Research Journal of Applied Sciences, Engineering and Technology 9(6): 471-477, 2015
DOI:10.19026/rjaset.9.1428
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 15, 2014 Accepted: November 3, 2014 Published: February 25, 2015

Research Article

A Qualitative and Quantitative Analysis of Multi-core CPU Power and Performance Impact on Server Virtualization for Enterprise Cloud Data Centers

[1]S. Suresh and [2]S. Sakthivel
[1]Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur-635109,
[2]Department of Computer Science and Engineering, Sona College of Technology, TPTC Main Road, Salem-636005, Tamil Nadu, India

Abstract: Cloud is an on demand service provisioning techniques uses virtualization as the underlying technology for managing and improving the utilization of data and computing center resources by server consolidation. Even though virtualization is a software technology, it has the effect of making hardware more important for high consolidation ratio. Performance and energy efficiency is one of the most important issues for large scale server systems in current and future cloud data centers. As improved performance is pushing the migration to multi core processors, this study does the analytic and simulation study of, multi core impact on server virtualization for new levels of performance and energy efficiency in cloud data centers. In this regard, the study develops the above described system model of virtualized server cluster and validate it for CPU core impact for performance and power consumption in terms of mean response time (mean delay) vs. offered cloud load. Analytic and simulation results show that multi core virtualized model yields the best results (smallest mean delays), over the single fat CPU processor (faster clock speed) for the diverse cloud workloads. For the given application, multi cores, by sharing the processing load improves overall system performance for all varying workload conditions; whereas, the fat single CPU model is only best suited for lighter loads. In addition, multi core processors don't consume more power or generate more heat vs. a single-core processor, which gives users more processing power without the drawbacks typically associated with such increases. Therefore, cloud data centers today rely almost exclusively on multi core systems.

Keywords: Analytical model, cloud computing, cloud workloads, server consolidation, simulation model

INTRODUCTION

Clouds are a large poll of hardware or software resources that can be accessed on-demand like a utility computing. These cloud services can be provided without any knowledge of the physical location of the servers and the systems that provide the computing services. It is continuously gaining popularity, due to its ease-of-use, on-demand resource provisioning, pay per use business model and ability to support execution of applications of diverse types (Armbrust *et al*., 2010; IBM, 2009).

Virtualization is the creation of virtual versions of a machine stack such as hardware platform, OS, storage device, or network resources. It is an art of partitions the IT hardware into slices by implementing hypervisors on top of the IT hardware and converting physical infrastructure into virtual servers, virtual networks, virtual storage etc. A Virtual Machine (VM) is a type of application that is used to create a virtual environment to run Operating Systems (guest OS). These VMs allows for a single hardware component to be used to run different operating systems and different applications (server consolidation), which may be used by multiple cloud customers. A virtual server is a VM that provides functionality just like that of a physical server. The virtual server can be located anywhere and may even be shared by multiple owners. When creating a VM in this type of system, the virtualization software (Virtual Machine Monitor) must manage the resources between the host OS and the guest OS. Because this extra layer causes additional overhead and complexity, application performance suffers (Smith and Nair, 2005; Suresh and Kannan, 2014).

A recent survey of datacenter applications show that some of the most common workloads targeted for virtualization are parallel computing applications, databases, web hosting, mail exchange and file hosting. These are multi-threaded that uses CPU, memory and Input/Output (I/O) heavily (Makhija *et al*., 2006). Thus, the success of next generation cloud computing infrastructures depends on how effectively these infrastructures, instantiate and dynamically maintains computing platforms, constructed out of cloud

Corresponding Author: S. Suresh, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur-635109, Tamil Nadu, India

resources and services. This meet varying resource and service requirements of cloud costumer applications are characterized by Quality of Service (QoS) requirements. A primary goal for data center operators is to provide good response times, in terms of Service Level Agreements (SLAs) to users. In addition, controlling data center energy consumption is also a growing challenge; reducing energy consumption lowers TCO and can also help avoid the potentially even more serious problem of reaching the limit of a data center's available power supply.

