Resource Management in Cloud Computing: Cost Efficiency with User-Centric Approach

Prabhat Kumar
M.Tech CSE
MUR1904041
Department of Computer Science
Mewar University NH 48, Gangarar, Rajasthan 312901

Abstract—**Context: Cloud computing is a model for enabling convenient and on-demand network access to a shared pool of configurable computing resources like networks, servers, storage, application, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction. These resources are pooled for the usage of customers. Its main purpose is to cater the elastic need of resources of customers due to varying workload. Clients need to pay only for the amount of resources they use. There are three type of cloud services: Infrastructure as a service, Platform as a service and Software as a service. In IaaS, User acquire computing resources such as Processing power, Memory, Storage etc. In PaaS, Users are provided with programming languages and tools with high level of abstraction to develop applications. In SaaS, Users use web browser to access software that others have developed and offered as a service over the web. These features and resources can be leveled up and down as per requirements. The resources need to be managed so that all the customers may get the services for which they demand and the resources can be efficiently utilized. The resources must be managed properly to save cost, provide functionality and increase efficiency of the cloud. In this thesis, I will work on automated resource management which will ensure efficient and cost aware utilization of resources.**

I. INTRODUCTION

Cloud computing is a model for enabling ubiquitin, convenient, on-demand network access a shared pool of configurable computing resources (eg., network, servers, storage, application and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. The essential characteristics of cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity and measured services. A cloud computing platform dynamically provisions, configures, reconfigures and de-provisions servers as needed. Servers in the cloud can physical machines or virtual machines. A client can access all programs and documents on cloud through any computer device via internet connection. With cloud computing, The software programs used by a client are not run from any personal computer but rather stored on server and accessed via internet. It provides pay as you go pricing model. There are various services provided by closed computing such as: Infrastructure as service (IaaS), Platform as service (Paas) and software as service (SaaS). All these services need high level of resource elasticity. Thus it a challenge for cloud provider to manage resources so that the meet service level agreement (SLA). Most of these services are service provider centric. This thesis focus on user centric cloud services.

Keywords—Challenges, Cloud Computing, costing, Techniques.

II. Essential Characteristics of Cloud Computing

On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations)

Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability (pay-per-use basis) at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

III. Cost-aware elasticity vs Cost optimized elasticity:

A. We present a cost-aware system that integrates multile elasticity mechanisms such as replication and migration and computes both a cost-optimized configuration for the desired capacity as well as a plan for transitioning the applicaton from its current setup to its new configuration. This algorithm can take into account price different server type to minimize the infrastrature cost of provisioning a certain capacity. It alos minimize the time to add extra capacity using different elasticity mechanisms (we call this time as transition cost). We formulate our provisioning problem as an interger linear program(ILP) to account for both infrastrature and transition cost for deriving appropriate elasticity decisions.

B.

We implement as prototype of this algorithm on could provisioning engine, using the CloudSim simulator in Eclipse, that in corporates our optimizations, and evaluate its efficacy on both a private laboratory-based Xen cloud. Our experimental result:
Demonstrate that cost-aware elasticity can reduce infrastructure costs comparison to cost-oblivious provisioning capproaches,

demonstrate that integrating multiple mechanisms such as migration and replication into a unified approach can double the cost savings, and

demonstrate how our transition-aware approach can be employed to quickly provision capacity in scenarios where an application workload surges unexpectedly.

C. *External* The prior work on dynamic provisioning has not been cost-aware. By being costoblivious, prior approaches assume that so long as the desired capacity is allocated to the application, the choice of exact hardware configuration is immaterial. That is, the unit cost per core is assumed to be identical, making an N-core system equivalent, from a provisioning perspective, to an N-core systems with single cores. In the cloud context, however, the choice of the configuration matters, since pricing per core is not uniform. Hence, this algorithm must take server infrastructure costsinto account during provisioning

D. IV. RESULTS AND ANALYSIS

In Cloud computing, application providers can allocate resources purely on-demand. On-demand (OD) computing is an increasingly popular enterprise model in which computing resources are made available to the user as needed. The resources may be maintained within the user's enterprise, or made available by a service provider. The on-demand model was developed to overcome the common challenge to an enterprise of being able to meet fluctuating demands efficiently. Because an enterprise's demand on computing resources can vary drastically from one time to another, maintaining sufficient resources to meet peak requirements can be costly. Conversely, if theenterprise cuts costs by only maintaining minimal computing resources, there will not be sufficient resources to meet peak requirements.

