Fear assessment: Why data center servers should be turned off

Article · January 2014

DOI: 10.3233/978-1-61499-452-7-253

CITATION
1

READS
1,549

5 authors, including:

Damián Fernández Cerero
Universidad de Sevilla
20 PUBLICATIONS **157** CITATIONS
SEE PROFILE

Alejandro Fernández-Montes
Universidad de Sevilla
50 PUBLICATIONS **305** CITATIONS
SEE PROFILE

Luis Gonzalez-Abril
Universidad de Sevilla
247 PUBLICATIONS **1,195** CITATIONS
SEE PROFILE

Juan A. Ortega
Universidad de Sevilla
190 PUBLICATIONS **911** CITATIONS
SEE PROFILE

Pattern Recognition Letters 67 (2015) 59–65

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Energy wasting at internet data centers due to fear☆

Alejandro Fernández-Montes[a,*], Damián Fernández-Cerero[a], Luis González-Abril[b], Juan Antonio Álvarez-García[a], Juan Antonio Ortega[a]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Av. Reina Mercedes s/n, Sevilla, 41012, Spain
[b] Departamento de Econoía Aplicada I, Universidad de Sevilla, Av. Ramón y Cajal s/n, Sevilla, 41018, Spain

ARTICLE INFO

Article history:
Available online 3 July 2015

Keywords:
Costs of fear
Energy efficiency
Data center
Grid computing
Cloud computing

ABSTRACT

The fear experienced by datacenter administrators presents an ongoing problem due to the low percentage of machines that they are willing to switch off in order to save energy. This risk aversion can be assessed from a cognitive system. The purpose of this paper is to demonstrate the extra costs incurred by maintaining all the machines of a data center executing continuously for fear of damaging hardware, degradating the service, or losing data. To this end, an objective function which minimizes energy consumption depending on the number of times that the machines are switched on/off is provided. The risk aversion experienced by these data center administrators can be measured from the percentage of machines that they are willing to switch off. It is shown that it is always the best option to turn off machines in order to reduce costs, given a formulation of the cognitive aspects of the fear experienced by datacenter administrators.

1. Introduction

A data center is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and various security devices. Large data centers are industrial-scale operations that can consume as much electricity as a small town and sometimes constitute a major source of air pollution in the form of diesel exhaust.

The main purpose of a data center is to run applications, perform tasks or store data. The many examples of internet and computing services performed by data centers include:

The spread of cloud and grid computing paradigms has increased the size and usage of data centers; today there are thousands of data centers worldwide, which means millions of machines in total.

The majority of these facilities are located in the USA (about 25% of the total energy consumption of data centers worldwide [20]) and to a lesser extent in Europe. However, large companies such as Google locate a number of their data centers in high latitudes near the north pole to minimize cooling costs, which represent almost 40% of total energy consumption of these infrastructures [1].

Energy consumption by data centers has grown in the past ten years to 1.5% of worldwide energy consumption [25]. Major companies have therefore addressed their energy-efficiency efforts to areas such as cooling [7], hardware scaling [8] and power distribution [9], thereby slowing down the growth in power consumption in these facilities in recent years as we can see in Fig. 1, which shows the latest predictions.

In addition to these areas of work, saving energy by switching on/off machines in grid computing environments has been simulated using various energy efficiency policies, such as turning off every machine whenever possible, and turning off a number of machines depending on workload [10].

Although it has been demonstrated that about 30% of energy can be saved by applying these energy-aware policies [11], big companies still prefer not to adopt such policies due to their potential impact on the hardware, the possibility of damaging machines, and the costs associated with this hardware deterioration.

The purpose of this paper is to compute the costs imposed by the risk aversion experienced by data center administrators on switching off machines, and to show that even when taking these fears into consideration, some servers of the data center should still be turned off to minimize energy consumption and overall costs.

1.1. Cognitive systems modeling emotions

In psychology [33], emotion is a subjective, conscious experience characterized primarily by psycho-physiological expressions, biological reactions, and mental states. It is influenced by hormones and neurotransmitters, such as dopamine, noradrenaline, serotonin, oxytocin, cortisol, and gamma-aminobutyric acid. Furthermore,

☆ This paper has been recommended for acceptance by Lledó Museros.

* Corresponding author. Tel.: +34 954 559 769; fax: +34 954 557 139.
E-mail address: afdez@us.es (A. Fernández-Montes).

