Co-funded by the
Erasmus+ Programme
of the European Union

Erasmus + **KA2: Cooperation for innovation and the exchange of good practices**

Knowledge Alliances

Enabling SMEs to gain competitive advantage from the use of HPC

D5/2 Development of region-specific HPC for SMEs Training Material

Ireland

Prepared by: P5 – Arctur d.o.o.

Version: v1.0 Final

Project Acronym	**SME/HPC**				
Project Number	588372-EPP-1-2017-1-IE-EPPKA2-KA				
Project URL	www.smehpc.eu				
EU Project Officer	Silvia.DE-BENEDETTI@ec.europa.eu				
Project Coordinator	**Name**	Prof. Bill O’Gorman		**Affiliation**	WIT
	Email	wogorman@wit.ie	**Phone**	+353.872429537	
Document Name	**Development of region-specific HPC training material** with links to other teaching materials, manuals and documentation for achieving advanced knowledge.				
Deliverable Number	D5/2				
Type	On line document	**Distribution Level**	Public		
Responsible Author(s)	P5: ARCTUR d.o.o.				
Contractual Date of Delivery	M24				
Status	Final				
Quality assurance readers	Assoc. Prof. Dr. Nadja Damij (P2)	Prof. Bill O’Gorman (P1)			

Version	Date	Status	Author (Unit)	Description
0.1	29th November 2019	Draft	Arctur (P5)	Generic version (D5.1) adjusted to the region-specific context of Ireland
0.2	13rd December 2019	Draft	Arctur (P5) WIT (P1)	Version updated and improved on the basis of specifications, comments and suggestions received from WIT
0.3	13th December 2019	Final	Arctur (P5)	Additional improvements by Arctur
0.4	18th December 2019	Final	UNN (P2) WIT (P1)	Quality Review
1.0	23th December	Final	Arctur (P5)	Final

Table of Contents

Introduction

High performance computing (HPC) is a rapidly developing field of expertise, where new systems and technologies are emerging with immense speed. Small and medium-sized enterprises (SMEs) from less developed regions have limited experience in teaching short-term training courses on HPC since their regions are not at the forefront of HPC development in addition to their weaker university-enterprise linkages. Therefore, staff capacity building through training in necessary to address existing gaps in relation to HPC expertise and utilization. This document is designed and developed for SME staff to gain knowledge about HPC, understand purpose of HPC usage, identify SME opportunities for HPC usage and get examples of problem solving when they are using HPC.

Study Learning Objectives

The purpose of this course is to increase SME staffs' knowledge of HPC domain, including the importance of HPC, application areas, technology components and systems, and how HPC could help SMEs.

After completing this self-study course, you will be able to:

- Understand the use and benefits of HPC for SMEs
- Know how to use HPC and be familiar with Linux Command Line Interface
- Be competent to promote available HPC systems and services
- Be able to identify and access additional relevant HPC re/sources
- Gain the understanding of HPC usability for SMEs
- See good practices of HPC usage and gain the knowledge to identify opportunities for SMEs
- Understand the administrative process of SME engagement with HPC providers
- Be able to formalize engagement with HPC providers
- Gain the understanding of HPC infrastructure
- Get hands on experience by remotely working on HPC (example problem solving)
- See specific use cases for domain environment

Organization of this Study Material

This self-study course material is organized in five distinct modules. Modules are targeted at providing knowledge at different levels: Basic, Intermediate and Advanced. Module names and referring level are mentioned below. The aims and desired learning outcomes for each module are described at the individual module level.

Module Number	**Module Level**	**Module Name**
1	Basic	Basic Understanding High Performance Computing (HPC)
2	Basic	Understanding HPC usage and identify opportunities for SMEs
3	Basic	SME/HPC provider relationship management
4	Intermediate	Exercises and solutions to practical problems by using HPC
5	Advanced	Complex (advanced) exercises and solutions to practical problems by using HPC

SME/HPC

Co-funded by the Erasmus+ Programme of the European Union

Module 1

Understanding High Performance Computing (HPC)

Basic

Aims of this Module:

This module aims to describe / explain:

- Understand the use and benefits of HPC for SMEs.
- Understand how to use HPC and be familiar Linux Command Line Interface.
- Be competent to promote available HPC systems and services.
- Be able to identify and access additional relevant HPC re/sources.

Prerequisites:

You do not need to have any previous knowledge of HPC but familiarity with basic computing concepts is desired.

Learning Outcomes:

At the end of this module you will be:

- Able to understand the concepts of HPC.
- Familiar with the history and evolution of HPC.
- Able to explain the importance of HPC for scientific and business applications.
- Able to identify the potential of HPC.

Module 1

- ✓ What is High Performance Computing?
- ✓ How is HPC different from regular desktop computing?
- ✓ Why is HPC important?
- ✓ History of HPC
- ✓ Reflections on this module learning
- ✓ Recommended Further Readings / Resources
- ✓ Basic linux CLI commands

1. Introduction to High Performance Computing (HPC)

1.1 What is High Performance Computing?

High-performance computing (HPC) involves the use of "supercomputers" and massively parallel processing techniques to solve complex computational problems through computer modelling, simulation, and data analysis (Technopedia, 2016). High-performance computing brings together several technologies, including computer architecture, programmes and electronics, algorithms, and application software under a single system to address advanced scientific and business computational needs quickly and effectively. HPC technology focuses on developing parallel processing algorithms and software so that programmes can be divided into small independent parts, and can be executed simultaneously by separate processors by incorporating both administration and parallel computational techniques. HPC systems have the ability to deliver sustained performance through the concurrent use of computing resources. The terms "high performance computing" and "supercomputing" are sometimes used interchangeably.

1.2 How is HPC different from regular desktop computing?

While a desktop computer / workstation / laptop today typically contains a dual or quad core processor, an HPC system essentially represents a network of processors, each of which contains multiple computational cores as well as its own local memory to execute a wide range of software programs (NICS, 2016). To put it into perspective, a laptop or desktop with a 3 GHz processor can perform around 3 billion calculations per second. While that is much faster than any human can achieve, it pales in comparison to HPC solutions that can perform quadrillions (1 quadrillion = 10^6 billion) of calculations per second.

One of the best-known types of HPC solutions is the supercomputer. A supercomputer contains thousands of compute nodes that work together to complete one or more tasks through parallel processing, which is similar to having thousands of PCs networked together, combining compute power to complete tasks faster. The software programmes that coders write to run on supercomputers are divided up into many smaller independent computational tasks, called "threads," that can be executed simultaneously on these cores. Modern can consist of over 100,000 "cores" or more and for supercomputers to operate effectively the 'interconnection' between the cores must be well designed. For example, Cray Titan, one of the fastest supercomputers in the world, contains just fewer than 300,000 cores, which are capable of operating more than 6,000,000 concurrent threads (Shimpi, 2012).

In other words, a supercomputer can be compared to tens of thousands of workstations performing together like a symphony orchestra to process billions and trillions of bits of data every second, sometimes for hundreds of users simultaneously. Some supercomputers are general or multipurpose machines that perform diverse tasks such as modelling and simulation or advanced business data analytics; others may be dedicated to specific tasks, such as operating cloud-based services, such as music streaming or managing telecommunications infrastructure (Carey, 2015).

1.3 Why is HPC important?

The Digital Economy is developing rapidly worldwide. It has been cited as the single most important driver of innovation, competitiveness and growth. Digital innovations such as supercomputing are an essential driver of innovation and spur the adoption of digital innovations across multiple industries and small and medium-sized enterprises, fostering economic growth and competitiveness. Applying the power of supercomputing combined with Artificial Intelligence (AI) and the use of Big Data provide unprecedented opportunities for transforming businesses, public services and societies.

Modelling and simulation have come to complement theory and experiment as a key component of the scientific method, and many of our scientific findings and technological advances rely on models simulated on high performance computers. The importance of this approach becomes clear with the following statement made by Andy Searle, Head of Computer Aided Engineering, Jaguar Land Rover. “It costs £500,000 (€610,000) to do each physical test of a car crash, and it’s not repeatable. It costs £12 (€14.64) to run a virtual simulation of a car crash, and it’s fully repeatable, so it can be used to optimize the design of a vehicle.” (BIS, 2013). This cost case is based on a study which is already around 6 years old and costs for the same quality of simulation have dropped since then. High-performance computing is mature technology mainstream for 40 + years, driven by academia and research, and it is based on well-understood mathematical, scientific and engineering principles, for solving advanced problems and performing research activities through computer modelling, simulation and analysis. However, HPC application areas are broader than simply modelling and simulation and cover large-scale data processing, real-time HPC and other applications. More recently, real-time HPC systems have appeared which can assist humans in instantaneous decision taking (for example, a medical simulation that supports a surgeon performing an operation).

The performance of HPC systems is increasing, as Figure 2 illustrates, growth in supercomputer operating speeds has increased exponentially over the past two decades (doubling in just 1.5 years). The digital revolution of the past two decades has led to unprecedented demands for high-performance data-processing systems. It is anticipated that the HPC will continue to be a vital technology, and demand for greater computing power will continue. One of the areas of HPC system application is in dealing with a wealth of data amassing to Terabytes/Petabytes of information. The Big Data development is relatively recent, in the last 10–15 years, but is evolving rapidly, and still developing its foundations and scientific underpinnings. Synergy of HPC and Big Data paradigms will enable new insights and ways the research and development of risk data analysis is undertaken (Holmes and Newall, 2016). Figure 1 demonstrates how HPC can accelerate the product development process cycle and reduce the “time to market” for a company providing competitive advantage. Some of the key application areas of HPC are described later in this module (and also in module 3) with case examples. These examples will help to further understand the importance of HPC in both scientific research and business application domains, to address some of the most advanced and complex problems.

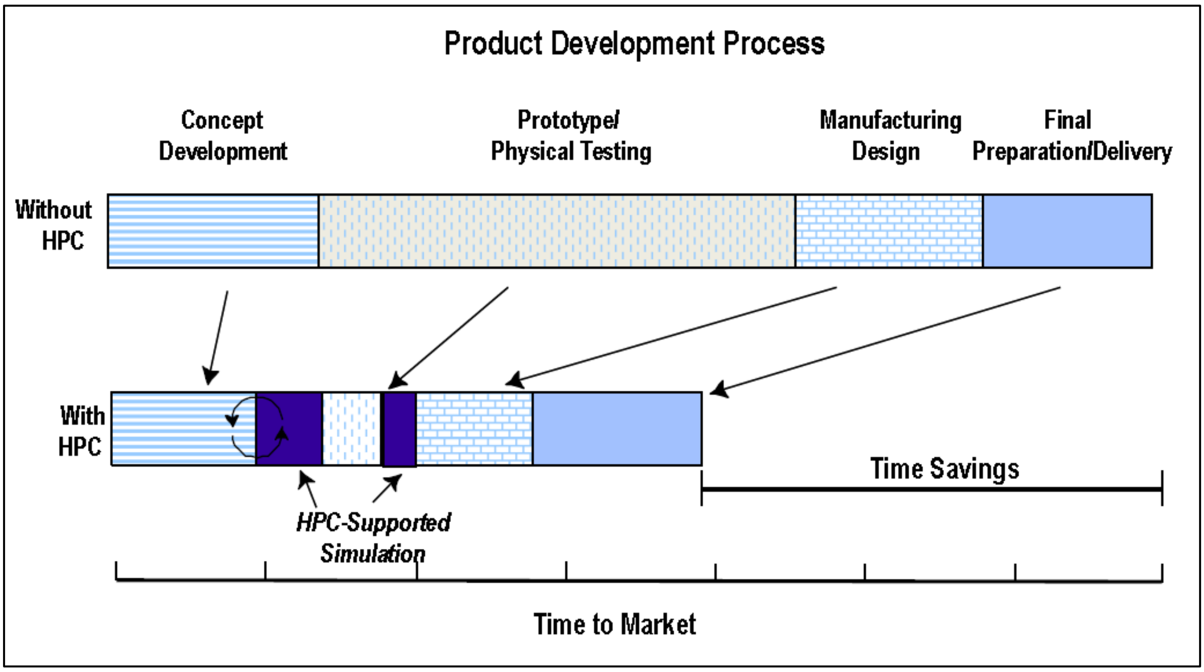

Figure 1: *HPC advantage: Reduction in time to market [Source: IDC]*

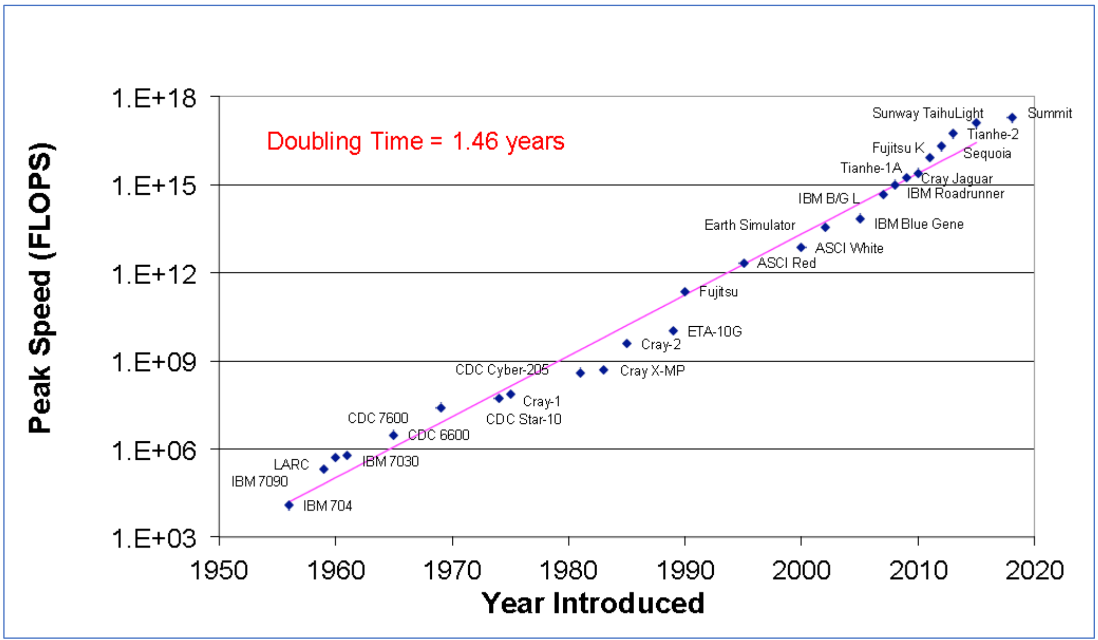

Figure 2: *Growth of Supercomputing (1950-2020)*
Source: https://www.top500.org/lists/2018/11/

1.4 History of High Performance Computing

The famous quote ``The Only Thing Constant Is Change" seems certainly to be true for the market of High-Performance Computing (HPC), If we look back on the last seven decades, it is clear how the marketplace for HPC has rapidly changed in terms of vendors, architectures, technologies and applications. Despite all these changes, however, the evolution of performance on a large scale seems to be a very steady and continuous process. Moore's Law is often cited to describe the progress in this context and as Figure 3 shows how well this law holds for a significant lifespan of modern computing between 1950 and 2000. On an average we see an increase in performance of 2 orders of magnitude every decade. See also Figure 4 *Speed of world's fastest and 500th-fastest Supercomputer (1993-2015)*

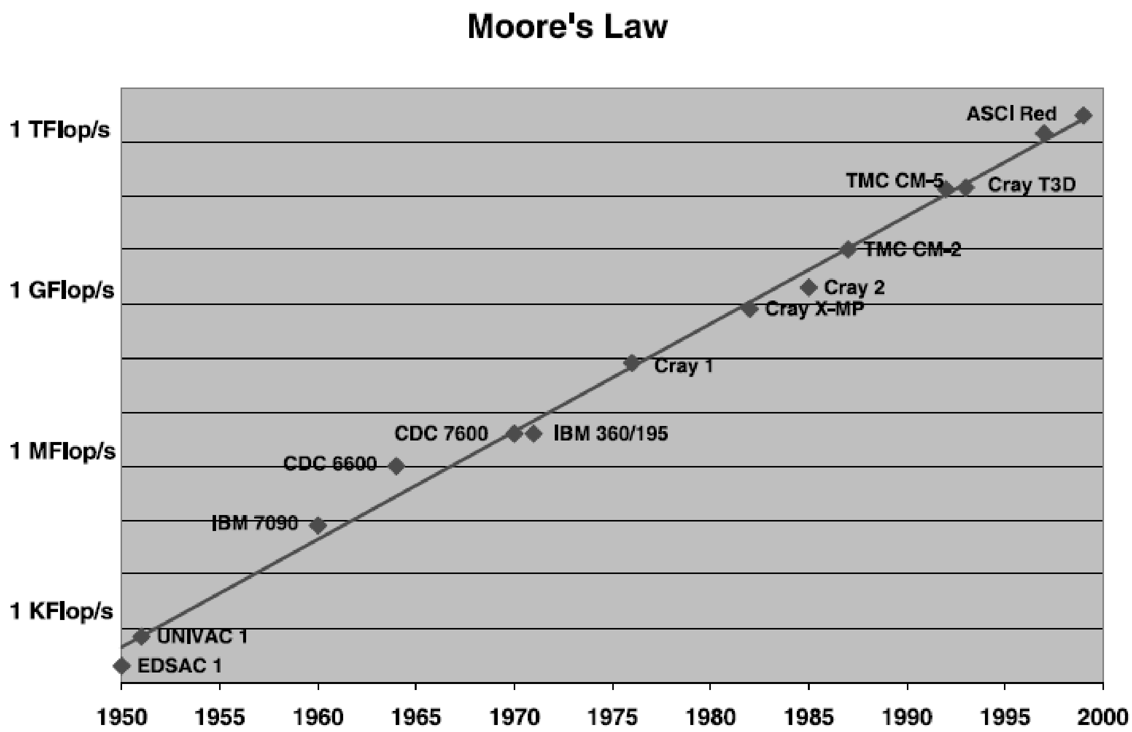

Figure 3: *Performance of fastest computer systems for last 5 decades compared to Moore's Law [Source: Strohmaier et al., 1999]*

1940s to 1960s: The first Computers

During this period the development of the computers was largely driven by aviation and military needs. Invented by a German scientist named Konrad Zuse, the Z3 was presented at the German Laboratory for Aviation in Berlin on May 12, 1941, as the world's first universal programmable, binary-based, electromechanical computer. Colossus was the first programmable, digital, electronic computer developed in the UK (Bletchley Park, England) in 1943 during World War II. It used vacuum valves and was designed for a single task of breaking Nazi codes like Enigma. In the USA, ENIAC (Electronic Numerical Integrator and Computer) was the first stored-program electronic computer designed and built at the University of Pennsylvania from 1943 to 1946. During the Cold War period, supercomputing was primarily used for design of Nuclear weapons, Aircrafts, Submarines etc., intelligence gathering and processing, and code breaking.

1970s to 1990s: Vector Era (Shared Memory Computers)

During this period speed in supercomputers was primarily achieved through two mechanisms:

- Vector Processors: these were designed using pipeline architecture to rapidly perform a single floating point operation on a large amount of data. Achieving high performance depended on data arriving in the processing unit in an uninterrupted stream.
- Shared Memory Multiprocessing: a small number (up to 8) of processors with access to the same memory space. Interprocess communication took place via the shared memory.

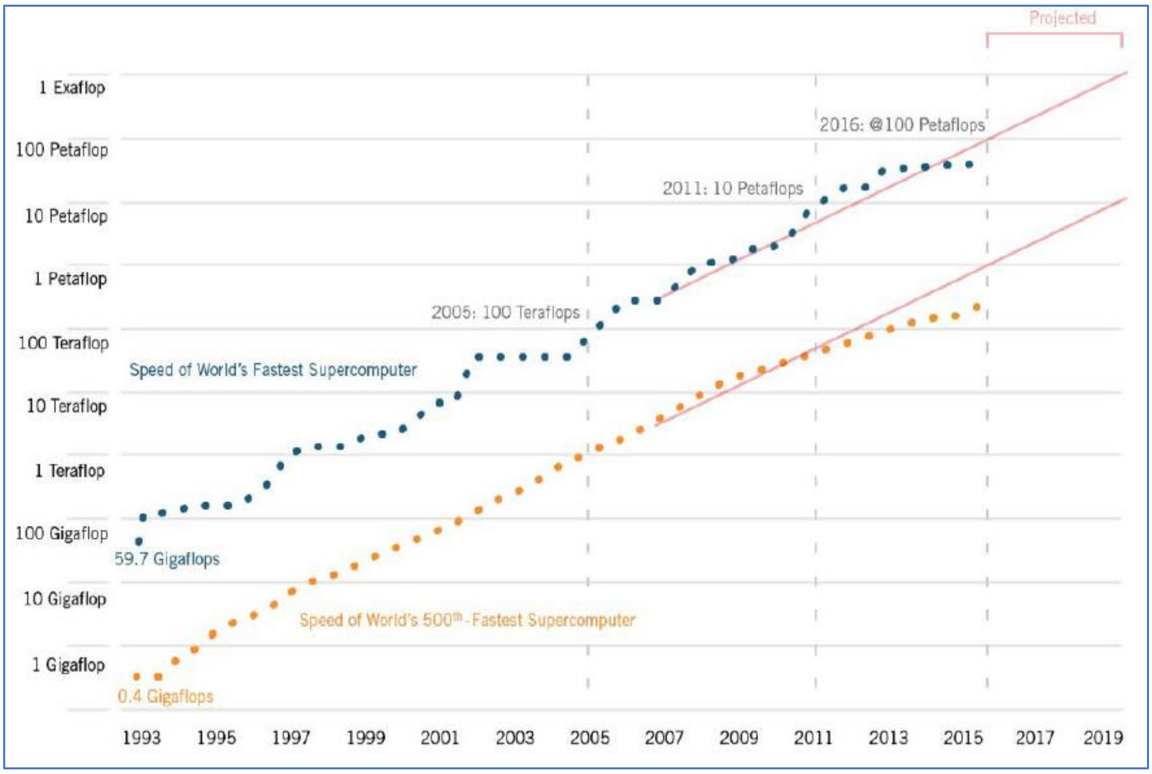

Figure 4: *Speed of world's fastest and 500th-fastest Supercomputer (1993-2015)*
Source: Ezell and Atkinson (2016)

1990s to 2000s: Cluster (Distributed Memory Computers)

The age of truly effective parallel computers had begun, but was already limited by access to shared memory. Memory contention was a major impediment to increasing speed; the vector processors required high-speed access to memory but multiple processors working simultaneously created contention for memory that reduced access speed. Vector processing worked well with 4 or 8 processors, but memory contention would prevent a 64 or 128 processor machines from working efficiently. The alternative to shared memory is distributed memory, where each processor has a dedicated memory space. The challenge became implementing effective processes communication - processes can't communicate with one another by writing data into shared memory; a message must be passed. During the 1990s there began to be a lot of interest in distributed memory computers.

2000 to Present: GPU and Hybrid Cluster era

During the 2000s the trend of increasing processor speeds was replaced with increasing the number of processor cores. This led to hybrid clusters with a large number of processors, each with a small number of core sharing RAM and some cache space. With the development of special purpose processor units such as GPU (Graphics Processing Unit) and other general purpose accelerator hardware, today's top supercomputers are hybrid clusters with a large number of standard processor nodes where each node has a multicore processor with some individual cache, some shared cache, and RAM shared by all cores.

