The Scientific Case for Computing in Europe 2018-2026

by the

PRACE Scientific Steering Committee

To Whom It May Concern

Brussels, 18 October 2018

Subject: Third Edition of The Scientific Case for Computing in Europe (period 2018-2026)

Dear Reader,

I am very pleased to present The Scientific Case for Computing in Europe 2018 to 2026, prepared by PRACE Scientific Steering Committee (SSC), an autonomous and independent advisory body for the PRACE Association.

This third edition continues a tradition established by the HPC in Europe Taskforce (HET) at the beginning of PRACE and supports the vision that the PRACE infrastructure will enable high-impact scientific discovery and engineering research and development across all disciplines in Europe. Thanks to the high level of expertise and tireless commitment of the members of the SSC, this publication provides a clear overview of the achievements of computational science across various domains, and is an excellent advocate for the investments required for state-of-the-art scientific computing to bring further benefits to society.

The importance of the Scientific Case was expressed by the PRACE Council during its 31st meeting in Brussels on 18 October 2018, by endorsing the scientific content of the document. At the same time, the PRACE Council received the recommendations included in the document by the SSC, and undertook actions to analyse if and how these could be taken into account to further fulfil PRACE's mission. We are thankful to the SSC for providing us these insights and we are proud to base our actions upon them.

Finally, I would like to mention and thank all those who made this publication possible in particular Professor Erik Lindahl, Editor-in-Chief, and Professor Sinéad Ryan, SSC Chair, who made this achievement possible.

I wish you an inspiring read,

Th. Lippert

Prof. Dr. Dr. Thomas Lippert
on behalf of the PRACE Council

The Scientific Case for Computing in Europe 2018-2026

The Scientific Case for Computing in Europe 2018-2026
www.prace-ri.eu

Editor-in-chief
Erik Lindahl – Professor of Biophysics, Science for Life Laboratory, Stockholm University & KTH

Authors and Editors
The Members of the PRACE Scientific Steering Committee compiled and edited the chapters.

Producer
Marjolein Oorsprong – PRACE Communications Officer

Proofreading
Harry O'Neill, Insight Publishers

Graphic Design
Crystal Mather, Insight Publishers

Published by Insight Publishers, Bristol, UK

The Partnership for Advanced Computing in Europe (PRACE) is an international non-profit association with its seat in Brussels. The PRACE Research Infrastructure provides a persistent world-class high performance computing service for scientists and researchers from academia and industry in Europe. The computer systems and their operations accessible through PRACE are provided by 5 PRACE members (BSC representing Spain, CINECA representing Italy, ETH Zurich / CSCS representing Switzerland, GCS representing Germany and GENCI representing France). The Implementation Phase of PRACE receives funding from the EU's Horizon 2020 Research and Innovation Programme (2014-2020) under grant agreement 730913. The EXDCI-2 project has received funding from the EU's Horizon 2020 Research and Innovation programmed under grant agreement no. 800957.

Published: October 2018
ISBN: 9789082169492

Table of Contents

Executive Summary

Applications of computing underpin all aspects of our lives. It is a cornerstone of academia and industry where everything, from fundamental physics and climate research to self-driving cars and financial trading, depends on algorithms to deliver solutions that are accurate, fast, safe and cheap. Data is driving a scientific revolution that relies heavily on computing to process, analyse, and translate information into knowledge and technological innovations. Simultaneously, computing is undergoing a tectonic change with the end of Moore's law for hardware and extensive deployment of accelerator technologies where traditional modelling is increasingly complemented by data-driven approaches and artificial intelligence. We are witnessing a revolution in humankind's ability to solve complex problems by relying on the synergy of advanced algorithms, data, and hardware. The US, China and Japan are making great strides in these frontiers, and we call attention to the urgent need for an expanded European advanced computing infrastructure to cover a broad range of applications, education and workforce training.

European researchers are leading in algorithm and software development in many fields, which has helped create a thriving computational community. However, the present infrastructure's compute capacity, as well as insufficient investments in algorithms and software, are currently the major bottlenecks to European leadership, applications and industrial impact. Moreover, enhanced synergies between scientists working on hardware, algorithms, and applications is required for advancing the frontiers of science and industry in Europe for the benefit of its citizens.

PRACE and EuroHPC are expected to invest in an ambitious programme for promoting such synergies and a critical aspect of this will be the promotion of technical staff to curate algorithms and software and their effective development on HPC infrastructures. The PRACE Scientific Steering Committee has formulated this third version of the scientific case in relation to a number of areas of major societal relevance, and identified both success stories and breakthroughs that will be possible with investments in next generation infrastructure:

- A world class European computational infrastructure will **Expand the Frontiers of Fundamental Sciences**, extending and complementing experiments. Researchers will be able to simulate formation of galaxies, neutron stars and black holes, predict how solar eruptions influence electronics, and model properties of elementary particles. This will explain the source of gamma-ray bursts in the universe, advance our understanding of general relativity, and help us advance quantum chromodynamics that govern the properties of matter. This fundamental research advances the state-of-the-art of scientific computing and helps attract new generations to science, technology, engineering and mathematics.

- Simulations are critical in **Climate, Weather, and Earth Sciences**. Exascale resources will enable sub-kilometre resolution instead of 10km, better mathematical models, and ensembles of simulations for uncertainty quantification. This will extend the forecasting ability, it will enable researchers to include soil and ocean effects into global models, and by quantifying the accuracy, climate simulations will increasingly be able to predict the efficiency of climate actions. Next-generation weather forecasting will provide local prognoses and longer lead times, with substantial financial impact, and researchers will increasingly be able to understand, image, and predict why earthquakes happen.

- In **Life Sciences & Medicine**, bioinformatics will have tremendous impact for personalised medicine. Researchers are already able to rapidly identify genetic disease variants, and it will become possible to identify diseases that are caused by combinations of variants, with treatments tailored both to the patient and state of the disease.

Structural biology will increasingly rely on computational tools, allowing researchers to predict how the flexibility and motion of molecules influence function and disease. Deep learning techniques will provide more specific diagnosis and treatment plans than human doctors, making medical imaging one of the largest future computing users.

- For **Energy** applications, the oil and gas industry is moving to full waveform inversion combined with neural networks for accurate detection, and exascale resources will make it possible to use full wave equation models. While still applied to fossil fuels, this will significantly reduce CO_2 emissions and air pollution compared to that produced by coal, which is still the dominant source of energy in the world. Simulations are of similar importance to improve the efficiency of nuclear power, hydropower, wind turbines and, not least, batteries and high-voltage cables to enable transmission and storage. In the longer term, magnetohydrodynamic simulations of plasma are critical for fusion energy in the ITER project.

- Computing is already used widely in **Infrastructure & Manufacturing.** Engineering applications based on fluid dynamics, combined with orders-of-magnitude-faster resources, will enable direct numerical simulations of the Navier-Stokes equations for better accuracy and significant fuel savings e.g. for cars and airplanes, while also helping us understand phenomena such as cavitation. New data-driven approaches will enable scientists in academia and industry to integrate all aspects of design in models, use information from from internet-of-things sensors, include uncertainty quantification in predictions, and consider the entire life cycle of a product rather than merely its manufacture.

- **Chemistry & Materials Science** will remain one of the largest users of computing, with industry increasingly relying on simulation to design, for example, catalysts, lubricants, polymers and liquid crystals. Traditional electronic structure and molecular mechanics methods are being complemented both with multi-scale models and data-driven approaches using deep learning to predict properties of materials. This will enable researchers to fulfil the grand challenge of designing and manufacturing all aspects of a new materials from scratch, which will usher in a new era of targeted manufacturing and so enable European industry to compete on know-how rather by having to reduce salary-related costs to make it more competitive.

- Advanced computing will increasingly be applied to handle **Complexity & Data**. Existing fields are starting to use deep learning to generate knowledge directly from data instead of first formulating models of the process, while next-generation infrastructure must be able to handle these applications with drastically increased data storage and I/O bandwidth capabilities. This will enable computing to be applied in a whole range of non-traditional areas such as the humanities, social sciences, epidemiology, finance, promoting healthy living, determining return-on-investments for infrastructure by considering behavioural patterns and, not least, in helping to develop society and secure democracy.

- With the break-down of Dennard scaling since 2006 and the end of Moore's law in sight, it is critical that infrastructures not only cater to existing applications but address the need for advances in **Next Generation Computing** with new algorithms that scale better and completely new approaches to solve mathematical problems in more energy-efficient ways. This will have significant impact on potential future technologies expected a decade or more from now, including, for example, quantum computing, neuromorphic computing, RNA computing and, not least, computing to help in understanding how the human brain itself computes.

The PRACE Scientific Steering Committee recommends:

- To be internationally competitive, Europe must build a strong joint computational infrastructure and community instead of loosely-coordinated national organisations. We must rectify the under-investment in computational infrastructure in both hardware, algorithms, and software. As these investments also compete with funding for academic research projects, it is essential that this joint computational infrastructure and community is highly cost-efficient, and that it provides a long-term vision and strategy to deliver the resources that academia and industry both require. Access to the infrastructures should be based only on excellence and technical readiness. It is imperative that the scientific peer review programme developed in PRACE is maintained and extended.

- PRACE has served the HPC community in Europe well. However, it is important that changes are introduced to ensure that significantly increased resources are steered by scientific impact. Internationally-respected scientists and industrial representatives must be directly involved at the highest level of governance of future infrastructures, not merely involved to provide advice. The governance of EuroHPC/PRACE must include academic and industrial directors, and the holders of both these positions and board chairs should have demonstrated world-leading expertise and scientific records, as is the case in other major infrastructures.

- The infrastructure must reflect diverse needs for architecture, bandwidth and network. While some applications will perform well as a single simulation on an exaflop machine, most achieve outstanding efficiency with ensembles of many simulations or high-throughput analysis. Data-driven research is becoming critical, and there is convergence where both need access to large infrastructure resources. Simulations generate unprecedented amounts of data that has to be analysed, and petabytes of information generated by data-focused research needs storage, extreme I/O, and accelerator processing power for training. A leading infrastructure has to cater for all these scientific needs.

- To use resources efficiently, a larger share of investments must be spent on software, algorithms and education – effectively the human side of the infrastructure. All fields of research are facing a bottleneck where access to programming and development expertise is a major obstacle to addressing the challenges described in this report. The H2020 Centre-of-Excellence action is one promising approach, but to achieve European leadership, resources must be reserved for infrastructure staff to ensure strong engagement in the development of compiler technology, new algorithms suitable for accelerators and many-core processors, and the development of the most widely-used codes. These are exceptionally talented experts and they are in high demand. To retain them it is critical that they see that the academic and infrastructure environment provides a competitive career path for them.

Acknowledgements

The production of the scientific case would not have been possible without numerous scientists, from both academia and industry, who have contributed with their time, facts, opinions and text as well as working on the editorial content. For this, we are particularly indebted to (in alphabetical order) Jean-Claude André, Marina Bécoulet, Sacha Brun, Luke Drury, Claudia Filippi, Zoltan Fodor, Laura Grigori, Heiner Igel, Frank Jenko, Sylvie Joussaume, Dimitri Komatitsch, Petros Koumoutsakos, Núria López, Ignacio Pagonabarra, Mike Payne, Matej Praprotnik, Luciano Rezzolla, and Claudio Zannoni. We would also like to acknowledge participants from the EXDCI project for sharing contributions as well as participating in a joint meeting in Brussels, and the EC Centres of Excellence for their valuable input. Katia Soto and Marjolein Oorsprong in the PRACE AISBL office as well as Susanne Lewis from ETH Zürich have patiently assisted with logistics and organising meetings and presentations, and we are also grateful to everyone at PRACE and in other organisations, who have shared their ideas. Thank you.

On behalf of the entire PRACE Scientific Steering Committee,

Erik Lindahl
Scientific Case Editor

Sinéad Ryan
SSC Chair

Computing drives Science – and Science drives Computing

It is only about 80 years since Konrad Zuse, Alan Turing, and John von Neumann introduced the concept of automatic machines that can manipulate symbols to solve problems through mathematical algorithms, which led to the birth of computer science. No other modern research field has had such an impact on our lives: two thirds of all humans carry at least one computer in their pocket, and children today grow up in a world where they cannot imagine not having instant access to a Library of Alexandria of information, not being able to publish their own ideas, not have access to education online or not be able to video call anyone around the world.

The reason for these advances is that science has become interwoven with computing over the last half-century. The fundamental physics discoveries advancing the frontiers of knowledge of the universe could not happen without immense amounts of computing. The fluid dynamics and adaptive control problems that were unimaginable challenges a few decades ago have become standard parts of airplane design, and formerly science fiction-like ideas of having computers predict weather, diagnose disease or recognise natural spoken language are all applications we now take for granted. What these advances have in common is that at one point they were all considered absurdly difficult and far beyond the capabilities of mathematics, models and available computers – but they became possible to solve when a large number of individuals invested decades of effort into using computing to model problems more difficult than anybody had imagined before.

The Partnership for Advanced Computing in Europe (PRACE) has a mission to enable high-impact scientific discovery and engineering research and development across disciplines, and the Scientific Steering Committee (SSC) is tasked with providing input into long-term strategic goals and visions that will benefit both science and society formulating the scientific case for investing in computing. This may seem a futile effort since it is virtually impossible to predict exactly what results will have the largest impact. However, in considering a longer-term view, it is astonishing how scientific computing has consistently not only delivered the envisioned impact but far exceeded the wildest expectations of scientists too. It is a humbling, but fitting, illustration that the previous version of the PRACE Scientific Case did not even consider the importance of high-end computing for the self-driving vehicles that might soon be the largest disruption to the transport sector in a century.

Just as it is the role of science to push the boundaries of knowledge, the case for scientific computing is focused on advancing the boundaries of computing and providing a vision for what will become possible with new algorithms and computers that are orders of magnitude more powerful than those of today. However, the importance and impact of computing on society is universal: High-performance computing (HPC) is not an isolated type of computing, but involves advanced resources being used to solve some of today's most demanding problems – that in a decade or two will be handled on laptops, phones, or the internet-of-things. For Europe to be a global leader in science, computing and technology, it is imperative to formulate an integrated, long-term vision and to have flexible budgets focused on the entire spectrum of computing and not only on, for instance, hardware or services. As important as fast computers are, gradual improvements in performance can easily result in incremental scientific advances, while virtually all significant advances have been realised by new generations of algorithms and methods that were suddenly enabled by new types of hardware, memory or storage. One such example is how the extremely parallel stream processors that first became popular in graphics cards resulted in a revolution for machine learning and artificial intelligence – which in turn would probably not have happened without researchers quickly and unexpectedly adopting these new architectures for scientific computing in chemistry and fundamental sciences.

The exponentially increasing advances of scientific computing can easily be taken for granted, but these accomplishments have only been realised because of substantial, and ongoing, investments in research and infrastructure. Computing has revolutionised and enabled our intellectual capacity to tackle complex problems, it is changing how we use our brains, and it is transforming the arts, sciences, and every aspect of our society. Advanced computing and machine learning are used to plan, optimise transportation and the storage and processing of our data, and they are instrumental in our efforts to analyse the human genome, to foster the resilience of urban areas, and contribute to the evolution of society. Our growing and ageing population requires the provision of food, clean water, energy, personalised health care and accessible education, while society needs to optimise its natural resources while simultaneously moving to a green and circular economy, which is no easy feat. Computing has become extremely valuable in how we access and handle the vast amounts of data required to enhance all of these processes and procedures.

Merely using computing will not magically provide the answer to all humankind’s challenges, but there is a consistent track record which demonstrates that the countries, regions, and businesses which invest in computing achieve better, faster, and more thorough outcomes including improvements in human health, competitive advantages in industry, increasing academic innovation resulting in spin-off and start-up companies, and a more modern education for an advanced knowledge-based economy – and all this is rooted in academic research supported by world-class infrastructure.

Rather than organising a disciplinary atlas of all scientific computing research, the following chapters in this report will highlight a handful of broad areas of current research and societal challenges where scientific computing is of paramount importance, and where academia and industry would be able to deliver specific breakthrough achievements with increased European investments in infrastructure. While the needs in these areas are diverse, there are also many common trends that we summarise, followed by specific recommendations for the current and long-term development of advanced computing in Europe.

Expanding the Frontiers of Fundamental Sciences

What did the universe look like in the fractions of a second after the Big Bang? What is the nature of the dark matter that makes up more than a quarter of the universe? Can we explain the nature and interactions of the visible matter in terms of the elementary building blocks? Why can solar flare eruptions lead to satellites failing? Some of the most fundamental questions about the nature of the universe, its evolution from the Big Bang and the matter that fills it are described by physics at extreme length scales – from the galactic to the sub-atomic – and the research has been revolutionised by advances in numerical simulation and data analysis. Answering these questions tends to be exceptionally computationally demanding and data-intensive, which not only puts them at the frontier of science, but frequently also at the forefront of HPC use and development with far-ranging impact. This includes curiosity-driven science not undertaken with direct applications in mind, but history has shown that posing and answering these questions resonates with society (the recent discovery of gravitational waves being a great example), it attracts young people to Science, Technology, Engineering and Mathematics (STEM) careers, and leads to many unexpected applications. Magnetic activity in space can have large impact on electronics, the local information-sharing needs at the European Organisation for Nuclear Research (CERN) resulted in the development of the World Wide Web, and current large-scale physics experiments such as the Square Kilometre Array (SKA) are pioneers of data science.

Astrophysics

Astrophysics and cosmology have unique impacts on society and the public at large since they are focused on existential questions about the origin and evolution of the universe following the Big Bang, dark matter and dark energy, the nature of gravity, chemical enrichment of the universe, formation of galaxies, the life cycles of stars and planets, the generation and role of magnetic fields, and the vibrant question of whether we are alone in the universe. Most of these cannot be addressed by direct experiments, but only answered with the aid of sophisticated computational simulations and the processing of extraordinary amounts of data from observational facilities or missions (e.g. LSST, EUCLID, JWST, Pan-STARRS, Solar Orbiter, PLATO, and Juice). Advanced computing is as essential to modern astrophysics and cosmology as the telescopes of the European Southern Observatory (ESO) or the space missions of the European Space Agency (ESA) but it lacks the European coherence, vision and stability that ESO and ESA bring to their domains. Despite this, Europe is a world leader in key areas.

Many simulations in astrophysics use e.g. combined fluid dynamics and kinetic approaches to understand the interstellar medium, galactic bodies and, not least, magnetohydrodynamics, light and gravitation. In spite of significant progress, it is still not possible to reach realistic Reynolds numbers, but next-generation HPC resources will make it possible to model both large-scale turbulence and small-scale magnetic properties in the same systems.

