ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

A Fast Parallel Data Processing Method in Cloud Computing

Miss Megha Rode, Mr .Mohasin Tamboli, Mr. Runwal Sachin

Lecturer, Department of IT, Vishwabharati Academy's College of Engineering, Ahmednagar, India.

Abstract: The Cloud Services such as Infrastructure as a Services (IaaS) provide fast data processing services for Parallel data processing in cloud to store,manage and processing resources. Many Cloud computing companies have in progress to use this framework for efficient parallel data processing in the cloud to make their product easy for customers to access these services and to deploy their programs. Consequently, the allocated compute resources may be increase processing time and cost. In this paper We have to discuss the opportunities and challenges for fast data processing in parallel data processing in clouds .The first data processing framework Nephele which allocate the resources parallel and schedule them and then execution for particular task is carried out in that the large amount of data is divided into number of several independent subtask and data is distributed among nodes and then lastly compute them parallel. Assigned to different virtual machines which automatically instantiated and completed during the job execution we present extended evaluations of Map Reduce processing job on cloud system and compare the effect to the data processing structure Hadoop

Keywords::Cloud Computing, data processing, parallel, resource allocation, task scheduling, many task computing, and nephele

I.INTRODUCTION

Cloud computing is a model for enabling convenient on demand network access to a shared resources that can be rapidly provisioned and released withminimal management effort or service provider interaction.Todaya growing number of companies have to process large amount of data in cost–efficient manner .Cloud computing is having following typesPlatform as a Service (PaaS), Infrastructure as a Service (IaaS), Software as a Service (SaaS)Software as a Service (SaaS)From end user's point of viewApps are located in the cloudSoftware experiences are delivered through the Internet .Platform as a Service (PaaS)From developer's point of view Cloud providers offer an Internet-based platform to developers who want to create services but don't want to build their own cloud. Infrastructure as a Service (IaaS).Cloud providers build datacentersPower, scale, hardware, networking, storage, distributed systems, etcDatacenter as a serviceCloud users rent storage, computation, and maintenance from cloud providers pay-as-you-go like utility Users can rent application software and databases, using Software as a Service. The infrastructure and platforms on which the applications run is managed by cloud providers. These companies are operators of internet to propagate data the companies which deal with large amount of data for that they have to deal with many numbers of commodities server for that they have made traditional database solutions prohibitively costly. Instead, these companies have an architectural standard based on big amount of product servers. Problems like processingcrawled documents or regenerating a web index are split data them into number of subtask, and then distributed among the available nodes, and lastly computed in parallel into original solution. In order to simplify the development of distributed application on top of such architectures many of these companies have also built data processing framework. Examples areterms like high-throughput computing (HTC) or many-task computing (MTC), depending on the amount of data and the number of tasks involved in the computation the processing framework takes care of distributing the program among of data and the number of tasks involved in the computation. The processing framework takes care of distributing the program among the available nodes and executes each instance of the program on the appropriate data fragment .although these systems differ in system design, their programming models share similar objectives namely hiding the bother of parallel programming .In cloud shared resources, software as well as information is provided. Virtual machine charges the companies for only amount of time they are used. The main goal of this paper is to decrease the overloads of the main cloud and to improve the performance as well as the quality of the cloud by dividing job into no of subtask in the cloud by using cloud storage, task manager and job manager that performs the different task using different resources

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

available in the network. The growing no of companies which run with minimal resources and take on the resources as they needed on demand and pay only when there is a use is enabled by IaaS (Infrastructure as a Service). This becomes achievable by the Virtual Machines (VM).we are going to split job manager into number of task manager and they are assigned to different employees in the team and resources are assigned dynamically using virtual machine which schedule the task and after performing operation it given to the main solution for that we first start virtual machine in the cloud and the resources needed for the tasks that going to be executed parallel.

