2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

# Modeling of Resource Granularity and Utilization with Virtual Machine Splitting

Hojjat Baghban[1], Jerry Chou[1], Ching-Hsien Hsu[2], Yeh-Ching Chung[1]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[2]Department of Computer Science and Information engineering, Chung Hua University, Hsinchu, Taiwan
hojjat.baghban@sslab.cs.nthu.edu.tw, jchou@lsalab.cs.nthu.edu.tw, chh@chu.edu.tw, ychung@cs.nthu.edu.tw

***Abstract*—The increasing trend in IT users and their needs for computational power in cloud data centers leads to noticeable growth in physical servers. It is a challenging issue which causes the dramatic burden of power consumption and the number of Physical machines. Virtualization is remarkable method for reducing the number of physical servers with appropriate processing performance and utilization. But, it is worth saying that the fulfilling the resource utilization is still one of the significant challenging issue, especially in in data centers environment. Actually, there are some applications situated on a large single virtual machine. One way to guarantee the reasonable physical server utilization is to let the application to be split and hosted on smaller virtual machines with the sufficient computational power. Although exploiting multiple small virtual machine instead of one large virtual machine benefits appropriate physical resources utilization and reducing the number of turn on physical machine, it is sustained penalty in terms of demanding extra resources due to map the applications on new virtual machines. However, existing research have not clarified precisely the reason in terms of that the data center is sustained extra resources and computational power overhead due to splitting the original application and exploiting more smaller virtual machines provided to preserve the criteria of the original application on the large virtual machine. This paper demonstrates through mathematical modelling that the physical resource providers, which are situated in cloud data center, endure the penalty in terms of extra physical resources. The mentioned mathematical modeling in this paper will be noticeable in cloud data center energy efficiency and physical resource utilization performance.**

***Keywords—Virtual Machine Splitting; Resource Granularity; Cloud Computing***

## I. INTRODUCTION

Cloud computing is a large-scale computing system, which consists of a collection of heterogeneous networked and virtualized resources. Computing virtualized resources such as virtual machines (VMs), storage, applications, and servers can be transferred on demand. Virtualization [8, 3] is a technology that has a remarkable benefit in the computational systems issues such as system reliability, resource management, and resource utilization improvement.

Datacenter is a critical computing resource in controlled environment and under centralized management [16]. These computing resources include mainframes, physical servers, web servers, applications, files and print servers, storage devices, computational components, software and operating systems. Mid and large organizations usually have one or more many datacenters, according to its organizational and computational needs. So, it would result in complex and distributed environment that expects an accurate management and maintenance which are demanding a remarkable budget. Datacenters should be capable to sustain rapid growth in performance and the number of devices which can be located in it. Also, it is feasible to increase the number of devices such as physical servers, routers, ports, switches, etc. So, there should be a flexibility and scalability in such a dynamic environment. Datacenters are growing according to computational demand to fulfill a safe and scalable environment for sensitive and computation-intensive applications.

According to the predetermined schedule, all of the presented services should be operated without interruption. In other words, datacenters present a safe computational environment for storage, information communication, computational services and accurate data processing [16, 10, and 14]. But, recently, the challenging issues which are situated in Datacenter is the energy efficiency and Physical server (PM) utilization. Researchers and related manufactures are being involved in enhancing the cost of energy and physical machine (PM) utilization which should be helpful in cloud economy. But, there are still open problems for this matter.

Virtualization is a well-known method which is used in several datacenters for reducing the physical server with appropriate processing performance and utilization. Historically, virtualization developed during 1960s [13, 9].The benefits of this method is concerned with hardware independencies, isolation, security [11] and consolidation. In other words, through virtualization fewer physical server, space, and power would be needed. Virtualization caused an improvement in system efficiency through increasing hardware and software utilization [15]. It is worth saying that energy efficiency is not just completely fulfilled via virtualization.

VM is considered as a software implementation of Physical Machine (PM) that executes the web server application like PMs [7]. Through VMs, multiple isolated operating systems can be run on the same PM. They provide an Instruction Set Architecture (ISA) which can be different from the real machine. But, it is worth saying that when several VMs are running simultaneously on the same physical machines, the

978-1-5090-4065-0/16 $31.00 © 2016 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.134

performance would be unstable due to concurrent access to resources.

Also, the increasing trend in number of running high load application on VM causes the remarkable workload on the physical server. Although this issue fulfills the server consolidation which is a solution for energy consumption, it should be mention that it would lead to PM overload problem and the potential bottleneck in one physical server due to numerous webserver application requests. So, a precise tradeoff between these matters[1] is expected.

There are different types of virtualization that can be exploited in the datacenter. Server virtualization [1] and storage virtualization are two important technologies that are widely used. Server virtualization brings cost-effectiveness in terms of having fewer physical servers, where VMs can share the same hardware. But, datacenter energy efficiency and server utilization cannot be completely resolved through virtualization. So, a comprehensive datacenter management approach is needed.