Recent years have seen a growth in Direct Connect Architecture that provides a direct connection between the processor, the memory controller and the I/O area to improve overall system performance. Hardware vendors (Uhlig, 2005; AMDI, 2006) have extended the use of Direct Connect Architecture to connect the cores on a dual or multi-core chip die and to connect each core to its memory controller. This multiple processing cores can execute multiple streams of instructions in parallel so that multiple threads of execution can be supported concurrently. Also, together connected processor cores lets data flow freely and reduces latency problems. This feature allows servers to exploit the parallelism inherent in many real life applications, especially server applications such as Web servers. Furthermore, hardware vendors building virtualization technology inside the processor, enables multi-core processors could then provide better overall performance vs. a single-core processor trying to run virtualization. With hardware virtualization along with dual or multi-core processor, there would be fewer layers and less complexity, improving application performance; it would use VMM as its virtualization software, which would manage the VM. VMM also would track the availability of physical hardware, letting applications take advantage of the hardware as it becomes available (Uhlig, 2005; AMDI, 2006).

Even though virtualization is a software technology, it has the effect of making hardware more important. As, performance study is an ongoing pursuit and hardware and software development getting matured day by day, it is desirable to do performance study in regular interval that often sheds new light on aspects of a work not fully explored in the previous publication. Thus, this study studies virtualization technology and the breakneck speed at which computer processor technology has progressed, enables the cloud computing environment to become viable and beneficial to customers. Hence, this study does analytical study and simulation model of, does incorporating multi core technology enables higher levels of performance and consumes less power typically required by a higher frequency single core processor with equivalent performance, to become viable and beneficial to customers. As part of this, the study focuses on a few simple questions:

- Whether the consolidated Physical Machine (PM) should have n multi cores, or a fat CPU n times that of each original server.
- How much smaller than the number of PM (m), can we make chosen cores (n) while preserving tolerable performance i.e., Can multi core VM cluster scale well; and finally.
- What is the impact on energy/power usage and whether this modifies the chosen policy?

The main contributions of this research work are as follows:

- Analytical performance modeling based on queuing theory for multi core impact on server consolidation for performance and power consumption.
- Analytical study of non-virtualized and virtualized model for core impact on performance and power consumption.
- Propose and developing a simulation model of non-virtualized and virtualized model and validate it for core impact on performance and power consumption.
- Evaluate and characterize the simulation model for cloud workloads whether multi core strategy is effective over fat CPU server in performance and power.
- Investigate factors that impact the effectiveness of consolidation in multi-core environments.

Evaluating new system model and techniques under various, controllable and repeatable conditions are hard in real Cloud. Because, system level implementation requires deep understanding on hardware architecture, as well as low level programming and debugging. In order to address this problem, the cloud virtualization environment, is simulated to evaluate the performance of the model using CSIM simulation toolkit (Schwetman, 2001), a C/C++ library that allows assembling a complete virtualization system with flexible configurations.

LITERATURE REVIEW

With respect to hardware advancements efficient power management and performance optimization in large-scale data centers and server clusters has gained much attention in the research community in recent years. There exists a huge body of literature on performance, energy efficient computing and communication. This section provides some view on them.

Pedram and Hwang (2010) present an analytical power and performance modeling for multi-tier internet applications based on the mean-value analysis .It is based on queuing theory to calculate the response time

of the clients based on CPU and I/O service times. However the values all are theoretical based.

Zheng and Cai (2010) proposed energy proportional M/G/1 queuing system model based service differentiation in server clusters, provides controllable and predictable quantitative control over power consumption with theoretically guaranteed service performance. In addition, it adapts vary-on vary-off mechanism that turns servers on/off to adjust the total number of active servers based on the workload.

Carlsson and Arlitt (2011) presented queuing theory based analytic modeling and simulations model to show that the effective utilization of a server running a delay sensitive application like a web server. Furthermore, by intelligently scheduling delay tolerant batch applications alongside delay sensitive web applications on the system they showed that the system responsiveness can be improved without significantly affecting the response times.

Shari *et al*. (2011) use simulations to evaluate a multi core server resource framework among applications such that individual performance requirements of these applications are met. The framework is evaluated on the SPECCOMP and the SPECJBB benchmark suite. Our paper is based on this model wherein it considers geographical cloud workloads.

Goudarzi and Pedram (2013) proposed a geographical load balancing solution for a multi-datacenter cloud system focused on response time sensitive interactive applications. It decides about VM assignment to datacenters or VM migration from one datacenter to another, by considering the heterogeneity of VMs and datacenters, cooling system inefficiency and peak power constraint in each datacenter.

