Performance of Private Clouds in Health Care Organizations

Richard Rauscher (PhD candidate)
Col. of Medicine and Dept. of Computer Science & Eng.
Pennsylvania State University
Hershey, Pennsylvania, USA
rauscher@psu.edu

Raj Acharya (advisor)
Department of Computer Science & Engineering
Pennsylvania State University
University Park, Pennsylvania, USA
acharya@cse.psu.edu

***Abstract*—This paper describes doctoral research in progress that is focused on the optimization of private computing clouds for use in health care. Three areas of research will be examined: appropriate levels of hardware provisioning, algorithms for virtual machine to physical machine mapping and appropriate levels of redundancy.**

Keywords—cloud computing, health care information systems, electronic health records, private cloud

I. INTRODUCTION

Attaining computational efficiency in health care is becoming increasingly important. Currently, 17.9% of the GDP of the United States is spent on health care. It is an information intense industry, and, through private and governmental incentives, is increasingly using digital information to manage care. This paper describes doctoral research that will examine the specific requirements for utilizing cloud computing in health care.

Cloud computing and computer virtualization have complicated the decision process around how to architect solutions. Health care decision makers must seek to minimize their spending while ensuring sufficient computational capacity, regulatory compliance and security. The intensity of the use of digital information in health care is expected to increase rapidly. One expected area of significant data growth is genomics. The cost of whole genome sequences for individuals is dropping precipitously; if current trends are sustained, by 2018, a full genome sequence will cost less than a magnetic resonance imaging (MRI) scan. Data collection will also increase when the use of wearable and implantable devices become widespread. Early research is currently being conducted that collect data from asymptomatic individuals for the purpose of detecting insipient disease states before clinical presentation [1]. These devices, with high sampling rates, may lead to longer lives but will also require vastly greater data and analytic capabilities.

Efficiently computing upon and storing massive amounts of health care data is vital. The expected exponential growth of health care data creation and the advent of cloud computing requires that we consider both when contemplating future states of health care computing infrastructures. As there remain many uncertainties regarding the legal complexities of public clouds, some health care organizations are turning towards private computing clouds. A private computing cloud is defined as a cloud where the entity is in possession of the computational resources, administers the computational resources and the resources are under control of that entity [2, 3]. These private computing clouds must be efficiently provisioned so as to meet the needs of the health care provider while minimizing waste. Unlike public clouds, where resources are set aside for just-in-time allocation, organizations that invest in their own private clouds are less likely to tolerate large amounts of idle resources.

This doctoral research focuses on three distinct areas of private cloud computing with specific health care considerations: hardware provisioning, application organization within the private cloud and levels of redundancy. In preparation for researching these areas, the authors have considered the performance characteristics of health care information systems as well as the attitudes of health care CIOs towards cloud computing models.

II. BACKGROUND

A. History

At its most basic level, this research focuses on minimizing the hardware necessary to reliably address a defined level of demand. The earliest disciplined mathematical examination of finite resource management where demand exceeded supply was performed by the Danish mathematician Agner Krarup Erlang at the dawn of public telephony [4]. Erlang's and subsequent contributions by Kleinrock [5] and scores of others to queuing theory and loss models are relevant today and used to model contentious queuing systems. Early work in performance and resource management was related to telephone and data networking but it has been repurposed for everything from operating systems to optimizing customer service in emergency rooms. Haverkort et al have examined performance issues with respect to degradable systems. They worked in the area of research now known as performability modeling [6].

Machine virtualization (and therefore, cloud technology) is not new. Although not called virtual machines, Supnik and others were using simulated computers to test new architectures in the late 1960s and early 1970s [7]. Early work in commercial virtualization took place at IBM within the VM/370 team that constructed the first hypervisor [8]. Work on virtualization and hypervisors was limited in the 1980s and 1990s and mostly confined to mainframes and

executing non-native operating systems (e.g. running Microsoft Windows on a PowerPC-based Macintosh). A query of the Association for Computing Machinery's database shows that there were no titles of papers with the word "hypervisor" between 1976 and 1995. Interest in virtual machines resurfaced in the late 1990s when several events converged: operating systems, originally constructed for consumers, began to be used in enterprises; the utilization of the computational machinery of those operating systems was relatively low; and operating environments and applications became more interdependent and required custom operating system configurations for each application.

