International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 15 Issue II, May 2019

MANAGEMENT OF DATA INTENSIVE APPLICATION WORKFLOW IN CLOUD COMPUTING ENVIRONMENT

Amrita Singh, Gyanendra Baghel

ABSTRACT

In cloud computing, there are few efficient algorithms in the literature for scientific workflow tasks allocation and scheduling for heterogeneous resources such as those proposed in grid computing context, they usually require a bounded number of computer resources that cannot be applied in Cloud computing environment. Indeed, unlike grid, elastic computing, such as Amazon's EC2, allows users to allocate and release computed resources on-demand and pay only for what they use. Therefore, it is reasonable to assume that the number of resources is infinite. This feature of Clouds has been called "illusion of infinite resources". However, despite the proven benefits of using Cloud to run scientific workflows, users lack guidance for choosing between multiple offerings while taking into account several objectives which are often conflicting.

KEYWORDS- Data intensive, workflow application.

1. INTRODUCTION

Cloud computing is internet based computing which provides web services through service providers. These services are provided to the users on rent like pay-as-use model in which user have to pay according to the access or use of the services [1]. The three basic types of services are provided: Infrastructure as a service (Iaas), Platform as a service (Paas) and Software as a service (Saas) as shown in fig.1.

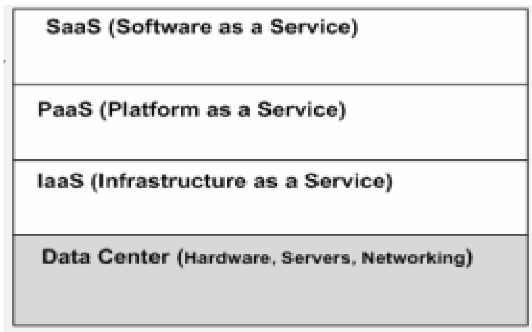

Figure 1: Services on cloud [2]

Iaas provides the physical resources such as memory, processor etc. Paas provides the framework or platform for developing their own applications by using cloud and there is no need to install any platform on their own machine. Paas services such as .Net etc. Saas is basically used for running the existing applications like facebook. The user does not compact with installation of any software on their physical machine. The cloud provides such software for running these types of applications.

2. DATA INTENSIVE APPLICATION

Data-intensive computing is a class of parallel computing applications which use a data parallel approach to processing large volumes of data typically terabytes or petabytes in size and typically referred to as big data. Computing applications which dedicate most of their execution time to computational requirements are deemed compute-intensive, whereas computing applications which necessitate large volumes of data and devote most of their processing time to I/O and manipulation of data are deemed data-intensive [3].

A data-intensive application workflow deals with the high workloads of data to control than its computations of the data. The another meaning of data intensive deals with the transferring the huge amount of data but computational part deals with the processing of the tasks. The transfer of data consumes more time, and also store that data, than the processing of the data. For characterizing the difference between the data-intensive and computer-intensive one aspect is used which is the CCR (Computation to Communication Ratio). The applications with the lower values of this ratio are data intensive applications in nature [4].

3. RELATED WORK

Allocation and preparing workflow tasks in Cloud has gained popularity in recent times and few algorithms are proposed to deal with this problem [5] [6] [7] [8]. In [5], the authors developed a model that uses particle swarm optimization (PSO) for task-resource mapping to minimize the overall cost of execution such that it completes within deadline that user specifies. To tackle the problem of choosing resource among different cloud providers, a binary integer program is projected in [5], where the objective is to decrease the total infrastructure capacity under budget and load balancing constraints.

In [6], the authors offered a model for formulation of the generalized federated placement problem and application of this problem to load balancing and consolidation inside a cloud, wherever one cloud can subcontract workloads to partnering clouds to meet peaks in stipulate. They used an Integer Program formulation of the placement program and provide a 2-approximation algorithm. In [7], the authors proposed a binary integer program formulation for cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads. In [8], the authors proposed a set of heuristics to cost-efficiently schedule deadline-constrained computational applications on both public cloud providers and private infrastructure. Although, a large amount of these studies consider unbounded number of resources, however, they convert the initial problem (bi-criteria) to constraints problem.

4. SCHEDULING ALGORITHMS

4.1 COST BASED ALGORITHM

In the cost-based approach, we focus only on minimizing the execution and communication costs of using a set of virtual machines incurred by the execution of a given workflow. However, for each obtained feasible solution by using an allocation strategy the overall completion time corresponding is computed. Recall that the intention is to assign tasks to virtual machines respecting the precedence constraints. The cost-based approach is an application allocation and scheduling algorithm for an unbounded number of virtual machines. As mentioned earlier, users can request and obtain sufficient resources at any time. The approach proposed has three most important phases, namely:

- **Tasks sorting phase**

Tasks sorting phase In order to group the tasks that are independent of each other, the given workflow (DAG) is traversed in a top-down fashion to sort tasks at each level. As a result, tasks belonging to the same level do not exchange data and can be executed in parallel (because they are not related by precedence constraints).

