Cloud Computing: A New Business Paradigm for Biomedical Information Sharing

Arnon Rosenthal, Peter Mork, Maya Hao Li, Jean Stanford, David Koester, Patti Reynolds

The MITRE Corporation[1]

Bedford, MA, McLean VA, USA

{Arnie, pmork, jstanford, haoli, dkoester, preynolds}@mitre.org

Abstract

We examine how the biomedical informatics (BMI) community, especially consortia that share 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. Substantial economies of scale potentially yield costs much lower than dedicated laboratory systems or even institutional data centers. Overall, even with conservative assumptions, for applications that are not I/O intensive and do not demand a fully mature environment, the numbers suggested that clouds can *sometimes* provide major improvements, and should be seriously considered for BMI. Methodologically, it was very advantageous to formulate analyses in terms of component technologies; focusing on these specifics enabled us to bypass the cacophony of alternative definitions (e.g., exactly what does a cloud include) and to analyze alternatives that employ some of the component technologies (e.g., an institution’s data center). *Relative* analyses were another great simplifier. Rather than listing the absolute strengths and weaknesses of cloud-based systems (e.g., for security or data preservation), we focus on the changes from a particular starting point, e.g., individual lab systems. We often find a rough parity (in principle), but one needs to examine individual acquisitions—is a loosely managed lab moving to a well managed cloud, or a tightly managed hospital data center moving to a poorly safeguarded cloud?

Keywords: Cloud computing, Data sharing, Bioinformatics, Security, Distributed computing, Cost-benefit analysis

Article Outline

[1] Approved for Public Release, Distribution Unlimited. To appear in Journal of Biomedical Informatics, April. 2010. This is a preprint version.

1. Introduction

"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, we describe the fundamentals of cloud computing 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 a 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). The consequent dollars and delays are often ignored in purchase decisions, but can be very substantial.

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.

For consumers, cloud computing is primarily a new *business* paradigm, as opposed to a new *technical* paradigm; a cloud vendor (a commercial company) provides hardware, a software infrastructure (platform), or an application as a service to its customers. In the simplest scenario, a cloud vendor allows its customers to gain the capabilities of a simple server—albeit a virtual one—in which the processing, network, and storage resources are controlled dynamically. More sophisticated clouds also provide useful datasets (e.g., genomic or census data), management capabilities, programming environments (e.g., **.**Net in Microsoft Azure), web service platforms (e.g., Google App Engine), or access to particular applications (e.g., BLAST [1]). Cloud users can acquire or relinquish processing power and storage, often in minutes, merely by sending a service request to the cloud vendor. The server (or storage, or communication channel) is "virtual" in the sense that the vendor provides capacity as needed—e.g., a server, or slice of a server, from its pool of machines.

The goal of this paper is to help decision makers 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.

Two aspects of our analysis bear mentioning. First, we steer around the un-resolvable debate about where to draw the boundary between "cloud" and "not a cloud" (or "grid" and "not a grid"). Authors have different concerns, and will persist in drawing different boundaries. Also, definitions involve a list of inclusions and exclusions, which a reader is unlikely to recall. So we present a feature list, rather than absolutely requiring or forbidding features. Technical analyses refer to systems having or lacking a particular feature, regardless of whether that system is categorized as a cloud, institutional data center, or consortium grid. The features are useful as information retrieval keywords—we call a system a cloud if it has a preponderance of the features that authors emphasize in systems *they* call clouds. Second, we clarify discussions of both costs and security by employing a *relative* approach. That is, rather than list pros and cons of clouds in isolation, we consider "before" and "after." By identifying issues that are not substantially changed, we greatly reduce the scope of comparison.

In Section 2 we present background information on grids and clouds. Section 3 provides an overview of consortium computing. Section 4 discusses cloud infrastructure for medical consortia and describes sample cloud vendors. The next two sections contain the central evaluations. Section 5 evaluates several different tradeoffs, and Section 6 discusses cloud security, a major concern of many potential adopters. Section 7 identifies properties that make a project amenable (or not) to cloud computing, and Section 8 presents conclusions.

2. Background

Powerful instruments, satellites, and sensor networks can easily generate terabytes to petabytes of scientific data in a day [2]. 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 era of spreadsheet-based research data storage is approaching its limits [3].

2.1 Distributed System Architectures

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 highly touted recent phenomenon. As noted, there is little hope of obtaining consensus or a standard definition regarding exactly what constitutes a “cloud” (and the term “grid” has been similarly overloaded). For example, [4] emphasizes quality of service contracts for a cloud, [5] contrasts social issues with technical infrastructure, while others focus on price or on the nature of the resources provided (e.g., storage, processors, platforms, or application services). Some writers emphasize what the cloud provides to its consumers, e.g., services on demand. Others emphasize what is underneath—a warehouse full of servers. No single definition is “best” for all purposes.

2.2 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 out-sourcing*: Instead of a consumer providing their own hardware, the cloud vendor assumes responsibility 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 [6].
- *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. Some clouds monitor their offerings for malicious activity.
- *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.
- *Parallel computing*: Map/Reduce and Hadoop are frameworks for expressing and executing easily-parallelizable computations, which may use hundreds or thousands of processors in a cloud. The system coordinates any necessary inter-process communications and masks any failed processes.

3. 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 interconnects 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, such as BIRN and caBIG. In this section, we provide an overview of biomedical computing infrastructure, paying particular attention to the needs of consortia.

3.1 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.

3.2 Biomedical Research Consortia

Today, one typically provides servers within a laboratory; institutional data centers provide a second option. However, a single institution cannot provide all the needed resources, and collaborations go beyond its boundary. The complexity of deploying computational infrastructure, especially across multiple institutions, has encouraged creation of many independent biomedical consortia to facilitate sharing data and software among labs. The consortium provides the skills and resources needed to support a rich set of capabilities, offloading some work from the laboratory. Individual laboratories can then focus on extending the higher, biomedical-specific layers.

Traditionally, these consortia have contributed to all layers of the computational stack. As surveyed in the next section, frequently, they create a *grid* that provides a unified interface, and some management capabilities, for a large set of machines.

3.3 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 [4], [7], and [8]. 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 [9].

Some notable grids rely on machines volunteered from the general public. These 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 [10] and [11]. 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 [12].

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 [13] and the Storage Resource Broker for large data sets [14]. (Commercial digital library systems from IBM, Microsoft, etc., provide rather similar capabilities to the latter [15]). 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 sequence similarity search [16] or image processing [17], tasks that require substantial computational power. Sometimes the code is tuned to particular processor and interconnect designs, making it difficult to port to other hardware.

The consortium often imposes minimum requirements on the participants' hardware and software configurations. For example, the BIRN requires participants to install standardized hardware racks [18]. These requirements (to be removed in next-generation BIRN) can represent a significant barrier to entry, especially for small laboratories. Overviews of the experiences of the BIRN and caBIG consortium grids appear in [15] and [19]. Several technologies and demonstration systems are surveyed in [20].

