A Survey on Cloud Computing Scheduling Algorithms

Waleed Abd Elkhalik[1], Ahmad Salah[3], Ibrahim El-Henawy[3]
[1,2,3]*Department of Computer Science, faculty of Computers and Informatics, Zagazig University, Al-Zagazig, Sharqiyah, Egypt*

Abstract

Resource allocation in cloud computing is comprised of two main functions: static resource scheduling and dynamic and it also includes subsequent activities like types of resource scheduling, resource scheduling algorithms and their evolution. It displays a vital character in efficient utilization of resources. For any resource scheduling algorithm, the cost, time and energy are most the important QoS parameters. Resource Scheduling Algorithm (RSA) plays an important role in scheduling and execution of most appropriate resources to workloads. It refers to the process of appropriate generation of the schedule that decides which tasks will be mapped on to which resources. In order to ensure QoS to the cloud workload according to the requirements of user. Sometimes RSAs adopt dynamic behaviour whereby resources are scheduled after resource provisioning. Such algorithms are called dynamic RSAs and are considered more efficient than the static resource scheduling. In this paper, we describe all the important resource scheduling approaches that aim at optimizing the user Quality of Service (QoS) metrics such as cost, makespan, reliability, priority and provide cost-effective executions and achieve objectives such as load balancing, availability and reliability in the cloud environment.

Keywords —*Cloud Computing, Resource management, Resource Scheduling, Need of Scheduling, Makespan, RSAs, QoS*

I. INTRODUCTION

Clouds provide a very large number of shared and scalable pool of virtual resources such as data centres, storages, Networks, firewalls and software in form of services, including stages for computation, and provides a convenient on-demand network access to a shared and scalable pool of virtual resources such as servers, storage, and services [1].

Cloud computing can be classified into the four master types [12] as shown in Fig. 6. :

- Public clouds. The cloud services are provisioned for open use by the public as everyone may register, owned managed or use services.
- ***Private clouds.*** The cloud services are provisioned for exclusive use by a single organization, whose data and processes are managed within the organization without the restrictions of the network bandwidth, security exposures and legal requirements.
- ***Community cloud.*** . The cloud services are provisioned for exclusive use by a specific community of consumers from different organizations, whose data and processes are managed within one or more of organizations
- ***Hybrid clouds.*** The cloud services are a composition of two or more distinct cloud services (private, community, or public)

The resources which provide the services are somewhere placed on the Internet rather than our local system. And all this is provided to end users just like any other utility which they can access anytime, anywhere in the world with the help of Internet. It frees the users from the burden of managing hardware, software, storage and networks by providing them with a pool of virtualized resources according to their need.

As shown in Fig. 2, Cloud services can be delivered or categorized in three models relying upon the service which they expect to give to the client. These classes are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [4].

