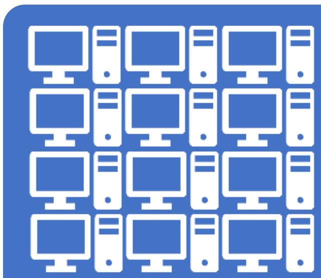
Government
Office for Science

- Key concepts in large-scale computing
- Users and uses of large-scale computing
- The global stage, data and hardware
- UK infrastructure, software and skills
- Achieving the UK's potential in large-scale computing

Large-scale computing: the case for greater UK coordination

A review of the UK's large-scale computing ecosystem and the interdependency of hardware, software and skills. This report provides evidence to support the establishment of a coordination function and sets out the key policy areas to consider.

September 2021



The UK currently has **12** systems in the Top500

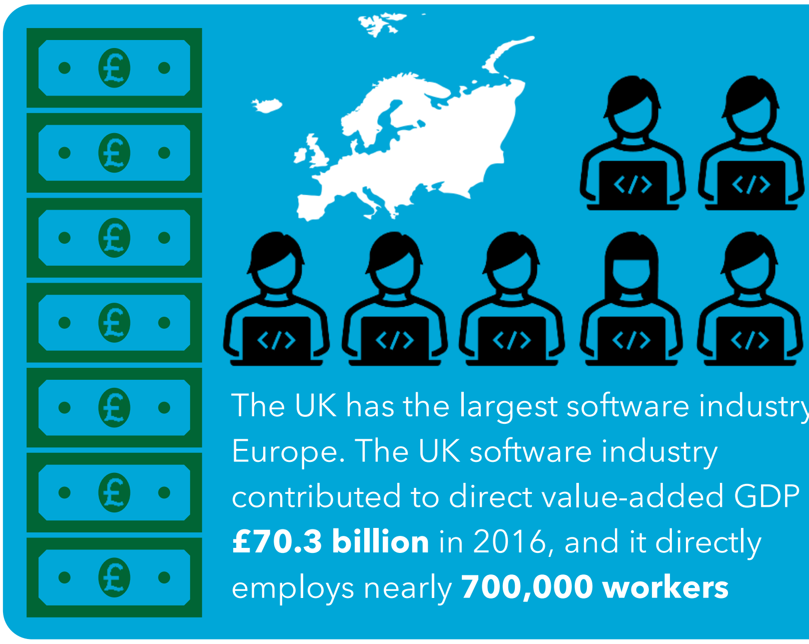
Representing **2.4%** of global installations and **1.4%** of total performance capacity

In **2018**, **74%** of high-performance computing sites were found to run some of their workloads in the **cloud**

Developments in large-scale computing are expected to open up a range of new use cases:

- **Smart cities**
- **Emergency preparedness and resilience**
- **Energy networks**
- **Public health**
- **Digital twins**

The Met Office supercomputer is the UK's highest-ranking machine at **#37**



The UK has the largest software industry in Europe. The UK software industry contributed to direct value-added GDP of **£70.3 billion** in 2016, and it directly employs nearly **700,000 workers**

Glossary of Terms *See Appendix A for full glossary of terms.*

Term	Meaning
Access models	A method by which users can access the capability of a computing system (for example, cloud computing).
Algorithms	A set of instructions that, when given to a computer, are used to solve problems.
Artificial intelligence	A broad area of computer science concerned with the use of computer systems to carry out functions analogous to human cognition.
Cloud computing	A rapidly emerging access model where computing infrastructure is accessed on-demand via the internet.
Exascale	A high-performance computing system that is capable of at least one Exaflop per second (i.e., a system that can perform more than 10^{18} floating point operations per second).
Large-scale computing	Computer systems where processing power, memory, data storage and networks are assembled at scale to tackle computational tasks beyond the capabilities of everyday computers. Often involves the widespread use of parallelisation (see Appendix A). An umbrella term encompassing terms such as high-performance computing, high-throughput computing, supercomputing and novel computing paradigms.
Petascale	A high-performance computing system that is capable of at least one Petaflop per second (i.e., a system that can perform more than 10^{15} floating point operations per second).
Quantum computing	An experimental form of computing which utilises the probabilistic qualities of quantum mechanics. It has the potential to massively accelerate certain computational tasks (also see Appendix C).

Preface

Large-scale computing (LSC) has revolutionised our lives in fields of national importance such as weather and climate modelling, and financial services. It is an important enabler of R&D as demonstrated by DeepMind's remarkable recent breakthrough on protein folding. Computing power also underpins key technologies such as Machine Learning and Digital Twins.

The pace of innovation in LSC is accelerating rapidly, the highest performing system today is 175,000 times more powerful than the largest system 20 years ago. This brings with it new opportunities and emerging applications in areas such as fusion energy and public health. With new opportunities come challenges the UK must address if it is to realise the true potential of LSC and the PM's ambitions for the UK as a Science Superpower.

This report describes the LSC landscape in the UK and sets out the building blocks to creating a world-class computing ecosystem. While quantum computing is likely to be of critical importance, we are some years from commercialisation and therefore is not covered in much detail. The core of this report is based around seven challenge areas and options to overcome each: the need for better national coordination; nurturing a healthy UK ecosystem; ensuring we have the right mix of hardware, skills, and software; minimising energy consumption; and the UK supply chain.

LSC forms part of the UK's national infrastructure and many of the key issues span multiple sectors. The Government therefore has a significant role to play in nurturing and supporting the UK ecosystem both as a consumer and a funder. Therefore, our primary recommendation is to establish a team within Government to take policy responsibility for large-scale computing and address the challenges that are identified in the report. Industry, as a major user and innovator of large-scale computing, will be part of the solution and need strong ties to the coordination team within Government.

This report has been informed by a wide range of academic, government and industry experts. I would like to thank them all for their invaluable contributions.

Sir Patrick Vallance

Government Chief Scientific Adviser

Executive summary

There is a strong case for continued public investment in large-scale computing. However, this needs to be coupled with improved long-term strategic planning, driven by close collaboration between Government, academia, and industry.

Large-scale computing is an essential tool for solving industrial and scientific problems. It is used to analyse highly complex or data-intensive problems involving simulation, optimisation, data processing and artificial intelligence (AI). Large-scale computing supports a range of R&D-intensive sectors and is used as a research tool at many of the frontiers of science. Several large-scale computing systems, such as those used for weather forecasting, form part of the UK's critical national infrastructure. Computing needs vary substantially across sectors, with users having different requirements in terms of both capability and support requirements:

- In the public sector, service reliability is a key consideration. For public sector users, cybersecurity and the physical locations of infrastructure are other key considerations.
- Academia requires access to systems of a range of sizes and architectures owing to the diversity of programs run.
- Industrial users of large-scale computing use a wide range of access models. Private sector use of public systems often includes a requirement to publish results and cybersecurity is a key concern.
- Small and medium enterprises (SMEs) often do not have an awareness of the range of business problems that large-scale computing can solve or the technical requirements to use it. In addition, SMEs often require support in adapting their software to run on large systems.

The UK is a global leader in a number of computing domains, including in areas of software development, computational modelling, data analytics, cybersecurity, AI and machine learning.[1,2] However, the UK's large-scale computing infrastructure lags behind other major global economies and would benefit from a refreshed, longer-term focus that takes into account interdependencies such as skills and software. As of November 2020, China and the US had 214 and 113 of the top 500 computer systems globally, and France and Germany had 18 and 17. In contrast, the UK had 12.[3] The UK's share of global high-performance computing capacity has decreased by three-fifths over five years, falling to 2.0% in 2019.[4]

Large-scale computing is a fast-changing field, where systems have a working life of five to eight years before becoming obsolete. We are entering an era of unprecedented hardware diversification, driven by the development of new computing architectures for specific tasks, in particular for AI research. Advances in data science have also driven demand for large-scale computing. As a society, we are generating far more data than ever before, and without the means to analyse it, its value is left unrealised.

Cloud computing has become an increasingly popular access model. Cloud access provides opportunities for new users to reap the benefits of many (but not all) forms of large-scale computing by offsetting upfront costs with cost of use. Users can assess their needs and choose a cost model to suit, whether that be on-site hosting or access via the cloud. Commercial cloud service providers are increasing their range of capabilities. This is still a nascent sector, and the market is currently dominated by a limited number of leading providers based in the US.

There is a strong case for continued public investment in large-scale computing. Access to world-class computing capabilities enhances UK competitiveness across a number of sectors. Lack of sufficient access to computing can create bottlenecks within research, development, prototyping and testing. The cost of leading-edge systems is greater than any single institution or business can afford. The UK Research and Development Roadmap describes supercomputing as a key part of our digital research infrastructure.[5] Large-scale computing plays an essential role underpinning UK R&D, helping to achieve ambitious 'moonshot' challenges, such as zero emission air travel.[6]

Investment in computer hardware alone will not be sufficient. To achieve the full potential of large-scale computing and allow the UK to compete at an international level, longer-term strategic planning is needed. Better coordination is required between Government and the publicly and privately funded research sectors. This report focusses on the functions that will be necessary for the UK to address these challenges and seize on the opportunities brought by large-scale computing.

Global trends spark important strategic questions about the types of national investments the UK should make. Our vision is to establish a dedicated oversight group tasked with providing effective coordination to the large-scale computing ecosystem and its users, to level up Britain as a science superpower.

Recommendations

1. **National coordination.** At present, there is no overall national roadmap for large-scale computing, and there is no single team within Government that provides coordination or carries out wider strategic assessment of UK needs. This has resulted in a wide variety of different procurement practices across departments.
   - **We recommend establishing a team within Government to provide policy leadership of large-scale computing**. This team would be responsible for developing a rolling long-term roadmap for large-scale computing. This roadmap should cover the whole UK computing ecosystem, including software, skills, and user needs. This would help to improve resource sharing between organisations and provide a conduit for industry engagement. Finally, this team could advise Government on computing procurement, helping the public sector become a more intelligent customer.

2. **Future systems.** Access to leading-edge computing is vital for many academic and industrial fields. Future computing capabilities will open up new application areas and help to solve problems that are currently intractable. Projects are underway around the world to develop exascale computing which would deliver systems roughly 140-times faster than the UK's fastest system as of November 2020. These systems could mark a step-change in certain capabilities. Meanwhile, novel computing architectures could greatly accelerate specific applications such as training AI algorithms.
   - **The UK should maintain access to leading-edge computing capability, as well as a diverse range of system architectures.** For certain applications, academia and the public and private sectors will require access to exascale systems. Exascale capabilities can be achieved through a balance between international collaboration and procuring new domestic systems.

3. **User needs and the future ecosystem.** Large-scale computing users have a range of different requirements – from hardware requirements to access and support. Making large-scale computing accessible to more users would help to solve new business and research challenges. However, many barriers exist that can prevent potential users from accessing these resources such as a lack of awareness, necessary expertise, or appropriate access-models.

- ➢ **The roadmap for computing should be designed to meet the diverse requirements of users.** This could include targeted support for SMEs to access large-scale computing, as well as taking steps to ensure that the UK marketplace for cloud computing is competitive.

4. <u>**Software.**</u> High-quality software is fundamental to realising the benefits of investments in computing. Software must be fit for purpose and be regularly tested, updated, and archived. The UK is a world leader in software development; to maintain this, software development must keep pace with advances in hardware.
   - ➢ **Best practices for developing robust and verifiable software should be adopted across industry, academia, and Government.** In academia, testing, sharing, and archiving of software should become an essential part of the research process, ensuring that scientific results can be replicated. New computing architectures will necessitate considerable reengineering of software.

5. <u>**Skills.**</u> There is an acute shortage of large-scale computing professionals, including system architects, data engineers, system operations professionals and software engineers. Exploitation of advanced computing requires skilled cross-disciplinary teams. Within Government and academia, career paths for computing professionals are often not well-defined.
   - ➢ **Career pathways should be developed to help attract, support, and retain these skill sets.** Job security, salary structures and progression opportunities should be improved to retain talent. The pipeline of talent could be bolstered through expanded apprenticeships and professional training programmes.

6. <u>**Energy and sustainability.**</u> Electricity is a significant proportion of the overall running costs of large-scale computing. Computing facilities are large consumers of energy, and they can put significant demands on regional electrical grids.
   - ➢ **Total cost of ownership should be considered during procurements, as well as whole life-cycle carbon emissions.** A long-term roadmap could help ensure that computing needs are factored into electrical grid planning.

7. <u>**The UK supply chain.**</u> The global large-scale computing sector has become increasingly consolidated through recent mergers and buyouts. To promote innovation, it is important to maintain a sector that is diverse and competitive. Most of the vendors of large-scale computing systems, and

their associated hardware, are based outside of Europe. By some estimates, the global market is forecast to grow to £38 billion by 2022[7]. However, due to a lack of domestic capability the UK is not currently well-placed to capitalise on this growth.

- **A long-term roadmap for procurement could provide clarity to the UK computing sector, helping to encourage domestic investment.** A strong domestic sector would strengthen the UK's influence in shaping global hardware trends and standards. This would also provide a platform for capitalising on export opportunities brought about by growing global demand for large-scale computing.

Table of Contents

Preface ... 4
Executive summary ... 4
Recommendations ... 7
1. Introduction ... 12
1.1. Scope and outline of this report ... 12
1.2. Key concepts in computing ... 13
1.3. Large-scale computing ... 14
2. Users and uses ... 17
2.1. Users of large-scale computing ... 17
2.2. Emerging users and use cases ... 18
2.3. Access models ... 20
3. The global backdrop ... 26
3.1. The international landscape ... 26
3.2. The data revolution ... 29
3.3. Developments in computing hardware ... 30
4. The UK's position in the global landscape ... 36
4.1. Current UK infrastructure ... 39
4.2. Software and skills ... 43
4.3. National coordination ... 46
4.4. The UK supply chain ... 47
5. Achieving the UK's potential in large-scale computing ... 49
5.1. The value of large-scale computing ... 49
5.2. The case for change ... 50
5.3. Recommendations ... 50
5.3.1. Challenge 1: National coordination ... 52
5.3.2. Challenge 2: Future systems ... 54
5.3.3. Challenge 3: User needs and access models ... 57
5.3.4. Challenge 4: Software ... 60
5.3.5. Challenge 5: Skills ... 62
5.3.6. Challenge 6: Energy and sustainability ... 65

5.3.7. Challenge 7: The UK supply chain 67

Appendix A: Glossary of terms........ 69

Appendix B: A schematic description of a typical large-scale computing system . 72

Appendix C: Hardware for large-scale computing 73

C.1 Types of processors........ 73

C.2 Emerging computing paradigms 75

Acknowledgements 77

References........ 79

1. Introduction

Since the early nineteenth century, the UK has played a leading role in developing the field of computing. UK scientists have played key roles in developing the first mechanical computer (the Difference Engine), the first programmable computer (Colossus) and the first commercial computer (Ferranti Mark 1). The UK was also home to many of the pioneers of computing, including Babbage, Lovelace, Flowers, and Turing.[8]

Early computing systems often filled large rooms and required several operators. Although modern personal computers have become much smaller, there are a number of specialist applications that still require warehouse-sized large-scale computing systems. Over several decades, we have seen exponential growth in system performance, the largest high-performance computing system today is 175,000 times more powerful than the largest system 20 years ago.[i]

1.1. Scope and outline of this report

This report discusses the large-scale computing landscape in the UK. It outlines several building blocks required for creating a world-class computing ecosystem and the coordination mechanisms needed to support this. The following areas are within the scope of this report:

- High-performance computing (HPC) and high-throughput (data intensive) computing (HTC), which throughout the report are encompassed by the term 'large-scale computing'.
- Access models for large-scale computing.
- Large-scale data handling, storage, and networking; and,
- The wider ecosystem around large-scale computing, including software, skills, and users.

Outside of the scope of this report are:

- Personal computing and workloads carried out on personal computers.
- Issues around data governance and ethics, except to the extent that they influence technical decisions concerning infrastructure provision.
- Cloud-based computing workloads that do not require HPC or HTC capability; and,
- Digital services such as video streaming and e-commerce.

This report has been prepared through interviews with stakeholders from across the large-scale computing ecosystem.

[i] Figures are calculated from LINPACK benchmark scores (R_{MAX}) for the most powerful HPC systems globally. Data range is June 2000 to June 2020. Source: TOP500 (top500.org).

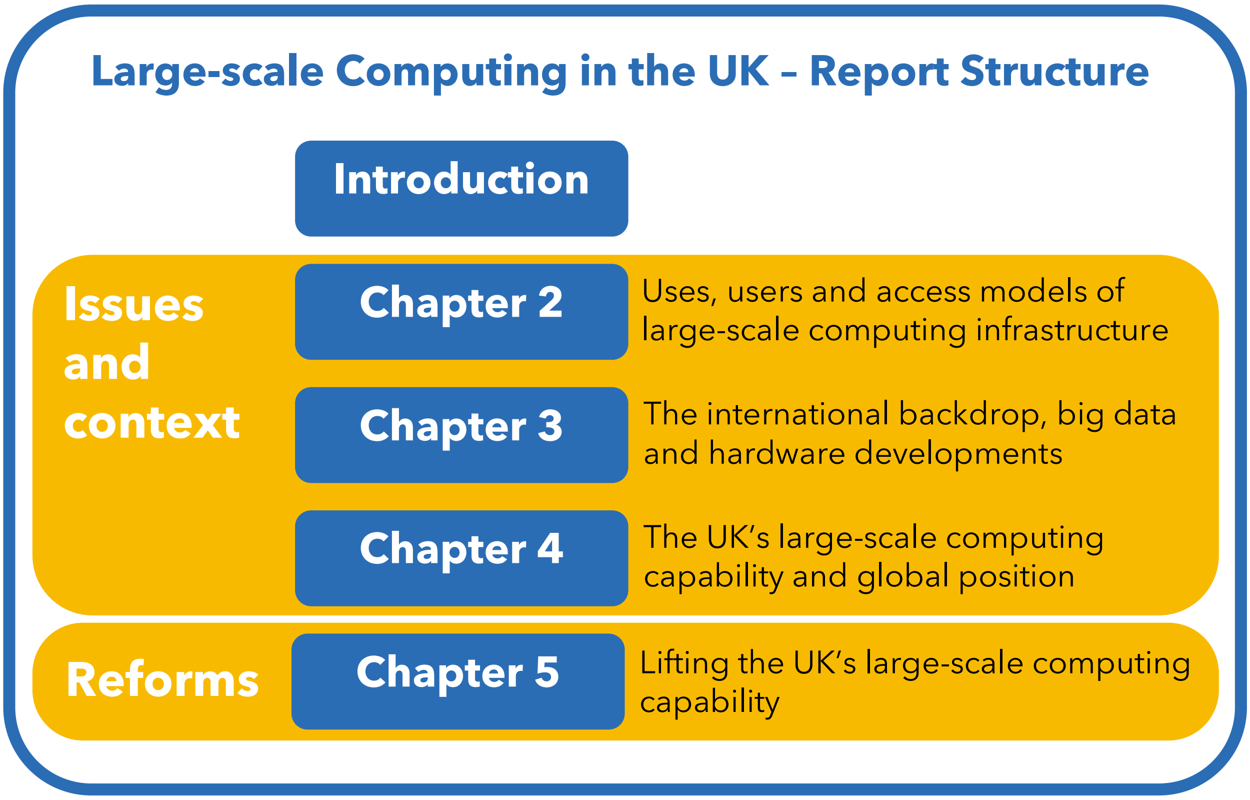
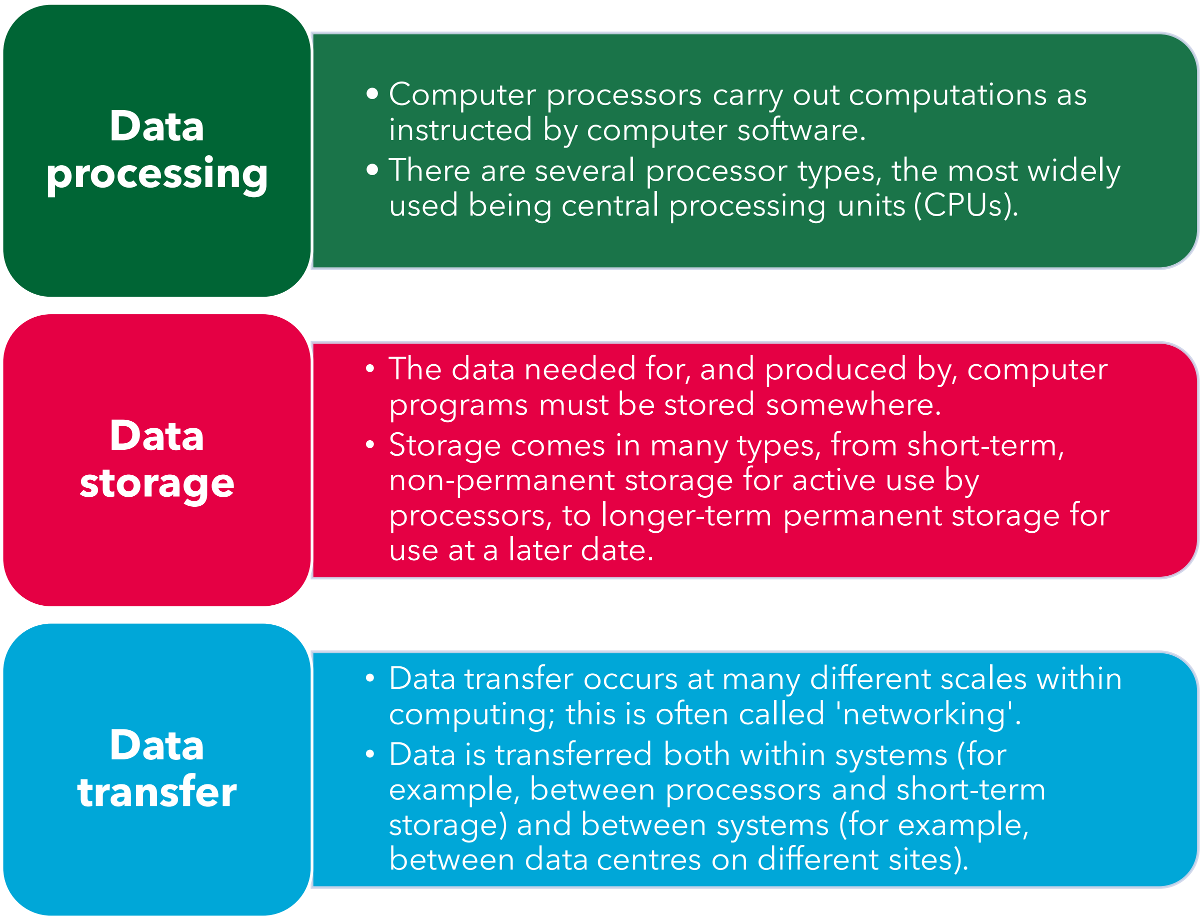

Large-scale Computing in the UK – Report Structure
Introduction
Issues and context
Chapter 2
Uses, users and access models of large-scale computing infrastructure
Chapter 3
The international backdrop, big data and hardware developments
Chapter 4
The UK's large-scale computing capability and global position
Reforms
Chapter 5
Lifting the UK's large-scale computing capability

Figure 1: Structure of this report.

