# On Improving Peak and Off-Peak Energy Consumption Pattern in Cloud Data Centers

Thusoyaone J. Moemi[1], Obeten O. Ekabua[2]

[1] Computer Science Department, North-West University
North-West University, Mafikeng Campus, South Africa
[1]17071100@nuw.ac.za, [2]obeten.ekabua@nwu.ac.za

**Abstract—Using the subscription basis and the pay as-you-go model, cloud computing delivers to consumer's infrastructure, platform and software applications as services. The initiative is to deliver the design of the next generation data centers that enables users to access and deploy applications on-demand from anywhere at competitive cost. Data centers are expensive to maintain and 'unfriendly' to the environment because they require massive amounts of energy during peak and off-peak periods. High carbon emissions in data centers lead to overheating which affects the machines lifetime and reliability. Therefore, to make adequate use of the precious energy resource, it is pertinent to know the amount of energy required per instance in a data center. Consequently, in this research article, we developed and implemented energy efficiency models and optimization algorithms for improving delivery of on-demand energy resources during peak and off-peak periods in a cloud computing environment. This was achieved by developing a load balancing model, called LBVMA model, which supports energy reduction in our data centers. The experimental results show significantly the efficiency derived from reduced energy consumption. The reduction and efficient energy usage (EEU) helps to improve delivery of on-demand energy resources in a cloud computing environment.**

**Keywords—** Cloud Computing, Energy Efficiency, Models, Algorithms, Data Center, Service Delivery.

## I. INTRODUCTION

In recent years, effective service provisioning and delivery in terms of delivering applications over networks through the internet has become the concern of network operators and engineers. For example, broadband has become more widespread and some of the popular applications used today by most internet users in broadband are Facebook, Gmail and twitter. This is because their service levels are acceptable in terms of money and time costs because computer facilities are shared by consumers or customers of the services. One of the other advantages of cloud computing is that, more than one company can share resources or computer facilities in a data center irrespective of where the data center is located [1].

The new features in this cloud computing paradigm are its acquisition model which is based on purchasing of services; its business model which is based on pay for use, its access model which is over the Internet to any device and its technical model which is scalable, elastic, dynamic, multi-tenant, & sharable [2].

Some of the services offered in cloud computing environment are software as a service (SaaS), platform as a service (PaaS) and Infrastructure as a service (IaaS) [3] Fig. 1 shows the relationship between these services in the cloud environment.

Paying for services like hardware, software or platform has been a challenge for companies over the last few years. Cloud computing has helped organizations providing services over the Internet to improve their business models, and relationships with their customers since they don't have to buy software, hardware or platform that is going to be out dated in time, but rather rent the services from a cloud provider for some time. Some of the advantages derived from this new paradigm are those derived from outsourcing while complexity is reduced by shifting some of the company or organizational responsibilities to a cloud provider that is an expert in that specific field [1, 3]. This is also advantageous as it helps organizations to start new products and services with less risk and less expenditure. For example businesses operating in a cloud environment can offer to rent more computing capacity to customers during peak periods.

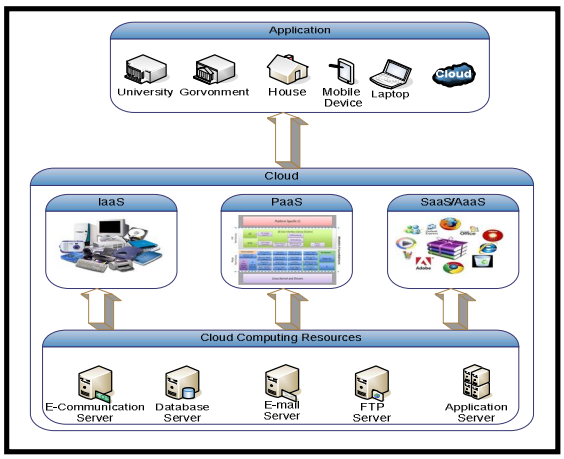


Fig. 1 Cloud Computing Environment

As nice as the idea of operating in a cloud is, the reality is that the services themselves whether they be software, infrastructure or platform are hosted in a data center [4]. A data center is a physical building that contains multiple servers that stores data. A cloud provider is a data center somewhere renting services to cloud consumers and they charge consumers mostly according to how much power they used. Therefore, to stay competitive, cloud providers have to find more efficient ways of consuming power in their data centers. Hence, many data centers are trying to reduce their carbon footprint and power consumption through the implementation of visualization and cloud computing [5]. Consequently, energy efficiency and the reduction of air pollution are challenges while providing cloud computing services [1, 6]. Running a single 300-watt server for a year costs about $ 338, and can emit up to 1,300 kg of carbon dioxide according to Duy et al. [7].