Potential breakthrough: As an example, the Sun's intense magnetic activity (space weather) has a large impact on life on Earth, but we still know surprisingly little of solar physics. Developing accurate models to simulate and understand why solar flares happen can be critical to avoiding electronics failures, in satellites or electrical grids, for example, and as recently as 2017, researchers at the University of Helsinki published some of the most accurate space weather simulations to date[1] with the aid of PRACE resources that enabled them to develop a new, more advanced code (Figure 1). As our world becomes increasingly dependent on technology and electronics, a

[1] Palmroth, et al. Ann Geophys. 35, 1269-1274 (2017)

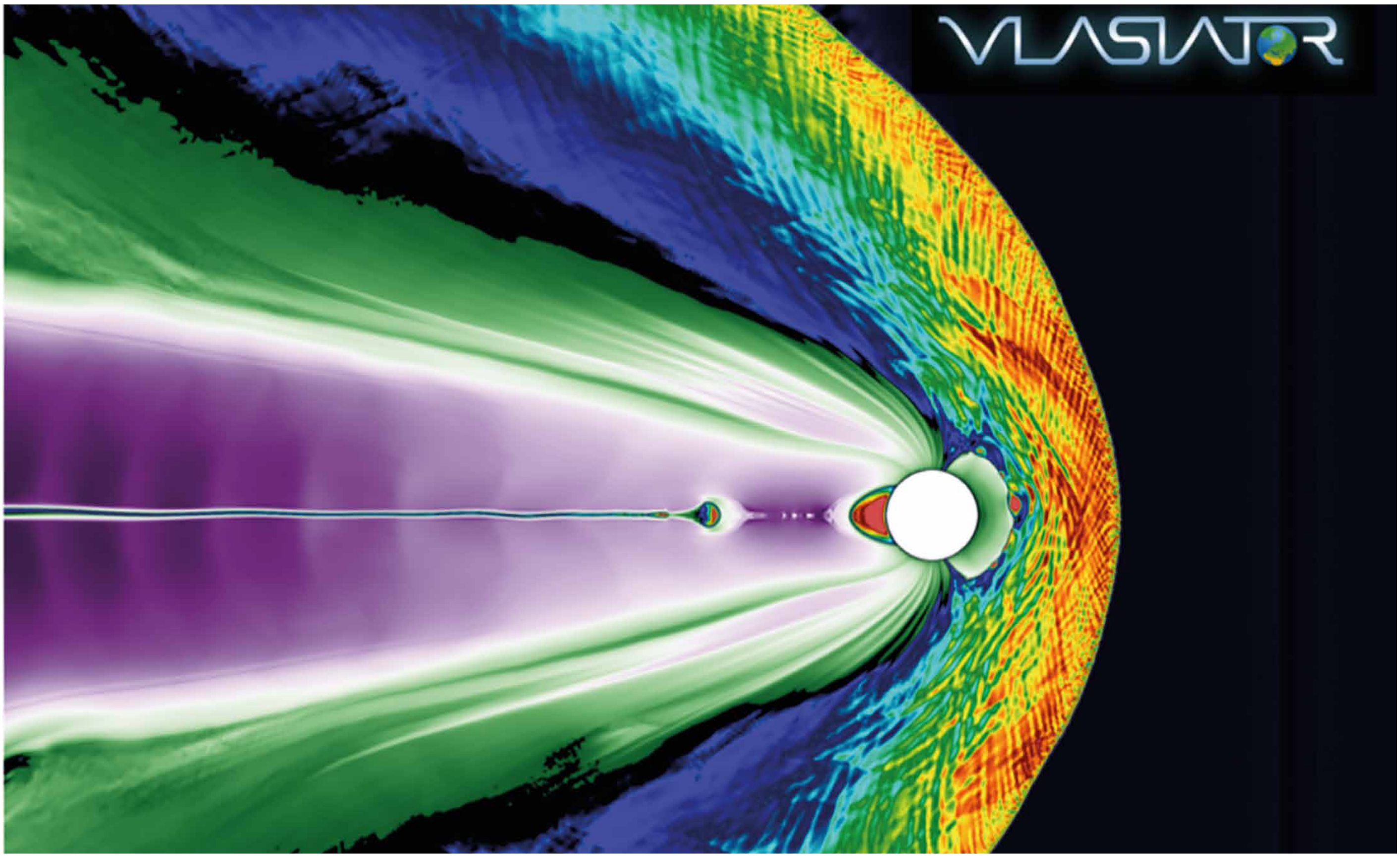

Figure 1: Simulations of "space weather" are important to understand and in the future increasingly predict how solar eruptions lead to magnetic storms that lead to electronics failures e.g. in satellites or power grids, and ERC and PRACE resources have enabled new generations of more advanced simulation codes.
Image: Minna Palmroth.

solar storm like the one recorded in 1859 would have devastating effects on society – so developing methods to predict them will have direct commercial applications.

For cosmological structure formation, some of the most sophisticated and large simulations have likewise been produced by European collaborations. Numerical relativity is an essential technology for gravitational wave observations, for example the spectacular first identification of a neutron star merger event in 2017, where European groups had leading roles. Perhaps less exciting but just as fundamental is the forthcoming third data release (adding to the already published first and second releases) from the ESA astrometric mission GAIA[2], that will provide tens of terabytes of data with unprecedented accuracy for almost two billion objects, including parallaxes, proper motions, photometry and spectrophotometry. This is already transforming our knowledge of the structure and formation of our own universe, and none of it would be possible without advanced computing and data analysis.

General relativity provides a precise formulation for the interaction between matter and the underlying spacetime which contains it. Einstein himself observed that the equations are essentially nonlinear wave equations, by which geometrical distortions of the spacetime propagate at the local speed of light. Because the coupling of gravity to matter is extremely weak, precise measurements are required in order to detect gravitational waves. It is only within the last decade that the required technologies have become available and implemented in large-scale

[2] http://sci.esa.int/gaia/

interferometric detectors such as LIGO and Virgo. Operating and tuning the detectors was a great challenge, and the actual detection of gravitational waves from noisy data, leading to the birth of gravitational-wave astronomy just a few years ago, would not have happened without advanced algorithms, software, and computing that made it possible to determine differences in the 4km mirror spacing smaller than 1/10,000 of the diameter of a proton. The first direct detection of gravitational waves from a binary system of black holes merging[3], in 2015, that the detection of gravitational waves is not only technologically possible, but that it is also a route to extracting physical information otherwise not available. Even more recently, in August 2017, LIGO and Virgo recorded the signal from the inspiral and merger of a binary neutron-star system[4]. The correlated electromagnetic signals recorded by about 70 astronomical observatories and satellites provided striking confirmation that such mergers can be associated directly with the observation of short gamma-ray bursts. This marks the birth of multi-messenger gravitational-wave astronomy and provides important clues to solving the long-standing puzzle of the origin of short gamma-ray bursts in the universe.

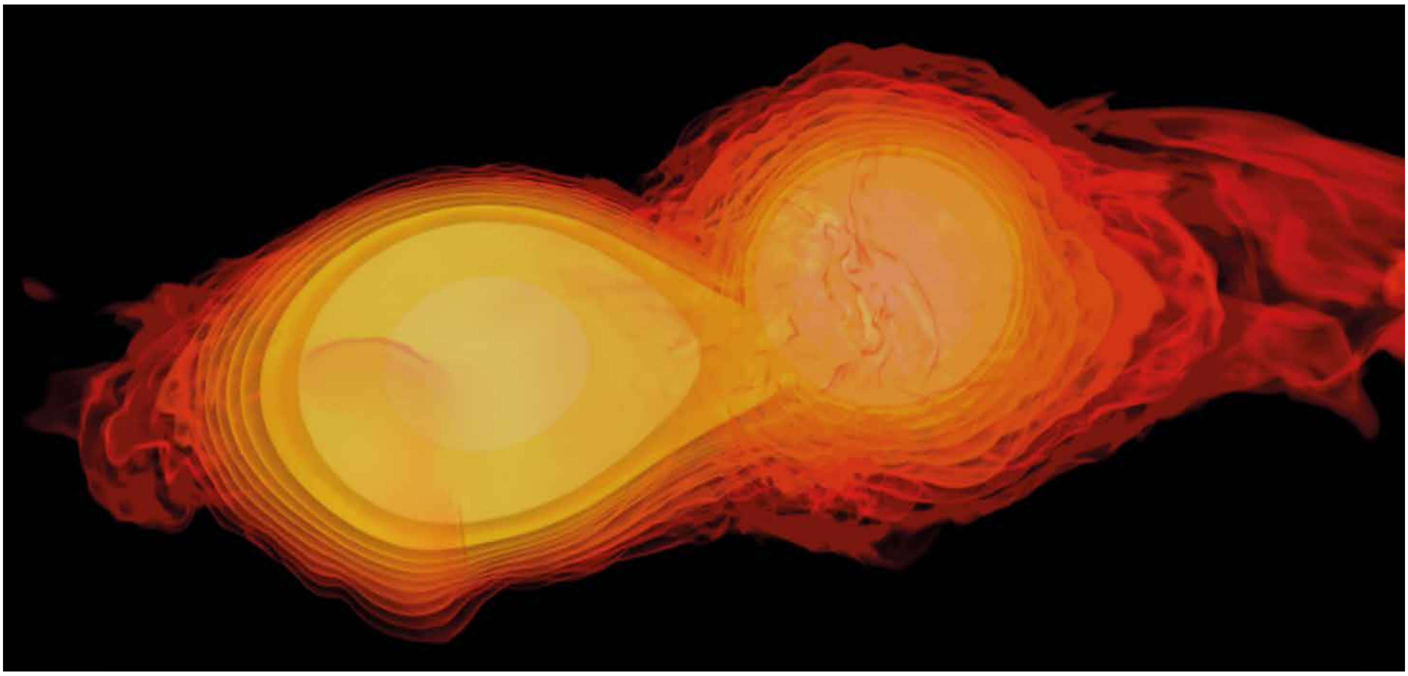

Figure 2: Visualisation of a simulation of binary neutron stars of unequal mass during the inspiral and final collapse. Advanced computer simulations make it possible to perform science on time and length scales in the universe completely inaccessible to experiments. Neutron stars are unimaginably dense objects (1017 kg/m3), and next-generation computing infrastructure will e.g. enable researchers to understand the character of matter at these densities corresponding to nuclear particles, and how the similar inspiral and merger of black holes give rise to short gamma-ray bursts observed in the universe.
Image: Luciano Rezzolla.

[3]LIGO Scientific Collaboration and Virgo Collaboration, Phys. Rev. Lett 116, 061102 (2016)
[4]LIGO Scientific Collaboration and Virgo Collaboration, Phys. Rev. Lett 119, 161101 (2017)

Potential breakthrough: New numerical methods and better treatments of the microscopic as well as macroscopic physics of compact objects will enable simulations with more accurate descriptions of the dynamics of black holes and neutron stars (Figure 2), but this will require order-of-magnitude-larger computational resources. This will make it possible to explore the vast space of parameters that characterise these systems (mass, mass-ratio, and spin) to sub-percent precision, which will result in highly accurate descriptions of their dynamics. Researchers will be able to determine the composition and radii of neutron stars with unprecedented accuracy and provide long-awaited information on the state of matter at nuclear densities. In addition, it will be possible to explore the dynamics of the massive object produced by the merger and understand how the interaction of a black hole with an accreting hot and dense torus can lead to the generation of short gamma-ray burst. Reaching this goal will lift the veil on a process that has remained obscure for almost forty years since the discovery of these phenomena.

The development of new numerical methods and algorithms will play a crucial role in this success. Exascale machines will enable new methods for the simultaneous solution of the Einstein equations, relativistic hydrodynamics, and magnetohydrodynamics (MHD). One useful example is the progress in new-generation codes implementing so-called “discontinuous Galerkin methods” that are attractive due to excellent wave-propagation properties and scalability. This allows the propagation of smooth linear and nonlinear waves over long distances with little dissipation and dispersion errors, which is particularly well suited for the solution of the Einstein equations and relativistic hydrodynamics. Since the data naturally only depends on its nearest neighbours no matter what order of accuracy is targeted, the algorithm enables highly efficient parallel implementations on exascale resources with millions of processor elements.

Particle Physics

Computing plays a similarly critical role in high-energy particle physics from accelerator design, storing and analysing experimental data, to theoretical predictions of properties of matter that feed searches for new physics. At the Large Hadron Collider (LHC), the data requirements are rapidly expanding: In 2017, CERN passed a milestone with 200 petabytes of data permanently archived. The current experiments are producing unprecedented amounts of new data – 73 petabytes in 2016 alone. As these volumes increase, the complexity of storage, access and data analysis must also evolve so that science potential is not limited, and doing so will likely advance the fundamental state-of-the-art in data science too.

One of the largest applications of supercomputing resources in particle physics is quantum chromodynamics (QCD), which is the theoretical framework for the interactions of the fundamental building blocks of all visible matter – called quarks and gluons. These particles are never found free in nature but are inextricably bound inside e.g. protons and neutrons – a process called confinement. This property causes QCD to behave very differently to the better-understood quantum theory of electromagnetism (QED) and in turn this means that numerical simulation is needed to provide the most precise and robust predictions of the properties of matter. This is accomplished with lattice QCD, an enormous computational challenge where space and time are discretised on a four-dimensional grid that must simultaneously be large enough to hold the particle of interest but also fine-grained enough to accurately reproduce the continuum physics. This field is one of the largest users of scientific computing worldwide, with several international consortia coordinating simulations as progress is limited by the available computational resources.

The lattice QCD community has sophisticated software suites, developed for a range of hardware. In the past it has played an important pathfinder role including the development of several generations of special-purpose hardware in the Array Processor Experiment (APE) project, up to the apeNEXT, extreme-scale parallelisation in collaboration with IBM for the Blue Gene systems, and the field has been an early adopter of disruptive technologies including e.g. GPU computing where it has played important roles in developing general algorithms suitable for this type of hardware.

Rapid improvements in algorithms and increasing amounts of available resources have led to dramatic scientific progress in recent years. One example is how the stability of atoms and the entire nature of our universe as we know it depends on neutrons being more massive than protons – and yet we did not for a long time know why this is the case (Figure 3). In 2015, Borsanyi and co-workers published a ground-breaking lattice QCD simulation of the neutron-proton mass difference, with all systematic uncertainties under control[5]. This research was not only able to reproduce the tiny measured difference in mass (merely 0.14% of the average of the two masses), but the lattice QCD calculations showed for the first time how the difference is a result from competition between electromagnetic effects and isospin breaking, which means we can now understand the mass difference from fundamental properties of quarks and their interactions, and it includes a number of new predictions that challenge new rounds of experiments.

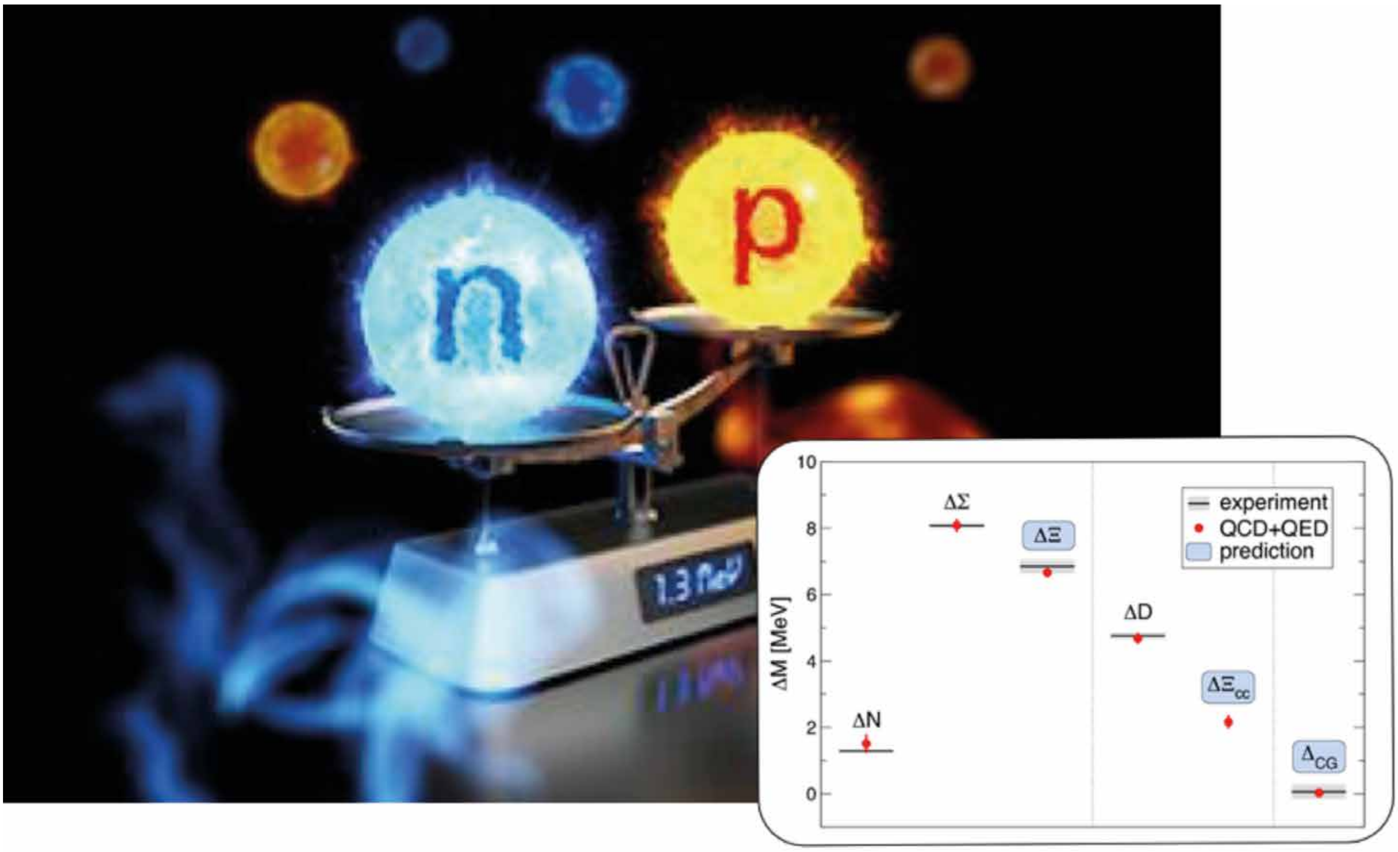

Figure 3: As recently as 2015, Borsanyi and co-workers were able to use massive lattice QCD simulations supported by PRACE and the European Research Council to successfully explain the origin of the tiny 0.14% relative difference in mass between neutrons and protons, which is the entire reason atoms are stable and our universe looks the way it does. This is due to a delicate competition between electromagnetic and isospin breaking effects, and from these simulations we now understand the stability of atoms from the first principles of quarks and their interactions. Future infrastructure with 1000-fold more powerful computers will enable researchers to extend this to other hadron particles and volumes closer to physical size to test physics beyond the standard model.
Images: Top left: University of Wuppertal, Bottom right: Science 347, 1452 (2015).
© Science 347, p. 1452, year 2015, DOI 10.1126/science.1257050

[5]Borzanyi, et al. Science 347, 1452-1455 (2017)

Potential breakthrough: In the next 5-10 years, it will become possible to perform most calculations of hadronic systems at or near the physical quark masses with large volumes - 6 fm3 and larger, to provide precise calculations of particle physics quantities that are crucial in searches for new physics (e.g. the QCD contribution to the muon's anomalous magnetic moment, g-2), and to extend these first-principles numerical techniques to nuclear physics simulations in terms of quarks and gluons. New algorithms and future architectures will similarly play crucial roles in this progress, and large efforts in software development will be required to tackle e.g. simulation methods for finite-density QCD which exhibits a sign-problem arising from a complex Boltzmann weight, handling the signal-noise that is currently an exponential problem, and the rapidly-growing computational complexity in matrix traces and propagator contractions particularly as problem scale and precision increases. Addressing these challenges will require an increase of at least 1000-fold compared to today's computational resources, but it will transform modern physics as we know it.

Summary

- Fundamental science deals with many problems that are simply not accessible to experiments, which leaves theory and computer simulations as the only tools. The computing needs are typically bleeding-edge, with many applications scaling well and using accelerators, and there are several high-impact scientific challenges that it will be possible to solve in the next few years already with an order-of-magnitude increase in resources.
- Handling large data sets is an increasing challenge both for large data-collecting infrastructures and many simulations, requiring new approaches, but this is also an important driver of research in modern data science itself.
- Computational research in fundamental science has historically been among the earliest adopter of new technologies, bespoke hardware and accelerators, as well as new types of data-focused science.
- Some of the greatest advances of the last few years – and expected future ones – are due to improvements in algorithms. To capitalise on these developments, it is critical that a significant proportion of European investment into advanced computing is focused on programming expertise and application support as well as the underlying numerical mathematics.

Climate, Weather, and Earth Sciences

High-performance computing plays a crucial role in climate modelling, weather forecasting, oceanography, and solid Earth sciences. The fields have a common background of dealing with exceptionally complex data as well as many chaotic processes, but recent advances in ensemble simulations and uncertainty quantification have had a groundbreaking impact on the quality of predictions, which now take the form of quality-controlled probability distributions taking both measurement fluctuations as well as parameter and model limitations into account. This type of simulation is finally making it possible to improve the understanding of processes as well as address important societal challenges associated with prediction of near to longer term changes under natural and anthropogenic forcing. Some of the key future challenges are related to including both the physical process and the biogeochemistry into account and coupling different parts of the climate system, moving towards so-called Earth system models. Climate is the most inclusive application, encompassing the atmosphere, oceans, land and the cryosphere, and it has the advantage of being a more stable physical process compared to the short-term weather forecasting at fine spatial resolution. However, for climate, weather and oceanography, the accuracy of the results is directly related to the spatial resolution, and exascale systems that enable kilometre-scale resolution would be able to explicitly resolve both eddies in the oceans and deep convection in the atmosphere. Similarly, solid Earth science like earthquake studies is limited by the size, resolution and frequencies that we can model today.

Understanding and Predicting a Changing Climate

Climate models are needed to understand climate variability and climate change under natural and anthropogenic forces. While there are still many challenges remaining to further improve the models, it is critical to recognise that the existing models have repeatedly proven their ability to reproduce past changes within their claimed accuracy, as well as predicting the current changes in the world – the science is largely settled. Climate models have been critical for providing information on climate change, with the current paradigm increasingly shifting towards providing reliable climate information at regional scale for adaptation.

One of the recent European success stories from this research field is the largest-ever computational resource allocation in PRACE to the UPSCALE project in 2012, which enabled a team led by Luigi Vidale from the National Centre for Atmospheric Science (NCAS) to produce a unique new ensemble of global atmosphere simulations covering a quarter of a century at resolutions ranging from 130km down to 25km by scaling simulations to tens of thousands of cores[6]. This has become an invaluable resource for the community to understand how different future anthropogenic emissions will change climate and pathways of storms, for example (Figure 4), and and it is also an excellent example of the benefits of providing open access to the resulting data. The resources required to complete the project would have been very difficult to secure in any national allocation system.

