International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume 2 Issue 4, PP : 239-244 1 April 2013

Cloud Computing for Biomedical Information Management

Mohana R.S.[1], Dr.P.Thangaraj[2], S. Kalaiselvi[3], B.Krishnakumar[4]

[1,3,4] Department of Computer Science and Engineering , Kongu Engineering College, Erode, Tamilnadu, India
[2] Department of Computer Science and Engineering , Bannari Amman Institute of Technology Sathyamangalam , Tamilnadu , India
mohana.pragash@rediffmail.com

***Abstract*— Biomedical informatics community shares data and applications, can take advantage of a new resource called "cloud computing". Clouds generally offer resources on demand. In most clouds, charges are pay per use, based on large farms of inexpensive, dedicated servers, sometimes supporting parallel computing. When clouds are used for biomedical information management, it costs much lower than dedicated laboratory systems or even institutional data centers. Biomedical applications that are not I/O intensive and do not demand a fully mature environment can use clouds which has major technological improvements. Instead of listing the strengths and weaknesses of cloud-based systems (e.g., for security or data preservation), this paper reviews the changes from individual lab systems to a grid and cloud based environment for Biomedical Informatics. Many observers believe that clouds represent the next generation of computing paradigm for biomedical application.**

Keywords- Cloud computing: Data sharing; Bioinformatics; Security; Distributed computing

I. INTRODUCTION

Life sciences make heavy use of the web as a medium for data access and computational analyses. Bioinformatics problems are solved by interacting with two or more machines which increases the need for computing. International Nucleotide Sequence Database Collaboration and many other institutions provide public accessible and machine-readable services to retrieve, submit, or analyze bioinformatics data [3-6]. Taverna [7], Cyrille2 [8] and BioPipe [9] are recent tools available to aggregate services, like workflow environments and scripting environments. This aggregation of services shows great potential, as exemplified in the successful experiment in [10] where microarray data, genomic sequence information, and pathway databases were integrated in a workflow to aid the search for candidate genes responsible for phenotypic variation.

Cloud computing is a convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be provided and released with minimal effort.

"Cloud" computing has been receiving much attention as an alternative to both specialized grids and to owning and managing one's own servers. Currently available articles, blogs, and forums focus on applying clouds to industries outside of biomedical informatics. In this article, the fundamentals of cloud computing is described and illustrate how one might evaluate a particular cloud for biomedical purposes.

Typically, laboratories purchase local servers for computation or data-intensive tasks that cannot be performed on desktop machines. Locally-hosted machines are also increasingly used to share data and applications in collaborative research, e.g., in the Biomedical Informatics Research Network (BIRN) and Cancer Biomedical Informatics Grid (caBIG), both funded by the National Institutes of Health (NIH).Meanwhile, image analysis, data mining, protein folding, and gene sequencing are all important tools for biomedical researchers. These resource-intensive shared applications often involve large data sets, catalogs, and archives, under multiple owners, often with bursty workloads. In response, biomedical consortia (often involving multiple institutions) have implemented their applications on top of laboratory-hosted servers in distributed grid architecture, as described in Section 2. To sustain such servers, laboratories and their institutions require space, cooling, power, low-level system administration, and negotiations (e.g., about software standards and firewalls between institutions).

Clouds shift the responsibility to install and maintain hardware and basic computational services away from the customer (e.g., a laboratory or consortium) to the cloud vendor. Higher levels of the application stack and administration of sharing remain intact, and remain the customer's responsibility.

The goal of this paper is to help users at biomedical laboratories, funding agencies, and especially consortia to understand where cloud computing may be appropriate and to describe how to assess a particular cloud. We focus on labs that need to share information with outsiders, such as consortia investigators—the rapidly-growing cloud literature suffices to guide labs that simply wish to acquire cheaper compute resources.

