



# ARCHER2 Net Zero Case Study

Lorna Smith, Alan Simpson, Laura Moran, Andy Turner

EPCC

The University of Edinburgh



## Document Information and Version History

<b>Version:</b>	0.5
<b>Status</b>	Draft
<b>Author(s):</b>	Lorna Smith, Alan Simpson, Laura Moran, Andy Turner
<b>Reviewer(s)</b>	

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	2022-10-04	Initial draft	Lorna Smith
0.2		Sections on ARCHER2 and energy	Andy Turner
0.3	2022-10-24	Added sections and an appendix	Alan Simpson
0.4	2023-03-16	Added section on power and energy efficiency	Alan Simpson
0.5	2023-03-16	Added various sections and tidied up report	Lorna Smith
0.6			
0.7			
1.0			

## Executive Summary

UK Research and Innovation (UKRI) is committed to becoming net zero by 2040. This commitment includes UKRI's significant digital research infrastructure (DRI), which entails large-scale High Performance Computing (HPC) resources, data storage facilities and software libraries. To facilitate this, the DRI scoping project is investigating how UKRI can achieve net zero computing and covers all of the DRI infrastructure.

ARCHER2 is the UK's National HPC service, operated by EPCC, the University of Edinburgh, at the Advanced Computing Facility. ARCHER2 provides an invaluable resource to UK academics to deliver world-class research, including research into the impact of climate change. As part of the wider DRI scoping project, this case study has investigated the emissions associated with ARCHER2 and makes a series of recommendations to move towards Net Zero for large-scale facilities of this type.

## Recommendations

As a result of this case study, we have the following recommendations:

### Societal Value

- High utilisation is vital for efficient use of resources as, for example, idle power utilisation is high and embodied carbon is a significant fraction of the lifetime emissions.
- Service Providers should evidence the societal and/or economic value of their service, to be evidenced against the carbon impact of their service.

### Best Practice

- Based on the PRACE infrastructure workshops reports, a set of best practice recommendations have been made for all data centres. Data centres should test their services against these recommendations, which include: planning for future power requirements; efficient water cooling; re-use of waste heat; broad instrumentation; energy efficiency; and energy grid-friendly scheduling.

### Power and Energy Efficiency

- Vendors should be required to ensure that idling power is reduced significantly.
- Instrumentation is vital. This includes, for example, accurate per job data, all hardware components, data centre power supply and all cooling elements.
- Data centre construction should use environmentally aware builders and should begin to make estimates of the cost of carbon construction.
- Data centres should be required to use green energy.
- Some, but not all, codes may benefit from a reduction in CPU frequency. Correct decisions require good information (e.g., instrumentation).

### Energy Consumption at the User Level

- Data centres are custodians of significant power and should therefore actively engage in good-citizen behaviour. For example, in times of potential nationwide power shortages, Data Centres could reduce power utilisation or donate generated power to the grid.
- To guide users towards making better energy/carbon decisions, Service Providers should consider reporting to users the energy use of their jobs and implementing charging mechanisms that include a component of energy use.

- Better training should be provided for users on energy efficiency and carbon emissions. This could potentially build on the work of the Green Software Practitioner open-source training: <https://learn.greensoftware.foundation/>
- Researchers should be encouraged to report their positive environmental impact.

#### Commissioning and Decommissioning

- For future procurements, vendors should be required to provide high quality information on the carbon impact of the manufacturing and delivery of the hardware.

## ARCHER2: Purpose and System Description

### ARCHER2

ARCHER2 is the UK national supercomputing service funded by UKRI (via EPSRC and NERC) with the HPC platform provided by HPE and the hosting and support provided by EPCC at The University of Edinburgh. The ARCHER2 system is a liquid cooled HPE Cray EX with 5,860 compute nodes giving a total of 750,080 compute cores. ARCHER2 is one of the largest HPC systems without GPU accelerators in the world and sits at #25 in the June 2022 Top500 list of supercomputers. As well as the hardware platform, the ARCHER2 service provides a range of support for users and beyond, including a team of expert RSEs at EPCC to provide support for the research community, an extensive training programme and an outreach component. ARCHER2 is a general-purpose service supporting a wide range of research communities and software with a corresponding wide range of technical requirements and use cases.

#### ARCHER2 system hardware

##### *Compute nodes*

The ARCHER2 compute nodes are CPU only (i.e. they do not have any GPU accelerators available) with a total of 128 physical CPU cores per node. There are 5860 compute nodes, each with:

- 2x AMD EPYC™ 7742 64c 2.25 GHz processors
- 256/512 GB DDR4 memory
- 2x 100 Gbps Slingshot 10 interconnect interfaces (one per processor)

##### *Interconnect*

- HPE Cray Slingshot 10 - ethernet based, dynamic routing
- Dragonfly topology - maximum of 3 hops between any two compute nodes
- 768 switches in the whole system

##### *Storage*

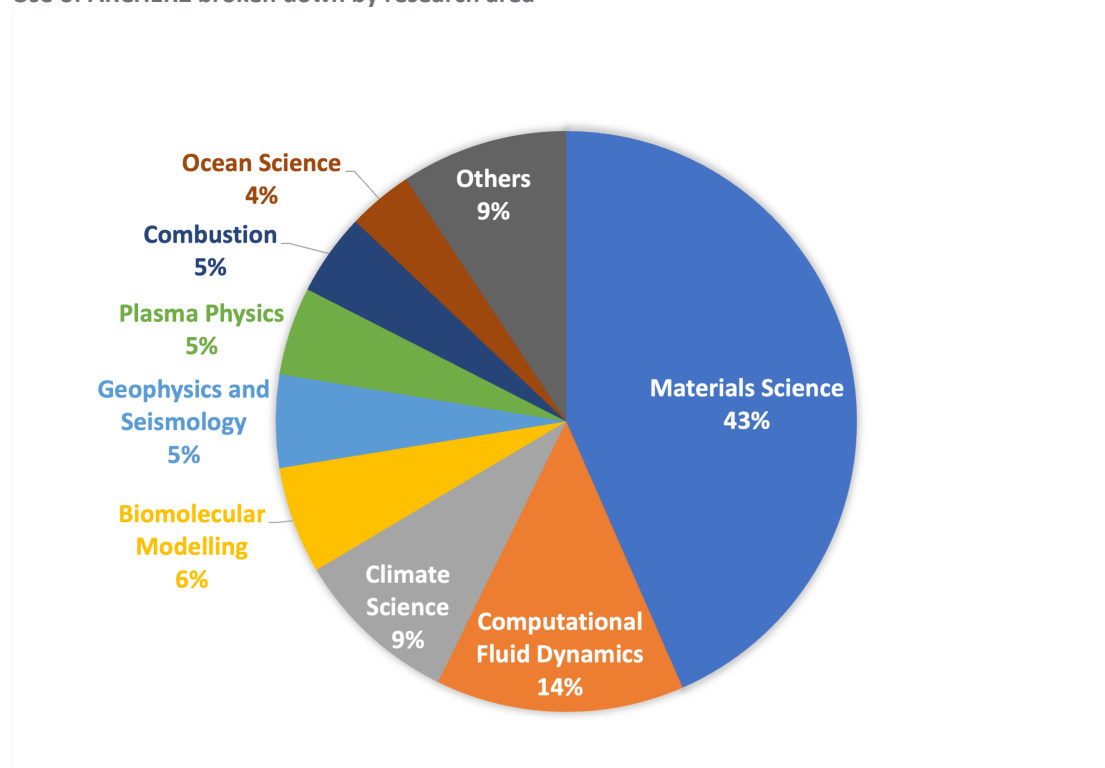
- 3x HPE ClusterStor L300 Lustre file systems, each 3.6 PB
- 1 PB HPE ClusterStor E1000 solid state storage
- 4x NetApp FAS8200A file systems, 1 PB total

### Scientific Endeavour on ARCHER2

ARCHER2 facilitates world-class science which in turn generates significant benefit to society and the economy. Modelling at this scale is only possible with systems such as ARCHER2 and it is important to recognise the essential benefit such systems provide, when considering the environmental impact of running such a system.

