Developing High Performance Computing Resources for Teaching Cluster and Grid Computing courses

Violeta Holmes[1] and Ibad Kureshi[2]

[1] University of Huddersfield, Huddersfield, UK
v.holmes@hud.ac.uk
[2] Durham University, Durham, UK
ibad.kureshi@durham.ac.uk

Abstract
High-Performance Computing (HPC) and the ability to process large amounts of data are of paramount importance for UK business and economy as outlined by Rt Hon David Willetts MP at the HPC and Big Data conference in February 2014. However there is a shortage of skills and available training in HPC to prepare and expand the workforce for the HPC and Big Data research and development. Currently, HPC skills are acquired mainly by students and staff taking part in HPC-related research projects, MSc courses, and at the dedicated training centres such as Edinburgh Universitys EPCC. There are few UK universities teaching the HPC, Clusters and Grid Computing courses at the undergraduate level. To address the issue of skills shortages in the HPC it is essential to provide teaching and training as part of both postgraduate and undergraduate courses. The design and development of such courses is challenging since the technologies and software in the fields of large scale distributed systems such as Cluster, Cloud and Grid computing are undergoing continuous change. The students completing the HPC courses should be proficient in these evolving technologies and equipped with practical and theoretical skills for future jobs in this fast developing area.

In this paper we present our experience in developing the HPC, Cluster and Grid modules including a review of existing HPC courses offered at the UK universities. The topics covered in the modules are described, as well as the coursework project based on practical laboratory work. We conclude with an evaluation based on our experience over the last ten years in developing and delivering the HPC modules on the undergraduate courses, with suggestions for future work.

Keywords: HPC Course Structure, HPC and Grid Resources for Teaching, Computer Clusters, Campus grids, Cloud Computing, Advanced Teaching Materials

1 Introduction

The availability of powerful computers and high-speed networks as low-cost commodity components are changing the way computers are used. This has led to the rise of large-scale