.

On-demand computing products are rapidly becoming prevalent in the marketplace. Computer Associates, HP, IBM, Microsoft, and Sun Microsystems are among the more prominent on-demand vendors. These companies refer to their on-demand products and services by a variety of names. IBM calls theirs "On Demand Computing" (without the hypen). Concepts such as grid computing, utility computing, autonomic computing, and adaptive management seem very similar to the concept of on-demand computing. Jason Bloomberg, Senior Analyst with ZapThink, says that on-demand computing is a broad category that includes all the other terms, each of which means something slightly different. Utility computing, for example, is an ondemand approach that combines outsourced computing resources and infrastructure management with a usage-based payment structure (this approach is sometimes known as metered services).

Many industry insiders expect on-demand computing to become the most pervasive enterprise computing model within the next few years. According to Irving Wladawsky-Berger, IBM's vice-president of technology and strategy (quoted in a ZDNet Tech Update article), "The technology is at a point where we can start to move into an era of on-demand computing. I give it between two and four years to reach a level of maturity."[56]

This ability to allocate resources on an as-needed basis which we refer to as elasticity, can yield significant cost savings, but also raises new challenges for the application providers, particularly in an Infrastructure as a Service (IaaS) cloud. In this chapter we present a User centric cost efficient system that provides efficient support for elasticity in the cloud by:

i Leveraging multiple mechanisms to reduce the time to transition to new configurations, and

ii Optimizing the selection of a virtual server configuration that minimizes the cost.

V .CONCLUSION

The parameters used in this model are:

- For High performance, 600 cloudlets are allocated.
- For Medium performance, 400 cloudlets are allocated.
- For Low performance, 200 cloudlets are allocated.

Table 1 shows the results for particular parameters and workload, for all three types of configurations i.e. high performance, medium performance and low performance. From this table, it is observed that the cost of high performance model is high while

- Total Cost of VMs in High performance configuration is approx. 247% more than low performance configuration.
- High energy consumption in high performance configuration.
- Service Level Agreement (SLA) is more achieved in High performance configuration as compared to other two configurations.

Thus this model provides a different approach to users so that they can configure their cloud as per their own requirements without any interference of service provider using an interface. This tends to bring more power in the hands of users.

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

We have identified the security techniques that are used in the case of when data resides in the Cloud in Systematic process. The identified challenges, mitigation techniques and compromised attributes are described in Appendix section. The few popular security methods are Secure Socket Layer (SSL) Encryption; Multi Tenancy based Access Control, Intrusion Detection System, Novel Cloud dependability model, Hadoop Distributed File System and Hypervisor. From the Analysis of results from survey we have identified the following security challenges.

VI. References

[1] Peter Mell, Timoty Grance. "The NIST definition of cloud computing." National Institute of Standards and Technology, Special publication 800-145, September 2011.

[2] Pankaj Sareen. "Cloud Computing: Types, Architecture, Applications, Concerns, Virtualizations and Role of IT Governance in cloud." IJARCSSE, Vol. 3, issue 3, pp 533-538, March 2013.

[3] Gujarati, D.N., Essentials of Econometrics. McGraw-Hill Education, 2009.

[4] SIGOPS Operating System Review Vol. 41, issue 5, pp. 205-220, October 2007.

[5] Armbrust, Michael, Fox, Armando, Griffith, Rean, Joseph, Anthony D., Katz, Randy, Konwinski, Andy, Lee, Gunho, Patterson, David, Rabkin, Ariel, Stoica, Ion, and Zaharia, Matei, ―A view of cloud computing. Commun.,‖ ACM, Vol. 53, issue 4, pp.50-58, Apr. 2010.

[6] ―Interoute from the ground to the cloud. What is SaaS.‖

http://www.interoute.com/what-saas, January 2014.

[7] ―Interoute from the ground to the cloud. What is PaaS." http://www.interoute.com/what-paas, January, 2014.

[8] Nurmi et al. in Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., and Zagorodnov, ―The eucalyptus open-source cloudcomputing system In Cluster Computing and the Grid, 2009,‖ CCGRID '09, 9th IEEE/ACM International Symposium, pp. 124-131, May 2009.