The first four chapters of this report set out the key concepts and issues surrounding large-scale computing, discuss the main uses of this infrastructure and describe the UK position within the broader global landscape. Following this, chapter five outlines seven building blocks for strengthening the large-scale computing ecosystem in the UK.

1.2. Key concepts in computing

Broadly, computing infrastructure can be grouped into three key functions which all computer systems require. These are *data processing*, *data storage* and *data transfer*, as explained overleaf.



Data processing

- Computer processors carry out computations as instructed by computer software.
- There are several processor types, the most widely used being central processing units (CPUs).

Data storage

- The data needed for, and produced by, computer programs must be stored somewhere.
- Storage comes in many types, from short-term, non-permanent storage for active use by processors, to longer-term permanent storage for use at a later date.

Data transfer

- Data transfer occurs at many different scales within computing; this is often called 'networking'.
- Data is transferred both within systems (for example, between processors and short-term storage) and between systems (for example, between data centres on different sites).

Figure 2: Three broad systems involved in computing.

1.3. Large-scale computing

Some computing tasks are so large or complex that they go beyond the capabilities of personal computers. These computational tasks can provide challenges in some, or all, of the following areas:

- **Processing:** Very large amounts of processor time are required.
- **Storage:** Very large amounts of data storage are required.
- **Transfer:** Very large amounts of bandwidth are required for data transfer.

Large-scale computing systems address these bottlenecks through parallelism: the simultaneous, coordinated use of multiple processors, storage media and communication links. Through parallelism, computational problems are split into smaller elements that are run concurrently on a number of processors, greatly reducing the overall time required. Large-scale computing is often subdivided into high-throughput computing (HTC) and high-performance computing (HPC) (see 1.3.3.).

This infrastructure typically has a high cost to purchase and operate. A typical system has a working life of just 5 to 8 years before becoming obsolete. The first twelve months of a large-scale computing system are when it is at its most competitive and therefore preparations to ensure rapid uptake of new resources are essential to maximise return on investment. The rapid rate of hardware development and improvement means that systems quickly become outclassed. The performance improvement offered by new systems, combined with the high operational and power costs which existing systems incur, results in short lifetimes for systems.

1.3.1. Architecture and set-up

Large-scale computing systems tend to be bespoke and specialised, being designed by vendors to efficiently solve a certain class of problems. Different sets of problems have different processing, memory, and networking requirements. Therefore, the design of the data processing, data storage and data transfer functions vary. A schematic description of a typical system is shown in Appendix B.

1.3.2. Software development and system operation

Running large-scale computing workloads requires software that has been designed to run efficiently in a highly parallel way – where workloads are divided into many smaller tasks that can be processed separately, with results being combined upon completion. This is more challenging than writing software that is designed to process instructions sequentially ('in serial').

Writing parallelised programs requires the software engineer to understand the system architecture and control the flow and distribution of data between individual processors. Significant training and support, in terms of both software development and system access, is required before inexperienced programmers can efficiently make use of parallel computing at scale. Domain scientists also play an essential role, mapping the real-world scientific problems to computational solutions.

Operating large-scale computing systems is also a complex task. Highly skilled system administrators and software developers are required to configure, tune, and deploy software for use on large-scale computing infrastructure. Typically, very few jobs will use 100% of a system, so multiple jobs are often run simultaneously using differing proportions or partitions of a system. Efficient job allocation is therefore essential to achieve cost-effective utilisation of systems. This creates complex scheduling and dependency challenges.

1.3.3. Computational workloads for large-scale computing

Different types of large-scale computing programs run efficiently on different types of systems. A commonly used distinction is between **high-performance computing (HPC)** and **high-throughput computing (HTC)** as compared in the table below.

	High-Throughput Computing	**High-Performance Computing**
Optimised for	Capacity – completing a large number of tasks	Capability – completing sizeable tasks quickly
Typical applications	Analysing and processing large volumes of data	Modelling and simulating complex systems
Coupling of code	Loose (i.e., individual nodes communicate with each other infrequently)	Tight (i.e., individual nodes communicate with each other frequently)

Table 1: Comparison of high-throughput computing and high-performance computing.

Typically, HPC workloads generate data while HTC workloads analyse data generated elsewhere. However, the distinction between HTC and HPC workloads is not clear-cut, and the two terms tend to be defined inconsistently. Furthermore, with the growing use of large-scale data analytics within modelling and simulation, some argue that a convergence is underway. Indeed, both types of workload can make use of similar large-scale computing systems, though it is more efficient to use system architectures that are tailored to the needs of the application. A well-designed large-scale computing system, tailored for particular science workflows, can deliver comparable science outputs to systems with much higher theoretical performance but that are less balanced in terms of memory, network bandwidth and processing power. For example, the high memory per node of the new DiRAC-3 (Distributed Research using Advanced Computing) Memory Intensive service means that it will be capable of performing cosmological simulations (see Section 2.1) on 49,000 cores – which, when run on the European PRACE (Partnership for Advanced Computing in Europe) machines require 200,000 cores. Therefore, rather than making a binary distinction between HPC and HTC is it more helpful to recognise that there are a range of workload types which might require large-scale computing (see Section 2).

2. Users and uses

Strategic decision making that considers user needs and future uses will have the greatest impact on society. Long term planning and coordination should work to make large-scale computing accessible to more users encouraging uptake, competitiveness and innovation.

The types of problems that are solved using large-scale computing fall into four broad areas:

- Simulation of large and/or complex systems.
- System optimisation.
- Processing and analysing very large volumes of data.
- AI and machine learning.

Problems do not always fall into a single one of the four areas above, and there is some degree of overlap between these categories.

2.1. Users of large-scale computing

Large-scale computing is a fundamental tool in many disciplines, enabling verification of experimental results, modelling of otherwise inaccessible systems or processing of massive quantities of data.[9,10]

There are a wide range of applications for large-scale computing across public sector bodies, academia, and industry. A selection of applications is shown below.

- **Chemicals and materials:** Chemistry and materials science are significant users of 'in silico' chemical modelling, for example using simulations to understand the structures and properties of materials, modelling molecular interactions, and designing new drugs.
- **Weather and climate modelling:** Large-scale computing is crucial for high-resolution weather and climate modelling. As well as the Public Weather Service, accurate weather data supports the transport, agriculture, energy, and defence sectors. Large-scale computing also underpins climate modelling to support the work of the Intergovernmental Panel on Climate Change (IPCC).
- **Life sciences:** Large-scale computing enables the processing and analysis of large biological datasets, such as population health data or human genome sequences. This facilitates the advancement of understanding of biological systems such as vascular system modelling and supports drug development for novel pandemics like COVID-19. [11]
- **Engineering and product development:** Businesses can improve some of their research, development, and prototyping activities through use of large-scale computing. For example, in the consumer products sector it can

enable product stability tests to be carried out in minutes rather than weeks.[12] In the aerospace sector it can facilitate more rapid and cost-effective prototyping of engines.[13]

- **Nuclear stockpile stewardship:** In the Ministry of Defence, large-scale computing provides modelling and simulation capabilities to support research into the performance and reliability of nuclear warheads.[14]
- **Financial services:** Large-scale computing supports a range of applications in the financial services industry. Simulation techniques including the Monte Carlo method can be used for risk characterisation and optimisation. Emerging techniques such as AI and deep learning can inform investment decisions.
- **Fundamental physics:** Large-scale computing is crucial for high-energy physics experiments such as those undertaken at CERN. Through the modelling of particle collisions, analysis of data from particle accelerators and comparison of both simulated and real data, large-scale computing supports new insights into particle physics.
- **Cosmology:** Simulations of galaxy formation and evolution are vital for generating the scientific return for UK investments in major international telescopes and satellite missions. They provide insight into the origin and evolution of all the structures in the universe.
- **AI and Machine Learning:** Training complex and advanced deep learning models requires the fast processing of large data sets, which large-scale computing allows. In recent years there has been exponential growth in the computing power required to train the most demanding AI algorithms.
- **Defence:** Large-scale computing underpins defence led modelling and simulation; ranging from low level physics modelling up to the analysis of mission effectiveness via campaign models.
- **The internet:** Large-scale computing can be used to analyse internet traffic, which is useful for informing internet policy, identifying and preventing outages, defending against cyberattacks, and designing more efficient computing infrastructure. [15] Most of social media, search tools and web services are powered by large-scale computing and underpin their business models, investment planning and advertising revenues.

2.2. Emerging users and use cases

Developments in large-scale computing such as exascale computing (see Section 3.3.1), as well as new, more accessible, access models are expected to open up a range of new use cases, including:

- **Smart cities**: Recent advances in sensor technology and Internet of Things (IoT) devices have stimulated the development of small, cheap mobileenabled sensors that collect large quantities of data. Large-scale computing can be used to process and create insights from the large volumes of data produced by smart cities.

- **Emergency preparedness and resilience**: There is growing interest in the use of large-scale computing to improve responses to emergency situations, for example the FireGrid project exploring the use of computing to inform and coordinate fire emergencies.[16] During the COVID-19 pandemic, large-scale computing resources have been used for epidemiology, bioinformatics, and molecular modelling to inform policy decisions and accelerate the discovery of treatments.[17]
- **Fusion energy:** The Governments Ten Point Plan for a Green Industrial Revolution has set a grand challenge to deliver fusion energy to the grid in the 2040s. This time scale eliminates conventional, real world, test based, iterative design as an option, instead requiring reactor designs to be developed 'in-silico'. This requires a paradigm shift in the way complex products are designed, exploiting exascale computing and state of the art data driven engineering design technologies.
- **Energy networks:** Power grids are being modernised through the introduction of IoT sensors, including smart meters. The data that these sensors provide can be used for real-time data analysis and optimisation through large-scale computing.[18]
- **Public health:** Healthcare providers are producers of enormous quantities of data, which historically have been difficult and costly to analyse. Large-scale computing facilitates research into medical data, giving insights into population health and improving public health interventions.[19]
- **Computer-aided diagnostics:** Through advances in image recognition and machine learning, large-scale computing can be used to provide near real-time analysis of medical scans. The Industrial Strategy Challenge Fund challenge 'From data to early diagnosis and precision medicine' is focussed on this area.[20]
- **Processing of unstructured data:** Advances in data science, AI and natural language processing could create opportunities to use large-scale computing in new fields. In the humanities and social sciences, there are large quantities of unprocessed, unstructured data from sources including books, videos, and social media. Large-scale computing can facilitate processing of this data, providing new insights in a range of fields.
- **Digital twinning:** Increasing computing power and the collection of real-time data from sensors opens up opportunities for the creation of digital replicas of physical objects in the built and natural environments. Studying the replica of the object allows for prediction of real-world behaviour and

the early identification of problems. Efforts such as the National Digital Twin Programme facilitate the development of digital twins.[21]

The use of large-scale computing can offer substantial benefits to new users, with increased diffusion of large-scale computing capabilities possibly playing a key role in addressing the UK's productivity gap.[22] However, uptake amongst many potential SME users remains limited due to multiple barriers[23]:

- **Awareness and expertise:** They may lack awareness of the benefits that it can bring as well as the specialised skills required to make use of large-scale computing and to map their real-world challenges to computational solutions. Intermediaries can play a critical role in raising awareness and offering specialised support which together provide a pathway to greater SME utilisation of large-scale computing.
- **Finance:** The high levels of investment required in infrastructure and human resources to establish large-scale computing capabilities and the perceived high-risk of such projects can exclude SMEs with relatively small overall R&D budgets and limited access to finance. Access to shared or on-demand cloud-based resources may therefore be an important enabler of uptake by SMEs.

2.3. Access models

The range of user requirements from large-scale computing leads to a variety of different access models. Some of the key considerations are:

- **Frequency and variability of use:** Usage requirements will determine which access model is most convenient and economical. Some users have infrequent or sporadic demand for large-scale computing, whereas others rely on it continuously for business-critical processes requiring back-up, redundant systems to maintain very high-levels of uptime.
- **Level of support required:** While some experienced users will be able to make use of large-scale computing facilities relatively independently, others may require a high degree of support to both understand the use cases of large-scale computing and to effectively write software for these systems.
- **Intellectual property and security concerns:** Some users make use of software or data that is sensitive or proprietary, which introduces limitations on which access models are viable.

2.3.1. "Classical" access models

Historically, access to large-scale systems has been achieved through the following access models:

- **On-site hosting:** Large companies and some public sector institutions often find it most efficient to host their own large-scale computing systems in-house. This represents a significant capital expense, but can be cost-effective if they have reasonably steady, high demand for large-scale computing. On-site hosting can also mitigate some security concerns.
- **"Bid-for-time" shared systems:** This access model is frequently employed for large systems across academia. Research councils or institutions fund systems for the use of the whole community, and users often bid for time through an assessment of scientific merit. A requirement to publish findings in peer-reviewed journals helps to ensure that allocated time is well used. Industry can gain access through research collaborations or on a pay-per-processor-hour basis.
- **Grid or scavenger computing:** In this access model, computing resources across multiple institutions and geographic locations are brought together via a network as a "grid".[24] Due to its distributed nature, these systems tend to be used for high-throughput, easily parallelisable workloads. In some cases, grids can be assembled from idle or underutilised computing resources in a process called cycle-scavenging. In this way distributed computing projects like Folding@home (which aims to help scientists develop new therapeutics) have enabled scientific challenges to be tackled using personal computers.[25]

2.3.2. Large-scale computing in the cloud

Cloud computing is a significant access model for a broad range of computing tasks. This refers to on-demand access to computing infrastructure owned by a cloud service provider (CSP) via the internet.[26] In this report, we cover specifically HPC and HTC in the cloud. There are also a range of other uses of cloud computing.

Cloud access has the potential to become a major provisioning model for large-scale computing. A key distinction from other access models is in the way computing resources are procured, generally as an on-demand service with a flexible pay-as-you-go model, without requiring users to make up-front capital investment or participate in a merit-based bidding process.

The typical features of different access models are described below. For comparison, a non-cloud access model is also shown:

	Public cloud	Managed private cloud	On-premises hosting (non-cloud)
Owner operator	Cloud service provider (CSP) owns infrastructure and manages system	CSP manages system and often owns the hardware	End user purchases and manages system
Location	Hosted at the CSP's data centre	Typically hosted at the CSP's data centre	Hosted by the user
Cost model	Pay-per-use	Ongoing subscription to CSP	Upfront capital and ongoing running costs
Exclusivity	Resources shared between many users	User has sole access	User has sole access
Hardware	Selection of hardware configurations may be available	Bespoke hardware designed for user needs	Bespoke hardware designed for user needs
Scalability	Scalable (additional processor cores can be made available for resource-intensive tasks)	Limited scalability within initial bespoke hardware, highly scalable to non-bespoke public cloud infrastructure	Not easily scalable (tasks are limited to the hardware specifications of the system)

Table 2: Features of different large-scale computing access models.

Hybrid models, utilising a mix of public cloud, private cloud, and on-site hosting, are also available. These models can be especially useful for managing peaks and troughs in demand.

Large-scale computing in the cloud offers a number of potential advantages:

- **Cost-per-core model:** Instead of large, up-front capital costs, users pay ongoing subscriptions based on their usage. System upgrades and system management become the responsibility of the CSP rather than the user. This may open up large-scale computing to new users who are unable to afford the upfront costs of large-scale computing systems. The cost-per-core of a cloud service provider drives purchase decisions and keeps the market competitive.

- **Enabling flexible demand:** As CSPs pool resources between many users, the cloud model allows for fluctuating levels of demand from individual users. Users with their own in-house systems can make use of "cloud bursting" where computing workloads are dynamically allocated to the cloud to meet spikes in demand.[27]
- **Access to different hardware architectures:** CSPs can offer access to a range of hardware architectures across their data centres. Through this, users can run programs on different computing architectures tailored to specific applications.
- **Support to users:** Cloud computing can offer 'large-scale computing-as-a-service', where users with limited expertise can submit jobs via an easy-to-use portal.[28] This may help overcome skills barriers that can limit uptake of large-scale computing.

However, there are several limitations and challenges around cloud computing:

- **Performance at scale:** The capability of commercial cloud systems to run complex workloads cost effectively has been improving over time. However, it does not yet meet the needs of the most demanding users, who require access to world-class large-scale computing systems, nor can it currently match the capability of traditional large-scale computers. The procurement and operation of these leadership-class computing systems is often prohibitively expensive. Furthermore, cloud facilities are often not equipped with the very high-speed interconnect networks required to run certain tightly coupled programs (see Section 1.3.3).[29,30]
- **Virtualisation:** The foundation of cloud computing. To utilise cloud computing, workloads must be able to run on virtual machines (an emulation of a physical computer). While there is a growing trend towards virtualisation in many organisations, it can also require significant adaptation of infrastructural set-up and associated specialist skills. Virtualisation also adds an additional layer of abstraction between the user and the physical hardware which can reduce efficiency.
- **Cloud "lock-in":** When using cloud service providers (CSPs) it is possible to become locked-in due to tight technical integration and reliance on a particular CSP's services. Additionally, the present cost-model for cloud computing includes significant costs for data egress (i.e., transferring data out of a CSP's data centre), which can make it prohibitively expensive for some users to switch suppliers. The Government Digital Service provides guidance on pragmatically managing this lock-in.[31]

- **Cybersecurity:** Cloud provision often involves moving data and code to remote servers for processing and storage. Many uses of large-scale computing systems involve sensitive data and/or software. In such cases, users need high confidence in the security of data during transfer, processing, and storage.
- **Critical national infrastructure (CNI):** For some users, continuous system availability is mission critical to deliver an essential service (for example, managing a national weather service, power grid or smart city infrastructure). In such cases, in-house management or private cloud models may be preferred.
- **Regulation:** The nature of the supply side of large-scale computing is changing with the use of CSP on the rise. Should a cloud service become so heavily used that it resembles CNI, it would need to be regulated to assure appropriate use, security and resilience. There is currently no mechanism to classify large-scale computing infrastructure as CNI which could lead to misuse and deter future users of the technology.
- **Support, software development and skills:** While cloud provision may open-up access to new users, it does not eliminate the need for skilled, in-house computing professionals.[32]

The CSP market is currently dominated by three US-based providers: Amazon Web Services (AWS), Microsoft Azure and Google Cloud. A 2019 paper by Hyperion Research, sponsored by AWS, show the extent of market consolidation: *"AWS is the primary CSP for 58% of surveyed HPC user organizations that run workloads in the cloud, more than double the percentage for the number two vendor (23%) and more than seven times higher than the third-place competitor (8%)."*[33]

Despite the challenges mentioned above, cloud computing provides increasingly viable solutions for the lower end of large-scale computing applications. It is a particularly promising access model for users who could not previously justify the large capital investments, inflexibility and related risks associated with other access models. The growth of cloud computing therefore represents one mechanism through which access to large-scale computing by SMEs could be expanded. For example, the Fortissimo project has created a successful marketplace for European SMEs to access large-scale computing resources, expertise, software, and tools on a pay-as-you-go basis.[34]According to 2018 analysis by Hyperion Research[35], sponsored by Google, 74% of large-scale computing sites were found to run some of their workloads in the cloud, compared to just 13% in 2011. However, despite the rapid proliferation, these users still only use the cloud for a relatively small proportion (~10% on average) of their workloads. In the coming years, we can expect cloud computing to increasingly complement and augment, rather than fully replace, existing access models for large-scale computing.[7]

3. The global backdrop

Many nations around the world have seen strategic planning lift their large-scale computing capability. The rapid growth of aggregated data and internet of things (IoT) will form the feedstock for innovation from countries with adequate capacity.

