

CloudAnalyst: A CloudSim-based Visual Modeller for Analysing Cloud Computing Environments and Applications

Bhathiya Wickremasinghe¹, Rodrigo N. Calheiros², and Rajkumar Buyya¹

¹The Cloud Computing and Distributed Systems (CLOUDS) Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne, Australia

²Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Brazil

Abstract—Advances in Cloud computing enable new possibilities to Internet applications developers. Previously, the main concern of Internet applications developers was deployment and hosting of applications, because it required acquisition of a server with a fixed capacity, able to handle the expected application peak demand, installation of the whole software infrastructure of the platform supporting the application, and configuration of the application itself. Furthermore, server was underutilized because peak traffic happens only at specific times. With the advent of the Cloud, solution of this problem became cheaper and easier with the use of pay-per-use, flexible elastic infrastructure services offered by Cloud providers. Because several Cloud providers are available, each one offering different pricing models and located in different geographic regions, a new concern of application developers is selecting providers and data center locations for applications. However, there is a lack of tools that enable developers to evaluate requirements of large-scale Cloud applications in terms of geographic distribution of both computing servers and user workloads. To fill this gap in tools for evaluation and modeling of Cloud environments and applications, in this paper we propose CloudAnalyst. It was developed to simulate large-scale Cloud applications with the purpose of studying the behavior of such applications under various deployment configurations. CloudAnalyst helps developers with insights in how to distribute applications among Cloud infrastructures and value added services such as optimization of applications performance and providers incoming with the use of Service Brokers.

Keywords-Cloud Computing; Simulation; Modeling

I. INTRODUCTION

Cloud computing is an area that is experiencing a rapid advancement both in academia and industry. This technology, which aims at offering distributed, virtualized, and elastic resources as utilities to end users, has the potential to support full realization of “computing as a utility” in the near future [15]. Along with the advancements of the Cloud technology, new possibilities for Internet-based applications development are emerging. These new application models are based in two parties: in one side, there are the cloud service providers that are willing to provide large-scale computing infrastructure at a

price based primarily on usage patterns. It eliminates the initial high-cost, for application developers, of environment set up an application deployment environment. On the other side there are large-scale software systems providers, which develop applications such as social networking sites and e-commerce, which are gaining popularity on the Internet. These applications can benefit greatly of Cloud infrastructure services to minimize costs and improve service quality to end users.

Previously, development of such applications required acquisition of servers with a fixed capacity able to handle the expected application peak demand, installation of the whole software infrastructure of the platform supporting the application, and configuration of the application itself. Furthermore, servers were underutilized because peak traffic happens only at specific times. With the advent of the Cloud, solution of this problem became cheaper and easier with the use of pay-per-use, flexible elastic infrastructure services provided by Cloud providers.

When trying to bring these two ends together several issues, which impact the net benefit of Cloud, arise. Some of these issues are geographic distribution of user bases, capabilities of the Internet infrastructure within those geographic areas, dynamic nature of the usage patterns of the user base, and capabilities of Cloud services in terms of adaptation or dynamic reconfiguration, among others.

A comprehensive study of the whole problem in the real Internet platform is extremely difficult, because it requires interaction with several computing and network elements that cannot be controlled or managed by application developers. Furthermore, network conditions cannot be predicted nor controlled, and it also impacts quality of strategy evaluation.

Study of such dynamic and massively distributed environments in a controlled and reproducible manner can be achieved with the use of simulation. CloudSim [5] is a Cloud simulation toolkit that allows modeling of infrastructures containing Cloud Data Centers, users, user workloads, and pricing models. CloudSim enables modeling and simulation of typical Cloud infrastructures, even though it has been developed without focusing any specific Cloud provider.

In this paper we propose a tool, called CloudAnalyst, which supports visual modeling and simulation of large-scale applications that are deployed on Cloud Infrastructures. CloudAnalyst, built on CloudSim, allows description of application workloads, including information of geographic location of users generating traffic and location of data centers, number of users and data centers, and number of resources in each data center. Using this information, CloudAnalyst generates information about response time of requests, processing time of requests, and other metrics.

By using CloudAnalyst, application developers can determine the best strategy for allocation of resources among available data centers, strategies for selecting data centers to serve specific requests, and costs related to such operations.

