



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TOWARDS BIG DATA COMPUTING:ENERGY-AWARE ALGORITHMS FOR MAPREDUCE WORKFLOWS

1st Rohit Arun Khalse,
ME Student,

Computer Science & Engineering,
Everest College of Engineering and
Technology,Aurangabad(MH), India,

2nd Dr. V. S. Karwande,
Assistant Professor,

Computer Science & Engineering,
Everest College of Engineering and
Technology,Aurangabad(MH), India,

Abstract— As a result of the exponential rise of data pertaining to millennials, it has become increasingly difficult to store, retrieve, and evaluate the data as required. For the purpose of ensuring that everyone has access to the data whenever they want it, it is kept in large clusters. This diversified collection of clusters is utilised for the replication and storage of files. There are two factors that impact the dependability of the data that is being requested: the file replication factor and the connectivity between clusters. As a result of the increased number of duplicates that are manufactured, the rising demand for replicas leads to an increase in the amount of energy that is consumed as well as the amount of time that is wasted. Unfortunately, as a result of this, the system may become unstable and may not immediately react to the user's request. The system recommends a way of file replication that is trustworthy, efficient, and energy-saving. This approach only duplicates files when there is a significant demand from users. This is done in order to conserve space on clusters. The process of storing files on clusters according to the demand for data has the potential to preserve cluster energy and choose clusters depending on the capacity of their replica storage spaces. A number of different strategies may be used in order to achieve performance goals. These strategies include streamlining the process of file read latency for users, minimising congestion for user requests, and developing a strategy for quick file replication in the case of a failure.

Keywords— *Replica, Energy-Efficient Adaptive File Replication (EAFR), Data Replication for Efficient Cluster Scheduling (DARE).*

I. INTRODUCTION

Within the scope of this investigation, an energy-efficient adaptive file replication system aims to improve data availability by taking into consideration the popularity of files and the efficiency of storage systems. The handling, processing, and transport of enormous volumes of files all at once is the responsibility of data-intensive clusters. Additionally, in order to increase the amount of data that is accessible, certain file systems produce and store three copies of each file on servers that are dispersed across a number of racks. To ensure that data is accessible during the random replica destination selection process, it is necessary for all

servers to remain active, which results in the use of power. The random selection of replica destinations does not take into account the bandwidth of the destination or the capabilities of the request processor. As a result, network congestions may occur as a consequence of connection capacity constraints, and servers may get overburdened with data requests. The handling, processing, and transport of enormous volumes of files all at once is the responsibility of data-intensive clusters. Additionally, in order to increase the amount of data that is accessible, certain file systems produce and store three copies of each file on servers that are dispersed across a number of racks. Both the availability of data and the efficiency of the file system may be further enhanced by making use of the heterogeneity of the file system and the server, which they tend to take for granted. Due to the fact that the popularity of files fluctuates greatly, it is likely that a strict minimum of three copies will not be able to reply quickly enough to an influx of read requests to "hot" files, while at the same time wasting resources (such as energy) on copies of "cold" files that are seldom used. Because servers differ in terms of network speed, hardware design, and capacity, it is essential to select replica servers in order to ensure that there is a minimum delay in both the replication process and the response time to requests. The greatest number of service requests that may be serviced at the same time is referred to as capacity. The EAFR, which stands for Energy-Efficient Adaptive File Replication System, that we have presented is comprised of three components. Through the process of adjusting to the fluctuating popularity of files over time, it achieves a balance between the availability of data and the efficiency of its operations. If a file is popular, there will be more copies of it, but if it is not popular, there will be less copies. In addition, servers are classified as either hot or cold servers according to the amount of energy they consume; cool servers are utilised for the storage of cold files respectively. This contributes to the efficiency of energy use. The next thing that EAFR needs to do is select a server that is capable of storing a copy and has sufficient resources, both physical and virtual. We recommend a few things in order to make EAFR even better: a method for selecting data nodes

that is aware of the network in order to reduce the amount of time it takes for files to be read; a method for maintaining replicas that are aware of the load in order to ensure that backup copies of files are made quickly in the event that a replica node fails; and a method for adjusting the transmission rate dynamically in order to prevent possible congestion while copying files to a server.