The value of ARCHER2 is evidenced by the very high utilisation of the system, averaging over 90% utilisation, and the large number of researchers, around 3000, utilising compute resources to further their scientific research. This is coupled with a broad spectrum of science enabled within the EPSRC and NERC remit, ranging from material science, through Computational Fluid Dynamics, to plasma physics and ocean science. Further, a significant proportion of the science enabled on ARCHER2 contributes directly to the global challenge of climate change, with, for example, research into the impact of climate change, extreme weather prediction, pollution reduction and improvements to renewal energy sources. The breakdown in scientific areas is shown in the Pi chart below.

Use of ARCHER2 broken down by research area



Finally, further evidence to the value of such a service comes from the scientific output itself. While the science on ARCHER2 is too extensive to evidence here, the following are recent highlights from the service.

### Coral reef connectivity in the southwestern Indian Ocean<sup>1</sup>

Researchers at the University of Oxford have used the power of ARCHER2 to simulate how billions of coral larvae are transported by ocean currents. Understanding the connectivity of coral reefs can help conservationists to identify the most vulnerable coral reefs and develop appropriate conservation measures. This research represents a major contribution towards coral reef conservation (amongst other applications) in the southwestern Indian Ocean, acts as a step towards tackling UN sustainable development goals 1, 2, 13 and 14 (<https://sdgs.un.org/goals>) and supports the United Kingdom's international development goals. This research was only possible thanks to the computational power of ARCHER2.

<sup>1</sup> Noam Vogt-Vincent (University of Oxford), April Burt (University of Oxford), Lindsay Turnbull (University of Oxford), Satoshi Mitarai (Okinawa Institute of Science and Technology), Helen Johnson (University of Oxford)

### Flow within and around a large wind farm<sup>2</sup>

Today, the need for renewable sources of energy is more urgent than ever. However, the clustering of turbines in existing wind farms leads to inefficiencies. Researchers at Imperial College London have used the power of ARCHER2 to perform one of the largest wind farm simulations to date to investigate this issue. The simulation enables the researchers to visualise and study the dynamics of the interacting turbine wakes in great detail, and provides insight into complex flow phenomena such as wake meandering, tip and hub vortex breakdown, and the interaction of the wind farm with the atmospheric boundary layer. It is hoped that the obtained insights will contribute towards maximising the efficiency of energy extraction from wind farms as these grow in size, number, and density.

### Predicting airborne pathogen spread indoors<sup>3</sup>

Whenever we exhale, we release tiny droplets which may contain pathogens. Researchers at the University of Birmingham have used ARCHER2 to model how the airborne transmission of these droplets is affected by factors such as ventilation and heating. Understanding this is key to making indoor spaces more resilient and preventing large-scale spread of diseases such as COVID-19, tuberculosis or measles. The findings from these simulations will contribute to addressing the current gap in knowledge on how airborne pathogens spread in different indoor scenarios, and how to best mitigate this with targeted evidence-based solutions for a particular indoor scenario to reduce risk of infection. Furthermore, the outcomes of this research could also be applied beyond COVID-19 to the spread and tracking of other indoor air pollutants.

### Recommendations

- High utilisation is vital for efficient use of resources as, for example, idle power utilisation is high and embodied carbon is a significant fraction of the lifetime emissions.
- Service Providers should evidence the societal and/or economic value of their service, to be evidenced against the carbon impact of their service.

## ARCHER2 in Context

This section compares ARCHER2 with other large HPC systems in terms of:

- Best Practice for HPC Data Centres
- Power Efficiency

### Best Practice for HPC Data Centres

PRACE (Partnership for Advanced Computing in Europe) is a funded by a series of significant EC awards to facilitate the access to a research infrastructure that enables high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE has 25 member countries and EPCC leads the UK activity. As part of the EC-funded work, PRACE organises an annual series of European Workshops on HPC Infrastructures (EWHPC) that aims to bring together specialists in High-Performance Computing (HPC) centre design and operation from around the world to discuss the latest trends in HPC infrastructure and supporting technologies for supercomputing centres. More details of this workshop series can be found in

<sup>2</sup> N. Bempedelis & S. Laizet (Imperial College London)

<sup>3</sup> Aleksandra Monka, University of Birmingham Bruño Fraga, University of Birmingham

*Appendix: PRACE Infrastructure Workshops*, including summaries of the best practices identified at recent workshops.

Recent PRACE Infrastructure Workshops stressed the importance of sustainability for data centres and highlighted the following key areas of best practice:

- Planning for Future Power Requirements
- Efficient Water Cooling
- Re-Use of Waste Heat
- Broad Instrumentation
- Energy Efficiency
- Energy Grid-Friendly Scheduling

The ARCHER2 Service has made progress to address each of these, as outlined below.

### Planning for Future Power Requirements

We have liaised with SPEN to ensure that a new electrical substation has been built on the ACF site. This provides 30MW of power to a brand-new, energy-efficient computer room (CR4), whereas ARCHER2 is currently housed in a 4MW computer room (CR3). We believe that the ACF now has access to enough power (and other infrastructure) to host a future Exascale system for the UK.

### Efficient Water Cooling

ARCHER2 and EPCC's other major systems have used water cooling for many years, and the ACF has significant infrastructure (pumps and chillers) for water cooling. Due to the cool climate in Edinburgh, for much of the year, we are able to pump warm water on to the roof to allow "free-cooling" to the atmosphere and minimise chiller overheads. We are currently working with HPE to increase the inlet water temperature further to maximise free cooling.

### Re-Use of Waste Heat

It is significantly more difficult to re-use waste heat in countries like the UK where there is little district infrastructure for heating. Nevertheless, EPCC has ambitious plans for re-using the waste heat from the ACF via an abandoned mine system, allowing the energy to be accessed via commercial and domestic heat pump technology. If successfully implemented the concept has the potential to provide low carbon, resilient, low-cost heating that is sustainable both in terms of heat pump performance and the shallow geothermal resource. More details are available in: **"The Geobattery Concept: A Geothermal Circular Heat Network for the Sustainable Development of Near Surface Low Enthalpy Geothermal Energy to Decarbonise Heating"** Andrew Fraser-Harris, Christopher Ian McDermott, Mylène Receveur, Julien Mouli-Castillo, Fiona Todd, Alexis Cartwright-Taylo, Andrew Gunning, Mark Parsons (<https://doi.org/10.3389/esss.2022.10047>).

### Broad Instrumentation

EPCC now collects significant information about ARCHER2 and the associated infrastructure. This is collated and visualised on a Grafana dashboard. The information includes: system load, total power, power per cabinet, user login sessions, down nodes, service outages, and local fuel split. This allows us to, for example, measure the power usage of individual jobs and link that to local fuel split at that time. This instrumentation work was given as a presentation and paper at a recent Cray User Group (NERSC, May 2022): **"Automated service monitoring in the deployment of ARCHER2,"** Kieran Leach, Philip Cass, Steven Robson, Eimantas Kazakevicius, Martin Lafferty, Andrew Turner, Alan Simpson.

