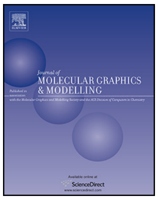
Journal of Molecular Graphics and Modelling 44 (2013) 177–187

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Topical perspectives

# The emerging role of cloud computing in molecular modelling

Jean-Paul Ebejer[a,b,1], Simone Fulle[a,1], Garrett M. Morris[a,c], Paul W. Finn[a,*]

[a] InhibOx Ltd., Oxford Centre for Innovation, New Road, Oxford OX1 1BY, UK
[b] Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
[c] Crysalin Ltd., Cherwell Innovation Centre, 77 Heyford Park, Upper Heyford, Oxfordshire OX25 5HD, UK

ARTICLE INFO

*Article history:*
Received 23 March 2013
Received in revised form 5 June 2013
Accepted 7 June 2013
Available online 17 June 2013

*Keywords:*
Cloud computing
Molecular modelling
Cheminformatics
Computer aided-drug design
Virtual screening
Docking

ABSTRACT

There is a growing recognition of the importance of cloud computing for large-scale and data-intensive applications. The distinguishing features of cloud computing and their relationship to other distributed computing paradigms are described, as are the strengths and weaknesses of the approach. We review the use made to date of cloud computing for molecular modelling projects and the availability of front ends for molecular modelling applications. Although the use of cloud computing technologies for molecular modelling is still in its infancy, we demonstrate its potential by presenting several case studies. Rapid growth can be expected as more applications become available and costs continue to fall; cloud computing can make a major contribution not just in terms of the availability of on-demand computing power, but could also spur innovation in the development of novel approaches that utilize that capacity in more effective ways.

## 1. What is cloud computing?

Cloud computing has emerged relatively recently as an approach to providing on-demand large scale computational infrastructure (according to Google Trends [1] interest starts to register around 2007). The specific characteristics of cloud computing, and the features that distinguish it from other high-performance and distributed computing approaches are, however, nebulous. Perhaps the only agreement on defining the term ‘cloud computing’ is the difficulty in finding common ground amongst the plethora of technical terms and marketing buzzwords. To generate a consensus view of the key concepts of cloud computing we have analyzed seven recent definitions (Table 1) and generated a “word cloud” from these (Fig. 1).

The first definition from Table 1 comes from the work of Mell and Grance [2] and is a popular working definition of cloud computing from the National Institute of Standards and Technology, U.S. Department of Commerce. Their definition focuses on computing resources which can be accessed from anywhere and may be provisioned online (as opposed to requiring some lengthy cloud vendor involvement). It also specifies five characteristics of cloud computing (i.e. on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service), three service models (i.e. Software as a Service, Platform as a Service and Infrastructure as a Service – explained later) and four deployment methods (i.e. private cloud, community cloud, public cloud and hybrid cloud). Most of the other definitions do not mention deployment methods. In contrast to other definitions, this one does not explicitly mention virtualization as a key technology.

Vaquero et al. [3] collected 22 excerpts from previous works and fused these into a single definition by studying the common properties of cloud computing. This emphasizes the importance of Service Level Agreements (SLA) in order to increase confidence in the cloud environment and defines virtualization as the key enabler of cloud computing. Both these definitions [2,3] mention a pay-per-use business model for cloud computing.

Marston et al. [4] offer a definition which is mostly covered by the previous ones (with the exception that they explicitly state that cloud computing lowers costs by offsetting capital costs with operational ones); but throughout their article they give an interesting, and different, business perspective to cloud computing which is missing in almost all the other related publications.

Buyya et al. [5] supply a definition with the vision that cloud computing will become the fifth utility and that cloud computing is the next generation data centre. This definition also extensively highlights the importance of SLAs and the lack of market oriented resource management (e.g. violation of SLAs, automatic allocation of resources to fulfil SLAs, etc.) in cloud environments.

[*] Corresponding author. Tel.: +44 1865261468.
*E-mail address:* paul.finn@inhibox.com (P.W. Finn).
[1] Both authors contributed equally to this work.

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.06.002

**Table 1**
Selected definitions for cloud computing.

| Author | Definition | Ref. |
|---|---|---|
| Mell et al. (2011) | "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." | [2] |
| Vaquero et al. (2009) | "Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically re-configured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLAs." | [3] |
| Marston et al. (2011) | "It is an information technology service model where computing services (both hardware and software) are delivered on-demand to customers over a network in a self-service fashion, independent of device and location. The resources required to provide the requisite quality-of service levels are shared, dynamically scalable, rapidly provisioned, virtualized and released with minimal service provider interaction. Users pay for the service as an operating expense without incurring any significant initial capital expenditure, with the cloud services employing a metering system that divides the computing resource in appropriate blocks." | [4] |
| Buyya et al. (2009) | "A cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers." | [5] |
| "Intel" (2010) | "Cloud computing is an evolution in which IT consumption and delivery are made available in a self–service fashion via the Internet or internal network, with a flexible pay-as-you-go business model and requires a highly efficient and scalable architecture. In a cloud computing architecture, services and data reside in shared, dynamically scalable resource pools, often virtualized. Those services and data are accessible by any authenticated device over the Internet. The key attributes that distinguish cloud computing from conventional computing are:<br>• Compute and storage functions are abstracted and offered as services<br>• Services are built on a massively scalable infrastructure<br>• Services are delivered on demand through dynamic, flexibly configurable resources<br>• Services are easily purchased and billed by consumption<br>• Resources are shared among multiple users (multi-tenancy)<br>• Services are accessible over the Internet or internal network by any device" | [6] |
| Grossman (2009) | "Clouds, or clusters of distributed computers, provide on-demand resources and services over a network, usually the internet, with the scale and reliability of a data centre." | [7] |
| Hill et al. (2013) | "Cloud computing is a means by which computational power, storage, collaboration infrastructure, business processes and applications can be delivered as a utility, that is, a service or collection of services that meet your demands. Since services offered by cloud are akin to a utility, it also means that you only pay for what you use." | [8] |

The fifth definition is taken from "Intel's Vision of the Ongoing Shift to Cloud Computing" [6]. It offers a particular perspective on cloud computing, from the vantage point of a major hardware manufacturer with a commercial interest in building clouds. It describes three main aspects that help realize cloud computing; federation (ease of moving data and services within the cloud), automation (cloud services should be provisioned with no human intervention) and client-awareness (cloud-based applications can take advantage of the capabilities of the end-point device). The definition also makes explicit reference to security; only authenticated devices may connect to the cloud environment.

Grossman [7] offers a general, working definition of cloud computing, which highlights that cloud computing environments should behave like local data centres, while Hill et al. [8] give a recent definition of cloud computing; which also presents cloud computing as a utility.

Thus we can summarize the main features of cloud computing:

*Service-oriented*: Computing services such as computing cycles, data storage, and networking are all offered as end-products to customers. This also implies that the administration and maintenance of these devices (e.g. faulty disks) fall under the responsibility of the cloud vendor. Under some definitions, cloud computing is offered as a utility where usage is metered and customers are charged for what they consume in a similar way to water, electricity or telephony.

