Bulletin de la Société des Sciences de Liège, Vol. 85, 2016, *p.* 1273 - 1282

Presenting a solution based on content distribution networks in order to decrease energy consumption in data center's infrastructure

Abdolreza Rasouli KENARI[1,*], Hossein HASSANPOUR[2], Vahid AHMADI[3]

[1]Faculty of electronic and computer science Qom University of technology, Qom, Iran
[*]Corresponding author: rasouli@qut.ac.ir
[2]Department of Computer and Electronic Taali Institute of Higher Education
Qom, Iran, H.hasanpour@taali.ac.ir
[3]Department of Computer and Electronic Taali Institute of Higher Education
Qom, Iran, v.ahmadi@taali.ac.ir

Abstract

Growing usage of electronic technologies such as internet has brought about creation of various platforms in this infrastructure. Data centers are exploited by different governmental and non-governmental firms within various locations according to some criteria like facility usage volume, shared information amount, political and geographic positions ands. These centers include different processing-storage and communicational segments, each needs its own energy and cost consumption which affects efficiency, productivity and profitability of these centers. Content delivery network is a solution to connect these data centers which are located in various geographical nodes. Applying this technological solution can lead to efficiency increase, delay decrease in sending information and other capabilities. Using digital technology platforms requires different sources of nonrenewable energies such as electricity, petroleum and etc. Consuming energy causes environmental degradation, beside leading firms to undergo extra costs in the process of providing services. In this paper, we are trying to present a method in order to decrease the energy amount using the suggested idea in connected content distribution networks and their usage in internet infrastructure and focusing on geographical differences of data centers. The results of this method are mentioned at the end of this paper which conclude the above model decreases energy consumption and increases efficiency based on its usage area.

Keywords: Content distribution Network, Datacenter, energy consumption, Cloud Computing.

1 Introduction

Internet infrastructure includes various technologies such as different communicational technologies in the main core or its secondary channels, various storage technologies as one of the most applicable present technologies. Network growth is evaluated based on accuracy and availability. Content distribution networks can be mentioned as some of the applied methods belong to this platform's communicational segment which are used in network platform in order to avoid flash crow event, provide content accessibility increase and information delivery delay in network platform. Data centers are regarded as the main components of this technology. Data centers which are scattered within internet network platform play request's processing and storing role. This segment aims at fulfilling all user's requirements and is also one of the most expensive segments of internet infrastructure which is exploited by private and state firms. Expenses are regarded as a significant factor in attracting users to a technology; because users expect an efficient infrastructure with justifiable expenses and this service can be calculated based on incoming expenses to servers. System usage cost is calculated based on different factors including the amount of expenses which are paid by providers (expenses such as infrastructure consuming energy). Optimizing the amount of consumption by this segment can also save the environment which is regarded as one of the most significant criteria in nowadays technological progress.

Reminder of this paper can be as follows: first section evaluates the above method's prerequisites. The next section is dedicated to examine the discussed model regarding its requirements and capabilities.

2 Literature review

In this section, we aim at defining the applied conceptions in this research to improve and enhance the proposed solution in next steps.

2.1 Content distribution network

A content distribution network (CDN) has a mirror effect in enlarging efficiency level by repeating a part of information on several web servers which are called surrogate servers in this practical scope and strategically scattered within different nodes of network[1]. This method of placing servers helps network to decrease its information delivery rate. Perhaps we can better define content distribution network by making an example. Imagine a user wants to receive a web page. Each webpage includes static segments like movies and pictures with slower change rate which are the

same for all users and dynamic segments that have been determined based on user's requirements. We consider two following methods:

2.1.1.1 The first method

Information request usually receives the retentive server after passing a network platform and after evaluating user's required information is sent via that system. There is usually possibility of high delay rate while sending data, because data should pass across different segments. Resending information is also required because of inadvertent changes and this problem leads servers to undergo extra expenses.

