Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090455

Anticipation-based Green Data Classification Strategy in Cloud Storage System

Xindong You[1,2,*], *Chi Dong*[1,2,3], *Li Zhou*[1,2,*], *Jie Huang*[1,2] *and Congfeng Jiang*[1,2]

[1] School of Computer ScienceHangzhou Dianzi University, HangZhou 310037 China
[2] Key Laboratory of Complex Systems Modeling and Simulation , Ministry of Education, China
[3] Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis, China

Received: 24 Nov. 2014, Revised: 24 Feb. 2015, Accepted: 26 Feb. 2015
Published online: 1 Jul. 2015

Abstract: The energy consumption problem is one of the critical issues to be addressed in current large-scale storage systems. In order to reduce energy consumption of cloud storage system and meet the performance requirements of users, this paper proposes a green data classification strategy based on anticipation (AGDC), which classify the data in cloud storage system: the hot data stored in the hot disk regions, the cold data stored in the cold disk regions. AGDC employ neural-network prediction on seasonal data, prediting the temperature of data in the next period, executing seasonal data migration in cold and hot regions. This paper also adopts a new correlating algorithm on new data, analyzes its relations with old data in the storage system and prediting the data temperature. New energy consumption model also established in this paper. Simulation experiments based on Gridsim showed that the cloud storage system with green data classification strategy based on anticipation has a good effect on reducing energy consumption. At the expense of average response time of 0.005s, proposed algorithm saved about 16% of energy consumption compared TDCS.

Keywords: Energy consumption, Cloud storage systems, Anticipation, Green Data classification, Forecasting

1 Introduction

Cloud storage is an emerging solution which store resources in the cloud for users. The user can at anytime, anywhere, through any device connected to the Internet cloud and easily access the data. Structural model of cloud storage system consists of four layers: 1) storage layer. Storage layer is the most basic part of cloud storage; cloud storage is often a large number of storage devices and the distribution of many different regions; 2) Basic management. Infrastructure management is the core part of the cloud storage, which is also some of the most difficult goal to achieve. Cloud-based storage management can achieve interoperability between multiple storage devices, so that a plurality of storage devices can provide same external service, and provide greater access to more data and better performance; 3) Application interface. Application interface layer is the most flexible part in cloud storage; 4) Access layer. Access layer provides the interface for users to access cloud storage services. Because of the distinct levels of cloud storage system, the system has good scalability which making cloud storage consumption possible.

With the increasing popularity of data-intensive applications and services, large-scale data centers consumed enormous power resource. The energy consumption of storage system accounting for 25% to 35% [12] of data centers energy consumption. With various application requiring storage devices of a 60% annual growth rate, the energy consumed of storage system will not be ignored. The cloud storage has become a trend in the future development of storage; Cloud Storage System has been widely used. Therefore, how to reduce cloud storage devices energy consumption in large data center is an urgent problem need to be solved. A lot of research has been done on cloud storage consumption, part of which is based on data classification, data backup, data placement, data files, such as lifting the policy. However, the classification using data-saving technologies are not related to the new data, also the seasonal data are not well positioned (the new data and seasonal data as defined in 3.1).Therefore, we propose a green data classification strategy based on anticipation

* Corresponding author e-mail: youxindong@hdu.edu.cn, zhouli@hdu.edu.cn

(AGDC), the new data and seasonal data were well positioned in this classification policy.

2 Related Work

A lot of research has been done on cloud storage consumption, part of which is based on data classification, data backup, data placement, data files, data lifting and other policies. Li, Hongyan proposed REST architecture [13]: by slightly changing the data layout strategies, most of the time REST can safely keep a lot of redundant storage node in the standby mode, even in case of power failure can ensure redundant nodes safety. It deploys advanced real-time workload monitoring, which provides flexible power standby or power down nodes to adapt to changes in load. Saiqin Long, Yuelong Zhao et al proposed TPES saving strategies [14]: they are based on variable factors backup management strategy, based on the best total cost of reconfiguring the cluster state transition strategy and tactics based on workload prediction and observation. Liu Jingyu .et proposed S-RAID structure which uses SSD disks mixed with ordinary [6] disk, by closing part of the idle disk to save energy. Experiments show that the hybrid S-RAID5 disk which composed of 12 general disks and two solid-state SSD disk compared with the same level RAID5, energy consumption of hybrid S-RAID5 disk is only 28%.AutoMig [7] Comprehensive parameters such as history file access, file size, equipment utilization .etc., dynamic classification of documents and using the LRU queue maintains files state which in flash memory device. In hierarchical storage systems show that compared with existing methods, AutoMig effectively shortening the foreground I / O response time. For storage system data access frequency differences, E.Pinhero et al proposed PDC (Popular Data Concentration) model [3]: it migrate data periodically to a few hot disk, and set the lower access frequency data on the left of the disk. It saving energy, but PDC has a bad effect on the performance of the system, since most applications in PDC will be focus on a small part of the request to the disk, increasing the fraction of the disk I / O load and the system's I / O response delay time. RiniT.Kaushik et al proposed Green HDFS named Lightning [8,9]: According to the characteristics of the data, the cloud storage system is divided into the Hot Zone and Cold Zone, it put data which not accessed long time in the Cold Zone, which allowing Cold Zone can be as long as possible in off or low state; Frequently accessed data will be placed in the Hot Zone, Hot Zone long period of high energy state. The simulation experiments show that: division of the hot and cold zones can reduce energy consumption by 26%.

