Analyzing Utilization Rates in Data Centers for Optimizing Energy Management

Michael Pawlish
Montclair State University
Dept. of Environmental Mgmt.
Montclair, NJ, USA
pawlishm1@mail.montclair.edu

Aparna S. Varde
Montclair State University
Dept. of Computer Science
Montclair, NJ, USA
vardea@mail.montclair.edu

Stefan A. Robila
Montclair State University
Dept. of Computer Science
Montclair, NJ 07043 USA
robilas@mail.montclair.edu

***Abstract*—In this paper, we explore academic data center utilization rates from an energy management perspective with the broader goal of providing decision support for green computing. The utilization rate is defined as the overall extent to which data center servers are being used and is usually recorded as a percentage. Recent literature states that utilization rates at many data centers are quite low resulting in poor usage of resources such as energy and labor. Based on our study we attribute these lower utilization rates to not fully taking advantage of virtualization and cloud technology. This paper describes our research including energy data analysis with our proposed equations for performance measurement and forecasting, corroborated by evaluation with real data in a university setting. We suggest that future data centers will need to increase their utilization rates and thus shift more towards the cloud in order to lower costs and increase services despite current concerns for security of cloud technology.**

Keywords - *Cloud, Data Centers, Forecasting, Green IT, Utilization Rates*

I. INTRODUCTION

Data centers have experienced rapid growth due the demands of society for faster and uninterrupted service. Within only five years (from 2000 to 2005) the energy usage due to data centers has doubled constituting 1.5-2% of the world's energy consumption [6]. While many analysts and researchers were predicting another similar increase to follow this growth, due to the economic downturn as well as increase in energy efficiency of the computing systems, and increased use of virtualization and cloud computing this did not materialize [1, 4, 6, 8]. While reduced, the energy usage still continued to grow at a 12% rate [6].

IEE International Green Computing Conference,
June 5-8, 2012, San Jose, CA, USA.

Perhaps a greater factor in the slowing down of the expansion of data centers was attributed to innovation in the form of virtualization and cloud technology development [5,6]. Virtualization is the ability to harness the power of many servers for numerous applications whereas prior to virtualization one or a few applications were assigned to each server. This frame of mind was a major factor that leads to the traditionally low utilization rates that wasted resources by having servers running without processing applications to the full potential of the equipment. A second innovation that greatly increased utilization rates was the development and maturing of cloud technology. The shift to the cloud encourages economies of scale by running servers on a virtual platform that has higher utilization rates and these large cloud providers tend to be located in areas with lower energy costs. Typically the server farms for large technology firms such as Google, Yahoo, Amazon and Facebook have been locating their servers in areas where the electricity is provided from cleaner sources such as hydro-electric power [13].

All these factors motivate our work on analyzing utilization rates in data centers with a view to providing better energy management. It fits into our broader goal of developing a decision support system for green data centers, a step towards achieving sustainable solutions for a greener planet. We provide a theoretical perspective on utilization rates, conduct data analysis by proposing relevant equations that serve as performance metrics for energy management helpful in forecasting, and summarize our preliminary evaluation of utilization rates with real data. We envisage future data centers using cloud technology to achieve greener solutions.

II. A THEORETICAL PERSPECTIVE

A. Historical Trends in Utilization Rates

The current literature on data centers reports that an area of improvement is the low utilization rates on servers. Recent research reported utilization rates

below 25% suggesting that servers could be switched to idle for most of the time [2, 3, 10]. This situation results in poor usage with respect to two areas:

- Most energy consumed is not used for a productive purpose. This means that large amounts of unnecessary carbon dioxide are released into the air if coal is used to generate the power.
- From a broad perspective, a large amount of natural resources are being tied up in wasted resources.

B. Optimizing Performance of Data Centers

As stated earlier, data centers are extremely energy intensive. When additional servers are added the two operational costs that will increase are cooling and power costs. A common problem in data center operations is to add more servers without fully taking advantage of the potential computing power on existing servers, this trend is what we call *server sprawl*. The policy of optimizing servers with greater utilization rates will reduce server sprawl and have the following advantages:

- Lower electricity usage that translates to cost savings and a reduced carbon footprint
- Decrease management costs to maintain and service additional servers
- Free up floor space on inefficient or phantom servers
- Provide greater efficiency in the use of resources, e.g., less server sprawl equates to lower cooling costs, decreased electrical bills and less use of natural resources to build additional servers

Given this theoretical discussion, we now proceed to conduct analysis by proposing equations that pertain to various aspects of data center utilization rates.