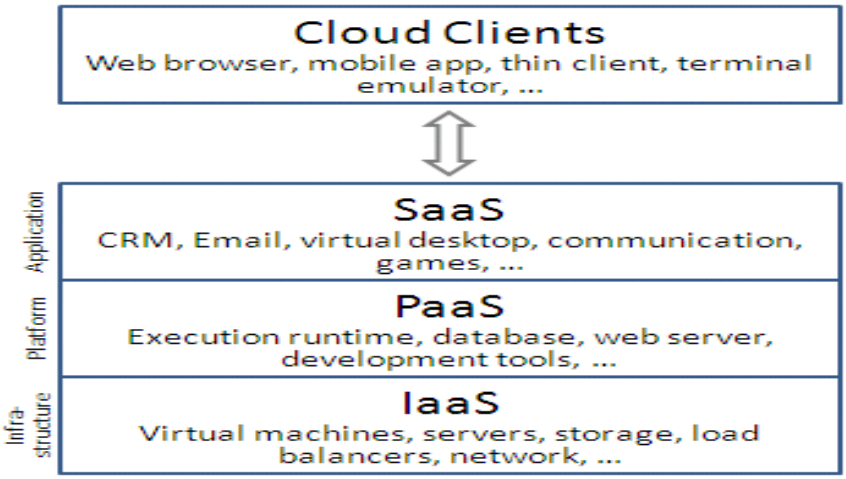

Fig.1Cloud computing structure

II.CHALLENGES AND OPPORTUNITIES

In this section we have to discuss how these new opportunities and challenges used for fast parallel data processing in cloud computing.

Opportunities: Today's processing structure typically assumes the resources they handle which consist of consistent compute nodes. It is designed to deal with individual nodes failures, in that we have to consider number of available machines to be steady particularly when scheduling the job's then we have to processing the job's and execution of job in the cloud can be used to create such group of nodes. New virtual machines can be allocated at any time machines which is no longer used can be terminated and the cloud customer will be charged for them no more. The processing workflow is dynamically allocated by the scheduler for that user must start single virtual machine which runs an incoming job and then advice the cloud to directly start the required virtual machine according to the job's processing phases. The design of the processing framework and the way is described first, the scheduler of the processing framework must become aware of cloud environment a job should be executed. It must know about the different types of available virtual machines and their cost. It can be allocate or destroy them on behalf of cloud customer second; the standard used to describe the jobs must be powerful enough to express dependencies between the different task managers and the job manager. The cloud must be aware of which task's output is required as another task's input. Otherwise the scheduler of the processing framework cannot decide what happen at that point at particular instant when virtual machine is no longer needed and reallocated it. Although at the end of a job only a small number of reducer tasks may still be running and it is not possible to shut down the idle virtual machines since it is unclear if they contain some intermediate result.

Challenges: The nature of cloud virtualization is to create virtual device the framework divides the resources into one or more execution environment as simple as partitioning. .However; it is a new challenge to group of node setups. Public cloud providers have diverse, often unclear, standards for security, compliance and governance

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

III.DESIGNING &IMPLEMENTATION OF NEPHELE

Nephele's architecture follows a classic master-worker pattern as illustrated in Fig 2.it is a Structural overview of Nephele running in an Infrastructure-as-a-Service (IaaS) in the cloud. In that we first submit a Nephele compute job, at that time a user must start a Virtual Machine(VM) in the cloud which runs the so it is called as Job Manager (JM). The Job Manager receives one or more client's jobs from the user which is responsible for scheduling purpose andafter that it coordinates with their execution. It is used to communicate withtheir interface in the cloud storage which provides the control to the instantiation of virtual machine and it uses the interface in the Cloud Controller. The Cloud Controller which perform the task such that a Job Manager can allocate or deallocate VMs according to the current job execution. In this paper we will comply with common Cloud computing terminology and refer to these VMs as instances. The term instance type will be used to differentiate between VMs with different hardware characteristics. E.g., the instance type "m1.small" could denote VMs with the one CPU core and one GB of RAM and a 128 GB disk while the instance type "c1.xlarge" could refer to machines with 8 CPU cores and 18 GB RAMand 512 GB disk storage. The actual execution of tasks which a Nephele job consists is carried out by a set of instances. In that each instance runs a so-it is called as a Task Manager (TM). The Job Manager divide the task into no of task manager and each task executed after that executes them and after that it informs the Job Manager about their completion if any possible errors are present then it inform to the cloud storage. Unless a job is submitted to the Job Manager at the last we expect that the set of instances to be empty. Upon job reception the Job Manager decides depending on the job's particular tasks. The job manager decidehow many andwhat type of instances the job should be executed on and when the respective instances must be allocated and deallocatedby the cloud controller to ensure a continuous but cost-efficient processing. The newly allocated instances give the data with a previously compiled VM image. The image is configured to automaticallywhich start a Task Manager and register it with the Job Manager. Once all the necessary task done then the Task Managers have successfully contacted the Job Manager, it triggers the execution of the scheduled job. As a result, we expect the cloud to offer persistent storage. This persistent storage is supposed to store the job's input data and eventually receive its output data at last. It must be accessible for both the Job Manager as well as for the set of Task Managers.