### Energy Efficiency

As part of this case study, EPCC has investigated on ARCHER2 the impact of processor frequency on total energy usage and performance of jobs. This investigation indicated that, for many codes, a reduction in processor frequency (2.25 GHz --> 2GHz) had a significant reduction in energy used (15-20%). Due to this positive result, we changed the default processor frequency in December 2022. Initial results showed that the overall power draw of ARCHER2 reduced from 3.1MW to 2.5 MW. While a more detailed analysis is ongoing, these results seem to indicate a significant improvement in energy efficiency.

### Energy Grid-Friendly Scheduling

As part of this case study, and an associated sandpit project (HPC-JEEP), EPCC has also calculated the energy use of different software applications and scientific areas. This allows a variety of different charging strategies that can encourage users to consider the energy usage of their research (see **“HPC-JEEP: Energy Charging on ARCHER2 and the DiRAC COSMA HPC services”** *Alastair Basden, Andy Turner*). Being able to predict the likely power usage of jobs in the queue will allow us to have control over power usage and we are investigating scheduling policies that, e.g., reduce power usage at peak times or when the local fuel mix is higher in carbon emissions.

### Recommendations

- Based on the PRACE infrastructure workshops reports, a set of best practice recommendations have been made for all data centres. Data centres should test their services against these recommendations, which include: planning for future power requirements; efficient water cooling; re-use of waste heat; broad instrumentation; energy efficiency; and energy grid-friendly scheduling.

### Power Efficiency

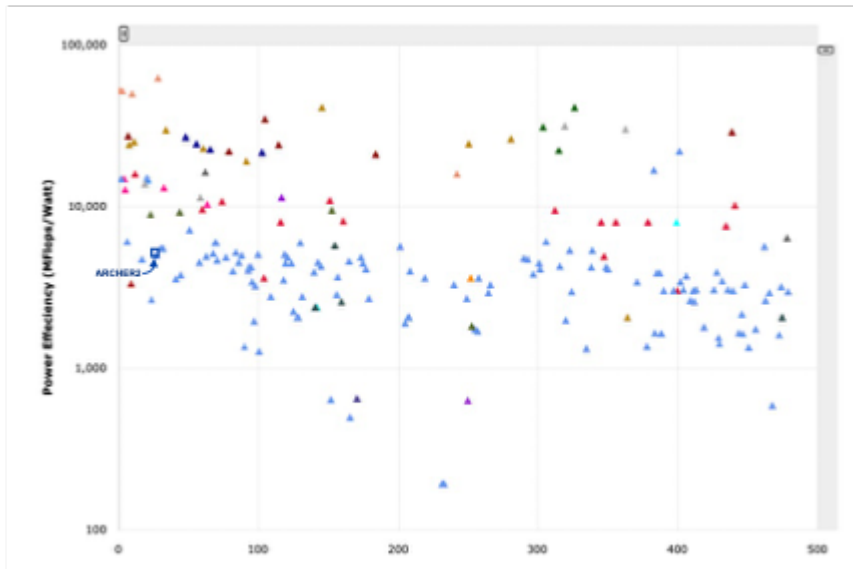
In this section, we compare ARCHER2 with other large HPC systems in terms of performance and power draw. The data comes primarily from the Top500 website (<https://www.top500.org/>) that collates information on the largest supercomputers in the world. The Green500 list on that site (<https://www.top500.org/lists/green500/>) ranks the major HPC systems by their Power Efficiency in GFlops/watt.

The ARCHER2 HPL run that was submitted to the Top500 list was run as part of the early testing process and ran for a much shorter time than usual (<2 hours), and only ran on 5600 nodes (out of 5860). This means that the measured performance was smaller than could have been achieved and that the Power Efficiency was also smaller. Nevertheless, ARCHER2 did achieve 19.54 GFlops to put it at Number 25 on the June 200 list. This run took a maximum of between 3.8 and 3.9 MW. Using even the higher power figure, gives a Power Efficiency of just over 5.0 GFlops/watt.

ARCHER2's rating of 5.0 GFlops/watt would put it at around number 75 on the Green500 list. Of the Top 25 systems on that list, 24 of them are GPU based, and the other system uses a custom processor.

The graph below plots the Power Efficiency of the Top500 systems (in MFlops/watt).





The systems are grouped by accelerator/co-processor: the CPU-only systems are blue triangles, while the various accelerators/co-processors are other colours of triangles. ARCHER2's position is marked with a dark blue square.

The graph indicates that most CPU-based systems have Power Efficiencies of between about 2 GFlops/watt and 6 GFlops/watt and that nearly all the systems that have more than 6 GFlops/watt have accelerators/co-processors (in fact, mostly GPUs). Even with the restrictions previously mentioned for the HPL run, ARCHER2's Power Efficiency is better than most other CPU-based systems.

## Power and energy efficiency of ARCHER2

This section considers power and energy efficiency of ARCHER2. In particular:

- the power draw of the different components of ARCHER2;
- the infrastructure of the Advanced Computing Facility (ACF) and how this impacts on the energy efficiency of ARCHER2;
- Improvements in power efficiency implemented during this case study;
- power and energy efficiency of application codes running on ARCHER2.

### ARCHER2: Power Draw by Component

Based on data from HPE, we have the following estimates of power draw for components:

- Slingshot interconnect switches: approx. 700 W per switch, 768 switches in the system, 540 kW in total
- ClusterStor Lustre file systems: approx. 8 kW per file system, 5 file systems in system, 40 kW in total
- Cooling Distribution Units (CDU): 16 kW per CDU, 6 CDU in the system, 96 kW in total

Given that each compute node has an estimated maximum power draw of around 700 W and there are 5860 compute nodes, this gives a total maximum power draw for the compute nodes of over 4,000 kW.

This, in turn, gives an estimated maximum power draw for the ARCHER2 system of over 4,500 kW with the following rough percentage breakdown by component:

- Compute nodes: 89%
- Slingshot interconnect switches: 12%
- CDU: 2%
- ClusterStor Lustre file systems: <1%

This breakdown indicates that the node energy use is by far the most significant component of energy use for jobs on the ARCHER2 system. For large jobs there may be some contribution from interconnect use, but file system energy use will likely be a negligible component for all jobs.