Virtualization is considered as a foundation of cloud computing. This technology allows the creation of intelligent abstraction layer which hides the underlying hardware and software complexities. For instance, through server virtualization, heterogeneous operating systems can share the same hardware components. And, the operating systems are enabled to be moved between different PMs without interrupting the running application. Also, the same is valid for storage virtualization. Virtualization makes cloud computing cost effective because it simplifies the delivery of services through a platform to optimize the resources in a scalable manner.

Performing a large application on a single virtual machine requires a noticeable amounts of physical resources to be allocated to the original virtual machine (OVM) with situated on a single PM. On the other hand, although it can preserve server consolidation criteria, due to potential overload it is expected to benefit the VM splitting techniques. So instead of allocating a large application on an OVM, by considering the type of original application, splitting and map into smaller virtual machines (SVMs) with appropriate free computational power of the PMs. It is not reasonable to considering VM splitting model for all of the applications. But, it is expected to benefit different model by considering the application's type and criteria which may consistent with one or more VM splitting model.

It should be mentioned that despite considering any VM splitting model, the significant point is concern with computing the total resource requirement of VMs after VM splitting by considering the arrival request to the system. So, this paper is trying to define and analyze two VM splitting model and physical resource penalty, due to incurring extra total latency, to preserve the response time of arrival request at least equal running them on OVM provided that fulfilling the PMs utilization.

[1] server consolidation & PM overload problem

## II. RELATED WORK

Various research have presented in terms of virtualization and PM utilization which has a particular features and capabilities. But few of them concentrated VM splitting techniques and potential models. In [6] the relationship between the amounts of required resources for the target virtual machine and the appropriate application performance is investigated. The authors proposed an online model estimator which is located in a feedback control system for resource allocation among the virtualized environment.

Also, in [4], the authors investigated the application splitting among smaller VMs with goal of preserving the response performance of the application run on virtual machine before splitting. This research did not mention splitting issue from VMs points of view to decrease the potential overhead and bottleneck on the OVM which situated on a single physical server.

In [5] the authors proposed an approach which is called VMSA for resource virtualization and allocation. The aims of this research is to decrease the number of PM and server utilization. They exploit the splitting approach from arrival applications points of view to fulfill the mentioned goals.

## III. VIRTUAL MACHINE SPLITTING MODELS

Generally, the application requestors must find the appropriate VM and dispatch toward them. For example, VMs can played the roles of web server where the application is hosted. Here, due to fulfill the physical server's utilization it is worth splitting the large application and map them into SVMs. In other words, instead of allocating the huge amount of computational power to a single OVM to provide required resources of a single VM, it is cost effective to benefit the free capacity of the other Physical servers (PMs); such as computational power. This issue is one of our goal for fulfilling a remarkable improvement in physical server's utilization. Also, it would reduce the needs for extra $PM_s$, which leads more energy consumption. Generally, VM splitting can be analyzed and implemented from application's points of view into two models. Before explaining the two potential splitting models it is expected to clarify the OVM situation. Originally, there is a single virtual machine. And, all of the requests refer to it for processing their jobs. So the arrival requests, which find busy OVM, are forced to be waited in the waiting queue until the required resource (for example computation capacity) will be free. This situation is followed the M/M/1 queuing system where there is one waiting requestors queue and one server for serving the arrival requests [12].

The Average arrival rate of requests $(R_{arrival})$ is according to the poison distribution and on the other hand the average service rate $(R_{service})$ of the allocated requests in the virtual machine follows the exponential distribution. According to the Little's theorem [2], the average latency time per request (T) in the system before performing splitting approach which depends the average waiting time in the queue and the average request service time $(RST)$ can be obtained through equation (1).

$$T = \frac{1}{R_{service} - R_{arrival}} \quad (1)$$

In the following subsection, two potential splitting model are defined and analyzed.

*A. Parallel Spliting Model (PSM)*

In the first splitting model the OVM is split equally in which each smaller virtual machine has the same capability in terms of servicing the request's resource requirement. In other world, it is worth saying that no matter which SVM is allocated to arrival request. The PSM model associates with M/M/C model [12] in queueing system. Here, 'C' is the number of SVM with their own request's queue. The computational power and physical resource which are allocated to each SVM can be the same. For simplicity, in the PSM model it can be assumed that all of the virtual machine can have the equal computational power which are provide the request's resource requirements. The dispatched requests on the SVMs can be serviced simultaneously. Different arrival request with different workload can be situated into each of the waiting queue until the free application is ready to be allocated.