[6]Hawkins, et al. Nature Climate Change 2, 574-575 (2012

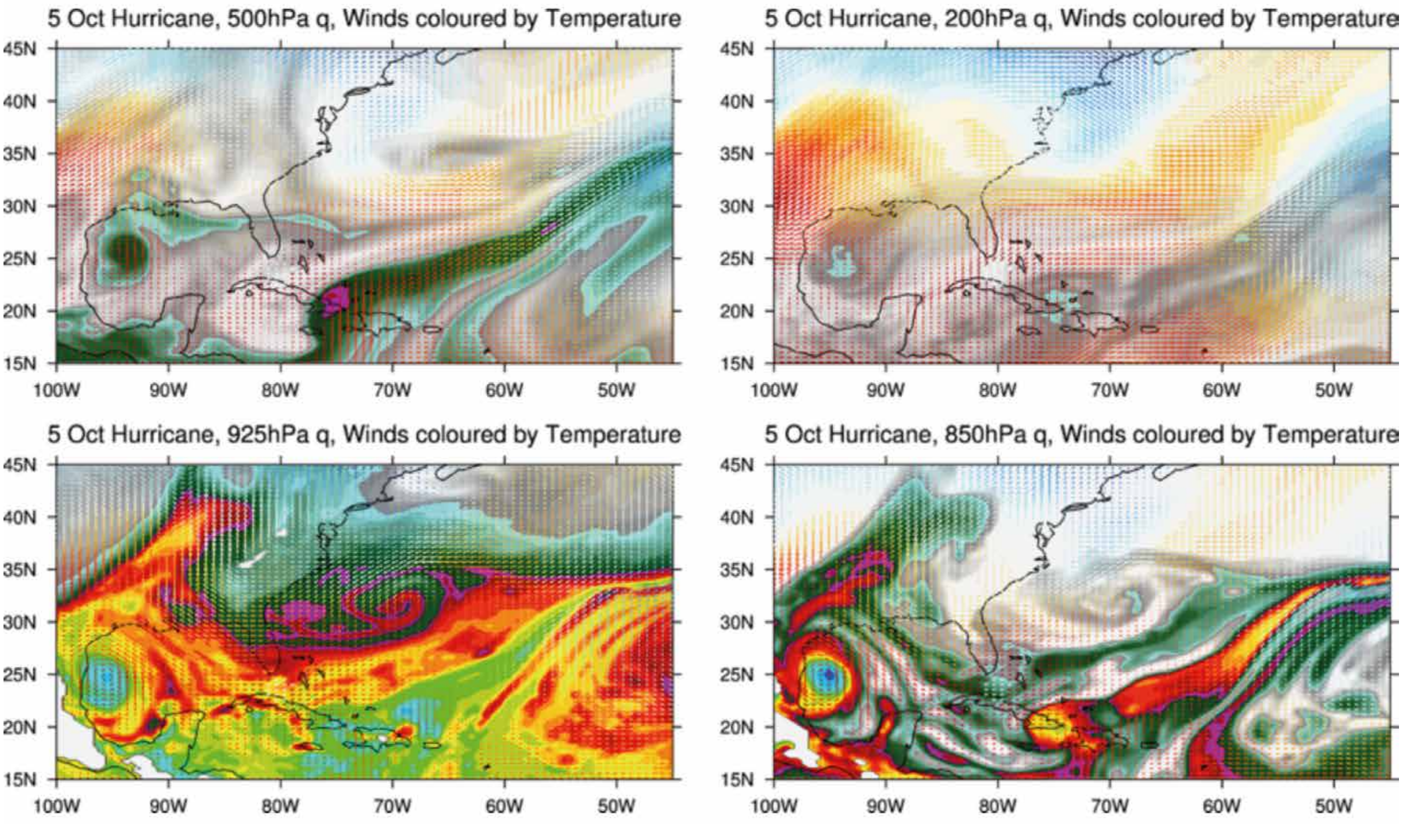

Figure 4: Animations of example twin cyclones in the Atlantic from the UPSCALE project, where one evolves into a full cyclone and the other moves north with reduced strength. Community-wide efforts and massive allocations in PRACE are enabling researchers to understand how climate change will affect e.g. affect future storms, for example, and publicly available data means these precious sets can be reused for a number of analyses and new models.
Image: Luigi Vidale/University of Reading.

Potential breakthrough: As emphasised within the European Network for Earth System modelling (ENES) 2017 infrastructure strategy[7], much better climate models are needed to address scientific challenges such as climate sensitivity and the role of cloud and carbon feedbacks, climate predictability (in particular of extremes), and representation of water cycles – but also to understand the dynamics of past climate changes such as glacial-interglacial cycles. While model predictions of global changes have become increasingly accurate and there is now little doubt the CO_2 level in the atmosphere is influencing our climate, there is an acute need for better understanding of how to best mitigate and address climate change. What types of actions will have largest effect? What areas in the world will be most sensitive to flooding, droughts or extreme temperatures? What safety margins should we apply when constructing new buildings close to water? Providing high fidelity in regional scale models will require a substantial increase in computing power, both to enable increasing spatial resolution, complexity, and ensembles of simulations to reduce the uncertainty. Similarly, improvements in the fundamental models requires much longer simulations, especially critical for modelling past climate to understand and tune models. These are typically community-wide efforts, and another major challenge is that the community is dependent on continuous access to high-performance computing resources. Even all the

[7]Joussaume S., et al. ENES Report Series 2, 20 (2017)

currently available computational resources at a European level fall far short of these needs and, given the competition for allocations, this is creating a significant bottleneck for climate research in Europe. The field has an urgent need for an order-of-magnitude expansion in the infrastructure.

Preparing for future architectures is highly challenging for climate application which relies on a number of important legacy codes. Community approaches are needed to share developments and to investigate new approaches such as domain specific languages and separation of concerns between the science and the HPC optimisation. However, despite many code development challenges, the field will be capable of utilising exascale computational resources with ensemble models that use a number of similar simulations with slightly different parameters or initial conditions to quantify the uncertainty of predictions. In terms of infrastructure, the exabyte data challenge is an even more severe bottleneck than the exaflop one. As the quality and number of measurements and amount of simulated data has increased, storing, distributing and analysing it has become a severe impediment. The community has only begun to address these issues, and better ways to efficiently process data remain to be explored in a much more detailed and complete manner. This will require more exchange, and possibly more organisation, both between research groups and researchers/infrastructure. From a climate data analysis perspective, a paradigm shift must be envisioned in the Earth System Grid Federation (the current international infrastructure for this field) that will handle both compute and data. Different approaches based on data-intensive facilities running high-performance analytics frameworks jointly with server-side analysis capabilities should also be explored.

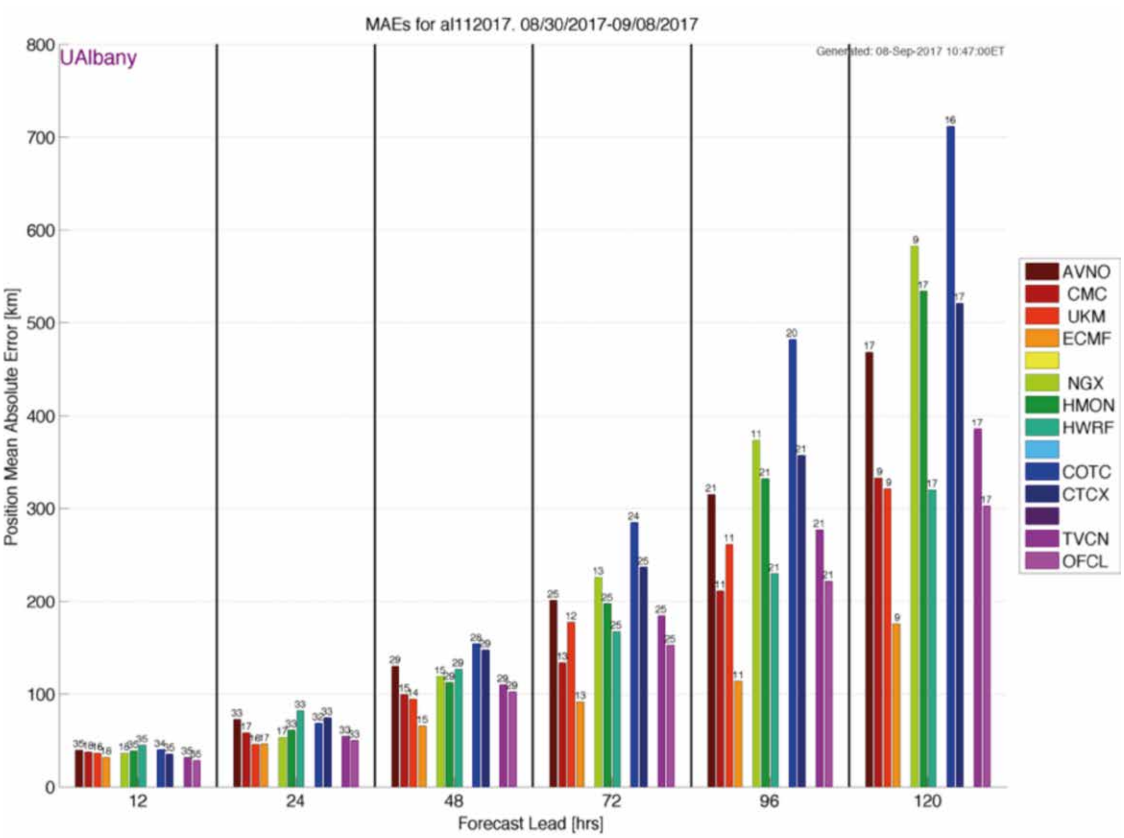

Figure 5: Mean error in position of cyclone trajectory predictions as a function of the forecast lead for a range of codes, where the ECMWF prediction model outranked the competition on all prediction scales in fall 2017. Faster computers and more data will make the models even more complex and likely necessitate global collaborations. It is of paramount importance that infrastructures worldwide provide adequate long-term support to code development.
Image: Brian Tang/University of Albany.

Accurate Weather Forecasting and Meteorology

Numerical weather prediction (NWP) is undergoing a fundamental scientific revolution as the current atmosphere-focused models evolve towards Earth system models with much enhanced detail of processes in the ocean, on land, sea-ice, snow, and regarding atmospheric composition. This evolution brings NWP models closer to the physical complexity already covered in climate prediction models but at much higher spatial resolution and using ensembles for predicting forecast uncertainty. A fundamental difference with climate is that NWP production schedules are exceptionally tight as the socioeconomic value of forecasts is strictly tied to the timely availability of forecast data. However, this also brings the societal value into sharp relief: while advanced computing infrastructure can amount to a substantial investment, this pales in comparison to the costs of making an evacuation order of a major city due to a potentially approaching hurricane. In this respect, it is important to note that while Europe has not been leading in infrastructure hardware investments, the European research software is internationally leading. One example is the accuracy of hurricane trajectory predictions, where the mean position error in the 120-hour forecast was recently assessed to around 180km for the simulations from the European Centre for Medium-Range Weather Forecasts (ECMWF), while no other model was below 300km (Figure 5).

Potential breakthrough: Increasing the accuracy and range of weather forecasting would directly translate to immense societal value that far outweighs the costs of the computational infrastructure and research. It will improve crop yields in agriculture, it will provide longer advance warnings for adverse weather conditions such as thunderstorms disrupting air traffic, and numerous businesses will gain a range of benefits. To achieve this, one of the single most important goals is to move to finer grids. By using models developed by the COSMO consortium, combined with state-of-the-art portable acceleration and massive parallelism, it has already been possible to reach ~1km grid resolution with 80 layers for some local European forecasts, but achieving fast enough forecasts for even finer grids, more layers, larger regions and longer forecasts will not be possible without exascale computing infrastructure (Figure 6). While there are many important methodological issues left to resolve, this is an area where the software is already capable of scaling to the largest resources available today, and many of the research goals could be achieved almost instantly when expanded infrastructure becomes available.

Given the increase in model complexity and the much-enhanced user communities feeding NWP model output into downstream services (e.g. for regional predictions, hydrology, air quality and marine services), data volumes are increasing rapidly, and they are expected to continue to do so in the future. This requires investment in efficient post-processing and data dissemination tools. Greater model complexity, spatial resolution and ensemble sizes are expected to increase both compute and storage needs by two to three orders of magnitude. This growth will need to be addressed along the entire forecast production chain, from observational data dissemination and pre-processing, the generation of initial conditions using complex mathematical algorithms and the forecast model, the forecast production itself, to the post-processing of model output and product dissemination.

[8] http://www.cosmo-model.org

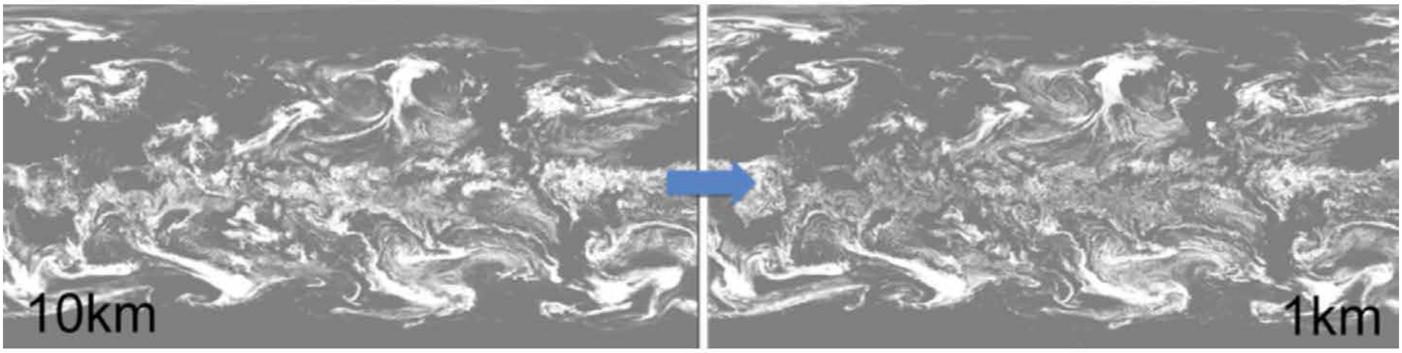

Figure 6: Moving from 10km (left) to 1km (right) grid resolution global models as well as more atmospheric layers for both climate simulations and weather forecasting will enable improved accuracy and longer forecasts, which in particular for weather prediction translates directly into concrete financial benefits for society and industry. However, while the software has already made tremendous advances, it still comes at extreme computational cost, and these benefits will not be possible to realise without European exascale resources and close collaboration between scientists and infrastructure staff.
Image: The ESiWACE Centre of Excellence in Simulation of Weather and Climate in Europe.

Oceanography

Progress in ocean science is linked to the computing power available because of the need for high resolution in models and data assimilation systems. Furthermore, these equations need to be coupled to wind, waves, tidal motion and marine biogeochemical numerical models, altogether leading to about 100 state variables. The latter are empirical numerical features and optimisation algorithms for the parameter values need to be developed for different parts of the world's oceans. Operational oceanography is a new and rapidly growing sector, providing key assessments for coastal water quality, fisheries and marine ecosystems, offshore industry, military operations, maritime transport, and more. The advent of satellite measurements of sea level (altimetry) and of the global Argo array of profiling floats has led to major breakthroughs by overcoming the historically sparse data coverage in the ocean.

Potential breakthrough: Ocean biogeochemistry is still undersampled and coupling of oceanic data with hydrological models is in its infancy, with uncertainties associated with the biogeochemical cycles. Wind waves and sub-mesoscales drive variations in the upper ocean vertical motion that affect nutrient supply and thus ocean biota, and these are not yet resolved by global or regional models. Another concern is that general circulation models do not yet consider tidal forcing, so they cannot resolve the mixing processes induced by tides that are not possible to parameterise. Coupled current and wave modelling has greatly advanced at the theoretical level in the past 15 years, but there are still no large-scale models coupled to surface waves and only a few coastal models. These concerns will be addressed by recent advances in ocean modelling that include more processes and multivariate data assimilation systems as well as higher-resolution and more accurate numerical models, but it will require order-of-magnitude-larger computational resources and substantial investments in code development.

Solid Earth sciences

Earth sciences cover a number of areas including understanding ground motion, studying the effects of potential earthquakes, modelling the structure of the Earth's interior and explaining the magnetohydrodynamics of how the planet's magnetic field is generated. A key issue is to better quantify uncertainties and estimate the probability of extreme events through simulation. For some problems, the underlying physics is adequately understood and the main limitation is the amount of computing and data capabilities available. Much larger computational resources and data-handling capabilities are also required to advance understanding, for example modelling and simulating earthquake dynamics, rupturing processes coupled with high-frequency radiation in heterogeneous soil or imaging based on high-frequency body waves.

Potential breakthrough: In the absence of deterministic prediction methods, the forecasting of earthquake ground motion based on simulation of scenarios is a key challenge in Earth sciences, and a promising tool to mitigate earthquake-related hazards. However, this type of forecast is still much more difficult and shorter-notice than weather prediction, which is a major reason why earthquakes still lead to severe and sudden losses of both life and property. The probabilistic model of earthquake recurrence has recently been challenged with the discovery of transient deformation processes called "slow earthquakes". Better understanding of these is urgently needed, which requires systematic analysis of large volumes of data recorded by seismic and geodetic instruments. Another important issue is to improve physical understanding of rupture processes and seismicity, which will require large simulations together with data assimilation and inversion. Coupling the modelling of earthquakes with high-frequency radiation in heterogeneous media will remain an exceptionally grand challenge even with the next generation of computers. Merely increasing the frequency above 5 Hz would mean a 64-fold increase in problem size. There is similarly an urgent need to enhance simulations, imaging, and improve realism by incorporating fundamental physics related to subsurface soil behaviour (up to 100x higher complexity). High-resolution models are required to develop and assess fast operational analysis tools for quasi-real-time seismology and early warning systems. Accurate simulations must span from metres near the earthquake source to hundreds of kilometres across the entire region, and from hundredths of a second to capture the greatest impact on buildings to hundreds of seconds for the full event. This will need at least an order of magnitude more resources, and inverting the problem to achieve prediction, for instance in imaging, will require repeated simulations of the forward problem, needing another 1-2 orders of magnitude of compute. Although the codes are typically limited by memory bandwidth and data handling rather than floating-point operations, this means ensembles of earthquake simulations will easily utilise a large fraction of future exascale resources.

Summary

- Research in climate, weather and Earth sciences requires next-generation computational resources, along with new types of infrastructure capable of handling, analysing and disseminating vast amounts of data either measured or generated from simulations.
- The societal and financial impact of such research is immense and goes far beyond the cost of the infrastructure: computational models are critical to reducing the impact of climate change by finding the most appropriate and cost-efficient counter-actions, and better weather forecasting is of paramount importance e.g. in agriculture.
- The software in the field is diverse, and ranges from legacy codes in urgent need of modernisation and improved scaling to state-of-the-art GPU-accelerated extremely parallel programmes. However, both require continued investments to maintain, port, and optimise in addition to implementation of new algorithms.
- In addition to a few codes with exceptional scaling, there is an acute need for resources to enable large ensembles of independent runs with different parameters to quantify the uncertainty in models. The field will easily be able to use exascale computing, provided future infrastructure is organised to handle this type of ensemble simulation.

Life Science – Improving Human Health

Life science is one of Europe's most competitive research areas and, in particular for biomolecular research, computing is a critical resource for high-impact results and commercial applications. New tools have transformed research in life science, in particular biochemistry, biotechnology and bioinformatics. In many cases, it is no longer possible to have a high-impact paper accepted without including computational studies. BioTeam Consulting made headlines in 2015 with a prediction that 25% of life science would require HPC – this came true and updated numbers forecast 50% by the end of 2018 , which likely makes it the fastest growing field in HPC. Life science, biology and medicine are in the middle of a revolution in which DNA/RNA sequencing and new generations of measurement devices are producing petabytes of data, and accurate models paired with fast computers have suddenly made it possible to model a whole range of phenomena from molecular level to cells and time-resolved tomography imaging of patients. We are seeing the emergence of new integrative computing approaches such as the Virtual Physiological Human project and others aiming to combine simulations, models and data to integrate all aspects related to the body, health, environment, and disease to investigate the body as a whole. Combined with advanced visualisation, this is of particular relevance for next-generation medical training, where students will be able to work both on realistic cases and perform a large number of virtual dissections before seeing their first live patient.

Bioinformatics & Systems Biology

The single most important trend over the last 10-15 years is the drastically reduced cost of genome sequencing, and the associated rapid growth of sequence databases. As of mid-2018, there is data for more than $3x10^{12}$ bases in 650 million sequences, and this is doubling roughly every 18 months. These sequences are useless without advanced bioinformatics methods to turn sequence data into biological knowledge. A key realisation in the 1990s was that the size of the databases is the most important factor to improve sensitivity and accuracy of predictions, which is why the recent growth has also lead to exponentially increasing impact of bioinformatics. New methods can detect coupled evolution of different residues to predict the structure of proteins involved in disease, identify what other proteins (from 20,000 in the body) a key disease target might interact with, or identify genetic variants in patients that increase the risk of e.g. heart disease, hypertension, diabetes, or various tumours. The need for high-performance computing in bioinformatics is exploding; in merely two decades it has moved from running sequence analyses in web browsers or desktops to being the fastest-growing HPC field in many countries (coupled with similarly large demands for storage and data analysis), and the continued advances in experimental sequencing means it will likely also soon be one of the largest, especially given the potentially dramatic impact it could have on health care.