Some challenges to overcome when implementing cloud computing in data centers are power, space, capacity and bandwidth [3, 7]. There is also the risk that closed privately owned and controlled cloud computing architectures could suppress innovations [3].

## II. PREVIOUS WORK

Moreno and Xu [8] presented dynamic resource provisioning mechanisms based on customer utilization patterns, to allocate capacity in real-time cloud data centers. They also analyse the impact of their model in fulfilling and energy efficiency. The goal of Moreno's and Xu's model is to "improve energy efficiency by reducing the waste of resource derived from customers' overestimations." They used empirical methods to compare three over allocation approaches over 24 hours. The approaches are over allocation LAF, over allocation FDF and without over allocation. The authors concluded that their model improves data center utilization as compared to simple DRR approaches.

Justice et al [9] identified energy efficiency metrics for reducing costs and implementation of green initiatives in data centers to be used by IY managers for measuring and maintenance purposes. They examined the strengths and weakness of metrics, PUE and DCP, two of the most commonly used metrics data center metrics. From their findings they concluded that there is a need for standards and metrics in the industry and farther recommended that future metrics should be normalized for all data centers across the industry

Xiaoli and Zhanghui [10] proposed a new model which is energy aware that considers the energy efficiency in cloud computing environments. They improved the Bin Packing algorithm. They used simulation with both C++ and Matlab to analyse their results. Based on their results they concluded that their algorithm makes good use of resources and that fragments of the active server can be used well.

Calheiros et al [11] introduced CloudAnalyst, a novel simulation tool for large scale applications in cloud computing environments. They explained in detail the architecture of the CloudAnalyst simulator and the various algorithms and policies contained in it. They further showed how the simulator can be used in different scenarios. They concluded that CloudAnalyst is not a comprehensive solution for all simulation needs and that their approach and tool will evolve over time.

Wang and Wang [12] proposed a new energy efficient multi-task scheduling model based on Google's framework that processes massive data called MapReduce, to improve energy efficiency of servers. For their model, they created a new practical decoding and encoding method for individuals. They used methods from intelligent systems. They also introduced a local search operator in their algorithm to improve its search ability. The authors used simulation methods to validate their model and from their results they concluded that their model is efficient and effective.

## III. ALGORITHM PROPOSED

Another objective for this research work is to develop energy optimization algorithm. The development of this algorithm requires the development of two different algorithms to effectively reduce the rate of energy consumption in the data centers. These two algorithms are load balancing virtual machine aware algorithm and defragmentation algorithm.

### *A. Load Balancing Virtual Machine Aware Algorithm*

The Load balancing is the process of taking complex or large work load that needs a lot of processing power to be processed and dividing it into modules then distributing it to different machines or notes for processing. In so doing, the processing time and processing power are reduced. For example, taking a large mathematical equation and using a distributed system to compute it just like in Grid computing. In cloud computing, the process is the same, but the only difference is that the process is done on a virtual plain, which is at virtual machine management or hypervisor level. In this work we determine which load balancing algorithm is more energy efficient in a virtual machine management level. Virtual machine management level is referred to as virtual machine queues in our designed model as can be seen in

There are different types of clouds in the Internet, as cloud computing is using computing resources over a network as a service. Hence, our LBVMA model is shows how different clouds and cloud users interact through the Internet. More importantly, our implemented cloud configuration includes a virtual machine queue which manages the distribution of loads to different virtual machines.

Increasing workload in the model shows that the cloud customers is requesting for services in the cloud through the use of the underlying network, the Internet. In our cloud configuration, we have a Service Broker Policy which ensures that the cloud customers and cloud providers understand Service Level Agreements (SLA) among themselves before they proceed with the transaction of money and services. All the services whether they be SaaS, PaaS or IaaS are all hosted in Data Centers, which is what follows the Service Broker Policy. Data Centers have servers that host the services and

most importantly the virtual machines. So, just before distributing the cloud customer requests to the virtual machines, there are virtual machine queues that are managed by load balancers. The load balancers use different load balancing algorithms to distribute work loads to the virtual machines.