According to the M/M/C queueing system, the utilization factor of PSM model can be obtained through equation (2).

$$U_c = \frac{R_{arrival}}{C \times R_{service}} \quad (2)$$

The average system Latency per request in the PSM model ($T_{PSM}$) which is concern with the requests queueing ($W_q$) and servicing delays can be obtained via equation (3).

$$T_{PSM} = W_q + \frac{RPR}{CP_{PSM}} \quad (3)$$

Where *RPR* is the processing requirements of the allocated request on the *SVM*. And, $CP_{PSM}$ concern with the average computational power of SVM in the PSM model. In other words, $CP_{PSM}$ is the average processors rate available in PMS model.

$W_q$ is the average delay that a request must wait in the Queue. And *RST* is the average request service time of the request in virtual machine which can be obtained through equation (4).

$$RST = \frac{RPR}{CP_{PSM}} \quad (4)$$

Using ' Erlang C' formula [12] and Little's Theorem the total latency time of the system in the PSA model can be obtained,

$$T_{PSM} = \frac{P_Q}{(C \times (\frac{CP_{PSM}}{RPR}) - R_{arrival}} + \frac{RPR}{CP_{PSM}} \quad (5)$$

Where the $P_Q$ is the probability that the arrival request finds the busy SVMs and has to wait in the Queue which can be obtained through equation (6).

$$P_Q = \frac{(C \times U_c)^C p_0}{C! \times (1 - U_c)} \quad (6)$$

In the idle case it is expected that $T_{PSM}$ and the total system latency before Splitting (T) are equal. So for comparing the total latency time between the PSM model and the original case it is expected to have the same condition in the OVM in terms of request service rate. So it is assumed that the service rate of the OVM is $C \times R_{service}$. By considering the M/M/C queuing system model, the total latency of the system in the original case is defined as $T'$.

$$T' = \frac{1}{(C \times R_{service}) - R_{arrival}} \quad (7)$$

By computing the total system latency in the both OVM case and the PSM model, the difference latency of these systems ($\Delta_{Latency_{PSM}} > 0$) is obtained as follow.

$$\Delta_{Latency_{PSM}} = T_{PSM} - T' \quad (8)$$

Where, $\Delta_{Latency_{PSM}} > 0$

In the idle condition it is expected to have the same total response time before and after splitting process ($\Delta_{Latency_{PSM}} = 0$). PSM model will be suffered extra response time. So, due to preserve the system performance equal to the case before splitting process, the system should incurred the cost in terms of extra physical resources to compensate the further latency.

It is worth saying that increasing the amount of number of smaller virtual machines due to more Application splitting, although the number of arrival requests waiting in the queue ($L_q$) would be decreased [12], it leads to more free computational capacity in physical severs which can affect the physical machines utilization. This issue concerns the resource owners. Also, it is worth mentioning that it is a matter that concerns with the behavior of M/M/C system [2]. So, according to application's type, a precise trade-off for setting this parameters is expected to fulfil physical machine utilization and energy efficiency appropriately.

$$L_q = \frac{U_C{}^{C+1}(1 - U_C)}{(C-1)!(C - U_C)^2} \quad (9)$$

*B. Concurent Splitting Model (CSM),*

Running a large application on a single VM can lead the challenges in terms of performance degradation of the OVM which host the large application. So to solve this matter, we can split the large application into smaller ones in such a way that each split application which is hosted on smaller VM responsible for a specific task. In other words, instead of running a large application with different obligations on a single VM (Original Virtual Machine), each SVM will be responsible for pre-arrange duty. The arrival request of each split application depends on the processed tasks of the prior SVM. For analyzing the total latency of such series virtual machines, it is expected to exploit the network of queueing systems [12]. For simplicity it is assumed that each split application mapped on a SVM which is situated on separated PM. The free capacity of computational power in each PM can be used is such a way to guarantee the physical server utilizations and decrease the number of PMs. It leads energy efficiency too. In CSM model each smaller virtual machines plays the role of a server which follows the M/M/1 queuing system. By considering the Little's Theorem and the Kleinrock independent approximation, the total Average latency of requests in the CSM model can be obtained from equation (10).

$$Latency_{CSM} = \sum_{j=1}^{C} \left( \frac{R_{arrival\ j}}{R_{service\ j} \times (R_{service\ j} - R_{arrival\ j})} + \frac{1}{R_{service\ j}} + Link_{Latency\ j} \right) \quad (10)$$

As each arrival request in each SVM depends on the processed job of the prior SVM (except the first SVM), the system sustained a communication overhead due to each request size and the communication link load which concern with the link traffic that is changed over times. This part of delay cannot be ignored, especially when the number of SVMs are increased.

