A Call for Energy Efficiency in Data Centers

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

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

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

Anand Ranganathan
IBM Thomas J. Watson Research Center,
Hawthorne, NY USA
arangana@us.ibm.com

Abstract

In this paper, we explore a data center's performance with a call for energy efficiency through green computing. Some performance metrics we examine in data centers are server energy usage, Power Usage Effectiveness and utilization rate, i.e., the extent to which data center servers are being used. Recent literature indicates that utilization rates at many internal data centers are quite low, resulting in poor usage of resources such as energy and materials. Based on our study, we attribute these low utilization rates to not fully taking advantage of virtualization, and not retiring phantom (unused) servers. This paper describes our initiative corroborated with real data in a university setting. We suggest that future data centers will need to increase their utilization rates for better energy efficiency, and moving towards a cloud provider would help. However, we argue that neither a pure in-house data center or cloud model is the best solution. Instead we recommend, from a decision support perspective, a hybrid model in data center management to lower costs and increase services, while also providing greater energy efficiency.

Keywords: Cloud, Data Centers, Green IT, Utilization Rates, Energy Efficiency

1. INTRODUCTION

The data center is the backbone of the Internet that has provided tremendous communication gains; however, at the same time energy efficiency in data centers is often a secondary concern. The management of data centers is increasingly becoming more complex from dealing with legacy equipment, developments in technology such as blade servers and virtualization, and the present push to outsource much of the data center through cloud providers; all while top management has been keeping budgets level or seeking cuts. Traditionally, energy efficiency has therefore not been a top priority with data center managers, due to the aforementioned challenges of operating a data center. In this paper, we claim that following a hybrid business model that takes advantage of cloud technologies and the existing in-house data center will assist in developing a more effective strategy for energy efficiency.

There are a number of reasons to seek energy efficiency in computing facilities. First, in many places in the world energy consumption is increasing at a faster rate than new energy sources are being developed. In the United States, there is a tremendous push back by the public to any type of new large-scale energy production facilities. This push back results in delays in construction of new facilities, and according to supply and demand should result in future elevated energy costs due to the increasing demand [15].

A second reason for seeking data center energy efficiency is the pure economics of squeezing out inefficiencies in current systems [2]. In the rush to build data centers in the first decade of the 21st century, energy efficiency had a low priority. Now that the market has matured, there is a need to find gains such as low hanging fruit, for example, increasing the temperature in the data center or placing the lighting on motion detectors. By making data centers more efficient, or lowering the cost and environmental impact, management will see improvement in their operating costs.

A third reason to pursue an energy efficiency strategy is to keep current with emerging technology advances. For example, virtualization that allows more applications to run on fewer servers is an important technological development from an energy efficiency perspective [3, 12]. Virtualization has allowed the retirement of a number of servers, or basically has permitted more processing power to be computed with less electrical consumption [7]. Servers are therefore continuing to be built that are smaller and more powerful from previous generations.

Finally, another reason to seek out energy efficiency is public perception. In a recent cover issue of the Sunday *New York Times*, the data center industry was presented as the next wasteful and polluting industry of the 21st century [5]. This perception of the Information Age is contrary to the positive reputation that many individuals hold towards the Internet, and the article exposed many efficiency problems, including particularly the low utilization rate in data centers that is addressed in this paper.

This paper presents a detailed analysis over a three-year period of energy usage, and documents the low utilization rate in a mid-size university data center similar to a typical computing facility described by previously published literature [1, 8, 11]. Data mining techniques such as Case Based Reasoning (CBR) and decision trees are provided as approaches for decision support in the management of the center. Results from the analysis and mining support the arguments in favor of a hybrid data center. Here, existing local capacity is combined with an outside cloud provider as the most efficient strategy to pursue for enhanced service, low cost, and a more energy efficient model.

2. PARAMETERS IN DATA ANALYSIS

We obtained our data from our university data center, typical of most organizational data centers, in that the servers are not homogeneous. As characteristic of most in-house data centers, legacy equipment is the norm with differing vintages of servers and cooling components. Sampling was conducted manually by visiting the data center and recording energy usage over a three-month period during the spring semester of each year for three years. The purpose of documenting server energy usage was to establish a base line study, and document the carbon emissions. We focus on certain parameters for analysis as described next.

