ISSN 2229-5518

QoS Aware Resource Scheduling in Internet of Things-Cloud Environment

Krishnapriya .S, Joby P. P

Abstract— Internet of things (IoT) is a novel idea of linking objects. Establishment and functioning of IoT rely on existing networks as well new network infrastructures. Cloud environmentis an critical factor for the establishment and functioning of an IoT system. When cloud infrastructure is used for the establishment of IoT, it has to deal with heterogeneous workloads. These workloads include diverse applications with various priorities, performance and resource demands. There are tasks that are limited to the kind of machine types they demand for execution. Also, majority of the IoT applications are time bound and therefore task response time is an important QoS factor. Sub-optimal resource scheduling and sub-optimal load balancing will cause performance degradation. In this paper, we have proposed the concept of a QoS aware resource scheduling technique called Load Balanced Particle Swarm Optimization scheduling, which can reduce application response time and ensures load balancing..

Index Terms— Cloud computing, Internet of Things, Quality of Service, Load Balancing, Optimization, PSO, Resource Scheduling.

---

1 Introduction

Internet of Things (IoT) is the next revolutionary technology in converting the present communication infrastructure into a completely future network. Tasks and workloads in the IoT - Cloud platform are of various types and priorities. Dynamic resource scheduling for heterogeneous workloads in IoT is critical for ensuring Quality of service (QoS). Quality of Service is a challenging requirement of cloud computing. Contents transmitting through the IoT system are of high priority, best effort, guaranteed, mission critical or real time applications. The problem of assigning various types of workloads into different machines is an np-hard problem. Resource allocation in cloud environment is analogous to an online algorithm where the algorithm is unaware about the list of incoming inputs. Since prior knowledge of input items are not known to the allocation problem, a dynamic capacity allocation suits best for the IoT - Cloud system.

2 Internet of Things

The traditional Human to Machine (H2M) communication system is gradually changing into the Machine to Machine (M2M) communication referred as Internet of things. In this scenario, all the devices in the communication system are considered as 'things' instead of computers. These 'things' could be vehicles, home appliances, sensing devices, military equipments, medical tools, container in a shipment etc.

The IoT introduces the concept of linking all physical things in an environment based on internet for communication and transmitting messages. The goal of IoT is to support intelligent identification, location, tracking, monitoring and management of objects. IoT is therefore based upon the integration of several communication solutions, identification and tracking technologies, sensor and actuator networks and distributed smart object [1]. Much more importance is for the way that IoT promises a smart environment that would offer immense savings of time, energy and resources. IoT is expected to contain huge numbers of sensors collecting and passing on data on environmental conditions, physiological measurements, machine operational data, etc.

IoT is the concept where the future environment will be surrounded by smart objects with interdependent networks of smart communication protocols, smart homes, smart traffic, smart city, smart health, etc. Enormous number of services will be made available to interact with these 'smart objects' over the Internet, like sending a query, changing their state, convey the information associated with them. While taking all this into account, Quality of Service is going to be very huge. It thus creates a world where the physical objects are perfectly integrated into the information network, and where the physical objects can become active participants in areas like business processes, life critical applications, etc. Moreover, from the technical point of view, the basic idea behind this concept is the pervasive presence of a variety of things or objects such as Radio-Frequency IDentification (RFID) tags, sensors, actuators, mobile phones, etc. These objects are able to interact with each other and cooperate with their neighbors through unique addressing schemes in order to achieve common goals [2].

IoT-cloud infrastructure is the interconnection of sensing and actuating devices providing the ability to share information across platforms. A number of challenges are there in establishing an environment like this - one of them is the ability to meet quality of service.

- *Krishnapriya .S is currently pursuing masters degree program in Computer science and engineering in PPG Institute of Technology,India, E-mail: krishnapriya.sivan@gmail.com*
- *Joby P.P is Associate Professor in Computer science and engineering in PPG Institute of Technology, India. E-mail: jobymone@gmail.com*

ISSN 2229-5518

3 Quality of Services

Quality of Service in cloud environment is the capacity to render priority to applications and thus guarantying a certain performance level. QoS requirements depend upon the application type and end users. Three levels of QoS in a heterogeneous system are best effort, Differentiated and Guaranteed. There are different approaches to ensure QoS in cloud platform such as scheduling management, network traffic management, dynamic capacity provisioning etc. The heterogeneous types of applications and the quality requirements in a cloud environment are shown in fig. 1.

