CAP 706: PAPER WRITING-I

A Priority and Shortest-job-First Scheduling for Resource Allocation In Cloud Computing

Submitted by:

Nitesh Parbhakar

Registration Number: 11303695, MCA(Hons)

Manjit Singh

Registration Number: 11306042, MCA (Hons.)

Supervised by:

Mr. Raj Singh (Asst. Professor)

School of Computer Application

Lovely Professional University

DECLARATION

I hereby declare that the Paper writing entitled," A Priority and Shortest-job-First Scheduling for Resource Allocation in Cloud computing" submitted for the Master of computer application Degree is entirely our original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

CERTIFICATE

This is to certify that Manjit Singh, Nitesh Prabhakar MCA(Hons.) paper writing titled, "A Priority and Shortest-job-First Scheduling for Resource Allocation in Cloud computing" below my direction and command. To the best of my knowledge, the present work is the result of their original investigation and study. No part of the paper writing has ever been submitted for any other degree or diploma. The paper writing is fit for the submission and the partial fulfilment of the conditions for the award of Master's Degree.

We further declared that we or any other person has not previously submitted this report to any other institution/university for any other degree/ diploma or any other person.

Date: 29 April 2015 Signature of Supervised

Mr. Raj Singh (Assistant professor)

ACKNOWLEDGEMENT

We would resembling to precise our obligations of thankfulness to our Guide Mr. Raj Singh who provided us the golden prospect to do this magnificent assignment on the research topic Wireless Sensor Network & Cloud Computing, it also aided us in undertaking a lot of Exploration & we arose to identify approximately so numerous belongings we are truly pleased to them. Secondly we would also like to thank our Parents and Friends who helped us a lot in finalizing this project within the limited time frame.

INDEX:-

- Introduction to Priority and Shortest-job-First Scheduling for Resource Allocation in Cloud
- Introduction to Cloud Computing
- Architecture of cloud computing
- Types of resources
- Types of Cloud Computing
- Problem Definition
- Objectives
- Techniques Used for resource allocation
- Literature Review
- Algorithm used
- Tools
- Conclusion
- References

A Priority and Shortest-job-First Scheduling for Resource Allocation in Cloud

Abstract

Background: Given the dynamics in resource allocation schemes offered by cloud computing, effective scheduling algorithms are important to utilize these benefits.

Aim: In this paper, we propose a scheduling algorithm integrated with task grouping, priority-aware and SJF (shortest-job-first) to reduce the waiting time and make span, as well as to maximize resource utilization. *Method*: Scheduling is responsible for allocating the tasks to the best suitable resources with consideration of some dynamic parameters, restrictions and demands, such as network restriction and resource processing capability as well as waiting time. The proposed scheduling algorithm is integrated with task grouping, prioritization of bandwidth awareness and SJF algorithm, which aims at reducing processing time, waiting time and overhead.

Keywords: Cloud Computing, Task Grouping, Scheduling, SJF.