B. Health Care and Cloud Computing

There is no evidence of health care organizations using private computing clouds prior to 2003. We published the first description of a hospital utilizing a private computing cloud in 2004 [9]. In the United States, the set of regulations known as the Health Insurance Portability and Accountability Act (HIPAA) define how health data are managed and secured. Due to HIPAA, the use of public cloud computing for health care has been fraught with legal uncertainty. Schweitzer performed an analysis of HIPAA requirements and how they may be met to make use of cloud computing [10]. Schweitzer identified and enumerated the elements that should be part of a contract with a cloud provider and to which the cloud provider would have a legal obligation to comply. Negotiating rigorous contracts with shared liability (particularly when the liability is not easily calculable) will indirectly increase the cost of using a public cloud service. Armbrust et al conducted an early analysis regarding the economics of using public cloud computing [11]. This report indicated that public cloud computing, if all costs were considered, was marginally more expensive than private cloud computing. Cloud providers have been unwilling with sign HIPAA-defined business associate agreements (BAAs) due to the liability associated with doing so. The January 2013 HIPAA Omnibus rules clarified the need for a BAA and the liabilities for public cloud providers who host protected health information [12]. Where there was previously a level of ambiguity regarding the need for public cloud computing providers to sign a BAA, none now exists: public cloud computing providers who knowingly host protected health information (PHI) for covered entities must sign a BAA and accept the associated liabilities. As stated previously, they may not be willing to sign these agreements and thus may not legally accept responsibility for hosting PHI. Although the number of cloud providers seemingly willing to sign BAAs is increasing [13], the individual cloud provider may not be willing to sign the particular health care entity's BAA depending upon the level of liability that is extended to the provider.

To measure the sentiment of health care IT leadership, the authors conducted an IRB approved (PSU IRB #41384) survey among members of the College of Health Care Information Management Executives (CHIME). The survey sought to understand the attitudes regarding the use of public and private cloud computing. The ultimate findings of the survey were that 63% of the CIOs surveyed were not pursuing public cloud solutions primarily due to security and legal concerns. By contrast, more than 61% were considering private cloud solutions.

C. Cloud Computing Performance

Cloud computing performance has been an active area of research since the earlier virtualization. Gambi and Toffetti considered models of the dynamic growth and shrinkage of computing clouds. They indicated that traditional linear and simple queuing systems are not sufficient for modeling complex and dynamic cloud systems [14]. They proposed using Kriging (Gaussian Process Regressions) as a model of cloud dynamism. Smith considered the use of standard industry benchmarks to measure a cloud's performance characteristics [15]. Yigitbasi et al built a system for creating load and measuring resources from inside the cloud and examined the performance of several cloud models [16]. Like Gambi and Toffetti, Brebner examined the elasticity of clouds and the performance implications and characteristics of growth and shrinkage in computing clouds [17]. Duong et al considered an approach for public cloud providers to optimize their revenue while maintaining or minimally sacrificing their quality-of-service agreement with their clients [18]. They specifically focused on applications with small latency tolerances such as on-line gaming. Cecchet et al proposed a system called Dolly to manage the stateful nature of databases in consideration of dynamic growth and shrinkage in cloud systems [19]. Liu and Wee determined that no single optimal configuration for cloud services supported all types of user behavior. They therefore proposed a dynamic switching architecture which optimizes an applications use of public cloud resources depending upon resource needs. Varadarajan et al discovered a novel attack vector for stealing resources from neighboring virtual machine for the benefit of the attacker [20]. Tan examined the performance of highly parallel scientific computing problems and their performance within public computing clouds [21]. Similarly, Ostermann, Iosup et al developed a cloud performance measurement methodology and found disappointing performance characteristics of Amazon's EC2 public cloud for solving scientific computing problems [22]. Zhao examined the use of cloud computing for coordinating multi-player games. Their work suggests that gaming may achieve higher performance if cloud based resources are utilized instead of client-based resources [23]. Deng et al considered another dimension of performance. They examined the environmental impact of data centers and methods to balance carbon emissions by shifting virtual machines between data centers in different parts of the world [24]. The VOLARE system provides a context awareness for mobile devices connecting to cloud services to offer differentiated services for varying capabilities of mobile devices [25].