*Massive scalability*: the cloud solution allows for flexible or 'elastic' allocation of resources where the limiting factor is typically cost

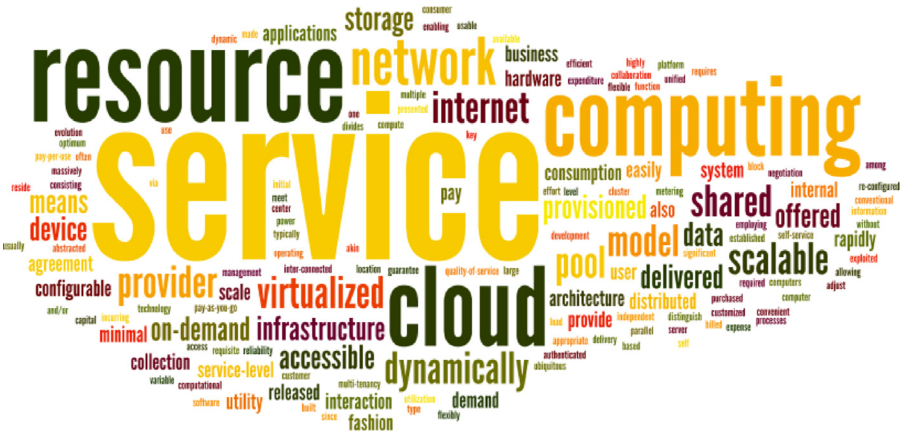


**Fig. 1.** Word cloud based on definitions for cloud computing. The word cloud was built based on the definitions in Table 1 after removal of the most common English words. The composition shows more common words in a larger font, and highlights the most common themes; service, resource, computing, cloud, and network.

as opposed to availability of computing resources. This flexibility can help manage spikes in computational requirements in an efficient way. For example, if the goal is to generate 50 conformers instead of 10 for each molecule in a virtual library, this can be achieved in roughly the same time by allocating five times more computing power. The massive scalability holds true for embarrassingly parallel tasks such as the example given above but is limited for applications where a vast amount of inter-processor communication are required (e.g. the exchange of atomic positions and inter-atomic forces in molecular dynamics (MD) simulations).

*On-demand resources (extensibility)*: Related to scalability, a distinguishing feature of cloud computing is that the resources can be added in real time, sometimes in an automated fashion (e.g. through a process that monitors CPU load and automatically adds more CPU resources once a usage threshold is reached).

*Virtualization*: Computer hardware is shared amongst multiple users in a common resource pool. Virtualization technology allows, thereby, for the abstraction of physical servers and of storage devices. This is, in particular, advantageous when using heterogeneous computing resources which need to be presented as one single platform or when different operating systems are to be run on the same hardware.

*Cloud API (Application Programming Interface)*: this is missing from the definitions in Table 1 but it is one which is becoming increasingly important. In order to avoid cloud vendor lock-in, some projects such as Apache Deltacloud [9], have abstracted from the different APIs of each cloud vendor to offer a “meta-API” that provides a unified interface, lowering the barrier for switching between cloud vendors. Also, a large number of cloud vendors are offering APIs where anything which can be done using a graphical user interface may also be done programmatically. This is the way forward for building robust tools and libraries which are not tightly coupled to a particular vendor in the cloud ecosystem.

These key properties of cloud computing help to distinguish it from two other, longer established, distributed computing paradigms for providing large scale computational resources, namely high performance computing and grid computing.

## 2. Differences between cloud, high performance, and grid computing

High performance computing (HPC), sometimes also referred to as “cluster computing”, is provided by a set of homogeneous machines which are connected locally by a dedicated low-latency high speed network such as Infiniband or Myrinet. Thus, HPC typically has better throughput (or “capability”) than either cloud or grid computing (Fig. 2). In contrast to cloud computing, HPC suffers from a fixed, or at best limited, extensibility (or “capacity”) [10]. HPC clusters are offered by some cloud computing vendors but are rare; even Amazon’s “cluster compute” instances do not have low-latency interconnect. In this discussion, the term HPC will refer to “traditional” HPC, i.e. a locally owned cluster with private access in which the processors are usually connected by low-latency high speed interconnect.

Grid computing infrastructure, on the other hand, is composed of a heterogeneous network of loosely coupled computers acting in concert to perform large tasks [11]. The key difference between grid computing and HPC is that in grid computing there is no requirement for homogeneity of computing devices, which may be spread over a geographically diverse area connected via different networks. Each node on the grid may be administered by different entities, but they typically have a common software component (middleware), to communicate the results of each individual computing node to a central location. An example of such middleware is MapReduce [12].

As can be seen, grid and cloud computing share many similarities, but the key difference is that in cloud computing the hardware is virtualized. This can allow for security by isolation as long as different virtual machines running on the same hardware are configured such that they cannot directly access each other’s files, even if they reside on the same physical device. Furthermore, cloud computing has usually an inherent business model (the sale of computing power). Improved usability is also another prevailing property of cloud computing compared to grid computing. More detailed differences between cloud and grid computing can be found in [10,13,14].

In principle, compute-intensive problems could be solved using a mixture of these approaches. Mateescu et al. [10] propose the term “hybrid computing”, which captures the advantages of HPC, grid, computing, and cloud computing while reducing weaknesses as “no single paradigm is the best solution”.

## 3. Vendors, providers, and end-users of cloud computing

The *cloud computing vendor* supplies the infrastructure, typically at a fee. Various service models exist, which are explained in more detail in the next section. Examples of cloud computing vendors include Amazon, who offer Amazon Web Services (AWS); Microsoft, who offer Windows Azure; Google, who offer Google Compute Engine (GCE) and Google App Engine; and Rackspace; it should be noted that this list does not constitute an endorsement of these particular vendors.

The *cloud solution provider* uses the products made available by the cloud computing vendor and builds software services on top of these. The tools developed by cloud solution providers can themselves be used by others to provide new or enhanced tools [15]. An example of this multi-level approach could be a cheminformatics company offering an online conformer generation method built on one cloud platform. A second company or academic research group could make use of the conformer generation software to generate initial atomic coordinates of a library of compounds for rigid ligand docking running on a different platform.

The *end-user* consumes the solutions offered by a cloud solution provider. This might be a human interacting with a website, e.g. a user writing an article in Google docs, or large scale computations which use the cloud resources.

In some cases the cloud solution provider can also be the end-user, e.g., where a company, government or academic research group both develops a cloud-based application and uses this application to execute a particular molecular modelling project. In the next section, we will summarize the general types of services provided by cloud computing.

## 4. Services provided by cloud computing: IaaS, PaaS, and SaaS

In recent years there has been a proliferation of “X-as-a-Service” terminology. The Wikipedia entry for cloud computing lists no less than fourteen such services (Table 2). While most of these are specific in nature and tackle a single area or technology in computing (e.g. databases), Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are the most generic terms which describe a cloud computing offering and are therefore discussed in more detail.

*Infrastructure as a Service* is the provisioning of computing power and resources. At a basic level this typically includes storage, networking, processing units and memory. The physical form of the computing power varies from vendor to vendor but the common

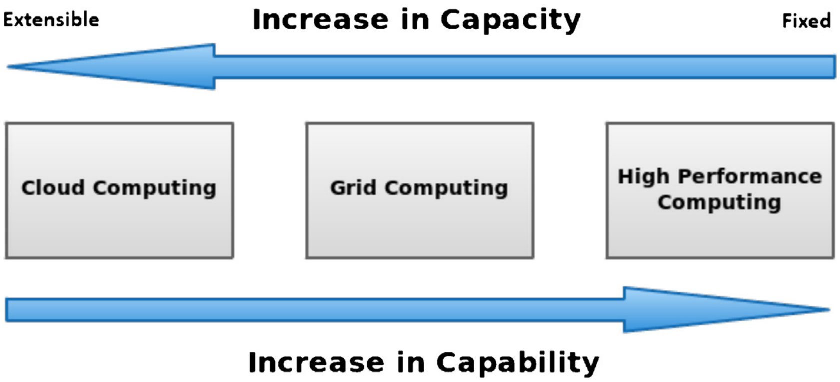


**Fig. 2.** Capacity and capability spectrum in cloud computing in comparison to grid and traditional high performance computing. By traditional HPC we mean a locally owned cluster with private access [10] which are usually connected by low-latency high speed interconnect, e.g. Infiniband or Myrinet. In contrast, typical cloud computing offerings are connected by standard Ethernet. It should be noted that clusters with low-latency high speed interconnect do exist in the cloud but are rarely available; even Amazon's "cluster compute" instances do not have low-latency interconnect.