Introduction of Cloud Computing

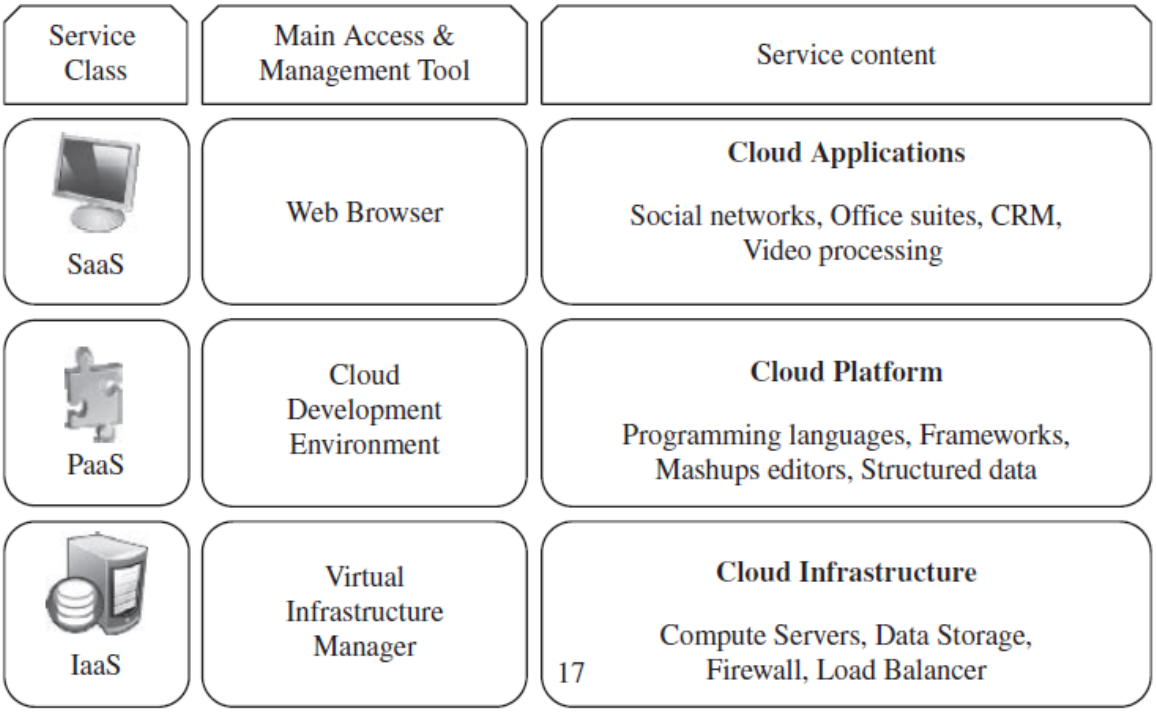
In classical desktop computing, software programs are running in computer that allows dynamic resource allocation on joined resources using a grouping of techniques from parallel, distributed, & platform virtualization technologies. The documents which create are stored on personal computer. It may generate a problem of lack of memory or we can say storage problem . All Document are achieved with other computers on the same available network but it cannot be access by computers from external network.

But now in modern era i.e. Cloud Computing any software programs not running in personal computer although these are stored on servers & any one can accessed with the help of Internet. it plays like an essential technology in the field of information technology. This behave likes computing as a service, software, distributed resources & information are given to the Personal computers and many System devices. User can precisely access from any web browser. On the servers, Business software and the data are stored at a remote location.

Examples like Yahoo . com, Gmail . com, Hotmail . com

There is no need to run an e-mail software in the computers. Any one can loged in remotely to a email account .Account's software and storage doesn't be on the computers. It is stored on the cloud computing.

Architecture of resources in cloud computing:



Types of resources

1).Infrastructure as a Service:

It includes resources like computation, storage space, & announcement on command are called as Infrastructure as a Service (IaaS).

Users have advantages to execute various performance of the server, for example: initial and stopping it, simple customize this by install software post & configured access authentication and firewalls rules.

2).Platform as a Service:

A cloud platform provides facility to the developers to create and deploy applications. There is no have to identify how much recollection and you can say how several process that applications/software will use. Building blocks to new applications are provided in multiple programming models and specialized services

3).Software as a Service:

In this end users can access by this layer with web portals. Users or client shiftted locally install in computer software to online available services that provide same type of work. For Example:- social networking websites, Office Suite etc.

Types of cloud computing:-

1. Private Cloud: In private cloud Framework is arranged and used & all available resource maintained and controlled by the Firm. Internally or externally hosting is also done in it and can be managed by third party.

2. Community Cloud: In Community Cloud, Infrastructure supports a specific community and with common approaches. Cloud are operated by several organizations in sharing.

3. Public Cloud: In Public Cloud Infrastructure Client can access services without any control and at specific rent. This cloud service is used by different types of firms.