Latest developments in computing and internet technology has led to Utility computing - the new Cloud Computing paradigm. This on-demand computing or service computing is a concept of connecting to the external computing resources via Internet. It is defined as: a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted virtualised, dynamically-scalable, managed computing power, storage, platforms and services are delivered on demand to external customers over the internet (Foster et al., 2008). Currently, a significant research focus lies on designing HPC solutions that integrate capabilities of High-performance computing with pay-per-use model of Cloud computing (Li, 2015).

Analysts measure the speed of computers by their ability to calculate floating-point operations per second (or “flops”). As Figure 4 illustrates, growth in supercomputer operating speeds has increased exponentially over the past two decades. See also Table 1, *List of the top 5 fastest Supercomputers available currently*.

Name	Peak Speed	Developed By	Location
Summit	148.6 PFLOPS	IBM	Oak Ridge National Lab, USA
Sierra	94.6 PFLOPS	IBM	Lawrence Livermore National Lab, USA
Sunway TaihuLight	93.0 PFLOPS	NRCPC	National Supercomputing Centre in Wuxi, China
Tianhe-2	61.4 PFLOPS	NUDT	National Super Computer Centre in Guangzhou, China
Fortera	23.5 PFLOPS	Dell EMC	Texas advanced computing center/univ. of Texas.

Table 1: *List of the top 5 fastest Supercomputers available currently*
Source: https://www.top500.org/lists/2018/11/

1.5 Some of the Application Areas for HPC

In section 1.3, it has been mentioned how HPC is used for solving advanced problems and performing research activities through computer modelling, simulation and analysis. While HPC has traditionally been implemented for scientific research and computational science through use of supercomputer, recently the focus is gradually shifting from supercomputers to computing clusters.

HPC is currently used to solve complex modelling problems in a spectrum of disciplines, addressing goals for both scientific research and business applications. Some of the key disciplines where HPC is been heavily used are:

- Artificial Intelligence and Machine Learning
- Climate Modelling and Weather predictions
- Automotive Engineering
- Agricultural Production
- Financial Markets
- Cryptographic Analysis
- Personalised / Precision Medicines
- Molecular Biology and Genetics
- Nuclear /Plasma /Quantum/Geo Physics
- Physical Oceanography
- Product Design

1.6 South East Ireland HPC usage

The region has one supercomputer named ‘Kay’ which is hosted in the Telecommunications Software and Systems Group’s (TSSG) data centre in Waterford Institute of Technology. In order to provide technical support to researchers across Ireland, Irish Centre for High-End Computing (ICHEC) has provided High Performance Computing (HPC) resources. This enabled a diverse range of advanced research which is illustrated by the numerous publications in high-impact journals (including, Nature, Science) acknowledging ICHEC's HPC and human resources[1].

ICHEC regularly organises roadshows across all the Third-Level Institutions (Universities and Colleges) in Ireland with the intention to engage the scientific community and user in order to create a vivid ecosystem between HPC and Science in Ireland. Additionally, ICHEC also offers extensive training and education programme that caters for researchers across third level institutions and the Irish workforce. The courses range from induction training courses to university education and mentorship opportunities that cover basic to advanced HPC skills. These academic courses are typically hosted at Irish third level institutions with participation open to all academic researchers without fees.

1.7 European Exascale Projects

High Performance Computing (HPC) is at the core of major advances and innovation in the digital age – as the European Commission (EC) stated in a recent Cordis article. In the current framework programme, Horizon 2020, Europe has committed €700M to supercomputing research. European ambitions in this field are high: the declared goal is to have a supercomputer based on EU technology featuring among the world top three by 2022.
To reach this aim, Europe can build on Exascale research efforts that got under way in the previous funding framework, FP7. Between 2011 and 2016, eight projects, with a total budget of more than €50 Million, were selected for this first push in the direction of the next generation

1 https://www.ichec.ie/academic/national-hpc/ns-activities

supercomputer: CRESTA, DEEP and DEEP-ER, EPiGRAM, EXA2CT, MontBlanc (I + II) and Numexas. The challenges they addressed in their projects were manifold: innovative approaches to algorithm and application development, system software, energy efficiency, tools and hardware design took center stage.

Over the last five years, the projects have joined forces and started developing a European Exascale Community. The idea was to leverage synergies between the projects and establish a platform for scientific exchange – but also to increase the visibility of European efforts in HPC on a global scale.

The birth of a new community

The collaboration started with individual researchers from the various projects discussing and exchanging ideas and concepts. Developing a more formal framework for the collaboration was the logical consequence of these fruitful exchanges. Already in 2012 DEEP and MontBlanc teamed up to host a training workshop together. In 2013, the cooperation was further intensified: the European Exascale Projects (EEP) Workshop was born. Since then, the growing community has been meeting once a year for a workshop in various formats to intensively discuss the projects' status, talk about lessons learned, and in this way identify synergies as well as cooperation opportunities.

Coming of age – the community grows

Soon the members involved wanted to use the opportunity and get in touch with interested audiences outside of their core circle. While the first Birds-of-a-Feather (BoF) sessions focused on European Exascale approaches and strategies, a more scientific and technical orientation followed. BoFs on topics like 'Programming Models for Exascale: Slow Transition or Complete Disruption?', or workshops on prototyping, were organized at international conferences and workshops such as ISC and SC. A special focus was also devoted to the 'beneficiaries' of Exascale research: HPC applications in critical fields of the European Research Arena. At PRACEdays15 the topic was approached from a more industry specific point of view whereas at EASC2016, scientific applications were in the spotlight.

Handing over the torch to the next generation

The European Exascale Projects bring together experts from world-leading companies, leading supercomputing institutions, and outstanding academics who were the first to embark on the European Exascale adventure. In their brochure, the frontrunner FP7 projects presented their results as well as shared their insights on the European Exascale Community. Last but not least, they used the opportunity to hand over the torch to the new generation of H2020 projects, represented via EXDCI – the European Extreme Data and Computing Initiative – to keep up the collaboration initiative, intensify it and streamline the development of the European HPC Ecosystem on the way to next-generation supercomputers.

1.8 Basic Linux commands

Open-source Linux is a popular alternative to Microsoft Windows, and if you choose to use this low-cost or free operating system, you need to know some basic Linux commands to configure, operate, and interact with your system smoothly. Linux is dominant operating system in HPC usage, so if you want to operate with most of HPCs, you will need to know some basic commands.

When dealing with the Linux operating system, commands are required as inputs to inform or direct a computer programme to perform a specific operation. Understanding the most basic Linux commands will allow you to successfully navigate directories, manipulate files, change permissions, display information such as disk space, and more. Obtaining basic knowledge of the most common commands will help you easily execute tasks via the command line. Find the most common Linux commands in this table:

Command	Description
cat [filename]	Display file’s contents to the standard output device (usually your monitor).
cd /directorypath	Change to directory.
chmod [options] mode filename	Change a file’s permissions.
chown [options] filename	Change who owns a file.
Clear	Clear a command line screen/window for a fresh start.
cp [options] source destination	Copy files and directories.
date [options]	Display or set the system date and time.
df [options]	Display used and available disk space.
du [options]	Show how much space each file takes up.
file [options] filename	Determine what type of data is within a file.
find [pathname] [expression]	Search for files matching a provided pattern.
grep [options] pattern [filesname]	Search files or output for a particular pattern.
kill [options] pid	Stop a process. If the process refuses to stop, use kill -9 pid.
less [options] [filename]	View the contents of a file one page at a time.
ln [options] source [destination]	Create a shortcut.
locate filename	Search a copy of your filesystem for the specified filename.
lpr [options]	Send a print job.
ls [options]	List directory contents.
man [command]	Display the help information for the specified command.
mkdir [options] directory	Create a new directory.
mv [options] source destination	Rename or move file(s) or directories.
passwd [name [password]]	Change the password or allow (for the system administrator) to change any password.

.

1.9 Reflections on this module learning

In this module you saw the benefits of HPC usage for SMEs, gained understanding of how to use HPC and got familiar with Linux Command Line Interface. The history and evolution of HPC was presented. Now you are able to explain the importance of HPC for scientific and business applications and identify the potential of HPC.

1.10 Recommended Further Readings / Resources

1. High Performance Computing for Dummies: Wiley Publishing, Inc. (http://hpc.fs.uni-lj.si/sites/default/files/HPC_for_dummies.pdf):
   In this document you will get all the basic information about HPC, hardware, software of HPCs, basically all information for a complete layman's understanding in the HPC field.
2. Handbook for getting started, Sesame net (https://sesamenet.eu/wp-content/uploads/2016/12/D4.2_fin.pdf):
   This handbook will explain the economic benefits of using HPC for people at managerial positions and will provide a path for attaining this benefits through collaboration with the HPC Competency Centers in SESAME Net. Apart from the purely economic reasons we consider the improvements in innovation and marketing and the competitive advantages stemming from the use of HPC to provide justification for taking the managerial decision to explore the use of HPC.
3. Europe Towards Exascale, exascale-projects (http://exascale-projects.eu/EuroExaFinalBrochure_v1.0.pdf):
   This document outlines the landscape of the European HPC Eco-system. An emphasis is placed on the current European ambitions in the development of Exascale level supercomputers and related solutions. It also identifies the areas of HPC technology that demonstrate a potential for international collaboration.
4. European technology platform for high performance computing, update on European HPC (https://www.etp4hpc.eu/pujades/files/Update%20on%20European%20HPC%20-%20March%202018%20-%201.pdf):
   This document outlines the landscape of the European HPC Eco-system. An emphasis is placed on the current European ambitions in the development of Exascale level supercomputers and related solutions. It also identifies the areas of HPC technology that demonstrate a potential for international collaboration.
5. Exascale computing in the United States

(https://www.researchgate.net/publication/328844092_Exascale_Computing_in_the_United_States):

This document outlines the use of Exascale computing in the United States and explains exascale computing project.

6. Exascale computing and big data: the next frontier (https://pdfs.semanticscholar.org/c6f1/42f5011ddb90921f1185b14a147b807119f9.pdf?_ga=2.50376745.1895536544.1567426056-1108196605.1567426056):

   In this document you will get an introduction to exascale computing about, advanced scientific computing, scientific and engineering opportunities, technical challenges in advanced computing and economic and political challenges.

Videos / Podcasts

1. https://www.youtube.com/watch?v=bkLVuNfiCVs (Introduction)
2. https://www.youtube.com/watch?v=A_i5kOlj_UU (HP Video)

MODULE 2

Understand HPC usage and identify opportunities for SMEs

Basic

Aims of the Module

This module aims to describe/ explain:

- Gain the understanding of HPC usability for SMEs.
- See some good practices of HPC usage.
- Gain knowledge to identify opportunities for SMEs.

Prerequisites

Self-study of Module 1

Learning Outcomes

At the end of this module you will

- Have an understanding of HPC usability for SME.
- Be familiar with good practices of HPC usage.
- Gain knowledge to identify opportunities for SMEs.

Module 2

- ✓ Aspects and benefits of using HPC technology
- ✓ Use cases
- ✓ Success Stories
- ✓ Roadmap for improving SME uptake HPC

2. Understand HPC usage and identify opportunities for SMEs

2.1 Aspects and benefits of using HPC technology

This paragraph explains why usage of HPC may benefit SMEs. You will see some of the reasons and aspects why HPC usage is good and how some SMEs made success with HPC usage.

2.1.1 Economic reasons:

For example, reducing designing and prototyping costs; detecting design errors in early stages of product development. It is important to always focus on the CEO-like argument, so in detecting design errors earlier it should be mentioned the positive economic impacts that this would have: early detected design errors are cheaper to overcome. Another example, if prototyping has a major role in the SME's value chain then HPC saves the costs of building many expensive physical prototypes which now can be built virtually, minimizing the number of physical prototypes to be built.

The three key economic reasons to consider integrating HPC in SMEs (and also in big enterprises) are:

- Cost reduction: Prototyping (wind tunnel, physical testing) and design costs;
- Reducing design times, reducing time to market (translating this time reduction in money);

- New product/service development that is not possible to develop without HPC.

Cost reduction

HPC can help the SME to save costs, for example in **prototyping and design procedures**, as well as in **operational costs**. Physical prototyping is to create a physical prototype (“An early sample, model, or release of a product build to test a concept or process or to act as a thing to be replicated or learned from” (4)). This process is very time and cost consuming, because of the time needed to produce an item outside of an assembly line and the cost of prime matters, and this only gets worse when multiple prototypes are needed, especially in the case of destructive physical testing when prototypes can only be used once. Physical prototypes can be substituted by virtual prototypes that are cheaper to develop and can be used again and again without destroying them.

Virtual prototypes once developed can be used to simulate the physical behavior or interactions under many circumstances: aero and hydrodynamics, substituting the need of complexity and expensiveness of wind tunnels or fluid tests, useful in: Aerospatiale, automotive, naval and constructive; Physical tests, virtual prototypes can be used to perform multiple virtual test using simulation, no matter if these tests simulate physical damage or full destruction of the prototype (Blackwell, A. H.; Manar, E., eds. (2015). "Prototype". UXL Encyclopedia of Science (3rd ed.). Design costs can be reduced using HPC, since HPC brings us the power to design prototypes of products without the need of any physical prototype and begin to get knowledge about it in the very early stages of design. This also improves the design process allowing detecting design errors or misconceptions in these early stages of design, where they can be corrected easily and cheaper than in later stages. Operational costs of the SMEs can also be saved by accurate and extensive simulation, modelling and optimization of certain kind of operations. For example, HPC can be used in optimization of public transport or logistics. It is widely used in the search of oil and gas or in energy production problems.

Reducing design times

In some sectors like aeronautics, engineering or energy, where the interaction of fluids with products like: aircrafts, shipbuilding, turbines, wind or offshore energy generators is a key factor of efficiency, HPC can be a key to success because it can **reduce design times drastically**.

The times needed to construct a virtual prototype and discover the full interactions with their surrounding or inner contained fluids are much shorter (Handbook for Getting Started) than the time needed to get this knowledge from physical ones. This is an uncontestable advantage that can make the difference between a successful and a ruinous project, as well as enable a project that otherwise is never started because HPC possibilities were not considered or simply were unknown when tight schedules were present and standard design times are not possible.

As an example of the prior two reasons, under the umbrella of the EU funded project **FORTISSIMO**, the Italian SME EnginSoft reduced the time needed for full design and optimization of a single pump turbine **from 2-3 years to 6 months** and this fact allowed recovery of the full SME investment in HPC Cloud based simulation and external expertise in less than six months.

Products impossible to produce without HPC

Another example of economic benefit from HPC is the possibility **to produce or study products that with standard technology limitations are impossible or too costly to design and produce**. The study of hazardous substances (explosive, toxic or mutagenic) has heavy limitations due to their inherent properties and security measures needed to manipulate them. The physical measures needed to overcome these limitations are prohibitive in terms of cost, so the physical experimentation with these substances will have a bad ROI in most cases and involve high risks in all of them. HPC can overcome these limitations and simulate the behavior of these substances with virtual models getting knowledge about them in a non-risky environment.

Limiting the number of physical tests only to experiments with high percentages of success

Another FORTISSIMO experiment shows how a Swiss SME, Lonza from the chemical sector, suppliers to the Pharma & Biotech sector and especially to the ingredients market, explored with success the possibility to determinate physical properties of compounds with desired precision using HPC. The use of HPC clusters reduced the amount of time needed to perform calculations from a six-months-long calculation in a 16- core cluster, to less than one day on an HPC system.

2.1.2 Innovative reasons:

Most of the advantages of using HPC technology can be obtained by SMEs that have substantial Research & Development departments, since the access to HPC resources can improve the possibilities to develop innovative products and services. By collaborating with an established HPC Competency Centre, the SME will not need to incur costly on-premises infrastructure. The costs of owning and operating such infrastructure are a heavy burden for most SMEs. By using flexible contracts SMEs can have a low-risk, low-effort, pay-as-you go access to HPC and leverage state-of-the-art software and tools to speed-up their development process.

Apart from the cost-related reasons, there are many other main advantages to using HPC, some of which are:

- Obtaining results much faster by employing a high amount of computing power;
- Achieving much more accurate representation of the product and its working environment;
- Exploring much wider search space for the parameters of the envisaged product.

By collaborating with a leading HPC Competency Centre, the personnel at the SME's Research & Development department acquire precious expertise that is also useful in other, more standard situations. For example, the increased use of multi-core devices in all kinds of products makes knowledge of parallel computing paradigms an important asset. The need to achieve a given results with as little use of energy as possible stresses the importance of using optimal algorithms and scalable programming approaches. Overall, the use of HPC for completing simulations, that would normally run for days or weeks on a desktop workstation, in a matter of hours enables much faster time-to-market and thus, possibly, first-mover advantage for the SME. However, it should not be underestimated that being in close contact with academia and HPC experts in industry, that SMEs can have advanced knowledge of new developments in HPC technology or applied sciences that will enable cross-fertilization and rapid incorporation of the new technologies into their product portfolio.

Best product build with best technology:

In our society, customers have a much-improved perception of most products when they are built with the latest technology. Even if the use of this technology only involves a very low percentage of improvement, there are clients that always want to own the "state-of-art" product with the last improvements, while others are more concerned about the environment and want to have "green" products with more efficiency and low power consumption. HPC technology can deliver all these properties, since it is an innovative technology. However, it currently has relatively low use with SMEs because its applications and benefits are not yet widely understood. Innovative technologies can lead to innovative products, and for SMEs willing to improve their products and processes this can be a competitive advantage. HPC can be applied successfully to improve products and processes in order to make them more "green".

Under the umbrella of FORTISSIMO there are situations where products were optimized using HPC that gained special characteristics differentiating them from other products. For example, in the aerodynamics sector, Pipistrel, a Slovenian SME, used HPC to improve the design of their aircraft and KEW optimized the aircraft wiring design not only to generate cost savings prime materials but also to reduce the weight of prime materials. Also, less weight means less fuel consumption and over the 30 to 50 years of an aircraft's life this has major economic relevance. In the civil construction sector, the Scottish SME, IES, performs building energy efficiency studies but making the jump to using HPC enables them make studies of much greater scale thereby allowing them to expand their energy analyses and services from just buildings to cities.

2.1.3 Marketing reasons

An important marketing reason for the use of HPC is to present the enterprise as a cutting-edge technological firm. This has a positive effect in its market position and on the perception of its products and services by customers. And of course, it results in the subsequent positive economic impacts, e.g. more sales, price premiums, and so forth.

Leading brands partnerships are a key:

When selling a product and you want to reach a market where a “brand” or “organization” is well known and with a good reputation, it is a good idea to be associated with a leading “brand” or “organization” so that your product gets an improved perception over other products that do not have this advantage. For example, SESAME Net is composed of very well-known and well-considered partners, therefore the collaboration with SESAME HPC partners can lead to a win-win situation, where the new product is associated with a well-known brand and on the other hand SESAME Net also gets improved visibility.

Be a leading brand in their sector:

HPC can be the difference between one leading brand and their competitors that don’t use this technology. To associate your brand with innovation and differentiate against competitors’ products, the use of HPC for product improvement is a good point that highlighted correctly can increase a product brand´s sales. End-customers always want to buy best cost-value products and including HPC in the value chain of the SME can be a technology to achieve these added-value products. This consideration is even more visible in the provision of services, where the use of powerful HPC equipment can lead to increased consumer trust. In all those cases it is important to be able to substantiate the marketing claims with scientific and technical arguments.

2.1.4 Competitive advantage reasons:

High Performance Computing (HPC) plus data science allows public and private organizations get actionable, valuable intelligence from massive volumes of data and to use predictive and prescriptive analytics to make better decisions and create game-changing strategies. The integration of computing resources, software, networking, data storage, information management, and data scientists using machine learning and algorithms is the secret sauce to achieving the fundamental goal of creating durable competitive advantage.

Being able to bring new products to the market, thanks to the HPC edge on the research and development process, allows SMEs to gain a competitive advantage over their competitors. This first-mover advantage, thanks to the technological leadership, has several positive impacts for a firm. For example, brand recognition, this is the automatic association in costumers' minds of the new types of products to its initial manufacturer, e.g. the copying machines and Xerox, the tissue paper and Kleenex. Another example of first-mover advantage is the pre-emption of scarce assets, which allows the first-mover firm with superior information to purchase assets at market prices; below those that will prevail later in the evolution of the market. Additionally, HPC capable firms to reduce their product cycle time, the period of time a manufacturer needs to complete the development and production of a new, or modified, product. It does not only reduce costs, but an improvement in the time to market response allows for bigger flexibility to adapt to the changing market conditions. Also the use of HPC enables the design of a more accomplished product with less defects and covering a wider specter of consumer needs.

In the today's economy it is increasingly important to find consumer niche and
to engage with potential new clients. With the use of HPC technology, SMEs can employ advanced algorithms and processes that will increase their market share and reach wider consumer audiences.

2.2 Use cases

Based on the excellent collaboration between HPC competence centers and SMEs, SESAME Net partners developed several examples (use cases) illustrating how real companies benefit from HPC technologies in different areas of business. Examples of these use cases are presented as follows:

2.2.1 Biomedicine

HPC-Competence Centre

The Institute of Information and Communication Technologies (IICT) at the Bulgarian Academy of Sciences hosts the Bulgarian largest HPC centre (http://www.hpc.acad.bg/) and provides knowledge, expertise and services to businesses and researchers to develop innovative applications and products.

Enterprise

Established in 1995 AMET Ltd. is a company dedicated to the development, modern manufacturing and distribution of electronic medical equipment and modules, mechanical parts and units for incorporation into other products/equipment. With ~100 employees, AMET Ltd achieved a strong leadership position in a highly competitive market and therefore needs scientific expertise in order to further develop products with enhanced power and precision.

How HPC makes the difference

AMET Ltd., (see http://amet-bg.com/en/), started to develop a medical device for radiofrequency ablation of hepatic tumours. It needed precise mathematical modelling and computer simulation of the heat transfer process in order to optimize the parameters of this low-invasive therapy technique. Adequate representation of the problem was achieved by FEM discretization for a time-dependent partial differential equation of a parabolic type, generated based on segmented medical images. The 3D simulations were performed using a supercomputer and allowed precise estimation of the ablation parameters and an increase in power efficiency.

2.2.2 Manufacturing and Materials

HPC-Competence Centre

The Fraunhofer Institute for Algorithms and Scientific Computing SCAI offers the application of sophisticated in-house and/or commercial optimization and robust design methods in combination with numerical simulations for complex flow problems.

Enterprise

The participating industrial company was Hennecke Polyurethane Technology. Hennecke designs and constructs high-performance machines and plant technology for polyurethane processing, enabling their customers to achieve high-quality and efficient production results.

How HPC makes the difference

The cooperation between SCAI and Hennecke featured the optimization of two different subdomains of polyurethane processing. The first case examined an injector for mixing the polyurethane components Polyol and Isocyanate. Based on a sensitivity analysis varying several geometric parameters, an optimization process was launched, which yielded a set of optimal parameters for a required range of operating mass flow. In the second case, slab stock foam plants used for the production of foam blocks have been optimized, to achieve a foam of high quality. In both tasks the fluid solver STAR-CCM+ from CD-adapco, the optimization software DesParO from Fraunhofer SCAI and a self-developed process chain environment was applied to optimize processes and simulate them in acceptable computation times on SCAI's HPC Cluster.

2.2.3 Virtual Prototyping

HPC-Competence Centre

IT4Innovations National Supercomputing Centre is an important part of the infrastructure of the Czech Republic focused on HPC research and services. The Centre operates the most advanced HPC technologies and services and makes them available to Czech and foreign research teams from both academia and industry. One of the key functions of the IT4Innovations National Supercomputing Centre is to support industry in the Czech Republic.

