Virtual Machine Provisioning for Cloud Scenarios: A Survey of Approaches and Challenges

Lincolin Nhapi
Dept. of Computer Science & Engineering
ITM University Gwalior
Madhya Pradesh, India
+918225819218
lbnhapi@gmail.com

Arun Kumar Yadav
Dept. of Computer Science & Engineering
ITM University Gwalior
Madhya Pradesh, India
+919584600832
arun26977@gmail.com

Ram Shringar Rao
Dept. of Computer Science & Engineering
Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
+919968408090
rsrao08@yahoo.in

ABSTRACT
In this study, we present a review on proposed techniques for Virtual Machine Provisioning in cloud data centers. We analyze each technique and present what we think are the pros and cons of that particular technique. Furthermore, we also highlight current research issues in Cloud Computing and give a conclusion.

Keywords
Virtual Machine, Cloud Computing, Virtualization, Server consolidation, Migration, Cloud service provider.

1. INTRODUCTION
Computing is going through a transformation process where the goal is to model computing services into commodities that can be offered to the public in a way similar to services like electricity, water, telephone and gas. In such kind of a model, users can access services depending on their requisite, irrespective of where the services are hosted. Utility computing has been a long vision of computing that other computing paradigms, like grid computing, have attempted to deliver. The recently emerging paradigm with the potential to deliver computing as utilities is cloud computing [1]. Cloud computing is advancement in technology that focuses on how computing systems are designed, how applications are developed, and takes advantage of existing technologies. Its basis is on the notion of dynamic provisioning, which is applicable not merely to services but also to storage, networking, compute capability, and generally to computing infrastructure. Computing vendors offer resources that are available through the internet and on pay-as-you-go. Today, any person in possession of a credit card can easily make a subscription for cloud services and set up servers for an application within a short time. The infrastructure can be scaled according to demand and users pay only for the resources used.

ICTCS '16, March 04-05, 2016, Udaipur, India
 ISBN 978-1-4503-3962-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2905055.2905174

This paper is organized as follows, section 2 gives the definition of cloud computing and how it came into existence. Section 3 briefly examines other technologies related to cloud computing. In section 4 we introduce cloud service models and in section 5 we review virtual machine provisioning approaches. Section 6 presents issues and challenges in virtual machine provisioning and finally in section 7 we conclude our findings and give future recommendations.

2. CLOUD COMPUTING
The following section presents a general outline of cloud computing, together with its definition and a contrast with related ideas follows.

2.1 Definition
The vision of utility computing has been in existence from long ago. During the 20th century, a man named John McCarthy by that time had a vision that computing amenities would be publicly provisioned as utilities [2]. During the 1990s era, the term "cloud" was used in different contexts like describing big ATM networks. Nonetheless, the term began to gain popularity in 2006, having been used by Eric Schmidt, Google's Chairman, in his description of a model for providing business over the internet. Ever since, the phrase "cloud computing" has been used in many contexts and predominantly in marketing.

In this study, we embrace a definition of cloud computing coined by NIST's (National Institute of Standards and Technology) [3]. According to NIST, Cloud Computing is a paradigm that enables sharing of computing resources over a network. The resources are provisioned on-demand and involve minimal interaction between service provider and consumer.

The existence of different views of cloud computing stems from the fact that cloud computing is not a novel innovation but rather a new way of doing business that leverages existing technological advancements to run businesses efficiently. In fact, among the technologies that are leveraged by cloud computing, virtualization has been in existence and the same goes for utility-based pricing. In other words, cloud computing makes use of virtualization as well as utility-based pricing in order to satisfy the financial as well as technological demand of information technology.

3. ASSOCIATED TECHNOLOGIES
The following technologies are usually associated with cloud computing because of their similarities to it and are oftentimes compared to cloud computing.

3.1 Grid Computing

Scientific computations, which are generally computationally intensive, brought about the birth of Grid computing. In this model, distributed or dispersed computing resources are conjoined in order to complete a single computational objective. The similarity between cloud computing and grid computing is that grid computing additionally utilizes networked computational tools to accomplish application-level goals. Be that as it may, cloud computing goes a step ahead by adopting virtualization as its core technology. Cloud computing leverages virtualization to achieve dynamic resource provisioning and sharing.

