Eurasia Journal of Science and Technology	ISSN: 2663-1024	Volume 2, Issue 2,pp 71-77	April 2020

APPLICATION OF PARALLEL COMPUTING IN CLOUD TECHNOLOGIES

Doudou Guo
School of Computer Science and Engineer, Harbin Normal University, Harbin, Heilongjiang, China
E-mail: 120926248@qq.com

Abstract
Every day multiple advancements are happening in the cloud computing technologies and parallel computing. There is a deeper connection in the evolution of cloud platforms and concepts like virtualization and storage in the cloud. How much ever the advancements are progressing, there are always challenges rising in one way or other. To tackle these cloud computing problems, we have proposed a solution that deals with the scheduling processes and manages the overall throughput in the system. It has included the energy-aware scheduling algorithm within the dynamic scheduling that constitutes for cloud efficiency.
Keywords: MapReduce, Hadoop, Azure, Dryad, Amazon S3 and cloud technology

1.INTRODUCTION

Cloud computing is a huge platform that renders an infrastructure for software applications to do multiple oper- ations and data storage services. It allows a wide number of users to support their businesses and organizations in highly approved and efficient m ethodologies. T here i s no cloud architecture not the concept in the cloud that can stand without the implementation of parallel jobs processing and their execution.
Various organizations use the cloud technologies that have different runtime framework environments like Hadoop, MapReduce, Dryad, Azure, Amazon S3, Twister4Azure and others. All these environments provide open-source infras- tructure for their required software development using dif- ferent resources. The main feature of the cloud environment is that it charges the vendors based on the utilization costs of the resources. And cloud also promotes multiple accesses to private and public business users to use their resources. This provides a great scope of parallel computing processes in the background of the runtime frameworks. Different forms of parallel computation modes like single centralization or decentralization control servers are tailored in the cloud envi- ronment characteristics. As most of all the cloud, these cloud technologies work on a large-scale distributed environment, they rely on the high performance and greater output yielding frameworks invoked in the internal jobs.
This project is going to research on the study of parallel computing techniques in the cloud runtime communications and environment for scheduling the jobs. Alongside, the performance of different cloud platform techniques will also be analyzed and studied. We also achieve a scope in learning the architecture changes from the traditional systems to the current models that are highly in trendsetting. Given a chance, on identifying the trends and challenges in these architectures, there is a greater possibility of determining the best performance model/environment using parallel pro- gramming in the cloud.The main objective of researching these papers is to un-derstand the background and basis of parallel computations and its methodologies in the most
demanded environment, i.e., the cloud architecture. On achieving better knowledge in the ongoing trends and challenges in the cloud computing environment, relevant approaches in parallel computing are identified and provided a proposed solution for optimal use of resources.

2.RELATED WORK

To understand the technologies and concepts of cloud computing in deeper, a wide range of papers have been thoroughly explored. In the paper [1], the authors talk about the scalable properties in the parallel computing processes that are applicable in the Azure cloud technology by using a decentralized and distributed iterative MapReduce called the Twister4Azure. Here the scalability in the cloud is increased and optimized by the implementation of iterative computations in the distributed framework that involves numerous sub-operations. These sub-operations include jobs like scheduling, intermediate communications between the data nodes, broadcast of certain messages and the cache systems. As memory space and load are a constraint to most of the businesses, usage of these techniques makes it more advanced and increase the computation speeds of the systems involved with these cloud systems.
The authors in paper [2] helped in the detailed understand- ing of the latest cloud technologies like Hadoop, Dryad and the role of the MapReduce framework in these technologies. They also included the usage of different tools like Cloud- burst, Cap3 and HEP and briefed about their performance analysis in measures like computation speed, responsiveness and scalability approaches. They have provided a high-level comparison of these tools usage and their results

to compare the model that provides the best performances in various circumstances. They also state that Hadoop is the only technology model that shows better efficacy when analyzed against other techniques.

