Cost-Efficient Virtual Machine Provisioning for Multi-tier Web Applications and Video Transcoding

Adnan Ashraf
adnan.ashraf@abo.fi
Department of Information Technologies, Åbo Akademi University, Turku, Finland
Turku Centre for Computer Science (TUCS), Turku, Finland
Department of Software Engineering, International Islamic University, Islamabad, Pakistan

***Abstract*—Infrastructure as a Service (IaaS) clouds provide virtual machines (VMs) under the pay-per-use business model. The dynamic on-demand provisioning of VMs allows IaaS users to ensure scalability of their web applications and web-based services from really low to really high loads. However, VM provisioning must be done carefully because over-provisioning results in an increased operational cost, while under-provisioning leads to a subpar service. In this research work, our main focus is on cost-efficient VM provisioning for multi-tier web applications and video transcoding. Moreover, to prevent provisioned VMs from becoming overloaded, we augment VM provisioning with an admission control mechanism. Similarly, to ensure efficient use of provisioned VMs, under-utilized VMs are consolidated periodically. Since cost-efficient VM provisioning is an optimization problem, we apply metaheuristic approaches to find a near-optimal solution.**

***Index Terms*—Cloud computing; web applications; video transcoding; virtual machine provisioning; admission control; server consolidation**

I. Introduction

Web applications are often deployed in a three-tier computer architecture that consists of client, application, and database tiers [1]. The client tier runs within the user web browser, while the application server and the database server tiers run in the remote server infrastructure. Both the application and the database tiers are implemented using a computer cluster to be able to process many user requests simultaneously. In this configuration, a load balancing subsystem distributes the user requests among the computers in the cluster. Traditionally, these clusters are composed of a fixed number of computers and are dimensioned to serve a predetermined maximum number of concurrent users.

A web-based video streaming service is also implemented using a cluster-based distributed system, which may consist of different types of servers, such as, video streaming servers and video transcoding servers. A video transcoding server converts a compressed video from one format to another [2]. It may change video format, bit rate, frame resolution, frame rate, or any combination of these [3]. Video transcoding is a compute-intensive operation. For an on-demand video streaming service, it may be necessary to transcode a large number of videos on-the-fly in realtime. Transcoding of a large number of simultaneous videos necessitates the need for a cluster of video transcoding servers.

Infrastructure as a Service (IaaS) clouds, such as Amazon Elastic Compute Cloud (EC2) [4], provide virtual machines (VMs) under the pay-per-use business model. The dynamic on-demand provisioning of VMs allows IaaS users to deploy and scale their web applications and web-based services without requiring to invest into large-scale IT infrastructures. With cloud elasticity, it is possible to create a dynamically scalable cluster of servers consisting of a varying number of VMs. However, VM provisioning must be done carefully because over-provisioning results in an increased operational cost, while under-provisioning leads to a subpar service, which may violate users' Quality of Service (QoS) requirements concerning performance, resulting in a loss of revenue.

Determining the number of VMs to provision for a cluster is an important problem as the exact number of VMs needed at a specific time depends upon the user load and the QoS requirements. In this work, our main goal is cost-efficient VM provisioning for multi-tier web applications and video transcoding. Moreover, to prevent provisioned VMs from becoming overloaded, we augment VM provisioning with an admission control mechanism. For cost-efficiency, it is also necessary to reduce under-utilization of servers in a cluster. As a recent study showed that $80\% - 85\%$ under-utilization of servers is common in enterprises [5]. Under-utilization of VMs can be reduced by using server consolidation techniques similar to those used in data centers for power-efficiency [5], [6], [7].

Cost-efficient VM provisioning with VM consolidation and admission control is a combinatorial optimization problem [8]. Therefore, we apply metaheuristic approaches [8], [9] to find a near-optimal solution.

II. Proposed Approach

The main problem that we intend to tackle is cost-efficient VM provisioning with augmented server consolidation and overload control on provisioned VMs. We seek solutions for multi-tier web applications and on-demand video transcoding. Although there are many similarities between VM provisioning for web applications and VM provisioning for video transcoding, each one of them also has its own challenges. In this section, we present the proposed approach while providing a brief overview of some of the most important challenges that it addresses.

We propose a cost-efficient VM provisioning approach for multi-tier web applications [1], [10], [11] and on-demand video transcoding [3]. Moreover, for preventing servers from becoming overloaded, the VM provisioning approach is augmented with an admission control mechanism [12], [13]. Similarly, the underutilization of VMs is minimized by providing a VM consolidation mechanism.

