Analysis of Data Centre Resources for Businesses

Govinda.K[#1], Saheel Sasi.A[*2]

[#1]*SCS,VIT University*
Vellore, India
[1]kgovinda@vit.ac.in

[*2]*SCSE,VIT University*
Vellore, India
[2]saheelsasi@gmail.com

***Abstract*— One of the practical concerns that have attracted significant attention is the efficient resource management in the virtualised data center. In order to maximize the revenue for commercial cloud providers, the economic allocation mechanism is desired. Nowadays, industries are seeking scalable IT solutions such as data centers, hosted in either in-house or a third party for the advancement of virtualization technologies and the benefit of economies of scale. It is ubiquitous to have data centers via cloud setting. Very little is known about the interaction of workload demands and resource availability. A large scale survey of in-production data center servers within a period of two years would fill this gap. The seasonality of resource demands and its affects by different geographical locations are the main focus. This paper presents a brief analysis of data center growth to meet business demands in different geographical locations.**

***Keywords*— Data Centre, Resources, Region, Asia, Analysis.**

I. INTRODUCTION

To provide reliable and scalable computing infrastructure for massive internet services, more and more data centres have been built across the world today. Modern virtualization based data centers could better support the increasing trend of cloud-based infrastructure with flexible resource management by taking advantage of the virtualization technology. Heterogeneous applications are supported by virtualization technology. Therefore, it can reduce amount of hardware in use and lead to energy-efficient system because of consolidation. Resources are collected into a virtual resource pool for sharing among users in the virtualized data center. Computer resources can be obtained on a pay as you go basis. VMs with different CPU, memory and disk capacity are the key specifications required in an application.

To set up a data center one has to invest a great amount, in which more than 45% of the total costs go the servers. According to a report, Google spent $2.6 billion for data centers in 2007. Therefore, useful work accomplishment per dollar invested is an important goal. But, utilization in the data center can turn out to be remarkably low, which is around 5-15 % on an average. It could greatly affect the revenue if there is low resource utilization. The two main reasons for the low average resource utilization could be; Firstly, we cannot represent the real runtime demand as the users' request for resources is always blind. Therefore, when the application doesn't use up its reserved share, the resources apparently tend to be wasted. This could lower the revenue eventually. Secondly, underuse of server resources can be caused due to unsuitable VM placement. There are several popular ways for VM placement. Performance reduces when allocating VMs randomly without considering network topology. The other method simply consolidates VMs for resource savings also ignores the network condition, thereby causing serious traffic along with large cost between few nodes.

It is an emerging and accelerating trend to migrate systems onto the cloud. to facilitate the development of cloud computing, data centers are considered as the back-bone. Cloud performance hinges on the computation, storage and network capacity provisioned at the data centers. Virtualization technologies are powered by data centers that enable multiple resources being multiplexed or shared. The data center management is challenging. It can be classified into two;

(i) Resource management, that focuses on dynamically controlling workloads given a resource pool and,
(ii) Capacity planning, which focuses on resource provisioning.

Data centers are further powered by virtualization technologies [1]–[4] that enable multiple resources being multiplexed/ shared among a large number of users with diverse time-varying access patterns [5], [6]. A large number of studies [3], [7], [8] have aimed at examining virtualization technologies to improve data center efficiency via workload consolidation. Capacity planning on the other hand often relies on time series methodologies [8]–[10] for workload forecasting in order to dimension resource capacity. In general, little is known about "real" workloads placing demands on data centers and how their combined demands on different resources evolve across time.

This paper fills this gap with effective capacity planning by providing hard data on the evolution of workload resource demands that are in production. A detailed performance survey across several thousand servers in different data centers located in four different continents from June 2009 to May 2017 is conducted in this paper. The centers picked up are mature data centers, i.e., the customer servers would be corporations due to which there would be no churn from casual users. Our main objective is to make an evolutionary view of the work load over a period of time. We aim at presenting data both from a global perspective but also from a more narrow point by taking the nature and the sheer quantity into consideration. We present some elementary economic analysis along with the presentations of the utilization of CPU, memory and disk in different time scales i.e. daily and monthly.

