Server Consolidation Schemes in Cloud Computing Environment: A Review

Aidan Ghojoghi and Mirsaeid Hosseini Shirvani

***Abstract*— Cloud computing is a new paradigm which deliver computing resources as utility. Datacenters as cloud infrastructure encounter with several issues such as power management for the sake of economic viewpoint. Researches show that high rate of power wastage in large scale datacenters is related to sprawl resource usage making low utilization. More recently, server consolidation techniques have been developed to maximize resource utilization in at least used number of physical servers. This technique is applied in virtualization environment which allows physical servers to host several operating systems and related applications. Server consolidation approach abstracts system under study into NP-hard bin-packing problem. Several works have been done in literature to solve server consolidation problem. This paper analyses the papers and compare them with parameters derived from research context. Commonalities and differences are argued. Then, open issues and challenges are concluded to work in future.**

***Index Terms*— Cloud Computing, Datacenters, Server Consolidation, VM Consolidation, Virtualization**

I. INTRODUCTION

Cloud computing is a new paradigm which deliver computing resources as utility [1-2]. Datacenters as cloud infrastructure encounter with several issues such as power management for the sake of economic viewpoint. Cloud providers try to find a way to reduce their overall costs. The total cost of ownership (TCO) includes fixed costs, capital expenditure, and variable costs, operational expenditures. Lager amount of operational expenditure is related to server sprawl [31]. Server sprawl is a phenomenon which resources are dispersed through systems with low utilization making high rate of power consumption. More recently, server consolidation techniques have been developed to maximize resource utilization in at least used number of physical servers. This technique is applied in virtualization environment which allows physical servers to host several operating systems and related applications. Server consolidation approach abstracts system under study into NP-hard bin-packing problem [3,18]. Several works have been done in literature to solve server consolidation problem [4-6]. Server consolidation leverages virtual machine migration to pack workload images on minimum number of used physical servers. So, the schemes used should be aware of user QoS requests to reduce SLA violation rates during consolidation schemes. Several papers paid attention to server consolidation technique as promising approach toward power management. This paper analyses the papers and compare them with parameters derived from research context. Commonalities and differences are argued. Then, open issues and challenges are concluded to work in future. The rest of the paper organized as follows. Section II is dedicated to literature review including sub section cloud and consolidation definitions. A study on several consolidation works is brought in section III. A comparison between schemes is placed in section IV. Section V concludes the current paper in terms of open issues and challenges for further research in future.

II. LITERATURE REVIEW

A. Cloud Computing

NIST definition of cloud computing states that cloud computing is "A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [7]. It seems to be agreed by researchers and for our work NIST definitions suffices. Moreover, the definition has noticeable terms and features in which tabulated in table1. It facilitates service delivery via internet including dynamically service provisioning on-demand, scalable, virtualized resources as a service and pay-per-use model with less management intervention without any interaction with providers.

TABLE I: CLOUD COMPUTING FEATURE ACCORDING TO NIST [7-8]

Feature	Description and characteristics
Essential characteristics	(i) On-demand self-service; (ii) broad network access; (iii) resource pooling; (iv) rapid elasticity; (v) measured service
Service models(SPI model)	(i) Infrastructure as a service (IaaS); (ii) Platform as a service (PaaS); (iii) Software as a service (SaaS)
Hosting Model	(i) External; (ii) Internal
Deployment models	(i) Private cloud; (ii) Community Cloud; (iii) Public Cloud; (iv) Hybrid Cloud
Roles	(i) Cloud Auditor; (ii) Cloud Service Provider; (iii) Cloud service carrier; (iv) Cloud Service Broker; (v) Cloud Service Consumer

Manuscript published September 30, 2016

Mirsaeid Hosseini is faculty member of IAU (Sari Branch) as Corresponding Author: (e-mail: mirsaeid_hosseini@iausari.ac.ir)