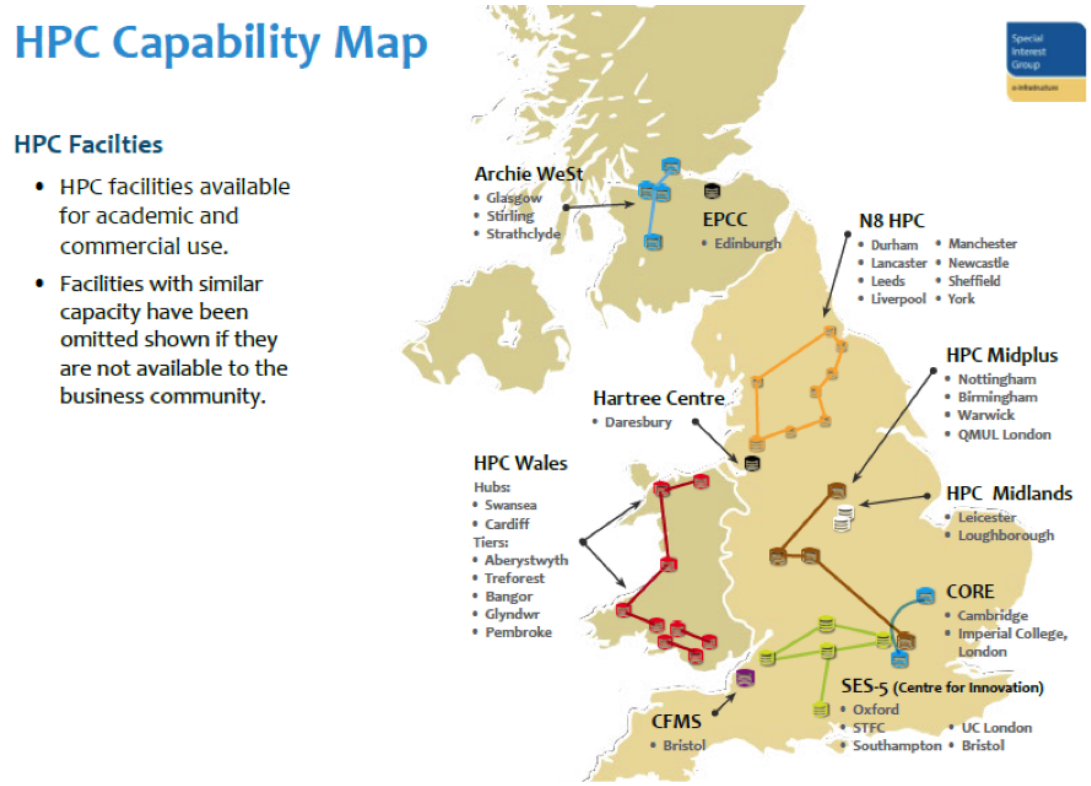

Figure 1: HPC Facilities [14]

distributed systems: Cluster computing in local area networks, Cloud computing in private and public networks, and Grid computing in wide-area networks. At the same time the technologies and software for such High-Performance Computing are continuously changing.

The HPC and Big data are new technologies vital for the advancement in science, business and industry. According to the Department for Business Innovation and Skills [13] It is estimated that the Big data market will benefit the UK economy by 216 billion and create 58000 new jobs before 2017. In its report UK Participation in HPC at the European level e-Infrastructure Leadership Council (2012) analysed the vision, policies and support for HPC at the European level and recommended the strategy the UK should take. It identified the strategic importance of HPC and e-infrastructure as drivers of economic growth and societal well-being. The report pointed out that training and advanced skills must be an integral part of postdoctoral training, but that the basic foundation must be laid at undergraduate level. In 2012 and 2013 the UK government invested extensively into the HPC systems infrastructure, spending millions of pounds in creating a Hartree Centre, STFC HPC facilities, and funded N8 Tier 2 facility encouraging collaboration between the universities at the north of England. Similar HPC centres were established in Scotland and Wales as seen in Figure 1.

To drive innovation in business and industry and support scientific research, it is not enough to invest only in the HPC systems hardware and software infrastructure, but also to invest in the education of skilled and knowledgeable workforce capable of using HPC technology and able to develop and build new HPC and Big Data infrastructures [13].

In order to bridge the gap between the need for HPC-proficient staff and the availability of skilled computing and engineering graduate courses, it is necessary to design and develop HPC courses and training as part of undergraduate, postgraduate and research programmes.

We have been working on developing the courses and modules for parallel and distributed systems since 2004 [11, 12], devised comprehensive undergraduate and postgraduate teaching and learning material for lectures, seminars, and practical laboratory exercises, and acquired dedicated computer hardware resources to support the delivery of theory and practice of HPC. The students joining the modules in HPC are expected to use prior knowledge acquired during their undergraduate studies in technologies that include computer system architecture, networking, operating systems, and programming, as the building blocks for acquiring new skills and knowledge in the HPC systems, cluster and grid middleware, and parallel programming.

The methodology for teaching and learning High-Performance Computing, Cluster, Cloud and Grid technology in the final year of our graduate course in computing or engineering is based on practical problem solving supported by hands-on laboratory work. The laboratory experiments offer necessary practical experience with Cluster and Grid middleware software required to construct, deploy and evaluate Cluster and Grid applications. To provide for such a practical, experiential approach in the delivery of High-Performance Computing on the undergraduate and postgraduate courses, we have used open-source software and developed HPC hardware resources Queensgate Clusters and Huddersfield Queensgate Campus Grid (QGG).

In designing the modules for teaching HPC at undergraduate level, in 2004 our aim was to develop a Cluster and Grid computing course for engineering and computing students who already have some knowledge in Cluster and Grid-related fields such as computer systems engineering, networking, operating systems and programming. We have used commercial-off-the-shelf (COTS) computer systems, open-source operating systems, cluster and grid middleware software, and MPI (message-passing interface) parallel programming environments to deliver theory and practical know-how in HPC technologies.