In Section 2 background information on grids and clouds is presented. Section 3 provides an overview of consortium computing. Section 4 discusses cloud infrastructure for medical consortia and describes sample cloud vendors. Section 5 evaluates several desirable features of cloud, and Section 6 discusses cloud security. Section 7 presents conclusions.

II. BACKGROUND

Terabytes to petabytes of scientific data are easily generated by powerful instruments, satellites, and sensor networks in a day [11]. As biomedical research transitions to a data-centric paradigm, scientists need to work more collaboratively, crossing geographic, domain, and social barriers. Interdisciplinary collaboration over the Internet is in demand, making it necessary for individual laboratories to equip themselves with the technical infrastructure needed for information management and data sharing. For example, a research group may need to include data from clinical records, genome studies, animal studies, and toxicology analyses. The use of spreadsheet in biomedical application for data storage is reaching its limits [12].

A. Buzzwords in Distributed system

Grids, virtualized data centers, and clouds constitute three approaches to sharing computer resources and data to facilitate collaboration. These architectures overlap in their implementation techniques and in the features they offer to biomedical consortia. Furthermore, systems of each category adopt good ideas from the others, and tradeoffs often depend on the presence of that feature, not on the overall categorization. We summarize these architectures briefly here and express detailed comparisons in terms of individual features.

Grid technology is popular in the scientific community. Grid participants typically share computational resources running on independently-managed machines, using standard networking protocols. Grid toolkits often provide management and security capabilities. When running computationally-intensive jobs, one frequently receives an entire machine, or several.

Data center virtualization products typically assume a dedicated pool of machines that are used to support a variety of tasks. They have become quite successful in commercial and government data centers. While one may occasionally allocate a whole machine (or cluster) to a single, computationally-expensive task, more often these products allow multiple virtual processors, storage systems, and networks to be supported over the same set of underlying hardware. Virtual machines can be quickly activated or deactivated. If each virtual machine is lightly utilized, one can consolidate many virtual machines onto the same physical hardware, thus improving utilization and cost. To compete with open source products (such as Xen), leading vendors (such as VMware) now include higher-level services, such as configuration management, workload orchestration, policy-based allocation, and accounting.

Cloud computing is a model which offers leasable computational resources on-demand over a network. The cloud computing model can simplify access to a variety of computing architectures, including large memory machines, while eliminating the need to build or administer a local computer network addressing challenges in access and deployment of infrastructure for bioinformatics.

B. Cloud features

The following features, especially the first three, are commonly associated with clouds. A *consumer* can be an individual lab, a consortium participant, or a consortium.

• *Resource outsourcing:* The cloud vendor is responsible for hardware acquisition and maintenance.

• *Utility computing:* The consumer requests additional resources as needed, and similarly releases these resources when they are not needed. Different clouds offer different sorts of resources, e.g., processing, storage, management software, or application services [13].

• *Large numbers of machines:* Clouds are typically constructed using large numbers of inexpensive machines. As a result, the cloud vendor can more easily add capacity and can more rapidly replace machines that fail, compared with having machines in multiple laboratories. Generally speaking these machines are as homogeneous as possible both in terms of configuration and location.

• *Automated resource management:* This feature encompasses a variety of configuration tasks typically handled by a system administrator. For example, many clouds offer the option of automated backup and archival. The cloud may move data or computation to improve responsiveness.

• *Virtualization:* Hardware resources in clouds are usually virtual; they are shared by multiple users to improve efficiency. That is, several lightly-utilized logical resources can be supported by the same physical resource.

III. CONSORTIUM COMPUTING

Clouds are candidates for several roles in biomedical computing, ranging from compute services to archival storage to acting as a neutral zone among laboratories in a consortium. Individual labs often include basic servers. Labs that engage in computationally expensive research (e.g., protein folding or simulations) may rely on clusters of high-performance machines with fast interconnect between processors. At the other extreme, international repositories (e.g., SwissProt and GenBank) require extensive storage, but less impressive computational power.