Table 1. Server Energy Usage

Date	PDU 1 kWh	PDU 2 kWh	PDU 3 kWh	PDU 4 kWh	Total
3/01/10-6/01/10	66,598	46,838	90,527	80,382	284,345
3/01/11-6/01/11	50,680	36,093	85,994	75,381	248,148
3/01/12-6/01/12	40,433	26,061	86,615	78,547	231,656

2.1 Utilization Rate

The utilization rate if defined as the extent to which the CPU is busy at any given instance of time, as stated in the Equation 1 herewith:

$$U = \frac{\sum_{n=1}^{T}(CPU\ Rate)}{T} \quad \text{--- 1}$$

Here U represents the utilization rate calculated as an efficiency ratio that sums up each instance of the CPU rate over a total time span T, such that CPU rate is the extent to which the CPU is busy at a given instance of time. Utilization rate gives management an idea of how much the data center is being used, and can be expressed as a percentage. Based on this, it is clear that it is desirable to increase the utilization rate for energy efficiency.

1) *Observations from A Data Center Host*

We consider a data center with two hosts that continually shift user demand for optimal performance. As an example we hereby present utilization rate calculation for a single day. The CPU rate per minute is emailed to us in a file based on continuous monitoring of data center hosts. We sum up this CPU rate and divide it by the total number of minutes per day to get the daily utilization rate.

Host 1-Thursday 6/14/12
$\sum$ *CPU Rate = 49,350*
Utilization Rate = 49,350/ 1440 =34%

Based on such calculations, Tables 2 and 3 give a broader picture of utilization rates for the first six months of 2012.

Table 2. Utilization Rates

2012	Host 1		
Month	*Average Utilization rate*	*Monthly low*	*Monthly high*
Jan.	*38%*	*7%*	*86%*
Feb.	*34%*	*10%*	*85%*
March	*30%*	*7%*	*60%*
April	*35%*	*8%*	*68%*
May	*35%*	*10%*	*63%*
June	*29%*	*9%*	*60%*

Table 3. Utilization Rates

2012	Host 2		
Month	*Average Utilization rate*	*Monthly low*	*Monthly high*
Jan.	*42%*	*20%*	*86%*
Feb.	*35%*	*25%*	*90%*
March	*38%*	*21%*	*87%*
April	*35%*	*9%*	*82%*
May	*38%*	*21%*	*84%*
June	*42%*	*18%*	*90%*

An initial observation is that average utilization rates are around 30% to 42%, which we believe is on the low side. To enhance energy efficiency, our argument is that data centers need to operate at higher utilization rates than these. From an economic perspective the cost of running data centers, as per our analysis, is that the data center is running at an optimal operation point only around 1/3 of the time. This is an apparent waste of resources that unnecessarily contributes to carbon emissions when fossil fuels are used for generating the required electricity. After examining the utilization rate, we delve further into our case study through a metric called Power Usage Effectiveness as explained next.

2.2 Power Usage Effectiveness

The Power Usage Effectiveness (PUE) is an efficiency ratio of data centers that was developed by the industry, and is defined in the following Equation 2:

$$PUE = \frac{Total\ Facility\ Power}{IT\ Equipment\ Power} \quad \text{--- 2}$$

In theory, if the PUE equaled 1.0, the data center would be considered perfectly efficient since Total Facility Power would equal IT Equipment Power. In reality or practice, a PUE slightly above 1.0 has been observed in some ultra efficient data centers, for example, Facebook's Prineville data center located in Oregon. Presently, a PUE of around 2.0 seems to be the industry average since there is power lost in Total Facility Power for energy use by such components as lighting and cooling. Using such measures as efficient design factors, for example, airside economizing (free cooling) that uses outside air to lower the data center room's temperature, and therefore uses less power than traditional air-conditioning is a typical method to lower the PUE and the energy usage. There also seems to be a growing trend of locating data centers in higher latitudes to take advantage of the cooler climates. One such example has been the growth trend in data centers in Sweden, due to such factors as a stable government with cheap electricity that is derived from hydropower that does not contribute to carbon dioxide emissions.