In paper [3], the author has extensively identified the factors that contribute to the scheduling process in specific. As the main aim of cloud computing is to utilize the most of their rendered services, the author has provided a novel approach of the proposed method that includes the priority,waiting times and other influencing factors in the scheduling process. Its primary focus is to gain in the utilization and optimization of maximum resources including the server. Other research papers help in understanding the intermedi- ate concepts of cloud, components in the cloud platform, its architecture, applications and the problematic areas of development.

3.CLOUD COMPUTING ARCHITECTURE

A.About Cloud

There are various definitions for cloud given by multiple authors and researchers. According to cloud Management services, it is defined as on demand availability of the resources like computation power, data storage services, management resources and networking in the computer sys- tems. An architecture is called as cloud if it has the following fundamental attributes like

Multi tenancy (it means the resources are capable for sharing at different levels in the cloud system)

Massive Scalability (the cloud should have greater bandwidth and storage capacities to promote global- ization of data)

Elasticity (this is the capability where users can rapidly increment or decrement their utilisation of computing resources)

Pay as you go (uses pay money only for the services that they have utilised)

Self-provisioning of resources (it means that users has the ability to add the remote services from any part of the world)

B.Cloud Computing Architecture related with Mobiles

There are multiple cloud-based services that have imple- mented their own unique cloud architectures. The considered system has mobile devices connected to cloud computing architecture as this is the process involved with mobile devices to access data every day.

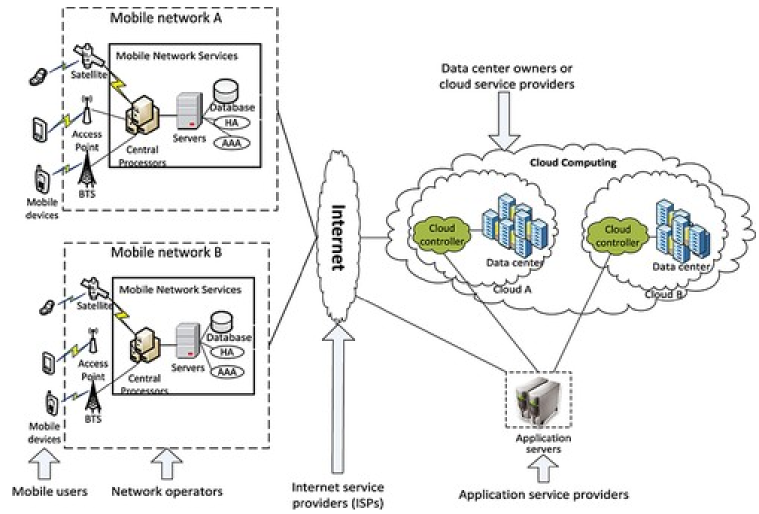

Figure 1 Mobiles included cloud computing architecture [4] .

From the Fig 1 diagram it is seen that the mobile devices are accessible to the base stations like trance transceiver stations (which are also the base stations)that connects the mobile networks in order to create a connection Between the network provider and phones. whenever a request is made from a mobile device, it is carried over the network to the central processing station with in the server of the network service provider. this is the main location where operations like authorization And identification are performed to vali- date the user of mobile devices. On successful authentication of the user, The request is carried through the Internet to the cloud server. The processing of data in the cloud can involve the methodologies of number of processors included in computation, Virtualisation And information providing services implemented in the cloud architecture. as this is one of the popular architectures that is involved in every ones daily life, load balancing and traffic management are also the regular services maintain and managed in the cloud services.

C.Cloud Services

Although there are cloud-based services that define their own distinct and unique cloud architectures, the reference model of the services would be the same. A cloud computing technology is defined as an integration of the computation components (hardware ,software and infrastructure) that has the capability of delivering the cloud services and manage- ment of different applications in different levels. so the cloud computing services are classified into the following three service levels.Software as a Service (SaaS),Platform as a Service (PaaS), Infrastructure as a Service (IaaS).