http://dx.doi.org/10.1016/j.patrec.2015.06.018
0167-8655/

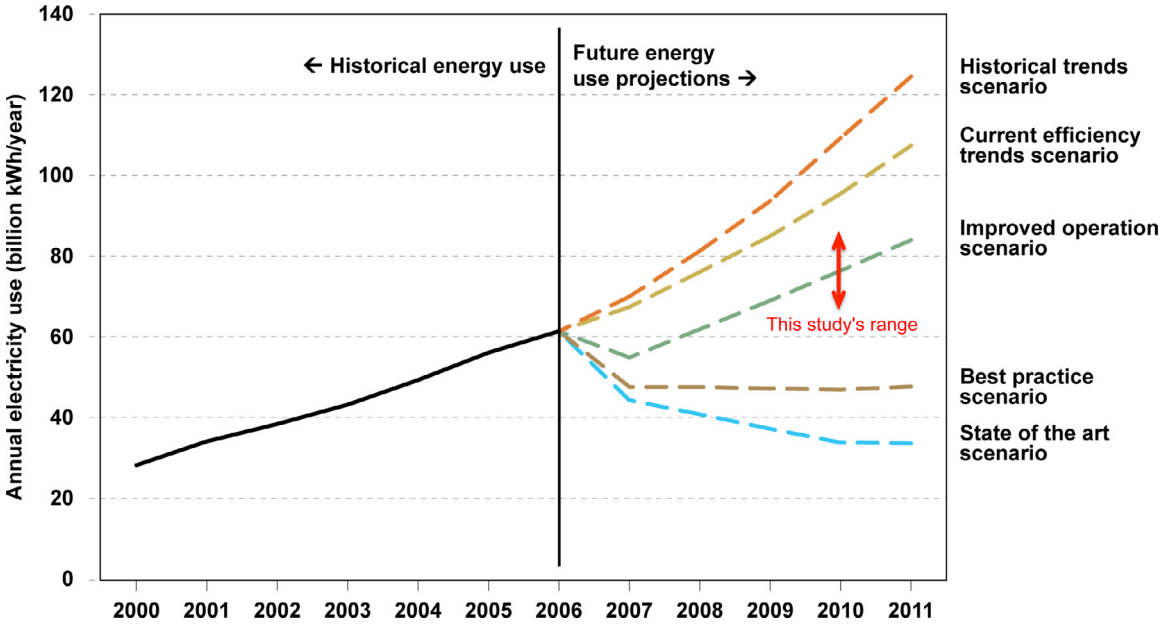

Fig. 1. Data center energy consumption worldwide [24].

neurologists [14] have made progress in demonstrating that emotion is as, or more, important than reason in the process of making decisions. Modeling emotions is a problem tackled from diverse knowledge areas: robot-based systems [6], music [30], videogames and virtual worlds [15] and domain-independent systems [16]. Moreover, emotion recognition systems [18] are on the rise in effective computing research. Data can be obtained from diverse sources: physiological signals (electromyogram, blood pressure, skin conductance, respiration rate and electroencephalogram rate), speech and facial expressions. Focusing on the emotional fear, it appears in response to a specific and immediate danger or a future specific unpleasant event. It can be measured and detected through biosignals such as irregular heart and respiration rate [5,19], visual signals (head gestures, nods and shakes) [17] and facial feature information [34]. Several studies [21] using optogenetic techniques have shown how aversive experiences trigger memories and suggest that combined hebbian and neuromodulatory processes interact to engage associative aversive learning.

Our interest in this paper is to model a function that quantifies the costs of the fear experienced by a datacenter operator on deciding whether a machine must be switched off. According to Michael Tresh, formerly a senior official at Viridity, a company that delivers energy-optimization to data centers: *"Data center operators live in fear of losing their jobs on a daily basis, because the business won't back them up if there's a failure."* The startup 'Power Assure' which is focused on energy management, marketed a technology that enables commercial data centers to safely power down servers when they are not needed, but, as the manager of energy efficiency programs at the utility, Mary Medeiros McEnroe, explains that, even with aggressive programs to entice its major customers to save energy, Silicon Valley Power, a not-for-profit municipal electric utility, failed to persuade a single data center to use that technology. *"It's a nervousness in the I.T. community that something isn't going to be available when they need it"* [13]. Moreover, Power Assure, was dissolved in october 2014. Its technology was based on algorithms that enabled optimal server capacity and application needs to be calculated and to automatically shut off unnecessary capacity or spin up more capacity based on actual application demand. Jennifer Koppy, research director for data center management at International Data Corporation (IDC), said Power Assure's energy management technology was *"extremely forward-looking … they had a superb idea, but I don't think the market is ready yet."*

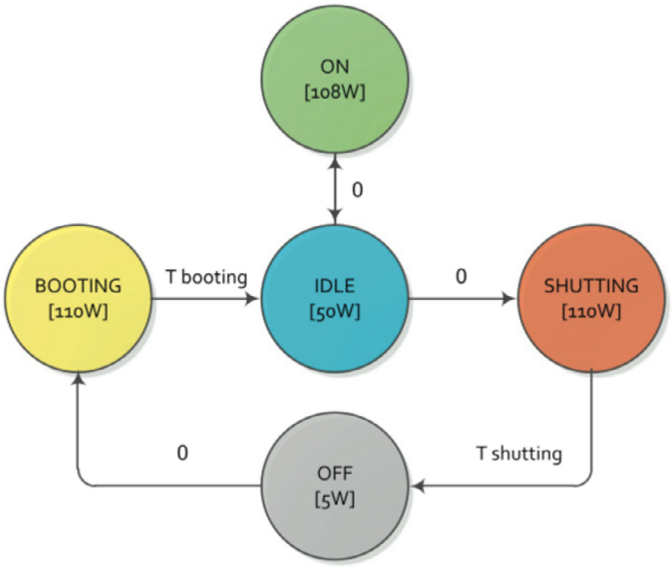

Fig. 2. Life cycle of a data center server [10].