infrastructures for compute-intensive tasks are compute clusters and/or GPU based cloud instances. The cloud computing vendor may also provide additional services on this layer such as automatic load balancing, diagnostic tools and Domain Name Server (DNS) resolution. The cloud solution provider must select the computational "instance type" (e.g. one with high memory), the supported operating system, and the number of required computational nodes or "instances". Some cloud vendors allow the cloud solution provider to create a disk image (or checkpoint) from a cloud instance. This has the advantage that upon starting a second instance all the libraries and applications installed in the first instance will already be available. Open source software such as Eucalyptus [16] and OpenStack [17] allows companies to build and manage their own public, private, and hybrid IaaS clouds. As an example, later in this article we describe our own experiences of a cheminformatics project where we used AWS IaaS to build a very large corporate virtual library and a concomitant ultrafast virtual screening platform.

*Software as a Service* is a cloud computing offering where the software application is provided through a front end such as a web server or through some other program interface. The advantage for the end-user is the facile access to the cloud resources as there is no need to worry about the hardware, installation or maintenance of the underlying methods, nor the intricacies of the cloud solution (load balancing, application stack, etc.). As a result, the end user only has to handle the data aspects, e.g. the upload of the data file. An example of SaaS could be a front end for docking calculations where the user uploads a protein structure and a set of ligands. A list of preinstalled software and currently available front ends developed for molecular modelling projects will be given later.

*Platform as a Service* occupies the middle ground between IaaS and SaaS: the cloud computing vendor supplies the hardware infrastructure, the programming infrastructure, and tools and libraries which make up a solution stack. Examples of PaaS include Google App Engine and AWS Elastic Beanstalk. These allow the cloud solution provider to place its own code on the cloud devices and to make these available to end users using the cloud infrastructure. Another PaaS example is KNIME (Konstanz Information Miner) as provided by the CloudBroker [18,19] platform. KNIME is a workflow management system that allows the graphical construction of workflows to execute data analysis, which includes many nodes of interest for molecular modelling. The CloudBroker platform allows the end-user to execute KNIME nodes and workflows in the cloud under a SaaS model while other cloud solution providers could integrate KNIME workflows into their own solution, building an enhanced platform which is offered as a service.

**Table 2**
Several X-as-a-Service terms.

| X-as-a-Service terms | Abbreviation |
|---|---|
| Infrastructure as a Service | IaaS |
| Platform as a Service | PaaS |
| Software as a Service | SaaS |
| Network as a Service | NaaS |
| Storage as a Service | STaaS |
| Security as a Service | SECaaS |
| Data or Desktop as a Service | DaaS |
| Database as a Service | DBaaS |
| Test Environment as a Service | TEaaS |
| API as a Service | APIaaS |
| Backend as a Service | BaaS |
| Integrated Development Environment as a Service | IDEaaS |
| Integration Platform as a Service | IPaaS |

## 5. Cloud computing for molecular modelling

### 5.1. Why consider cloud computing for molecular modelling projects?

Some molecular modelling tasks, such as molecular dynamics simulations, quantum mechanical calculations or 3D virtual library construction are very computationally intensive [20] and are often associated with the generation or processing of massive amounts of data. Very often, especially in small or medium sized organizations, the data and computational requirements vary over time. Three main reasons why cloud computing might be more compelling for these types of molecular modelling projects than more traditional in-house computing facilities, are scalability, reliability, and lower cost.

*Scalability*. Computing resources can be allocated dynamically to fit the needs of a particular project. This scalability of computing resources has the advantage that large-scale applications that are parallelizable can be finished in a short period of time by allocating the required computing resources. As soon as the project is done, the cloud resources can be deallocated. In contrast, traditional HPC clusters usually underutilize the computing resources or have resource bottlenecks (Fig. 3). The use of cloud computing is, therefore, particularly useful for smaller biotech companies which can perform computationally intensive projects on demand without having the need to buy and maintain HPC clusters. Of course, the degree of scalability that can be achieved depends on the particular task and the cloud hardware employed.

*Reliability*. This refers to both the hardware and software aspects of cloud computing. Cloud vendors like AWS guarantee 99.999999999% durability (this corresponds to an average annual expected loss of 0.000000001% of objects) and 99.99% availability

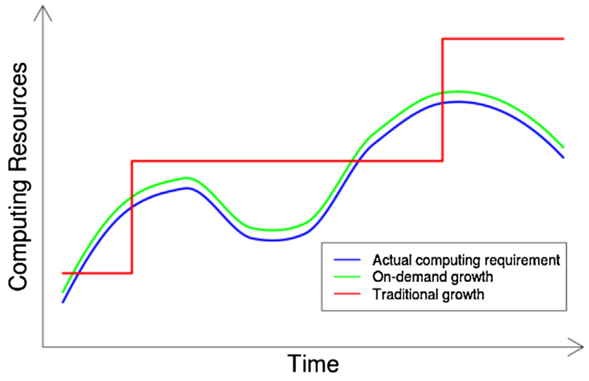


**Fig. 3.** Schematic comparison of traditional data-centre and cloud models for provisioning computational resources. Traditionally, computing resources are grown stepwise, for practical reasons, with the aim of being able to meet peak demand (red line). These resources are usually under-utilized and there may be a lag in responding to increased demand (blue line). With on demand growth as provided by cloud computing (green line), the required resources can be added on the fly and closely coupled to the actual computing requirement. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

for stored objects over a given year in their data storage service, Amazon Simple Storage Service, S3 [21]. These levels of service can be achieved because data is stored at multiple locations and checksums are used to make sure file data packets are not corrupted upon network transfer. Such levels of reliability can be challenging to achieve using in-house infrastructure. In PaaS and SaaS offerings, the cloud solution provider is also responsible for maintaining the cloud software stack and that it interoperates well with the underlying hardware.

*Lower cost*. Cloud computing can reduce the total cost of computational resources. First of all, it decreases capital expenditure (capex) as well as, potentially, operational expenditure (opex), because there is no initial capital investment required to buy the hardware or to pay for the ongoing hardware and software maintenance, electricity consumed, data centre insurance, etc. Of course, there is a balance between these savings and the costs of purchasing the cloud computing service, but in practice the cost of cloud resources can be modest in comparison to the in-house alternative. Sophisticated solution providers and end-users can take advantage of the "Spot Market" in cloud computing resources provided by some vendors. Users bid a price for spare resources, which are allocated to them if their bid is above the "Spot Price". The Spot Price can be significantly lower than the on-demand price, but workflows need to be robust to abrupt deallocation of resources when the price bid is exceeded. It should be noted that estimating the total cost is not always straightforward as there are some components such as those associated with I/O operations and network transfers that are hard to quantify. Some cloud vendors do provide calculators to help estimate these costs [22].

However, no technology solution is without its disadvantages and cloud computing is no exception. Cloud computing offers some additional challenges over traditional computing, mostly arising from not having a local computing infrastructure.

*Selection of vendor*. It can be hard to compare cloud vendors because their base definitions and offerings, such as what comprises a computing unit, vary. The only reliable way to determine the best cloud solution is to run the same benchmarks with different cloud vendors, which is time-consuming. Also, developing solutions based on one provider's platform may create a barrier to migrating to another provider that subsequently makes available a better service offering.