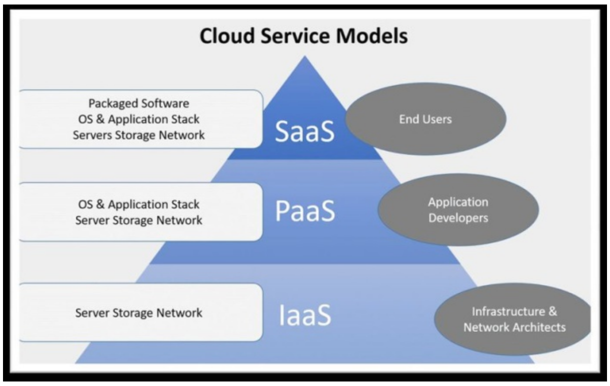

Figure 2 Types of cloud computing [5] .

In the IaaS platform of computing, there is a deployment of different types of virtual machines, load balancers, hosting services for servers and storage applications, network land- scape managements. It is highly integrated with the services and applications involved with the business providers and organizations at a higher level. In PaaS system, the back- end technologies of any application will reside. It is the location for deployment of servers, web application servers, third party tools that are used for back end tracking, and management servers. As the name indicates, it provides a platform to let the application stand on itself in the cloud architecture by making connections and integrations with other services in the cloud. Based on the specialization of the application, this level of cloud is classified for integration Services or data provision services. In SaaS, users are able to access the applications that are hosted in the cloud. They also have terminals that can support the use of emulators, mobile applications and thin client applications. they also provide interfaces for the management tools like CRM applications. All these service models also called the types of cloud computing.

4.ROLE OF PARALLEL COMPUTING IN CLOUD COMPUTING

There are multiple features that are invoked from the concepts of parallel computing and implemented in the cloud computing platform. It is defined t hat p arallel c omputing is a form of computation In which many calculations or sub- events are carried out simultaneously. in this environment, a given problem is broken down into discrete parts where multiple CPU's execution and resources are involved to achieve it. the main advantages of parallel computing are seen in the time conservation and effective management of resources (cost of resources), concurrency in solving a problem and ability to solve more complex problems. These are some of the fundamental attributes for the promotion of application development in the cloud system. As parallel computing is needed to increase the computational power of the systems in relation to the memory frequencies, maximum resources utilisation And virtualisation in the cloud. these features are also the goals that are to be achieved in the cloud systems in an efficient manner. The common concepts that establish a bridge between parallel computing and cloud computing are the following:

Methodology application (For example in the problem-solving approaches and solution development strategies)

Multiple execution units (where an application is mod- ularized and developed by users on different platforms)

Use of non-local resources (this mainly helps in The de- velopment of an application concurrently from various sources available in the cloud)

Provision of concurrency (complex problems and their computations are performed on different processors for better executions)

5.CLOUD TECHNOLOGIES

Cloud technology is defined based on the communicating frameworks and software's that are running on it. most common examples of the cloud that are widely used to nowadays are Amazon S3, Hadoop systems and Dryad. Nowadays these cloud technologies a providing open-source platform for the infrastructure needed further development of different applications for private cloud systems. some of the examples of These open-source software's are Eucalyptus, Nimbus, Cloud Stack, open shift, FOSS cloud etc. Another main purpose of these frameworks is to deploy and help in the management of virtual machine networks, the capability to pool the computation results between the data centres and other resources like network and storage. The inclusion of these cloud systems in the cloud technologies have been created with the new approaches that were developed in the parallel computing concepts. here is the detailed information about the widely demanded cloud systems.