2.1.1.2 The second method

In this method, we are trying to present the conceptual model of content delivery network. This model selects closest servers to users based on various alternatives and statics segments which usually allocate a high volume of information to themselves are repeated in these servers. This repetition leads high volume segments which are the main reason behind the presence of delay and congestion within the network to be delivered to user with the lowest cost and the fastest rate. Dynamic segments of information should be sent to user via the main server. This segment of information will be delivered to user in faster speed and with lower network traffic[2].

Now, relying on this example we can better define the general concept of content delivery network and then, better discuss this method.

CDN is an optimized network which is able to deliver a content to a network, this platform can be internet, intranet and etc. In CDN literature, content is defined as each digital data source which includes encrypted and metadata media. Encrypted media include attached dynamic-static media such as document, image, voice, picture and metadata. Metadata offers a description for contents and information and leads to better identification, management and description of multimedia data[2].

2.1.2 Content distribution network providers and their geographical scattering

Three most significant components of content distribution networks are content provider, CDN provider and the end user. Content provider or CDN customer is someone who has an address of web objects. These objects are held in the main server of content provider. CDN provider is an organization which presents secure content with the lowest delay within different nodes of its coverage area. The end user is regarded as a part of model who is also exposed to contents. CDN providers use cashing servers in different parts of their network which are called surrogate servers and respond to user's request by selecting the best servers based on determined criteria. The number of surrogate servers can differ in different provides coverage area based on their geographical areas[3]. For example table 1 demonstrates a list of providers of content distribution network services and the number of applied surrogate servers within their supported geographical platforms. Figure 1 illustrates a schematic view of geographical circumstances of CloudFlare content distribution network servers. This figure evaluates scattering rate of the above server within different geographical nodes and the defined criteria for this scattering is the number of users who are receiving offered facilities.

Figure 1: A schematic view of geographical circumstances of Cloud Flare content distribution network

2.1.3 Various methods for presenting content distribution networks

Different methods have been presented in order to replace surrogate servers in this technology by considering demand growth from using firm's side. We can mention a single or multiple firms which will be elaborated later in this paper.

2.1.3.1 Replacing surrogate servers via single firm method

In this method, service providers place several servers in central points of their network based on network alternatives such as geographical width of network users, incoming workload rate to equipment and serve their user's in this way. In this method, users receive information from the closest server containing information from their relevant firm.

Table 1: A list of providers of content distribution services and their characteristics

CDN Service Providers	Service Type	Content Distribution	Fees	Customers
Akamai www.akamai.com	Multi-ISP, partial-site Request servicing, peering	More than 12,000 surrogate servers spanning 1,000 networks in 62 countries	US$1,995 per month for each Mbps of delivered content	Covers 70 percent of the market, with more than 3,600 customers including Apple, CNN, MSNBC, Reuters, and Yahoo
Adero www.webvisions.com/	Multi-ISP, full-site request servicing, peering	Surrogate servers in more than 30 countries	Depends on resellers (CDNs that buy Adero services)	Serves 30 customers, including resellers Exodus and UUNET
Digital Island www.sandpiper.net	Multi-ISP, partial-site request servicing, peering	2,500 surrogate servers spanning 327 networks in 35 countries	Starts at US$1,500 per month	More than 900 customers including AOL, Canon, Cisco Systems, Microsoft, and Hewlett Packard
Mirror Image www.mirror-image.com	Multi-ISP, partial-site request servicing, peering	22 surrogate servers in North America, Europe, and Asia	US$2,100 per month for each Mbps of delivered content	More than 200 customers including Creative, Open Systems, and SiteRock
Inktomi www.inktomi.com	Single-ISP, full-site request servicing, peering	10 surrogate servers across China	Starts at US$4,000 per month	13 CDNs including Adero and Digital Island and more than 200 Web sites

2.1.3.2 Placing surrogate servers via multiple firm method

As it is obvious, in the previous method delay rate will increase by increasing the distance between user and the center which is supported by the contactor company. Thus, this method decreases the technological efficiency of content delivery networks and provides the possibility of defining a new method with the title of replacing multiple surrogate servers by contracting determined service levels. In this method, content delivery network firms use their own surrogate servers based on provisions to meet user's requirements. In this way, when user demands increase, and they are out of firm's geographical area, firm places information in closest surrogate servers of other counterpart firms and then, delivers it to users and in this way, avoids efficiency decrease and delay rate increase in system. This method is called intert_connected content distribution networks method[3].