2. Problem analysis

It makes sense that one of the most effective ways to achieve considerable energy savings is to turn off computers that are not being used. Although this idea is generally accepted by users, and hence most personal computers are turned off at night or during periods of low usage, it is seldom implemented in data centers or at enterprise level.

Although the average server utilization within data centers is very low (typically between 10% and 50% [4]), very few companies prefer to turn off the machines that are not in use rather than leaving them in an idle state. While idle servers consume half the energy of those in a state of intensive use [24], this remains a high direct and indirect energy cost due to the increased need for cooling. The several different states through which a machine can pass are shown in Fig. 2. In this state diagram the average power consumption of a common server per CPU in each state is also shown, and the time needed to

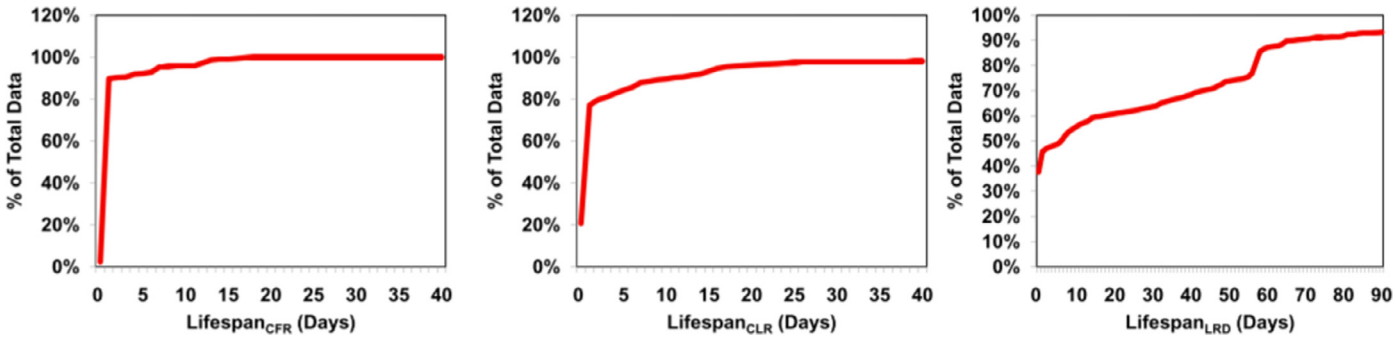

Fig. 3. File access pattern in Yahoo cluster [23].

change from one state to another. Notice that a server with 4 CPUs spends 4 times the energy shown in each state, i. e. 432 Watts*h in ON state.

It has to be noted that very few machines cannot be switched off. Some of the machines from the data center act as master nodes, while the vast majority of machines act as slave nodes which are candidates to be switched off.

The main reasons why IT departments generally prefer to keep machines idle are for fear of:

- **Hardware damage**: It is known that due to a high number of switching on/off cycles, some computer hardware components suffer stress, which can lead to computer deterioration. We incur this as a cost: **The repair cost**. The component that is usually damaged is the hard drive [29], which has other implications besides simply their repair or replacement costs. However, due to the constant improvements of these components and the new SSD hard drives, it can be expected that the failure rate of these pieces of hardware will diminish over time, and therefore these new drives will reduce this type of fear.
- **Service degradation**: When a task needs the service of this damaged computer which can no longer perform a service, in a new cost is incurred due to the worsening in service quality, response times, etc.: **The opportunity cost**. Despite this potential opportunity cost, as we have seen, the server utilization within data centers is very low therefore, in a distributed environment, is highly unlikely that no other machine in the data center can provide the service that this machine was providing.
- **Data loss**: This is a critical issue in a data center infrastructure. If the machine (and its hard drive) that has been damaged was the only one that stored certain data and this data has been lost, certain critical operations could not be performed and it would entail very high operation costs. However, as mentioned above, distributed systems such as data centers typically replicate their data between multiple machines across the data center servers, and therefore, it is highly unlikely for information to be lost. Data loss will only happen if data has just been created and has not had time to be replicated.

Due to these fears experienced by the IT staff from big internet companies, file distribution policies within data centers are designed to minimize the possibility of losing any data, thereby maximizing the availability of data and the available computing capacity to perform tasks associated with it.

These distribution policies do not aim at energy efficiency. To achieve this energy efficiency, data center managers rely on hardware systems that work by: switching off some components - mainly the hard drive - to a state of inactivity; improving cooling systems; adopting chiller-free cooling strategies; or by raising operating temperature [7].