4. Hybrid Cloud: This Hybrid Cloud can be combined of Private, Community and Public Cloud Infrastructure. But the characteristics are unique in this Cloud that provide advantages to the users. In this, data and applications sent through one device to another Cloud device.

Problem Defination:

The cloud contain enormous resources which are required for different users. The allocation of resources in cloud is the big issue. Our main concern is to find the various issues and resource allocation techniques which can reduce the overhead of resources on cloud so that each resource can be made available to users and to reduce the overall delay.

Objectives:

1.To conduct a study of resource allocation & monitoring on the cloud computing.

2.To describe cloud computing and its properties, research issues in resource management mainly in resource allocation and monitoring

3.To study current solutions approach for resource allocation and monitoring)

4. to find the problem in the existing frame work and improving it

Techniques used in Resource Allocation:-

1. Round robin:-

In this algorithm concept of time quantum or slices is used. This algorithm divided time into multiple quantum and particular time quantum or time interval is given to each node and the node will perform its operations. On the basis of time quantum or slices resources are provided to the clients.

Advantages:-

-Maximize all the available resource in a reasonable manner.

- In the same manner or sequence the VMs are installed to provide all the nodes which provides equality.

Disadvantages:

-Power Consumption is far above as much as nodes will be reserved turn on for taking too much time .

2. Preemptive Priority

In cloud computing , there are various scenario where different features needed to be considered as some job can wait until other important jobs completed. Pre-emption process shows that it can remove previous scheduled activities which according to certain constraints like priority. Priority increases the weight age of the jobs which helps in decreasing the complexity.

3 Shortest Response Time First

The basic principal is that jobs having less priority is going to be executed first, if there are two jobs with same priority then first cum first serve rule will be applied.

LITERATURE SURVEY

Ronak et al [1] proposed a study on dynamic resource allocation to cloud computing that is based on TARA stand for Topology Aware Resource Allocation. Empirical study shows many Resource allocation system that focus on memory resource but those lack in other factors, Future work can be extended with smarter and secured optimal resource allocation algorithm.

Vignesh et al[2] composed a study on analysis of resource scheduling algorithms and three algorithms Preemptive Priority, Round Robin, and Shortest Remain Time First have been used it has been calculated that SRTF gives the lowest time parameters and maximum effective algorithm for resource scheduling. In virtual, this work figure out disk space organised is critical matter. Current scheduling algorithm provides high throughput and cost effect but there is no reliability and availability. Future work can be to use a algorithm which increase availability and reliability in cloud computing environment and also comparative study with existing algorithm.

CHEN Yibo et al [3] To build up CDSMS, a framework for customer-driven dynamic resource allocation in cloud computing , this decrease the total relocation times of adjusting the value of argument of reply time dynamically on clients profiles moment, this may decide the most excellent resource terms algorithm routinely in different scenarios to get better resource utilization. new and analytic results show that CDSMS effectively captures the dynamic nature of client and reduces 11.6% resource wants as compared with QuID. Future work is to improve the precision of client accepted response time for improve the experience of client

Abbas Horri et at [4] planned the new QoS-aware VMs consolidation advance for cloud environments which take the process based on resource utilization history of virtual machines. planned algorithms for implemented and evaluated using CloudSim simulator.

planned algorithms for implemented and evaluated using CloudSim simulator is prove on QoS metrics and energy expenditure or show. Future research direction will be the investigation for more work- load models that provide more best result.

Chunlin Li et. al [5] studies resource allocation to optimize objects of cloud users, IaaS, SaaS supply in Cloud computing. This paper proposes the structure of different layers in the cloud computing that are IaaS and SaaS for efficient resource allocation through an iterative algorithm. The experiments are conducted to compare the performance of proposed joint optimization algorithm for efficient resource allocation with other related works.