Suresh and Kannan (2013) studies performance behaviors of full virtualization models of different architectures such as hosted (Virtual Box) and bare Metal (KVM) virtualization using micro level, macro level and application level benchmarks with the current hardware advancements and software advancements in the cloud environment. In effect it yields that Virtual Box outperforms KVM in CPU and thread level parallelism and KVM outperforms in all other cases. Both are very reluctantly accepted in disk usages comparing with native system.

Chen *et al*. (2014) evaluated energy consumptions theoretically for data centers using load balancing and server consolidation. In effect, they conclude that server consolidation helps in improving resource utilization by consolidating many VMs residing on multiple under-utilized servers; Load balancing can help in decreasing energy consumption by regularly dispensing the load and decreasing the resource consumption, hereafter decreasing energy consumption.

Junwei *et al*. (2014) present a strategy, formulates an optimal power allocation and load distribution for multiple servers in a cloud as optimization problems. In specific, it is defined for multiple multi core server processors with different sizes and certain workload; addressing two important research problems that explore the power performance tradeoff in large-scale data centers from the perspective of optimal power allocation and load distribution.

Suresh and Sakthivel (2014) proposed a novel system using meta-heuristic combinatorial search techniques that automatically regulates the VMM CPU scheduler related to the applications on-the-fly with dynamic changes in the environment to maximize throughput and minimize response time. Their evaluation, for various scenarios with synthetic workloads shown that application response time in cloud get impacted with adaptive CPU resource allocation with changing customer workloads.

Several researchers have explored the impact of consolidating multiple workloads on non-virtualized multi-core systems using simulation and queuing theory model. However, no work is aimed at Time Zones and geographies specific variable workloads. To the best of our knowledge, multi core impact on power and performance trade off has not been analyzed and investigated before for the various cloud workloads. Thus, our investigation in this study makes initial effort to analytical and simulation study of power-performance tradeoff in data centers with multi core vs. fat CPU virtualized servers for various geographies specific cloud workloads. Our results in this study provide new theoretical and practical insights into power management and performance optimization in cloud computing.

METHODOLOGY

Analytic models: As improved performance is pushing the migration to multi-core processors, this section explores power and performance implications and limits of server consolidation through a simplified model. It provides an analytic modeling to evaluate the number of VMs has a direct effect on how the available resources (CPU cores) are shared among all VMs. Assume, consolidating applications running on m Physical Machine (PM) of capacity C on to one PM, given as $n*C$. The paper utilizes a M/G/1queuing model (Kleinrock, 1975) to calculate the Response Time (RT) exhibited when processing requests as a function of computational capacity and request arrival rate. The model assumes an exponentially distributed request inter-arrival time with mean $1/\lambda$ and a server which process requests with a constant service time with mean $1/\mu$. Based upon these assumptions, the model defines the average normalized response time = (average response time of an algorithm/average response time of the baseline algorithm), as RT = T λ (T-the average time spent in the system). The queuing theory yields the normalized response time (RT_o) using m PM as:

$$RT_o = \rho/ (1-\rho) \quad (1)$$

For each of the PM (where ρ (server workload) = λ/μ*)*. When consolidating these VMs into one PM (m processors), the queue becomes one with m machines working at the same rate μ, servicing an arrival rate of m*λ. Subsequently, the normalized Response Time (RT_p) for one server with *p* processors is:

$$RT_p = m*\rho + Q_P/ (1-\rho) \quad (2)$$

(Here Q_P is the 'queuing probability', Q_P<< for light loads). Suppose, a single PM that is m times faster, once again it is modeled as a single server queue but with service rate $m*\mu$. Since ρ remains intact, the normalized RT in this case (RT_c) remains the same as RT_o in 1. Thus it is apparent that for Q_P<< case, the consolidation onto a multi-core machine versus fat CPU can result in significant degradation in performance, at least as measured by average normalized RT. For heavy loads, on the other hand, the second term is poor in both cases.