3.2 Utility Computing

Utility computing models a computing paradigm that aims to provide resources on-demand and billing users according to usage as compared to fixed pricing. In a like manner, cloud computing adopts a utility based pricing model and because of that, it is often viewed as an attainment of utility computing. By using a pricing model that resembles basic utilities (e.g. water, electricity etc.), cloud service providers can surely decrease their operating costs and make the most of their resources.

3.3 Virtualization

The technological foundation upon which cloud computing is built is virtualization [4]. It is a computing paradigm that abstracts the functions of computing and their implementation details from physical hardware. Virtualized Servers are generally called Virtual Machines (VMs). Virtualization allows the separation among software and hardware, resources and processes as well as among users. Through virtualization, a virtual environment can be created on which different operating systems can be hosted on a single physical machine. Two main categories of virtualization are as follows. They are dependent upon the layer where virtualization occurs.

3.3.1 Hypervisor based Virtualization

This type of virtualization makes use of a software layer called the hypervisor. The responsibility of the hypervisor is to manage the resources of the physical hosts and to ensure that the services required for virtual machines to run are provided. The hypervisor also allocates resources such as memory, CPU cores, storage, and bandwidth to virtual machines. All VMs have no direct access to the underlying hardware layer and thus VMs direct their requests for resources to the hypervisor layer. There are two major implementations of this kind of virtualization:

a) Full Virtualization

This type of virtualization completely imitates system hardware. The operating system (OS) does not require any adjustments and so as the applications [5]. Virtualization is implemented transparently at hardware level. Common examples comprise VMware Workstation, VirtualBox, KVM and Microsoft Virtual PC.

b) Paravirtualization

In paravirtualization, there is need of adjusting the OS and perhaps its applications so as to completely benefit from virtualized hardware layer optimizations [5]. Hence, it achieves better performance than Full Virtualization. For example, Xen offers a paravirtualization solution.

3.3.2 Container based Virtualization

Container-based virtualization is an approach to virtualization where the virtualization layer runs inside the operating system. It is also known as operating system virtualization. The guest virtual machines are called containers. This technology does not imitate the entire hardware [6]. In container-based virtualization, each VM is allocated its own network space and process as opposed to being managed by a hypervisor. This kind of virtualization also permits running of numerous isolated containers. Some examples are OpenVZ and LinuxVServer [7]. This type of virtualization is more effective compared to legacy virtualization styles since virtualization is done within the operating system meaning that there is only one OS making the hardware calls. Limitations are not inevitable however, as an example; each guest must use the same operating system as sued by its host. Thus consequently, workload migration is very complex in comparison with an environment that supports hypervisor-based virtualization.

4. CLOUD SERVICE MODELS

Altogether, there exist three types of cloud service models (Fig.1). These are distinguishable by the type of service offered to customers.

4.1 Software as a Service (SaaS)

In this model, software is availed to consumers as a service over web interfaces or web browsers. The software service runs on infrastructure managed by the SaaS provider or by another third party infrastructure provider [8].

4.2 Platform as a Service (PaaS)

In this model, services such as web servers, operating system, database server's application programming interfaces and an execution environment are delivered as a computing platform. Application developers maintain control over their developed and deployed software and perhaps configure settings for the software-hosting environment [8]. The underlying software and hardware layers continue to be managed by the PaaS provider. This includes operating systems, storage, networks and servers.

4.3 Infrastructure as a Service (IaaS)

In this case, the services provided are typically informed of virtualized hardware. Hardware computing resources like network and storage are provisioned as services. The clients can run and deploy operating systems and applications of choice. The hardware resources are provisioned from a pool of servers distributed across several locations from data centers. The IaaS provider, as in all the other models, manages the physical infrastructure and clients are permitted to manage or control their virtual resources. The virtual infrastructure is typically made of virtual machines.

