PRACE

EuroHPC Joint Undertaking

EuroHPC Summit Week 2021 & PRACEdays21

Executive Summary

EuroHPC SUMMIT WEEK 2021 & PRACEdays21

EUROHPC Summit 2021 Executive Summary

www.prace-ri.eu

Editor-in-chief

Marjolein Oorsprong – PRACE Communications Officer

Proofreading

Marjolein Oorsprong – PRACE

William Davis – Insight Media

Article editor

Harry O'Neill, Insight Media

Graphic design

Alison Davis, Insight Media

Published by Insight Media, 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 823767.

ISBN: 9789082942286

Foreword

Serge Bogaerts, Managing Director of PRACE aisbl and Chair of the EuroHPC Summit Week 2021 Organisation and Programme Committee (OPC)

This year's EuroHPC Summit Week was certainly one to remember. When COVID-19 forced us to cancel the event planned to be held in Porto in March 2020, we were still confident we could shift everything by one year and welcome you to that wonderful Portuguese city in 2021. But the pandemic lasted, and more restrictions followed, and towards the end of 2020 we realised we would have to change the set-up of the conference completely, from a live event with limited livestreaming, to a fully digital event.

And, just like organising a live event entails a lot more than just renting a few meeting rooms, setting up event space is only a small step in the entire script of an online event.

The effort was huge – and I would like to thank all the teams behind the scenes who worked tirelessly with us to achieve this amazing feat. We attracted almost triple the number of participants of any edition that has gone before, and the experience has certainly taught us some valuable lessons and raised the bar for the coming years.

To keep the momentum of EHPCSW 2021 going for a while longer, we have created this executive summary of the EuroHPC Summit Week 2021 and PRACEdays21. From green HPC to diversity, and from quantum computing to COVID-19, the EHPCSW 2021 included a wide variety of HPC-related speeches, presentations, panel discussions, and posters, and the articles on the following pages include interviews with some of our high-level speakers and recaps of session held throughout the week.

I wish you happy reading and hope to meet you again in person at EHPCSW 2022, which will be held in Paris in March next year.

"The effort was huge – and I would like to thank all the teams behind the scenes who worked tirelessly with us to achieve this amazing feat"

Table of Contents

PRACE User Forum 2021

The PRACE User Forum is an independent entity where PRACE users can discuss their experiences and express their future needs as well as give feedback on the current services and resources provided by PRACE. The aim is to provide an effective mechanism through which the Tier-0 user community can give feedback to PRACE.

At PRACEdays21, the PRACE User Forum held a session where they presented a number of points of interest about the feedback they receive, as well as providing an open session for users to voice their concerns and give suggestions in the presence of PRACE representatives.

After an initial talk by Catarina Simões & Klara Meštrović, in which they described the finer details of the peer review process that PRACE uses to select the best applications based on scientific excellence, Troels Haugbølle presented some of the key findings of a survey of all past PIs of the PRACE infrastructure carried out in 2019, and also addressed many of the pressing questions surrounding the future role of PRACE with regards to the incoming EuroHPC infrastructure.

What is a typical PRACE user?

The User Forum survey showed that the typical user group of PRACE uses between 10-100 million core hours, with 10% of groups using more than that. Research groups tend to use codes that are adapted for multicore processors. Some groups run part of their science using accelerated nodes or hybrid setups, but this is still a minority of users. Another interesting result from the survey is that the majority of user groups use in-house developed codes, especially the groups using more than 100 million core hours.

What do users think about PRACE?

Participants in the survey were asked their opinions on PRACE. One response, which mirrored that of many, said that "there is no alternative to PRACE if one needs an allocation in the order of 100 million core hours". This shows that at the European level, HPC resources are really crucial for the digital agenda. Other key views coming from the survey were that PRACE is a very well-respected organisation for their fair and balanced application and review process. One of the keys to the success of PRACE has been that its infrastructure has been developed bottom-up with input from scientists and computing centres.

"There is no alternative to PRACE if one needs an allocation in the order of 100 million core hours."

EuroHPC and PRACE

EuroHPC will start delivering core hours to researchers this year, and will supersede PRACE peak performance by a factor of five by 2022. This represents a great increase in computing capacity, but the pressing question is: where does PRACE stand in this picture? One thing that is crucial to remember is that EuroHPC is spending all of its money on accelerated architectures, with a 90-10 split between GPUs and CPUs, meaning there is a real need for a transition in the software that people are running. PRACE will still retain a similar amount of performance when it comes to multicore architectures, and therefore still very much has a part to play.

For instance, the PRACE peer review process is recognised as the de facto gold standard. As well as this, PRACE also delivers a lot of other things besides computing hours that are important for European HPC users. The PRACE training centres produce human capital that can use the European data infrastructure. Thousands of graduates and post-graduates have been trained in programming techniques and domain-specific applications in a variety of fields by the best experts in Europe. This was made possible by opening these courses at computing centres across Europe to everyone who is part of the PRACE infrastructure.

As well as this, the PRACE High-Level Support Teams have been playing an essential role in enabling new and established research groups to transition to new architectures and in making codes more efficient. This function will be mirrored in EuroHPC by some new competence centres, which will be crucial components for getting people ready to use new infrastructure.

PRACE has also been exemplary at establishing new collaborations, one example being its work with

Troels Haugbølle, vice-chair of the PRACE User Forum, presented some key findings from a survey of past PIs

the Fenix infrastructure to support analysis and data-intensive science. Finally, PRACE will remain relevant for all researchers from countries who are not yet members of EuroHPC, the UK being a notable example. PRACE will be able to secure Tier-0 resources for people working in those countries.

The user survey indicated that the majority of research groups are still unable to carry out their science using GPUs. This is probably the single biggest challenge that will need to be tackled at European level. Some scientific applications run on GPUs, others will eventually, but an important subset does not match the underlying hardware. This is why PRACE will still remain relevant to many scientists, by offering the same amount of pure CPU performance as EuroHPC.

The history of PRACE has shown that having a diversity of architectures is important for supporting the diversity of scientific applications. Input by applied scientists from academia and industry, who are the users of the infrastructure, helps to broaden the scope and improve access rules. Therefore, the PRACE User Forum believes it would be constructive for EuroHPC to establish a similar user forum, to give users a channel for their feedback. This is a common practice in all large science infrastructures, including CERN, ESO, ESA and ESS.

"The PRACE User Forum believes it would be constructive for EuroHPC to establish a similar user forum, to give users a channel for feedback."

Broad usage of the EuroHPC capacity will require sustained high-level support or funding programmes aimed at transitioning current Tier-0 research groups and communities to accelerated architectures. Otherwise, there will be a real risk of undersubscription. It will not be enough to just focus on middleware and optimised libraries – there is a real need to maintain European excellence in software and application development. Human resources are present throughout industry and academia, and software support is much less costly than hardware.

Urgent computing: HPC for emergencies

Urgent computing refers to the use of HPC resources during an emergency situation. These can include natural hazards such as earthquakes, tsunamis, volcanic eruptions and pandemics, but also anthropogenic hazards such as marine oil spills. The ChEESE Centre of Excellence's *European Urgent Computing Workshop*, which took place on the first day of EHPCSW, set out to explore some of the work being carried out to develop state-of-the-art workflows that will help mitigate the damage of future emergencies. Over the next five pages, we hear from six different researchers about their work on specific types of emergencies and how high-performance computing is helping.

Radiological dispersion in the atmosphere

Antonio Cervone, Italian National Agency for New Technologies

Atmospheric dispersion of radioactive pollutants can be unpredictable, or in the best case can be predicted with extremely short notice. Working out the best way to protect people from this hazard can be tricky. Counter-measures include sheltering inside and iodine prophylaxis, but these take time to implement and have their drawbacks. Counter-measures therefore need to be carefully matched to the levels of risk posed by the pollutant dispersion, which is where the role of high-performance computing comes in.

Risk of receiving a dose of radiation can be calculated by predicting both the atmospheric dispersion and the dilution in air of the radioactive material. Often, these things are not precisely known, so best estimates and uncertainty calculations are needed. Real time measured data and high-resolution numerical weather prediction data need to be coupled to help in these calculations.

There are three possible scenarios which can lead to the atmospheric dispersion of radioactive pollutants. The first of these is accidents in nuclear-related installations, such as the Chernobyl disaster in 1986 and the Fukushima Daiichi disaster in 2011. The second potential scenario is the use of malicious dispersion devices, such as the use of dirty bombs by terrorists. The final scenario type is known as a hidden emergency, of which one of the better-known examples is the cloud of radioactive isotope ruthenium-106 that was detected over Europe in 2017.

Each of these scenarios require different levels of resolution, domain size and response times when considering urgent computing. A terrorism event using a dirty bomb would likely need a finer resolution and a quicker response time, but the domain size needing to be addressed would be smaller. A reactor accident, however, would require less resolution and a slower response time but over a larger (continental or even global) domain size. In any case, calculation times of these scenarios should be around one tenth of the ideal response time, which requires robust data streamlines, automatisation and 24/7 preparedness.

Antonio Cervone

Recent simulation tools such as Eulerian models, Lagrangian models and CFD provide much higher accuracy and performance than was possible in the past, but they are computationally demanding and require longer execution times. A lot of measured data is needed to calibrate models and check results. For large scale applications, data networks for have existed for many years, but for microscale applications, streams of data are less available.

Weather data is a necessary ingredient in this work. For microscale simulations, this must be interpolated and adapted to the specific 3D computational domain. The uncertainties that affect numerical weather prediction data are taken into account using ensembles. Additional, slow-varying databases – population distribution, emission sites, land-use – are required for some features of simulations, as is data about buildings in 3D domains.

With these kinds of simulations, it is very important to have as high degree of fidelity as possible in order to better protect people. Timely and best-estimate responses require fast and efficient HPC implementation of codes and the development of integrated platforms. Although this theme is recognised as a top priority by various parts of the EC, adequate funding schemes have not yet been made available. There is therefore a need for better recognition of urgent computing within EuroHPC.

Tsunami computation

Steven Gibbons, Norwegian Geotechnical Institute

Destructive tsunamis are infrequent, with most exposed regions often going many generations between catastrophic events. The 2004 Indian ocean tsunami was the first big tsunami of the modern media age, in which 230 000 deaths and images spread across the world brought global awareness about tsunamis. Later, the 2011 Tohoku earthquake and tsunami provided a textbook example of a cascading disaster, with the subsequent catastrophic failure of the Fukushima nuclear power plant causing many years of difficulties.

Tsunamis can come from landslides, volcanoes and even severe weather, but 80% come from earthquakes. Tsunami waves propagate very efficiently across oceans, and the earthquake source and shape of the ocean floor both have a big effect on where a tsunami will hit hardest, so both need to be modelled accurately and at the right scale.

Fast and accurate numerical simulations of tsunamis require efficient use of HPC, and for this reason it is a significant area of focus for the ChEESE CoE. The computational engine used is the Tsunami-HySEA code developed at the University of Malaga. Even on a single GPU, it can perform tsunami simulations with substantial dimensions and resolution faster than real time.

Optimal splitting of the spatial domain allows simulations to be split over multiple GPUs, further reducing the computation time. However, scalability depends considerably on the physical parameters of the seafloor and the coastline needing to be resolved. A hypothetical tsunami in the Caribbean area would scale less efficiently due to the complex nature of the multiple island coastlines in the region, which introduces code imbalance.

A number of pilot demonstrators in this ongoing project focus on different aspects of tsunamis. In faster than real time tsunami simulations, the code is pushed to the extremes, priming GPU optimised code to simulate tsunamis on vast and complex spatial domains as rapidly and as accurately as possible.

Steve Gibbons

A basic workflow for a faster than real-time tsunami simulation starts with specification of the earthquake source. For a given magnitude, location and focal mechanism, the resulting seafloor displacement is calculated. This is guided by knowledge of the region's tectonics and preliminary seismic data. The tsunami wave travels from the source to the coast of interest, where a far finer resolution is used to capture the coastal inundation.

The ability to visualise the output is crucial for validating the code, for research into tsunami generation and impact, and for conveying the hazard to stakeholders. Properties of the wave can be displayed over wide areas for a snapshot in time, or you can display how wave height at a given location changes over time. It is also possible to show maximum levels of inundation to a high spatial resolution, or look at inundation for a given scenario in relation to infrastructure and planned evacuation zones.

Probabilistic tsunami forecast is one of the ultimate tests of urgent HPC. This is where a large earthquake has been reported and there is an urgent need to predict, ahead of real time, the likely tsunami inundation. With these real emergencies, information is usually quite sparse. It may be many minutes after the event before an earthquake is located accurately. Its size, focal mechanism and seafloor displacement are subject to great uncertainty – just like a weather forecast with limited information, many different outcomes are possible. A practical tsunami forecast must therefore accommodate this uncertainty so that probabilities can be ascribed to each of the outcomes.

As more data comes in, different scenarios become more likely or less likely, and some will be ruled out. Depending on the computing resources available, various scenarios are calculated, and the likelihood of the effects of inundation along the coast are calculated. Decisions can then be made with alert levels, using rules set out by civil protection authorities.

The probabilistic tsunami forecast workflow is now operational as a prototype. It was developed and evaluated using past earthquakes and tsunami events in the Mediterranean, and is now being trialled on new earthquakes as and when they happen anywhere in the world. Finally, a third tsunami pilot demonstrator in the ChEESE CoE is the probabilistic tsunami hazard analysis. Computationally infeasible until fairly recently, this can show why certain parts of coast are constantly inundated due to seafloor shape and the topography of the land and shore.

Seismic hazard assessment

Marta Pienkowska, ETH Zurich

Large earthquakes are rare and thus unpredictable. In the last decade there have been around 125 earthquakes of large magnitude, and they are always a surprise. The regions and areas where they can happen are known, and experts try to assess probabilities of such large events in specific areas, but knowing when a large earthquake might actually strike is beyond our capabilities at present.

The goal with within seismic urgent computing is to explore the possibilities of urgent supercomputing to obtain fast (within a few hours) physics-based ground shake maps for regions that are affected by recent earthquakes. Right after an earthquake, it is important to know what areas were likely subjected to the highest intensities and what the probable impacts were in those areas.

At present, shake maps combine ground motion prediction equations (GMPEs) with observed data from a dense network of sensors that supply information in real time as the event happens. The real data collected immediately after the event significantly reduces the uncertainty of these maps, but in areas that are not as well instrumented, shake maps are significantly less reliable. For this reason, researchers are still looking for and exploring ways that would help to better assess damage in the aftermath of an event and improve rescue management decisions.

One option is to use phyiscis-based simulations to contribute to hazard assessments immediately after an earthquake. For a domain of about 100 x 100 x 50 kilometres, simulating elastic wave propagation with a dominant frequency of about 10Hz would require in the order of 90 000 core hours. Simulating these high-frequency waves is most certainly an HPC-sized challenge, and it is these frequencies that are most relevant for structural engineering and damage assessment. Although not possible now on such short timescales, this is the ultimate aim of seismic urgent computing.

An urgent seismic simulation workflow has been developed by the ChEESE CoE. Part of this is the automatic alert service, an ongoing background process that screens earthquake databases for high impact events. Once a potential high-impact event is registered an alert is triggered which starts the workflow manager, which then sets into place various independent tasks in parallel.

Marta Pienkowska

Part of the workflow involves deciding whether the event is a candidate for an urgent simulation in a more robust way. This is done using machine learning algorithms, which look to see whether the earthquake is greater than or equal to seven on the Modified Mercalli intensity (MMI) scale, which categorises earthquakes by its effects at a given location rather than via direct seismic measurements (as in the Richter scale).

It is then necessary to prepare the input parameters for the wavefield HPC simulation. These parameters include the location, depth and magnitude of the earthquake, which are all provided by external agencies, as well as the focal mechanism. Typically, this particular piece of information is not available immediately after the event, so a quick estimation of the focal mechanism must be done using either a statistical method or a clustering algorithm.

Once the input parameters have been finalised, they are prepared in the YAML file format so they can easily be plugged into different wave propagation solvers. Wavefields can then be created, showing the earthquake's kinematic rupture propagating through time and space, which can then be processed to provide the final shake maps.

Volcanic ash cloud forecasting

Arnau Folch,
Barcelona Supercomputing Centre
Sara Barsotti,
Icelandic Meteorological Office

Volcanic ash clouds are generated during eruptions, where large amounts of particulate material and aerosols are injected into the atmosphere. Coarser material rapidly falls to the ground, which has a big impact on the surrounding area. However, the finer particles remain entrapped in the atmosphere and are transported downwind, forming large ash clouds that pose a hazard to civil aviation.

Systems are in place for forecasting ash clouds every time a volcanic eruption happens. These are overseen by the Volcanic Ash Advisory Centres (VAACs) that cover the whole planet and issue forecasts for the trajectory of those clouds. However, these forecasts have low space-time resolution, only provide semi-quantitative information (ash or no ash), and do not provide any forecast uncertainty metrics.

There are several reasons for these shortcomings. Quantifying what is happening in real time with a volcanic eruption is very difficult, as is providing information at the space-time resolution needed by civil aviation authorities. As such, the ChEESE CoE is now working to develop a pilot demonstrator that can provide increased-resolution ash dispersal forecasts, which will also include the use of ensemble-based data assimilation.
This new workflow will combine the FALL3D dispersal model with high-resolution geostationary satellite retrievals to provide higher resolution (km grid size) forecasts at continental scale. It will be fully automated, running at all times with a forecast window of three days in the future, and will be able to deliver its forecast products within a maximum of one hour from when an alert is triggered.


Arnau Folch

Ensemble-based forecasts can provide two types of products. Firstly, they can provide deterministic products based on a combination of the ensemble members (e.g. ensemble mean). In the context of ash cloud forecasts, an example would be providing the concentration of ash found at a particular flight level, or the ash cloud top height. The second type of product is probabilistic products, which are based on counting the fraction of ensemble members that verifies a certain threshold. For example, the probability of ash concentration at a certain height being above an engine safety threshold could be given.

