International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 03 Issue 01 | 2015

MapReduce performance in heterogenous Environment:A Review.

Ms. Shital D.Cheke, Prof. Lokesh Bijole
Department of Computer Engg., V B Kolte Collge of Engineering, Malkapur (M.S.), India

ABSTRACT

Mapreduce is distributed processing model for large scale data intensive application data mining and web indexing.Mapreduce is programming model for processing generation of large data sets with parallel distributed algorithm on a cluster .It is composed of map() procedure and reduce() procedure .Hadoop is open source implementation of mapreduce invented by yahoo,widely used for short job requiring low response time. The current Hadoop implementation assumes that computing nodes in the cluster are homogeneous in nature.mapreduce and hadoop both can efficiently work in homogenous environment. In this article, we survey some of approaches that have been design to improve the mapreduce performance in heterogenous environment.Hadoop is seriously limited by its MapReduce scheduler as it does give well performance in heterogenous environment.LATE is important extension of Hadoop,bt still it falls short to solve crucial problem.It gives poor performance as it uses static manner to comput progress of task,to overcome this shortfall SAMR is invented[8].SAMR calculates progress of task dynamically and having capability to adapt continuously varying environment automatically.

***IndexTerms*-Mapreduce,heterogenous environment,hadoop,SAMR, LATE.**

I. INTRODUCTION

Mapreduce is a programming model for processing and generating large data sets with parallel,distributed algorithm on cluster.A MapReduce program is composed of map() procedure that perform filtering and sorting.Reduce() procedure that performs a summary operation.The MapReduce framework is for processing parallelizable problems across huge datasets using large number of computers.collectively known as “Cluster”, if all nodes are in local area network and known as “grid”, if geographically distributed. The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.The important contribution of MapReduce framework is not only actual map and reduce,but data locality and fault tolerance achieved for variety of application by optimizing execution engine. In past decades ,the WWW has been adopted an ideal platform for developing data intensive application like data mining and web indexing which need to access large amount of data ranges from gigabytes to several petabytes.Example is Google,it uses mapreducing framework as it processes approximate 20 petabytes of data per day.As mentioned earlier apreduce is an attractive programming model for data processing in high performance cluster and provides fault tolerance.Hadoop is an open source implementation of mapreducing model developed by yahoo.It processes on large means hundreads of terabytes of data on atleast 10000 cores[13],[12].It is also used in Amazon and facebook[14],[12].

The mapreduce model runs on homogenous cluster.Homogenous cluster is network of nodes with homogenous architecture that is with identical configuration.As dealing with cluster , mapreduce model also contain 1 master and atleast 1 slave.Mapreduce takes care of partitioning large amount of input data,scheduling of program execution on several machines,data locality,fault tolerance,inter machine communication.

But it is difficult to keep cluster in homogenous environment as nodes breaks then it is necessary to replaced nodes or upgrade cluster.

The homogeneity also has some drawbaks which is need to consider and bring heterogenous enviornmrnt into practice :

- It is not always possible and desirable to have machines with identical architecture.
- It is not possiblr to distribute different workload when making scheduling decision.
- It is not possible in virtulised data centers.

Thus ,it becomes necessary to think about mapreduce model in heterogenous environment.The heterogenous cluster means machines with different configuration.The mapreduce model in homogenous environment gives better performance as compare to heterogenous environment.So the big challenge is to improve

performance of mapreduce model in heterogenous environment. Many researches [3], [10], [8], [9] and [11] has discussed how the heterogeneity affects the Mapreduce performance and de- veloped algorithms to improve the performance of Mapreduce in the heterogeneous environments. In this paper ,there is complete survey of all algorithms that designed to improve performance in heterogenous environment.

II. BACKGROUND A.MAPREDUCE OVERVIEW

The Mapreduce model is a programming model for processing large amount of data even a petabytes.It provides the framework for processing parallelizable problems across huge datasets using large number of computer that is nodes referred to as cluster if all nodes in same local area network otherwise referred as Grid.

J. Dean and S. Ghemawat[1] told that the Mapreduce model was developed by google.It uses this mapreduce framework to process 20 petabytes of data per day.Google uses mapreduce to run data intensive application on distributed infrastructure like commodity cluster.Many functional languages for example Lisp which uses map and reduce procedures and this thing inspired mapreduce[1].Mapreduce is a model which enables any person without having distributed programming knowledge to create his own mapreducing function running in parallel across multiple nodes by specifying two fundamental functions which is mentioned before in paper.Mapreduce program composed of two fundamental functions Map() procedure and Reduce() procedure.Map() procedure perform filtering and sorting.Reduce() procedure performs summary operation. A map function that processes key/value pairs to generate a set of intermediate key/value pairs, and a reduce function merges all the intermediate values associated with the same intermediate key. Mapreduce executes the map and reduce in parallel across the nodes of the cluster.