In the fall of 2013, the second phase of our study was initiated with the installation of meters to measure the energy consumption of the data center. Due to relatively large PUE values observed and considering the installation of temperature/relative humidity sensors, a future research question would focus on how to lower the PUE. A next step will be to raise the temperature in the data center by 2 degrees Fahrenheit. The research team feels confident with the sensors in place to prevent hot spotting, and the team is curious of the savings in the carbon footprint and electricity cost. The ultra efficient cloud data centers are able to operate with a PUE slightly above 1.0 and that further supports our argument that hybrid computing is more energy efficient as discussed later in this paper. In the next sub-section, the carbon footprint of the data center is analyzed in order to assess its carbon dioxide emissions.

2.3 Carbon Footprint

From an energy management perspective, perhaps the most important parameter is the carbon footprint of an organization that represents the atmospheric carbon dioxide emissions that directly correlates with energy usage. More specifically, the carbon footprint of an organization is the estimated total of the output of carbon dioxide released in the atmosphere from primarily burning fossil fuels to supply the power for operations. In this case, we refer to the operations of the data center. Currently, the estimated amount of CO_2 released from data centers worldwide is approximately 2% which is a growing concern [4]. The standard formula to calculate the carbon footprint is given in Equation 3 as follows:

$$C = \frac{E*N}{T} \quad \text{--- 3}$$

Where C represents carbon footprint, E represents electrical usage in kWh per year, N represents national CO_2 emissions, and T represents metric tons (1 metric ton equals 2,204.6 lbs.) In our evaluation, we have recorded the energy usage of our data center servers and calculated the carbon footprint using the given formula. These values are summarized for a three month period in Table 1. Based on this, the total carbon footprint for data center servers at our university is calculated per year as stated in Equation 4 below. Consider that:

$$E_{year}=E_{sample}*4 \quad \text{--- 4}$$

Where E_{year} represents total yearly energy used in 2012, E_{sample} represents a sample of the total energy consumption over the three month period. The results are thus as follows for the energy usage of the servers in 2012.

$$E_{2012}=231{,}656\ kWh * 4 = 926{,}624\ kWh$$

The carbon footprint for the servers C_S is therefore calculated using Equation 1, considering N = 1.34 lbs/kWh as the national average of US CO_2 emissions [13].

$$C_S = 926{,}624\ kWh * 1.34\ lbs/kWh * 1\ metric\ ton/2{,}204.6lbs = 563\ metric\ tons/year$$

A metric ton conversion ratio is used because CO_2 emissions are commonly expressed in the international community in metric tons. Now consider the carbon footprint for cooling or air conditioning. The estimated electrical usage is 58 kW per hour with three air conditioning units running 7 days a week, and 365 days per year. The electrical power usage for air conditioning is 1,524,240 kWh/year. Thus, for example, the total carbon footprint for air conditioning C_{AC} in our data center is calculated as:

$$C_{AC} = 1{,}524{,}240\ kWh/year * 1.34\ lbs/kWh * 1\ metric\ ton/2{,}204.6lbs = 926\ metric\ tons/year$$

From Table 1 and the air conditioning power usage calculation presented above, we also obtain the combined power usage for 2012 including data center servers and air conditioning. This is calculated as 926,656 kWh/year (servers) + 1,524,240 kWh/year (cooling) = 2,450,864 kWh/year. Therefore, based on our measurements and estimations, our data center is contributing approximately 1,500 metric tons per year of CO_2 into the atmosphere that is not a good indicator. Especially considering that due to low utilization rates presented in the next section of this paper, the majority of the time CO_2 emissions are being wasted on idle servers and the concerned cooling.

Given this analysis of parameters, we now consider case based reasoning and decision trees in addressing the problem of energy efficiency in data centers.

3. DEPLOYMENT OF CASE BASED REASONING

The data-mining paradigm of Case Based Reasoning (CBR) has been deployed in our work. CBR discovers knowledge from previous cases or examples and uses that for reasoning about other similar cases in the future. A typical CBR model uses the R4 cycle: Retrieve, Reuse, Revise and Retain. In R4, we retrieve a similar past case, reuse it to fit the current scenario as far as possible, revise it using methods in the field of "adaption in CBR", and then retain the adapted learned case as for future cases.