Enterprise

The company BORCAD cz was established in 1990 as a construction and development studio. Today BORCAD cz is a leading European producer of railway and medical technologies. The company employs 190 people, exports to more than 80 countries worldwide and thanks to its unique design and original construction solutions is one of the most innovative companies in this field.

How HPC makes the difference

Numerical modelling and simulations are commonly used in the research and development of new products because they significantly shorten the time necessary to bring a new product to the market and also save a lot of money. The old-fashioned way of product development by trial and error, where usually quite costly prototypes have to be built and tested, is increasingly being replaced by numerical modelling of virtual prototypes. IT4Innovations has cooperated with BORCAD since June 2013. The company searched for a solution of a problem they had with entering the UK market with passenger seats for regional and long-distance rail transport. BORCAD needed certification that their passenger seat complied with regulations. The certification process involves a crash test which checks not only seat integrity but also bio-mechanical criteria measured on crash test dummies. Since physical tests are very expensive, BORCAD decided to use numerical modelling and simulation to perform a virtual crash test on new designs and perform only two physical tests. The cooperation helped BORCAD to obtain the certificate after successfully passing the crash tests with much less resources and time invested.

2.3 Use cases of Ireland region

South East Ireland

The region plays a minor role in the Irish economy accounting €17,899 million for regional GDP in 2015 (Eurostat, 2018) which is 6.83% of the country's GDP. The regional average GDP share is 10.2%. However, Dublin region alone accounts for 38% of Ireland's GDP. Even though it has suffered in the economic downturn since 2008, the region's economic recovery has begun. Despite the region having the highest unemployment rate in Ireland, it has decreased from 19.2% in 2011 to 9.4% in 2016. However, 78% of employment in the region is based on tertiary sector within which 'wholesale and retail trade; transport; accommodation and food services activities' employs 36% of the total tertiary sector in the region[2]. Nonetheless, Information and Communication sector only employs 5% of the region's workforce. While 2 regions do not have any employment in the sector, 9.33% of Dublin regional workforce is employed in Information and Communication Sector:

Border	0%
Dublin	9.33%

2 dated 22/03/2018 extracted on 08/07/18

Mid-East	4.31%
Mid-West	2.35%
Midland	0%
South-East	3.44%
South-West	3.62%
West	4.49%

However, the South East region has a growing base of companies involved in a variety of life sciences related activity, principally but not exclusively, manufacturing (Abbott Ireland Vascular Division, Bausch & Lomb, EirGen Boston Scientific, Sanofi Genzyme, Nypro, GlaxoSmithKline, Lake Region Medical, Merck Sharp & Dohme, Pinewood Laboratories and Waters Technologies). The overseas industries employed 12,445 in over 70 firms in 2014 and concentrated mainly in electronics and precision engineering, pharmaceuticals and healthcare and in internationally traded services (SEAPJ, 2015). Accordingly, pharmaceutical and healthcare sectors are the relevant sectors in the regions for the development of HPC. As regards infrastructure, the region has one supercomputer 'Kay' which is based in TSSG data centre which gives the region an added advantage over other regions that do not have access to supercomputer in their regions.

2.4 European Success stories

In this chapter you will see some examples of benefits of SMEs using HPC in their business. The success stories were presented for Fortissimo marketplace, which offers novel solutions to companies' challenges, enables them to discover new business opportunities and brings together all the necessary actors to construct a solution that matches their business requirements.

The Fortissimo Marketplace enables companies to:

- Greatly reduce time-to-results using pay-per-use SaaS.
- Develop complete and innovative solutions in collaboration with an easily available team with diverse expertise.
- Analyse large data-sets better, pushing any limits far beyond current capabilities.

- Simulate in greater detail and get more reliable results.
- Be ambitious by exploring new ideas.

All Fortissimo success stories are presented in Fortissimo Success Stories Booklet: (https://www.fortissimo-project.eu/news/final-edition-fortissimo-success-stories-booklet)

2.4.1 Cloud-based optimization of water turbines for power generation

Organization

Zeco is an Italian SME in the renewable energy sector. It specializes in the production of different types of water turbine. SMEs like Zeco must develop and innovate their products to remain competitive. High-fidelity simulation using CFD (Computational Fluid Dynamics) has become an essential tool for turbine designers because it results in better designs for less effort and lower cost. However, for Zeco and, in general for SMEs, full exploitation of CFD tools is often not possible as they lack the necessary computing power, and the skills to exploit it effectively. The objective here is to demonstrate how all the necessary resources can be assembled to give ZECO a one-stop-shop for the simulation of turbines leading to business benefits across the whole value chain.

The Challenge

Current practice in the design of hydro-power plants is to determine the most suitable design in a series of time-consuming experiments. However, SMEs in this sector face stiff completion and tight deadlines to sell their turbines in both national and global markets. The challenge facing Zeco was to improve its design processes by the use of HPC-based high fidelity simulations of flow in its turbines through the use of CFD-based tools

The Solution

A CFD-based HPC application has been developed which enables the design of a small hydro power plant in a very fast and reliable way, compared to current practices. The use of this application can contribute significantly to savings in time and money in the development of new water-turbine systems. High-fidelity simulations and the availability of HPC significantly reduce the development costs of prototypes, so the time to market is also significantly lower

Business impact

HPC-based CFD calculations have reduced the design time of a turbine from 1 year to 3 months. As manufacturing the turbine takes 8 months, the time to market can be reduced from 20 to 11 months. Without the use of HPC, the development process could take up to two years, which is no longer a competitive time frame in this sector. Using such turbines, a medium-sized hydropower plant costing €1.5 million can reduce operational costs by €350,000 per installation over two years. Furthermore, the optimization through HPC leads to a 1% increase in plant efficiency, with a 50% reduction in the total number of days required for maintenance. This means an increase of the revenue related to energy production of up to €40,000 per year per installation. Due to these improvements, Zeco expects to increase its market share by at least 5% with an additional profit of €50,000 per year. As a result of the increased market for advanced simulation using HPC, EnginSoft expects a growth of 10% in business related to the turbomachinery market sector, which means an additional profit of around €50,000 per year. CINECA estimates potential revenues for the HPC service of €100,000 per year. The workflow developed here is applicable to other sectors as well, so there is a large potential market. CINECA's target is to acquire two customers for this service per year for the next three years, with an estimated increased revenue of €900,000 and a profit of around €100,000.

Benefits

- Reduction of design time from 1 year to 3 months.
- Reduce operational costs by €350,000 per installation over 2 years.
- 1% increase in plant efficiency, 50% reduction in total number per days of maintenance.
- Increase by at least 5% in market share

Video Presentation:

https://www.youtube.com/watch?v=VS4OVduxRzY

2.4.2 HPC Cloud-based simulation of flange tightening

Organization

Texas Controls is a Spanish SME offering tightening and sealing solutions to large industrial facilities in a number of sectors. These are especially important to customers in the oil and gas industry, where it is imperative to avoid leaks in pipes, pressure vessels, or reactors that are under extreme pressures and temperatures. In these situations, preventing leaks is much less costly than dealing with the consequences of leaks. Therefore it is crucial to be able to predict the behavior of flanged joints, and to understand elastic interactions between them. Texas Controls has used computer simulations previously, however, the computational demands of modelling the behavior of flanges were high, and so an HPC-based solution was necessary. This was the first time they had used such technology.

The Challenge

To seal a joint properly, a gasket is placed inside a groove located on both flanges. The challenge here was to simulate and optimise tightening of the flanges - during the closing and tightening process, the gasket and the flanges may be damaged through deformations and high levels of stress. If the joint is damaged, it could fail, which could seriously harm workers, the surrounding community, and the environment.

The Solution

A computer model was developed which represented all the functional parts of a flange, including the gasket and the tightening bolts. This model was driven by a user interface, which enabled different tightening scenarios to be evaluated. Using the model, Texas Controls could accurately simulate and improve the design of the tightening process. The model was implemented using both open-source and proprietary simulation codes. Several sizes of case studies were run.

Business impact

Based on previous experience in the field (when no simulations were carried out), a non-optimised tightening of a 24 stud bolt flange took 108 man-hours; using simulation, Texas Controls reduced the process to 72 man-hours - a 33% time saving per flange. Whilst this represents considerable savings in labor costs, the most important outcome is the reduction in downtime of industrial installations such as refineries. The cost of “down time” for a medium-sized hydrocracker is about €21,000 per hour (€500,000 per day). Using advanced simulation, flange tightening can be reduced from 27 hours to 18 hours. This means a saving to the end-user of ~€180,000, because the shutdown path is shortened by the same amount that the tightening process is optimised. HPC-based simulation also allows technicians to avoid damage to flanges during the tightening, which is not possible using the usual experience-based method. This can also have significant cost implications. These benefits give Texas Controls a significant competitive advantage in a highly technical industrial sector, which should result in winning major, international commissioning contracts. This is expected to result in an increase in revenue of €2 million over the next 3 years, and a related 15% increase in staff employed. Texas Controls has a range of other services where the use of cloud-based HPC tools have the potential to similarly enhance Texas Controls' competitive advantage.

Benefits

- 33% time saving per flange for Texas Controls.
- Optimisation of flange tightening process.
- Increased revenue of €2 million over the next 3 years.

Video Presentation:

https://www.youtube.com/watch?v=a8pwi-8PLg0

2.4.3 Cloud-based HPC optimization of manufacturing processes

Organization

EMO is a Slovenian SME specialising in the production of tools and dies for stamping, particularly sheet metal, in the automobile and aerospace industries. EMO utilizes laser metal deposition (LMD) technology. AIMEN is a Spanish not-for-profit organization with expertise in laser technologies, manufacturing processes and the development of monitoring systems for industrial applications.

The Challenge

LMD is an additive manufacturing technology that enables the generic 3D printing of large metal parts. Additive manufacturing is a rapidly growing sector, as it allows for complex components to be produced with short lead times. However, the lack of sufficient control remains a barrier, as it can result in unnecessary reworking, waste and an increase in 3D printing times. This reduces both profits and efficiency. This experiment aimed to overcome the current deficiencies in online monitoring and control of laser processing, so that EMO can realize the full benefits of additive manufacturing and create better quality products

The Solution

EMO currently uses LMD. As part of this process, EMO needs to gather and analyse significant amounts of data. A more efficient workflow would enable it to complete more projects without extra investment. A new system has been developed in this experiment that exploits recent advances in AI for image analysis and in data acquisition from images of the process. The new technology, called CyPLAM, is a novel approach to the online monitoring of LMD. It uses deep-learning principles, working on the Fortissimo infrastructure, to enable online and real-time quality control and monitoring of key features such as dilution and clad height. CyPLAM has been validated by testing on a repair application using LMD.

Business impact

CyPLAM addresses the quality issues of LMD-created parts, putting EMO at an advantage as they are first to use the technology. Future CyPLAM users can expect to cut production times by over 30%, as well as producing a higher quality product needing less reworking. Overall, end users can expect a 20% saving in operational costs and a 30% reduction in lead-time, compared to traditional approaches. AIMEN will use CyPLAM technology to support its recently launched CLAMIR system, a commercial process control system for Laser Additive Manufacturing. The experience and knowledge acquired during this experiment have allowed CESGA to obtain new projects and contracts within the industry. They have also created a training course on Machine Learning. 3D printing and other additive manufacturing technologies have had a major impact on the European manufacturing industry, allowing fast and flexible prototyping and part creation. This industry is home to many SMEs due to the comparatively low cost of entry and is growing fast. The worldwide 3D manufacturing industry is growing at a rate of 25% per year and is expected to be worth 6.5 billion USD (€5.92 billion) in 2019.

Benefits

- Using HPC, EMO can reduce operational costs by 20% and save over 2.000 machine hours per year.
- Users of CyPLAM can stay competitive in the global Additive Manufacturing market.
- As a result of the expertise gained in this experiment CESGA is offering a new Machine learning service (SaaS) based on TensorFlow.

Video Presentation:

https://www.youtube.com/watch?v=so-onjsZPZY

2.4.4 Cloud-based optimization of a multi-body wave energy device

Organization

Zyba is a UK registered SME with expertise in digital and numerical modelling. Since 2014, Zyba's primary goal is to engineer simple systems that work with nature. Together with BioRock Technology, Zyba is harnessing wave energy to protect coasts from erosion, enhance coral reefs and deliver sustainable power. HPC resources and expertise in this experiment were provided by Arctur.

The Challenge

Zyba's core product, CCell, is a curved wave energy converter designed to be efficient and simple to assemble and operate. It moves with the waves to simultaneously extract their energy for electricity and reduce their impact on the beach. BioRock reefs use the low-voltage charge from CCell to form limestone rock from seawater minerals that serve as habitat for corals to grow at an accelerated rate, creating an active breakwater to protect shores from erosion. Optimisation of the CCell for each site is heavily dependent on the local environment and is a computationally intensive task, which is too expensive for most SMEs' budgets as is physical testing.

The Solution

Simulations of the CCell device in different incoming wave conditions were carried out using the OpenFOAM software package on an HPC machine. In addition, an easy-to-use GUI (Graphic User Interface) was developed, which allows simulations to be set up quickly and to show a series of scripts and tools written to streamline of the workflow on the HPC system. The increased computation power allowed the whole CCell system to be modelled for the first time, including power-takeoff hardware, software, and the intended control logic. This information provided a completely new insight into how the different pieces of the system work together.

Business impact

As a result of the experiment, Zyba has been able to increase its productivity and deploy a series of pilots offshore. The automated and streamlined design process enabled it to rapidly optimise each design as well as reduce design costs and development time. With the new GUI, less experienced users can now complete design tasks on their own, which empowers individuals within the team and increases the rate of innovation. Overall the new system has enabled Zyba to significantly speed up its market entry with saving around 9-12 months and will continue to facilitate the evolution of future products and service offerings. Provision of resources from an HPC centre transforms an imposing capital expense to a more manageable operating cost. Administrative tasks associated with IT management and upgrades are also effectively outsourced, alleviating pressure on a small SME team and allowing them to focus on their design work.

Benefits

- Reduced simulation set up time from 2 hours to less than 1 minute.
- Mitigated use of physical modelling, providing a nine-fold reduction in cost.
- Reduced HPC costs from €0.09/core-hour to €0.05/core-hour as all software used was open source.
- Ability to concurrently run simulations improved productivity by a factor of seven.

Video Presentation:

https://www.youtube.com/watch?v=CI-AG6ts_gc

2.4.5 Cloud-based multiphysics simulation for MEMS micro speaker design and development

Organization

USOUND GMBH based in Graz, Austria, is a fast-growing audio company, developing and producing the most advanced audio systems for personal applications based on MEMS (Micro Electro Mechanical System) technology. HLRS is the High-Performance Computing Center Stuttgart of the University of Stuttgart in Germany.

The Challenge

USound GmbH develops unique MEMS miniature speakers based on the piezoelectric effect. Such speakers have all the advantages of MEMS technology; automated production and assembly, a smaller form factor, negligible heat generation and improved linearity compared to the classical loudspeaker. A MEMS production run is expensive and time consuming. Very powerful multi-physics simulation tools are needed as an alternative to physical prototypes to predict the acoustic performance of the speaker and to reduce the number of production runs needed.

The Solution

The USound in-house multi-physics simulation, running on the commercial tool COMSOL Multiphysics, has been used to predict the Sound Pressure Level produced. This approach had restrictions and simplifications due to limited computational power available. The calculation time for this model is 10 to 20 hours, which slows down the development process and does not allow fast parameter variation. HPC offers the possibility to use a more detailed geometry without compromises, analyse and consider acoustical losses with thermosviscous methods and speed the development process. The COMSOL Multiphysics tool was adapted to run on the HLRS cluster system directly over the native user interface.

Business impact

USound is pioneering MEMS speakers. In the area of microphones, MEMS have replaced traditional microphones almost completely by now. USound is expecting the same to happen to micro-speakers. To address the mobile communications market, further technology and product development is needed. The development for the next-generation mobile speaker product will be 50% faster using the cluster computing methods developed in this experiment.

It will therefore allow USound GmbH to gain market share and react to customer needs in a timely manner. In the growing sector of MEMS devices, powerful multi-physics simulations are mandatory because experimental evaluation is very expensive. As devices become more complex, all sorts of mechanical, electrical and thermal influences have to be considered. These lead to very demanding simulation models. MEMS developers are limited by their conventional workstations. The experiment shows that for a demanding FEM (Finite Element Method) MEMS model, COMSOL cluster computing offers a step forward in increasing the complexity and reducing computation times of the FEM model.

Benefits

- Ability to use highly detailed models in fast simulations.
- Using exact models, €400,000 can be saved by reducing the number of prototyping runs from 4 to 2 (one run is in the order of €200,000).
- Faster time to market by up to 1.5 years.

Video Presentation:

https://www.youtube.com/watch?v=dApYR1ET8YU

2.4.6 Optimization of steel structure manufacturing

Motivation

The current process of designing a new steel fabrication plant encompasses several steps and involves personnel with different competencies. At the beginning, the commercial crew interacts with the customer, in order to gather the plant requirements. Subsequently, the technical team prepares a first draft plant layout, based both on the customer requirements and on the previous experiences related to very well-known plant layout templates. Based on a typical production mix, a production optimisation is run in order to verify the expected plant performance.

Since the needed simulation and optimisation tools require powerful hardware, this task has to be done by the technical office and cannot be executed at the customer premises. Therefore, a video of the simulation run is recorded by the technical office and sent to the customer. By watching the video, the customer develops a better understanding of the process and, typically, he wants to apply several changes to the layout or to the machines used. Those changes are sent back to the technical team, where a new simulation and optimisation task is executed and a new video is generated. This loop is usually repeated until the required level of maturity of the solution is achieved. Thus, the existing process is time consuming and inefficient because of many iterations, where only the technical team can run simulations and assess the plant productivity, using dedicated workstations. Each production optimisation of a typical plant of medium complexity (composed of 4 machining stations, 2 loading bays, 2 unloading bays and the automatic handling system) requires approximately 8 minutes on a high-end, 8 cores desktop PC, while it requires 30 minutes on a normal laptop. Clearly, a 30 minutes window for each optimisation is prohibitive in a negotiation with the customer. Each month, at least 10 requests for early design modifications and simulation are sent to the technical office, to start and carry on the negotiation phase and the mentioned iterations (within the average of 20 new negotiations per year). The experiment is meant to optimise this process and to enable quicker and faster simulation and optimisation even at the customer’s site. This vision requires the implementation of two cloud-based services to simulate and optimise the production of a complex manufacturing system, composed of several machines and conveyors. The two services are coupled with a client application meant to streamline the access and the steps required to successfully simulate and optimise a production plant (i.e. upload of the simulation model, customisation of the layout, selection of the production mix and visualisation of the results). The technical objective of this experiment is thus to provide the functionalities of the simulation and optimisation tools as cloud-based HPC services, in order to achieve the main business goal to empower a wider range of user (i.e. the commercial crew) with quicker simulation and optimisation solutions to be deployed at the customer's premises.

Technical impact

The implementation of the experiment was successful in reducing the time needed to perform an optimisation for a layout of medium complexity from 30 minutes to approximately 3 minutes on portable devices, blowing away the barrier that made impractical the use of such tools during the negotiation phase, at the customer's premises. Modifications can now be applied showing directly the effects of the changes, streamlining the interaction towards the best configuration. With the achieved implementation of the experiment, several direct economic benefits are expected over a short to mid-term period.

Economic impact

FICEP benefits from a more efficient proposal phase due to the streamlined interaction between the technical team and the commercial crew, now endowed by quick-everywhere simulation and optimisation capabilities (for the average plant afore mentioned, the number of iterations is reduced from a minimum of 6 – where each iteration takes 4 man-days – to 2, quantifiable in € 4.800 savings, not taking into account the improved quality of the service offered). Taking into account the number of negotiations processes initiated per year, which were estimated before in 20 negotiations, this leads to an estimation of €96,000 /year savings. Furthermore, collaboration between different FICEP teams located worldwide is boosted, as the results of different layout simulation are stored in the cloud, further increasing the capability to properly address the customer's needs. The cloud-based configuration also allowed TTS to develop a new business (and pricing) model: a monthly €100 fee in a pay-per-use model allows TTS to reach a wider number of SMEs having a limited expenditure capacity but a strong necessity of simulation functionalities especially during the machine design phase. These companies would benefit from a usage of the platform purchased as a service on demand. Such SMEs usually operate in niche markets providing specialty high-performing machines in small (also one-of) lots. This will result in an increased number of active customers for TTS, with more than 20 additional machine manufacturers using TTS cloud-based solutions, resulting in € 80,000 of additional sales over a 3 year’s time horizon starting from the project conclusion, with the creation of 2 new jobs over the same time period

2.4.7 Improving fire safety of buildings by simulation in the cloud

Motivation

Fire safety of buildings is of relevance for each European citizen. Who does not want to feel safe when shopping in a mall or visiting a theatre? In case of an emergency event, people have to get out of the building in a fast, streamlined but not chaotic manner, sprinklers have to be placed in the planning and construction phase so to keep fire under control or even extinct it, fire fighters have to have sufficient access routes and water supply, etc. This all should be simulated and optimized before starting the building process. The market size in Spain (Europe) for fire simulation tools and services is estimated to amount to 1.5 million (€ 20 million). The goal of this experiment is to improve fire safety designs in the building sector using CFD (Computational Fluid Dynamics). Specifically, this means to integrate a CFD tool called CYPE-FDS into the CloudFlow platform for detailed fire simulation scenarios focused on the building design industry. As a show case, a real shopping center located in Spain will be used for a complete fire safety design process. The expected technical impact is an improvement of the complete workflow of the fire safety design stage, a reduction in time for the model preparation and results analysis, and higher accuracy in the prediction compared to the traditional process through leveraging the open source solver FDS (Fire Dynamics Simulator). Economically this will show the following effects: a) an increased number of fire-safety design projects, b) reduced costs of the fire-safety facilities due to optimization (30 percent cost reduction as an average, approximately € 80,000 in a case similar to the one solved in this experiment), and c) reduced costs for hardware and software for users applying fire simulation due to a pay-per-use concept for HPC resources and open source software. Currently, fire protection analysis and design is based on scalar and prescriptive models (simple and fast approach) according to codes like CTE DB SI 6 code, Eurocode (EN 1992-1-2:2004 and EN 1993-1-2:2005). But these models do not supply detailed information about the movement of smoke and the temperature evolution. Although the usage of CFD tools would be possible, the technology is not commonly used in industry because of the hardware and software costs, the limited number of CFD specialists and the time constraints for defining the fire safety design. In the reference case, this represents approximately 30 percent of the total time for the complete building design (structural and facilities). The complete fire safety design using CFD tools must be solved in no more than three weeks in order to be competitive with prescriptive models. The challenges lie in providing a user-friendly tool with fast response times using HPC resources to facilitate fire simulations for AEC engineers that are not experts in fire simulation software. The approach is to cloudify the open source solver FDS to obtain the necessary accuracy for the fire protection analysis and design. CYPE-FDS, the pre-post tool, which facilitates the case definition, will be available in the workflow through a virtual machine. Several fire scenarios could be executed in parallel thanks to HPC resources inside CloudFlow. Finally, the results obtained in the case experiment will be summarized in a Best Practice Guideline for FDS use in building industry. This guide could be used to establish strategies for estimating a good balance between the cell size, the solving time, and the accuracy of results; HPC efficiency (number of cores, solving time reduction, simulation cost ratios); and numerical considerations in order to define subdomains.