[9] ―Open Nebula: The Open Source Toolkit for Data Center Virtualization.‖

http://www.opennebula.org, January 2014.

[10] ―Openstack: Cloud Software.‖ http://www.openstack.org., January 2014.

[11] ―What is IaaS.‖ http://www.interoute.com/what-iaas, January 2014.

[12] Hamilton, James R. ―Architecture for modular data centers,‖ CoRR

abs/cs/0612110(2006).

[13] Al-Fares, Mohammad, Loukissas, Alexander, and Vahdat, Amin. ―A scalable, commodity data center network architecture.‖ ACM SIGCOMM 2008 conference on Data communication (New York, NY, USA, 2008), SIGCOMM '08, ACM, pp. 63–74.

[14] Katz, R. H. ―Tech titans building boom.‖ IEEE Spectr., Vol. 46, issue 2, pp. 40-54, Feb. 2009.

[15] Nanda, Susanta, and Chiueh,Tzi. ―A survey of virtualization technologies.‖

Tech.rep., Department of Computer Science, SUNY at Stony Brook, 2005.

[16] Chen, Shigang, and Nahrstedt,Klara. ―Hierarchical scheduling for multiple classes of applications in connection-oriented integrated-service networks.‖ IEEE International Conference on Multimedia Computing and Systems, Volume 2 (Washington, DC, USA, 1999), ICMCS '99, IEEE Computer Society, pp. 9153, 1999.

[17] Demers, A., Keshav, S., and Shenker, S. ―Analysis and simulation of a fair queueing algorithm.‖ SIGCOMM Comput.Commun. Rev., Vol. 19, issue 4, pp 1-12, Aug. 1989.

[18] Duda, KennethJ., and Cheriton, David R. Borrowed

―virtualtime(BVT)scheduling: supporting latency-sensitive threads in a generalpurpose scheduler.‖ SIGOPS Oper. Syst. Rev,.Vol 33,issue 5, pp. 261- 276, Dec. 1999.

[19] Jones, Michael B., Ros¸u, Daniela, and Ros¸u, Marcel-Catalin. ―Cpu reservations and time constraints: efficient, predictable scheduling of independent activities.‖ 16th ACM symposium on Operating systems principles (New York, NY, USA, 1997), SOSP '97, ACM, pp. 198–211, 1997.

[20] Waldspurger, C. A. ―Lottery and stride scheduling: Flexible proportionalshare resource management.‖ Tech. rep., Cambridge, MA, USA, 1995.

[21] Matthews, Jeanna Neefe, Hu, Wenjin, Hapuarachchi, Madhujith, Deshane, Todd,Dimatos, Demetrios, Hamilton, Gary, McCabe, Michael, and Owens, James. ―Quantifying the performance isolation properties of virtualization systems.‖ Workshop on Experimental computer science

(NewYork,NY,USA,2007), ExpCS '07, ACM.

[22] Gupta, Diwaker, Cherkasova, Ludmila, Gardner, Rob, and Vahdat, Amin.

―Enforcing performance isolation across virtual machines in xen.‖ ACM/IFIP/USENIX 2006 International Conference on Middleware (New York, NY, USA,2006), Middleware '06, Springer-Verlag New York, Inc., pp. 342–362.

[23] Cerbelaud, Damien, Garg, Shishir, and Huylebroeck, Jeremy. ―Opening the clouds: qualitative overview of the state-of-the-art open source vm-based cloud management platforms.‖ 10th ACM/IFIP/USENIX International Conference on Middleware (New York, NY, USA, 2009), Middleware'09, Springer-Verlag New York, Inc., pp. 22:1–22:8.

[24] Chase, Jeffrey S., Anderson, Darrell C., Thakar, Prachi N., Vahdat, Amin M., and Doyle, Ronald P. ―Managing energy and server resources in hosting centers.‖ 18th ACM symposium on Operating systems principles (New York, NY, USA, 2001), SOSP '01, ACM, pp. 103–116.

[25] Vahdat, Amin. ―Future directions in distributed computing: Dynamically provisioning distributed systems to meet target levels of performance, availability, and data quality‖ Springer-Verlag, Berlin, Heidelberg, pp. 127– 131, 2003.