3.4 Coping with Institutional Concerns

Institutional authorities need to be satisfied that sharing arrangements are appropriate and secure. Also, institutions may require adherence to hardware, software, or governance standards, which may conflict with the standards required by a consortium. Such constraints can lead to laborious negotiations, delays, missing capabilities, and vulnerabilities. This section describes three major areas that concern institutional authorities, and separates out issues that are unaffected by whether a cloud is used.

Data privacy: The institution is obligated to protect data that it generates or receives from partners. To do so, review boards must ask whether the planned usage for the data is appropriate (e.g., ethical and covered by patient consent), and whether the external recipient seems trustworthy.

We can now provide two substantial simplifications for analyzing the effect of clouds on privacy. First, while vetting the appropriateness of proposed usage is important, it can be handled as a separate process, independent of the mechanisms used to achieve sharing. It will thus not be further discussed. Second, at the top level, we can treat trustworthiness of the sharing mechanism much as we would treat trustworthiness of an external research partner. For example, similar *top level* questions (below) apply to either, "Is a pharmaceutical company in France a trustworthy partner to receive our data?" or "Is a sharing mechanism implemented at a data center hosted in France a suitable recipient?"

We formulate our discussions of Trustworthiness in terms of three questions. First, is the recipient *legitimate* (i.e., do we think they mean well)? The recipient's reputation, including organizational affiliation and certifications, may guide such decisions. Harvard or IBM might be acceptable, respectively, for research or cloud; unknown unaffiliated researchers or startup companies might not. Second, to avoid misunderstandings, has the recipient made appropriate promises (accepted *obligations)* about degree of system protection and about enforcing the owner's policy about sharing the data onward? (A recipient laboratory might simply promise not to pass the data onward, but a sharing mechanism will need to enforce a complex policy. Each recipient might be required to maintain firewalls, to limit staff access, and conduct regular audits). Third, are the recipient's technical and human systems *able* to meet their obligations to protect data against attacks and carelessness?

Protecting other systems: When a lab hosts consortium or other externally-accessible resources, external traffic must traverse the institution's networks and firewall.[2] This traversal increases risks of congestion and malware, especially if the firewall is loosened to accommodate the traffic (e.g., to allow database accesses from outside the institution). Also, whenever the consortium's services and membership expand, risks may need to be reexamined.

Efficiency and standards: Institutions often seek to reduce costs by reducing heterogeneity. For example, site licensing agreements or chief information officer (CIO) mandates at one institution may require Oracle databases on Sun servers. These institutional policies may conflict with consortium requirements to use PostgreSQL on HP. If a laboratory does not get the necessary waivers, the multi-institution data-sharing consortium will thus have heterogeneous hardware and software. Some applications may fail, or run very slowly, and extra costs will be incurred for training, software conversion, and configuration management.

4. 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 [21], 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.

Like a lab's cluster from Sun or HP, 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.

4.1 Cloud Infrastructure for Biomedical Consortia

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

[2] A *firewall* prevents unwanted traffic from crossing a perimeter, usually by filtering a message header based on local policy. Firewalls understand networks, ports, and servers, but not individual users or stored data items.

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). Public clouds contain data from multiple customers and problem domains; the consequent security tradeoffs are discussed in Section 6. The cloud can be owned either by the vendor (creating control and legal issues, discussed in Section 6.3), 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.

4.2 Sample Cloud Vendors

We now provide sample data points—gleaned from company announcements, blogs, and other sources—about current cloud capabilities and the directions cloud computing seems to be headed. Of course, the landscape of offerings is likely to change rapidly. Clouds are offered externally, or used internally, by the following:

- **Internet companies** may offer space for rent on clouds they run to support their normal operations or create new clouds for customer use.
    - Amazon, the current leader, sells virtual servers on its cloud (EC2) [22], along with simple message queuing (SQS) [22], file space (Simple Storage Service -S3 [23]), an n-tuple store (SimpleDB) [24], an announced UNIX file system, and several other services [25]. These support commonly used virtual machines (e.g., Linux, Windows), can run many popular software products (e.g., databases, though performance needs deeper investigation), and present an idiosyncratic interface for storage and management.
    - Other Internet companies such as Google [26], Yahoo! [27], and Microsoft MSN [28] already use clouds to support their own operations [29], including extensive parallelism. Some of their publically available cloud applications (e.g., search, gmail) were written to match their own clouds' interfaces (e.g., Google's cloud facilitates parallelism). Multiple such interfaces are expected. IBM and other major vendors are expected to offer Amazon-like infrastructure capabilities, together with enterprise-quality management, security, and robustness.
- **Enterprise-internal clouds:** Many computer companies are expected to help large enterprises set up their own clouds, internal to their own firewalls. Such an arrangement may alleviate worries about control and liability (e.g., requirements of the Sarbanes-Oxley law), but will not help facilitate cross-institutional data sharing. The US Department of Defense has contracted to create a private cloud that follows military security practices [30]; advocates tout improvements in speed of procurement. Hybrid

clouds may soon federate a public cloud with a private cloud that hosts more sensitive data.

- **Small players:** Several relatively small companies host clouds already, as well as help enterprises acquire their own clouds. For example, 3Tera claims to provide many management services absent in Amazon and to already host MySQL comfortably [31].
- **Application providers:** Rather than running their own server farms, these companies and consortia provide versions of their products that run on clouds. DBMSs[3] now available on the cloud include Oracle, DB2, Vertica, and MySQL. On the other hand, the robust, distributed S3 storage poses problems for DBMS capabilities [32]. For parallel computing, there is Apache's Hadoop, an open source analog of Google's MapReduce parallelization facility. This facility allows one to easily deploy a highly parallel biomedical research service such as BLAST [16].

The cloud vendor's business proposition is that, as a service provider (e.g., Google, Microsoft, Amazon, IBM, or a smaller player), they can buy, power, manage, and repair a massive array of rather uniform servers in a large warehouse, at a much lower unit cost than can a single university, or consortium that spans geographically distributed laboratories.

Our specimen cost analysis below shows that this cost proposition is very plausible. Current prices for resources on commercial clouds are very attractive for some applications, and our calculations suggest that these prices are based on real low costs, not marketing ploys. [21] and [33] suggest even larger savings. The technical strengths and emerging competition suggest that these favorable trends will continue [34]. Nonetheless, there are applications where today's clouds are more costly; e.g., Amazon charges heavily for moving data on and off the cloud, and if inactive users remain connected, continues to charge for their virtual machines.