Technical impact

CYPE, the ISV partner, will be able to offer new cloudified solutions for their software to new and existing clients. Furthermore, improvements in the CYPE FDS workflow are realized due to the use of HPC resources. Adopting the same procedure for other software, e.g. structure analysis, energy efficiency and construction management, CYPE will be able to increase the cloud based product offerings allowing not only for solving more complex problems in the domain of fire simulation but also other domains. Cottés and Itecam, both end-users in the experiment, profit from the user interface improvements, especially for pre- and post-processing the model. The training time for the CFD tool can be reduced significantly by providing application-specific tailored functionality. Additionally, access to HPC resources is streamlined by the workflow and does not require special know-how on the user side. Concerning the fire safety facilities design the end users benefits from the increasing number of fire scenarios that could be simulated. Finding the optimum design speed is increased by up to 30-40 percent due to the usage of the cloud solution because it is accessible anytime, has no idle time and runs with good stability and low risk in simulation interruption.

Economic impact

CYPE will address a bigger target market with the expectation of 2,000 clients within one year after the experiment completion. Furthermore it is likely to start 200 projects with existing clients, one year after the experiment completion, with a perspective of 400 projects in the third year. Concerning new clients, CYPE expects 100 projects in the first year and 300 projects in the third year, respectively. The financial benefit with a total income for the first year is estimated by € 57,000 and by € 192,000 in the third year in software license sales. CYPE is planning to employ two new software developers. At the moment only 20 percent of the Fire Safety Design Projects from the end user Cottés require the usage of CFD tools. This corresponds to € 350,000 - € 400,000 per year given a total revenue of € 3.5 -4.0 Mio. Using the cloudified solution of the experiment, Cottés expects to double the number or projects resulting in a budget of approx. € 750,000 per year. Cottés is also expecting to reduce the costs of the active and passive mechanisms and gaining more probability of winning the tendering process. Today, Cottés trades only in the Spanish market, developing 20 fire safety design projects in 2016. Gaining competitiveness through the cloudified version, the number of projects could be doubled. Furthermore, Cottés customers would benefit from obtaining optimized solutions in shortened time period, with a 30-40 percent cost reduction, as an average € 80,000 per project. Another benefit is the potential access to new markets like major civil engineering construction projects (airports, tunnels) and the forensic engineering sector. Another sector could be opened if insurance companies will ask for a CFD solution in order to know the origin of a fire. If 0.1 percent of these fire scenarios required a detailed analysis through CFD simulations, Cottés would increase its number of projects in a range from 15 to 25. After the CloudFlow Project, Cottés is planning to employ two more CFD specialists in fire simulation and one engineer expert in Fire Safety Codes. In 2016 ITECAM has developed four CFD projects for private companies using commercial software such as ANSYS-CFX, Solidworks Flow, etc. focusing on CHT analysis, pressure drop simulations, and coupled multi-physics problems. Thanks to the “Fire in the Cloud” application experiment, ITECAM expects to expand its CFD simulation capabilities to other areas, such as: HVAC analysis, FSI simulation under fire scenarios, human evacuation analysis, etc. At the end of the third year, ITECAM is planning to consolidate a specific CFD department with 2-3 engineers, dedicating 50 percent of its activity to FDS simulations. In 2017, ITECAM expects to achieve revenues close to € 50,000 through the use of “Fire in the Cloud”, collaborating with Cottés and CYPE in consulting activities, benchmarking, and training some of our associated companies. In 2020 the CFD department is expected to achieve revenues close to € 250,000, approximately 45 percent of this amount would be generated through FDS activities. Arctur being an infrastructure provider, expects to have mostly an economic impact form this experiment. This is reflected through the increased sale of Cloud and HPC resources. The increase of the sales follows the increase of new users using the services at offer.

2.5 Examples of HPC usage in South East Ireland business sector

In Ireland, ICHEC leverage their knowledge in a multi-disciplinary approach to technical computing, data management and software engineering. This has result in thriving programmes of public sector and industrial engagement, with particular expertise and success in the energy sector. ICHEC has also worked with a number of leading national energy companies on improving the accuracy of wind and solar weather forecasts. Additionally, ICHEC provides regular training to independently manage and run the forecasts on the ICHEC systems with an overall objective for the client to develop the skills required. Other sectors also include engineering and automation, pharmaceutical and healthcare, animation, financial and precision agriculture sector.

2.6 Applications of HPC in the industry for R&D in South East Ireland

In Ireland, academic and industrial R&D collaboration supported more than 1,400 researchers across sectors including high-resolution weather forecasting, bioinformatics, oil & gas exploration and remote observation for precision agriculture and land planning. Additionally, in collaboration with Intel engineers, ICHEC are investigating advanced parallel programming methods, models and tools to modernise and re-factor complex and mission-critical software applications across a wide range of domains, including materials science, weather forecasting, ocean-wave modelling, seismic imaging and cryptanalysis, and the core aim is to significantly boost software performance at lower energy costs. There is also a development towards a point-of-care technology for heart disease that monitors the behaviour of small particles in the blood called platelets through which the patient provides a small amount of blood, which is then videoed as it flows on a vascular protein-coated surface within a microfluidic chip.

2.7 Cooperation between academia and industry in South East Ireland

While cooperation between industry and academia exist, there are no reports or data for the South East region as regards HPC. As nationwide, there exist HPC cooperation between academia and industries in different sectors. Along with Tullow Oil plc, ICHEC and DND are researching and developing leading-edge data management software for handling large seismic datasets. This initiative was funded by Science Foundation Ireland for three years and is expected to make an impact in the energy industry through more cost-efficient and time-effective exploration. Additionally, ICHEC is also working with Irish technology company Skytek to provide an archive for the European Space Agency for Sentinel and other data which will provide archived products, real-time data and on-demand processing for public sector and commercial users in Ireland. Along with Intel Parallel Computing Centres (IPCCs) by Intel, ICHEC has positioned itself at the leading edge of many-core R&D, HPC software design and end-user exploitation. None of these projects are related to South East Ireland.

Furthermore, The Irish Centre for High-End Computing (ICHEC) has launched a European network called SESAME (Supercomputing Expertise for Small and Medium Enterprises) which aims to give Irish SMEs affordable access to HPC technology.

2.8 Other pilot region-specific relevant aspects of the HPC development

While HPC activities in Ireland exist, there are no reports or data for the South East region as regards HPC. As nationwide, an initiative was funded by Science Foundation Ireland to make an impact in the energy industry through more cost-efficient and time-effective exploration. Additionally, the Irish Centre for High-End Computing (ICHEC) is also working with Irish technology company Skytek to provide an archive for the European Space Agency for Sentinel and other data which will provide archived products, real-time data and on-demand processing for public sector and commercial users in Ireland. However, South East region specifically does not have development of HPC. Even though cooperation between academia and industries in different sectors exists in the region, HPC is still underdeveloped with lack of awareness especially among the SMEs.

2.9 Roadmap for improving SME uptake of HPC

In this chapter you will get information about SESAME Net programme, which aims to create an open and inclusive European network of Competence Centers and Organizations joining forces in order to raise SMEs' awareness on HPC and to demonstrate its features and benefits.

The network on the one hand enables its members to share experiences, exchange best practices, learn from each other, identify similarities and differences in the circumstances faced in each of the regions, to gain from synergy effects and on the other hand builds a significant means to strengthen the European middle class. By pooling expertise ranging from classical simulation through high performance data analytics to machine learning, offering consulting services and providing interested SMEs to acquire and build internal expertise SESAME Net aims to become an entry point to HPC for SMEs even for SMEs from countries that do not currently have such centers. More information of this project is available on https://sesamenet.eu.

2.9.1 Introduction

The value of computing in science and engineering is well established, and the potential of computational technologies is uncontroversial. Leading companies, especially in automotive and aerospace, have been some of the early adopters of High Performance Computing (HPC) in the last century. Computational structure modelling and computational fluid dynamics (CFD) have become integral to product design in the industry for decades. It is worth pointing out that this adoption of HPC was also supported by European projects, especially at the end of the nineties (e.g. EUROPORT-1, EUROPORT2 and EUROPORT-D2 and other FP4 projects).

In addition to the services of HPC centres, small and large commercial players in cloud computing offer HPC Cloud Resources where the best known worldwide are Amazon with its Amazon Web Services (AWS) offering, such as Elastic Compute Cloud (EC2) and Microsoft with Azure. The latter supports, via its HPC software pack, seamless access from the desktop or a cluster on premises to the Azure HPC environment. Today, the marketing points to readily available large-scale cloud computing on demand for Big Compute and Big Data, in other words, tackling compute-intensive and data intensive problems that have been a mainstay for the HPC community.

There is also a growing number of companies and start-ups offering services on top of worldwide Cloud HPC infrastructures to orchestrate HPC solutions on premises with regional and/or worldwide Cloud offerings e.g. Cyclecomputing.com. Other companies like Rescale.com provide solutions based on new business models for ISVs, software providers and cloud hardware providers around the globe. In addition, technological advances mean that a modest server or cluster with the equivalent performance of the best supercomputer in 2000 (https://www.top500.org/list/2000/11/) is now available for less than €100,000. A medium-sized HPC cluster is no longer a large capital investment for enterprises who have the choice to deploy compute intensive workloads internally or externally in a Cloud environment. As a result, we might assume HPC is ubiquitous.

Nevertheless, it is well known that the adoption of HPC among SMEs (small and medium enterprises with fewer than 250 employees) is not as high as among large enterprises. Public HPC centers (As SESAME Net, Fortissimo) can play a significant role in the SME uptake of HPC. They are often geographically close to the SME, accessible and speak the native language of the individual SMEs. Barriers to HPC adoption arising from a lack of trust can often be easily overcome because the HPC centres do not act as competitors in the market and are public institutions. They can provide staff skilled in HPC technology and applications often supported by university departments. They know about and may have in place cutting edge HPC technologies, are able to provide assistance to design new and innovative products, resolve real-world problems relating to performance and code optimisation, and play an instrumental role in up-skilling the SME, hence an effective Technology Transfer programme. HPC centres may offer their own compute resources but can also offer consultancy to use commercial Cloud HPC. SMEs in engineering for example are often running applications on their own servers, but even when the application theoretically scales to millions of processors, they are not aware of how many processors they can utilize effectively with their chosen modelling parameters. Therefore, public HPC centers represent an attractive intermediate step before SMEs gain access to HPC Cloud resources.

In the following chapters, we will introduce ideas to foster the adoption of HPC by SMEs.

2.9.2 Are SMEs ready to use HPC services?

Digital Transformation Scoreboard 2017 (http://ec.europa.eu/DocsRoom/documents/21501) offers an overall image on the digital technology landscape in Europe. It illustrates, from the point of view of digitalization, that there is significant variability between geographical regions in Europe. This is also true from the point of view of SMEs' readiness to use HPC. Beyond this heterogeneity, there are several aspects common to all countries. At a first level of analysis of the feedback collected from SME/HPC partners, SMEs that have a potential for HPC usage can be grouped in several categories:

1. Aware of the benefits of HPC technology and are already using it.
2. Aware of the existence of HPC technology but do not use it. The main identified causes are that the SME:
    - Is not convinced about the benefits of HPC
    - Considers that it cannot afford the cost of using HPC
    - Does not have skills/knowledge to embrace HPC
    - Does not trust the HPC provider with respect to the reliability of resources, security and confidentiality of sensitive data
3. Not aware of the benefits of HPC technology.

Currently the amount of SMEs in the last two categories is significantly larger than those in the first one (the SESAME Net estimation is that overall less than 10% of SMEs are in the first category) which means that most of the SMEs are not ready, at least not without targeted support, to efficiently uptake HPC technology. Of course, some spin-offs or engineering companies are ready to use HPC when they have extra workloads or peak needs. There are success stories on SMEs across various domains such as automotive and aerospace design, chemistry and materials synthesis, planning in civil engineering and energy markets, environment monitoring, modelling for health & pharma, design of medical devices etc. SMEs operating in the field of design and manufacturing are more inclined and interested in developing innovative products and undertake R&D projects. However, it is expected that the HPC adoption will be much more challenging for an SME from sectors where HPC has not gained mainstream popularity, or where prior HPC use has been mostly confined to academic research efforts. Furthermore, the size of the SME might be an important factor, e.g. where comparatively large work force is required for the implementation of a HPC-based solution without being directly business-critical. There are success stories involving SMEs with a small number of employees (ten or more) as well as larger size SMEs (up to two hundred). Spinoffs, start-ups and engineering companies are in general more open to explore features and benefits of HPC in all European countries.

The main questions to be asked when assessing the readiness of SMEs to use HPC are:

- Do they know? This is related to ***knowledge and awareness.***
- Do they need? This is related to the type of activity and domain challenges.
- Do they want? This is mainly related to trust and business strategy.
- Do they have resources? This is related to the availability of skilled staff and certain technical prerequisites (e.g. connectivity).

In the following section, some of these aspects are outlined in detail based on the information gathered by SESAME Net and related initiatives.

2.9.1.1 Knowledge and awareness

"97% of the companies that employ HPC consider it indispensable for their ability to innovate, compete, and survive" (SESAME Net, 2017. Moreover, the report emphasizes that the returns on investment in HPC are extremely high, and many companies that invest in HPC become economically successful. However, many SMEs are not aware of these benefits or they perceive HPC as a technology, which is beneficial only for large companies. In some countries, e.g. Germany, almost every major university has an HPC system, suggesting that the percentage of companies being aware of HPC should not be that low. In addition, many technically interested SMEs in these regions seem to be informed about new compute technologies used in HPC. However, there are SMEs that cannot identify the benefits of HPC or are not even aware of the technology. Unfortunately, such SMEs are difficult to reach. The preliminary SESAME Net questionnaire indicates that around 40% of SMEs either do not know the possible advantages or feel that they do not need HPC. Raising awareness is one of the actions which should be continued by HPC centres in collaboration with other stakeholders (e.g. industry associations), considering that the level of awareness has not reached a critical mass for making the "democratization of HPC for SMEs" a reality yet. The actions should focus on industry sectors, which can demonstrate the benefits of HPC in a relatively short time.

2.9.1.2 Price

The availability of workforce with an adequate educational background or training in using HPC remains scarce. One of the reasons for this situation is the limited interaction in some countries between academia and industry on the exploitation of high-end computing systems. Moreover, many SMEs seem to lack a long-term training strategy that would allow staff to be trained in new technologies not currently perceived to be critical for the company. Reassigning staff from existing projects in order to prepare them for envisaged future projects, using HPC technologies, might not be feasible for small companies. According to the preliminary assessment of SME needs conducted by SESAME Net, more than 60% of SMEs who are aware of HPC consider that they would need help in modelling and/or HPC consultancy and would be interested in participating at training activities focused on applications of HPC technologies, particularly at those targeting their specific needs. In addition, more than 80% of SMEs are not aware of existing training activities in their regions. Concerning the perceived lack of skills there is an almost homogeneous distribution over various HPC topics such as applications usage, code development/optimization, infrastructure administration and introduction to HPC. Providing targeted training materials, particularly "how to" nuggets of knowledge, tutorials, online video etc. are mandatory in order to bridge the gap between the current and the required skills of SMEs.

2.9.1.3 Affordability

Many SMEs perceive HPC technologies as expensive and affordable only by big companies because they perceive that they cannot afford the specific software and hardware and that direct and indirect cost are too high. One of the reasons of this perception is that they do not understand the increase in value they can get with HPC and consequently are not willing to invest in HPC without having clear benefits for day-to-day business operations. However, cloud-based HPC services offered on a pay-per-use basis makes access to powerful computing resources, often combined with appropriate software solutions, financially viable for an SME thereby reducing software, hardware and maintenance costs. Access to HPC services can be facilitated by ensuring the flexibility and affordability of software licensing models, by offering some free services (for example, consultancy, training, and best practice guides) and by the creation of targeted funding programmes.

2.9.1.4 Usability

Currently there is a gap between user demands and HPC solution offerings, particularly with respect to the ease of accessing and using HPC services. Flexibility of access to HPC resources is critical for HPC uptake. According to feedback received from the experienced partners, there seems to be a knowledge gap for the general usage and features of HPC, particularly from a technical point of view. For SMEs interested to explore the benefits, usability in terms of tailored HPC solutions is very important. The usability problems can be overcome by delivering tailored front-ends for HPC back end solutions. The offered solutions should be accompanied by targeted support (more experienced users may need less cosmetic add-ons to interact with the system or service and less support than less experienced ones).

2.9.1.5 Connectivity

Robust connectivity (e.g. reliable and high-speed Internet connection) is a prerequisite for cloud-based HPC services. As illustrated by the Digital Transformation Scoreboard (2017), broadband connectivity has become prevalent in most European countries, and improving the quality of access is one of the priorities of some national strategies for digitalization. Concerns about security have a significant influence on the decision of a company to use remote services such as cloud based HPC. Unlike connectivity, which is more related to infrastructure reliability, e-security involves trust in software as well as in staff. The preliminary study conducted by SESAME Net illustrates that currently around 15% of questioned SMEs use data storage in Cloud but more than 50% would be willing to move and share data related to their products, services and/or processes over the Internet as long as security is ensured and that the Internet service provider can be trusted. In order to address such SMEs concerns, HPC centres should demonstrate their security management capabilities. This can be done in different ways depending on the technical skills of the SME interlocutor. For instance, interlocutors with zero to basic IT knowledge could be presented with a number of case studies showing how specific HPC centers dealt with security issues in the past, i.e. its record of accomplishment or reputation in security. In addition, detailed security solutions and specifications should be made available to interlocutors that are more knowledgeable.

2.9.2 Are HPC Centres ready for SMEs?

As in the case of SMEs, there is heterogeneity amongst the HPC centres with respect to their readiness in working with SMEs. The readiness of a HPC centre can be assessed at several levels: technical, legal and business. At a as regards the technical level, almost all HPC centres are ready for SMEs as they usually have both the computing resources and expertise to service SMEs. However, they should be flexible enough for some restructuring to better address SMEs' needs. Engagements with industry typically require tight schedules and/or different time-to-response scales relative to research activities, the centres should be prepared for this. Predictive performance, i.e. foreseeable estimated times to solution and measures of system resilience are important indicators for industrial users. In order to meet such requirements, HPC centres should be well prepared with respect to compute resources, back-up procedures and provide consultancy. From a legal point of view, some HPC centres are non-commercial organizations, e.g. associated with universities, and are thus limited in how they may collaborate with SMEs, e.g. projects that receive public funding may require the results to be made public, in other words adherence to an open research policy; this might impede the involvement of SMEs. In some countries, there is a growing number of commercial HPC providers offering specialised services (e.g. in engineering or for rendering) on small and medium HPC clusters, which can compensate for this legal issue. In other countries, like Croatia and Romania, most of the HPC competence centers are non-commercial (government) organizations with restrictive business models that do not allow these centres to work with SMEs. Transnational cooperation with centres that can accommodate such contracts, e.g. through networks like SESAME Net, can possibly solve this issue for SMEs. From the business point of view, centers organised as commercial entities (like Arctur in Slovenia, Yotta in Croatia and IT4I in Czech Republic), have appropriate business models and their systems are prepared in order to facilitate collaboration with SMEs. However, providing HPC services to SMEs remains limited in terms of business activity. The main limit is imposed by the lack of skilled human resources. In most countries, HPC centers do not have enough resources to deal with a wide variety of SME requirements and to ensure adequate response times. Due to the research-oriented nature of the centres, their staff profile is predominantly technical and there is a general acknowledgement of the absence of staff dedicated to marketing or business development activities. Currently there seems to be low capacity in the centres to provide strong SLAs (Service Level Agreements) and strict service definitions, as this is usually not required when serving the education and research communities. There is also often a lack of formal service methodologies that might be demanded by enterprise customers (e.g. proven by ISO (International Standards Organisation) certifications). The number of HPC centres that are ready across all three perspectives in SESAME Net includes partners in Belgium, Germany, Ireland, Slovenia, Bulgaria, Czech Republic, Poland, Greece and Spain. These HPC centres are ready to work with SMEs as they have prepared business models, well-defined support for SMEs in terms of infrastructure, technologies and expertise and also offer training activities for young entrepreneurs. However, in some cases, the collaboration with SMEs relies on support from European and national funds because SMEs usually don't have much resources to research new technologies. In other cases, many of the centers attached to universities (as is mainly the case in countries like Croatia, Lithuania and Romania) have little or no prior experience in working with SMEs in a commercial context (four of the thirteen SESAME Net partners have no such experience) and lack the human resources for effective communication with companies.

2.10 Reflection on this module learning

In this module you get the information about HPC usability for SMEs, get experience of good practices of HPC usage and saw some success stories of HPC combined with SMEs.

2.11 Recommended Further Readings / Additional Resources

1. Best Practice guidelines for HPC in industry and in particular to SMEs, SESAME Net (https://sesamenet.eu/wp-content/uploads/2016/12/D3.1_fin.pdf):

   This document gives comprehensive information about best practices in HPC industrial use that were collected and documented within the project. This information is mainly addressed to HPC centres active on regional or national levels, working, or planning to work with SMEs. The deliverable concludes with a "best practice set of guidelines" and "contract templates" that cover all aspects of the process of providing HPC services for SMEs, so that they can be used comprehensively by HPC centers.

2. Fortissimo success stories – fortissimo's small & medium, manufacturing enterprises (https://www.fortissimo-project.eu/sites/default/files/Fortissimo_SS_Booklet_web_0.pdf):

   In this document, you will get information about the marketplace which simplifies access to HPC services for European companies enabling them to be more competitive. You will see what benefits HPC gives to SMEs and get some examples of success stories of SMEs using HPC.

3. Cloudflow success stories, computational cloud services and workflows for agile engineering

   (https://eu-cloudflow.eu/files/2017-0421_CloudFlow_Broschuere_EN_web.pdf):

   In this brochure you can see application experiments that have been an integral concept of the CloudFlow project. They are SME-driven use cases for the CloudFlow platform. CloudFlow is designed to execute application experiments in three waves, generating a total number of twenty experiments. They each have their own success story which are presented in this brochure.

4. Roadmap for improving SME uptake of HPC, SESAME Net:

   The role of this Work package is to support and coordinate various activities to raise awareness for the advantages of using High Performance Computing *among SMEs. It aims to achieve the following general objectives as described in the proposal:*

• *Raising awareness for the benefits of HPC usage by SMEs*

• *Development of a road-map to help increase the uptake of HPC among SMEs*

• *Collection and preparation of dissemination material and awareness raising events*

• *Coordination and organization of outreach events*

5. Hyperion Research - Research Highlights In HPC, HPDA-AI, Cloud Computing, Quantum Computing, The Global Exascale Race, and Innovation Award Winners (https://hyperionresearch.com/wp-content/uploads/2019/02/Hyperion-Research-SC18-Breakfast-Presentation.pdf):

   In this presentation, you will get information of HPC market, HPDA-AI and ML/DL update, Cloud and quantum computing update and information about global exascale race.

6. High performance computing for energy, HPCE (https://hpc4e.eu/sites/default/files/hpc4e_project_files/HPC4E_Whitepaper.pdf):

   In this document you will find information about the use of exascale computers in Oil & Gas, Wind Energy and Biogas Combustion industries.

7. The Exascale Effect: the Benefits of Supercomputing Investment for U.S. Industry (https://www.compete.org/storage/images/uploads/File/PDF%20Files/Solve_Report_Final.pdf):
   This document outlines research to assess how government investment in HPC benefits U.S. industrial competitiveness and what areas of continued investment would provide the greatest benefit moving forward. During a six-month research period, Intersect360 Research conducted 14 in-depth interviews with forward-thinking representatives of industrial HPC-leading organizations and then

gathered 101 responses to a comprehensive online survey of U.S.-based, HPC-using companies.