A performance penalty is imposed on hardware components left in a state of inactivity and the entire data center has to assume a delay of up to several seconds for inactive drives. In addition, we must take into account that there is a trend among these infrastructures that involves the utilization of multiple hard drives – ranging from 4 to 6 – rather than RAID systems, which are less energy efficient. In this type of system, hard disk consumption only accounts for 10% of energy consumption; the bulk of the energy is consumed by harder scalable hardware components such as RAM or CPU, which consume about 63% of the total energy [28].

To achieve this high availability of data stored in the data center, many parallel-computing frameworks and distributed file systems such as Hadoop [31] and GFS [12], make use of data replication as a strategy to maximize its availability and fault-tolerance, distributing it in accordance with policies that minimize the possibility of corruption in all stored replicas and thereby the irretrievable loss of any data.

The above policies meet the requirements satisfactorily, since they minimize the risk of data loss within the data center. However, these kinds of policies have some disadvantages, including:

- **Location and status of data are not taken into account**: Temporal data locality is essential to building operating optimization policies for the data center due to the usage of and access to file patterns. Therefore, the computation required to execute the related tasks follows a pattern as shown in Fig. 3.
  In the case study of the Yahoo! Hadoop cluster that serves as a base for GreenHDFS [22], 60% of this cluster total space was being used by data that is not often accessed. The current average lifetime during which a piece of data is often used is 3 days in 98% of cases, and even exceptions to this pattern of use, 80% of the files were used intensively for fewer than 8 days. Within the group of files that are not frequently used, non-access periods varied between 1 and 18 days [23].
- **File distribution policies are not efficient-friendly**: Current distribution policies scatter data blocks between the largest possible number of machines with the aim of minimizing the risk of losing any data due to hardware failure on the machine, failure of facility components at rack level, etc. Servers are therefore constantly underused as mentioned above, which in turn results in low power usage in data storage and associated computing, as well as making impossible an orderly shutdown of these servers impossible without jeopardizing the proper functioning of the data center. Moreover, these distribution policies are based on the static and constant replication model, where all file blocks have the same number of copies and are distributed following the same rules, regardless of the access or computing needs.

For the reasons discussed (the low rates of storage and computing power utilization of these facilities), it seems that if efficient distribution policies are applied in conjunction with switching on/off policies, then not only will data center performance be free from compromise in achieving greater energy efficiency, but also substantial improvements in both aspects can be achieved due to the

inefficiency of the current data distribution policies. Of course, this kind of efficient distribution and switching on/off policies can never jeopardize the availability and integrity of data, but must minimize (if not improve) impact on overall data center performance. Within these distributions and machine power on/off policies we can highlight:

- **Covering subset**: These policies are based on splitting the data center into many disjoint areas so that a number of replicas of each file are stored. The goal of systems that implement these policies is to switch off the maximum number of sectors in the data center to achieve greater energy savings, without affecting the correct operation [26] [35]. The disadvantages of the systems that implement these policies are:
  - The worsening write rate due to write-offloading associated with writing on machines that are not running at the time that writing occurs [3].
  - The number of replicas of each file is constant and static.
  - Neither data time locality nor file utilization pattern are taken into account.

  Systems like Sierra [32] and Rabbit [2] obtain a very high energy proportionality with virtually no impact on the availability and only a slight impact on the overall performance of the data center.
- **Data temperature**: Systems that apply these policies are based on the temporal locality and frequency of use of the files stored in the data center to consistently assign them a temperature (the more frequently used the file is, the hotter the temperature) and redistribute them into two areas: a hot zone aimed at maximizing the performance and availability of data stored on it; and a cold zone whose aim is to minimize the energy consumption of the machines assigned to this area. In such systems, such as GreenHDFS [22], the ultimate goal is to efficiently distribute the machines between these different areas, maximizing the overall performance thanks to improvements in the hot zone, minimizing the overall energy consumption thanks to improvements in the cold zone, increasing the time response as little as possible when reading files from machines switched off (in GreenHDFS, only 2.1% of the readings were affected by this temporary penalty due to switching on the machine at the time of the reading), thereby significantly reducing the energy consumption of servers: 24% in the case of GreenHDFS [23].
- **Dynamic replication**: Other solutions, such as Superset [27], take the above strategies as a starting point, but also take into account the "temperature" of the data above a threshold, not only to power on/off machines, but also to increase or decrease the number of copies of stored data, thereby preserving the availability of data and reducing overall energy consumption thanks to the switching on/off policies and improved performance. This is achieved by transferring storage space and computing power from the cold files that are not frequently used, to those files that need these resources, i.e, the hottest files.

As we have discussed, the problems related to the server shutdown are not critical and do not endanger the proper operation of these infrastructures. Therefore, this paper studies the costs caused by risk aversion, and the energy savings and reduced environmental impact that could be achieved if this fear is overcome.

3. Theoretical analysis

A function that quantifies the costs of fear, i.e. the costs associated with the belief that turning off data center machines imposes a greater cost than the energy savings achieved, is proposed. From this function, an assessment of the risk aversion to switching off machines is provided.