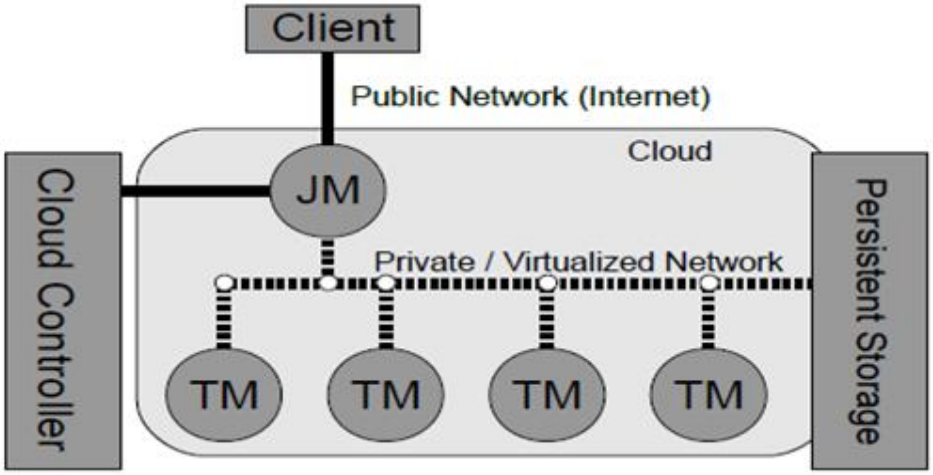

Fig.2.Structural overview of Nephele running in an Iaas

B.**ARCHITECTURE DIAGRAM:** The nodes involved user interface for the system. In Fig 3Architecture Diagram for fast parallel data processing in the cloud the actual execution carried out by set of instance, each instance runs so it is called as task manager. The job manager receives the user jobs which are responsible for scheduling and execution .It communicates with cloud controllerthe cloud controller can allocate and DE allocates virtual machine according to the current job execution. A task manager receive one or more task from job manager at the same time, execution is carried

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

out and after that it informs the job manager about their job completion or any possible error .when job is submitted to the job manager at that time set of instance should be empty .upon the job response the job manager decide which instance should run and when the respective instance must allocate to ensure a continuous but cost –efficient fast parallel data processing.

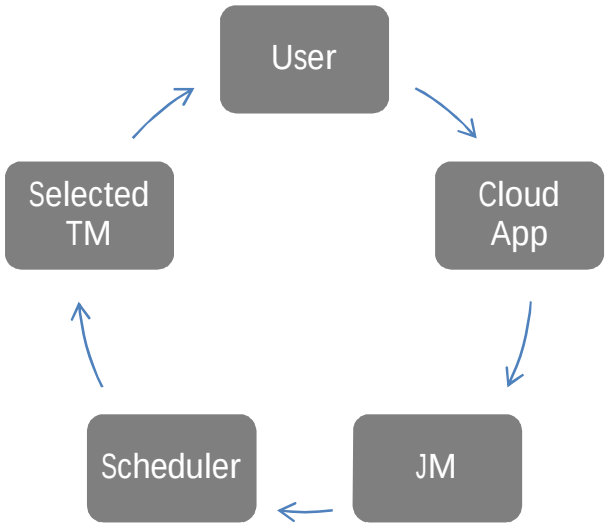

Fig.3 Architecture Diagram for fast parallel data processing in the cloud

C.**JOB ESCRIPTION:** Jobdescription based on directed acyclic graph (DAGs) in that vertices and edges are used vertices is used to represent job's individual tasks and edges denote communication channels. Figure 4 illustrates the simplest possible Job Graph. It consist of one input, one task, and one output vertex one major goal of job graphs has been simplicity Users shouldbe able to describe tasks and their relationships on an conceptual level. Therefore, the Job Graph does not explicitly model task parallelization and the mapping of tasks to instances. However, users who want to wish it influence these aspects can provide annotations to their job description. The Job graph focuses on simplicity and in that no explicit modeling of parallelization and no explicit assignment to virtual machine and channel types is used. Users can provide annotations to influence construction of schedule.