Volcanic ash clouds remain one of the biggest challenges for urgent computing, and this demonstrator is showing what HPC is now capable of in these situations. In March 2021, a VOLCICE exercise was scheduled by the Icelandic Meteorological Office (IMO) to practice the response to an explosive eruption at Beerenberg volcano in Norway. The pilot demonstrator developed through ChEESE was used to test out how it might work in reality. When compared with other similar forecasts, it has better time latency, much higher time-space resolution, uses ensembles instead of single run forecasts, and provides a wider array of quantitative products such as probabilistic forecasts.

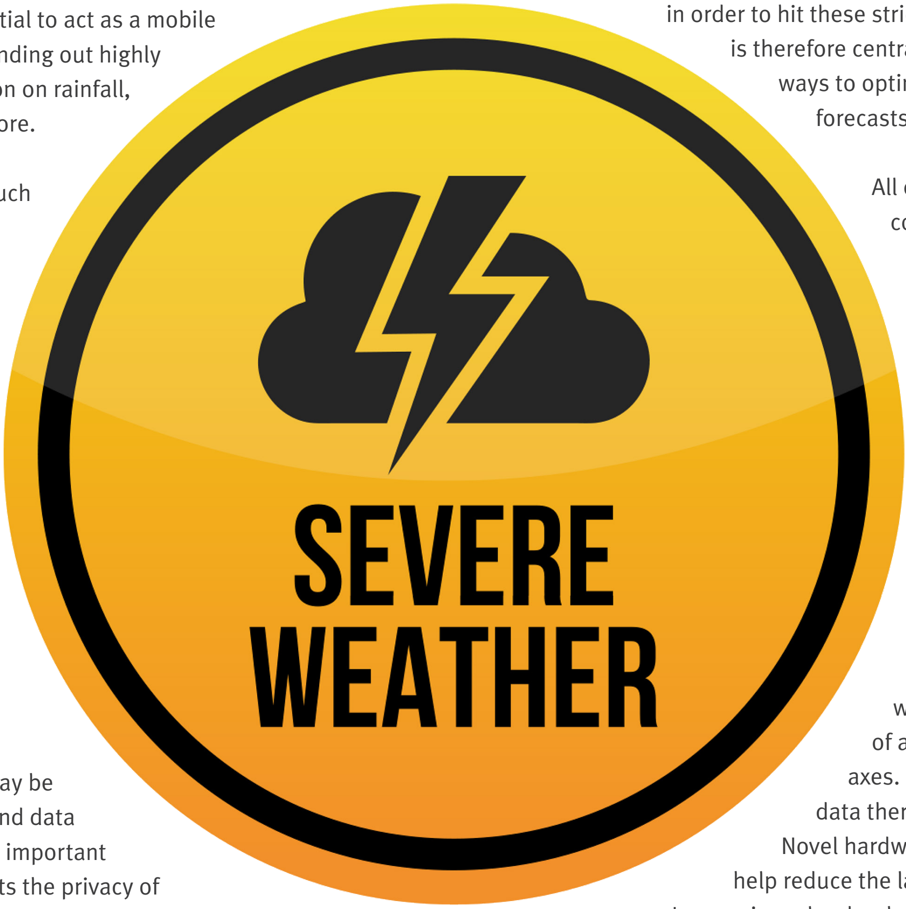
Advanced prediction of weather extremes

Tiago Quintino, ECMWF

Extreme weather events can have impacts way beyond what we normally consider. A heat wave doesn't just cause more people to go to the beach – it can lead to crop failure, which can then cause the price of crops to increase, which in turn can cause a rise in deaths and social instability. Extreme weather events are therefore well worth being able to predict.

The European Centre for Medium-Range Weather Forecasts (ECMWF) provides products that are used in many ways. The organisation is now facing a number of challenges as the methods for gathering, storing and using data change. For instance, the rise of the internet of things – the myriad of connected sensors that surround us – means that incoming data now arrives in many forms.

Nearly every mobile phone now has a pressure and temperature sensor inbuilt, and when people use weather apps, the user's phone is often roped in to becoming another node in the worldwide network of weather sensors. Looking to the future, as driverless cars become more common, each of them has the potential to act as a mobile weather station, sending out highly localised information on rainfall, temperature and more.



For organisations such as the ECMWF, all of this means that they now need to deal with many different types of data of varying levels of accuracy and reliability, and little harmonisation. Mobile phones in pockets will give the wrong temperature. Personal weather stations set up in people's gardens may be calibrated wrong. And data anonymisation – an important process that protects the privacy of users – purposefully adds errors to data locations to prevent people from being tracked. Good for privacy – bad for weather forecasting.

Another challenge for weather forecasting is the increasing demand for higher definition earth system models. The first way one can increase definition is by increasing model resolution, conceptually fairly simple. Another axis of definition is model complexity. By introducing more physics – better descriptions of soil, better models of radiation – the definition of the model increases. Finally, ensembles have the ability to increase the reliability and therefore definition of a forecast – the more ensemble members, the better. For extreme weather prediction, all three of these axes must be developed.

Tiago Quintino

However, the increasing definition of earth system models presents a challenge to urgent computing in one of its defining traits – it must be done quickly. Forecasts need to be delivered within one hour, and the parallel file systems used in HPC are put under huge strain when reading and writing at the same time in order to hit these strict deadlines. Access to data is therefore central when thinking about ways to optimise the speed of weather forecasts.

All of this brings about the conundrum of the best way of providing access to data in a timely manner. Weather forecasts can be looked at as 6D-hypercubes, with dimensions of latitude, longitude, variables, timesteps, probabilistic perturbations and more. The problem is that different users want to look at and access that data in different ways, carrying out a variety of analyses across multiple axes. New methods for accessing data therefore need to be explored. Novel hardware, such as NVRAM, can help reduce the latency of data access. In-premises clouds where users can access data and run their own workflows could also help. As the size of these data hypercubes reaches petabyte scale, ongoing work on software and specialised hardware will be needed to maintain their usefulness to extreme weather forecasting.

Understanding supercapacitors

The PRACE Ada Lovelace Award for HPC 2021

Dr Céline Merlet of CNRS is the winner of the PRACE Ada Lovelace Award for HPC 2021. Speaking at EHPCSW, she outlined her research which has led to a deeper understanding of the functioning of supercapacitors: high-power electrochemical storage devices that have a number of applications for enabling a sustainable future.

Dr Céline Merlet

Launched in 2016, the PRACE Ada Lovelace Award for HPC is awarded annually to a female scientist who makes an outstanding contribution to and impact on HPC in Europe and the world, and who serves as a role model for women who are at the start of their scientific careers.

The award is named after the Countess of Lovelace, a British mathematician who lived in the 19th century and among else worked with Charles Babbage on the machine they called the Analytical Engine — one of the first precursors of computers. Many historians regard Ada Lovelace's contribution to this mechanical calculator as the very first algorithm — and herself as the first person to be rightly called a programmer.

Earlier this year, PRACE announced that the winner of the 2021 award was CNRS researcher Céline Merlet of the Centre Inter-universitaire de Recherche et d'Ingénierie des Matériaux (CIRIMAT) in Toulouse. Through her research, Merlet has decisively advanced the investigation of supercapacitors with numerous innovations. As well as this, she is active in "Femmes & Sciences", an association that promotes women in science and technology.

Dr Merlet received her PhD degree in 2013 from Université Pierre et Marie Curie in Paris where she had worked on molecular simulations of carbon-carbon supercapacitors. She then joined the University of Cambridge as a postdoctoral researcher, working on simulating NMR spectra and diffusion of ions in energy storage materials such as porous carbons and lithium manganese oxides. Since 2017, she has been working at the CIRIMAT laboratory.

At this year's EHPCSW, Dr Merlet outlined her work in a talk entitled "Multi-scale models and HPC for a better understanding and performance prediction of electrochemical energy storage systems". Improving the performance of energy storage systems is an important aspect of moving towards a world where renewables are the predominant source energy. Everyone is familiar with batteries, which are present all around us in our laptops and mobile phones, but less well-known are supercapacitors. These systems charge and discharge extremely fast in comparison to batteries. One current use for them is regenerative braking in hybrid buses, where energy is stored from the braking and stopping and then used later to start the bus, saving up to 30% of fuel.

One of the main issues with supercapacitors is their relatively low energy density when compared to batteries. Electrodes in supercapacitors are made from porous carbon, and Merlet's career so far has focused on investigating the details of how they work so that they might be improved upon. High-performance computer simulations have enabled her to look more closely at the details of how they function. Classical molecular dynamics at the scale of nanometres and nanoseconds show how the anions, the cations and the solvent move through the pores of the carbon, while mesoscopic simulations help to bridge the gap between findings from molecular simulations and experiments.

One of Dr Merlet's key publications was in Nature Materials in 2012 on the molecular origin of supercapacitance in nanoporous carbon electrodes, which has become a key reference with more than 600 citations. Back then, simulations of supercapacitors were only feasible using simplified and inaccurate electrode structures, with the materials being represented by crude approximations resembling plain sheets or tubes. An unanswered question from these simulations was why a large increase in capacitance was seen when the size of the pores in the carbon was below one nanometre. This was a surprising result as this pore size is close to the size of the ions, which many presumed would make it difficult for ions to enter. Dr Merlet's contribution to the investigation of these high-power electrochemical storage devices was to develop efficient and accurate simulations of the complex porous structures of carbon electrodes in operating systems. This resulted in a much more realistic representation of supercapacitor's operation and provided important insights into the device's molecular mechanisms.

In her postdoc at the University of Cambridge Department of Chemistry, Dr Merlet focused on new lattice models to interpret nuclear magnetic resonance experiments on supercapacitors. This required her to expand the scale of her simulations drastically to be able to calculate physical properties that are measured in experiments. In the process, she developed a new simulation method to examine the electrode/electrolyte interface, which is decisive for supercapacitor efficiency. Her newly written simulation code was 10 000 times faster than previous simulation software.

"In the process, Dr Merlet developed a new simulation method to examine the electrode/ electrolyte interface, which is decisive for supercapacitor efficiency."

Dr Merlet is now a CNRS researcher with her own group, and in her most recent projects she has focused on the development of systematic multi-scale models to predict key properties of supercapacitors built out of different materials, such as energy density and power density. "Through her simulations, Dr Céline Merlet has shown how HPC can help to find new classes of energy materials for a sustainable future," Professor Matej Praprotnik, Chair of the PRACE Scientific Steering Committee, said earlier this year. "Dr Merlet is a remarkable young scientist," he adds. "She develops high-performance computing methods for crucial scientific questions and is on her way to become one of the global leaders in computational materials science."

Dr Merlet regards gender equality as a fundamental personal concern. That is why she is involved in "Femmes & Sciences", an association that promotes women in science and technology through various activities. Among other things, the chemist was instrumental in the development of "Mendeleieva", a board game that encourages pupils and grown-ups to discover 125 female scientists.

To promote the game and gender equality, she toured science fairs and visited high school classes. "There are still too few women in scientific and technological professions and careers, and I would like to help change that by showing the kids that a successful career in science and technology is possible for both men and women," she says.

Asked if she had a message for any women or girls looking to get into the world of HPC, Dr Merlet said: "In my case, I did not have any problem in the field of HPC as a woman, so I did not experience any barriers in this sense. I think it is important to remember that not everyone wants to do HPC or science and that is OK – everyone should be happy to do what they want. But if you find HPC fun, even though not many people around you think the same, don't be put off – as long as you like it, it doesn't matter what other people think!"

Scientific keynote

HPC vs COVID-19: Glycans and the spike protein

As part of the global efforts to fight the COVID-19 pandemic, PRACE last year opened a fast-track call to provide urgent computing access to researchers from a variety of backgrounds. **Elisa Fadda** of Maynooth University, one of the recipients of computing time through this call, spoke at this year's EHPCSW about her work on the role of glycans in the infamous spike protein of the virus.

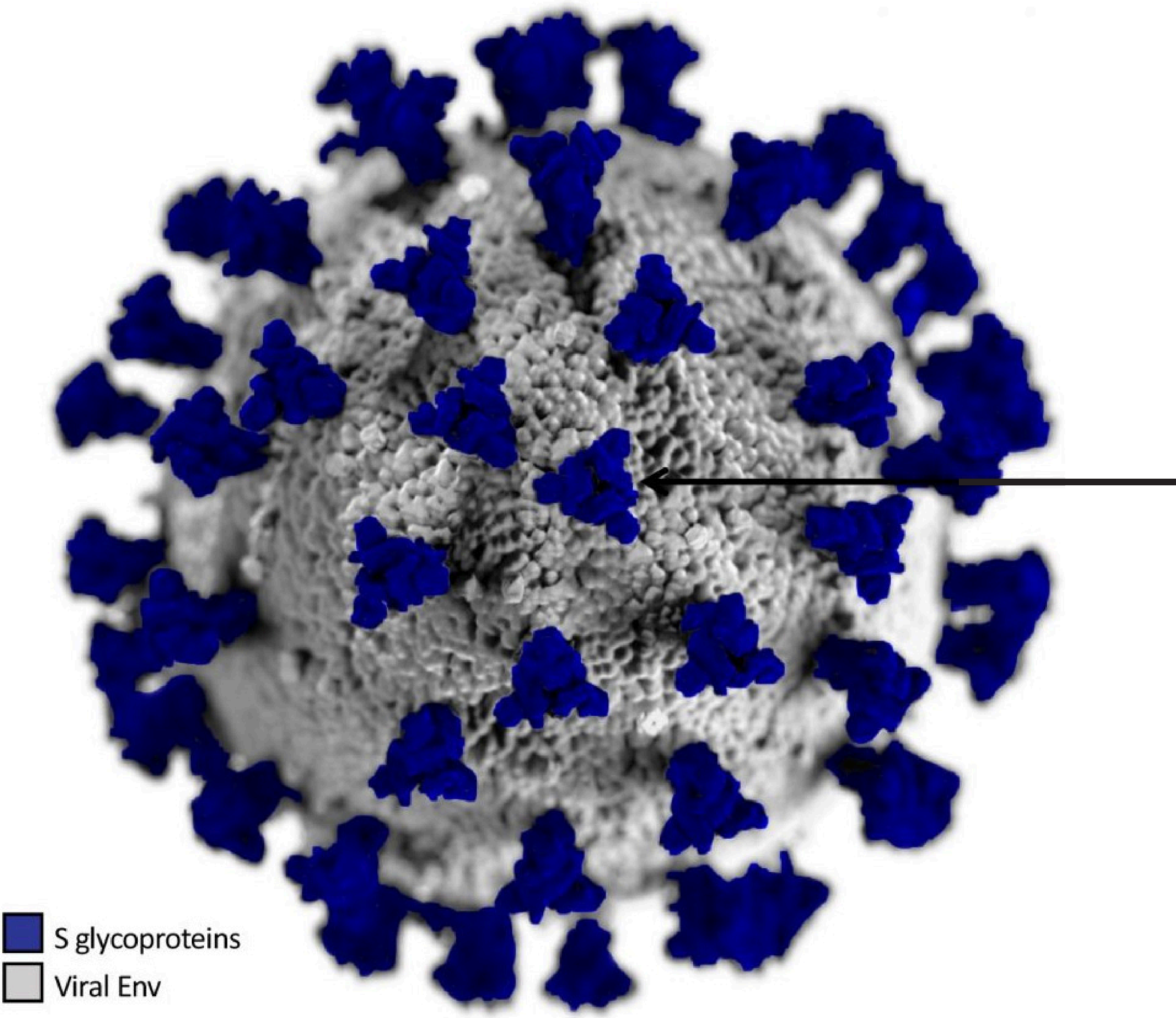
Elisa Fadda

Complex carbohydrates are the most abundant biomolecules in the natural world. Most proteins are covered in a thin layer of these branched structures, which can be made up of any number of hundreds of different units. However, despite their ubiquity, they have historically been almost entirely disregarded when considering the structure of the molecules they are attached to, seen as unnecessary details that do not contribute to the overall function.

Massive progress in the field of complex carbohydrate research over the past 20 years has now changed the perception of these molecules from that of redundant decorations to essential biological enablers. As it turns out, the fine layer that they form around proteins, like moss on a stone, often plays a crucial role in the protein's function – including that of the SARS-CoV-2 spike protein.

Understanding how these carbohydrates behave is a tricky process. They are highly dynamic and highly flexible, meaning that most traditional experimental methods are unable to capture any useful information about their function. Instead, computational biophysicists like Elisa Fadda of Maynooth University painstakingly reconstruct these biomolecules digitally so that they can see how they move and react in their own environment.