Between these extremes are biomedical consortia that facilitate the exchange of data and applications among its participants. In this section, we provide an overview of biomedical computing infrastructure, paying particular attention to the needs of consortia.

A. Laboratory infrastructure

To meet its research needs, a laboratory must build or acquire computational infrastructure. As illustrated in Fig 1. the most basic capabilities include computation, storage, and network bandwidth. These resources are managed by an operating system, which also provides simple mechanisms for coordinating application requests (e.g., to register and invoke services) and for enforcing policy. On top of the operating system, one layers complex generic infrastructure (such as a database management system–DBMS, catalog, digital library, or workflow manager) and complex policies. Uniquely biomedical infrastructure (e.g., BLAST) leverages this generic infrastructure. Finally, one deploys biomedical applications built atop the underlying layers.

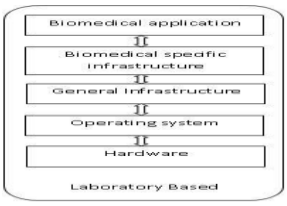

Fig 1. Infrastructure employed at local laboratories

B. Grid infrastructure for consortia

Grid technologies have proved useful in the scientific community, enabling researchers to employ computation, data, and software across a range of machines. Surveys appear in [14], [15], and [16]. Underneath the interface that consumers see, grid implementations typically connect independently owned and geographically distributed servers. Naturally, there is also a need to federate across several grids or clouds [17].

Some notable grids use machines volunteered from the general public to provide cheap computational power for long-running computations that require more resources than one institution can afford, e.g., large, decomposable problems in protein folding or astronomical signal analysis [18] and [19]. The price is unbeatable (machine time is free, the grid software is open source, and Internet traffic is cheap). However, this approach does not guarantee fast response, or provide robust, always-available storage. Worse, it cannot be used with sensitive data – since an untrustworthy host machine can easily bypass grid security [20].

Several biomedical consortia have built their own grids, federating the data and applications contributed by their members. Such grids often employ sophisticated open source software such as Globus for computation [21] and the Storage Resource Broker for large data sets [22]. Such grid software offers substantial management capabilities, such as catalogs for discovery (e.g., find images based on metadata values), and mechanisms for ensuring data security and privacy. The catalog and security services face demands (unmet in some initial releases) for high availability and for rapid scale-up to handle surges when large numbers of new images need to be registered and processed. As they mature, clouds will be an attractive candidate. Grids also often support such as sequence similarity search [23] or image processing [24], tasks that require substantial computational power.

C. Clouds

Cloud vendors effectively sell computation and storage resources as commodities, providing users with the illusion of a single virtual machine or cluster, implemented over thousands of the vendor's computers (in some cases, virtual and physical machines correspond 1-to-1). Some cloud vendors and third parties sell higher-level resources, such as the GoogleApp application platform, relational DBMSs [25], or the SalesForce application. Underneath, the virtual resources are mapped transparently to the underlying physical resources, optionally subject to constraints on geographic location (e.g., replicate at a remote site, but stay within the European Union). The customer controls the virtual machine's capacity (computational and storage) by sending the cloud vendor a service request to add or subtract resources as needed. The time to gain or release capacity (for small fractions of the provider's inventory) is typically measured in minutes, not months.

Fig 2.illustrates graphically the layers that cloud offerings often allow to be offloaded. Note that this diagram is essentially identical to the server architecture described above in Fig 1. The difference lies in who is responsible for providing the lower-level capabilities.