III. ANALYSIS WITH EQUATIONS

We consider important issues in data center management and propose equations that can be used for analysis. The equations we formulate here serve as performance metrics for various aspects that are important for energy management and forecasting.

A. Equation for Utilization Rate

We first state an equation (Equation 1) to determine the utilization rate based on its definition as:

$$Utilization\ Rate = \frac{\sum_{n=1}^{T}(\text{CPU Rate})}{\text{T}} \qquad (1)$$

In this equation the CPU rate is the extent to which the CPU is busy at any given instance of time. The utilization rate is thus calculated in this formula as an efficiency ratio that sums up each instance of the CPU rate over a total time span T and divides by the value of T. Utilization rate gives management an idea of how much of the time the data center is being used. *It is desirable to maximize the utilization rate to enhance performance*.

B. Equations for Cost per Server

An important metric that we propose from an energy management perspective relates to a breakdown of the costs from a per server basis. We assert that by examining the data center from a per server basis we can gain further insight into performance analysis. Accordingly, we put forth a proposition to analyze the cost per server, which we define in terms of the following five components:

1. Air conditioning cost per server, $P_{cooling}$, is defined as a metric that can be used to estimate the air conditioning cost per each additional server added to the data center where:

$$P_{cooling} = \frac{\sum_{n=1}^{T}\text{cooling costs}}{\sum_{n=1}^{T}\text{servers}} \qquad (2)$$

   Here, $\sum$ cooling costs is the air conditioning power usage to cool the data center over a time span T and $\sum$ servers gives the total number of servers over that time span. *Since this metric represents an additional expense, is important to lower the cooling cost in order to get better performance*.
2. The cost of running all the servers can be calculated on a yearly basis from historic records from the Power Distribution Units. We propose that the energy cost per server denoted as P_{server} provides a general cost structure for adding each additional server where:

$$P_{server} = \frac{\sum_{n=1}^{T}\text{server energy costs}}{\sum_{n=1}^{T}\text{servers}} \qquad (3)$$

   In this equation $\sum$ server energy costs denotes the total electrical cost for the data center for time T and $\sum$ servers again represents the total number of servers. *Hence, we argue that the energy cost per server needs to be reduced to maximize efficiency and thus enhance performance*.
3. Another metric we formulate is the administration cost per server denoted as A_{staff} which is attributed to running the data center. This provides a cost structure on a per server basis depending on the number of staff members working on the data center where:

$$A_{staff} = \frac{\sum_{n=1}^{T}\text{administration costs}}{\sum_{n=1}^{T}\text{servers}} \qquad (4)$$

In this metric ∑ administration costs depicts the total cost for supporting the management and staff attributed to the data center over the time period T and ∑ servers, as in the other equations above, represents the total number of servers. *It is certainly advisable to keep these administrative costs low as a step towards achieving better performance. In other words, it is advisable to reduce the number of staff members if possible.*

4. An important performance metric is also the fixed cost related to basic utilities such as rent, heat and water needed to run the data center. This is denoted as F_{fixed} and provides a general idea of the total fixed annual cost per server where:

$$F_{fixed} = \frac{\sum_{n=1}^{T} \text{fixed costs}}{\sum_{n=1}^{T} \text{servers}} \quad (5)$$

Here ∑ fixed costs gives the total fixed cost for the data center for time T (and ∑servers gives the total number of servers). *Wherever possible, the fixed costs should also be minimized to serve as a positive indicator of performance.*

5. Yet another significant performance metric that we propose is the replacement costs per server denoted as R_{server}. This relates to the cost associated with replacing a server. It examines the fixed cost per year for owning the individual server where:

$$R_{server} = N_{server}/L_{span} \quad (6)$$

In this equation, N_{server} is the average cost of a new server while L_{span} represents the estimated product life span of the new server. It can be seen that this performance metric is a little different from the other metrics pertaining to server costs. *While replacing servers is important, it is obviously desirable to minimize the replacement costs for performance enhancement. Thus, it is desirable to lower the costs of new servers and try to obtain replacement servers with greater life spans.*

Considering all the cost components as formulated above, the total cost per server, C_{server} is calculated as a summation of these five components. This is an important performance metric given as:

$$C_{server} = P_{cooling} + P_{server} + A_{staff} + F_{fixed} + R_{server} \quad (7)$$

Since this cost is a summation of the individual costs, needless to say, it important to reduce this as a step towards performance enhancement. We can therefore keep track of this combined metric C_{server} *as a general indicator of performance.*

We claim that by gaining the full cost on a per server basis, data center managers can better determine the real cost of the addition or the retirement of the individual server. This is very useful.