The choice is not binary. An institutional data center exhibits some cloud characteristics (e.g., virtualization, services on demand, collocated servers) that may sometimes be an attractive alternative to laboratory-based computing, especially when data is not shared with outsiders. They may offer greater local knowledge and perhaps lower communication costs and fewer legal complications, and are considered in our tradeoff discussions below.

5. Evaluating the Tradeoffs of Using Clouds

Advocates expect that clouds will soon become the default way to host highly flexible shared data repositories. Still, each organization must perform a comparison for its needs. This section describes areas where an organization needs to understand and evaluate the changes that a cloud would bring them—dollar costs to be considered (Section 5.1), and qualitative changes, such as reducing delay in expanding a sharing arrangement (Section 5.2). Security comparisons appear in Section 6,

5.1 Capacity, Often at Low Cost

This section examines three major cost drivers: system administration, idle capacity, and power usage and facilities. At each step, we provide specimen cost figures for conventional systems, extracted from our organization and from web postings. The specimen analysis is a coarse approximation, because environments vary greatly, e.g., electricity rates can differ by a

[3] Since DBMS efficiency and failure tolerance depends on low level interactions with disks, one must both run performance benchmarks and ensure that virtual disks truly persist.

factor of four within the USA, and administrative loads per server differ enormously. Our calculations assume *very* conservatively that research organizations procure hardware, bandwidth, and facilities (buildings and power) at the same price as cloud vendors. Others, with access to more detailed data, have estimated factors of roughly 5 to7 in favor of giant purchasers (such as cloud vendors) [33]. However, our sample organizations did have relatively high administration costs; others may do better. With these figures, we see a very large gap (factor of three) in underlying costs between cloud-based and conventional solutions. We conclude from this rough analysis that, despite our plentiful margin of error, the fundamentals seem very favorable as an alternative to new laboratory machines; well managed data centers fall somewhere in the middle.

5.1.1 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 (the lower levels of Fig. 1—hardware maintenance, spare parts, adding new machines, and infrastructure software). 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.[4] 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. We describe several data points for specimen low-level administration costs, assuming salary cost of $100K per administrator staff year. Administration costs seem most significant with either loose management, volatile requirements, or hardware scattered around many rooms on a campus.

- Using anecdotal evidence about some MITRE systems, we estimated that 1/3 of administrators' time is spent on low level administration. The 2/3 spent on user management and local applications is excluded from our cost estimates. This facility supports prototyping projects, and their frequent reconfigurations may account for a relatively high cost. Each administrator handled about 30 processors, so low level infrastructure and software distribution work comes to 1.1% of a staff year per server, or $1.1K per server year. (Assuming a three year server lifespan, low level administration costs slightly more than the hardware.)
- One government organization has about 8 servers per administrator. Assuming the same 1/3 ratio of low level administration, this costs $3.75K per year per server.[5]
- The BIRN consortium suggests that backup will consume 10% of an administrator per rack[6], and that hardware maintenance will cost extra. Software distribution is managed efficiently by the central staff across dozens of homogeneous racks, and costs little. (A multiprocessor rack, switching, and cabling may come to $30K, while just the backup component of low level administration over a three year life matches this figure). Our cost estimate is conservative, omitting several costs that were not publicly reported—power, hardware setup

[4] Note that open source or user-developed applications may be hosted in this way. Business models for licensing commercial applications (such as Oracle) on the cloud are immature and evolving.

[5] The organization is rolling out a new offering, which should be more efficient.

[6] http://www.nbirn.net/cyberinfrastructure/acquire_rack.shtm downloaded 7/15/08

and maintenance, and negotiating institutional firewall issues. We estimate administrative costs as being at least equal to the purchase cost of a server.

For some laboratories, our estimates of current practice may be pessimistic. Hamilton [35] estimates 140 servers per administrator for moderate scale institutional data centers (much less than hardware costs). There are also qualitative advantages to local staff, who understand people, practices, and priorities. However, institutional centers still represent a loss of control by the laboratory. Also, for an organization experiencing high costs, advice to get better management and more skillful staff in the lab is hard to follow. Many labs may find it preferable to outsource to institutional data centers or clouds, for more professional management.

5.1.2 Idle Capacity

In conventional systems, system resource utilization is low, estimated at 15–20% for data centers [36]; 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. (The analysis in [21] adds an extra charge for requests that were not served because load exceeded capacity). In contrast, a cloud (or institutional data center) smoothes these effects across many customers, and today may attain 40% utilization [37], 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.

5.1.3 Power Usage and Facilities

Server power is expensive, and cooling and other overhead power consumption is assessed to be at least comparable [38]. 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.

5.1.4 Specimen Cost Comparison

We now give a specimen analysis of the cost of supporting a biomedical application on Amazon web services. Echoing many others, we conclude that cloud computing is already very cost-effective in some settings. When one reaches an acquisition stage, one needs to redo the cost calculation for the specific system being built, and with current cost quotes from cloud vendors, and then bring in qualitative and security issues.

Consider a grid that includes 23 TB of data and 60 processors, with uploads of 40 GB per month and downloads of 13GB per month—roughly comparable to the size of the system managed by BIRN. A conventional system needs 60 processors that cost approximately $1K per year, or $60K total, in early 2009. Storage for 60TB costs about $6K, or only $2K per year. Assuming that one administrator can manage 30 machines (and that one third of the administrator's time is spent on low-level maintenance), there is an additional maintenance cost of $66K per year. The purchase and administration cost of a conventional system is $128K each year billed to the laboratory, plus an additional ~$60K in energy costs (though these may be hidden in institutional overhead) and undetermined costs for space and network bandwidth.

Of course, many of the processors are frequently idle; assuming 16% utilization (vs. 40% for a cloud), only 24 processors would need to be rented from a cloud vendor. Using Amazon's online EC2 calculator [22] in May 2009, a cloud-based system would cost $3.4K per month for data storage and bandwidth (uploads and downloads). The processors cost an additional $1.7K per month. Thus, the cost of using the cloud is $61K each year, which includes hardware, power, operating system, basic security and infrastructure administration, backup of the persistent store, and application replication.

Though this cost comparison is an estimate, it demonstrates that for new systems, clouds' rental costs look quite attractive. Even omitting power costs, our specimen estimate shows clouds to be superior by roughly a factor of 3 for providing infrastructure and replicating applications.

5.2 Qualitative Benefits

This section addresses ways in which a system built using clouds can reduce the burden on laboratory managers, be more scalable and resilient (so users get better service), and make it easier to share data and tools.

5.2.1 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 as illustrated in Fig. 2. For physical security (protecting your disks from theft), 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). The net effect, subject to caveats in Section 6, is that the systems burden on principal investigators or consortium managers is reduced.

Chargeback policies are a complex area, and we will not examine them in depth. Whatever policy is chosen, explicit charges per use make it more transparent, but managers may wish to impose limits.