The famous spike protein that allows SARS-CoV-2 to attach to our cells has been studied in great detail over the past year by researchers looking for ways to target it in the treatment of COVID-19. Studying the carbohydrates that cover the surface of the protein is essential for understanding the disease

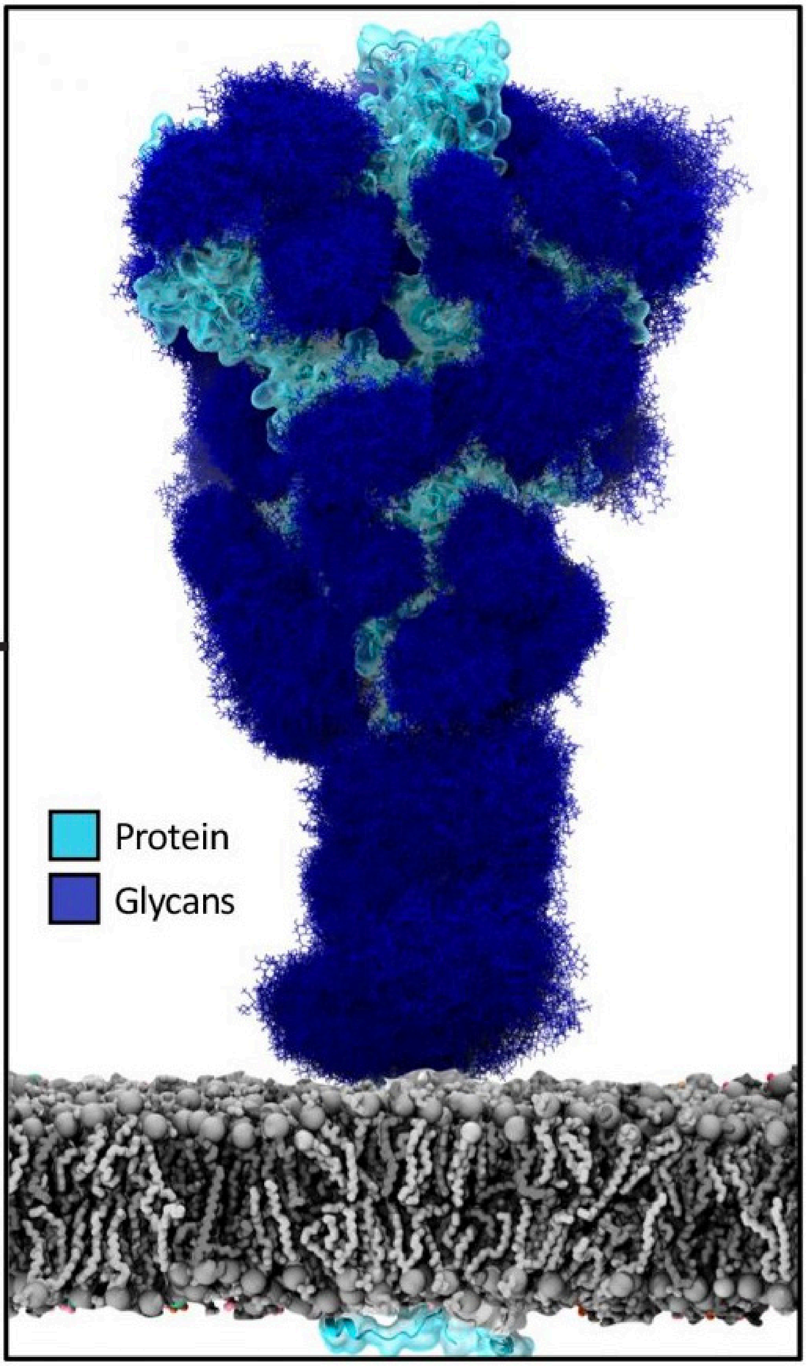


Graphical representation of the SARS-CoV-2 virus surface (grey) with the spike (S) proteins highlighted in blue. The atomistic model of the fully-glycosylated SARS-CoV-2 S embedded in the viral membrane is shown on the right-hand side.

Image source: Courtesy of Dr. L. Casalino, Amaro Lab, UCSD. Reference: Casalino et al., ACS Central Sci (2020), DOI: 10.1021/acscentsci.0c01056.

better, as Fadda explains. "Working alongside my colleague Rommie Amaro in San Diego, we have shown that SARS-CoV-2 is entirely unique from other viruses due to what is known as its glycan shield." This coat of carbohydrates that hides the virus from our immune system is present in all enveloped viruses, including HIV and influenza. "What makes it unique in this coronavirus is that specific glycans within the shield are intrinsically involved in the mechanism of the spike protein that allows it to latch on to our cells. Without these glycans, the spike protein would be useless and the virus would not be contagious or dangerous in any way."

This feature of SARS-CoV-2 could explain another unusual phenomenon that it displays. Viruses replicate by hijacking the cellular machinery of their host. As such, the glycan shield of SARS-CoV-2 is made by whatever cell its predecessor infected. Where this becomes interesting is that the different types of cells in our body – cells in our lungs, cells



"Without these glycans, the spike protein would be useless and the virus would not be contagious or dangerous in any way"

in our intestine – produce different glycan shields. This leads to variation in the infectivity of the virus depending on what type of cell it has come from. Analysing this viral quirk experimentally is an almost impossible task. However, Fadda and her colleagues have been able to construct various types of glycan shields and use high-performance computers to see how they behave. This could pave the way for some new therapeutic approaches. "It's possible that instead of targeting the spike protein, the glycosylation machinery that creates the glycan coats could be hindered in some way so that any viruses created would have defective spikes. "Of course, any therapy targeting this would have to be very selective as 70% of the proteins we create in our body are glycosylated. It will probably be a case of refining the strategy so one or two enzymes are being targeted."

Fadda's studies on the role that glycosylation plays in SARS-CoV-2 have partially been made possible by a PRACE allocation of 15.84 million core hours on the Marconi100 supercomputer hosted by CINECA, Italy. Using molecular dynamics simulations, her team has studied five spike protein models with different patterns of glycosylation, observing how the carbohydrates react and rearrange as the spike binds to a human cell. "These simulations allow us to see the whole system moving and show us the role that the carbohydrates play," says Fadda. "It's like using a computational microscope that provides detail down to the level of atoms in real time."

With data collection now finished, a paper on the work carried out in the PRACE project is due to be published shortly. The project builds upon work carried out earlier in 2020 on the NSF Frontera computing system at the Texas Advanced Computing Center. The unique perspective that Fadda's work provides, along with the extremely high-profile nature of the COVID-19 pandemic, also led to her work featuring in the New York Times last year. "The whole experience of continuing our work on this with PRACE has been fabulous," she says. "We were able to access our allocation straight away and, when we needed help, it was available immediately."

Diversity, equity and inclusion

Lea Lønsted and **Lucye Provera** of consulting firm Mercer spoke at this year's EHPCSW on the impacts of COVID-19 on different demographic groups in the workplace, as well as providing an overview of the factors driving diversity, equity and inclusion (DEI) in the tech industry and the ways the progress can be made in these respects.

DEI and COVID-19

Mercer is an HR consulting firm obsessed with data. When the COVID-19 pandemic hit the world and spread across the globe, it quickly started collecting data on its impacts on the working world, focusing particularly on whether companies were taking notice of variation in how it impacted different demographic groups of their employees. When Scandinavian companies were asked this back in March and April 2020, only 29% were keeping track of this kind of data. A similar survey in the UK showed only 21% of companies tracking these differences.

Since then, Mercer has been investigating this issue more closely and has gained a number of insights. "We can see that there's a definite skew in how COVID-19 has impacted women and people of colour in terms of job security," says Lea Lønsted, head of Diversity, Equity and Inclusion (DEI) consulting for Mercer in Denmark. "Typically, women and people of colour are in more vulnerable positions and more frontline positions, and as such many more people from these groups have lost their jobs as a consequence of COVID-19."

A similar trend was seen in terms of financial safety, not only in the obvious form of being employed, but also because those groups tended to work jobs which were more difficult to transfer to a remote working environment. Finally, large differences were also seen in terms of caregiving. For instance, women working from home have often not only had to do their job but also take responsibility for organising family life, looking after children, and in some cases providing home schooling.

Of course, with any crisis there also comes opportunities, and the COVID-19 pandemic is no exception to this. "If companies are in a situation where they need to cut down their workforce, this is a good opportunity from them to examine whether they are inadvertently cutting down on positions that have more women or people of colour or other demographics," says Lønsted. "Companies should use this time to stress test their DEI strategy when it comes to workforce composition."

Other opportunities have presented themselves in terms of pay equity, where some companies who have had to reduce pay have chosen to reduce men's pay more and consequently cut any pay gaps that may have existed. As well as this, work flexibility is a factor that affects all demographics, and is something that a majority of companies have had to embrace to a certain extent due to lockdown restrictions. This mass movement towards homeplace working represents a unique opportunity to revise and rethink the way the world thinks about flexibility, not just in terms of putting policies in place to allow more flexibility, but also in terms of workplace cultures embracing the ability to do things at different times and in different places.

DEI in the tech industry

Since 2014, Mercer has been carrying out a large-scale worldwide study "Let's get real about equality" on the key drivers behind DEI in organisations. By 2020, it had almost 1 200 organisations participating globally, of which 132 are in tech, covering almost 1 000 000 tech workers in 54 countries across all regions of the world.

One major finding has been a much greater focus sustainability in terms of environmental, social and corporate governance. Regulations and quotas have also found their way on to the agenda more and more, particularly in the EU. Mercer recently participated in a set of working sessions with the EU's new cabinet on gender equality, brainstorming ways to reduce the gender pay gap across the union. "I would say watch this space in terms of seeing more regulations, but already we have seen increasingly strong regulations related to gender in terms of pay and workforce composition in countries like France, Germany, Sweden, Iceland, the UK and Norway," says Lønsted.

An increase in voluntary disclosure has also been seen. The summer of 2020 in the USA was marked by several tragic incidents, such as the death of George Floyd, but a positive outcome of this has been a shift towards companies no longer thinking it is OK to just say "we're against racism". Consumers and employees now expect companies to have anti-racist and anti-sexist policies and cultures in place, and they expect them to articulate these and disclose them to a much greater degree than has ever been seen before. Although Europe has not followed this trend to such an extent, it will likely follow suit in the coming years.

Finally, there is now a marked difference in terms of expectations between employees from the Millennial and Gen Z generations and employees from the Gen X and baby boomer generations. The older generations typically expect to work at a company for at least seven years and therefore put greater value on pensions and other such benefits. In contrast, younger generations expect much more fleeting employment of around two years, and this puts pressure on companies to articulate their purpose and values so that their employees feel comfortable with what they represent.

Making DEI work

The three components of DEI – diversity, equity and inclusion – go hand in hand. Of these three, diversity is the easiest to measure. Gender, age, race, ethnicity, disability, LGBTQ+, socioeconomic status and more are all tangible and easy to place metrics upon. But having a diverse culture or diverse representation is not the same as being inclusive.

"Already we have seen increasingly strong regulations related to gender in terms of pay and workforce composition in countries like France, Germany, Sweden, Iceland, the UK and Norway"

Inclusion is about fostering belonging. How employees feel when they go to work, whether they agree with what the company stands for, and whether they feel they can bring their whole selves to work are just some of the ways that inclusion can be conceptualised. Finally, equity involves having policies and processes in place that ensure that there is equal access to opportunities, experiences and pay.

"Through our work advising companies on developing their DEI strategies, our advice is to start by solving issues around inclusion first," says Lønsted. "Start out by looking at your workplace culture, and ensuring you have policies and processes to ensure opportunities, experiences and pay are fairly distributed among employees, and the rest will follow."

Lucye Provera, a member of Mercer's Global DEI consulting team, adds: "There is a perception that the tech industry is a male industry, and this leads to challenges in recruiting in women and other minorities. What is required is a shift in culture and a focus on inclusion in order to see real change."

Update

After the DEI in the tech industry presentation was recorded, the EU Commissioner of equality presented new legislation related to this topic to the European Parliament. You can find out more about this by following this link:

https://www.mercer.com/our-thinking/law-and-policy-group/eu-employers-face-pay-transparency-disclosure-mandates.html

Women in tech

Data from the "Let's get real about equality" study shows that improvements are happening in the tech industry, but at a very slow pace. The overall gender split reported was 66-34% male-female, which became even more unbalanced at the executive level where an 81-19% split was seen. Examining promotion rates, hire rates, and exit rates is also important if the aim is to have an overall higher representation of women, both in general and in leadership positions. Predictive progression analyses have shown that even if the male and female flows in this respect were the same, in 10 years the gender split would only have moved to 64-

36%. "Essentially, if we want to move the needle, we need to be a little bit more aggressive when it comes to adjusting these parameters."

Every organisation in the tech industry therefore needs to start examining their talent flows. Recruitment is one aspect of this, but retention and internal promotions are also important to consider. By understanding where the choke points are, organisations can focus on the highest areas of opportunity for driving change.

Mercer proposes a four-step approach. First up comes diagnosis, looking at the data to find out what the problems are. Then comes engagement, not just at board level and leadership level, but also at employee level. Following this comes the step in which actions are taken, where programmes can be created to solve the problem. Finally comes accountability. This means ensuring that these changes happen, setting goals and sharing the level of success in meeting these goals in a transparent way.

Flexible working

The concept of flexible working has been thrust into the limelight as large portions of the global workforce have had to work from home over the past year. But flexible work is not just about working from home. It is also about how people work and at what time they work. Even those people whose jobs require them to work shifts in a specific location can be offered flexibility in terms of the composition of their shifts.

In Mercer's survey, 77% of tech organisations said they offered a variety of flexible work options, but only 58% of organisations valued remote working as much as working in person, and only 48% of leaders promoted the uptake of flexible work. It has to be mentioned that this survey was done before the pandemic and so the figures have likely changed, but issues still exist in this area. "We still hear from women that they fear taking the opportunity to work flexibly as men who don't will progress further," says Lønsted. "We also see to some extent with older employees that there's a fear that working a little bit less will lose them some of their entitlement or respect, so there's still much to be done around the culture and how we articulate flexibility."

Bridging skill gaps

Over the past year, Lucye Provera has been working with an association that represents telecommunications companies in Europe, focusing on how these organsations are aiming to support the digital upscaling of women and the aging workforce. The world is now at a pivotal moment in time, because by 2025 the amount of work being done by humans and machines will be equal. People are being pushed out of jobs they had done previously, and the skillsets required for future workers are changing rapidly. "Critical thinking, analytics, problem solving, and self-management skills such as active learning and agility are the areas that companies definitely need to focus on, in addition to the gender perspective where we still see big gaps between men and women and older and younger people."

"Flexible work is not just about working from home. It is also about how people work and at what time they work."

Companies are quickly starting to change the way things are done, for instance setting up academies to focus on the pipeline of talent even before it reaches the organisation, especially in the tech industry. Reverse mentorship programmes are ensuring knowledge is shared and passed on amongst different generations, or in some cases even up into the organisation.


Lucye Provera


Lea Lønsted

Moving forwards in the tech industry

So, how can progress be made? Women thrive and DEI programmes work when leaders at all levels are engaged. People in organisations also need to be personally engaged in the goals of DEI programmes and policies, rather than going along with them because they are being told to. It also has to be remembered that there are no quick fixes. "People have to realise these things will take time, but that they are worth it for bringing about long-term sustainable change."

From an organisational perspective, looking at the data is key. Examining what is helping and what is hurting, and moving things away from an opinion-based perspective to a data-based perspective is what will help drive change most effectively. Processes that will help promote and support women and other underrepresented groups are also needed. Finally, programmes that support women's unique needs around financial wellness and financial health are especially important in areas where national welfare systems are not as prevalent.

HPC in Portugal – A national overview

Before the COVID-19 pandemic swept the globe, EHPCSW 2020 was due to be hosted in the city of Porto. Postponed for a year, it was sadly not possible to have the conference there this year either, but at this year's online event, participants were led through an overview of Portugal's ongoing HPC strategy for the coming years by three of its main stakeholders.

Paulo Quaresma, a director of the Portuguese Science and Technology Foundation FCT, gave an overview of the Portuguese policy on advanced computing. In 2018, it approved a national strategy that will support and promote initiatives at a European level that can have an impact on society and its citizens. A key part of its strategy is to use renewable sources of energy and achieve a sustainable infrastructure. It is a difficult and multidisciplinary challenge that will involve the development and deployment of new and innovative solutions.

Being part of EuroHPC is a key part of Portugal's strategy. As well as this, its interaction with all of its national supercomputing infrastructures has allowed FCT to open a call for scientific projects where researchers can apply for CPU time. The Iberian Network of Advanced Computing has a role to play too, as do green computing solutions. There is an ongoing expansion of petascale supercomputers, including those focused on AI solutions and deep-learning algorithms.

Portugal was one of the seven countries that in 2017 signed the declaration that brought EuroHPC into being, which will ensure the implementation of a world-class European infrastructure for advanced computing. The two new Portuguese petascale computers have a peak performance of 10 petaflops, which will help with these goals.

In a more general view, Portugal's advanced computing policy will allow it to support the improvement of its digital processing capacity, and Quaresma was keen to emphasise the care and effort that has gone into the development of high-value human resources that are a key point for advanced computing. One priority domain area for Portugal is health, covering topics such as personalised medicine, image processing and omics. Green energy is another key domain, as well as the Internet of Things, climate change, the environment, Industry 4.0, finance, and natural language processing.

Joao Nuno Ferreira, general coordinator of the National Scientific Computing Unit of FCT, then talked about the national advanced computing network, its standing today and the plans for the near future. FCT has been engaged with PRACE for many years, and a lot of its HPC activities have been carried out through this collaboration. Its national node is located at the University of Coimbra at the Laboratory for Advanced Computing, which is now a well-established part of the wider European network. Portugal has also been involved in distributed computing for a long time, focusing mainly on the high energy physics community. It also has a longstanding collaboration with Spain in this regard.

The 2017 declaration in Rome that brought about EuroHPC was important for Portugal, triggering fast change in HPC. Soon after the declaration, FCT established two partnerships, one with the University of Minho and the other with the Texas Advanced Computing Centre in the USA. This second partnership is one of the pillars of the larger partnership of the University of Texas and Portugal that also covers a number of other areas.

Another important agreement that has supported Portugal's HPC

The MACC machine in the Riba D'Ave data center in Minho, based on approximately 40 STAMPEDE 1 racks, assigned to the Foundation for Science and Technology under a partnership of Portugal with the University of Texas at Austin, including the collaboration of the Texas Advanced Supercomputing Center and the University of Minho.

efforts was the Portugal-Spain Summit 2018 in Valladolid, which secured a memorandum of understanding between the two countries to collaborate closely in the calls for hosting the new supercomputers funded by EuroHPC. This was successful, with a bid being won to host a petascale computer, as well as a bid to host a precursor to an exascale computer. Collaborative work has been ongoing with the other hosting countries ever since.