Now consider the case where the single PM onto which we consolidate the workload is only n times faster than the original servers. In this case, it is obvious that the normalized response time RT_n is:

$$RT_n = m*\rho/ (n-m*\rho) \quad (3)$$

From the Eq. (3), it is apparent that it is possible to use a PM far less powerful than the consolidation of the *m* original PM, as long as *n/m* remains reasonably large as compared to ρ; and if indeed $n>>m*\rho$ then the average normalized RT degrades only linearly by the factor of n/m.

Thus it is obvious that the examination yields limits to server aggregation using virtualization. The theoretical maximum benefit, in terms of a reduction in number of servers, is *n/m* = ρ, at which point the system becomes unresponsive. In practice it is possible to get fairly close to this, i.e., if n/m = ρ (1+Є), then the average normalized RT becomes 1/Є. In consequence, whatever the initial inefficiency, one can decide on an acceptable average normalized RT and plan the consolidation strategy accordingly.

Subsequent to the aforementioned case, in the case of consolidation onto an n-core server, average normalized RT_p, as:

$$RT_p = m*\rho + Q_P/ (1 - ((m*\rho) /n)) \quad (4)$$

It is apparent that the RT remains the same as RT_p in (2) for Q_P<<. Thus the RT still degrades by a factor of m, independent of n, as compared to fat CPU case (1). However, in the case of (Q_P>>) heavy load, where the second term dominates, there is a degradation in performance in the multi-core case (n<<m), over (m = n), say (2).

When Q_P<<, the response time same as 2. It shows that RT degrades by a factor of *m*, independent of *n* over 1. In addition when Q_P>>, there is a marked degradation in performance in the multi core processor case if *n*<<*m*, as compared to 2. It directs to the end that the cores allocated to a VM directly impact the hosted application's performance. Further, analytical model proves that multithreaded application servers can exploit multi-core architectures efficiently thus clouds rely almost exclusively on multi-core, systems.

As power consumption of chips is given as, $P \alpha V^2$ (P-power, V-voltage), a system that runs at a clock speed n times faster than a 'base' system, will consume n^2 the power of the base system; thus it is apparent that the multi core (n core) system will consume only n times the power. Thus there is a trade off, between reducing power usage by VMs aggregation onto multi core CPU systems, versus improved performance on systems with fat CPU but at the cost of nonlinear growth in power consumption per PM.

Simulation modeling of the system: To support the analytical study, a simulation model is build with server clusters (each cluster is assumed a VM) and experiment with policies to switch on and switch off CPU cores based on its usage. The server clusters have 12 server machines (for a single core virtualized machines case). The request service time is deterministic and is 200 msec for each machine in the cluster. A load balancer (VMM) controls the server cluster, which distributes arriving requests to server cluster in a random fashion, executes a policy to switch on/off the used and unused CPU cores in response to usage at the machines. The policy is, switch on core consumes 200 Watts and a switch off core consumes 5 Watts. The time to switch on and switch off a machine is instant. A machine must be switch on/off for a minimum of 1 min before it can change its power state. The SLA for the system is, the server cluster must maintain an SLA based on measured response time. The SLA states that the mean response time must not exceed 250 msec and that the 99% response time must not exceed 500 msec. A figure of the simulated server consolidated system is shown in Fig. 1.

One of the key areas to be looked at in the cloud is how a cloud service provider handles various capacity requirements in different time zones at a given time. To make experiments reproducible, it is important to rely on a set of input traces to reliably generate the workload, which would allow the experiments to be repeated as many times as necessary. It is also important to use diverse workload traces collected from a real system such as slowly varying (ITA, 1998), Large variations (NLANR, 1995) and artificially generated quickly varying (synthetic workload-5 customers/sec) workload, as this would help to reproduce a realistic scenario. Furthermore, this implies that the overall system workload is composed of multiple independent heterogeneous applications, which also corresponds to a cloud environment.