- **Number of subtasks:** It is used to represent adeveloper it can declare developer Task to be suitable for parallelization.
- **Number of subtasks per instance:** Each Subtask is assigned by default in that separate instances are used.
- **Sharing instances between tasks:** Different tasks are usually assigned to different sets of instances unless prevented by another scheduling Restriction.
- **Channel types:** Each edge is used to connect two Vertices the user can determine a channel type.
- **Instance type:** A subtask can be executed on Different instance types which may be more or less suitable for the considered program.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

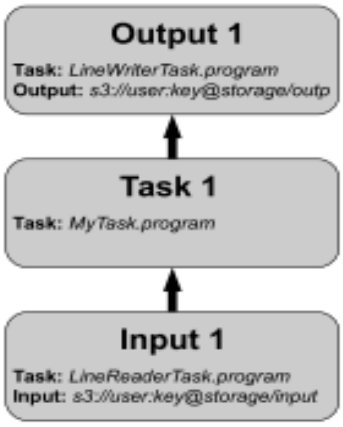

Fig 4an example of job Graph Nephele

D.**EXECUTION GRAPH**: Primary scheduling data structure consists of explicit parallelization in that tasks can be declared parallelizable and in that tasks specify wiring of subtasks. In second explicit assignment to virtual machines it specifies by ID and type in that type refers to hardware profile. Data processing in the cloud computing dealing with on demand virtual machine allocation in that some issues with on-demand allocation is when to allocate and deallocate virtual machines .There is no guarantee of resource availability in that stages ensure three properties first is virtual machine of upcoming stage are available. Second is all workers are set up and ready to use and last is data of previous stages is stored in persistent manner.

In Fig 5shows Execution Graph constructed from job Graph .Execution Graph is pure direct acyclic graph (DAG).Execution Graph equals the user's job graph. Every vertex of the original job graph there exist a Group vertex in the execution graph in that group vertices also represent distinct tasks. During the job processing edge between groups vertices do not represent any physical communication paths.Nephele separates the execution graph into one or more instances called as execution stage.Nephele having three different types of channels

- Network channels:
  In network channel two subtask exchange data via TCP connection .It allow pipelined processing. The consuming subtask are immediately transported by producing subtask.As a result two subtasks connected via network channel may be executed on different instances. In Network channels vertices must be in same stage.
- In-Memory channel:
  It is same like network channel difference is only that instead of using TCP connection. The subtasks exchange data using the instances of main memory. In-Memory channel vertices must run on same virtual machine and vertices must be in the same stage.
- File channel: In that two subtasks exchange record via the local file system .The records first entirely written to intermediate file then afterwards read into consuming file. In file channels vertices must run on same virtual machine and vertices must be in different stages.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

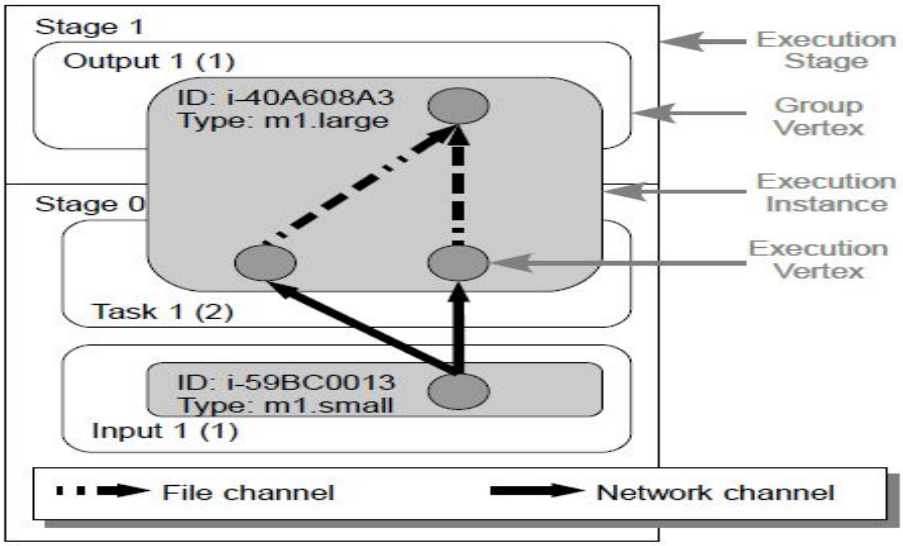

Fig 5 An Execution Graph created from original job Graph

IV. MODULES DESCRIPTION

A.Network module:
It is a Client-Server computing model in which data is distributed and it is usedtodivide the data into number of tasks or workloads between service providers (servers) and service requesters, called clients. In that clients and servers operate over a computer network in that it uses separate hardware. A server machine is a high-performance host which is running on one or more server programs which share its resources with clients. A client also shares its resources; Clients therefore initiate communication sessions with servers which listen to incoming requests.