II. RELATED WORK

Cloud computing can be defined as “*a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers*” [1].

The level on which computing services are offered to consumer varies according to the abstraction level of the service. In the lowest level, Infrastructure as a Service (IaaS), services are supplied in the form of hardware on which consumers deploy virtual machines, software platforms to support their applications, and the application itself. An example of an IaaS service is Amazon EC2 [9].

In the next level, Cloud consumers do not have to handle virtual machines. Instead, a software platform for hosting applications (typically, web applications) is already installed in an infrastructure and offered to consumers. Then, consumers have to use the platform to develop their specific application. This strategy is known as Platform as a Service (PaaS). Examples of this case are Google App Engine [10] and Aneka [13]. Finally, in Software as a Service (SaaS), an application is offered to consumers, which do not have to handle virtual machines and software platforms that host the application.

Reproducible and controlled experiments on any of these levels require the use of alternative experimentation methodologies than real execution in a real platform. Simulation is one of such alternatives and this is the focus of this work.

There have been many studies using simulation techniques to investigate behavior of large scale distributed systems, as well as tools to support such research. Some of these simulators are GridSim [2], MicroGrid [3], GangSim [14], SimGrid [4] and CloudSim [5]. While the first three focus on Grid computing systems, CloudSim is, for the best of our knowledge, the only simulation framework for studying Cloud computing systems. Nevertheless, grid simulators have been used to evaluate costs of executing distributed applications in Cloud infrastructures [11][12].

GridSim toolkit was developed to address the problem of performance evaluation of real large scaled distributed environments (typically Grid systems but also P2P networks) in a repeatable and controlled manner. The GridSim toolkit is a Java based simulation toolkit that supports modeling and simulation of heterogeneous Grid resources and users spread across multiple organizations with their own policies for scheduling applications. It supports multiple application models and provides primitives for creation of application tasks, mapping of tasks to resources, and managing of tasks and resources.

CloudSim enables seamless modeling, simulation, and experimenting on Cloud computing infrastructures. It is a self-contained platform that can be used to model data centers, service brokers, and scheduling and allocation policies of large scale Cloud platforms. It provides a virtualization engine with extensive features for modeling creation and life cycle management of virtual machines in a data center, including policies for provisioning of virtual machines to hosts, scheduling of resources of hosts among virtual machines, scheduling of tasks in virtual machines, and modeling of costs incurring in such operations. CloudSim framework is built on top of GridSim toolkit.

CloudSim allows simulation of scenarios modeling IaaS, PaaS, and SaaS, because it offers basic components such as Hosts, Virtual Machines, and applications that model the three types of services.

CloudAnalyst is built directly on top of CloudSim toolkit, leveraging the features of the original framework and extending some of the capabilities of CloudSim. CloudAnalyst design and features are presented in the next section.

III. CLOUD ANALYST

Even though Clouds make deployment of large scale applications easier and cheaper, it also creates new issues for developers.

Because Cloud infrastructures are distributed, applications can be deployed in different geographic locations, and the chosen distribution of the application impacts its performance for users there are far from the data center.

Because Internet applications are accessed by users around the world, and because popularity of applications varies along the world, experience in the use of application will also vary. Quantifying impact of number of simultaneous users, geographic location of relevant components, and network in applications is hard to achieve in real testbeds, because of the presence of elements that cannot be predicted nor controlled by developers.

To allow control and repeatability of experiments, simulators such as CloudSim can be used. Simulation experiments apply models of both applications and infrastructures [7]. So, simulation requires some effort from application developers to model both the target infrastructure and the software in a language that is interpreted by the simulator. Even though simulators offer support to model such scenarios, they are conceived to be

applied in general experiments, and so modeling of specific scenarios may be time demanding.

One of the main objectives of CloudAnalyst is to separate the simulation experimentation exercise from a programming exercise, so a modeler can focus on the simulation complexities without spending too much time on the technicalities of programming using a simulation toolkit. The CloudAnalyst also enables a modeler to repeatedly execute simulations and to conduct a series of simulation experiments with slight parameters variations in a quick and easy manner.

The main features of CloudAnalyst are the following.

Easy to use Graphical User Interface (GUI). CloudAnalyst is equipped with an easy to use graphical user interface that enables users to set up experiments quickly and easily. CloudAnalyst GUI is shown in Figure 1.