The national network has been developing rapidly, based in four operational centres and ten national competence centres. One of these operational centres will be supporting the forthcoming Square Kilometre Array radioastronomy project. The Minho Advanced Computing Centre is also set to play a big role in Portugal's HPC strategy. The competence centres have had large screens installed to help researchers with the visualisation of their computing projects, will provide support for local users, and specialise in a number of specific domains.

In terms of the future, Ferreira emphasised the need to set up the new national petascale and pre-exascale resources as soon as possible. EuroHPC project activities will also be prioritised, as will participation in the Centres of Excellence for HPC applications, which Portugal has so far had little involvement with. Finally, knowledge and skills in quantum computing will be improved.

The final talk on Portugal's HPC strategy came from Rui Oliveira, Director of the Minho Advanced Computing Centre (MACC), which has three main areas of operation. It aims to support and foster research and innovation in the codesign of HPC systems, from their implementation to optimisation and exploitation, a major priority for EuroHPC. It will also foster the creation, expansion and consolidation of research teams on national and European high-priority scientific domains that require large-scale digital simulations and data-intensive systems and applications. Finally, it will offer advanced computing resources and consultancy to science and higher education institutions, public administrations, and companies.

An important capability for Portugal is the ability to use advanced computing facilities remotely. To do so, it is no longer enough to have remote access to a main console where users can submit jobs. There is a need for visualisation facilities that allow for complex visualisations, which MACC is helping to facilitate.

Portugal is now an important part of Europe's HPC ecosystem. It will become a prominent node of the European pool of exascale HPC capacity with the implementation of its new EuroHPC supercomputers, and its strong relationship with Spain will help with this. Oliveira explained the importance of being central to bringing new capacities to the south Atlantic countries, as well as collaborating internationally to achieve global sustainability goals.

The EuroHPC supercomputer Deucalion, a state-of-the-art petascale system, will be installed and begin operations in 2022. Its use will be broader than just supporting science and traditional HPC domains, with 10% of the system using GPUs mainly for data science applications. It will also incorporate experimental technologies such as FPGAs, and it will never usea single watt of energy derived from fossil fuels, instead relying completely on renewable energy.

Industrial keynote

Big Science projects and their advances for the industry

Big Science projects help us to advance the very pinnacle of human knowledge using state-of-the-art technology and computing power. However, they also have the potential to have much further reaching impact in unexpected areas of our lives. Speaking at EHPCSW, Mauro Gameiro of Critical Software gave an overview of the company's involvement in some of the most important science projects in the world, and how the advancements being made could provide huge and immediate benefits for many areas of society and industry.

Big Science projects are often unfairly regarded as huge drains of taxpayer money that don't give back anything tangible to society. They are often presented in mainstream media as providing big leaps in scientific knowledge, but without having any immediate applications for improving the lives of ordinary people.

Industry, however, sees these big science projects in a different way. The scientific advances made in them can be used by industry to create new solutions and new products that will improve people's lives. With the advent of big data and artificial intelligence, HPC capabilities are now being used more regularly in day-to-day applications and across many different markets. The advances in technology being made in Big Science projects are therefore ushered in with a lot of enthusiasm by industry.

Critical Software

Critical Software was founded at the University of Coimbra – one of the oldest universities in the world – by three PhD students who were working on fault injection techniques, a method used for understanding how systems behave when stressed in unusual ways. In 1998, the NASA Jet Propulsion Laboratory was looking for software capabilities for fault injection, detection and recovery, and thus became Critical Software's first customers.

The company's mission statement is to develop innovative and reliable technologies for critical systems. Starting in the space sector with NASA, it soon expanded to other markets such as aerospace, transportation, energy and finance. Today, it is organised into three divisions: high-integrity systems, smart technology solutions and digital engineering systems.

For a small company starting in Portugal, a country with little technology and engineering tradition at the time, Critical Software had a big credibility obstacle to overcome. Their strategy from the beginning therefore has been to set the bar very high in terms of quality. It is now part of a small group of companies worldwide that have managed to combine CMMI level 5 with agile methodologies.

"The SKA telescope will generate the same amount of data every day as the entire planet does in a year"

Across all of the industries it operates in, Critical Software applies the same standards of performance and reliability, investing heavily in its quality processes and innovation. Because of the small size of Portugal's domestic market, the company has been outward looking from day one, and now has further offices in the UK, Germany and the USA. "What could have been a disadvantage for us – starting in a small country – actually worked in our favour," says Mauro Gameiro, the company's principal engineer for space and big science projects. "We also have a strong reinvestment policy in terms of our people, business, infrastructure and growth. That's how we have managed to scale to where we are today, while maintaining the company as private and independent."

Software for satellites

Critical Software has worked on a number of projects in the space domain, covering systems engineering, embedded software development, safety critical verification and validation, software security and product assurance, and simulation platforms. Working alongside space agencies like NASA and the European Space Agency (ESA), as well as companies like Airbus and ThalesAlenia Space, it has over the last 15 to 20 years worked on a wide range of satellite missions.

Earlier this year, the Solar Orbiter satellite was launched with the mission of studying solar and heliospheric physics. Critical Software was responsible for the development of the equipment management components of the flight software on the central onboard computer of the spacecraft, and provided on-site support to Airbus, the prime contractor of the Solar Orbiter mission, for software specification, design implementation, testing and validation. As well as this, it provided software system engineering expertise, namely in terms of the requirements baseline specification, functional avionics support, technical support and other activities.

Another satellite project the company is involved in is the ExoMars Trace Gas Orbiter run by the European Space Agency. Launched in 2016, it is currently orbiting Mars. As well as conducting scientific investigations into the biological or geological origin of trace gases on the planet, it is also acting as a relay for information being sent from rovers that are currently exploring the planet's surface.

A slightly different field of work in space involves the detection and classification of sea vessels using data from the TerraSAR-X imaging radar Earth observation satellite. Critical Software developed algorithms to process radar images and detect and classify vessels according to a predefined nomenclature. This information is then integrated with data acquired by vessel tracking systems.

Forest fires are a huge problem in many countries worldwide, including Portugal. Critical Software has implemented several projects aimed at developing the service chains required to provide operational forest fire risk maps to civil protection agencies and to the pulp and paper industry. The risk maps are supported by land cover data, daily weather forecasts, daily vegetation status data, and burnt areas maps.

Finally, another project from ESA called Desertwatch, which aims to monitor the progression of desertification, is being supported by the company. Critical Software has been been involved in the design and implementation of service chains capable of producing land cover data and certification related indicators to support the monitoring activities.

The Extremely Large Telescope

Critical Software is part of some of the most ambitious science projects in the world, one of which has involved working with the European Southern Observatory (ESO) on the Extremely Large Telescope (ELT) project. The ELT is a revolutionary scientific project involving the construction of a 40m-diameter telescope in the Atacama Desert of northern Chile. It will address many of the most pressing unsolved questions about our universe. The ELT will be the largest optical near-infrared telescope in the world and will gather 13 times more light than the largest optical telescopes that exist today. It will be able to correct for atmospheric distortions, providing images 16 times sharper than those from the Hubble space telescope.

The ELT will vastly advance astrophysical knowledge by enabling the detailed study of planets around other stars, the first galaxies of the universe, supermassive black holes, and the universe's dark sector. Since 2015, Critical Software has been involved in the project as the sole provider of the verification and validation activities of all components of the telescope. Besides this, the company has been involved in the development of the local control system of the M1 mirror, the ELT's primary mirror and the centrepiece of the revolutionary astronomy machine.

The Square Kilometre Array

The Square Kilometre Array (SKA) will combine signals received from thousands of small antennas spread over a distance of several thousand kilometres to simulate a single giant radio telescope capable of extremely high sensitivity and angular resolution. When completed, the SKA will be made up of over half a million antennas, the combined areas of which will span over a square kilometre.

The project will provide continuous frequency coverage from 50 megahertz to 14 gigahertz in the first two phases of its construction, followed by a third phase that will extend the frequency range up to 30 gigahertz. The capabilities of the SKA will be designed to address a wide range of questions in terms of astrophysics, fundamental physics, cosmology and particle astrophysics, as well as extending the range of the observable universe.

The scale of the figures describing the computing challenges of the project are almost hard to believe. The computing power required for each of the two SKA supercomputers will match that of the best supercomputers in the world. In terms of data, the telescope will generate the same amount of data in one day as the entire planet does in one year.

Mauro Gameiro

Critical Software has been involved in three work packages of the SKA project. In the "Telescope manager" and "Signal and data transport" work packages, they performed independent analyses of the element system requirements and the reliability, availability and maintainability plans. However, it is in the "Science data processor" work package where their work has truly shone.

Here, they developed what is known as the Slow Transients Pipeline (STP) prototype. It was implemented using C++, a high-level language that also provides low-level fine control of the operations, using the Linux operating system. In this project the main bottlenecks and limitations for the algorithms were identified and analysed. They then investigated computationally efficient solutions for the most complex algorithms present in the pipeline to reduce the number of computations, memory readings and writings, and improve the parallel scalability. For the parallel programming the used Intel Threading Building Blocks, a C++ template library that supports several parallel patterns. Starting with algorithms that took 160 seconds to execute, they were able to speed up the process by a factor of 60.

The knowledge acquired and the technology developed in the scope of this project in terms of improvements in algorithm performance has a wealth of potential for reuse in other markets and domains. Energy consumption profiles for price optimisation in production planning, and image processing for quality control in highly sensitive industries are just two examples where process optimisation could be done by reusing the knowledge and work done in this project. Vegetation monitoring and water level assessment could also benefit, and the financial sector will be able to improve fraud detection in financial transactions. The HPC techniques developed could enable a speed up of the processing of DNA chains, and monitoring in railway stations for real-time detection of dangerous actions could be improved.

"This has been one of the most interesting projects we have ever worked on in terms of what we have developed and the potential it will all have for other sectors," says Gameiro. "All of the engineers that worked on it cannot forget it because it involved the use of high-level techniques to improve the algorithm performance, the results of which were outstanding and have left everyone with a great sense of pride."

HPC in industry
Computational materials modelling for sustainable technologies

Chemicals and materials company Johnson Matthey uses materials modelling to enable state-of-the-art sustainable technologies. Speaking at EHPCSW 2021, **Glenn Jones** outlined how HPC lies at the heart of their current R&D endeavours, while also describing their planned forays into cloud and quantum computing.

Johnson Matthey is a global leader in sustainable technologies, and their science impacts areas such as low emission transport, pharmaceuticals, chemical processing and making the most efficient use of the planet's natural resources. JM uses their expertise to develop products for lower automotive emissions, batteries, fuel cells, blue and green hydrogen generation, clean energy, efficient chemicals manufacture, and complex active molecules for pharmaceuticals.

Supporting all of these areas are the Johnson Matthey Technology Centres. Glenn Jones is the head of the physical and chemical modelling team, which brings together physicists, chemists, material and chemical engineers to look at materials from the atomic-scale all the way up to their macroscopic applications. They use computational chemistry to screen for new materials, scanning phase spaces to find new catalysts and materials, then whittling down those materials to a few target candidates for experimental colleagues to work on. High-performance computing has a big role to play in much of the research carried out at Johnson Matthey. Quantum mechanics can be involved, looking at nanometre and femto- to picosecond scales with molecular dynamics, while continuum methods will use computational fluid dynamics. Engineering design then links different components of the process together, moving towards process modelling. Each of these aspects require different types of hardware.

High memory may be needed for some of the continuum methods to be able to store some of the 3D structures. Batch high throughput, such as parameter sweeps of an engineering design, can also be used. GPUs are particularly good at doing discrete elements or classical molecular dynamics. But the biggest use, particularly in Jones' group which carries out a lot of density functional theory calculations, is with traditional low latency network HPC.

Johnson Matthey faced many barriers in terms of bringing computational chemistry into its work. In the 1980s, overselling of codes led to some scepticism about its value. In the early 1990s, density functional theory didn't really work for chemists – they would try to simulate a hydrogen molecule and it would fall apart. By the mid 1990s these issues had been resolved, but it was only by the early 2000s where, using the same algorithms but with more computer power, chemists could find activation barriers and map out reaction pathways for multiple materials. This allowed them to pick out trends and start to be able to predict characteristics. However, despite the now proven usefulness of computational methods and HPC for materials modelling, companies investing in their own in-house infrastructure are often hit by the same problems. "Teams in companies start with a few servers sat in the back of an office somewhere, which then turns into a cluster or a group of clusters," says Jones. "Different groups in the organisation then might pop up with their own resources, and things continue to grow in this very inhomogeneous way. So, we asked ourselves: is there a better way of doing things?"

Cloud HPC: Enabling science across the business

To meet the wide-ranging computational requirements of its various research endeavours, Johnson Matthey has in recent years been investigating the use of cloud HPC, as Jones explains. "There are a number of reasons we began to look into the capabilities of cloud computing," he says. "First and foremost, we want to improve the complexity of our investigations while maintaining or improving accuracy. If you buy a cluster, you're always going to be limited by the capacity of that cluster. Cloud in principle allows us a certain amount of flexibility to burst to a number of processors that we couldn't normally use. This helps us reduce queue times, opens up the possibility of high throughput calculations and helps us to tackle things that we wouldn't normally do. "Of course, another key part of the appeal of cloud HPC is the democratisation of computing power. If you have one cluster that sits with one research group, then only that research group will tend to use it. We wanted to take steps to open up our HPC capabilities for everybody within the organisation in the best way possible using modern technology."

Back in 2017, the company approached a number of cloud providers to see if they could provide what was needed. Initial scepticism at the ability of the cloud to carry out task such as density functional theory calculations proved well-founded, as every vendor in turn failed to pass their benchmarking tests. It wasn't until 2019 that Microsoft Azure was finally able to give performance that the company was happy with. A couple of fairly simple chemical systems were used to test the capabilities of the cloud. Looking at a palladium oxide cluster, good scaling was achieved when compared with Johnson-Matthey's own on-site machine. This result gave them the encouragement to move forwards from this proof of concept to a full pilot. "Within our cloud pilot, we were essentially looking at two main points" explains Jones. "First, the scientific requirements – how do things work, and will it work for us? And secondly, what is the user experience like? As scientists, we spend a lot of time working on these machines, and we wanted to make sure that we would be comfortable working with this new system."

The pilot was a success in terms of scientific requirements, with the cloud proving faster in terms of time to solution than their on-site machine. Furthermore, it was able to do stretch calculations and enable the company to do calculations that were previously impossible. "We looked at doing some simple molecular calculations and also some classical calculations on a GPU, and here we saw a ten-fold speed increase from our current state to running on the cloud GPU, which gives a clear advantage for the business."

One use case in the pilot involved understanding mechanisms. "We often have lots of fragments of molecules on materials that we need to calculate, and this leads to a large batch of calculations that we need to do," says Jones. "To explore this reaction network might normally take us several weeks or months, but on the cloud we were able to do this in a matter of days, giving us a huge boost in the throughput of our calculations. The reasons for that increase were firstly that the queueing times on our cluster are quite bad because we over-utilise them. As well as this, the burst capability of the cloud means that there are more resources available when the calculations do come through."

Another use case looked at monolith channels found in car exhausts for filtering small particles of carbon from diesel engines. Jones and his team wanted to understand how these filters work, the dynamics of the carbon with them, and how they might be better designed in the future. Carrying out the use case task in-house on a legacy HPC machine took almost 25 hours to run, but the cloud was again able to drastically cut this running time. "For us, this changed these kinds of simulations from a curiosity that might overall take us between three to six months to run as a proof of concept, to actually being able to deliver them in a timescale that is useful for the business," says Jones.

Johnson Matthey is now at the rollout phase of its cloud HPC project, and having just finished their user acceptance testing, are about to go live with it. Working alongside their IT department to turn it into a service and make it compliant with the whole enterprise, it is set to start being used this year.

Quantum computing: A new paradigm with potential to disrupt

Looking further into the future, the company is now taking its first steps into investigating the possibilities of quantum computing. Quantum computing has a number of different applications, including cryptography, optimisation and quantum machine learning, but the one Johnson Matthey have been looking at is quantum chemistry. Metal oxides are at the heart of many Johnson Matthey technologies. They are complex materials that come in lots of different crystal structures and can be used for battery and fuel cell applications, but current techniques are not capable of characterising them properly. For example, palladium oxide, a technologically useful material for the company, is predicted to be a metal by standard density functional theory calculations, but in reality it has a band gap. This initial error in the description of the electronic structure means that any subsequent reaction profiles and predictions based upon it will be wrong.

Likewise, looking at battery materials such as lithium nickel oxide with standard codes gives slightly erroneous answers when compared to experimental data. So how can materials science be done more reliably by bringing quantum computing into the equation? When attempting to improve upon the accuracy of standard density functional theory calculations to move towards full configuration interaction, computational costs scale exponentially with the number of electrons in the system. What quantum computing promises instead is polynomial scaling, so even though calculations will still get more expensive as more and more electrons are involved, there will be a point at which a quantum computer can do more than a classical machine.

Johnson Matthey is now in the midst of two projects exploring the possibilities of quantum computers. The first of these is looking at battery materials, working with the University College London (UCL) chemistry department and a spinout from UCL called Phasecraft. This is an eighteen-month feasibility study in which Phasecraft is developing quantum algorithms, while Johnson Matthey is working alongside UCL to work out which systems are important, when it will be possible to calculate them on a quantum computer, and what resources will be needed to make this kind of quantum computing useful.