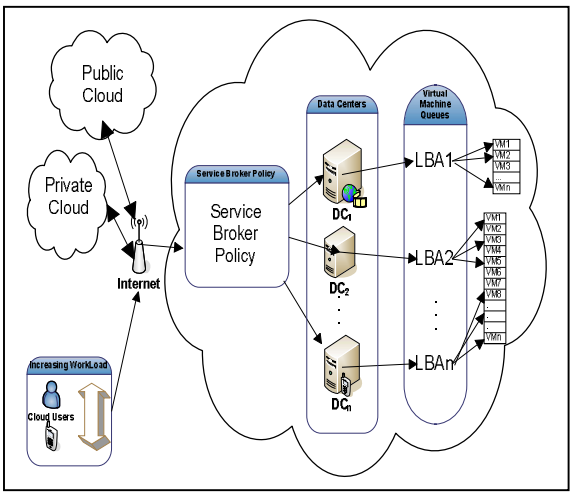


Fig. 2 (Load Balancing Virtual Machine Aware) LBVMA Model.

### *B. Defragmentation Algorithm*

No A server is a computer that manages centrally stored data or network communication resources. A server also provides and organizes access to these resources for other computers linked to it.

```
1.          Quicksort(HD,p,t) {
2.          if (p < t) {
3.            q <- Partition(HD,p,t)
4.            Quicksort(HD,p,q)
5.            Quicksort(HD,q+1,t)
   }
}
6.          Partition(HD,p,t)
7.            x <- HD [p]
8.            i <- p-1
9.            j <- r+1
10.         while (True) {
11.           repeat
12.                 j <- j-1
13.             until (HD [j] <= x)
14.             repeat
15.                 i <- i+1
16.             until (HD [i] >= x)
17.           if (i<-=""> A[j]
18.             else
19.                 return(j)
   }
}
```

Fig. 3 Defragmentation Algorithm

Storage devices in most servers are a collection of hard drives or hard discs. As people or nodes connected to the server in the network insert and delete into the server, the hard drives get fragmented. This fragmentation causes processing speed to be slow because the Central Processing Unit (CPU) collects and stores its data and information in the hard drive or Random Access Memory (RAM).

The reason for is that, the hard drive after fragmentation has variable distance between data or information stored in it in terms of space in bytes. Defragmentation is a process of removing this distance and bringing information in the hard drive closer together, and in so doing making processing time faster because searching time is reduced. Virtual machine migration is when virtual machines are moved from one physical machine to another. This migration also causes fragmentation. So to solve this problem we developed and implemented a defragmentation algorithm, shown in Fig. 3, which should be active after virtual machine migration, just before the server turns off, to optimize processing time and energy consumption.

The algorithm is based on the assumption that the hard drives found in servers of the data centers are designed like an array. So, basically the defragmentation algorithm is a quicksort algorithm. The reason for choosing the quicksort algorithm is that it has a worst case performance of O($n^2$) and a best case performance of O($n$ log $n$) in terms of time complexity. From line one to line five is the main function of the quicksort algorithm which contains a pivoting element and three functions, one to sort the left hand side of the pivot and the other to sort the right hand side of the pivot. The last function is to reposition the pivot during the sorting process. Line 6 to line 14 shows how and in what conditions the function that will reposition the pivoting element will do so. Line 15 to 19 is recursive of the process. Therefore, we have input and output as:

**Input:** HD fragmented array HD, a pivot element p and a traversing element t.

**Output:** HD defragmented array HD.

## IV. Energy Efficiency Model

Part of the objective for this research article is to develop an energy efficiency model. Therefore, a linear relationship between CPU utilization and electrical power is assumed for our model. For example, say for a given job $j_1$, information of the processing time and the processor utilization is enough to calculate its power consumption. We define the consumption of a resource $r_i$ at any given time as:

$$C_{\mathrm{i}} = \sum_{j=1}^{\mathrm{n}} c_{\mathrm{i,j}} \qquad (1)$$

Where n = number of task running at that time and $c_{i,j}$ is the resource usage of job $j_j$.

We also define energy usage, Pi, of a resource at any time as:

$$P_{\mathrm{i}} = (P_{\max} - P_{\min}) * C_{\mathrm{i}} + P_{\min} \qquad (2)$$

Where $P_{\max}$ refers to the peak load consumed and $P_{\min}$ refers to active mode minimum power consumption usually as low as one (1%) percent.