In order to support experiential problem-solving based teaching, and to promote HPC research, we have initiated a project to establish a university-wide High-Performance computing (HPC) resource starting in the School of Computing and Engineering (SCE). Working with the researchers from the School of Applied Sciences (SAS) and Art, Design and Architecture (ADA), and staff from Computing and Library services. We have deployed a number of computer clusters, and unified our HPC resources to form the Queensgate Grid QGG (Figure 2). The QGG is also linked to the external HPC resources at STFC Daresbury laboratories, and UK national e-Science resources.

The university HPC resources currently support the delivery of Computer Clusters and Grids courses/modules to both undergraduate and postgraduate students in the School of Computing and Engineering, and the SAS and SCE final year student projects requiring high processing power. The QGG is also supporting over 200 users from the schools and research institutes across the university. The work on deploying the QGG campus grid demonstrated that Higher Education institutions can satisfy the demand for the HPC resources, and prepare the students for future jobs in the area of HPC, Service Oriented Architecture, Cluster, Cloud and Grid computing and Big Data, without purchasing expensive supercomputing facilities.

In the following sections we will review the existing undergraduate and postgraduate courses in the HPC and big data at the UK universities. We will focus on the structure of the undergraduate module - Computer Clusters and Grids and describe the lecture material and laboratory activities we developed. The hardware and software resources for the module will be described in detail and the environment for the practical coursework activities. We will reflect on the methodology in presenting the HPC material and evaluate effectiveness of the module delivery

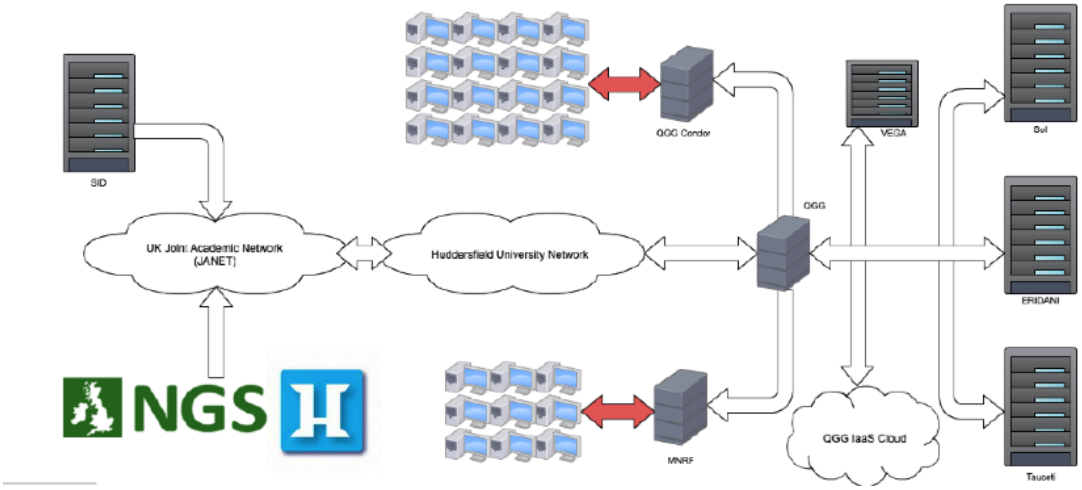

Figure 2: QGG - Queensgate Campus Grid

using as evidence student module feedback. Finally, we will assess the modules main objective educating future HPC professionals, and consider examples of successful graduates who are currently employed in HPC related industries or have continued their research work in the HPC and big data.

2 Review of the existing HPC related courses in the UK

There are a number of Masters Postgraduate courses in the UK Universities that offer modules related to High Performance computing and Big Data. Imperial College London MSc Advanced Computing Areas of Specialism/HPC offers study of methods applied to the design of software for complex, parallel systems [4]. The University of Edinburgh offers an MSc in High Performance Computing at the EPCC Training Centre and Specialist Area course for Postgraduate students in Computer Systems, Software Engineering and HPC which embraces both the theory and practice of designing programmable systems and prepares students for PhD and for careers in the software industry. The MSc course was originally funded by EPSRC and included how to write high-quality computer programs, how to speed up program execution on supercomputers, how to use Grid technologies, visualise scientific data and mathematical tools that underpin computational science and engineering [1]. Their current MSc course in HPC offers broad-based coverage of PC and parallel computing.