Laboratories still have the right, and the requirement, to manage who accesses their virtual machines. To do so, they may employ firewall, authentication and authorization systems from the cloud vendor, or, for greater sophistication, from third parties (as applications on their virtual servers and virtual firewalls).

5.2.2 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 (such as for access to Swine Flu clinical data). 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.

However, some users have had unpleasant surprises about costs associated with unexpectedly heavy use of cloud resources. With conventional hardware, one knows how much money is committed; with resources on demand, programs may spend unexpectedly large sums of money if I/O volume is unexpectedly high, or users silently fail to release unneeded servers and storage. These effects are difficult to monitor. We expect some cloud vendors to offer suitable throttling services soon; until then, administrators need to be vigilant.

5.2.3 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 beyond those included in the "conventional" costs in Section 5.1; and additional risks (users who manage recovery poorly may lose all their data, e.g., in a flood). A cloud potentially reduces all three[7]. Even for a laboratory that opts to retain its own servers, a cloud can still be useful for archiving and remote data backup.

5.2.4 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.

5.2.5 Fewer Issues to Negotiate with Institutional Authorities

We now reconsider the concerns raised in Section 3.4, from the perspective of cloud computing. The institution's concern that noncompliant products in a lab may increase the cost of institutional support does not apply when the products are instead part of an externally hosted consortium service, so no negotiations will be needed.

Negotiations about protecting other systems in the lab or the institution are likely to be significantly reduced. When consortium resources are hosted inside the institution, traffic involving those resources may put other systems at the institution at risk. The lab may need to negotiate exceptions from the institution's firewall to allow the traffic in, and to negotiate increases in institutional bandwidth. Unfortunately, if the lab gets its way, the institution's firewall protections are weakened and congestion may result. If the lab cannot negotiate the changes, data sharing is blocked. Either way, both researchers and institutions must devote substantial time and skill [12], [39], and [40], and collaborative research must wait.

Cloud-hosted resources cut the Gordian knot by keeping the new, potentially malicious traffic *outside* the institution, benefiting both the institution and the laboratory, reducing both risk and negotiations. In the same vein, no negotiation is needed if computations on the cloud wish to employ other services available externally, e.g., data mining or BLAST. Yet another positive scenario results if the lab hosts computations on external researchers' sensitive data. In fact, one may wish to reorganize workflows to minimize traffic impinging on the various institutions.

[7] Amazon's cloud has experienced well publicized downtime. While this may be a sign of immaturity, an acquirer should certainly look at their vendor's track record. We are not aware of any loss of persistent data.

Hosting data externally avoids the risk that external requests will place a heavy load on the institution's network. There are two small countervailing factors. First, the laboratory needs some bandwidth to post its data to the cloud. Also, if a laboratory needs to perform extensive *internal* processing of the cloud-hosted data, it may keep a local replicate to avoid transferring the data repeatedly. Fortunately, Post traffic requires only that the institution supply low priority bandwidth (batch is tolerable), and the storage cost for replication is low (Section 5.1.4). Thus, cloud-hosted systems seem to require less negotiation of bandwidth.

One also needs to negotiate firewall policy changes just enough to allow data and security information to be sent to the laboratory's own virtual machine on the cloud. This opening seems much narrower than allowing a variety of service calls from a variety of partners. Again, the need for negotiation seems reduced.

A laboratory may then take advantage of a cloud to add collaborators more rapidly. New collaborators no longer require greater internal processing resources, nor do they need to negotiate bandwidth increases and firewall changes. As mentioned in Section 3.4, the cloud does not remove a laboratory's responsibility to manage who can access what resources. Security policies and enforcement software are a necessary part of the infrastructure and need attention from the laboratory, whether on conventional servers or in a cloud.

Service level agreements tend to be more formal with a cloud, unless a customer accepts the provider's default. Thus, outsourcing requires the customer to be more explicit about requirements, and then to negotiate guarantees or choose among the provider's offerings. When systems staff understands the needs, a cheaper informal process might suffice. When problems arise, a laboratory head has great leverage on her staff, but there may be limited machine and human resources to respond, and no explicit guarantees.

The remaining criterion was to "protect the laboratory's data." Trustworthiness of the sharing mechanism on the cloud raises the same top level questions (see Section 3.4) as for a new research collaborator, e.g., how well the recipient protects against hackers. However, institutions may be reluctant to approve hosting in clouds until vendors have accumulated a substantial history, showing no more breaches than ordinary systems. Hence negotiations will increase. The next section further explores data security.

6. Security of Data Stored in a Cloud

Security is one of the major concerns when laboratories consider moving sensitive information to machines they do not own. [41] 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, while also considering the tradeoffs in the previous section.

Our security analysis considers two scenarios that differ in terms of how much is to be outsourced: (1) Just the data and applications intended for external access (while maintaining unshared data locally); or (2) All of the data and applications on the lab server. Intermediate points and redundant hosting are possible, but not discussed.

As the number of partners and shared resources increase, one will face extra labor to manage permissions. There is also extra risk of inappropriate data release, due to having more users who may misunderstand policy or be careless or malicious. However, this increase is not greatly affected by where the laboratory resources are hosted. For example, when an authorized recipient sells patient health records to a tabloid, the problem was not in the technical system. Hence, as in Section 3.2, we omit issues that seem not to vary with hosting.

We decompose the analysis into several parts. Section 6.1 addresses several operational issues. Section 6.2 deals with external intrusions by hackers, a risk that concerns many decision makers but is perhaps not increased as greatly as some think. Section 6.3 examines nontechnical risks of outsourcing from a laboratory, and Section 6.4 summarizes security issues. (See also [42].)

6.1 Security Management

First, a laboratory must continue to manage security. Machines on a cloud still need firewalls, virtual private networks, and so forth. The laboratory will still need to acquire security management software (commercial or open source). Thus, one must examine whether one's chosen security software actually runs well on the cloud, including potential technical or licensing difficulties. Also, one may need additional approvals to place sensitive security metadata (e.g., user identities and relationships) on clouds; products that use encrypted or hashed metadata are to be preferred. On the other hand, outside the institutional firewall, it may be easier to provide access from other institutions. Finally, if requirements are rudimentary, e.g., that all consortium members can share all posted data, they may be able to use cloud vendors' built-in security mechanisms.

Second, system administrators often possess excessive privileges—a significant risk. Compared with laboratories, practices in virtualized data centers (institutional or cloud) are likely to have greater formality, separating the administration of different aspects of a system. In particular, while laboratory administrators and security staff may be allowed to read and change the data they administer, a cloud vendor will tend to treat each customer's virtual machine as a private preserve. On the other hand, institutional and especially cloud administrators will have more difficulty distinguishing illegitimate access or understanding laboratory priorities—outsourcing can break a valuable human network.