2.2 Data Center

Data center refers to a set of servers, security/communicational infrastructures and electronic equipment which are applied for presenting, preserving and supporting web-based services (Internet/Intranet/extranet). Organizations, firms and people can set up information and network-based services on internet (intranet, extranet) by applying presented services from website data centers. Data center can be defined as a processing center, storage center and data collection center or all of them based on its practical type. Plentiful various data centers are serving customers all over the internet network. Some of these data centers are commercially restricted in usage and just applied inside the organization; while some of other centers are commercially or generally accessible and usable via internet. This segment of infrastructure is one of the most costly parts of serving demands in massive networks like internet. We can mention energy consumption cost as another costly factor in this segment, accordingly various methods are used to optimize this segment[4].

2.2.1 Energy consumption in data centers

Data centers are a set of powerful servers which are used for presenting, preserving and supporting web-based networks. Individuals and organizations can set up their web-based information and services in their desired network by applying data centers services. Data centers consumption level is 100 times more than their counterpart buildings in terms of size. Different components can affect data centers energy consumption. As figure 2 which is derived by Liebert company researches demonstrates cooling department consumes more than all other departments in a data center and this level of consumption can differ based on data centers hardware characteristics, computational load and input/output rate in different times. It should be noted that other factors can also affect this alternative (energy consumption).

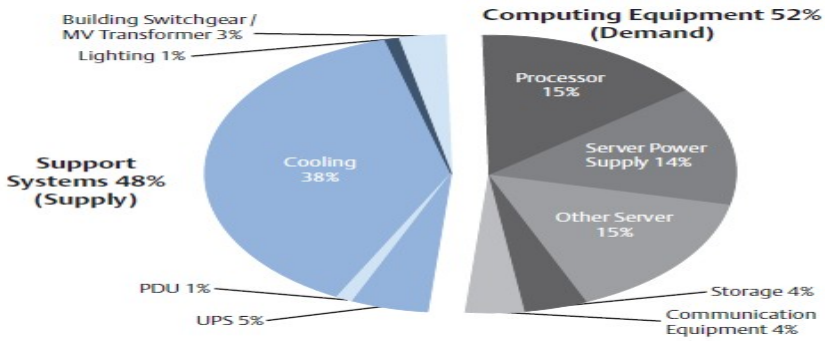

Figure2: A schematic view of partitioning energy consumption in data centers

As Computerworld website reported[5], we expect 4 percent growth in energy consumption of this hardware equipment from 2014 till 2020 despite the fact that data centers had experienced 24 percent growth in electric power consumption from 2005 to 2010. This issue will be highlighted more than previous time regarding the growing demands for exploiting computational resources.

2.2.1.1 Energy consuming components in data centers

Energy consumption in data centers can be divided to building and IT department. Energy consumption in IT and information storage departments include: computers, servers, information storage devices and related hardware to network and they usually consume a constant level of energy[6]. Energy consumption in building consists of: cooling system, fans, pumps, lighting and fire extinguishing. The whole consumed energy in data centers is obtained by these two sources.

2.2.1.2 Data centers efficiency in terms of energy

As formula 1 demonstrates the partial effect of processing a data center is calculated by dividing the process amount to the energy consumption amount. It also depends on equipment and server productivity which are shown in formula 2 and3, respectively[4].

Efficiency= $\frac{Computation}{Total\ Energy} = \frac{1}{PUE} * \frac{1}{SPUE} * \frac{Computation}{Total\ Energy\ To\ Electronic\ Components}$ (1)

Facility Efficiency: Power Usage Effectiveness (PUE)=$\frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy}$ (2)

Server Efficiency: SPUE = (Total Server Power) / (Useful Power) (3)

According to the fact that the most amount of energy is wasted in heat form (as figure 2 illustrates), finding methods to decrease the generated heat in data centers plays a significant role in decreasing data centers consumed energy. We will mention several methods to economize energy in the following[4]:

- Selecting an appropriate capacity for data center
- Applying racks which can be cooled with water
- Using lightening systems with lower heat waste such as LIDs
- Applying UPSs with capacities more than 60 percent
- Appropriate design of building with lowest energy waste
- Choosing optimized cooling system
- Permanent energy monitoring using energy management software and comparing it with standard quantities
- Using the cold weather in cold seasons and cold regions.
- Using servers with lower consumption
- Appropriate arrangement of racks

3 The proposed model

Data centers are exploited in various locations by organizational and non-organizational firms based on alternatives such as internet infrastructure facilities volume, shared information level and political and geographical characteristics as the main component of internet infrastructures.

These centers include different communicational and storage-processing segments which their usage can lead to different energy and cost consumption and these results affect these center's efficiency, productivity and profitability. Consuming energy in different segments brings about heating. This heating which uses the most percentage of energy consumption according to figure 2 to be eliminated is an undeniable component of our discussed environment. If this heating is not kept in the desired temperature can lead to failure and decrease in efficiency and productivity rate of

system. Load volume of different segments affects the amount of generated heat. A part of this heat is usually generated via energy flow in orbits. We aim at decreasing energy consumption by falling data centers to sleep.

In our proposed model, used data are duplicated in different data centers based on their previous type and time applications to help firm to select the best location in order to meet users demands based on obtained previous information, time analysis and geographical aspects. We need to apply content distribution networks technology and connecting different data centers to each other to present our desired infrastructure and provide this desired infrastructure to implement our proposed model.