5. LITERATURE REVIEW

Zhang and Li [9] proposed two solutions for large scale virtual machine provisioning in clouds. One is uses an image booting mechanism while the other makes use of memory forking. Firstly, the proposed image booting solution was characterized by a committed image transferring architecture which reduces the time taken to copy a VMI file (Virtual Machine Image) from storage to a host. VMI file usually comprises of several gigabytes of data.

The second proposed solution takes advantage of memory-forking mechanism. They named it VM Thunder Plus. It is an improvement of the image booting method. The idea behind VM Thunder Plus is to reduce the delay that comes about as a result of applications initializing after booting. The authors were able to

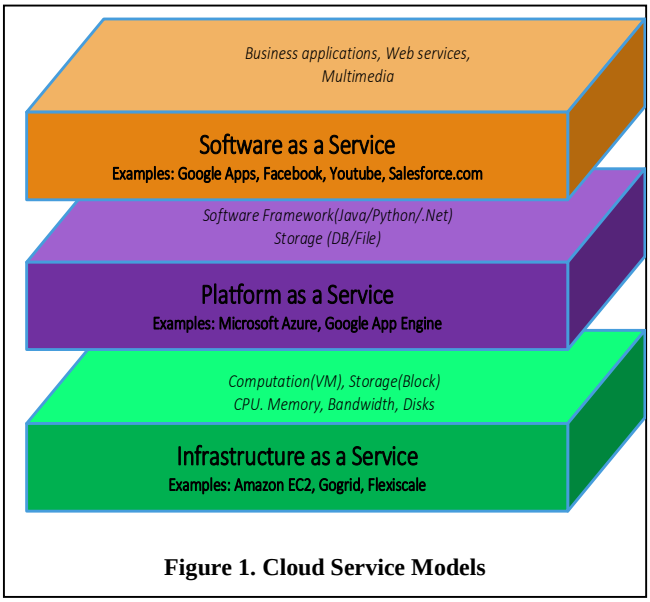

Figure 1. Cloud Service Models

achieve this by copying the state of memory when applications are initialized. The memory-state is forked from a template Virtual Machine to large scale. This method focuses on memory data flows. The advantage of these two solutions is the ability to scale and fast VM provisioning. The use of a peer-to-peer image streaming architecture also facilitates the fast and on-demand provisioning of VMs. The image booting method also enables bootstrapping of VMs at large scale without the need to modify the hypervisor. Additionally, good I/O performance of VM during runtime is also supported. The downside of this approach is each host and VM needs to maintain a cache repository of VMI file and so the method may require large amounts of memory and thus becomes costly to implement. The cache hit ratio is also not specified by the authors. In addition, the proposed solution assumed only same type of VMI file whereas in reality, VM requests are for different OS types.

Lo, Fan and Wu [10] looked at how VM provisioning mechanisms can be used to lower data center operational costs particularly saving electricity. They proposed a tri-schemed solution with the following three components. (1) An overload detection mechanism that determines the moment at which migration of VMs should begin, (2) A VM selection mechanism with the responsibility of pin pointing VMs to migrate from an overloaded host, and (3) A VM placement algorithm which selects a host that should house a migrated VM. This tri-schemed VM provisioning mechanism proposed by Lo et al took into consideration the issue of energy consumption and violation of Service Level Agreements (SLA). In comparison to other techniques, the authors were able to achieve greater CPU usage per host. The number of VM migrations was also reduced as well as violation of SLAs. However, there still remained a large number of idle hosts in their results which leads to high energy consumption. In addition, there were still several violations of SLAs together with numerous VM migrations which are resource intensive.

De et al [11] focused on speedup of virtual machine provisioning by reducing booting time. Transfer of the large image template file from an image repository to a compute node, and booting are the main causes of delay in the provisioning workflow. In order to solve this problem, De et al did a comparison of three methods for pre-provisioning VM instances. Each method comes up with the composition of the inventory, given a fixed size cache space. Using request logs, the incumbents determine the image templates which will be high in demand, and also approximate the number of requests for each VMI type. Upon receiving a matching request, a cached VM is promptly provided to the client. The cache or VM inventory space is freed up when a VM instance is delivered to a user. The inventory is replenished periodically with new VM instances. With this method, average service time was reduced compared to other caching techniques. The drawback of this method is that keeping VMs in standby mode may result in high power consumption as VM hosts will remain active in anticipation of requests. The authors also did not mention how runtime performance of hosts is impacted when this technique is applied. Thus the number of hosts required to run VMs may increase.