One observation from ARCHER2 is that the compute node idle power draw is a large fraction of the peak power draw. An idle compute node on ARCHER2 typically draws 200 to 250 W (around 30% of the 700 W peak power draw). The processors typically draw around 40 W on an idle node and the memory around 115 W. We assume that the remaining idle power draw (around 100 W) is mostly due to the Slingshot interconnect cards. On ARCHER2, this does not lead to issues with excess energy use as the utilisation is so high (over 90%) but, if the system was less busy, this would lead to a large power draw even if the system was not being used. The ARCHER2 service is working with HPE to understand how this high idle compute node power draw can be reduced. The majority of the difference between idle and active power use on an ARCHER2 compute node is the change in power draw of the processors: each processor has an idle power draw of around 20 W and, when loaded, can draw up to 225 W<sup>4</sup>. Typical breakdown of power draw by node component for a fully powered node:

- Processors: 450 W (64%)
- Memory: 120 W (17%)
- Interconnect NIC: 100 W (14%)

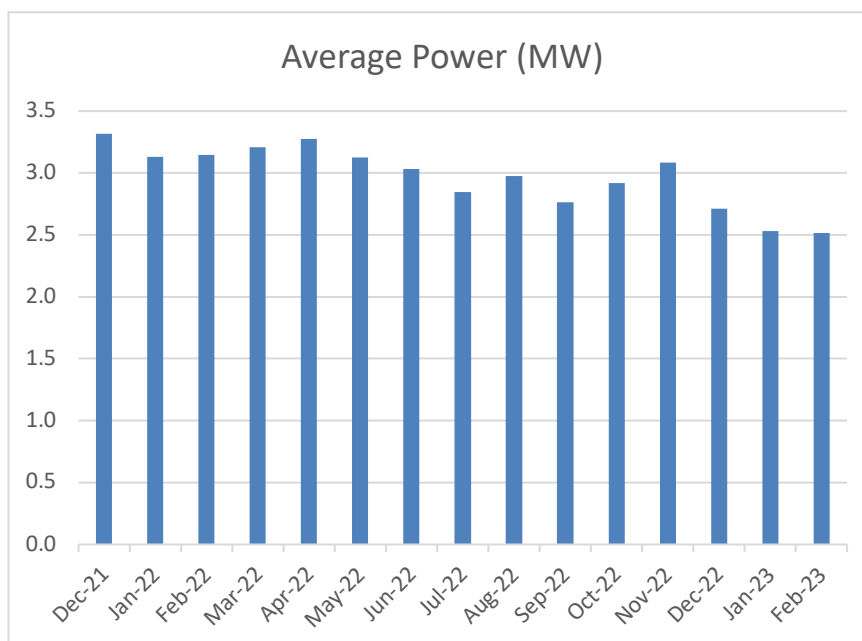
Of course, the actual power draw at any one time will vary by the software being used – typically variation in the power draw of the processor determines the variation in the power draw of individual nodes.

### Monitoring power draw/energy use of ARCHER2

We can see that software on ARCHER2 does not generally draw the maximum possible power on compute nodes as the mean power draw of the ARCHER2 compute cabinets since it came into service in November 2021 is under 3.0 MW. The plot below shows the power draw of the ARCHER2 compute cabinets per month.

.

<sup>4</sup> <https://www.amd.com/en/products/cpu/amd-epyc-7742>



Changes in power draw over the lifetime of the service are due to changes that have been made to improve the energy efficiency and reduce the power draw. These changes are described in more detail in the *Improvements in Power Efficiency* section below.

### Monitoring framework

Through our monitoring framework we can monitor the power draw or energy use of the ARCHER2 compute provision at several different levels:

- Incoming power: meters on power feeds from the national grid
  - Captured and stored by the University of Edinburgh Estates and Buildings Department
  - Meter data is captured and logged every 30 mins
- Power draw of individual ARCHER2 compute node groups
  - Captured from ARCHER2 compute cabinet PDU monitoring and stored in EPCC's Graphite monitoring database
  - Visualised using the Grafana monitoring tool
  - Power draw log resolution is 1 minute
- Total compute node energy use on a per-job basis
  - Captured by the Slurm scheduler
  - Analysis of this data is being performed as part of the HPC-JEEP activity<sup>5</sup> within the UKRI DRI Net Zero Scoping Project

### Recommendations

- Vendors should be required to ensure that idling power is reduced significantly.
- Instrumentation is vital. This includes, for example, accurate per job data, all hardware components, data centre power supply and all cooling elements.

<sup>5</sup> <https://net-zero-dri.ceda.ac.uk/hpc-jeep/>

## Advanced Computing Facility (ACF)

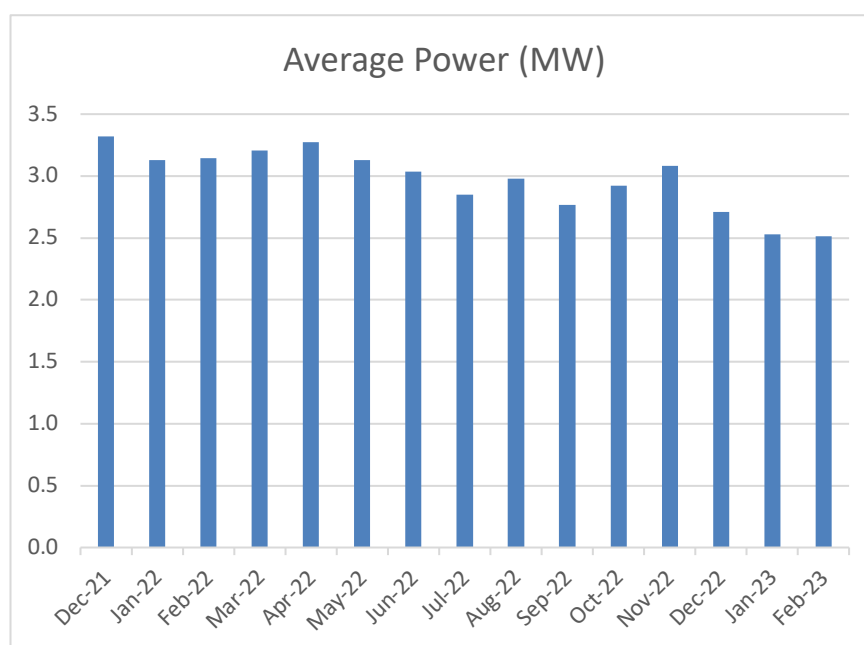
ARCHER2 is housed in the University of Edinburgh's Advanced Computer Facility (ACF) at the Bush Estate on the outskirts of Edinburgh. The ACF is specifically designed to house large supercomputers securely and efficiently and is powered via a 100% green, renewable energy tariff.

ARCHER2 is housed in the ACF Computer Room 3 (CR3) and is cooled efficiently via direct liquid cooling. The cooling system for CR3 is designed to take advantage of free-to-air cooling as much as possible. In free-to-air cooling, warm water is pumped up to the roof of the ACF building where the heat is exchanged with the cooler outside air with cooled water returned to extract heat from the ARCHER2 system. Air temperature in Edinburgh is generally lower than the UK average, so ARCHER2 is able to benefit from more of this energy efficient means of cooling than would be possible in warmer locations. When the outside air temperature is too warm to cool the water in the CR3 cooling loop on its own, we make use of chiller units to remove the excess temperature. The ability to gradually employ chiller units while still making use of a proportion of free-to-air cooling means we cool water efficiently and minimise the use of chiller units. CR3 dynamically responds automatically to the load on ARCHER2 (the amount of heat being generated) and the external atmospheric conditions to adjust cooling water flow and chiller operation to try to maximise the efficiency of the cooling plant supporting the system. The improvements in power efficiency discussed below also reduce the cooling load and allow CR3 to operate even more efficiently.

## Improvements in Power Efficiency

Over the first 12-18 months of the ARCHER2 service, EPCC have implemented a number of changes targeted at reducing the power draw of ARCHER2 and increasing the energy efficiency of applications. These include:

- Moving from High Performance Mode to Performance Determinism Mode: this should reduce the power draw with negligible impact on the performance of large jobs [May 2022]
- Reduced default processor frequency: this should again significantly reduce the average power draw [December 2022]
- Increase in coolant temperatures: we are currently progressing a reconfiguration of ARCHER2 to increase the normal operating temperature of the facility water and so increase the use of free cooling, with an expected consequent reduction in power costs. This is ongoing and should show benefits from 2Q2023.