The success of the high-throughput approach in sequencing is now mirrored in several other areas. New techniques such as RNA sequencing can monitor the expression of different proteins as a function of time and place in the cell. High-throughput proteomics can identify molecules, metabolomics unravels the entire spectrum of chemical processes, and automated semantic text analysis of scientific papers can be applied to create advanced mathematical models of higher-level biological processes in the entire organism with a top-down data-

[9]https://www.hpcwire.com/2017/01/04/berman-charts-2017-hpc-trends-life-sciences/
[10]http://www.vph-institute.org

driven approach. One example is how the epidemiology of the 2009 A/H1N1 influenza pandemic could be traced in close to real time by quickly sequencing and then calculating phylogenetic trees of virus samples. Systems biology is already applied to optimising processes in biotechnology, and the concept of networks describing functional coupling is becoming a powerful tool to understand how different components and cells interact. This is critical to understanding why particular genetic variants induce disease, to identifying what components in the network can be used as targets for drugs or gene therapy-based treatments, and not least in helping us understand how our genes interact with the environment and why sensitivity to various types of diseases depends so much on the individual (Figure 7).

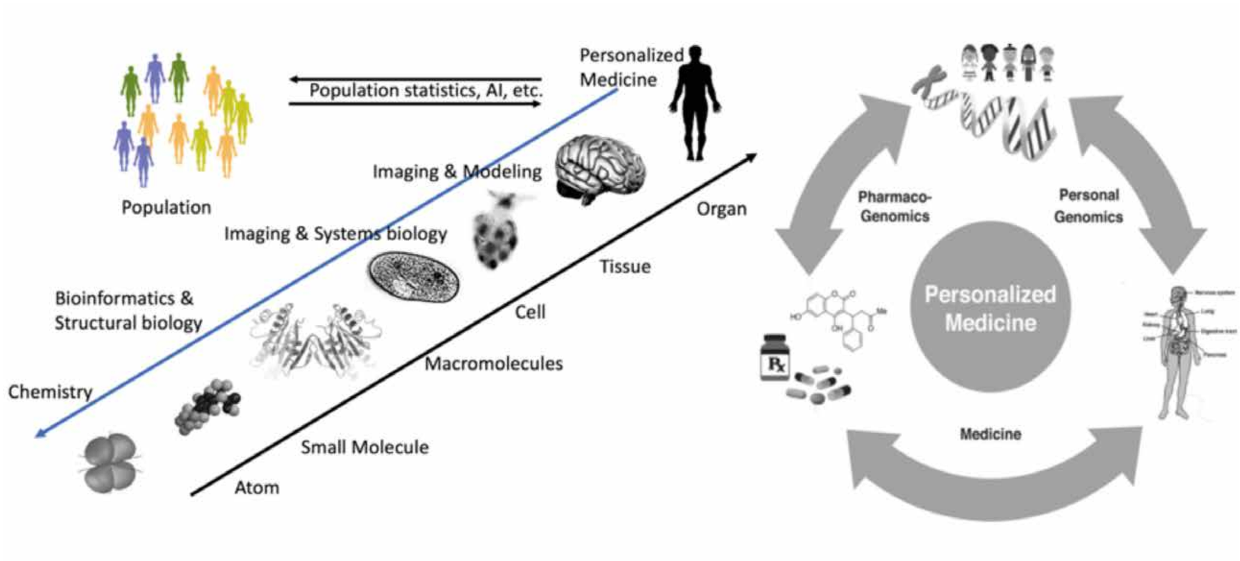

Figure 7: Advanced computing plays a key role in all levels of modern medicine and health. Left: Bioinformatics has long been critical to understanding the structures of proteins, but is increasingly expanding into systems biology, population-level studies, and integrating both structural biology, imaging, informatics and models to predict health outcomes. Microscopy and imaging technologies are similarly undergoing revolutions with new deep learning-based artificial intelligence methods that are frequently better than doctors at diagnosing disease. *(Image: BioExcel Centre of Excellence for Biomolecular Computational Research).* Right: Integration of high-throughput sequencing, genomics and systems biology methods to study drug effects on populations and genetic variants is leading to new personalised medicine where both the drugs and treatments are optimised to the individual and state of the disease. With genomics increasingly being applied clinically, there are rapidly increasing needs for time-critical on-demand access to large advanced computing resources.
Image: Haskin Fernald et al., Bioinformatics 27, 1741-8, CC-BY-NC.

The extreme volumes of sequence and other types of high-throughput data make them natural candidates for parallelism, and the field will theoretically be able to utilise even the largest exascale resources. However, the historical usage profile of independent single-node jobs, large reliance on script languages, and researchers without strong computing backgrounds are becoming obstacles to further progress. Urgent assistance with software development, application support, training and new access modes is needed to resolve the computing bottleneck. However, as this is achieved, the needs in sequence data analysis are likely to be one of the strongest justifications not only for moving to exascale, but the generations beyond it too.

Potential breakthrough: The value of high-throughput sequencing is not only proportional to the amount of sequence data, but also to possibilities of carrying out more advanced analyses. Bioinformatics is currently limited by the available computing power. Increasing the resources 50-100x will make it possible to include epigenetic information such as samples of the same sequence from many different individuals, to identify correlations e.g. between human sequences, the microbial genome in our guts and disease. It will be possible to move our understanding of disease beyond the traditional view of identifying a single genetic variant to studying complex diseases that result from higher-order interactions of a whole network of genes where the individual variants are harmless, but their combination results in disease. Systems biology models will be able to identify previously unknown health threats at the population level (where it is far cheaper to prevent disease than curing individuals), enable better treatments for isolated areas with disperse populations, predict how new drugs will interact with different genetic variants, and introduce personalised medicine where the entire treatment and drug combination are optimised on both the patient's genome and the exact state of the disease. There will also be completely new and challenging needs: as high-throughput genomics moves into the clinic, there are examples of severe diseases e.g. in new-born infants where it can be a life or death matter to be able to sequence and analyse a complete genome in a few days. This is difficult to handle with present national and international infrastructure, while the the demand is too large to be addressed in hospitals, but new modes of time-critical on-demand access to advanced computing will have an extremely large impact on future healthcare and active ageing.

Molecular Life Science & Structural Biology

Molecular life science is one of the fastest growing users of computational technologies with a clear need for exascale, existing world-leading European research, and direct commercial applications as described in the EESI-2 report[11]. As pointed out in the EPT4HPC strategic research agenda[12], exascale systems will advance the area by enabling calculation of protein functions, structural protein-protein interactions, distributions of binding free energies, and simulations to understand electron transfer. This will have a transformational influence on our daily lives through health and medical applications, development of new drugs, methods for efficient medication delivery, biotechnology, environment, agriculture, the food industry and not least education – biomolecular computation has changed the view of molecules in our cells from static to dynamic entities.

Even traditional biochemistry research such as structural biology is becoming highly computational. Instead of manually working on individual samples, both X-ray crystallography and nuclear magnetic resonance (NMR) increasingly use automated pipelines for structure determination, and new methods such as free-electron lasers (XFEL) as well as single-particle electronic cryo-microscopy (cryo-EM) rely on capturing extreme amounts of scattering and electron-microscopy images, respectively, after which advanced reconstruction algorithms are employed to build computed electron densities based on which model is most likely to have generated the ›100,000 experimental images. Not only does this mean structural biology is rapidly moving from the wet lab to computers, but in doing so it is fundamentally altering both our view of structures and how structural biology is performed. With modern cryo-EM, it is neither necessary to obtain extremely pure samples of molecules nor grow crystals, and the computational reconstruction will provide 2-3Å resolution not only of a single average state, but also a whole range of conformations that describe the flexibility and different functional states of a molecule (Figure 8). Similar techniques will be critical for the emerging XFEL and neutron scattering fields.

[11] http://www.eesi-project.eu/wp-content/uploads/2015/05/EESI2_D3.3_Final-report-on-application-grand-challenges.pdf
[12] http://www.etp4hpc.eu/sra.html

In parallel, purely computational approaches such as molecular modelling, docking and simulations have become universal tools to understand how the molecular-level processes in our cells are related to molecules interacting and moving, for instance how pumps transport molecules across cell membranes, how ion channels in our nervous system respond to pharmaceutical compounds, how virus capsids are assembled or how complex organs such as the skin are formed from lipid building blocks. Pharmaceutical companies and the agricultural sector rely on designing molecules through docking and high-throughput screening, and more recently they are also applying free energy calculations at a large scale. Beyond individual molecules, the intricate network of protein interactions is a mainstay of cellular processes. Dissecting these at atomic detail is invaluable, but when looking at increasingly complex molecular machineries there is no single technique that provides all the answers. As a consequence, researchers increasingly rely on computational tools combined with experimental data to provide atomic-level information about interactions. Many of the leading codes are developed in Europe, scale quite well and are able to use the latest-generation accelerators efficiently, but there are also substantial challenges. Since the system size is often fixed (e.g. a given protein), better performance on exascale systems will require much better strong scaling. The field has made substantial progress on new techniques to sample extremely complex systems by using Markov state models, free energy calculation or other ensembles of simulations. This class of new approaches will make it possible to put exascale resources to efficient use even for quite small systems, to integrate experimental and computational methods, and to replace high-throughput screening with more predictive methods.

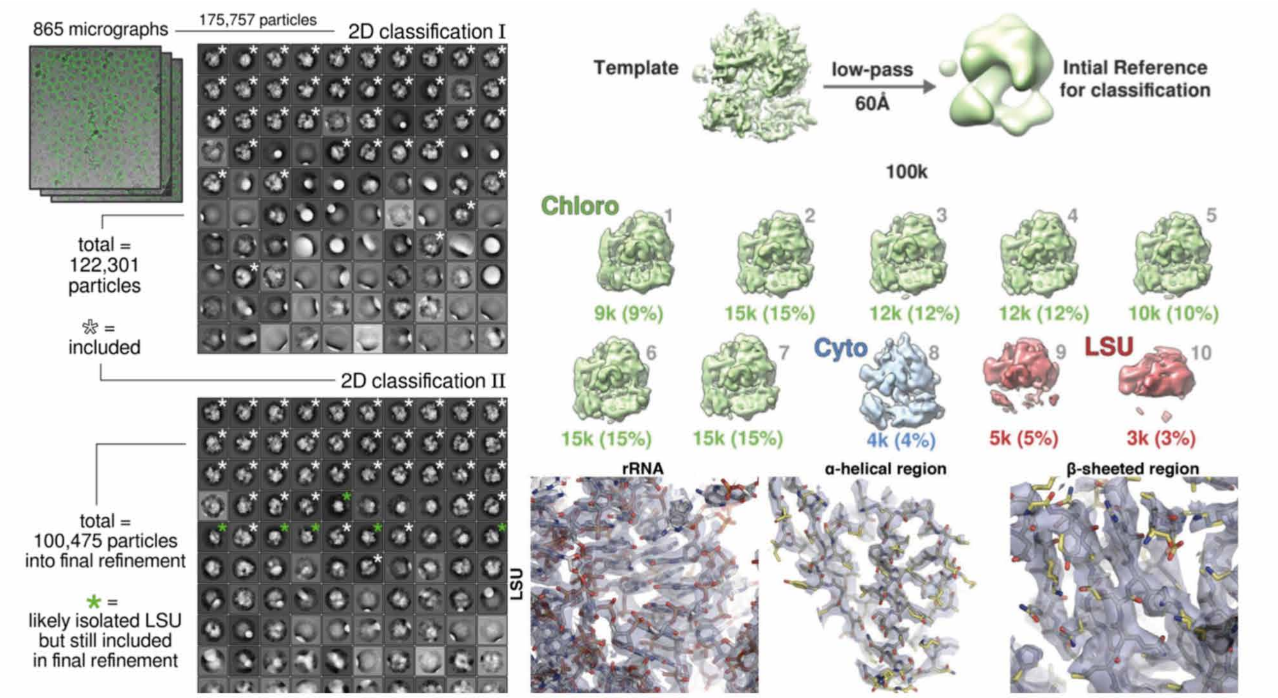

Figure 8: Structural biology is increasingly computational. Modern cryo-EM foregoes the need for highly purified samples and growing crystals, instead using new generations of electron microscopy detectors to collect hundreds of thousands of noisy images of proteins from all directions. Advanced computing is used both to identify particles and classify them into 2D-classes (left), to reconstruct many low-resolution 3D classes that resolve both different conformations of the main target as well as separating out other molecules ("in silico purification"), and finally reconstruct 2-3Å high-resolution structures. Next-generation accelerators and advanced computing will make it possible to perform these types of reconstructions on-the-fly in front of the microscope and enable new types of reconstruction algorithms e.g. for time-dependent free-electron-laser data.
Images: Forsberg et al., IUCrJ 4, 723-7 (2017), CC-BY.

Potential breakthrough: An increase in both computational performance and data-handling capacity by 50-100x will enable all structural biology methods to treat proteins as fully flexible entities, and will allow for the development of new methods that experimentally resolve how proteins move and how this is related to their function. The field will change from manual model building to automated on-the-fly optimsation of structures based on input from several sources of data (e.g. Cryo-EM, XFEL, NMR, and neutrons), and it will become possible to rapidly determine how hundreds of different small drugs bind to a protein. One of the largest differences between sequence- and structure-based methods today is the much higher computational demands of the latter, which means they cannot be applied on massive scale to e.g. screen all-vs-all interactions in genomes, or even hundreds of thousands of drugs. This will change with exascale computing, and massive-throughput computations on structure will move much closer to the state of bioinformatics today. This will have immediate impact on the identification of potential drug-like compounds and their subsequent optimisation, as well as the design of therapeutically used peptides and proteins (e.g. antibodies or nanobody fragments) that are engineered to increase their affinity towards the corresponding binding partners and stabilise active conformations of the biomolecules. More accurate free energy calculations will complement high-throughput docking for determining solubility and affinity of specific drug compounds to receptors (and optimise them), while the docking methods will suddenly be capable of screening entire genomes for protein-protein interactions, including effects of genetic variants. This has large commercial relevance for drug design in the pharmaceutical industry and also for the development of new generations of drugs to detect biomarkers, which are traceable substances indicating various physiological states, in particular related to disease conditions.

Imaging, Artificial Intelligence & Integrative Modelling

All types of imaging technologies are advancing extremely rapidly in clinical settings. Traditional two-dimensional still-image X-rays are being replaced with computer tomography (CT) technology that can resolve tissue in three-dimensional detail as well as provide real-time movies e.g. as joints are moving. New detectors and software enabled by much greater computational power have made it possible to combine this with reduced radiation doses. Similarly, new generations of magnetic resonance imaging (MRI) and positron emission tomography (PET) generate orders of magnitude more data that is making it possible to understand functional processes in the brain, our nervous systems and how signals and chemicals propagate through a patient's body – but the value of these tools both in clinical and research settings is directly related to the possibility of rapid access to very large amounts of computational power to analyse the data. One example of how this is changing the nature of medical imaging is traditional ultrasound used for diagnosing the state of abdominal organs and fetal development. In the last few years, artificial intelligence methods using convolutional neural networks have been introduced to automatically classify images and identify organs, with accuracy already being better than the degree of agreement between clinical experts[13]. Similar techniques are enabling computers to reconstruct three-dimensional live views of a fetus or an organ, and advanced visualisation using shadows and segmentation are providing a completely different user interface for what was formerly a relatively primitive technique.

Artificial intelligence techniques are of equal importance for cellular and organ imaging where one of the major success stories is accurate classification of tumours in clinical samples. This is important as an aid to a skilled pathologist so that they do not miss a potential tumour, and it will ensure that almost every clinic in the world has access to classification accuracy on the same level as the best pathologists in the world. Increasingly, the artificial

[13] Xu, et al. MICCAI 2018 Proceedings, https://arxiv.org/abs/1805.10376

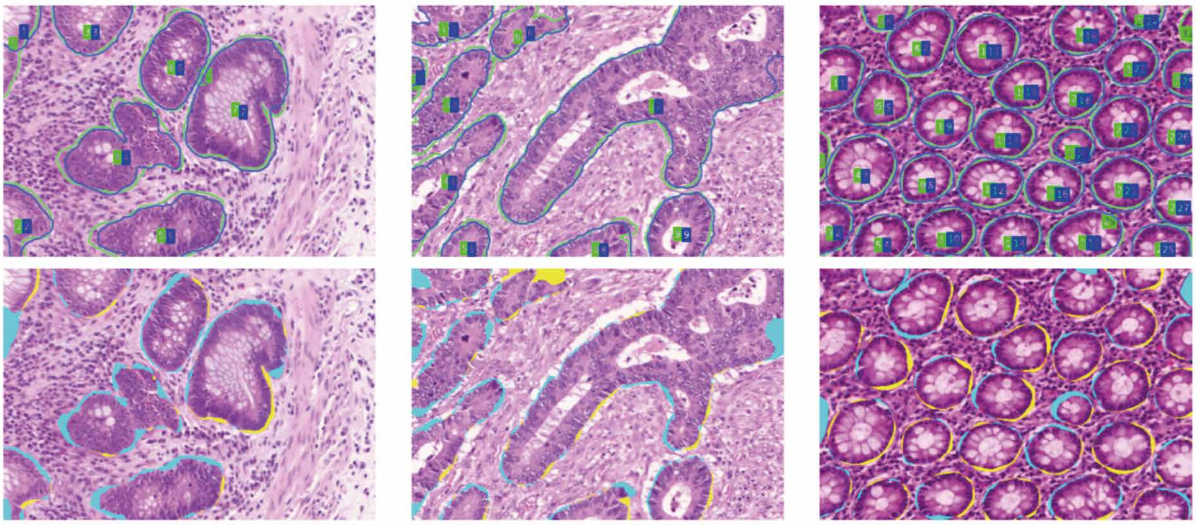

Figure 9: Tumour classification and segmentation with deep convolutional neural networks. Reference extension in green, segmentation shown as blue. For the second row, false-negative areas are shown as cyan, and false-positive in yellow. This is currently an extremely active area of research where increased computational power and training data sizes go hand-in-hand. *Image: Kainz P, Pfeiffer M, Urschler M. PeerJ, DOI 10.7717/peerj.3874 (2017), CC-BY.*

intelligence-based neural networks are becoming better than the best pathologist. In addition, modern networks are now trained not only to locate tumours for very early detection (when treatment is much more efficient), but also to classify their extension (Figure 9), annotate the state of the disease and how the tumour is likely to respond to various treatments[14]. The training of these deep learning networks depends on vast computational power as well as extreme amounts of data that requires access to very large infrastructure resources. Even larger data sets and more advanced algorithms will require next-generation computational resources. However, once the neutral network training is complete, the inference stage where it is used to predict properties of new samples is much less computationally intensive, and is possible to perform in the clinic already with the resources available today. In particular for remote areas, or a countryside clinic where it is not possible to have all expertise, such instant access to treatment recommendations would tremendously improve healthcare.

Potential breakthrough: Imaging, gigantic data sets, and artificial intelligence are already providing a revolution for healthcare, and this will continue – but it will not be possible to use orders-of-magnitude-larger datasets without similar increases in computational resources and infrastructure to handle the data. Better methods for automated segmentation of tomography data will turn autopsies into rapid non-invasive techniques, and advanced visualisation will lead to significantly-improved medical training since students everywhere will have access to accurate visualisations for all concepts and organs, and can perform unlimited amounts of virtual dissections while reducing the ethical concerns of handling human remains. The artificial intelligence approach is expected to spread far beyond imaging, and in the near future we expect systems biology, drug design, whole-body models, and the design of medical cases for student training to increasingly be constructed with deep learning using clinical training data. In imaging, the networks will be used for differential diagnosis and treatment planning rather than mere classification. Initially, the typical model will be to use high-performance computing

[14]Tang, et al. https://arxiv.org/pdf/1806.09507.pdf

resources for training networks or rendering visualisation, while the inference or user interaction can happen offline, but as the amount of data collected in the clinic grows rapidly the requirements for time-critical on-demand computational resources will grow.

Summary

- Life science is one of the fastest-growing users of high-performance computing both in Europe and worldwide, with a wide range of uses from chemistry, bioinformatics, and structural biology to diagnosis and treatments in clinical settings.

- Bioinformatics and systems biology have extreme – and growing – needs for computing and storage due to the increasing size of sequence databases and other high-throughput techniques. This is an area of tremendous potential impact on human health, but with major challenges due to the lack of computing expertise in application researchers who rely on single-node jobs and script languages. Helping bioinformatics and systems biology to better use the next-generation resources through assistance with modern software development, I/O handling, and parallelisation is critically important for computing in Europe.