In order to make sure that QoS (Quality of Service) is met when provisioning virtual machines in cloud computing, Das and Adhikary [12] suggested the idea of recycling of VMs. They attempted to tackle the problem of making virtual machines repeatedly. The authors made an assumption that QoS can be achieved by monitoring the number of admitted VM requests. This, by the authors, will ensure that the system does not get overloaded. Here, several input queues were created which groups similar requests according to resource requirements. Each request in the queue or group of requests is served by the same virtual machine. This results in minimization of VM creation and destruction time to a certain extent. The authors also introduced priority among VM requests where higher priority requests are served on a newly created VM and they went on to simulate their model. By their results, they were able to achieve the targeted QoS and their model performs steadfastly. The highly probable limitation with this approach is, when forecasting the volume of computational resources needed as well as the type of queue where a request should be sent, the forecast may be incorrect. To add on, consider a scenario where all requests differ, this implies that a new VM should be created for serving each request. This leads to performance degradation of the proposed model.

Nejad et al [13] took into consideration the scarcity and heterogeneity of cloud resources while addressing the challenge of VM provisioning. They proposed a “truthfully greedy” methodology so as to solve the problem of VM provisioning and allocation. They named it G-VMPAC-II. The proposed mechanism determines the VM provisioning and allocation as well as user payments. In addition, they determined an estimate of ratio for the proposed methodology, following which they designed “truthful greedy mechanisms” regardless of common knowledge that greedy algorithms do not meet the requirements needed to ensure truthfulness. By this, the allocation and payment determination were designed so as to fulfil the truthfulness property. They claimed that their approach allows dynamic provisioning of VMs, and has no need of pre-provisioning virtual machines.

According their experimental outcome, the proposed greedy methods attain near optimal solutions while at the same time apprehending the volatile market. The proposed mechanism also provisions computing resources to meet the demand and generates more income. Additionally, the execution time is very little. This approach zeroed on maximizing revenue whereas reducing costs especially electricity can also increase profits and has an added benefit of reducing carbon dioxide emissions into the atmosphere.

Kochut and Karve [14] put forward an algorithm for VM Provisioning that is based on similarity between virtual machine images. The key idea here is to lessen amount of data that has to be transferred from the storage device to the hypervisor when booting VMs. The proposed system uses direct attached storage to store virtual machine temporary data. This approach minimizes the burden on the storage server and reduces bandwidth used. When a portion of a VM image to be provisioned is available on the host (as a result of other similar VMs on the host), this mechanism reconstructs that part of the image without the necessity of transferring the image from the storage device.

A testbed was created for the implementation of the proposed algorithm. After validating their algorithm by making use of discrete event simulations, Kochut and Karve went on to propose a VM placement strategy which is based on the extent of similarity between images. Their system was able to achieve a significant reduction (upto 80 percent) of data copied to the hypervisor from the storage device. Furthermore, this mechanism is specifically suited for big and extremely used hypervisor clusters. The negative side of their proposed approach is that it may not be suitable for handling large number of similar requests.

Ashraf [15] proposed a Virtual Machine allocation mechanism that is particularly suited for web applications as well as video transcoding. The emphasis was on cost reduction. In order to avoid servers overload, Ashraf proposed an admission control mechanism. In a similar way, the VM consolidation mechanism minimizes the number of underutilized VMs. This publication was still an ongoing research and results were not available. However the foreseeable disadvantage for this technique is the frequency of virtual machine migrations which may impact runtime VM I/O performance.