All of the above related work shows that the use of data classification technology put different data into different regions of the disk to save energy is feasible and effective. However, all the above classifications using data-saving technologies are not related to the new data, and the seasonal data are not well positioned (the new data and seasonal data as defined in 3.1).Therefore, we propose a green data classification strategy based on anticipation (AGDC), the new data and seasonal data were well positioned in this classification policy. The green data classification strategy based on anticipation compared file access rate with the threshold value. If access rate is greater than the hot threshold, it is hot data. If access rate is less than the cold threshold value, it is the cold data. While access rate is between hot threshold and cold threshold, it will be predicted. When it is hot, it will be moved to the hot regions, on the contrary put it into the cold regions. If new data is written, and the data not exist in the original cloud storage system, we analyze the related degree about property value of new data and property value of old data in hot and cold disk. Find the most relevant old data. The new data will be written into disks which are the same type of old data.

3 Green Data Classification Strategy Based on Anticipation (AGDC)

3.1 Mathematical Model

In this paper, the data will be divided into five category, they are new dataold data hot datacold dataseasonal data. The following is the definition of these different kinds of data.

Data Temperature: The higher average number of data access operation is, the higher data temperature is, on the contrary, the lower data temperature is. Factors that directly affect the data temperature is the times of data to be operated.

New data: data which does not exist or less time in cloud storage system, and the time that data existence is less than threshold.

Old data: data that exists a period of time in cloud storage system.

Cold data: old data that the average number of data access operation is less than cold temperature threshold.

Hot data: old data that the average number of data access operation is greater than cold temperature threshold.

Seasonal data: old data the averages of data access operation is greater than hot threshold some times, data temperature is fluctuating.

Factor that directly affects the data temperature is times of data operation. That factor associates with human behavior. If no major changes occur, human behavior factors and the behavior of the data attribute is generally not much change. So we consider that a data property value indirectly presents a data temperature value. For new data, we use data temperature which the most relevant with that datas property value in old data to determine the temperature.

A data property value contains the name of the data, keywords, content, and storage form and so on.

Time period: time period is defined as a short time interval, time period division will affect system performance, too short time period lead to frequent data movement, and frequent data movement will increase system consumption. Too long time period division lead to the result that effect of system consumption is not obvious.

Time zone: time zone is composed by some time periods; data season is reflected in a time zone. We assume a time zone composed by time periods.

Here are the steps of our data classification algorithm:

1) Determining whether the existence time of data access is less than the existence threshold. If the result is no, the data is old data, go to stage 2); Otherwise, the data is new data, go to stage 4);

2) Comparing average visit frequency value and hot-cold threshold when old data in time zone. If average visit frequency value is greater than hot threshold, the data is hot data, the data will be written into hot disk. If average visit frequency value is less than cold threshold, the data is cold data. The data between hot and cold threshold is judged seasonally, cold data and seasonal data will be written into cold disk. If the data is seasonal data, after it is written into cold disk, go to stage 3);

3) Predicting average access rate of seasonal data at next time by neutral network, determine temperature of the next time. If average access time of the next time is greater than or equal hot threshold, the data will be written into hot disk. Otherwise, the data will be written into cold disk;

4) Analysising the related degree about property value of new data and property value of old data in hot and cold disk. Find the most relevant old data. The new data will be written into disks which the same type of old data.

We assume that $F = \{f_1, \dots, f_u, f_v, \dots, f_m\}$, F represents file collection, $f_i = \{p_i, m_i\}$ where p_i represents monthly access rate in the two years before current date of file named f_i,m_i represents property values of f_i ,which include hash values of file keywords (there are five), hash value of file name hash value of file content. p_{ij} represents the no. j elements in array p_i, ap_i represents average of array p_i and $ap_i = \frac{1}{24}\sum_{j=1}^{24} p_{ij}$.$th_c$ represents the threshold that has been judged as cold data, th_h represents the threshold that has been judged as hot data.