Aidan Ghojoghi is from Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran (e-mail: aaa.ghojoghi@yahoo.com

Everything is as a service in cloud, XaaS, where X is software, hardware, platform, infrastructure, data, business, etc. [9]. But all of them is in the form of SPI model as depicted in figure1. In software as service model, a pre-made application such as e-mail can be provisioned to user. In PaaS, user can installs operating systems and developing tools to develop cloud based software, as such he/she can purchase his/her products such as Google query language (GQL), Microsoft Azure and etc. The IaaS model provides virtual machines, network, hardware and user can install operating systems and deploy his applications but does not control over infrastructure such as Amazon EC2 instances [10].

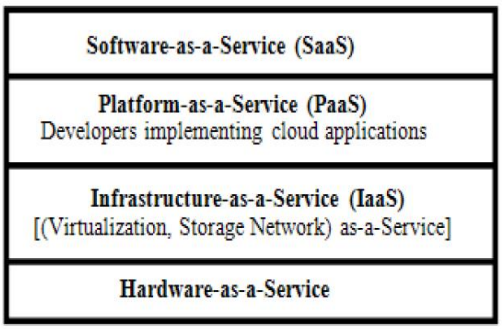

Fig. 1. Cloud computing SPI model

Cloud service deployments are typically available depending on requirements via public, private, community and hybrid [10]. Public clouds are provisioned via internet which are owned and operated by a provider. Public clouds are available for general public such as social networks, storage and email services etc. Also, organizations can publish their web services over public clouds. Private cloud, on the other hand, is owned and operated by specific organization. Infrastructure of cloud can be operated both on-premises by organization and third party. In community cloud, infrastructure of cloud is shared between numbers of organizations which have common interests. It can reduce capital costs. Also, infrastructure control can be both by organizations and third party. Hybrid cloud contains combination of number of different clouds making ability to move data and services between clouds. It commonly helpful to deal with server spike [7, 10]. Organizations and enterprises lower their costs by getting services from cloud providers such as Amazon EC2, Azure, IBM, Salesforce, etc. by pay-per-use model.

B. Server Consolidation

Server consolidation is an approach to the efficient usage of (physical) servers in order to reduce the total number of servers that an organization requires. The practice developed in response to the above-described server sprawl, a situation in which multiple underutilized servers take up more space and consume more energy than can be justified by their workload [31]. Server virtualization provides technical means to consolidate multiple servers leading to increased utilization of physical servers [11]. A significant chance for power optimization will be presented by consolidation of applications in cloud computing environment. There are significant inter-relationships between resource utilization, performance of consolidated workloads and power consumption. Clusters in datacenters in idle or low utilization status consume large amount of electricity as energy. For example, the energy consumption of non-operative, but turned on, accounts for approximately 70% of the full loaded server energy consumption [12]. Virtualization technique is widely used in cloud datacenter allowing different virtual machines co-host on same physical machine; it also applies consolidation approach to pack virtual machines over minimum number of physical machines [4]. Qos-aware schemes is very important in this ambit because awkward consolidation algorithms neglecting migration cost and applications affinity on special resources may nullify the benefit of consolidation yielding high SLA violation rate. To deploy the requested jobs in cloud environment, the user makes a request to a resource broker, specifying the number of processing units required and the associated memory requirements. If the requested CPU and memory resources are available, the job is accepted. This static strategy ensures that all jobs accepted into the cluster will have sufficient processing units and memory to complete their work. Nevertheless, it can lead to a waste of resources, as many workloads proceed in phases, not all of which use all of the allocated processing units at all times. Consolidation technique dynamically declines number of active servers by releasing unnecessary machines in the current computing phase [32]. Fig. 2. Illustrates consolidation scheme.