Zaman et al [16] proposed an auction based mechanism so as to alleviate the challenge of VM allocation. The proposed mechanism extended their earlier work to include dynamic configuration of VM instances and price reservation. In this mechanism, computing resources are seen as "liquid" resources that can be configured and provisioned as different types of VMs. Any interested user/client places a bid for a specific VM bundle. Virtual machine allocations are based on the users' bids and involves a minimum price set to ensure that the cloud service provider is able breakeven. Real workload data was used by the authors to perform simulations and results were compared to the performance of an auction-based static virtual machine provisioning mechanism. The results indicated that dynamic VM provisioning satisfies resource demand and effectively increases revenue than static VM provisioning specifically in high demand cases. However, when demand is low, static VM provisioning performed better than dynamic VM provisioning profit wise. They concluded that by supplementing the two auctions based techniques, a highly effective model can surely be built.

Calcavecchia [17], introduced a new mechanism namely Backward Speculative Placement (BSP). The mechanism avails past demand pattern of a virtual machine to a potential host. The host in turn uses this historical demand behaviour as a basis for relocation or deployment. The method predicts future demand of virtual machines by taking into account the demand relationship between VMs i.e. correlation among virtual machines. VM placement decisions were divided into two stages: (i) continuous deployment, where a placement decision is made immediately after the arrival a request and migrations are forbidden, and (ii) re-optimization, where the existing VM placement is re-optimized by allowing virtual machines to relocate to a different host. Results show that the BSP technique has the ability to place virtual machines in such a way that satisfies demand. Though the BSP technique is unique, it does not consider energy consumption and cost reduction. Also relocating of virtual machines to other hosts is resource intensive and thus may affect runtime I/O performance of VMs.

Halder et al [18] proposed a novel iterative placement algorithm to consolidate virtual machines on available hosts. They proposed a metric, "risk", which specifies the extent to which SLAs are violated. Their goal was to consolidate virtual machines on as few as possible physical hosts and at the same time taking into account the "risk" associated with violating resource boundaries of every application. Their proposed algorithm took into consideration the current placement of VMs and calculated a correlation among an application`s resource needs and used aggregated resource levels. The authors went on to evaluate their algorithm by way of experimenting and results showed that their algorithm was able to place VMs in a way that decreases the number of active hosts. The performance drawback of this technique, as outlined by the authors, was noticed when evaluating a dataset comprising of largely positively correlated virtual machines. The algorithm's performance was reduced by about 50%.

6. ISSUES AND CHALLENGES

Despite the fact that cloud computing is maturing at a faster rate and has also been well received by industry in general, there is still very active research in this domain of cloud computing. Many prevailing issues have not been entirely addressed, while on the other hand fresh problems keep rising from industrial applications. The following section provides a summary of some of the research matters in the cloud computing paradigm.

6.1 Dynamic Resource Provisioning

In cloud computing, resources are acquired and released on-demand. In this case, the goal of a service provider is to provision and de-provision cloud resources to satisfy its Service Level Objectives (SLOs), at the same time lowering the cost of operations. Nevertheless, it's vague how a service provider can be able to attain this goal. In particular, it's a handful task to try and associate Service Level Objectives to tangible computing resources like memory, CPU etc. Mapping SLOs and computing resources is very challenging. Moreover, in order to curtail volatilize nature of demand, an agile response to resource provisioning requests is required.

6.2 Virtual Machine Migration

The migration of virtual machines from one host to another facilitates the agility and robustness of resource provisioning in data centers. Virtual machine migration has its roots in process migration mechanisms [19]. A few years back, VMware and Xen introduced live VM migration techniques that require very less stoppage times less than a second [20]. The work in [21] highlighted that in order to circumvent the complexities associated with process level migration; it is advisable to migrate the entire operating system as a distinct entity together with its applications. The main advantage of VM migration is to counter sudden increase in workload or hotspots. As of today, discovering hotspots and subsequently commencing VM migration is short of the robustness needed to react to rapid changes in workload. In addition, the memory-state has to be relocated efficiently and in a consistent manner, taking into account application resource requirements and servers.