D. Cloud Computing Provisioning

Provisioning resources in cloud computing to match computational requirements has also been an active area of research. Although provisioning goes hand-in-hand with performance analysis, performance analysis subject matter has focused on how well cloud models perform under certain application loads while the provisioning literature has focused on how resources are allocated to a cloud. Much of

the research examines how clouds grow and shrink dynamically. Tan et al examined how I/O paths limit dynamic provisioning of cloud infrastructure and modeled the system using a traditional Jackson network [26]. Chapman created a new language construct and tested it in the RESERVOIR system to manage cloud elasticity [27]. From a service provider's perspective, Rao et al built a system where the cloud organization was determined by a distributed learning algorithm [28]. Rao claimed to have used such a learning algorithm to efficiently provision a cloud configuration with low overhead and few required samples. Tan et al modeled resource provisioning in a cloud using a traditional telephony/queueing system and assumed discrete Erlang admin/no-admit rules for provisioning cloud services [21]. The CloudNet system considers the resources required to migrate "live" virtual machines over wide area networks (WANs) [29]. CloudNet improves the migration of virtual machines over WANs by 65%. Bayllocator is a system that allocates RAM between different virtual machines running in a cloud based on a Bayesian analysis and prediction of memory utilization [30].

E. History as a Predictor of the Future

Significantly, Stokely et al examined the predictability of resource usage in a cloud environment. They found that individual behavior was difficult to forecast. However, they found that *in aggregate*, they were able to successfully forecast the future using historical information with a margin of error of 12% [31]. This research contributes to our algorithm for computing an efficient level of hardware provisioning.

III. PREPATORY RESEARCH

A. Survey Results

As indicated above, we have conducted a survey of Chief Information Officers of health care entities in January 2013. The ultimate finding of the survey was that health care CIOs were much more likely to utilize private than public cloud computing technologies.

B. EMR Simulation and Resource Utilization

To gain insight into the performance and resource requirements, we examined the characteristics of a OpenEMR, an open-source electronic medical record system. We choose OpenEMR because the source code is available and there are no commercial constraints. Furthermore, many of the commercial database vendors prohibit or restrain the publication of benchmarking data without their consent [32].

1) Experimental Design

The architecture of OpenEMR is similar to the model represented in Figure 1 [33]. The following elements were measured with varying simulated user loads: input/output operations, CPU loads, RAM loads (excluding file caching), RAM loads (including file caching). Each server within the OpenEMR system was configured as a virtual computer on a Ubuntu 10.4 workstation with 12 processing cores and 12 GB of RAM. The hypervisor was Oracle VirtualBox 4.1.10. Each virtual machine was allocated one virtual processor. The virtual appliance (pre-configured virtual machine) was downloaded from the main OpenEMR website. It was tested and shown to be functional and subsequently cloned and configured such that the original virtual appliance represented the web and application server and the second virtual machine managed only the database. The two parts of the application were split to facilitate finer measurement of application and database computational requirements.

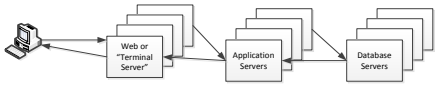

Fig. 1: Typical EMR Application Architecture

OpenEMR is a web-based application, thus a web load inducer was used to simulate the load of multiple users on the application. The web load inducer selected was Apache Jmeter [34] which executed manually generated traces. Jmeter was run from the host operating system. To avoid extraneous traffic, the guest operating systems which housed the OpenEMR framework were technically limited such that they could only communicate to each other and the host machine. Wherever possible, the traces used to drive the simulations were created by recording "normal" EMR activity. This was done by recording HTTP transactions through proxy server that was built into Jmeter. The OpenEMR software was adjusted to use plain-text transfer methods (as opposed to the default of SSL) so that the transactions could be observed by the proxy server such that they could be repeated. The recorded activities were replayed to simulate human behavior. The data collected were CPU load average, memory utilization, page fault rate (to ensure other statistics weren't confounded by page faults), I/O operations per second and network bandwidth utilization. The experiment was run with 1, 5, 10, 15, 25, 50 and 100 simulated users.