II. LITERATURE REVIEW

A. Data Center Network Architectures

The current data centers, access tier, aggregation tier and core tier are known as the three tier architecture. The bottom level is access tier, in which each server connects to one (or more, for redundancy purposes) access switch. Each access switch connects to one (or more) switches at the aggregation tier, and finally, each aggregation switch connects with multiple switches at the core tier. This approach supports multiple heterogeneous applications with a familiar management infrastructure. The inter-VM communication has to transfer data from access tier through aggregation tier to achieve core tier (if necessary), then return. The more tiers packages pass, the more bandwidth consumption as well as the more communication latency. Thus, the data center network may become unbalanced depending on the communication patterns. That is, the higher tiers have much more traffic than the lower ones because of less links and bandwidth sharing among multiple switches, especially severe in the core tier.

Several new data canters network architectures have been proposed regarding core tier network bottleneck in the last two years. VL2[2] also has three-tier architecture but with main difference that the core tier and the aggregation tier form a Clos[4] topology, and designing valiant load balancing protocol to relieve the core tier workload. PortLand [3] is another three-tier architecture that shares with the VL2 the same Clos topology feature. It makes use of fat-tree topologies and evenly distributing the up-links between all the aggregation switches. Both VL2 and PortLand improve network load balancing by designing novel network topology and protocols. However, it would lead to much more complex link connection that increases the operation and maintenance expense along with difficult identified name space problem.

B. Overview of Revenue Management and Overbooking

In various industries like airlines, hotels and car rentals [5], the revenue management has been widely adopted. The data centers which wish to focus on maximizing the revenue can also apply the revenue management technique. To increase the profit, various industries sell more than its capacity as a strategy of revenue management which is known as overbooking. However, it differs from the fields of application. A seat in an airplane cannot be occupied after the flight has taken off. In the same way the grid does not require fixed starting times for a resource, Overbooking for high-performance computing (HPC), cloud, and grid computing has been introduced in [6][7]. DRIVE [18] studies various economic resource allocation strategies, including the applicability of overbooking in the Grid computing, but fails to provide specific solutions. Sulistio[19] proposes overbooking strategies to mitigate the effects of application cancellation and no-show focusing on the time scale. It doesn’t deal with the under-usage resource during the runtime.

C. Resource Allocation in Data Centers

Various capacity tools such as VMware Capacity Planner [8], IBM WebSphere CloudBurst [9], Novell PlateSpin Recon [10] and Lanamark Suite [11] normally decides the VM placement. These tools seek to consolidate VMs for CPU, physical memory and power consumption savings, thus can lead to situation in which VM pairs with heavy traffic among them are placed on host machines with large network cost between them. Application level optimization techniques [12,13] alleviate the problems based on the current resource allocation state. The original VM placement is used to determine the network consumption. The initial resource allocation that could be responsible for various performance anomalies is critical to be determined. Tara[14] proposes a new architecture for optimized resource allocation in Infrastructure-as-a-Service (IaaS)-based cloud systems. The allocation decision taken by the IaaS is strategized by a what-if methodology. To estimate a the performance for a given resource allocation strategy, a prediction engine with a light weight simulator is used. This paper deals with exclusive deploying of application to servers along with focusing on MapReduce.

III. EXAMINED ANALYSIS

This paper describes the analysis of resource requirements for businesses at various regions with respect to different sizes of data canters to meet business needs from cloud consumers described as follows. We first describe about the scenarios of Asia-pacific region, Canada, Eastern Europe and Japan in terms of single data center, rack/computer room, mid-size data centers, enterprise data centers and large data canters respectively.

A. Asia Pacific Region

In the Asia pacific region, when the statistics are examined, the single data centers have a drastic growth between the years 2010 and 2016 to meet business requirements as shown in figure1.

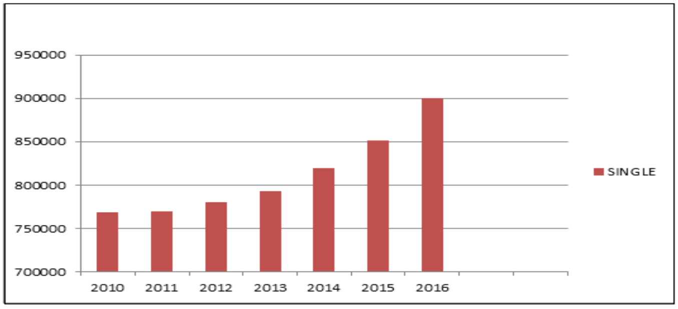

Fig1. Year Vs Size of Single Data Center

There is a gradual growth in the number of rack/computer rooms between the years 2010 to 2016 as shown in figure2.