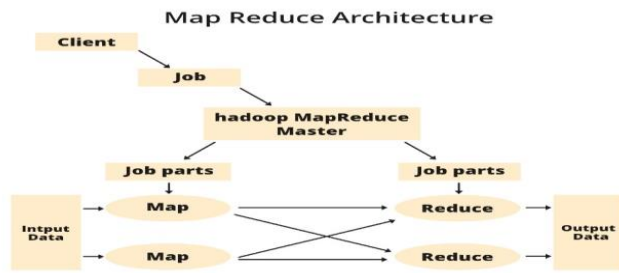


Figure 1.1: Map-Reduce Architecture.

II. LITERATURE SURVEY

This article efficient distribution of resources is of the utmost importance in clusters that are comprised of economical Single-Board Computers (SBCs) that have a restricted amount of processing capabilities. In resource-constrained contexts, where energy efficiency and cost-effectiveness are of the utmost importance, these clusters are increasingly being implemented in edge computing environments. One of the most significant difficulties associated with Hadoop YARN scheduling is load-balancing. This is because frugal nodes within the cluster have the potential to become overloaded, which may lead to a decline in performance as well as numerous instances of out-of-memory problems, which can eventually result in task failures. Within the scope of this research, we provide an Adaptive Multi-criteria Selection for Efficient Resource Allocation (AMS-ERA) in Frugal Heterogeneous Hadoop Clusters. Our criteria takes into account the requirements for tasks in terms of CPU, memory, and disc space, and then matches those requirements with the resources that are available in the cluster in order to achieve optimal resource allocation. We build a heterogeneous SBC-based cluster that is comprised of 11 SBC nodes in order to verify our strategy. We then conduct many tests in order to evaluate the performance of the cluster by utilising the Hadoop wordcount and terasort benchmark for a variety of workload settings inside the cluster. On the basis of these findings, the Hadoop-Fair, FOG, and IDaPS scheduling techniques are evaluated and contrasted. Using terasort and wordcount as benchmarks, our findings indicate that the proposed AMS-ERA leads in a considerable increase in performance, with a reduction in execution time of 27.2%, 17.4%, and 7.6% correspondingly.[1]

The efficient processing of enormous datasets in heterogeneous Hadoop clusters offers a significant problem in this age of ever growing accumulation of vast amounts of data. Traditional job schedulers frequently have difficulty adapting to the dynamic and diversified nature of big data workloads as well as the complexity of heterogeneous settings. This results in inefficient resource utilisation and longer task execution durations. A unique approach that is meant to tackle these issues is presented here in the form of Adoptive Load Balancing schedule, which is the result of this study. The model that we have presented makes use of a dynamic scheduling algorithm that is aware of resources and intelligently adjusts to the peculiarities of real-time workloads and the capabilities of nodes. It is equipped with a strong optimisation engine that includes a Node Profiler, Adaptive

Scheduler, Task Classifier, Performance Monitor, and Load Balancer, all of which work together in a seamless manner to improve task execution and resource utilisation. We have conducted testing on both traditional and current models, such as the default Hadoop schedulers, as well as with our suggested model, and the results have shown that our proposed model is far more effective than the traditional methods. It beat these models in a variety of parameters, including the amount of time it took to execute the work, the efficiency with which data was allocated, the amount of time it took to transmit data, and the amount of resources that were used. It may be concluded that the Adaptive Load-Balancing Job Scheduler Approach represents a significant advancement in the realm of distributed computing for the purpose of big data analytics. By optimising resource utilisation, it not only improves the performance and efficiency of Hadoop clusters, but it also makes a contribution to the economic and environmental sustainability of the distributed computing environment. This technique demonstrates how to provide more responsive, efficient, and scalable solutions for the processing of large amounts of data in contexts that involve heterogeneous computing.[2]

Execution time, computational complexity, and efficiency are some of the primary challenges that traditional parallel computing for power management systems faces. These challenges include process time and delays in power system condition monitoring, particularly consumer power consumption, weather data, and power generation for the purpose of detecting and predicting data mining in centralised parallel processing and diagnosis. Because of these limitations, data management has emerged as an important factor to address in research and a bottleneck in the process. Cloud computing-based approaches have been created for the purpose of handling data in power management systems in an effective manner. This was done in order to deal with the limits that were imposed. The purpose of this study is to examine the notion of cloud computing architecture, which is developed to fulfil the multi-level real-time needs for improving monitoring and performance. This architecture is built for various application scenarios for power system monitoring. After that, cloud computing solutions are explored against the backdrop of big data, and developing parallel programming models like Hadoop, Spark, and Storm are briefly introduced in order to analyse the advancements, restrictions, and innovations that have occurred in these models. A model was developed by using related assumptions in order to estimate the major performance metrics of cloud computing applications. These metrics include core data sampling, modelling, and the analysis of the competitiveness of big data. In conclusion, it presents a novel design idea that is based on cloud computing, and it goes on to provide some recommendations that are centred on cloud computing infrastructure, as well as ways for handling real-time big data inside the power management system, which are necessary in order to overcome the difficulties associated with data mining.[3]