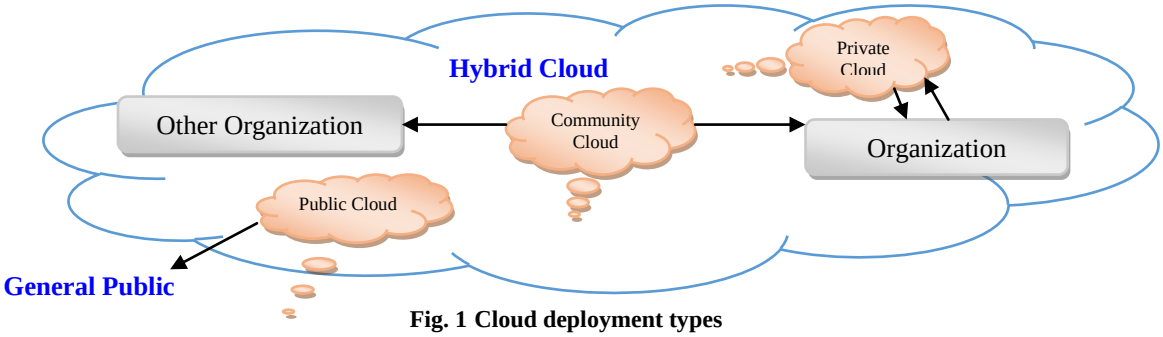

Fig. 1 Cloud deployment types

- Infrastructure as a Service (IaaS): are self-service models for accessing, monitoring, and managing remote datacenter Infrastructures such as storage, networking, and networking services (e.g. firewalls). Instead of having to purchase hardware, uses can rent just for use similar to electricity or other utility billing.
- Platform as a Service (PaaS): are used for applications, and other development, while providing cloud components to software supporting the full ‘‘Software Lifecycle’’ that allows cloud consumers to develop cloud services and applications (e.g. SaaS).
- Software as a Service (SaaS): represent the largest cloud market and are still growing quickly, which delivers special-purpose software that is remotely available by consumers.

On one hand, the service providers aim at maximizing their profit and return on investment, while on the other hand, users demand for cheapest, fastest and the most reliable services. In order to fulfil both the ends’ demands, an efficient and proper resource management mechanism must be founded [5].

The most challenging problem in cloud is the resource scheduling, and considered the second stage of resource management after resource provisioning.

Resource provisioning is defined to be the stage to identify proper resources needed to accomplish a given workload based on QoS requirements described by cloud consumers whereas resource scheduling is mapping and execution of cloud user workloads based on selected resources from resource provisioning stage as shown in Fig. 3 and Fig. 4. Workloads or tasks are been submitted to be executed with their details and based on these details the broker allocates the adequate resource(s) for a given tasks and determines the feasibility of resources provisioning based on QoS requirements [8]. Broker is also responsible for monitoring the performance to add or release any extra resources to resource pool. After successful provisioning of resources scheduling is done in second phase as workloads has been sent to the scheduler and processed in workload queue, scheduling agent maps the provisioned resources to given workload(s), execute the workload(s) and release the resources back to resources pool after successful completion of workload(s).

Based on the incoming workloads’ details (cloudlet), scheduler schedules them and then mapped them with the available and appropriate resources based on the scheduling policies and according to the QoS parameters mentioned SLA the role of dispatcher is coming to dispatch the workloads to be executed.

Within the execution of the workloads comes the monitoring processes for resources such as checking resource scheduling status like number of resources needed is available or not has been checked , also QoS monitor to check whether all the workloads are executing within their specified range or not such as checking whether workloads are executed before desired deadline or not. There is violation of SLA if workload executes after desired deadline [9].

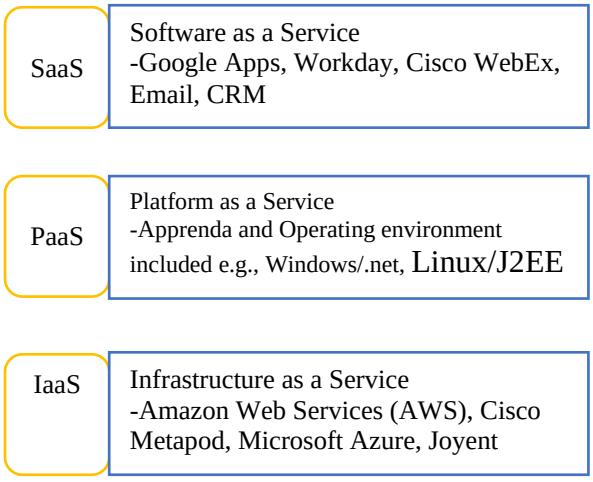

Fig. 2 Cloud Deployment Models

Based on QoS requirements, scheduling of resources for adequate workloads is a challenging issue. For an efficient scheduling of resources, it is necessary to consider the QoS requirements [6]. There is a need to uncover the research challenges in resource scheduling to execute the workloads without affecting other QoS requirements.

Varieties and differences of resource scheduling criteria and parameters make different categories of Resource Scheduling Algorithms (RSAs). This research work will discuss resource scheduling, the second phase of resource management and show how effective resource scheduling reduces execution cost, execution time, energy consumption and considering other QoS requirements like reliability, security, availability and scalability.