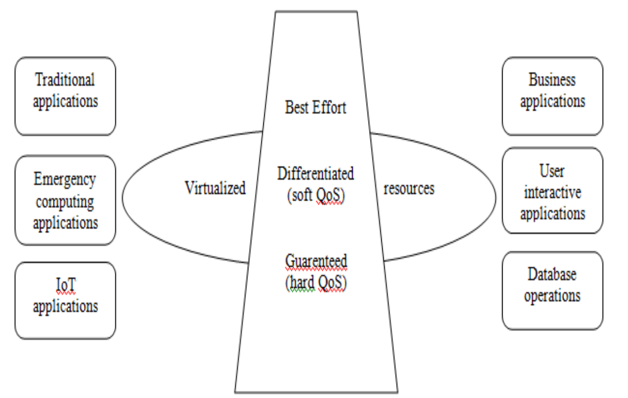

Fig. 1. Applications and Quality of Service

Quality requirements for IoT application include efficient energy utilization, latency, network resource consumption, optimal utilization of computational recourses etc. In this work, we are focusing on the optimal resource allocation with balanced load and minimum response time in IoT-Cloud platform. Efficient allocation of resources is important to optimize resource utilization.

4 Related Works

Paper [3] proposes a data storage framework for cloud platform that renders storage of huge IoT data efficiently and also integrates unstructured and structured data. This framework also combines multiple databases and manages diverse data generated and collected by sensors. The devices involved in IoT generate massive data that would rapidly increase and therefore the IoT storage solutions must also be efficient to handle and manage this huge data. The proposed framework has four main modules namely; file repository, database module, service module and resource configuration module. The database management module is developed for structures data, which aggregates different databases. To handle unstructured data, the file repository module uses hadoop distributed file system. Also, for supporting the access of remote data, this architecture generates RESTful service to render HTTP interface independent of platform.

Paper [4] focuses on message scheduling which points to service provisioning. Here, messages are grouped into either high priority or best effort messages and then the suitable QoS scheduling algorithm is suggested. The different sensor nodes are categorized into IoT subgroups, each subgroup with a broker delivering for all nodes and handles separate queues for best effort and high priority messages. For message scheduling, network layer routing algorithms are also considered in this architecture.

Paper [5] proposes a dynamic resource allocation scheme for cloud environment, that has considered both task and machine diversity. Here, workloads are grouped into different classes having same characteristics using K-means clustering. Then the algorithm dynamically set the number of active servers to save energy.

5 Problem Description

One of the biggest challenges in this integral part of the future internet is the caliber to manage heterogeneous types of tasks in this system along with ensuring Quality of Services. IoT may rely on cloud computing for its functioning. When cloud environment supports such a system, it has to handle diversified data. Also, majority of applications in IoT are time bound or real time applications. Therefore, dynamic resource provisioning is another critical area that has to be addressed by cloud system. Moreover, when analyzed, resource types in a cloud infrastructure are also heterogeneous. Heterogeneous system consists of set of heterogeneous processors with different memory capacity, processing speed etc. IoT workloads include diverse applications with different priorities, performance and resource requirements. The various tasks will experience different execution cost in different machines, so task scheduling is a crucial issue to improve the efficiency of this architecture. The failure in considering the heterogeneity of both workloads and machines may cause sub-optimal results and scheduling delays. The proposed algorithm considers both machine heterogeneity and workload heterogeneity.

6 Proposed System

Particle Swarm Optimization is a metaheuristic algorithm that tries to find an optimal solution to a problem from a set of feasible solutions. In order to provide an optimal resource scheduling with balanced load and minimal response time, we have proposed a Load Balanced Particle Swarm Optimization scheduling technique. This algorithm dynamically schedules diversified tasks on to appropriate processors, along with equalizing load across machine. Here, optimal resource allocation with balanced load and minimum response time is the Quality service requirements that we have considered. Workloads will be assigned to virtual machines in an iterative manner, so that when a new task arrives rescheduling of tasks are carried out to find the best solution.

In this algorithm, a particle represents a solution for the resource scheduling. Let M1 and M2 be two machines and W1, W2, W3 and W4 be the workloads that need to be allocated to

ISSN 2229-5518

the machines. Three possible candidate solutions to the allocation task are shown in fig. 2. Each of these candidate solutions represents a particle in the algorithm.

