Cloud computing is currently on everyone’s lips – it is of interest to researchers (to speed up computations), managers (to reduce costs), and also the general public (to access music everywhere). The information technology research and advisory firm Gartner Inc., for example, announced that cloud computing will be one out of ten top information technology topics in the next years. At least in 2015 cloud computing is predicted to be a mass product and will likely be the preferred solution for many application areas [1]. In view of these developments, the idea of cloud computing sounds promising for the communities of Medical Biometry and Medical Informatics as well. Possible benefits include cost reductions in hardware, maintenance and software licensing, access to latest software and hardware for everyone from any place in the world, standardized and centralized analyses platforms, and new ways of collaboration between researchers and practitioners.

Biomedical research, as many other scientific disciplines, currently experiences the increased focus on high performance computing (HPC) – mainly driven by the three aspects of 1) larger data sets and the “big data” hype, 2) increased computational requirements stemming from more sophisticated statistical methodologies, and 3) new computing environments including multi-core processors and the use of graphics processing units to do general purpose computing. On this account, the German working group for Statistical Computing of the German Region of the International Biometric Society (IBS-DR) and the German Association of Medical Informatics, Biometry and Epidemiology (GMDS) organized two workshops in the year 2011 (March 23 in Munich, and September 27 in Mainz). In the course of these meetings, whose organizing committees consisted of the authors of this editorial, leading community experts discussed how HPC technologies can be used to improve and accelerate biomedical analysis.

The present Focus Theme summarizes the results and the outcomes of the workshops. Specifically, the four papers highlight a selection of benefits and challenges for grid and cloud computing, and the general use of graphic processors in the communities of Medical Biometry and Medical Informatics. Additionally, they contain evaluation of technical solutions and business model aspects. Before going into details about the four papers, we briefly introduce common terms and relations of HPC technologies.

1. Terms and Relations

High performance computing in the sense of grid and cloud computing is a relatively young field, and no definite standard literature has emerged until now. For detailed technical introductions, we suggest the selected publications [2–7]. At any rate, due to the fast developments in these technologies it is necessary to check for the latest
publications and user manuals. The most important terms and relations in the field of grid and cloud computing include the following ones:

High-Performance Computing (HPC), which is a buzzword for efficient computing environments. HPC covers all aspects of computing that require a lot of computing power and memory. Traditionally, computing centers as the Leibnitz-Rechenzentrum (LRZ, Munich, Germany; [8]) provide access to HPC by operating supercomputers. A supercomputer is – per definition – a computer at the frontline of current processing capacity, particularly with respect to the speed of calculation. However, scientific organizations and institutes often maintain bold in-house computer technology as well. This is due to the fast developments in hardware technology, the decreasing hardware costs, and possibly induced by specific problem domains (e.g., sensitive medical data).

Common HPC computer environments are:

- Multi-core processors: a single computing component with two or more independent actual processors (called “cores”), which are the units that read and execute program instructions.
- Computer clusters: a set of loosely connected computers (today, equipped with multi-core processors) that work together. They can, in many respects, be viewed as a single system. Note that not all machines of a computer cluster have to be of the same configuration (computer model, operating system, etc.). Such clusters are usually referred to as “heterogeneous” clusters.
- Grid computing: a huge computer cluster that is more loosely coupled, more heterogeneous, and especially geographically dispersed. Generally, grids are computer resources stemming from multiple administrative domains connected to reach a common goal.
- Computing on graphics processing units: a computer cluster where the computing components are graphics processing units (GPU). Traditionally, GPUs are used as visual processing units, but today’s graphics hardware can perform complex computations as well and has less acquisition cost. One prominent manufacturer is NVIDIA offering their parallel computing architecture Compute Unified Device Architecture (CUDA).

All computing environments described above have one thing in common: serial code will not run faster on these computers than on any other “normal” computer. The power of HPC environments is in fact gained through the possibility of parallel or distributed computing. This is the concept of dividing one large problem into many smaller ones, that are then solved concurrently (“in parallel”). Well-established software tools for parallelization are MPI, OpenMP, or Pthreads. A new(er) development is the MapReduce framework for distributed computing on large data sets (introduced by Google in 2004).

In computing, virtualization is the creation of a virtual version of an operating system such as Windows or Linux. It can be viewed as part of an overall trend in information technology to centralize administrative tasks while improving scalability and overall hardware-resource utilization. Virtualization is one corner stone of the idea of utility computing. Here, the computer processing power is seen as a utility that clients can pay for only as needed. This concept culminates in the today’s hype topic cloud computing, which can also be interpreted as the next generation of grid computing.