Let us present the problem. Given a set of tasks to be computed in a period of time T, it is assumed that the minimum power consumption, *min*, is achieved by turning off the machines whenever possible, and that the maximum power consumption, *Max*, is obtained in the case that the machines never are turned off. Hence, the extra expense imposed due to the consumption from all those machines remaining turned on without interruption is given by

$$M = Max - min$$

Let us suppose that a datacenter has n machines, all of them equal. This act is justified since actually, data center machines are grouped by racks of identical machines. Even machines from different racks share the same components or at least components are produced by the same manufactures.

Let N_j be the maximum number of times that a machine j, $j = 1, \ldots, n$, can be turned on given an operation time T. This value is computed as a maximum that depends on operation time T, shutting down time (T_{off}, time needed to switch off a machine) and turning on time (T_{on}, time needed to switch on a machine from the off state) as follows:

$$N_j = \frac{T}{T_{off} + T_{on}}$$

Therefore, by considering that all machines are equal ($N_j = N$), the maximum number of times that the machines of the datacenter can be turned on given an operation time T is

$$N_1 + \cdots + N_n = n \cdot N$$

Let X_i^j be the random variable which takes the value 1 if a computer j breaks down on power switching i and 0 otherwise. Hence, if the probability of $X_i^j = 1$ is p_i, that is $P(X_i^j = 1) = p_i^j$, then X_i^j follows a Bernoulli model and, hence $E[X_i^j] = p_i^j$. With respect to p_i^j, some considerations must be given:

- As aforementioned, all machine of the datacenter are supposed equal, therefore $p_i^j = p_i$ for any $j = 1, \ldots, n$.
- p_i depends on the power switching i and this values can be considered constant within a horizon of the framework T. Clearly, $p_i = p_i(t)$ and $\frac{d\,p_i(t)}{dt} > 0$, that is, the probability of malfunction of a machine is going to increase during its life. Nevertheless, the technology of this machine provides that this probability decreases slowly, $\frac{d\,p_i(t)}{dt} \approx 0$, and the considered operation time, T, is short compared to its lifetime. Hence, $p_i = p \approx constant$ can be considered.
- With respect to the value of p. The advance in the technology indicates that the real value of p_i is to be very close to 0. Nevertheless, from a cognitive point of view, the data center administrator can consider it is a high value and this is the reason why it would never be a good idea to switch off machines.

It is worth noting that if a machine breaks down, there are other machines of the datacenter available to replace its operational requirements. Let n_j be the number of times that a machine j should be switched off, $0 \le n_j \le N$.

From here, if x denotes the number of power cycles, a new random variable

$$S(x) = \sum_{j=1}^{n} \sum_{i=1}^{n_j} X_i^j, \qquad x = 0, 1, \ldots, n \cdot N$$

where $x = \sum_{j=1}^{n} n_j$, that is, x is the number of power cycles performed in all machines. Thus, $S(x)$ is a random variable that represents the number of machines broken down in x power cycles and, hence, $0 \le S(x) \le n$ for any x.

The average number of damaged machines after x switching on/off cycles, that is, the expectation of the random variable $S(x)$ is calculated as follows:

$$E[S(x)] = E\left[\sum_{j=1}^{n} \sum_{i=1}^{n_j} X_i^j\right] = \sum_{j=1}^{n} \sum_{i=1}^{n_j} E[X_i^j] = p \cdot \sum_{j=1}^{n} n_j = x \cdot p$$

Furthermore, the cost of repairing computers damaged by the switching on/off cycles has also to be taken into account. Let $C_r > 0$ be the average cost of repairing the computer. Hence, the costs of fear, derived from switching on/off machines, denoted by C_{fear}, can be given as follows:

$$C_{fear}(x) = x \cdot p \cdot C_r, \quad x = 0, 1, \ldots, n \cdot N$$

In addition, if a computer is turned off and then there is a request that requires the machine to be turned on, then the client will need to wait until the computer is turned on. Considering C_o as the opportunity cost that measures the value that a customer gives to that lost time, and T_{on} as the time needed for a computer to be turned on. Then, the turn on costs, denoted by C_{on}, can be quantified as follows:

$$C_{on}(x) = x \cdot T_{on} \cdot C_o, \quad x = 0, 1, \ldots, n \cdot N$$

Therefore, the total cost of turning off x machines, denoted by $C(x)$, is given as $C(x) = C_{fear}(x) + C_{on}(x)$, that is,

$$C(x) = x \cdot (T_{on} \cdot C_o + p \cdot C_r) \qquad x = 0, \ldots, n \cdot N \tag{1}$$

From the above function, the cost of switching off the machines is as follows:

$$C(n \cdot N) = n \cdot N \cdot (T_{on} \cdot C_o + p \cdot C_r)$$

Nowadays due to the different aspect of the life among them, the cognitive aspect, most companies prefer not to turn off machines so this decision implies that $C(n \cdot N) > M$. The main aim of this paper is to show that this is not an optimal decision.