[26] Raghavendra, Ramya, Ranganathan, Parthasarathy, Talwar, Vanish, Wang,

Zhikui, and Zhu, Xiaoyun. ―No ‗power' struggles: coordinated multi-level power management for the data center.‖ 13th international conference on

Architectural support for programming languages and operating systems (New York,NY, USA, 2008), ASPLOS XIII, ACM, pp. 48–59.

[27] "Search data center: Green Computing", http://searchdatacenter.techtarget.com/definition/green-computing, February 2014.

[28] Mell, Peter, and Grance, Tim. ―Effectively and Securely Using the Cloud Computing Paradigm.‖ May 2009.

[29] Perilli, Alessandro, Manieri, Andrea, Algom, Avner, Balding, Craig. ―Cloud Computing Risk Assessment‖ ENISA. Tech.rep., ENISA, Greece, Nov.2009.

[30] Chen, Y., Paxson, V., and Katz, R. ―What‗s New About Cloud Computing Security.‖ University of California, Berkeley Report No. UCB/EECS-2010-5 January 20, 2010 (2010), 2010–5.

[31] Subashini, S., and Kavitha, V. ―Review: A survey on security issues in service delivery models of cloud computing.‖ J. Netw. Comput. Appl. 34, 1 (Jan. 2011), 1–11.

[32] "Search data center: cloud provisioning." . http://searchcloudprovider.techtarget.com/definition/cloud-provisioning, Februrary, 2014.

[33] Armbrust, Michael, Fox, Armando, Griffith, Rean, Joseph, Anthony D, Katz,Randy H, Konwinski, Andrew, Lee, Gunho, Patterson, David, Rabkin, Ariel, Stoica, Ion, and Zaharia, Matei. ―Above the clouds: A berkeley view of cloud computing.‖ Tech. Rep. UCB/EECS-2009-28, EECS Department, University of California,Berkeley, Feb. 2009.

[34] Duda, Kenneth J., Cheriton, David. ―Borrowed-virtual-time (BVT) scheduling: supporting latency-sensitive threads in a general-purpose scheduler.‖ SIGOPS Oper. Syst. Rev., Vol. 33,Issue 5, pp.261-276, Dec. 1999.

[35] Goyal, Pawan, Vin, Harrick M., and Cheng, Haichen. ―Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks.‖ IEEE/ACM T rans.Netw.5,Vol. 5, pp. 690-704, October 1997.

[36] Gulati, Ajay, Shanmuganathan, Ganesha, Ahmad, Irfan, Waldspurger, Carl, and Uysal, Mustafa. ―Pesto: online storage performance management in virtualized datacenters.‖ SOCC (New York, NY, USA, 2011), SOCC '11, ACM, pp. 19:1–19:14.

[37] Shen, Zhiming, Subbiah, Sethuraman, Gu, Xiaohui, and Wilkes, John.

―Cloudscale: elastic resource scaling for multi-tenant cloud systems.‖ SOCC '11 (New York, NY, USA, 2011), ACM, pp. 5:1–5:14.

[38] Wood, Timothy, Shenoy, Prashant, Venkataramani, Arun, and Yousif, Mazin.

―Sandpiper: Black-Box and Gray-Box Resource Management For Virtual Machines.‖ Computer Networks: The International Journal of Computer and Telecommunications Networking Vol. 53,Issue 17, Dec. 2009.

[39] Wood, T., Tarasuk-Levin, G., Shenoy, P., Desnoyers, P., Cecchet, E., and Corner,M. ―Memory buddies: Exploiting page sharing for smart colocation in virtualized data centers.‖ International Conference on Virtual Execution

Environments (VEE'09), pp. 31-40, April 2009.

[40] Shivam, Piyush, Marupadi, Varun, Chase, Jeff, Subramaniam, Thileepan, and Babu, Shivnath. ―Cutting corners: workbench automation for server benchmarking.‖ USENIX2008 Annual Technical Conference on Annual

Technical Conference (Berkeley, CA, USA, 2008), ATC‗08, USENIX Association, pp. 241–254, 2008.

[41] Zheng, Wei, Bianchini, Ricardo, Janakiraman, G. John, Santos, Jose Renato, and Turner, Yoshio. ―Justrunit: experiment-based management of virtualized data centers.‖ USENIX Annual technical conference (Berkeley, CA, USA, 2009), USENIX‗09, USENIX Association, pp. 18–18, 2009.

[42] Sotomayor, Borja, Montero, Ruben S., Llorente, Ignacio M., and Foster, Ian.