Cloud computing is the delivery of computing as a service rather than a product. Shared resources, software, and information are provided to computers and other devices as a commercial service over the Internet. The service is provided in the form of three fundamental models: Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). IaaS is the most basic service model and offers access to computers as virtualized machines, storage, firewalls, load balancers, databases, networks, and other IT technologies. Launched in July 2002, Amazon Web Services is currently the most famous IaaS provider. Its access to computers in the cloud is called Amazon Elastic Cloud (EC2), and by using Amazon Machine Images, any kind of operating system can be started on EC2 resources. This service is available to the general public and therefore is called public cloud computing. In contrast, private clouds are infrastructures operated solely for a single organization, whether managed internally or by a third-party and whether hosted internally or externally.

Next to the hardware, software also plays an important role – it has to be adapted to parallel computing as well. In biomedical research, one commonly used software environment is the statistical programming language R [9]. Following the need for parallel computing, R (as of version 2.14) ships with a directly integrated parallel computing framework (see, for example, [10, 11] for an overview and an introduction). In addition, repositories provide access to a large number of extensions. The two most important ones are: (1) the Comprehensive R Archive Network for extensions in all kind of (statistical) directions [12]; and (2) the Biocductor project for extensions in terms of the analysis and comprehension of genomic data [13].

2. Overview

The four papers of the Focus Theme “Grid and Cloud Computing Methods in Biomedical Research” discuss applications as well as software-oriented advantages and disadvantages when using grid, cloud computing and GPUs for biomeasure analyses. The papers add up to a solid reference for researchers and departments which contemplate doing statistical computing in the cloud, on a local grid or using a supercomputer.

Bernau et al. [14] focus on the technical aspects of microarray analyses using grid and cloud computing. They investigate nested resampling procedures in the context of supervised classification of high-dimensional microarray data and different normalization methods of gene expression microarray data. Two specific projects, recently implemented at their department, are used to compare the technical realizations in the cloud to common computer cluster implementations. The authors discuss the substantially different workflows.
and show that these statistical projects can be efficiently realized on cloud resources.

Knaus et al. [15] focus on the financial aspects of grid and cloud computing. The authors discuss and compare the costs of a local cluster maintained by a department to a comparable cloud offer. They systematically list all direct costs and estimate all indirect costs. Furthermore, with the help of a simulation study they estimate the scale of costs for typical methodological research projects in biomedicine. In their case, they conclude that using a cloud is a viable option for projects with limited maturity, or may serve as a supplement for short peaks in demand.

Piotrowski et al. [16] provide a discussion of high performance computing in the cloud. Genomic analyses, for example, require high performance computing infrastructure and massive parallelization. The authors introduce their R package SPRINT and use, for example, parallel implementations of the Pearson correlation and the partitioning around medoids algorithm to identify the strengths and weaknesses of a computational cloud for HPC analyses. By submitting jobs from different locations in the world they can prove a good and scalable performance using their HPC framework in the cloud.

Pütz et al. [17] focus on genome-wide epistasis calculations. The inherently parallel structure of the problem makes the analyses a perfect candidate for massive parallelization. However, due to several data security laws and confidential patient data, the authors have to use an in-house grid solution. Therefore, they compare the classical parallelization using multiple CPU nodes to corresponding GPU counterparts. The result is a cost-effective solution with highly parallel GPUs that outperforms current cluster-based systems on a price/performance criterion.

Obviously, beside the technical and business related aspects of cloud and grid computing discussed in this Focus Theme, further aspects matter. Of particular importance, for example, are legal foundations of data analysis and, consequently, data security in the cloud. Supercomputers will be required in the future for high efficient and very large applications, but most of the small to medium sized computer cluster environments can be outsourced to the cloud [18]. Therefore, at least an uniform cloud usage law and data handling law for Europe is necessary. Recently, the European Commission took on this requirement and set up a working party that analysed all relevant issues for cloud computing service providers operating in the European Economic Area and their clients [19].

No doubt, these technologies and their usage for biomedical research are in the early stages. The current main limitations are the time and costs for uploading and storing data in the cloud as well as unclear security laws. Cloud environments have to be standardized to enable an easy connection between cloud resources and locally running software environments. However, the four papers of the Focus Theme prove that cloud computing is a reasonable alternative to an in-house computer cluster. The performance differences between cloud and local environments are small, and without huge personal extra work, research code can be ported into the cloud. In summary, all important technologies are available and ready to use.

Acknowledgment

We thank the working group “Statistical Computing” of the IBS-DR and the GMDS, the organization committees of the workshops and the anonymous reviewers for improving the Focus Theme articles.

References