Third, physical security protects against threats such as stealing disks or adding tapping devices (attached or remote) to the hardware and networks hosting the biomedical data. On balance, cloud and institutional data centers seem better on this criterion. Data centers are generally quite secure physically, while laboratories' security levels differ drastically. Also, unencrypted CDs and laptops have led to high profile breaches. When data is available on the cloud, there is less impetus for lab personnel to travel with their own copy or to share by shipping a CD. Also, while a large data center (institutional or cloud) is a richer target, targeted attacks within the cloud against a specific laboratory's database are difficult, since it is hard to determine which server or disk holds the data. On the other hand, if one physical machine in the data center is penetrated, eventually it may host something the attacker wants.

6.2 Risks Due to Hackers

Wherever a laboratory stores its data, internally or externally, outside hackers pose a threat. This section considers how the hacker risk and security management labor change if one moves data from a laboratory to a cloud or to an institution’s central data center.

The laboratory will need to decide what hacker risks are acceptable, in return for the other promised advantages. For example, neither clouds nor institutional data centers are as hacker-proof as a laboratory server without Internet access, which does not need to share biomedical data with outside users.

A cloud is shared among many users, at both the macro level (open to many users) and a micro level (multiple virtual resources on each physical one). An institutional data center is also shared, though on a smaller scale. Section 6.2.1 considers risks due to such sharing (called *multi-tenancy*). Section 6.2.2 considers advantages when one splits among virtual machines.

6.2.1 Multi-Tenancy Risks

Virtual machines share physical resources, relying on a software *hypervisor* to keep them appropriately separate. (Multi-tenancy can also arise at the application level, and the application provides the separation among users.) The cloud thus provides less separation than when one has separate servers in a laboratory.

Like all complex software, hypervisors can be hacked, after which an attacker can directly access the shared physical CPU, network, or storage. He then can deny service, destroy data, or steal confidential data. Researchers have demonstrated many ways to hack a hypervisor, and virtualization vendors have provided extensive analyses of ways to reduce the risk [43 and 44]. As of December 2008, no malicious exploits had been reported [45].

A laboratory machine has the significant advantage that an attacker has little legitimate access. An institutional data center or private cloud makes its capability available to many hundred users; a public cloud allows anyone with a credit card to run arbitrary programs. The need to arrange payment is still a barrier against automated, broadcast attacks.

To further assess the risk, note that targeted attacks seeking specific lab’s biomedical data seem the most dangerous. Fortunately, it may be difficult for attackers to know which physical machine to attack, if they are targeting a specific lab’s data. To make it more difficult for an attack that subverts one of a lab’s systems to find the others, one might wish to scatter them to different physical servers, if the virtualization system permits.

Institutional data centers and clouds do have some countervailing defenses. Both are likely to have a professional security staff, unlike a laboratory. Clouds that provide only an application framework with limited interfaces (e.g., just web service calls) are somewhat easier to secure. For comparison, laboratories’ conventional infrastructure—operating systems, DBMSs, and web servers—already have many, many known vulnerabilities. The key, then, is to estimate the *incremental* risk.

6.2.2 Protections at Virtual Machine Boundaries

Security professionals traditionally recommend partitioning a system as a means of protection. One can put a firewall on any laboratory server, wherever the server is hosted. The ease of creating new virtual machines provides ways to improve the security of virtually-hosted data by creating new boundaries. In this section, we examine the utility of partitioning resources into separate areas, conferring protections against attacks that do not break the hypervisor to intrude into other virtual machines.

Consider that if two data items are on the same system, then that system must be accessible to anyone who accesses either item. For sake of example, suppose item D_1 is to be shared with selected collaborators, and D_2 is to be made *publicly* accessible. Now suppose we place D_1 and D_2 on the laboratory's server. Due to our desire to share, especially for D_2, we have vastly increased the set of people who can access the laboratory server. There is increased risk for all the other data on that server. Avoiding this phenomenon may be the greatest security benefit of hosting in a cloud.

If instead one hosted the shared resources on a cloud, there is less risk to the laboratory's other resources. Next, one might be able to partition the resources so that D_1 and D_2 reside on separate virtual machines, each with a more restrictive firewall and fewer user accounts. Now an attacker who reaches D_2 does not threaten D_1. Further, the VM in the cloud is not acting as a general purpose machine, so one can create a firewall that rejects unneeded types of access (ports, protocols, services, etc.); it need only provide for the intended sharing arrangements.

An intermediate approach is to outsource to a new virtual machine in your institution's virtualized data center, proceeding as above. Now the laboratory obtains the benefits, but the institution's risks actually increase, as more users have accounts on its data center virtual machines. Also, the institution's firewall may cause difficulties (as discussed in Section 3.4) while providing only modest protection—that firewall may allow traffic for often-hacked applications (e.g., email) and there are thousands of potentially malicious or playful employees or students inside. The institutional firewall's net security effect can even be negative if the illusion of protection encourages laboratories to neglect their own security measures.

6.3 Nontechnical Outsourcing Risks

To round out the picture, we now describe nontechnical risks to cloud-based systems, and the risks' common sense ameliorations. The ideas here constitute conventional wisdom, not novelty, but are important to consider. Further anecdotes and in depth discussions appear in [42]. Our aim is to show organizations nontechnical threats they need to address, and that these threats can be overcome.

When a laboratory outsources hosting, it (or its consortium) still "owns" its virtual machines and the resources at the cloud or data center. Permissions, resource limits, and priorities must be administered by lab or consortium administrators who can recognize legitimate usage, and have a human network that enables rapid resolution of ambiguities. Still, outsourcing implies loss of control in several ways.

When the cloud provider is a separate company, behavior may become very adversarial. Agreements must be more carefully formalized, especially with respect to business disputes and closure. Until the legal environment matures and standard practices emerge, experience with commercial software provides some useful analogies and practices. First, as a primary protection, choose a cloud provider with a strong reputation and business, not an unknown startup (except perhaps for short term usage). Beyond that, choose suppliers whose contract language suits your needs, in areas such as how they may use your data and request logs, protection from them freezing your data and applications in a business dispute, and a structure that lets them guarantee advance warning before cutting off service (even if they are sued by their suppliers, or go bankrupt). Also, require your provider to provide sufficient documentation so you can port your system to an alternative, if the provider cannot meet their obligations, or if competitors become more attractive.

Multi-tenancy causes several nontechnical risks, in addition to the hacking vulnerabilities discussed earlier. First, it is not yet clear whether the legal system prohibits law enforcers or litigants from seizing a multi-tenant system (by analogy, an apartment building) to punish one of the tenants. We also need to hope that spam filters and other site-reputation services are extended so they can distinguish among tenants and blackball only specific ones that have been alleged to engage in malfeasance.