Phuoc et al [6] proposed present a novel structural design that takes benefits of teamwork of thin-thick clients, particularly aiming at utilizing and data distribution Cloud Computing resources for meet the accepted Quality-of-Service requirements. Simulation result shows that the proposed schemes can improve resource allocation efficiency and achieve better performance than the existing ones. Through carefully implemented simulations, the planned method helps in to enlarge the effective of resource allocation and utilized with best strategy. Current work discussed an algorithm for selecting best resource allocation scheme in direction to assure many Service Level Agreement.

Daji Ergu et al [7] planned a resource allocation task oriented model on cloud computing . Resource allocation task was levelled in pairwise judgment matrix method & AHP served the accessible resources and user preferences. Weights of tasks are computed using the Analytic Hierarchy Process method & the equivalent calculated resources are given on terms of weights of task. The result point out that it is significant to additional find the not consistent elements and get better the consistency ratio when the weights of tasks are allocate dynamically resources in cloud computing. Future Allocating the resources dynamically & correctly for the tasks.

[8]. Cloud computing focus on research community and the industry over recent years because of its elasticity in software deployments. Software engineering for cloud platform systems is a new dominion in research requiring careful considerations on its characteristic with respect to outdated software development paradigms. The research focuses on the part of effective scheduling run time tasks. The aim of cloud computing is to recognise cooperation work and resource sharing. This makes resource management very complex when different types of resources reproduce isomerism, dynamic nature and a variety of user demands. So scheduling problem is an important research area. The utilization rate of enormous resources in datacenter is connected to scheduling. The essential mechanism of this new type software system is to scheduling the applications to the resources pool.

[9]. Scheduling is a decision process, and its content is deploying resources to applications of different clients at a suitable time, or during a specific period of time. The target of optimizing scheduling considers one or two factors that include cost, task completion time, task priority, profit and so on. In the premise of guaranteed resource utilization rate, scheduling policies mainly focus on allocation management of resources and satisfy the resource demands of users. Eventually, scheduling policies should effectively improve the number of completed applications, increase profit of service party, reduce cost which is

undertaken by service party when accepting applications and guarantee QoS (Quality of Service) demand of clients. In cloud computing system, there exists some applications with a great deal of light-weight tasks. Dispatching these fine-grained tasks to a pool of resources that provide high processing capability is not economical and consumes extra waiting time and turnaround time by comparing a coarse grained task allocation to the resource [10]. Since that the overall turnaround time includes each task scheduling time, execution time and transmission time. A large amount of fine-grained tasks will spend a lot of time on scheduling and transmission. It is rather impractical to consume the resource processing capability, and lowers resource utilization rate when a fine-grained task is allocated and executed to a resource with high processing capability. The total turnaround time of fine-grained tasks can be further reduced by grouping these fine-grained tasks as coarse-grained tasks in the entire scheduling process.

BACKGROUND: Cloud computing consists of a group of computing resources that are delivered over a network, which is accomplished by using virtualization technologies to combine and allocate resources suitable for various different software applications. It provides a platform for solutions requiring different configurations, matching physical hardware combinations in a virtualized cloud environment managed by cloud platform software to deliver improved services. The strategies used in the cloud platform software become important, which directly influence the runtime performance of software applications running on its platform. Therefore the effective scheduling policies to maximize the utilization of the virtualized resources are the primary focus of this study. This section provides a summary of related scheduling approaches applied in cloud computing.

A. Scheduling models in Cloud Computing

In traditional distributed environment, the aim of optimizing scheduling is mainly focusing on system performance, such as system throughput, CPU utilization rate and almost never considering QoS. In cloud computing environment, we are not only emphasizing resource utilization rate and system performance, but also requiring a guaranteed QoS of users based on different demands. Users can choose the resource in the cloud by themselves according to their own requirements.