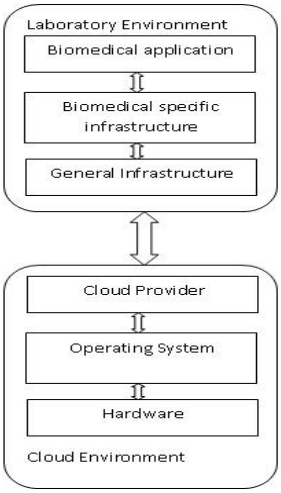

Fig 2. Cloud based biomedical application

A cloud provides a base upon which customers build their own applications. The general infrastructure layer provides

capabilities needed by application builders (e.g., databases) and system administrators (e.g., security mechanisms). The next layer provides capabilities widely needed in biomedical informatics. Finally, each laboratory will need to add capabilities and applications to meet its own needs. As Fig 2. shows, many additional layers of capabilities still need to be provided by a consortium, a system integrator, or biomedical software environment vendor. Regardless of the underlying infrastructure, customers still need to provide everything specific to their own application.

IV. CLOUD INFRASTRUCTURE FOR BIOMEDICAL CONSROTIA

As discussed above, biomedical researchers are beginning to rely on consortium grids, due to the difficulties of managing laboratory silos when researchers from multiple institutions need to share data. However, laboratories still acquire their consortium-support hardware conventionally, with substantial delays, need for physical space, and limited economy of scale. They still face the management difficulties of either heterogeneous underpinnings or being forced to acquire uniform systems. Labs small resource pool makes it hard to rapidly increase or decrease capacity.

Clouds offer many management services similar to grids, but their underpinnings have a "mass production" flavor. They typically use large data centers with many thousands of processors, acquired and managed by one organization, often kept fairly uniform. Within a data center, the network bandwidth is usually high, allowing the underlying computers to share data with one another efficiently (though not as fast as a specialized cluster).

The cloud can be owned either by the vendor or for private clouds, possibly by the customer organization. Compared with scientific data centers, clouds offer economies of scale and the ability to adjust to workload variations. They have attracted wide interest, going beyond the scientific community.

V. DESIRABLE FEATURES OF CLOUD

In this section we will provide insight on various desirable features of cloud.

A. System administration

Low-level system administrative costs can be quite high for laboratory systems scattered around an institution, often far greater than raw hardware costs. A cloud lets an organization offload three sorts of *low-level* administration. First, the cloud vendor is responsible for system infrastructure. Second, once a backup policy is specified, the cloud vendor executes it. Finally, an application can be installed once, and becomes available to all authorized users. At higher levels, administrators deal with many application-support and upgrade issues, as well as user management. Moving to a cloud should not greatly change such work, so in keeping with our "relative" approach, we do not include it.

In severe cases, the low-level administration costs can be greater than the *total* cost for a cloud service.

B. Idle capacity

In conventional systems, system resource utilization is low, estimated at 15–20% for data centers [26]; other estimates are lower. There are multiple causes for low utilization. Systems managers tend to buy for near-peak and future loads, and thus do not use the whole capacity all the time. Differences in work schedules and project maturity will lead to peaks and valleys. In contrast, a cloud (or institutional data center) smoothes these effects across many customers, and today may attain 40% utilization [27], with higher values plausible in clouds (e.g., as load sharing over time zones becomes more mature, and exploiting more diverse user bases). One virtual server seems likely to do the work of at least 2.5 typically-utilized servers. We expect similar figures for bandwidth utilization. For storage, the utilization savings will be less dramatic—data must be stored even when not in use.

C. Power usage and facilities

Server power is expensive, while cooling and other overhead power consumption is assessed to be at least comparable [28]. Together, they at least equal server purchase costs, for typical servers today. Cloud vendors can do much better than the typical laboratory, or even institutional data center, based on better management of voltage conversions, cooler climates and better cooling, and lower electricity rates (cloud vendors tend to cluster near hydropower). They also often locate where real estate is cheap.

D. Less to manage

Today, managers of laboratories or biomedical consortia need to manage physical systems, capital expenditures, and acquisitions of multiple kinds of hardware and software. This task can become significantly simpler when hardware and network acquisition, maintenance, and management are offloaded to the clouds. For physical security, outages, or disaster recovery, the laboratory or consortium must specify a level of service and a vendor capable of implementing it (vendors, like in house staff, must be chosen carefully, and are fallible).