## V. ENERGY EFFICIENCY MODEL

For the purpose of our experiments, we use CloudAnalyst to simulate our data. Cloud Analyst is a simulation tool designed to simulate real cloud environments and scenarios. It is built on CloudSim and designed on a java programming language and iText 2.1.5.

On the other hand, cloud analyst has all the capabilities of CloudSim with a user friendly Graphic User Interface (GUI) [10, 13]. The experiment was run on a machine having core *i*5 intel processor and 4Gig RAM. The simulation tool used was Cloud Analyst. The experimental design map is as shown in Fig. 4.

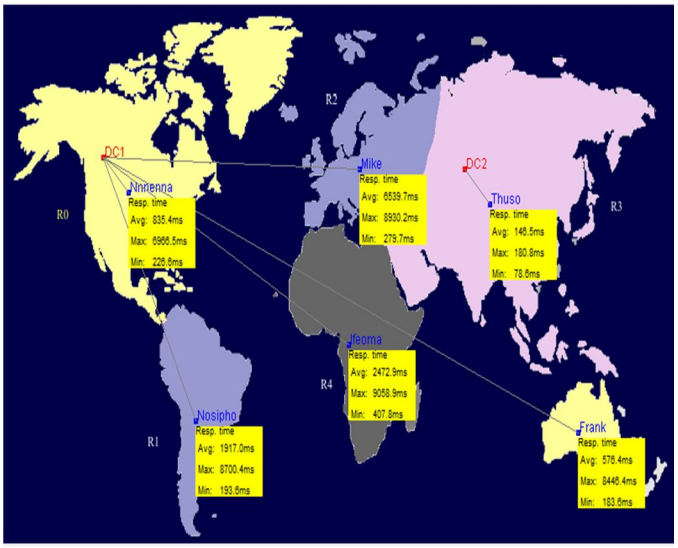


Fig. 4 Region Showing Map

Six geographically located user bases were created and two data centers as shown in Table I and Table II respectfully.

TABLE I

USER BASES

| Name | Region | Requests per User per Hr | Data Size per Request (bytes) | Peak Hours Start (GMT) | Peak Hours End (GMT) | Avg Peak Users | Avg Off-Peak Users |
|---|---|---|---|---|---|---|---|
| Ifeoma | 4 | 20 | 150 | 15 | 20 | 300000 | 80000 |
| Mike | 2 | 20 | 150 | 12 | 24 | 450000 | 300000 |
| Nnnenna | 0 | 20 | 150 | 14 | 18 | 250000 | 20000 |
| Nosipho | 1 | 20 | 150 | 16 | 20 | 500000 | 60000 |
| Thuso | 3 | 20 | 150 | 13 | 22 | 350000 | 5000 |

Shown in Table I is a collection of input parameters that form parts of managing user bases and these input parameters are manipulated to produce the desired outcome. As you scroll up or down, one input parameter is not visible. The output results obtained from these input parameters are reported and discussed in Table V of section 6.

TABLE II

DATA CENTERS

| Data Center | # VMs | Image Size | Memory | BW |
|---|---|---|---|---|
| DC1 | 5 | 10000 | 512 | 1000 |
| DC2 | 5 | 10000 | 512 | 1000 |

Shown in Table II is a collection of input parameters that form parts of managing user bases and these input parameters are manipulated to produce the desired outcome. The output results obtained from these input parameters are reported and discussed in Table VI of section 6.

As important information, for each data center, the physical machines uses x86 architecture and runs in a Linux operating system and Xen virtual machine manager. Each physical machine has four processors and their speed is 10 000Hz. Table III and Table IV shows the characteristics for the delay and bandwidth matrix.

TABLE III

DELAY MATRIX

| Region\Region | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 25 | 100 | 150 | 250 | 250 | 100 |
| 1 | 100 | 25 | 250 | 500 | 350 | 200 |
| 2 | 150 | 250 | 25 | 150 | 150 | 200 |
| 3 | 250 | 500 | 150 | 25 | 500 | 500 |
| 4 | 250 | 350 | 150 | 500 | 25 | 500 |
| 5 | 100 | 200 | 200 | 500 | 500 | 25 |

Table III shows the delay matrix between the different geographically located regions as shown in Fig. 4. At different regions, we varied the input of bandwidth data to be transmitted. The varied input is as shown in Table III.