Algorithm 1 green data classification strategy based on anticipation (AGDC)

Require: $p_i, m_i, th_c, th_h, th_{ex}, n$

for each time window T do **do**
  for each file **do**
    if the time data exists is bigger than existence threshold th_{ex} **then**
      The data tag for the old data;
      $ap_i = \frac{1}{2n}(j=1)^{2n} p_i j$;
      if $ap_i \le th_c$ **then**
        The data tag for the cold data;
        put the files in the cold node;
      else
        if $ap_i \ge th_h$ **then**
          The data tag for the hot data;
          put the files in the hot node;
        else
          The data tag for the seasonal data;
          execute neural network prediction for seasonal data, get y_i;
          if $y_i \ge th_h$ **then**
            put the files in the cold node;
          else
            put the files in the hot node;
          end if
        end if
      end if
    else
      if the time data exists is smaller than existence threshold th_{ex} and $p_i != ''$ **then**
        The data tag for the new data;
        if $ap_i \ge th_h$ **then**
          put the files in the hot node;
        else
          put the files in the cold node;
        end if
      else
        uses the correlation analysis for the new data;
        predicts the temperature of new data;
        if the temperature of new data $\ge th_h$ **then**
          put the files in the hot node;
        else
          put the files in the cold node;
        end if
      end if
    end if
  end for
end for

3.2 Real-time Monitoring Model

Real-time Monitoring Model is also designed in this paper, which monitored the entire system to prevent the occurrence due to some uncertain factors that lead to cold data suddenly gets hot and data from hot to cold.

Real-time Monitoring rules:

1) For data in cold disk suddenly gets hot, once monitoring data becomes hot, the data will be migrated and must be marked. When the time that data becomes cold is equal with time threshold named t_{tc}, the data will be migrated.

2) For data in hot data suddenly gets cold, once the cooling time reaches the threshold data t_{hc}(t_{hc} is the sum of a plurality of time periods size), the data will be migrated.

3.3 Energy consumption model

According to the green data classification strategy based on anticipation (AGDC) model, we established the corresponding model of energy consumption.

$F = \{f_1,...,f_u,f_v,...,f_m\}$,F is a set of files,$F_h = \{f_1,...,f_h,...,f_u\}$,$F_h$ is a set of popular files.

$F_c = \{f_v,...,f_c,...,f_m\}$,$F_c$ is a set of unpopular files,

$F = F_h \cup F_c$ and $F_h \cap F_c = \emptyset$. $f_i = (s_i,\lambda_i)$,s_i is the size of the file f_i,$\lambda_i$is the frequency of arrived of the file,$\lambda_c^{'}$ is the average frequency of arrived of the cold file ,$\lambda_h^{'}$is the average frequency of arrived of the hot file. In this paper we assume that access to the file obey Zipf-like distribution. $\theta = \frac{\log A/100}{\log B/100}$, the ratio of popular files and unpopular files is $\varphi = (1-\theta)/\theta > 1$, then

$$\frac{|F_h|}{|F_c|} = \varphi, |F_c| = \frac{|F_h|\theta}{(1-\theta)} \tag{1}$$

$R = \{r_1,...,r_b,r_p,...,r_x\}$,$R$ is a set of task request,$R_h = \{r_1,...,r_h,...,r_b\}$,$R_h$is a set of request of hot files.$R_c = \{r_p,...,r_c,...,r_x\}$,$R_c$ is a set of request of cold files,where $R = R_h \cup R_c$ and $R_h \cap R_c = \emptyset$,the ratio of request of popular files and request of unpopular files is η,then

$$\frac{R_h}{R_c} = \eta, \eta > \varphi \tag{2}$$

In this paper we assume that each disk may be configured to high speed rotation mode and low speed mode. Once mode is set, it cannot be dynamically scheduled to another mode in the middle of the disk servicing, but can be assigned to another mode by administrator.

$D = \{d_1,...,d_e,d_f,...,d_n\}$,$D$ is a set of disk,$D_h = \{d_1,...,d_h,...,d_e\}$,$D_h$ is a set of rapidly rotating disk. $D_c = \{d_f,...,d_c,...,d_n\}$,$D_c$ is a set of relatively slow rotating disk,$D = D_h \cup D_c$ and $D_h \cap D_c = \emptyset$.$t^h(Mb/s)$ is the transmission rate of rapidly rotating disk, $p^h(J/Mb)$ is positive energy consumption of rapid disk, $i^h(J/s)$ is energy consumption of rapid disk at ideal moment, t^c (Mb/s) is the transmission rate of slow rotating disk, $p^c(Mb/s)$ is positive energy consumption of slow disk, i^c(J/s) is energy consumption of slow disk at ideal moment. From relevant literature we know that the transmission rate between disks is related to positive energy consumption and energy consumption at ideal moment. [1], so making $\frac{t^h}{t^c} = \delta, \delta > 1$, $\frac{p^h}{p^c} = max(1,\frac{2\delta}{3})$, $\frac{i^h}{i^c} = max(1,\delta/2)$. Making $S_h^{'}$ is average size of hot request files, $S_c^{'}$ is average size of cold request files. Assuming that $\frac{S_h^{'}}{S_c^{'}} = k, 0 < k < 1$, as a result of the popularity of files is inversely proportional with the size of a file, making $S_h^{'}=\xi/R_h$,$S_c^{'} = \xi/R_c$,then

$$\frac{S_h^{'}}{S_c^{'}} = \frac{\xi/S_h^{'}}{\xi/S_c^{'}} = \frac{\varphi}{\eta} = k \tag{3}$$

$$|D_h|/|D_c| = \frac{|F_h| * S_h^{'} * 3}{|F_c| * S_c^{'} * 1} = 3\varphi k = \gamma \tag{4}$$