A. VM Provisioning Delay

In practice, it takes a few minutes to provision a VM from an IaaS provider [1], [10]. Due to the inevitable VM provisioning delay, handling of a sudden spike in the incoming user load becomes a challenge. Some of the strategies that we use to overcome this drawback of public IaaS clouds include provisioning multiple VMs at a time [1], [10], using additional VM capacity [1], [10], and using load prediction to provision proactively [3], [10].

B. Admission Control

Resource allocation alone does not prevent servers from becoming overloaded [12], [13], [14], [15]. One of the main reason is that the load balancer may always direct some new requests or load to an already overloaded server. Another reason is the VM provisioning delay, which may lead to over-admission on existing VMs. Therefore, to prevent servers from becoming overloaded, VM provisioning needs to be augmented with an admission control mechanism. Traditional admission control approaches that are designed for a fixed number of servers may be evaluated based on server overload prevention and gain in throughput. However, the main challenge for dynamically scalable clusters is to devise an admission control mechanism that leverages cloud elasticity to provide a good tradeoff between cost and QoS. We use session-based adaptive admission control for web applications [12]. Likewise, for video transcoding, we use a stream-based admission control approach [13].

C. Load Prediction and Proactive Provisioning

Many traditional VM provisioning approaches, such as [1], [16], [17], [18], [19], use reactive provisioning. However, the primary shortcoming of reactive provisioning is that it starts a provisioning operation only after a significant increase in the load is detected [20]. Therefore, the new VMs can only be used instantly if the VM provisioning is instantaneous [10]. However, due to the VM provisioning delay, the reactive approach may fail to handle increased load, especially under sudden load spikes. Alternatively, some approaches use prediction of future load to provision preemptively [20], [21]. We use a proactive approach for video transcoding [3] and a hybrid approach for web applications that assigns certain weights to reactive and proactive provisioning [10]. The main challenge in prediction-based approaches is in making predictions with high prediction accuracy under realtime constraints [10]. We use a two-step load prediction method [22] with a simple linear regression model [12], which predicts a few steps ahead in the future with high prediction accuracy under realtime constraints.

D. Reduced Oscillations in Number of VMs

Another important challenge is to reduce oscillations in the number of provisioned VMs. This is desirable because oscillations in the presence of VM provisioning delay may lead to deteriorated performance [1]. Moreover, since some IaaS providers, such as Amazon EC2, charge on hourly basis, oscillations in the number of provisioned VMs may result in a higher provisioning cost [3]. We use a few strategies to counteract oscillations in the number of VMs, such as, delaying new provisioning operations until previous provisioning operations have been realized [1], [19] and terminating only those VMs that are constantly under-utilized for a longer period of time [1] and whose renting hour approaches its completion [3].

E. Sharing of VM Resources for Improved Utilization

For cost-efficient VM provisioning to deploy and scale multiple web applications, the proposed approach should provide a finer deployment granularity than the smallest VM provided by the contemporary IaaS providers [10]. This is especially important when deploying a large number of web applications, most of which may have very few users, while a few of them may have many users. We use shared hosting, which deploys one or more web applications on each VM [1], [10]. Moreover, popular applications would often be deployed in many VMs, while sporadically used applications would not be deployed at all in order to save resources. Therefore, at any given time, an application may be deployed in zero, one, or more VMs [1]. Thus, instead of provisioning at least one full VM per application, shared hosting effectively supports provisioning a fraction of a VM per application, resulting in a reduced number of total VMs. Deployment of multiple web application in a shared hosting environment enables two levels of scaling, namely server-level scaling and application-level scaling [1]. Server-level scaling provisions and terminates VMs from an IaaS cloud to create a dynamically scalable cluster of servers, whilst application-level scaling deploys and removes web applications from each virtualized server.

For video transcoding, we use video segmentation at Group of Pictures (GOP) level, which splits video streams into smaller segments that can be transcoded independently of one another [3]. It allows transcoding of multiple video streams concurrently on a single VM. The sharing of VM resources among multiple concurrent streams improves VM utilization, which helps in reducing total number of required VMs.