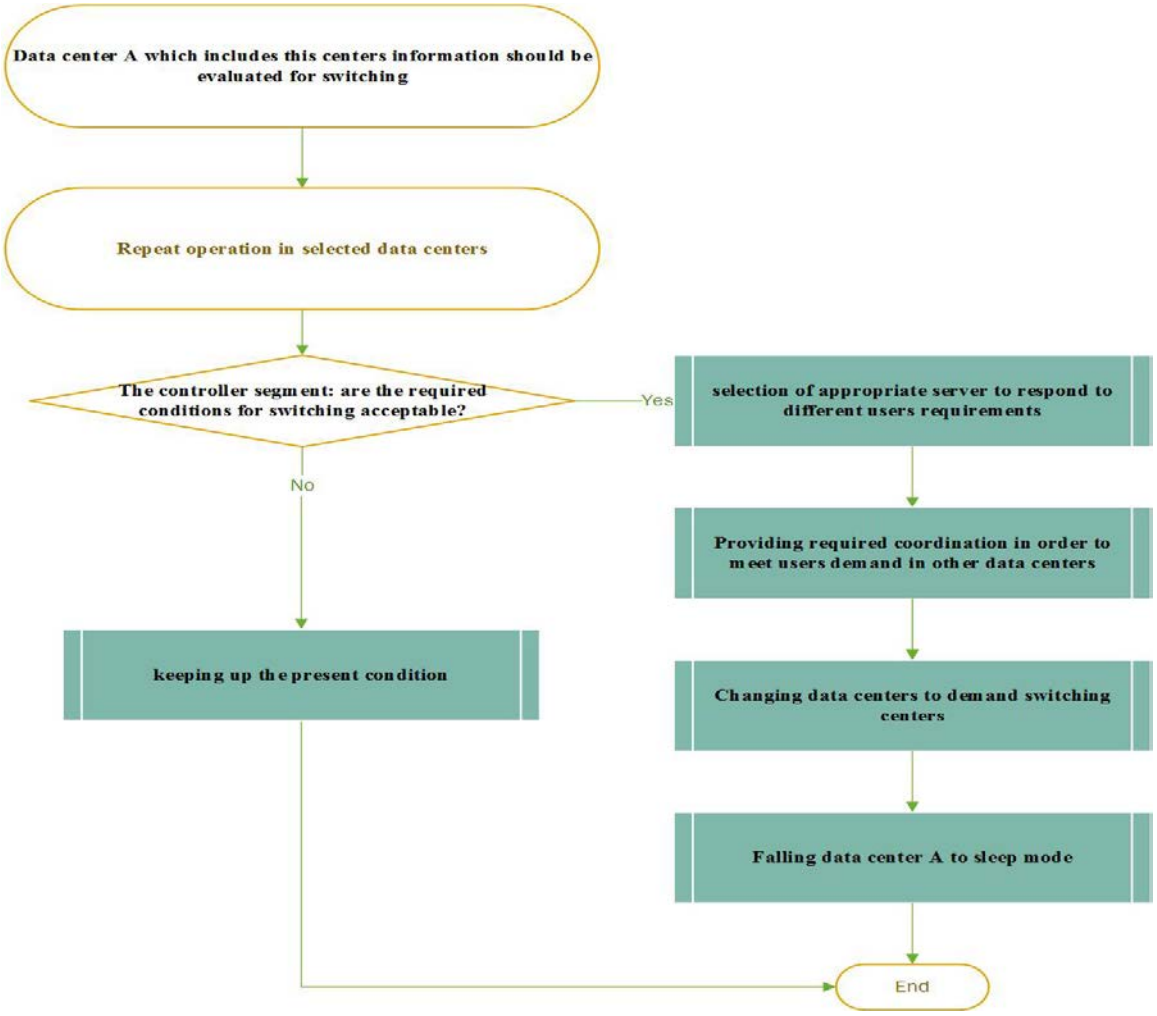

Figure3: Modelling the proposed model in applicants side

As it is demonstrated in figure 3, data center mode can be defined after information duplication based on different status analysis in controller segment.

After determining system status, if the selected mode is falling equipment to sleep, the following steps should be taken before performing this action.

1-If there is any difference between two data centers while performing mirroring information, at first information should be updated and differences in data centers should be eliminated.

2-Ruuning workloads should be transferred to a specified segment of data center which sleeps after a while.

3-Changing the above data center to an information switching center in order to change the direction of user's request to an active data center

4-Falling system to sleep

5-Responding to user's requests by applying facilities of another selected location

Taking these steps falls a part of used infrastructure to sleep and in this way, affects different segment's energy consumption.