A.Amazon S3

Amazon S3 is one of the popular services from the group of Amazon Web Services(AWS). hey basically S3 is the main domain That is responsible for the storage services provided by the Amazon cloud drive. It is also called as an extension of SaaS type of cloud computing as it provides the storage services to the applications that are deployed in the SaaS cloud layer. The main feature of Amazon S3 is scalability due to its file system type of storage. the files that are stored in the drive are stored in the form of Apache Parquet files, that helps in the file management easier by implementing the columnar format for the storage of files. the main idea of this type of file system is that the visibility of data stored in these columns is increased and helps in the querying of data, accessing remote files easier [6]. In the advanced systems, the data is converted into binary format present in the columns, which helps in the better organization of whole hey data in the file, and its columns that enables to do lossless compressions. It also helps to transfer huge files from one region of the cloud system (data centre) to another location. The entire file system is called S3 bucket and the short form for this type of data storage is called Variant Call Format (VCF) to access the data from these files. It is known that S3 is an extended version of a data storage format which makes it more secure for data alongside its durability and scalability features. It also helps in the delivery of high performance in applications with the help of retrieving the objects using the known keys. this is one of the qualities of Amazon S3 to attract larger applications in adopting this way of storage and data access services provided by AWS.

B.Hadoop

Apache Hadoop is known for its wide range of open-source framework and the software utilities provided for a huge number of computers to reach their objective in data computation and storage in the security systems. This open- source framework is established in a distributed fashion. This feature helps in the processing of big data and it's storage using the MapReduce tool [2]. The 2 main components of The Apache Hadoop are the storage and the Hadoop distributed file systems (HDFS), that plays a significant role in the processing of data and its management. It follows a traditional file system approach for data storage and handles the entire system in a master-slave working system. hear the master note is responsible for maintaining the namespaces of the file systems where is the slave's help in retrieving the actual data nodes that are requested by the Master node. The slave nodes can also help in performing the read and write operations requested from the client systems. The authenticated clients can directly request data from the slave nodes as shown in the 3. One of the main modules in the Hadoop systems is Yet Another Resource Navigator (YARN) framework. This helps in the management of computer resources like job applications to processors, managing the execution of these jobs and scheduling the processes. It is a part of the framework and is known to have great scalability which in turn helps in the management of clusters and resource utilisation. Here is the detailed explanation of the distributed framework and data processing tool of Hadoop i.e., the MapReduce. As MapReduce is evolved from the distributed type of data framework, it is widely used in the open-source implementations present in the Google framework [2]. The MapReduce serves in the data administration using the HDFS techniques. It is needed to govern the data accessibility across multiple local and non- local disks that are requested in the HDFS API systems. It has got reliable capabilities to manage and deploy data in a different type of network structures. Implementation of MapReduce in the Hadoop makes it more advanced with the handling of data communications and multiple program execution management. It is highly helpful in the data op- erations within different nodes, disk management operations and computations required for process scheduling.

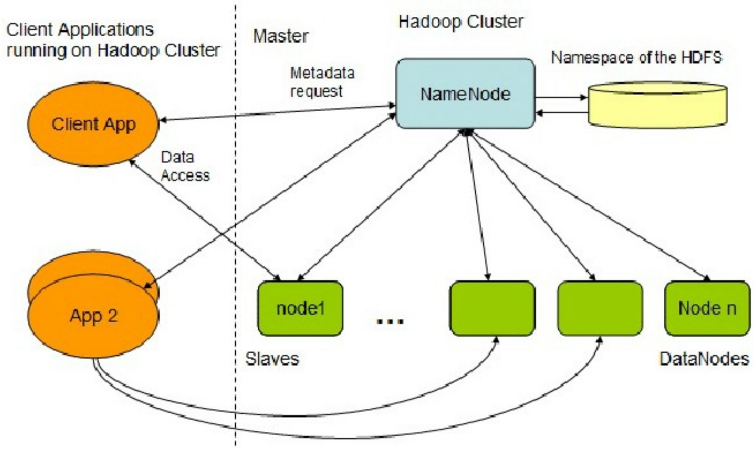

Figure 3 Types of cloud computing [2] .