E. Scalability

When the workload experiences significant change, a cloud can add or release resources in minutes. A cloud can provide extra processing resources during the peaks (within limits) when the transaction load spikes. One can improve response time on large, parallelizable tasks by applying many servers, as opposed to running a single laboratory server for hours. Further, one pays for resources actually used, not for capacity.

F. Superior resiliency

Cloud vendors store backups of users' applications and data in multiple geographical locations. If a machine fails, others can take over, at the same location, or between locations (for disaster recovery).

A laboratory that implements its own fault tolerance and disaster recovery requires management effort (mentioned above); additional software, hardware, and space and additional risks (users who manage recovery poorly may lose all their data, e.g., in a flood). A cloud potentially reduces all three. Even for a laboratory that opts to retain its own servers, a cloud can still be useful for archiving and remote data backup.

G. Homogeneity

A consortium system implemented in a cloud can give all authorized investigators access to the same tools, such as workflow tools to process images taken from biomedical scanners. In contrast, peer to peer sharing without consortium managers is unlikely to provide all relevant tools, and keep them up to date. In a grid implemented over a heterogeneous environment, the consortium cannot easily manage tools that run natively over the different operating systems. Alternatively, while a consortium grid built over homogeneous lab-hosted resources can distribute and manage tools effectively, the dedicated system increases cost and will deter translational science collaborations that need only occasional access.

VI. SECURITY OF DATA STORED IN CLOUD

Security is one of the major concerns when laboratories consider moving sensitive information to machines they do not own [29]. This section examines the security impact of outsourcing a laboratory's data to either a data center, to a cloud, or to a conventional managed consortium grid over lab-hosted systems. We emphasize confidentiality, because that seems the greatest barrier to sharing arrangements; however, some comments also apply to other aspects of security (integrity, denial of service). We find that some risks decrease and some increase, with neither side of the argument overwhelming the other. Thus, each laboratory or consortium will need to assess security for its environment.

VII. CONCLUSION

Cloud architectures for biomedical informatics is introduced to users who may wish to build applications using a cloud, and for investigators who want to share data with collaborators. The previous sections demonstrated that hosting on clouds sometimes offers large benefits, significant flexibility and ease-of-administration benefits, and comparable security. It seems desirable to begin pilot efforts in which organizations examine the most current cloud offerings. Decision criteria need to go beyond straightforward dollar costs, to include risk reduction (e.g., of data loss or service unavailability), increased flexibility and scalability, and protection of an institution's other systems. We reiterate that the biomedical organization retains the right to set and enforce its own sharing policy. Many observers believe that clouds represent the next generation of server computing.

REFERENCES

[1] Arnon Rosenthal, Peter Mork , Maya Hao Li, Jean Stanford, David Koester, Patti Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing," Journal of Biomedical Informatics,vol 43, April 2010 , pp 342-353.

[2] Johannes Wagener1, Ola Spjuth2, Egon L Willighagen2 and Jarl ES Wikberg2 ," XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services," BMC Bioinformatics 2009, vol 10,pp 279-281.

[3] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J" Database resources of the National Center for Biotechnology Information," Nucleic Acids Res 2009, vol 37,pp D5-15.

[4] Miyazaki S, Sugawara H, Ikeo K, Gojobori T, Tateno Y," DDBJ in the stream of various biological data," Nucleic Acids Res 2004, vol 32, pp D31-34

[5] Labarga A, Valentin F, Anderson M, Lopez R," Web services at the European bioinformatics institute," Nucleic Acids Res 2007, vol 35,pp W6-11.

[6] Sugawara H, Miyazaki S," Biological SOAP servers and web services provided by the public sequence data bank," Nucleic Acids Res 2003, vol 31,pp 3836-3839.