An MSc course in Big Data is offered at the University of Sterling, focusing on analytics, machine learning and data visualisation in Parallel and Distributed Systems [16], and MSc Big Data at the University of Liverpool is focusing on big data problem in the context of HPC [15].

An MSc course at Cranfield University combines software engineering with high performance computing, teaching tools and techniques that employers are looking for [20]. An MSc in Multi-Core computing at the University of Manchester [22] is centred around programming multi-core systems, whilst the MSc in Computer Science at the University of Lancaster [21] is focusing on design and development of computer systems in Cloud Computing, Big Data and Data Mining.

However, there are only two Universities (according to www.ucas.com) which offer under-

graduate courses with HPC: Computer Science with HPC at Cardiff University, and Mathematics with HPC at Plymouth University [19].

The above summary of HPC-related courses offered at the UK universities clearly shows that there is a need for the undergraduate courses which will focus on the skills shortages in this important strategic area for the UK economy.

Furthermore, this justifies the design and development of Cluster and Grid computing modules offered in the SCE at the University of Huddersfield. The modules teach the theoretical underpinning of the HPC technologies and skills necessary for designing, deploying and managing the HPC infrastructure, using dedicated laboratory hardware and software and the QGG campus grid resources.

3 Development of HPC resources for teaching Cluster and Grid Computing on UG and PG level

The possibility to build inexpensive super-computers from COTS components initially inspired us to design and implement an experimental HPC cluster for teaching parallel computing theory in 2004. The motivation for the Parallel Computing Architecture undergraduate module development was an observation that there was an improvement in students performance and engagement when teaching parallel computer theory in the lectures when this was supplemented with practical experience during laboratory exercises. The development of the East Lancashire Institute of HE (ELIHE) cluster enabled hands-on approach in teaching parallel computer architecture, programming environments, tools and libraries for development of parallel applications using MPI. The ELIHE cluster consisted of 9 PCs running commodity software Linux, and CLIC Mandrake cluster middleware [11, 12].

The teaching material, lectures and laboratory activities, and laboratory HPC resources were developed further during 2007 and 2009 [9] in the SCE at the University of Huddersfield. At that time the laboratory equipment for Cluster and Grid modules were not available and a decision was made to involve the students in building their own cluster devices from recycled laboratory PCs and networking equipment, using open-source Linux operating system and cluster and grid middleware. This further increased the students engagement, and inspired some of the students to repeat the experiments at home using COTS equipment and knowledge and skills gained on the course.

In 2010 the school invested in a dedicated computer cluster built from 33 PCs, providing a 132 core system for teaching Cluster and Grid UG and PG modules. Additional HPC resources were made available to students in the QGG campus grid which was developed as part of URF HPC funding, and through membership of the NGS [8, 10]. This has provided scientific and engineering HPC resources that can be used in practical laboratory work. It enabled benchmarking of different computer architectures and platforms and evaluation of their performance. This new HPC resource has provided an excellent platform for teaching HPC related subjects in real-life scenarios.

The Cluster and Grid lecture and laboratory material was updated to incorporate practical work on the new resources.

The initial objective of designing the modules with a practical problem-solving approach to teaching Cluster and Grid systems was even more important in every subsequent modules delivery.