C. EMR Simulation Results

As the number of users increased, the CPU load average of the database server increased linearly. When 100 users were simulated, the load average on the database server was sustained at about 90. The CPU load average of the application server was minimal and always less than one. The network utilization of the database server and the application server were similar with the application server using slightly more overall bandwidth. This is to be expected since the application server was communicating with the simulated clients and the database server. Both increased linearly with increases in simulated users and peaked at about 12 Mbits/sec. I/O operations per second were nearly zero on the application server and varying linearly with the number of simulated users on the database server. Page faults were absent as the physical RAM allocated to each virtual machine was not exhausted. Due to space restrictions, graphs have been omitted.

IV. AIM 1: EFFICIENT PROVISIONING OF PRIVATE CLOUD

A. Problem Statement

Given that health care systems are likely to use private clouds, are predictable and have well-defined application behaviors, determine a system to efficiently provision hardware to the private health care cloud.

B. Motivaton

The first aim of this research was motivated by a practical problem at the Penn State Milton S. Hershey Medical Center: what is the appropriate level of resources to allocate to our private cloud? The need to execute this research has been further motivated by a survey of US-based health care chief information officers and their preference towards the use of private computing clouds.

C. Background

This research examines private (infrastructure-as-a-service) cloud computing for health care. Private cloud computing is an alternative to public cloud computing in that it provides many of the benefits of public cloud computing without the onus of executing a HIPAA-compliant business associate agreement with a vendor. To optimize efficiencies, health care organizations must ensure that their private cloud is properly sized given a set of parameters (operational characteristics, room for growth and redundancy). This section discusses a novel tool for evaluating sizing options for efficiently creating a private cloud given the unique nature of health care computer operations.

D. Cloud Simulators

There have been several systems constructed to provide estimates of potential cloud utilization (eg. [35] [36] [37]). These systems have focused on simulating public cloud use, network issues, virtual machine instantiation and destruction or other aspects primarily focused on public clouds. These general purpose cloud simulators do not address the unique nature of health care computing. Specifically, they do not take into account the predictability of private health care clouds.

E. Unique Aspects of Health Care Computing

Hospital-based health care is an around-the-clock operation with well-defined patterns of care. Nursing staff, who are the primary deliverers of care in in-patient settings, typically work either three eight hour shifts or two twelve hour shifts in twenty-four hour periods. Ambulatory, low-acuity clinics operate with standard office hours. Time-of-day and day-of-year patterns in care provisioning cause reasonably predictable patterns of computer usage. Ambrust et al, indicated that public cloud computing is best suited for usages where the resource consumption was unpredictable due to its elasticity. By contrast, health care operations and its computer use are remarkably predictable.

Health care and the practice of medicine is remarkably conservative and resistant to change [38]. This culture of change resistance permeates health care information technology departments and application vendors. Additionally, many hospitals use closed-source applications that cannot be easily modified for a specific environment.

F. Methods

A novel algorithm for determining the most efficient provisioning of resources was developed by the authors. This algorithm was presented in [39]. The algorithm takes into consideration the unique cyclical nature of health care computer operations. Unlike the previous cloud simulators, this system relies heavily on a trace-driven analysis using historic resource consumption behavior as the main predictor of future utilization. Historic resource utilization has been shown to be an accurate predictor of future performance [31]. This new method uses a minimally-invasive measurement system based on the host resources (RFC 2790) simple network management protocol (SNMP) interface [40]. The SNMP interface is reasonably ubiquitous, supported on most operating systems and avoids the pragmatic issue of potential conflicts with vendor-provided application software. The function of the algorithm is described below. At a basic level, the algorithm works as follows:

1. Collect data.
2. Divide and discretize the resource utilization data.
3. Sum the use of homogenous resources across all applications for each short time period.
4. Combine the use of each time of day for each resource into a "bucket".
5. Perform statistics on the data in the bucket and compute a point that is 99.5% greater than the sample set.
6. Find the maximum for each homogenous resource across all buckets.