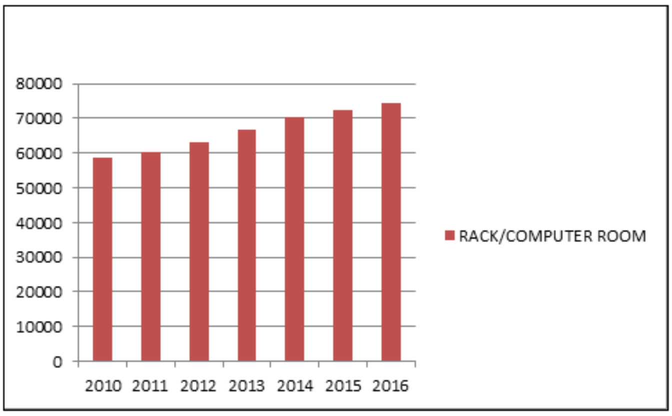

Fig2. Year Vs Size of Rack

Similarly the mid size data centers also grow from 2007 to 2016 as shown in figure3.

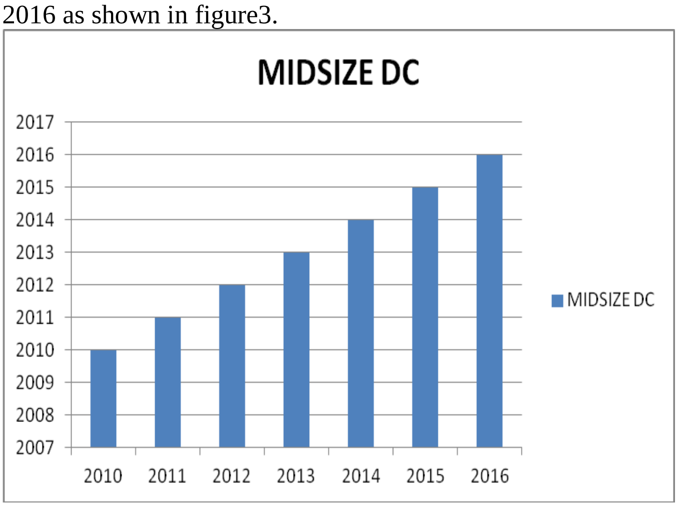

Fig3. Year Vs Mid-Size Data Center

When it comes to the enterprise data centers the numbers range from 1000 to 1600 between the years 2010 to 2016 as shown in figure4.

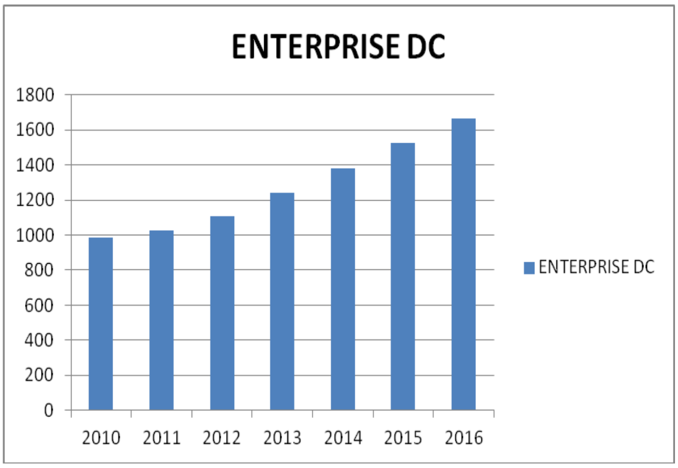

Fig4. Year Vs Enterprise Data Center

Large data centers grew from 100 to 200 within the same span of years.

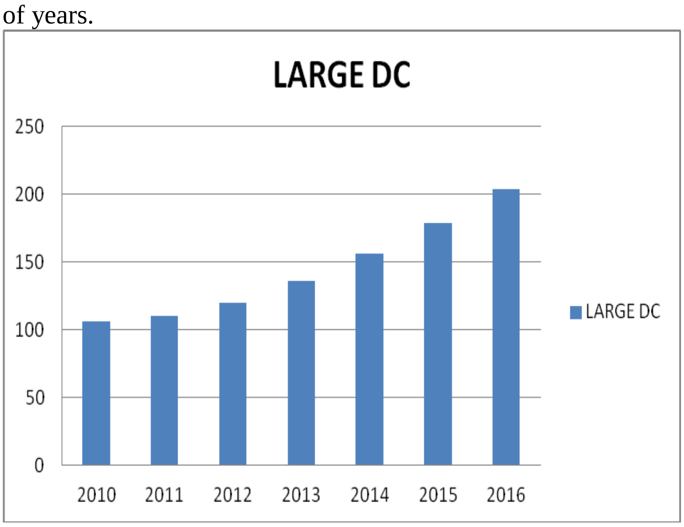

Fig5. Year Vs Large Data Center

B. Canada

In Canada the statistics witness sudden crests and troughs within a short span of 7 years as shown in figure6.