IV. PRELIMINARY EVALUATION

We conducted the evaluation with real data from a large data center on our campus which represents a fairly typical academic setting. This is the campus of Montclair State University, the second largest public school in the state of New Jersey, located approximately 15 miles from New York City. The results we present can be applied to other campuses with suitable situations.

A. Data Collection

A big challenge in the process of data collection was acquiring access to real data given various privacy and administrative concerns. To deal with it, we had to go through the formalities of obtaining permissions from several authorities within the university and explain to them the critical need for this real data in order to conduct effective analysis and provide green solutions that would be useful to the campus. We are also in the process of trying to acquire such data from external sources which would further strengthen our work. Another major challenge was to monitor the data continuously which presented logistic and cost issues. For example, the cost of the real time monitoring of energy usage by the PDUs (Power Distribution Units) and wireless monitoring of temperature and relative humidity that we originally considered ordering to monitor was $35,000 U.S. dollars. This was not an expense we could feasibly cover through our research grant and nor did we consider this a cost-effective method to gather data. This is given the fact that we are in the process of proposing our entire strategy as a green computing solution and the means of data acquisition has to be reasonable. In other words, we cannot suggest to an organization that if they wish to adopt greener solutions, they must first incur a huge capital expense for purchase of additional equipment.

B. Observations from Data Center Servers

Next we present one example of the calculation for the Utilization Rate, and in Table 1 we present the summary of Utilization Rates for the first week of classes. Using the data from figure 1 we calculate the average Utilization Rate on January 16th, 2012 for Server 1 host. Recall from our fundamental Equation 1, that the utilization rate is the summation of the

CPU rate for each minute divided by the total number of minutes for the time span considered. We consider this over a time span of 1 day which is 1440 minutes. Also note that the CPU rate is recorded every minute which gives 1440 data points that need to be summed up. Thus, for example on our Server 1 on 1/16/12, the 1440 data points that give the CPU rate per minute are summed up to get 41,433 and this number is divided by 1440 to give a utilization rate of 28%. Table 1 gives a broader picture of utilization rates for the second week of January 2012. An important observation is that of the 14 utilization rates presented, there were only four times when the utilization rate exceeded 50%.

Table 1. Utilization Rates

	Server 1		Server 2	
Day	*∑ CPU Rate*	*Utilization Rate*	*∑ CPU Rate*	*Utilization Rate*
Sun.	24,656	17%	125,505	87%
Mon.	41,433	28%	40,261	28%
Tues.	97,177	67%	52,393	36%
Wed.	74,832	52%	47,583	33%
Thur.	124,959	87%	45,867	32%
Fri.	39,503	27%	51,328	36%
Sat.	22,005	15%	46,981	33%

C. Forecasting with Performance Metrics

We now present further evaluation for our MSU data center using some of the performance metrics we proposed earlier in equations 2 through 7. We use estimated values here serving to forecast the performance of our data center. Consider first the air conditioning costs. As we observed above, the estimated electrical power usage for air conditioning is 1,524,240 kWh/year. An estimated cost of electricity in the state of New Jersey is approximately 0.14 cents/kWh. Our data center has approximately 500 servers and hence the following would be the forecasted as the air conditioning or cooling cost per server for a time span of 1 year using Equation 2.

$$P_{cooling} = (1{,}524{,}240 \text{ kWh/ year} \times 0.14 \text{ cents/kWh}) / 500$$
$$= \$426 \text{ per server}$$

Similarly, the electrical power usage on a per server basis can be forecasted considering the same estimated cost of electricity with total electrical power usage obtained from monitoring the four PDUs which equals 888,516kWh/year. Given 500 servers in our data center, we get the following from Equation 3.

$$P_{server} = (888{,}516 \text{ kWh/year} \times 0.14 \text{ cents/kWh}) / 500$$
$$= \$249 \text{ per server}$$

We have not shown the administration and fixed costs per server since we are in the process of acquiring data or estimates on these. Finally, to forecast the replacement costs per server, R_{server}, we use an estimated cost of a new server to be approximately $3000 US dollars and an estimated life span of 5 years. With these values in Equation 6, we get:

$$R_{server} = C_{server}/L_{span}$$
$$= \$3000/5$$
$$= \$600 \text{ per year}$$

These forecasted values indicate that our data center is functioning fairly well but there is considerable scope for improvement in performance with respect to various factors. We claim that it is important to enhance utilization rates and minimize carbon footprint emissions in order to strive for a greener environment. In the future, we thus propose to consider cloud based solutions as a step towards achieving our goal.