[7] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P,"Taverna: a tool for the composition and enactment of bioinformatics workflows," Bioinformatics 2004, vol 20,pp 3045-3054.

[8] Fiers MWEJ, Burgt A, Datema E, de Groot JCW, van Ham RCHJ," High-throughput bioinformatics with the Cyrille2 pipeline system." BMC Bioinformatics 2008, vol 9,pp 96.

[9] Hoon S, Ratnapu KK, Chia J, Kumarasamy B, Juguang X, Clamp M, Stabenau A, Potter S, Clarke L, Stupka E,"Biopipe: a flexible framework for protocol-based bioinformatics analysis," Genome Res 2003, vol 13 ,pp 1904-1915.

[10] Fisher P, Hedeler C, Wolstencroft K, Hulme H, Noyes H, Kemp S, Stevens R, Brass A," A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis." Nucleic Acids Res 2007, vol 35,pp 5625-5633.

[11] Markram H. Industrializing neuroscience. Nature 2007;445:160–1.

[12] Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: needs and barriers. JAMIA 2007,vol 14, pp 478–88.

[13] Ross JW, Westerman G.," Preparing for utility computing: The role of IT architecture and relationship management." IBM System Journal 2004,vol 43,pp 5–19.

[14] Special issue on life science grids for biomedicine and bioinformatics. Future Generation Computer Systems 2007,vol 27.

[15] Krasnogor N, Shah A, Barthel D, Lukasiak P, Blazewicz J, "Web and grid technologies in bioinformatics, computational, and systems biology: a review." Curr Bioinf 2008, vol 3,pp 10–31.

[16] Special issue on grid technology in biomedical research. IEEE Transaction Information Technology Biomed 2008,vol12.

[17] Buyya R, Ranjan R, guest editors," Special issue on federated resource management in grid and cloud computing

systems." International journal of grid computing: theory, methods, and applications (FGCS), Elsevier Press;2009.

[18] Vijay SP, Baker I, Chapman J, Elmer S, Larson SM, Rhee YM," Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing," Biopolymers 2002,vol 68,pp 91–109.

[19] Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D, "SETI@home: an experiment in public-resource computing," Comm ACM 2002,vol45,pp 56–61.

[20] Krefting D,"Medigrid: towards a user friendly secured grid infrastructure,"Future Generation Computer Systems 2009,vol 25,pp 326–336.

[21] Foster I, Kesselman C,"Globus: a metacomputing infrastructure toolkit." Int J Supercomputer Applications 1997,vol 11,pp 115–28.

[22] Moore R, Sheau-Yen C, Schroeder W, Rajasekar A, Wan M, Jagatheesan A,"Production storage resource broker data grids. "Second IEEE international conference on e-science and grid computing, Dec., 2006, pp. 147.

[23] Stephen F. Altschul,Warren Gish,Webb Miller,Eugene W. Myers,David J. Lipman," Basic local alignment search tool." Journal of Molecular. Biology.1990,vol 215,pp 403–410.

[24] Sharma A, Pan T, Cambazoglu BB, Gurcan M, Kurc T, Saltz J," VirtualPACS – a federating gateway to access remote image data resources over the grid," Journal of Digital Imaging 2009,vol 22,pp 1–10.

[25] Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A,"Above the clouds: a berkeley view of cloud computing." EECS Department, University of California, Berkeley Technical Report No. UCB/EECS-2009-28.

[26] Evdemon J, Liptaak C. Internet Scale Computing: MSDN Blog, Oct 17, 2007.

[27] Vogels W. Beyond Server Consolidation. Queue 2008;6:20–26.

[28] Belady CL. In the data center, power and cooling costs more than the IT equipment it supports. Electronics Cooling 2007.

[29] Langella S, Hastings S, Oster S, Pan T, Sharma A, Permar J," Sharing data and analytical resources securely in a biomedical research grid environment." JAMIA 2008,vol 15,pp 363–73.