C.Dryad

Dryad is a distributed computing system that was designed and developed during a research project conducted at Mi- crosoft Research. It was created during the evaluation of runtime execution that was conducted on the parallel data applications. it comprises of a group of language extensions that helps in The usage of different modes in programming in the distributed framework of large scale applications.it is defined as a domain for a new language that uses C++ libraries and distributed parallel program outputs in the execution of a dryad implemented application. It basically helps in the graph-structured data flow diagram where the vertices of the graph are used in calculating the runtime execution of the application. It is also structured by the use of TCP pipes for the communication channels. As dryad is a distributed framework implemented on large scale applications, it has other features like job monitoring, process management through visualization and transportation of data between these communication channels (that connects the vertices in a graph) that are arranged on the individual systems. The DAG systems are further combined with the best algorithms that derive job priorities and further assign these scheduling orders to the job manager.

Table 1 LED COLOUR FOR EACH SENSOR.

	Hadoop	
Data Processing model	Map Reduce	Directed graphs(DAG)
Process Scheduling	FIFO schedul- ing	Paths determined by graphs in different topologies
Storage type for data	HDFS	NTFS, Cosmos DFS
Data Handling	HDFS	Shared directories and sometimes local nodes or disks
Type of data Commu- nication	Point to point using Http	TCP pipes and files

D.Trends

The trends observed over a period in his field deals with the changes regarding Operating Systems.

Dynamic Parallelism in the operating system has created a new movement in the use of virtualization and virtual machines. It has benefited in the expression and de- scription of parallel execution in various programming languages and in the interface

The spread of internal parallelism is the execution and workflows processed at the server-side applications.

In the process of increasing the throughput of the server systems, a single job is substituted into sub-level jobs at the same time and processed. It also has to make sure that there is no exhaustion caused to the CPU cores with the increased load on these server systems.

New programming model address concerns of implicit or explicit creation of parallelism.

Virtualization has been a driving force for every trend in the world of cloud-based parallel computing.

Transporting of these algorithmic approaches and ideas from these models to the cloud is very vital.

The two main factors that contribute to the greater utilization of CPU cores are the implementation of explicit multithreading and the control level parallel programming.

E.Challenges

There is a high dependency on the multi-core CPUs to invoke the concept of virtualisation in Many levels of the operating systems and in the cloud systems.

The cloud services have to invoke constant standard- ization of the protocols, interfaces with the virtual machines and resource managers to benefit in the quick and economical consumption of the cloud services.

Effective communication and synchronization among processes in the job queues.It is always a difficult job To implement a highly scalable computing unit in the large systems within a limit cost.

6.DYNAMIC PARALLELISM

Dynamic parallelism is the concept of Heterogeneous computing that is highly adapted in the future trends of the server environment. this is a type of scheduling management done at the level of the operating system. This is invoked mainly to achieve better utilisation of computer resources and CPU cores and increase the throughput from servers of the computing systems. as OS kernel is open to adding up to new programming models, different parallelization approaches can be included in the user or kernel libraries of a system. This type of computing is mainly invoked with the use of multithreading libraries in the GPU programming models. The implementation of dynamic parallelism is done in 2 different ways. They are:

Dynamic Partitioning: this is the partitioning of resources generally used in the concepts of virtualisation at various levels. In the latest there can be 4 different types of virtual- ization techniques they are:

Hardware Emulation: This is referred to as one of the traditional approach S of virtualisation. Here it involves the emulation of infrastructures (hardware) like chipsets, processor relevant components used in both instruction management and devices.

Full Virtualisation: It is also known as the native way of virtualisation. This type of virtualisation Creates an interface between the VM included guest OS and the native type of hardware unit in systems. The VM Ware is also called the hypervisor because it creates an instruction set such that it provides a sense of combined unit of OS and the hardware items.

Paravirtualization: This is a similar type of full vir- tualisation but also the most recommended version of virtualization than the remaining. The main difference is that it uses a function to distribute the access between the hardware unit through the integration of virtualization responsible code on the operating system itself. This hybrid system of virtualization has lots of advantages bundled in the distributed framework combined computations.