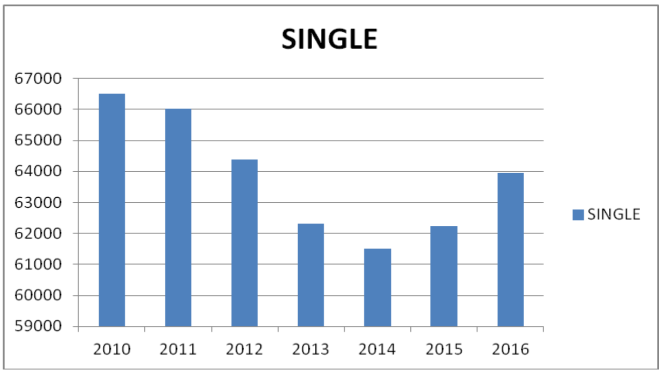

Fig6. Year Vs Size of Single Data Center

In Canada the statistics witness sudden crests and troughs within a short span of 7 years for rack as shown in figure7, Through this examination of the chart, we notice a sudden fall of racks from more than 14000 to less than 12000.

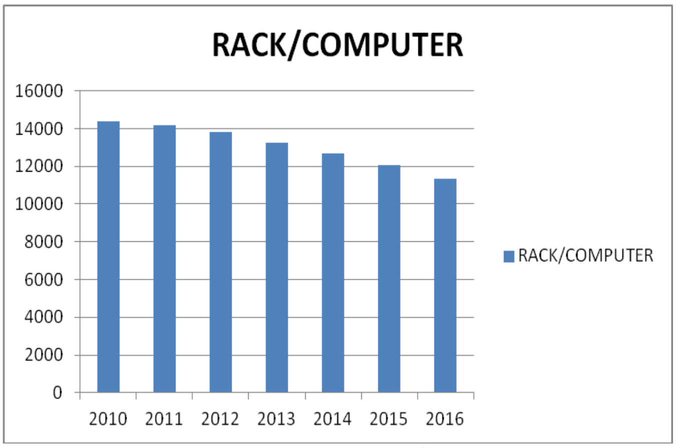

Fig7. Year Vs Size of Rack

Mid size data centers decreased from 650 to 564 as shown in figure8.

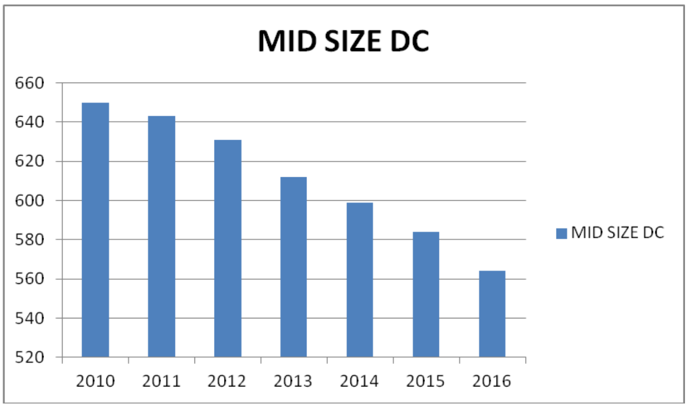

Fig8. Year Vs Mid-Size Data Center

The enterprise data centers decreased eventually between 2010 and 2012 but had a sudden raise from 2013 to 2016 as shown in figure9.