V. SHIFT TOWARDS THE CLOUD

Based on our analysis with preliminary evaluation and a study of the literature, we offer several suggestions for optimizing utilization rates in the future. Our main suggestion here is that shifting applications from the data center to a cloud provider does offer the potential to raise utilization rates as long as older servers are decommissioned or retired. The cloud is increasingly viewed by management as an offsite back up, a solution to process extra demand in activity, and an area to place some of the redundant applications. Therefore we claim that data centers of the future will incorporate more aspects of cloud computing.

A. Concerns with Cloud Computing

Currently, the main challenge is for management to become comfortable with running sensitive data over the cloud. While cloud providers argue that their data centers are more secure than the traditional data center, there have been instances of security breaches in cloud providers, and undoubtedly will be future risks due to bugs and hackers. Below are some of the main concerns of shifting to the cloud that should be considered prior to moving data to the cloud.

- Security of data is the most often cited concern by management and prior to shifting to the cloud the

legal department should be consulted for possible breaches of security.
- Cloud provider continuity is an important concern in the event that the provider is acquired, merged or faces insolvency.
- Data lock-in is a real concern where the cloud provider has a unique format on the data that makes transferring the data challenging or costly.
- There would be a cost to reconfigure the data center system and retire older servers.

B. Advantages of the Cloud in Data Centers

There are some clear drivers for shifting some data center applications to the cloud to moderate spikes in utilization rates of traditional data centers. Some advantages are listed here.

- Flexibility is achieved by not having to worry about either over provisioning or under provisioning for services or user demand.
- Redundancy will always be critical to running a data center; however data centers can be refigured to include hosting and back up provided by the cloud.
- A relative low cost structure for computing and storage when compared to traditional data centers.
- The "pay as you go" or metered cost structure that allows for purchasing of actual computing time.

C. Trends in Cloud Technology

Cloud technology is a paradigm shift in that it enables the ability to put together massive computer infrastructure on demand [7]. The cloud can be viewed as a disruptive technology due to the tremendous impact it will have on the Information Technology sector. Many researchers are predicting that the cloud will be the next utility in the sense that an organization will pay for computing and storage capacity similar to an electric or other utility bill. The shift of resources will cause adjustment of labor markets as smaller data centers are eliminated.

Companies such as Greenqloud, Iceland [14] and CloudSigma, Switzerland [15] have implemented free cooling strategies, used renewable geothermal and hydropower energy, and adopted carbon neutral policies as steps towards greenness. They claim that due to such factors their cloud technologies are among the greenest in the world, as per GPUE (Green Power Usage Effectiveness) indicators.

An important advancement in cloud technology is the use of virtualization that allows for efficient management of resources, but presents a challenge for the proper metering when implementing virtualization [5]. Currently, virtualization technology has significantly raised utilization rates; however there are challenges when assigning costs due to inadequate metering. We believe that this present challenge will be adequately solved in the near future.

A further development in cloud technology is the shift of placing more backup and storage on the cloud which saves on maintenance costs while providing offsite storage [11]. Disaster recovery is a main driver for organizations to backup or store data on a cloud provider. Combined with the economies of scale offered by cloud providers in the sense of labor and energy costs, we believe that there will be a natural shift by the market towards cloud technology.

Also, a growing trend is the push for private clouds that provide the benefits of cloud technology while still maintaining control over security of data [12]. The issue of whether to use private or public clouds remains debatable as they both offer significant advantages but also represent trade-offs with respect to issues such as cost and security.

D. Discussion

From a managerial perspective we have shown that the utilization rates are low in our data center at MSU, and from the current literature, utilization rates are quite low on many other data centers throughout the world [2, 3, 10]. The result is that many data centers are over built and that these data centers are running at idle speed the majority of the time. From our analysis, our data center seems to be nearing capacity when scheduled batch processing jobs are being completed during the evening hours. We believe that resources could be better optimized by turning towards a cloud provider, especially when running the batch processing jobs.