In our study we use CBR in various examples, one of which is shown in Figure 1. In this example, we examine the case where there is inefficient use of energy in data centers. Following Figure 1 in a clockwise rotation based on the R4 cycle yields a four step process as follows. The first step in this cycle retrieves relevant information pertaining to the potential to lower CO_2 and energy usage by $1/50^{th}$ by shifting email operations to a cloud provider. This estimation is calculated by considering that this data center has approximately 50 data racks, and the student email system takes up about one full rack. (Note that the employee and faculty email were not outsourced earlier due to legislation and privacy issues). The second step in the R4 cycle involves reusing the information that recommends the use of higher energy efficiency in cloud providers that will result in more efficient resource use. The third step in the cycle is to revise the case with the recommendation of our main argument for a hybrid model. This suggests using the existing data center through higher in-house server utilization, plus backup provided by a third party cloud company. A hybrid model will resize existing data centers, and shift spikes in demand to an outside cloud provider. The final step in the CBR cycle is to retain the new knowledge for the future as the learned case. This places an emphasis in continual data center management that measures and monitors metrics such as server sprawl, energy usage and utilization rates, while using a portfolio management approach to determining which applications are candidates for a cloud provider.

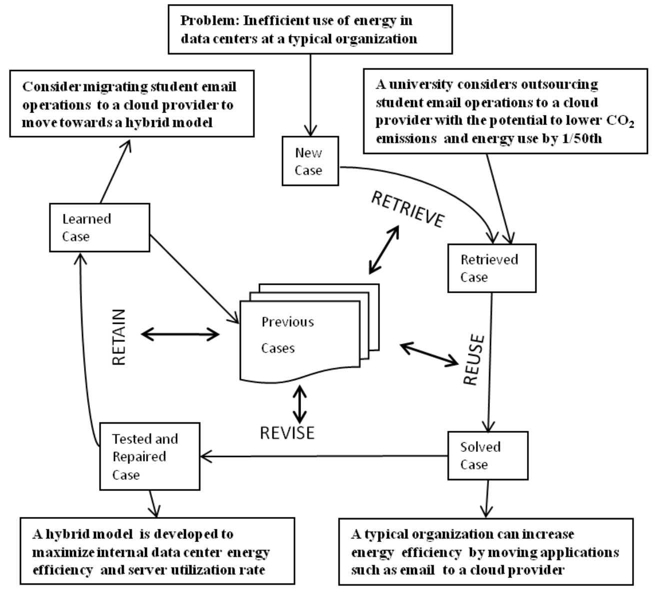

Figure 1: CBR for Energy Efficiency in a University Data Center

Based on our CBR model where $1/50^{th}$ of the electricity and resulting carbon emissions could be transferred to a cloud provider by outsourcing the student email, the question remains if the cloud provider could be more energy efficient than the internal data center. If there was greater energy efficiency in a cloud provider by utilizing, for example, hydro-electric power or by utilizing resources more efficiently, there would be a net benefit. The equation for this translation of the net carbon benefit would be the following, i.e. Equation 5:

$$C_{Benefit}=1/50*1500\ CO_2\ tons=30\ CO_2\ tons \quad ---5$$

The net carbon benefit, $C_{Benefit}$ would result in $1/50^{th}$ of 1,500 calculated metric tons of CO_2 from our data center which translates to approximately 30 metric tons per year of CO_2 savings (minus the addition of any CO_2 from the cloud provider). Currently, the data on a rack level or server basis is not provided by cloud companies, and we realize that our argument is based on the assumption that cloud providers are more resource-efficient, since that is a key operating goal of cloud providers. However, in all scenarios this may not be the case, e.g., when the cloud provider is using fossil fuels as an energy source.

4. ANALYSIS WITH DECISION TREES

While CBR examines specific cases, a decision tree follows a logical path on more of a general problem. Thus, decision trees have a specific starting point and flow through a series of questions to a recommended strategy. In the decision tree in Figure 2, the starting point examines whether the PUE is greater than 2.5, which is set as a baseline. This is because it has been found from our discussions with data center personnel that industry standards for PUE are usually below this number. Energy usage is increasingly becoming an important factor for management to measure in order to achieve a more energy efficient data center, and the PUE is an efficiency ratio of energy use.