As a framework for the analysis and processing of large amounts of data, Hadoop YARN has emerged as the dominating option. When it comes to energy economy, however, the scheduler that is built into the Hadoop YARN architecture is not meant to be efficient. This article provides an update to the existing Hadoop YARN architecture that is more energy-efficient in order to solve the problem that has been presented. In addition, we describe the MapReduce scheduling in the heterogeneous Hadoop YARN cluster as an energy consumption optimisation issue, and we offer a

heuristic approach to solve this optimisation problem. This algorithm is then used to optimise the energy consumption. In order to enhance the Hadoop YARN cluster's performance and achieve greater energy economy, the suggested technique makes use of load balancing as well as dynamic voltage/frequency scaling. In order to determine whether or not our approach is successful, we conduct exhaustive tests on a real Hadoop YARN cluster that is comprised of five servers. The results demonstrate that our approach is capable of achieving superior performance and delivering substantial reductions in energy consumption when compared to three different approaches that were applied to issues that were comparable.[4]

Companies in the information technology (IT) industry who are working towards the goal of lowering their carbon footprint while also delivering large-scale cloud services face a significant obstacle as energy conservation. Especially in light of the growing amount of data that is dispersed around the globe, these businesses frequently rely on the technology of data replication in order to fulfil the objectives of their tenants, such as performance. We present a static and multi-objective data replication approach (E2ARS) in this research. The purpose of this strategy is to minimise the amount of energy that is consumed as well as the amount of money that is spent by the provider. E2ARS makes use of the heterogeneity of the cloud as well as technologies that are efficient with energy. The first thing that we do is compare the various policies that we have designed for our plan, which range from merely taking into account energy consumption to only taking into account spending. What should come as no surprise is that the more you wish to limit the amount of energy that is consumed, the less you duplicate. After that, we examine E2ARS in relation to several tactics that have been published in the past. In situations when those solutions only accomplish one of the two goals, E2ARS is able to lower both energy consumption and expense thanks to its implementation.[5]

Regarding the success of a nation's economy, the energy policy of that nation is of the utmost importance. Dependence on imports, diversification of energy sources, rising global energy demand, energy efficiency, climate effect, and other issues are among the many things that should be taken into consideration with regard to energy. The estimation of energy consumption and the levels of energy efficiency within the economy are the foundations upon which effective energy planning is built. The article describes the findings of research that was carried out for these two indicators, with the energy system of the Republic of Bulgaria serving as the topic of the investigation. For the purpose of forecasting the energy intensity of the economy, mathematical models are offered. These models are developed via the study of the trend model with time series and by the use of artificial neural networks. Additionally, a neural network model has been constructed for the purpose of forecasting the final energy consumption. This model is based on data regarding the gross domestic product as well as the energy intensity of the environment.[6]

It is the responsibility of the Hadoop distributed file system, often known as HDFS, to store extremely huge data sets in a dependable manner on several clusters of commodity equipment. The High-Density File System (HDFS) offers high throughput data service by utilising replication to fulfil the requests of customers. Replicating data is a compromise that involves sacrificing greater data availability in exchange for increased disc consumption. In recent research, many data replication management frameworks have been proposed.

These frameworks are designed to dynamically adjust the replication factor of files in accordance with the popularity of the data. This allows for the retention of more replicas for data that is in high demand, which ultimately improves the system's overall performance. As the popularity of the data decreases, these methods limit the replication factor, which in turn causes a shift in the distribution of the data and ultimately results in an imbalanced distribution of the data. The cluster experiences hot spots, low data locality, and high network use as a result of an imbalanced distribution of data. The first thing that we have established in this paper is that lowering the replication factor results in an imbalanced distribution of data when the default replica deletion mechanism of Hadoop is utilised. Then, we demonstrate that even maintaining a balanced data distribution by the use of WBRD (data-distribution-aware replica deletion strategy), which we suggested in earlier work, results in sub-optimal performance on heterogeneous clusters. We suggest a heterogeneity-aware replica deletion strategy (HaRD) as a solution to this problem in order to avoid further complications. When removing replicas, HaRD takes into account the processing capabilities of the nodes; as a result, it stores a greater number of replicas on the nodes with more processing capability. We carried out a performance study on a specialised heterogeneous cluster consisting of 23 nodes following the implementation of HaRD on top of HDFS system. When compared to Hadoop and WBRD, respectively, our findings indicate that HaRD was able to minimise execution time by as much as sixty percent and seventeen percent.[7]