Making energy consumption that the requestor r_h of a set of request R_h request to visit file $f_h(f_h \in F_h)$

$$e_h^{active} = S_h * p^h$$

The average service time of requestor r_h request to visit file f_h

$$at_h^{active} = S_h/t^h$$

Total time of test is t_{ceshi}, total energy consumption that the set of hot request R_h visit files is:

$$e_{R_h}^{active} = \sum_{h=1}^{|R_h|} e_h^{active}$$

The total service time of the set of hot request R_h to visit files is

$$at_{R_h}^{active} = \sum_{h=1}^{|R_h|} at_h^{active}$$

The energy consumption of hot disk at ideal moment in test time is:

$$e_{hot}^{idle} = i^h * (|D_h| * t_{ceshi} - at_{R_h}^{active})$$

The energy consumption of hot disk is

$$\begin{aligned} e_{hot} &= e_{R_h}^{active} + e_{hot}^{idle} \\ &= \sum_{h=1}^{|R_h|} S_h * p^h + i^h * (|D_h| * t_{ceshi} - \sum_{h=1}^{|R_h|} S_h/t^h) \end{aligned} \tag{5}$$

Similarly, the energy consumption of cold disk is

$$\begin{aligned} e_{cold} &= e_{R_h}^{active} + e_{hot}^{idle} \\ &= \sum_{c=1}^{|R_c|} S_h * p^c + i^c * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c) \end{aligned} \tag{6}$$

Then the total energy consumption when we distinguish cold disk from hot disk is

$$e_{total} = e_{hot} + e_{cold}$$

The formula 1–6 into which available

$$\begin{aligned} e_{total} &= \frac{3+2\delta\varphi}{2\delta\varphi}|R_h|S_h^{'}p^h + \frac{2+\gamma\delta}{\gamma\delta}i^h|D_h|t_{ceshi} \\ &- (1+\frac{2}{\varphi})\frac{i^h|R_h|S_h^{'}}{t^h} \end{aligned} \tag{7}$$

Similarly, the total energy consumption when we not distinguish cold disk from hot disk is

$$\begin{aligned} e_{total}^{'} &= (1+\frac{1}{\varphi})|R_h|S_h^{'}p^h + (1+\frac{1}{\gamma})i^h|D_h|t_{ceshi} \\ &- (1+\frac{1}{\varphi})\frac{i^h|R_h|S_h^{'}}{t^h} \end{aligned} \tag{8}$$

Then we can confirm that the saved energy of the temperature of data by data sorting algorithm is

$$\begin{aligned} e &= e'_{total} - e_{total} \\ &= \frac{1}{\varphi}(1-\frac{3}{2\delta})|R_h|S'_h p^h + \frac{1}{\gamma}(1-\frac{2}{\delta})i^h|D_h|t_(ceshi) \\ &+ \frac{1}{\varphi}\frac{i^h|R_h|S'_h}{t^h} \end{aligned} \quad (9)$$

3.3.1 Theoretical analysis of energy saving

$$e_{total} = e_{hot} + e_{cold}, e'_{total} = e'_{hot} + e'_{cold}, e'_{hot} = e_{hot}$$

$$e_{cold} = \sum_{c=1}^{|R_c|} S_c * p^c + i^c * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c),$$

$$e'_{cold} = \sum_{c=1}^{|R_c|} S_c * p^h + i^h * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^h),$$

As $t^h > t^c$,then

$$\sum_{c=1}^{|R_c|} S_c/t^h < \sum_{c=1}^{|R_c|} S_c/t^c),$$

So

$$|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c < |D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^h$$

Because $i^c < i^h$,So

$$i^c * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c) < i^h * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^h);$$

As $p^c < p^h$,then

$$\sum_{c=1}^{|R_c|} S_c * p^c < \sum_{c=1}^{|R_c|} S_c * p^h$$

And

$$i^c * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c) < i^h * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^h),$$

So $\sum_{c=1}^{|R_c|} S_c * p^c + i^c * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^c) < \sum_{c=1}^{|R_c|} S_c * p^h + i^h * (|D_c| * t_{ceshi} - \sum_{c=1}^{|R_c|} S_c/t^h)$, so $e_{cold} < e_{cold}$;

So $e_{cold} < e_{cold}$,then $e > 0$ so the percentage of energy saving must be greater than 0. Through the above proof we know that the green data classification strategy based on anticipation(AGDC) which positioning different types of data to different disk regions has a great consumption-saving performance.

4 Simulation results and performance analysis

In this paper the above algorithm is integrated into the GridSim simulator and the corresponding energy consumption parameters are added to the node in order to evaluate the performance of the algorithm. To verify the effectiveness of the algorithm, this paper used the performance of cloud storage system of not integrated data classification algorithm(HDFS), integration general classification algorithm (TDCS) and combined green data classification strategy based on anticipation(AGDC).By comparing the performance impact of the different request arrival rate, the number of disk, different tasks arrival (a Zipf like distribution is subject to different index),and the new data of different ratio ,which evaluate our proposed strategy.