Co-funded by the
Erasmus+ Programme
of the European Union

MODULE 3
SME/HPC provider relationship management

Basic

Aims of this Module:

This module aims to describe / explain:

- Global market landscape for HPC.
- Who are the leading providers of HPC solutions.
- The administrative process of SME engagement with HPC providers.
- Be able to formalize engagement with HPC providers.

Prerequisites:

Self-study of Module 1 and 2

Learning Outcomes:

At the end of this module you will be

- Familiar with the global HPC market landscape, including market drivers, trends, segments and key players.
- Able to define an SME and get a glance of European SME sector.
- Able to articulate the value proposition of HPC for SME businesses.
- Familiar with the administrative process of SME engagement with HPC providers.
- Able to formalise engagement with HPC providers.

Module 3

- ✓ HPC Global Market Landscape
- ✓ Key drivers for HPC market growth
- ✓ HPC market segments
- ✓ Key players in HPC market
- ✓ How could HPC solutions benefit SMEs
- ✓ Initiatives on HPC adoption initiatives for SMEs: EU&USA
- ✓ Implementation of new types of SLAs
- ✓ SLA/SLIs (Service Level Agreements/Service Level Indicators) templates

3. HPC for Small and Medium Enterprises

3.1 HPC Global Market Landscape

Some key market statistics and trends are shown below (Intersect360 Research, 2018; Grand View research, 2018)

- Global HPC market size was valued at **USD 35.4 billion (€32,26 billion)** in 2017
- HPC market is expected to expand at a **CAGR of 7.2%** from 2018 to 2025 with a projected market size of **USD 59.6 billion (€54,32 billion)** by 2025
- Growth is driven significantly by **commercial HPC solutions**
- HPC market includes **Servers, Storage, Software, Networks, Services and Cloud**
- Revenue-wise **Commercial HPC solutions** constitute 56% of the market, followed by solutions for **Government** (26%) and **Academic** (18%) projects
- While the USA is the **leading market for HPC** (30% of global market share), the fastest growth is happening in Asian countries (China, India, Indonesia etc.)

- By deployment of HPC solutions, **“on-premises”** segment currently holds the top position. The **“on-cloud”** deployed solutions are gaining higher adoption and this segment is currently exhibiting fastest growth.
-

3.1.1 Key drivers for HPC market growth

In this section, we will present the factors that are currently driving demand for HPC solutions and consequently, fueling substantial growth in the HPC market.

- **Scientific research** in the areas of space exploration, energy, weather forecasting etc. that demands high performance computing. For example, the National Nuclear Security Administration (NNSA) and the Office of Science, both of which fall under the U.S. Department of Energy, have embarked upon a collaborative effort called the Exascale Computing Project.

- HPC is finding **widespread application in diverse scientific and engineering areas** such as military defense, national security, training and simulation, navigation technologies, high-resolution image processing, cryptographic analysis, satellite mapping, autonomous vehicle design, among others.

- **Rising popularity of cloud computing coupled with the digitization initiatives** adopted by several governments will drive the market in coming years. For instance, the Government of United Arab Emirates (UAE) has launched digital transformation projects, such as Smart Abu Dhabi and Smart Dubai. Such projects would encourage adoption of cloud computing and trigger the demand for HPC systems.

- The ability of HPC systems to **process large volumes of data at higher speeds** is prompting government agencies, defence agencies, academic institutions, energy companies, and utilities to adopt HPC systems. Enterprises are generating massive amounts of data, which also creates the need for computational solutions and capabilities to manage or handle them. Delivering **"Big data" solutions with HPC systems** is currently one of the top focus areas in today's computing world

- Scientific research as well as business applications involving **Artificial Intelligence (AI) and Machine Learning** technologies necessitate high performance computing solutions

- The emergence of **Exascale Computing** (Fiore et al., 2018) is expected to open new opportunities in the HPC market in the future. Exascale computing refers to computing systems capable of at least one exaFLOPS, or a billion billion (i.e. a quintillion 10^{18}) calculations per second. It is considered to be a significant achievement in computer engineering, for it is estimated to be the order of processing power of the human brain at neural level and accordingly, is being used in the Human Brain Project.
- Systems not running true HPC applications but requiring HPC technologies to run services at supercomputing scale (e.g., Google, Amazon, Baidu, etc.) – referred as **"Hyperscale systems".**

3.1.2 HPC Market Segments

At a broad level, HPC market can be categorized into following two segments:

High-Performance Technical Computing (HPTC):

Applications in Science and Engineering that typically includes following areas:

- Bio/Life Sciences/ Pharmaceuticals
- Manufacturing and Industrial Design
- Geosciences and Geoengineering
- Meteorology
- Energy
- Defense and Aerospace
- National Security
- Government, Academic and Not-for-profit

High Performance Business Computing (HPBC)

Non-scientific Business applications that typically includes following areas:

- Financial Services
- Media and Entertainment
- Retail
- Transportation and Logistics
- Healthcare
- Business Analytics

Table 2 below illustrates these two market segments - HPTC and HPBC, with some example application areas within each segment.

High Performance Technology Computing (HPTC)	High Performance Business Computing (HPBC)
University/Academic - Basic and Applied research	Economics/Financial: Risk analysis, Fraud detection, automated trading, portfolio management, pricing etc.
Government Labs – Basic and applied research	Digital Content Creation and Distribution: Computer-aided graphics in film, media and entertainment (gaming)
Geosciences and Geoengineering – Oil & Gas exploration and Reservoir modelling	Retail: Customer analytics, product design, supply chain management
Chemical Engineering: Process and Molecular design	Transportation
Defense and Energy – Basic and Applied research in areas such as Nuclear, Aerospace, Alternative energy etc.	Training and Simulation
Weather forecasting: Near term and Climate / Earth modelling	Information Technology (IT) and Telecommunication
Computer-aided Engineering: Automotive design and testing, transportation, structural, mechanical design	Healthcare- Data analytics,
Electronic Design and Automation: Electronic component design and verification	
Mechanical Design & Drafting: 2D/3D design and verification, mechanical modelling	
Bio-sciences and the Human Genome : Drug discovery, disease detection / prevention	

Table 2: *HPTC and HPBC Segments*

Figure 5, 6 and 7 are shown with the aim to provide a glance of the HPC market landscape, in terms of market revenues by verticals (Industry / Government / Academic), leading vendors for HPC servers (on % market share) and companies (on % of global revenue earnings from HPC solutions). However, these data could be a bit out-of-date and also, might not be accurate due to the mechanism market research firms use to categorise HPC solutions and calculate market performance.

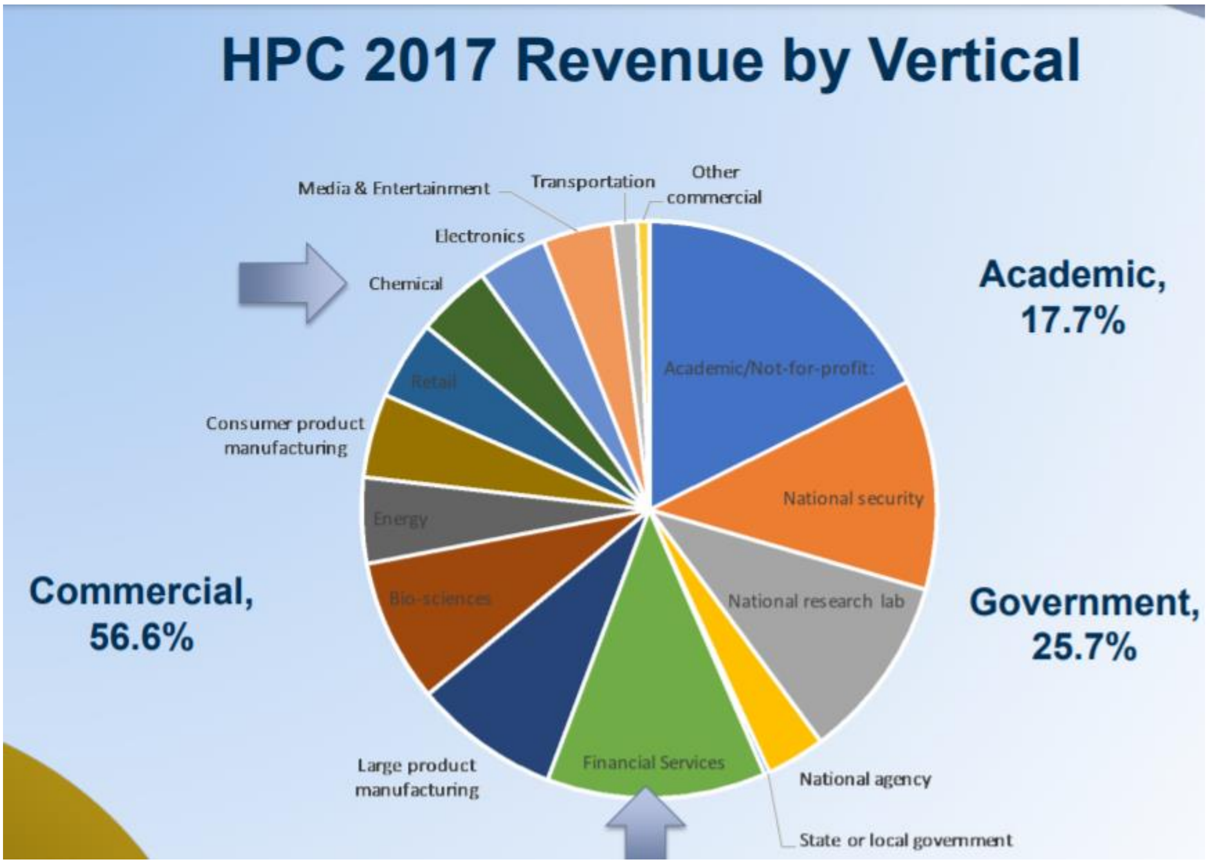

Figure 5: *Global HPC Market Revenue by Verticals 2017*
[Source: Intersect360 Research, 2018]

3.1.3 Key Players in HPC Market

Globally, the top HPC manufactures / solution providers are:

- Hewlett Packard (HP) Enterprise (US)
- Dell (US)
- Lenovo (China)
- IBM (US)
- Cray (US)
- Fujitsu (Japan)
- NEC (Japan)
- Atos SE (Europe, France)
- Silicon Graphics International (SGI), US
- Advanced Micro Devices (AMD), US
- Intel Corporation, US
- Cisco

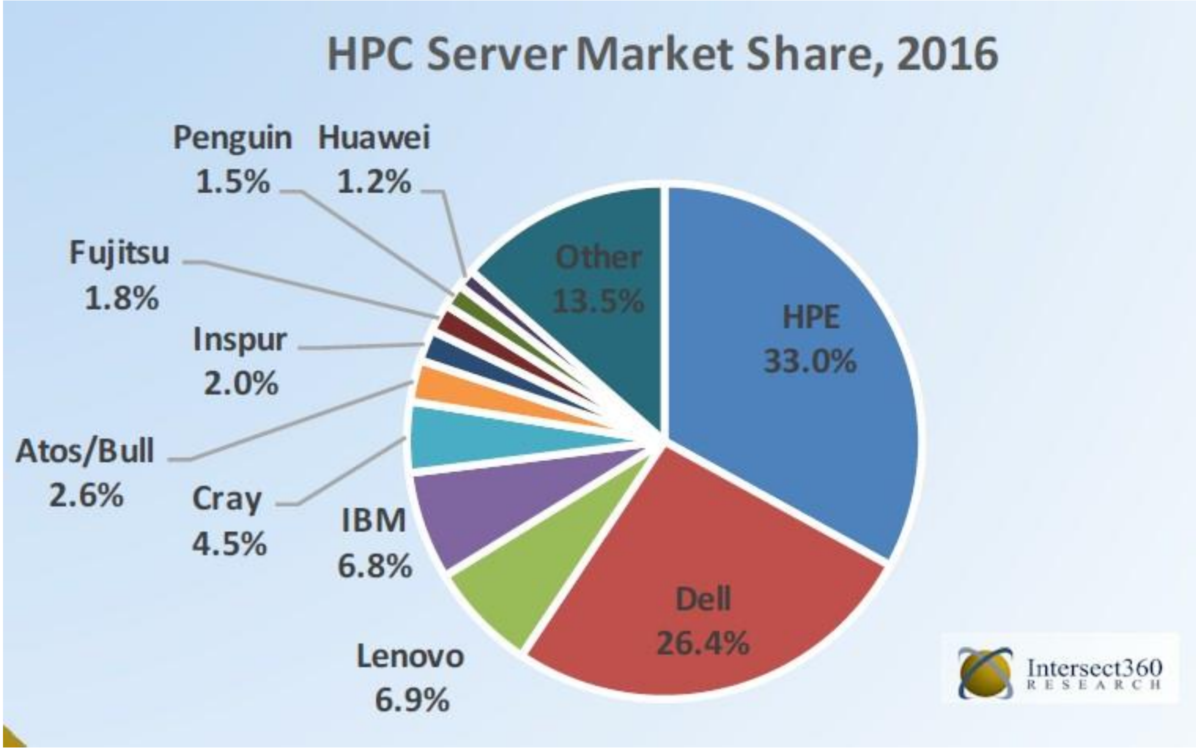

Figure 6: *HPC Server Market Share 2016 [Source: Intersect360, 2017]*

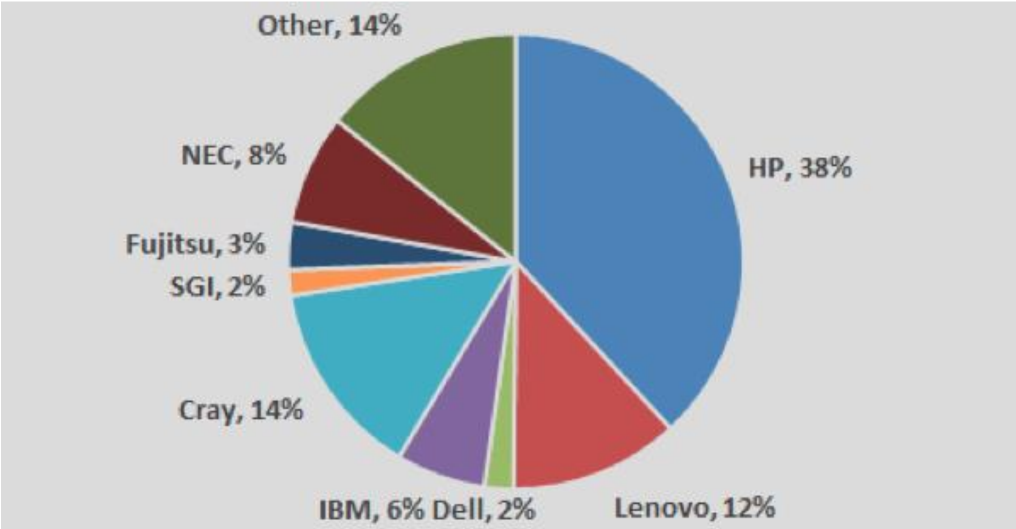

Figure 7: *Company share of Global HPC Revenues Q1 2015 [Source: IDC]*

3.1.4 How could HPC solutions benefit SMEs?

Making HPC solutions accessible to SMEs can be a tremendous differentiator, as it can boost productivity and innovation in the SME sector to an unprecedented level and provide businesses with huge competitive advantage. In addition to running simulations and modelling, HPC can also enable SMEs perform data analytics such as "Big Data" services in a cost-efficient way. An international study of industrial value chains has shown that reputed OEMs (original equipment manufacturers, the firms such as Boeing or GM) ensure that SMEs (located in the value chain) have adequate access to and facility with OEM's high-performance computing environment. One of the key reasons for this decision is to be certain that OEM's key supplier base can interface with their product development systems and also produce the most innovative and cost-effective parts and components of their own.

In this section, we will illustrate some real-world examples from three European SMEs providing services across different industry sectors – Industrial Design, Media and Information Technology. The examples should provide an understanding of how HPC resources can help these SMEs in driving innovation that can have a profound positive impact, not just on national economies, but at a larger level societies and the quality of human life.

Example #1 – RECOM Services [source: European Commission, 2015]

RECOM Services, a 12-person SME based in Stuttgart, Germany, started out doing combustion modelling for large-scale power plants and now does designs and process optimisation for a wide range of industrial furnaces and boilers. This micro-enterprise could not justify buying an HPC system and instead performs simulations on supercomputers at the nearby High Performance Computing Centre Stuttgart. This has enabled RECOM to grow revenue quickly, including recent moves into the U.S. and Asian markets.

Example #2 – Sicos / M.A.R.K. 13 [source: European Commission, 2015]

Sicos subsidiary M.A.R.K. 13 is a 45-person German media company focused on designing movies. This business requires swift reaction time and extremely high quality. When approached to do work for the Australian-German animated movie based on the internationally known book "Maya the Bee," M.A.R.K. 13 entered into collaboration with Germany's High Performance Computing Centre Stuttgart (HLRS) to guarantee high-quality, on-time production of 3D pictures for the 79-minute movie. Although the movie only required about 1% of HLRS resources, it could not have been done without high investment costs by such a small company. Using HLRS resources not only helped speed up the works but also substantially reduced the financial risk for the customer.

Example #3 – SME S1 in Slovenia [source: Marolt et al. 2016]

SME S1 was founded in the beginning of 1990's as an internet service provider and has transformed over the years into a high-tech IT service and solutions provider in the field of HPC. Today it offers services of supercomputer infrastructure to their clients and provides them system administration, optimisation and parallelisation of code, cloud computing services, web and mobile SaaS services. SME S1 provides their clients end to end IT solutions for their problems through either offering services on their own HPC platform or they guide clients for building their own HPC infrastructure. It also offers its clients classical IT system administration, software application development services. Initially they started with leasing only infrastructure for HPC, but this did not generate enough revenue to cover the high maintenance costs, let alone to create profits. Therefore, they started to lease HPC along with specialised services in the cloud and focused mostly on manufacturing SMEs. This redesign of their business model helped SME S1 to adjust the value proposition to its potential customers.

3.2 Initiatives on HPC Adoption by SMEs: Regional and International Perspectives

3.2.1 HPC Adoption Initiatives for SMEs: EU Countries

In February 2012, the European Union released a report titled “High Performance Computing: Europe’s Place in the Global Race”, which articulated a goal (and set of action plans) to achieve HPC leadership including by “acquiring at least one Exascale computer in the same timeframe as the U.S., Japan, and China (European Commission, 2012). Europe’s HPC strategy recognises that “HPC is a strategic resource for Europe’s future” with “computational science already the ‘third pillar’ of science” and “industry relying more and more on HPC to innovate in products and services”. In accordance with this initiative, the European Union has significantly increased funding for HPC systems research and development, including launching the European Technology Platform on High-Performance Computing (ETP4HPC) in 2012, part of a €700 million public-private investment in HPC through the Contractual Public Private Partnership (cPPP) initiative. That investment has been complemented by €400 million in commitments (mostly in-kind contributions) made through the Partnership for Advanced Computing in Europe (PRACE). Reviewing Europe’s efforts since 2012 to bolster its HPC capabilities, market research firm IDC noted in late 2015 that, “The European Commission, PRACE, and ETP4HPC have done an admirable job of advancing Europe’s position in the few years since the 2012 Communication”. IDC notes that, historically, “Europe’s HPC community has been more oriented toward science than industry and industrial access to Europe’s leading supercomputers has grown but remains limited.” However, Europe is working to address this, with the PRACE, SME HPC Adoption Programme in Europe (SHAPE), and Fortissimo initiatives focused on expanding industrial use. Thus, while Europe does “have some world-leading HPC centers for collaborating with industrial users, including SMEs (such as HLRS, Teratec, SURFsara, CINECA, and LRZ)” democratising access to HPC resources for industrial purposes remains as much a challenge in Europe as it is in the United States. The major thrust of the European HPC strategy remains positioning its HPC investments so that it can become a leader in HPC software and applications development. In other words, Europe appears to be trying to position itself to lead in HPC adoption broadly, and in terms of production, to focus on the software rather the hardware side.

Among the world's premier national HPC data centers, the High Performance Computing Centre Stuttgart (HLRS) is situated in the heart of Germany's auto industry. HLRS is seeing increasing demand for HPC from SMEs. Especially in this region; SMEs serve as technology solution providers for larger companies. Increasingly, these large clients require a validation of their technology through simulation. In certain fields, simulation can play a crucial role but is not well known inside tier 1 companies. Very small companies with very special knowledge in modelling and simulation make a living in these small market niches, but they need access to large-scale systems for the computational part of their portfolio (European Commission, 2015).

Recently, the PRACE Council has decided to make SHAPE (SME HPC Adoption Programme in Europe) a permanent service. There have been several initiatives undertaken by this SHAPE programme to help make HPC use more pervasive among Europe's SMEs, in order to increase their competitiveness and ability to contribute to the European economy. This initiative addresses the widespread concern among European HPC stakeholders surveyed that greater outreach to SMEs in needed because relatively few seem to be aware of the value of HPC or the opportunities available to SMEs through PRACE. The projects the SMEs propose for HPC use would not need to be of special scientific or technical interest—they would only need to be important for advancing the SMEs' business prospects. Awards would be based primarily on this importance factor, along with the SMEs' commitment to the process. The initiative would be governed by the same groups that govern the SHAPE programme and would be given access to compute cycles and expertise sufficient to accommodate, say, a dozen SMEs in the first year of the initiative's operation. The experiences would be monitored and reviewed to determine whether the initiative should be continued for a second year, with or without changes.

3.2.2 HPC Adoption Initiatives for SMEs: USA

While large manufacturers have made great progress in leveraging HPC for innovation, the penetration of HPCs into America's SME manufacturing base has been rather slow and sporadic (Ezell and Atkinson, 2016). SME manufacturers (those with fewer than 500 employees) account for about 98.5 percent of U.S. manufacturing companies as of 2013. Moreover, 94 percent of all U.S. manufacturers employ 100 or fewer workers. This vast number of SMEs constitutes the so-called "missing middle" of HPC adoption in U.S. industry. This term "missing middle" refers not directly to company size but rather to a company's computing capacity; the term specifically refers to the group of HPC users between low end, mostly workstation-bound HPC users, and the kind of high-end HPC uses typically performed at national labs and some universities. Nevertheless, in industry parlance, the term has come to refer to the wide swath of small- and mid-sized manufacturers who could be leveraging HPC in their product development or manufacturing processes but are not. For example, one 2013 study estimated that only 8 percent of U.S. manufacturers with fewer than 100 employees are using HPC. Earl Joseph, an HPC analyst at IDC, estimates that at least 25,000 U.S. manufacturers, the vast majority of these SMEs, would benefit from having access to HPC-empowered modelling and simulation tools in design, prototyping, and testing of their parts, components, and finished products.

Three major barriers that prevented America's SME manufacturers' adoption of HPC solutions are identified. First, there exists a general lack of knowledge about how to apply HPC tools to solve engineering challenges, an especially acute problem because many of the engineers working in America's SMEs simply were not exposed to computational sciences in their electrical or mechanical engineering training. Second, taking those engineers "off the line" to train them in modern modelling and simulation tools takes them away from the urgent needs of the business and represents an expense many SMEs cannot incur. Third, and more subtly, many existing modelling and simulation packages (e.g., designed to model aircraft and engines) are often too complex or overdesigned for the needs of smaller manufacturers.