*Selection of cloud vendor instance types*. When a vendor offers different types of computational nodes, or instance types, it may not be immediately obvious which one is best suited to a particular job and therefore most cost effective. In a ligand-based virtual screening experiment, would higher processor speeds be of more benefit than larger instance memory sizes (where the ligand descriptors may be stored in memory to avoid slow I/O)? Again, time-consuming benchmarking on a subset of the data will probably be required.

*Transfer of data*. Typical cheminformatics projects, such as virtual library building or virtual screening, can produce huge amounts of data, while data mining applications may consume vast databases of information. Transferring potentially terabytes of data to or from the cloud computing vendor over the internet may be unfeasible. The most common solution to this problem is to ship external hard-disks to the cloud computing vendor and have the required data either imported or exported and shipped back; this is informally known as 'sneakernet'. This is a time consuming and manual process, presents potential logistical problems (such as lost or damaged disks, or customs delays if shipping across borders) and can easily take longer than the actual computation.

### 5.2. Current use of cloud computing for molecular modelling projects

As discussed above, the case for the use of cloud computing in molecular modelling for large-scale and data-intensive applications appears to be compelling [23]. This holds true particularly for embarrassingly parallel tasks such as the generation of combinatorial databases, virtual screening of millions of compounds, and the analysis of the huge datasets generated by genome sequencing or chemogenomics projects (Fig. 4). In order to use cloud computing, however, many tedious or technical tasks must be performed by the user such as configuring the compute nodes, installing the required software, and launching, monitoring and terminating the computation [24]. Such expertise is frequently lacking in biotech SMEs and this might be one reason why the use of cloud computing for molecular modelling applications is still in its infancy.

Encouragingly, front ends have been developed in recent years that facilitate user access to these massive computational resources as a Software as a Service (SaaS) (Table 3). In the following section, we will first introduce academic and commercial initiatives that have developed SaaS front ends for molecular modelling programs that facilitate the use of cloud computing resources. Then, case studies of cloud computing will be presented to demonstrate the scope and potential of these computational resources for molecular modelling.

### 5.3. Cloud computing SaaS front-ends for molecular modelling programs

Probably the most sophisticated SaaS front ends for cloud computing are provided by commercial providers such as CycleCloud [25], which offers clusters with preinstalled applications that are widely used in the area of bioinformatics, proteomics, and computational chemistry. The most interesting from a molecular modelling perspective are the availability of Gromacs, to perform molecular dynamics (MD) simulations, and ROCS to perform shape-based screening of compounds. Accelera recently announced AceCloud [26], which is a commercial on-demand GPU cloud computing service. At the time of writing, information about this service, such as performance and costs, is limited, but the optimization of MD simulation packages for running on GPUs, such as Amber, Gromacs, Lammps, and NAMD is reported to result in large performance increases compared to calculations run on CPUs [27], and thus could be of interest to groups performing free energy calculations.

*Rosetta@cloud*. The Rosetta software suite [28] facilitates molecular modelling tasks such as prediction of protein and RNA 3D

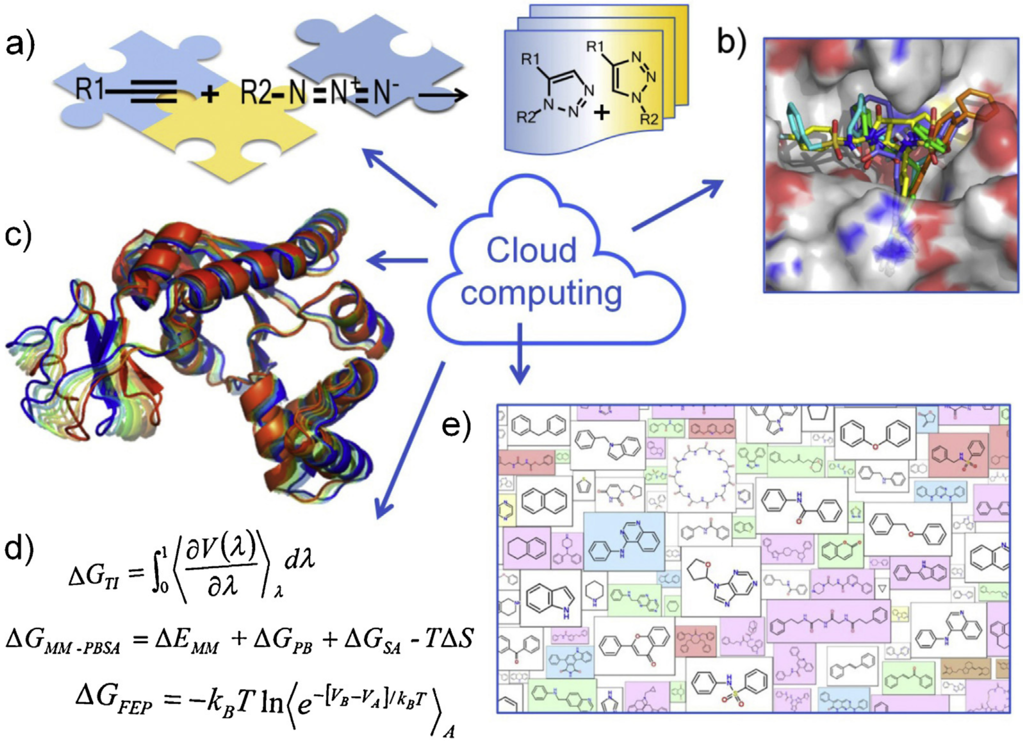

a)
R1
+
R2–N=N+=N-
R1
R2
R1
R2
+
b)
c)
Cloud computing
d)
$\Delta G_{TI} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda$
$\Delta G_{MM\text{-}PBSA} = \Delta E_{MM} + \Delta G_{PB} + \Delta G_{SA} - T\Delta S$
$\Delta G_{FEP} = -k_B T \ln \left\langle e^{-[V_B - V_A]/k_B T} \right\rangle_A$
e)

**Fig. 4.** Example applications of cloud computing for molecular modelling projects. (a) Generation of large compound databases, (b) structure-based screening (docking solutions against a thrombin structure), (c) molecular dynamics simulations (structural ensemble of the adenylate kinase generated by using the NMSim web server [59]) followed by (d) free energy calculations, and (e) analysis of huge datasets generated e.g. by chemogenomics projects (the molecule cloud of the ChEMBL database, taken from Ref. [60] with permission of the authors. The colour coding indicates the preference of the corresponding scaffold for a particular target class).

structures, protein-ligand and protein-protein docking, antibody modelling, and enzyme design. Because of the large search spaces of these problems, finding good solutions is highly computationally demanding. To meet this challenge a biomedical software company developed Rosetta@cloud, which offers the Rosetta software suite as a cloud-based service hosted on Amazon Web Services. As of March 2013, current Rosetta@cloud usage fees range from US $0.13 for “Small instance” to US $6.20 for “High I/O Quadruple Extra Large Instance” per hour [28]. For users with limited budgets and modest computational requirements, a free of charge web service provided by Rosetta is a compelling alternative computational source [29].

*AutoDockCloud* [30] is a workflow system that enables distributed screening on a cloud platform using the molecular docking program AutoDock [31]. AutoDockCloud is based on the open source framework Apache Hadoop, which implements the MapReduce paradigm [12] for distributed computing. In addition to the docking procedure itself, preparation steps, such as generating docking parameter and ligand input files, can be automated and distributed by AutoDockCloud. Initial testing was carried out using a subset of the DUD dataset [32] namely the set of oestrogen receptor alpha (ER-α) agonists of 2637 compounds, using a private cloud platform called Kandinsky (48 nodes with 16 cores per node) running up to 570 map tasks (i.e. docking jobs) in parallel. The preparation steps, however, have been only reported on a small dataset of 115 compounds [30]. Thus, it will be interesting (and necessary) to see the performance and usability of AutoDockCloud on a larger dataset and on public cloud resources.