The other project is with a spinout from UCL called Rahko, as well as Kings College London and the National Physics Laboratory. "This project is very similar in many ways to the other project, except that rather than looking at battery materials, we're looking at heterogeneous catalyst materials," says Jones. "Rahko are building a platform for quantum machine learning, and in this project we actually have access to a quantum computer at the National Physics Laboratory to test things out. We're currently in the process of finishing a paper looking at hybrid classical-quantum algorithms and their applications to industrially relevant processes, with a particular focus on metal oxides, so hopefully that will be out in the near future."

As Jones' talk demonstrated, computational modelling is integral to companies like Johnson Matthey, and many of the methods they use are only possible because of HPC as well as the algorithms that have been developed by the community. Cloud HPC will in the future be providing the infrastructure and the flexibility that is needed within the organisation without having to build a datacentre or administer a supercomputer, while quantum computing is helping to push their research into previously untrodden territories.

Q&A

Anders Dam Jensen, Executive Director of the EuroHPC JU, answers our questions

Can you please start by introducing what it means to you, as Executive Director of EuroHPC JU, for the EHPCSW 2021 to still be taking place during these difficult times?

The JU became an autonomous EU body in the midst of the pandemic and, not surprisingly, I have yet to physically meet many of my staff. We have yet to all come to the EuroHPC JU offices or meet in the same room. All processes and procedures for the JU have therefore been established within the constraints of the pandemic and using technology at every available opportunity. This is the context in which we worked with our partners to organise the first fully digital EuroHPC Summit Week (EHPCSW) in 2021. Creating an online platform, coordinating live and recorded talks and liaising with many speakers are just some of the unprecedented challenges we had. We have to thank PRACE, ETP4HPC and all colleagues involved in making this first experiment a great success. With 1337 registrations and 165 speakers and sessions, it gave us an overview of the European HPC ecosystem and insight into the latest achievements and future challenges.

This success also shows how resilient my team and the HPC community in Europe are and how, with a bit of imagination, hard work, decent connectivity and goodwill, we were all able to come together and make this conference a success. As Executive Director, I naturally can't wait for the pandemic to be over, and for us all to be able to physically meet each other and have the chance to get to know many of the brightest minds that are enriching HPC in Europe and around the world. I look forward to seeing everyone in Paris for EHPCSW 2022, which I hope will finally allow us all to meet again.

As has been highlighted in the COVID-19 call and through the many successful projects being presented at EHPCSW, HPC has never been more important. What do you see as the biggest achievement of HPC over the last year?

One of HPC's great achievements over the past year has without a doubt been its contribution to the race against time to contain the spread of COVID-19, the management of the pandemic and the search for a cure. It has been impressive to see how the HPC community has come together to undertake the ground-breaking research to isolate the COVID-19 virus and identify drugs that can treat the symptoms of COVID-19 as well as simulate the spread of the virus to better protect public health.

A good example of how HPC has helped to combat the pandemic is EXSCALATE4CoV, a project supported by the European Commission's Horizon 2020 that harnesses the potential of supercomputing to better address pandemic situations of supranational interest. The project aims to identify the safest and most promising drugs for the immediate treatment of the already infected population, followed by the identification of molecules that can inhibit the pathogenesis of the coronavirus to counter future conditions. EXSCALATE4CoV is based on the Exscalate platform (EXaSCale smArt pLatform Against paThogEns), already used in the study of the Zika virus, and is only one of 17 projects founded by the European Commission in this framework.

As none of the EuroHPC JU machines were operational at the time, EuroHPC cannot claim any credit other than to have witnessed this work closely and with interest. But as soon as EuroHPC systems are up and running and in production, they will certainly contribute to the development of vaccines and drugs in infectious and chronic diseases and any other application that can change and sometimes save people's lives.

How important is the annual EHPCSW 2021 for the high-performance computing ecosystem and why?

The EHPCSW is an important annual event to showcase the latest achievements and challenges in the European and international HPC world, not only in Europe but also globally. Thanks to the shared efforts of PRACE, ETP4HPC, the EuroHPC JU and the host country (Portugal this year) the event brings together key European HPC stakeholders, from HPC technology and infrastructure providers to scientific and industrial HPC users in Europe. It brings together experts from academia and industry presenting their advances in HPC-supported science and engineering. It is also an important opportunity to network, share ideas and create new alliances for future HPC projects, as we have seen again this year despite the event being virtual.

EHPCSW 2021 also proved more than ever to be an important high-level political moment for the whole European Union. We were honoured to have with us the Portuguese Minister of Science, Technology and Higher Education, Manuel Heitor, representing both the host country and the Presidency of the Council of the European Union and as the first speaker of the EuroHPC Day. EuroHPC is strategic for Europe's competitiveness and economy at global level. Since the last State of the Union, building on Europe's success in next-generation high-performance computing has been identified as a strategic investment

Anders Dam Jensen

priority, and will underpin the entire digital strategy, from big data analytics and artificial intelligence to cloud technologies and cybersecurity. The EHPCSW is an important event in this context, allowing European HPC stakeholders to meet and discuss how they are implementing this strategic priority together.

Could you tell us about what EuroHPC JU has achieved over the last year and its plans for the future?

Several speakers during EHPCSW 2021 stressed their surprise at how quickly EuroHPC was built. The last four years have been amazing in terms of the speed with which the commission and the member states have been able to create a unique and innovative joint venture and build one of the largest computer grids in the world in less than two years. This is an impressive achievement, even without the added complications caused by COVID-19. So, I can certainly say that the autonomy of EuroHPC JU is one of our greatest achievements of the last year.

Since then, and after the hundreds of millions of euros already invested, we are starting to see some progress. Much has already been achieved with the launch of all 20 research and innovation actions. Much more will come with most of our seven HPC infrastructures becoming operational and in production by the end of 2021, of which Vega will be the first. Once completed, they will add more than 670 petaflops of computing power to the European grid. This is an incredible achievement in such a short time – and it is just the beginning. We are jointly achieving our goal of acquiring top-of-the-range supercomputers for Europe that will redefine the Top500, consolidate a world-class European HPC ecosystem, and pave the way for strategic EU autonomy in high-performance computing. We have managed to get the most out of our current regulation and we are very much looking forward to the new one, so we can continue working for the next seven years and more towards important new achievements for all our citizens.

Looking ahead, the new MFF programmes for 2021-2027, with Horizon Europe, the Digital Europe Programme, the Connecting Europe Facility funds, and the new regulation that is currently being discussed at EU level, will be a further step in our efforts to create a world-leading European HPC ecosystem. The proposal will allocate for the period 2021-2033 approximately €3.5 billion of EU funding, which will be matched by funds from EU Member States and our private members for a proposed total budget of €8 billion. With these funds, we will not only develop a world-class exascale and post-exascale HPC infrastructure, but also a quantum computing infrastructure which will be accessible to public and private users across Europe. We will also continue to support the development of technologies and applications to support the supercomputing ecosystem, and exploit the synergies of HPC with AI, big data and cloud technologies. We will improve awareness, knowledge and training in HPC. All in all, we will ensure the development of top-of-the-line HPC infrastructures and technologies in Europe for the next decade, to maintain Europe's position in the global race towards exascale, post-exascale and quantum computing capabilities.

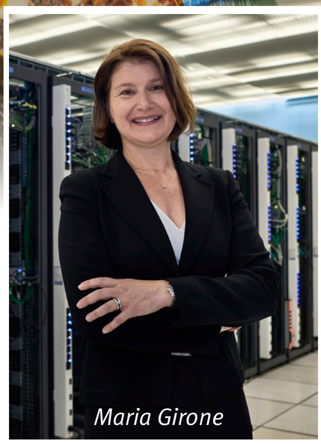
How important is the relationship between PRACE and the wider European HPC ecosystem?

PRACE's experience and work to date in the broader HPC ecosystem is extremely valuable. PRACE has, up until now, managed the access to existing computing and data management resources and services for large-scale scientific and engineering applications at the highest level of performance, and providing a broad education and training effort for effective use of research infrastructures have been important first steps in creating and coordinating the embryo of a truly European HPC ecosystem.

With the new strategic priorities of the European Union and the advent of a new financial framework, EuroHPC will need partners in Europe to ensure that new HPC resources are well-used in order to step up efforts and play at the highest level of global HPC. The EU has tasked the JU to equip it with a world-class federated, secure, and hyper-connected supercomputing and quantum computing service and data infrastructure. The JU will, once the regulation is adopted, select partners who will assist us to develop the necessary technologies, applications and skills for reaching exascale capabilities and a quantum computing innovation ecosystem. In the next few years, guaranteeing Europe's leading role in the data economy, scientific excellence, and industrial competitiveness will increasingly depend on our capability to develop key HPC technologies and excellence in HPC applications. To make this happen, a pan-European strategic approach and leadership by EuroHPC JU will be essential.

Is there anything you would like to add?

I would like to thank PRACE for the huge work they did in organising EHPCSW 2021.



Maria Girone

Scientific keynote

Computing challenges at the Large Hadron Collider

As the Large Hadron Collider gears up to enter an even more powerful form from the end of 2027, high-energy physicists are eagerly anticipating glimpses of a world beyond the Standard Model. Speaking at EHPCSW, **Maria Girone,** CTO of CERN openlab, explained the many challenges surrounding the project and how a multi-science collaboration with the Square Kilometre Array will help them and the science community at large move towards the world of exascale computing.

CERN, the largest particle physics laboratory in the world, was founded following the Second World War as an act of peaceful scientific collaboration in order to understand the fundamental structure of the particles that make up everything around us, as well as unveiling the secrets of the Big Bang and the origins of the universe.

It is now most famous for being the home of the Large Hadron Collider (LHC), a 26.7km concrete-lined circular tunnel, around which particles are accelerated close to the speed of light and then collided with each other. These collisions produce rare particles, such as the famous Higgs boson, and by measuring their properties, scientists are increasing our understanding of matter.

From its inception, the planning behind the LHC has been both immaculate and meticulous. It has been designed to follow a carefully set out programme of upgrades over the course of three decades, with active experimental phases known as runs punctuated with 'shutdown' phases during which upgrades and maintenance take place. Now, the LHC is gearing up for its third run to begin in early-mid 2022.

In 2027, the fourth run of the LHC will begin, marking the beginning of operations for what will then be known as the High-Luminosity Large Hadron Collider (HL-LHC). This new machine will crank up the luminosity (a measure linked to the number of collisions occurring in a given amount of time) by a factor of 10.

The HL-LHC will produce around 15 million Higgs bosons per year, five times what the LHC is capable of now. Maria Girone, CTO of CERN openlab, a public-private partnership that facilitates CERN's collaborations with leading ICT companies and research institutes, explains: "The HL-LHC upgrade will change our situation from looking for a needle in a haystack to having many more needles available to find. We will be able to study Higgs bosons with a much higher precision, and also possibly observe rare phenomena that we have not yet seen so far."

A new era of data challenges

The amount of data produced already by the LHC is astounding by any standard. 150 million sensors deliver data 40 million times per second which, if written out, would be at the scale of petabytes of data per second. Filtering and data selection occurs just microseconds after the collisions in order to reduce the amount of data that must be processed, an extremely delicate process that permanently discards much of the data produced.

The transition from raw to reconstructed data is then performed on a distributed computing infrastructure called the Worldwide LHC Computing Grid, which is composed of about 161 sites in 42 countries. On top of all of this, simulation work is also carried out so that data can be compared to theoretical predictions during the analysis step. As the transition towards the HL-LHC gathers momentum, new ways of dealing with the increasing amounts of data will be essential.

"We are anticipating a need for about six to ten times more processing power and about three to five times more data storage at the HL-LHC," says Girone. "While we expect distributed computing to continue to play a vital role, these growth rates exceed improvements that we might expect in the existing general-purpose processing and storage technologies that we use today on the Worldwide LHC Computing Grid."

Speaking at this year's EHPCSW, Girone outlined the computing challenges being faced by the high-energy physics community, and the important role that high-performance computing will play in meeting them. CERN is currently pursuing an ambitious programme of research and development to modernise and optimise its software, adapt its codes to high-performance computers, reduce its storage footprint, and introduce more efficient techniques like AI and machine learning into its workflows.

"High-performance computing falls at the intersection of several important areas for us," says Girone. "It is potentially an invaluable tool for helping us close the resource gap we are facing, with the ever-growing power of supercomputers due to keep up with the bigger challenges being presented by the HL-LHC. As well as this, an enormous pool of expertise exists in the HPC community in developing applications and efficiently adopting AI and machine-learning techniques. That is why we believe engaging with the HPC community will play a critical role in the success of our future work."

Challenges towards exascale

Specific experiments at CERN, such as ATLAS and CMS, have already shown encouraging progress in demonstrating how HPC systems can be used in this context for a number of years, but in order to make a significant dent in the resource gap that will be faced as the pinnacle of data-intensive science moves towards exascale, the further use of HPC sites and the exploitation of heterogeneous computing architectures will be essential.

One of the key activities towards optimising high-energy physics applications for exascale is to have applications engineered for the hardware that will be used. Amongst other things, the EU-funded DEEP-EST project is working on an energy efficient architecture on the path to exascale for HPC as well as high-performance data analytics workloads.

Once the applications have been reengineered, challenges will still remain in bringing data to and from supercomputing centres. The Worldwide LHC Computing Grid is currently developing a data lake model that separates processing and storage functions. The system relies heavily on the network and intermediate caches to deliver the data. This architecture will be ideal for HPC systems, and a series of data challenges are currently being carried out to demonstrate the feasibility of this data lake model on the path to exascale.

The combined storage volume of the Worldwide LHC Computing Grid is currently one exabyte, with almost all data in ROOT files. Processing data intensive workflows in HPC involves moving a large number of these files from storage locations to processing resources. It is a complicated end to end service challenge involving many components. The network tools are now being modernised to use multiple 100 gigabit per second links, with a goal of eventually demonstrating the ability to process up to 10 petabytes a day.

An exascale project for an exascale problem

The hugely ambitious Square Kilometre Array (SKA) radio telescope project, which will be able to survey the sky more than ten thousand times faster than currently possible, is due to begin collecting data around the same time as the HL-LHC and faces similar challenges in terms of exascale computing and storage. Recognising their similar needs, CERN and the SKA Observatory (SKAO) launched a multi-science collaboration last summer alongside PRACE and GÉANT in order to tackle the technical challenges of HPC integration. "CERN and SKAO both possess in-depth knowledge of the applications and the challenges of data distribution, which is an important element for both sciences," says Girone. "PRACE is offering up its considerable expertise in system adaptation and in software environments, while GÉANT is providing the infrastructure that is needed.

"I think this collaboration will allow us to tackle the different challenges of HPC integration, with our programme of demonstrators that we have established from the start of the collaboration and that we are executing this year in 2021. I hope that this is really going to help us, the larger science community, to understand our path towards exascale. In a way we have come together to solve an exascale problem with an exascale project, and I hope in some way we can also show and pave the way for how collaborations in science should work in the future."

Scientific and industrial research is taking ownership of the revolution – started a decade ago by online enterprises such as Google and Facebook – of data-centric discovery as a complement to the simulation-centric approach traditional to the HPC community. While data processing has always been at the heart of scientific discoveries relying on large-scale instruments, the need to increase the performance of the associated data processing with extreme computing and the interdependence of accurate simulations and efficient design of experiments, has become more pressing.

The advent of data science, resulting from the access to and ability to extract information from huge detailed sets of data from research and other sources such as the Internet of Things, has extended the range of scientific communities that can benefit from large cyberinfrastructures to including social sciences, humanities and more. The COVID-19 pandemic has intensified the need for data gathering and analysis at unprecedented scales and speeds. This convergence of interests comes with the challenge of providing an infrastructure that is suited for this wider range of research topics, and that supports new discoveries efficiently.

Architectures of exascale computers relying on computing accelerators are a key element in this evolving landscape in providing, for instance, efficient platforms for AI-supported research. The challenge for the decade to come is to enable researchers to leverage the value of data from the edge where a significant part of the data needs to be collected, cured and filtered, to the data centre where it can be further processed at extreme scale. This is associated with the need to train researchers to use new transverse disciplines such as AI, machine learning, and data-mining, and to provide them with tools that manage the associated data logistics and implement large-scale workflows across the future computing continuum. These developments need to take into account the exponentially growing energy consumption of such large computing infrastructures. Making HPC greener should become a priority for hardware developers.

At this year's EHPCSW, the panel discussed the movement towards blending traditional simulation-centric research with new data-centric paradigms for scientific and industrial discovery and innovation, and what this will look like 10 years from now.

PRACEdays21 panel discussion

Will HPC, AI and data science be the same in 10 years' time?

HPC, AI and data science

The first question from the moderator Guy Lonsdale asked which aspects of simulation-centric HPC will still be here and which will have disappeared in 10 years' time. Mark Asch's work for BDEC has examined this type of question, and he thinks that simulations are on the way out. The way science is being done is changing due to the data revolution occurring. The extremely large volumes of data coming from instruments like satellites and telescopes, medical imaging, the Internet of Things, are changing the way things are done. Artificial intelligence and machine learning will play a bigger role. HPC centres will need to open themselves in terms of access and vision to play their parts in this new data-driven and model-driven science. The future will be the integration of exascale HPC into much more complex workflows.

Joost VandeVondele agreed, but he also sees a role for simulations that will remain. Models based on the laws of physics are going to remain important. Of course, more and more data will be associated with this, in terms of input and output. Simulations will be part of complex workflows where they might be generating data or used to analyse data in more detail. It's also important to look back 10 years – almost all systems then were CPU-based, so it is likely that the hardware side of things will evolve dramatically over a similar time period moving forwards. Systems will become much more heterogeneous and specialised, and this will impact simulation-based research.