Ability to define a simulation with a high degree of configurability and flexibility. Simulation of complex systems such as Internet Applications depends on many parameters. Typically, values for those parameters need to be arbitrarily assumed or determined through a process of trial and error. CloudAnalyst provides modelers with a high degree of control over the experiment. By modeling entities and configuration options such as: Data Center, whose hardware configuration is defined in terms of physical machines composed of processors, storage devices, memory and internal bandwidth; Data Center virtual machine specification in terms of memory, storage and bandwidth quota; Resource allocation policies for Data Centers (e.g., time-shared vs. space-shared); Application users as groups and their distribution both geographically and temporally; Internet dynamics with configuration options for network delays and available bandwidth; Service Broker Policies that control which segment of total user base is serviced by which Data Center at a given time; and simulation duration in minutes, hours or days.

Repeatability of experiments. CloudAnalyst allows modelers to save simulation experiments input parameters and results in the form of XML files so the experiments can be repeated. The underlying CloudSim simulation framework ensures that repeated experiments yield identical results.

Graphical output. CloudAnalyst is capable of generating graphical output of the simulation results in the form of tables and charts, which is desirable to effectively summarize the large amount of statistics that is collected during the simulation. Such an effective presentation helps in identifying the important patterns of the output parameters and helps in comparisons between related parameters. In the current version of CloudAnalyst, the following statistical metrics are produced as output of the simulation: Response time of the simulated application; overall average, minimum and maximum response time of all user requests simulated; Response time arranged by user groups, located within



Figure 1. CloudAnalyst GUI.

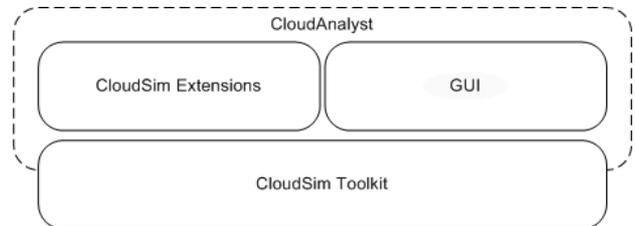


Figure 2. CloudAnalyst architecture.

geographical regions; response time arranged by time, showing the pattern of changes in application usage during the day; usage patterns of the application; number of users using the application arranged by time or regions of the world, and the overall effect of that usage on the data centers hosting the application; time taken by data centers to service a user request; overall request processing time for the entire simulation; average, minimum and maximum request processing time by each data center; response time variation pattern during the day as the load changes; and details of costs of the operation.

Use of consolidated technology and ease of extension. CloudAnalyst is based on a modular design that can be easily extended. It is developed using the following technologies: Java (the simulator is developed 100% on Java platform, using Java SE 1.6); Java Swing (the GUI component is built using Swing components); CloudSim (CloudSim features for modeling data centers is used in CloudAnalyst); and SimJava [6] (some features of this tool are used directly in CloudAnalyst).

A. CloudAnalyst design

As depicted in Figure 2, CloudAnalyst is built on top of CloudSim toolkit, by extending its functionalities with the introduction of concepts that model Internet and Internet Application behavior. The design of CloudAnalyst is shown in Figure 3. The main components and their responsibilities are discussed below:

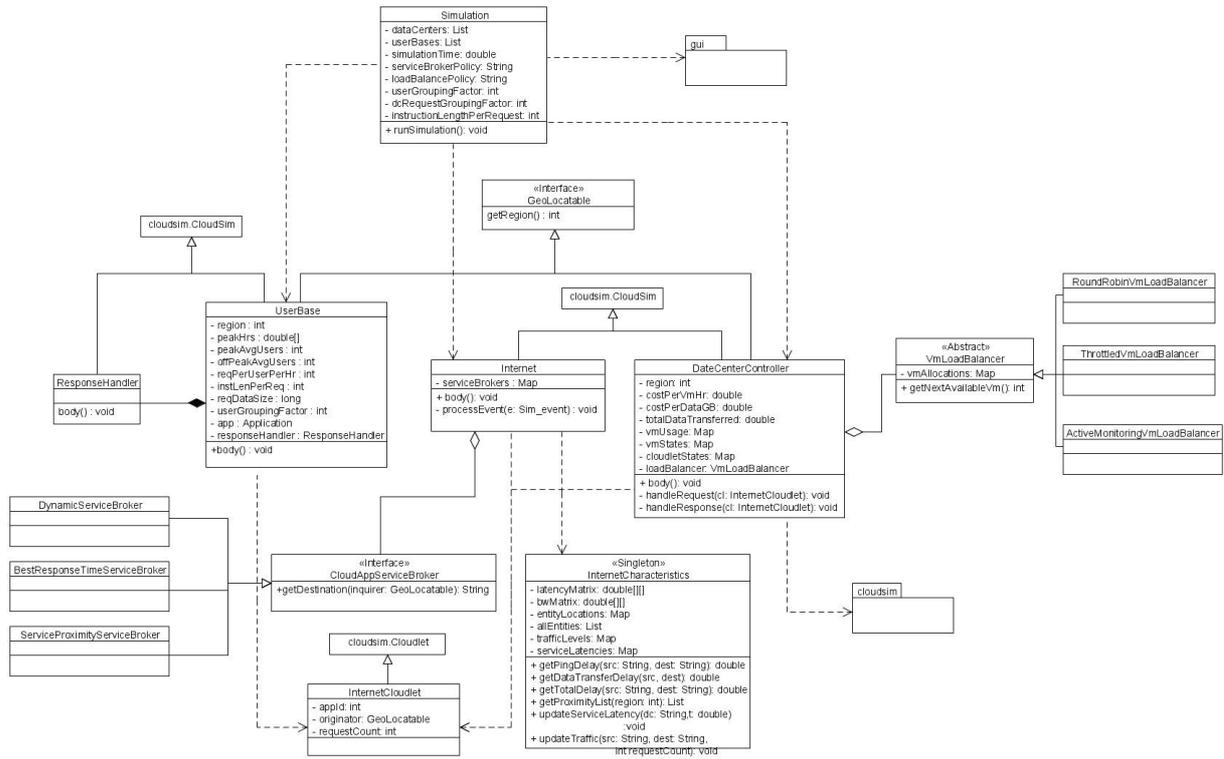


Figure 3. CloudAnalyst Class Diagram.

GUI Package. It is responsible for the graphical user interface, and acts as the front end controller for the application, managing screen transitions and other UI activities.

Simulation. This component is responsible for holding the simulation parameters, creating and executing the simulation.

UserBase. This component models a user base and generates traffic representing the users.

DataCenterController. This component controls the data center activities.

Internet. This component models the Internet and implements the traffic routing behavior.

InternetCharacteristics. This component maintains the characteristics of the Internet during the simulation, including the latencies and available bandwidths between regions, the current traffic levels, and current performance level information for the data centers.

VmLoadBalancer. This component models the load balance policy used by data centers when serving allocation requests. Default load balancing policy uses a round robin algorithm, which allocates all incoming requests to the available virtual machines in round robin fashion without considering the current load on each virtual machine. Additionally, it is possible application of a throttled load balancing policy that limits the number of requests being processed in each virtual machine to a throttling threshold. If requests are received causing this threshold to be exceeded in all

available virtual machines, then the requests are queued until a virtual machine becomes available.

CloudAppServiceBroker. This component models the service brokers that handle traffic routing between user bases and data centers. The default traffic routing policy is routing traffic to the closest data center in terms of network latency from the source user base. In addition an experimental brokerage policy for peak load sharing is implemented on CloudAnalyst. This routing policy attempts to share the load of a data center with other data centers when the original data center's performance degrades above a pre-defined threshold.

An important design consideration we took was grouping simulation elements to improve the efficiency of simulation. In CloudAnalyst the events are grouped at three levels. In the first level, there are user bases, which represent a cluster of users which are handled as a single unit. In the next level, user requests generated from each regional user base are grouped based on a grouping factor, which is kept independent of the user base size. In the last level, requests simultaneously processed by a single virtual machine are grouped. The last two grouping factors are configurable by CloudAnalyst users, and it is also possible not to group simulation elements.

Table 1. User bases used in the experiment.