4.1 Parameters Explanation

The experiment is based on the realization of hot and cold disk array simulator and parameters of disk is shown in table 1. Characteristics workload characteristics and disk

Table 1: disk related parameters

Average positioning time of node disk	5.4ms
Storage capacity	128TB
Shaft speed	10000RPM
High speed disk transfer rate	31M/S
Energy consumption of task of high-speed disk	78.08J/Mb
High speed disk energy consumption when the task does not exist	5.26J/S
Average bandwidth among nodes	4MB/S
Low speed disk transfer rate	9.3M/S
Energy consumption of task of low-speed disk	55.04J/Mb
Low energy consumption of disk when the task does not exist	2.17J/S

drive, the two kinds of parameters will directly affect the final data in our experiments. And a large number of parameters affecting the characteristics of the workload, we identify five key characteristics:

(1) ***The number of files***

The total number of files directly determines the load disk array is assigned a parallel disk array, and set it to 500, with 16 disk drive array so that each disk can hold about 312 case files. The number of each file on disk is determined to be based to the situation of imitating the real world the reality.

(2) ***Request arrival rate***

Due to the request arrival rate directly affect the work at the time of testing times and storage system, it affects the system energy consumption. This paper assumes that *the request arrival rate was 20 25 30 35 40 45 (/S), the default value is 35.*

(3) ***The proportion of hot and cold requests***

Because the proportion of hot and cold file access directly affects the reading times of hot and cold regions, and affects the entire memory system energy consumption. The different indexes of a Zipf like distribution affect different request, while the index, which represents the percent of A access is access to B percent of the file, so we will be the value of A:B and the value is 60:40, 70:30, 80:20, the default value is 70:30.

(4) ***The coverage of file system***

The percentage coverage of the entire file system Defines as accessing request workload of the file system file. We set up the system coverage and the value is 100%, which means all the files of the file system in the parallel disk array system access had at least one.

*(5)****The ratio of hot and cold disk number***

Setting reasonable heat cold disk number ratio can effectively save the energy consumption, according to the previous formula, we set the hot and cold disk and the value is 1:3.

Table 2: the related experimental data description

The total number of files	1000
The request arrival rate	35(20,25,30,35,40,45)
The request arrival distribution (Zipf like distribution)	Inclination(A:B=75: 25)-(60:40 70:30 75:25 80:20 85:15)
The total number of disks	16(12 16 20 24 28 32 36)
The file size of distribution random(random distribution)	(1,20)-(20,30,40,50)
The proportion of new data for the the entire system	10%(1%,5%,10%, 15%, 20%)
The simulation time (time zone * time * each time the simulation time)	3*5*200 S (3 means 3 time zones;5 means 5 time periods;)

4.2 The Effect of the Number of Disks on the System Performance and Energy Consumption

The following experiments are about the effect of disk number on different system performance and energy consumption. The total number of disks are respectively 12 16 20 24 28 32 36.We default hot disk ratio was 3:1 in the experiment, while the results are shown in Fig.1 and Fig. 2.

As shown in Fig.1 and Fig. 2, along with the increase of number of disks, the average response time of three systems were gradually reduced, and the response time is getting closer. The HDFS system response time is the lowest, while the AGDC system is the longest, but

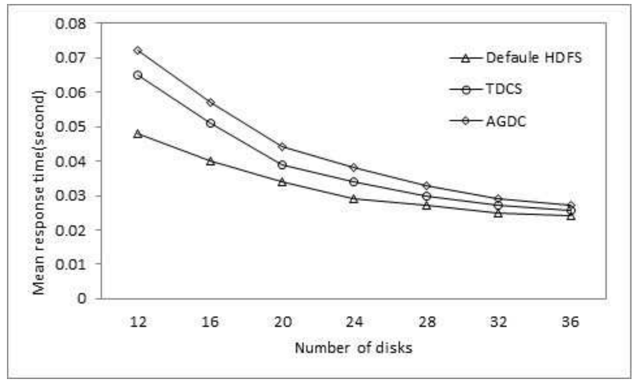


Figure 1: The number of disk impact on system performance.

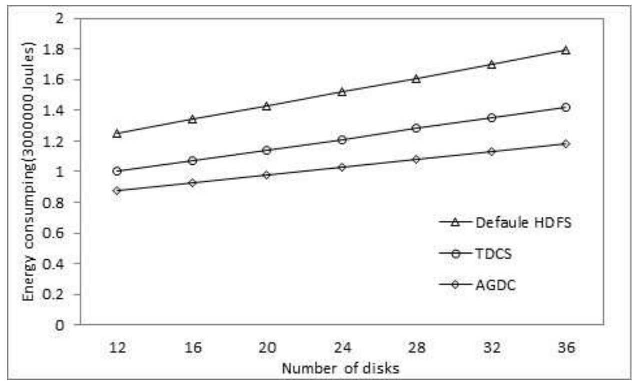


Figure 2: Number of disks impact on the system energy consumption.

between AGDC system and TDCS system response time is little difference.