Guy Lonsdale then posed a related question, asking whether discussions around AI and data science will still be talking about convergence with HPC in 10 years, or will HPC just be the standard tool that they use. Maria Girone spoke from the perspective of large data science like high-energy physics, where she hopes that in 10 years the technical challenges related to convergence of AI and data science with HPC will be solved. In the last few years, Girone's work with CERN openlab has involved a collaboration with the Square Kilometre Array, GÉANT and PRACE, in which they have been trying together to solve their future challenges surrounding data. Large science projects will have huge amounts of data to deal with, and HPC will be the tool to help. The software landscape in these projects will also have to move towards embracing heterogenous architectures.

Maria Pérez spoke about her work on the convergence between big data and HPC, which has been ongoing for a number of years. She believes that this convergence will be a reality in 10 years' time. Initiatives like BDEC have demonstrated the feasibility of this challenge, and many facets have in fact already been solved. She also mentioned that there is a need to distinguish between the real convergence of AI and HPC, which involves using the same solutions for both AI and HPC models, with using HPC as a standard tool for AI and data science. The latter of these two is happening more and more nowadays – HPC helps to deal with large volumes of data using GPUs and specialist hardware, and also uses simulations to work with the data produced. Pérez believes this will be a common reality for both science and industry in 10 years.

A question from the virtual audience was then taken, asked by Managing Director of PRACE aisbl Serge Bogaerts. He asked whether, beyond the underlying technologies, AI and data science will develop towards becoming new pillars of science? Maria Girone took the question, mentioning that in her environment of big science projects, they are hungry for data scientists to come to them and help them improve their capabilities for using AI everywhere in their data processing chains and in infrastructure. AI is therefore certainly becoming a pillar of science in her environment.

The role of hardware

Following on from this, Guy Lonsdale moved the discussion to the realm of hardware, asking whether

The panel

Moderated by Guy Lonsdale, SCAPOS

Maria Girone
CTO CERN openlab, CERN

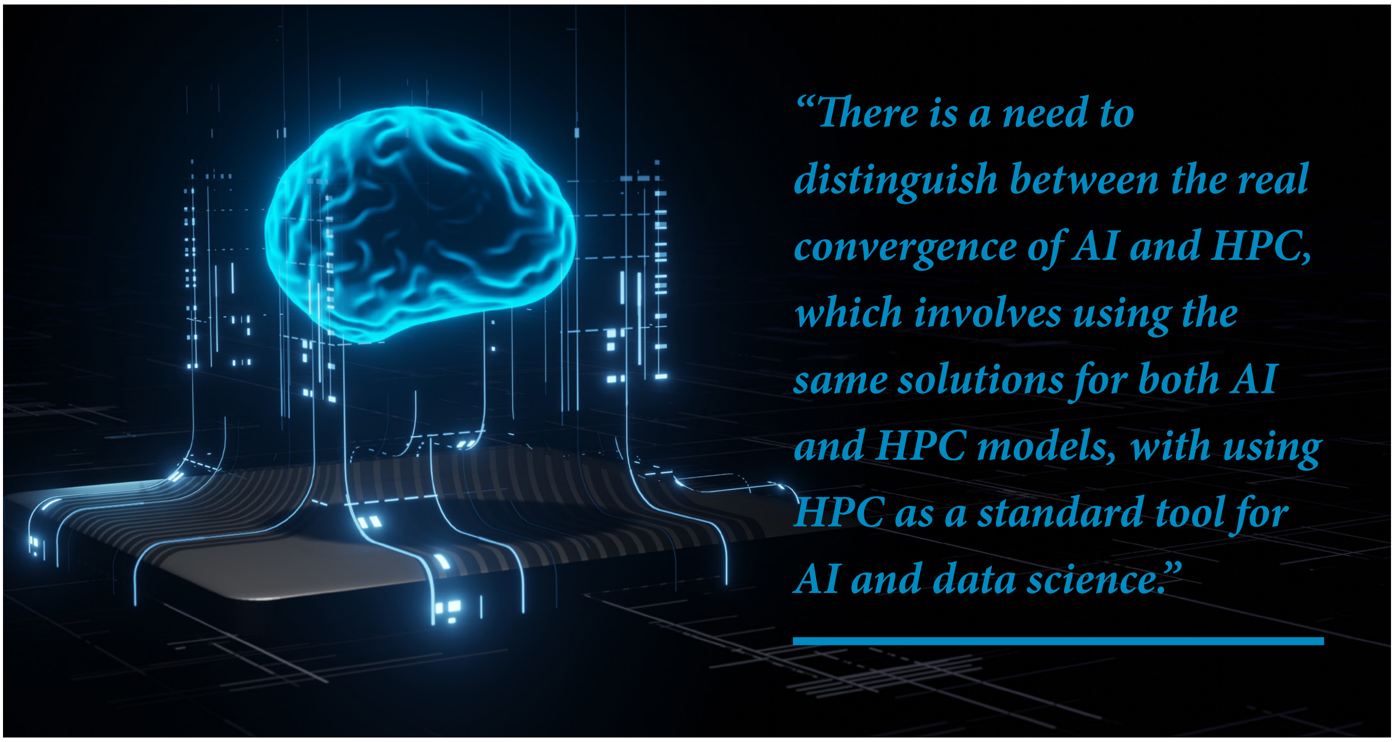
Céline Merlet
CNRS researcher and winner of the Ada Lovelace Award for HPC 2021

Maria Pérez
Full professor, University of Madrid

Mark Asch
Professor of Applied Mathematics, University of Picardie

Mauro Gameiro
Principal engineer, Critical Software

Joost VandeVondele
Associate director, CSCS



the panel expected there to be a converged compute node architecture in 10 years' time, or whether instead they thought there would be an increasing heterogeneity in the compute nodes being used? Mark Asch mentioned that there had been a lot of reflection about this question around networking applications and other cloud-based architectures. He believes there will indeed be diversity in the compute nodes as well as the applications. But can there be something more universal between the architectures and the applications that helps them communicate easily?

Asch believes that this will be extremely important, especially in international projects that deal with big societal challenges like global warming and big science challenges like radioastronomy. A lot of research will need to be done to work out how to formulate an internationally agreed standard, in the same way that the internet has HTTP, which will enable universal communication between the diversity of applications and hardware.

The discussion then moved on to the role of scientific industrial research community in influencing the direction of AI and data science – will this be a different path to those driven by the likes of Google and Facebook? Céline Merlet, who works on energy storage research, believes that there needs to be more thought put into how scientific and industrial research can converge. Mauro Gameiro explained how he thinks AI and data science in industry are mostly used for social networking and streaming and are therefore driven by those parts of the industry. However, in the shadow of giant brands such as Google, AI is also being used in the automotive, aerospace and other industries, but the targets are completely different. The same path as these companies can be trodden, as it has been a great success for them, but it will have to be adapted for each business case.

Another question then came from the audience, asking how the panel believed data assimilation would evolve in the coming decade. Maria Girone answered, saying that data is certainly the element that makes big high-energy physics projects stand out compared to other sciences at present. Having to deal with various forms of data has forced the community to learn many techniques about collecting and distributing it. There is a fundamental difference between how industry and science deal with data: in science, the end goal is to try and make discoveries, whereas industry use it to make money. Each community has its own challenges with data, and these are only going to grow over the coming years.

The digital continuum

Lonsdale then moved the conversation on, asking whether there are examples of scientific communities using complex workflows across large cyber infrastructures today, and how they would be expected to develop in the future. Joost VandeVondele brought up the weather and climate community as a prime example of a community that has a complex workflow, with inputs coming from sensors in the field which then have to be turned into products in a short space of time. This community is increasing its data volumes as it improves and is an example of a traditional HPC field that will be using a lot more AI and machine learning to extract and compress their data and provide it in an understandable way to decision makers.

Maria Girone used the opportunity to talk about the high-energy physics community and the work at the LHC, where they have been creating a large cyber infrastructure that is devoted to their research. They are now some of the leading experts in complex workflows for such systems. In 10 years, there will be much more complex data with the advent of the High Luminosity LHC, and much more storage and computing power will be need, but workflows will be very similar to those being used today. Collaborating together on the challenges towards exascale will be of paramount importance, and there have been successful meetings such as those run by BDEC which bring together communities to discuss the path to the future.

Communities outside of pure science, such as social sciences and humanities, have typically lagged behind science and engineering in the use and uptake of HPC. Guy Lonsdale asked the panel if they expected that to change, or whether they expected other aspects of the digital continuum to be the big game changers for those communities. Joost VandeVondele mentioned the use of data analytics by social sciences, which uses large amounts of computing power to analyse social structures. Natural language processing is also an area which has seen a lot of use. Activities are happening in areas that are considered social sciences, and although they have not yet fully penetrated into the HPC domain, it is likely that this will become more commonplace in the future.

Training and education for the future of computing

Measures are now being put into place to provide more training and education in the field of HPC, but if future researchers are also expected to learn about machine learning, data mining, artificial intelligence and more, is there a need to rethink undergraduate and graduate education?

Maria Pérez believes that the current programmes are insufficient for teaching expertise in all of these fields. Finding people who are proficient in all of them is almost impossible. Although some steps are being taken in some universities, more will have to be done for the future, and this will fall upon not just academia but also industry. Data science itself requires an understanding of computer science, mathematics and statistics. Then, specific domain knowledge is needed on top of all of this.

"Everyone will to some extent have to move towards a computational way of thinking when working in teams with complex workflows."

Mauro Gameiro agreed, saying that universities teach these things, but from the experience of Critical Software, there is a gap between what people learn and what is needed to work in industry. The solution will be to somehow reduce this gap by increasing collaboration with professors that teach these technologies, perhaps helping them to teach using real examples so that the techniques being learned can be applied in practice.

The discussion then turned to whether a move towards multidisciplinary research teams as a standard approach might be an alternative to training domain experts in all of the digital continuum technologies. Céline Merlet believes that there will indeed need to be some changes in university teachings as it is not practical for people working in science to have no knowledge of these things. However, she does think that multidisciplinary teams are important, because when working on a day-to-day basis, it is good for people to be able to ask questions to people with a bit more knowhow.

Merlet has always worked with experimentalists and has always valued being able to tap into their knowledge, and believes the same applies for domain experts being able to ask questions of digital continuum experts. Joost VandeVondele agreed, saying that there will need to be a joint language between all of these experts, and that everyone will to some extent have to move towards a computational way of thinking when working in teams with complex workflows.

Moving towards exascale computing

Finally, the panel was asked whether they think exascale computing will be standard in 10 years' time. Joost VandeVondele answered yes, because if we look back 10 years then we can see a similarly large change in the power of computers available to researchers. The driving force is of course that good science needs these resources. Every researcher knows in the back of their mind what they could do better with their computational work if they had more resources available, and this will be the same 10 years from now, so the demand for ever more powerful computers is not going to change.

A big congratulations on winning the best poster award. Could you tell us a bit about your background as a researcher and how you ended up working with high-performance computers?

Thank you very much. I basically started out as a physicist with an interest in cosmology, particle physics and computing. Although all my history as a physicist had an impact on who I am as a researcher I think there are two moments which explain the motivation behind my current work. During my bachelor's I worked with my now-PhD-supervisor Professor Carlos Martins in cosmological simulations and during my master's I was exposed to GPU computing during a computational subject lectured by Professor Ariel Guerreiro. For my PhD I simply had the idea of exploring GPU computing in an area where it had not been used previously as it deeply intrigued me from a physics point of view.

What was your poster about?

The poster is about improving the performance of cosmic string simulations. Before I continue allow me to explain what a cosmic string is by analogy. Think of a glass of water. As it cools down, there is a specific temperature where water turns to ice but this does not happen homogeneously. As a consequence, the resulting ice contains fault lines or "cracks". Some theories of new physics propose the existence of fields which, as the Universe cools down, after a phase transition, end up forming filaments. These filaments - cosmic strings - are then somewhat analogous to the fault lines in ice. When a network of strings evolves, strings can interact and as a result, structures such as sharp kinks or closed loops of string appear, which inevitably will leave some observational imprint.

Simulations of these objects, when one does not consider Input/Output of data, are often bottlenecked by the amount of memory and memory bandwidth available. This poster is an attempt at tackling these bottlenecks, by exploiting the inherent high bandwidth of GPU memory and using in-situ visualisation techniques to reduce the amount of data to be output to disk.

Why was high-performance computing needed for your research?

The study of defects in cosmology (and by consequence their observational footprints) can be done via semi-analytical models, which allow one to predict the evolution of a cosmic string network throughout the entirety of cosmic history although they have an unavoidable caveat: such models can free parameters which cannot be calculated ab-initio. Curiously simulations cannot simulate a network for such a long period of time, however they can be used to calibrate the free parameters of the aforementioned model - either via an average characterisation of the network or via the positions of strings. In a sense there is a symbiotic relationship between these two.

The calibration however can be highly dependent on the ability of a simulation to accurately resolve subtle structures along the strings, which means we have a need to target larger swaths of spacetime. High-performance computing plays an important role as without

José Correia

Q&A

José Correia, winner of the PRACE Best Poster Award 2021

adequate hardware resources and advanced computing paradigms one cannot simulate such large lattices with the necessary resolution to study and improve these models. This is even worse when we add more than one string type to the simulation and more resources are necessary to calibrate models with even more parameters.

What does winning this award mean to you?

I am honoured and happy to have received this award. As a recognition of my efforts, it gives me even more motivation and confidence to continue research in cosmology.

How did you find the experience of EHPCSW this year?

It was a curious experience for mainly two reasons. It showcased at a European and national level the scientific research conducted in different fields via HPC resources. Additionally it displayed the current and future resources available to European researchers. As an HPC user and developer it is exciting to see the field evolve and witness new supercomputing centres with different architectures and new training units being announced.

Scientific keynote

AI for improving HPC efficiency and energy footprint

The influence of artificial intelligence grows every year and can now be seen in all aspects of the digital continuum. **Michela Milano** delivered a scientific keynote presentation on her work in using AI for improving the efficiency of and energy footprints of HPC machines.

The recent success and developments of AI algorithms, in particular machine learning and deep learning, require computing systems to be scalable, flexible and have low latency at all levels to effectively tackle these challenges. On the other hand, hardware is becoming increasingly heterogeneous, with many domain-specific processors such as GPU and FPGA.

To extract performance from large-scale heterogeneous architectures in a seamless ecosystem, new programming paradigms, languages, compilers, operating and runtime systems will be needed to provide new abstractions and services over highly heterogeneous hardware. Michela Milano's talk provided an overview of how AI techniques can support computing systems to be efficient, to consume less energy while staying within error and time constraints, and to detect anomalies.

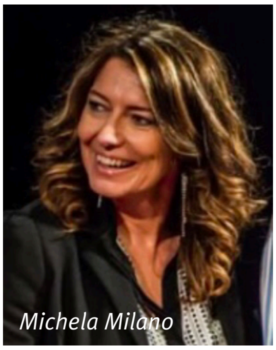
AI-based power capping

One of the main benefits that AI has brought to the table is power aware systems for minimising power consumption. Today's supercomputers are power limited – an acceptable value for exascale computer is around 20 megawatts, and so power capping is a widespread method for dealing with energy-related concerns.

Power capping supported by hardware limits the performance of the computing nodes without affecting the quality of service. Dynamic voltage and frequency scaling slows down different cores to reduce power consumption, but increases the duration of jobs, affecting deadlines and costs. With current pricing schema based on the effective runtime, slowdowns would increase the price for the user, so accounting schema will need to be redesigned to apply discounts for users that have slowed down jobs.

A very clear research question to ask is how power capping can be enforced by acting on the job execution order. If the power consumption is known before the execution of the job, then the power consumption can be used as any other resource and can be scheduled accordingly and shifted in time depending on power caps. But how can this be done?

Two AI based components are needed: (1) a mechanism to predict



the job power consumption before the task has run, and (2) a power-aware job dispatcher. The first one can be based on a machine learning model trained on data collected on job power consumption. Once this is in hand, a power aware job dispatcher can be created, which has been part of Milano's work using constraints programming and heuristics carried out on the EURORA HPC system at CINECA. Milano's work required a dataset that compares job features with power consumption, and this was obtained through monitoring frameworks that collect system information, available on most HPC systems nowadays, using hardware sensors that record data on core load, temperature, power consumption and more every five seconds. This data has been used to create and train a machine learning model that predicts how much power a job will use.

Different regression methods have been tested for this task, the most successful of which was based on Random Forest. The time required to train the model is very short, and the model will then have an accuracy of around 90%. A dispatcher based on a combinatorial optimisation model can then be used to determine where to place jobs and at what time they should start. This requires combining a fast heuristic technique with constraint programming. Two different approaches have been explored for this. The first one is a priority rule-based scheduler, while a second hybrid decomposition approach has also been developed.

Compared with state-of-the-art approaches to power capping, excellent results have been achieved with the new approach when job arrival frequency is high. Less benefits are seen when job arrival frequency is not so high, but the results are still very good. Overall, merging hardware-based power capping with the new decomposition approach provides a number of benefits.

AI for anomaly detection in HPC

Detecting anomalies and faults in HPC systems is difficult. For example, if a machine stops in a manufacturing plant then it is easy to

understand that an anomaly has occurred. In HPC systems, anomalies are more subtle. For instance, they can be due to wrong configurations that do not stop the machine, making them difficult to detect. There are hundreds of thousands of possible sources for malfunction in an HPC system, so the ideal solution would be to automatise the detection process of such anomalies.

Milano and her team have used two approaches, with their results based on a case study of the DAVIDE systems at CINECA, which is equipped with a detailed monitoring infrastructure developed at the University of Bologna called EXAMON. This collects data on the system by measuring 200 metrics every five to ten seconds, providing a large historical dataset that can be used for research purposes.

Three types of metrics are collected: node related information, CPU-core related information, and system-wide information. All of this data is used by a deep learning model that helps to understand whether there is an anomaly or not.