We illustrate the use of this algorithm by example. Let's assume we're examining a single homogenous resource (e.g. RAM). Let's also assume a twenty-four hour period and a time quantum of one minute. Let's assume ten operating environments and a data collection period of two weeks. For each of the ten computers, there should be utilization data for each minute of the fourteen day period. Due to data collection errors, this is not always true (e.g. a machine is down or data are not recorded properly). Thus, a normalization step is required to correct for missing or extraneous data. Then, for each of the time quanta across all of the time periods, sum the utilization. This step characterizes the actual resource utilization over the entire sampling period. For example, if each of the ten computers had an average of 2GB of utilization at 8:03am on day 3, the sum would be 20GB of usage for that minute. This is repeated for every minute (quantum) over the entire collection period. Then for each minute of each day, place the value of the previous step into a "bucket" such that the data from minute one of each day are grouped and minute two of each day are grouped, etc. Calculate statistics upon each bucket and determine a utilization amount that is greater than 99.5% of the samples. The maximum of these calculations across all buckets is a point that is statistically greater than 99.5% of all probable utilization scenarios.

G. Explanation

This example is a specific description of a generalizable algorithm. This method will produce the level of resources that should, based on history, satisfy at least 99.5% (or any arbitrary level) of all resource demands. Unlike other methods which simply sum the maximum resources potentially demanded by each application, this method takes into account staggered use of resources over time. In the near

future, we plan to compare this method of hardware provisioning with other approaches.

H. Expected Contributions

In addition to the contribution of the algorithm above, we expect to create a tool that data center managers and systems administrators can either download or use via a web interface. This tool will assist them in collecting data in order to appropriately size their private cloud. We believe that this will improve the efficiency of computing in private clouds in health care and other organizations with similar requirements.

V. Aim 2: Virtual Machine organization

A. Problem Statement

Given a fixed set of hardware and applications with well-defined and predictable behavior, determine the efficient static or dynamic mapping of virtual machines to physical machines. This problem has been studied in generalized terms. This research will examine the specific aspects and optimizations possible in a highly predictable environment.

B. Motivation

Given the assumptions that health care systems will tend towards the use of private clouds and that the utilization of health care information systems is predictable, this research will examine how virtual machines should be efficiently mapped to physical machines within a private cloud. It will examine not just the relative location of virtual machines to each other but also internal cloud network optimizations (e.g. Xenloop [41]). It will also contemplate the use of a hybrid (private/public) cloud but continue to assume that most health care providers must be able to operate "off the grid".

C. Research Approach

We will create discrete event simulations that will characterize performance under varying circumstances. This will include varying machine placement, hardware architectures and specific private hypervisors. We will also implement novel experimental network stacks within specific open source health care information systems and note the different user-perceived performance characteristics given varying loads and varying architectures.

D. Expected Contributions

We expect that the contributions from this research will help to determine if health care is a special case and worthy of further research or specific approaches. We will also determine if special purpose protocols are useful for improving the user-perceived performance of health care applications in a private cloud.

VI. Aim 3: Determining appropriate redundancy

A. Problem Statement

Given that health care providers are likely to use self-administered private clouds to host and manage their data, build a methodology for determining the appropriate level of redundancy.

B. Motivation

The level of redundancy required to support a given system is a complex decision and there is little guidance for data center and network managers to use to affect a reasonable balance. Additionally, increased redundancy typically increases the level of complexity. The increased complexity can paradoxically have an overall negative effect on reliability.

C. Research Approach

This research, which is admittedly in its early stages, will draw heavily on the performability work of Trivedi, Havenhort and others. Ultimately, we would like to construct a calculus that will provide guidance to system architects given a set of reliability constraints. We expect to consider the Trivedi continuous Markov models that represent the "degree of failure" of the system, the transition probabilities within the state models and the economics of increased hardware redundancy.