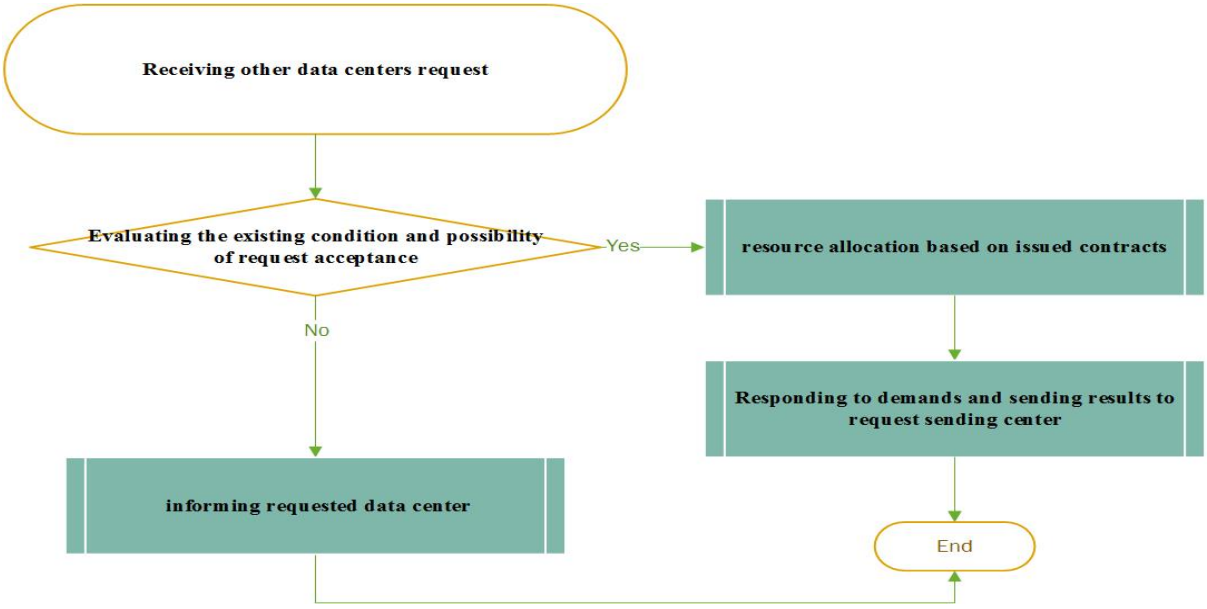

Figure4: Modelling the proposed model in selected data center

Figure 4 has shown different steps of responding to another data center's request by request receiver. In this condition, the responder data center responds to another data center's request by analyzing its status. If another data center accepts to respond, received requests will be responded by it after undergoing a mode change. One of the requirements of this model is to preserve the information owner's identity. As figure 3 illustrates the above segment acts as a switching center after changing mode. So the final response will be sent to that segment and then, to the mentioned user.

3.1 Designing controller segment

Status change with high volume of frequency brings about plentiful failures in equipment. It is also capable of decreasing system efficiency remarkably because of including high computational and communicational volume; accordingly, we can consider status change as one of the most important segments of the proposed model. Applied alternatives in this segment can differ based on infrastructures requirements. In this paper, we have applied the most effective alternatives among the others which will be explained later.

Two main responsibilities of controller segment are defined as follows:

1-Should system go to sleeping mode or continue its operation in normal mode

2- Which data center should be selected among other alternatives to burden the workload of our data center.

For selecting data center's mode, we can apply one of the following alternatives:

Boarding mode: one of the most significant alternatives which engages in system workload rate is boarding mode; because system workload rate differs in different hours. Results of evaluated data in the next section prove this claim.

Evaluating system workload in a similar status in the past: evaluating obtained statistics about system's previous workload and concluding based on these statistics help firms to implement the proposed scheme.

Period of sleeping mode which can be tolerated by system: one of the key alternatives which should be defined in the area of examining system's previous workload is period of tolerable sleeping mode by system. As we mentioned before, frequent mode change can bring about efficiency decrease. Thus it is rational to evaluate system's previous status and assure system's efficiency rate by initializing this characteristic. We create decision table based on specified alternatives in control center later in this section and then, information analyzers respond to mode change by evaluating this request.

In another segment of control center we should determine which data center should be used for performing this data center's operations. We have to use a system set for placing mirrored copies of information in initial phase of this segment. After determining and placing information in these centers, we should consider the issue of selecting the best surrogate data center. In next section, we will elaborate on these alternatives in detail.

Determining alternatives for selecting the best surrogate data center

This factor is one of the most effective criteria in selecting the best surrogate data center for responding to user's different requirements. In general, the best responder server to user's request, is the most similar one to user in terms of proximity topology.

Proximity topology is an abstract criterion which considers a variety of alternatives such as velocity, physical distance, reliability, transformation expenses in order to calculate proximity to users. We can combine this criterion with other ones. Some of the content delivery networks apply a combination of surrogate server's electors. For instance, we can mention a combination of proximity topology and the workload of existing servers in network. In this method, closest servers to user are selected based on proximity alternatives at first, then, it can decided that weather the best server is in overload mode (is usable) or not. If workload of the closest selected server is high and it takes time to use this server, the next close server will be selected as surrogate server and called the best surrogate server for responding to requests at the end. When the distance between user and the best selected data center increases, service provider can resolve the problem of existing delay by allocating a high speed communicational link to communicational route to user.

4 Elaborating on results of the proposed scheme

In this section, we examine different statuses of service providers in a wide variety in order to elaborate on results of the proposed model and then, we calculate efficiency level of model based on obtained results. We have applied different organization's workload information for analyzing the proposed model.

4.1 The NASA Ames iPSC/860 log

Existing information in this file includes obtained data by user's accounts in a three month interval and within 128 existing nodes of aerodynamic numerical systems of NASA AMES research center[7]. Figures 5 and 6 demonstrate system workload status chart within various hours and days, respectively.