- Structural biology and molecular modelling are going through revolutions where experiments increasingly become computational reconstruction problems (similar to other imaging applications), and future exascale resources will make it possible to perform massive-scale screening of interactions, make accurate calculations of binding affinity for drug design, and understand how the flexibility and motion of molecules are related to their interactions and potential disease.

- New reconstruction algorithms and deep learning neutral networks are revolutionising medical imaging. Interactive visualisation and better diagnosis as well as methods for e.g. tumour classification that rival or beat the best pathologists are expected.

- Some applications in medicine, in particular in clinical settings, will have significant needs for new modes of access to time-critical on-demand advanced computing, e.g. to identify the genetic variant of a gravely ill newborn child in the clinic in mere days.

Energy

Rapid economic development, intensive urbanisation, and constant growth of the world's population has resulted in a substantial increase in energy demand. There is a broad desire to accelerate the move to renewable energy sources but, since it does not yet meet demand, society is also faced with the challenges of balancing requirements to improve the efficiency and output of current sources (and reducing their environmental load), while also investing aggressively in renewables to reduce emissions of greenhouse gases and air pollution to combat dramatic climate change. Computing research in the energy domain goes far beyond pure academic studies, and is applied broadly in industry to produce the energy required to provide light, heat, transportation and healthcare.

Improving Output of Current Lower-CO_2-Impact Energy Sources

The environmental impact of burning fossil fuels can be substantial, but in the short term it is still an indispensable part of the energy mix, e.g. for the airline industry where alternatives such as biofuels cannot yet be produced at large scale. In addition, the largest source of energy worldwide is still coal; it is the source of a third of all energy and some 40% of electricity generated, and its consumption is expected to stay at least on the same level through the mid 2020s[15]. However, replacing large parts of this with modern natural gas is a feasible goal, and while still a fossil fuel it would cut CO_2 emissions in half and drastically reduce air pollution.

Despite the drop in price of crude oil in 2014, oil/gas companies are investing in new technologies including high-performance computing to explore new ultra-deep offshore, or non-conventional, oil fields (including shale gas) by using better seismic algorithms and more accurate reservoir modelling methods. Oil/gas has become the second largest profitable market for HPC (just after finance) with an increase of the compound annual growth rate (CAGR) of 9.2% between 2012 and 2017[16]. Most of the major companies and contractors in Europe, the US, Canada, South America and China have invested in their own multi-petascale HPC resources. Notably, the first non-public systems in the Top-500 list[17] are all owned by these companies; it is a unique field in the sense that both research and development of relatively traditional HPC usage is strongly driven by industry, despite high demands for return-on-investments. They have a clear roadmap towards exascale, primarily for the development of efficient and accurate novel seismic processing methods for exploration (such as full waveform inversion (FWI) or separated wavefield imaging (SWIM)) as well as production (4D seismics coupled to reservoir modelling, uncertainty quantification, and multi-scale modelling from the pore to the reservoir scale).

Considering only the progress of seismic acquisition (more sensors, seismic sources, more streamers, higher frequencies of acquisition, multi-component data, etc.), the computational demands will increase by at least one order of magnitude before 2020. In addition, the integration of more physics, complex approximations and more iterations will lead to an increase of another 1-3 orders of magnitude. With the massive increase of sensors and recording of data from heterogeneous format and sources, artificial intelligence deep learning technologies is a

[15]https://www.iea.org/topics/coal/
[16]HPC Trends in the Oil/Gas Sector, International Data Corporation (IDC), now Hyperion Research
[17]http://www.top500.org/

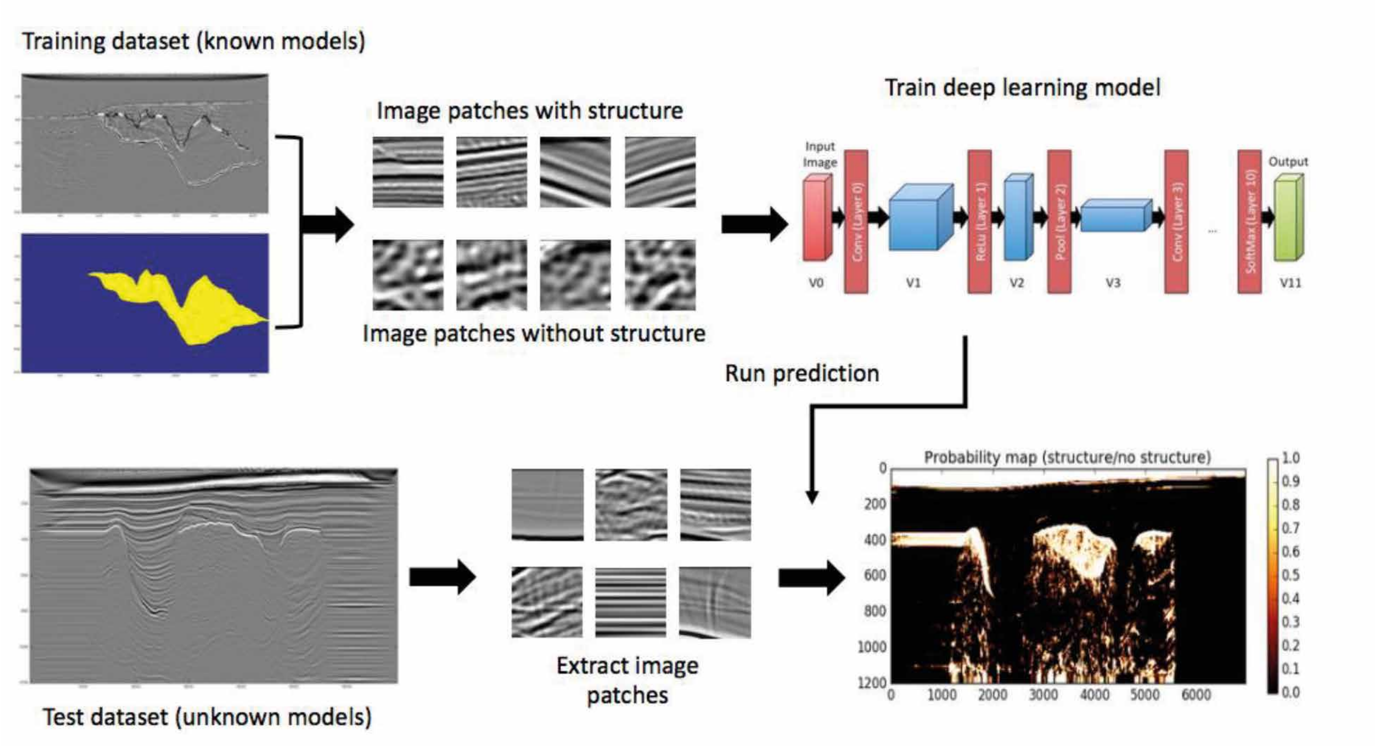

Figure 10: Full waveform inversion coupled with convolutional neural networks for deep learning makes it possible to include detection of challenging salt geological bodies in the workflows, which provides drastically faster results compared to manual inspection and much more accurate results than traditional FWI-only workflows.
Image: Schlumberger & Denes Vigh, WesternGeco DOI: 10.1190/segam2017-17627643.1.
© DOI: 10.1190/segam2017-17627643.1

promising path for classification, segmentation, pattern recognition, and exploration of temporal series. For oil/gas, this is focused on seismic interpretation, real time analysis of data from wells in production, bio-stratigraphic analysis, satellite images in case of oil leakage, forecast of production, well planning before production in existing fields, smart forecast of production, and anticipation of failures.

Potential breakthrough: The oil/gas industry is already one of the largest customers of accelerators such as GPUs, and the rise of new hardware architecture providing AI or neuromorphic computing will provide a promising short-term perspective, with other technologies such as quantum computing already on the radar. Currently, the industry starts drilling at around 20% probability of finding oil resources; increasing this ratio will improve the economics and open new exploration possibilities. The acceleration of AI problems through quantum computing could be one of the first applications, but porting e.g. Darcy laws for flow modelling through a porous medium (used in basin and reservoir modelling) will be a real challenge. In just a few years, this will raise difficult challenges: either completely rewrite existing codes to scale beyond 300 petaflops (if at all possible), or make a clean break to instead start focusing on next-generation disruptive technologies such as quantum computing.

Several recent research results have highlighted the coupling between HPC and AI techniques for FWI. For example, reconstructing complex salt geological bodies poses a huge challenge to FWI due to the absence of low frequencies in the data needed to resolve such features. Currently, a skilled seismic interpreter has to interpret these bodies and manually insert them into the Earth model and repeat this process several times in the workflow. As an alternative, the community is now moving to deep learning algorithms, mainly convolutional neural networks (CNNs), to generate useful prior models for FWI by learning features relevant to Earth model building from a seismic image (Figure 10).

Even for traditional architectures, there is a direct correlation between access to larger computational resources, introduction of more advanced and accurate models for finding resources, and profitability. This has created a large interest in HPC and data analytics, and exascale resources will make it possible to introduce new models that take viscoelastic and anisotropic effects into account, least squares/residual shot reverse time migration (L2RTM), and eventually full waveform imaging (Figure 11). The industry is also seeing expanded use of HPC to scale out basin and reservoir modelling, and more recently molecular dynamics and chemistry applied to processes of refining or global optimisation of plants including multi-disciplinary interactions, e.g. simulating full plant lifetime and control systems, emulation of operations, and risk analysis. HPC in conjunction with high-performance data analytics is becoming crucial for real time monitoring by coupling simulations and inputs from networks of sensors or seismic acquisitions when drilling or operating a well in order to better steer the process, and reduce costs and risks. Finally, there is great activity in research on merging data-driven and physics-driven modelling, in particular using deep learning to allow testing parameter impact and scenarios for optimisation of the reservoirs. While this does not qualify as green energy, over the next decade it is potentially the single most important large-scale development that would enable us to drastically cut CO_2 emission by reducing reliance on coal.

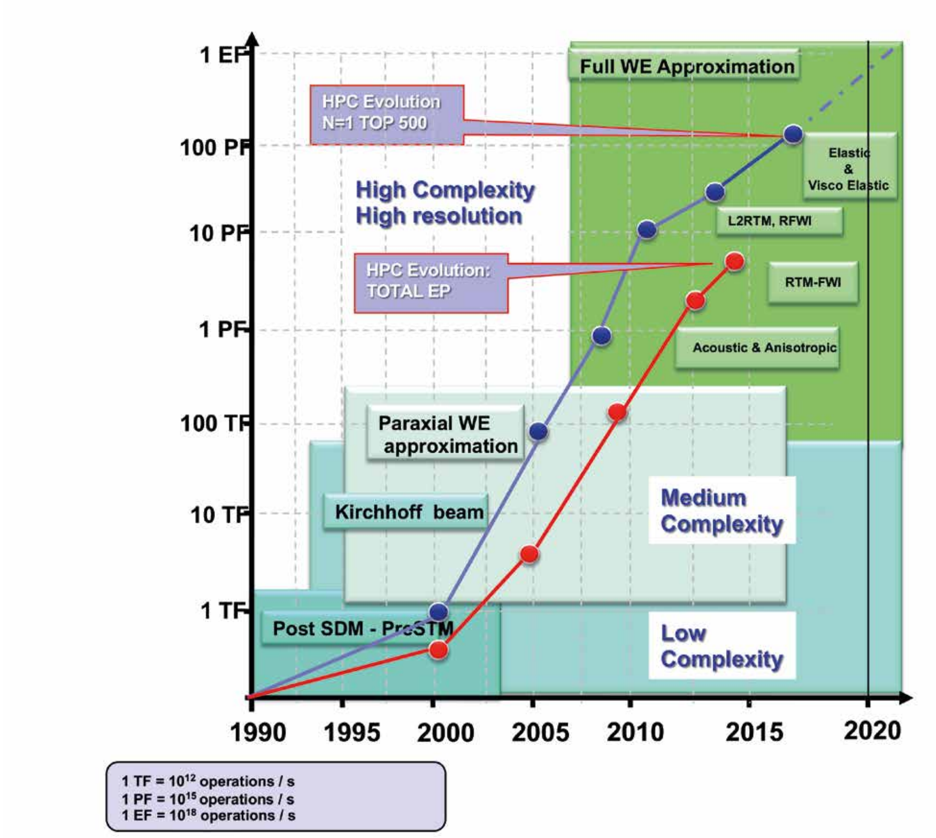

Figure 11: History and roadmap of computational models used in seismic modeling and their computational requirements. Exascale resources will make it possible to introduce new generations of more accurate models with orders-of-magnitude larger computational requirements than today.
Image: Henri Calandra, Total E&P.

Figure 12: Advanced computer simulations are used in a number of areas to optimise the output of renewable carbon-free energy sources. Simulations are used to predict the flow and optimise turbine design e.g. for hydropower, computational fluid dynamics is used both to optimise the placement of wind turbines relative to each other, the shape of their blades, and photovoltaics have seen massive efficiency improvements due to new materials designed with computational chemistry. Simulations of material properties have been instrumental to develop more efficient insulators for underwater transmission cables that now make it possible to approach 1,000,000V in high-voltage cables to drastically reduce transmission losses, which is the key bottleneck when energy is produced in one location but consumed in another.
Images: Wikimedia Commons. Left: Kim Hansen/Richard Bartz, middle: Christoffer Riemer, right: Walter Dvorak, CC-BY-SA.

HPC is of similar importance for a number of other energy sources, for instance nuclear power where advanced computing is applied in virtually all domains. Both computational chemistry and finite element methods are used to design materials to better withstand aging degradation from radiation and understanding crack initialisation and propagation. Simulations of neutron flux are used to predict the properties and placement of fuel and moderation rods to improve reactor efficiency, and fluid dynamics is used both to improve the energy uptake in the coolant (reducing boiling losses) and to design more efficient turbines. European researchers and industry are leading internationally in nuclear power, and new generations of failsafe reactors are unique in their ability to provide massive amounts of energy with very low amounts of CO_2 emission.

Safe & Renewable Energy Sources

The amount of energy produced from renewable carbon-free sources like solar (photovoltaic) power, wind power, hydropower, tidal power, or geothermal energy is growing rapidly, but due to the similarly rapid increases in demand, it has not yet allowed much reduction in the use of gas, oil, and especially coal. The development of clean renewable sources of energy to transform the global energy balance is tied to requirements of significant increase in their efficiency and reduction of cost. A large amount of research is being invested in this area, both in academia and industry, in particular related to improving efficiency of turbines, development of new materials for higher solar energy efficiency, and not least solving issues of storage and transmission of the produced energy (Figure 12).
The complexity of the scientific research and new technological developments in these domains is strongly linked to the modelling in physics, chemistry, and optimisation of energy distribution. Use of advanced computing is enabling higher levels of complexity and realism in modelling that lead to radical improvement in evolving cheap and clean energy technologies on much faster time scale. One example is the predictive modelling of more efficient photovoltaic materials that both exhibits higher efficiency combined with better properties for strength, flexibility, transparency, and resistance to high or low temperatures – combined with changes in processes making them much cheaper to produce.

Modelling and simulations are also used to improve storage and transmission of energy. The energy losses due to resistance in transmission cables are proportional to the current, so the easiest way to address this is to reduce the current by instead increasing the voltage. However, the voltages that can be used are in turn limited by the properties of polymer insulator materials. European companies such as ABB are now using computer simulations of materials to develop insulators with better properties, which has contributed e.g. to the latest-generation extruded cross-linked polyethylene (XLPE) megavolt cables. Not only does this improve the ability to transfer energy over long distances, which is critical to increase the use of renewables, but it also a rapidly growing business worldwide – leading research in more advanced materials is an important competitive advantage for European industry.

Longer-term, one of the most interesting research areas for development of unlimited clean sources of energy is fusion power production based on hydrogen isotopes. This can either be performed in magnetically confined plasma devices or inertial fusion facilities where lasers are used to trigger nuclear fusion. Fully ionised gas, or plasma, is the state where electrons have separated from the atomic nuclei, and this is actually the most common state in our universe. Cold or hot plasmas have a wide range of technological applications including the destruction of toxic materials, modification surfaces coating, food processing, and cancer treatment. However, the arguably most challenging goal of plasma physics for the next 20 years will be to demonstrate feasibility of an energy source based on nuclear fusion. A new era in magnetic fusion started in 2007 with the construction of the ITER tokamak in France within the framework of a large international partnership between the EU, China, Russia, the US, Japan, India, and Korea (Figure 13).

Figure 13: The ITER site under construction, March 2018.
Image: ITER.

Potential breakthrough: Magnetic fusion research is facing tremendous challenges and obstacles in the development of high-confinement plasma scenarios combined with high levels of fusion plasma control. Theory and predictive modelling play crucial roles for ITER operation and construction. In terms of HPC resources, the most challenging are the so-called first-principles modelling codes that concentrate on the detailed understanding of phenomena like electromagnetic small-scale instabilities responsible for turbulent transport and hence the plasma confinement or large-scale magnetohydrodynamic (MHD) instabilities that define operational limits for a given plasma scenario, mechanisms of plasma heating, and plasma-wall interactions. The computing challenges for this non-linear physics are extreme, involving 3D complex geometry and kinetic effects (Figure 14), but also quite similar to those in other fields such as astrophysics, fluid dynamics, and weather forecasting. The challenge stems from the wide range of spatial and time scales to be covered: from metres and seconds at the machine scale down to millimetres and milliseconds or microseconds for MHD and turbulence. A large group of first-principle codes based on global 5D gyro-kinetic approaches can explore small scale instabilities responsible for turbulent transport in tokamaks, but their accuracy is severely limited by the available resources. Access to exascale resources will enable new generations of these simulations to make progress in modelling both ion and electron dynamics, which is highly challenging due to large differences in mass, electromagnetic effects, and representation of the ITER-size plasma enclosure.

Existing MHD codes mainly use a fluid approach, but kinetic effects have increasingly been introduced to better represent high energy particles and impurities, and these models are much more advanced with respect to the realistic 3D geometry and the ITER-size enclosure. However, to fully reach realistic ITER plasma parameters and relevant Reynolds numbers, much larger computational resources are required. Still, MHD simulations have already provided valuable contributions to understanding the strongly non-linear behaviour typical for global MHD modes in order to propose effective plasma control methods required for ITER. With access to exascale computational infrastructure by the early 2020s, as well as further development of MHD codes including plasma-wall interaction in steady-state and during transient large heat fluxes (due to instabilities), first-principle models should approach the realistic plasma parameters for ITER within a decade. The goal of this magnetic fusion modelling is a "numerical tokamak", where all aspects of a plasma scenario and control are considered self-consistently. These global MHD-electromagnetic turbulence models with plasma core-edge-surface integration will be a critical tool for modelling and optimisation of fusion experiments in general and for ITER in particular, which will commence operation 2025-30.

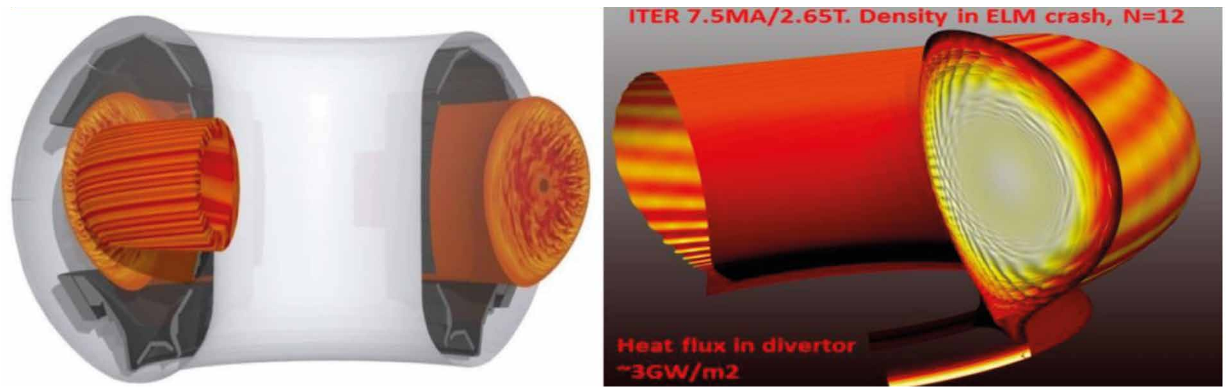

Figure 14: Left: Simulation of plasma turbulence with the gyrokinetic GENE code. Right: Simulation of edge magnetohydrodynamics in ITER with the non-linear code JOREK
Images: Frank Jenko, Marina Bécoulet

Summary

- Energy HPC research spans an exceptionally broad range of methods, including fluid dynamics to improve wind turbines, seismic modelling in the oil/gas industry, electronic structure calculations for higher photovoltaic efficiency, optimisation of material properties to reduce transmission losses, and not least magnetohydrodynamics to predict and control fusion plasma properties.
- Exascale resources will permit researchers to use new generations of much more detailed computational models, in particular for seismic modelling, fluid dynamics and magnetohydrodynamics, where there are also a number of codes with excellent scaling properties.
- The societal and commercial impact of energy-related HPC is obvious and immense. Large investments are being made in industry, covering both hardware and algorithms, and for European research and industry to stay competitive it is critical with close academic collaborations and state-of-the-art infrastructure.
- In particular for seismic modelling, data-focused deep learning approaches are rapidly gaining in importance, but rather than replacing compute-focused approaches they are typically combined. Future European infrastructure needs to be able to cater better to this type of usage.
- Some of the largest long-term impact might come from new forms of energy, in particular fusion, but realising this will require large investments in magnetohydrodynamics research and code development.