The expectations that consumers have for processing power are growing on a daily basis as a result of the development of technology. Computing in the cloud is one of the most cutting-edge technologies that provides users with the ability to fulfil their computationally intensive needs. In order to enable on-demand provisioning of resources, cloud computing has taken use of virtualisation technology. This has led to a growth in the complexity of cloud infrastructure, which means that failures are absolutely unavoidable. Mistakes like this might lead to failure, which would result in significant losses for the organisations. The techniques that are utilised for fault management typically necessitate the utilisation of more resources, which in turn increases the intake of energy. In addition, cloud infrastructure involves a significant amount of energy consumption and is the primary contributor to the amount of carbon content. The combination of rising demand and a finite supply of renewable resources have resulted in severe energy crises. Therefore, energy-efficient fault-tolerant solutions are required in order to tolerate faults and ensure dependable, scalable, and flexible availability of cloud services. This will avoid the failure of the system while simultaneously reducing the amount of energy that is consumed. When it comes to cloud computing services, fault tolerance and energy efficiency are two of the most important factors that need to be taken into consideration concurrently in order to guarantee availability, performance, and dependability. Within the scope of this work, the fundamental ideas of faults, mistakes, and failures are discussed. In addition to this, it addresses the various solutions for fault tolerance as well as the differences and similarities between energy efficiency and fault tolerance.[8]

Hadoop and MapReduce are playing a significant role in the storing and processing of parallel data in shared distributed systems. This includes the processing of distributed data as well. When it comes to the quantity of work and clusters, Hadoop was primarily developed for environments that are similar to one another. Heterogeneity becomes a major difficulty to confront when new technologies are brought to Hadoop. This is owing to the fact that Hadoop-schedulers have produced undesirable outcomes and performance has deteriorated. There is a persistent effort being made by researchers to mitigate the negative impacts that heterogeneity might have. By classifying some of the current schedulers according to the parameters that impact the heterogeneity factor, the purpose of this research is to shed some insight on the aforementioned factor.[9]

When it comes to dealing with data (or storage that holds these data) in large-scale distributed systems, data replication offers an effective technique from which to handle the situation. It is generally accepted that data replication is one of the optimisation strategies that have been implemented in distributed systems, such as database communities. This strategy involves maintaining multiple copies of data at two or more resource sites in order to achieve high levels of performance, availability, and reliability in distributed systems. As a consequence of this, it has come to light that the process of selecting resources, which may be considered a component of resource management, is intricately intertwined with the techniques for data replication and resource management in distributed systems. Therefore, the purpose of this study is to explore the various data replication mechanisms that are now in use in distributed systems, including distributed database management systems (DBMS), data grids, and cloud computing environments. In addition, a study of resource management in distributed systems is presented, with a particular emphasis placed on Grid computing as well as cloud computing.[10]

MapReduce and distributed file systems such as HDFS have emerged as the paradigm for the creation of large-scale applications that require a significant amount of data. This is due to the recent growth of services that are based on the Internet. The amount of data that is produced by social media platforms is really substantial, and as a result, it necessitates a more rapid execution, which in turn presents a significant challenge to the task throughput. In order to complete large-scale data processing, it is necessary to handle a variety of difficulties, including the distribution of data, the scheduling of tasks, and the execution of jobs. Through the use of a hybrid data placement method and an adaptive cluster reconfiguration algorithm, a unique technique is given with the intention of enhancing the distribution of data inside the MapReduce framework. A number of data placement tactics, such as profiling and pipelining in rack aware data placement methodology, are included in the approach. A faster reaction time will be achieved by the implementation of the proposed technique, which will result in increased efficiency for replica placement.[12]

III. PROPOSED METHODOLOGY

This picture depicts the EAFR that has been proposed for this project, which aims to reduce the amount of time it takes for replication to finish, as well as the amount of power that is consumed. EAFR dynamically increases the number of replicas for files that are regularly accessed in order to handle

high request loads and, as a result, reduce the amount of time it takes for files to be read into memory. At the same time, it decreases the number of replicas for files that are viewed infrequently without compromising the read efficiency of those files.