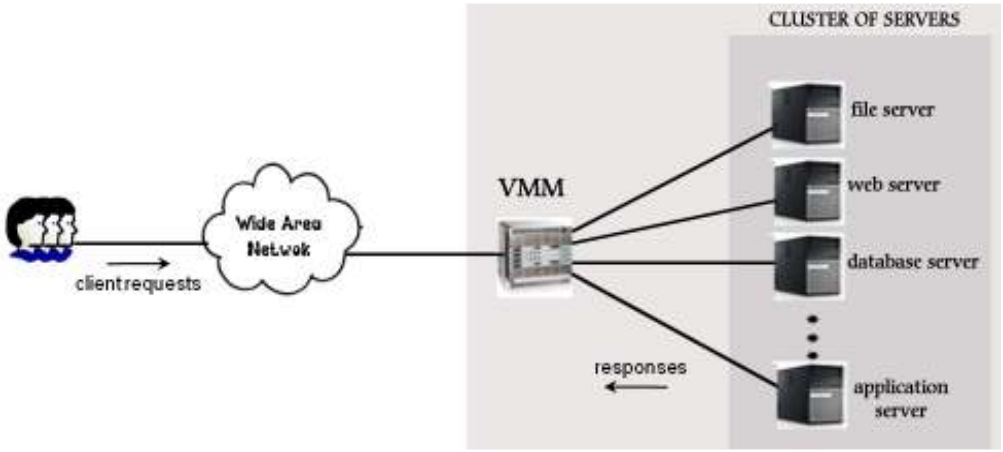

Fig. 1: Simulation modeling of the system

We want to compare three different systems in terms of mean response time (mean delay) and power vs. offered load. It is assumed that clusters servers are equipped with single or multi core CPUs. A multi-core CPU with n cores each having m MIPS is modeled as a single-core CPU with the total capacity of nm MIPS in the single fat server.

Scenario 1: Single physical machine with clock speed m times that of each VM-this can be modeled as a single M/G/1 server with the service rate m*μ and the arrival rate is λ. It is considered as the base condition.

Scenario 2: m consolidated VMs each having a single core i.e., 12 VMs each with single core. This can be modeled as M/G/m system -system where m queues of M/G/1 type with service rate μand the arrival rate is λ are in parallel; such that every customer enters each system with the same probability.

Scenario 3: m/n consolidated VMs each having n = {2, 3, 4} cores. In the sense, scenario 3.1 (6 VMs each VM with two core), scenario 3.2 (4 VMs each VM with 3 cores) and scenario 3.3 (3 VMs each VM with 4 cores) represent consolidated virtualized environment. This can be modeled as an M/G/m system and a system where m queues of M/G/n type with service rate n*μ and the arrival rate λ are in parallel; such that every customer enters each system with the same probability.

RESULTS AND DISCUSSION

Table 1 gives performance in terms of mean response time (mean delay) and power consumption results, vs. offered load of server consolidation strategy for multi-core CPU machine vs. one with faster clock speed for the aforementioned scenarios. For large variations workload (heavy workload), VMs with 3 and 4 cores (scenario 2 and 3) works much better than the fat server. This is justified since applications, as well as VMs, is not tied down to processing cores and can be executed on an arbitrary core in parallel. The consolidation of VM is here about 3 to 4 VMs In addition scenarios 2 and 3 gives good power usage over fat single server. For quickly varying synthetic workload (moderate workload) all multi core scenarios (single core, double core and 3 cores VMs) budge with SLA and gives good power savings. Especially, for the scenario 3.1, 3.2 and 3.3 the response time is 0.20219, 0.20034 and 0.20007 respectively (too good). However, the server clusters with single core budge to SLA and gives good energy savings over all other cases. As it allows the high consolidation density (consolidation density refers to the ratio of virtual machines to physical machines) of 12, scenario 2 is preferable for the moderate workload. In the case of low workload, all scenarios satisfy the SLA. However, single fat server works well in this case in terms of response time (Min.-0.0166, Max. -0.0530). However, CPU usage and utilizations are very good in multi core cases for all workloads than in the fat server case (very poor). Similarly for the power metric, it is obvious that single fat server consumes roughly the same power (1728 KWh) irrespective of the cloud workloads. However, multi cores are highly promising and assure power savings in all cases by switch off/on unused cores based on the demand. In summary consolidating workloads on the same multi core physical node reduces the number of active servers and increases the utilization of individual nodes, improving the energy efficiency. Subsequently, the performance and energy efficiency of a consolidated system considerably varies depending on the potential resource availability on the VMs.