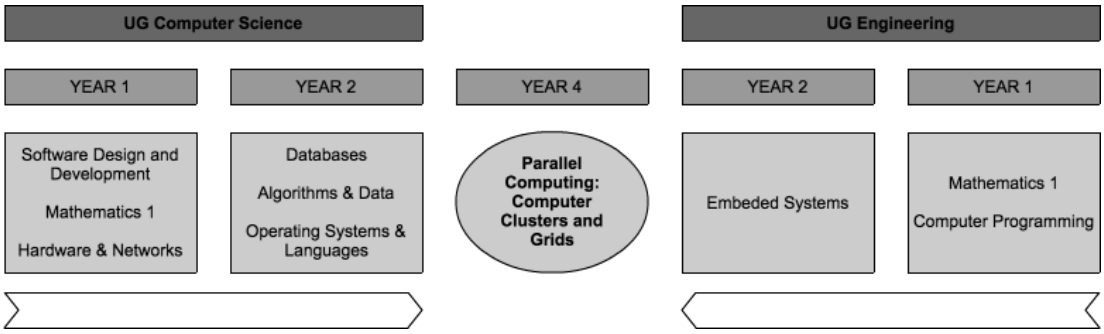

Figure 3: Prior Knowledge of Student Base

4 Undergraduate Clusters and Grid Module design

The undergraduate module in HPC Clusters and Grids, the main topic of this paper, was designed for computing and engineering students in the final year of their courses, and expected prior knowledge and skills included C/C++ programming, communications, computer systems architecture, and operating systems. Each program has a different pathway so students do not start the course with matching backgrounds (see Figure 3). The initial lectures bridge the gap between the students.

The module includes the following topics: Revision of computer systems architecture, Multi-core systems - Graphic Processing Units (GPUs), Linux operating system, Computer networking and storage, Cluster OSCAR middleware, Grid Globus toolkit, Condor middleware, Message Passing Interface (MPI), basic parallel programming using MPICH and OpenMP, speed-up and efficiency of parallel programming.

In addition, the MAUI/Torque resource management and the Portable Batch System (PBS) were used to demonstrate parallel and serial job submission to the clusters and grids. The overall delivery was focused on imparting knowledge and skill necessary to build and run HPC systems for processing large data and task-intensive scientific and engineering applications. The students were encouraged to research existing supercomputing systems as listed in top500 [5] ranking, compare and contrast them with the small scale university clusters and campus grid, and present their findings in a seminar session of the module.

The starting point for designing and developing lecture material was books HPC Linux Clusters [17] and The Grid 2: Blue Print for a New Computing Infrastructure [7] , and paper on Cluster Computing in the Classroom [2]. There are many resources available for teaching cluster and grid computing, however it is still challenging to design a module that will cover relevant topics and cater for the students varied backgrounds. It is not possible to assume that all the necessary prerequisites are covered within their previous courses. This requires an additional effort to ensure that all students have a consistent level of knowledge and understanding of Cluster and Grid technologies, and are able to attain the learning outcomes of the module.

The teaching material, lecture notes and laboratory worksheets are designed to be delivered over two 11 week terms. The lectures given in 1.5 hours slots, and the practical work in 1.5 laboratory sessions.

4.1 Lectures Topics

This section outlines the lecture topics in order they are usually presented over 24 weeks in terms 1 and 2.

Term 1 topics:

1. Introduction to different computer architectures supporting HPC; definition of Cluster, Grid and Cloud Computing
2. Review of Linux operating systems, structure and basic commands
3. Clusters Overview and Building Blocks: PCs, network interconnects, switches; Operating systems
4. Cluster middleware OSCAR;
5. Cluster Single System Image
6. MPI standard and implementations
7. MPI Parallel programming using MPICH; Programming on laboratory and university clusters
8. Measuring Performance of Parallel Computers; Amdahls law;
9. Applications of HPC
10. Workflow Management and Resource management: MAUI/Torque and PBS
11. Introduction to Graphic Processing Units: GPU clusters

Term 2 topics:

1. Introduction to Grids
2. Concepts of Grids, Virtual Organisations
3. Grid Architecture and Technologies
4. Grid Security: Data Integrity and PKI paradigm. Application level tools, Languages and Compilers
5. Globus and Condor middlewares
6. gLite, Unicore and VDT middlewares
7. Introduction to Utility Computing evolution from the Grid to Cloud computing
8. Cloud Computing
9. Cloud based Services
10. Cloud Computing Providers
11. Summary of Cluster, Grid and Cloud Technologies

4.2 Practical Laboratory Activities

The practical laboratory work is closely linked with the theory presented in the lectures, and further expanded and explored in the coursework projects.