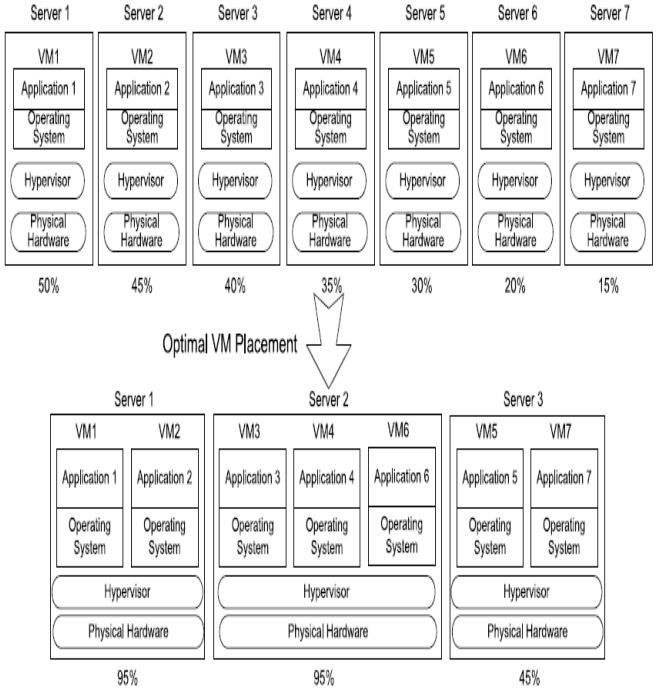

Fig. 2. VMs Consolidation Scheme [13].

After server consolidation by VMs migration, servers numbered 4 through 7 will be set into lower power consumption state or hibernate mode to save energy [13,31]. Server consolidation is tantalizing approach, because effective consolidation is not as trivial as packing the maximum workload in the smaller number of servers with keeping each resource (CPU, Disk, Network, etc.) on every server near 100% utilization. In practice, at least two concerns makes consolidation problem become very hard decision-making problem. Firstly, consolidation methods must carefully decide which workload should be combined on the

same physical server. In fact, understanding the nature of workload and their affinity to resources are crucial. Secondly, the migration cost in terms of additional resource usage and migration time may lead performance degradation and prolongs workload execution. Therefore, the consolidation problem is more complex than bin packing problem. In literature, some researchers abstracts server consolidation problem as multi-dimensional bin packing problem where servers are bins with each resource (CPU, disk, network, etc.) being one dimension of the bin. The bin size along each dimension is given by the energy optimal utilization level. Each hosted application with known resource utilizations can be treated as an object with given size in each dimension. Minimizing the number of bins should minimize the idle power wastage [33]. However, that is not true in general, according to aforementioned reasons causing the energy aware consolidation problem to differ from traditional vector bin packing.

III. A Study over Server Consolidation Schemes in Literature

In [4], a VM consolidation approach has been applied on OpenStack, an open-source platform for Cloud computing to show effectiveness of VM consolidation on power consumption reduction. The proposed approach makes decision to map VMs into the minimal number of physical servers to reduce the runtime power consumption. On the other hand, server consolidation technique is severely resource intensive which may cause service degradation. So, proposed algorithm considers CPU, network and other resource features to avoid performance degradation and SLA violation. Its experimental results show the effectiveness [34].

Author in [5] have defined new concept of min, max and shares driven from virtualization technology like Xen to specify at least and at most amount of resources that can be allocated to each VM respectively and Shares provides advice to hypervisor to distribute spare resources among racing VMs. Setting a min for a VM ensures that it receives at least that amount of resources when powered on and setting a max for a low priority application ensures that it does not use more resources, thus keeping them available for high-priority applications. Shares provide advice to the virtualization scheduler on how to distribute resources between contending VMs. Author have applied series of techniques for placement and power consolidation of VMs in data centers by taking consideration the fact that different application have different priority and different affinity to resources. It provides power-performance compromises in modern datacenter by applying fine-grained approach to well mange ensuring that high priority applications get the most resources. In this way, the algorithm keeps data center in appropriate utilization level and consequently manages power consumption without making performance degradation.