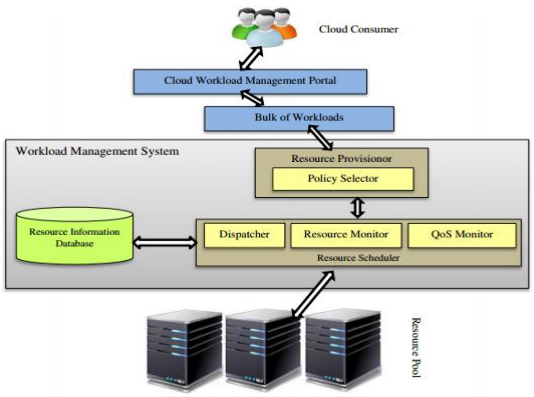

Fig. 3 Cloud Resource Scheduling [6]

In cloud environment, cloud consumer and cloud provider are two parties. Cloud consumer submits workloads while cloud provider provides resources for execution of workloads. Both the parties have different requirements: provider wants to earn as much profits as possible with lowest investment and maximize utilization of resources while consumer wants to execute workload(s) with minimum cost and execution time. However, executing number of workloads on one resource will create interference among workloads, which leads to poor performance and reduces customer satisfaction. To maintain the service quality, providers reject the requests that result in unpredictable environment [2].

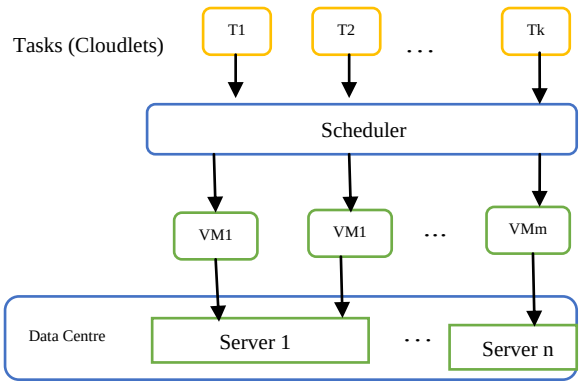

Fig. 4 Cloud Task Scheduling

One of the challenges of resource scheduling is that both cloud consumers and providers are not willing to share their information with each other, the challenges include also security, uncertainty, dispersion, and heterogeneity of resources. All of these challenges and limitations make task and resource scheduling more difficult and an NP-hard optimization problem and to resolve it we need untraditional RSAs [3].

The organization of rest of this paper is as follows: Section 2 presents the objectives of resource scheduling and need of scheduling in cloud. Section 3 presents the background of resource scheduling in the cloud environment. Section 4 describes the resource scheduling algorithms. Finally, Sect. 5 concludes with future research remarks.

II. NEED OF SCHEDULING IN CLOUD

One of the objectives of scheduling the resources on the cloud is to identify the suitable resources for scheduling the appropriate workloads (cloudlets) on time and to increase the effectiveness of resource utilization. Also to minimize the amount of resources for a workload without violating the level of service quality, or minimize workload completion time (or maximize throughput) of a workload. For better resource scheduling, best resource workload mapping is required. Other objective of resource scheduling is to identify the adequate and suitable workload that supports the scheduling of multiple workloads, to be capable of achieving the QoS requirements such as CPU utilization, availability, reliability, security, privacy etc. for cloud workload [7]. Other objectives such as Service Level Agreement (SLA), Quality of Service (QoS), deadline constraints, load balancing and the profit for both cloud providers and consumers as shown.

III.SCHEDULING IN THE CLOUD

Resource scheduling is the second stage of resource management process after resource provisioning as shown in Fig. 5.

Resource provisioning comes in the first stage of resource management and comprises of two main procedures.

- ***Resource Detection*** **:** the process of finding the list of available resources
- ***Resource Selection :*** the process of choosing the best resource from list generated by resource detection based on SLA or QoS requirement described by cloud consumer

After provisioning of resources, workloads (cloudlets) are submitted to resource scheduler. Then the resource scheduler will ask to submit the workload for provisioned resources. Resource scheduling is done after resource provisioning comprises of three main procedures Resource Mapping, Resource Execution and Resource Monitoring.