F. Reduced Number of VMs and Migrations

When deploying and scaling a large number of web applications in a shared hosting environment, it is important to consolidate under-utilized VMs from time to time in order to reduce under-utilization of VMs and consequently total number of provisioned VMs. In this case, server consolidation [5], [6], [7] should migrate all active sessions and web applications from the least loaded under-utilized VMs to other VMs and then terminate the least loaded VMs. This is achievable with live VM migration [6]. However, live migration is a resource-intensive operation. Therefore, the main challenge here is to augment VM provisioning with a server consolidation mechanism, which uses a reduced number of VMs along with a reduced number of VM migrations.

For video transcoding, server consolidation may not require live migration because video segments can be transcoded independently of one another and transcoding of segments takes relatively less time to complete [3]. Therefore, it may be more reasonable to let the segments complete their execution and then terminate the VM when there are no more running and pending segments on it.

G. Automatic Adjustment and Adaptability

To ensure cost-efficiency and QoS under diverse load conditions, it is necessary that the proposed VM provisioning, admission control, and server consolidation approaches automatically adjust and adapt themselves according to the load conditions. For admission control, we use a weighting coefficient, which is automatically adjusted and tuned based on four different parameters that represent load conditions [12]. We also use a similar weighting coefficient for prediction-based VM provisioning [10]. It is calculated based on the prediction error.

H. Selection of Appropriate Parameters

Selection of appropriate parameters to represent different load conditions, QoS, and server performance is also an important challenge. For instance, one of the most intuitive performance measure for web applications is response time [23]. However, an application may have different expected response times for different types of user requests [1]. Therefore, expected response time may be difficult to define for a given application. Alternatively, it may be reasonable to use server-level and application-level resource utilization metrics as an indication of load conditions and performance [10]. We use commonly used resource utilization metrics, which include CPU load average, memory utilization, and network utilization [1], [10], [12].

On-demand video transcoding is a compute and memory intensive operation. Consequently, resource utilization metrics, such as CPU load average and memory utilization, are not appropriate indicators of load conditions and performance. Therefore, we use video transcoding rate and video play rate for VM provisioning for on-demand video transcoding [3].

III. Related Work

Most of the existing works on VM provisioning and dynamic resource allocation for web-based systems can be classified into two main categories: Plan-based approaches and control theoretic approaches [24], [25], [26]. Plan-based approaches can be further classified into workload prediction approaches [20], [21] and performance dynamics model approaches [16], [17], [18], [19], [27]. One common characteristic of all of these existing works is that they do not use shared hosting. Another common characteristic is that they only provide a server-level scaling mechanism. Whereas, our proposed approach for web applications [1] also provides a separate mechanism for scaling of individual web applications.

There are currently only a few approaches for cloud-based distributed video transcoding, such as [28], [29]. However, they do not address VM provisioning problem for on-demand video transcoding.

Server consolidation approaches, such as [5], [6], [7], dynamically reallocate VMs to physical nodes with the aim of reducing total number of required nodes. However, in the context of cost-efficient VM provisioning from an IaaS cloud, we require a different type of server consolidation. It should periodically migrate all active web applications and user sessions from the least loaded under-utilized VMs to other VMs. Thus, releasing the least loaded VMs for termination. Therefore, our goal is to reduce number of provisioned VMs and their renting durations, rather than reducing number of physical nodes.

Admission control approaches, such as [14], [15], [30], [31], [32], [33], [34], [35], aim to prevent server overloading under high load situations. One common characteristic of these traditional approaches, except [35], is that they make decisions only on acceptance or rejection of incoming user load. The approach in [35] has its own disadvantages. The discount-charge model of [35] requires additional web pages to be included in the web application and it is only effective for e-commerce web sites when more users place orders. In the context of cloud computing, it may also be possible to defer the incoming load until some new VMs are provisioned or some existing VMs become less loaded. Therefore, there is an opportunity to develop an admission control mechanism, which may choose between using an existing VM or provisioning a new VM for accommodating new incoming load [12], [13].

IV. Expected Contributions and Obtained Results

The main expected contribution is a cost-efficient VM provisioning approach for multiple multi-tier web applications and on-demand video transcoding. Some more specific contributions include VM provisioning approach for web applications [1], [10], [11] a session-based adaptive admission control approach for web applications [12], an approach for automatic load generation on web applications [36], a VM provisioning approach for video transcoding [3], a video stream based admission control approach for video transcoding [13], and server consolidation approaches for web applications and video transcoding. The proposed cost-efficient VM provisioning approach is validated with discrete-event simulations and prototype implementations.