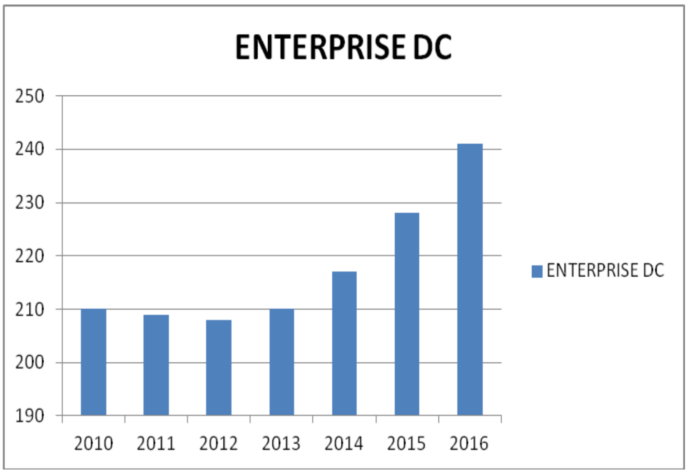

Fig9. Year Vs Enterprise Data Center

In case of large DC increased from more than 20 in 2010 to more than 35 by 2016 as shown in figure10.

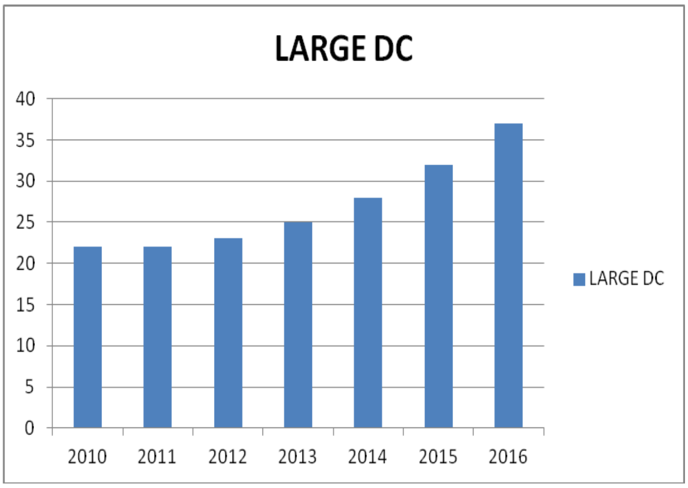

Fig10. Year Vs Large Data Center

C. Eastern Europe

Single data centers ranged from 147274 to238792 between 2010 to 2016 as shown in figure11.

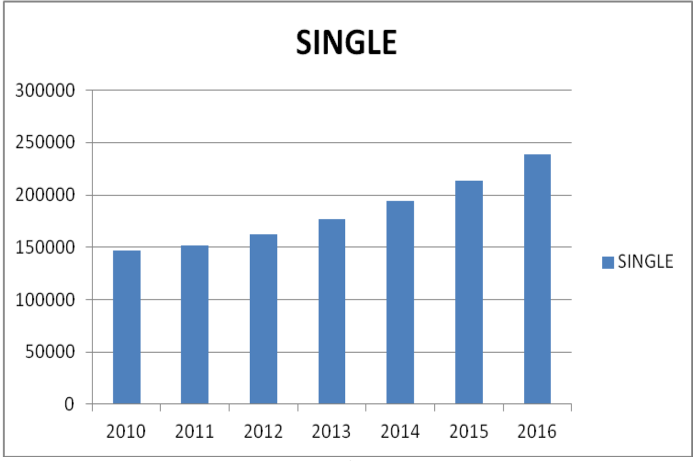

Fig11. Year Vs Size of Single Data Center

As shown in the figure12 the number of racks/computer rooms fluctuated between the years 2010 and 2016.

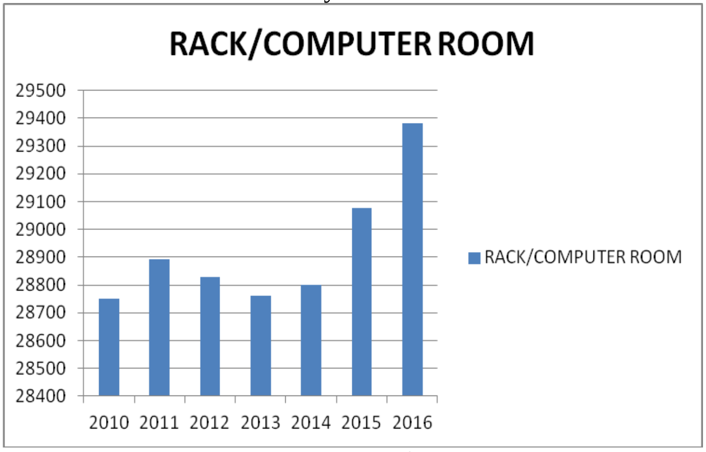

Fig12. Year Vs Size of Rack

It fell drastically from 380 to 286 between the years 2010 and 2016.

