of both women and men. This Action will be able to benefit from the results in FESTA through cross participation. Other best practices, e.g. the Athena Swan initiative (UK), will be considered.

This COST Action will take steps to further improve the gender balance issue by ensuring that both women and men are present in the MC and the different WGs, as well as the other groups and committees that are formed. When defining invited speaker lists for workshops and training schools, both women and men will always be included. This will be a simple but effective measure to provide a range of role models and to project inclusivity.

Early-stage researchers (ESR) will be involved in several ways, including the Training schools, which will contribute to their formation, and special ESR activities at the Action workshops and meetings. These can be special sessions with feedback by more senior researchers, individual career and research mentoring and social events aimed at networking. Furthermore, the STSM provide excellent opportunities for ESR to benefit from the broad competence within the consortium.

F. TIMETABLE

	Year 1		**Year 2**		**Year 3**		**Year 4**	
Kick-off Meeting	Mo1							
Plenary meetings			Mo13		Mo25		Mo37	Mo48
MC meetings	Mo1		Mo13		Mo25		Mo37	Mo48
SG Meetings (at		Mo7		Mo19 --		Mo31 --		Mo43 --

conferences / remotely)		-- Mo9		Mo21		Mo33		Mo45, Mo48
WG meetings	Mo1		Mo13		Mo25		Mo37	Mo48
WG individual meetings (at conferences)		Mo7-- Mo9		Mo19 - - Mo21		Mo31 -- Mo33		Mo43 -- Mo45, Mo48
STSM (10 per each year)	Mo1 – Mo12 (min. 4 STSMs)		Mo13 -- Mo24		Mo25 -- Mo36		Mo37 -- Mo47	
Training Schools					Mo25		Mo37	
Final meeting and peer reviewed symposium								Mo48
Management	Mo1 -- Mo48							
Reporting and Evaluation		Mo7	Mo13	Mo19	Mo25		Mo37	Mo43, Mo48
Dissemination	Mo1 -- Mo48							

(MoXX = XX-th month of the Action)

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: BG, DE, ES, FR, IE, IT, LU, PL, PT, RO, SE, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 48 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN

H.1 Who?

According to the EC guidelines on scientific information in the digital age (COM(2007) 56 final,

Feb. 2007) "Europe must improve the production of knowledge through research, its dissemination through education, and its application through innovation." Therefore, this Action will target three complementary audiences as follows:

Research: Propagate the Action results among the scientific community with particular emphasis on "Computing Methodologies" (based on 2012 ACM Classification system):

- Parallel/Distributed computing methodologies. This is specifically relevant to WG2, and related to WG3 and WG4.

- Modelling and simulation - Model development and analysis, algorithmic performance; Simulation types, techniques; support systems, and evaluation. This is primarily aimed at WG1 and intrinsically related to WG3 and WG4.

That is to say, in terms of research, this Action is bringing together two topics of the Computing Methodologies area of the ACM Classification System.

Education: Produce outreach training and educational materials aligned with the EC initiative on "Increasing the Attractiveness of Science, Engineering & Technology Careers". It is important to highlight that the four working groups are tackling complementary objectives in Science, Engineering and Technology. Whilst WG1 deals with engineering and technology challenges arising from infrastructures and middleware, WG2 concentrates on programming abstractions related to computer science, and WG3 and WG4 deal with a broader spectrum of scientific models from life, social, and physical sciences. Such materials will therefore increase the general visibility of the Action and, ideally, reach new audiences such as ICT professionals, commercial software developers, and, above all, the general public. It is important to highlight that Big Data has started to emerge as a buzzword in the ICT industry, and neutral informative training is required by the distinct stakeholders. On the other hand, while modelling, simulation, and HPC have a more established place among different scientific communities, educational materials with a multidisciplinary approach to link such communities can only help to reach out to additional groups in industry and society in general. These materials will be used by the Action partners consortium not only in the Action's Training Schools but also in clustering events, the different EC–sponsored CORDIS information channels, trade fairs, and other diffusion activities. Following an Open Access model, all materials will be available through the Action website. Action partners also intend to integrate Bachelor, Master, Diploma, and PhD students into the Action endeavours through student projects, courses, seminar talks, and theses.