*VMD plugin for NAMD simulations.* A software plugin for VMD [33] has been developed that provides an integrated framework for VMD to be executed on Amazon EC2 [24], thereby facilitating MD simulations of biomolecules using the NAMD program [34]. This plugin allows use of the VMD Graphical User Interface to: (1) create an compute cluster on Amazon EC2; (2) submit a parallel MD simulation job using the NAMD software; (3) transfer the results to the host computer for post-processing steps; and (4) shutdown and terminate the compute cluster on Amazon EC2 [24]. In a case study by the authors, a simulation of the satellite tobacco mosaic virus (∼1 million atoms in the solvated system) for 30 ps using a compute cluster with 64 cores on Amazon EC2 took 4.6 h and cost US $64; thereby, achieving a speedup of 32 compared to a single CPU. Using the same set up and scaling the numbers, a simulation trajectory of 100 ns length would need about 2 years and cost approximately $200,000. Clearly MD simulation of such a large system is at the high end of what is currently achievable. Fortunately, most researchers

**Table 3**
Molecular modelling programs that are offered as SaaS front-ends.

| Program | Application | Front end name/Provider | Ref. |
|---|---|---|---|
| ROCS | Shape based screening | CycleCloud[b] | [25] |
| Rosetta | Structure prediction, protein–protein docking | Rosetta@cloud[b] CloudBroker[b] | [19,28] |
| AutoDock | Molecular docking, structure based screening | AutoDockCloud[a] CloudBroker[b] | [19,30] |
| Gromacs | MD simulations | CycleCloud[b] CloudBroker[b] | [19,25] |
| VMD & NAMD | MD simulations | VMD[a] | [24] |
| BLAST | Sequence alignment | Windows Azure[b] CloudBroker[b] | [19,36] |

[a] Academic provider.
[b] Commercial provider.

in the drug design area simulate much smaller systems (around 60,000 atoms for a protein of 300 amino acids with explicit solvent).

The case study on the satellite tobacco mosaic virus shows that the use of an HPC cluster provided by AWS, which is recommended for tightly coupled parallel processes, is indeed suitable for running communication-bound parallel applications such as MD simulations. Furthermore, this study implies that the VMD plugin will provide facilitated access of the AWS cloud for MD simulations. For applications such as MD simulations that are not "embarrassingly parallel", and for which high levels of inter-processor communication are required, it is important to investigate scalability of the calculation and therefore the selection of the number of processors carefully – calculations that use more processors may be less cost-effective if scaling is significantly sub-optimal.

### *5.4. Data storage and analysis in the cloud*

Besides providing computational power, cloud-based services also show promise for "big data" storage and easy access to data analysis software without the need for local installation and maintenance. The ability to host big data on the cloud is particularly useful in bioinformatics, as the gap between data generated by next-generation sequencing and the ability to store and analyze such data locally, is growing [35]. AWS offers access to multiple public databases [36] for a variety of scientific fields including the PubChem library and a 3D version of PubChem. PubChem provides information on the biological activities of small molecules, and is a component of NIH's Molecular Libraries Roadmap Initiative. All public datasets in AWS are hosted free of charge and can be accessed from Amazon EC2 instances and then integrated into cloud-based applications [35,37]. Another example, NCBI BLAST on Windows Azure [38] is a cloud-based implementation of the Basic Local Alignment Search Tool (BLAST) of the National Center for Biotechnology Information (NCBI). The BLAST program can be used to search sequence databases for similarities between a protein or DNA query sequence and known sequences. The Windows Azure implementation allows researchers to rent the required cloud computing time and automate the data analysis. For this, a web interface is provided where the user can initiate, monitor, and manage their BLAST jobs (i.e. upload the input file and specify BLAST parameters and the number of job partitions). Cloud-based platforms, which allow the use of such facilities via a public-facing web server are, to the best of our knowledge, currently only available for the analysis of high-throughput sequencing data (a list of available services is provided in reference [35]) but are missing for small molecule data. To set up such a service one could use the e-Science Central platform which is discussed next.

*e-Science Central* [39] is a cloud based platform for data analysis developed at Newcastle University, UK, which allows scientists to store and share data or software in a secure and controlled way, as well as to analyze data using their own workflow scripts. This can be done entirely through a web browser allowing easy access, independent of the location of the workplace. e-Science Central has been designed to be independent of any specific cloud infrastructure and can run on both private (e.g. Eucalyptus) and public clouds (Amazon EC2 and Microsoft Windows Azure), thus providing the user with some degree of independence in their choice of cloud provider [39].

To demonstrate the suitability of e-Science Central several case studies were presented [39], including the migration of an existing QSAR model pipeline called 'Discovery Bus' to the cloud [39,40]. Discovery Bus allows automatic generation or updating of hundreds of QSAR models for a given dataset, out of which the most predictive model is finally selected. Because of the parallel nature of many underlying aspects, such as descriptor calculation and model training (e.g. via multiple linear regression, neural network, classification tree, etc.), the cloud has great potential to speed up the QSAR model building process. To use the cloud computing power, the Discovery Bus workflow had to be re-implemented as a hierarchy of e-Science workflows [40]. The resulting implementation allowed processing of the ChEMBL dataset (database of bioactive drug-like molecules) in ∼14 h by using 100 Windows Azure nodes, each containing two cores. The speed-up in processing the QSAR workflow was nearly linear (88.2% of the ideal) for up to 200 nodes [40]. An important step in achieving this scalability was to move almost all data communications from the central e-Science data repository to the Azure Blob Storage Service. According to the authors, this also has the advantage that the system communicates via local disks on the cloud rather than by shared file systems, thereby reducing the costs which have to be paid by users for network operations [40].

Many scientists will probably have access to 200 nodes via a local HPC cluster. Nevertheless, this example is a good use of cloud resources as it permits the updating of the QSAR models by paying only for the necessary compute power required whenever a new version of ChEMBL is released. Furthermore, it provides a framework for scientists to access a centralized data repository, and one could imagine that different users could extend the QSAR model building procedure by adding new training models without having to re-implement or install those already available.

## 6. Applications of cloud computing for molecular modelling

The number of published studies in molecular modelling where cloud resources were used is still small. In fact, the only published studies we could find where cloud resources were used to boost molecular modelling projects come from the Baker laboratory at the University of Washington [41–43], although there may be other studies that have used cloud computing but which do not explicitly describe this in the publication. In both studies, the Windows Azure computing resource was used to enable protein structure predictions using the Rosetta modelling software: Loquet et al. [42,43] used this resource to predict the structure of the *Salmonella typhimurium* T3SS needle protein, while Thompson et al. [41] used the Azure resource to show that protein structure prediction can be facilitated by incorporating backbone chemical shift data and distance restraints derived from homologous proteins of known structure and an all-atom energy function. The projects were enabled by the donation of cloud computing resources by Microsoft and by implementing a modified version of the BOINC (Berkeley Open Interface for Network Computing) interface which was originally designed for volunteer grid computing [44]. For a list of further research projects supported by the Microsoft Research Cloud Research Engagement project see [45]. To accomplish the protein folding calculations 2000 cores on Windows Azure were run for just under a week [46]. Initially an equivalent number of cores and computing hours were used to check the algorithms and process. Overall, these calculations demonstrate the potential of cloud computing to impact protein folding research projects.

Several biotech and pharmaceutical companies (e.g. Eli Lilly [47]) have started to experiment with cloud resources to carry out computationally intensive molecular modelling R&D projects. Two case studies will now be presented to further demonstrate the scope and potential of cloud-based services and applications for molecular modelling projects.