First, in order to simplify the function given in (1), the variable $y = \frac{x}{n \cdot N}$ is considered which indicates the proportion (per unit) of the number of switching on/off cycles applied against the natural maximum applicable. Hence,

$$C(y) = y \cdot n \cdot N \cdot (T_{on} \cdot C_o + p \cdot C_r) \tag{2}$$

From the definition of y, it can be seen that $(1 - y)$ represents the percentage of switching on/off cycles not applied to the machines. Assuming that the cost of having all the extra machines turned on, M, is proportional to the percentage of switching on/off cycles applied as represented by y, then $(1 - y) \cdot M$ represents the cost of having the machines switched on. From these latter two costs, the cost for having a percentage of machines turned off, is given by

$$f(y) = y \cdot n \cdot N \cdot (T_{on} \cdot C_o + p \cdot C_r) + (1 - y) \cdot M \quad 0 \le y \le 1 \tag{3}$$

Since M is not null, and $C(1) = n \cdot N \cdot (T_{on} \cdot C_o + p \cdot C_r) > M$ due to current fear experienced by most companies, the value

$$A = \frac{C(1)}{M} > 1$$

is considered and the cost function, denoted by $f_{current}$, is written as follows:

$$f_{current}(y) = A \cdot y + (1 - y) \qquad 0 \le y \le 1,\ A > 1 \tag{4}$$

Following the current hypothesis which assumes that switching off any machine implies more cost (the so-called fear cost), this objective function reaches its minimum when $y_0 = 0$, i.e., when no machines are turned off and they maintain continuous execution, and $f_{current}(y_0) = 1$ (An example of this kind of function is given in Fig. 4).

4. Fear cost

In this section, a new cost function is given by assuming that the switching off/on of machines in moderation may have a benefit.

First, let us indicated that the function $f_{current}(y)$ verifies that $\frac{d f_{current}(y)}{dy} = A - 1 = cte$, that is, the increment of the emotional cost caused by the modification of the percentage of power cycles is constant for the datacenter administrator. However, this is not a realistic hypothesis since by taking into account the pessimism (cognitive aspect) of the administrator, the $f_{current}(y)$ function must verify that $\frac{d^2 f_{current}(y)}{dy^2} > 0$ since, for instance, the incremental cost to change of 0.1–0.2 must be smaller than the incremental cost to change of 0.7 to 0.8.

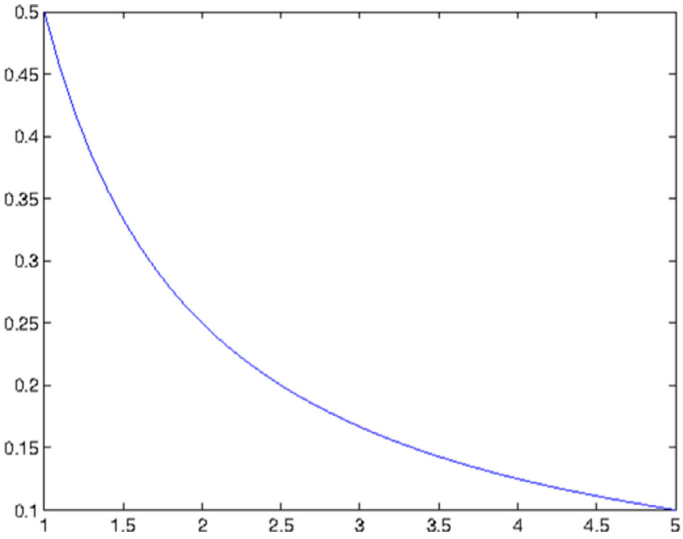

Fig. 4. Graphical representation of the point where the minimum of function (5) is attained.

Hence, the new function cost, denoted by f_{prop}, must verify that

- If y is near to zero, then $f_{prop}(y) < f_{prop}(0)$ since the switching off/on of machines in moderation may have a benefit. Furthermore, in order to provide a regular function it is imposed that $\frac{d f_{prop}(y)}{dy}$ exists for any y.
- By following the commentary of the $f_{current}(y)$ function with respect to the second derivative, the $\frac{d^2 f_{prop}(y)}{dy^2} > 0$ is required.
- Furthermore, a similar to $f_{current}(y)$ functional form is required for $f_{prop}(y)$.

Thus, the simplest function with these conditions is:

$$f_{prop}(y) = Ay^2 + (1 - y) \qquad 0 \le y \le 1,\ A > 1 \tag{5}$$

Moreover, another justification of this new function is that it lends less weight to the value of $C(y)$ given in (2) in the function (3). Hence the importance of switching off machines is relaxed.

The function (5) is convex (see Fig. 5) and reaches its minimum at the point:

$$y_0 = \frac{1}{2A} \tag{6}$$

and $f_{prop}(y_0) = 1 - \frac{1}{2A}$. This means that the ideal situation is to switch off $\frac{1}{2A}$% of machines, and the savings, as a percentage, are equal to the percentage of machines switched off.

Thus, if A has a high value, this favours the shutdown of the servers in the data center. And, if A is close to 1 it favours keeping 50% machines on/idle which is a consequence of the supposition that the 'switching off/on of machines in moderation may have a benefit'.