6.3 Server Consolidation

Survey has revealed that data center servers spend much time operating between 10% and 50% of their maximum capacity, and that idle servers consume approximately 50% of their peak power consumption. Server consolidation is a tactic that is used to maximize resource usage/utilization at the same time lessening electricity usage in data centers. In order to maximize resource usage, VMs often need to be carefully packed onto a single host so that no host will be underutilized. The resulting idle hosts can then be switched to an energy-saving state. Live Virtual Machine migration enables the consolidation of VMs. A problem of concern here is how to optimally consolidate servers in a data center so that idle servers are put in sleep mode. This problem is time and again expressed as a modified version the vector bin–packing problem [22]. Nonetheless, application performance should be prioritized over server consolidation. Resource usage of individual VMs (also known as footprint) varies periodically [23]. When there is a sudden change in the footprint of a VM, the contention of shared resources (e.g. memory, bandwidth etc.) among VMs also increases. In case a host is consolidated to its fullness, sudden change in footprint results in extreme contention of resources. This makes it necessary to take a look at variations of virtual machine footprints so as to make use of observed patterns when designing server consolidation algorithms.

6.3 Energy Management

One more significant concern in cloud computing is improving energy efficiency. Estimates indicate that cooling and powering expenses contribute more than half (about 53%) of the total expenses incurred by data centers [24]. As indicated by the Natural Resources Defense Council (NRDC) of United States of America, U.S. server farms in 2013 devoured roughly 91 billion kilowatt-hours of power [25]. By 2020 server farm power use is anticipated to reach to approximately 140 billion kilowatt-hours every year, costing US$13 billion every year in power charges and radiating about 100 million metric tons of carbon contamination every year. Subsequently this calls for cloud vendors to review their electricity usage in order to diminish expenses and lower carbon contamination.

Lowering electricity consumption is not the only thing of paramount importance here, but also to safeguard our environment and adhere to government regulations concerning environmental standards. As a result, combined efforts to minimize carbon emissions by whatever way has received considerable attention world over. As an example, Power-efficient hardware that slows down CPU speeds [26]. More so, studies are being conducted to study power-aware networking protocols as well as infrastructure [27]. Power-efficient data centers have also received no less courtesy. One major concern among all these ways and means is to strike a perfect balance between application performance and electricity savings.

6.4 Security

Security is one more vital research domain in cloud computing. Physical security systems of data centers are normally not accessible to clients. This means clients/users have to depend on infrastructure providers (IaaS) to realize full data security. The clients however have access to configure their own security stings remotely, without knowledge of its full implementation. In this case, the IaaS provider has to ensure security objectives such as: (i) confidentiality, for secure data transfer and access, and (ii) auditability, for proving that application security settings have or have not been tampered with.

7. CONCLUSION AND FUTURE CHALLENGES

Cloud Computing is slowly maturing into a favorable model for provisioning of computing resources in a swift manner. These resources can be allocated and de-allocated with less interaction between consumers and service providers. In other words, resources such as storage, servers, networks services and applications are provisioned with little administration effort. Cloud resources are provisioned to consumers in form of virtual machines (VMs) that are hosted on the cloud providers' infrastructure. However, from the perspective of a cloud provider, virtual machine provisioning and allocation becomes a challenging issue as cloud providers need to control their costs (particularly energy) and at the same time ensure quality of service (QoS) by not violating service level agreements (SLAs). Providing more than enough resources leads to high administrative costs while providing less resources results in poor service delivery. Thus Cloud providers will realize profits by cutting energy costs and also minimize carbon dioxide emissions.

8. REFERENCES

[1] Buyya, R. et al. 2013. Mastering cloud computing. McGraw Hill.

[2] Parkhill, D. 1966. *The challenge of the computer utility*. Addison-Wesley Pub. Co.

[3] NIST Definition of Cloud Computing v15, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[4] Zhang, Q. et al. 2010. Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*. 1, 1 (2010), 7–18.

[5] VMware. Understanding Full Virtualization, Paravirtualization, and Hardware Assist. http://www.vmware.com/files/pdf/VMware_ paravirtualization.pdf.

[6] Docker: The Linux Container Engine 2015. http://www.docker.io. Accessed: 2015- 11- 30.

[7] Li, W. 2014. Algorithms and Systems for Virtual Machine Scheduling in Cloud Infrastructures. Umea University Sweden.

[8] Mell, P and Grance, T. The NIST Definition of Cloud Computing. National Institute of Standards and Technology (NIST), 2011.