―Virtual infrastructure management in private and hybrid clouds.‖ IEEE Internet Computing, Vol. 13,issue 5, pp. 14-22, Sept. 2009.

[43] Telecommunications Magazine, Special Issue on Service Creation. January 2000.

[44] N.Krishnaveni, G.Sivakumar. ―Survey on Dynamic Resource Allocation Strategy in Cloud Computing Environment.‖ International Journal of Computer Applications Technology and Research, Vol. 2,Issue 6,pp. 731 - 737, 2013.

[45] V.Vinothina, Dr.R.Sridaran, Dr.Padmavathi Ganapathi. ―A Survey on Resource Allocation Strategies in Cloud Computing.‖ (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, issue.6, 2012.

[46] Dr. V. Vaithiyanathan, S. Ram Kumar and M. Lavanya. ―A Review on Different Scheduling Algorithms in cloud environment.‖ International Journal of Applied Engineering Research, ISSN 0973-4562, Vol. 8, issue 20, 2013.

[47] Saeed Parsa and Reza Entezari-Maleki. ―RASA: ANew Task Scheduling Algorithm in Grid Environment.‖ World Applied Sciences Journal, Vol 7, pp.152-160, 2009.

[48] H. Topcuoglu, S. Hariri, and M.-Y. Wu, ―Performance-Effective and Low- Complexity TaskScheduling for Heterogeneous Computing‖, IEEE Trans. Parallel and Distrib.Sys., pp. 260–274, 2002.

[49] Arash Ghorbannia Delavar, Mahdi Javanmard etc.al, ‖RSDC (Reliable Scheduling Distributed in Cloud Computing)‖ International Journal of ComputerScience, Engineering and Applications (IJCSEA) ,June2012.

[50] L. F. Bittencourt and E. R. M. Madeira, ―HCOC: A Cost Optimization Algorithm for Work flow Scheduling in Hybrid Clouds‖, J. Internet Svcs. AndApp., pp. 207–227, December 2011.

[51] [Dr. M. Dakshayini, Dr. H. S. Guruprasad. ―AnOptimal Model for Priority based Service Scheduling Policy for Cloud Computing Environment‖ International Journal of Computer Applications, October 2011.

[52] Shamsollah Ghanbari, Mohamed Othman , ―A Priority based Job Scheduling Algorithm in Cloud Computing‖, International Conference on Advances Science and Contemporary Engineering, ICASCE,2012.

[53] S. Abrishami,etc al.,‖ Cost-Driven Scheduling of Grid Workflows Using Partial Critical Paths‖, 11thIEEE/ACM GRID, pp.81-88, October 2010.

[54] [El-Sayed T. El-kenawy, Ali Ibraheem El-Desoky,etc.al.,―Extended Max-Min Scheduling Using PetriNet and Load Balancing‖, International Journal of SoftComputing and Engineering (IJSCE),September 2012.]

[55] S. Pandey, etc al., ―A Particle Swarm Optimization-Based Heuristic for Scheduling WorkflowApplications in Cloud Computing Environments‖,24th IEEE AINA, pp.400-407, April 2010.

[56] ―Class Cloud Information Service‖ http://www.cloudbus.org/cloudsim/doc/api/org/cloudbus/cloudsim/core/ Cloud InformationService.html, March 2014.

[57] Upendra Sharma, ―Elastic Resource Management in Cloud Computing Platforms‖ University of Massachusetts-Amherst.

[58] Urgaonkar, Bhuvan, Pacifici, Giovanni, Shenoy, Prashant, Spreitzer, Mike, and Tantawi, Assar. ―An Analytical Model for Multi-tier Internet Services and Its Applications.‖ ACM SIGMETRICS Conf., Banff, Canada, June 2005.

[59] Kailasam, Sriram, Gnanasambandam, Nathan, Dharanipragada, Janakiram, and Sharma, Naveen. ―Optimizing service level agreements for autonomic cloud bursting schedulers.‖ ICPP Workshops, pp. 285–294, 2010.

[60] ―The Cloud Computing and Distributed Systems (CLOUDS) Laboratory, University of Melbourne" http://www.cloudbus.org/cloudsim/, June 2014.

[61] ―Class Cloud Information Service‖ http://www.cloudbus.org/cloudsim/doc/api/org/cloudbus/cloudsim/core/ Cloud InformationService.html, June 2014..