An intelligent dynamic resource management framework has been proposed based on the IaaS cloud environment in [31]. As illustrated in Fig. 3 the framework can be applied for system and resource management monitoring.

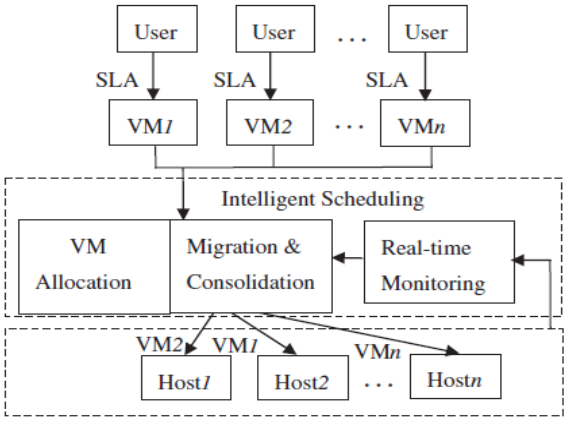

Fig. 3. Servers Utilization Monitoring Framework [31].

The framework dispatches VMs as quickly as possible to do VM placement right after user applies VM resources to cloud and determines SLA. By doing so it ensures response time and then by taking benefit of real-time monitoring it consolidates based on host's running condition. In this paper, framework performs dynamic consolidation with heuristic approaches based on two conditions (i) if hosts cannot comply with SLA then by reducing upper level of threshold value of processor utilization, some VMs will be transferred to other hosts for the sake of offloading and assuring that SLA would be met. (ii) for the hosts in which their utilization are below the lower threshold , all of their VMs will be transferred to another hosts and switching the hosts in hibernating mode to save energy. Moreover, the framework does dynamic adjustment process in certain time interval and tries not to have frequent VM migration making performance degradation.

The proposed algorithm in [6], iMeter, uses prediction module to obstruct higher power consumption for the sake of unplanned resource allocation datacenters encountering with unprecedented instantaneous burst workloads. Method in [6] leverages kernel-based performance counter to profile VM power meter. It considers six types of VM performance counter such as CPU, memory, physical disk, cache, process and network interfaces. Also, for power meter it uses program interface in which was embedded in their firmware by manufacturers for administrators to query the power usage at editable intervals. The time interval must not be very short and very long., because both of them have their own drawbacks. In this paper the best time interval has been adjusted with 2 seconds by applying some experiments. The paper's framework profiles kernel-counter to finds out relationships between VMs resources need and power consumption. To do so, it exploits super vector regression (SVR) to estimate the relationship between selected performance counters and the power consumption of a VM and then based on this prediction it consolidates VMs to appropriate minimum number of servers to rise server utilization and sets others in low power status like power off and the power consumption significantly levels off. Experimental results show that its prediction accuracy totally has 5% error at most.

A power-aware application placement has been presented in [14]. It considers power and migration costs along with placing the application VMs to physical machines. The

proposed framework was named Pmapper consisting three components (i) Performance manger (ii) Power Manager and (iii) Migration manager. The performance manager globally keeps QoS in the negotiated SLA. The power manager applies monitoring engine to surveillant current power consumption, it exploits CPU throttling by adjusting CPU clock rate whenever needed. Migration manger launches VMs consolidation as abstracted to bin packing problem with variable bin sizes based on performance manager and power manger. Moreover, it takes VMs cost into account to make decision in case of migration.

Benjamin et al. in [11] have studied a mathematical programming approaches for server consolidation problems in virtualized large scale data centers. They have shown aforementioned problem is strongly NP-Hard and assumed it is multidimensional bin packing problem [29]. The authors formulated problem as static server allocation problem in large datacenters and presented a heuristic method called SSAPv, static server allocation problem with variable workload. They considered two widespread type of services being used by users in cloud environment (i) W/A/B services, namely, web, application and database services respectively and (ii) ERP services to show the working of their proposed method.