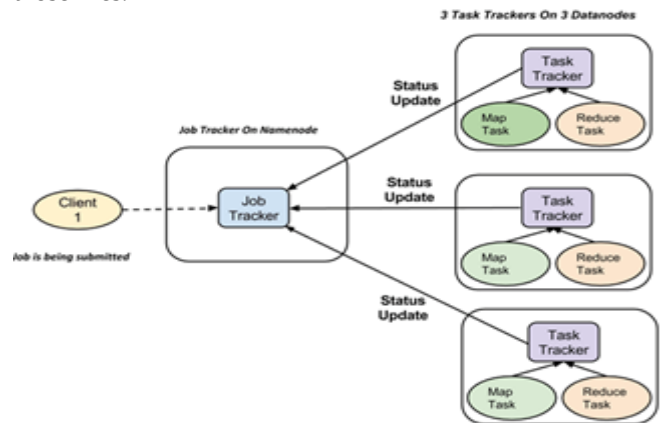


Figure 3.1: System Architecture

In order to improve the availability of data, we have developed a solution for EAFR that takes into consideration the popularity of files as well as the efficiency of the file storage system. Clusters that are heavy on data are characterised by the simultaneous storage, processing, and transportation of many files. Certain file systems produce and keep three replicas of each file on servers that are randomly selected and scattered across many racks. This is done in order to improve the availability of your data. Because the selection of replica sites is completely random, it is necessary to keep all servers running continuously in order to ensure that data is always accessible. This leads to an exorbitant amount of power usage. The random selection of replica destinations disregards the bandwidth of the destination as well as the capability of the request to be processed. This might potentially lead to network congestion due to the capacity constraints of some connections, which could result in the server being overloaded with data requests.

IV. EXPERIMENTAL RESULTS

We have created a system for EAFR with the intention of enhancing the availability of data while taking into account the significance of file popularity and the effectiveness of the file storage system. A significant number of files are concurrently stored, processed, and transported in clusters that are characterised by their high data density. A number of file systems generate and store three copies of each file on servers that are randomly picked from a variety of racks. This is done in order to improve the amount of data that is available. The random selection of replica destinations necessitates the continuous operation of all servers in order to guarantee the availability of data; nevertheless, this increases the amount of power that is used. There is a possibility that network congestions may emerge as a result of capacity limitations on some connections, and the server may become overburdened with data requests. This is because the random selection of replica destinations does not take into account the bandwidth and capacity of the destination to handle requests.

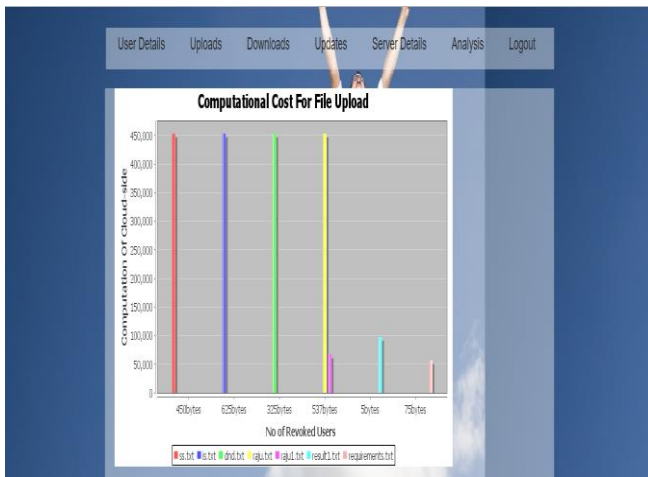


Figure 4.1: Time Analysis

V. CONCLUSION

The needs of data-intensive clusters, which require file systems to have fast file read latency and low power consumption, have been met by the implementation of a system that has been designed to meet those expectations. The system makes use of a dynamic transmission rate adjustment strategy to reduce the likelihood of congestion occurring during the process of file replication to a server. Additionally, it employs a network-aware data node selection strategy to minimise the amount of time it takes for files to be read. Furthermore, it employs a load-aware replica maintenance strategy to efficiently create file replicas in the event that replica nodes fail. The results of experiments conducted on a large cluster that is representative of the actual world reveal that EAFR and the recommended techniques are successful in meeting the requirements of file systems that are present in big clusters. In the future, it will be possible to increase file security by introducing server failure tolerance into the system infrastructure. In addition to improving efficiency, data partitioning methods also improve security. As a consequence of this, the most recent transmission speed of a server may be used as a prediction for the pace at which it will transmit data in the near future. It is possible for EAFR to assign a new file replica without taking into account the usage of links or monitoring the condition of network congestion. The data transfer rate of recent files is taken into consideration while making the decision about the location of the replica. Future study will concentrate on improving data localisation in replica placement and determining the appropriate number of cold servers to maximise energy savings while preserving file read efficiency. Both of these subjects will be addressed in the future.