Next, laboratories may need to restrict where the cloud will physically host their data and applications. For example, they may wish to avoid countries whose governments are intrusive or whose intellectual property laws seem inadequate. Amazon and others have begun providing such controls for their cloud environments.

Finally, academic researchers have argued that before hosting sensitive data externally, one should encrypt it for fear that the data will be stolen or modified by the cloud provider (as a business strategy or rogue staff). The cost of doing so is high—strong encryption makes it difficult to index the data, multiplying access costs. Encryption resists some technical attacks (stealing files), but attackers can still come in the front door, by subverting a legitimate requestor or the access control system. The nontechnical reasons for distrust seem exaggerated. We trust banks not to dip into individual customers' accounts. Analogously, if a cloud vendor were found to be violating their customers' data as a matter of corporate policy, they would instantly lose their business. Their staff may have individual miscreants, but the same is true of a university, hospital, or consortium. Furthermore, the cloud vendor is likely to have better monitoring in place to prevent such activity.

6.4 Summarizing the Security Tradeoffs

Moving data to a cloud *improves* security for the systems that remain inside a laboratory or institution. At a cloud, both data and server backups can be arranged easily; if high availability is required (e.g., for 24/7 sensor data feeds), recovery to a second cloud might be desired. The move also provides strong physical protection of the machines, and enables creation of separate virtual machines and firewalls for each independent laboratory application (or honey pot). The cloud will also firmly separate system administration from data and application administration, and make available a security staff and tools. A well managed virtualized institutional data center will provide all but the first advantage, to some extent. Disaster recovery becomes easier to manage (once one decides how much protection to pay for).

On the other hand, remote administrators may understand less of the local situation, and clouds present large attractive targets. On a public cloud, any attacker with a credit card can establish an account on *some* virtual machine in the cloud, to begin hacking through the hypervisor, a risk that does not apply in conventional systems. The contractual and legal issues become worse with a cloud. Some leading vendors, e.g., Amazon, have not yet demonstrated (or, to be fair, promised) high availability.

Neither approach seems uniformly superior, and experiences are still sparse, but we can highlight a few observations. Risks need to be assessed against the "background" risks: any Internet-connected machine is vulnerable to many attacks, and authorized recipients may fail to protect data. If a laboratory is not sharing its data, replacing an internal server by one on a cloud seems to increase the hacker risk to data confidentiality and integrity—the threat of hypervisor attacks probably outweighs extra security staff. However, if partners already access the laboratory machines, then the benefits of good fences (Section 6.2.2) may outweigh the cloud risks. Overall, the extra risks seem moderate, and may not dominate the cost and convenience issues.

7. Moving Forward

The subsections below respectively consider what makes a good target application for cloud computing, identify some poor targets, and discuss difficulties in the transition process.

7.1 Good Targets for Near-Term Cloud Initiatives

Clouds tends to be preferable when service demands are variable or demand is unknown in advance, and where the cloud vendor passes on large economies of scale in procuring servers, power, and space, and in supplying specialized staff and tools. However, even with favorable winds, one also needs to consider issues of technology insertion. Informed by the above analyses, we identify some promising areas for initial exploitation of cloud technology for bioinformatics. (These recommendations assume the conclusion of Section 6, i.e., that security should not be a show-stopper). The following factors make a project an attractive candidate for cloud computing in the near term:

- The project has high costs for computing, administration, space, and electric power in its current or envisioned state.
- The members wish to share with outsiders, but find that institutional policies block outsiders' access to their local system.
- The project requires highly variable amounts of processing and storage resources. For example, some workloads spike when new data arrives; other sites may suddenly become highly popular (e.g., in the event of an epidemic). In addition, a system that is being reengineered may need extra capacity during development and testing, and later to run the existing and the replacement system simultaneously.
- The system requires off-site backups for data and for processing.
- The applications have easily parallelized code (contrasting with section 7.2).
- One wants long-term repositories to outlive the laboratory that now hosts the data.

General software management criteria apply as well. For example, it is easier to introduce new technology (e.g., a cloud) packaged within a new capability that benefits the biomedical community, such as more secure and rapid data sharing across a consortium. In contrast, users and business managers resist technology-driven replacements of systems they see as running smoothly.

Informed by the above analyses, we identify some promising areas for initial exploitation of cloud technology for bioinformatics. With new technologies, one usually wants to implement new capabilities or solve major existing difficulties. If a system already serves its users satisfactorily and is not being reengineered, the net payoff (after cost of change) will tend to be lower, while resistance may be high. Therefore, below we look at new functionality.

Archiving, backup, and fault tolerance: Whether data are private to a laboratory or shared in a consortium, they need long-term archiving (possibly outliving project funding or the Principal Investigator's career), and protection from permanent failure (e.g., disk crashes) and natural disasters. Even in more routine circumstances, important resources such as catalogs should be able to run in two places, to avoid temporary outages.

Sharing data and tools across a consortium: As discussed above, clouds seem able to support cost effective storage, access, and tool execution, with suitable enforcement of access policies, and easier management.

High performance computing (HPC): Some biomedical applications require extensive computation, often with uneven workloads (e.g., submitting a batch of images). Good candidates for clouds include applications with many small, independent requests where cost is a major driver ("capacity computing"), plus *some* large problems where one wants faster response ("capability computing"). For example, distributed BLAST [1]—and in general, computations where the Hadoop model is appropriate (significant data parallelism and reduction phases with relatively few stages)—are candidates for a cloud. Also, even if the raw computation is unsuitable, one might wish to use a cloud for sharing results, subject to the usual cost and security tradeoffs.

7.2 Less Suitable Targets

For comparison with the highly suitable targets above, this section identifies criteria that make a system a poor candidate for transition to cloud. The last items refer to the environment rather than the system itself.

First, some HPC applications (e.g., protein folding and high-end image processing) exploit detailed physical characteristics of the underlying hardware and require substantial data movement among processors. In a cloud, the physical characteristics of the hardware are not revealed by the vendor. HPC applications that rely on such detailed knowledge are therefore likely to perform poorly.

Second, if one gets unfavorable results from the cost comparison (e.g., with Amazon today, due to heavy network traffic) or the security comparison (e.g., your local staff is highly skilled and trustworthy, and you expect determined attacks on the virtualization software), then clouds are unsuitable. Also, the legal barriers to allowing a third party to manage the data may be insuperable in some situations, at least for now. For very large users, such as a government agency, a private cloud may be an attractive alternative.

Third, if communications fail, the cloud becomes unavailable. For applications that must be highly available *and* that need only local data, a local solution seems better.

Next, cloud advocates may oversell, promoting a vision of perfectly shared data, workflows and repositories, displays, reports, tools, etc. Merely changing how data is hosted will not improve integration among your databases, create new applications, or make investigators (who retain ultimate control) more willing to share data. Achieving all the promised features will take considerable time and management resources and is therefore high-risk. It may be wiser to begin with simple data sharing using off the shelf tools.