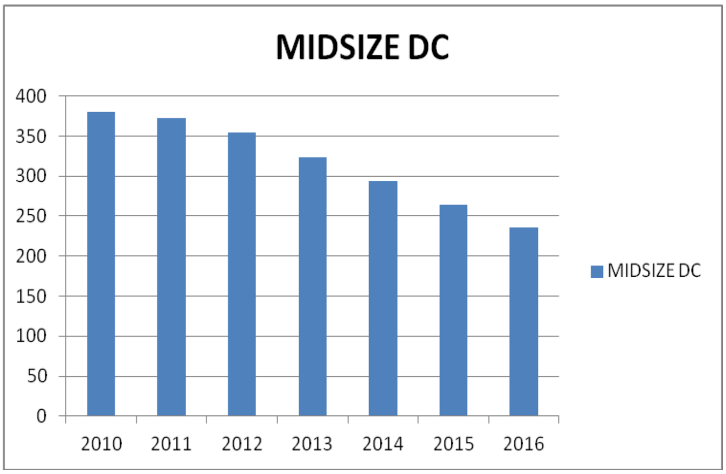

Fig13. Year Vs Mid-Size Data Center

Enterprise DC eventually increased from 196 to 243.

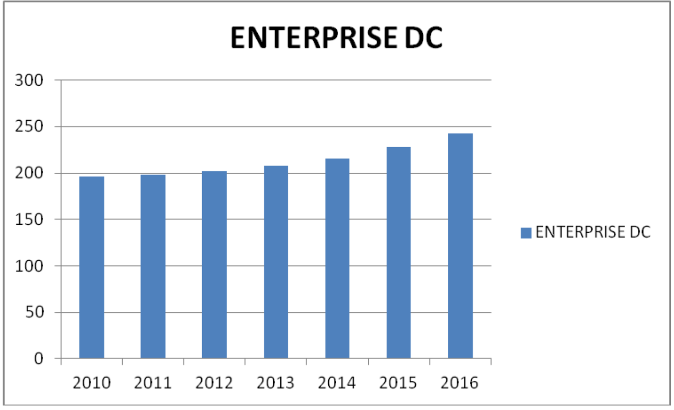

Fig14. Year Vs Enterprise Data Center

There were 44 large data centers in 2010 and increased to 65 by 2016.

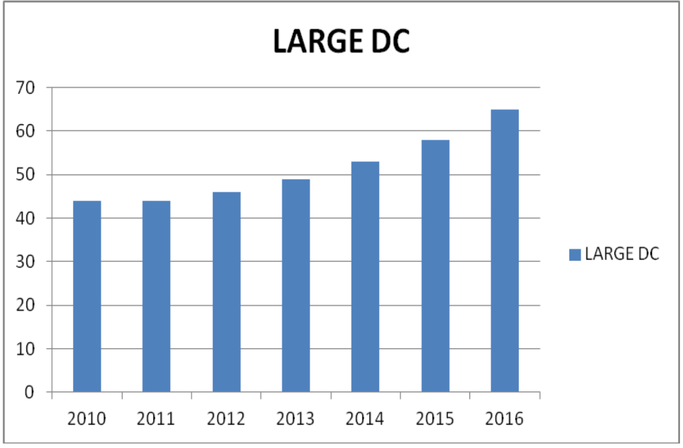

Fig15. Year Vs Large Data Center

There were 44 large data centers in 2010 and increased to 65 by 2016.

D. Japan

Single data centers increased in 2011 compared to 2010 but experienced a sudden fall by 2016.

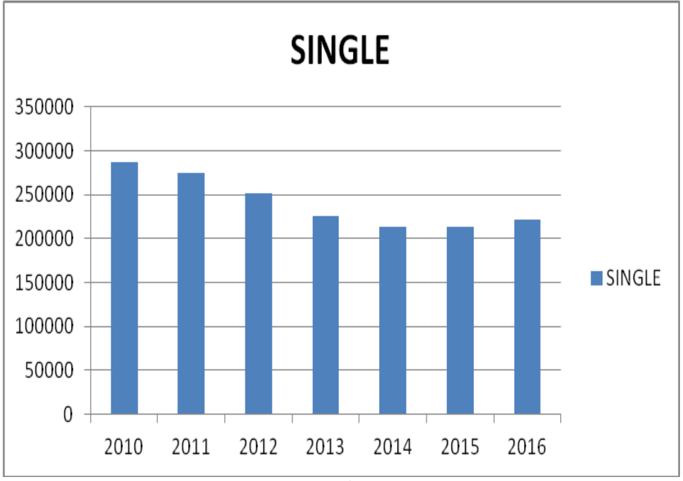

Fig16. Year Vs Size of Single Data Center

They drastically fell from 27532 in 2010 to 15706 in 2016.