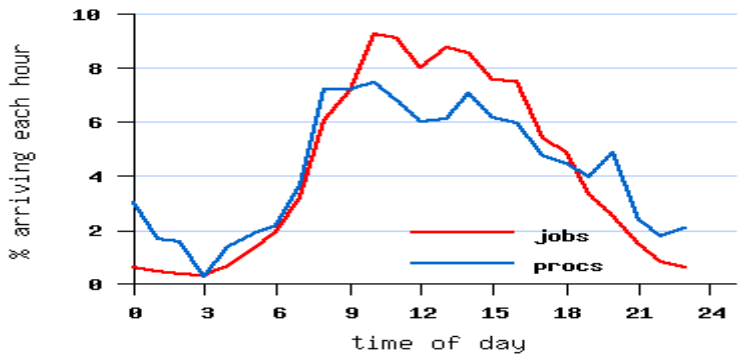

Figure 5: System workload status chart in different hours a day using The NASA Ames iPSC/860 log data

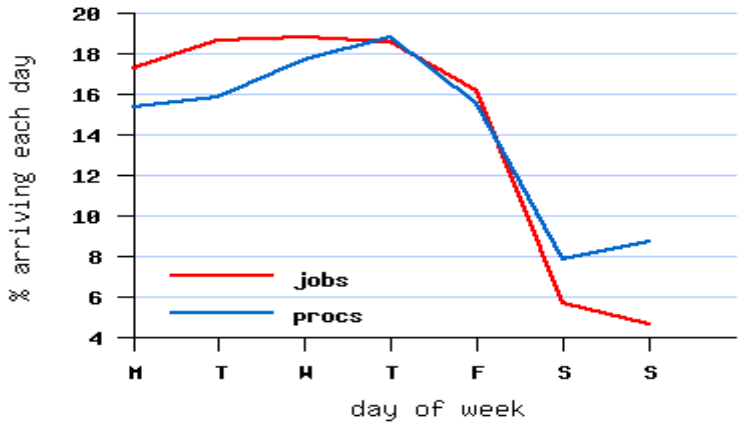

Figure 6: System workload status chart in different days using The NASA Ames iPSC/860 log data

4.2 Intel Netbatch Grid

This file includes Intel Netbatch Grid user account's data in time period of one month. This grid environment contains several connected clusters in different places of world, each includes 10000 nodes itself. This data is collected by Ohad Shai, Edi Shmueli, and Nir Antebi and are used in[8]. Figures 7 and 8 illustrate analysis results of these files, respectively.

Figure 7: System workload status chart in different hours a day using Intel Netbatch Grid data

Figure 8: System workload status chart in different days using Intel Netbatch Grid data

4.3 The RICC Log

This information includes RIKEN Integrated Cluster of Clusters user account data from May to Sep 2010. Different segments of this cluster platform contain 1024 nodes, each includes 12 GIG memory and 4 CPU cores and so they totally include 12 Terabyte memory and 8192 processor's core. RIKEN is used for scientific researches of japan universities and research centers. This data has been used in [9, 10]. Figures 9 and 10 illustrate status analysis of this center respectively.

Figure 9: System workload status chart in different hours a day using The RICC log data

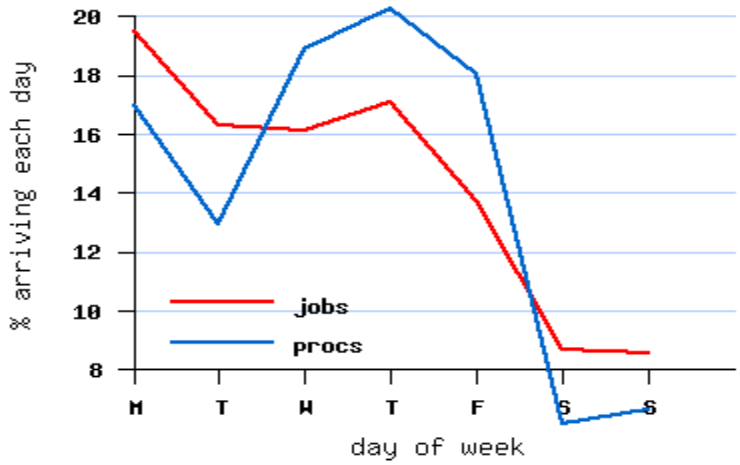

Figure 10: System workload status chart in different days using The RICC log data