The first outcome of this work is an approach for automatic load generation for performance and scalability testing of web applications called ASTORIA [36]. Cost-efficient VM provisioning with integrated admission control and server consolidation mechanisms is an ongoing research project. We have proposed an approach to create dynamically scalable application server tiers to deploy and scale multiple web applications per VM [1], [10], [11], a prediction-based VM provisioning approach for video transcoding [3], a session-based adaptive admission control approach for virtualized application servers [12], and an approach for video stream based admission control for video transcoding [13]. We have been currently working on server consolidation approaches for web applications and video transcoding. Moreover, applying metaheuristic approaches [8], [9] to optimize cost-efficiency is also part of our ongoing research.

Acknowledgements

The author wants to thank Professor Ivan Porres at Åbo Akademi University for his guidance and encouragement.

References

[1] A. Ashraf, B. Byholm, J. Lehtinen, and I. Porres, "Feedback control algorithms to deploy and scale multiple web applications per virtual machine," *38th Euromicro Conference on Software Engineering and Advanced Applications*, September 2012.

[2] A. Vetro, C. Christopoulos, and H. Sun, "Video transcoding architectures and techniques: an overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 2, pp. 18 – 29, mar 2003.

[3] F. Jokhio, A. Ashraf, S. Lafond, I. Porres, and J. Lilius, "Prediction-based dynamic resource allocation for video transcoding in cloud computing," *21st Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pp. 254–261, 2013.

[4] "Amazon Elastic Compute Cloud." [Online]. Available: http://aws.amazon.com/ec2/

[5] W. Vogels, "Beyond server consolidation," *ACM Queue*, vol. 6, no. 1, pp. 20–26, Jan. 2008.

[6] A. Murtazaev and S. Oh, "Sercon: Server consolidation algorithm using live migration of virtual machines for green computing," *IETE Technical Review*, vol. 28, no. 3, pp. 212–231, 2011.

[7] E. Feller, C. Morin, and A. Esnault, "A case for fully decentralized dynamic VM consolidation in clouds," *Cloud Computing Technology and Science, IEEE International Conference on*, pp. 26–33, 2012.

[8] C. Blum, J. Puchinger, G. R. Raidl, and A. Roli, "Hybrid metaheuristics in combinatorial optimization: A survey," *Applied Soft Computing*, vol. 11, no. 6, pp. 4135 – 4151, 2011.

[9] M. Harman, K. Lakhotia, J. Singer, D. R. White, and S. Yoo, "Cloud engineering is search based software engineering too," *Journal of Systems and Software*, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2012.10.027

[10] A. Ashraf, B. Byholm, and I. Porres, "CRAMP: Cost-efficient resource allocation for multiple web applications with proactive scaling," *4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, December 2012.

[11] T. Aho, A. Ashraf, M. Englund, J. Katajamäki, J. Koskinen, J. Lautamäki, A. Nieminen, I. Porres, and I. Turunen, "Designing IDE as a service," *Communications of Cloud Software*, vol. 1, no. 1, 2011.

[12] A. Ashraf, B. Byholm, and I. Porres, "A session-based adaptive admission control approach for virtualized application servers," *5th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*, November 2012.

[13] A. Ashraf, F. Jokhio, T. Deneke, S. Lafond, I. Porres, and J. Lilius, "Stream-based admission control and scheduling for video transcoding in cloud computing," in *13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 2013.

[14] S. Muppala and X. Zhou, "Coordinated session-based admission control with statistical learning for multi-tier internet applications," *Journal of Network and Computer Applications*, vol. 34, no. 1, pp. 20 – 29, 2011.

[15] L. Cherkasova and P. Phaal, "Session-based admission control: a mechanism for peak load management of commercial web sites," *Computers, IEEE Transactions on*, vol. 51, no. 6, pp. 669 –685, jun 2002.

[16] A. Wolke and G. Meixner, "TwoSpot: A cloud platform for scaling out web applications dynamically," in *Towards a Service-Based Internet*, ser. Lecture Notes in Computer Science, E. Di Nitto and R. Yahyapour, Eds. Springer Berlin / Heidelberg, 2010, vol. 6481, pp. 13–24.

[17] Y. Hu, J. Wong, G. Iszlai, and M. Litoiu, "Resource provisioning for cloud computing," in *Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, ser. CASCON '09. New York, NY, USA: ACM, 2009, pp. 101–111.