3.1. The international landscape

Large-scale computing is a global endeavour, with most advanced economies making national investments in large-scale computing to support academic research, private sector research and development and national security. Global trends are leaning towards scale up, international centralisation and collaboration of large-scale computing capability. However, there are a number of related issues still to be analysed such as governance, international use agreements, regulation and cybersecurity.

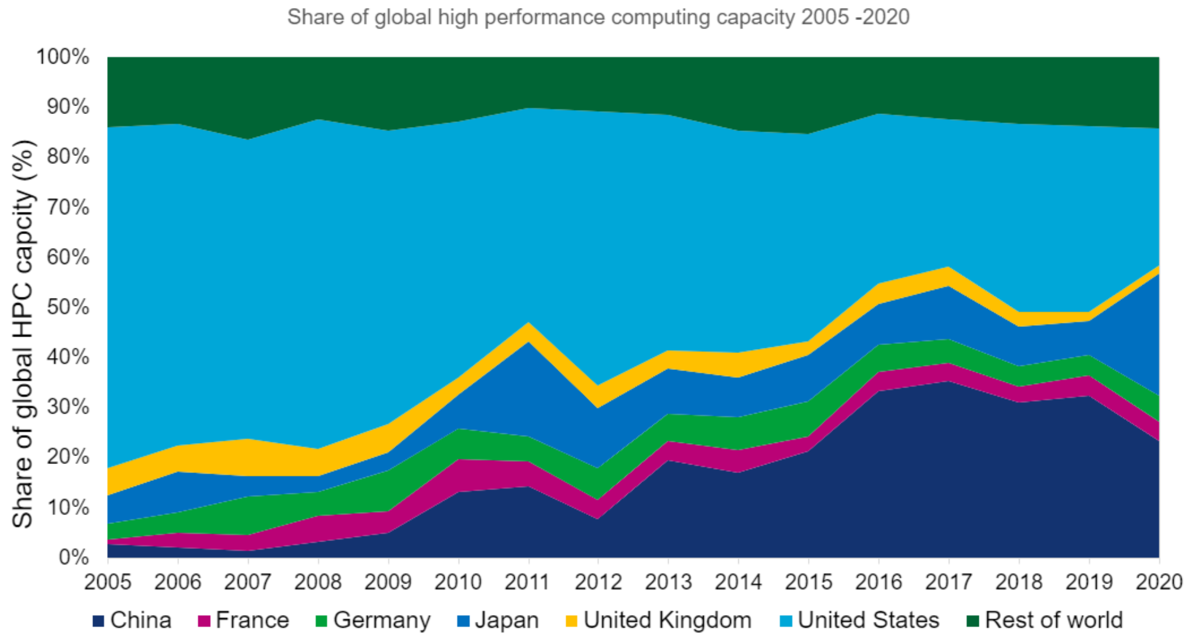


Figure 3: Global high-performance computing capacity (2005 to 2020)[ii]

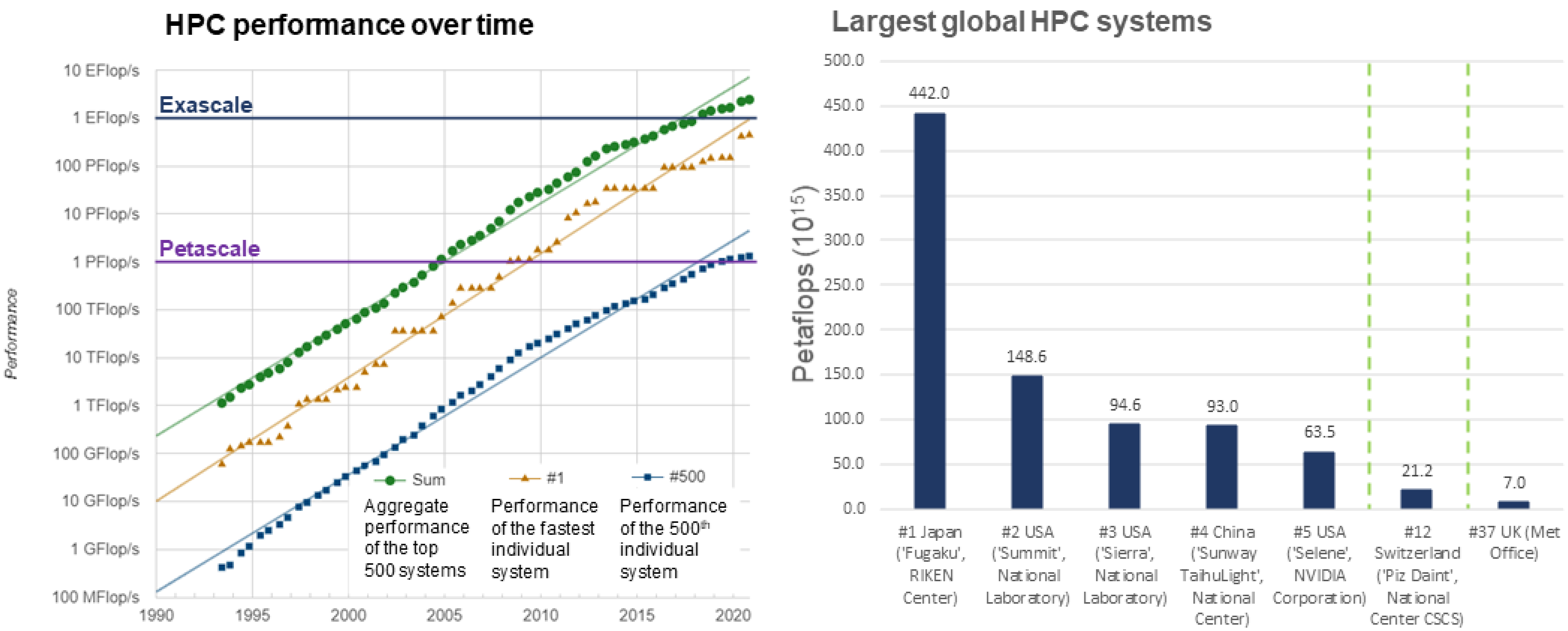
The capabilities of large-scale computing have grown substantially year on year. Since 2005, there has been a 696-fold increase in the aggregate performance of the world's 500 most powerful HPC systems. The world's fastest HPC system, 'Fugaku' at the RIKEN Centre for Computational Science in Japan, has a benchmark performance of 442.0 petaflops,[ii] and is 22 times more powerful than

[ii] Fugaku's R_{MAX} performance is 442.0 x 10^{15} floating point operations per second achieved using the LINPACK benchmark. (See the glossary for description of this metric.) Source: TOP500 (top500.org).

the expected peak performance of ARCHER2 which is set to shortly replace ARCHER as the UK's National Supercomputing Service.

Computing is increasingly seen as an international benchmark for scientific superiority, which has led to an increasing emphasis on ranking and league tables. A key trend of the last decade has been the emergence of China as a major global player in HPC, alongside the US (Figure 3). China has expanded its performance share of the global top 500 systems from 2.6% in 2005 to 32.3% in 2019.

The aggregate performance of the top 500 systems have already exceeded the exascale ceiling and it will not be long before the top system reaches this benchmark. With the current top system being such a step change in performance, the UK and PRACE systems seem to lag behind (see figure 4).



Exponential growth in supercomputing power has occurred over the last two decades. The first Petascale system was realised in 2008. Exascale systems are predicted for 2021/22.

The UK's fastest system, at the Met Office, is the 37th most powerful globally, The UK has access to the Swiss system 'Piz Daint' (ranked 12th) through the PRACE programme.

Figure 4: Largest global HPC systems[iii]

[iii] Figures are calculated from LINPACK benchmark scores (R_{MAX}) for the 500 most powerful HPC systems globally. Source: TOP500 (top500.org), November 2020 review.

3.2. The data revolution

Unprecedented volumes of data in a wide variety of forms are being produced globally. Large-scale computing resources are essential to process and analyse this data to gain insights and extract value. As well as the often-discussed growth of social media data, current and emerging sources of this data include:

- Large science projects like the Large Hadron Collider (LHC) and the Square Kilometre Array (SKA).
- Increased gathering and collation of health data from patient records, medical imaging, and genome sequences.
- Data from IoT sensors and the development of smart networks and cities.
- Growing digitisation of library, archive, and museum materials.
- Increasingly detailed results from scientific and industrial simulations.

3.2.1. New opportunities and challenges

The National Data Strategy sets out our vision to harness the power of responsible data use to boost productivity, create new businesses and jobs, improve public services, support a fairer society, and drive scientific discovery, positioning the UK as the forerunner of the next wave of innovation. One specific opportunity to highlight is that, as a result of the increasing availability and diversity of data, a range of novel application areas for large-scale computing are emerging (see Section 2.2). Meanwhile, the growing availability of computing power is considered a major driver of advances in AI and machine learning, enabling the training of more complicated algorithms on ever growing datasets.[36] These new use cases have potentially significant implications for the large-scale computing ecosystem, changing the balance of computing infrastructure required through an increase in demand for both HPC and HTC systems. Many potential new users may not have the prerequisite awareness or skills to seek out or make use of large-scale computing resources.

As well as increasing demand for resources, in some areas there is also a convergence of big data analytics and AI with more traditional modelling and simulation applications of large-scale computing.[37] Emerging applications may make use of both analysis and simulation concurrently. As data volumes become greater, network and input-output bandwidth may become an increasingly significant bottleneck for making use of large-scale computing resources for data analytics. Having sufficient and readily accessible data storage facilities to enable the long-term preservation and use of valuable data is also essential. The Department for Digital, Culture, Media and Sport (DCMS) have developed a National Data Strategy (NDS) which aims to ensure the security and resilience of the infrastructure on which data use relies.

A move towards more real-time data processing and analysis may help to keep long-term storage and network requirements more manageable.[38] Edge computing (computing carried out near the source of data) is likely to play an important role in enabling these developments and the real-time insights they could provide.[39] These developments may see large scale systems embedded within distributed networks of computing resources (for example, IoT sensors, edge processors) where one of the key challenges will be "data logistics".[40]

3.2.2. Privacy, security, and governance implications

The NDS sets out the importance of safeguarding responsible data use and ensuring the security and resilience of our data assets, alongside seeking to capitalise on the opportunities and yield the benefits associated with it. There are clear public benefits accrued by the analysis of large volumes of data by users across academia, industry, and the public sector. However, the sensitive nature of much of this data, such as patient records or genome sequences, raises concerns around its maintenance and transfer. Cybersecurity and encryption will become more important to protect data and give users reassurance.

This includes exploring and implementing new approaches for data governance and data sharing. Edge computing could improve security by enabling more distributed networks which are more resilient to attack or disruption and reduce both the movement of sensitive data and the need to hold it in centralised databases.[41] Authentication, Authorisation and Accounting Infrastructure (AAAI) can enable secure and transparent access to complex and geographically distributed datasets through federated authorisation and authentication systems. The UKRI Infrastructure Roadmap included AAAI as a key consideration for e-infrastructure.[42] The Government has committed to publishing a new National AI Strategy later this year, which will aim to build on the excellent progress the UK has made through the AI Sector Deal, in order to realise the potential economic, productivity and societal opportunities presented by AI. The widespread implementation of these federated networks under a common framework could greatly improve the ability to extract value from sensitive data.

3.3. Developments in computing hardware

3.3.1. Exascale computing

The current largest global HPC systems operate in the 'petascale' range (10^{15} to 10^{17} floating point operations per second (FLOPs)). Exascale systems are those that are capable of 10^{18} floating point operations per second. Exascale systems will lead to a step change in capability for specific applications through a substantial increase in capacity.

While milestones such as exascale are useful, principal consideration should be given to the use cases of the infrastructure (i.e., the real-world problems that the system will solve) and the improvement they offer over previous systems, rather than simple benchmarks of hardware performance. Most systems achieve a small percentage of their theoretical maximum performance under real-world testing.[iv]

The first exascale systems are expected to be running in 2021 or 22, with China, Japan, the European Union, and the US all developing exascale initiatives. The European High-Performance Computing Joint Undertaking (EuroHPC) is the European Union's exascale initiative. Operating from 2018 to 2026, three pre-exascale systems (including one with a theoretical peak performance of 550 petaflops[43]) and five smaller petascale systems are planned, with two exascale systems being delivered circa 2023.[44] The UK is not a member of EuroHPC.[45]Achieving exascale capability will require sustained efforts in developing the necessary skills, software and infrastructure. The US Exascale Computing Project includes funding for software and application development on a similar scale to – and in coordination with – their hardware initiatives.[46] Networking, storage and archiving will have to be scaled-up to account for the increased data required and produced. New models of data movement, processing and storage may be needed to prevent new bottlenecks. An emerging example is the link between digital twin technology and digital threads – the bank of data used when developing a digital twin. This technology needs warehousing of data that is not yet available in the UK. To take full advantage of digital twins, the UK will need to develop a Tier-0, Big Data Centre that will curate data assets either as part of a system for data analytics, or as a long-term archive.

Improvements in power efficiency will also need to be achieved; extrapolation of power consumption from current top-tier systems would lead to untenable energy costs.[47] The US Department of Energy (DOE) has set a cap of 40MW for its own exascale projects.[48] This power demand also has implications for both running costs (each 1MW of power demand adds c. £1 million per year to running costs) and environmental impact (see Section 5.3.6). Electrical grid infrastructure is also a

[iv] The LINPACK benchmark is the most commonly quoted measure of the performance capability of systems. It is based around solving a system of linear equations, a common, but simple, task which is not fully representative of the operations performed in scientific computing. The High-Performance Conjugate Gradients (HPCG) benchmark is an alternative measure of HPC performance. It uses computations and data access patterns that are more representative of those used in scientific research. On average, systems achieved only c. 2% of their LINPACK performance using the HPCG benchmark. Source: TOP500 (https://www.top500.org/hpcg/lists/2019/11/).

key consideration, as large amounts of power are required for single sites, which can limit where systems can be located.

Expected uses for exascale systems[49]

Exascale systems will lead to a step change in capability for specific applications, both dramatically improving existing computational work and allowing for expansion into previously intractable areas. The increased capacity of exascale systems enables higher resolution modelling of more complex systems over longer time periods, better uncertainty characterisation and the processing and analysis of very large datasets.

Some of the areas which will benefit from exascale systems are:

- **Weather and climate modelling:** the production of more accurate, more reliable and longer-term forecasts and climate models at a greater granularity of scale.
- **Quantum mechanics:** more precise modelling of particle phenomena and the exploration of previously intractable subatomic physics problems, such as calculating neutrino mass.
- **Materials science advances:** increasing the accuracy, size and timescales of materials simulations, enabling modelling of previously unfeasible systems.
- **Tackling astrophysics questions:** simulation at the scale of cosmological events and simulation of gravitational processes which could help solve fundamental questions about the universe.
- **Digital twins:** enabling the detailed simulation of digital replicas of complex physical objects or processes, such as entire aircraft engines, to optimise performance while reducing development and testing costs.[50]
- **Understanding fundamental biology**: high-resolution modelling of molecular interactions and processes, supporting drug development and enzyme design.

3.3.2. Diversification of hardware technologies

The power of computer processors has increased exponentially over the last 50 years. However, fundamental physical limits in transistor density are now being reached. Computing approaches have become increasingly tailored to application needs, with processors being designed for particular uses.[51]

This trend is leading to diversification in system architectures. Heterogenous systems, which use more than one kind of processor, are becoming more commonplace with hardware accelerators, such as Graphics Processing Units (GPUs), Tensor Cores or Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs), increasingly being used to augment CPUs (see Appendix C for more information). 29% of the global top 500 HPC systems use GPU accelerators.[52] GPUs are particularly adept at performing machine learning tasks.[53]

Other new computing paradigms are emerging which may introduce additional diversity into the hardware landscape such as:

- **Neuromorphic computing** which mimics the function of the human brain through the use of electricity 'spikes' rather than constant current. It is particularly adept at modelling the human brain and performing machine learning tasks.
- **Quantum computing** uses the properties of quantum mechanics to perform computations. It has a number of potential applications across machine learning, as well as in cryptography, modelling quantum systems, and solving linear systems of equations.

We can expect to see increased diversification in system makeup as hardware becomes less general purpose and is more tailored to the application in question. This is likely to further complicate procurement decisions, with systems increasingly being built to solve specific types of problems and transfer of workloads between systems becoming more difficult.

Hardware diversification also has implications for software development. Employing different types of hardware accelerators requires development of new software or extensive updates to existing software. This trend towards increased hardware diversification will require software that can run more flexibly across different architectures. Co-design of large-scale computing systems, while difficult to implement, offers many benefits in terms of delivering stakeholder-driven design goals, significant reduced risk of non-performance and ultimately a better end result constituting better value for money. They also deliver additional benefits in terms of supporting the UK skills-building agenda in key digital areas and providing opportunities to innovate in areas such as sustainability.

3.3.3. Green computing

Globally, running computing infrastructure consumes vast quantities of energy. Carbon emissions are generated from the production and disposal of large-scale computing hardware, and during operation through the powering and cooling of systems. The proportion of overall greenhouse gas emissions from ICT infrastructure is expected to increase sharply over the next few decades.[54,55]

Driven by technological developments, the power efficiency of computation has historically doubled every year and a half.[56] However, as the scale of systems has grown, so too has their energy footprint. El Capitan at the Lawrence Livermore National Laboratory, one of the planned US exascale systems, will have an energy budget of 30 to 40 MW;[57] equivalent to the electricity consumption of c. 70,000 to 90,000 UK homes.[58].

While representing only a small relative percentage of global power consumption and emissions, there is growing interest in how the expansion of large-scale computing can continue to be delivered in a manner that is sustainable and in line with Paris Agreement targets. An example of this is the recent procurement by the Met Office for a supercomputer from Microsoft which will deliver 18 times the capability of the current system over ten years and will be run on 100% renewable energy. The Green500 list tracks the top 500 supercomputers in the world ranked by energy efficiency rather than computing power.[59]

There are several means through which energy consumption and waste can be minimised, such as optimising resource utilisation and innovative cooling infrastructure. These approaches, as well as reducing carbon emissions, can also save money. The Royal Society's '*Digital Technology and the Planet'* report highlights that the technology sector is in a position to lead by example and manage its own carbon footprint. Not only can large-scale computing optimise utilisation, it can also contribute to the decarbonisation of the energy grid through scheduling – allocating tasks to times of peak renewable energy generation. With the UK hosting COP26 and assuming presidency of the G7, there is an opportunity to demonstrate leadership in the digitisation of the net zero transition.

3.3.3.1. Optimising utilisation

Optimised utilisation means that more research can be delivered for the same carbon footprint, thus the research output per tonne of carbon dioxide increases. Examples of how this could be achieved include improved software development practices (see Section 5.3.4); minimising data movement and unnecessary storage; dynamic workload scheduling (including "bursting" workloads to the cloud during spikes in demand)[60]; and ensuring resource allocation is matched to workload requirements.

A range of research avenues are also being explored to improve system efficiency, such as:

- Designing **novel system architectures** to reduce energy usage. The Horizon 2020-funded ECOSCALE project is studying integrating FPGAs (see Appendix C) and CPUs in architectures that reduce data transfer requirements.
- **Mixed-precision computing** (see Appendix C) can reduce energy consumption by minimising the wasteful use of overly precise calculations.[61]

- **Flexible applications,** which can tolerate a variable power supply could allow the use of renewable energy to power systems.[62] This approach could both reduce running costs and assist with electricity grid management.

It is worth noting that employing the above approaches requires either developing new software or making extensive updates to existing programs.

3.3.3.2. Cooling infrastructure

Large-scale computing produces significant amounts of heat during operation. Energy is continuously required to cool systems to keep them within their operating temperatures; for traditional air-cooled systems this is a significant proportion of the overall power consumption of these systems. There are a number of strategies that can be employed to reduce the power costs and environmental impact of cooling:

- Infrastructure can be cited in locations where there is a surplus of renewable energy.[63]
- Improvements can be provided by continuous monitoring and use of analytical techniques such as machine learning.[64] The layout of computing infrastructure can also be designed for thermodynamic efficiency, which can be aided by computer simulation.[65]
- Warm wastewater from system cooling can be used in district heating schemes.

4. The UK's position in the global landscape

Investment in hardware alone cannot bring about the most effective use of systems. The UK is a leader in software development and sought after, skilled professionals. Despite this, demands outweighs supply and the workforce lacks the diversity required to produce software that encompasses the needs of society.

The UK has a reputation as a highly skilled and innovative economy which make it an attractive place to carry out research and to do business. This is underpinned by strong legal and regulatory frameworks, as well as protections for intellectual property. A world-class national large-scale computing ecosystem can build on this foundation, enhancing the attractiveness of the UK for research and business investment. Meanwhile, a strong position globally can ensure that UK priorities in terms of privacy, cybersecurity and openness are incorporated into international standards and initiatives. It can also help strengthen the UK's role in helping to tackle pressing global challenges, for example: by supporting research into climate change, pandemics, and extreme weather events.