Infrastructure and Manufacturing for Humankind

Engineering has long been at the forefront of success stories in computing. Today, every car and craft on the market has been produced after advanced simulations that examine all aspects of their design, from structural components to aerodynamics and safety assessment. These simulations are typically replacing or complementing expensive experiments and provide quantitative information for the full design process that is translated to significant economic benefits to manufacturers. Similarly, after an aircraft is deployed in operation, computing identifies optimal routes by accounting for factors such as wind direction and cost of airspace, and before this computing has been used to determine what aircraft to deploy on what route and how to determine ticket pricing. Similar optimal routing computations are envisioned as applicable to self-driving cars and drones. The design cycle is no longer limited to the product, but it extends to its full operational cycle thanks to the capabilities offered by computing.

Engineering and Simulation

These types of advances are now being seen making rapid progress in all sectors of manufacturing, from electrical and mechanical equipment to food and drink. Simulations allow explorations for new batteries and new materials. The advent of 3D printing would not have happened had it not been for advances in computing. Modern engines reaching the EURO6 emission standards would not be possible without combustion simulations used to simultaneously optimise for lower fuel consumption and reduced NOx emission (notably, this conflict between reducing CO_2 and NOx emissions was the reason for the recent manipulation scandals in the car industry). Insight from simulations has led to the development of hybrid nanostructures for ultrafast cooling of microchips, bioinspired multi-objective optimisation algorithms have led to innovations in turbomachines and design of aircraft wings and data-driven uncertainty quantification has provided us with new multi-scale models for the design of artificial organs. Direct numerical simulation (DNS) is finally becoming feasible with next-generation supercomputers, which will make it possible to directly solve the Navier-Stokes equations numerically without any turbulence model, resulting in much greater accuracy. Naturally, today's supercomputers are also used to design the chips of tomorrow's supercomputers. Such advances have relied on high-performance and high-throughput computing, and in the coming years we are destined to experience an ever-increasing demand for such resources.

Traditionally, only a handful of European corporations have relied on the highest-end supercomputing, but this is changing, in particular with research projects involving both academia and industry. One recent success story of new manufacturing is how researchers at the Technical University of Denmark used PRACE supercomputing resources to perform computational morphogenesis reaching billions of voxels and design an entire airplane and optimise the strength-to-weight ratio without restriction to the traditional rib-and-spar wing architecture[18], published in Nature 2017 (Figure 15). This achieves weight savings of 2-5%, but more importantly the resulting structures show remarkable similarity to naturally occurring bone, e.g. in bird beaks. These types of structures would be difficult to manufacture with traditional means, but when combined with metal 3D printing it could open doors to entirely new landscapes of engineering.

[18]Aage, et al. Nature 550, 84-86 (2017)

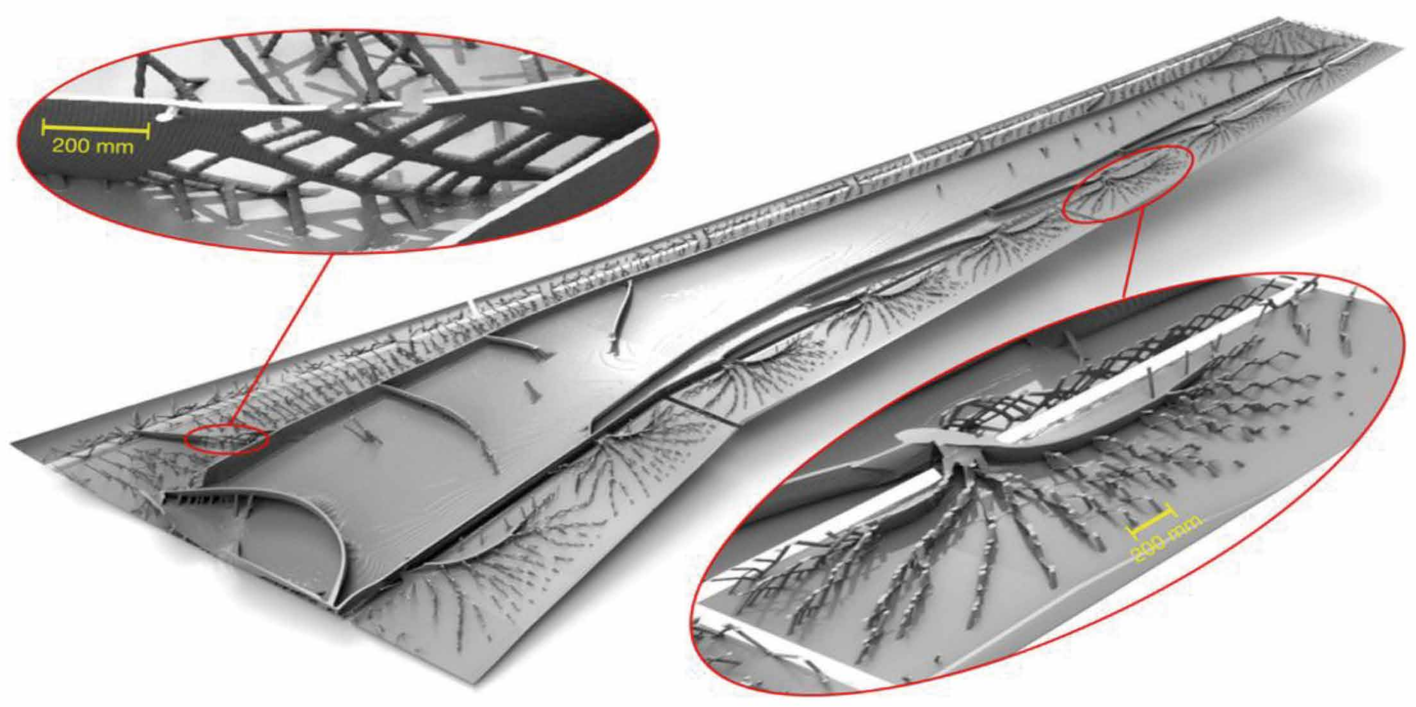

Figure 15: Design optimisation using computational morphogenesis with billion-voxel resolution in PRACE made it possible for researchers at the Technical University of Denmark to optimise the design of an entire airplane wing without limitation to the usual rib-and-spar structure. This would save at least 2-5% of weight, and combined with new computer-controlled manufacturing technologies, such as metal 3D printing, this type of academic-industrial collaborations could open doors to entirely new landscapes of engineering and manufacturing.
Image: Nature 550, 84-86 (2017).
©Nature 550, p. 84, year 2017, DOI 10.1038/nature23911

Potential breakthrough: State-of-the-art simulations are similarly finally making it possible to understand well-known complex phenomena such as the cavitation responsible for huge losses in efficiency and equipment material damage e.g. in control valves, pumps, or propellers (Figure 16). In fact, this process is also highly similar to boiling and cooling, which are other major source of energy efficiency losses in industry – new methods to design surfaces to minimise these would both advance fundamental understanding of the behaviours of liquid and flow and have major industrial impact by designing shapes to minimise cavitation.

Integrative Design, Manufacturing, and Operations

Today, there is a confluence of manufacturing, design and materials engineering that hinges on their common threads: computing and data science. These provide the foundations necessary for innovations that are in turn critical for some of the most important challenges of our society such as energy, health, poverty, and clean water supply. Efforts to solve such grand challenges have driven the transformation of engineering in terms of its methods of inquiry and its target applications. Computing has a prominent role in this transformation, as it has revolutionised our intellectual capacity to tackle complex problems and provides us with a unique way to complement experiments and empirical models.

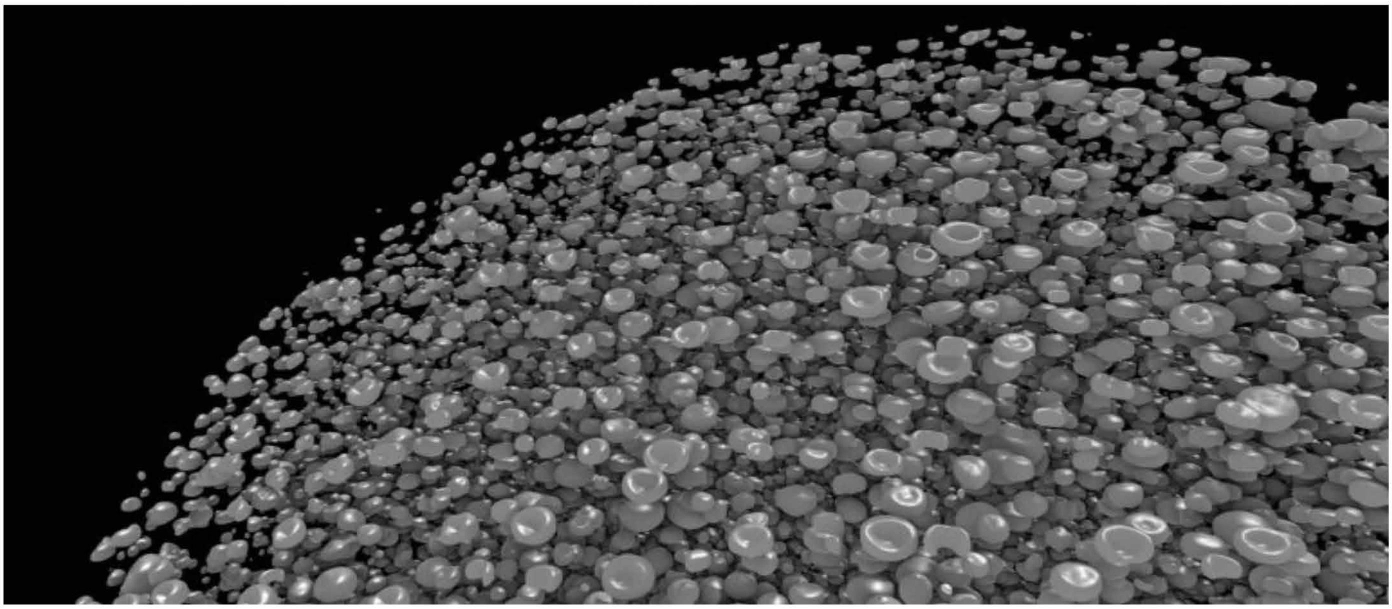

Figure 16: Cloud cavitation collapse refers to the cavitation process where vapor bubbles form in low-pressure regions of liquids, and the violent collapse of these bubbles leads to material erosion on nearby surfaces that considerably reduces the expected lifespan of equipment in marine propulsion, turbomachinery, or fuel injection systems. It is only the last few years that it has become possible to simulate these phenomena on large scales to understand the process, in this case using an unprecedented 13 trillion grid points to simulate 50,000 bubbles at 14.4 petaflops on 1 million cores of the Sequoia supercomputer.
Image: Petros Koumoutsakos.

Potential breakthrough: The majority of manufacturing processes involve modelling and simulation. Design is inherently linked to optimisation and inverse problems, while projects such as the Materials Genome hold the promise of unprecedented innovations. Despite significant advances, there remains huge potential for innovation in these domains by further incorporating advanced computing and data science methodologies. Such methods include multi-scale and predictive simulations, stochastic optimisation and data-driven uncertainty quantification, machine learning and decision algorithms as well as high-performance and cloud/commodity computing. Engineering and information technology are also interfacing in the big revolution of our times: data science. Sensors of every scale are becoming increasingly available and their capabilities are largely responsible for the boom in data that we are experiencing. Sensors are being used in applications from monitoring highway traffic to providing failure warnings for critical engineering systems such as bridges. In turn, this data is used by artificial intelligence systems to design smart cities and better infrastructures from energy-positive buildings to new highways and wind farms – and public availability of the data has led to a huge number of start-ups with new business models focused on delivering faster, cheaper, and more efficient information.

Computing has become a fundamental mode of scientific inquiry and engineering innovation, and at the same time it increasingly influences every aspect of our society. In addition to developing HPC infrastructures it is important to educate a new generation of scientists and engineers that are aware of computational thinking as a way to solve problems, rather than merely using computers as an improved tool to process old-style models.

We believe that engineering education and research programmes must:
- Develop suitable models, algorithms and software that will take advantage of exascale HPC architectures being deployed in EuroHPC.
- Appreciate the capabilities offered by high-throughput computing for optimisation and uncertainty quantification for quantifiable decisions at a system level.
- Understand the potential of HPC in answering complex engineering problems and fundamental questions that provide insight that can then be used for engineering innovations.

Computing alone is not the answer to every manufacturing, design, or materials problem. Moreover, the energy demands of computers are soaring while vast amounts of data are challenging classical methods of inference and the foundations of scientific computing on natural laws. However, tackling complex problems inherently implies breaking disciplinary boundaries and computing presents a fertile ground for the integrative thinking of engineers. The successful research programmes of the future will be the ones that curate disciplinary strengths and at the same time adopt effective transdisciplinary problem-solving approaches. Such trans-disciplinary approaches are likely to be the most valuable not only for advancing manufacturing, design, and materials for the benefit of humankind, but to create impact in all fields of science, engineering, and industry.

Summary

- Europe has internationally leading research in a number of engineering disciplines, and in particular computational fluid dynamics simulations are already entirely limited by available resources. The codes scale well, and exascale capability would enable unprecedented accuracy and detail.
- New algorithms, models, and multi-scale techniques combining particles and continuum are making a much broader set of applications accessible by HPC, and simulations are increasing for all types of design and optimisation in manufacturing.
- Data science is growing exceptionally fast in importance, but rather than replacing simulations, artificial intelligence and deep learning approaches are increasingly used to process both measurements and simulation-generated data to increase the accuracy both of the models and predictions. It is critical that the next-generation European infrastructure is capable of handling both simulations requiring millions of processors and data-science focused projects requiring accelerators and vast amounts of storage.
- To further increase the impact of HPC on design, manufacturing, and civil infrastructure, it will be critical to foster transdisciplinary approaches that integrate all parts of product life-cycles.
- One of the most urgent needs is to redesign European engineering education programmes to promote computational thinking as the core of modern problem-solving rather than merely as a powerful calculational tool to handle traditional models.

Future Materials: From Molecules to Machines

The rapid increase in computational power over the last few decades combined with enormous advances in methods and algorithms has been instrumental in the development of efficient strategies for modelling the properties of materials and compounds from first principles and to address systems with increasing degrees of complexity. These software and hardware developments have allowed us to move beyond a qualitative understanding of materials and processes, and to obtain quantitatively accurate results that demonstrate the predictive nature of simulations in the areas of materials and chemistry.

Material Properties from Atomic and Electronic Structure Simulations

A huge part of the technological development of society is driven by the successful preparation of novel materials and by the effective use of their properties incorporated into product design. To date, however, even for simple materials, many properties are still poorly characterised and understanding the functioning of a material all the way up to its performance within a device has rarely been accomplished. The increase in computational power has however allowed us to investigate materials of increasing complexity and our ability to predict properties of simplified yet realistic molecular and solid systems has often led to revolutionary concepts with appealing potential applications in medicine, energy, and catalysis. One example of this are the so-called supercapacitors used instead of batteries to store energy for short periods of time, e.g. for stop/start systems in hybrid cars and the Kinetic Energy Recovery System (KERS) used in Formula 1 racing, with potential applications for all types of vehicles. Although they have been used commercially for over 50 years, only incremental advances have been made in terms of their performance. As recently as 2016, PRACE resources enabled researchers from EPFL Lausanne to perform one of the most comprehensive simulations of supercapacitors to date, and shed new light on how charge is built up and released in nanoporous carbon-based supercapacitors, opening new venues for the development of more efficient devices (Figure 17). This progress has attracted considerable industrial interest because of the importance of these materials, which also display good performance as membranes to improve separation processes.

Another example of the tremendous current impact of HPC and materials is the electronic structure calculations, the majority of which are based on density functional theory (accounting for a large share of the HPC use in Europe), that have allowed understanding of the mechanisms underlying the electronic, magnetic, and mechanical properties of materials of significant structural and compositional complexity. The improvements in functionals (key ingredients in these simulations) have led to the position where a large variety of materials properties now can be predicted ab initio with great accuracy. The exploitation of systematic production of highly accurate material data with artificial intelligence and machine learning methodologies, together with the application of high-throughput protocols, opens the possibility of efficiently describing complete phase diagrams for new sets of materials, to understand and predict the behaviour of materials under extreme conditions which are impossible to reach experimentally, or computational discovery of new materials with targeted properties and functionalities. As an example of the importance of HPC for this type of work, recent PRACE allocations were instrumental in allowing researchers from EPFL Lausanne to develop schemes to identify phases and morphologies formed by binary mixtures of inorganic compounds. Combining extensive computer simulations with the Pauling File, the largest collection of reference experimental data on inorganic crystals, has allowed a massive search to take place for all current missing structures and stoichiometries for binary compounds, and this type of study is now being added to publicly available databases[19].

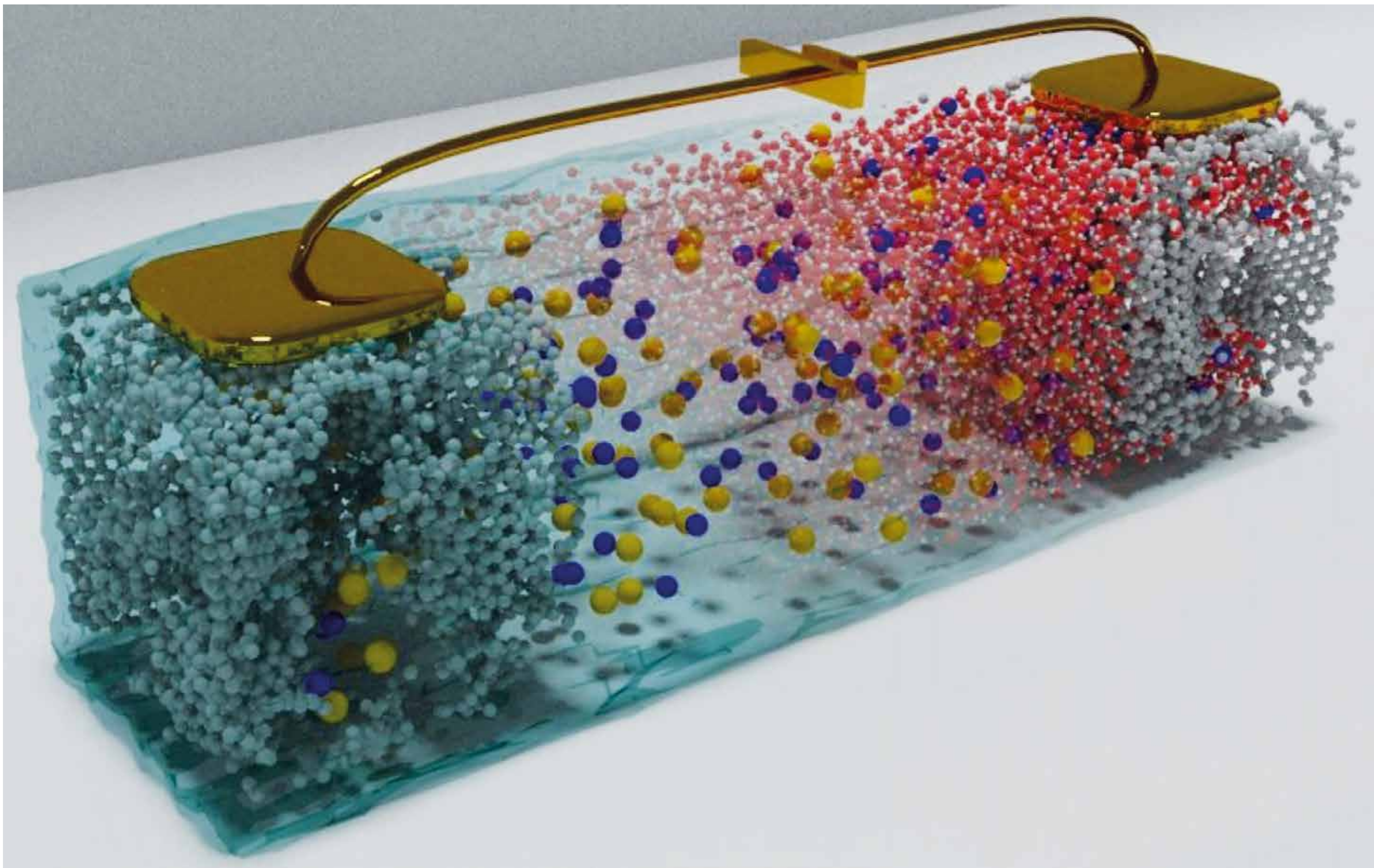

Figure 17: Molecular mechanics simulations of supercapacitors enabled by PRACE in 2016 have enabled researchers at EPFL Lausanne to understand the factors that determine the capacity of the devices, in particular how charge builds up in the nanostructure. This has important potential applications both for kinetic-energy recovery systems and possible uses as membranes for separation or desalination.
Image: Mathieu Salanne, EPFL.