The challenge is that these batch processing jobs contain sensitive data such as a student's information and other important data. And while many organizations have moved sensitive data to the cloud, our organization and probably many mid to large sized organizations have been hesitant to shift due to the perceived or real threats to moving to the cloud. Many applications such as payroll, email and customer relations software are already deployed on cloud platforms. In the future as the cloud continues to enjoy such economies of scale in energy and labor costs over traditional data centers, management will be forced to shrink the data center size while incorporating spikes in demand for services through a cloud provider.

VI. CONCLUSIONS

In this paper we investigated data center management taking into account utilization rates and related factors with the goal of providing energy efficiency for enhancing performance. We provided a theoretical discussion on utilization rates and proposed suitable equations for performance measurement that are useful in forecasting and decision support. These developments will be incorporated into our decision support system for data centers which is being developed as an outcome of this whole effort.

The paper includes preliminary evaluation with real data from our university data center and summarizes our findings. Our initial hypothesis that servers were being underutilized has been confirmed. Ongoing work includes continuing to monitor the servers on our campus and also trying to get external data while seeking ways to optimize utilization rates.

In summary, a few ways in which organizations can seek higher utilization rates are the following;

- Provide better information to the concerned offices of information technology running data centers in the form of current utilization rates that can be used for enhanced decision making
- Implement virtualization technology to place more applications on less servers
- Encourage management to explore options of using cloud technology
- More effectively schedule batch processing jobs to better utilize the data center's resources

We envision the future of data centers for mid to large sized organizations to incorporate a gradual shift to the cloud. As the information technology industry works towards solving and proving cloud technology, management will be more comfortable to shift to the cloud. From a cost structure or economic perspective, operational costs in the form of labor and energy for cloud technology are lower than traditional data centers and will result in a shift towards optimizing resource management. This work has the broader impact of developing sustainable solutions for a greener planet.

VII. ACKNOWLEDGMENTS

This research has been supported by a grant from the PSE&G Company. We would like to thank PSE&G for their encouragement and support.

VIII. REFERENCES

[1] Anderson, S. (2010) Improving Data Center Efficiency, *Energy Engineering*, Vol. 107, No. 5, pg 42-63.

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. (2010) A View of Cloud Computing, *Communications of the ACM*, April, Vol. 53, No. 4.

[3] Forge, S. (2007) Powering Down; Remedies for unsustainable ICT, *Foresight*, Vol. 9, No. 4, pg 3-21.

[4] Judge, J., Pouchet, J., Ekbote, A., And Dixit, S. (2008) Reducing Data Center Energy Consumption, *ASHRAE Journal*, November, pg 14-26.

[5] Kansal, A., Zhao, F., Liu, J. Kothari, N. and Bhattacharya, A.A, (2010) Virtual machine power metering

[6] and provisioning. ACM SOCC, pg. 39-50

[7] Koomey, J. (2011) *Growth in Data Center Electricity Use 2005 to 2010*, Analytics Press, Oakland, CA.

[8] Pedersen, T.B. (2010) Research challenges for cloud intelligence: invited talk. EDBT/ICDT Workshops 2010.

[9] Ruth, S. (2009) Green IT-More than a 3 Percent Solution? *IEEE Internet Computing*, July/August, pg 74-78.

[10] Ruth, S. (2011) Reducing ICT-related Carbon Emissions: An Examplar for Global Energy Policy? *IETE Technical Review*, Vol. 28, Issue 3, May-June.

[11] Siegele, L. (2008) Let It Rise: A Special Report on Corporate IT. *The Economist*, October.

[12] Taft, D.K. (2011, December 5). *IBM's Top 12 Tech trends for 2012*. [Online]. Available: http://www.eweek.com/c/a/Cloud-Computing/IBMs-Top-12-Tech-Trends-for-2012-Include-Cloud-Analytics-Mobile-221458/

[13] Robles, L. (2011, November 29). *Four Trends that Shaped Cloud Computing in 2011*. [Online]. Available: http://venturebeat.com/2011/11/29/four-trends-that-shaped-cloud-computing-in-2011/

[14] Cook, G. & Van Horn J. (2011, May 24). *How dirty is your data?* [Online]. Available: http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/

[15] Greenqloud (2012, January 26). *Greenqloud* [Online]. Available: http://www.greenqloud.com/

[16] CloudSigma (2012, January 26). *Why choose cloud servers from CloudSigma?* [Online]. Available: http://www.cloudsigma.com/