Finally, existing projects will have inertia, and will require a major cost advantage to motivate a transition to a cloud. Costs already incurred, ranging from hardware purchase to building a staff, will not be recovered.

7.3 Transition Obstacles

The first big obstacle is the discomfort of stakeholders (scientists and institutional review boards) as two changes are proposed simultaneously: allowing more external sharing and using a cloud as the host. A biomedical researcher does not surrender control of his data by placing them in a cloud—but managing this control will require considerable work, as described in Section 5. Nevertheless, these changes are likely to intertwine in stakeholders' minds, and the separation may need to be explained repeatedly. Other technical obstacles include:

- *Software portability:* Before one switches to a new environment, one needs to ensure that critical applications (biomedical and security) will continue to run, despite technical and licensing issues. This is part of traditional transition planning and cannot be ignored when moving to a cloud. For example, some cloud offerings offer a non-standard programming environment or lack persistent storage. While applications designed natively for a cloud may not have difficulties, existing ones may. Thus, most laboratories and consortia should seek a vendor who offers a close match to conventional UNIX, Linux, or Windows servers.

- *Cloud unfamiliarity and immaturity.* Virtualized data centers, including clouds, require additional skills to maintain security. For example, when virtualizing existing servers, one must not deploy sensitive data on the same virtual machine as widely-accessible data [44]. The products are immature, have experienced outages, and lack some desirable capabilities (e.g., as of mid-2008, Amazon's S3 product does not support firewall configuration based on IP address). However, cloud offerings are improving rapidly; for example, GoGrid claims very high reliability [34]. One will need to identify one's needs and evaluate vendors' track records.

Clouds simplify some management tasks (load projections and capital budgeting) but do require some new management practices:

- *Transitioning to a cloud will change the ways in which biomedical systems are built, managed, and funded.* This change may require that project or consortium PIs expand their skills in contracting effectively, including service-level guarantees and Help facilities for developers.
- *The models used for costing computational acquisitions need to be changed, to better reflect true costs.* When doing cost comparisons, PIs will need to assess the cost of hosting a system in a cloud, and also to expand conventional systems' cost analysis to include oft-omitted costs such as systems administration and facilities (space, electric power, and cooling equipment). Institution-level accounting will also need to change, to account for facilities costs. However, the move to clouds need not await all these developments—in many settings, the cost benefits are sufficient that even a rough analysis will point toward clouds.

8. Conclusion

We introduced cloud architectures for biomedical informaticists 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 financial benefits, significant flexibility and ease-of-administration benefits, and comparable security.

While not definitive, the case seems strong enough to justify management attention from consortium leads, laboratory directors, and university CIOs. It seems desirable to begin funding 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. While one must be cautious with maturing technologies, we expect that clouds will soon be suitable for many biomedical research needs.

Acknowledgement: The paper has benefited *greatly* from the reviewers' suggestions, which were unusually detailed and insightful. A portion of this work was supported by the National Center for Research Resources of the National Institutes of Health, under Contract No. TIRNO-99-D-00005 Task Order No. 20188.

References

[1] Schatz MC. BlastReduce: High Performance Short Read Mapping with MapReduce. Available at http://www.cbcb.umd.edu/software/blastreduce/ and http://www.umiacs.umd.edu/~jimmylin/cloud-computing/speakers/project-presentations.html.

[2] Markram H. Industrializing neuroscience. Nature 2007;445:160-61.

[3] Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, Tarczy-Hornoch P. Issues in biomedical research data management and analysis: needs and barriers. JAMIA 2007; 14:478-88.

[4] Special Issue on Life Science Grids for Biomedicine and Bioinformatics. Future Generation Computer Systems 2007;27.

[5] Szolovits P. What Is a Grid? JAMIA 2007;14:386.

[6] Ross JW, Westerman G. Preparing for utility computing: The role of IT architecture and relationship management. IBM Systems Journal 2004;43:5-19.

[7] Krasnogor N, Shah A, Barthel D, Lukasiak P, Blazewicz J. Web and Grid Technologies in Bioinformatics, Computational, and Systems Biology: A Review. Current Bioinformatics 2008;3:10-31.

[8] Special Issue on Grid Technology in Biomedical Research. IEEE Trans Inf Technol Biomed 2008;12.

[9] 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.

[10] Vijay SP, Baker I, Chapman J, Elmer S, Larson SM, Rhee YM, Shirts MR, Snow CD, Sorin EJ, Zagrovic B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers 2002;68:91-109**.**

[11] Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D. SETI@home: An Experiment in Public-Resource Computing. Comm ACM 2002;45:56-61.

[12] Krefting D. Medigrid: Towards a user friendly secured grid infrastructure. Future Generation Computer Systems 2009;25:326-336.

[13] Foster I, Kesselman C. Globus: A Metacomputing Infrastructure Toolkit. Int J of Supercomputer Appl 1998;11:115-29.

[14] 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 2006; Dec:147.

[15] Moore RW, Rajasekar A, Wan M. Data Grids, Digital Libraries, and Persistent Archives: An Integrated Approach to Sharing, Publishing, and Archiving Data. Proceedings of the IEEE 2005;93:578 – 588.

[16] Altschul SF, Gish W, Miller W. Basic Local Alignment Search Tool. J. Mol Bio 1990; 215:403–10.

[17] 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. J Digit Imaging 2009;22:1–10.

[18] BIRN - Biomedical Research Network. Available at http://www.nbirn.net/.

[19] Oster S. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. JAMIA 2008;15:138-49.

[20] Shah A, Barthell D, Lukasiak P, Blacewicz J, Krasnogor N. Web & Grid Technologies in Bioinformatics, Computational Biology and Systems Biology: A Review. Current Bioinform 2008;3:10-31.

[21] Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M. Above the Clouds: A Berkeley View of Cloud Computing. EECS Department, University of California, Berkeley Technical Report No. UCB/EECS-2009-28.

[22] Amazon Web Services Simple Monthly Calculator. Available at http://calculator.s3.amazonaws.com/calc5.html.

[23] Amazon Simple Queuing Service (SQS). Available at http://www.amazon.com/Simple-Queue-Service-home-page/b?node=13584001.

[24] Amazon SimpleDB. Available at http://www.amazon.com/SimpleDB-AWS-Service-Pricing/b?node=342335011.

[25] Amazon Web services. Available at http://www.amazon.com/Web-Services-AWS-home-page/b?node=15763381.

[26] Google App Engine. Available at http://code.google.com/appengine/.

[27] Yahoo! and Computational Research Laboratories collaborate on cloud computing research. Available at http://www.901am.com/2008/yahoo-and-computational-research-laboratories-collaborate-on-cloud-computing-research.html.