The above graph shows the impact of the first 2 changes on the average power draw of ARCHER2. As can be seen the power reduced in May 2022 from around 3.22 MW to 2.94 MW, and then again in December 2022 to 2.53 MW. This is an overall reduction in power usage of more than 20%. More details of these changes, including their impact on applications are given in the next subsection.

## Power and energy efficiency of applications

The most useful way to consider the efficiency of a system, such as ARCHER2, is to consider the power and energy efficiency of applications running on the system as these reflect normal usage. When UKRI procured ARCHER2, the performance was evaluated using a set of applications benchmarks agreed with key representatives of the user community. EPCC has subsequently been using a suitable subset of these benchmarks to monitor the applications performance, power and energy use of the ARCHER2 system. Initially we generally only re-ran the benchmarks whenever a significant change was made to the system. More recently, we have begun to run the tests on a regular basis, to help identify unexpected changes in the operation of the system.

The ARCHER2 system has two different modes for processor power: High Performance mode (where each processor uses the highest power it can) and Performance Determinism mode (where the processors each use the same power setting). The ARCHER2 system was supplied configured in High Performance mode. Based on discussions with HPE, it seemed likely that this may not be optimal for energy usage for large jobs and so we implemented a change to Performance Determinism mode in May 2022. We evaluated the change using benchmarks on multiple nodes and were able to demonstrate a 6-11% reduction in energy used with negligible change in performance. Overall, ARCHER2's power usage reduced from about 3.22 MW to 2.94 MW which is a 9% improvement in both power and energy efficiency.

The job scheduler (Slurm) on ARCHER2 provides the ability for each job to choose its own frequency for the processors to be 2.25 GHz, 2.0 GHz or 1.5 GHz which should also have a significant impact on overall power draw of the system. We investigated the performance and energy use of a number of benchmarks as the processor frequency was changed. Many high-use applications are not CPU-bound, and so for the majority of benchmarks, it was possible to reduce the requested processor frequency from 2.25 GHz to 2.0 GHz with a significant saving in both energy and power and only a more modest reduction in performance (reducing further to 1.5 GHz was not beneficial).

For nearly all applications, the reduction in processor frequency led to a significant reduction in energy-to-solution of 8-21%. While the benchmarks did generally take longer to complete, for many applications the increase in residency time was only a few percent. Given the concerns over the UK's 2022-2023 Winter power supply, significantly reducing the overall power draw of ARCHER2 was a responsible course of action. More details on this work are available in the *Energy Consumption at the User Level* section below.

In December 2022, we therefore changed the default processor frequency from 2.25 GHz to 2.0 GHz. For applications which did not see benefits to this reduction, we reset the default processor frequency to 2.25 GHz in their corresponding module. All users were given simple instructions in how to set the processor frequency for their jobs and were encouraged to test out the change for their own application. This has resulted in an overall reduction in power from about 2.94 MW to 2.53 MW, i.e., around 14% lower power draw. A further, more detailed analysis of the impact of this change will be completed in April 2023.

In summary, during the first 15 months of the ARCHER2 Service, we introduced two significant changes that have together seen a reduction in power draw (and hence energy costs) of over 20%, and a corresponding improvement in research output per kWh of around 15%. Planned future work includes:

- increasing the facility coolant temperature to increase free cooling;

- analysing the impact of the change in default processor frequency.

### Recommendations

- Data centre construction should use environmentally aware builders and should begin to make estimates of the cost of carbon construction.
- Data centres should be required to use green energy.
- Some, but not all, codes may benefit from a reduction in CPU frequency. Correct decisions require good information (e.g., instrumentation).

## Energy consumption at the user level

### Impact of CPU frequency on application energy use and performance

Frequency is one of the key factors that control the rate at which the CPU can perform operations. Typically, the higher the frequency the CPU is running at, the more times it can execute (potentially multiple or parallel) operations per second and this can lead to higher performance for applications - particularly those that are critically dependent on floating point performance. However, the higher the CPU frequency, the more power it draws and so it consumes more energy per second. This means that part of the energy consumption of an HPC job is determined by the ratio between runtime (or performance) of the application and power draw (which can be controlled by setting CPU frequency). As the relationship of CPU frequency to application performance and CPU frequency to power draw may not follow the same trends, it may be possible to reduce the power draw by a certain amount while reducing the performance by a smaller amount and this could lead to an overall energy consumption saving for a particular job (even though the job itself may run for longer than at a higher frequency).

There are a number of benefits and potential benefits to improving the energy efficiency of the service. In terms of energy consumption, the primary benefit is that it reduces the overall cost of the ARCHER2 service - using less energy leads to reducing the energy bills for running ARCHER2 per amount of research output. A reduction in power draw of ARCHER2 makes the service more friendly to the UK National Grid, allowing the additional power capacity to be redistributed for other uses - a consideration that is particularly important this Winter where there are concerns that there may be periods where the UK National Grid capacity will be under severe pressure. As custodians of significant amounts of power we see this, and other forms of good citizen behaviour, as key to good data centre management. Reducing the power draw to the processors also has the potential to reduce the cooling overheads for the service as a whole which both reduces energy use by the service further and makes the service more resilient to high air temperatures such as the extreme heat events we saw during Summer 2022.

To estimate the potential impact on performance and energy use of changes to the ARCHER2 CPU frequency, the CSE team ran a number of application benchmarks with varying CPU frequencies and measured the performance (as reported by the applications) and the total compute node energy use as recorded by Slurm in the ConsumedEnergyRaw job property (the energy use data in Slurm comes from the pm\_counters plugin). We also recommended that all users examine the variation of energy use with CPU frequency for their jobs so they can choose the right setting for their applications. The processors on ARCHER2 support three frequency settings: 1.5 GHz, 2.0 GHz and 2.25 GHz. We used the `--cpu-freq` Slurm option to control the processor frequencies as described in the ARCHER2 User Guide<sup>6</sup>.

<sup>6</sup> <https://docs.archer2.ac.uk/user-guide/energy/#controlling-cpu-frequency>

Once we had the raw measurements for each of application benchmark, we computed the mean performance and energy consumption at 2.25 GHz and then calculated the relative performance and energy consumption for each benchmark run so we could plot relative performance against relative energy consumption for each of the CPU frequency settings.

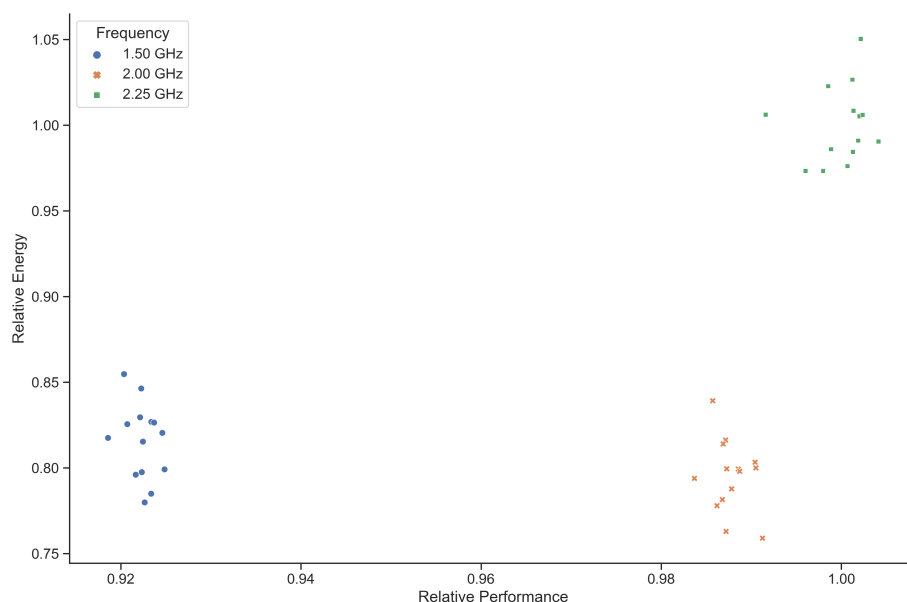
We ran four different application benchmarks:

- VASP  $\text{TiO}_2^7$  - using the ARCHER2 central VASP 6.3.1 installation
- OpenSBLI Taylor-Green Vortex 512<sup>3</sup> strong scaling<sup>8</sup> - compiled using CCE 11
- GROMACS 1400k atoms<sup>9</sup> - using the ARCHER2 central GROMACS 2021.3 installation
- CASTEP Al3x3 (Al Slab)<sup>10</sup> - using the ARCHER2 central CASTEP 21.1.1 installation

All benchmarks were run as pure MPI using 128 MPI processes per node. We ran using both 1 node (128 MPI processes) and 4 nodes (512 MPI processes).

The plots below show the results of the single node benchmark runs for each of the applications - each plot shows relative performance against relative total energy use. The performance and energy are relative to the mean values at 2.25 GHz.

#### Relative performance against relative total energy use for the VASP $\text{TiO}_2$ benchmark



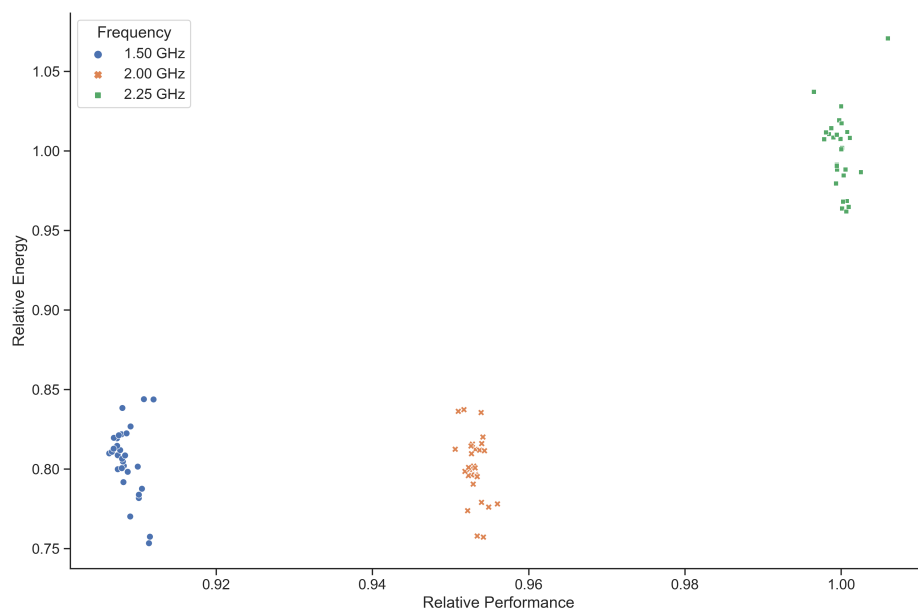
#### Relative performance against relative total energy use for the OpenSBLI TGV512ss benchmark

<sup>7</sup> <https://github.com/hpc-uk/archer-benchmarks/tree/main/others/VASP>

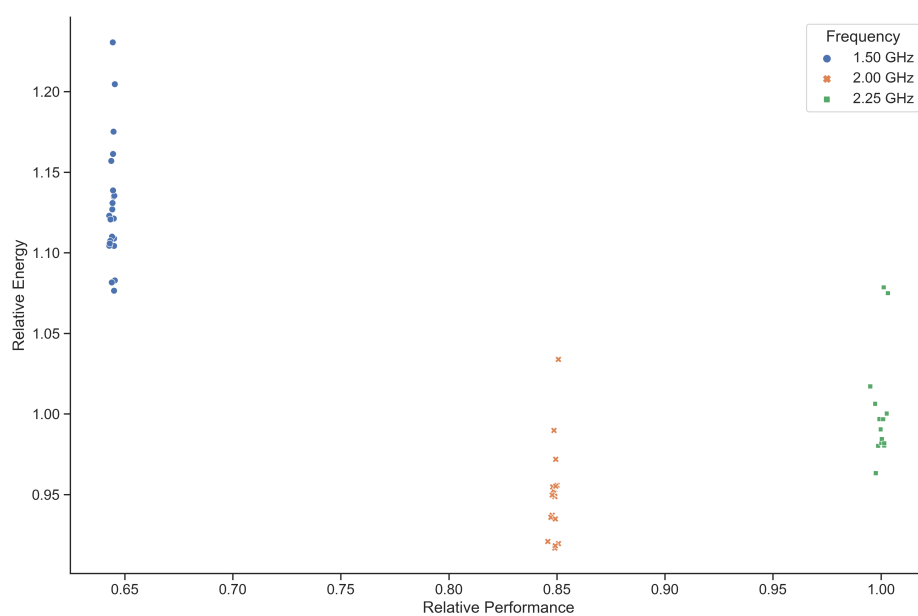
<sup>8</sup> <https://github.com/hpc-uk/archer-benchmarks/tree/main/apps/OpenSBLI>

<sup>9</sup> <https://github.com/hpc-uk/archer-benchmarks/tree/main/apps/GROMACS>

<sup>10</sup> <https://github.com/hpc-uk/archer-benchmarks/tree/main/apps/CASTEP>

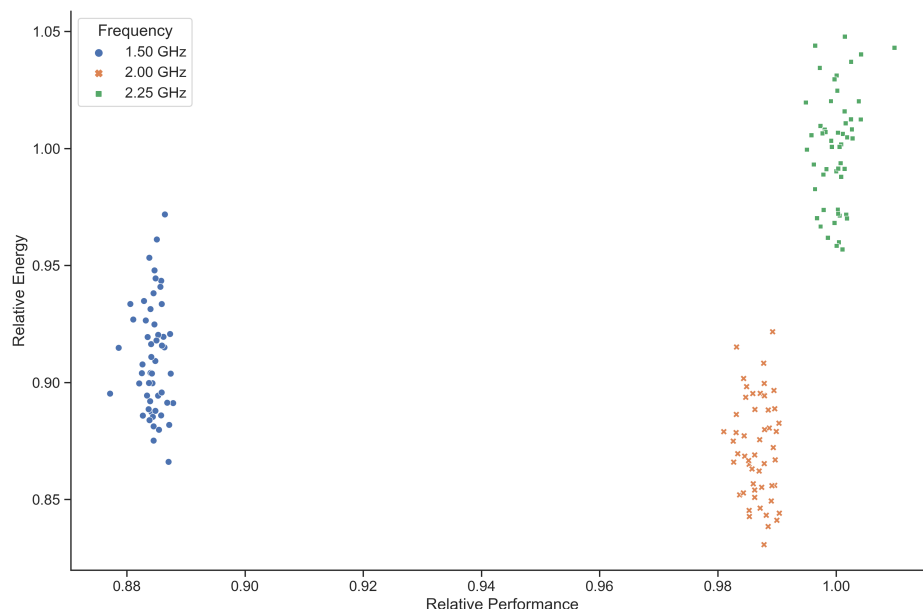


Relative performance against relative total energy use for the GROMACS 1400k benchmark



Relative performance against relative total energy use for the CASTEP Al3x3 benchmark





From the plots we can see that, for all of the benchmarks, the energy consumed reduces as we move from the default, highest (2.25 GHz) CPU frequency to 2.0 GHz. A further reduction to 1.5 GHz leads to an increase in energy consumption from 2.0 GHz. It is clear that for these benchmarks, running at 2.0 GHz minimises the energy consumption.