At the same time, along with the increase of number of disks, the energy consumption gradually increased, energy consumption of HDFS system is the biggest. Energy consumption of AGDC system is the least, while the saving effect obviously and the good performance of the system meet the user's demand. Thus, along with the increased number of disks, the advantages of AGDC system became clear, visible AGDC system has a good effect in saving disk utilization is not high cloud storage system, and good system performance.

4.3 The Impact of Aggregate Access Rate on System Performance and Energy Consumption

The followings experiment the impact of the type of request arrival rate on system performance and power consumption. The experiments take the request arrival rate respectively as 20, 25, 30, 35, 40 and 45. The results are shown in Fig.3 and Fig. 4.

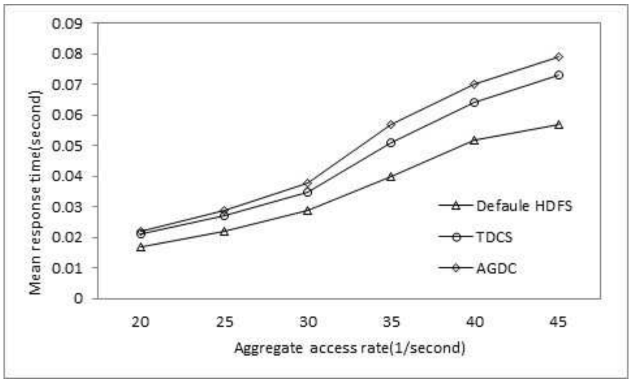


Figure 3: The aggregate access rate impact on system performance.

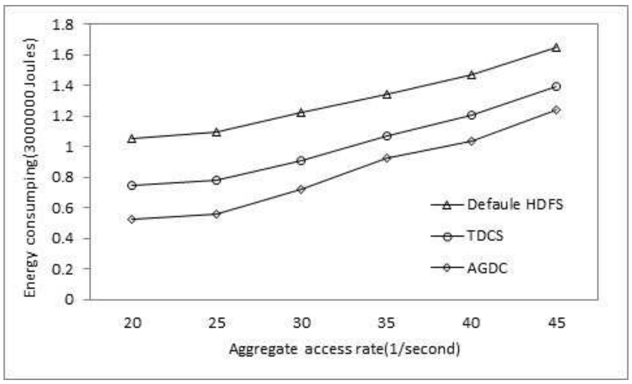


Figure 4: The aggregate access rate impact on system energy consumption.

The Fig.3 and Fig.4 show that, with the request arrival rate increasing, the average response time of three systems increases also the increasing trend gradually slows, and the response time difference between AGDC and TDCS system gradually increases but slowly that is not significant. On the other hand, the energy difference between AGDC and TDCS system is still very large although there is a slowing trend, where the energy consumption of the HDFS system is the highest. In summary, under the premise of meeting the system performance requirements, the energy reduce effect of AGDC system is obvious.

4.4 The Effect of Skew Degree Values on System Performance and Energy Consumption

The followings experiment the effect of different slope value A: B on system performance and energy consumption. The experiments take the values of A: B respectively as 60:40, 70:30, 75:25, 80:20, 85:15 and the default total number of the disk is 16. The results are shown in Fig.5 and Fig.6.

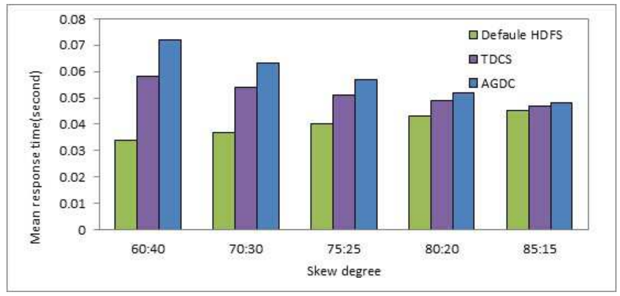


Figure 5: The skew degree value impact on system performance.

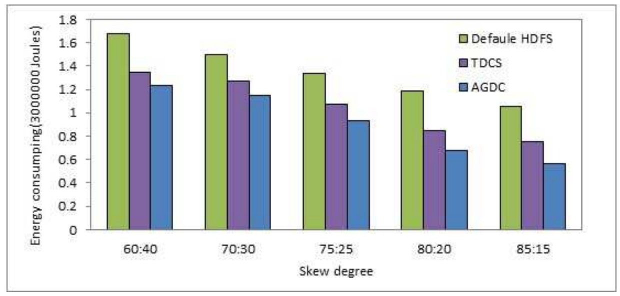


Figure 6: The skew degree value impact on system energy consumption.