B. Scheduling Task: The client starts the task to be processes to the job manager. The job manager which performs the function of to read the task and dispatches thetask, and it coordinates and schedules the task .The task manager which is used to allocate and deallocate the resources.

C. Client module:the client which sends the request to the job manager then the job manager execute the task the job manager will schedule the task and process the task and coordinates the task and wait for the replyand after completion it will give the response. The client is the one who initiates the request to the job manager.

D.Job manager Module: The job manager will wait for the task from the client and coordinates the process and it checks the availability of the server if the server is available for the task to be done. It canallocate the resource for execution and wait for the completion response.

F. Task manager module: The task manager will wait for task to be executed and then executes the task and sends the complete response to the job manager in turn to the client.The actual execution of the task it done in the task manager.

V. THE SCHEDULINGALOGORITHM

The scheduling algorithm consists of following point the arrival of new task, the completion of task and last is current execution task. When the current task is active at that that it immediately discarded and the task with highest efficiency is selected for execution.

We chose FCFS algorithm with time complexity concern for scheduling:

Algorithm:

```
Scheduler (process)
{
    If tm1.timespan<=0 then
      Tm1.timespan=process.timeComplexity;
```

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

```
        Tm2.timespan--;
        Tm3.timespan--;
Else if tm2.timespan<=0 then
        Tm1.timespan--;
        Tm3.timespan--;
        Tm2.timespan=process.timeComplexity;
Else if tm3.timespan=0 then
        Tm1.timespan--;
        Tm2.timespan--;
        Tm3.timespan=process.timeComplexity;
End If
}
```

When new job come it first inserted on the ready queue at that time starting time would be the expected finishing time of the current active task .Based on this starting time we then compare its expected utility with the rest of the task in the queue. If it is less than that of the following time span then we reinsert this job to the queue according to its new expected utility we calculate the result according to the following algorithm.

VI. CONCLUSION

In this paper we have discussed the challenges and opportunities for scheduling fast data processing in cloud computing. It is represented with the help of first data processing framework Nephele.We have described Nephele's basic architecture which represent the ability to assign specific virtual machine depend upon type of specific task .processing job which automatically allocate and deallocate virtual machine. In that job execution is helpful to improve resources utilization and to reduce producing cost.

In general we think that our work represent an important contribution to growing fields of parallel data processing

REFERENCES

[1] Amazon Web Services LLC. Amazon Elastic Compute Cloud (Amazon EC2). http://aws.amazon.com/ec2/, 2009.
[2] Amazon Web Services LLC. Amazon Elastic MapReduce.http: //aws.amazon.com/elastic a reduce/, 2009.
[3] AmazonWeb Services LLC. Amazon Simple Storage Service. http://aws.amazon.com/s3/, 2009
[4] D. Batter's, S. Ewen, F. Hueske, O. Kao, V. Markl, and D. Warneke. Nephele/PACTs: A Programming Model and Execution FrameworkFor Web-Scale Analytical Processing. In SoCC '10: Proceedingsof the ACM Symposium on Cloud Computing 2010, pages 119–130, New York, NY, USA, 2010. ACM.
[5] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib,S. Weaver, and J. Zhou. SCOPE: Easy and Efficient ParallelProcessing of Massive Data Sets. Proc. VLDB Endow. 1(2):1265–1276, 2008.
[6] H. chih Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker. Map-Reduce-Merge: Simplified Relational Data Processing on LargeClusters. In SIGMOD '07: Proceedings of the 2007 ACM SIGMODinternational conference on Management of data, pages 1029–1040, New York, NY, USA, 2007. ACM.
[7] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang.Maximum Likelihood Network Topology Identificationfrom Edge-Based Unicast Measurements. SIGMETRICS Perform.Eval. Rev., 30(1):11–20, 2002.
[8] R. Davoli. VDE: Virtual Distributed Ethernet. Testbeds and ResearchInfrastructures for the Development of Networks & Communities, International Conference on, 0:213–220, 2005.
[9] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processingon Large Clusters. In OSDI'04: Proceedings of the 6th conferenceon Symposium on Opearting Systems Design & Implementation, pages10–10, Berkeley, CA, USA, 2004.USENIXAssociation.
[10]E. Deelman, G.Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman,G. Mehta, K.Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob,and D. S. Katz. Pegasus: A Framework for Mapping ComplexScientific Workflows onto Distributed Systems. Sci.Program., 13(3):219–237, 2005.