A network-aware VM consolidation algorithm called inter-and-intra datacenter VM-Placement has been proposed by burak et al. in [15] to aim energy savings in large data centers in which include multiple medium size datacenters geographically distributed and connected via backbone network. They have formulated VMs consolidation by new mixed integer linear programming (MILP) to place VMs to appropriate hosts by considering users requirement expressed in SLA and CPU, memory, network bandwidth, delay and the distance of users to be hosted the selected datacenter. Their results show the improvements in terms of power consumption, resource utilization and fairness for many connected medium size datacenters.

Authors in [16] have propounded server consolidation algorithm in which their objectives are application performance and system utilization. It leverages virtualization to run away from server sprawl in case of increasing server equipment in demanding IT service rising in which server sprawl makes low server utilization and high system management cost. Current paper uses key performance metrics in their monitoring framework that triggers to consolidate VMs into minimum number of physical hosts. Also, it considers migration cost not to have negative affection on user application performance such as SLA violation.

A heuristic based server consolidation algorithm named RFAware has been implemented in [17]. Despite other researches in server consolidation field in which have focused only on reducing the number of active physical servers (PMs) using Virtual Machine (VM) Live Migration to reduce the number of underutilized servers, the author proposed an algorithm along with reducing the number of active PMs considers a consolidation approach to reduce residual resource fragmentation, the residual resources can be efficiently used for newVM allocations, or VM reallocations, and some future migrations can also be reduced. Datacenters were suffered from VM sprawl whereas each proportion VM's demand in resource differs from other VM's demand. This paper by applying heuristic algorithm comply with objectives such as Reducing Residual Resource Fragmentation, Minimizing the number of live migrations, Reducing SLA violations and Ensuring Scalability of the consolidation approach as well. Its algorithm rigorously increases power savings with minimize the number and amount of underutilized resources.

A linear program and heuristic algorithm has been presented in [30] that controls the number of migration and reduces server energy consumption. It is cost-aware and migration controller algorithm because precludes unnecessary migrations due to unpredictable workloads that require VM resizing. The algorithm prioritizes VMs with steady capacity to obstruct residue migrations. The authors have formulated it as linear programming problem. Their heuristics a little improves the famous first-fit decreasing (FFD), best-fit decreasing (BFD), worst-fit decreasing (WFD), and almost worst-fit decreasing (AWFD) algorithms [18].

IV. A Comparison between Variety Server Consolidation Schemes

Fig. 4 illustrates our subjective comparison framework on published papers derived from literature.

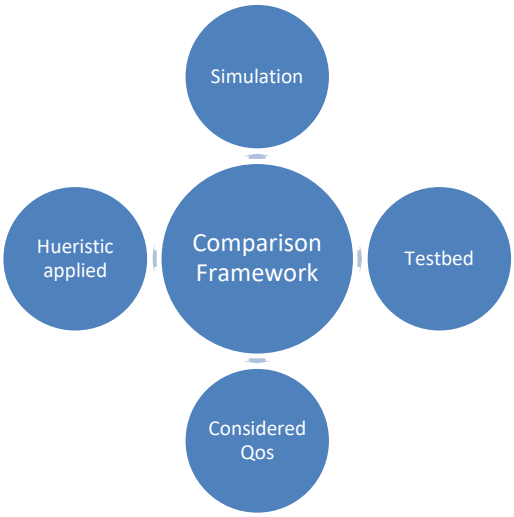

Fig. 4. Our subjective Comparison Framework

Different researches using server consolidation technique that have been segregated in terms of simulation environment, Testbed/Validation, quality of service metrics and technique used in literature. Moreover, greedy algorithm is representative to not mentioned technique by author(s). Comparison of them are tabulated in Table 2.