B. HADOOP

Hadoop[6] is an open source framework that implements the MapReduce parallel programming model[7].Hadoop implementation of MapReduce closely related with Google[2].Hadoop is open source implementation of mapreduce developed by yahoo.The Hadoop,open source project which processes vast amount of data in parallel on large clusters in reliable and fault tolerance[4],[5].There is single master which dealing with number of slaves.The input file which is resided on distributed system throughout nodes in cluster is divided into even size chunks which replicated for fault tolerance.Hadoop divides each MapReduce job into set of task.Each chunk is then processed by the map task and produce key value pair as output by using user defined function Map().Then that output is split into buckets according to the key.Once all Map are finished,Reduce task apply reduce function to the list of Map output[12].Fig 1 shows MapReduce computation.

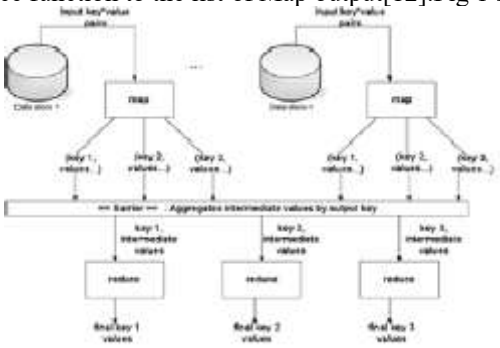

Fig1:A MapReduce computation[3] Hadoop assumes homogenous cluster which means that all nodes in cluster have same processing power and capability.Hence all nodes finish computation at same time.If machine is slower than other faster machinre then that slower machine is considered as faulty machine.In order to run data intensive application ,it is required to build large cluster.And it is not always possible to have large cluster with same machine. assumption is broken when deploying Hadoop in a heteroge- neous cluster:

- Hadoop assumes homogenous workloads.
- It doesn't take the different of workload charactereristics between jobs into account when making scheduling decision.
- Hadoop is not satisfied in virtualised data centers.

III MAPREDUCE APPROACHES IN HETEROGENOUS ENVIORNMENT

A. FAULT TOLERANCE ALGORITHM

A key benefit of MapReduce is that it automatically handles the fault and hides complexity from programmer.

The open sorce project Hadoop is most well MapReduce framework,which processes vast amount of data even petabytes of data in parallel on large clusters that is thousands of nodes in reliable and fault tolerant manner[4],[5].Its MapReduce scheduler just considers scheduling in homogenous environment.The task scheduler assumes that all nodes are in homogenous environment and task makes progress linearly.But this assumption got fail in case of Heterogenous environment.

In this section ,survey is on fault tolerance algorithm to improve fault tolerance support in heterogenous environment.

1. LATE:LONGEST APPROXIMATE TIME TO END ALGORITHM

M. Zaharia et al.[3] designed a new scheduling algorithm which is highly robust to heterogeneity. LATE is based on three principles[12]:

- Priortizing task to speculate.
- Selecting fast nodes to run on.
- Capping speculative tasks to prevent thrashing.

When node has an empty task slot ,Hadoop chooses a task for it from on three categories:

- Any failed tasks are given the highest priority.
- Non running task are considered
- The tasks which need to execute speculatively.

Hadoop monitors progress depend upon progress score between 0 and 1 to select speculative task.As there are two types of task Map task and Reduce task.For Map task,the progress score is the fraction of input read data.For reduce task ,the execution is divided into three phases:copy phase,sort phase,reduce phase.Each of which represents $1/3^{rd}$ of the progress score[12].LATE scheduling algorithm always launch backup task for those task which are having more remaining time than other task.LATE always speculatively executes the task which will finish farthest in future.LATE estimates finish time based on the progress score given by Hadoop.Hadoop estimates the progress rate of each task as progress sscore /T,where T is the amount of time the task has been running for,and it estimates the task's finish time as (1-ProgressScore)/ProgressRate[8].

LATE tries to improve Hadoop by attempting to find real slow task by calculating remaining time of all other task.Then LATE scheduler will select a task with long remaining time and considers this set if tasks as slow tasks. However it does not compute the remaining time for tasks correctly, and can not find real slow tasks in the end.Also it often launches backup task for inappropriate task because LATE cant find total time remaining for running task,that is LATE uses static way to find performance.