The UK's large-scale computing ecosystem is complex and interlinked, with a wide range of users and providers. It relies on a diverse and accessible physical infrastructure base, with large-scale computing systems of varying size and type. High-quality software and skilled workers are necessary to ensure that these systems can run effectively.

The UK is recognised for its capabilities in computer science and software development, and has strengths across several computing domains, including AI, computational fluid dynamics and bioinformatics. The UK has a highly skilled large-scale computing workforce, however demand for computing professionals currently outstrips supply.

Within HPC, the UK currently possesses 1.4% of overall capacity world-wide (Figure 5). The UK's largest system, at the Met Office, is the 37th most powerful in the world.[66] The UK has access to the 12th largest global system, based in Switzerland, through the research consortium PRACE. On the hardware side, the UK supply chain is limited, with few domestic vendors or hardware suppliers.

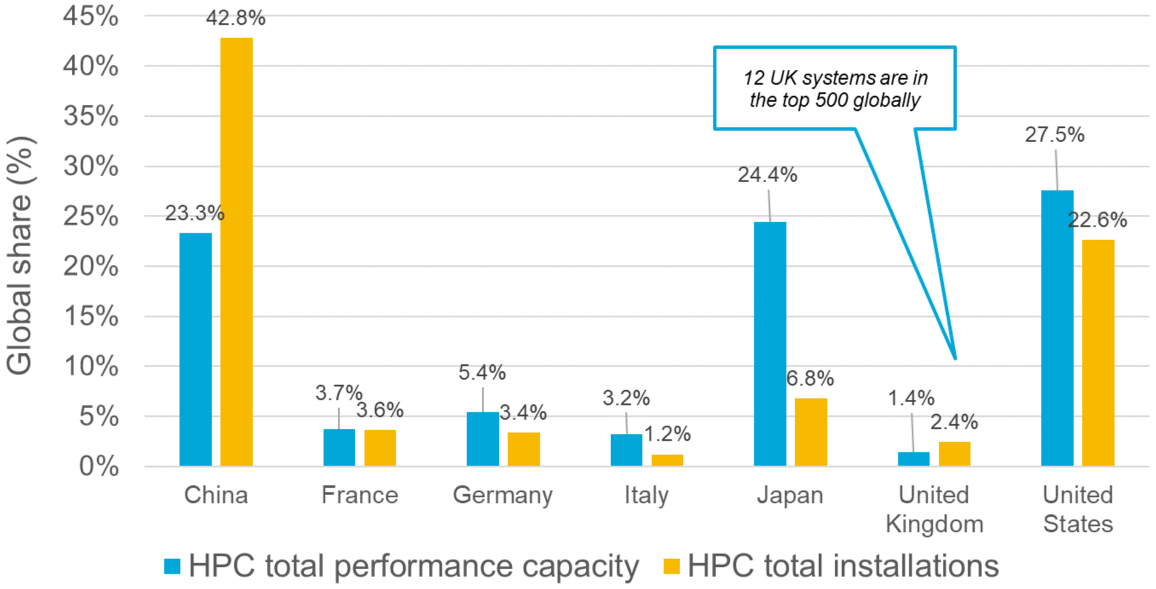


Figure 5: Share of global HPC capacity (selected countries)[v]

[v] Figures are calculated from LINPACK benchmark scores (R_{MAX}) for the 500 most powerful HPC systems globally, November 2020. Source: TOP500 (top500.org).

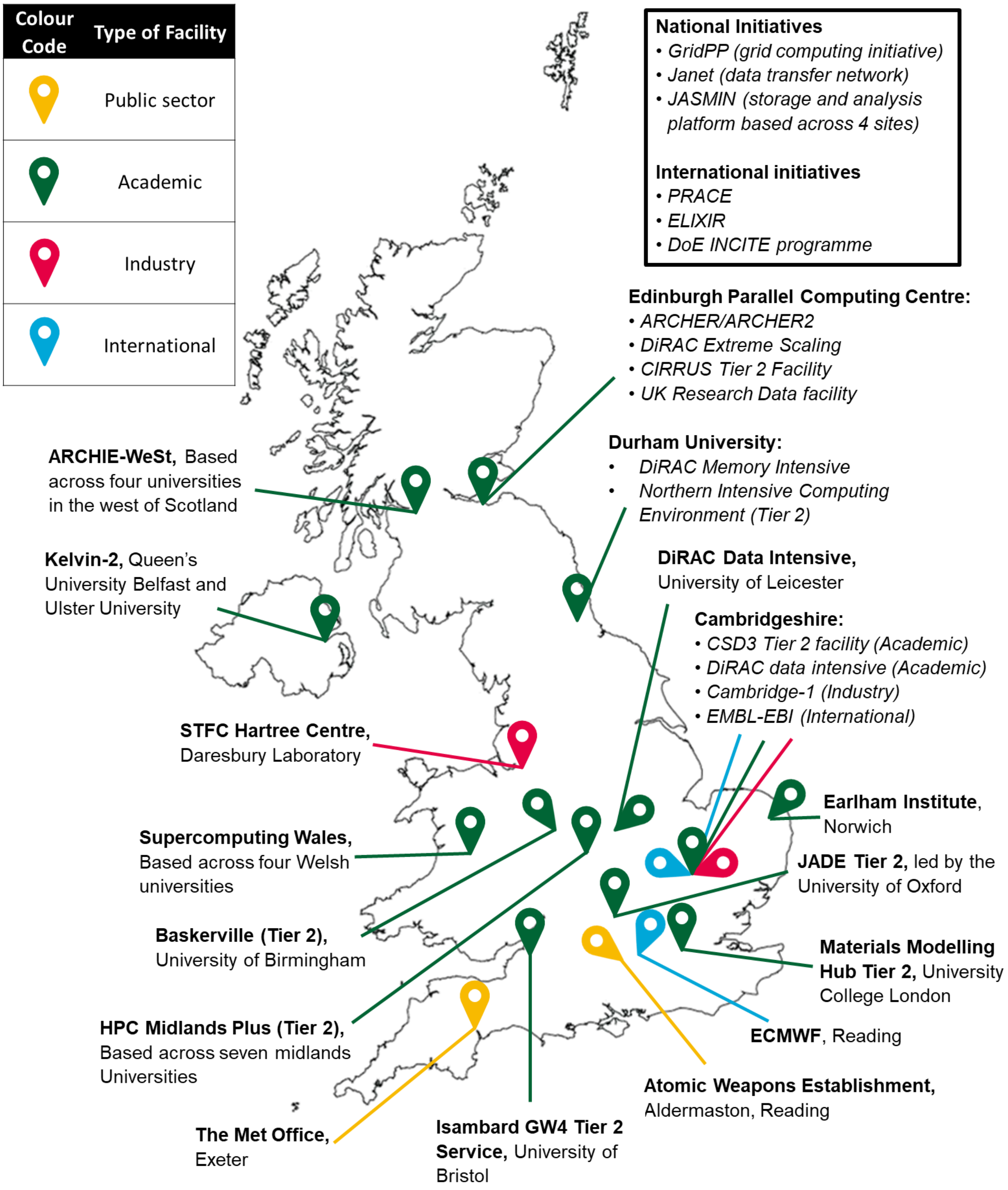

Colour Code
Type of Facility
Public sector
Academic
Industry
International
National Initiatives
• GridPP (grid computing initiative)
• Janet (data transfer network)
• JASMIN (storage and analysis platform based across 4 sites)
International initiatives
• PRACE
• ELIXIR
• DoE INCITE programme
Edinburgh Parallel Computing Centre:
• ARCHER/ARCHER2
• DiRAC Extreme Scaling
• CIRRUS Tier 2 Facility
• UK Research Data facility
Durham University:
• DiRAC Memory Intensive
• Northern Intensive Computing Environment (Tier 2)
DiRAC Data Intensive, University of Leicester
Cambridgeshire:
• CSD3 Tier 2 facility (Academic)
• DiRAC data intensive (Academic)
• Cambridge-1 (Industry)
• EMBL-EBI (International)
Earlham Institute, Norwich
JADE Tier 2, led by the University of Oxford
Materials Modelling Hub Tier 2, University College London
ECMWF, Reading
Atomic Weapons Establishment, Aldermaston, Reading
Isambard GW4 Tier 2 Service, University of Bristol
The Met Office, Exeter
HPC Midlands Plus (Tier 2), Based across seven midlands Universities
Baskerville (Tier 2), University of Birmingham
Supercomputing Wales, Based across four Welsh universities
STFC Hartree Centre, Daresbury Laboratory
Kelvin-2, Queen's University Belfast and Ulster University
ARCHIE-WeSt, Based across four universities in the west of Scotland

Figure 6: UK large-scale computing landscape map. The map includes UK supercomputers in the global top 500 and other large-scale computing infrastructure of a national or regional significance.

4.1. Current UK infrastructure

The UK's large-scale computing infrastructure comprises a range of systems of varying sizes and hardware architectures to support a range of different applications (Figure 6). Large-scale computing infrastructure comprises not just computing systems, but also storage facilities, networks, and other supporting facilities.

Large-scale computing infrastructure can broadly be divided between public sector, academic and industry systems. However, these boundaries are fluid, and many sharing arrangements exist. The UK also has access to international large-scale computing systems through partnerships and sharing agreements. Large-scale computing infrastructure comprises not just systems, but also storage facilities, networks, and other supporting facilities.

A summary of some of the UK's key large-scale computing infrastructure is outlined below. Many universities, companies and research institutes also host their own in-house facilities which are not listed. UKRI has recently carried out a landscape analysis of the UK's research and innovation infrastructure, which includes large-scale computing.[67]

4.1.1. Public sector

Some of the largest large-scale computing users are in the public sector and their facilities represent critical national infrastructure.

- **Met Office:** The Met Office has the UK's fastest supercomputer alongside two smaller systems for resilience.[68] Portions of these computers are made available to the academic community.[69] A new system has been approved and expected to come online in 2022 and deliver a six-fold increase in capability with a further three-fold uplift planned later in the 2020s.[70]
- **Atomic Weapons Establishment (AWE):** AWE runs HPC to support nuclear warhead stewardship.[71]

In addition to these larger systems, smaller high-security systems also exist at the Defence Science and Technology Laboratory (DSTL) to support a wide range of diverse defence and security workflows. The Ministry of Defence is keen to investigate opportunities for extending use of these systems to UK SMEs and academic organisations to better exploit the UK's technical expertise for sensitive workflows.

4.1.2. Academic research infrastructure

A range of national and regional systems are in place to facilitate access for users across the UK academic community. A selection is described below:

- **National large-scale computing systems:**
    - **ARCHER2:** Forms the general-purpose national supercomputing service for industry and academic researchers in the UK based at the Edinburgh Parallel Computing Centre (EPCC). The transition from ARCHER to ARCHER2 is currently underway. This system is designed to be "capable on average of over eleven times the science throughput of ARCHER".[72]
    - **DiRAC:** Provides more targeted large-scale computing services for the Science and Technology Facilities Council (STFC) theory community with a range of tailored computational infrastructure based across four sites for data-intensive, memory-intensive and "extreme-scaling" workloads.[73]
    - **EPSRC Tier 2 systems:** The Engineering and Physical Sciences Research Council (EPSRC) currently funds a mixture of Computing Centres of Excellence to provide a diversity of computing architecture at a mid-level of performance capability. In the latest round in 2019 £34 million was invested in nine centres:[74]
        - Cambridge Service for Data Driven Discovery (CSD3) (led by the University of Cambridge).
        - The Materials and Molecular Modelling Hub (led by University College London).
        - JADE the Joint Academic Data Science Endeavour (led by the University of Oxford).
        - Sulis: An EPSRC platform for ensemble computing delivered by HPC Midlands+ (led by the University of Warwick).
        - GW4 Tier 2 HPC Centre for Advanced Architectures (led by the University of Bristol).
        - Northern Intensive Computing Environment (led by Durham University).
        - Baskerville: a national accelerated compute resource (led by the University of Birmingham).
        - Cirrus Phase II (led by the University of Edinburgh); and,
        - Kelvin-2 (led by Queen's University Belfast and Ulster University).

- **Regional large-scale computing systems:**
    - **Supercomputing Wales:** Provides computing facilities for academic users from Cardiff, Swansea, Bangor, and Aberystwyth Universities. Part-funded by the European Regional Development Fund (ERDF) through the Welsh Government, with support from university partners.[75]

  - **ARCHIE-WeSt:** Funded by EPSRC and operating in partnership with the Universities of Glasgow, Glasgow Caledonian, West of Scotland, and Stirling, it provides large-scale computing services to academia, industry and the public sector in the west of Scotland.[76]

- **HTC facilities:**
  - **Earlham Institute:** A Biotechnology and Biological Sciences Research Council (BBSRC) funded research institute focussed on genomics and computational biosciences. It hosts a range of large-scale computing resources tailored to the specific needs of bioinformatics workflows.[77]
  - **EMBL-EBI:** The European Bioinformatics Institute (EMBL-EBI) is a UK-based component of the European Molecular Biology Laboratory (EMBL) and aims to "help scientists realise the potential of 'big data' in biology". As well as hosting a wide range of databases, the institute also possesses significant HTC capacity, some of which is made available to the research community through a range of analysis tools.[78]
  - **GridPP:** GridPP is the UK's contribution to the Worldwide Large-Hadron Collider Computing Grid (WLCG) to allow the processing and analysis of data arising from particle physics experiments at CERN. Funded by STFC and established in 2001, it brings together computing resources and expertise across 18 institutions and has allowed the UK to play a leading role in the WLCG.[79]

- **Data transfer, storage, and analysis infrastructure:**
  - **JASMIN:** A storage and analysis platform based across four sites provided by the Natural Environment Research Council (NERC) and STFC for climate and earth sciences applications.[80]
  - **Janet network:** Using 8,500km of fibre optic cabling, Janet provides high-speed networking to the academic community.[81] Janet is connected to other similar research networks across Europe through GÉANT (the Gigabit European Academic Network).[82]
  - **UK Research Data Facility (RDF):** Based at the EPCC, the EPSRC- and NERC-funded facility provides high-capacity disk and tape storage for data from national large-scale computing facilities, ensuring its long-term preservation.[83]

4.1.3. International access

The UK currently has access to a range of world-class large-scale computing systems through international agreements. Some of the key partnerships are below.

- **PRACE:** The Partnership for Advanced Computing in Europe has 26 member countries and provides shared large-scale computing infrastructure across Europe.[84] Access to leading-edge systems is allocated based on scientific merit. The programme is being continued through 2021 but its long-term future, and how it will relate to the European High-Performance Computing (EuroHPC) Joint Undertaking, remains uncertain.
- **ELIXIR:** The European life-sciences infrastructure for biological information (ELIXIR), with its central hub on the Wellcome Genome Campus, Cambridge, brings together life-science research institutions from across Europe to provide support in sharing, storing, and analysing biological data.[85] This includes a dedicated scheme of work to integrate large-scale computing infrastructure for life-sciences across Europe as part of a federated system.[86]
- **ECMWF:** The European Centre for Mid-range Weather Forecasting is a multinational collaboration currently based in Reading, whose primary role is to provide forecasts to member states in the 10-day range. It hosts two systems in the UK, although this large-scale computing capacity is moving to Bologna, Italy.
- **Department of Energy INCITE Program:** Access to the world's most powerful supercomputers at the US's DOE network of national laboratories is available to UK researchers and companies.[87] Applicants must demonstrate their work tackles problems at the frontiers of science and engineering, and that their software can effectively scale to make use of large-scale computing systems.

4.1.4. Industry

Some industry users maintain in-house large-scale computing capacity, making use of individual large-scale systems, grid computing approaches or cloud access. Industry users may also obtain access to the above public systems, either through bidding for use on the basis of scientific merit or by paying for system time.

Many industry users work with the **Hartree Centre,** an STFC-funded centre tasked with helping UK industry to take advantage of large-scale computing opportunities. The Hartree Centre offers access to computational resources including a data analytics platform, two GPU accelerated hybrid systems for big data or AI applications and Scafell Pike, an HPC system capable of delivering 1.8 petaflops of performance (R_{MAX}, LINPACK benchmark).[88] It also provides training, develops software tools, and engages in collaborative R&D to support uptake in industry.

NVIDIA, inventor of the GPU is currently building a new supercomputer called Cambridge-1 which promises to achieve 400 petaflops of "AI performance" through its GPU accelerators and 8 petaflops of LINPACK benchmark performance. The system is intended to form part of its AI Centre of Excellence in Cambridge and is expected to be operational by the end of 2021. It will focus on supporting joint industry research targeted at AI applications in the health and life sciences sectors while machine time will also be donated to university researchers and start-ups.[89]

4.2. Software and skills

4.2.1. The UK software base

The UK has the largest software industry in Europe. The UK software industry contributed to direct value-added GDP of £70.3 billion in 2016, and it directly employs nearly 700,000 workers (c. 2.2% of the UK workforce).[90]

It is difficult to overemphasise the interdependent nature of software, skills, and hardware. Without the necessary skills, good software cannot be written or maintained. Without the necessary software, hardware cannot be made use of effectively. The UK is recognised for its capabilities in computer science and software development, alongside strengths in the research domains that make use of large-scale computing. 92% of academics make use of some type of research software, with 69% regarding it as fundamental to their work.[91]

Producing software for large-scale computing requires the use of specialist programming languages and tools, as well as constant updating and optimising to ensure software can run on the latest systems. With the increased scale and diversity of hardware, commonly used software will require significant updates to utilise future generations of supercomputers. Ensuring software is sustainable and will continue to be useable, available, and efficient in the future is critical to ensuring benefit from large-scale computing. This requires the development and embedding of best practice in software development across the computing ecosystem.

A range of initiatives funded by UKRI exist to tackle issues of software sustainability, several of which are described below:

- **CoSeC** (Computational Science Centre for Research Communities) supports the development and improvement of research software and undertakes training and outreach to share knowledge and expertise.[92]

- The **Software Sustainability Institute**, funded by several UK universities and UKRI, works to promote best practice in research software development.[93]
- **Collaborative Computational Projects** (CCPs) harness UK expertise to develop, maintain and distribute the software code which is essential for scientific computing.[94] CCPs help to ensure that the UK's software base keeps pace with a rapidly changing computing landscape so that UK researchers can take advantage of available hardware infrastructure to tackle research challenges.
- **Embedded Computational Science and Engineering** (eCSE) projects support the optimisation and further development of many software applications on the national large-scale computing services by research teams across the UK.

The existence of these initiatives points to a recognition of the importance of sustainable, high-quality software. However, for the next generation of systems, including exascale systems, popular UK software will require substantial updating or rewriting to make use of these new capabilities. Thus, programmes to prepare the software base need to begin several years before any new system comes online. The ExCALIBUR Programme, led by the Met Office and EPSRC, aims to redesign high-priority codes and algorithms for exascale systems.[95] Funding for the ExCALIBUR Programme is £45.7 million over five years.[96]

4.2.2. Large-scale computing skills

Good software requires skilled software developers. Good software engineers not only produce high-quality and efficient software, but also ensure the reliability and reproducibility of research through good archiving and updating practices. The UK has a strong skills base in areas such as the science of computing and software development fed by an internationally renowned university system. Many research software engineers (RSEs) come from either computer science or other scientific disciplines and often possess advanced degrees.[97] They cite the ability to work at the forefront of science as a strong draw to the profession.[98]

International competition for individuals with these skill sets is fierce and they can attract lucrative salaries from private sector employers. For RSE's, the widespread use of short-term temporary contracts and lack of a defined career track can lead to retention issues for these workers in academic settings.[99]

Several existing initiatives in the UK aim to improve retention in academia:

- The EPSRC has developed a **Research Software Engineer Fellowships** scheme to support and provide better career pathways for RSEs.[100]

- The **Society of Research Software Engineering** exists to improve recognition of research software engineers and push for more defined career paths. They also organise conferences, workshops, and webinars for the RSE community.[101]
- The emergence of **research software engineering groups** within universities can offer a means of improving job security and expanding access to RSEs by researchers. In RSE groups, engineers are employed on a longer-term basis by a university and then loaned out (typically at cost) to individual research groups.[102]

4.2.3. Diversity in the large-scale computing workforce

A diverse workforce is crucial to ensure a robust sector supported by a range of perspectives and experiences. However, diversity in the large-scale computing workforce is understudied, with limited data available.

A survey by the Software Sustainability Institute found that RSEs, which make up part of the large-scale computing workforce, fell behind on gender, ethnicity, age and disability diversity.

Of the UK research software engineering workforce in 2018:[103,104]

- 15% were female (compared to 48% of the wider UK workforce[105]).
- 21% were over 45 years old (33% of the wider UK workforce are over 49 years old[106]).
- 5% were from ethnic minorities (compared to 12% of the wider UK workforce[107]).
- 6% had a disability (compared to 19% of the working age population[108]).