In a system like this, the number of states that are considered anomalous is much lower than the number of states that are considered correct. A model is needed that is capable of learning relationships between groups of features. Two different approaches have been used for the deep learning and machine learning model for anomaly detection. The first one is called semi-supervised, while the second one is totally unsupervised.

In the semi-supervised method, an auto-encoder is used which comprises a deep network that tries to reconstruct the input provided through an encoding and decoding stage, minimising the loss function, i.e. the difference between the input and the reconstructed input through an encoding and decoding step. It is able to grasp the core salient features of the input data and reconstruct the normal behavior of the HPC system by minimising the loss. When anomalous behaviour is inputted, the reconstruction loss is very high, acting as an indicator of such anomalous behaviour.

Milano's team has also explored a fully unsupervised method, which tries to cluster different traces from the EXAMON monitoring infrastructure. The goal here is not just to detect an anomaly, but also classify the type of anomaly observed. In this method, a variational auto-encoder is used. Standard auto-encoders create a latent space for mapping features of the input data, but these latent spaces are non-continuous. In contrast, variational auto-encoders can create continuous latent spaces, making it easier to identify clusters after training.

The variational auto-encoder can be trained using all training data with no need for labels – this is what makes it "unsupervised". There is no need to discriminate between normal and abnormal anomalous detection, and the latent representation contains clusters for each class of behaviour. Then, a clustering algorithm can be applied to the latent representation to identify the clusters.

The green

HPC session

The Green HPC session at EHPCSW, moderated by Pedro Alberto of the University of Coimbra, brought together experts from a number of fields to talk about the different ways in which high-performance computing can be made as environmentally friendly as possible. With the dawn of exascale computing on the horizon, it is more important than ever that every facet of these machines, as well as the software that is run on them, is explored for ways to reduce energy consumption.

Sustainable HPC datacentres in Kajaani

Opening the session was Jukka-Pekka Partanen from CSC, a Finnish company that provides IT services for science. He used his time to talk about the company's sustainable solutions for datacentres in Kajaani, a town in north-eastern Finland. CSC has modern and reliable infrastructure, safe data storage and 50 years of expertise in serving HPC customers. Core network connectivity from Kajaani is also excellent. CSC runs the Finnish research network FUNET, and its datacentres are connected to both NORDUnet in the five Nordic countries and GÉANT in the rest of Europe.

Kajaani is an excellent location for sustainable datacentres for a number of reasons. Tens of thousands square metres of free space is available in greenfield, brownfield and ready-made white space solutions. Four national grid links come to the site, the power from which is scalable up to hundreds of megawatts based on customer needs. Green energy production in the region is excellent, with three hydroelectric power plants and a wood burning biomass power station. Finland's cool climate is a huge advantage for reducing energy consumption in datacentres, with those in Kajaani having some of the best annual power usage effectiveness ratings in the world. CSC also has a full team of skilled staff on site who have been instrumental in running successful datacentre operations and HPC services since 2012.

Their operations have been granted the ISO 27001, making them one of the first datacentres in Finland to achieve this goal. This and many other factors, such as the political stability of the country and lack of threat from natural disasters, mean that Kajaani is perfectly placed for delivering world class eco-efficiency.

With LUMI, one of EuroHPC's world-class pre-exascale supercomputers, set to be built on the site of the CSC datacentres, the plan is to use waste heat generated by the HPC system in the district heating network of Kajaani, replacing 20% of the district's annual heating production from fossil fuels. In total this this will reduce annual CO_2 emissions by about 13 500 tonnes – equal to the emissions of about 4 000 cars – meaning that the carbon footprint of the site will actually be negative.

Overall, Kajaani will bring over €10 million savings annually compared to most of the HPC datacentres locations in Europe. This is mainly because of the price of the electricity, which makes up around 60% of the annual cost of running an HPC facility.

Hardware and software codesign for energy efficiency and performance

Next came a three-pronged presentation from E4 Computer Engineering, the University of Bologna and CINECA. Fabrizio Magugliani of E4 began by explaining the company's approach to designing systems for energy efficiency and performance. They work from the notion that these two goals are two sides of the same coin, and that one must hold both in the mind if they are to achieve them successfully.

Customers of E4 now expect energy efficiency to go hand in hand with high performance, and this can only be achieved by incorporating energy efficiency from the very initial steps of design. The company believes that the best way of achieving this is by incorporating three main groups of stakeholders: hardware developers, system level developers, and the the end users.

As hardware developers, E4's company mantra is to build on proven technology, prepare and drive for the next generation of technology, and lead in innovation. This mantra is implemented through five steps. Start with the customers overall requirements, select the best technology and the best providers, test to ensure that the solution is what the customer is really looking for, validate the solution with the technology providers and with the customer, and finally deploy the solution.

E4 has strong connections with the major global silicon vendors, talking on a daily basis and providing real input about customer requirements. They are also proud partners of the European Processor Initiative (EPI) as part of the science team, and they intend to have the EPI processor in their upcoming product lines. E4 also codesigned the cooling solution for DAVIDE, a system hosted at CINECA developed within phase III of PRACE's pre-commercial procurement. Looking to the future, E4 is collaborating with a startup InQuattro on the next generation of two-phase cooling equipment, prototypes of which will be completed in 18 months as part of the TEXTAROSSA EuroHPC call.

Providing input from the system level developer side was Benedetta Mazzoni from the University of Bologna, who spoke about scalable and multi-scale energy monitoring. Measuring power consumption at multiple scales and granularity enables important services that ease the maintenance and operation of datacentres, helping reducing energy associated costs and making the datacentres more sustainable and greener.

The complex nature of datacentres makes measuring their power and energy consumption an intricate task. To take full advantage of this information, it is important to correlate it with other measurements like computing nodes, activities, jobs running and facility state. Mazzoni and her colleagues have for the last year been working on the design of solutions for holistically monitoring datacentres.

EXAMON, which stands for exascale monitoring framework, is a scalable framework developed by the University of Bologna alongside CINECA and E4 engineering for applying IoT, big data and AI technology to the holistic monitoring of datacentres. In applying it to the DAVIDE supercomputer, they were able to reduce sampling time of power consumption down to milliseconds and microseconds.

The team discovered that when measured at this fine granularity, power consumption enables the extraction of effective application fingerprints, which can then be used to counteract cyber hazards. To ease the application of this methodology on the Marconi 100 system, they integrated fine grained energy monitoring functionality directly into the board management controller firmware without the need for additional hardware. Furthermore, in the context of the European Processor initiative, they are working on the design of a novel open-source power controller capable of leveraging fine grained power consumption information at the micro architecture level to an enhanced energy proportionality.

Daniele Cesarini, an HPC specialist at CINECA, then spoke about energy-efficient runtime systems. Codesign is the way to achieve energy efficiency without sacrificing performance, but addressing one component at a time won't result in maximising the entire ecosystem-wide efficiency. An approach based on system runtime is therefore mandatory for optimising the overall energy consumption of an HPC system.

Cesarini and colleagues have developed COUNTDOWN, a tool that identifies and automatically reduces the power consumption of CPU cores when running MPI based applications. It is completely transparent from the applications, so does not require modification of the source code. It is able to collect information on overall power and energy consumption of the compute node and several other performance metrics. It can work together with EXAMON to reduce power consumption and monitor the power variation of a compute node in real time.

"Codesign is the way to achieve energy efficiency without sacrificing performance, but addressing one component at a time won't result in maximising the entire ecosystem-wide efficiency."

Optimising numerical precision to improve energy efficiency

The final talk of the session was delivered by Fabienne Jézéquel of the LIP6 laboratory at Sorbonne University, who spoke about the optimisation of numerical precision for improving energy efficiency. When a numerical simulation is performed, floating point arithmetic is used, of which there are various formats. Half precision is now increasingly supported by new hardware, including some recent GPUs and ARM CPUs. By decreasing precision, execution time is decreased, volume of results-exchange in parallel applications is decreased, leading to improved energy efficiency.

Energy consumption is in fact proportional to P^2, where P is the number of bits in the mantissa (the part of a floating-point number which represents the significant digits of that number). Moving from double to single precision reduces energy consumption by five times, while moving from single to half precision provides similar reductions. However, doing so can cause computed results to become invalid because of rounding errors.

Several approaches exist for rounding error analysis. Interval arithmetic consists of replacing scalar values by intervals, providing guaranteed bounds for each computed result, but the error may be overestimated, and so interval arithmetic cannot be applied in a naive way in a large code. Another approach is static analysis, a rigorous analysis that requires no execution of the code and takes into account all possible input values. However, it is not well suited to large programs. Finally, the best method is to use a probabilistic approach that estimates the number of correct digits of any computed result, can be used in HPC programs, and requires no algorithm modification in the user code.

Jézéquel described the principles of stochastic arithmetic, where each arithmetic operation is executed three times with a random rounding mode i.e. each result is rounded up or down with the same

probability. The differences between these results are due to rounding errors, and the number of correct digits in the results can be estimated using a Student's test with a confidence level of 95%. Each operation is executed in a synchronous way up, meaning each operation is executed three times before going to the next operation, so that numerical instabilities can be detected.

Stochastic arithmetic is implemented in the CADNA library, which can be used to control rounding errors in C, C++ and Fortran codes. CADNA provides stochastic types consisting in three floating point variables and one integer. Stochastic types exist for half, single, double and quad precision. All operators and mathematical functions are overloaded for these stochastic types, so CADNA requires only a few modifications in user programs. It can be used in parallel codes that are based on MPI, OpenMP, GPU and vectorised codes. In one CADNA execution, the accuracy of any result can be obtained along with a complete list of numerical instabilities.

Another tool is the SAM library, which implements stochastic arithmetic in arbitrary precision. SAM provides a new stochastic type that can be used to control rounding errors in arbitrary precision. Recently, a multi-native version of SAM has been developed in order to control operations that mix different precisions. For instance, this allows one to control rounding errors in operations involving three variables with different mantissa lengths. This is particularly interesting for accuracy estimation on FPGA, where the mantissa length is crucial for performance and also for energy efficiency.

With these numerical validation tools, it is possible to use reduced or mixed precision to improve performance and energy efficiency. There are existing mixed precision linear algebra algorithms, for instance for the solution of linear systems with iterative refinement. The design of mixed precision algorithms is a very active area of research. Precision autotuning tools for large codes also exist, but these rely on comparisons with the highest precision result, which can give completely different results depending on the level of precision.

To solve some of these issues, Jézéquel and her colleagues have developed PROMISE which performs floating point autotuning based on CADNA. PROMISE provides a mixed precision code that can use half single, double and quad precision, and takes into account the required accuracy of the results. PROMISE uses CADNA to validate a type configuration and uses the Delta Debugging algorithm to search for a valid type configuration. It does not do an exhaustive search, which would be very costly, but is able to provide a valid type configuration with a reasonable complexity.

Jézéquel concluded that to optimise precision and improve energy efficiency, numerical validation tools such as CADNA, precision autotuning tools such as PROMISE, and existing mixed precision algorithms can all be used. Her team plans to extend floating point autotuning to arbitrary precision, and also plans to combine mixed precision algorithms and floating point autotuning. In floating point autotuning tools, the mixed precision versions of algorithms could be used automatically when they exist.

Taking EuroHPC Summit Week 2021 digital

Marjolein Oorsprong has been organising the EuroHPC Summit Week including PRACEdays successfully for many years - until 2020 that is, when the pandemic closed Europe down and the live event was cancelled. After much thought, planning and late nights, however, Marjolein and her team managed to deliver the event digitally for 2021, putting together a programme that showcased the very best of Europe's HPC to more delegates than ever before. Here she takes us through the ups and downs of making EuroHPC Summit Week 2021 digital.

Marjolein Oorsprong

Planning started for this event in 2019 because that was when we decided to have the EuroHPC Summit Week 2020 in Porto. I worked with the team in Portugal to get everything organised for this – the venue, hotels, meeting rooms , catering and the programme, even the location for the gala dinner. Everything was in place – and then the pandemic hit.

So, we couldn't go to Porto in March 2020, and there was no chance to do anything digital due to the short notice between the start of the lockdown and the planned event, so we decided to postpone and host the event back in Porto in 2021, thinking all would be fine by then.

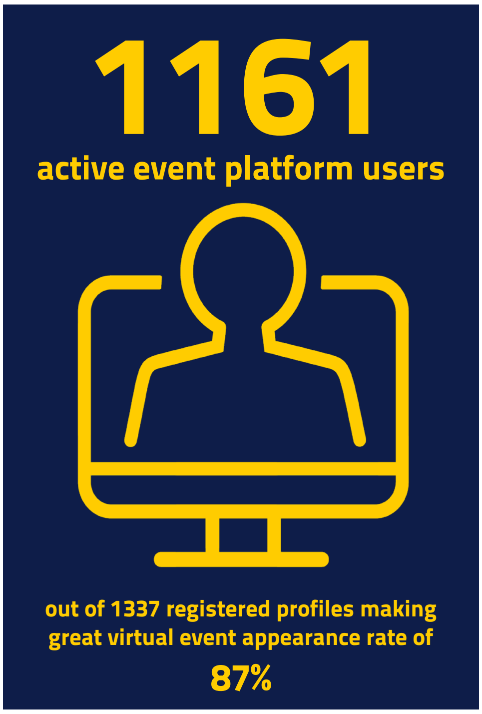
Of course, the pandemic dragged on, lockdowns were not lifted as expected and travel restrictions were still in place – but we decided to wait and see how things developed. When the summer of 2020 passed and March 2021 was getting closer, I knew time was running out to get the planning started and everything in place for whatever we were going to do. We had to make a decision.

I was sure we were not going to be able to have a live event and even if we could to some extent, we would have to do something for those who could not be there. So we looked at a hybrid event, with limited meetings onsite for those in Portugal and livestreaming for everyone else. But by the time we got to October, I realised that this was just not going to work and I pushed for the Organisation & Programme Committee to make a decision to go fully digital. My reasoning was that doing a hybrid event was going to be like doing two conferences at the same time and there just was not enough time to do this. There was not even scope to move the event and buy more time as the EHPCSW follows the EU presidency calendar, so if we moved we would have to go to the next presidency country. Portugal would not be happy!

Once we had agreed to go fully digital, I had to find a supplier for the platform to host the event and I wanted something bespoke – branded, lots of bandwidth and flexible enough to host a large and simultaneous programme. So I put out a call to tender for the platform and get the supplier lined up before the end of 2020 in order to have things ready for the March 2021 event.

And that was just the platform. I then had to reorganise the speaker programme and organise how we were going to get everything online and interactive in time. The speakers were all in place, but the programme needed to fit in a new timescale and account for the fact that audiences are just not going to sit in front of their screens from 8am to 8pm, so we had to spread things out over five days instead of three and run many sessions in parallel.

While finding the right platform and getting that designed, I worked in parallel on setting up a new programme, redesigning the schedule to fit five



days from 10am to 4 or 5pm every day. That's when I created the themed days that we had in the end – with Monday for EuroHPC, Tuesday for industry, Wednesday for COVID-19, Thursday for science and Friday for the European HPC ecosystem.

I found a platform by the end of December, with five bids in total and only two being eligible, choosing the one we had over a more complex 3D system that would give delegates the illusion of walking into the Porto venue online and being able to interact in this virtual world.

That, of course, would have been wonderful, but I felt it might distract from the planning needed and we would be spending all our time working on what the platform could do rather than the content of what we would be presenting, how we would be doing that and how the audience could participate.

So, we chose the platform we ended up using and had an agency handling this development, programming it and branding it for us. We then made the decision to ask speakers to pre-record their presentations to simplify things further. I knew that by going digital we would have many more attendees – and this was indeed the case with 1 337 registrations as opposed to around 450 in our previous event – so I wanted to minimise any complications with live broadcasts on the day. Not everyone wanted to pre-record, so we had about a 50/50 split between live and pre-recorded, while the Q&As were live too.

Another complication was the sessions in which we had live speakers and pre-recorded speakers presenting alongside one another, so I got another team in to run a sort of virtual TV studio in the background for each of the sessions and this worked really well, with them making sure that every speaker, live or otherwise, was lined up and ready to go at the right time and that the sound was on and no one muted, even during the Q&A sessions. And, as everyone saw, it all worked very well.

The greatest challenge I faced in all this was getting everyone involved in all this to respect the deadlines. I had people sending me recordings just days before the event went live and this was a big headache, especially when the recordings that arrived were not of great quality! There was a lot of chasing and working through the night to get everything ready.

Overall though, I think the digital event was a great success. I would even say that it was a lot more inclusive because it allowed a lot of people who would not normally have the funds to travel in person to the event to attend. For example, we had quite a large contingent from Africa and these were people who can't come to Europe unless we fund or sponsor them.

The other value of the digital event is that people seemed to be more active in the sessions as they were able to pick and choose what to attend more easily. I think when you are attending an event in person all week, you may feel obliged to attend sessions that

don't interest you that much because you are there and you may easily be distracted in these sessions. In the digital event, people seemed to choose carefully what to attend, block off the time and really participate and stay engaged in the sessions that interest them.

Of course, one big difference between live and digital for me is that I spend a lot of time at live events fixing things with 'duct tape and paperclips' – putting up posters, fixing a microphone, getting new chairs in place. With a digital event, you can't do that and so you need to have everything planned. There is little margin for fixing last-minute problems.

While the inclusive nature of the digital event was a real benefit, I feel what it lacks is the scope for personal interaction – meeting for a coffee, bumping into people you haven't seen for a while, getting together in groups for meetings. You can't do this in a digital form. I would always choose a live event where possible and I look forward to doing our next one all together in one place. I miss that. I miss my team on the ground and I miss seeing all the delegates.

Having said that, I believe this year's event was a huge success. We had great numbers, of registrations and also of people attending on the day and remaining active throughout. In the end, I got so much positive feedback about the event: people really did seem to enjoy it. And, of course, the content of our presentations and discussions was amazing again this year. The COVID-19 day was very interesting and everyone commented how good it was to see how much science is going on in the background, when all we normally see on the news is the end result – the vaccine or the new treatment. That was a real highlight.