TABLE IV

BANDWIDTH MATRIX

| Region\Region | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 2,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| 1 | 1,000 | 800 | 1,000 | 1,000 | 1,000 | 1,000 |
| 2 | 1,000 | 1,000 | 2,500 | 1,000 | 1,000 | 1,000 |
| 3 | 1,000 | 1,000 | 1,000 | 1,500 | 1,000 | 1,000 |
| 4 | 1,000 | 1,000 | 1,000 | 1,000 | 500 | 1,000 |
| 5 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 2,000 |

Table IV shows the input bandwidth matrix between the different geographically located regions as Fig. 4. At different regions, we varied the input of bandwidth data to be transmitted. The varied input is as shown in Table IV.

## VI. EXPERIMENTAL RESULTS

Using the develop energy efficiency model and the various inputs from the user bases, the delay and bandwidth matrix, the simulator generates the user base response time from the two data centers. The generated result is as shown in Table V. The output contains the minimum, maximum and average user base time response in terms of mille seconds.

TABLE V

USER BASE RESPONSE TIME

| User Bases | Min (ms) | Max (ms) | Avg (ms) |
|---|---|---|---|
| Frank | 172.83 | 1718.89 | 461.95 |
| Ifeoma | 401.4 | 7688.57 | 1526.57 |
| Mike | 254.16 | 7304.98 | 2464.76 |
| Nnnenna | 49.23 | 736.05 | 292.74 |
| Nosipho | 160.75 | 453.54 | 848.1 |
| Thuso | 50.43 | 5461.83 | 370.14 |

The output produced after running the configuration in Tables I, II, III and IV in terms of response time is shown in Table VI as follows

TABLE VI

DATA CENTER RESPONSE TIME

| Data Centers | Min (ms) | Max (ms) | Avg (ms) |
|---|---|---|---|
| DC1 | 0.11 | 7076.86 | 1831.7 |
| DC2 | 8.71 | 3295.29 | 1106.04 |

## A. *Discussions*

When the amount of data processed at user base level as shown in Table I is compared to the amount of user base response time shown in Table V and Fig. 5, it is evident that quality of service in terms of response time is much better for data centers that have more physical machines.

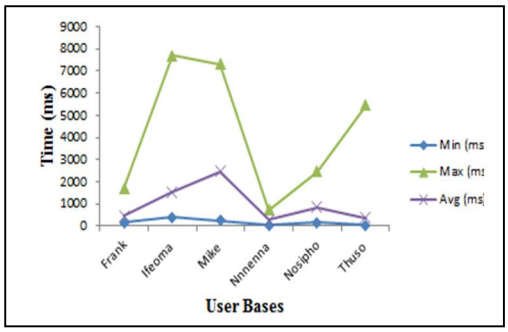


Fig. 5 User Base Response Times

Fig. 5 shows a correlation of user bases and their corresponding response time, which is used to determine the min, max and avg response time for the two (DC1, DC2) data centers reported in Table VI with the results obtained, it was possible to graphically show the min, max and avg response time for the data centers as shown in Fig. 6. The user base response time is an indication of the quality of service provided and the data is obtained from experimentation result in CloudAnalyst.

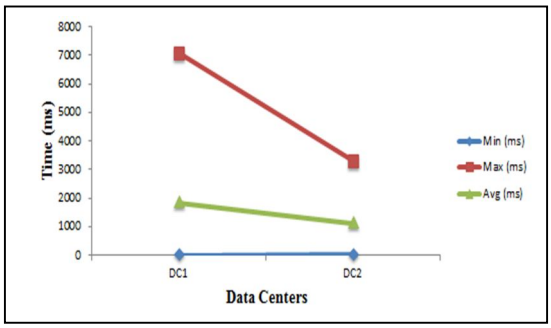


Fig. 6 Data Center Response Times

Using the same configuration in Table V and Table VII, Fig. 7 shows a comparison of energy consumed by a data center facility using the LBVMA Model, where the load balancing policy is throttled and service broker policy is set to optimal time response. Therefore Fig. 7 shows a correlation between execution time and the energy consumed. From LBVMA model and the chosen parameters, it can be seen that energy consumption is less.