Potential breakthrough: Future materials, such as metamaterials, can display behaviours and structural properties determined by the particular dynamical processes that characterised their formation. Therefore, self-assembly, interfaces, and boundaries must be included for computational modelling to be predictive. For materials with complex structure, heterogeneity in size or chemical composition, simulations will need to cover a wide range of spatial and temporal scales. In the coming years, we expect to see a significant development of integrated multiscale methods to bridge from atomic or even electronic structure resolution to the continuum regime[20](Figure 18). This has large potential for designing membrane pores and biological systems able

[19]Pizzi, et al. Comp. Mat. Sci 111, 218-230 (2016)
[20]Delgado-Buscalioni, et al. Eur. Phys. Special Topics 224, 2331-2349 (2015)

to respond to external stimuli (e.g. temperature or pH) to be used in biomedical, diagnostic, and industrial applications. From a modelling and theoretical perspective, the major challenge is to describe complex matter in a unifying way over the wide range of scales inherent to such systems. This will require new algorithms and software to allow for simulations of open systems exchanging energy and matter with the exterior, and protocols to model behaviour over different scales depending on the resolution. The combination of several computational approaches and high-throughput protocols will open new venues to study non-equilibrium behaviour of complex materials and will shed light on how to use self-assembling systems.

The advances in hardware we are witnessing and the expected increase in computational power open qualitatively new possibilities for material modelling and place simulations as drivers for design and innovation even at the level of the in silico synthesis of novel compounds. However, the computational approach to study future materials constitutes a complex ecosystem of codes covering a wide variety of length scales, time scales, and degree of accuracy, which requires constant innovation in the development of appropriate conceptual and algorithmic tools to take advantage of the existing underlying hardware. Advances in algorithms and new computational tools have provided a much larger increase in our capability to address complex materials challenges than the transistor density increase predicted by Moore's Law. Future rapid advances in our ability to tackle materials of increasing complexity will require new algorithmic developments that can benefit from and exploit future growth in hardware infrastructures and heterogeneous architectures.

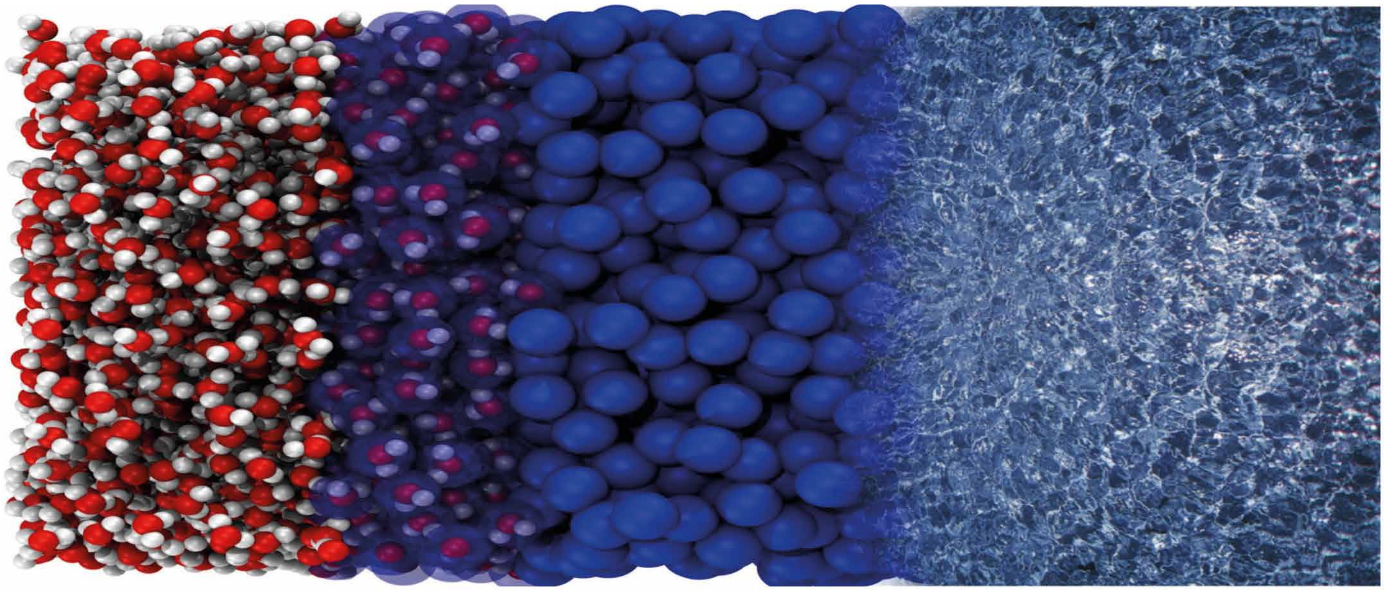

Figure 18: A triple-scale simulation of liquid water, in which atomistic water is concurrently coupled with a coarse-grained model (blue spheres) which is in turn coupled to continuum hydrodynamics. As we move toward our region of interest, i.e., the atomistic molecular dynamics domain on the left, the resolution and associated computational costs are progressively increased whereas the opposite is true while moving away toward the computational fluid dynamics domain on the right. *Image: Matej Praprotnik.*

This identification, development, and exploitation of new classes of materials is key for European industry and competitiveness. To illustrate a handful of examples:

- The use of ab-initio, Born-Oppenheimer molecular dynamics to understand the catalysis of petroleum cracking and other important reactions on transition metal surfaces (Johnson Matthey, BASF).
- The use of mesoscale simulation methods in the prediction of phase diagrams of multicomponent surfactant and polymer mixtures heavily marketed by companies such as Unilever and Procter & Gamble. This will require rapid chemical potential calculation by particle insertion methods.
- The prediction and modification of crystal habit by simulation, in particular attachment energies and entropies. This is used by major pharmaceutical companies such as Pfizer in the prediction of drug solubility and delivery.
- Calculating accurate relative free energy changes as drugs are transported from solution to the active site of proteins, as well as their binding affinity. The relative binding constant of two drugs can be calculated from the free energy cycle as one drug is transformed into another in both environments, and with exascale computational resources these methods will be able to use even more accurate models and forcefields to replace the current more approximate docking methods.
- The prediction of lubrication and friction coefficients between two solid surfaces including ionic liquids in factory machines at extreme loads (›1GPa), physiological lubrication of joints by polymers at loads of 7MPa, and tertiary oil recovery (British Petroleum).
- The growth of gold nanowire using kinetic Monte Carlo calculations combined with the computational fluid dynamics of spray jets (Merck solutions).
- Mesoscale simulation of the effect of electric fields on director reorientation and orientational phase transitions in liquid crystal phases in the production of new flat screens (Mark Global).
- Simulation of polymer melt mixtures to create channel structures for use in membranes for osmotic water purification (Fuji Films).
- The use of quantitative structure-activity relationship (QSAR) and machine learning techniques for the design of novel antimicrobial peptides for cleaning solutions (Unilever).
- The prediction of mechanical properties of co-block polymer melts to create new and efficient tyre composites (molecular and mesoscale dynamics) for Michelin.

Data-driven Materials Design

The development of new materials has played a historical role in improving the capabilities of our societies, allowing larger populations to live with increasing better conditions and prosper. The population is growing, but resources are limited and the development of new materials will have to be based on principles of circular economy rather than the previous make-make-dispose extractive industrial model. However, this in itself has potential to enable further growth and societal benefits since materials design now also has another set of severe constraints based on the materials they can be sourced from. For instance, roughly 50% of today's products are designed for single-use. To formulate, optimise and design the new materials and fabrication methods to furnish a circular economy with particular control over the use of critical raw materials will only be possible through much more advanced computational models that, in turn, require exascale hardware resources.

While data has always played a pivotal role in material design and discovery, its importance as a major driver for innovation is increasingly appreciated, as evidenced by the many current efforts worldwide in material data such as the NOMAD[21] and MAX[22] EC Centres-of-Excellence, independent initiatives such as ioChemDB[23], and the US Materials Project[24]. New machine learning technologies are providing substantial added value by enabling the extraction of more information from these new data projects and from existing databases – the community is increasingly generating knowledge directly from the databases. Furthermore, the transferability of different concepts through suitable metadata enhances (i) the identification of materials and compounds with desired properties and (ii) the transferability and conversion of the gained molecular understanding into mesoscopic models and ultimately devices, following the concepts developed in the European Materials Modelling Council and their "materials MOdelling DAta". The improvement of codes and their integration plays an important role in this progress as has been recognised through the Virtual Materials Market Place, an initiative that demonstrates materials research is predictive enough for wide industrial application, and that there is industrial demand for materials design using both high-performance computing and artificial intelligence.

Potential breakthrough: For example, future predictive atomistic simulations of organic materials will be able to determine physical properties like phase transition temperatures, dielectric constants and viscosities starting from molecular structures. They have recently proved to be successful in doing so, not only for simple molecules but also for rather complicated systems such as liquid crystals and functional materials for organic electronics. However, a key to success is the development of a suitable, validated set of intermolecular interactions (force fields) and to couple this to artificial intelligence (AI) techniques. The problem here is that the optimisation of a force field for a particular type of materials can take months, making it next to impossible to generate enough cases to teach an AI system. Reducing the time to generate the needed results by a factor of 100 or more would be key to producing force fields which are general enough to fulfil the vision of predictive simulation tools on the bench of every chemist. Achieving such a reduction requires new algorithms that take advantage of AI and the existing expertise on ab initio and multi-scale modelling, but there are highly promising signs of deep learning neural networks being able to predict energies of molecules with accuracy similar to electronic structures codes. These efforts will allow us to predict crystals from chemical structure, to move from synthesis to activity prediction, and when combined with accelerated dynamics and appropriate sampling techniques, to recover complete and quantitatively predictive phase diagrams. The computational design of future materials will require high-throughput evaluation of different compositions under different boundary conditions. Exploring the phase space of relevant configurations is an almost intractable problem, but exascale computing will for the first time enable us to sample it in a quantitative manner.

Potential breakthrough: Building on the remarkable progress in material modelling over the last 30 years, we are in a position to envisage a grand challenge to develop a holistic approach to material production. Such a perspective constitutes a qualitative shift compared to the ways we have considered material modelling to date. Indeed, most predictive atomistic modelling efforts have been concerned with materials at equilibrium, while many industrial processes involve non-equilibrium stages where different fabrication protocols can dramatically change material properties. To make materials modelling a quantitatively predictive strategy, it is necessary to integrate all aspects involved in the design, from identification of the properties of their fundamental constituents

[21]https://www.nomad-coe.eu/
[22]http://www.max-centre.eu/
[23]http:/www.ioChem-BD.org/
[24]https://materialsproject.org/

at the molecular level, to the industrial processing and the assessment of the material's long-term stability and behaviour in its operative environment.

It is worth noting that, to date, while it has often successfully aided the design of materials, computer modelling alone has yet to create a complete material that has been produced industrially. This is due to the fact that both the material production process and the long-term real-world behaviour, involving evolution of the composition and microstructure, and the large-scale failure mechanisms, are beyond our current modelling capabilities. Addressing these challenges will require significant enhancement of modelling methods as well as leadership-class computing capabilities. A computational predictive study for the industrial production of new materials requires computational analysis of the integrated procedure and steps that determine their fabrication. Such studies are challenging as they encompass non-equilibrium processes at realistic conditions that need to be properly implemented, in particular extending the sample sizes and time windows that can now be studied by at least a factor of 100. The use of computational methods will only play a leading role in the fabrication of new materials if they ensure a sufficiently high level of accuracy and if all relevant time scales can be covered.

Summary

- Computational chemistry and materials design have long been among the largest users of HPC worldwide, both in academia and industry, and there are a number of success stories where it has significantly contributed to both understanding and design. Exascale resources should finally make it possible to address the grand challenge of full computational design, in which all aspects of a material and its production properties can be optimised computationally and which would have major long-term industrial impact.
- Many molecular mechanics and electronic structure codes are highly accelerated and run well in parallel on current architectures, but exascale computing will lead to extreme new demands as these codes will need continuous improvements to keep up with the advances in hardware and new algorithms required to enable computational design. There are also a number of critical high-accuracy algorithms that cannot yet be made to scale as efficiently, but where high-throughput calculations are extremely important to aid data-driven materials design.
- There has been a rapid increase in the usage of artificial intelligence methods (in particular deep learning) to predict energies and structure of new molecules and materials based on extensive databases of e.g. electronic structure calculations. This is a trend that is expected to become much stronger over the next few years, and the field is likely to have large needs for data storage and processing as well as hardware suitable for deep learning training.
- The next generation of researchers will need to master a wide range of methodologies, ranging from identification of effective potentials starting from ab initio studies to the behaviour and response of systems with complex composition over long length and time scales. We therefore need a concerted effort to educate a new generation of computational material scientists and direct resources towards software development in addition to the ongoing investment in computational infrastructure.

Complexity and Data

The preceding chapters have highlighted the increasing importance of data-driven research as a universal trend in all areas of science. This is closely coupled to similar changes in industry, and driven by the renaissance of neural networks, in particular in the form of deep learning. Until recently, a prerequisite for solving almost any problem would be to first develop a simplified representation of the process, but this is now frequently replaced with the approach of measuring, collecting or generating extreme amounts of data, and then training neural networks to not merely classify data, but predict properties such as atomic coordinates, energies, or even complex actions such as steering a car or an industrial robot. This type of advanced computing has been highly dependent on access to hardware optimised for training neural networks, typically by providing extremely high-performance (10-100x more than normal processors) for half-precision (16 bit floating-point), or even 8-bit integer operations formulated as tensor products, but with the latest generation of hardware this is now provided by accelerators such as GPUs present in normal compute clusters.

These applications have sometimes been based on radically different usage models (e.g. the cloud) compared to that typically provided by compute centres. However, as the amount of training data available continues to grow the network training is increasingly performed in parallel using multiple accelerators sharing memory, large secondary solid state storage on fat nodes, and increasingly multi-node parallelism with the Message-Passing Interface (MPI) standard – and this is a usage model exceptionally well suited for HPC resources, in particular when combined with requirements to handle petabytes of data and extreme I/O requirements. Rather than seeing this as less advanced parallelism, researchers, infrastructure operators, and policymakers would be wise to take heed: by June 2018, it was announced that a team led by Dan Jacobson had sustained 1.88 exa-ops (reduced-precision operations, rather than normal floating point) for a neural network-training work load targeting gene interactions in biofuel crops[27] on the entire Summit supercomputer (4000 multi-accelerator nodes), several years before any traditional applications will even have access to any machine capable of an exaflop of normal-precision floating-point operations. This is one of many indications of a change that will force both users and centres to adapt to new usage models.

However, this new approach to advanced computing also has implications far beyond the computational models used. It is putting data in the front row for all studies of complex processes. A central task of future computing infrastructures will be to handle extreme amounts of collected data, for instance as described in the many application chapters in this report, but also from the internet-of-things, new types of sensors in research equipment, and mined from the text of scientific literature. Raw output data of scientific computation needs to start being treated as the valuable resource it is. Both the European Open Science Cloud[28](EOSC) and many other organisations have stressed the importance of data being Findable, Accessible, Interoperable, and Reusable (FAIR principles), and providing open access to shared data is something both scientists and centres will increasingly need to integrate into the long-term planning of scientific work –examples of this approach include the NOMAD and AiiDA consortia for novel materials discovery (as well as many others) where researchers are moving beyond individual electronic structure calculations to rather focus materials discovery on the entire process of finding results in databases. Planning the availability and dissemination of data will be central for future projects, and funding/services for this will have to be part of the infrastructure, at least for major community projects.

[27]Weighill, D., et al. Front. Energy Res. 6, 30 (2018)
[28]https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

The move to data-driven rather than model- or computation-driven research is leading to the application of advanced computing in non-traditional areas: digitised data in social sciences is opening everything from archaeological excavations to language patterns and comparative literature analysis to new types of research, and computers are making it possible to instantly trace potential epidemics outbreaks (e.g. flu) by nurses adding metadata about the number of patients calling in with sore throat or fever symptoms throughout a country. Neural networks trained on massive datasets using large computational resources are revolutionising the automated translation available on our cell phones, not to mention computers' ability to understand natural language user interfaces.

Large-scale computational analyses of social media networks and correlations between users, posts, and events were instrumental in uncovering the attempts to systematically influence elections both in the US and Europe, and in a similar way many banks and credit card companies employ screening algorithms for early fraud detection (e.g. stolen cards), and insurance companies use statistics both of previous pay-outs and projections of future events to determine premiums. The financial industry is a massive user of computational resources for arbitrage and high-frequency trading; while this might seem a nuisance for the little investor, overall it guarantees both liquidity and correct pricing for financial instruments, which makes it possible to buy/sell an investment at arbitrary times – it increases the efficiency of the market for everyone. Governments and central banks are similarly increasing users of computing to simulate and predict short- and long-term effect of policy decisions such as changes in interest rates, taxes, or welfare programs, and even police authorities are using computing to identify criminal networks based on previous convictions.

Potential breakthrough: Next-generation hardware capable of training neural networks from much larger data sets (and corresponding infrastructure to handle storage and I/O requirements) holds the potential to apply computing in entirely new areas. Used in the right way, complex data analysis can be an exceptionally powerful tool for good, just as it already helped detect undue election influence. It will help us to better understand finance, society, and political action. What means have historically been most efficient to fight poverty or famine? How should we increase (or at least maintain) participation in the democratic process? With limited resources, what new infrastructure will provide the best return-on-investment when changed behaviour in the population is considered? How can we provide advanced education to millions of students worldwide? How will new taxes or welfare programmes influence growth, the willingness to work, and the well-being of citizens? How should foreign aid be planned to improve the local economy instead of hurting it? What fraction of research funding should be spent on fundamental research vs. applied research vs. pure innovation programmes? How should different emission rights be priced and traded to ensure the largest effect on the climate, while enabling industry to transform business models rather than going bankrupt?

There are a number of projects worldwide now building longitudinal databases of individuals: their life choices, and health over the span of their lives. Computers will help us determine how different decisions and environmental factors influence health if a patient is diabetic, 10% overweight, with a certain genetic variant causing hypertension, but also a vegetarian – what long-term effects will exercise vs. hypertension medication and insulin have for their health? Gigantic amounts of money will be saved, and human health improved, if

[29]https://www.lifegene.se/For-scientists/About-LifeGene/

society could move more to preventive health care – and there are indications that widespread adoption of data-collecting devices such as smart watches or fitness bands combined with data analysis and feedback provides a type of behavioural economics "nudging" as described by Thaler[30] that promotes better health choices.
It is instructive to compare with biological sciences, where the introduction of cheap high-throughput sequencing combined with efficient bioinformatics has completely changed the nature of many areas of biology from descriptive qualitative fields to data-focused quantitative ones, which has resulted in fantastic research advances. Embracing the complexity of data and introducing computing more broadly in the social sciences and humanities will potentially have even larger positive effects on society – because they deal with more complex questions ultimately rooted in human behaviour and thoughts. Providing advanced computing resources to realise this will be of paramount importance for society.

Summary

- All major European projects, infrastructures, and users will need to adapt to the increasing focus on data-driven research and discovery. Infrastructures can no longer deliver only compute solutions, but will also have to aid in handling, maintaining, and disseminating data over longer periods.
- Neural network-based methods are making rapid inroads in all fields of science, using both measured, collected, and simulated data as input. While these usage models can be different, they are rapidly converging with traditional HPC requirements, and there are already examples of deep learning applications that have been able to sustain almost 2 exa-ops for real workloads – extreme amounts of complex data combined with artificial intelligence methods are enabling new types of parallelism.
- Embracing complexity and the move to data-driven instead of model-driven research will enable completely new applications in areas such as social sciences, humanities, finance, politics, and policymaking – but this in turn will lead to new method development challenges for researchers to not only deliver predictions, but also be able to explain why the input conditions will generate a particular outcome.
- There will be large demands on infrastructure to handle new types of I/O intensive workloads, massive numbers of short jobs, and parallelism based on ensembles or merely splitting training data to provide hardware-accelerated neural network training, and there is a large need for new generations of researchers and infrastructure specialists trained both in traditional parallelism and deep learning applications.
- With the explosive growth of sensitive data, it is important to anticipate the ethical, social, and political issues that will inevitably arise. While the infrastructure must provide means to handle sensitive data, the community also has a responsibility to develop strict ethical guidelines.