The inexpensive and sustainable sourcing of the hardware used in the laboratory work is achieved by using re-cycled laboratory PCs, with onboard Ethernet and additional Ethernet card (for the head node), and 100KB switches. The cost of software is low since the operating system and cluster and grid middlewares are open-source (free?). In 2014/15 OS used was Linux CentOS 5.4. The cluster middleware software used was OSCAR version 5.1.

OSCAR (Open Source Cluster Application Resource)[3] is an example of computer cluster middleware. The Globus toolkit [6] is a de-facto standard for the grid middleware a fundamental enabling technology for the Grid. The HTCondor middleware for High Throughput computing [18], is enabling the use of dormant laboratory and library machines to be used to increase high performance computing facilities for scientific and engineering applications.

During practical laboratory exercises the students are working in groups of two or three.

In term 1, each group is given 3 PCs (2014/15 issue was a 3 Intel dual core machines, each with 2 GBytes of RAM and 320 GByte HD), a single 100KB switch and a selection of Ethernet cables.

The students are expected to build, install and test their laboratory cluster in the first 6 laboratory sessions.

To test and profile their clusters the students use a set of prepared MPI programs and a graphical web monitoring tool Ganglia, giving detailed reports of CPU, memory and network utilisation.

The remaining five laboratory sessions in term 1 are focused on MPI programming, such as Pi constant calculation, running on the groups laboratory clusters, and on the university clusters. The students are taught to write PBS scripts and submit the jobs to the university resources. The job submission was done from the Linux clusters, and standard laboratory Windows machines. Having an opportunity to try different cluster architectures and middlewares is essential for acquiring broad knowledge of available systems, their performance, limitations and suitability for particular applications.

In term 2, first four laboratory sessions are dedicated to practical introduction to Grid middleware Globus and Globus Requirement Specification language (RSL) for submission of parallel jobs to the QGG campus grid. The importance of Grid/Cloud security, authentication and authorisation is conveyed to the students by issuing and using students unique certificates/keys for working on the QGG.

Labs five and six in term 2 are used for practical introduction to Condor middleware, using Huddersfield University Condor pool The remaining 5 laboratory sessions are allocated for coursework project work in building the groups Condor pool, which is a practical problem solving activity relevant to campus grids.

4.3 Coursework projects

The assessment for the Clusters and Grids module was designed to reflect the considerable practical work involved in the module, such that exam and coursework contribute 50% each to the final module mark.

The tasks in the coursework projects are typically divided in the cluster and grid tasks, term1 and 2 respectively:

- The cluster building and performance testing is a term 1 task. The students are expected to evaluating a speed-up of a program execution (in a MPI environment) using multiple

cores in the laboratory OSCAR cluster they have created, and on the university clusters. This activity is designed to motivate them to compare different system architectures, identify possible bottlenecks and improvements in the systems performance.

- The Globus grid and HTCondor projects are term 2 tasks. The students are using Globus tools to connect and move files between different parts of the QGG campus grid, and submit and run application using Globus commands. Also, they are using HTCondor based tools to view resources and mange jobs on universitys HTCondor pool. Coursework task is to implement HTCondor grid middleware to create a Condor pool using their laboratory machines.

5 Discussion

The Parallel Computer Architectures, Clusters and Grids module has been offered since 2008 in SCE with the student numbers increasing every year.

The content of the lecture and laboratory material is updated every year due to the rapid changes in the Cluster, Grid and Cloud Computing, keeping up with the latest changes in the HPC technology.

Continuous development of the laboratory hardware and software resources is challenging, and there were a number of different hardware and software solutions offered since the beginning of the module development.

The main challenge is keeping up with open-source development for Linux operating system and cluster and grid middlewares, which are changing continuously. As a result of this development, previous (obsolete versions) of software cannot be used in the next year delivery of the module.

In order to keep the cost of hardware resources low, re-cycled laboratory PCs are used to source hardware for practical laboratory work, and they often vary in their system architecture, reliability and quality which can cause problems in the laboratory. Because the hardware specification of the PC and networking equipment are changing, they are not always supported by the operating system and Cluster and Grid middleware currently used in the lab.

The module is challenging for students due to a broad computing and engineering knowledge required when building laboratory clusters and grids. In addition, the open-source tools and framework for clusters and grids are buggy and hardly documented.