TABLE II: COMPARISON PROPOSED VARIETY SERVER CONSOLIDATION APPROACHES

Proposed Scheme	Simulator / Programming Language/ Benchmark	Testbed/Validation	Considered QoS metrics	Techniques
VM consolidation: A real case based on OpenStack Cloud[4]	The last OpenStack release[26] called Diablo[ref8 of VM-consol] and Apache	testbed made by physical servers with CPU Intel Core Duo E7600 @ 3.06 GHz, 4 GB RAM, and 250 GB HD, and used	power consumption, host resources, and networking optimization	Greedy

	HTTP Server Benchmarking Tool (ab) [94]	KVM as hypervisor		
Shares and Utilities based Power Consolidation in Virtualized Server Environments[5]	RedHat Enterprise Linux(RHEL] using HPCC benchmark [5]	The testbed consists of three VMware ESX 3.5 servers with 3 GHz, 4.8 (2x2.4) GHz and 6 (2x3.0) GHz total capacity	Power Cost and Application Utility(Amount of SLA violation)	Greedy algorithm based Multi bin packing constraints
Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers[31]	CloudSim[19-21]	configured with 1000 hosts, including 500 ProLiant DL360 G4p hosts (configured with 3400 MHz _ dual-core, 6 GB memory, 1 GB network bandwidth) and ProLiant ML110 G3 hosts (configured with 3000 MHz * dual-core, 4 GB memory, 1 GB network bandwidth).	Resource utilization, response time and power consumption	Hybrid shuffled frog leaping heuristic based
iMeter: An integrated VM power model based on performance profiling[6]	Benchmarks such as NASA NPB, IOzone [22] and Cachebench[23]	It applies two platforms Dell and Hp: One is a Dell PowerEdge M610 blade server with two Intel Xeon E5620 quad core with 24 GB RAM and 2 SAS 250 GB hard. The other platform is an HP BL680G7 blade server with two Intel Xeon E7530 six core with 32GB RAM and 2 SAS 500 GB hard drives	CPU, Memory and I/O Utilizations, VM performance and VM power usage	Support Vector Regression
pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems[14]	Trade6 application as well as the two HPC applications daxpyand fma[24-25]	First setup is IBM HS-21 Bladecenter with mutiple blades Each blade has 2 Xeon5148 dual-core Processors and runs two compute-intensive applications from an HPC suite, namely daxpyand fma and a Linpack benchmark HPL [27] on VMWare ESX Hypervisor and the second setup consists of 9 IBM x3650 rack servers running VMWare ESX with L2 cache of 4MB. Each server has a quad-core Xeon5160 processor running at 2.99 GHz	Power Consumption and Performance	Bin Packing FFD[28]
A Mathematical Programming Approach for Server Consolidation Problems in Virtualized Data Centers[11]	Open Source Solver: LP-Solve	workstations running Windows XP Professional (Service Pack 2) on anAMDAthlon XP 2;100 Model 8 with 2.1 GHz	Response time, Service Time, Server Utilization and Energy Cost	Linear Programming with FF and FFD heuristic
Inter-and-Intra Data Center VM-Placement for Energy-Efficient Large-Scale Cloud Systems[15]	OpenStack [26]	A datacenter to house 1120 physical hosts where at most 5 VMs run on each physical host	Power Consumption, Resource Utilization and Fairness	MILP (Mix Integer Linear Programming)
Application Performance Management in Virtualized Server Environments[16]	*Websphere Workload Simulator* (WSWS)	IBM BladeCenter environment with blades as our physical machines. VMWare ESX server (hypervisor) is deployed on three bare HS-20 Blades (Intel architecture)	Response time, throughput, server utilization	Heuristic PTAS , polynomial time approximate solutions
Heuristics based server consolidation with residual resource defragmentation in cloud data centers[17]	CloudSim[20]	800 simulated PM with their capacity and determined simulation time and consolidation time interval	Resource Utilization(CPU, Memory, Bandwidth), performance and power consumption	Heuristic
Server consolidation with migration	Python	Intel Core 2 Duo processor with	Response time, Throughput,	Linear Programming and

control for virtualized data centers[30]	2.4 GHz and 4 GB of memory	utilizatio n	Heuristic s