- Resource Mapping : the process of mapping of workloads with appropriate resources based on QoS requirements described by the cloud consumers in terms of SLA
- Resource Execution: the process of executing appropriate resources to the suitable workloads on time to achieve the efficient utilization of resources.

- Resource Monitoring: the process of checking the current workload during the executing of particular cloudlet. If the value of required resources are more than the value of provided resources them as for more resources.

After successful execution of cloud workloads, releases the free resources to resource pool and scheduler is ready for execution of new cloud workloads.

Two main aspects of resource management

- Consumer: that wants to execute his workload with the minimum time and minimum cost without violation of his SLA.

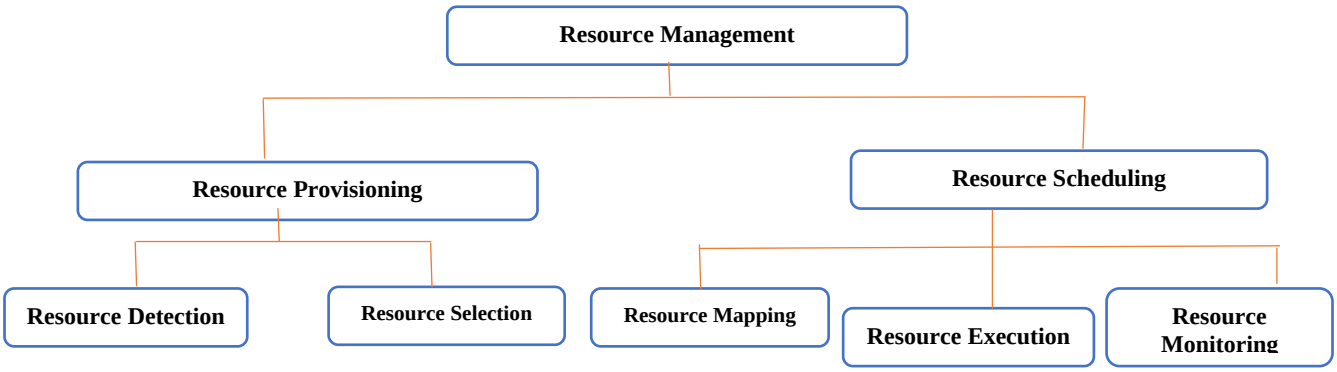

Fig. 5 Resource Management Process [10]

- Provider: that wants to execute the workload with the minimum number or resources and with maximum throughput.

In the SLA for both parties (Provider and consumer), there is a possible deviation to achieve QoS requirements for both of them. Resource monitoring used to measure the SLA deviation, QoS requirements and performance of the resource utilization for both of physical and virtual infrastructures. Performance optimization can be best achieved by efficiently monitoring the utilization of computing resources. So, we need a comprehensive intelligent monitoring agent to analyse the performances of resource execution.

A. Types of Scheduling Algorithms [11]

1) Static scheduling v/s dynamic scheduling

It based on the information related to the status of the task and resources as their introduced in advance the scheduling is done in that's the static scheduling, but in the dynamic scheduling the tasks are allocated in runtime.

2) Online v/s Batch mode scheduling

It is dependent on time of executing the tasks. In the tasks is arriving it is executed immediately, but in batch mode scheduling is or disconnected planning in which tasks are executed in the specific time interval.

3) Preemptive v/s Non-preemptive scheduling

It is dependent on ability of occurring interrupted to the running tasks during the scheduling process. In Preemptive scheduling the scheduling process can be interrupted if a high priority tasks enters the queue, but in Non-Preemptive scheduling the running tasks cannot be interrupted by any other tasks even if it have a higher priority. Any other tasks witch enters the queue has to wait until the current tasks finish it's execution.

Fig. 6 shows that there are many and different resource scheduling algorithms and their research percentage [Bargaining, Compromised Cost Time, Cost, Time, Profit, Dynamic, Priority, Hybrid, VM, Nature Inspired and Bio Inspired, Optimization, Energy and QoS and SLA based RSA]. Notes that the energy research based scheduling algorithms has the maximum research (18 %) and SLA and QoS based resource scheduling algorithms (18 %) while priority, profit and compromised cost & time based resource scheduling algorithms has only (3 %). Nature inspired and bio inspired based resource scheduling algorithms contributes (9 %) and bargaining, optimization, and cost based resource scheduling algorithms contributes (8 %), (8 %) and (7 %) respectively.