### *6.1. Case study I: Virtual screening of 21 million compounds*

The power of cloud infrastructure for screening purposes was demonstrated by a project performed by Cycle Computing in collaboration with Schrödinger and Nimbus Discovery [48,49]. The

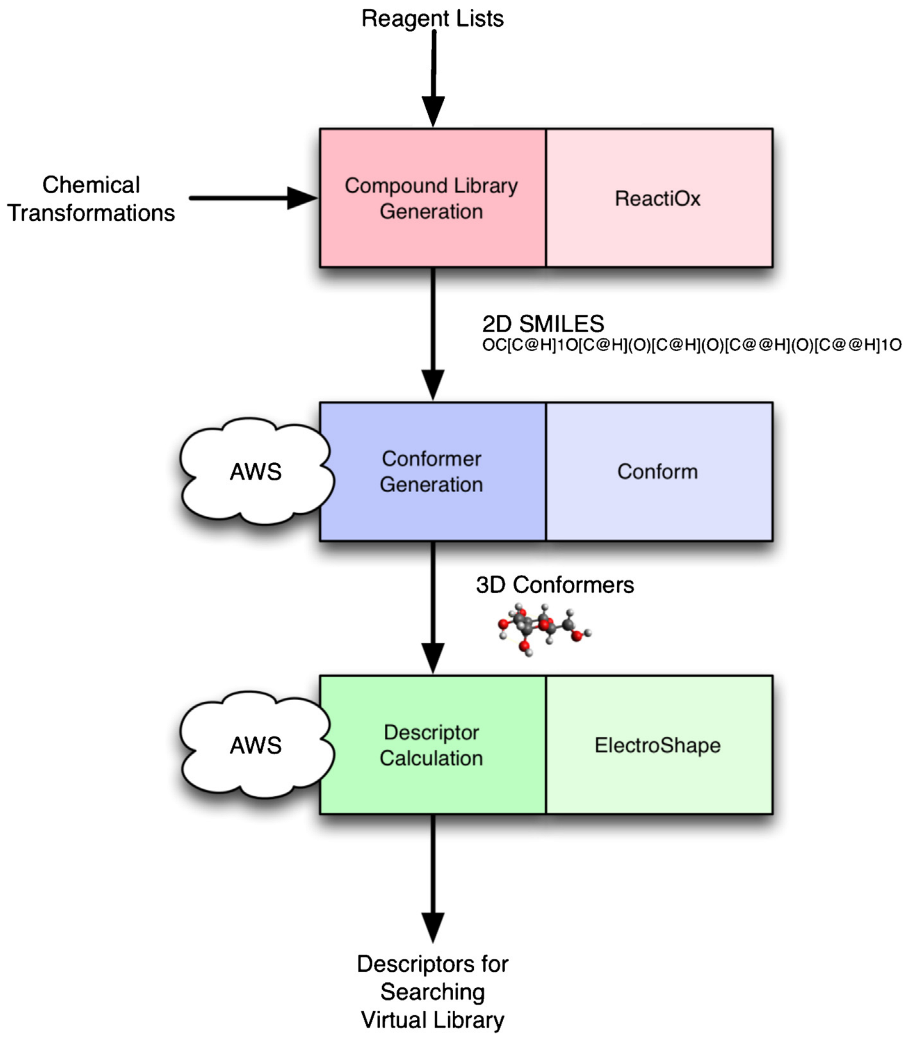


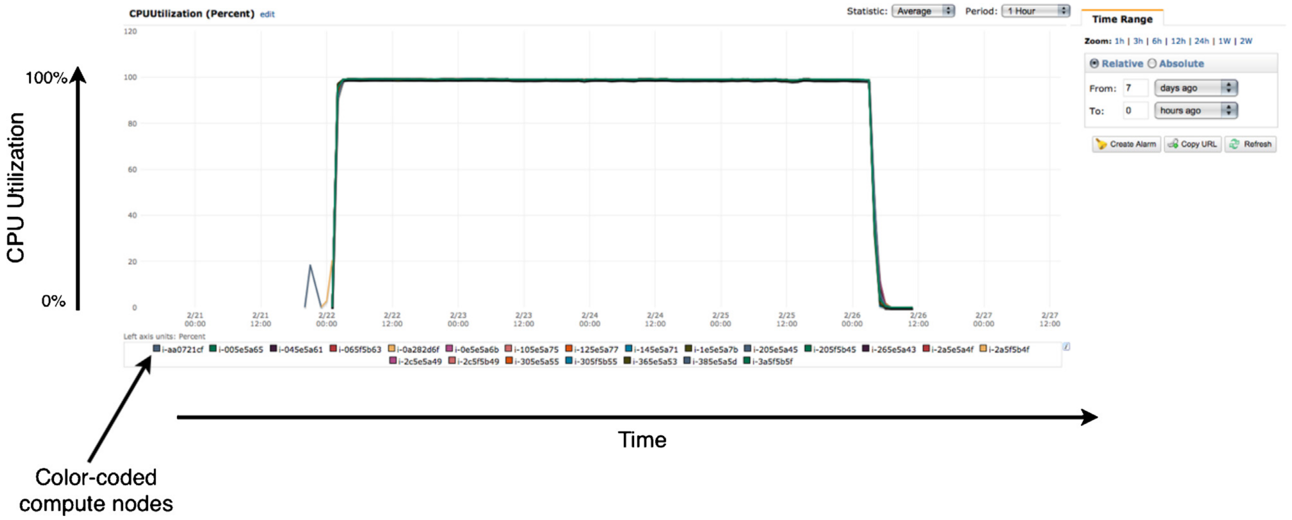
**Fig. 5.** Flowchart showing the generation of a virtual library. The example shows virtual library enumeration, conformer generation and ultrafast ligand-based screening descriptor calculation using the cloud. InhibOx's proprietary tools, ReactiOx, Conform [55] and ElectroShape [50–52], were used to put together a customized version of E-Scape, an ultrafast ligand-based virtual screening system, for one of its corporate clients.

companies launched a 50,000 core cluster on AWS to screen 21 million compounds in less than three hours (costs approximately US $4850 per hour). The screening was performed with Schrödinger's Glide docking software using SP (Standard Precision) mode, which is more accurate but computationally more intensive than its HTVS (High Throughput Virtual Screening) mode usually used for initial screening purposes. The basis for these high-performance calculations was the CycleCloud software developed by Cycle Computing which enabled the creation of compute clusters in the cloud and the automation and management of large scale calculations. In February of 2013, Cycle Computing announced the set up of a cluster consisting of approximately 10,600 instances that was also used for compound screening purposes, with a total cost of about US $4400 [48].

### 6.2. *Case study II: Generation of a database of ∼30 million compounds*

Lead identification for new projects remains challenging, and both experimental and computational high-throughput screening approaches often fail to find suitable starting points for optimization. One weakness of the current approaches is their reliance on historical compound collections (either commercially available or from previous drug discovery programmes), which represent a small and biased sample of chemical space. Therefore, combining a greatly expanded and diversified chemical space with a very fast screening method is highly desirable, especially if the compounds are synthetically accessible.