2 .SAMR:SELF ADAPTIVE MAPREDUCE SCHEDULING ALGORITHM

Q.Chen et al[8]proposed SAMR scheduling algorithm,which calculates the progress dynamically.It adopted some features from LATE scheduler and proposed modified version to improve MapReduce in terms of saving time of execution and the system's resources.SAMR incorporates historical information recorded on each node to tune parameters and final slow tasks dyanamically.In order to save system resources, SAMR divides slow nodes into Map slow nodes and Reduce slow nodes.SAMR defines fast nodes can finishes a task in shorter time and slow nodes takes longer time.Map/Reduce slow nodes executes Map/Reduce task using longer time.

SAMR mainly based on three given aspects:

- SAMR takes historical information recorded on every node to tune weight of each stage dyanamically.
- SAMR takes the two stages characteristics of map tasks into consideration for first time.
- SAMR classifies slow nodes into Map slow node and Reduce slow nodes.

When job is committed ,SAMR splits the job into lots of fine grained map and reduce tasks ,then assign them to series of nodes in heterogenous cluster.SAMR reads all historical information stored on each node and updates information after every execution. SAMR can get more accurate PSs of all task by using historical information[8].It finds real slow task and decreases execution time and finds which task need backup task.It identifies slow nodes and ensure that backup task will not be launch on that.It gets final result of fine grained tasks when slow task or back up task finishes first.SAMR signifiacantly decreases the time of execution up to 25% compared with Hadoop scheduler and 14% compared with LATE scheduler[8].

IV. CONCLUSION

In this article,an detail review of key feature that improves MapReduce performance in Heterogenous environment is presented.SAMR algorithm improves fault tolerance and decreases execution time better as compared LATE.LATE tries to improve Hadoop by attempting to find real slow task by calculating remaining time of all the task.But LATE does not find the remaining time accurately along with it uses static way to compute progress of task.Hence SAMR is invented as it calculates progress of task dynamically.It uses historical information and classifies slow nodes into map slow nodes and reduce slow nodes. The algorithm decreases the execution time ofMapReduce jobs compared to Hadoop scheduler and LATE schedular , especially in heterogeneous environments.The algorithm selects slow tasks and launch backup replicas accordingly while classifying nodes correctly, and saving a lot of system resources.

V. REFERENCE

[1] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. InProceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation -Volume 6, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[2] Hadoop, http://lucene.apache.org/hadoop, (last view June 30,2012).

[3] M.Zaharia, A.Konwinski, A.Joseph, Y.zatz, and I.Stoica. Improving mapre-duce performance in heterogeneous environments. In OSDI'08: 8th USENIX Symposium on Operating Systems Design And Implementation, October 2008.

[4] Yahoo,"Yahoo!hadoop tutorial" http://public.yahoo.com/gogate/hadoop-tutorial/start-tutorial.html.

[5] Hadoop,"Hadoop home page ",http://hadoop.apache.org/.

[6] Hadoop,http://hadoop>apache,org.2009.

[7] J. Dean and S.Ghemawat Mapreduce:simplified data processing on large clusters.In OSDI2004 proceedings of yhr 6th symposium on operating system design and implementation,pages10-10,2004

[8] Q.Chen, D.Zhang, M.Guo, Q.Deng and S.Guo. SAMR: A Self adaptive MapReduce Scheduling Algorithm In Heterogeneous Environment. In Pro-ceedings of IEEE 10th International Conference on Computer and Infor-mation Technology, 2010.

[9] X.Zhang, Y.Feng, S.Feng, J.Fan and Z.Ming. An Effective Data Locality Aware Task Scheduling Method for MapReduce Framework in Heterogeneous En-vironments. In International Conference on Cloud and Service Computing, 2011.

[10] J.Xie, S.Yin, X.Ruan, Z.Ding, Y.Tian, J.Majors,A.Manzanares, and X.Qin. Improving MapReduce Performance through Data Placement in Heteroge-neous Hadoop Clusters. In proceedings of IEEE International parallel and dis-tributed Processing Symposium, 2010.

[11] Z.Guo, G.Fox. Improving MapReduce Performance in Heterogeneous Net-work Environments and Resource Utilization. In proceeding of IEEE/ACM 12th International Symposium on Cluster, Cloud and Grid Computing, 2012.

[12] Salma Khalil, Sameh A.Salem, Salwa Nassar and Elsayed M.Saad,Mapreduce performance in heterogenous Environment-review, International Journal of Scientific & Engineering Research, Volume 4, Issue 4, April -2013.

[13] Yahoo! launches worlds largest hadoop production application http:/developer.yahoo.com/blogs/hadoop/posts//2008 // 02/yahoo-worlds-largest-production-hadoop/, (last view June 30, 2012).