Hence, a coefficient, denoted by *fear*, which measures the risk aversion to switching off the machines, is modeled as follows:

$$fear = 1 - \frac{1}{A}$$

This value verifies $0 \le fear \le 1$ and satisfies:

- $fear = 0$ $(A = 1)$ implies low risk aversion, and under the hypothesis of 'switching off/on machines in moderation may have a benefit', means switching off 50% of the machines.

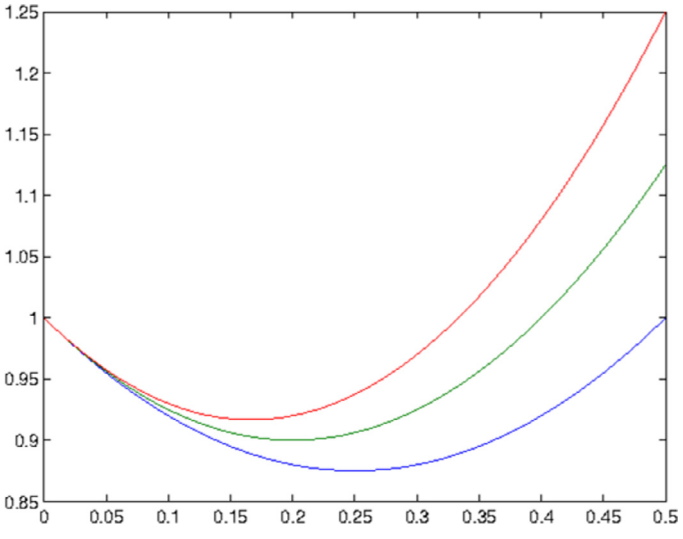

Fig. 5. Graphical representation of the function (5) for $A = 2$ (blue), 2.5 (green) and 3 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

- $fear = 1$ ($A = \infty$) implies maximum risk aversion, and therefore machines are never switched off.

As aforementioned, most data center companies currently do not shut down servers, so the value of A is set to ∞ in the proposed function and hence the number of machines switched off is 0 ($y_0 = 0$).

Based on these developments, it is possible to model the risk aversion experienced by data center companies by posing a simple question: *What percentage of machines are you willing to switch off?* For instance, if the answer is 10%, the equation $0.1 = \frac{1}{2A}$ is resolved, which means that $A = 5$, thus the $f_{prop}(y) = 5y^2 + (1 - y) \qquad 0 \leq y \leq 1$ and from this point the emotion of the fear experienced by the company is as follows:

$$A = 5 \Rightarrow fear = 1 - \frac{1}{5A} = 0.8$$

In contrast, if the answer is 40% of machines, then $A = \frac{5}{4}$, and therefore: $fear = 0.2$.

5. Conclusions

In this paper we have presented the cost of risk aversion to which most companies currently subscribe due to the false belief that turning off machines in data centers involves more costs than savings.

In order to demonstrate this, an objective function has been proposed which determines that a lower total cost can always be attained by turning off data center servers a number of times, showing that the current belief is a mistake that should be corrected by applying shutting on/off policies.

As future work, we plan to measure the extra costs associated with turning off the machines in terms of hardware damage and to measure the energy savings that could be obtained by building a software system which implements policies for energy efficiency in data centers.

Acknowledgments

This research is supported by the project Simon(TIC-8052) and Context-Learning (P11-TIC-7124) of the Andalusian Regional Ministry of Economy, Innovation and Science and by the Spanish Ministry of Economy and Competitiveness through the project HERMES - Healthy and Efficient Routes in Massive Open Data-based Smart Cities (TIN2013-46801-C4).