[30] Richard Thaler, Prize Lecture, Swedish Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2017

Next-Generation Computing

The role of computing and fundamental mathematics and computer science in the past, in particular their role in realising the advances mentioned above, is mirrored by similar opportunities for the future. Even the most advanced hardware advances pale in comparison to algorithmic discoveries such as the fast Fourier transform, Verlet's linked-cell algorithm, the particle-mesh Ewald method for long-range electrostatics, linear-scaling methods to solve electronic structure problems, Krylov subspace and multigrid methods, Markov state models to parallelise time and sampling, or using adaptive meshes instead of brute force calculation. The reason computing has become such a powerful tool to address the entire spectrum of environmental, scientific and industrial issues is because researchers have repeatedly developed algorithms and implementations that make it possible to utilise new generations of hardware. The root of the advances in healthcare, global warming, understanding our universe – and even deep learning – was the development of mathematics and computer science to tackle the problems.

In the same way, the future potential breakthroughs mentioned in the previous chapters will only be possible if scientists are able to use models that are larger, more accurate, ones that cover a wider range of scales in space and time, and not least using much more data. This is the core of the case for investing in more powerful computers, and unless scientists are able to use such new hardware efficiently, the societal impact will not be realised. Computer hardware is continuously evolving, and Moore's law still holds with a doubling in transistor density of integrated circuits roughly every two years. However, since 2006 the observations (Dennard scaling) describing how transistor power density would stay constant (with absolute power per transistor going down rapidly due to Moore's law) has broken down completely. This has resulted in a fundamental break with the previous development. Instead of researchers being provided a "free lunch" with compute cores of rapidly increasing frequency, individual cores no longer get faster (the clock frequency is sometimes even going down) and instead increased performance is delivered by providing many more functional units, or even accelerators that are yet more specialised.

This in turn is making large supercomputers extremely complex and heterogeneous. The cost of transferring data between different parts of hardware to follow an algorithm is now surpassing the cost of performing arithmetic operations by several orders of magnitude, and this gap is steadily growing on all levels from the processor cores through different levels of cache, main memory, and network connecting multiple nodes. Some computational applications are still able to hide this because their communication frequency is low (e.g. only once every few seconds in some cases), or the research application might be able to derive better results by moving to larger problem sizes (say, finer grids covering an airplane wing), but these are only stop-gap solutions: as the size of the resources increase, almost all current applications will eventually reach regimes where it no longer makes sense to increase the size of the simulation model, and their performance will then be limited by communication latency rather than compute performance.

Potential breakthrough: Due to hardware development, most current algorithms only reach a tiny fraction of the theoretical peak performance of massively parallel computers. Accelerators in particular have forced researchers to completely revisit fundamental algorithms such as sorting and random number generators. While this has been highly successful, it will not be sufficient to move old codes to new architectures. There is an acute need to develop new technologies and formulate alternative methods and algorithms for a whole range of scientific problems to address the challenges posed by current and emerging technology. This requires research covering the whole spectrum from new methods and algorithms for simulating complex systems to new types of runtime environments, tools such as debuggers or profilers, and not least data repositories to ensure reproducibility of results. While this will require substantial effort and long-term financial commitments, there is a clear track record of this type of endeavour leading to revolutionary advances such as the streaming computing paradigm of accelerators that is now driving deep learning.

To make matters worse, while Moore's law will likely hold for a while we will eventually reach the end of CMOS development as we know it since the architectures are a couple of generations away from fundamental physical limits related to the size of atoms. Some of this can be countered by moving to larger chip silicon dies (although more difficult and expensive to produce), but the energy demands of CMOS-based computers are soaring. Current and foreseeable demands on computing capacity, storage, and communications might soon reach limits of physical laws in terms of reliability and economics. At the same time, the vast amount, veracity, velocity, and variety of data are challenging classical methods of thinking and the very foundations of computing. In a sense, while new needs are emerging in science and society, computing still largely relies on the same concepts and principles set forward by Zuse, Turing, and Neumann eight decades ago.

Even for medium-range perspectives, scientific computing will need to explore new types of technologies such as quantum computing, analogue circuits, neuromorphic/bio-inspired computing, and distributed computing. For the first time in 75 years, scientists are presented with a golden opportunity to restructure and reinvent computing as an enabling technology and a new intellectual challenge. Beyond advances that require harnessing these emerging computing paradigms, there are opportunities to make substantial contributions based on lessons learned. In the past, computing developed across distinct branches of mathematics, hardware and software. Later efforts aiming for integration largely failed due to disciplinary and structural boundaries, but future developments will have to transcend disciplinary boundaries to find ways of solving pressing problems for science and society. The energy demands are even challenging the classical methods of inference and the foundations of scientific computing on natural laws – the existing modes of computing are likely not sustainable in the long run.

In the last few decades, computing research has raised a number of important questions: are there fundamental computing patterns that are shared across scientific fields? How are these patterns employed in hardware ranging from laptops to supercomputers? What is the role of non-numerical estimations, such as human intuition, in such simulations? How much accuracy do we need in a simulation to understand a natural system or to reach a decision? Are existing architectures sufficient, given the advances in fields such as robotics and data sciences? How do existing architectures address the vast amounts of data permeating our society? In answering these questions, scientists have been developing computational methods that can be broadly classified as multi-scale modelling and inference. Simulations that range from molecular to galaxy levels have been performed by exploiting state-of-the-art supercomputers, but as they begin to stretch physical limits of reliability and economic sustainability it gives rise to more fundamental questions about whether the traditional floating-point models used throughout science are inherently imprecise manifestations of the real world. What accuracy is needed in simulations of models with uncertainty? What is the best trade-off between imprecision and energy savings? How does the human thought process of problem-solving interface with computer hardware?

Potential breakthrough: There is a fundamental chicken-and-egg problem whereby scientific computing applications wait for the emergence of new hardware, while new hardware will not be successful without some applications showing critical advantages on it. It took almost a decade for the current accelerator paradigm to catch on, and had it not been for the gaming industry driving hardware development, it is doubtful that it would have been sustained. Application-focused researchers in turn will not even start to consider completely new types of hardware until fundamental computer science algorithms work well. To be leading in future generations of computing, Europe needs an ambitious programme to develop the algorithms that will be used by quantum computers, neuromorphic computers, high-latency but extremely energy efficient potential solutions such as using RNA for computing, and better understanding of algorithms humans use to compute. Although efficient compared to many alternatives, current artificial intelligence techniques and chips are still far behind the human brain in terms of computational power – and the latter only requires approximately 20W. While much of this work should not be expected to have much impact on applications the next few years, it is likely to define computing in general, and scientific computing in particular, in the future. While current and future needs must be balanced, investments in mathematics, algorithms and leading computer science research is ultimately much more important than hardware to make Europe a leader in future computing technologies, and these efforts have to be integrated in the infrastructure and community.

Summary

- While faster computers have been of central importance for scientific computing, the largest advances have consistently come from better algorithms.

- As Dennard scaling is breaking down, applications will need to scale over many more functional units to achieve performance. While this is possible for many codes, as hardware continues to evolve virtually all current applications will eventually be latency-limited, and there is thus a critical need to develop new types of methods that can exploit millions to billions of processor elements. Many of these already exist, but it will require large efforts to adapt them to new problems and implement them in existing codes.

- In a medium-term future, Moore's Law will likely reach the limits of physics. Although it is theoretically possible to instead add more transistors by increasing chip sizes, the power consumption will be unsustainable if we expect the same continued improvements in performance. The long-term success of scientific computing will require researchers to explore more energy efficient alternatives such as quantum computing, neuromorphic computing, or RNA computing – but none of these yet have the fundamental mathematical and computer science algorithms needed. To be leading in next-generation computing, it is imperative for Europe to have an ambitious programme of engaging in this method development, and this has to be integrated in the infrastructure handling current hardware.

Key Infrastructure Requirements

Compute & Storage Hardware

All the applications in the previous chapters have one common feature: an urgent need for access to much larger computational resources. This is not limited to a single aspect or scaling, but currently available compute power restricts the applications at stake in several ways. They require higher resolution, better sampling, longer simulations to reach relevant timescales, more computationally expensive mathematical models to improve accuracy, ensemble methods for uncertainty quantification, simulations taking more processes (e.g. both atmosphere and oceans) into account, processing keeping up with exponential growth in experimental data collection, and hardware to address new needs such as training of deep learning neural networks.

The need to improve these multiple aspects implies very demanding computational requirements. Many scientific applications are still dominated by floating-point usage, and the increased resolution of grids, larger system sizes, and more accurate computational methods needed to deliver the challenges set out in this case will require at least 50-100x larger compute performance over the next few years, corresponding to exascale, and there is a roadmap of the needs at least one order of magnitude further.

Many applications are also severely limited by memory bandwidth or communication latency. Next-generation resources will need to at least maintain, and ideally increase, the amount of memory per core, and the communication latency both between chips and nodes has to be drastically reduced (at least an order of magnitude) for algorithms based on fast iterations to achieve significantly improved performance.
However, this does not mean a single exaflop machine reserved for one application at a time is an urgent need. Apart from a few niche cases (such as rapid training of neural networks), few if any applications will exhibit perfect scaling, and with large competition for resources it will be much more efficient to have hundreds of applications sharing a machine rather than having access to the entire system 1/100th of the time. New algorithms such as Markov state models or ensembles for uncertainty quantification achieve significantly better scientific answers by using collections of many loosely coupled simulations to provide standard error estimates. In particular, since these methods exhibit close to perfect scaling (or even super-scaling when combined with adaptive sampling), the most likely high-efficiency usage model of next-generation systems will be simulations that scale over 5-10x more processor elements than today, combined with running 10x more loosely coupled jobs. This will naturally address many of the challenges with fault tolerance in extremely large machines, since individual members of ensembles can be restarted without affecting the rest – but it will require significantly more flexible queue systems than available today. However, as a consequence, it might be more efficient for next-generation systems to be designed as loosely coupled “islands” of hardware rather than aiming for full bisectional bandwidth between all nodes – because the total cost efficiency must be competitive with smaller resources.

Although the name stuck several years ago, the strong focus on “exaflop” is unfortunate. Infrastructure investments of this size should rather be decided entirely based on the performance improvement delivered for the scientific computing applications used, with the goal of achieving 50-100x better results than current 10-20 PFLOP systems. The field has a long tradition of focusing on the usage of the floating-point units and networks, but this is highly artificial today. Few of us worry about not using the branch prediction units in a processor enough, that we are leaving some memory unused, or that we might not be using the full memory bandwidth.

In the same way, the only relevant measures should be (i) the importance of the scientific problem, and (ii) how fast/efficiently a given machine can solve it. Leaving part of the network underutilised is no more severe than not exercising the entire memory bandwidth.

The fraction of codes that can use accelerators is increasing rapidly. This is a highly valuable addition for problems that depend on rapid iterative algorithms since the accelerators achieve much higher performance per node. However, as nodes increasingly come equipped with 2, 4 or even 8 accelerators, the gap in bandwidth and latency inside a node vs. between them will become even more severe, and many codes might require significant development efforts to keep up with this. For codes that program accelerators directly rather than using standard libraries such as BLAS/LAPACK, it is also a concern that the current accelerator landscape has a fairly heavy bias towards a single vendor. Some of these codes might require large porting efforts if Europe decides on a different type of accelerator architecture, which in turn will take several years to implement and test. For the next 1-2 generations, this means there is a need for a balanced and diverse infrastructure landscape providing accelerators to codes for which it is cost-efficient, but also more traditional CPU resources for codes not yet accelerated or which rely on script languages.

While it is easier to move and port codes between different CPU architectures, a large number of scientific computing applications are already highly accelerated and tuned for the most common processors. Extending this to more architectures is a non-trivial effort that can take years to accomplish. Furthermore, one experience from the last few years is that a huge diversity in architectures, combined with strict requirements on showing scaling for the specific architecture, creates an artificially segmented infrastructure where a typical application might only be eligible to run on a single machine. The priority then becomes more a matter of applying to the right machine than formulating the best proposal. To avoid this and improve the mobility of awards, the number of architectures in future European infrastructures should probably be limited to one accelerator and one non-accelerator architecture, and for both of them there has to be a natural progression for users with a smaller system available in the 2019-2020 timeframe, and a larger machine a few years later.

Storage and I/O requirements are expected to grow even faster than compute needs, with 100x higher needs than today within the next five years, in particular with much larger data sets being used e.g. for data-driven research and deep learning. This must be coupled with provisioning of a large-scale end-to-end data e-infrastructure to collect, handle, analyse, visualise, and disseminate petabytes to exabytes of data, for instance by providing high-bandwidth gateways to future EOSC-provided dissemination.

Operations & Environments

The scientific computing world is 100% standardized on Linux today, and it is imperative that any future architecture provides a full Linux environment including both POSIX-compliant file systems, the common development tools with free C/C++/Fortran compilers from GNU and LLVM, as well as potential vendor-provided compilers, parallel debuggers, and all common libraries used in HPC – all of which need to support the latest language standards. Nodes must also support execution of scripting languages such as Perl/Python, and accelerator hardware must provide support for some sort of standardised open API to enable long-term portability of codes.

As scientific computing has become more diverse, high-end resources will also be expected to cater to much broader categories of users. Queue systems will need to be adapted to handle orders of magnitude more jobs than previously possible to enable ensembles of simulations as well as new users carrying out many independent tasks e.g. in deep learning or bioinformatics, and we expect rapidly increasing demands for high-performance computing environments to support containerised applications using e.g. Docker images with MPI support. Supercomputer centres, with their profound expertise of computing architectures and programming models, have to recast their service activities in order to support, guide, and enable scientific program developers and researchers in refactoring codes and re-engineering algorithms, influencing the development process at its root, and also making sure highly skilled staff develop close relations with key user communities. These services should be provided for longer periods, and the resulting codes adapted to the whole spectrum of machines, with users of the community codes determining what resources are most suitable to apply for. For Europe to have a leading infrastructure, it is also critical that compute centres have staff participating in the development of open compilers and language standards.

These trends are not limited to the high end of supercomputing but apply to all tiers of the HPC ecosystem. The compute node architectures will follow current technology trends and integrate thousands of cores, including accelerators. The effort involved in software refactoring is substantial and requires a deep knowledge of the algorithms, of the codes and, in many cases, of the science at stake. Therefore, a successful practice is to place these activities in community projects where substantial code development activities are common and long-term software development activities can be sustained.

Finally, one of the single most successful aspects of PRACE is the comprehensive training programme that includes activities targeting users from novices to experts. For scientists to perceive the European-level organisation as their primary community, it is imperative that these programmes are expanded, and that there is a clear progression where users start with introductory courses, continue with more advanced activities, and eventually engage and start to influence the community through masterclass-level education.

Summary

- Applications are ready to use at least 100x more computational resources within the next few years, but also require similar increases of storage and I/O bandwidth.
- There is a clear need for exascale computing, using ensembles of many simulations for higher efficiency, but few applications have any need to execute on an entire single exaflop machine for very brief periods of time.
- There is a need for both accelerator and traditional CPU architectures, but too large a diversity creates artificial segmentation where applications will only be eligible for one system.
- Queue systems will need to evolve to support both containers and much larger sets of active and queued jobs such that fault tolerance can be achieved with automatic checkpointing and restarting of individual tasks instead of a complete ensemble.

Recommendations

Computing in general and scientific computing in particular have outstanding track records of providing breakthrough research results, advancing society and providing a strong basis for commercialisation and growth. There is a huge diversity of research and applications in Europe, ranging from fundamental physics to direct applications of seismic modelling or artificial intelligence in industry. These areas all have clear roadmaps with expected breakthroughs over the next few years that will only be possible by moving to 50-100 times more powerful infrastructure resources both for computing and data.

Many European countries have invested in national computing resources, which can be highly cost-efficient, but the insufficient infrastructure on the joint European level is a major weakness. To be competitive in the world, Europe needs to develop synergies, promote clusters of excellence, build a critical mass of researchers from several countries collaborating on challenges, enhance industrial use of computing, and identify areas where Europe can take an internationally leading role, in particular in software. This requires a paradigm shift where researchers and industry perceive the joint European-level infrastructure and community as their primary point of engagement, excellence, and innovation.

For Europe to achieve this, it is imperative to move beyond 3-5-year funding cycles, primarily involving hardware, and rather focus on what type of infrastructure and long-term collaborations are required. While computational hardware accounts for a significant share of the cost, the long-term success and sustainability of a computational ecosystem is more dependent on our decisions about how we invest in human infrastructure to ensure the most creative and skilled individuals want to dedicate themselves to careers in computational infrastructure, academia, and industry.

Despite the costs, it is critically important to recognise that not investing, or relying on national initiatives, is itself an active decision and one that will have direct, grave, consequences for Europe. European academia and industry would be increasingly fragmented, uncompetitive, and irrelevant in an age where the US, Japan, and China are making gigantic investments in computing. Much of the research would still be accomplished – possibly even by the same individuals working abroad – but it will be coupled with a massive brain-drain and loss of advanced European innovation in the most rapidly growing field of technology. While some results will no doubt disrupt existing business models, they create much more in terms of new markets and employment opportunities in the highly-skilled, technology-focused, domains which Europe must focus on to avoid competing only within the level of the lowest salaries.

Based on this, the PRACE Scientific Steering Committee recommends:

- Europe must urgently rectify its significant under-investment in joint computational infrastructure compared to the US, Japan, and China. EuroHPC is one important way to achieve substantial investments in hardware, technology, and software, but as these investments also compete with funding for research projects, it is essential that it is highly cost-efficient, that it provides a long-term vision and strategy to deliver the resources on which academia and industry depend, and that it results in deep engagement of computing researchers to build excellence at a European level.
- While PRACE has served the community well, to justify significantly increased resources (in particular given the competition with grant funding) and ensure decisions are steered by scientific impact, highly-merited active scientists must have direct influence on the highest level of governance of future infrastructures, rather than merely providing advice. Similar to many other research infrastructures, this is best achieved by having both a chair, vice chair, and scientific/executive director who are professors with strong scientific track records e.g. by having been awarded prestigious research grants for their academic work, and the scientific users must have equally strong representation in all decision-making bodies. An industrial director with extensive experience from industry would similarly strengthen the industrial involvement.
- Access to resources should be based only on scientific excellence (including industrial scientific needs for such programmes) combined with technical readiness. It is imperative to maintain the high-quality scientific peer review programme developed in PRACE, but it must be combined with a reduction in the number of architectures and simplification of the resource allocation process so that there is a better correlation between scientific ranking and awards.
- Even with fewer different systems, the infrastructure needs to reflect the diverse needs of users, provide both accelerator and non-accelerator systems, and support different job patterns. While some applications can scale exceptionally well as a single parallel simulation, others achieve better efficiency and outstanding impact with ensembles of simulations, massive deep learning training, or high-throughput analysis of medical images. All of these are excellent approaches. The justification for access to highly expensive infrastructure must be based solely on the expected impact and suitability of the approach. Queue systems must be adapted to fit the research, not vice versa.
- Providing solutions for processing vast amounts of data is becoming just as important as computation, and there is a clear path of convergence with both needing access to the same infrastructure resources. New simulations and faster computers generate unprecedented amounts of data that has to be analysed, and the petabytes of information generated by data-focused research needs advanced storage, extreme I/O capability, and large amounts of cutting-edge accelerator processing power for training. A leading European infrastructure has to cater to all these needs.
- For users to keep pace with next-generation technologies, it is imperative that a larger share of investments is spent on software, algorithms, and education – the human side of the infrastructure. All fields of research are facing a bottleneck where access to programming and development expertise is a major obstacle to addressing the challenges described here. The H2020 Centre of Excellence action is one promising approach, but to achieve European infrastructure leadership there have to be resources reserved for compute centre staff to allow them to have long-term engagement in development of compiler technology, new parallel algorithms suitable for accelerators and many-core processors, as well as development of the most widely used scientific codes, where European research groups are internationally leading in many areas. These are exceptionally talented experts who are in high demand in industry, and to retain them it is critical that they perceive the academic and infrastructure environment provides clear career paths with opportunities for more responsibility.

- The rapid advances in data collection and artificial intelligence will enable new uses of information, much of which could be sensitive. It is important to anticipate the ethical, social, and political issues that will inevitably arise (and which to a limited extent we can already see happening). The community must develop ethical guidelines and standards for the use of HPC; with great power always comes great responsibility and HPC technology is no exception.

In closing, few (if any) other infrastructures provide the type of return-on-investment achieved by advanced computing. It is ubiquitous in both science and society, and Europe is currently at a crossroads where scientists in both academia and industry have clear roadmaps of the applications and impact that would be achievable in just a few years, but which are stymied by currently available infrastructure resources. Investing in an ambitious, cost-efficient, and user-focused comprehensive European computational infrastructure will not only address this, but also provide the commercialisation, technology development, training, and workforce education necessary for Europe to stay competitive in a world increasingly dominated by computing and artificial intelligence.

www.prace-ri.eu