I was also so happy with the involvement of EuroHPC and this attracted a great deal of additional interest which I am sure increased attendance at all the other sessions from delegates who may not have come otherwise.

The diversity track was also a huge success. I was sad that we couldn't do this live as I had all sorts of exciting plans for this one (wait for next year), but I had many people approaching me to be involved and that doesn't normally happen.

The session on the impact COVID-19 has had on diversity (see page 49) was also very interesting and a highlight for me. We will attempt to do this again next time, enmeshing the diversity track with the other content rather than bolting it on.

Would I do it all like this again? Yes. I now know what I am up against and I would give myself more time. Of course, I probably will be doing it again, because now the bar has been raised: we will have a digital element to the live event next year, with live streams and digital interaction, and what we have learned this year will make all that possible.

European HPC Ecosystem Summit

At this year's EHPCSW, the European HPC Ecosystem Summit brought together some of the major players in European supercomputing to discuss the development of new organisations and initiatives that will help improve the quality of and access to HPC resources in Europe.

The European Open Science Cloud

John Womersley began by speaking about the European Open Science Cloud (EOSC). The goal of EOSC is to facilitate the transition in the way that science is done. This is already underway, between a publishing paradigm of science to science based on electronic access to open data. This will enable a much broader participation in the research agenda. To achieve this, skills and rewards and appropriate metrics will need to be provided, so the goal of EOSC is to stimulate and support this new ecosystem and facilitate this transition in a coherent way that works across disciplines. EOSC is working to make data open, but also to make sure that it is FAIR (Findable, Accessible, Interoperable and Reusable).

At its heart, EOSC will be a way of accessing digital objects – data, code and other research outputs. It will enable them to be accessed by associating them with identifiers, providing standards and code that work with the digital objects, and providing contextual information that enables them to be discovered. It is important to note that EOSC is almost like the twin sibling of the European e-infrastructure organisations, which will offer the storage, compute and connecting services needed to service data and create interoperability.

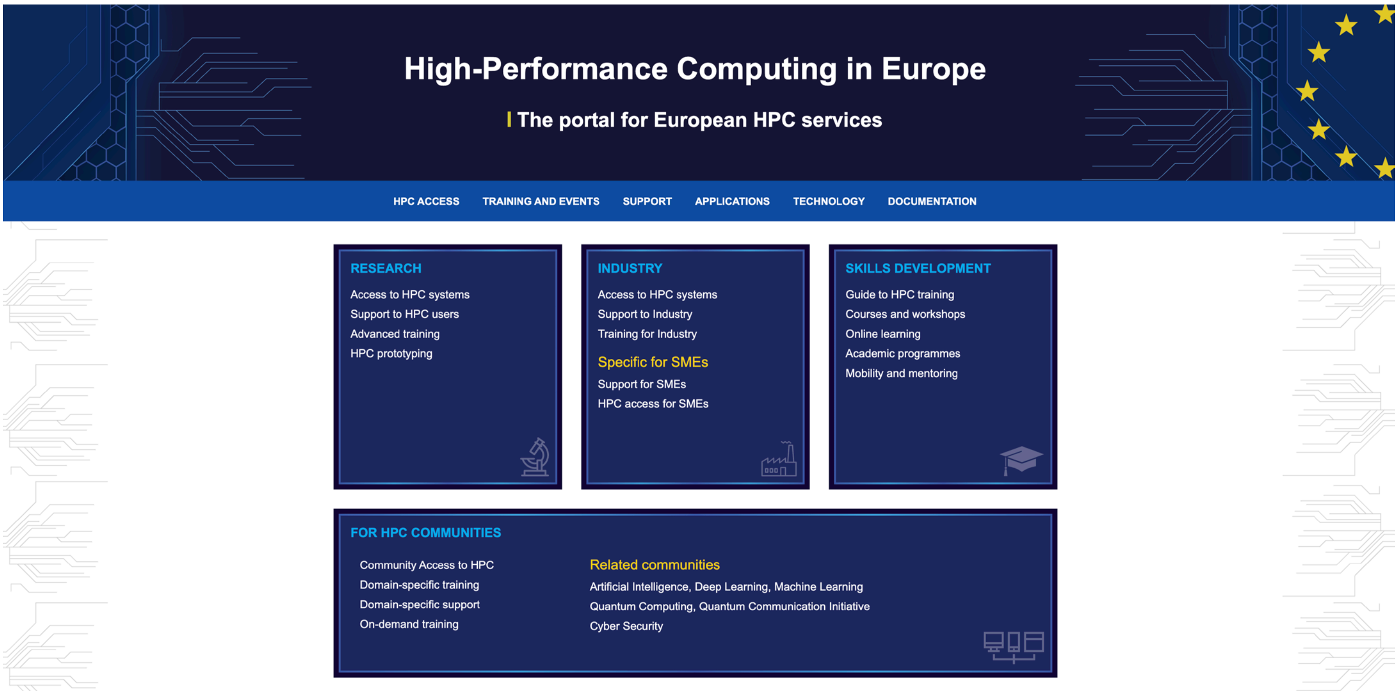
The overarching principle for EOSC is that research has to be at its centre. It will follow a multi-stakeholder approach and ensure research artifacts are as open as possible and only as closed as necessary. The FAIR principles of making data findable, accessible, interoperable and reusable will be followed, and existing and upcoming data and e-infrastructures will be federated. Finally, EOSC will strike the right balance between machines and people to deliver the services needed by European scientists.

A number of critical success factors for EOSC have been identified. Researchers who perform publicly funded research need to make relevant results available as openly as possible, and this data must be FAIR by design. Professional data stewards will need to be available in research-performing organisations in Europe to support open science, and researchers must be incentivised to perform open science. The scope of EOSC will eventually need to be widened to serve both the public and private sectors, and will need to support a wide range of digital objects including data, software and other research artifacts. The EOSC Association certainly has a big task ahead of it. Some of the challenges facing it include ensuring the sustainability of accessing data beyond the Horizon Europe timeline, maintaining a common approach within a very heterogeneous funding landscape and many national initiatives, and integrating the data infrastructures with the compute and storage infrastructures in a way that allows useful services with low latency.

The European HPC Ecosystem

Florian Berberich of PRACE then took to the virtual floor to give an update on the European HPC ecosystem with a focus on the links to EOSC and the impact of EuroHPC. There are four main cogs of the European HPC ecosystem that need to be maintained to ensure that it runs smoothly. Infrastructure involves making the best high-performance computers available for science and industry.

Technology involves the development of new hardware such as exascale computers and processors. Applications is about ensuring that Europe maintains excellence in developing applications for HPC. Finally, pan-European coordination now widely rests on EuroHPC, which will bring a lot of funding for new HPC systems and also aim to widen the usage and skills of people in academia and industry.



The home page of the HPC in Europe Portal, which coordinates European HPC services. Visit https://www.hpc-portal.eu/

The European Data Initiative (EDI) and EOSC will both play key roles in the European HPC ecosystem in the future, and Berberich outlined some ideas about how they will work together in the future. Both initiatives are linked in a number of ways via networks, data and their users, and aim to enhance Europe's science, economy and society.

However, the do have a number of differences. For instance, although they share the principle of FAIR data, EOSC focuses on this in terms of data and EDI focuses on computing. Berberich believes that cloud access to HPC resources should be a common goal for both organisations.

Collaboration between the two organisations should lead to a win-win situation. Broad science communities, outside of the well-established HPC communities, will have access to HPC, creating new user communities and the development of new applications. There must be a continuum of computing and data resources for academic and industrial users, including SMEs, from laptops up to high-performance computers.

Moving on to the impacts of EuroHPC, Berberich described how procuring petascale, pre-exascale and exascale systems for Europe will be one of its main goals, but that its role will not stop there. A research and innovation programme has been launched that will cover extreme-scale technologies and applications, innovating and widening the use of HPC skills, integrating building blocks, an EPI follow-up call, and a call on education and training.

Two important additions will be added to the European HPC ecosystem as part of EuroHPC. The CASTIEL project, which started last year, will coordinate and support the national competence centres that will be set up as part of EuroHPC, and EuroCC will act locally to map available national HPC competences and identify existing knowledge gaps, encouraging integration and collaboration between the competence centres to strengthen the European ecosystem and serve HPC users in Europe.

The HPC in Europe Portal

The final talk of the session came from Oriol Pineda of the Barcelona Supercomputing Centre. The objective of the HPC in Europe Portal is to collect and group European HPC services through a bottom-up approach and the creation of a centralised database that can be accessed by all stakeholders. HPC access, training and events, applications, technology and documentation will all be available through the portal. Target groups for the portal include researchers, industry, communities, students and the general public.

The HPC in Europe Portal will need to be user-oriented, easily adaptable to future changes, convenient to administrators, allow automated imports, have neutral European branding, and provide acknowledgment to contributors. The overall objective is to develop and make available a framework for all stakeholders to use and benefit from, and then promote and facilitate the use of this framework. It will not just be a simple collection of external links, instead providing lots of useful contextual information around everything held in the framework.

The platform is now up and running, and so the next goal for those working on it is integrating it with other initiatives. Major stakeholders have been identified, and work is ongoing on the development of intercommunication between databases, so that information posted in one place will automatically be updated in all of the relevant areas.

PRACE and COVID-19

How HPC has fought the pandemic

The onset of the COVID-19 pandemic took the whole world by surprise, but PRACE was quick to act, forming a fast-track call for research which has now awarded over half a billion core computing hours. Here, **Matej Praprotnik**, Member of the Scientific Committee of the Fast Track Call for Proposals, tells tells the story of how events unfolded, which eventually led to this year's EHPCSW session on the crucial research that has been conducted to help mitigate the effects of the disease.

Matej Praprotnik

Last year, when the severity of the COVID-19 pandemic became clear, we proposed to organise a fast-track call for research on COVID-19, after which we quickly moved to define the scientific topics of the call. I am one of five scientists (together with the former Chair and Vice-Chair of SSC and the Chair and Vice-Chair of Access Committee) who acted as the selection committee for the incoming projects.

We selected the projects that we felt fitted the call and then sent each of them to at least two scientific external reviewers. We then made a final decision to accept or reject projects, and finally awarded the accepted projects with core computing hours on PRACE machines across Europe.

We have welcomed a huge range of projects connected to COVID-19 research. These have included biomolecular research that helps to understand the mechanism of the disease, bioinformatics research that studies how different mutations affect the behaviour of the virus, research into developing medicines and vaccines, epidemiological studies that examine how the disease spreads among the population, and computational fluid dynamics studies of how the virus is transmitted through the air in droplets released when we speak or cough.

Researchers answering the call had to show not only that they had a relevant scientific problem to solve, but also that they were ready with the right supercomputing methods to begin work right away. Some research areas were more ready than others. For instance, the epidemiological community was not really ready to run their models in a highly parallelised way on Tier-0 resources.

Evaluation was very swift – one of our conditions was that we made a decision to accept or reject a project within one week, so that the scientists and engineers would gain access to the computing resources as fast as possible. All of the knowledge gained through these projects has also been made publicly available to aid the global efforts in fighting this disease.

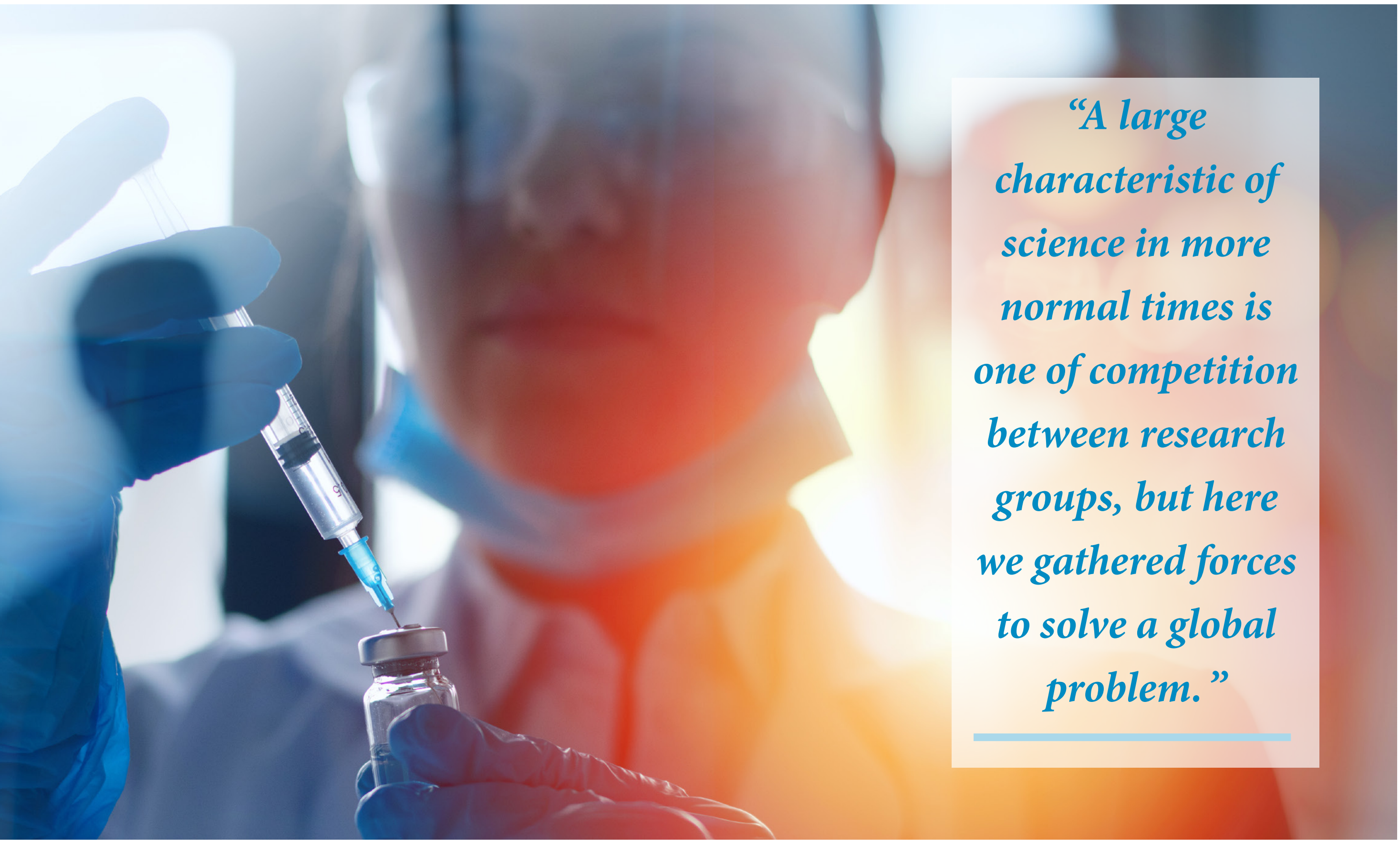
The call is still ongoing, although it will slowly be integrated into the Project Access Calls which are done twice per year by PRACE. It has been a lot of work for us and the scientific community at large, including the researchers and the computing centres. A lot has been asked from everyone and we were extremely happy with the way that people from all disciplines responded. It is not a normal situation for scientists

having to push through a proposal so quickly, and there was no drop or compromise in quality.

The success of this call has relied on the trust that we have built with the scientific community over many years. One of the main strengths of PRACE is the rigour of our application process and the high regard in which it is held among scientists.

This whole call has in essence been an exercise in urgent computing, and although we are prepared for this type of thing, a worldwide pandemic is not what we were expecting to have to deal with. We are proud to say that we have now awarded over 0.5 billion core hours through this call to around 35 projects, and the results of this work are coming in thick and fast.

Another thing we organised was in collaboration with our USA counterparts XSEDE. Monthly webinars with PIs from both sides were held in which updates from their respective projects were given, creating an environment for free exchange of information and ideas. A large characteristic of science in more normal times is one of competition between research groups, but here we gathered forces to solve a global problem. Modelling allows us to tackle problems that cannot be addressed through experimental work. For



example, glycolysation, which describes the forest of sugars that are found on the surface of all proteins, cannot be determined experimentally as the dynamics are too fast. Through our projects, we have helped to contribute by filling these gaps where experimental work falls short.

The fact that we have managed to create not just one but many vaccines for COVID-19 in the space of a year is nothing short of exceptional. Nothing was known about this virus a year ago. A whole community has now flourished out of this, which has brought together people from fields who would never have even communicated before.

I think a long-term effect of this will be that new solutions are found to other problems, so there are some positives to come out of this situation. The lessons we have learned will be invaluable in so many fields moving forwards.

The session at this year's EHPCSW aimed to bring together the scientists who have produced the most prominent results from research done through our COVID-19 call. We have tried to present an array of different topics to give people a taste of the breadth and depth of what we have achieved over the last year.

We hope that people who attended were able to get some ideas about how HPC can be used in so many different ways to address a problem.

Computing power is increasing at an incredible rate, and we are developing more and more tools that can work in tandem with this power. Artificial intelligence, although not so prominent at the beginning of this research into COVID-19 due to the lack of data, is likely to play an increasingly important role as we move forwards. Together with HPC, it is a powerful tool for addressing many of the problems that society faces today and in the future such as climate change.

The work on COVID-19 using HPC this year has reiterated to the community that one of the biggest challenges it faces moving forwards relates to software engineering, and the people who sit between expertise in computer science and domain-specific knowledge. At present, these people are either scientists who are really talented at computing, or coders who have a good feel for science. At present, there is no clear career path for these important roles, so there is a real need for us to put training and education in place for this as these people will be in high demand as the way we do science moves more and more towards exploiting computing power.

Executive Summary

www.prace-ri.eu