OS-level Virtualization: In this type, the virtualization is provided on the Operating system itself. It con- tributes its virtualization being connected to different individual systems through the same OS kernel level. It also helps in the natural boosting of the processing which further reflects in the performance of the system.

Dynamic Provisioning: This is generally the management of resource consumption done for the utilities accessed under the constraints and billing.

7.PROPOSED SOLUTION

The work that has been done mainly focuses on maintain- ing the overall throughput of the system while improving the efficiency of job allocation to the processors at hand.

This process of manipulating the job process queues would help in reducing the overall time consumed for the process execution in the cloud. To achieve this we focus on the base issue of efficient job allocation and move towards a smarter version of doing this. I have identified a workflow or an algorithm that can be followed for an efficient imple- mentation of the aforementioned idea. We first work with dependencies between the jobs and the impact they have on each other as the first examining factor to be considered for the implementation. The dependency of jobs can be defined as the relationship between jobs where if one job is waiting for a set of instructions from another job, or waiting for one job to finish first and moves to the waiting state. I am trying to work with a solution where we take this into account while distributing jobs among processors we check for any jobs that have high dependencies and as these would go into a waiting state and may cause a deadlock of the processes we try to get them executed of scheduled on a higher priority basis. Thus, we have avoided the problem of higher wait times for job scheduling and distribution.

The next factor that has been considered is the switching time. Switching times in job allocation can be defined as the time taken for the process to get all the prerequisites loaded up and start the execution of the process. If the process requires a huge number of start-up services it has a longer switching time. This time is usually considered for the scheduler. The other influential factor is execution time. Execution time is the total time taken for the job to finish executing and return the result back to the assigning system. In this, we keep a track of jobs that have a high execution time. Now when we consider working with the Switching and Execution time while scheduling jobs onto the processors we do the following:

The jobs having high switching time is given low priority as it delays the execution of other quicker jobs. The jobs having low switching times are considered first.

The jobs having high execution time goes into further analysis of why the time is so high, is it due to the dependencies or due to the importance of the task. If the task if important we execute it first hence having a high priority. If there is a dependency that is delaying the job execution we follow the above-mentioned scheduling for dependencies.

Now as the problem of job distribution or schedule has been solved, we move ahead to work with the use of resources. As we are working with cloud technologies and platforms we can tap on the benefit of sending different jobs to different servers for execution. On this server, the job is broken down in to further parts and assigned to processors. I am trying to demonstrate how we chose or assign which job goes to which server. As we cannot randomly assign jobs to the servers, we consider a few key factors to assign these jobs, namely the transmission speed, uplink and downlink time and server load tolerance etc. Here the considered factors not only deal with the properties of the job but rather deal with the server properties as well. If the server is not in a position to take up more load we do not assign the job that has a huge switching time to it as it may load up the server. In my ideology, I have considered the optimal load cut off for each server to be 70% of its overall load tolerance. For example, if the load tolerance of a server is 100, any job that needs more than 70 blocks of resource space is ignored. The additional 30% of the space is always reserved for future and computational purposes so that the job doesn't stop executing mid-way or throw memory management errors in run time. Given a particular server, if the uplink and downlink time to and from it is less for a job we execute that job first and give it more priority in the scheduling processes.

We can also consider using different load balancing methodologies that are available which can be further ex- perimented on. Once we have addressed the issue of server selection, we turn towards parallel processing of these jobs. Parallel processing of the jobs consume an enormous amount of energy, to regulate the energy consumption we consider the work of Fredy Juarez titled "Dynamic energy-aware scheduling for parallel task-based application in cloud com- puting" [7] where a model that deals with this problem as a multi optimization problem is presented. Similarly, we can consider the execution time and the energy consumption as two factors to optimize the Multi-Objective Optimization (MOO) problem. Here we try to strike a balance between them by considering the most important factor in each situation or scenario.