[18] T. Chieu, A. Mohindra, A. Karve, and A. Segal, "Dynamic scaling of web applications in a virtualized cloud computing environment," in *e-Business Engineering, 2009. ICEBE '09. IEEE International Conference on*, oct. 2009, pp. 281 –286.

[19] W. Iqbal, M. N. Dailey, D. Carrera, and P. Janecek, "Adaptive resource provisioning for read intensive multi-tier applications in the cloud," *Future Generation Computer Systems*, vol. 27, no. 6, pp. 871–879, 2011.

[20] Y. Raivio, O. Mazhelis, K. Annapureddy, R. Mallavarapu, and P. Tyrväinen, "Hybrid cloud architecture for short message services," in *Proceedings of the 2nd International Conference on Cloud Computing and Services Science*, ser. CLOSER '12, 2012.

[21] D. Ardagna, C. Ghezzi, B. Panicucci, and M. Trubian, "Service provisioning on the cloud: Distributed algorithms for joint capacity allocation and admission control," in *Towards a Service-Based Internet*, ser. Lecture Notes in Computer Science, E. Di Nitto and R. Yahyapour, Eds. Springer Berlin / Heidelberg, 2010, vol. 6481, pp. 1–12.

[22] M. Andreolini and S. Casolari, "Load prediction models in web-based systems," in *Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*, ser. valuetools '06. New York, NY, USA: ACM, 2006.

[23] H. Liu, *Software Performance and Scalability: A Quantitative Approach*, ser. Quantitative Software Engineering Series. John Wiley & Sons, 2011.

[24] X. Dutreilh, N. Rivierre, A. Moreau, J. Malenfant, and I. Truck, "From data center resource allocation to control theory and back," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pp. 410–417.

[25] W. Pan, D. Mu, H. Wu, and L. Yao, "Feedback control-based QoS guarantees in web application servers," in *High Performance Computing and Communications, 2008. HPCC '08. 10th IEEE International Conference on*, sept. 2008, pp. 328 –334.

[26] T. Patikirikorala, A. Colman, J. Han, and L. Wang, "A multi-model framework to implement self-managing control systems for QoS management," in *6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2011, pp. 218–227.

[27] R. Han, L. Guo, M. Ghanem, and Y. Guo, "Lightweight resource scaling for cloud applications," in *12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 2012, pp. 644 –651.

[28] Z. Huang, C. Mei, L. E. Li, and T. Woo, "CloudStream: Delivering high-quality streaming videos through a cloud-based SVC proxy," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 201–205.

[29] Z. Li, Y. Huang, G. Liu, F. Wang, Z.-L. Zhang, and Y. Dai, "Cloud transcoder: Bridging the format and resolution gap between internet videos and mobile devices," in *22nd ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2012.

[30] J. Almeida, V. Almeida, D. Ardagna, I. Cunha, C. Francalanci, and M. Trubian, "Joint admission control and resource allocation in virtualized servers," *J. Parallel Distrib. Comput.*, vol. 70, no. 4, pp. 344–362, Apr. 2010.

[31] C.-J. Huang, C.-L. Cheng, Y.-T. Chuang, and J.-S. R. Jang, "Admission control schemes for proportional differentiated services enabled internet servers using machine learning techniques," *Expert Systems with Applications*, vol. 31, no. 3, pp. 458 – 471, 2006.

[32] X. Chen, H. Chen, and P. Mohapatra, "ACES: An efficient admission control scheme for QoS-aware web servers," *Computer Communications*, vol. 26, no. 14, pp. 1581 – 1593, 2003.

[33] T. Voigt and P. Gunningberg, "Adaptive resource-based web server admission control," in *Computers and Communications (ISCC), 2002 Seventh International Symposium on*, pp. 219–224.

[34] A. Robertsson, B. Wittenmark, M. Kihl, and M. Andersson, "Admission control for web server systems - design and experimental evaluation," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 1, dec. 2004, pp. 531 –536 Vol.1.

[35] Y. A. Shaaban and J. Hillston, "Cost-based admission control for internet commerce QoS enhancement," *Electronic Commerce Research and Applications*, vol. 8, no. 3, pp. 142 – 159, 2009.

[36] N. Snellman, A. Ashraf, and I. Porres, "Towards automatic performance and scalability testing of rich internet applications in the cloud," *Software Engineering and Advanced Applications, Euromicro Conference*, pp. 161–169, 2011.