User base	Region	Time Zone	Peak Hours (Local time)	Peak Hours (GMT)	Simultaneous Online Users During Peak Hrs	Simultaneous Online Users During Off-peak Hrs
UB1	0 – N. America	GMT – 6.00	7.00–9.00 pm	13:00-15:00	400,000	40,000
UB2	1 – S. America	GMT – 4.00	7.00–9.00 pm	15:00-17:00	100,000	10,000
UB3	2 - Europe	GMT + 1.00	7.00–9.00 pm	20:00-22:00	300,000	30,000
UB4	3 - Asia	GMT + 6.00	7.00–9.00 pm	01:00-03:00	150,000	15,000
UB5	4 - Africa	GMT + 2.00	7.00–9.00 pm	21:00-23:00	50,000	5,000
UB6	5 - Ocenia	GMT + 10.00	7.00–9.00 pm	09:00-11:00	80,000	8,000

Another important component of CloudAnalyst is the network model. Modeling of bandwidth is probably the most complex task, especially considering the nature of a network such as the Internet. In the current version of CloudAnalyst we use a hypothetical parameter, *available bandwidth*, which is assumed to be the quota of Internet bandwidth available for the application being simulated, ignoring other external factors. Events such as traffic generation are produced using a Poisson distribution.

IV. A CASE STUDY: SIMULATION OF A LARGE SCALE SOCIAL NETWORK APPLICATION

A typical large scale application on the Internet that can benefit from Cloud technology is social networking applications. These applications may benefit from Clouds because they typically present non-uniform usage patterns. Access to such services varies along the time of the day, and geographic location from sources of service requests also varies. Furthermore, a new functionality in the service may cause a sudden increase in interested by the service, leading to an increase in number of requests arriving to servers that may be only temporary. Cloud allows infrastructures to dynamically react to increase in requests, by dynamically increasing application resources, and reducing available resources when the number of requests reduces. So, SLAs between Cloud providers and consumers are met with a minimal cost for consumers.

One well-known social networking site is Facebook [8], which has over 200 million registered users worldwide. On 18/06/2009 the approximate distribution of the Facebook user base across the globe was the following: North America: 80 million of users; South America: 20 million of users; Europe: 60 million of users; Asia: 27 million of users; Africa: 5 million of users; and Oceania: 8 million of users.

In this case study, we model the behavior of social network applications such as Facebook and use CloudAnalyst to evaluate costs and performance related to use of Clouds to host such an application.

A. Simulation Configuration

We defined six user bases representing the six main regions of the world with parameters described in Table 1. For our simulation we chose a similar hypothetical application at 1/10th of the scale of Facebook.

For the sake of simplicity each user base is contained within a single time zone and it is assumed that most users use the application in the evenings after work for about 2 hours. It is also assumed that 5% of the registered users are online during the peak time simultaneously and only one tenth of that number of users is on line during the off-peak hours. Furthermore, each user makes a new request every 5 minutes when he or she is online.

In terms of the cost of hosting applications in a Cloud, we assume a pricing plan which closely follows the actual pricing plan of Amazon EC2 [9]. The assumed plan is: Cost per VM per hour (1024Mb, 100MIPS): \$ 0.10; Cost per 1Gb of data transfer (from/to Internet): \$0.10.

Size of virtual machines used to host applications in the experiment is 100MB. Virtual machines have 1GB of RAM memory and have 10MB of available bandwidth. Simulated hosts have x86 architecture, virtual machine monitor Xen and Linux operating system. Each simulated data center hosts 20 virtual machines dedicated to Facebook. Machines have 2 GB of RAM and 100GB of storage. Each machine has 4 CPUs, and each CPU has a capacity power of 10000 MIPS. A time-shared policy is used to schedule resources to VMs. Users are grouped by a factor of 1000, and requests are grouped by a factor of 100. Each user request requires 250 instructions to be executed. User bases used in the experiments are described in Table 1.

B. Simulated Scenarios

Several scenarios are considered in our case study. The simplest one consists of modeling the case where a single, centralized Cloud data center is used to host the social network application. In this model, all requests from all users around the world are processed by this single data center. Data center has 50 virtual machines allocated to the application.

Table 2. Simulation settings and experiment results.