Map Reduce methodology usually works with components of Map and Reduce with key-value pairs on a two-level hierarchy model. We now try to use a different and more advance framework that may outperform the Map-Reduce method that has been adopted in this case. We consider the Azure Batch framework. Azure Batch creates and manages a pool of computer nodes, which installs the applications we need and schedules the jobs on the nodes. The batch can be used to build Software as a Service (SaaS) applications where there may be a need for huge computations. This framework, when combined with all the above-discussed idea of workflow, would give us a faster and energy-efficient solution that maintains a good throughput and maybe exceed the previous value as well. The newly identified model can possibly be able to run tightly coupled distributed work- loads where there needs to be an inter-application or inter-process communication to reduce the involvement of the master process in execution and hence reduce the overall wait time for the instructions to execute. AS we are running the applications on server load tolerance with a cut off of 70%, rendering software's like Autodesk Maya and Arnold can be run without any glitches and issues.

As we are using the Azure Batch as an integral part of our work is also possible to exploit the benefits they have to offer as well. One major benefit can be identified as the Monte Carlo risk simulation in the execution of embarrassing parallel computations, which can provide greater help in multiple applications. These techniques can be highly im- plemented in the cloud-based financial data services, image rendering risk analysis applications and alternative solution suggestion systems. They have a greater role in the parallel executing and risk mitigation focused applications in the cloud technologies.

8.CONCLUSION

The proposed solution is a combination of workflows in the cloud to improve efficiency. As cloud computing is a huge branch involving multiple concepts of parallel computations and resources utilization, at every branch of cloud service there are respective approaches to maximize the throughput of the service. Alongside, there are also multiple drawbacks and solvable challenges at every level of framework structure. So, the proposed solution can guarantee the efficiency of the solution and increase the performance and scalability of the deployed services in the application. Better scheduling methodology alongside the consideration of all the influential factors in the cloud services invocation for an application will serve the key attributes and complete the whole purpose of cloud implementation.

Cloud technologies have multiple frameworks and inter- nal structures are open to exploring them in deeper. The energy consumption and awareness are permitted just to the scheduling algorithm and workflow computation process. This can further be stretched to areas like in data commu- nications between the nodes, energy utilization calculations with different load balancer applications in the cloud system. Interconnectivity of these measurable features in the cloud with working in different workflow conditions can help in the generation of research work to grow.

REFERENCES

[1] Gunarathne, Thilina, et al. "Scalable Parallel Computing on Clouds Using Twister4Azure Iterative MapReduce." Future Genera- tion Computer Systems, vol. 29, no. 4, June 2013, pp. 1035–1048, 10.1016/j.future.2012.05.027. Accessed 27 Nov. 2019.

[2] Kannadasan, R., et al. "High Performance Parallel Computing with Cloud Technologies." Procedia Computer Science, vol. 132, 2018, pp. 518–524, 10.1016/j.procs.2018.05.004. Accessed 3 June 2020.

[3] Tripathy, Lipsa, and Rasmi Ranjan Patra. "Scheduling in cloud com- puting." International Journal on Cloud Computing: Services and Archi- tecture (IJCCSA) 4.5 (2014): 21-7.

[4] Dinh, Hoang T., et al. "A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches." Wireless Communications and Mobile Computing, vol. 13, no. 18, 12 Oct. 2011, pp. 1587–1611, 10.1002/wcm.1203.

[5] Azam, Md Gulnawaz. (2019). Application of cloud computing in library management: innovation, opportunities and challenges. 10.5281/zen- odo.2536637.

[6] R, Brundesh. "Introduction To Parquet File Format with a Parquet Format Example." AcadGild, 21 June 2016, acadgild.com/blog/parquet- file-format-hadoop.

[7] Juarez, Fredy, et al. "Dynamic Energy-Aware Scheduling for Par- allel Task-Based Application in Cloud Computing." Future Gen- eration Computer Systems, vol. 78, Jan. 2018, pp. 257–271, 10.1016/j.future.2016.06.029.