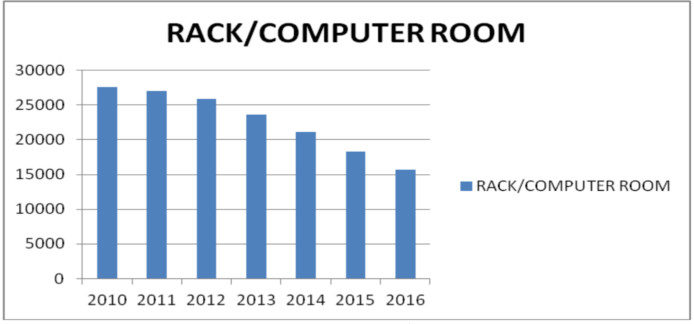

Fig17. Year Vs Size of Rack

The mid size data centers also observed a sudden fall from 380 in 2010 to 236 in 2016.

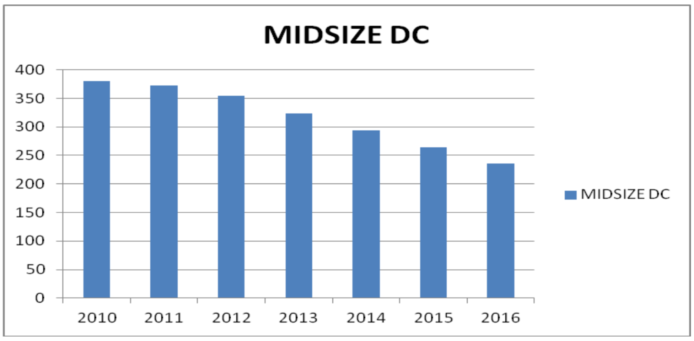

Fig18. Year Vs Mid-Size Data Center

They fell from 346 in 2010 to 292 in 2016.

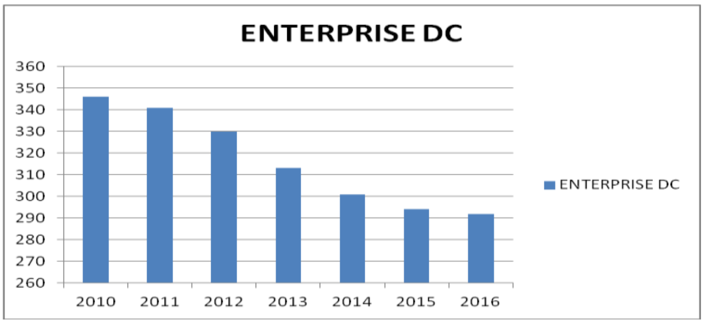

Fig19. Year Vs Enterprise Data Center

They increased from 85 in 2010 to 101 by 2016.