As shown in Fig.5, with the increase of the ratio of the slope value, the response time of HDFS system increases, while the response time of the other two systems is reduced. When the heat request is intensive, the response time of the three systems is very close. As shown in Fig.6, with slope values increasing, the difference of energy consumption between the three systems is still very obvious, where the HDFS's energy consumption is highest, the AG-DC's is lowest, and so that the energy reduce effect of AGDC system is obvious. Fig.5 and Fig. 6 illustrate that in the case of intensive heat request, the AGDC system can still achieve a good energy saving effect under the premise of meeting the system performance requirements. Thus in the case of heat-intensive requests, system performance between AGDC system, TDCS system and HDFS system is very close, and AGDC system saving effect is obvious, AGDC system has great advantages.

4.5 The Effect of the Proportion of New Documents on System Performance and Energy Consumption

The followings experiment the effect of different proportion of new files on system performance and power consumption. The experiments take the new file ratio

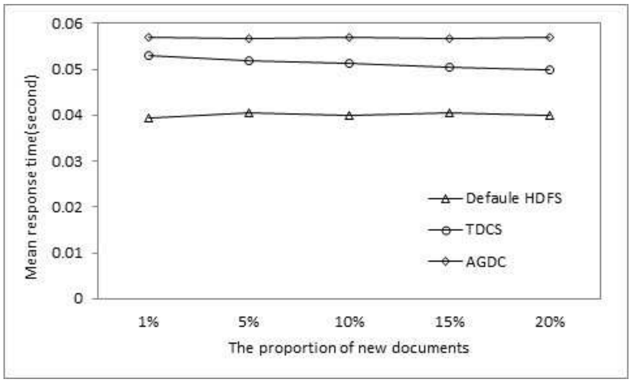


Figure 7: The proportion of new documents impact on system performance.

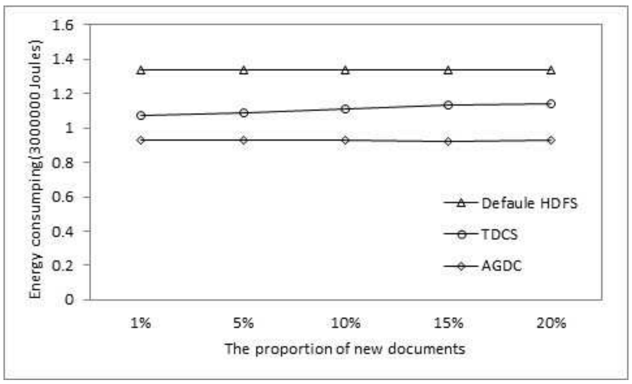


Figure 8: The proportion of new documents impact on system energy consumption.

respectively as 1%, 5%, 10%, 15%, 20%, and the default total number of the disk is 16. The results are shown in Fig.7and Fig.8.

Fig.7 shows that, with the increase of the proportion of new document, the response time of HDFS and AGDC system is substantially constant, while the response time of TDCS system decreases gradually, and the reduce tendency tends to flatten. The response time of the three systems all can meet users needs. As shown in Fig.8, the energy consumption of both AGDC and HDFS system are basically unchanged, while the energy consumption of TD-CS increases gradually, where, the energy consumption of AGDC system is lowest, and the saving effect is obvious. Fig.7 and Fig. 8 show that the AGDC can still achieve good energy saving results under the case of meeting user requirements with the increase of new files. The review, AGDC system in a large proportion of the new file system has a good saving advantage.

5 Conclusions and Future work

This paper presents a green data classification strategy based on anticipation (AGDC). Simulation experiments results show that the proposed classification strategy classifys new data and seasonal data ideally. We also establish an energy consumption model. Based on the model, energy consumption is compared between two cloud storage systems, one with classified the other without. Simulation results show that this model has a very good effect on reducing energy consumption, the established model saving substantial energy under the following condition :intensive request, large proportion of new documents and intensive heat requests and other. AGDC classified the data files reasonably, through the rational and effective classification, almost all the energy consumption and the increase of document reading delay are avoided when the hot files in cold area or the cold files in hot area. However, this strategy also resulted in some additional cost and some imprecise data locating. When compared to TDCS, our proposed algorithm saved about 16% energy consumption at the expense of increasing 0.005s average response time. Our future work is to minimize the additional cost as much as possible while reducing energy consumption.

Acknowledgment

This work is supported Natural Science Foundation of China (Grant No. 61402140, 61202094,6110004,61100193),by Natural Science Foundation of Zhejiang Province (LY13F0-20045, LY13F020047), Research Project of Education Office of Zhejiang Province(No. Y201222937,Y201432390), National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2012BAH24B04).

References

[1] Tao Xie. 2008. SEA: A Striping-Based Energy-Aware Strategy for Data Placement in RAID-Structured Storage Systems. IEEE Transactions On Computers. Vol 57, No 6, June 2008.

[2] Zhu, Q. and Zhou, Y. 2005. Power-aware storage cache management. IEEE Trans. Computer. 54, 587-602.