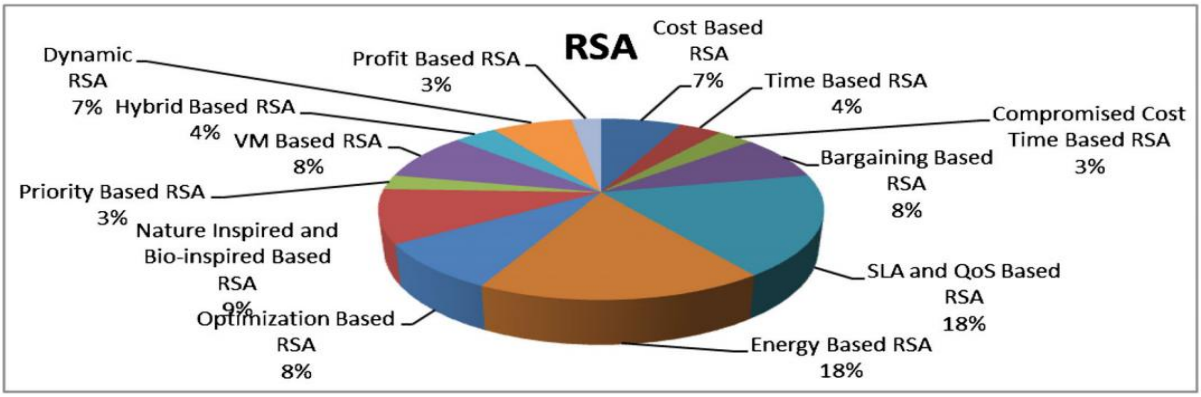

Fig. 6 Resource Scheduling Algorithms in Cloud [9]

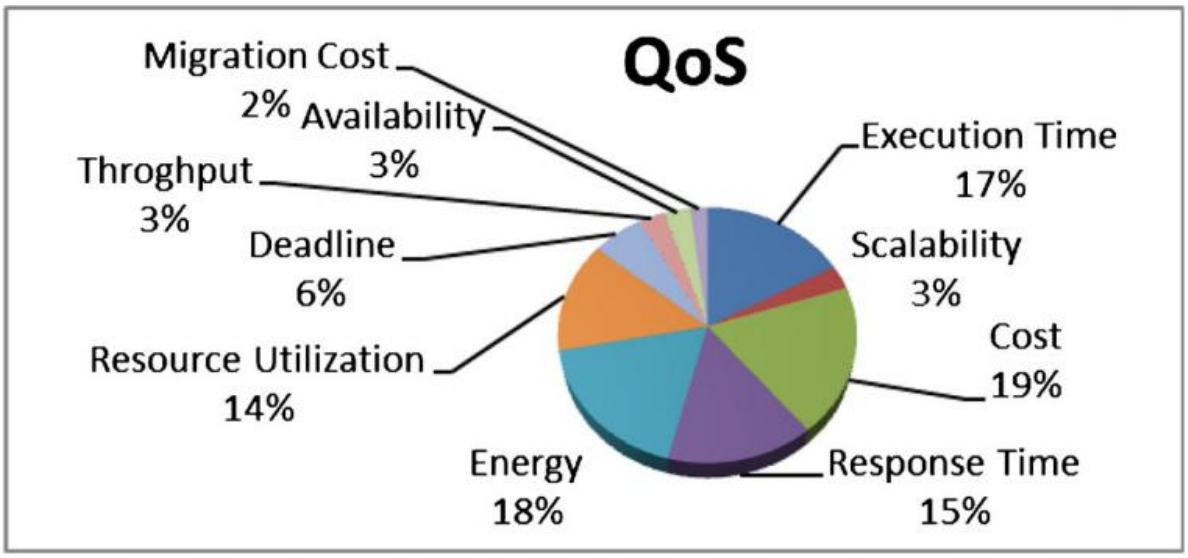

Fig. 7 QoS Parameters in RSAs [9]

B. Quality of Service (QoS).

To present the variant and large type of cloud services there are four types of cloud deployment available as shown in Fig. 1.