In recognition of these gaps in addressing the high-performance computing needs for the "missing middle" segment of SMEs and also, to leverage the opportunities for a potentially big market for HPC solutions, a few important initiatives have been launched at both regional and national levels. The initiatives primarily focus to help remedy the lack of availability, accessibility, or approachability to HPC tools for SME manufacturers. For example, the National Centre for Manufacturing Sciences (NCMS) has created a dozen centres throughout the United States (located near universities and national labs to tap into local expertise) to connect manufacturing firms with HPC resources. NCMS's network of "Predictive Innovation Centres" represents public-private collaborations providing U.S. manufacturers with HPC tools aimed at decreasing product design cycles, improving manufacturing processes, and reducing the need and costs of laboratory testing of new products. Likewise, the Ohio Supercomputer Centre's (OSC's) AweSim programme is a partnership among OSC, simulation and engineering experts, and industry to assist SME manufacturers with simulation-driven design to enhance innovation and strengthen economic competitiveness. As AweSim Director Alan Chalker explains, "Simulation-driven design replaces physical product prototyping with less expensive computer simulations, reducing the time to take products to market, while improving quality and cutting costs. Smaller manufacturers largely are missing out on this advantage". AweSim levels the playing field by giving smaller companies equal access to HPC technologies. AweSim invites SMEs to bring in their technical challenges and then work with experts to understand how HPC-enabled modelling and simulation tools can help solve their problems. The National Centre for Supercomputing Applications, a hub of transdisciplinary research and digital scholarship led at the University of Illinois at Urbana-Champaign, has also played a pivotal role in helping U.S. enterprises, large and small alike, understand how they can leverage HPC tools to bolster their competitiveness. Likewise, the Chicago-based Digital Manufacturing and Design Innovation Institute (DMDII), one of the institutes within the National Network for Manufacturing Innovation (NNMI), has developed a new cloud-based system to democratise SME manufacturers' access to HPC resources. DMDII envisions its Digital Manufacturing Commons (DMC) as a free, open-source software project to develop a collaboration and engineering platform that serves as an online gateway for digital manufacturing. Akin to an "app store for manufacturing," the DMC sits in a digital services marketplace with a software development kit and collaboration platform at its core, essentially equipping SME manufacturers with the modelling and simulation tools they need to address technical design challenges as well as access to shared HPC resources.

3.3 Implementation of new types of SLAs

A Service Level Agreement (SLA) is typically defined as an official commitment that prevails between an HPC service provider and a customer. In our case we consider a service provider as an HPC Centre which offers typically the access for SMEs to HPC Infrastructure as Service and more advanced or, today, even fully automated HPC Platforms and HPC Applications as presented in Figure 8. The aim of this section is to show key recommendations and guides for SLAs in new contractual models to help define better formal agreements between HPC Centers and SMEs or improve the existing or future business contracts.

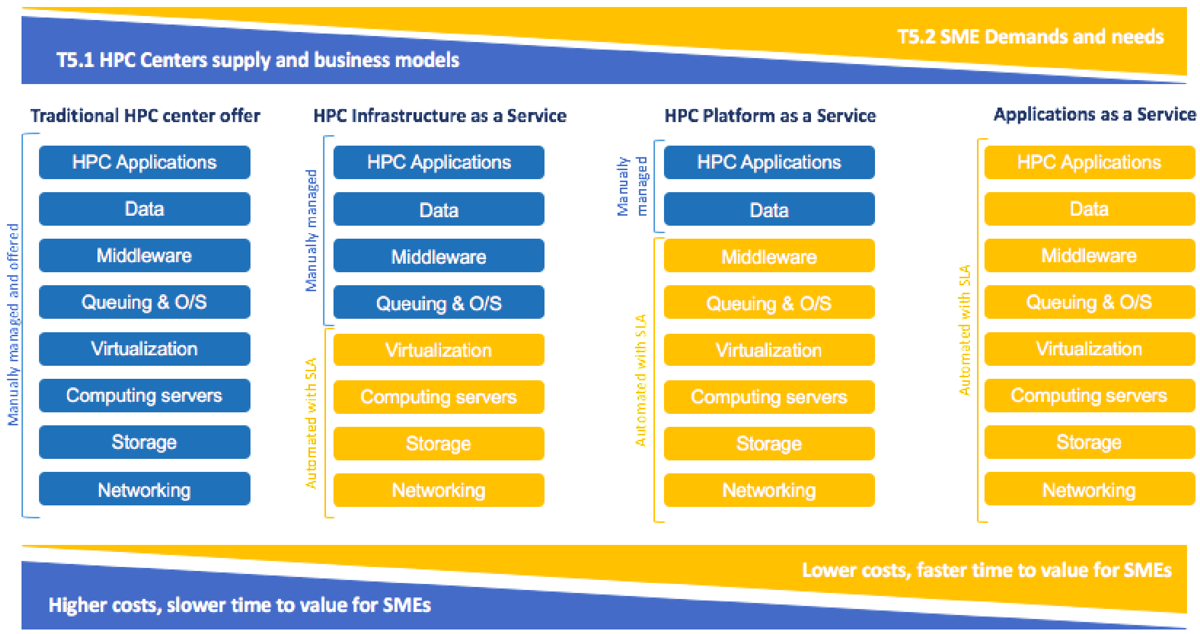

Figure 8*. The ongoing transformation of traditional HPC center offer into lower costs and more flexible HPC Infrastructure, Platform and Applications as a Service*

All performed analyses have, to date, been based on various existing measured parameters and characteristics of HPC infrastructures provided by HPC Centres that can be offered (or will be offered in the near future) in a fully automated way to customers as a service with SLAs attached. Today, or in the near future, as presented in Figure 9, a set of potential SLA parameters can be defined and fully automated, managed and enforced in a contract for SMEs in relationship not only relating to a traditional HPC resources, including Networking, Storage and Computing, but also to upper layers in PaaS and SaaS architectures, including virtualized HPC clusters on demand, middleware and back-end services, databases and data analytics tools and finally HPC applications. Before we formulate a set of basic assumptions for SLA (Service Level Agreement) let us introduce basic terms and definitions.

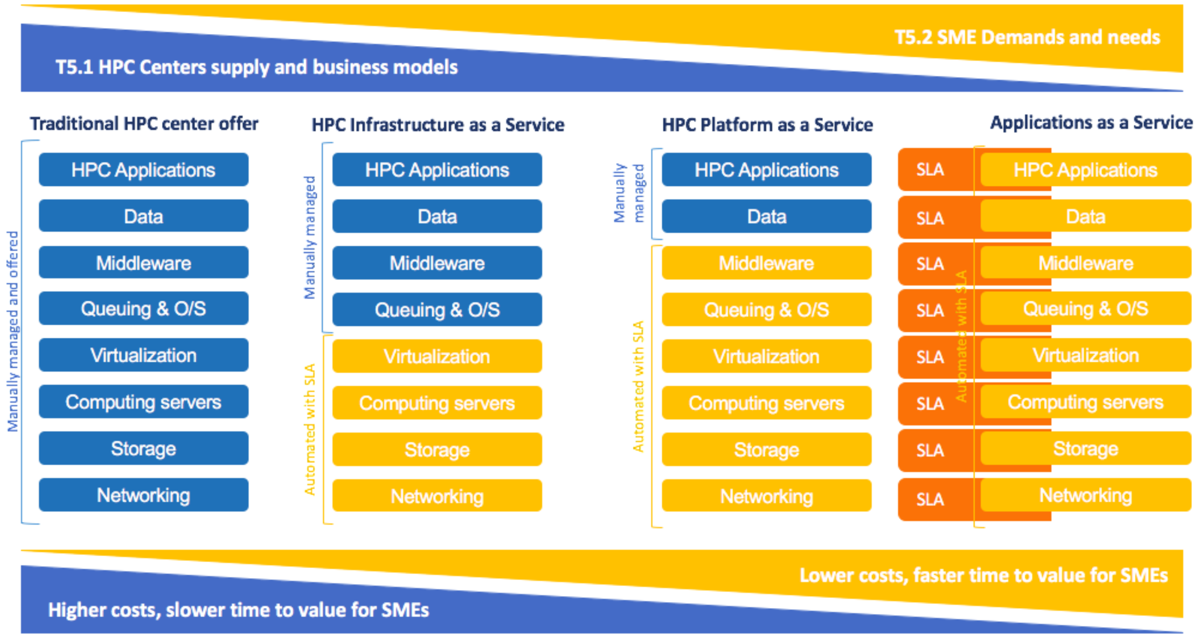

Figure 9*. The implementation of new types of SLAs in contractual models defined in the relationship with all the layers in service-oriented HPC architectures.*

In many existing contractual models which fit into the typical SLA definition introduced above Service Level Indicators are used to gauge a service against its target Service Level Objectives (SLO - Service Level Agreement). One or more SLOs form part of as SLA. Both SLOs and SLAs can be used for internal purposes as well as external ones.

For example:

- SLI (Service Level Indicators): % of free space in a HPC/storage cluster;
- SLO_internal: Minimum 15% of the HPC/storage clusters offered by HPC Center total space should be unallocated;
- SLO_SME_ customer: 5% of the HPC/storage cluster’s total storage space should be unallocated;
- Error budget: SLO_SME_ customer - SLO_internal = 10%.

When the error budget starts to be consumed, corrective measures must be initiated, definitely before the SLO_SME_customer level is reached. Breach of SLO_internal is a paging event, someone wakes up due to it, and therefore the nominal operation is to never breach even the SLO_internal. Other SLI (Service Level Indicators) definitions used in the market place are:

- Aggregation intervals: “Averaged over 1 minute”;
- Aggregation regions: “All the tasks in an HPC cluster”;
- How frequently measurements are made: “Every 10 seconds”;
- Which requests are included: “HTTP GETs from black-box monitoring jobs”;
- How the data is acquired: “Through our monitoring, measured at the server”;
- Data-access latency: “Time to last byte”.

Additionally, it is worth mentioning a set of recommendations for SLAs and contractual models taken from the biggest service providers in the market, e.g. Google:

- Don’t pick a target based on current performance - While understanding the merits and limits of a system is essential, adopting values without reflection may lock you into supporting a system that requires heroic efforts to meet its targets, and that cannot be improved without significant redesign;
- Keep it simple - Complicated aggregations in SLIs can obscure changes to system performance, and are also harder to rationalise;
- Avoid absolutes - While it’s tempting to ask for a system that can scale its load “infinitely” without any latency increase that is “always” available, this requirement is unrealistic. Even a system that approaches such ideals will probably take a long time to design and build, and will be expensive to operate - and probably turn out to be unnecessarily better than what users would be happy (or even delighted) to have;
- Have as few SLOs as possible - Choose just enough SLOs to provide good coverage of your system’s attributes. Defend the SLOs you pick: if you can’t ever win a conversation about priorities by quoting a particular SLO, it’s probably not worth having that SLO. However, not all product attributes are amenable to SLOs; it is difficult to specify “user delight” with an SLO;

- Perfection can wait - You can always refine SLO definitions and targets over time as you learn about a system's behaviour. It is better to start with a loose target that you tighten than to choose an overly strict target that has to be relaxed when you discover it's unattainable.

Above best practices regarding SLAs show that it is not very useful to expose a large number of detailed SLIs, that although useful to keep the totality of a distributed in good health, to SMEs. One can therefore speak in general terms of internal SLIs vs external SLIs. Thus, for SMEs it is optimal to find and expose SLIs that are short, descriptive, understandable and highly representative, i.e. provide a good coverage of the total HPC system health, in a single metric. For internal use it is useful to have, in addition to the SME facing ones, also a range of SLIs and SLOs that is more narrow and detailed on particular components of a system, rather than the whole. These SLIs can foretell a looming problem, which may be captured at the higher level, SME chosen ones, but can in doing so also point to a particular component of a system.

3.3.1 SLA/SLIs templates

The four main categories of Service Level Indicators in various SLAs are typically defined:

- Availability
- Latency
- Throughput
- Durability

The following example definitions can be used as references or templates for new contractual models implemented by HPC Centres to match SME needs.

Availability - is the fraction of time a service for SMEs offered by HPC Center is usable. What constitutes usable is sometimes somewhat subjective. Clearly, a system which does not respond at all, even with infinite latency, is not available. But even a system that responds, but with perhaps 1000x the normal latency, may be classified as unavailable, because it is so out of specification that its users cannot use it for its intended purpose.

Latency – latency measurements are difficult to formulate in SLA/SLIs, but important, to get right, as there are two conflicting interests:

- Potential SMEs require aggregated information;
- Complex systems have widely disparate typical latencies, making it difficult to make aggregate numbers representative.

For typical HTTP based API:s (and similar), the metric data should be labelled with PATH and OP, so that the reporting system can differentiate between operation and get deeper insight into a system.

Throughput - for an offered HPC service is measured as the number of requests handled in aggregate during a time period. A throughput SLO on a service means the system should be able to perform up to such a level, but it may not be performing to that level at all times. The SLO is only breached if there is a client request load which isn't being served, i.e. operations issued aren't serviced properly (either client-side or server-side, by timeouts or other similar errors). A high throughput usually translates into a high per-op latency, meaning these two SLI categories are somewhat contradictory. Realistic load tests are necessary to determine what actual throughput values can be serviced, while still staying within the SLO of the system component.

Durability - this SLA/SLI category applies mostly for storage systems which are typically attached to computing clusters offered by HPC Centers. It measures the durability of data, i.e. the absence of loss of any data objects. Durability of the data is measured by referring to the industry adopted terms, i.e. RPO (Recovery Point Objective) and RTO (sRecovery Time Objective).

3.3.2 Reference SLA/SLIs for SMEs

HPC infrastructure component	Latency test	Throughput test
HPC applications and compute	KPI1: Time needed to create and run an HPC application or virtual machine [time]	KPI2: CPU & memory speed benchmark [unit to be defined]
Compute platform / storage	KPI3: Latency of the compute platform / storage [time]	KPI4: Throughput to the compute platform / storage [data volume/time period]
Datasets object storage	KPI5: Access latency to datasets from HPC compute platform [time-to-first-byte]	KPI6: Transfer throughput between object storage and HPC compute platform [data volume/time period]
Network	KPI7: Latency between HPC Center and SME [time]	KPI8: Throughput between HPC Center and SME [data volume/time period]

Table 3: *HPC infrastructure component*

All the above SLIs also require SLOs corresponding to SMEs. All measurements in time and throughput will be measured according to a point in the statistical distribution (percentile). Availability of an KPI (Key Performance indicator) in the SLA is the fraction of samples of an SLI that are meeting its SLO. Expected order of magnitude of the SLOs defined in the following table, for performance metrics, and binary objective for pure availability metrics.

KPI #	**Expected SLO value order of magnitude**
1	O(10s)
2	O(xx [unit TBD])
3	O(100ms)
4	O(100MB/s)
5	O(1s)
6	O(50MB/s)
7	O(50ms)
8	O(5Gbps)

Table 4: *KPIs*

The availability of each KPI is measured as the fraction of SLI samples that meets the SLO value. The overall flow and its structure for SLA within HPC Centres and its relationship to a business model for a certain number of SMEs together with a typical business relationship established in a formal contract, and then monitored during day-to-day operations are presented in Figure 10.

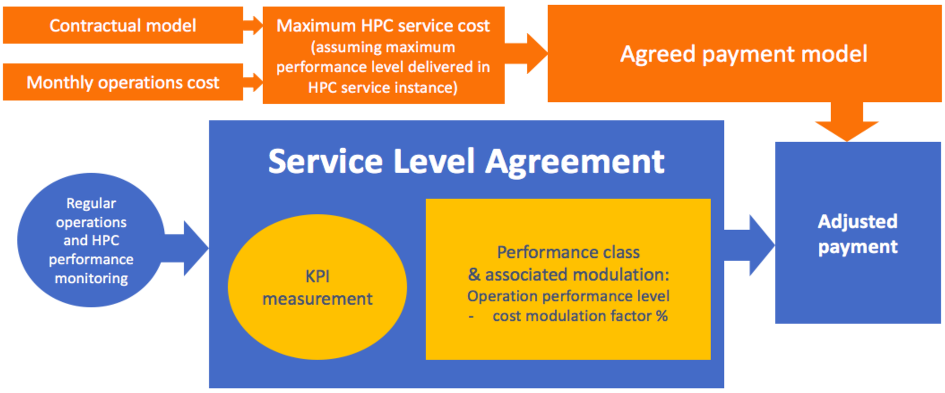

Figure 10: *Key components in new contractual models for HPC Centers taking into account various SLA parameters offered in service-oriented IaaS, PaaS and SaaS paradigms.*

3.4 Reflections on this module learning

This module presents the HPC global market landscape, including market drivers, trends, segments and key players. It shows how HPC solutions could benefit SMEs. The implementation of new types of SLAs are shown as well as SLA/SLI templates.

3.5 Recommended Further Readings / Resources

1. SME Demands and Needs (SESAME Net) (https://sesamenet.eu/wp-content/uploads/2016/12/D5.1_fin.pdf):
   This document describes the approaches and tools used to gather feedback from European SMEs about their knowledge, use and barriers in connection with HPC and related services (Cloud, Storage, High End Visualization, etc.).
2. Review Skills Gaps and Training Needs of SMEs (SESAME Net) (https://sesamenet.eu/wp-content/uploads/2016/12/D6.1_fin.pdf):
   Presented in this chapter are some of the training demands of SMEs. This is examined in light of existing provision of HPC training in Europe. Areas are also identified where there are opportunities for SESAME when developing training material in the remainder of the project.
3. European high-performance computing (ETP4HPC) (https://www.etp4hpc.eu/pujades/files/ETP4HPC_Handbook_2018_web_20181210.pdf):
   In this document you will get information about European Exascale projects, basic technology projects started in 2015, basic projects starting in 2018/2019, Co-design projects and other HPC projects. You will also get a quick overview of project, objectives and motivation about each project. In the second half of the document, centres of excellence in computing applications started in 2015 and 2018 are described and presented.
4. HPC Supply - SME matching measures (SESAME Net) (https://sesamenet.eu/wp-content/uploads/2018/01/D5.3_HPC-Supply-SME-matching-measures.pdf):
   The goal of this document is to determine a set of measures to be implemented that will increase the engagement of HPC Centers with SMEs. It also analyses and determines the mechanism for maximizing the interaction between SMEs and HPC Centres, including the understanding of SME needs and demands, the current provisioning and business models and finally determining the best engagement model with computing centres.

SME/HPC

Co-funded by the
Erasmus+ Programme
of the European Union

MODULE 4
Exercises and solutions to practical problems by using HPC

Intermediate

Aims of this Module:

This module aims to describe / explain:

- HPC infrastructure.
- Show examples of remote work on HPC.

Prerequisites:

Self-study of Module 1, 2 and 3

Learning Outcomes:

At the end of this module you will

- Be familiar with HPC infrastructure.
- Have hands on experience by remotely working on HPC.
- See specific use cases for domain environment.

Module 4

- ✓ HPC Terms and technologies
- ✓ HPC Intro – Exercises

4. Exercises and solutions to practical problems by using HPC

4.1 HPC Terms and technologies

4.1.1 Clusters

A computer cluster is a single logical unit consisting of a set of connected servers, working together as a single powerful machine. The individual servers (aka nodes) of a computer cluster are usually connected through very fast network connections (called the interconnect). The nodes of a cluster often have different purposes, e.g. the cluster is often composed of compute, login, storage and other specialised nodes e.g. for visualisation or with extra-large memory.

A computer cluster can accommodate much more processors and provide much more storage capacity than a single computer and therefore can provide much faster processing speed. The higher speed is mainly achieved by processing numerous tasks in parallel. It means that a number of processors cooperatively solve a problem at the same time (parallel computing). An example of a computer cluster is the HPC system “Avitohol” which forms the core of the computing infrastructure in the Institute of Information and Communication Technologies (IICT) in Bulgaria. It consists of 150 HP Cluster Platform SL250S GEN8 servers, each with 2 Intel Xeon E2650 v2 CPUs and 2 Intel Xeon 7120P coprocessors.

4.1.2 Supercomputers

Currently the fastest systems worldwide are often called Supercomputers. There are different constructions and architectures for Supercomputers existing. While in 2000 diverse architectures for supercomputers were popular, today most supercomputers are built as computer clusters consisting of mostly standard servers often equipped with faster network connections.

Another interesting development concerns the number of processors a Supercomputer provides. While around 1970 supercomputers used only a couple of processors, in the 1990s the first machines with thousands of processors appeared. By the end of the 20th century, massive supercomputers with tens of thousands of processors were the norm. Currently the fastest system at the time of writing, “Summit” of the DOE/SC/Oak Ridge national Laboratory United States which is reported to employ a combined number of 2414,592 compute cores

As a measure of the computation power of a system, the number of floating point operations it can perform in one second (FLOPS) is often used. The theoretical maximum that the system can achieve in the best case, based on its hardware limits, is called the “theoretical peak performance” and can simply be calculated using the technical specifications of the vendor. However, for practical measurements and especially for the inter-comparison of the performance of these systems, the High-Performance Linpack Benchmark (HPL) has prevailed. This benchmark measures the performance of the system solving a dense system of linear equations.

A list of the highest results achieved using this HPL benchmark is regularly compiled as the “Top500 list of Supercomputers”. From this list one can obtain lots of information regarding types of supercomputing systems, their hardware, software, etc. For example, at the time of writing, all of 500 fastest supercomputers documented in the list work at more than one Petaflops or 1015 floating-point operations per second.

However, problems that arise in industry rarely require such high amounts of computational power. An HPC Clustre of medium size, even though it has never been in the TOP500 list, may be entirely appropriate to handle most problems.

4.1.3 Shared memory

When a system is said to have shared memory, this means that although the system may have hundreds of processors, the memory space is not divided, but is considered shared between these processors. This means that changes made by one processor are visible to the others and that explicit communication is not required. Nevertheless, programming for shared memory systems requires some way of synchronisation to be implemented. Nowadays, even desktops and smartphones can be considered as shared memory parallel machines (figure 11). The most used Application Programming Interface (API) is the OpenMP® API specification for parallel programming.

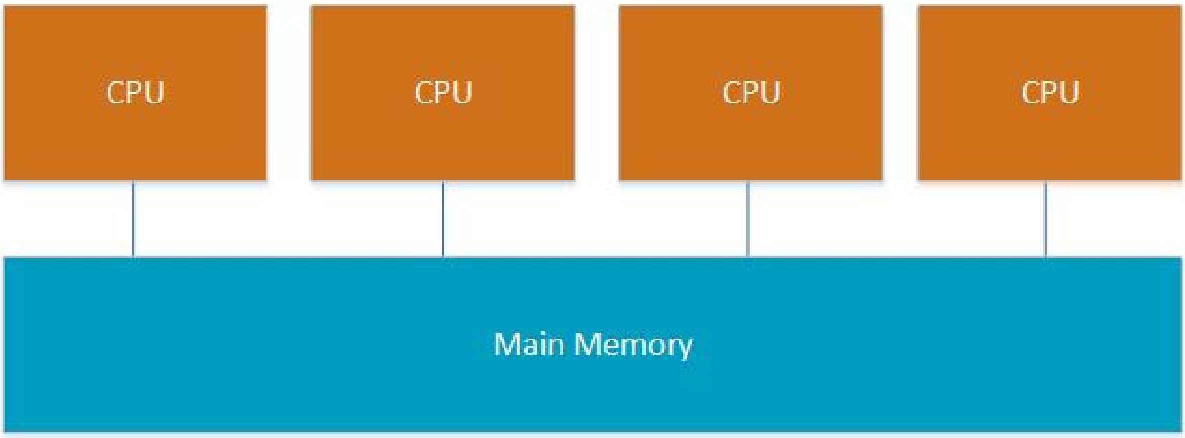

Figure 11: *Shared memory systems*

4.1.4 Distributed memory

In systems with distributed memory (figure 12) the memory is distributed between the processors. Each processor can only see its own memory. Changes made by one processor need to be communicated explicitly if other processors need to see them. The Message Passing Interface (MPI) consists of standardised libraries for message-passing and allows applications to explicitly transfer data between processors and synchronises the computations.