Innovation: Directed at promoting the Action results and best practices among professional, standard, and governing bodies. Specifically, those in charge of defining high–performance computing, simulation and Big Data standards such as the HPC Advisory Council, EuroSys, the Simulation Interoperability Standards Organization, the Analytics and Big Data Committee of SNIA Europe, and the Society for Modeling and Simulation International

H.2 What?

The following dissemination methods will be used for:

1. Action participants:

- An internal, protected website with descriptions of completed, current and future activities, including working documents and publications, reports from workshops, and information about grant applications. The site will foster an Open Access model to all relevant documents and publications.

- An internet-based communication network for interaction between participants, with special focus on the needs of PhD students and early-career researchers not only in Computing Methodologies but also in general science and engineering areas which can benefit from the Action objectives.

- Short-Term Scientific Missions by PhD students and early-career researchers within the Action. Ideally, these Missions will nurture cross-pollination between working groups. As an example, a computational scientist with a focus on WG2 activities can arguably benefit from a Mission to a research or industrial group with a focus on social science modelling within WG4.

- 3 training schools (months 16, 25 and 37) for PhD students and early-career researchers.

- 4 annual plenary meetings which will serve as an open forum to all Action participants as well as a peer-reviewed symposium.

- Publications in international peer reviewed journals and conference proceedings as a result of the scientific projects and annual meetings of the Action, annual reports by the Working Groups and by the Management Committee.

- A final report that describes the outcomes and successes of the Action.

2. Other Computer/Computational Science researchers (WG1 and WG2):

- Publication of state-of-the-art reports, annual reports, case study reports, workshop proceedings, software documentation, and the final report.
- A public website with general information about the Action as well as a repository of publications by the Action.
- The annual workshops, and the two training schools, organised by the Action.
- Tutorials and Articles in peer-reviewed scientific and technical journals, conferences, and symposia.

3. HPC, Modelling and Simulation, and/or Big Data software developers/practitioners (WG3 and WG4):

- Trials and training of novel programming languages, tools and methodologies developed by the Action.
- Release of open-source software tools for system analysis, based on the research carried out by the Action. The developed software should be available at commonly used public source code repository such as github.com for a possible extension of the collaboration with the external (to the Action) research and industrial partners.
- Encouraging PhD students within the Action to apply for internships with companies who can benefit from software produced by the Action.
- A public website with general information about the Action as well as a repository of publications by the Action.

4. General public:

- Overview articles in general science and technology or national funding agencies publications emphasising the impact of the Action results.

- Non-technical electronic notes, presentations, posters, and electronic/paper brochures freely available on the public website.
- Participation in events aimed at public understanding of science.
- Integration with Linux distribution.

H.3 How?

Crucially, the Action will strive to use and exploit all Action results through Open Science/Data practices, i.e. open access publication, open access to data repositories, and open-source software development. Aligned with EC regulations the Action will nurture a balanced support to both 'Green Open Access' (immediate or delayed open access that is provided through self-archiving) and 'Gold Open Access' (immediate open access that is provided by a publisher). In particular, Action participants will endeavour to:

- deposit peer-reviewed articles into an online repository located at the Action website;
- make their best effort to ensure open access to articles within 6-12 months as recommended by the EC. The Action will also reach out to the OpenAIRE project to inform decisions on Open Access; and,
- store code and data to in publically accessible repositories in order to foster reproducibility of research results.

Additionally, the Action foresees cross-pollination with well-established European efforts. Namely, the Network of Excellence on High Performance and Embedded Architecture and Compilation (HiPEAC), the Partnership for Advanced Computing in Europe (PRACE), any other relevant ongoing and forthcoming FP7 projects (e.g. funded FP7-ICT-2013-11-4.2 Scalable Data Analytics) and Horizon 2020 projects.