Obtained information by analyzing the above data demonstrates system's workload rate has been changed in different data centers (different hours and days). We can conclude that stopping the usage of equipment in deferent days and hours with low workload level brings about decreasing the total efficiency of the above system. Overlapping this obtained information registers that workload sharing based on the mentioned method can set performance level of system to desired level in different times. We fall some segments of equipment to sleep based on the discussed model and in this way, we take steps in decreasing energy consumption and increasing efficiency.

4.4 Advantages of the proposed scheme

Using the proposed scheme has many advantages for the using infrastructure which are defined as followings:

Decreasing energy consumption- according to *Decreasing energy consumption* in various modes of using parts, we can conclude that falling different segments to sleep leads to less energy consumption.

Increasing part's efficiency- in general, when the utmost power of parts are exploited we can claim that the required efficiency has been met. In the above method, when parts have low workload volume, we can fix their productivity in a determined level and in this way, fall another segment to sleep.

Decreasing service presentation cost-different alternatives play significant role in serving customers and these alternatives depend on server's expenses in order to achieve the required level of request. Thus, firms can decrease server's expenses and consequently customer's expenses by decreasing energy consumption, increasing efficiency and applying other firm's equipment.

Decreasing the amount of environmental degradation- destructive effects of technology have been one of the most bothering concerns of governments and organizations within recent years. Using nonrenewable energies such as petroleum, gas and electricity can remarkably affect environment in many aspects; this method helps us to take effective steps in saving environment by decreasing energy consumption in data centers.

Possibility of implementation in data centers with different infrastructure- the above method can be used in different hardware infrastructures because of applying sleep mode. The time of sleeping mode and running mode can differ in different infrastructures and it is changeable by applying the mentioned alternatives in controller segment.

5 Conclusion

Growing progress in information technology has led human life to remarkable changes which differ from way of life to ecosystem and environment. Among all other factors, incoming expenses to users have played primary role in this technology's usage. Data center is one of the most applicable infrastructures in information technology which burdens storage, processing and communicational responsibilities. Regarding energy consumption in the above section, we have defined a model to decrease this factor and then, we conclude the proposed model is able to be run in different circumstances with the above infrastructure configuration.

References

1. Li, Y., Y. Shen, and Y. Liu. Utilizing content delivery network in cloud computing. in Computational Problem-Solving (ICCP), 2012 International Conference on. 2012. IEEE.
2. Hosanagar, K., et al. Optimal pricing of content delivery network (CDN) services. in System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. 2004. IEEE.
3. Held, G., A practical guide to content delivery networks. 2010: CRC Press.
4. Alger, D., The Art of the Data Center: A Look Inside the World's Most Innovative and Compelling Computing Environments. 2012: Prentice Hall Press.
5. Cloud computing slows energy demand, U.S. says. 2016; Available from: http://www.computerworld.com/article/3089073/data-center/cloud-computing-slows-energy-demand-us-says.html.
6. Djemame, K., et al. Energy efficiency embedded service lifecycle: Towards an energy efficient cloud computing architecture. in CEUR Workshop Proceedings. 2014. CEUR Workshop Proceedings.
7. Feitelson, D.G. and B. Nitzberg. Job characteristics of a production parallel scientific workload on the NASA Ames iPSC/860. in workshop on job scheduling strategies for parallel processing. 1995. Springer.
8. Shai, O., E. Shmueli, and D.G. Feitelson. Heuristics for resource matching in intel's compute farm. in Workshop on Job Scheduling Strategies for Parallel Processing. 2013. Springer.
9. Dorier, M., et al. CALCioM: Mitigating I/O interference in HPC systems through cross-application coordination. in Parallel and Distributed Processing Symposium, 2014 IEEE 28th International. 2014. IEEE.
10. Heine, F., et al. On the impact of reservations from the grid on planning-based resource management. in International Conference on Computational Science. 2005. Springer.