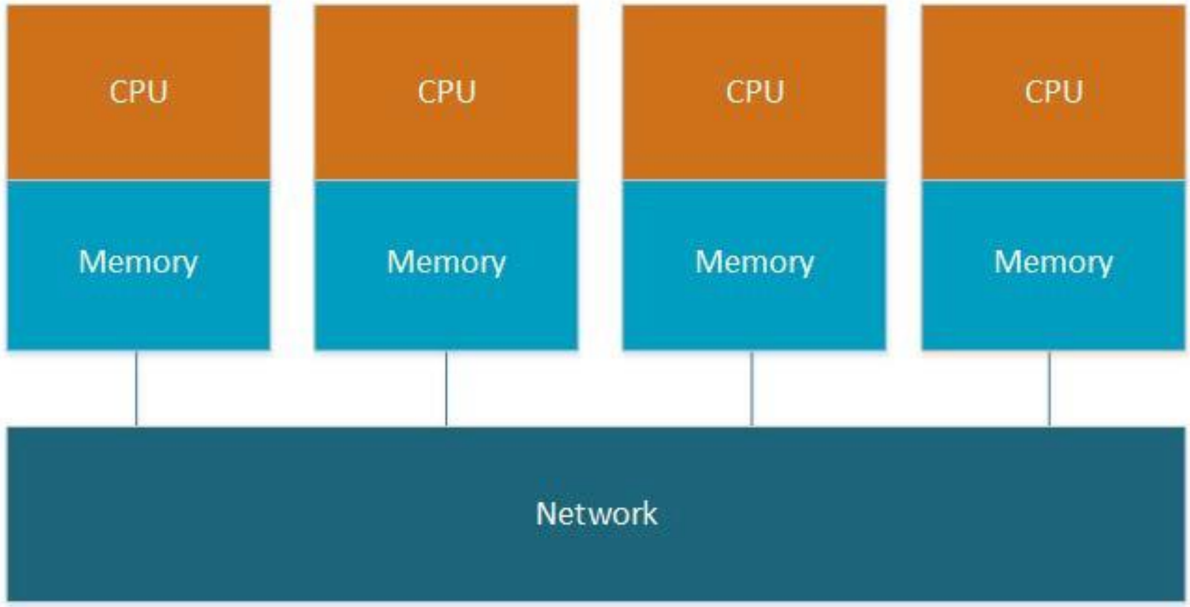

Figure 12: *Distributed memory systems*

4.1.5 Hybrid Systems

Most parallel computers are neither only a shared memory nor a purely distributed memory type but a mixture or hybrid of both. To get as much as possible performance out of a single compute node they are sometimes combined with accelerators.

Examples of such accelerators are graphics cards or Graphics Processing Units (GPUs as opposed to the general CPUs), general floating-point accelerators or Field Programmable Gate Arrays (FPGAs). Even CPUs for general purposes have vector instructions that are advantageous for HPC. Software that exploits vector processing offers much higher efficiency.

4.1.6 Parallel Programming

OpenMP

Open Multi-Processing (OpenMP) is a set of compiler directives, library routines and environment variables. OpenMP gives programmers a simple and flexible interface for developing parallel applications for platforms ranging from the standard desktop computer to the supercomputer. OpenMP is used on shared memory machines. On high performance clusters OpenMP is often combined with Message Passing Interface (MPI). This is called hybrid programming.

MPI

Message Passing Interface (MPI) is a standard to allow communication between the processors by exchanging messages on and between the compute nodes. It works on a variety of architectures often also called parallel computers. MPI is generally considered to be the industry standard and forms the basis for most communication interfaces adopted by HPC programmers. The standard defines the syntax and semantics of a core of library routines.

4.1.7 Usage example

Connecting to the HPC Centre

While not all use cases actually require the SME to manually perform computations on the HPC system, here we describe a common way of doing so. Typically, the HPC Centre will provide access to the system at user-level, meaning without any kind of administrative rights, and through an interactive command-line interface for general purpose access.

The most typical scenario, when we have access via MS Windows, would proceed as follows:

- An account request is submitted to the HPC Centre via a web interface or similar.
- The account is validated and confirmed. A username and password or alternative login credentials are provided to the user.
- The user can then login to a server which is connected to the Cluster (login node).

Via a program implementing the Secure Shell (SSH) protocol, also called an SSH client. SSH client software is available for all popular operating systems, including mobile systems such as Android. The SSH protocol also accommodates file transfers, so data can easily be transferred to and from the HPC system. Software to provide more comfortable file transfers through SSH exists for several operating systems as well, e.g. free WinSCP software for Windows. Optionally, to use programmes with a Graphical User Interface (GUI) on the HPC system, the client computer, i.e. your desktop PC, will need the installation of additional software. The HPC systems typically use the “X Window System”, which requires a counterpart on the client PC to function over network connections. The method of displaying a GUI window of an application running on the remote server on your PC is called “X-forwarding”, and typically needs to be enabled in the SSH client. Usually Linux PCs provide this functionality out-of-the-box, and there are free applications that provide such functionality on MS Windows.

Once connected to the HPC system, the user can interactively discover what kind of software packages and libraries are installed. If the particular software intended to be used is not readily available on the HPC system, it can be deployed by the HPC Centre. Typically, the centre provides web-based documentation, as well as a support helpdesk for these kinds of questions and general help. Because the HPC systems may support numerous users at the same time, they usually provide a way to “enqueue” the execution of programmes in a kind of waiting line system. A single call of an application in such a waiting system is then called a “(compute) job” or old fashioned “batch job”. A practical example of this mechanism is given in the next section.

Running jobs

Usually the computational workloads on an HPC system are divided into jobs. One job takes a particular subset of the available processors (or processor cores) and executes an application on them for a period of time, until completion or until its allotted time expires. Most of the jobs on an HPC system are parallel, taking much more than one CPU core for their execution. Jobs are sent to queues, where they are executed following some prioritisation strategy. Every user is given information about the suggested queues. A job is usually executed without need for user interaction, but if required, users can also request an interactive type of job. Interactive jobs are usually used for development and testing. The status of a job can be queried and usually e-mail notifications may be requested to be notified of status changes.

Some HPC Centres are able to provide higher levels of security, by creating virtual (or real) private sub-clusters, for example created for exclusive use by a company. Because of the exclusive access, the company may then organise the execution of workloads differently.

4.2 HPC Intro – Exercises

4.2.1 Connecting to the HPC system

Opening a Terminal

Connecting to an HPC system is most often done through a tool known as “SSH” (Secure SHell) and usually SSH is run through a terminal. So, to begin using an HPC system we need to begin by opening a terminal. Different operating systems have different terminals, none of which are exactly the same in terms of their features and abilities while working on the operating system. When connected to the remote system the experience between terminals will be identical as each will faithfully present the same experience of using that system.

Here is the process for opening a terminal in each operating system.

Linux

There are many different versions (aka “flavours”) of Linux and how to open a terminal window can change between flavours. Fortunately most Linux users already know how to open a terminal window since it is a common part of the workflow for Linux users. If this is something that you do not know how to do then a quick search on the Internet for “how to open a terminal window in” with your particular Linux flavour appended to the end should quickly give you the directions you need.

A very popular version of Linux is Ubuntu. There are many ways to open a terminal window in Ubuntu but a very fast way is to use the terminal shortcut key sequence: Ctrl+Alt+T.

Windows

While Windows does have a command-line interface known as the “Command Prompt” that has its roots in MS-DOS (Microsoft Disk Operating System) it does not have an SSH tool built into it and so one needs to be installed. There are a variety of programmes that can be used for this, two common ones we describe here, as follows:

PuTTY

It is strictly speaking not necessary to have a terminal running on your local computer in order to access and use a remote system, only a window into the remote system once connected. PuTTy is likely the oldest, most well-known, and widely used software solution to take this approach. PuTTY is available for free download from www.putty.org. Download the version that is correct for your operating system and install it as you would other software on your Windows system. Once installed it will be available through the start menu or similar.

Running PuTTY will not initially produce a terminal but instead a window full of connection options. Putting the address of the remote system in the "Host Name (or IP Address)" box and either pressing enter or clicking the "Open" button should begin the connection process. If this works you will see a terminal window open that prompts you for a username through the "login as:" prompt and then for a password. If both of these are passed correctly then you will be given access to the system and will see a message saying so within the terminal. If you need to escape the authentication process you can hold the control/ctrl key and press the c key to exit and start again. Note that you may want to paste in your password rather than typing it. Use control/ctrl plus a right-click of the mouse to paste content from the clipboard to the PuTTY terminal. For those logging in with PuTTY it would likely be best to cover the terminal basics already mentioned above before moving on to navigating the remote system.

Logging onto the system

This will differ for each HPC system. For this lesson, we will use examples for connecting to Arctur2 - an HPC system at Arctur. Although it's unlikely that every system will be exactly like Arctur2, it is a good example of what you can expect from an HPC system. To connect to our example computer, we will use SSH (if you are using PuTTY, see above for instructions). SSH allows us to connect to Linux computers remotely, and use them as if they were our own. The general syntax of the connection command follows the format: ssh yourUsername@some.computer.address Let's attempt to connect to the HPC system now:

```
ssh yourUsername@hpc-login.arctur.si -p 8022
```

Examining the nodes and Arctur2 system

Now we can log into the Arctur2 HPC system where we will look at the nodes. You should remember that there are at least two types of node on the system: login nodes and compute nodes.

We can use the *lscpu* command to print information on the processors on the login nodes to the terminal.

4.2.2 Arctur-2 system overview

Firstly, we will take a quick look how Arctur-2 is organised and go over the general specifications.

Currently, we have 3 different partitions:

- compute
- gpu
- bigmem

The '**compute**' partition is made up of 14 nodes, with the names node01 to node14. This is the default partition and your jobs will run on it if not specified otherwise. Each of these nodes has 2 Intel Xeon E5-2690v4 processors, together having 28 cores clocked at 2.60 GHz. Every node also has 128GB of fast DDR4 RAM. They also have 480GB of local SSDstorage.

The '**gpu**' partition consists of 8 nodes (gpu01 to gpu08). The only difference with compute nodes is that they each have 4 NVIDIA Tesla M60 GPUs. The Tesla M60 is a very powerful GPU as it is made up of two physical NVIDIA Maxwell GPUs with combined 16GB of memory. Many applications perceive the card as two separate GPUs, essentially having 8 GPUs per node. The '**bigmem**' nodes also have 28 CPU cores, but they are clocked a bit lower, at 1.7GHz. As you can imagine, they make up for it by having 1TB of DDR4 RAM and 2 480GB local SSD disks.

Your home folder is located on a shared NFS, which consists of SSD cached HDD disks. The default limits are 100GB per user, but if you need more feel free to contact support. All the nodes and the filesystem are interconnected with a 2x25GbE connection.

Units

A computer's memory and disk are measured in units called bytes. The magnitude of a file or memory use is measured using the same prefixes of the metric system: kilo, mega, giga, tera. So 1000 bytes is a kilobyte (kB), 1000 kilobytes is a megabyte (MB), and so on.

You may also see units that use multiples of 1024 rather than 1000. So, 1024 bytes is a kibibyte (KiB), 1024 KiB is a mebibyte (MiB), 1024 MiB is a gibibyte (GiB) and so on.

For small amounts of storage the differences between these two unit systems are negligible but as the size increases the differences can be significant. We can now repeat this for the compute nodes.

All interaction with the compute nodes is handled by a specialised piece of software called a scheduler, in this case SLURM.

4.2.3 SLURM basics

Before we dive in into SLURM, it would be good to consult the official quickstart guide: https://slurm.schedmd.com/quickstart.html

We are going to explain and show how to use SLURM and some of our custom tools.

SLURM is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters. It is used on Arctur-2.

- It allocates exclusive access to the resources (compute nodes) to users during a job or reservation so that they can perform their work
- It provides a framework for starting, executing and monitoring work
- It arbitrates contention for resources by managing a queue of pending work
- It permits to schedule jobs for users on the cluster resource

Commonly used SLURM commands

sacct is used to report job or job step accounting information about active or completed jobs.

salloc is used to allocate resources for a job in real time. Typically this is used to allocate resources and spawn a shell. The shell is then used to execute ***srun*** commands to launch parallel tasks.

srun is used to run a parallel job, it will also first create a resource allocation if necessary.

There are two types of jobs:

- interactive: you get a shell on the first reserve node
- passive: classical batch job where the script passed as argument to sbatch is executed

We will now see the basic commands of SLURM. Connect to Arctur2. You can request resources in interactive mode like this:

```
$ srun --pty bash
```

You should now be directly connected to the node you reserved with an interactive shell. Keep in mind that only you have access to the node, and it will be billed as you are running a job. Now exit the reservation:

```
$ exit  # or CTRL-D
```

When you run exit, you are disconnected and your reservation is terminated (billing stops). Currently, there are no time limits enforced on the reservations or jobs. To run a passive job, use srun or sbatch. One example is the following:

```
$ srun -N2
```

If you use the command 'hostname' (which prints the hostname of the host it is running on) on two nodes, you should see which nodes were allocated to you, this should be a very short job.

Be sure to check out all optional arguments srun can take by typing in **'man srun'** or by looking at the official documentation on https://slurm.schedmd.com/srun.html

Using **srun** like this will give the job output in you terminal session, and you can't really do anything else in that session until the job is done. A better approach for submitting jobs is to use **sbatch**.

The command **sbatch** takes a batch script as an argument and submits the job. In the script you specify all options such as the partition you want resources from, the number of nodes and similar.

An example of a simple batch script which runs a command (for example the command 'hostname') on 2 compute nodes is this:

```
#!/bin/bash -l
#SBATCH --account=
#SBATCH --partition=compute
#SBATCH --nodes=2
#SBATCH --time=00:20:00
#SBATCH --job-name=my_job
srun
```

Using your favourite text editor, save this as myjob.sh and use sbatch to run it:

```
  $ sbatch myjob.sh
```

You will get a jobID back from sbatch, which you can use to control your job (we cover this later). Unless specified otherwise, the output still be stored in a text file in the same folder in which the script is.

Running jobs on different partitions

Just define the partition name in the appropriate place in the job submission script like this:

```
#SBATCH --partition=
```

As shown before, available partitions are: compute, gpu and bigmem.

4.2.4 Job management

To check the state of the cluster (idle and allocated nodes) run:

```
$ sinfo
```

This is useful to see the state of the resources, and how many are available to you immediately. All the *idle* nodes are ready for use. If you need more nodes than currently available (if some other jobs are running in the system), just submit your job and it will wait in queue until requested resources are available. Sometimes, we will run our internal low priority jobs on the cluster too. They will run in a low priority queue, and will be suspended when you start your jobs. Unfortunately, with ' **sinfo**' you won't be able to determine how many nodes run low priority jobs. For that, we have developed another tool called '**savail**'. Try it and check if there are any nodes running low priority jobs:

```
$ savail
```

You can check the status of your (and only your) running jobs using squeue command:

```
$ squeue
```

Then you can delete your job by running the command:

```
$ scancel JOBID
```

You can see your system-level utilization (memory, I/O, energy) of a *running* job using:

```
$ sstat
```

In all remaining examples of reservation in this section, remember to delete the reserved jobs afterwards (using scancel or CTRL-C)

Pausing, resuming jobs

To stop a waiting job from being scheduled and later to allow it to be scheduled:

```
$ scontrol hold
$ scontrol release
```

To pause a running job and then resume it:

```
$ scontrol suspend
$ scontrol resume
```

4.2.5 Transferring files

If you are linux user you can use the command:

```
scp -P 8022 (-r) file
username@hpc-login.arctur.si:
```

which will copy the file to your home folder, or

```
sftp -P 8022 username@hpc-login.arctur.si:
```

which will open an interactive ftp session. If you want to place things in subfolders, remember, the path to your home folder is '/home/users/username'. The command will then look like this:

```
scp -P 8022 (-r) file
username@hpc-login.arctur.si:/home/users/username/subfolder
```

Alternatively, if you are more comfortable with a GUI and would like to drag and drop files to our system, you can connect via your file manager. For example, in Nautilus (default in Ubuntu and Fedora) you click on 'Other locations' or 'Browse network' (depending on version) and click 'Connect to server'. In the server address just enter:

```
sftp://hpc-login.arctur.si:8022/home/users/username
```

and click connect.

4.2.6 Accessing software

Before we start using individual software packages, we need to understand the why multiple versions of software are available on HPC systems and why users need to have a way to control which version they are using. The three biggest factors are:

- software incompatibilities;
- versioning;
- dependencies.

Software incompatibility is a major headache for programmers. Sometimes the presence (or absence) of a software package will break others that depend on it. Two of the most famous examples are Python 2 and 3 and C compiler versions. Python 3 famously provides a *python* command that conflicts with that provided by Python 2. Software compiled against a newer version of the C libraries and then used when they are not present will result in a nasty *'GLIBCXX_3.4.20'* not found error, for instance.

Software versioning is another common issue. A team might depend on a certain package version for their research project - if the software version was to change (for instance, if a package was updated) it might affect their results. Having access to multiple software versions allows a set of researchers to prevent software versioning issues from affecting their results.

Dependencies are where a particular software package (or even a particular version) depends on having access to another software package (or even a particular version of another software package). For example, the VASP materials science software may depend on having a particular version of the FFTW (Fastest Fourer Transform in the West) software library available for it to work.

Environment modules

Environment modules are the solution to these problems. A *module* is a self-contained description of a software package - it contains the settings required to run a software package and, usually, encodes required dependencies on other software packages. There are a number of different environment module implementations commonly used on HPC systems: the two most common are *TCL modules* and *Lmod*. Both of these use similar syntax and the concepts are the same; so learning to use one will allow you to use whichever is installed on the system you are using. In both implementations the module command is used to interact with environment modules. An additional subcommand is usually added to the command to specify what you want to do. For a list of subcommands you can use module -h or module help. As for all commands, you can access the full help on the *man* pages with man module.

On login you may start out with a default set of modules loaded or you may start out with an empty environment, this depends on the setup of the system you are using.

Listing currently loaded modules

You can use the module list command to see which modules you currently have loaded in your environment. If you have no modules loaded, you will see a message telling you so.

```
$ module list
No Modulefiles Currently Loaded.
```

Listing available modules

To see available modules, use module avail

```
$ module avail
------------------------------------ /usr/share/Modules/modulefiles --------------------------------
dot         module-git  module-info modules     mpt/2.16    null        perfboost   perfcatcher use.own
------------------------------------ /lustre/sw/modulefiles ----------------------------------------
abinit/8.2.3-intel17-mpt214(default)   hdf5parallel/1.10.1-intel17-mpt214     molpro/2012.1.22(default)
allinea/7.0.0(default)                  intel-cc-16/16.0.2.181                  mpt/2.14
altair-hwsolvers/13.0.213               intel-cc-16/16.0.3.210(default)         namd/2.12(default)
altair-hwsolvers/14.0.210               intel-cc-17/17.0.2.174(default)         ncl/6.4.0
amber/16                                intel-cmkl-16/16.0.2.181                nco/4.6.9
...
[removed most of the output here for clarity]
```

Loading and unloading modules

To load a software module, use module load. In this example we will use Python 3.

Initially, Python 3 is not loaded. We can test this by using the which command. which looks for programmes the same way that Bash does, so we can use it to tell us where a particular piece of software is stored

```
$ which python3
```

We can load the python3 command with module load:

```
$ module load anaconda/python3
$ which python3
```

Let's take a closer look at the gcc module. *GCC* is an extremely widely used C/C++/Fortran compiler. Lots of software is dependent on the GCC version and might not compile or run if the wrong version is loaded. In this case, there are three different versions: *gcc/6.2.0*, *gcc/6.3.0* and *gcc/7.2.0*. How do we load each copy and which copy is the default? In this case, *gcc/6.2.0* has a *(default)* next to it. This indicates that it is the default - if we type *module load gcc*, this is the copy that will be loaded.

```
[remote]$ module load gcc
[remote]$ gcc --version
gcc (GCC) 6.2.0
Copyright (C) 2016 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

So how do we load the non-default copy of a software package? In this case, the only change we need to make is be more specific about the module we are loading. There are three GCC modules: gcc/6.2.0, gcc/6.3.0 and gcc/7.2.0 To load a non-default module, we need to make add the version number after the / in our module load command

```
[remote]$ module load gcc/7.2.0
gcc/7.2.0(17):ERROR:150: Module 'gcc/7.2.0' conflicts with the currently loaded module(s) 'gcc/6.2.0'
gcc/7.2.0(17):ERROR:102: Tcl command execution failed: conflict gcc
```

What happened? The module command is telling us that we cannot have two GCC modules loaded at the same time as this could cause confusion about which version you are using. We need to remove the default version before we load the new version.

```
[remote]$ module unload gcc
[remote]$ module load gcc/7.2.0
[remote]$ gcc --version
gcc (GCC) 7.2.0
Copyright (C) 2017 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

We now have successfully switched from GCC 6.2.0 to GCC 7.2.0.