Very few initiatives exist to address diversity issues in the large-scale computing workforce specifically. One global network, established by the EPCC, which seeks to improve gender diversity within the sector is 'Women in High Performance Computing', which provides fellowship, education, and support to women in large-scale computing as well as support to the organisations who employ them.[109]

More broadly, initiatives which look to increase diversity and foster inclusion in STEM fields, or in the IT sector will likely impact on the large-scale computing workforce. These can include:

- Increasing uptake of STEM, computer science and data related subjects by under-represented groups in primary, secondary, and tertiary education. For example, 'Gender Balance in Computing' aims to increase the number of girls choosing to study a computing subject at GCSE and A level while

the charity 'Code Club' works to broaden opportunities for young people to learn to code through free after-school clubs for 9 to 13-year-olds.[110]
- Providing alternative entry pathways to careers in large-scale computing including traineeships and apprenticeships. For example, the STFC apprenticeship programme which includes engineering, computing and ICT apprenticeships.[111]
- Improving job retention, pay rates and career progression of employees from under-represented groups. For example, the Athena SWAN and Race Equality charters which encourage and recognise racial and gender equality amongst staff and students in higher education and research. [112,113]

4.3. National coordination

The UK large-scale computing community has various formal and ad-hoc coordination groups. The collaborative projects and initiatives highlighted in Section 4.2, among many others, help to ensure that the wide range of actors within the UK ecosystem are well-connected and talking to each other. UKRI is playing an active role in promoting a more coordinated ecosystem and has recently created the position of Director of Digital Research Infrastructure to provide strategic leadership in the space.

Various committees, technical steering groups and advisory boards exist to address the concerns and needs of different user groups. Complemented by more informal relationships, these groups promote collaboration and coordination throughout the ecosystem. These groups include:

- **IRIS**, a consortium to help coordinate the e-infrastructure facilities and programmes of STFC.[114] IRIS coordinates approaches to provision of physical infrastructure, facilitates resource sharing and provides information on e-infrastructure requirements to STFC.
- **The HPC Special Interest Group** aims to encourage awareness of HPC, promote best practice and share knowledge across the UK academic community. The group has members from 44 universities and 17 affiliate research centres and organisations.[115]
- **The High-End Computing Consortia** are seven consortia funded by EPSRC to cover different topics within EPSRC's remit. These provide resources to members on making use of ARCHER, the national HPC facility, as well as support for software development and a community network to share knowledge.[116]

A large number of Government departments and arm's length bodies have some responsibility for or involvement with large-scale computing. No single team within Government has overall responsibility for coordinating policy on large-scale computing.

4.4. The UK supply chain

Recent mergers and buyouts have led to consolidation in the large-scale computing marketplace, with most of the industry based outside of Europe. This dependence on single supply chains leaves the ecosystem exposed to shocks whereas a wider variety of computing techniques builds resilience and provides users with confidence. Of the twelve UK HPC systems in the top 500 globally, eleven use Intel processors and six were built by HPE Cray. By contrast to the US and China, which place strategic importance on developing domestic microprocessor capacity, the UK has a very limited domestic supply chain. The EU's exascale initiative, EuroHPC, also places emphasis on developing the European computing hardware sector.[117]

Despite its limited presence in the microprocessor market, the UK is a significant consumer of computing resources, which gives it some influence in the global large-scale computing supply chain. The UK also has some presence in the hardware market for large-scale computing and is playing a leading role in the development of novel computing paradigms and architectures:

- **Large-scale computing system vendors**
  - A number of international large-scale computing vendors have a presence in the UK. This includes **HPE Cray**, **Intel**, **IBM** and **Atos**.[118,119] Some vendors host R&D facilities in the UK.

- **Semiconductors**
  - **ARM Holdings**, a semiconductor design company, maintain their headquarters in the UK, although they were acquired by a Japanese technology firm in 2015.[120] ARM processors are of growing interest for use in large-scale computing systems and are used in the Japanese Fugaku system, currently the world's fastest large-scale system.[121]
  - ARM, along with several other semiconductor designers, does not fabricate its own chipsets but instead licenses its IP to semiconductor foundries. The majority of global semiconductor foundries are based in China and south-east Asia.
  - The **Compound Semiconductor Applications Catapult** works to support the development, prototyping and application of novel compound semiconductor technologies.[122]

- **Novel computing architectures**
  - UK start-up **Graphcore**, based in Bristol, develops hardware accelerators for AI and machine learning.[123]
  - Several UK companies have established a strong reputation in using FPGAs for computing applications. For example, the start-up **AccelerComm** is developing FPGA-based solutions for 5G applications. Basingstoke-based **Omnitek** was a successful developer of visual processing applications using FPGAs before being acquired by Intel in 2019. Edinburgh-based **Alpha Data** have been working in this area for almost 30 years and produce FPGA accelerator cards for many vendors.[124][125][126]
  - The **SpiNNaKer** project at The University of Manchester is pioneering the development of Neuromorphic computing architectures using ARM-designed processors.[127]

5. Achieving the UK's potential in large-scale computing

For investments to be effective, one must consider the interconnected nature of users, software, skills, the hardware supply chain and future systems. Putting in place a national coordination function will ensure that investments are in line with a long-term, strategic roadmap.

5.1. The value of large-scale computing

Large-scale computing underpins many of the fields where the UK has comparative strength, including bioinformatics, climate research, materials science, pharmaceuticals, machine learning and fluid dynamics. Driven by advancements in data science, modelling and AI, large-scale computing is being employed across an increasing range of disciplines where the UK is an early world leader.

Large-scale computing can help to tackle major societal challenges and enhance the UK's global influence. The UK's involvement in the COVID-19 High Performance Computing Consortium, an international initiative pooling computational capacity in service of projects which address the challenges of the pandemic and recovery, is a key example of this. UKRI leads the UK involvement in this consortium. Meanwhile, elements of large-scale computing, such as the Met Office's weather modelling, form part of the UK's CNI. Effective use of information from the Public Weather Service contributes an estimated £1.5 billion to the economy per year, in addition to lives saved.[128]

Large-scale computing can enhance business competitiveness, by enabling faster and cheaper product design and testing. For example, large-scale computing is used in the automotive industry to reduce product development times.[129] Access to leading-edge computing capabilities, and associated support, can make a place attractive to invest in and do business. Large-scale computing is an enabling tool for research and innovation. Given the range of direct and indirect benefits from these projects, it can be difficult to arrive at robust figures concerning the return on investment for large-scale computing. Furthermore, the time lags between investment in scientific research and the realisation of commercial and/or societal benefit can be longer than those often considered in an econometric time series.[130] More generally, there is well-established evidence that R&D has a positive impact on productivity and economic growth.[131, 132] There is also strong evidence that public investment in R&D stimulates private sector expenditure on R&D.[133,]

5.2. The case for change

Major economies around the world have developed long-term national strategies around large-scale computing and are investing significantly in their computing capabilities. Future exascale systems will lead to step changes in performance capacity and capability (see Section 3.3.1). Exascale computing will help to solve problems which are currently intractable, such as whole system modelling of jet engines and other complex digital twins. A number of countries are already building these systems or have announced plans to do so.

We are entering an era of unprecedented hardware diversification, driven by the development of new computing architectures. Many of these architectures are tailored for specific applications, such as novel hardware accelerators for machine learning (see Appendix C). For users, it is increasingly important to have access to a diverse range of architectures.

The cost of the largest leading-edge systems is greater than any single institution or business can afford. From our engagement, a number of users have reported difficulties accessing UK large-scale computing facilities. Issues include the limited overall capacity of the UK ecosystem and the sometimes restrictive access requirements for using public systems.

Increasingly, HTC and HPC computing are being performed in the cloud. This can make large-scale computing accessible to organisations that are unable to afford the high capital costs of purchasing systems. However, cloud access models create a range of other new challenges (see Section 2.3.2).

These global trends create important strategic questions about the types of national investments to make. However, at present, there is no UK-wide roadmap for large-scale computing, and there is no single team within Government to provide coordination.

5.3. Recommendations

Through engagement with sector stakeholders, we have identified seven key building blocks for a computing ecosystem that meets users' needs.

Our recommendations on the following pages are grouped into these seven areas:

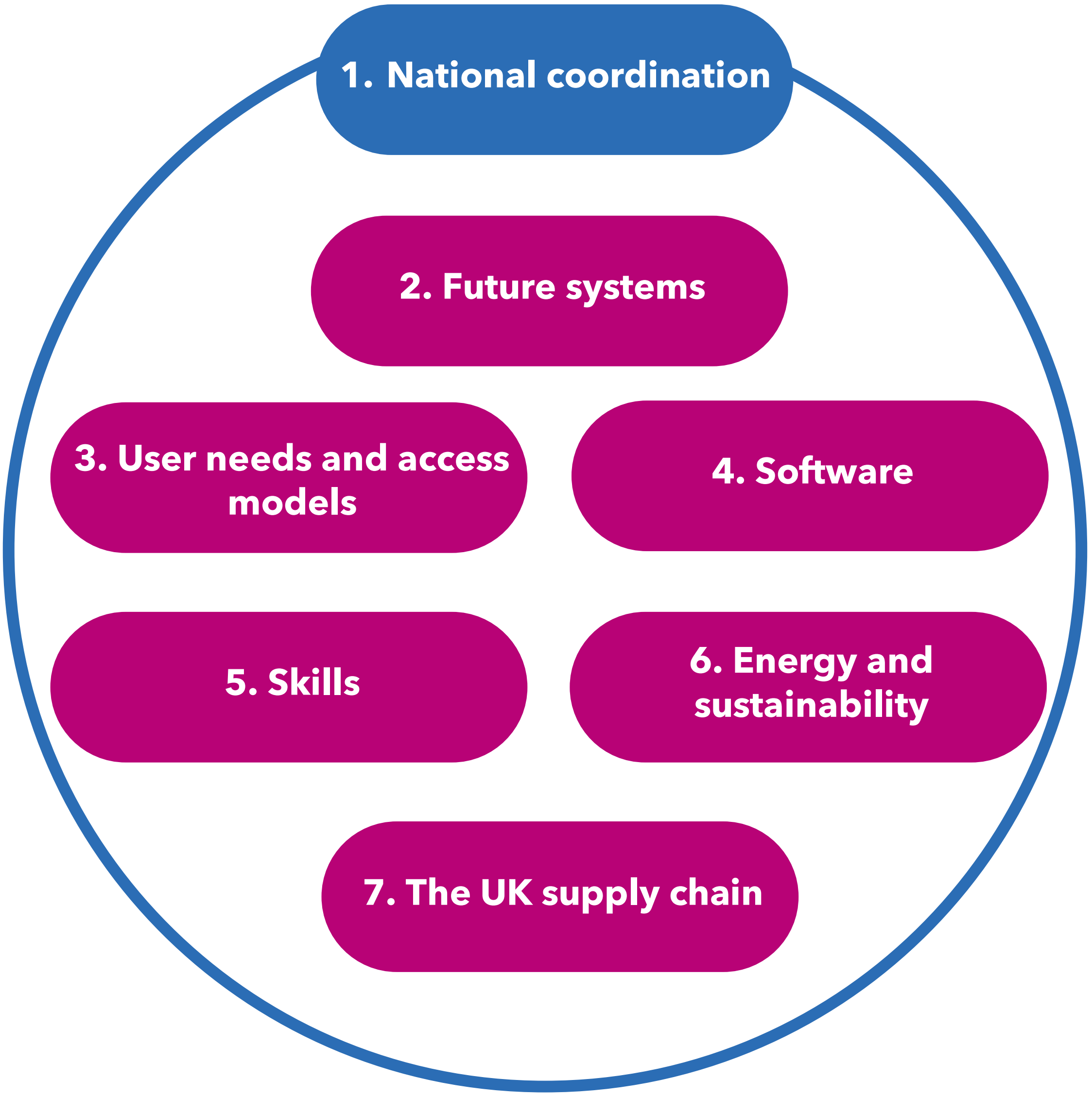


Figure 7: Schematic diagram demonstrating the importance of national coordination and how it underpins the remaining 6 recommendations.

The first section discusses national coordination, which is our central recommendation. Developing improved coordination mechanisms underpins all of the other recommendations in this report. New investment alone will not be able to deliver the improvements to the ecosystem described in the following sections. Without strong coordination, the UK risks procuring systems that do not fully meet users' needs, or those systems could lack the necessary skills, software and supporting infrastructure to use them effectively.

5.3.1. Challenge 1: National coordination

The UK's approach to computing is uncoordinated, introducing inefficiencies into procurement and limiting sharing of resources.
Digital research infrastructure, of which large-scale computing is one component, forms part of the UK's national infrastructure. Many of the key opportunities and challenges around computing span multiple sectors – common issues around skills, software and cybersecurity need to be considered in a joined-up way. Numerous projects and initiatives within the UK ecosystem seek to promote collaboration and coordination in order to address common issues and requirements across diverse users (see Sections 5.3.2 to 5.3.7).

Government has a significant role to play in nurturing and supporting the UK ecosystem both as a consumer and funder. However, at present there is no single team within Government that provides coordination or leadership. This limits strategic thinking across the whole UK ecosystem, including initiatives to support software, skills and the industrial base.

Across Government departments, procurement and investment practices are very varied and are often carried out in an ad-hoc way, without a cohesive overarching strategy. Large-scale computing procurements tend to be considered individually by officials without specific expertise in the area and without a view of the wider UK landscape. With the current global trend towards diversification in hardware (see Section 3.2.2), there is a risk that the Government could over-invest or under-invest in certain architectures without this broader view of the ecosystem. Government lacks a single, expert function that can support departments in scrutinising computing business cases.

- **We recommend establishing a team within Government that has policy responsibility for large-scale computing.**
    - This team would have **ownership of policy for** large-scale computing, working to address the opportunities and challenges identified in following sections (5.3.2 to 5.3.7).
    - This team would be responsible for **developing a rolling long-term roadmap** for the large-scale computing ecosystem. The roadmap should cover both supply and demand-side considerations.
    - This team should comprise of **both civil servants and secondees** from the public sector, industry and academia. It should provide a forum for major users to share ideas and keep up with international trends. To be effective, this team would need to have knowledge of both Government procurement processes and maintain domain expertise.

- This team could play a role advising departments and agencies on procurement, helping Government to become **a more intelligent customer** of computing.
- This team could also play a role in analysing the **security and resilience** of the large-scale computing ecosystem and **providing policy guidance** to Government.

5.3.2. Challenge 2: Future systems

The UK does not currently have a strategic plan for accessing world-class computing, including exascale.

Access to world-class computing systems is essential for maintaining international competitiveness in academia and industrial R&D. At present, the UK has only 1.4% of the world's HPC performance capacity and does not host any of the top 25 most powerful systems globally.[134]

UK users can currently gain access to leading-edge systems through programmes including PRACE (in Europe) and the US's DOE INCITE Program (see Section 4.1.3). Access to internationally hosted systems is often subject to international agreements, and computing time is often allocated on the basis of scientific merit. The UK currently pays a subscription via EPSRC for access to PRACE.

In addition to hardware, key to achieving strong performance from large-scale computing is efficient software (see Section 5.3.4). Skilled professionals are also essential, including software developers, system architects and administrators (see Section 5.3.5).

Many leading economies are developing exascale computing projects and installing pre-exascale systems now with performance in the hundreds of petaflops (see Section 3.3.1). There are long lead times involved in procuring large-scale systems and putting in place the infrastructure, skills and software to support them. It is therefore essential that the UK makes long-term, strategic commitments and sets out a clear roadmap up to and beyond the advent of exascale computing capabilities in the UK. If the UK is to deliver an exascale system in the early-2020s a decision will need to be made imminently.

➢ **The UK needs to maintain access to world-class computing capability.**

  - The public sector, researchers and academics require access to leading-edge computing systems. This can be achieved through a balance between international collaboration and procuring new domestic systems – becoming a UK partner for bilateral international computing strategies where applicable.
  - The coordination function should carry out regular capacity and use analysis for hardware, software, services and skills.
  - The UK should develop a national plan for developing exascale computing capability to ensure it can reap the benefits exascale systems will bring (see Section 3.3.1).

- To achieve the full benefits of next-generation systems, a long-term roadmap going beyond hardware is vital (covering skills, software and user engagement).

The hardware landscape is rapidly diversifying.
A wide range of architectures and hardware accelerators are increasingly used across large-scale computing (see Section 3.2.2 and Appendix C). Many of the largest global systems make use of both conventional CPUs and GPUs. Specialist hardware, such as FPGSs and TPUs are increasingly used for AI and machine learning applications. New computing paradigms such as probabilistic, quantum and neuromorphic computing are also emerging.

It is very difficult to predict which of these technologies, if any, will become most prevalent. It is likely that hardware will become increasingly tailored for specific applications. In the US, the DOE strategically invests in a range of architectures to ensure diversity in the market.

Given these trends, it is increasingly important for the UK to have access to a diversity of hardware architectures. Architecture diversity supports software developers by creating platforms for testing, evaluation and capability building. EPSRC's Tier-2 programme performs this function for the UK academic community.[135]

- **The UK must respond to the diverse and dynamic hardware landscape.**
  - Government should undertake regular horizon scanning in order to continuously respond to the latest technology developments. The coordination body proposed in Section 5.3.1. could fulfil this function, supported by UKRI, the research community and industry.
  - The UK should embrace diversity in the UK computing ecosystem and should strategically pursue access to a variety of architectures to meet diverse user needs, and to allow testing and evaluation of different systems.

The growing volume and diversity of data requires the hardware infrastructure to store, transfer and analyse it.
The growing volume and diversity of data being produced globally presents many opportunities (see Sections 2.2 and 3.2). Yet, these opportunities cannot be realised without the computing infrastructure to store, process, analyse and thus extract insights from this data. Most of these infrastructure needs can feasibly be dealt with by smaller, localised systems within individual organisations, businesses, and universities or through cloud-provision (see Section 5.3.3). In other cases, such as training computationally intensive AI algorithms, they may be able to make use of existing or forthcoming national facilities.[136]

However, there will be specific areas where the UK Government may want or need to intervene strategically to ensure the large-scale computing resources are in place to take advantage of these opportunities. One such area is where the data volumes produced are particularly large, sensitive and/or valuable, for example large science projects like the SKA or in health (for example, genome sequence data). The UK may also benefit from intervention in cases where users are not otherwise making optimal use or their data or where valuable data may be lost. For example, the Edinburgh International Data Facility provides secure data storage and analysis platforms, mostly through a "data service cloud", to various users as part of the city region deal.[137]

- **The UK must ensure that it has the hardware in place to reap the benefits of the data revolution.**
  - Horizon scanning should be used to keep a watching brief of opportunities and challenges in terms of the use of large-scale computing for data processing and analysis.
  - The UK should ensure that, where appropriate, its existing and future national computing infrastructure continues to be open to new and emerging data-driven use.
  - When necessary, the Government should invest strategically in large-scale computing infrastructure in accordance with a data services strategic roadmap for large-scale computing to support emerging data-driven applications. These investments may include data analytics platforms, high-capacity networks, data storage or archive facilities across the UK and should be complemented with supportive access models (see Section 5.3.3).

5.3.3. Challenge 3: User needs and access models

Users have a diverse range of needs from large-scale computing.
Computing needs vary substantially across sectors, with users having different requirements in terms of both capability and support requirements (see Section 2.1):

- In the public sector, service reliability is a key consideration. For example, the systems used for weather modelling at the Met Office require a very high-level of reliability and are continually in use. For public sector users, cybersecurity and the physical locations of infrastructure are other key considerations.
- Academia requires access to systems of a range of sizes and architectures owing to the diversity of programs run. Many problems across certain fields, from cosmology to quantum chromodynamics, drug design to nuclear fusion, will require access to the largest systems (including future exascale systems).
- Industrial users of large-scale computing use a wide range of access models (including on-premises systems, cloud computing and public infrastructure). Private sector use of public systems often includes a requirement to publish results, which can act as a barrier to industry access, where intellectual property protection is a major concern. Cybersecurity is another key concern for industrial users.
- SMEs often do not have an awareness of the range of business problems that large-scale computing can solve or the technical requirements to use it. In addition, SMEs often require support in adapting their software to run on large systems.

➢ **The long-term roadmap for computing should reflect the wide range of different needs of users.**

  - Identifying users with similar requirements is vital for making good investment choices. This ensures that systems are used by the widest range of potential users. Flexible funding mechanisms are required to achieve this, as are mechanisms that encourage sharing of resources.
  - The UK should invest strategically in a range of system architectures to support varying user requirements (see Section 5.3.2). The EPSRC Tier-2 centres offer a good example of this.
  - The UK should explore mechanisms to encourage potential SME users to engage with large-scale computing. The Prime Minister's Council for Science and Technology has outlined a number of recommendations on harnessing science and technology for economic benefit, including:

- mapping local science, technology and innovation assets.
- establishing an Innovation and Growth Place Fund; and,
- reviewing R&D tax credits to include technological developments and R&D for software.[138]

Many smaller users have difficulty accessing large-scale computing, or they may not understand the benefits it can provide.

Many potential users have a limited understanding of how large-scale computing could support their work. Increased engagement is needed to improve awareness of the benefits of large-scale computing. Different user groups require different levels of support:

- Experienced users, such as large engineering and technology companies, will be more comfortable with "raw" access to large-scale computing systems.
- Smaller users, such as SMEs, may require more assistance with understanding the capabilities and limitations of large-scale computing and writing parallelised software. The STFC Hartree Centre and the EPCC are involved in initiatives to encourage SMEs to use large-scale computing.