The table below summarises the benchmark results when moving from 2.25 GHz to 2.0 GHz - given by the mean relative energy use/relative performance at 2.0 GHz compared to the values at 2.25 GHz:

Benchmark	Relative Energy Use	Relative Performance
VASP TiO <sub>2</sub>	-20%	-1%
OpenSBLI TGV512ss	-20%	-5%
GROMACS 1400k	-5%	-15%
CASTEP Al3x3	-13%	-1%

## Enabling users to monitor the energy use of their jobs

To assist users in assessing the energy efficiency of their jobs on ARCHER2, we have documented how user can extract their per job energy use of the service in the ARCHER2 User Guide<sup>11</sup>. As an extension to this work, we are looking at way we can provide users and projects with information on their aggregate energy use of the service and place it into context against the electricity use of a typical UK household. For example, we are working on functionality that will provide individual users with a report similar to the one below on a monthly basis (this example is for the single largest energy user in November 2022 so most users will use much less energy than this example):

Total energy consumed: 31226 kWh

This is the equivalent to the monthly electricity use of 129 typical UK households

Of this, 98 kWh (0.3%) corresponds to energy from failed jobs

You are in the 90-100th percentile range of ARCHER2 users in terms of energy used in the period

Job size distribution (cores):

	Minimum	Q1	Median	Q3	Maximum
All users	1	512	1600	6464	131072

<sup>11</sup> <https://docs.archer2.ac.uk/user-guide/energy/>

	user1	128	1024	2048	4096	12288
Mean node power distribution (W):						
		Minimum	Q1	Median	Q3	Maximum
All users		0	418	484	521	834
	user1	0	499	517	529	611

## Charging Strategies for HPC Services

Historically, nearly all HPC services have charged based on usage of compute time, or more accurately, residency of compute resource. So, a job running on 128 cores for 2 hours, would be charged for 256 core-hours. Moreover, there is a general aim to apportion the total cost of the system across all jobs run on it. So, a base notional cost for each core hour could then be calculated as the total cost of the HPC system (capital plus running costs) divided by the total number of core-hours that it will deliver. This approach is reasonable if the capital cost is much greater than the cost of the energy used. While this may have been true in the past, it is not the case for most modern HPC systems.

Fortunately, many systems now provide better records of energy used per job, and so it is possible to incorporate the energy used into the charge of the job. This is valuable as it increases user awareness of the energy their job uses and then it is more likely that users will plan their research in a more energy-aware manner. Different strategies include:

1. Charge solely on residency but report the energy used as well to increase awareness.
2. Estimate the actual monetary cost of the job based on both residency and energy use.
3. Charge solely on energy usage to increase the incentivisation for users to minimise their energy usage. In this approach, the capital costs are considered as already committed.

Using the terminology of *HPC-JEEP: Energy-based charging on the ARCHER2 HPC service* (Alastair Basden, Andy Turner), the general formula for the calculation of the charge for a job based on the class of models we are considering is:

$$C_T = wC_E + (1 - w)C_R$$

In this formula:

- $w$  is the weight: a value in the interval [0,1] that defines the relative weight of charging for energy use of the job to the residency of the job
- $C_E$  is the computed charge for the energy use of the job - proportional to the energy use (e.g in kWh, J)
- $C_R$  is the computed charge for the residency of the job - proportional to the residency (e.g. in node-h, coreh-h, GPU-h)

With  $w=0$ , this is the first strategy above, i.e., charging solely on residency. With  $w=1$ , the charges are entirely based on energy usage (strategy 3). For a system like ARCHER2, the total energy could be a third of the total cost of the system. In that case, we could use  $w=0.33$ .

In summary, energy costs are an increasing fraction of the total cost of ownership of a system. It is therefore important to make users aware of their energy usage. Moreover, it is relatively straightforward to incorporate the energy usage of a job into charging strategies that can focus primarily on energy usage or can reflect the actual monetary cost of the job.

### Recommendations

- Data centres are custodians of significant power and should therefore actively engage in good-citizen behaviour. For example, in times of potential nationwide power shortages, Data Centres could reduce power utilisation or donate generated power to the grid.
- To guide users towards making better energy/carbon decisions, Service Providers should consider reporting to users the energy use of their jobs and implementing charging mechanisms that include a component of energy use.
- Better training should be provided for users on energy efficiency and carbon emissions. This could potentially build on the work of the Green Software Practitioner open-source training: <https://learn.greensoftware.foundation/>
- Researchers should be encouraged to report their positive environmental impact.

## Commissioning and Decommissioning ARCHER2

Various authors (for example<sup>12</sup>) have raised the importance of emissions associated with the manufacture and delivery of large-scale HPC resources. The HPC-Jeep sandpit project<sup>13</sup> also investigated this area of the lifecycle, estimating that as much as 40% of the lifetime emissions of the system come from their manufacture and delivery. Attempts within this case study to understand the emissions associated with the current ARCHER2 system have failed, due to a lack of available data from the vendor, HPE.

There is a however growing recognition that this aspect of the service is essential when looking to understand the overall emissions of a service, and a corresponding recognition from vendors that this sort of information will be a requirement in future procurements.

Emissions from the decommissioning of systems cannot be ignored. Waste generates 3.3% of global greenhouse gas emissions<sup>14</sup> and is therefore a crucial part of delivering a Net Zero service. Striving for zero waste during decommissioning is key, eliminating emissions from waste as well as saving many precious materials through re-use and recycling. The University of Edinburgh takes the following approach to the decommissioning of services:

- where possible, components are re-use within the ACF;
- where possible, components are then offered for re-use to appropriate suppliers and third-parties;
- remaining components are re-used and recycled via a specialist WEEE recycling service, with a commitment to maximising re-use and recycling while maintaining appropriate security.<sup>15</sup>

### Recommendations

- For future procurements, vendors should be required to provide high quality information on the carbon impact of the manufacturing and delivery of the hardware.

## Conclusions

<sup>12</sup> Chasing Carbon: The Elusive Environmental Footprint of Computing, DOI 10.1109/MM.2022.3163226, IEEE Micro

<sup>13</sup> <https://net-zero-dri.ceda.ac.uk/hpc-jeep/>

<sup>14</sup> <https://www.weforum.org/agenda/2022/11/waste-emissions-methane-cities/>

<sup>15</sup> <https://www.cclnorth.com>

This case study of the ARCHER2 system has investigated the emissions associated with ARCHER2 and makes a series of recommendations to move towards Net Zero for large-scale facilities of this type. The review includes a comparison of the service against best practice in the area, consideration of the data centre infrastructure, examples of actions taken to improve energy efficiency options to guide users towards making better energy/carbon decisions and consideration of the role of manufacture, transportation and decommissioning.

## Appendices

### Appendix: PRACE Infrastructure Workshops -- Best Practice

The mission of PRACE (Partnership for Advanced Computing in Europe) is to facilitate the access to a research infrastructure that enables high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE seeks to realise this mission by offering world class computing and data management resources and services through a peer review. PRACE also seeks to strengthen the European users of HPC in industry through various initiatives. PRACE has a strong interest in improving energy efficiency of computing systems and reducing their environmental impact.