As switching between different versions of the same module is often used you can use module swap rather than unloading one version before loading another. The equivalent of the steps above would be:

```
[remote]$ module purge
[remote]$ module load gcc
[remote]$ gcc --version
[remote]$ module swap gcc gcc/7.2.0
[remote]$ gcc --version
gcc (GCC) 6.2.0
Copyright (C) 2016 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.


gcc (GCC) 7.2.0
Copyright (C) 2017 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

4.2.7 Using resources effectively

We now know most of the basic mechanics of getting research up and running on an HPC system. We can log on, submit different types of jobs, use preinstalled software, and install and use software of our own. What we need to do now is understand how we can use the systems effectively.

Estimating required resources using the scheduler

Although we covered requesting resources from the scheduler earlier, how do we know how much and what type of resources we will need in the first place?

Answer: we don't. Not until we've tried it ourselves at least once. We'll need to benchmark our job and experiment with it before we know how much it needs in the way of resources.

The most effective way of figuring out how much resources a job needs is to submit a test job, and then ask the scheduler how many resources it used. A good rule of thumb is to ask the scheduler for more time than your job can use. This value is typically two to three times what you think your job will need.

4.2.8 Using shared resources responsibly

One of the major differences between using remote HPC resources and your own system (e.g. your laptop) is that they are a shared resource. How many users the resource is shared between at any one time varies from system to system, but it is unlikely you will ever be the only user logged into or using such a system. We have already mentioned one of the consequences of this shared nature of the resources: the scheduling system where you submit your jobs, but there are other things you need to consider in order to be a considerate HPC citizen, to protect your critical data and to transfer data.

Be kind to the login nodes

The login node is often very busy managing lots of users logged in, creating and editing files and compiling software! It doesn't have any extra space to run computational work. Don't run jobs on the login node (though quick tests are generally fine). A "quick test" is generally anything that uses less than 4GB of memory, 4 CPUs, and 10 minutes of time. If you use too much resource then other users on the login node will start to be affected - their login sessions will start to run slowly and may even freeze or hang.

Login nodes are a shared resource

Remember, the login node is shared with all other users and your actions could cause issues for other people. Think carefully about the potential implications of issuing commands that may use large amounts of resource.

Test before scaling

Remember that you are generally charged for usage on shared systems. A simple mistake in a job script can end up costing a large amount of resource budget. Imagine a job script with a mistake that makes it sit doing nothing for 24 hours on 1000 cores or one where you have requested 2000 cores by mistake and only use 100 of them! This problem can be compounded when people write scripts that automate job submission (for example, when running the same calculation or analysis over lots of different input). When this happens it hurts both you (as you waste lots of charged resource) and other users (who are blocked from accessing the idle compute nodes).

On very busy resources you may wait many days in a queue for your job to fail within 10 seconds of starting due to a trivial typo in the job script. This is extremely frustrating! Most systems provide small, short queues for testing that have short wait times to help you avoid this issue.

Test job submission scripts that use large amounts of resource

Before submitting a large run of jobs, submit one as a test first to make sure everything works as expected. Before submitting a very large or very long job submit a short trunctated test to ensure that the job starts as expected

Have a backup plan

Although many HPC systems keep backups, it does not always cover all the file systems available and may only be for disaster recovery purposes (*i.e.* for restoring the whole file system if lost rather than an individual file or directory you have deleted by mistake). Your data on the system is primarily your responsibility and you should ensure you have secure copies of data that are critical to your work.

Version control systems (such as Git) often have free, cloud-based offerings (e.g. Github, Gitlab) that are generally used for storing source code. Even if you are not writing your own programmes, these can be very useful for storing job scripts, analysis scripts and small input files. For larger amounts of data, you should make sure you have a robust system in place for taking copies of critical data off the HPC system wherever possible to backed-up storage. Tools such as rsync can be very useful for this.

Your access to the shared HPC system will generally be time-limited so you should ensure you have a plan for transferring your data off the system before your access finishes. The time required to transfer large amounts of data should not be underestimated and you should ensure you have planned for this early enough (ideally, before you even start using the system for your research).

In all these cases, the helpdesk of the system you are using should be able to provide useful guidance on your options for data transfer for the volumes of data you will be using.

Your data is your responsibility

Make sure you understand what the backup policy is on the file systems on the system you are using and what implications this has for your work if you lose your data on the system. Plan your backups of critical data and how you will transfer data off the system throughout the project.

Transferring data

As mentioned above, many users run into the challenge of transferring large amounts of data off HPC systems at some point (this is more often in transferring data off than onto systems but the advice below applies in either case). Data transfer speed may be limited by many different factors so the best data transfer mechanism to use depends on the type of data being transferred and where the data is going. Some of the key issues to be aware of are:

- **Disk speed** - File systems on HPC systems are often highly parallel, consisting of a very large number of high performance disk drives. This allows them to support a very high data bandwidth. Unless the remote system has a similar parallel file system you may find your transfer speed limited by disk performance at that end.

- **Meta-data performance** - *Meta-data operations* such as opening and closing files or listing the owner or size of a file are much less parallel than read/write operations. If your data consists of a very large number of small files you may find your transfer speed is limited by meta-data operations. Meta-data operations performed by other users of the system can also interact strongly with those you perform so reducing the number of such operations you use (by combining multiple files into a single file) may reduce variability in your transfer rates and increase transfer speeds.
- **Network speed** - Data transfer performance can be limited by network speed. More importantly it is limited by the slowest section of the network between source and destination. If you are transferring to your laptop/workstation, this is likely to be its connection (either via LAN or wifi).
- **Firewall speed** - Most modern networks are protected by some form of firewall that filters out malicious traffic. This filtering has some overhead and can result in a reduction in data transfer performance. The needs of a general purpose network that hosts email/web-servers and desktop machines are quite different from a research network that needs to support high volume data transfers. If you are trying to transfer data to or from a host on a general purpose network you may find the firewall for that network will limit the transfer rate you can achieve.

As mentioned above, if you have related data that consists of a large number of small files it is strongly recommended to pack the files into a larger *archive* file for long term storage and transfer. A single large file makes more efficient use of the file system and is easier to move, copy and transfer because significantly fewer meta-data operations are required. Archive files can be created using tools like tar, cpio and zip. We are going to look at tar as it is the most commonly used.

The tar command packs files into a “Tape ARchive” format intended for backup purposes. To create a compressed archive file from a directory you can use:

```
$ tar -cvWf mydata.tar mydata
```

(In bash, rather than include all options with their own - indicator, we can string them together so the above tar command is equivalent to tar -c -v -W -f mydata.tar mydata.) The second step *compresses* the archive to reduce its size.

The options we used for tar are:

- -c - Create new archive
- -v - Verbose (print what you are doing!)
- -W - Verify integrity of created archive
- -f mydata.tar - Create the archive in file *mydata.tar*

To extract files from a tar file, the option -x is used. If the tar file has also been compressed using gzip (as we did above) tar will automatically uncompress the archive. For example:

```
$ tar -xvf mydata.tar.gz
```

Consider the best way to transfer data

If you are transferring large amounts of data you will need to think about what may affect your transfer performance. It is always useful to run some tests that you can use to extrapolate how long it will take to transfer your data. If you have many files, it is best to combine them into an archive file before you transfer them using a tool such as tar.

MODULE 5
Complex [advanced] exercises and solutions to practical problems by using HPC

Advanced

Aims of this Module:

This module aims to describe / explain:

- Parallel programming, OpenMP, MPI and use of some programming frames.
- Advanced data management and data formats used on HPC.
- Some GPU programing with CUDA.

Prerequisites:

Self-study of Module 1, 2, 3 and 4

Learning Outcomes:

- Get experience with parallel programming, OpenMP, MPI, use of some programming frames.
- Be able to see some advanced data management and data formats used on HPC.
- Get basic knowledge of GPU programming with CUDA.

Module 5

- ✓ Basics
- ✓ Serial and Parallel applications
- ✓ Types of parallelism
- ✓ Message passing interface
- ✓ CUDA

5 Introduction to complex exercises

Parallel programming and the design of efficient parallel programmes is well established in high performance, scientific computing for many years. The simulation of scientific problems is an important area in natural and engineering sciences of growing importance. More precise simulations or the simulation of larger problems need greater and greater computing power and memory space. In the last decades, high-performance research included new developments in parallel hardware and software technologies, and a steady progress in parallel HPC can be observed. Popular examples are simulations of weather forecasting based on complex mathematical models involving partial differential equations; or crash simulations from the car industry based on finite element methods. Other examples include drug design and computer graphics applications for film and advertising industries. Depending on the specific application, computer simulation is the main method to obtain the desired result or it is used to replace or enhance physical experiments. A typical example for the first application area is weather forecasting where the future development in the atmosphere has to be predicted, which can only be obtained by simulations. In the second application area, computer simulations are used to obtain results that are more precise than results from practical experiments or that can be performed with less financial effort. An example is the use of simulations to determine the air resistance of vehicles. Compared to a classical wind tunnel experiment, a computer simulation can get more precise results, because the relative movement of the vehicle in relation to the ground can be included in the simulation. This is not possible in the wind tunnel, since the vehicle cannot be moved. Crash tests of vehicles are an obvious example where computer simulations can be performed with less financial effort. Computer simulations often require a large computational effort. A low performance of the computer system used can restrict the simulations and the accuracy of the results obtained significantly.

In particular, using a high performance system allows larger simulations which lead to better results. Therefore, parallel computers have often been used to perform computer simulations. Today, cluster systems built up from server nodes are widely available and are now often used for parallel simulations. To use parallel computers or cluster systems, the computations to be performed must be partitioned into several parts which are assigned to the parallel resources for execution. These computation parts should be independent of each other, and the algorithms performed must provide enough independent computations to be suitable for a parallel execution. This is normally the case for scientific simulations. To obtain parallel programming, the algorithms must be formulated in a suitable programming language. Parallel execution is often controlled by specific runtime libraries or compiler directives which are added to a standard programming language.in hardware technology. Major chip manufacturers have started to produce processors with several power-efficient computing units on one chip, which have an independent control and can access the same memory concurrently. Normally, the term core is used for single computing units and the term multicore is used for the entire processor having several cores. Thus, using multicore processors makes each desktop computer a small parallel system. The technological development toward multicore processors was forced by physical reasons, since the clock speed of chips with more and more transistors cannot be increased at the previous rate without overheating. Multicore architectures in the form of single multicore processors, shared memory systems of several multicore processors, or clusters of multicore processors with a hierarchical interconnection network will have a large impact on software development. In 2012, quad-core processors became the standard for normal desktop computers and 10 and 12-core chips are already available for use in high end systems. Another trend in parallel computing is the use of GPUs (Graphic Processing Units) for compute-intensive applications. GPU architectures provide many hundreds of specialised processing cores that can perform computations in parallel.

5.1 Basics

A first step in parallel programming is the design of a parallel algorithm or programme for a given application problem. The design starts with the decomposition of the computations of an application into several parts, called **tasks**, which can be computed in parallel on the cores or processors of the parallel hardware. The decomposition into tasks can be complicated and laborious, since there are usually many different possibilities of decomposition for the same application algorithm. The size of tasks (e.g., in terms of the number of instructions) is called **granularity** and there is typically the possibility of choosing tasks of different sizes. Defining the tasks of an application appropriately is one of the main intellectual work in the development of a parallel programme which is difficult to automate. The **potential parallelism** is an inherent property of an application algorithm and influences how an application can be split into tasks. The tasks of an application are coded in a parallel programming language or environment and are assigned to **processes** or **threads**, which are then assigned to physical computation units for execution. The assignment of tasks to processes or threads is called **scheduling** and fixes the order in which the tasks are executed. Scheduling can be done by hand in the source code or by the programming environment at compile time or dynamically at runtime. The assignment of processes or threads onto the physical units, processors or cores, is called **mapping** and is usually done by the runtime system but can sometimes be influenced by the programmer. The tasks of an application algorithm can be independent but can also depend on each other resulting in data or control dependencies of tasks. Data and control dependencies may require a specific execution order of the parallel tasks. If a task needs data produced by another task, the execution of the first task can start only after the other task has actually produced these data and provides the information. Thus, dependencies between tasks are constraints for the scheduling. In addition, parallel programmes need **synchronisation** and coordination of threads and processes in order to execute correctly. The methods of synchronisation and coordination in parallel computing are strongly connected with the way in which information is exchanged between processes or threads, and this depends on the memory organisation of the hardware. A coarse classification of the memory organisation distinguishes between **shared memory** machines and **distributed memory** machines. Often, the term thread is connected with shared memory and the term process is connected with distributed memory. For shared memory machines, a global shared memory stores the data of an application and can be accessed by all processors or cores of the hardware systems. Information exchange between threads is done by shared variables written by one thread and read by another thread. The correct behavior of the entire programme has to be achieved by synchronisation between threads, so that the access to shared data is coordinated, i.e., a thread reads a data element not before the write operation by another thread storing the data element has been finalised. Depending on the programming language or environment, synchronisation is done by the run-time system or by the programmer. For distributed memory machines, there exists a private memory for each processor, which can only be accessed by this processor and no synchronisation for memory access is needed. Information exchange is done by sending data from one processor to another processor via an interconnection network by **explicit communication** operations.

Exercise:

The Edinburgh Parallel Computing Centre (EPCC) has prepared an excellent parallel application example that we will use to learn more about parallel execution.

Log in to your HPC system and download the files along:

```
wget https://github.com/EPCCed/hpc-intro/tree/gh-pages/files/parallel_files.tar.gz
```

This is a compressed file. Try to uncompress it using the linux tool `tar`

Compiling codes

All of these short example codes are written in Fortran. They need to be **compiled** to an executable file so that we can run them on Cirrus. To do so, we need to load a compiler module:

```
module load gcc/6.2.0
```

and to compile the `serial_pi.f90` code we would do:

```
gfortran -o serial_pi.x serial_pi.f90
```

This will create a new file, `serial_pi.x` which you can execute.

Create and submit a job script to run the serial_pi.x code and examine the result.

5.2 Serial and Parallel applications

The basic idea of parallel computing is simple to understand: we divide our job into a number of tasks that can be executed at the same time so that we finish the job in a fraction of the time that it would have taken if the tasks were executed one by one. With serial computing, tasks are completed one at a time after each other.

Implementing parallel computations however, is not always easy or possible. Some HPC users will be able to write their own code and make use of one or more **parallel frameworks** to enable their code to run in parallel; however other HPC users will be using codes and applications written by other people and are dependent on the choices made by those application developers.

Even if the code or application you want to use is a **serial** application, HPC will often offer you some alternatives to running on your desktop PC such as access to nodes with more memory or even the opportunity to run many analyses concurrently (this is another form of parallelism).

Consider the following analogy:

Suppose that we want to paint the four walls in a room. We'll call this the problem. We can divide our problem into 4 different tasks: paint each of the walls. In principle, our 4 tasks are independent from each other in the sense that we don't need to finish one to start another. We say that we have 4 concurrent tasks; the tasks can be executed within the same time frame. However, this does not mean that the tasks can be executed simultaneously or in parallel. It all depends on the amount of resources that we have for the tasks. If there is only one painter, the painter could work for a while on one wall, then start painting another one, then work for a little bit on the third one, and so on. The tasks are being executed concurrently but not in parallel. If we have two painters for the job, then more parallelism can be introduced. Four painters could execute the tasks truly in parallel.

5.3 Types of parallelism

As we said before, there are basically two forms of parallelism:

Shared Memory Parallelism: Tasks share the resources on a compute node and communicate between themselves by reading and writing data to shared memory. Common terms to look for are *OpenMP* and *Threading* to help identify this type of parallelism. Using this form of parallelism, applications are (usually) limited to the number of cores and the amount of memory on a single compute node.

Distributed Memory Parallelism: In this mode of parallelism, tasks can be distributed onto compute cores all over the cluster; these cores do not necessarily need to be on the same compute node. The tasks (and cores) communicate with each other using a special protocol over a high-speed network (*an interconnect*). The typical term you will see that identifies this type of parallelism is *MPI*.

5.4 Message Passing Interface

The message-passing programming model is based on the abstraction of a parallel computer with a distributed address space where each processor has a local memory to which it has exclusive access. There is no global memory. Data exchange must be performed by message passing: to transfer data from the local memory of one processor A to the local memory of another processor B, processor A must send a message containing the data to B, and B must receive the data in a buffer in its local memory. To guarantee portability of programmes, no assumptions on the topology of the interconnection network is made. Instead, it is assumed that each processor can send a message to any other processor. A message-passing programme is executed by a set of processes where each process has its own local data. Usually, one process is executed on one processor or core of the execution platform. The number of processes is often fixed when starting the programme. Each process can access its local data and can exchange information and data with other processes by sending and receiving messages. In principle, each of the processes could execute a different programme (MPMD, multiple program multiple data). But to make programme design easier, it is usually assumed that each of the processes executes the same programme (SPMD, single program, multiple data). In practice, this is not really a restriction, since each process can still execute different parts of the programme, selected, for example, by its process rank. The processes executing a message-passing programme can exchange local data by using communication operations. These could be provided by a communication library. To activate a specific communication operation, the participating processes call the corresponding communication function provided by the library. In the simplest case, this could be a point-to-point transfer of data from a process A to a process B. In this case, A calls a send operation, and B calls a corresponding receive operation. Communication libraries often provide a large set of communication functions to support different point-to-point transfers and also global communication operations like broadcast in which more than two processes are involved for a typical set of global communication operations. The most popular one is MPI.

The message-passing interface (MPI) is a standardisation of a message-passing library interface specification. MPI defines the syntax and semantics of library routines for standard communication patterns. MPI is an interface specification for the syntax and semantics of communication operations, but leaves the details of the implementation open. Thus, different MPI libraries can use different implementations, possibly using specific optimisations for specific hardware platforms. For the programmer, MPI provides a standard interface, thus ensuring the portability of MPI programmes. Freely available MPI libraries are MPICH (seehttp://www.mpich.org) and OpenMPI (see http://www.open-mpi.org).

A MPI programme consists of a collection of processes that can exchange messages. Normally, each processor of a parallel system executes one MPI process, and the number of MPI processes started should be adapted to the number of processors that are available. In principle, each process can read and write data from/into files. For a coordinated I/O behavior, it is essential that only one specific process performs the input or output operations. To support portability, MPI programmes should be written for an arbitrary number of processes. The actual number of processes used for a specific programme execution is set when starting the programme.

On many parallel systems, an MPI programme can be started from the command line. The following two commands are common or widely used:

```
mpiexec -n 4 programname programarguments
mpirun -np 4 programname programarguments
```

This call starts the MPI programme name with p = 4 processes. The specific command to start an MPI programme on a parallel system can differ. A significant part of the operations provided by MPI are operations for the exchange of data between processes. The official description of the MPI standard provides many more details that cannot be covered in our short description. Most examples given in this chapter are taken from these sources. Before describing the individual MPI operations, we first introduce some semantic terms that are used for the description of MPI operations:

• **blocking operation**: An MPI communication operation is blocking, if return of control to the calling process indicates that all resources, such as buffers, specified in the call can be reused, e.g., for other operations. In particular, all state transitions initiated by a blocking operation are completed before control returns to the calling process.

• **non-blocking operation**: An MPI communication operation is non-blocking, if the corresponding call may return before all effects of the operation are completed, and before the resources used by the call can be reused. Thus, a call of a non-blocking operation only starts the operation. The operation itself is completed not before all state transitions caused are completed and the resources specified can be reused.

The terms blocking and non-blocking describe the behaviour of operations from the local view of the executing process, without taking the effects on other processes into account. But it is also useful to consider the effect of communication operations from a global viewpoint. In this context, it is reasonable to distinguish between synchronous and asynchronous communication:

• **synchronous communication**: The communication between a sending process and a receiving process is performed such that the communication operation does not complete before both processes have started their communication operation. This means in particular that the completion of a synchronous send indicates not only that the send buffer can be reused, but also that the receiving process has started the execution of the corresponding receive operation.

• **asynchronous communication**: Using asynchronous communication, the sender can execute its communication operation without any coordination with the receiving process.

Exercise:
https://mpi4py.readthedocs.io/en/stable/tutorial.html

5.5 CUDA (Compute Unified Device Architecture)

Manycore GPUs (Graphics Processing Units) are available in almost all current hardware platforms, from standard desktops to computer clusters and thus, provide easily accessible and low cost parallel hardware to a broad community. Originally, these processors have been designed for graphics applications. However, today there is an increasing importance for applications from scientific computing and scientific simulations. Especially for data parallel programmes there can be a considerable increase of efficiency. This efficiency improvement is mainly caused by the specific hardware design of GPUs, which has been optimised for large data of graphics applications and high throughput of floating-point operations to be executed by a large number of threads. Nowadays, GPUs may comprise many hundreds up to thousands of cores executing these threads. At the beginning, the amount of effort required to use GPUs for general non-graphics applications and simulations was extremely high due to programming environments, such as DirectX or OpenGL, designed for graphics applications. More recent programming environments for GPUs are CUDA and OpenCL. CUDA (Compute Unified Device Architecture) is a more generic parallel programming environment supported by NVIDIA GPUs for new generations since 2007 and can also be emulated on CPUs. More detailed information is out of scope of this study material, so please consult the additional materials, but you can still try out the exercise.

Exercise:

https://documen.tician.de/pycuda/tutorial.html

References

Almeida, S. (2013). An introduction to high performance computing. *International Journal of Modern Physics A*, *28*(22n23), 1340021.

Attaran, M., & Woods, J. (2018). Cloud Computing Technology: A Viable Option for Small and Medium-Sized Businesses. *Journal of Strategic Innovation and Sustainability*, *13*(2).

Barney, B. (2014). Introduction to Parallel Computing [online]. Available from: https://computing.llnl.gov/tutorials/parallel_comp/

BIS. (2013). Department for Business, Innovation and Skills, E-infrastructure: The ecosystem for innovation. 2013, London, UK.

Brochard, L. (2006). High performance computing technology, applications and business. *Informatik-Spektrum*, *29*(3), 191-200.

Eurich, M., Calleja, P., & Boutellier, R. (2013). Business models of high performance computing centres in higher education in Europe. *Journal of computing in higher education*, *25*(3), 166-181.

European Commission (2012). High-Performance Computing: Europe's place in a Global Race [online]. Available from: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0045:FIN:EN:PDF [Accessed: 23 April 2019]

European Commission (2015). Study on "High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy" [online]. Available from: https://ec.europa.eu/digital-single-market/en/news/study-high-performance-computing-eu-progress-implementation-european-hpc-strategy-final-report

European Commission (2018). High Performance Computing - best use examples [online]. Available from: https://ec.europa.eu/digital-single-market/en/news/high-performance-computing-best-use-examples

European Investment Bank (2018). Financing the future of supercomputing: How to increase the investments in high performance computing in Europe [online]. Available from: https://www.eib.org/en/infocentre/publications/all/financing-the-future-of-supercomputing.htm

EPCC (2015). Bringing cloud-based HPC to business [online]. Available from: https://www.epcc.ed.ac.uk/sites/default/files/EPCC%20News%2077.pdf

Ezell, S. J., & Atkinson, R. D. (2016). The vital importance of high-performance computing to US competitiveness. Information Technology and Innovation Foundation, April, 28.

Foster, I., Zhao, Y., Raicu, I., & Lu, S. (2008). Cloud computing and grid computing 360degree compared. Published in: IEEE conference proceedings of Grid Computing Environments Workshop 2008: 1–10. IEEE, New York.

Holmes, V., & Newall, M. (2016). HPC and the Big Data challenge. Paper presented at the Safety and Reliability.

Hong, L., Zhong-hua, L., & Xue-bin, C. (2010). The applications and trends of high performance computing in finance. Paper presented at the 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science.

Li, L. (2015). HIGH PERFORMANCE COMPUTING APPLIED TO CLOUD COMPUTING.

Livingstone, P. (2011). High Performance Innovation [online]. Available from: https://www.rdmag.com/article/2011/02/high-performance-innovation [Accessed: 23 April 2019]

Marolt, M., Lenart, G., Maletič, D., Borštnar, M. K., & Pucihar, A. (2016). Business model innovation: Insights from a multiple case study of Slovenian SMEs. *Organizacija*, *49*(3), 161-171.

Needleman, T. (2009). SMALL BUSINESS: Accounting in the clouds
'Cloud' computing is all silver lining for small businesses (and their accountants, too)(News). *Accounting Today*, *23*(10), 1.

Spiers, B., & Wallez, D. (2010). High-performance computing on wall street. *Computer*, *43*(12), 53-59.

Sridhar, T. (2015). Cloud Computing - A Primer [online]. Available from: https://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back-issues/table-contents-46/124-cloud2.html

Sterling, T., & Stark, D. (2009). A high-performance computing forecast: Partly cloudy. *Computing in Science & Engineering*, *11*(4), 42-49.

STFC (2015). Accelerating the product discovery process at Unilever [online]. Available from: https://stfc.ukri.org/innovation/success-stories/accelerating-the-product-discovery-process-at-unilever/

Strohmaier, E., Dongarra, J. J., Meuer, H. W., & Simon, H. D. (1999). The marketplace of high-performance computing. *Parallel Computing*, *25*(13-14), 1517-1544.

Wince-Smith, D. (2009). High-Performance Computing for All. *Issues in Science and Technology*, *25*(4), 67-70.

Rauber, T. and Rünger, G. (2010). *Parallel programming*. Berlin: Springer.

Introduction to High-Performance Computing : https://epcced.github.io/hpc-intro/

U.S. Department of Energy by Lawrence Livermore National Laboratory, Parraler computing, (2019): https://computing.llnl.gov/tutorials/parallel_comp/