➢ **Government should seek to minimise barriers that prevent new users from accessing large-scale computing.**

  - Access models need to be flexible to ensure that smaller users can access national computing infrastructure. The UK should consider offering targeted support to academics and SMEs in using large-scale computing.
  - Active mapping and engagement with potential users should be pursued and barriers that prevent SMEs from accessing public large-scale computing systems reviewed, including access models and requirements to publish results.
  - Providing researchers with avenues to seek support from large-scale computing professionals can help to encourage diffusion of key skills (see Section 5.3.5).
  - Greater support should be given to academics to convert scientific software designed for small-scale systems, so that they can run efficiently on large-scale systems.

Cloud computing offers an easy access model for users, but it has limitations.

Cloud-based large-scale computing has become an increasingly popular access model (see Section 2.3.2). As an on-demand service with a pay-as-you-go structure, it eliminates the often prohibitive upfront costs of computing infrastructure and shifts costs to ongoing operational expenses.

Cloud computing offers the ability to flexibly use different architecture types and capacity levels, allowing for more agile and variable computing use. For infrequent users, or those with highly variable workloads, it can be more cost effective. Cloud computing may therefore expand access to large-scale computing to a wider range of applications and users, such as SMEs. However, outsourcing computing to a CSP does not eliminate the need for in-house expertise or external support to use it effectively. For example, the virtualisation of computational workloads required to make use of cloud services can be a complex task.

Large-scale computing in the cloud is still an emerging access model and several challenges remain. It is unable to address the needs of the most demanding users, who require access to the very largest systems. Cloud provision also involves transferring data off-site, which can be impractical for very large workloads or introduce an unacceptable risk for certain particularly sensitive applications.

As described in Section 2.3.2, the CSP market is currently dominated by three US-based providers. The standard model of cloud computing includes significant costs for data egress (transferring data off of a cloud platform) which can make it prohibitively expensive to switch providers. Significant technical barriers also make transferring workloads between CSPs difficult.

- **The UK should take a pragmatic approach to cloud computing.**
  - Academic grants should have flexible terms to allow users to choose between procuring their own systems and subscribing to cloud platforms.
  - The UK should actively explore areas where cloud computing can expand access to large-scale computing to new users and applications.
  - Moving to the cloud does not eliminate the need for in-house computing expertise or external support. Managing cloud computing workloads requires skilled staff.
  - The UK should explore mechanisms to ensure that users can readily switch between cloud access models and vendors.

5.3.4. Challenge 4: Software

Popular software is not ready for next generation systems.
High-quality software is fundamental to using hardware effectively. The UK has strengths in software development, across areas including machine learning, climate science and software for financial services (FinTech). The speed of hardware development means software must continually be updated to run efficiently on the latest hardware. In particular, the diversification of hardware creates new challenges for software engineering.

Much of the software run on large-scale computing will require significant updates to run efficiently on next-generation hardware. Current software will not be capable of harnessing the full potential of upcoming systems, including exascale computing. Some initiatives exist to tackle this issue (see Section 4.2.1), including the ExCALIBUR Programme, which aims to redesign high-priority codes and algorithms for exascale systems.[139] In the US, the Exascale Computing Project has taken a long-term view of this challenge, and they have made substantial investments in software over a ten-year period to ensure that software will run efficiently on exascale systems.[140]

The UK has strength in many emerging areas of computing, including quantum and neuromorphic computing. These new computing paradigms will require very specialised software to be developed concurrently with hardware.

- **The UK must update its software base to realise the benefits of next-generation hardware and become a centre for excellence.**
  - The increasing diversification of hardware creates a need for more specialised software.
  - Investments in next-generation systems should be coupled with associated investment in developing and modernising software to run on these systems. For example, software that is designed specifically to take advantage of exascale systems.
  - Active dialogue and coordination between the user community and hardware manufacturers should be encouraged to ensure their efforts are aligned.

Software design practices in academia require modernisation.
In academia, the roles of software developer and researcher are traditionally performed by the same individual. This can lead to poor software development practices, such as lack of rigorous testing, updating and archiving. This can result in less efficient programs, which take longer to run. Poor archiving practices

sometimes result in software becoming lost, which threatens the ability to reproduce and verify results, a process fundamental to scientific research.

Several initiatives and organisations within the UK ecosystem, such as the Software Sustainability Institute, are already working to tackle these issues (see Section 4.2). Greater separation of the roles of domain scientist and programmer, and professionalisation of the software development role, may also help address the problem. Specialist skillsets are required to develop the highly parallel programs used in large-scale computing. Having RSEs embedded in research groups can help improve software quality and ensure it is well-maintained.

➢ **Best practices in software development should be embedded within research methods.**

  - Independently tested and well-archived software is fundamental to producing high-quality and verifiable results.
  - Good software design practices should be encouraged across disciplines through increased awareness and training. These practices should seek to ensure that software is accessible, updated and archived.
  - RSEs in academia can help to professionalise software development practices (see Section 5.3.5).
  - Efforts should be made to expand and enhance existing UK initiatives already addressing these issues (see Section 4.2).

5.3.5. Challenge 5: Skills

Career pathways for computing professionals within academia and the public sector are not well-defined.

Software development and administration of large-scale computing systems are complex tasks that require highly skilled workers (see Section 1.2.2). Skills such as the ability to write parallelised code, manage complex job flows and maintain systems are fundamental to the running of large-scale computing. Efficient software and system operation are essential for maximising value from large-scale computing.

The UK has a strong reputation for producing skilled workers in large-scale computing fields, however private sector competition for workers results in difficulties retaining staff in academia and the public sector. Stakeholders have described challenges in retaining mid-career professionals in the public sector and academia. It is essential to make sure that career pathways and incentives are in place to retain people with the necessary skills.

The public sector and academia struggle to provide salaries that are competitive with the private sector. In academic fields, RSEs do not have a clear career path, progression opportunities are often limited, and the use of temporary contracts is common. EPSRC Research Software Engineer Fellowships[141] and the Society of Research Software Engineering[142] are working to overcome these challenges (see Section 4.2.2).

- **Steps should be taken across academia and the public sector to retained large-scale computing professionals.**

  - Staffing costs should be fully accounted for during procurement to ensure facilities have adequate staffing for the life of a computing system.
  - The UK public sector and academia should consider new career frameworks, fellowships, base lines and salary structures to retain large-scale computing administrators and research software engineers.
  - Establishing RSE groups within universities (see Section 4.2.2) offers a model for broadening access to research software engineers and improving job security.

The UK skills base is not large enough to keep up with the growing demand for large-scale computing professionals.

While the UK produces highly skilled computing professionals, they are in extremely high demand. This leads to the difficulties in retaining the types of roles noted above. Ensuring the UK can fully capitalise on the benefits of large-scale computing requires nurturing and growing the overall skills base. The Institute of Coding, funded by the Department for Education, works across universities and employers to address digital skills gaps.[143]

Obtaining large-scale computing skills is strongly centred around formal education. Often these skills are gained through computer science degree courses or obtained by students over the course of a PhD project. This has resulted in a highly skilled, but small, workforce. To take advantage of opportunities provided by large-scale computing, the UK will need to nurture a balanced and diverse workforce of talented individuals through an appropriate range of entry-levels, qualification types and career pathways.

Domain scientists need to have the relevant skills to be able to map scientific problems to computational solutions. It is also important that they work closely with RSEs and other computing professionals to produce efficient and reliable software.

There is little data on the diversity of the large-scale computing workforce, however, the data that exists suggests the large-scale computing workforce is substantially less diverse than the overall UK workforce. While a small number of initiatives exist to address gender diversity in large-scale computing, and many exist to address gender diversity in technology, IT and STEM professions, very few initiatives exist to target other aspects of workforce diversity.

- **The UK needs to expand and diversify its large-scale computing skills base.**

    - Engagement with universities can help ensure that relevant topics are included in university curricula and help to raise students' awareness of careers in large-scale computing.
    - Non-degree entry routes and retraining opportunities should be explored and supported, such as apprenticeships, trainee schemes, conversion courses and continuing education.
    - The UK should ensure that it can attract and retain international talent.
    - The UK should consider further initiatives to improve the diversity of its large-scale computing workforce. The UK should seek-out improved

data on the diversity of its large-scale computing workforce to inform these decisions.

5.3.6. Challenge 6: Energy and sustainability

Large-scale computing has substantial energy requirements, which place limits on where large systems can be built.
Large-scale computers use large amounts of power both to process data and for system cooling. ARCHER, the UK's National Supercomputing Service, draws just under 2 MW of power when running at full capacity, equivalent to c. 4,700 homes.[144] The power demand of large-scale systems can put significant strain on local energy grids. Upcoming systems will require close integration with local and national grid planning to ensure their power needs can be accommodated.

Energy usage has increased with each successive generation of large-scale computers, and the exascale systems currently in development in the US will be subject to a cap on electricity usage in the 20 to 40 MW region.[145] While an exascale system would be a significant regional electricity consumer, power consumption would be relatively small when compared to overall UK electricity demand (in the region of 30,000 to 40,000 MW). The Horizon 2020 ECOSCALE project, which includes UK university partners, aims to reduce the power consumption of next generation HPC systems.[146]

- **The UK must ensure that it can meet the power infrastructure requirements of large-scale computing.**
  - A long-term national roadmap would help to ensure that power requirements of large-scale computing are integrated into energy grid planning.
  - Business cases for large-scale computing infrastructure should consider whole-life costs, including power consumption.

Advances in large-scale computing could affect efforts to decarbonise the electrical grid.
Intelligent siting of computing facilities can help to reduce their environmental footprint. Many global technology companies are locating data centres in Nordic countries due to a growing surplus of renewable hydropower and a cooler climate, reducing the cost of cooling.[147] With appropriate infrastructure, waste heat has the potential to be repurposed in district heating schemes.

The combined power demands of the processors, data storage, networking and cooling systems for large-scale computers can be substantial. A single exascale system in the 40MW range would consume roughly 0.1% of the UK's current electricity supply, the equivalent of the domestic consumption of c. 94,000 homes. The UK has a commitment to achieve net zero emissions by 2050. While the UK energy grid is rapidly decarbonising, it will continue to require non-renewable sources (particularly natural gas) for at least the next decade.[148]

Green computing, which seeks to minimise the environmental impact of computing, offers a means of mitigating this (see Section 3.2.1). Power consumption (both from computation and cooling) and the environmental cost of producing and disposing of hardware (e-waste) should be considered. UKRI's recent Environmental Sustainability Strategy commits the organisation to integrate environmental sustainability criteria into all procurement, capital, and infrastructure investment decisions.[149]

Large-scale computing procurements should include life cycle assessments of environmental impact. It is important that these assessments consider the full range of effects, including those that are a result of outsourcing to CSPs and other contractors. Government could play an important role in developing templates for these types of assessment, building on the relevant international standards (ISO 14040:2006 and 14044:2006).[150,151]

- **The UK must ensure that advances in computing are compatible with green electricity targets.**
    - Procurement decisions should seek, where possible, to mitigate the environmental impact of new computing infrastructure. This could be achieved through improving computing efficiency, intelligent siting of systems and increased use of renewable power.
    - Users should be incentivised to make efficient use of large-scale computing resources, for example by restricting core-hours, sharing and pooling of resources or rewarding efficient applications.
    - Large-scale computing procurements should include environmental life cycle assessments.
    - Engagement with large-scale computing vendors, as well as innovation grant competitions, could help to encourage developments in energy efficient computing, as well as stimulating the UK supply chain.

5.3.7. Challenge 7: The UK supply chain

The global large-scale computing market is highly consolidated, which reduces competitiveness.
Recent mergers and buyouts have led to consolidation of the large-scale computing supply chain. Of the twelve UK HPC systems in the global top 500, six were built by HPE Cray and eleven use Intel processors. World-wide, 459 of the 500 most powerful HPC systems use Intel processors.[152] The small number of vendors poses risks to the overall health of the global supply chain and limits competition. In the US, the DOE strategically invests in computing hardware from a number of different manufacturers to hedge risks and prevent vendor lock-in.

- **The UK should make efforts to ensure a healthy and competitive marketplace for hardware.**

    - Procurement practices can be designed to encourage innovation within the supply chain. System specifications should avoid over-specifying the hardware make-up of systems, instead describing the real-world problems that systems should solve. The Office for AI's procurement guidelines provides advice on this type of challenge-led procurement.[153]

The UK's hardware industry is fairly limited.
The small domestic supply chain for computing limits the ability of the UK to influence global hardware development. The vast majority of hardware is procured from outside the UK, and currency fluctuations can inflate costs.

One significant UK hardware vendor is Graphcore, which produces hardware accelerators for AI.[154] ARM Holdings, a leading semiconductor design company, was acquired by a Japanese technology firm in 2015, although their R&D operations remain based in the UK.[155]

It is likely not feasible for the UK to be able to compete in all areas of the supply chain. For example, entering into the field of semiconductor fabrication would require a multi-billion-pound investment in a semiconductor foundry, which would be difficult to justify. However, the UK has the potential to become a leader in certain areas of the market. One area is the development and prototyping of novel compound semiconductor technologies (see Section 4.4).[156] Other areas where the UK is well-placed to capitalise on future commercialisation opportunities include hardware acceleration and novel computing paradigms (such as neuromorphic computing and quantum computing). These areas represent new and growing markets.

- **The UK should explore initiatives to support its domestic large-scale computing industry.**

  - Setting a long-term roadmap for large-scale computing would provide clarity to the supply chain, helping the domestic hardware sector develop.
  - The global market for HPC is forecast to grow to £38 billion by 2022.[157] A stronger domestic sector would be better placed to capitalise on these export opportunities.
  - Initiatives should be considered to support UK hardware development and manufacturing. This could include investments in accelerators and testbeds for benchmarking, design and testing.
  - Areas of UK strength, such as hardware accelerators, novel computing paradigms and compound semiconductors, should be supported through to commercialisation.

Appendix A: Glossary of terms

Term	Definition
Artificial intelligence	A broad area of computer science concerned with the use of computer systems to carry out functions analogous to human cognition.
Authentication, Authorisation and Accounting Infrastructure (AAAI)	A security framework enabling secure and transparent access to complex and geographically distributed datasets through federated authorisation and authentication systems.
Central processing unit (CPU)	The unit within a computer which contains the circuitry necessary to execute software instructions. CPUs are suited to completing a wide variety of workloads, distinguishing them from other types of processing unit (for example, GPUs) which are more specialised.
Cloud bursting	Dynamic allocation of computing workloads to the cloud to deal with spikes in demand.
Cloud computing	A rapidly emerging access model where computing infrastructure is accessed on-demand via the internet.
Computer hardware	The physical components of a computer such as the processors, storage devices and physical connections or wiring.
Computer processors	The physical hardware that computers use to carry out calculations, manipulations and computations as instructed by software code.
Computer software	The written instructions that are necessary for computer hardware to carry out calculations, manipulations and computations.
Cloud service provider (CSP)	A company that provides computing resources as a service through the cloud.
Digital twins	A computer model of a physical entity or system (for example, a human brain, or a jet engine). The digital twin is updated from real-time data gathered from sensors on the physical object.
Edge computing	An approach to computing which involves data storage and computation close to the 'edge' of the network, nearer to users and sources of data. This can reduce response times and make large data volumes more manageable.
Exascale	A high-performance computing system that is capable of at least one Exaflop per second (i.e., a system that can perform more than 10^{18} floating point operations per second).
Field-programmable gate arrays (FPGAs)	Integrated circuits designed to be dynamically configured and reconfigured by a customer after manufacture.
FLOPS	Floating Point Operations per Second, i.e., the number of calculations involving rational numbers that a computer can perform per second.
Graphics processing unit (GPU)	A type of computer processor designed to accelerate image processing and other applications. GPUs now augment CPUs in some large-scale

	computing systems to accelerate certain highly parallel computational tasks.
Green computing	The environmentally responsible use of computing resources, often focussed on energy efficiency and lowering the environmental impact of manufacturing and disposal.
Grid computing	The use of a widely distributed network of computing resources, brought together as a 'grid', to achieve a single objective. Due to its distributed nature, these systems tend to be used for high-throughput, easily parallelisable workloads.
Hardware accelerators	Novel forms of computer processors which can perform certain specialised tasks faster and more efficiently than standard CPUs, thereby accelerating computation (see Appendix C).
High-performance computing (HPC)	A form of large-scale computing where parallelisation is used to deliver high overall capability and processing speed. Often applied to the modelling and simulation of complex systems.
High-throughput computing (HTC)	A form of large-scale computing where parallelisation is used to deliver high overall processing capacity. Often applied to the analysis and processing of large volumes of data.
Internet of Things (IoT)	The interconnection of large numbers of devices via the internet allowing them to exchange data. IoT devices often contain sensors to allow them to gather information on physical systems. Processing data from large numbers of IoT devices often requires large-scale computing.
Large-scale computing	Computer systems where processing power, memory, data storage and network are assembled at scale to tackle computational tasks beyond the capabilities of everyday computers. Often involves the widespread use of parallelisation (see separate entry). An umbrella term encompassing terms such as high-performance computing, high-throughput computing, supercomputing and novel computing paradigms.
Machine learning	The use of computer algorithms to produce mathematical models from input or "training" data. Machine learning is often considered a form of artificial intelligence.
Memory	Typically used to refer to a specific type of temporary data storage, for holding information for immediate use by computer processors.
Monte Carlo simulation	A widely used technique for uncertainty characterisation where simulations are re-run many times while randomly varying uncertain input parameters. This enables uncertainty in input variables to be accounted for and enables probabilistic outputs from deterministic modelling.

Network	The digital infrastructure which connects computing devices to allow the exchange of data.
Neural networks	A type of machine learning approach that is inspired by the behaviour of neurons in the brain.
Neuromorphic computing	An emerging paradigm in computing where hardware architectures are designed taking inspiration from the neural structure of the brain (also see Appendix C).
Parallelisation	Parallelisation is where large computational problems are divided up into many smaller ones which can then be run concurrently on separate processors.
Petascale	A high-performance computing system that is capable of at least one Petaflop per second (i.e., a system that can perform more than 10^{15} floating point operations per second).
Probabilistic computing	An alternative to conventional, deterministic computing where computation is performed using stochastic simulation and probability distributions.
Quantum computing	An experimental form of computing which utilises the probabilistic qualities of quantum mechanics. It has the potential to massively accelerate certain computational tasks (also see Appendix C).
Research software engineers (RSEs)	Specialists working within research establishments who combine scientific domain expertise with computer programming skills.
Small and medium-sized enterprise (SMEs)	Defined (in the UK) as businesses with fewer than 250 employees or a turnover of less than £25 million.
Software sustainability	An approach to software development which ensures that software is made accessible, improved and continually supported in a manner that maintains its long-term value into the future.
Tensor cores or tensor processing units	An AI hardware accelerator designed for machine learning tasks using neural networks (defined above).
Virtualisation	A mechanism by which a virtual version of computing components is created, with users seeing an abstraction of the physical component. For example, they appear to use a single computer which is in fact a composition of numerous systems. Multiple abstractions, or virtual machines, can run on a set of computational resources, for example, the cloud.

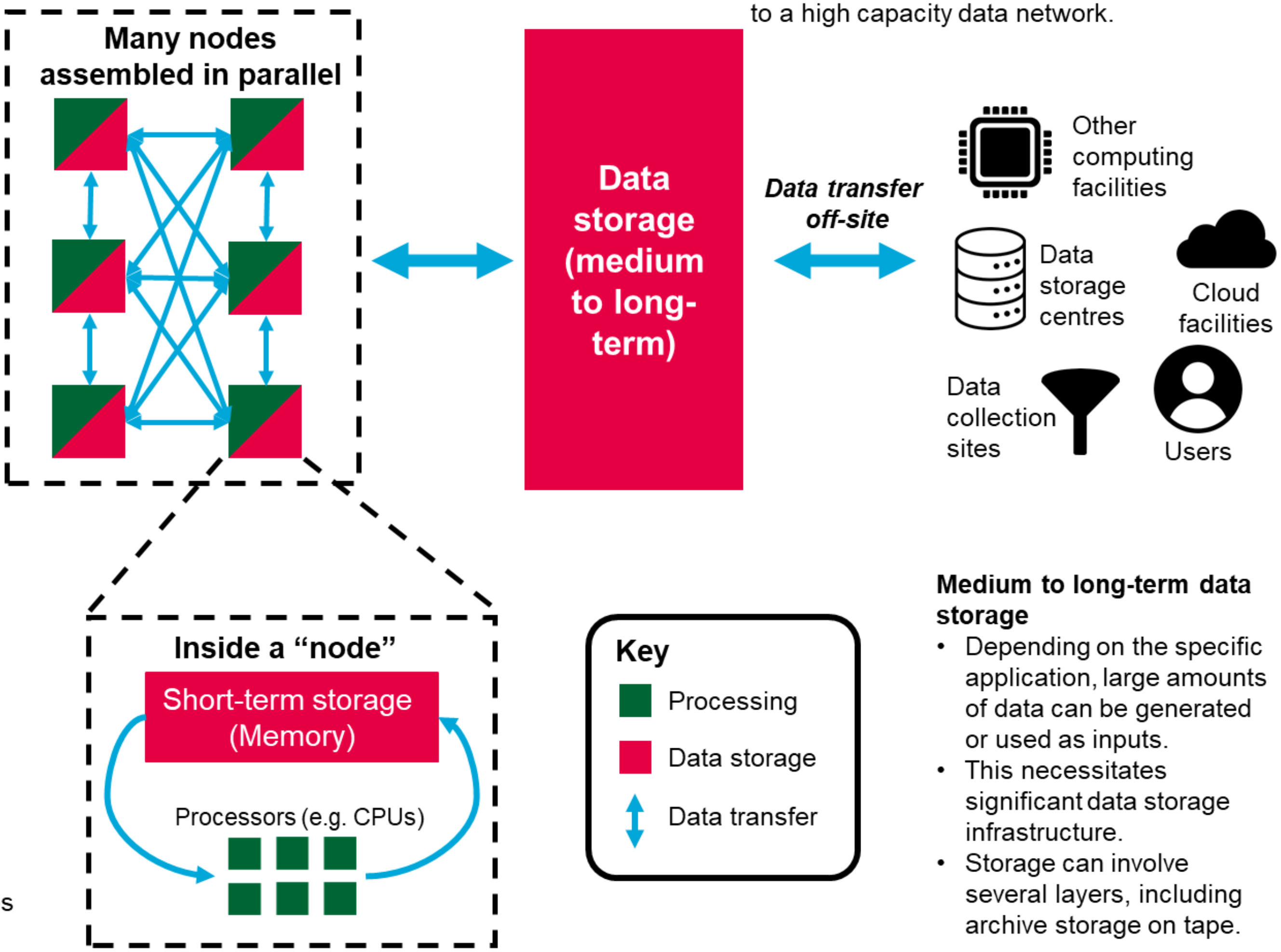
Appendix B: A schematic description of a typical large-scale computing system

Parallel processing

- Assembling many nodes (each with multiple processors) in parallel increases processing power.
- Some programmes need regular communication between nodes, requiring high-speed connections between them. These are known as tightly coupled programmes.