In addition to continuous increase of cloud consumers and requests to the cloud services, the level of agreements to their requirements must not be degraded. Quality of service refers to the ability of cloud providers to attain the SLA between consumers and cloud providers. Because of the complex and dynamic nature of Cloud, QoS is become more challenge issue in cloud computing. There are many metrics that will affect the quality of services (QoS Parameters). Fig. 9 shows the percentage of research papers that are considering different QoS parameters.

- Migration Cost
- Availability
- Throughput
- Deadline
- Resource Utilization
- Energy
- Execution Time
- Scalability
- Cost
- Response Time

IV. RESOURCE SCHEDULING ALGORITHMS

In order to attain QoS to the cloud workload according to the consumer requirements, the algorithm that performs the scheduling process of workloads to the resources must be efficient enough. Resource Scheduling Algorithms have an important role to achieve the SLAs between the cloud provider and consumers.

Resource Scheduling Algorithms can be categorized into:

- Cost Based RSA
- Time Based RSA

- Compromised Cost Time Based RSA
- Bargaining Based RSA
- Profit Based RSA
- SLA and QoS Based RSA
- Energy Based RSA
- Optimization Based RSA
- Nature Inspired and Bio-inspired Based RSA
- Priority Based RSA
- VM Based RSA
- Hybrid Based RSA
- Dynamic and Adaptive Based RSA

V. CONCLUSION AND FUTURE WORK

In this research paper, A brief introduction to the cloud concepts, cloud deployment types and models, cloud resources scheduling in the cloud environments, need of scheduling, scheduling procedure in the cloud, Types of scheduling algorithms and it's classification in the cloud, quality of services and it's parameters, and finally the varies type of resource scheduling algorithms and it's classification have been presented.

Our future work is to produce an improved flower pollination based task scheduling algorithm in cloud computing to minimize the makespan.

REFERENCES

[1] Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." Journal of internet services and applications 1.1 (2010): 7-18.

[2] Singh, Sukhpal, and Inderveer Chana. "Cloud based development issues: a methodical analysis." International Journal of Cloud Computing and Services Science (IJ-CLOSER) 2.1 (2012): 291-302.

[3] Singh, Sukhpal, and Inderveer Chana. "QRSF: QoS-aware resource scheduling framework in cloud computing." The Journal of Supercomputing 71.1 (2015): 241-292..

[4] Manvi, S. S., & Shyam, G. K. (2014). Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. Journal of Network and Computer Applications, 41, 424-440.

[5] Varshney, Shweta, and Sarvpal Singh. "A Survey on Resource Scheduling Algorithms in Cloud Computing." International Journal of Applied Engineering Research 13.9 (2018): 6839-6845..

[6] Chana, Inderveer, and Sukhpal Singh. "Quality of service and service level agreements for cloud environments: Issues and challenges." Cloud Computing. Springer, Cham, 2014. 51-72..

[7] Singh, Sukhpal, and Inderveer Chana. "QoS-aware autonomic resource management in cloud computing: a systematic review." ACM Computing Surveys (CSUR) 48.3 (2016): 42.

[8] Singh, Sukhpal, and Inderveer Chana. "Q-aware: Quality of service based cloud resource provisioning." Computers & Electrical Engineering 47 (2015): 138-160.

[9] Singh, Sukhpal, and Inderveer Chana. "A survey on resource scheduling in cloud computing: Issues and challenges." Journal of grid computing 14.2 (2016): 217-264..

[10] Singh, Sukhpal, and Inderveer Chana. "Cloud resource provisioning: survey, status and future research directions." Knowledge and Information Systems 49.3 (2016): 1005-1069..

[11] Chen, Huankai, et al. "User-priority guided Min-Min scheduling algorithm for load balancing in cloud computing." Parallel computing technologies (PARCOMPTECH), 2013 national conference on. IEEE, 2013.

[12] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).