VII. Conclusion

Attaining computational efficiency in health care is vital as the amount of data collected, stored, and analyzed continues to grow at an increasing rate. The use of public clouds comes with some risk; on the fore-front are the issues of security and liability. Because of this, health care IT leadership is considering a move towards the use of private clouds instead. With this future need in mind, the authors set forth to develop methods to ensure computational efficiency with the use of private clouds in health care settings (as well as other organizations with predictable patterns of computer usage). Our aims are to develop novel methods to: (1) determine appropriate levels of hardware provisioning, (2) develop algorithms for virtual machine to physical machine mapping, and (3) determine appropriate levels of redundancy. Most of the research performed to date is related to Aim 1. As we move forward, we are grateful for the opportunity to present our ideas and to receive feedback/guidance from the reviewers and attendees.

References

[1] J. Cohen, "The Patient of the Future," in *MIT Technology Review* Cambridge, MA: MIT Press, 2012.

[2] Z. Shuai, Z. Shufen, C. Xuebin, and H. Xiuzhen, "Cloud Computing Research and Development Trend," in *Future Networks, 2010. ICFN '10. Second International Conference on*, pp. 93-97.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, pp. 50-58, 2010.

[4] "Agner Krarup Erlang." vol. 2013: Wikipedia, 2013.

[5] "Leonard Kleinrock." vol. 2013: Wikipedia, 2013.

[6] B. R. Haverkort, *Performability Modelling : Techniques and Tools*. Chichester, UK ; New York: Wiley, 2001.

[7] R. Supnik, "Debugging Under Simulation," in *Debugging Tecniques in Large Systems*, R. Rustin, Ed.: Prentice-Hall, 1971.

[8] C. J. Young, "Extended architecture and Hypervisor performance," in *Proceedings of the workshop on virtual computer systems* Cambridge, Massachusetts, USA: ACM, 1973.

[9] R. Rauscher, "Server virtualization. There's a way around supporting multiple servers and operating systems," *Healthc Inform*, vol. 21, pp. 66, 68, Oct 2004.

[10] E. J. Schweitzer, "Reconciliation of the cloud computing model with US federal electronic health record regulations," *J Am Med Inform Assoc*, vol. 19, pp. 161-5, Mar-Apr.

[11] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," EECS Department, University of California, Berkeley UCB/EECS-2009-28, February 10 2009.

[12] "Modifications to the HIPAA Privacy,Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules " in *45 CFR Parts 160 and 164*, 2013.

[13] "HIPAA BAA Agreement, Omnibus rules new as of Jan 2013," in *Amazon Web Services Forum*. vol. 2013, 2013.

[14] A. Gambi and G. Toffetti, "Modeling cloud performance with kriging," in *Proceedings of the 2012 International Conference on Software Engineering* Zurich, Switzerland: IEEE Press.

[15] W. D. Smith, "Characterizing Cloud Performance with TPC Benchmarks," in *4th TPC Technology Conference, TPCTC 2012*, Berlin, Germany, pp. 189-96.

[16] N. Yigitbasi, A. Iosup, D. Epema, and S. Ostermann, "C-meter: a framework for performance analysis of computing clouds," Piscataway, NJ, USA, 2009, pp. 472-7.

[17] P. C. Brebner, "Is your cloud elastic enough?: performance modelling the elasticity of infrastructure as a service (IaaS) cloud applications," in *Proceedings of the third joint WOSP/SIPEW international conference on Performance Engineering* Boston, Massachusetts, USA: ACM.

[18] T. N. B. Duong, X. Li, R. S. M. Goh, X. Tang, and W. Cai, "QoS-Aware Revenue-Cost Optimization for Latency-Sensitive Services in IaaS Clouds," in *Proceedings of the 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications* Dublin, Ireland: IEEE Computer Society.

[19] E. Cecchet, R. Singh, U. Sharma, and P. Shenoy, "Dolly: virtualization-driven database provisioning for the cloud," *SIGPLAN Not.*, vol. 46, pp. 51-62.

[20] V. Varadarajan, T. Kooburat, B. Farley, T. Ristenpart, and M. M. Swift, "Resource-freeing attacks: improve your cloud performance (at your neighbor's expense)," in *Proceedings of the 2012 ACM conference on Computer and communications security* Raleigh, North Carolina, USA: ACM.