Given a list of high-yield one-step chemical reactions, and the set of all suitable reagents available in a corporate collection, it is possible to enumerate computationally a very large virtual library of synthetically accessible lead-like or drug-like molecules. By taking such virtual libraries, calculating low energy conformations for each molecule, and then generating the appropriate ElectroShape descriptors [50–52], it is possible to construct an ultra-fast ligand-based virtual screening system for a corporate collection that can be used for library design and ligand-based discovery of lead-like and drug-like molecules. InhibOx used Star Cluster from MIT [53] and cloud computing resources purchased from Amazon Web Services (AWS) to do just this [54]. AWS was used in two stages (Fig. 5), first to perform the enumeration of diverse sets of low-energy conformers of each molecule in a virtual library of just over 28 million compounds, and second to generate the ElectroShape descriptors for use in E-Scape, InhibOx's ultra-fast ligand-based virtual screening system. The virtual compound libraries were enumerated locally using InhibOx's proprietary virtual library generation tool, ReactiOx, using a wide range of chemical transformations and matched reagent lists. This produced sets of SMILES



Statistic: Average

Period: 1 Hour

● Relative ○ Absolute

From: 7 days ago

To: 0 hours ago

**Fig. 6.** Amazon CloudWatch screenshot. The graph shows the CPU utilization percentages for a subset (22/64 nodes) of the AWS compute cluster used to enumerate low energy conformers for an entire corporate virtual library. These CPUs effectively came into existence only for the period they were needed, and were shut down after the computational job completed.

strings defining the topology and chirality of the products that also had to pass physicochemical filtering criteria to ensure that the library compounds were suitable as leads. Conformer generation was performed on AWS using another of InhibOx's proprietary tools, Conform [55]; this produced diverse ensembles of up to 200 low energy conformers for each compound, the ensemble sizes growing with increasing conformational flexibility. Once the 3D conformers were generated, the descriptor calculation was performed, again on AWS, using InhibOx's ElectroShape, to produce descriptors for virtual screening with InhibOx's E-Scape.

Input SMILES files and the generated conformers were stored using Amazon Simple Storage Service, S3. Approximately 550 GB of bzip2-compressed GPG-encrypted SDF data was generated, which unencrypted and uncompressed consisted of about 5.7 TB of SDF files, containing just over 28 million molecules, each molecule having up to a maximum of 200 low energy conformations. A similar computational technique was used to calculate the various ElectroShape descriptors for each of the enumerated conformers in the virtual collection. Following completion of the calculations, the data was transferred of the cloud using AWS Export.

Before deciding on which of the various instance types offered by AWS should be used, a benchmarking exercise using a tiny subset of the corporate virtual library was carried out to determine which was most cost-effective for the conformer generation. It was found that the cc2.8xlarge instance type was most efficient for this particular problem. The final production run used 64 cc2.8xlarge instances [56] for a period of just over 4 days (Fig. 6). Each cc2.8xlarge has 2 Intel Xeon processors with 8 hardware cores and Hyper-Threading enabled, which allows each core to process a pair of instruction streams in parallel; each instance had 60.5 GB of RAM and 3.37 TB of instance storage. Each cc2.8xlarge instance was connected to a 10 Gigabit network, offering low latency connectivity with full bisection bandwidth to other CC2 instances. Each cc2.8xlarge node thus effectively has 32 cores, so in total 2048 cores were used simultaneously.

It is worth noting that the same job could have been completed more quickly by simply requesting more instances. Had InhibOx asked for five times as many instances (320) the same job would have completed in less than a day, it would have cost the same, and using performances numbers reported by AWS for the cc2.8xlarge instances [22] we estimated the compute cluster could have ranked as a Top500 supercomputer (estimated 70.2 TFlop/s). At the time the job was run, according to the then rankings of the Top500.org web site, a machine needed to be faster than Rmax = 50.9 TFlop/s to rank in the top 500.

## 7. Conclusions

Cloud computing is a recent phenomenon whose growth has been partly facilitated by the emergence of new technologies such as virtualization and open source web protocols like web services, REST, SOAP. Its adoption has been fuelled in part by the rapid growth of "Big Data", particularly where publically available big datasets are hosted by cloud computing vendors. The use of encryption in daily life, to access one's bank account or make a purchase through a web browser, has increased confidence in transmitting and storing sensitive data on the cloud, but data security remains a concern to some potential users. Some would argue that the significant spare capacity that has been created by big web companies like the search giant Google and the online retailer Amazon has also been a stimulus for affordable cloud computing. All new technologies suffer from an initial period of hype, but cloud computing appears to be past the "Peak of Inflated Expectations" described in Gartner's "Hype Cycle for Emerging Technologies" of 2012, enabling a more balanced assessment to be made [57]. Nothing is perfect, and some prominently publicized outages of cloud services remind us that backup plans should be made, and cloud solutions need to be architected so as to be robust in the face of such rare events, and to avoid vendor lock-in.

In the life sciences, the literature suggests that the cloud resources have so far been most widely used in bioinformatics, particularly for the analysis of data from next-generation sequencing projects. A number of excellent reviews already exist which summarize the use of cloud computing in bioinformatics and the interested reader is referred to these [35,58]. The uptake of cloud computing for cheminformatics and molecular modelling has been slower, perhaps in part because of the generally more proprietary nature of molecular modelling software and data.

Nonetheless, several biotech and pharmaceutical companies have started to experiment with cloud computing to perform compute-intensive molecular modelling projects. The on-demand availability of a huge computational resource at an ever decreasing price and without large upfront costs for purchasing computing equipment is very attractive, particularly when this equipment cannot be guaranteed to be used to its fullest capacity for the expected life of the hardware, or for calculations that occasionally require massive amounts of compute power beyond what's available in-house. The adoption of cloud computing by researchers from academic groups for molecular modelling is less apparent in the literature, perhaps because of the cost models of cloud computing and the existing access academic researchers have to centrally provided supercomputing and HPC facilities. It may be expected that more academic researchers will use cloud resources in the future for molecular modelling projects as costs decrease and these technologies mature and develop further, budgeting for large one-off cloud computations as a line item in a grant proposal.

In conclusion, coupled with the increasing availability of big datasets in biology and chemistry, there is little doubt that the use of cloud computing technologies will continue to grow, spur innovation, and enable more exciting discoveries in the years ahead.

## Acknowledgements

Jean-Paul Ebejer and Simone Fulle acknowledge funding by the EC Marie Curie Initial Training Network “Scientific Training in Antimicrobial Research Strategies” (grant agreement 238490).

## References

[1] Google Trends, http://www.google.com/trends/ (accessed on 12.03.13).

[2] P. Mell, T. Grance, The NIST definition of cloud computing, NIST Special Publication 800 (2011) 145.

[3] L.M. Vaquero, L. Rodero-Merino, J. Caceres, M. Lindner, A break in the clouds: towards a cloud definition, ACM SIGCOMM Computer Communication Review 39 (2009) 50–55.

[4] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, A. Ghalsasi, Cloud computing—the business perspective, Decision Support Systems 51 (2011) 176–189.

[5] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems 25 (2009) 599–616.

[6] Intel's Vision of the Ongoing Shift to Cloud Computing, http://www.intel.ie/content/dam/www/public/us/en/documents/white-papers/cloud-computing-intel-cloud-2015-vision.pdf (accessed on 12.03.13).

[7] R.L. Grossman, The case for cloud computing, IT Professional 11 (2009) 23–27.

[8] R. Hill, L. Hirsch, P. Lake, S. Moshiri, Guide to Cloud Computing, Springer, 2013.

[9] Deltacloud Web Page, http://deltacloud.apache.org/ (accessed on 12.03.13).

[10] G. Mateescu, W. Gentzsch, C.J. Ribbens, Hybrid computing—where HPC meets grid and cloud computing, Future Generation Computer Systems 27 (2011) 440–453.

[11] A. Marinos, G. Briscoe, Community cloud computing, Cloud Computing 47 (2009) 2–484.

[12] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, Communications of the ACM 51 (2008) 107–113.

[13] I. Foster, Y. Zhao, I. Raicu, S. Lu, Cloud computing and grid computing 360-degree compared, in: Grid Computing Environments Workshop, 2008. GCE'08, 2008, pp. 1–10.