V. CONCLUSION

Power usage is the first class concern in power hungry datacenters regard to economic viewpoint. The main reason causes power wastage is related to low resource utilization. So, to deal with the problem several works have been proposed with consolidation approach. After studying over published paper, we have presented our comparison framework. Then we reviewed papers and finally compared based on our subjective comparison framework. Server consolidation problem is abstracted to well-known NP-Hard bin-packing problem which items must be packed into minimum number of bins. Therefore, miscellaneous combinatorial algorithms have been propounded to figure out the aforementioned problem. One of the biggest challenges is related to not pay attention user QoS in this environment because the schemes which executes combinatorial algorithm take long time making high rate of user SLA violation. On the other hand, big portion of research work on CPU utilization as only resources which nullify server consolidation schemes in real conditions especially when there is not any meaningful relation between CPU usage and other resources. Future direction can be toward solving open issues such as considering resource vector utilization instead of considering limited number of resources along with development of new combinatorial algorithm to make fast decision in cloud fragile environment.

REFERENCES

[1] Armbrust M, Fox A, Griffith R, D. Joseph A and Katz R, ''Above the Clouds: A Berkeley View of Cloud Computing''. Technical report EECS-2009-28, UC Berkeley, 2009.
[2] Y. Jararweh, M. Jarrah, M. kharbutli,Z. Alshara, M. N. Alsaleh, M. Al-Ayyoub: CloudExp: A comprehensive cloud computing experimental framework, Simulation Modelling Practice and Theory 49 (2014) 180–192.
[3] J. Békési, G. Galambos, H. Kellerer, A 5/4 linear time bin packing algorithm, J. Comput. System Sci. 60 (1) (2000) 145–160.
[4] A. Corradi, M. Fanelli, L. Foschini, VM consolidation: A real case based on OpenStack Cloud, Future Generation Computer systems, 32(2014) 118-127.
[5] M. Cardosa, M. Korupolu, A. Singh, Shares and utilities based power consolidation in virtualized server environments, in: Proceedings of IFIP/IEEEIntegrated Network Management (IM'09), 2009, pp. 327–334.
[6] H. Yang, Q. Zhao, Z. Luan, D. Qian, iMeter: An integrated VM power model based on performance profiling, Future Generation Computer Systems, In Press.
[7] P. Mell, T. Grance, The NIST definition of cloud computing, Natl. Inst. Stand. Technol. 53 (6) (2009) 50.
[8] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, D. Leaf, NIST Cloud Computing Reference Architecture NIST Special Publication 500-292, 2011.
[9] B.P. Rimal, E. Choi, A Toxonomy and Survey of Cloud Computing Systems,Fifth International Joint Conference on INC, IMS and IDC (2009).
[10] Cloud Computing: Paradigms and Technologies (springerlink)
[11] B. Speitkamp, M. Bichler, A mathematical programming approach for server consolidation problems in virtualized data centers, IEEE Trans. Services Comput. (2010) 266–278.
[12] Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., & Jiang, G. (2009). Power and performance management of virtualized computing environments via lookahead control. Cluster Computing, 12(1), 1–15. 975.
[13] Y. Gao, H. Guan, Z. Qi, Y. Hou, L. Liu, A multi-objective ant colony system algorithm for virtual machine placement in cloud computing, Journal of Computer and System Sciences, In press.
[14] A. Verma, P. Ahuja, A. Neogi, pMapper: power and migration cost aware application placement in virtualized systems, in: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, 2008, pp. 243–264.
[15] B. Kartarciet. Al., Inter-and-Intra Data Center VM-Placement for Energy-Efficient Large-Scale Cloud Systems, First International workshop on Management and Security technologies for Cloud Computing 2012.