Through equation (11), the $Link_j$ latency in terms of transferring the request (i.e. the partial processed request) in $SVM_j$ is acquired.

$$Link\_Latency_j = \frac{PR_j}{link\_load_j} \quad (11)$$

Where $PR_j$ is the partial processed request that is transferred from $SVM_{j-1}$ to $SVM_j$ . Although the proposed model benefits better throughput, as it can be seen, the equation (12) demonstrates that in CSM model the system sustained the penalty in terms of communication overhead between the SVMs which leads to have response time more than running the original application request on the OVM.

$$\Delta_{Latency\ CSM} = \left(\sum_{j=1}^{c}\left(\frac{R_{arrival\,j}}{R_{service\,j}\left(R_{service\,j}-Rarrival_j\right)} + \frac{1}{R_{service\,j}} + \frac{PR_j}{link\_load_j}\right) - \frac{1}{R_{service}-R_{arrival}}\right) \quad (12)$$

Where, $\Delta_{Latency\ CSM} = (Latency_{CSM}\text{-T}) > 0$

Due to communication overhead, the CSM model cannot preserve the original latency which situated in the OVM. For preserving the total latency, same as running the large application on the OVM, the CSM system sustain penalty in terms of extra physical resources. But, as the throughput of proposed model is increase, it is expected to trade-off between these two parameters; to bring appropriate physical servers utilizations and energy efficiency.

## IV. Conclusion

Virtualization is an inevitable parts of cloud computing environment especially in data centers. Several appropriate techniques should be applied to fulfill the physical machines utilization among physical servers. This paper presented and analyzed PSM and CSM model which are two virtual splitting model. The performance of the proposed splitting models analyzed through application points of view consideration. The system latency is one of the parameters which is taken into account. Also a brief discussion in terms of system throughput of PSM and CSM are mentioned. Generally, the main goals of this paper are to bring the appropriate PM utilization and energy efficiency. Definitely, we cannot state which proposed models is the most effective approach. In other words, the significant decision making for choosing one of the proposed VM splitting model can be made provided that considering the applications type and their criteria to fulfill the mentioned goals.

## References


[1] Bernard Golden, Clarck Scheffy, Virtualization for Dummies, Sun and AMD Special Edition ed., Jennifer Bingham and Rev Mengle, Eds. USA: Wiley Publishing, Inc, ch. 1, pp. 6-12. 2008.

[2] Dimitri P. Bertsekas, Robert G. Gallager, "Data Networks, " Prentice Hall, 1992.

[3] K. Sato, H. Sato, S. Matsuoka. Model-based Optimization for Data-intensive Application on Virtual Cluster. In The 9th IEEE/ACM International Conference on Grid Computing (Grid'08), Tsukuba, Japan, pp.367-368.

[4] Liu Liu; Jie Xu; Hongfang Yu; Lemin Li; Chunming Qiao, "A novel performance preserving VM Splitting and Assignment Scheme, " 2014 IEEE International Conference on Communications (ICC), pp. 4215 – 4220, 2014.

[5] Liu Liu, Jie Xu, Hongfang Yu, Lemin Li, Chunming Qiao, " VMSA: a performance preserving online VM splitting and placement algorithm in dynamic cloud environments, " The Journal of Supercomputing, pp 1-25, December 2015.

[6] Padala, P., Hou, K. Y., Shin, K. G., Zhu, X., Uysal, M., Wang, Z., Merchant, A. , "Automated control of multiple virtualized resources, " In *Proceedings of the 4th ACM European Conference on Computer Systems, EuroSys'09.* (pp. 13-26), 2009.

[7] Prakash H R, Anala M.R, Shobha G, "PERFORMANCE ANALYSIS OF TRANSPORT PROTOCOL DURING LIVE MIGRATION OF VIRTUAL MACHINES," Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166, Vol. 2 No. 5 Oct-Nov 2011.

[8] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization. In Proceedings of the nineteenth ACM symposium on Operating Systems Principles (SOSP'03), Lake George, New York, USA, October 19-22, 2003, pp.164-177.

[9] P. Gum. System/370 Extended Architecture: Facilities for Virtual Machines. IBM Journal of Research and Development, 1983.

[10] Storage Considerations in Data Center Design, http://www.snia.org/, 2011 STORAGE NETWORKING INDUSTRY ASSOCIATION, November 2011.

[11] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," In the proc. of the workshop on Power Aware Computing and Systems (HotPower '08), December 2008.

[12] Sheldon M. Ross, "Introduction to Probability Models, " ELSEVIER, 2010.

[13] R. Creasy. The Origin of the VM/370 Time-Sharing System. IBM Journal of Research and Development, 1981.

[14] R&M Data Center, Handbook, www.datacenter.rdm.com, Reichle & De-Massari AG (R&M), 2011.

[15] XenMotion. Citrix Systems, Inc. [Online]. Available: http://www.xenserver5.com/ xenmotion.php

[16] Mauricio Arregocess, Maurizio Portolani, " Data Center Fundamental," CCIE No. 3285, Cisco Press, ISBN: 1-58705-O23-4, 2004.