- **Resource allocation phase**

In this phase the selection of an ”optimal” virtual machine for each task is decided. In other words, the virtual machine which gives minimum execution and communication costs for a task is selected and the task is assigned to that virtual machine. Following strategies: top- down, bottom-up and mixed exploration and portion strategy.

The top-down strategy consists of starting by the allocation of the initial task (level 1) to the virtual machine which gives minimum execution cost. After this assignment, the graph is traversed in a top-down fashion from level 2 to level L.

The bottom-up strategy consists on starting by the allocation of the finish task (the last level). After this allocation, the graph is traversed in a bottom-up fashion from level L − 1 to level 1. The mixed strategy starts by assigning the tasks belonging to the intermediate level, i.e. k ∈ {2, ...,L − 1}.

The mixed strategy starts by assigning the tasks belonging to the intermediate level, i.e. k ∈ {2, ...,L − 1}. Given starting level k, therefore the assignment of the tasks belonging to this level is only based on the execution cost

- **Pareto selection phase**

In this phase, we first compute the general completion time corresponding to each assignment and then only non-dominated solutions are maintained.

4.2 TIME BASED ALGORITHM

The time-based approach attempts to minimize the overall completion time (i.e. execution and communication time). As the cost-based approach, the time-based approach is an application matching and scheduling algorithm for an ”unbounded” number of VMs, which has three major phases, namely: a tasks sorting phase i), ii) an allocation phase and iii) Pareto selection phase.

- **Tasks sorting phase**

This phase is the same as for the cost based algorithm. Recall that this phase allows to group the workflow application tasks that are independent of each other.

- **Resource allocation phase**

The cost-based and the time-based approaches differ mainly at resource allocation phase. Indeed, the first one approach focus on minimizing the cost function while the second approach attempts to minimize the time function. The objective of the resource allocation phase of the time-based approach is to choose the virtual machine that minimizes the finish time of each task and the task is assigned to that virtual machine. The three allocation strategies, detailed above, are also applied to explore this approach.

- **Pareto selection phase**

Recall that at the end of the previous phase L assignment of all tasks is obtained. In this phase, we first compute the overall completion time corresponding to each assignment and then only non-dominated solutions are maintained.

5. CONCLUSION

In this paper, we have proposed three bi-criteria complementary approaches to tackle the allocation and scheduling workflow problems in Cloud environments. Moreover, in order to assess the quality of obtained solutions by our approaches we have proposed two lower bounds for each considered criterion.

Unlike existing works, these approaches take into account two conflicting criteria simultaneously: i) execution cost and ii) execution time. Moreover, they offer more flexibility to consumer to estimate their preferences and choose a desired schedule from the obtained efficient solutions. More precisely, the first one focused on the cost incurred by using a set of resources, while the second approach attempts to minimize the overall execution time. The third approach is based on the two first approaches for selecting only the Pareto solutions.

Moreover, we plan to extend the proposed work to take into account others criteria like carbon emission and energy cost. In addition, it is interesting to adapt the proposed approaches to the case where a task cannot be executed on all virtual machine types.

REFERENCES

[1] R. Buyya, A. Beloglazov, J. Abawajy, "Energy-aware resource allocation heuristic for efficient management of data centers for cloud computing", Future Generation Computer Systems, pp.755-768, 2012.

[2]http://blogs.msdn.com/b/jmeier/archive/2010/02/11/software-as-a-service-saas-platform-as-a-service-paas-and-infrastructure-as-a-service-iaas.aspx

[3] http://en.wikipedia.org/wiki/Data-intensive_computing

[4] J. Becla, "Real Life data intensive applications-challenges and solutions", Salishan Conference on high speed computing, 2010.

[5] J. Tordsson, R. Montero, R. Moreno-Vozmediano and I. Llorente, Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers, Future Generation Computer Systems, Volume 28, Issue 2, pp. 358-367, 2012.

[6] D. Breitgand, A. Marashini and J. Tordsson, Policy-driven service placement optimization in federated clouds, Technical report, IBM Haifa Labs, 2011.

[7] R. V. den Bossche, K. Vanmechelen and J. Broeckhove, Cost-optimal scheduling in hybrid iaas clouds

for deadline constrained workloads, In Proceedings of the 3rd IEEE international conference on cloud computing, pp. 228-235, 2010.

[8] R. Van den Bossche, K. Vanmechelen and J. Broeckhove, Cost-Efficient Scheduling Heuristics for Deadline Constrained Workloads on Hybrid Clouds, on Cloud Computing Technology and Science, IEEE International Conference, pp. 320-327, 2011.