Data transfer

- Given the large volumes of data involved, transferring data between sites requires access to a high capacity data network.



Many nodes assembled in parallel

Data storage (medium to long-term)

Data transfer off-site

Other computing facilities

Data storage centres

Cloud facilities

Data collection sites

Users

Inside a "node"

Short-term storage (Memory)

Processors (e.g. CPUs)

Key

🟩	Processing
🟥	Data storage
↕	Data transfer

Short-term data storage (Memory)

- Running memory intensive applications requires access to large amounts of short-term data storage.
- This memory can be shared (but only for smaller systems), or it can be distributed amongst individual nodes (as shown here).

Medium to long-term data storage

- Depending on the specific application, large amounts of data can be generated or used as inputs.
- This necessitates significant data storage infrastructure.
- Storage can involve several layers, including archive storage on tape.

Appendix C: Hardware for large-scale computing

This appendix gives more detail on the diversity of currently utilised hardware architectures as well as emerging paradigms that may be relevant for the large-scale computing landscape in the future. This annex focusses specifically on hardware used for **computer processing**.

C.1 Types of processors

The below described hardware types are all currently in use to varying degrees in large-scale computing systems.

	Key features	**Maturity and use in large-scale computing**	**Relevant tasks and applications**
Central Processing Unit (CPU)	• The traditional type of processor used for most general-purpose computing. • CPU cores carry out processing tasks sequentially, in serial. • A single computer "chip" will routinely contain multiple CPU cores that work in parallel.	• CPUs are the dominant processor type used in most large-scale computing systems. • Many large-scale computing systems still solely rely on the CPUs due to their versatility and ease of writing software for them.	• Extremely versatile for use in a wide range of tasks. • For some highly parallel tasks, other types of computer processors (noted below) can offer greater efficiency.
Graphics Processing Units (GPUs)	• Originally designed to accelerate graphics tasks like image rendering. • Their structure, with hundreds or thousands of cores assembled in parallel, allows them to efficiently carry out processing of large blocks of data (where the data structure is appropriate).	• General-purpose graphics processing units (GPGPU) have been developed for use in large-scale computing systems. • They can provide significant acceleration for certain applications and many large-scale computing systems now use GPUs alongside CPUs.	• Particularly applicable for use cases involving analysis of large amounts of data. • Examples include big-data analytics, artificial intelligence and large-scale image processing.

Tensor Cores or Tensor Processing Units (TPUs)	- A range of related computer architectures developed to efficiently tackle AI problems. - Google markets a tensor processing unit (TPU) while NVIDIA integrates so-called "tensor cores" into their Volta chip architectures.[158,159]	- Some of the most powerful large-scale computing systems already contain these accelerators. - Oak Ridge National Laboratory in the US has claimed to have run the world's fastest ever computation, making use of Tensor cores integrated into their Summit supercomputer to analyse genomics data. [160]	- These specialised processors are specifically designed for rapid matrix multiplication to efficiently run neural network-based AI algorithms.
Field Programmable Gate Arrays (FPGAs)	- A type of computer chip designed to be configurable by the user and can be optimised for specific applications. - The specific architectural set-up of the chip can be almost instantaneously reprogrammed and can even be reconfigured dynamically while the chip is in use.	- A very different approach to programming is required and their use in large-scale computing remains largely experimental. - Some systems are starting to integrate FPGAs, such as the Noctua supercomputer at Paderborn University in Germany.[vi] - FPGAs may be able to deliver performance and energy efficiency improvements compared to CPUs.[vii]	- Application areas are still being explored across fields including bioinformatics, financial modelling and computational biology.[viii] - FPGA could potentially outperform GPUs for certain deep learning applications where the level of parallelisation is variable.[ix]

[vi] https://www.top500.org/news/german-university-will-deploy-fpga-powered-cray-supercomputer/

[vii] https://www.nextplatform.com/2018/04/04/another-step-toward-fpgas-in-supercomputing/
https://arxiv.org/abs/1904.04953

[viii] https://www.researchgate.net/publication/258381392_High-Performance_Reconfigurable_Computing

[ix] https://insidehpc.com/2019/07/fpgas-and-the-road-to-reprogrammable-hpc/

C.2 Emerging computing paradigms

The table below describes novel approaches to computing being developed. These are at a fairly early stage of development, but they have the potential to dramatically improve certain types of computations.

	Key features	**Maturity and relevance for large-scale computing**	**Relevant tasks and applications**
Quantum computing	• This approach uses properties of quantum mechanics (namely superposition and entanglement) to perform computations. • Current quantum computers are incredibly sensitive to external interference and require operating temperatures close to absolute zero.	• While its applicability and speed of development remains highly uncertain, it could have quite revolutionary implications for large-scale computing. • In advance of the development of "general purpose" quantum computers, intermediate systems (called noisy intermediate-scale quantum devices) may provide benefits for certain applications.	• Quantum computers have the theoretical potential to solve certain classes of computational problems much faster or more efficiently than a conventional computer. • The capability to efficiently factorise very large numbers could render currently employed encryption techniques obsolete. • Quantum computing could also accelerate a range of scientific computing applications, for example the improved modelling of natural phenomena at the atomic scale and new forms of system optimisation.[x]
Neuromorphic computing	• This technique involves computational architectures that process information in a way that is analogous to neurons and synapses.	• The mimicking of the human brain could help to improve energy efficiency and overcome bottlenecks involved in the	• Opportunities are being explored to exploit neuromorphic architectures

[x] https://link.springer.com/article/10.1007/s10676-017-9438-0

	• Information processing and storage are integrated into the same operation.	transfer of data between processors and storage.[xi] • The SpiNNacker project, based at the University of Manchester, involves the development of a large-scale neuromorphic computing platform for 'spiking' neural networks.[xii]	for massively parallel and energy efficient information processing.[xiii] • Neural networks could also improve our understanding of the human brain.
Mixed-precision computing	• Recently, interest has grown in alternative computing approaches which do not rely on the same level of high-precision deterministic modelling.[xiv] • Having lower levels of precision for encoding values reduces computational and energy demands but also introduces rounding errors and therefore "noise" into simulations.	• Where levels of uncertainty in input data and/or model parameters are high, lower levels of precision could be utilised for representing specific variables without adversely affecting the overall accuracy of the computation.[xv] • Future energy-efficient supercomputers could include processors with variable levels of precision that could be tailored to specific requirements.	• Potentially appropriate for workloads where there is high-level of uncertainty in input variables. Examples include combustion and climate models.[161] • The noise introduced by low-precision computation can be useful in certain applications. For example, in stochastic optimisation methods which rely on the use of random variables.

[xi] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6444279/

[xii] http://apt.cs.manchester.ac.uk/projects/SpiNNaker/

[xiii] http://apt.cs.manchester.ac.uk/projects/SpiNNaker/

[xiv] https://www.nature.com/news/modelling-build-imprecise-supercomputers-1.18437

[xv] https://www.nature.com/news/modelling-build-imprecise-supercomputers-1.18437

Acknowledgements

The Government Office for Science would like to thank the many officials, experts and industry stakeholders who contributed to this project by providing expert advice and constructive reviews of drafts.

This project was led by a steering group of Chief Scientific Advisers from across Government:

- Professor Dame Angela McLean, University of Oxford and Chief Scientific Adviser for the Ministry of Defence;
- Professor Anthony Finkelstein, University College London and Chief Scientific Adviser for National Security;
- Professor Carole Mundell, University of Bath and Chief Scientific Adviser for the Foreign and Commonwealth Office;
- Professor John Aston, University of Cambridge and Chief Scientific Adviser for the Home Office;
- Professor John Loughhead, former Chief Scientific Adviser for the Department of Business, Energy and Industrial Strategy;
- Professor Paul Monks, Chief Scientific Adviser for the Department of Business, Energy and Industrial Strategy;
- Professor Mark Thomson, Executive Chair of the Science and Technology Facilities Council;
- Dr Mike Short, Chief Scientific Adviser for the Department for International Trade;
- Professor Robin Grimes, Imperial College London and Chief Scientific Adviser (Nuclear) for the Ministry of Defence;
- Sir Patrick Vallance, Government Chief Scientific Adviser;
- Professor Sheila Rowan, University of Glasgow and Chief Scientific Adviser for Scotland;
- Professor Stephen Belcher, Chief Scientific Adviser for the Met Office;
- Professor Tom Rodden, University of Nottingham and Chief Scientific Adviser for the Department for Digital, Culture, Media and Sport.

The project team at the Government Office for Science was led by James Selwyn and Martin Glasspool and included Mohamed Barre, Juanita Bawagan, Beatrice Freeman, Dr Catherine Hasnain, Sarah Jennings, Nick Matthews, Melanie Carlin, Isabel Robinson, Peter Sellen, Dr Arianna Sorba, Lewis Weinberger, Dr Tom Wells and Dr Simon Whitfield.

We are also grateful for the valuable time and input from:

- Dr John Baldoni, ATOM Consortium;
- Dr Lee Braine and Richard Barnes, Barclays;

- Prof John MacWilliams, Columbia University;
- Gordon Wai and Lindsay Taylor, Competition and Markets Authority;
- Dr Andrew Poulter, Defence Science and Technology Laboratory;
- Prof Mark Parsons, Edinburgh Parallel Computing Centre;
- Dr Richard Gunn, Engineering and Physical Sciences Research Council;
- Justin O'Byrne, Science and Technology Facilities Council;
- Dr Peter Bauer and Dr Martin Palkovic, European Centre for Medium-Range Weather Forecasts;
- Dr Ewan Birney and Dr Steven Newhouse, European Molecular Biology Laboratory;
- Dr Luke Mason, Hartree Centre;
- Dr Ben Bennett, Hewlett Packard Enterprise;
- Prof Sir Peter Knight, Imperial College London;
- Ian Dabson and Phil Saw, Infrastructure and Projects Authority;
- Jim Brase and Dr Patricia Falcone, Lawrence Livermore National Laboratory;
- Dr Paul Selwood, Met Office;
- Antony Bastiani, Ministry of Defence;
- Dr Tim Littlewood, Natural History Museum;
- Claire Chapman, Office for Artificial Intelligence;
- Dr Paul Kent, Oak Ridge National Laboratory;
- Owen Thomas, Red Oak Consulting;
- Dr Henner Wapenhans, Rolls-Royce;
- Dr Thomas Schulthess, Swiss National Supercomputing Centre;
- Dr James Hetherington, Nicola Mitchell and Dr Emily Swaine, UK Research and Innovation;
- Barbara Helland, US Department of Energy;
- Prof Richard Kenway and Prof Peter Clarke, University of Edinburgh;
- Prof David Britton, University of Glasgow;
- Prof Tim Palmer, University of Oxford;
- Claire Wyatt and Professor Simon Hettrick, University of Southampton;
- Prof Jack Dongarra, University of Tennessee.

References

[1] Government Office for Science (2018) *Computational Modelling: Technological Futures*, viewed 26 May 2020, https://www.gov.uk/government/publications/computational-modelling-blackett-review

[2] House of Lords Artificial Intelligence Select Committee (2018), *Report of Session 2017-19 - AI in the UK: ready, willing and able?*, viewed 26 May 2020, https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/publications/

[3] https://www.top500.org/statistics/list/

[4] Figures calculated from LINPACK benchmark scores (R_{MAX}) for the 500 most powerful HPC systems globally. Data range is November 2014 to November 2019. Source: TOP500 (top500.org).

[5] HM Government (2020) UK Research and Development Roadmap, viewed 19 August 2020, https://www.gov.uk/government/publications/uk-research-and-development-roadmap

[6] HM Government (2020) Principles for science and technology moon-shots, viewed 25 August 2020, https://www.gov.uk/government/publications/principles-for-science-and-technology-moon-shots

[7] Conway, S., Joseph, E., Sorensen, R. and Norton, A. (2019) *The Business Value of Leading-Edge High Performance Computing: 2019 Update*. Hyperion Research LLC, viewed 26 November 2019, https://www.hpe.com/uk/en/resources/solutions/hyperion-hpc-value.html

[8] BCS – The Chartered Institute for IT (2010) *A brief history of British computers: the first 25 years (1948 - 1973)*., viewed 9 December 2019, https://www.bcs.org/content-hub/a-brief-history-of-british-computers-the-first-25-years-1948-1973/

[9]European Commission (2018) *High Performance Computing - best use examples*. European Commission, viewed 9 December 2019, https://ec.europa.eu/digital-single-market/en/news/high-performance-computing-best-use-examples

[10] PRACE Scientific Steering Committee (2018) *The Scientific Case for Computing in Europe 2018 – 2026*. Bristol, UK: Insight Publishers, viewed 9 December 2019, http://www.prace-ri.eu/third-scientific-case/

[11] CompBioMed Related Projects (that make use of HPC and HTC), viewed 15 April 2021, https://www.compbiomed.eu/about/related-projects/

[12] Hartree Centre (2017) *Accelerating the product discovery process at Unilever*, Science & Technology Facilities Council, viewed 9 June 2019, https://stfc.ukri.org/innovation/success-stories/accelerating-the-product-discovery-process-at-unilever/

[13] Hartree Centre (2016) *Code optimisation for aero-engine innovation*, Science & Technology Facilities Council, viewed 24 June 2019. http://www.stfc.ac.uk/files/rolls-royce-case-study/

[14] AWE (2019) *Using Supercomputers*, viewed 9 December 2019, https://www.awe.co.uk/what-we-do/nuclear-warheads-lifecycle/science/using-supercomputers/

[15]Kepner, Jeremy; Cho, Kenjiro; Klaffy, KC; Gadepally, Vijay; Michaleas, Peter; Milechin, Lauren (2019) *Hypersparse Neural Network Analysis of Large-Scale Internet Traffic*, viewed 5 May 2021, https://arxiv.org/abs/1904.04396

[16] EPCC (n.d.) *FireGRID: HPC for fire emergency response systems*, viewed 8 September 2020, https://www.epcc.ed.ac.uk/projects-portfolio/firegrid-hpc-fire-emergency-response-systems

[17] UKRI (2020) *UK joins COVID-19 High Performance Computing Consortium*, viewed 8 September 2020, https://www.ukri.org/news/uk-joins-covid-19-high-performance-computing-consortium/

[18] Green, R.C., Wang, L. and Alam, M. (2011) *High performance computing for electric power systems: Applications and trends*. IEEE Power and Energy Society General Meeting, 1-8, viewed 17 December 2019,https://www.researchgate.net/publication/252049333_High_performance_computing_for_electric_power_systems_Applications_and_trends

[19] Data-Driven Innovation (n.d.) *DataLoch*, viewed 17 December 2019, https://ddi.ac.uk/projects/dataloch/

[20] UK Research and Innovation (n.d.) *From data to early diagnosis and precision medicine*, viewed 11 December 2019, https://www.ukri.org/innovation/industrial-strategy-challenge-fund/from-data-to-early-diagnosis-and-precision-medicine/

[21] https://www.gov.uk/government/collections/the-national-digital-twin-programme-ndtp

[22] Council for Science and Technology (2020) *Diffusion of technology for productivity*, viewed 9 June 2020, https://www.gov.uk/government/publications/diffusion-of-technology-for-productivity

[23] Gigler, B., Casorati, A., and Verbeek, A. (2018) *Financing the future of supercomputing: How to increase the investments in high performance computing in Europe*, European Investment Bank, viewed 9 June 2020 https://www.eib.org/en/publications/financing-the-future-of-supercomputing

[24] Foster, I (2002) *What is the Grid? A three point checklist*, viewed 9 June 2020, https://www.mcs.anl.gov/~itf/Articles/WhatIsTheGrid.pdf

[25] Folding@home (2018) About, viewed 9th June 2020, https://foldingathome.org/about/

[26] Mell, P and Grance, T. (2011) *The NIST Definition of Cloud Computing*. National Institute of Standards and Technology, viewed 9 December 2019, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[27] Netto, M. A., Calheiros, R. N., Rodrigues, E. R., Cunha, R. L., & Buyya, R. (2018). *HPC cloud for scientific and business applications: Taxonomy, vision, and research challenges*. ACM Computing Surveys, 51(1), 8

[28] Netto, M. A., Calheiros, R. N., Rodrigues, E. R., Cunha, R. L., & Buyya, R. (2018). *HPC cloud for scientific and business applications: Taxonomy, vision, and research challenges*. ACM Computing Surveys, 51(1), 8

[29] Gupta, A. and Milojicic, D. (2011) *Evaluation of HPC Applications on Cloud*. 2011 Sixth Open Cirrus Summit, IEEE, 22-26, viewed 11 December 2019, https://ieeexplore.ieee.org/document/6200551

[30] Yelick, K., Coghlan, S., Draney, B., & Canon, R. S. (2011) *The Magellan Report on Cloud Computing for Science*. US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR), 3.

[31] Government Digital Service (2019) *Guidance: Managing technical lock-in in the cloud*, viewed 9 October 2020, https://www.gov.uk/guidance/managing-technical-lock-in-in-the-cloud

[32] Yelick, K., Coghlan, S., Draney, B., & Canon, R. S. (2011). *The Magellan report on Cloud computing for science*. US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR), 3.

[33] Conway, S., Norton, A., Sorensen, B. and Joseph, E. (2019) *Technology Spotlight: Cloud Computing for HPC Comes of Age*. Hyperion Research LLC, viewed 18 December 2019, https://d1.awsstatic.com/HPC2019/Amazon-HyperionTechSpotlight-190329.FINAL-FINAL.pdf

[34] Fortissimo (2020) *Success Stories*, viewed 7 September 2020, https://www.fortissimo-project.eu/success-stories

[35] Conway, S., Norton, A., Sorensen, B. and Joseph, E. (2019) *Technology Spotlight: The Future of HPC Cloud Computing*. Hyperion Research LLC, viewed 13 December 2019, https://services.google.com/fh/files/misc/gcp_hyperion_tech_spotlight_aug_2019.pdf

[36] OpenAI (2018) *AI and Compute*, viewed 9 October 2020, https://openai.com/blog/ai-and-compute/

[37] Asch, M., Moore, T., Badia, R., Beck, M., Beckman, P., Bidot, T., ... & Deelman, E. (2018). *Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry*. The International Journal of High Performance Computing Applications, 32(4), 435-479, https://doi.org/10.1177%2F1094342018778123

[38] SKA (n.d.) *Software And Computing*, viewed 10 June 2020, https://www.skatelescope.org/software-and-computing/

[39] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L. (2016) *Edge Computing: Vision and Challenges*, IEEE Internet of Things Journal, 5(3), 637-646, https://doi.org/10.1109/JIOT.2016.2579198

[40] Asch, M., Moore, T., Badia, R., Beck, M., Beckman, P., Bidot, T., ... & Deelman, E. (2018). *Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry*. The International Journal of High Performance Computing Applications, 32(4), 435-479, https://doi.org/10.1177%2F1094342018778123

[41] The Royal Society (2019) *Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis*, viewed 10 June 2020, https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/

[42] UKRI (2019) *UKRI Infrastructure Roadmap Progress Report, viewed 9 June 2020, https://www.ukri.org/files/infrastructure/progress-report-final-march-2019-low-res-pdf/*

[43] EuroHPC (2020) *LUMI: a new EuroHPC world-class supercomputer in Finland*, viewed 9 November 2020, https://eurohpc-ju.europa.eu/news/lumi-new-eurohpc-world-class-supercomputer-finland

[44] EuroHPC (n.d.) *Leading the way in the European Supercomputing*, viewed 11 December 2019, https://eurohpc-ju.europa.eu/

[45] EuroHPC (n.d.) *EuroHPC Members, viewed 11 September 2020, http://eurohpc.eu/members-list*

[46] Exascale Computing Project (2019) *Overview of the ECP*, viewed 9 December 2019, https://exascaleproject.org/about/

[47] Energy-Efficient Heterogeneous Computing at Exascale (2015), *Project description*, http://www.ecoscale.eu/project-description.html

[48] Feldman, M. (2018) *DOE Shifts Exascale Plans into High Gear, Asks Supercomputing Vendors to Submit Proposals*. TOP500, viewed 9 December 2019, https://www.top500.org/news/doe-shifts-exascale-plans-into-high-gear-asks-supercomputing-vendors-to-submit-proposals/

[49] The Exascale Computing Project (2019) *Addressing a National Imperative: Application Development Update, September 2019*. US Department of Energy, viewed 11 December 2019, https://exascaleproject.org/the-ecp-2019-application-development-report-is-available/

[50] University of Bristol (2018) *Industry partnership to create the world's most accurate simulation of an aircraft jet engine*, viewed 8 September 2020, https://www.bristol.ac.uk/news/2018/september/asimov-prosperity-partnership.html

[51] Waldrop, M. M. (2016) *The chips are down for Moore's law*. Nature (530) 144-147, viewed 11 December 2019, https://www.nature.com/articles/530144a

[52] Top500 (2019) *Top 500 Lists*, viewed 11 December 2019, https://www.top500.org/lists/

[53]Mittal, S. (2019) *A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform*. Journal of Systems Architecture (97) 428-442, viewed 12 December 2019, https://www.researchgate.net/publication/329802520_A_Survey_on_Optimized_Implementation_of_Deep_Learning_Models_on_the_NVIDIA_Jetson_Platform

[54] A.S.G.; Edler, T. (2015) *On Global Electricity Usage of Communication Technology: Trends to 2030.* Challenges (6) 117-157.