An intuitive explanation for the behavior of the systems is the following: in the case of 12 parallel M/M/1 queues (scenario 2) there is always a nonzero probability that some servers have many customers in their queues while other servers are idle. In contrast to that, in the M/M/m case (scenario 3.1, 3.2 and 3.3, respectively) this cannot happen. In addition to that, the

Table 1: Performance in terms of mean response time (mean delay) and power consumption vs. offered load of server consolidation strategy for multi-core CPU machine vs. one with faster clock speed

	Performance (sec)								
	Response time								
Scenario	Min.	Max.	SLA (Avg.)	50%	98%	SLA (99%)	CPU usage	Utilization	Power usage (KWh)
Large variations									
1.0	0.0167	71.3420	1.8834	0.0316	16.246	28.178	5.52035	0.208	1728
2.0	0.2000	192.3060	2.7226	0.3931	25.511	40.221	6.98844	2.490	1285
3.1	0.2000	74.1480	1.9927	0.2638	16.924	28.381	6.83243	2.489	1413
3.2	0.2000	69.9890	1.8468	0.2072	16.326	27.859	6.81243	2.489	1488
3.3	0.2000	67.1520	1.6204	0.2000	15.952	27.749	6.68442	2.489	1527
Quickly varying (synthetic) workload									
1.0	0.0167	0.1136	0.01738	0.0167	0.02963	0.032038	4.5523	0.021	1728
2.0	0.2000	28.6739	0.25258	0.2000	0.48393	0.666198	5.2520	0.248	518
3.1	0.2000	2.2675	0.20219	0.2000	0.22726	0.288071	5.2003	0.250	796
3.2	0.2000	0.6861	0.20034	0.2000	0.20000	0.200000	5.0603	0.250	962
3.3	0.2000	0.4503	0.20007	0.2000	0.20000	0.200000	5.0723	0.248	1084
Slowly varying workload									
1.0	0.0166	0.05300	0.016660	0.0166	0.01660	0.01660	4.4803	0.002	1728
2.0	0.2000	1.14180	0.206340	0.2000	0.33852	0.37363	4.7323	0.025	197
3.1	0.2000	0.54180	0.200150	0.2000	0.20000	0.20000	4.6403	0.025	350
3.2	0.2000	0.38510	0.200003	0.2000	0.20000	0.20000	4.6120	0.025	504
3.3	0.2000	0.26032	0.200000	0.2000	0.20000	0.20000	4.5520	0.025	656

fat single server (scenario 1) is especially for lighter loads better than the M/M/12 system, since if there are only $k<12$ customers in the system the M/M/12 system has a smaller overall service rate $k*\mu$, while in the fat server all customers are served with the full service rate of $12*\mu$.

Simulation results prove that a processor with more cores works faster and more efficiently than a single fat CPU processor for several reasons: When multitasking, users will experience fewer bottlenecks than with single fat CPU processor. When running two processor intensive applications, each one can access its own core. When running a single application, multi cores can share the processing load, improving overall system performance. However, when multiple VMs concurrently run on the same physical host, they share the available physical resources, causing unpredictable drop in their performances when some of them have compute-intensive peaks. Service providers have therefore to carefully provision the amount of resources required in order to meet user performance requirements. To this end, they need performance models in order to evaluate the performances that can be achieved from a given hardware configuration. As different applications have different resource usage traits, collocate different application instances together on a physical server is more beneficial and ensures good resource usage. Say, collocate a CPU-intensive application with a memory-intensive application rather than collocating two CPU intensive applications. The system also can shut down portions of the cores that aren't in use, saving power and generating less heat. Apparently multi core processors don't consume more power or generate more heat vs. a single-core processor, which will give users more processing power without the drawbacks typically associated with such increases.

CONCLUSION

As the computing industry enters the multi-core era and more and more CPU cores are integrated into one single die, this study did analytic and simulation study, on does incorporating multi core technology enables higher levels of performance and consumes less power typically required by a higher frequency single core processor with equivalent performance. Analytical study shows that there is a trade off, between reducing power consumption by consolidating onto multi core CPU systems, versus improved performance on systems with faster clock speeds but at the cost of nonlinear growth in power consumption per server. Simulation model for the diverse workloads prove that all virtualized models for varies cores provided major improvements in performance and power consumption per workload over single fat CPU servers. With each scenario, runtimes remained flat until the number of VMs reached the number of cores, then increased in a predictable, linear way. Multiple processor cores in a single package that delivers parallel execution of multiple software applications; thus it is obvious that effective usage of the core resources becomes promising future. In effect, whatever the initial inefficiency, one can decide on an acceptable average normalized response time and plan the consolidation strategy accordingly. Lastly, apart from consolidation, it is important to note that individual applications implemented using multi-threaded application servers can also exploit multi-core architectures efficiently. Therefore, cloud data centers nowadays rely almost completely on multi-core, multi-processor systems.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the reviewers for their comments on improving the presentation and structure of the study. This work was supported by University of Grants Commissions, New Delhi, India, under Minor Research Project Scheme Grants No. MRP-4896/14 (SERO/UGC).