1) Cloud computing scheduling model: Cloud computing scheduling model is mainly constructed by Client, Broker, Resources, Resources supporter and Information Service. Fig.1 shows the scheduling model structure [11]. The tasks that users need to implement usually can be divided into serial application, parallel application, parameter scan application, cooperation application and so on. System allows users to set up resource demand and parameter preference. Different clients use resources at different prices, which may vary from time to time. Broker is a middle interface between clients and resources as well as used to find resources, choose resources, accept tasks, return scheduling results, and exchange information between clients and resources. Broker supports different scheduling policies, which can allocate resources and schedule tasks in accordance to the demands of clients. Broker is constituted by Job Control Agent, Schedule Advisor, Explorer, Trade Manager and Deployment Agent.

• **Job Control Agent**: It is responsible for monitoring jobs in the software system, such as schedule generation, jobs creation, status of jobs and communicating with clients and schedule advisor.
• **Schedule Advisor**: It is used to determine resources, allocate available resources which satisfy the demands of clients such as deadline and cost, as well as to allocate jobs.

• **Cloud Explorer**: It is a tool that communicates with cloud information service to find resources and identifies the list of authorized machines as well as records resources status information.
• **Trade Manager**: It determines resources access cost and tries to communicate with resources at a low cost under the guidance of schedule advisor.
• **Deployment Agent**: It uses scheduler instruction to activate the execution of tasks as well as to update the status of execution sending back to Job Control Agent in regular intervals.
During the transaction between clients and service providers, service providers register resource information at first. After clients submit tasks to broker, broker searches resources in information service and deploys tasks to appropriate resources in accordance to the corresponding scheduling algorithms. Before execution of tasks, broker evaluates completion time and cost of tasks. If the time exceeds deadline or the cost is higher than budget of clients, the broker will deny tasks. If the execution of tasks is accomplished, broker will return the deployment results to clients and gain relevant profits, otherwise, send error message back to clients.*2) Basic scheduling methods:* Scheduling methods always consider two aspects: one is characteristics of tasks, and the other is characteristics of datacenter resources [12] [13]. Tasks submit on the resources which are free and where the input data is available or on the other hand, tasks submit on some specific resources based on some criteria [14]. *3) Resource allocation:* The resources in the cloud computing can be allocated in many different ways. Traditional and simple method of task scheduling in cloud environment uses the client tasks as the overhead application base [1]. The allocation of resources that need to consider maximum utilization rate of resources are FCFS (First-Come-First-Service), SJF (Shortest-Job-First) scheduling, priority scheduling, RR (Round-Robin) scheduling, random, greedy, Genetic Algorithm [15]. The scheduling of tasks can also be FCFS, SJF, priority-based, RR, job grouping and so on. Scheduling algorithms choose a task to be performed and corresponding