[3] Pinheiro E, Bianchini . Energy Conservation Techniques for Disk Array-Based Servers [C]. Proceedings of the 18th International Conference on Supercomputing (ICS).New York, NY, USA: ACM, 2004: 68-78.

[4] Andrew Krioukov, Sara Alspaugh, et al.2012. Design and Evaluation of an Energy Agile Computing Cluster. Electrical Engineering and Computer Sciences University of California at Berkeley. Technical Report No UCB/EECS-2012-13. Jan 17, 2012

[5] Weddle COldham MQian Jinet ak PARAID: A Gear-Shifting Power-Aware RAID [C].Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST). Berkeley, CA, USA: USENIX. 2007: 245-260

[6] Liu Jingyu, Zheng Jun, Li Yuanzhang, Sun Zhizhuo, Wang Wenming. Hybrid S-RAID: An Energy-Efficient Data Layout for Sequential Data Storage [J]. Journal of Computer Research and Development. 2013, 50(1)

[7] Zhang Guangyan, Qiu Jianping. An Approach for Migrating Data Adaptively in Hierarchical Storage Systems [J]. Journal of Computer Research and Development. 2012,49(8)

[8] Kaushik, R. T. and Bhandarkar, M.2010.GreenTDCS: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster. In Proceedings of the 2010 international conference on Power aware computing and systems.HotPower '10. USENIX Association, Berkeley, CA,USA, 1-9.

[9] Kaushik, R. T., Cherkasova, L., Campbell, R., and Nahrstedt, K.2010. Lightning: self-adaptive, energy-conserving, multi-zoned, commodity green cloud storage system. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. HPDC '10.ACM, New York, NY, USA, 332-35.

[10] Kaushik, R. T. and Bhandarkar, M.2010. GreenTDCS: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster. In Proceedings of the 2010 international conference on Power aware computing and systems. HotPower '10. USENIX Association, Berkeley, CA, USA, 1-9.

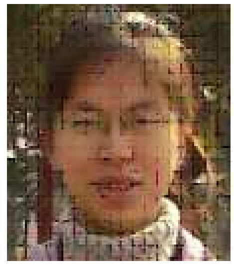
[11] Zhu, Q. and Zhou, Y. 2005. Power-aware storage cache management. IEEE Trans. Computer. 54, 587-602.

[12] Kim, J. and Rotem, D.2011.Energy proportionality for disk storage using replication. In Proceedings of the 14th International Conference on Extending Database Technology. EDBT/ICDT'11. ACM, New York, NY, USA, 81-92.

[13] Li H. REST: A Redundancy-Based Energy-Efficient Cloud Storage System[C]//Proceedings of the 2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE Computer Society, 2012: 537-542.

[14] Saiqin Long, Yuelong Zhao, Wei Chen, A three-phase energy-saving strategy for cloud storage systems. Journal of Systems and Software, v 87, n 1, p 38-47, January 2014.

[15] Long S Q, Zhao Y L, Chen W. MORM: A Multi-objective Optimized Replication Management strategy for cloud storage cluster[J]. Journal of Systems Architecture, 2013.



Xindong You is a lecturer of School of Computer Science and Technology, Hangzhou Dianzi University, China. She is with the Grid and Service Computing Lab in Hangzhou Dianzi University.Before joining Hangzhou Dianzi University, she was a PhD candidate in Northeastern University from 2002 to 2007. She received her PhD degree in 2007. Her current research areas include virtualization, distributed computing, etc.

Chi Dong a Master candidate of School of Computer Science and Technology, Hangzhou Dianzi University, China. She is with the Grid and Service Computing Lab in Hangzhou Dianzi University. Before joining Hangzhou Dianzi University, she was a University student in JIAXING University from 2008 to 2012. Her current research areas include distributed Cloud Storage, Energy Management, etc.

Li Zhou is an associate professor of School of Computer Science and Technology, Hangzhou Dianzi University, China. She is with the Grid and Service Computing Lab in Hangzhou Dianzi University. Before joining Hangzhou Dianzi University, she was a Master candidate in Hangzhou Dianzi University from2003 to 2008. Her current research areas include cloud computing, virtualizationetc.

Jie Huang is a lecturer of School of Computer Science and Technology, Hangzhou Dianzi University, China. He is with the Grid and Service Computing Lab in Hangzhou Dianzi University .Before joining Hangzhou Dianzi University, he was a PhD candidate in Zhejiang University from 2003 to 2008. He received his PhD degree in 2008. He current research areas include cloud computing, virtualization, geo-computation, etc.

Congfeng Jiang is an associate professor of School of Computer Science and Technology, Hangzhou Dianzi University, China. He is with the Grid and Service Computing Lab in Hangzhou Dianzi University.Before joining Hangzhou Dianzi University, he was a PhD candidate in Huazhong University of Science and Technology from 2002 to 2007. He received his PhD degree in 2007. His current research areas include power aware computing system, virtualization, grid computing, etc. Dr. Jiang is a member of China Computer Federation (CCF), IEEE and IEEE Computer Society.