However, the positive outcome of this challenge is that it encourages the students to search for information from a variety of sources in order to complete Cluster and Grid middleware building tasks, developing their research skills and confidence in the process, and making them independent learners. The student feedback was consistently good over the years, and it is evident from module reports that the students enjoy practical problem-cantered learning. They are also aware of the currency of the module content and its relevance to the HPC industry.

Inspiring and motivating students to become researchers in the HPC is one of the outcomes of the delivery of this module that we are most proud of. Since it was first offered at the University of Huddersfield, at least one student per year who has completed this module has continued as Master by Research (MRes) or a PhD student in the HPC Research Group. So far there were seven successful MRes students completing their HPC research degrees, and five continued further on PhD studies in HPC-related subjects at the University of Huddersfield or elsewhere. Some of the destinations of these graduates are in the HPC system administration of research institutes and Russell Group universities.

In future the module will be updated to reflect new developments in HPC and will rely closely on the HPC research at the University of Huddersfield.

Using COTS hardware and free open-source software in teaching Clusters and Grids demonstrates that universities do not need expensive national and international supercomputer resources to delivery HPC training.

To respond to the Big Data challenge, the GPU and Hadoop clusters already developed as part of the HPC research work, and integrated into QGG campus grid will be utilised during practical laboratory work and used to develop skills and knowledge in processing and visualising Big Data, the next strategic goal of UK business, industry and academia.

References

[1] EPCC (2006-2010). Msc in high performance computing, 2010.

[2] Amy Apon, Rajkumar Buyya, Hai Jin, and Jens Mache. Cluster computing in the classroom: Topics, guidelines, and experiences. In *Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on*, pages 476–483. IEEE, 2001.

[3] Michael J Brim, Timothy G Mattson, and Stephen L Scott. Oscar: open source cluster application resources. In *Ottawa Linux Symposium*, 2001.

[4] Imperial College. High performance computing, 2014.

[5] Jack Dongarra. June 2014 | TOP500 supercomputer sites, 2014.

[6] Ian Foster. The globus toolkit for grid computing. In *Cluster Computing and the Grid, IEEE International Symposium on*, pages 2–2. IEEE Computer Society, 2001.

[7] Ian Foster and Carl Kesselman. *The Grid 2: Blueprint for a new computing infrastructure*. Elsevier, 2003.

[8] Neil Geddes. The national grid service of the uk. In *e-Science*, page 94. Citeseer, 2006.

[9] Violeta Holmes. Grid enabled e-learning. 2008.

[10] Violeta Holmes and Ibad Kureshi. Huddersfield university campus grid: Qgg of oscar clusters. In *Journal of Physics: conference series*, volume 256, page 012022. IOP Publishing, 2010.

[11] Violeta Holmes and Terence McDonough. The elihe high-performance cluster. In *Cluster Computing, 2005. IEEE International*, pages 1–1. IEEE, 2005.

[12] Violeta Holmes, Terence Mcdonough, and Stephen Pickles. The elihe high-performance cluster for parallel computing. 2007.

[13] Coveney P. McGuire S. Parchment O. Kenway, R. and M. Parsons. UK participation in high performance computing (HPC) at the european level - publications - GOV.UK, 2012.

[14] Craig Kirkwood. Uk high performance computing an overview of capability., 2014.

[15] University of Liverpool. Msc big data, 2014.

[16] University of Sterling. Msc course in big data, 2014.

[17] Joseph D Sloan. *High performance Linux clusters with OSCAR, Rocks, openMosix, and MPI*. Rodopi, 2009.

[18] Condor Team. Htcondor. *HYPERLINK http://research. cs. wisc. edu/htcondor/htc. html http://research. cs. wisc. edu/htcondor/htc. html.*

[19] Universities and Colleges Admissions Service. Ucas directory, 2014 entry, 2014.

[20] Cranfield University. Msc software engineering in technical computing, 2014.

[21] Lancaster University. Msc in computer science, 2014.

[22] Manchester University. Msc in multi-core computing, 2014.