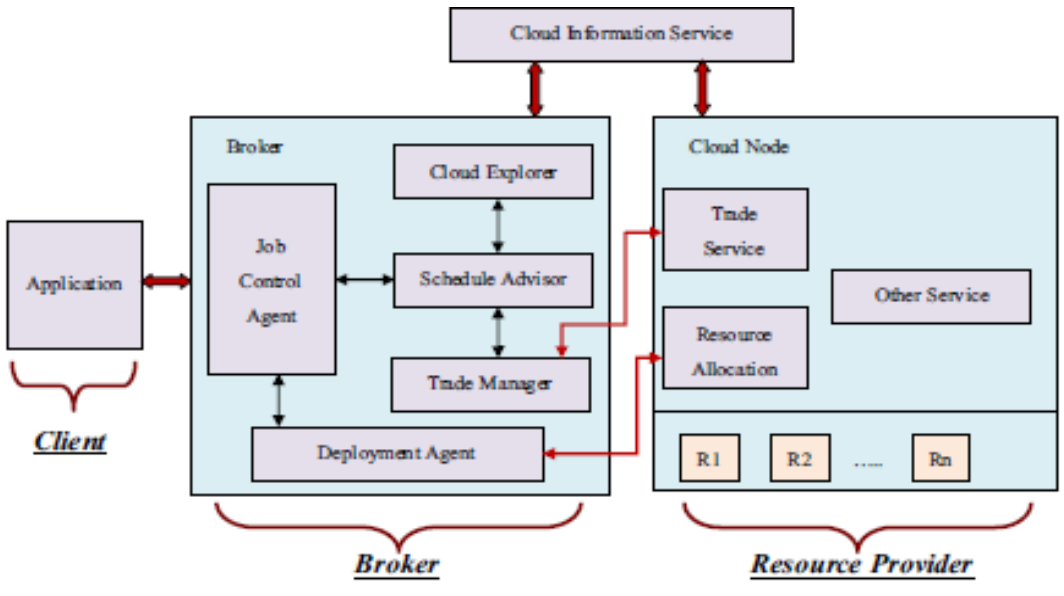
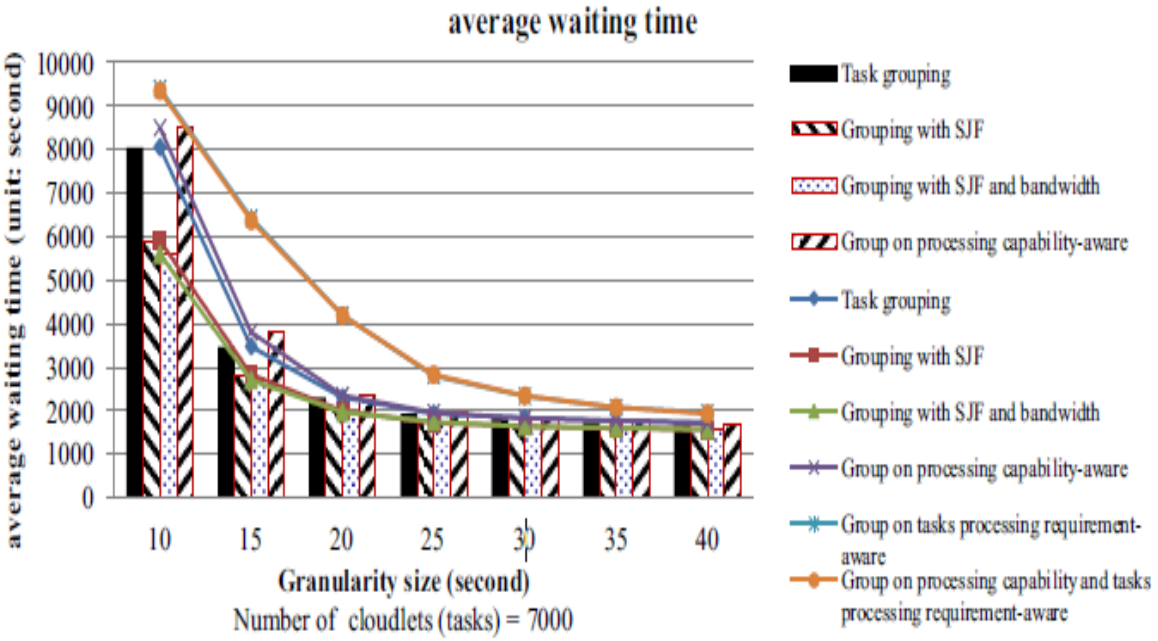


Fig. 1. Chart of scheduling model structure

resource in which the task will be executed, according to different characters of resource such as bandwidth, processing capabilities, cost, load balancing, and so on, as well as base on clients requirements of deployment. In this paper, we both focus on resource allocation and task scheduling, as well as take into account some specific criteria or priorities of tasks and