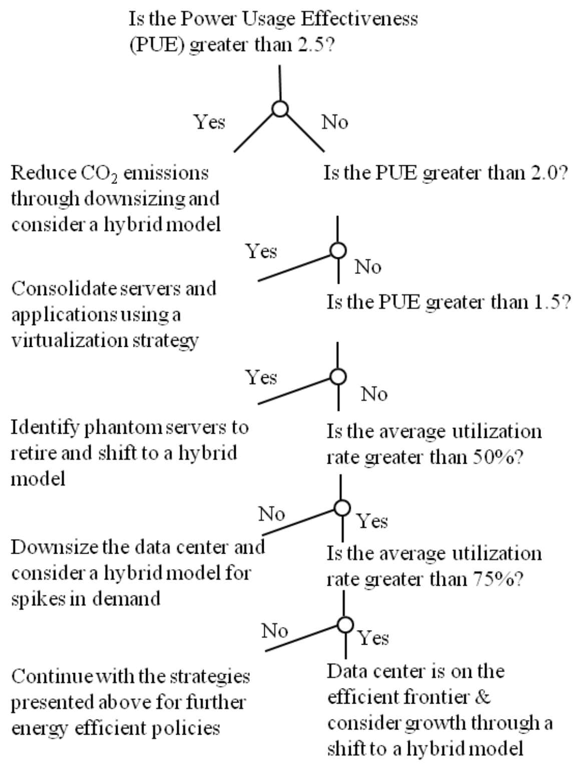

Figure 2: Decision Tree Examining PUE & Utilization Rate

As illustrated in this decision tree, a PUE of 2.5 was selected for initial comparison since currently most in-house data centers are operating at a higher level than cloud or external data centers. Next, the decision tree moves on to follow paths to achieve greater energy efficiency, with a second step to see if a virtualization strategy has been developed to reduce servers by moving more applications to fewer servers. Traditionally in the past, the general rule of thumb was to have one application per server, but this has proved to be costly and inefficient from a natural resource perspective. The consolidation of applications to fewer servers is a first step in a series of solutions that can be implemented simultaneously with other strategies, such as retiring phantom servers. Examples of phantom servers are servers that are still in operation that are not completing useful work, that were typically left on from previous administrators.

The decision tree moves on from the PUE analysis to analyze the utilization rate in blocks of 25% higher utilization rates. In each decision, further strategies are identified while optimizing the data center towards a hybrid strategy. The more efficient data centers will operate at around a 75% utilization rate with applications and spikes in user demand shifted to an external data center, i.e., cloud.

5. PROPOSAL FOR A HYBRID MODEL

Increasing utilization rates and lowering the PUE in data centers for enhanced energy efficiency is important for lowering the carbon footprint of organizations. In addition we propose a new paradigm of running a data center on a hybrid model as presented in the CBR example in Figure 3. The model begins with the first step of reducing the number of servers by 25%, through shifting spikes in demand to a cloud provider in step two. The goal in the third step is to increase the utilization rate to 70-80%, with the final objective achieved by greater virtualization and lowering the number of physical servers.

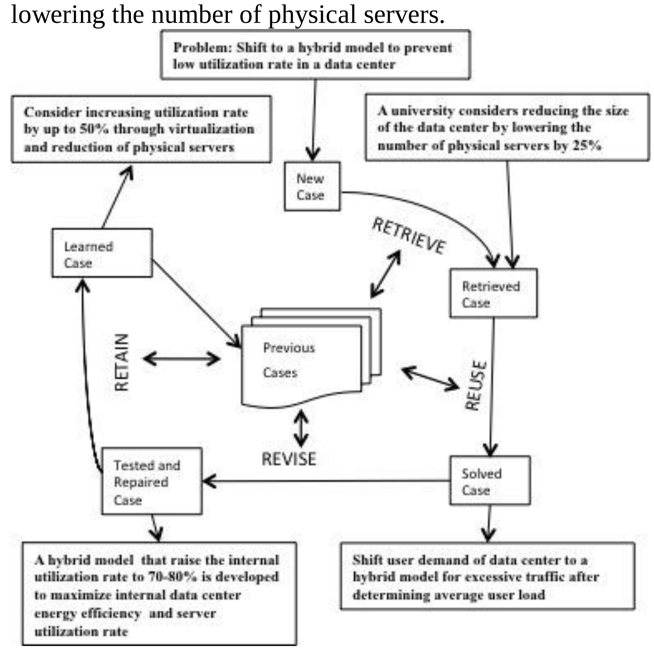

Figure 3: CBR for Shifting to a Hybrid Model

The reasons for using a hybrid model are the following based on our analysis between the trade-offs of an in-house data center and an external data center, i.e. cloud:

- Due to economies of scale most cloud providers can operate with a lower PUE and a higher utilization rate through having data centers geographically distributed.
- The PUE of in-house data centers tend to be higher than PUEs found in a typical external or cloud data center.
- An argument for keeping an in-house data center essentially boils down to security and privacy issues that industry and society will continue to develop into the future. For example, if health care records were kept on the cloud it may be more efficient, but people would be concerned that insurance

companies could obtain their records and deny coverage.

- Resistance to change by personnel is yet another issue. For example, data center managers are familiar with SQL based packages in a traditional database setting. Migrating to the cloud and using packages such as Hadoop/Hive could involve additional training.
- Our final argument is that current in-house data centers are overbuilt, since these data centers have been designed for peak usage. Typically, peak usage only occurs a few days of a year such as at the end of an accounting period, or during peak shopping seasons.

Therefore, while the usage of cloud computing is presently debated, we believe that cloud computing presents the next large wave in information technology. The economies of scale of cloud computing has brought forth an age where it is no longer necessary to provision computing needs for the future combined with elastic demand while all being instantaneous. In many cases, these benefits of the cloud outweigh the fixed costs of owning expensive capital and the operational costs of internal data centers depending on the organization. One of the most important factors is the flexibility provided by cloud computing which could lead to a competitive advantage in organizations depending on implementation of strategy.

We thus put forth a proposition that the answer for mid to large size organizations is a hybrid model of operating a data center, and we present the idea in both bullet point and a decision tree format. To transition to a hybrid model, we recommend four strategies as stated below:

1. First determine the rate of growth of the data center. To accomplish this energy usage needs to be recorded. For example, in the decision tree in Figure 4 an arbitrary number of 5% growth is selected, and each organization can select a goal to contain its energy usage accordingly.
2. Phase out 25 to 50% of servers due to low utilization rates, depending on organizational goals and objectives. The objective would be to match average utilization rates per month with actual server usage. Once again, in cases of excess demand an outside cloud provider would be secured.
3. Develop a data center strategy of keeping mission critical information on local servers and down size the data center by shifting non-critical information or applications to a cloud provider. To provide for back up in the local data center, a secure strategy would call for a cloud provider to additionally provide support for mission critical data.
4. Shift to public applications that are run on the cloud. For example, many applications such as payroll, human resource management, email, and customer service management are now provided by cloud software. We believe that this trend of cloud-based software will be the future technology that will have implications on the local data center by decreasing demand on present operations. We suggest that operations involving high security and privacy issues be retained on the internal data center servers.