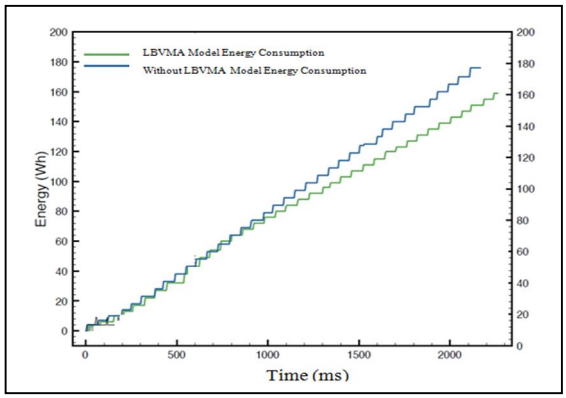


Fig. 7 Consumed Energy for Power Model

Fig. 8 shows power consumption due to deferent memory configurations. The memory configurations are 1000 MHz, 1200 MHz and 1500 MHz respectfully. The x-axis shows time in mille seconds and the y-axis shows the power in kilo Watts. From the figure it is clear to see that memory configuration at higher frequencies consume more power.

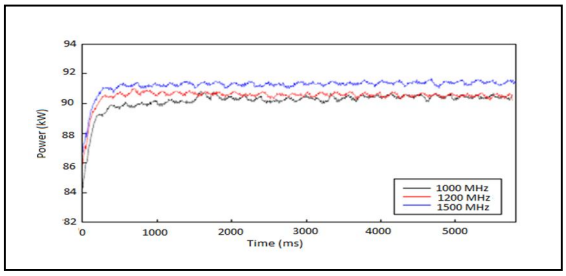

Fig. 8 Power Consumed by Memory

Recall that, in section 4 we mentioned we were going to determine which load balancing algorithm is more energy efficient in virtual machine management level. This next experiment shows exactly that. We used the same configuration in Table V and Table VII for all algorithms.

TABLE VII

OVERALL TIME RESPONSE

| | Round Robin | Equally Spread Current Execution Load | Throttled |
|---|---|---|---|
| **Avg (ms)** | 3739.52 | 3613.4 | 1996.66 |
| **Min (ms)** | 75.28 | 72.77 | 49.23 |
| **Max (ms)** | 7673.36 | 7721.21 | 7688.57 |

Table VII shows overall response time of requests processed from user bases to data centers and vice versa. Fig. 9 shows a graphical representation of the data in Table VII and represent a correlation of the load balancing algorithm with respect to response time represented in mille seconds. From Fig. 9 it is clear that the throttled load balancing algorithm performs better in terms of response time.

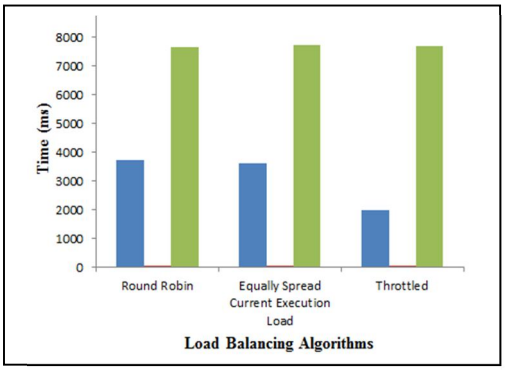


Fig. 9 Overall Time Response

## VII. CONCLUSIONS

While it is important to understand how to minimize energy consumption in data centers that host cloud computing services, it is also important to consider the energy required for migrating data to and from the end user and the energy consumed by the end-user interface. What is reported in this research is about the energy consumed in cloud computing data centers. Consequently, we developed a power or energy efficiency model, a load balancing virtual machine aware (LBVMA) model and an efficient energy usage (EEU) metric to enhance the calculation of power consumption in data centers and to determine whether or not the energy used in a data center is used efficiently.

More so, we developed and implemented a defragmentation algorithm as an optimization algorithm to optimize processing time in cloud data centers after virtual machine migration. From the experimental results, the quality of service in terms of response time is much better for data centers that have more physical machines than for those with less machines, but there was an observable higher energy consumption for memory configuration with higher frequencies. We can therefore conclude that, the energy consumed by a cloud service is directly proportional to the type of services it provides, the number of users serves, and the usage pattens of those users.

## VIII. FUTURE WORK

As a future application in order to improve or buttress the assertion and validity of our concept, we shall improve the existing algorithm code. This is primarily to accommodate the use of more algorithms since in its present state; both the Service Broker and Load balancing modules in CloudAnalyst have only three algorithms respectively. Moreover, because the research conducted here did not consider the application of genetic algorithms and neural networks, as a future work, we shall attempt these implementations.

## IX. ACKNOWLEDGEMENTS

We wish to acknowledge the support of our sponsors, Telkom Center of Excellence and Thrip for their support and the North-West University for access to resources needed to complete this work

## X. REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.