[9] Zhang, Z. et al. 2015. Large-scale virtual machines provisioning in clouds: challenges and approaches. Frontiers of Computer Science. (2015).

[10] N. Lo, P. Fan, and T. Wu, "An Efficient Virtual Machine Provisioning Mechanism for Cloud Data Center," pp. 703–706, 2014.

[11] De, P., Gupta, M., Soni, M. and Thatte, A. 2012. Caching VM instances for fast VM provisioning: A comparative evaluation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 7484 LNCS, (2012), 325–336.

[12] Das, A.K., Adhikary, T., Razzaque, M.A. and Hong, C.S. 2013. An intelligent approach for virtual machine and QoS provisioning in cloud computing. *The International Conference on Information Networking 2013 (ICOIN)*. (2013), 462–467.

[13] Nejad, M.M., Mashayekhy, L. and Grosu, D. 2015. Truthful Greedy Mechanisms for Dynamic Virtual Machine

Provisioning and Allocation in Clouds. *IEEE Transactions on Parallel and Distributed Systems*. 26, 2 (2015), 594–603.

[14] Kochut, A. and Karve, A. 2012. Leveraging local image redundancy for efficient virtual machine provisioning. *Proceedings of the 2012 IEEE Network Operations and Management Symposium, NOMS 2012*. (2012), 179–187.

[15] Ashraf, A. 2013. Cost-efficient virtual machine provisioning for multi-tier web applications and video transcoding. *Proceedings - 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013*. (2013), 66–69.

[16] Zaman, S. and Grosu, D. 2013. A Combinatorial Auction-Based Mechanism for Dynamic VM Provisioning and Allocation in Clouds. IEEE Transactions on Cloud Computing. 1, 2 (2013), 129-141.

[17] Calcavecchia, N. et al. 2012. VM Placement Strategies for Cloud Scenarios. 2012 IEEE Fifth International Conference on Cloud Computing. (2012).

[18] Halder, K., Bellur, U. and Kulkarni, P. 2012. Risk Aware Provisioning and Resource Aggregation Based Consolidation of Virtual Machines. *2012 IEEE Fifth International Conference on Cloud Computing* (2012), 598–605.

[19] Osman S, Subhraveti D et al (2002). The design and implementation of zap: a system for migrating computing environments. In: Proceedings of OSDI.

[20] Elastic Compute Cloud (EC2) Cloud Server & Hosting AWS: 2015. https://aws.amazon.com/ec2/. Accessed: 2015- 11- 30.

[21] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, and Andrew Warfield. 2005. Live migration of virtual machines. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation - Volume 2*(NSDI'05), Vol. 2. USENIX Association, Berkeley, CA, USA, 273-286.

[22] Chandra Chekuri and Sanjeev Khanna. 1999. On multi-dimensional packing problems. In *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms* (SODA '99). Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 185-194.

[23] Wood, T., Shenoy, P., Venkataramani, A. and Yousif, M. 2007. Black-box and gray-box strategies for virtual machine migration. *NSDI'07 Proceedings of the 4th USENIX conference on Networked systems design & implementation*. (2007), 229–242.

[24] Hamilton, J. 2009. Cooperative Expendable Micro-Slice Servers (CEMS): Low Cost, Low Power Servers for Internet-Scale Services. 4th Biennial Conference on Innovative Data Systems Research (CIDR)*Power* (2009), January 4-7, 2009, Asilomar, California, USA. 1-8.

[25] Energy Efficiency, Data Centers | NRDC: 2015. http://www.nrdc.org/energy/data-center-efficiency-assessment.asp. Accessed: 2015- 11- 30.

[26] Shekhar Srikantaiah, Aman Kansal, and Feng Zhao. 2008. Energy aware consolidation for cloud computing. In *Proceedings of the 2008 conference on Power aware computing and systems* (HotPower'08). USENIX Association, Berkeley, CA, USA, 10-10.

[27] IEEE P802.3az Energy Efficient Ethernet Task Force: 2015. http://www.ieee802.org/3/az/. Accessed: 2015- 11- 30.