[16] G.Khanna, K. Beaty, G. Kar, A. Kochut, Application Performance Management in Virtualized Server Environments, Network Operations and Management Symposium , 2006. NOMS (2006).
[17] K. S. Rao, P. S. Thilagam ,Heuristics based server consolidation with residual resource defragmentation in cloud data centers, Future Generation Computer Systems. In Press.
[18] L.T. Kou, G. Markowsky, Multidimensional bin packing algorithms, IBM Journal of Research and Development 21 (5) 1977.
[19] R. Calheiros, R. Ranjan, C. De Rose, R. Buyya, Cloudsim: a novel framework for modeling and simulation of cloud computing infrastructures and services. arXiv:0903.2525.
[20] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A.F. De Rose, R. Buyya, Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms, Software: Practice and Experience 41 (1) (2011) 23–50.
[21] Buyya, R., Ranjan, R., & Calheiros, R. N. (2009). Modeling and simulation of scalable cloud computing environments and the CloudSim Toolkit: Challenges and opportunities. In Proceedings of the seventh high performance computing and simulation conference (HPCS 2009, ISBN: 978-1-4244-49071), Leipzig, Germany (pp. 21–24). New York, USA: IEEE Press.
[22] W. Norcott, D. Capps, IOzone filesystem benchmark. http://www.iozone.org, 2003.
[23] P. Mucci, K. London, J. Thurman, The Cachebench Report, University of Tennessee, Knoxville, TN, 1998,
[24] IBM Active Energy Manager, http://www-03.ibm.com/systems/management/director/extensions/actengmrg.html
[25] HPL-A Portable Implementation of the High Performance Linpack Benchmark for Distributed Memory Computers, http://www.netlib.org/benchmark/hpl/
[26] OpenStack Cloud: http://www.openstack.org/, 2011.
[27] N.Kim, J.Cho and E.seo, Energy-credit scheduler: An energy-aware virtual machine scheduler for cloud Systems, Future Generation Computing Sysytems. In Press.
[28] Murray G Patterson. What is energy efficiency?: Concepts, indicators and methodological issues. Energy Policy, 24(5):377 – 390, 1996. ISSN 0301-4215. doi: 10.1016/0301-4215(96)00017-1.
[29] M. Garey and R. Graham, "Resource Constrained Scheduling as Generalized Bin Packing," J. Combinatorial Theory, Series A, vol. 21,pp. 257-298, Nov. 1976.
[30] T. C. Ferreto, M.A.S. Netto, R. N. Calheiros, C.A.F. De Rose, Server consolidation with migration control for virtualized data centers, Future Generation Computer Systems 27 (2011) 1027–1034.
[31] L. Jian-ping, X.Li, C. Min-rong, Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers, Expert Systems With Applications, in press.
[32] Power Efficient VM Consolidation using Live migration-A step towards Green computing: A white paper.
[33] A. Blaglazov et al., A taxonomy and Survey of Energy-Efficient Data Centers and Cloud computing Systems. White paper.
[34] S. Cash et al., Managed infrastructure with IBM Cloud OpenStack Service, **Published in:** IBM Journal of Research and Development (Volume: 60, Issue: 2-3, March-May 2016)

Authors Biography

Aidan Ghojoghi was born on july 9 1985. She received her B.S. and M.Sc. in computer software engineering from Mirdamad institute of Gorgan at 2013 and IAU (Sari-Branch) at 2016 respectively. Her research interests are in the areas of cloud computing, image processing and artificial intelligence

Mirsaeid Hosseini Shirvani received his B.S. and M.Sc. in computer software engineering from IRAN University, Tehran. Currently, he is candidate of Ph.d in computer engineering at university of science and research, Tehran, IRAN. Also, he is faculty member of IAU (Sari- Branch, IRAN) in computer engineering department. His research interests are in the areas of distributed systems, parallel processing and evolutionary computing.