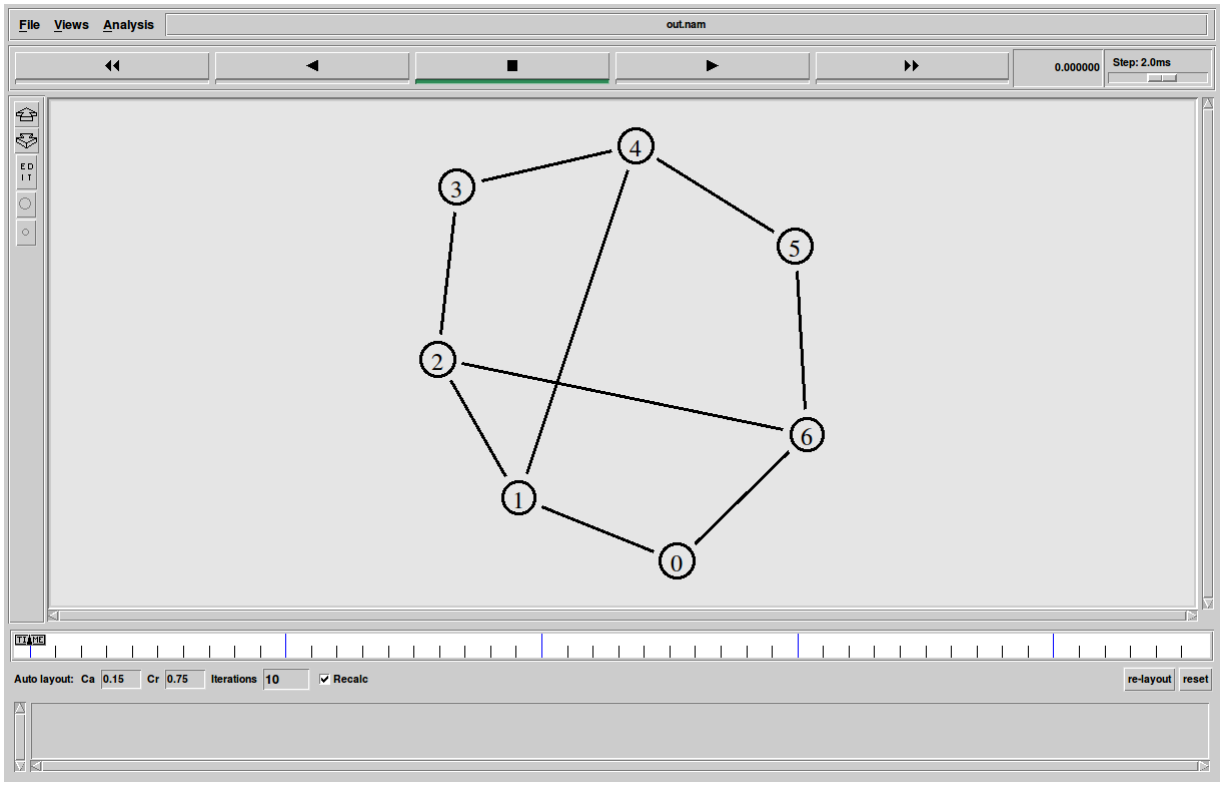
resources, such as resource bandwidth and processing capability, task granularity (fine-grained and coarse-grained) and deadline.

Algorithm : the improved task scheduling algorithm
1: groupid = 0
2: for i: = 0 to cloudletList size-1 do
3: m = i
4: for j: = 0 to ResList size-1 do
5: total jobMI = 0
6: max cloudletFileSize = 0
7: max cloudletOutputSize = 0
8: pre cloudletFileSize = 0
9: pre cloudletOutputSize = 0
10: total resMI*j* = ResList*j* MIPS ∗granularity size
11: while total jobMI≤total resMI*j* and i≤cloudletList size-1 do
12: total jobMI = total jobMI+cloudletList*i* MI
13: if max cloudletFileSize<cloudletList*i* FileSize then
14: max cloudletFileSize = cloudletList*i* FileSize
15: end if
16: if max cloudletOutputSize<cloudletList*i* OutputFileSize
then
17: max cloudletOutputSize = cloudletList*i* OutputFileSize
18: end if
19: i++
20: end while
21: i−−
22: if total jobMI>total resMI*j* then
23: total jobMI = total jobMI-cloudletList*i* MI
24: max cloudletFileSize = pre cloudletFileSize
25: max cloudletOutputFileSize = pre cloudletOutputFileSize
26: i−−
27: end if
28: if (m-1) == i then
29: i++
30: total jobMI = cloudletList*i* MI
31: end if
32: Create a new task whose MI equals total jobMI
33: Set a unique ID (groupid) to the created task
34: Insert this task into groupedcloudletList
35: Put the task on the groupedcloudletList*groupid*
36: Insert the corresponding resource ResList*j* into groupedvmList
37: Put the corresponding resource ResList*j* on groupedvmList*groupid*
38: end for
39: end for