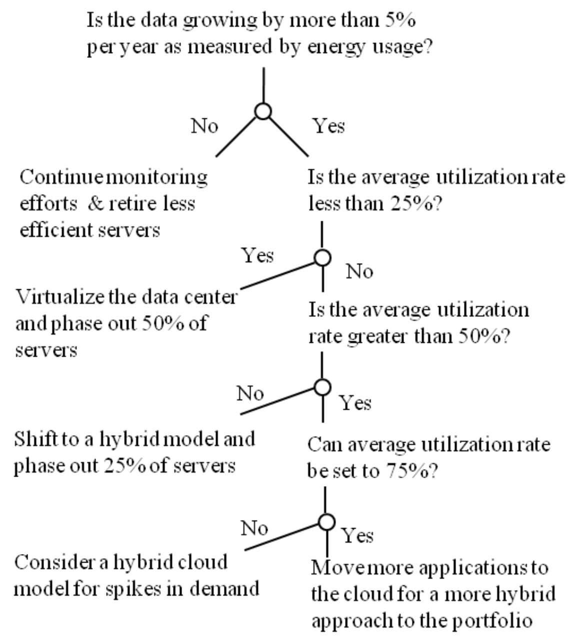

Figure 4: Decision Tree for Moving To a Hybrid Model

A more robust hybrid model as we envision it, would combine all four of the above mentioned strategies, and as we also envision future internal data centers operating at higher average utilization levels of 70 to 80% with spikes in demand and redundancy for backup supported by a cloud provider such as Amazon, Rackspace, Microsoft, Google or similar. These companies have cloud facilities that are geographically diverse while shifting demand to operate at higher utilization rates that support more efficient energy management. These commercial cloud providers generally do not make available information on their energy use or utilization rate performance due to releasing strategic information to competitors, but it would be expected that these cloud providers would be efficiently operating their facilities to reduce such factors as server sprawl, and increase such factors of virtualization, since that is their main operational goal.

6. CONCLUSIONS AND FUTURE WORK

From our analysis of a typical data center, utilization rates have been documented as operating on the low side, and the literature on this subject also documents other data centers operating with low utilization rates. From a broad perspective this is a societal problem, since resources in the form of energy and materials are being wasted, and the energy used is producing unnecessary carbon dioxide emissions when fossil fuels are the fuel source. To solve this problem, we recommended a shift in thinking of data center operations to a hybrid model with the following advantages:

- A shift to a hybrid model is that existing data centers are more fully developed by gaining higher utilization rates, or in other words the servers are more efficiently run.
- This hybrid strategy would involve the increasing use of virtualization with more applications running on fewer machines.
- The strategy would also rely on cloud providers to provide backup for mission critical operations, as well as providing for increased spikes in user demand.
- Operating the data center from a hybrid model would enhance energy efficiency, and we believe contribute to enhanced use of natural resources.

Finally, from a strategic perspective, the most important characteristic of implementing the cloud is the flexibility gained. The ability to have a variable cost instead of a fixed cost or asset will provide growth for innovation and experimentation on different business models. While it is impossible to predict new businesses that may develop in the future, the ability to be flexible and innovative have proven over time to be successful characteristics of organizational growth.

7. 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.

8. References

[1] Anderson, S. (2010) Improving Data Center Efficiency, *Energy Engineering*, Vol. 107, No. 5, pp. 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., pp. 50-58.

[3] Donnellan, B., Sheridan, C., and Curry, E. (2011) A Capability Maturity Framework for Sustainable Information and Communication Technology, *ITPro*, January/February, pp. 33-40.

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

[5] Glanz, J. (2012) Power, Pollution and the Internet, *New York Times*, Sept. 23, pp A.1.

[6] Greenberg, A., Hamilton, J., Maltz, D., and Patel, P. (2009) The Cost of the Cloud: Research Problems in Data Center Networks, *ACM SIGCOMM Computer Communication Review*, Vol. 39, No. 1, Jan., pp. 68-73.

[7] Kansal, A., Zhao, F., Liu, J., Kothari, N., Bhattacharya, A. (2010) Virtual Machine Power Metering and Provisioning, *Proceedings from ACM Symposium on Cloud Computing*, June 2010, pp. 39-50.

[8] Kapoor, R., Porter, G., Tewari, M., Voelker, G., Vahdat, A. (2012) Chronos: Predictable Low Latency for Data Center Applications, *ACM Symposium on Cloud Computing*, Oct 2012.

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

[10] Pawlish, M. & Varde A. S. (2010) A Decision Support System for Green Data Centers, *ACM CIKM Workshops Program*, Oct 2010, pp. 47-56.

[11] Pawlish, M., Varde, A. S., & Robila, S. A. (2012) Cloud Computing for Environment-Friendly Data Centers, *ACM CIKM Workshops Program*, Oct 2012, pp. 43- 48.

[12] Pedersen, T.B. (2010) Research challenges for cloud intelligence: invited talk, *EDBT Workshops Program*, March 2010, pp. 6:1.

[13] Schulz, G. (2009) *The Green and Virtual Data Center*, Auerbach Publications, Boca Raton, Fl., pg. 11.

[14] Siegele, L. (2008) Let It Rise: A Special Report on Corporate IT. *The Economist*, October, pp 1-14.

[15] World Commission on Environment and Development (1987) *Our Common Future*. Oxford: Oxford University Press.