REFERENCES

1. Basit Qureshi, "Towards Improving YARN Performance for Frugal Heterogeneous SBC-Based Edge Clusters", PREPRINTS, 3 May 2024.
2. Sunkari Mahesh, Dr K. Ram Mohan Rao, "Adaptive Load-Balancing Job Scheduler Approach for Optimizing Big Data Processing in Heterogeneous Clusters", IJCST, Volume 12 Issue 4, Aug 2024.
3. Ahmed Hadi Ali AL-Jumaili, Ravie Chandren Muniyandi, Mohammad Kamrul Hasan, Johnny Koh Siaw Paw and Mandeep Jit Singh, "Big Data Analytics Using Cloud Computing Based

Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations", MDPI, Sensors 2023.

4. Yongqiang Gao and Chong Huang, "Energy-Efficient Scheduling of MapReduce Tasks Based on Load Balancing and Deadline Constraint in Heterogeneous Hadoop YARN Cluster", IEEE, 2021.
5. Morgan Segu, Riad Mokadem, Jean-Marc Pierson, "Energy and Expenditure Aware Data Replication Strategy", IRIT, October 14, 2021.
6. Kostadin Yotov, Emil Hadzikolev, Stanka Hadzikoleva, "Forecasting Energy Efficiency and Energy Consumption in Bulgaria by Examining the Energy Intensity Indicator Using Neural Networks", IEEE, 2020.
7. Hilmi Egemen Ciritoglu, John Murphy and Christina Thorpe, "HaRD: a heterogeneity-aware replica deletion for HDFS", Springer, 2019.
8. Kamaljit Kaur, Kuljit Kaur, "A Survey on Energy-Aware Fault Tolerant Strategies in Cloud Computing", IJCSE, Vol.-7, Issue-5, May 2019.
9. Dr. Anil Sharma and Gurwinder Singh, "Issues and Challenges of multi-node Heterogeneous Hadoop Cluster", IJRAR, Volume 6, Issue 1, January 2019.
10. Wan Nor Shuhadah Wan Nik, Bing Bing Zhou, "A Survey on Data Replication and Resource Management in Distributed Systems", IJATICSE, Volume 8, No.5, Oct 2019.
11. Manal Alalawi and Herbert Daly, "Designing a Hadoop MapReduce Performance Model using Micro Benchmarking Approach", ICAICNF, July 2019.
12. Rutuja J. Deshmukh, Prof. R.A. Kulkarni, "A Survey on Efficient Data Distribution by using Hybrid Data Placement Strategy in MapReduce Framework", IRJET, Volume: 03, Issue: 05, May-2016.
13. Tong Shu and Chase Q. Wu, "Energy-efficient Mapping of Big Data Workflows under Deadline Constraints", Workflows in Support of Large-Scale Science, November 2016.
14. Yuhua Lin and Haiying Shen, "EAFR: An Energy-Efficient Adaptive File Replication System In Data-Intensive Clusters", IEEE, 2015.
15. Rajneesh Kumar, Dr.S.Govindarajan, "Scheduling Techniques for Workload Distribution in YARN Containers", IJEDR, Volume 3, Issue 2, 2015.
16. Lena Mashayekhy and Daniel Grosu, "Energy-Aware Scheduling of MapReduce Jobs for Big Data Applications", IEEE, VOL. 25, 2014.
17. Navin Sharma, David Irwin, and Prashant Shenoy, "A Distributed File System for Intermittent Power", IEEE, 2013.
18. Nitesh Maheshwari, Radheshyam Nanduri, Vasudeva Varma, "Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework", Elsevier, 2011.
19. Wenjun Li, Hailong Yang, Zhongzhi Luan, Depei Qian, "Energy Prediction For MapReduce Workloads", IEEE, 2011.

20. Thomas Wirtz and Rong Ge, "Improving MapReduce Energy Efficiency for Computation Intensive Workloads", IEEE, 2011.
21. Jaliya Ekanayake, Shrideep Pallickara, and Geoffrey Fox, "MapReduce for Data Intensive Scientific Analyses", IEEE, 2008.