Tool used:-

NS2(stands for network simulator). It is Designed by UC Berkeley and maintained by USC. NS2 is designed in C++ and object oriented version of TCL that called OTCL.



CONCLUSION

To improve the scheduling capability on cloud computing based software systems, simulations are used to ease the evaluations on different methods under various runtime situations in a cloud environment. The study suggests a task grouping scheduling algorithm combined with shortest job first and bandwidth awareness algorithms in an attempt to reduce the waiting time and its related processing costs.

References:-

[1] Ronak Patel and Sanjay Patel, "Survey on Resource Allocation Strategies in Cloud Computing", 2013

[2] Jaisankar N, Vignesh V, Sendhil Kumar KS,"Resource Management and Scheduling in Cloud Environment", international Journal of Scientific and Research Publications, 2013

[3] ZHOU Jingcai1, ZHANG Huying1, CHEN Yibo2:An Optimal Resource Provision Policy in Cloud Computing Based on Customer Profiles:Wuhan University and Springer-Verlag Berlin Heidelberg 2014

[4] : Abbas Horri · Mohammad Sadegh Mozafari · Gholamhossein Dastghaibyfard:,"Novel resource allocation algorithms to performance and energy efficiency in cloud computing", Published 2014

[5] Li·Layuan Li, "Efficient resource allocation for optimizing objectives of cloud users, IaaS provider and SaaS provider in cloud environment ",Published online: 7 February 2013

[6] Pham Phuoc Hung • Tuan-Anh Bui • Mauricio Alejandro Go ´mez Morales • Mui Van Nguyen • Eui-Nam Huh: "Optimal collaboration of thin–thick clients and resource allocation in cloud computing".

Received: 21 December 2012/Accepted: 15 March 2013/Published online: 21 May 2013 Springer-Verlag London 2013

[7] Ergu·Gang Kou et al: "The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment", Published online: 19 May 2011

[8] S. Selvarani and G. Sadhasivam, "Improved cost-based algorithm for task scheduling in cloud computing," in *Computational Intelligence and Computing Research (ICCIC), IEEE International Conference on*, 2010.

[9] B. Xu, C. Zhao, E. Hu, and B. Hu, "Job scheduling algorithm based on berger model in cloud environment," *Advances in Engineering Software*, vol. 42, no. 7, pp. 419–425, 2011.

[10] Q. Liu and Y. Liao, "Grouping-based fine-grained job scheduling in grid computing," in *Education Technology and Computer Science. First International Workshop on*, 2009.

[11] R. Buyya, D. Abramson, J. Giddy *et al.*, "An economy driven resource management architecture for global computational power grids," in *Proceedings of the 2000 International*

Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000), 2000, pp. 26–29.

[12] V. Korkhov, J. Moscicki, and V. Krzhizhanovskaya, "Dynamic workload balancing of parallel applications with user-level scheduling on the grid," *Future Generation Computer Systems*, vol. 25, no. 1, pp. 28–34, 2009.

[13] S. Gomathi and D. Manimegalai, "An adaptive grouping based job scheduling in grid computing," in *Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on*. IEEE, 2011, pp. 192–196.

[14] S. Singh and K. Kant, "Greedy grid scheduling algorithm in dynamic job submission environment," in *Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on*. IEEE, 2011, pp. 933–936.

[15] S. Sivanandam and S. Deepa, *Introduction to genetic algorithms*. Springer Publishing Company, Incorporated, 2007.