REFERENCES

AMDI (Advanced Micro Devices Inc.), 2006. AMD I/O-Virtualization Technology (IOMMU) Specification. Advanced Micro Devices Incorporation.

Armbrust, M., A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, 2010. A view of cloud computing. Commun. ACM., 53(1): 50-58.

Carlsson, N. and M. Arlitt, 2011. Towards more effective utilization of computer systems. SIGSOFT Softw. Eng. Notes, 36(5): 235-246.

Chen, F., J. Grundy, J.G. Schneider, Y. Yang and Q. He, 2014. Automated analysis of performance and energy consumption for cloud applications. Proceeding of the 5th ACM/SPEC International Conference on Performance Engineering, ACM, pp: 39-50.

Goudarzi, H. and M. Pedram, 2013.Geographical load balancing for online service applications in distributed datacenters. Proceeding of the IEEE International Conference on Cloud Computing (CLOUD'2013), Santa Clara.

IBM, 2009. The Benefits of Cloud Computing: A New Era of Responsiveness, Effectiveness and Efficiency in IT Service Delivery. Dynamic Infrastructure. Retrieved from: http:// www.informationweek.com/whitepaper/Software/Server- Virtualization/the-benefits- of-cloud-computing-a-new-era-of-res- wp1294274216447? articleID=177100033.

ITA, 1998. The Internet Traces Archives: World Cup 98. Retrieved from: http://ita.ee.lbl.gov/html/ contrib/WorldCup.html.

Junwei, C., L. Keqin and I. Stojmenovic, 2014. Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers qualitative performance study. IEEE T. Comput., 63(1).

Kleinrock, L., 1975. Queuing Systems: Theory. Vol. 1, John Wiley and Sons, NY.

Makhija, V., B. Herndon, P. Smith, L. Roderick, E. Zamost and J. Anderson, 2006. VMmark: A scalable benchmark for virtualized systems. Technical Report, VMware-TR-2006-002.

NLANR, 1995. National Laboratory for Applied Network Research. Anonymized Access Logs. Retrieved from: ftp://ftp.ircache.net/Traces/.

Pedram, M. and I. Hwang, 2010. Power and performance modeling in a virtualized server system. Proceeding of the 39th International Conference on Parallel Processing Workhops (ICPPW).

Schwetman, H., 2001. CSIM19: A powerful tool for building system models. Proceeding of the Winter Simulation Conference, pp: 250-255.

Shari, A., S. Srikantaiah, A.K. Mishra, M. Kandemir and C.R. Das, 2011. Mete: Meeting end-to-end qos in multicores through system-wide resource management. SIGMETRICS Perform. Eval. Rev., 39(1): 13-24.

Smith, J.E. and R. Nair, 2005. Virtual Machines: Versatile Platforms for Systems and Processes. 1st Edn., Morgan Kaufmann Publishers, Amsterdam, Boston.

Suresh, S. and M. Kannan, 2013. A performance study of hardware impact on full virtualization for server consolidation in cloud environment. J. Theor. Appl. Inf. Tech., 60(3).

Suresh, S. and M. Kannan, 2014. A study on system virtualization techniques. Int. J. Adv. Res. Comput. Sci. Technol., 2(1).

Suresh, S. and S. Sakthivel, 2014. SAIVMM: Self adaptive intelligent VMM scheduler for server consolidation in cloud environment. J. Theor. Appl. Inf. Tech., 69(1).

Uhlig, R., 2005. Intel virtualization technology. Computer, 38(5): 48-56.

Zheng, X. and Y. Cai, 2010. Achieving energy proportionality in server clusters. Int. J. Comput. Netw., 1(2): 21-35.