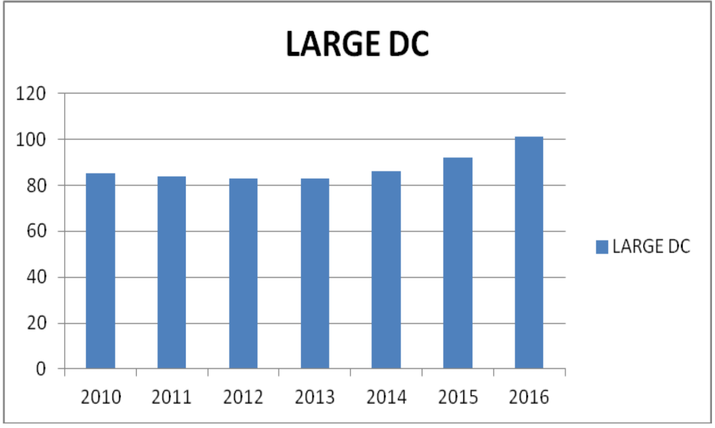

Fig20. Year Vs Large Data Center

IV. CONCLUSIONS

Continuous monitoring and performance measurement would be an efficient way to manage the resources. Measuring losses along the electrical power chain equipments such as transformers, UPS etc., would be considered to be a more detailed assessment. Supply of air temperature and humidity should be monitored for each CRAC or CRAH unit as well as the dehumidification/humidification status to ensure that integrated control of these units is successful. Normally, relocating or consolidating, or even optimizing data centers is a massive task. There is a significant risk that the processes will drag on for years to come, and what is even more disturbing is the fact that the task will never finish, leaving us with more data centers than we started with. One way of avoiding this is to implement fit-for-purpose key performance indicators (KPIs) that support business case, focusing on measuring activities and progress that will eventually allow you to turn off the switches in the old data centers.

REFERENCES

[1] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri , D. A. Maltz, P. Patel , and S. Sengupta, "VL2: A scalable and flexible data center network," in SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on Data communication, 2009.

[2] N. F. N. H. P. M. S. R. V. S. Radhika Niranjan Mysore, Andreas Pamboris and A. Vahdat, "PortLand: A scalable fault-tolerant layer 2 data center network fabric ," in SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on Data communication, 2009.

[3] W. J. Dally and B. Towles., " Principles and practices of interconnection networks," in Morgan Kaufmann Publishers, 2004.

[4] J. I. McGill and G. J. V. Ryzin. Revenue Management: Research Overview and Prospects. Transportation Science, 33(2):233- 256, 1999.

[5] Andrieux, A., Berry, D., Garibaldi, J., Jarvis, S., MacLaren, J., Ouelhadj, D., Snelling, D.: Open issues in grid scheduling. UK e-Science Report UKeS-2004-03, April 2004

[6] Hovestadt, M., Kao, 0., Keller, A. , Streit, A.: Scheduling in hpc resource management systems: Queuing vs. planning. Proceedings of the Job Scheduling Strategies for Parallel Processing (JSSPP) (2003)

[7] VMware Capacity Planner,.. *http://www.vmware.Comlproducts!capacity-*planner/."

[8]IBM WebSphere CloudBurst, *http://www-Ol.ibm.comlsoftware/*web servers/cloudburstl."

[9] Novell PlateSpin Recon, .. *http://www.novell.comlproducts/recon/.*

[10] Lanamark Suite, .. *http://www.lanamark.coml ...*

[11] M. Isard, V. Prabhakaran, J. Currey, U. Wieder, K. Tal war, and A. Goldberg. Quincy: Fair scheduling for distributed computing clusters. In Proceedings of 22nd ACM Symposium on Operating Systems Principles (SOSP '09), Big Sky, MT, Oct. 2009.

[12] M. Zaharia, D. Borthakur, J. S. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica. Job scheduling for multi-user mapreduce clusters. Technical Report *UCB/EECS-2009-55* , University of California, Berkeley, Apr.2009.

[13] G. Lee, N. Tolia, P. Ranganathan, R. Katz. Topology-Aware Resource Allocation for Data-intensive workloads. In APSys, 2010.

[14] HELLERSTElN, J. L. Google Cluster Data. http://googleresearch.blogspot. *coml20 1010* IIgoogle-cluster-data.html.

[15] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L.Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In Proceedings of the 9th IEEEI ACM CCGRID ' 09, pages 124- 131, Shanghai, China, May 2009.

[16] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage. Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds. In Proceedings of the ACM Conference on Computer and Communications Security, Chicago, 1L, Nov. 2009.

[17] K. Chard, K. Bubendorfer, and P. Komisarczuk, "High Occupancy Resource Allocation for Grid and Cloud Systems, a Study with DRIVE," in Proc. of ACM IntI. Symposium on High Performance Distributed Computing. New York: ACM, 2010, pp. 73-84.

[18] Vazirani, Vijay V. (2003), Approximation Algorithms, Berlin:Springer, ISBN 3540653678

[19] D.K. Friesen, M.A. Langston, Variable si zed bin packing, SIAM J. Comput. 15 (I) (1986) 222-230.

[20] Chandra Chekuri , Sanjeev Khanna, A PT AS for the multiple knapsack problem, Proceedings of the eleventh annual ACM- SIAM symposium on Discrete algorithms, p.213-222, January 09-11 , 2000, San Francisco, California, United States