	Scenario	Overall average response time (milliseconds)	Overall average time spent for processing a request by a data center (milliseconds)	Virtual Machine Cost	Data Transfer Cost
1	1 data center with 50 VMs	284.98	46.79	\$120.05	\$512.74
2	2 data centers with 25 VMs each	249.20	119.97	\$120.05	\$512.74
3	2 data centers with 50 VMs each	183.85	54.65	\$ 240.10	\$512.74
4	2 data centers with 50 VMs each with peak load sharing	184.92	54.60	\$ 240.10	\$512.74
5	2 data centers with 50 VMs each with peak load sharing and throttling	157.56	28.45	\$ 240.10	\$512.74
6	3 data centers with 50 VMs each with peak load sharing and throttling	124.12	29.12	\$ 360.15	\$512.74
7	3 data centers with 75,50,25 VMs, with peak load sharing and throttling	121.07	23.96	\$ 360.15	\$512.74

The second scenario consists of the use of two data centers, each one with 25 virtual machines dedicated to the application.

The third scenario consists of two data centers, each one with 50 virtual machines. In the next scenario, the two data centers share the load during peak time, whereas in the fifth there are peak load sharing and throttling.

In the sixth scenario three data centers with 50 virtual machines is used. In this case, there are also peak load sharing and throttling.

Finally, in the last scenario, data centers have different amount of virtual machines (25, 50, and 75). In this case, there are also peak load sharing and throttling.

Each of these scenarios was evaluated with execution of the workload previously described. Results are discussed next.

C. Results

Table 2 summarizes the results, while Figure 4 depicts variation of average response time. Results show that bringing the service closer to users improves the quality of service (response time in this case). It is an expected effect, because users experiment less effects from Internet issues (high latency, low bandwidth) when they are geographically close to the application server.

Results also show that service quality can be further improved with the application of load balancing in the application across data centers, which are supposed to be managed by different service brokerage policies, and also at virtual machine level within data centers. But the levels of improvement achieved depend largely on the load balancing algorithms employed. So, application of good load balancing strategies is paramount for large-scale distributed applications such as social networks.

For such improvements to be effective, sufficient capacity is required in the data centers to meet the peak

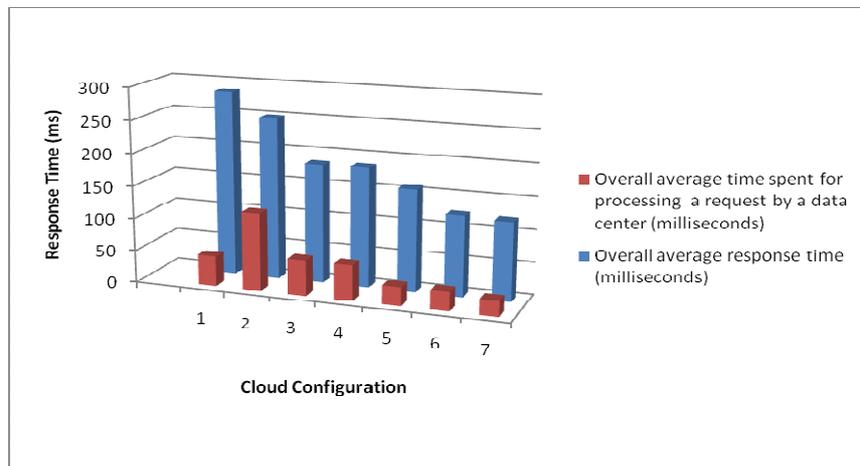


Figure 4. Summary of experimental results.

demand. It is not a matter of great concern in Cloud data centers, which apply economy of scale to make their business profitable and so they can offer more resources during peak traffic.

On the other hand, if provisioning for the peak capacity is allocated throughout the data centers, there is a significant proportion of the time in which capacity allocated for applications is not fully utilized, what decreases profits of application developers because they pay for unused resources. Once again, elastic cloud providers solve this problem by charging consumers proportionally to the amount of resources used. At the same time provides offer tools to automatically increase and decrease resources available to applications in order to meet established service level agreements.

Based on the above observations, a reasonable solution to an economic and efficient provisioning of resources to large scale distributed applications such as social networking is a setup where the resources are dynamically allocated by geographic location depending on the workload. E.g. the highest load from region 0 (North America) occurs from 13:00-15:00 GMT and during this time the data center servicing these requests (usually the data center located in region 0 itself) should have a higher number of virtual machines allocated to the application. But once the peak has passed in this region, the number of virtual machines is dynamically reduced in region 0 and the number of virtual machines in a data center in a region where the peak time is arriving is dynamically increased.