PRACE has 25 member countries (including the UK) whose representative organisations create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level. PRACE has had more than €125M funding from the European Commission (EC).

PRACE undertakes software and hardware technology initiatives with the goal of preparing for changes in technologies used in the Research Infrastructure and provide the proper tools, education and training for the user communities to adapt to those changes. One goal of these initiatives is to reduce the life-time cost of systems and their operations, in particular the energy consumption of systems and the environmental impact.

PRACE organises an annual series of European Workshops on HPC Infrastructures (EWHPC) that aims to bring together specialists in High-Performance Computing (HPC) centre design and operation from around the world to discuss the latest trends in HPC infrastructure and supporting technologies for supercomputing centres. The workshops cover a broad range of topics, ranging from cooling technologies to the provisioning of electricity, and from overviews of standards and certification to the management, storage, and analyses of the large streams of heterogeneous metrics that are crucial for the efficient operation of modern HPC datacentres. In the rest of this section, we summarise the main findings from two recent workshops. Detailed reports from the workshops are available at: <https://prace-ri.eu/infrastructure-support/european-workshops-on-hpc-infrastructures/>. For up-to-date information about future workshops, please refer to the EWHPC official website: <https://www.euhpcinfrastructureworkshop.org/>.

#### HPC Infrastructure Workshop 2022

EWHPC12 was held as a hybrid meeting on 7-9 June 2022, hosted by CSC in Finland. There were more than 90 participants (2/3 in person) from 33 different HPC sites across Europe, US, Japan and Australia. Given that Data centres and the rest of the computing sector account for between 5% and 9% of the total use of electricity in Europe, the workshop focused on energy efficiency, sustainability and carbon footprint.

Key technical topics covered included:

- Sustainability of supercomputing centres
- Quantum computing (installation and operation)
- Cooling efficiency
- Energy efficiency improvement
- Procurement strategy to address energy efficiency and sustainability

The workshop identified the following lessons learned and key trends:

- Sustainability is a major concern for sites of all sizes as well as for vendors. Reducing the carbon footprint (including not only the operation phase but the life cycle as a whole) as well as heat reuse when possible are the main actions taken toward this goal.
- Direct liquid cooling (DLC), most of the time in a chiller-less mode, contributes to reducing the power consumption. This cooling technique is now routinely implemented in most HPC centres regardless of their size.
- Increased inlet temperature does not seem to have a negative impact on the reliability of components and makes chiller-less cooling and heat reuse easier. However, there is a risk that the reduction of maximum T-case announced by the technology suppliers for future generations of compute components would lead to the necessity of keeping the inlet temperature lower than anticipated.
- The high price of electricity has a major impact on HPC centres. Solutions to keep the electricity bill within acceptable limits include power capping and, when no other solution can be found, shutting down part of the compute resources. Advanced power capping solutions, aiming at minimizing the impact of power capping on users, are actively developed.
- In order to reduce the cost and delay of HPC centre infrastructure construction or upgrade while avoiding over sizing solutions such as modular architecture and brown field installation are being implemented in HPC centres in Europe.

## HPC Infrastructure Workshop 2021

This was held online on 18-19 May 2021 and had more than 100 participants from 53 different sites. Conclusions were drawn in a number of areas, including: Trends with Respect to Power, Trends with Respect to Cooling and Heat Re-use, and Monitoring and Control. The full report is available online (at [https://prace-ri.eu/wp-content/uploads/WP311\\_HPC-Infrastructures-Workshop-2021.pdf](https://prace-ri.eu/wp-content/uploads/WP311_HPC-Infrastructures-Workshop-2021.pdf)) and relevant points are summarised below.

### Trends with Respect to Power

- Despite the progress that has been made in terms of flops per watt in modern CPUs and GPUs, most centres envisage that they will need more power in the coming years to supply a level of HPC capacity that meets the demands of researchers from both academia and industry.
- There may be financial (and emissions) benefits in scheduling jobs in a more predictable grid-friendly way by keeping the power consumption within a specific window.
- HPC services can use monitoring and analysis to predict the power usage of the applications that are sitting in the queue waiting to be scheduled and can take this into account to adjust their power usage and carbon footprint.
- With the introduction of more power coming from solar and wind sources in national and regional power grids, more fluctuations in the amount of power supplied by the grid are to be expected, and data centres need to take this into account.

### Trends with Respect to Cooling and Heat Re-use

- Direct liquid cooling solutions are predominant for compute systems in larger installations, while air-cooling is still dominating for storage.
- Re-use of waste heat has been employed in northern Europe for a long time due to the cold climate and the existing district heating. Re-use of waste heat for district heating can have significant benefits for overall carbon emissions.
- Some new installations use adsorption chillers that can transform waste heat into cooling for buildings or other computers.

### Monitoring and Control

- The predominance of warm-water-cooled high-density systems has led to tighter coupling between the management of the actual facility and the management of the HPC systems.

- Broad instrumentation of all components of the data centre and the HPC system and its operating system and job scheduling software is a basic requirement for improving energy efficiency.
- Most centres build their own integrated data centre management tool, relying heavily on open-source projects to provide core blocks of functionality.
- There are prospects for automating control, including machine learning for optimising decisions for facility control systems and/or job scheduling.

## HPC Infrastructure Workshop #10

The 10<sup>th</sup> European Workshop on HPC Centre Infrastructures was held in Poland on 20-23 May 2019. The workshop brought together 74 participants from across Europe and the rest of the world including both experts from the HPC facility management side as well as experts from companies working in the area of energy-efficient HPC. The workshop themes were immersion-cooling, standards relevant for HPC and HPC facility management, energy, future technologies, tooling for facility management, and developments concerning EuroHPC. The detailed report of the workshop is available on the PRACE website at <https://prace-ri.eu/wp-content/uploads/Best-Practise-Guide-WP292-HPC-Infrastructures-Workshop-10.pdf>. The main findings are summarised below.

### Trends with Respect to Cooling

- Direct liquid cooling solutions predominate over air-cooled solutions for the compute part of current state of the art HPC systems.
- Immersion cooling has not achieved comparable popularity. However, newer immersion cooling systems are more serviceable than earlier systems. However, drawbacks remain, such as lower density per square meter, resulting form racks being lower and “horizontal” rather than “vertical.”

### Trends with Respect to Power

- Despite the ongoing increase in flops per watt, most data centres expect to need increasing amount of power to supply enough HPC capacity to satisfy ever increasing demand.
- HPC datacentres could be a cause of instability in the power grid if there are large fluctuations in power draw. Maximising energy-efficiency for HPC may be constrained by the need to operate in a more grid-friendly way (or measures may need to be taken to improve grid resiliency).
- There may be benefits in scheduling in a more predictable way, i.e., keeping the power draw between a predetermined lower and upper bound. The workshop showed that this could be done via sites monitoring and analysing runs to predict the power usage of queued jobs and then taking this into account in scheduling policies.

### Monitoring and Control

- Predominance of water-cooled high-density systems has led to tighter coupling of facility management and HPC system management.
- Broad instrumentation of all components of the datacentre and HPC system is a key requirement for improving energy-efficiency. However, these data streams must be aggregated into meaningful overviews.
- Most major sites invest in tooling that they can customise to their datacentre and HPC system.
- Improved data capturing and analysis can allow automated control, feeding back decisions into facility control systems and/or job scheduling.