References

[1] N. Ahuja, C. Rego, S. Ahuja, M. Warner, A. Docca, Data center efficiency with higher ambient temperatures and optimized cooling control, in: Proceedings of the Semiconductor 27th Annual IEEE Thermal Measurement and Management Symposium (SEMI-THERM), IEEE, 2011, pp. 105–109.
[2] H. Amur, J. Cipar, V. Gupta, G.R. Ganger, M.A. Kozuch, K. Schwan, Robust and flexible power-proportional storage, in: Proceedings of the 1st ACM Symposium on Cloud Computing, ACM, 2010, pp. 217–228.
[3] H. Amur, K. Schwan, Achieving power-efficiency in clusters without distributed file system complexity, in: Computer Architecture, Springer, 2012, pp. 222–232.
[4] L.A. Barroso, U. Hölzle, The case for energy-proportional computing, IEEE computer 40 (12) (2007) 33–37.
[5] G. Chanel, K. Ansari-Asl, T. Pun, Valence-arousal evaluation using physiological signals in an emotion recall paradigm, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2007 (ISIC), IEEE, 2007, pp. 2662–2667.
[6] M. Díaz, J. Saez-Pons, M. Heerink, C. Angulo, Emotional factors in robot-based assistive services for elderly at home, in: Proceedings of the IEEE RO-MAN, 2013, IEEE, 2013, pp. 711–716.
[7] N. El-Sayed, I.A. Stefanovici, G. Amvrosiadis, A.A. Hwang, B. Schroeder, Temperature management in data centers: why some (might) like it hot, ACM SIGMETRICS Performance Evaluation Review 40 (1) (2012) 163–174.
[8] X. Fan, W.-D. Weber, L.A. Barroso, Power provisioning for a warehouse-sized computer, in: ACM SIGARCH Computer Architecture News, volume 35, ACM, 2007, pp. 13–23.
[9] M.E. Femal, V.W. Freeh, Boosting data center performance through non-uniform power allocation, in: Proceedings of the Second International Conference on Autonomic Computing, 2005. ICAC 2005, IEEE, 2005, pp. 250–261.
[10] A. Fernández-Montes, L. Gonzalez-Abril, J.A. Ortega, L. Lefèvre, Smart scheduling for saving energy in grid computing, Expert Syst. Appl. 39 (10) (2012a) 9443–9450.
[11] A. Fernández-Montes, F. Velasco, J. Ortega, Evaluating decision-making performance in a grid-computing environment using DEA, Expert Syst. Appl. 39 (15) (2012b) 12061–12070.
[12] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, ACM SIGOPS Oper. Syst. Rev. volume 37 (2003) 29–43.
[13] J. Glanz, Power, pollution and the internet, NY Times 22 (2012).
[14] D. Goleman, S. Sutherland, Emotional Intelligence: Why it can Matter more than IQ, Bloomsbury, London, 1996.
[15] J. Gratch, S. Marsella, Tears and fears: Modeling emotions and emotional behaviors in synthetic agents, in: Proceedings of the Fifth International Conference on Autonomous Agents, ACM, 2001, pp. 278–285.
[16] J. Gratch, S. Marsella, A domain-independent framework for modeling emotion, Cogn. Syst. Res. 5 (4) (2004) 269–306.
[17] H. Gunes, M. Pantic, Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, in: Intelligent virtual agents, Springer, 2010, pp. 371–377.
[18] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: a survey, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 827–834.
[19] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: first steps towards an automatic system, in: Proceedigns of the Affective dialogue systems, Springer, 2004, pp. 36–48.
[20] N. Hayes, DatacenterDynamics Global Industry Census 2012, Technical Report, DatacenterDynamics, 2012.
[21] J.P. Johansen, L. Diaz-Mataix, H. Hamanaka, T. Ozawa, E. Ycu, J. Koivumaa, A. Kumar, M. Hou, K. Deisseroth, E.S. Boyden, et al., Hebbian and neuromodulatory mechanisms interact to trigger associative memory formation, Proc. Natl. Acad. Sci. 111 (51) (2014) E5584–E5592.
[22] R.T. Kaushik, M. Bhandarkar, Greenhdfs: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster, in: Proceedings of the USENIX Annual Technical Conference, 2010, p. 109.
[23] R.T. Kaushik, M. Bhandarkar, K. Nahrstedt, Evaluation and analysis of greenhdfs: a self-adaptive, energy-conserving variant of the hadoop distributed file system, in: Proceedings of the Cloud IEEE Second International Conference on Computing Technology and Science (CloudCom), IEEE, 2010, pp. 274–287.
[24] J. Koomey, Growth in data center electricity use 2005 to 2010, A report by Analytical Press, completed at the request of The New York Times (2011).
[25] J.G. Koomey, Worldwide electricity used in data centers, Environ. Res. Lett. 3 (3) (2008) 034008.
[26] J. Leverich, C. Kozyrakis, On the energy (in) efficiency of hadoop clusters, ACM SIGOPS Oper. Syst. Rev. 44 (1) (2010) 61–65.
[27] X. Luo, Y. Wang, Z. Zhang, H. Wang, Superset: a non-uniform replica placement strategy towards high-performance and cost-effective distributed storage service, in: Proceedings of the International Conference on Advanced Cloud and Big Data (CBD), IEEE, 2013, pp. 139–146.
[28] D.A. Patterson, The data center is the computer, Commun. ACM 51 (1) (2008) 105.
[29] E. Pinheiro, W.-D. Weber, L.A. Barroso, Failure trends in a large disk drive population, in: FAST, volume 7, 2007, pp. 17–23.
[30] E. Schubert, Modeling perceived emotion with continuous musical features, Music Percept. 21 (4) (2004) 561–585.

[31] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010, IEEE, 2010, pp. 1–10.
[32] E. Thereska, A. Donnelly, D. Narayanan, Sierra: practical power-proportionality for data center storage, in: Proceedings of the Sixth Conference on Computer Systems, ACM, 2011, pp. 169–182.
[33] P.A. Thoits, The sociology of emotions, Annu. Rev. Sociol. (1989) 317–342.
[34] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S.S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling, in: Proceedings of the INTERSPEECH, 2010, pp. 2362–2365.
[35] Z. Zeng, B. Veeravalli, Do more replicas of object data improve the performance of cloud data centers? in: Proceedings of the 2012 IEEE/ACM Fifth International Conference on Utility and Cloud Computing, IEEE Computer Society, 2012, pp. 39–46.