V. CONCLUSIONS AND FUTURE DIRECTIONS

With the rapid advances of Cloud technologies, there is a new demand for tools to study and analyze the benefits of the technology and the best practices to apply the technology to large-scaled applications. CloudAnalyst is a new tool developed to address this demand. It is based on top of mature simulation frameworks such as SimJava and CloudSim.

We demonstrated how CloudAnalyst can be used to model and evaluate a real world problem through a case study of a social networking application deployed on the cloud. We have illustrated how the simulator can be used to effectively identify overall usage patterns and how such usage patterns affect data centers hosting the application. Furthermore, we showed how those observations provide insights in how to optimize the deployment architecture of the application. A possibility in this direction is introduction of dynamic configurability through a global Cloud Service Broker, which increases or decreases the amount of resources available to the application in different geographic locations depending on the load at a given time.

Our work is the first attempt towards developing a tool and an approach for studying large scale distributed applications behavior by simulation in Cloud computing environments. Therefore, the tool will evolve over the time, and the result of this process will improve quality of the model and of the analysis it supports. In the long term this type of simulation experiment has a big potential to

support testers to identify new features and issues, model them, and develop and evaluate new mechanisms and algorithms for resource management, this way improving performance of emerging Cloud applications.

REFERENCES

- [1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype and Reality for Delivering Computing as the 5th Utility," *Future Generation Computer Systems*, vol. 25, June 2009, pp. 599-616.
- [2] R. Buyya, and M. Murshed, "GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing," *Concurrency and Computation: Practice and Experience*, vol. 14, Nov. 2002, pp. 1175-1220.
- [3] L. X. Song H, Jakobsen D, Bhagwan R, Zhang X, Taura K, A. Chien, "The MicroGrid: A scientific tool for modeling computational Grids," *Proc. of the ACM/IEEE Supercomputing Conference*, IEEE Computer Society, Nov. 2001.
- [4] A. Legrand, L. Marchal, and H. Casanova, "Scheduling distributed applications: the SimGrid simulation framework," *Proc. of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 07)*, May 2001, pp. 138-145.
- [5] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities," *Proc. of the 7th High Performance Computing and Simulation Conference (HPCS 09)*, IEEE Computer Society, June 2009.
- [6] F. Howell and R. Macnab, "SimJava: a discrete event simulation library for Java," *Proc. of the 1st International Conference on Web-based Modeling and Simulation*, SCS, Jan. 2008.
- [7] J. Gustedt, E. Jeannot, and Martin Quinson, "Experimental methodologies for large-scale systems: a survey," *Parallel Processing Letters*, vol. 19, Sep. 2009, pp. 399-418.
- [8] "Facebook," <http://www.facebook.com>.
- [9] "Amazon Elastic Compute Cloud (Amazon EC2)," <http://aws.amazon.com/ec2/>
- [10] "Google App Engine," <http://code.google.com/appengine/>
- [11] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the Cloud: the Montage example," *Proc. of the 2008 ACM/IEEE Conference on Supercomputing*, IEEE, Nov. 2008.
- [12] M. Assunção, A. di Costanzo, and R. Buyya, "Evaluating the Cost-Benefit of Using Cloud Computing to Extend the Capacity of Clusters", *Proc. of the 18th International Symposium on High Performance Distributed Computing (HPDC 2009)*, ACM Press, June 2009.
- [13] C. Vecchiola, S. Pandey, and R. Buyya, "High-Performance Cloud Computing: A View of Scientific Applications", *Proc. of the 10th International Symposium on Pervasive Systems, Algorithms and Networks (I-SPAN 2009)*, Kaohsiung, Taiwan, Dec. 2009.
- [14] C. Dumitrescu, and I. Foster. "GangSim: a simulator for grid scheduling studies," *Proc. of the 5th International Symposium on Cluster Computing and the Grid (CCGrid 05)*, IEEE Computer Society, May 2005.
- [15] A. Weiss, "Computing in the Clouds," *netWorker*, vol. 11, Dec. 2007, pp. 16-25.