[21] Y. Tan, Y. Lu, and C. H. Xia, "Provisioning for large scale cloud computing services," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems* London, England, UK: ACM.

[22] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "A performance analysis of EC2 cloud computing services for scientific computing," Munich, Germany, 2011, pp. 115-131.

[23] Z. Zhao, K. Hwang, and J. Villeta, "Game cloud design with virtualized CPU/GPU servers and initial performance results," in *Proceedings of the 3rd workshop on Scientific Cloud Computing Date* Delft, The Netherlands: ACM.

[24] N. Deng, C. Stewart, D. Gmach, M. Arlitt, and J. Kelley, "Adaptive green hosting," in *Proceedings of the 9th international conference on Autonomic computing* San Jose, California, USA: ACM.

[25] P. Papakos, L. Capra, and D. S. Rosenblum, "VOLARE: context-aware adaptive cloud service discovery for mobile systems," in *Proceedings of the 9th International Workshop on Adaptive and Reflective Middleware* Bangalore, India: ACM.

[26] J. Tan, H. Feng, X. Meng, and L. Zhang, "Heavy-traffic analysis of cloud provisioning," in *Proceedings of the 24th International Teletraffic Congress* Krakow, Poland: International Teletraffic Congress.

[27] C. Chapman, W. Emmerich, F. G. Marquez, S. Clayman, and A. Galis, "Software architecture definition for on-demand cloud provisioning," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing* Chicago, Illinois: ACM.

[28] J. Rao, X. Bu, K. Wang, and C.-Z. Xu, "Self-adaptive provisioning of virtualized resources in cloud computing," *SIGMETRICS Perform. Eval. Rev.*, vol. 39, pp. 321-322.

[29] T. Wood, K. K. Ramakrishnan, P. Shenoy, and J. v. d. Merwe, "CloudNet: dynamic pooling of cloud resources by live WAN migration of virtual machines," in *Proceedings of the 7th ACM SIGPLAN/SIGOPS international conference on Virtual execution environments* Newport Beach, California, USA: ACM.

[30] E. Tasoulas, H. Haugerund, and K. Begnum, "Bayllocator: a proactive system to predict server utilization and dynamically allocate memory resources using Bayesian networks and ballooning," in *Proceedings of the 26th international conference on Large Installation System Administration: strategies, tools, and techniques* San Diego, CA: USENIX Association, 2012.

[31] M. Stokely, A. Mehrabian, C. Albrecht, F. Labelle, and A. Merchant, "Projecting disk usage based on historical trends in a cloud environment," in *Proceedings of the 3rd workshop on Scientific Cloud Computing Date* Delft, The Netherlands: ACM.

[32] O. Corp., "Oracle Technology Network Developer License Terms." vol. 2012, 2012.

[33] "OpenEMR." vol. 2012, 2012.

[34] A. Foundation, "Apache Jmeter." vol. 2013, 2013.

[35] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience."

[36] M. N. Rodrigo Calheiros, Cesar De Rose, Rajkumar Buyya, "EMUSIM: An Integrated Emulation and Simulation Environment for Modeling, Evaluation, and Validation of Performance of Cloud Computing Applications," *Software -- Practices and Experiences*, vol. 00, 2012.

[37] A. Nunez, J. L. Vazquez-Poletti, A. C. Caminero, J. Carretero, and I. M. Llorente, "Design of a new cloud computing simulation platform," in *Proceedings of the 2011 international conference on Computational science and its applications - Volume Part III* Santander, Spain: Springer-Verlag.

[38] E. J. Topol, *The creative destruction of medicine : how the digital revolution will create better health care*. New York: Basic Books.

[39] R. Rauscher, "Cloud Computing Considerations for Biomedical Applications," in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pp. 142-142.

[40] IETF, "RFC 2790: Host Resources MIB." vol. 2790: Internet Engineering Taskforce, 2000.

[41] J. Wang, K.-L. Wright, and K. Gopalan, "XenLoop: a transparent high performance inter-vm network loopback," in *Proceedings of the 17th international symposium on High performance distributed computing* Boston, MA, USA: ACM, 2008.