[55] Elmeligi A. (2018) *Assessing ICT global emissions footprint: trends to 2040 & recommendations.* J Clean Prod 2018; 177: 448–63

[56] J. Koomey, S. Berard, M. Sanchez and H. Wong, (2011) *Implications of Historical Trends in the Electrical Efficiency of Computing*, IEEE Annals of the History of Computing, (33) 3:46-54, doi: 10.1109/MAHC.2010.28

[57] HPC Wire (2020) *Exascale Watch: El Capitan Will Use AMD CPUs & GPUs to Reach 2 Exaflops*, viewed 10 June 2020, https://www.hpcwire.com/2020/03/04/exascale-watch-el-capitan-will-use-amd-cpus-gpus-to-reach-2-exaflops/

[58] Department for Business, Energy & Industrial Strategy (2018) *National Statistics: Sub-national electricity and gas consumption summary report 2017*. UK Government, viewed 11 December 2019, https://www.gov.uk/government/statistics/sub-national-electricity-and-gas-consumption-summary-report-2017

[59] Top500 (n.d.) *The Green 500*, viewed 28 February 2020, https://www.top500.org/green500/

[60] Morris, A. (2019) *Five Tips to Reduce Your HPC Carbon Footprint*, HPCwire, viewed 10 June 2020, https://www.hpcwire.com/solution_content/ibm/cross-industry/five-tips-to-reduce-your-hpc-carbon-footprint/

[61] Palmer, T. (2015) *Build imprecise supercomputers*, Nature (526) 32-33, viewed 11 June 2020, https://www.nature.com/news/modelling-build-imprecise-supercomputers-1.18437

[62] Cong Wang, M. Zink and D. Irwin (2015), "Optimizing parallel HPC applications for green energy sources," 2015 Sixth International Green and Sustainable Computing Conference (IGSC), Las Vegas, NV, 2015, pp. 1-8, viewed 11 June 2020, https://ieeexplore.ieee.org/document/7393694

[63] BBC News (2017), *Record-sized data centre planned inside Arctic Circle*, viewed 11 June 2020, https://www.bbc.co.uk/news/technology-40922048

[64] DeepMind (2018), *Safety-first AI for autonomous data centre cooling and industrial control*, viewed 11 June 2020, https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control

[65] Massachusetts Green High Performance Computing Center (n.d) *Energy Efficiency*, viewed 28 February 2020, https://www.mghpcc.org/about/green-design/energy-efficient-design/

[66] Top500 (2019) *Top 500 – November 2019*, viewed 11 December 2019, https://www.top500.org/lists/2019/11/

[67] UKRI (2019) *The UK's research and innovation infrastructure: Landscape Analysis*, viewed 17 June 2020, https://www.ukri.org/research/infrastructure/

[68] Met Office (n.d.) *The Cray XC40 supercomputer*, viewed 11 December 2019, https://www.metoffice.gov.uk/about-us/what/technology/supercomputer

[69] Met Office (n.d.) *Met Office and NERC joint supercomputer system (MONSooN)*, viewed 11 December 2019, https://www.metoffice.gov.uk/research/approach/collaboration/jwcrp/monsoon-hpc

[70] Met Office (2020) *Up to £1.2billion for weather and climate Supercomputer*, viewed 22 April 2020, https://www.metoffice.gov.uk/about-us/press-office/news/corporate/2020/supercomputer-funding-2020

[71] TOP500 (n.d.) *Damson - Bull Sequana X1000, Xeon E5-2697v4 18C 2.3GHz, Infiniband EDR*, viewed 11 December 2019, https://www.top500.org/system/179409

[72] ARCHER2 (n.d.) *ARCHER2 Hardware & Software*, viewed 22 April 2020, https://www.archer2.ac.uk/about/hardware.html

[73] Distributed Research utilizing Advanced Computing website, viewed 11 December 2019, https://dirac.ac.uk

[74] EPSRC (2019) *Tier 2 HPC Interview Panel 2019*, viewed 9 November 2020, https://gow.epsrc.ukri.org/NGBOViewPanelROL.aspx?PanelId=1-7V5C1S&RankingListId=1-7V5C4F

[75] Supercomputing Wales website, viewed 17 December 2019, https://www.supercomputing.wales/

[76] ARCHIE-WeSt website, viewed 17 December 2019, https://www.archie-west.ac.uk/

[77] Earlham Institute (n.d.) *About us*, viewed 22 April 2020, https://www.earlham.ac.uk/about-us

[78] EBI-EMBL (n.d) *EBI-EMBL services*, viewed 22nd April 2020, https://www.ebi.ac.uk/services

[79] GridPP (n.d) *History*, viewed: 12 May 2020, https://www.gridpp.ac.uk/about/history/

[80] JASMIN website, viewed 11 December 2019, http://www.jasmin.ac.uk/

[81] Jisc (n.d.) *Janet Network*, viewed 11 December 2019, https://www.jisc.ac.uk/janet

[82] GÉANT (n.d) *GÉANT pan-European network*, viewed 22 April 2020, https://www.geant.org/Networks/Pan-European_network

[83] EPCC (n.d.) *UK Research Data Facility*, viewed 22 April 2020, https://www.epcc.ed.ac.uk/facilities/uk-research-data-facility

[84] Partnership for Advanced Computing in Europe website, viewed 11 December 2019, http://www.prace-ri.eu

[85] ELIXIR (n.d.) *About us*, viewed 22nd April 2020 https://elixir-europe.org/about-us

[86] ELIXIR (n.d) *Compute Platform*, viewed 22 April 2020, https://elixir-europe.org/platforms/compute

[87] US Department of Energy INCITE Leadership Computing website, viewed 11 December 2019, http://www.doeleadershipcomputing.org/

[88] TOP500 (n.d.) *scafellpike - Bull Sequana X1000*, viewed 17 June 2020, https://www.top500.org/system/179163

[89] NVIDIA (2020) *NVIDIA Building UK's Most Powerful Supercomputer, Dedicated to AI Research in Healthcare*, viewed 9 October 2020, https://nvidianews.nvidia.com/news/nvidia-building-uks-most-powerful-supercomputer-dedicated-to-ai-research-in-healthcare

[90] Software.org: the BSA Foundation (2018) *The Growing €1 Trillion Economic Impact of Software, accessed 19 December 2019,* https://software.org/reports/2018-eu-software-impact/#united-kingdom-acc

[91] Hettrick, S. (2014) *It's impossible to conduct research without software, say 7 out of 10 UK researchers*. Software Sustainability Institute, viewed 11 December 2019, https://software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers

[92] Science and Technology Facilities Council (n.d) *CoSeC - Computational Science Centre for Research Communities*, viewed 11 December 2019, https://www.scd.stfc.ac.uk/Pages/CoSeC.aspx

[93] Software Sustainability Institute (n.d) *About the Software Sustainability Institute*, viewed 11 December 2019, https://www.software.ac.uk/about

[94] Collaborative Computational Projects (n.d.) *About the CCPs*, viewed 28 February 2020, http://www.ccp.ac.uk/about.html

[96] Department for Business, Energy & Industrial Strategy (2019) *£88 million to help unleash the productive power of the UK economy*. UK Government, viewed 12 December 2019, https://www.gov.uk/government/news/88-million-to-help-unleash-the-productive-power-of-the-uk-economy

[97] Philippe, O. (2018) *What do we know about RSEs? Results from our international surveys*, Software Sustainability Institute, viewed 17 December 2019, https://www.software.ac.uk/blog/2018-03-12-what-do-we-know-about-rses-results-our-international-surveys

[98]Cannam, C., Gorissen, D., Hetherington, J., Johnston et al. (2013) *Ten reasons to be a research software engineer*, Software Sustainability Institute, viewed 17 December 2019, https://www.software.ac.uk/blog/2013-08-23-ten-reasons-be-research-software-engineer

[99] Philippe, O. (2018) *What do we know about RSEs? Results from our international surveys*, Software Sustainability Institute, viewed 17 December 2019, https://www.software.ac.uk/blog/2018-03-12-what-do-we-know-about-rses-results-our-international-surveys

[100]The Engineering and Physical Sciences Research Council (n.d) *Research Software Engineer Fellowships II*, UK Research and Innovation, viewed 11 December 2019, https://epsrc.ukri.org/funding/calls/research-software-engineer-fellowships-ii/

[101] Society of Research Software Engineering (n.d.) *About*, viewed 11 December 2019, https://society-rse.org/about/

[102]Society of Research Software Engineering (n.d.) *RSE Groups*, viewed 11 December 2019, https://society-rse.org/community/rse-groups/

[103] Philippe, O. (2018) *What do we know about RSEs? Results from our international surveys*, Software Sustainability Institute, viewed 17 December 2019, https://www.software.ac.uk/blog/2018-03-12-what-do-we-know-about-rses-results-our-international-surveys

[104] Hettrick, Simon (2018), *International collaboration on RSE survey – socio demography,* viewed 17 July 2020 https://github.com/softwaresaved/international-survey/blob/master/analysis/2018/I.%20Socio%20demography.ipynb

[105] ONS (2020) *EMP01 SA: Full-time, part-time and temporary workers (seasonally adjusted)*, 'All in employment' Feb-Apr 2020 dataset, accessed 9 July 2020, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/datasets/fulltimeparttimeandtemporaryworkersseasonallyadjustedemp01sa

[106] ONS (2020) *A05 NSA: Employment, unemployment and economic inactivity by age group (not seasonally adjusted)*, 'Employment level – aged 16 and over' Feb-Apr 2020 dataset, accessed 9 July 2020, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/datasets/employmentunemploymentandeconomicinactivitybyagegroupnotseasonallyadjusteda05nsa

[107] ONS (2020) *A09: Labour market status by ethnic group*, 'Employment by ethnicity: People (not seasonally adjusted)' Jan-Mar 2020 dataset, accessed 9 July 2020, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/datasets/labourmarketstatusbyethnicgroupa09

[108] House of Commons Library (2020), *People with disabilities in employment*, published 3 January 2020, accessed 9 July 2020, https://commonslibrary.parliament.uk/research-briefings/cbp-7540/

[109] Women in HPC (n.d.), accessed 9 July 2020, https://womeninhpc.org/

[110] Code Club (n.d.), *About*, viewed 20 August 2020, https://codeclub.org/en/about

[111] STFC (2020), *Apprenticeships*, viewed 15 July 2020 https://stfc.ukri.org/about-us/work-with-us/apprentice-training-scheme/

[112] https://www.ecu.ac.uk/equality-charters/athena-swan/

[113] https://www.ecu.ac.uk/equality-charters/race-equality-charter/about-race-equality-charter/

[114] IRIS (n.d) *What is IRIS*, viewed 11 December 2019,https://www.iris.ac.uk/about/what-is-iris/

[115] The High Peformance Computing Special Interest Group for UK Academia (n.d), *About*, viewed 22nd April 2020 https://hpc-sig.org.uk/index.php/about/

[116] EPSRC (n.d) *High End Computing Consortia*, viewed 26 June 2020 https://epsrc.ukri.org/research/facilities/hpc/access/highendcomputingconsortia/

[117] EuroHPC (n.d.) *EuroHPC Leading the way in the European Supercomputing*, viewed 11 December 2019 https://eurohpc-ju.europa.eu/index.html

[118] Brueckner, R. (2015) *Cray Opens EMEA Headquarters in Bristol*. insideHPC, viewed 11 December 2019, https://insidehpc.com/2015/06/cray-opens-emea-headquarters-in-bristol/

[119] Intel (n.d.) *Intel® Parallel Computing Centers*, viewed 11 December 2019, https://software.intel.com/en-us/ipcc/centers

[120] BBC (2016) *ARM chip designer to be bought by Japan's Softbank*, viewed 11 December 2019, https://www.bbc.co.uk/news/business-36822806

[121] TOP500 (2019) *JUNE 2020*. TOP500.org, viewed 8 September 2020, https://www.top500.org/lists/top500/2020/06/

[122] Compound Semiconductor Applications Catapult (n.d.) *About Us*, 11 December 2019, https://csa.catapult.org.uk/about-us/

[123] Graphcore website, viewed 11 December 2019, https://www.graphcore.ai/

[124] AccelerComm (n.d.) *Our Products*, viewed 8 September 2020, https://www.accelercomm.com/products

[125] Intel Newsroom (2019) *Intel Acquires Omnitek, Strengthens FPGA Video and Vision Offering*, viewed 8 September 2020, https://newsroom.intel.com/news/intel-acquires-omnitek-fpgas/

[126] Alpha Data Website, viewed 13 May 2021, https://www.alpha-data.com/markets/data-center/

[127] The University of Manchester (n.d.) *SpiNNaker Home Page*, viewed 9th March 2020 http://apt.cs.manchester.ac.uk/projects/SpiNNaker/

[128] Met Office (n.d.) *How valuable is the Met Office?* Viewed 8 October 2020, https://www.metoffice.gov.uk/about-us/what/pws/value

[129] Joseph, EC., Conway, S., Ingle, C., Cattaneo, G., Meunier, C., Martinez, N. (2010) *A strategic agenda for European leadership in supercomputing: HPC 2020*, IDC, viewed 10 June 2020, https://op.europa.eu/en/publication-detail/-/publication/732405b8-00f4-44eb-b9ea-937f9b792f48/language-en

[130] Haskel, J., Hughes, A., Bascavusoglu-Moreau, E. (2014) The Economic Significance of the UK Science Base: A report for the campaign for science and engineering, UK-Innovation Research Centre, viewed 10 June 2020, http://www.sciencecampaign.org.uk/asset/4567DD2A-0604-42E5-AF8EEA248D3DCE1B/

[131] Salter, A. J., & Martin, B. R. (2001). *The economic benefits of publicly funded basic research: a critical review.* Research policy, 30(3), 509-532, https://doi.org/10.1016/S0048-7333(00)00091-3

[132] Halterbeck, M., Conlon, G., Julius, J. (2017) *The economic impact of Russell Group universities: Final Report for the Russell Group*, London Economics, viewed 10 June 2020, http://russellgroup.ac.uk/news/economic-impact-of-russell-group-universities/

[133] Economic Insights Ltd (2015) *What is the relationship between public and private investment in R&D? A report commissioned by the Department for Business, Innovation and Skills*, viewed 10 June 2020, https://www.gov.uk/government/publications/research-and-development-relationship-between-public-and-private-investment

[134] TOP500 (2019) *November 2019*. TOP500.org, viewed 11 December 2019 https://www.top500.org/lists/2019/11/

[135] The Engineering and Physical Sciences Research Council (n.d.) *Tier-2 High Performance Computing Centres*, viewed 11 December 2019, https://epsrc.ukri.org/research/facilities/hpc/tier2/

[136] Science and Technology Facilities Council (2020) *Hartree Centre - Case studies: Simulation and machine learning for future medicine*, viewed 8 October 2020, https://stfc.ukri.org/about-us/our-impacts-achievements/case-studies/simulation-and-machine-learning-for-future-medicine/
[137] Edinburgh International Data Facility (n.d.) *What we offer*, viewed 8 October 2020, https://www.ed.ac.uk/edinburgh-international-data-facility/services
[138] The Council for Science and Technology (2017) *Harnessing science and technology for economic benefit across the UK,* UK Government, viewed 19 December 2019 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653231/CST_letter_on_science_and_place_-_formatted.pdf
[140] The Exascale Computing Project (2019) *Addressing a National Imperative: Application Development Update September 2019*. US Department of Energy, viewed 11 December 2019, https://exascaleproject.org/the-ecp-2019-application-development-report-is-available/
[141] The Engineering and Physical Sciences Research Council (n.d) *Research Software Engineer Fellowships II*. UK Research and Innovation, viewed 11 December 2019, https://epsrc.ukri.org/funding/calls/research-software-engineer-fellowships-ii/
[142] Society of Research Software Engineering website, viewed 11 December 2019, https://society-rse.org/
[143] Institute of Coding website, viewed 11 December 2019, https://instituteofcoding.org/
[144] Department for Business, Energy & Industrial Strategy (2018) *National Statistics: Sub-national electricity and gas consumption summary report 2017*. UK Government, viewed 11 December 2019, https://www.gov.uk/government/statistics/sub-national-electricity-and-gas-consumption-summary-report-2017
[145] CORAL Collaboration (2018), *CORAL-2 ACQUISITION,* viewed 11 December 2019, https://procurement.ornl.gov/rfp/CORAL2/
[146] Energy-Efficient Heterogeneous Computing at Exascale website, viewed 11 December 2019, http://www.ecoscale.eu/index.html
[147] Øyvann, S. (2019) *Microsoft, Google and VW pile in, so what's behind Norway's data center boom?*. ZDNet, viewed 11 December 2019, https://www.zdnet.com/article/microsoft-google-and-vw-pile-in-so-whats-behind-norways-data-center-boom/
[148] National Grid (2018) *Future Energy Scenarios*. National Grid, viewed 11 December 2019, http://fes.nationalgrid.com/media/1363/fes-interactive-version-final.pdf
[149] UK Research and Innovation (2020) UKRI Environmental Sustainability Strategy, viewed 9 June 2020, https://www.ukri.org/news/ukri-launches-its-environmental-sustainability-strategy-and-sets-a-path-to-a-net-zero-future/
[150] International Organisation for Standardization (2006), *ISO 14040:2006 Environmental management – Life cycle assessment – Principles and framework*, viewed 14 July 2020, https://www.iso.org/standard/37456.html
[151] International Organisation for Standardization (2006), *ISO 14044:2006 Environmental management – Life cycle assessment – Requirements and guidelines*, viewed 14 July 2020, https://www.iso.org/standard/38498.html
[152] TOP500 (2020) *June 2020*, TOP500.org, viewed 9 October 2020 2019 https://www.top500.org/lists/top500/2020/06/
[153] Office for Artificial Intelligence (2019) *Draft Guidelines for AI procurement*. UK Government, viewed 11 December 2019, https://www.gov.uk/government/publications/draft-guidelines-for-ai-procurement/draft-guidelines-for-ai-procurement
[154] Graphcore website, viewed 11 December 2019, https://www.graphcore.ai/
[155] British Broadcasting Corporation (2016) *ARM chip designer to be bought by Japan's Softbank*, viewed 11 December 2019, https://www.bbc.co.uk/news/business-36822806
[156] http://www.compoundsemiconductorcentre.com/
[157] Conway, S., Joseph, E., Sorensen, R. and Norton, A. (2019) *The Business Value of Leading-Edge High-Performance Computing: 2019 Update*. Hyperion Research LLC, viewed 26 November 2019, https://www.hpe.com/uk/en/resources/solutions/hyperion-hpc-value.html
[158] Google Cloud (n.d.) *Cloud Tensor Processing Units (TPUs)*, viewed 17 July 2020 https://cloud.google.com/tpu/docs/tpus

[159] NVIDIA (n.d.) *NVIDIA VOLTA: The Tensor Core GPU Architecture designed to Bring AI to Every Industry,* viewed 17 July 2020 https://www.nvidia.com/en-gb/data-center/volta-gpu-architecture/

[160] Oak Ridge National Laboratory (2018) *Genomics code exceeds exaops on Summit supercomputer,* viewed 17 July 2020 https://www.olcf.ornl.gov/2018/06/08/genomics-code-exceeds-exaops-on-summit-supercomputer/

[161] Palmer, T. (2015) *Modelling: Build imprecise supercomputers*, viewed 17 July 2020, https://www.nature.com/news/modelling-build-imprecise-supercomputers-1.18437