[28] Parastatidis S. The Web as the Platform for Research, Grid Computing Environments (GCE) workshop, SuperComputing 07. Available at http://savas.parastatidis.name/web/talks/2007.11.12%20-%20SC07%20-%20Grid%20Computing%20Environments%20(GCE)%20workshop%20-%20The%20Web%20as%20the%20Platform%20for%20Research.pdf.

[29] Iskold A. Reaching for the Sky through the Compute Clouds. Available at http://www.readwriteweb.com/archives/reaching_for_the_sky_through_compute_clouds.php.

[30] Brygider, J. DISA’s Race to the Cloud, Defense Information Services Agency. Available at http://www.disa.mil/news/stories/cloud_computing.html.

[31] 3tera. Available at http://www.3tera.com/.

[32] Brantner M, Florescu D, Graf D, Kossman D, Kraska T. Building a database on S3. In: ACM SIGMOD International Conference on Management of Data. New York: ACM; 2008; 251-64.

[33] Hamilton J. Cost of Power in Large-Scale Data Centers [online]. November 2008. Available at http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx.

[34] Brodkin J. 10 cloud computing companies to watch. Network World, May 18, 2009. Available at http://www.networkworld.com/supp/2009/ndc3/051809-cloud-companies-to-watch.html?netht=rn_051809&nladname=051809dailynewspmal.

[35] Hamilton J. Internet-Scale Service Efficiency. In Large-Scale Distributed Systems and Middleware (LADIS) Workshop, September 2008. Available at: http://mvdirona.com/jrh/TalksAndPapers/JamesRH_Ladis2008.pdf.

[36] Evdemon J, Liptaak C. Internet Scale Computing: MSDN Blog, Oct 17, 2007. Available at http://blogs.msdn.com/jevdemon/archive/2007/10/24/internet-scale-computing.aspx.

[37] Vogels W. Beyond Server Consolidation. Queue 2008;6:20-26. Available at: http://portal.acm.org/citation.cfm?id=1348590&coll=Portal&dl=ACM&CFID=78225754&CFTOKEN=19192256&ret=1#Fulltext.

[38] Belady CL. In the data center, power and cooling costs more than the IT equipment it supports. Electronics Cooling 2007. Available at: http://electronics-cooling.com/articles/2007/feb/a3/.

[39] Who will build and test your BIRN rack. Available at http://www.nbirn.net/cyberinfrastructure/acquire_rack.shtm.

[40] Welch V, Mulmo O. Using the Globus Toolkit with Firewalls. April 2006, Cluster Monkey. Available at http://www.clustermonkey.net//content/view/122/32/.

[41] Langella S, Hastings S, Oster S, Pan T, Sharma A, Permar J, Ervin D, Cambazoglu BB, Kurc T, Saltz J. Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment, JAMIA 2008; 15: 363-73.

[42] Cloud Security Alliance. Security Guidance for Critical Areas of Focus in Cloud Computing. April 2009. Available at http://www.cloudsecurityalliance.org/guidance/csaguide.pdfhttp://www.cloudsecurityalliance.org/guidance/csaguide.pdf.

[43] Shackleford D, Neal J, Elpers T. Virtualization Security Essentials, (a Configuresoft white paper), 2008. Available at http://www.configuresoft.com/webparts/CMS/ViewDocument.aspx?ItemID=7e86ef55-c94c-476c-8597-39694ec73560.

[44] Amazon Web Services. Overview of Security Processes. September 2008. Available at: http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1697&categoryID=152.

[45] Creeger M. CTO Virtualization Roundtable, Part II. Comm ACM 2008;51:43-49.

Figures

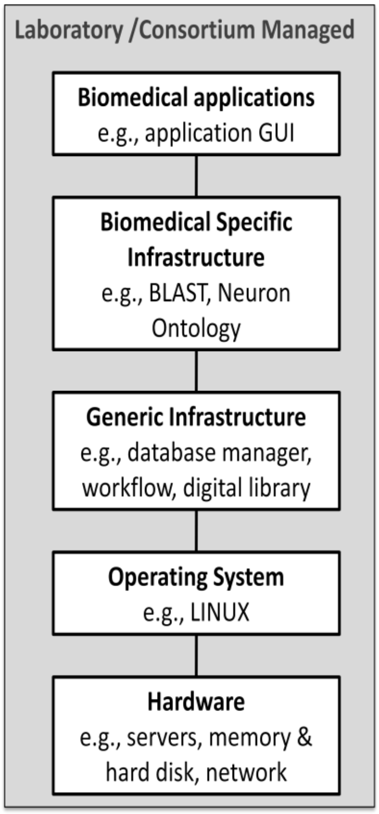
Laboratory /Consortium Managed
Biomedical applications
e.g., application GUI
Biomedical Specific Infrastructure
e.g., BLAST, Neuron Ontology
Generic Infrastructure
e.g., database manager, workflow, digital library
Operating System
e.g., LINUX
Hardware
e.g., servers, memory & hard disk, network

Fig. 1. A generic computing infrastructure employed at local laboratories, managed by the laboratory itself or a consortium for data sharing.

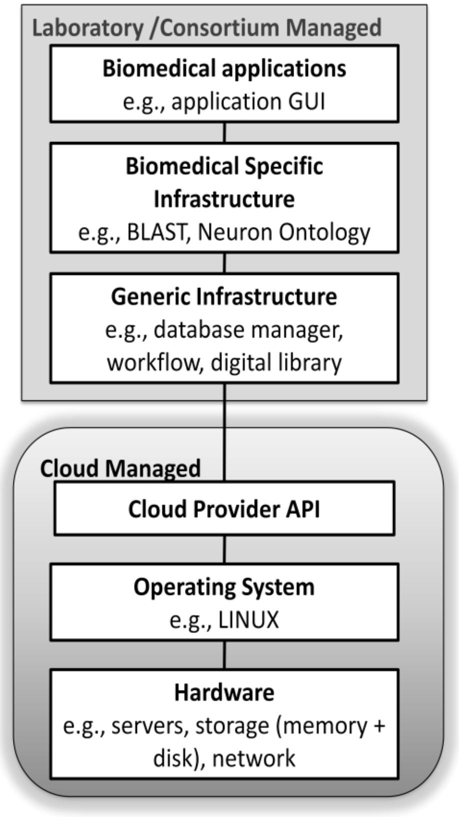
Laboratory /Consortium Managed
Biomedical applications
e.g., application GUI
Biomedical Specific Infrastructure
e.g., BLAST, Neuron Ontology
Generic Infrastructure
e.g., database manager, workflow, digital library
Cloud Managed
Cloud Provider API
Operating System
e.g., LINUX
Hardware
e.g., servers, storage (memory + disk), network

Fig. 2. Clouds can offload the responsibility of the bottom two layers of a basic computing infrastructure.