[14] C. Gong, J. Liu, Q. Zhang, H. Chen, Z. Gong, The characteristics of cloud computing, in: Parallel Processing Workshops (ICPPW), 2010 39th International Conference on, IEEE, 2010, pp. 257–259.

[15] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, et al., A view of cloud computing, Communications of the ACM 53 (2010) 50–58.

[16] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, et al., The eucalyptus open-source cloud-computing system, in: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009, pp. 124–131.

[17] OpenStack, http://www.openstack.org (accessed on 18.03.13).

[18] CloudBroker, http://www.cloudbroker.com (accessed on 18.03.13).

[19] Ported Applications by CloudBroker, http://si-se.ch/2013/presentations/hpc_cloud.pdf (accessed on 18.03.13).

[20] S. Fulle, H. Gohlke, Flexibility analysis of biomacromolecules with application to computer-aided drug design, Methods in Molecular Biology 819 (2012) 75–91.

[21] Amazon S3, http://aws.amazon.com/s3/ (accessed on 12.03.13).

[22] HPC on AWS, http://aws.amazon.com/hpc-applications (accessed on 18.03.13).

[23] Q. Zhang, L. Cheng, R. Boutaba, Cloud computing: state-of-the-art and research challenges, Journal of Internet Services and Applications 1 (2010) 7–18.

[24] A.K.L. Wong, A.M. Goscinski, The design and implementation of the VMD plugin for NAMD simulations on the Amazon cloud, International Journal of Cloud Computing and Services Science (IJ-CLOSER) 1 (2012) 155–171.

[25] Applications Provided by CycleCloud, http://www.cyclecomputing.com/cyclecloud/applications (accessed on 12.03.13).

[26] ACECloud – Acellera's On-Demand GPU Computing Service, http://www.acellera.com/products/acemd/acecloud/ (accessed on 12.03.13).

[27] M.S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A.L. Beberg, et al., Accelerating molecular dynamic simulation on graphics processing units, Journal of Computational Chemistry 30 (2009) 864–872.

[28] Rosetta@cloud on AWS, http://rosetta.insilicos.com (accessed on 12.03.13).

[29] Rosetta Online Server, http://rosettaserver.graylab.jhu.edu/ (accessed on 12.03.13).

[30] S.R. Ellingson, J. Baudry, High-throughput virtual molecular docking with AutoDockCloud, Concurrency and Computation: Practice and Experience (2012), http://dx.doi.org/10.1002/cpe.2926.

[31] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, et al., AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility, Journal of Computatioal Chemistry 30 (2009) 2785–2791.

[32] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking sets for molecular docking, Journal of Medicinal Chemistry 49 (2006) 6789–6801.

[33] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, Journal of Molecular Graphics 14 (1996) 33–38.

[34] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, et al., Scalable molecular dynamics with NAMD, Journal of Computational Chemistry 26 (2005) 1781–1802.

[35] L. Dai, X. Gao, Y. Guo, J. Xiao, Z. Zhang, Bioinformatics clouds for big data manipulation, Biology Direct 7 (2012) 43.

[36] Public Data Sets on AWS, http://aws.amazon.com/publicdatasets/ (accessed on 12.03.13).

[37] V.A. Fusaro, P. Patil, E. Gafni, D.P. Wall, P.J. Tonellato, Biomedical cloud computing with Amazon Web Services, PLoS Computational Biology 7 (2011) e1002147.

[38] NCBI BLAST on Windows Azure, http://research.microsoft.com/en-us/projects/ncbi-blast/default.aspx (accessed on 12.03.13).

[39] H. Hiden, S. Woodman, P. Watson, J. Cala, Developing cloud applications using the e-science central platform, Philosophical Transactions of the Royal Society A 371 (2013) 20120085.

[40] J. Cała, H. Hiden, P. Watson, S. Woodman, Cloud computing for fast prediction of chemical activity, in: 2nd International Workshop on Cloud Computing and Scientific Applications (CCSA), Ottawa, 2012.

[41] J.M. Thompson, N.G. Sgourakis, G. Liu, P. Rossi, Y. Tang, J.L. Mills, et al., Accurate protein structure modeling using sparse NMR data and homologous structure information, Proceedings of the National Academy of Sciences of the United States America 109 (2012) 9875–9880.

[42] A. Loquet, N.G. Sgourakis, R. Gupta, K. Giller, D. Riedel, C. Goosmann, et al., Atomic model of the type III secretion system needle, Nature 486 (2012) 276–279.

[43] A. Loquet, N.G. Sgourakis, R. Gupta, K. Giller, D. Riedel, C. Goosmann, et al., Corrigendum: Atomic model of the type III secretion system needle, Nature (2012).

[44] D.P. Anderson, BOINC. A system for public-resource computing and storage, in: Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing, 2004, pp. 4–10.

[45] Research Projects on Microsoft Cloud, http://research.microsoft.com/en-us/projects/azure/projects.aspx (accessed on 12.03.13).

[46] Case Study on Windows Azure, http://blogs.msdn.com/b/windowsazure/archive/2011/06/16/windows-azure-helps-scientists-unfold-protein-mystery-and-fight-diseas.aspx (accessed on 12.03.13).

[47] Eli Lilly on Cloud Computing, http://www.informationweek.com/hardware/data-centers/qa-eli-lilly-on-cloud-computing-reality/228200755?pgno=2 (accessed on 20.03.13).

[48] Blog of Cycle Computing, http://blog.cyclecomputing.com/ (accessed on 05.03.13).

[49] Cycle Computing Provisioned a 50,000 Core Cluster on AWS, http://www.marketwire.com/press-release/Cycle-Computing-Ramps-Global-50000-Core-Cluster-for-Schrodinger-Molecular-Research-1646214.htm (accessed on 12.03.13).

[50] M.S. Armstrong, G.M. Morris, P.W. Finn, R. Sharma, L. Moretti, R.I. Cooper, et al., ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics, Journal of Computer-Aided Molecular Design 24 (2010) 789–801.

[51] M.S. Armstrong, P.W. Finn, G.M. Morris, W.G. Richards, Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension, Journal of Computer-Aided Molecular Design 25 (2011) 785–790.

[52] M.S. Armstrong, G.M. Morris, P.W. Finn, R. Sharma, W.G. Richards, Molecular similarity including chirality, Journal of Molecular Graphics and Modelling 28 (2009) 368–370.

[53] StarCluster@MIT, http://star.mit.edu/cluster/about.html#id9 (accessed on 28.02.13).

[54] AWS case study, InhibOx, https://aws.amazon.com/solutions/case-studies/inhibox (accessed on 28.02.13).

[55] J.P. Ebejer, G.M. Morris, C.M. Deane, Freely available conformer generation methods: how good are they, Journal of Chemical Information and Modelling 52 (2012) 1146–1158.

[56] Next Generation Cluster Computing on Amazon EC2 – the CC2 Instance Type, http://aws.typepad.com/aws/2011/11/next-generation-cluster-computing-on-amazon-ec2-the-cc2-instance-type.html (accessed on 28.02.13).
[57] Gartner's Hype cycle, http://www.infoq.com/news/2012/08/Gartner-Hype-Cycle-2012 (accessed on 12.03.13).
[58] A. Shanker, Genome research in the cloud, OMICS 16 (2012) 422–4228.
[59] D.M. Krüger, A. Ahmed, H. Gohlke, NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins, Nucleic Acids Research 40 (2012) W310–W316.
[60] P. Ertl, B. Rohde, The Molecule Cloud – compact visualization of large collections of molecules, Journal of Cheminformatics 4 (2012) 12.