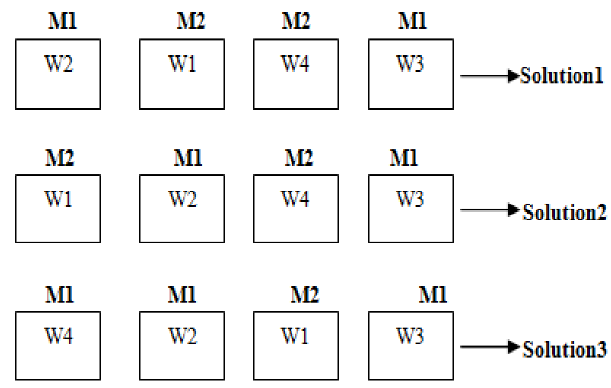

Fig. 2. Candidate solutions for task.

The solution depends upon the current resource capacity, amount of workloads, number of available machines etc. Each particle has its own position and a velocity factor. In the load balanced PSO algorithm, each particle adjusts its position dynamically, depending upon its previous position and the other particles' position. The gBest and lBest values are called fitness values where lBest is the particles personal best position and gBest is the best particle's position in the neighborhood. Fig. 3 represents searching of a new position for a particle, where Pi is the current position, Vi is its current velocity, Vi+1 is the new velocity and Pi+1 is the updated position of the particle.

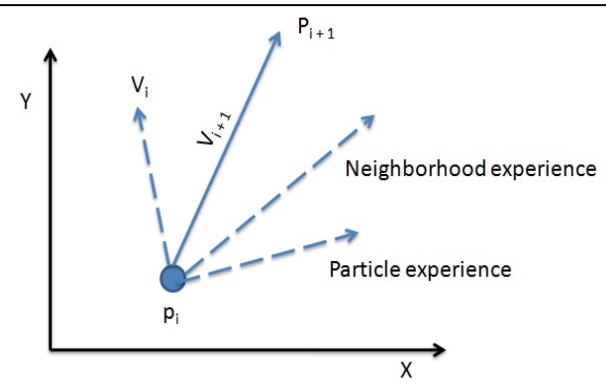

Fig. 3. Update location for each task

Equation (1) is used to calculate new velocity,

Vi+1 = Vi + c1*r*(lBest-pi) + c2*r*(gBest-pi) (1)

and new position is updated using (2),

Pi+1 = Pi + Vi+1 (2)

The pseudo code for the algorithm is shown below.

Step 1: For each particle,
Initialize its position and velocity
Step 2 : For each particle,
i) Calculate fitness value,
ii) If current fitness value better than previous fitness value,
iii) Update lBest= current value
Step 3 : For each particle,
i) Find best fitness value in the neighborhood,
ii) If lbest is better that gBest,
iii) Update gBest = lBest,
iv) Find velocity using equation 1,
v) Update position of particle using equation 2

Repeat step 2 and step3 till all iterations are reached.

In this method, the new position of the particle depends upon the previous position and position of surrounding particles, the workloads will be allocated to the appropriate machine only. Also, since the location of the tasks is rescheduled in each iteration based on the best fitness value, this algorithm ensures balancing of load across machines.

7 Conclusion

IoT system includes different types of applications with various quality service requirements and resource constraints. Cloud infrastructure provides resources with different processing capacities and memory spaces. In order to achieve an optimal scheduling, the diversity of both applications and resources has to be taken into account. Also, we have to avoid situations where few resources are overloaded while others are underutilized. A Particle Swarm Optimization algorithm is a simple and fast in responding technique. In this work, we have taken reduced application response time as the quality requirement. Here, the concept of a scheduling technique called load balanced particle swarm optimization scheduling is proposed that will allocate tasks to appropriate machines along ensuring load balancing and reduced application response time. As future work, we plan to extend our techniques to support multiple levels of QoS requirements.

References

[1] Giusto.D, Iera.A, Morabito.G, and Atzori,L., "The Internet of Things," *Springer*, ISBN: 978-1-4419-1673-0, (Eds.). 2010.
[2] D. Culler, D. Estrin, and M. Srivastava, "Guest Editors' Introduction: Overview of Sensor Networks," vol. 37, no. 8, 2004.
[3] Lihong Jiang, Li Da Xu, Hongming Cai, Zuhai Jiang, Fenglin Bu and Boyi Xu, "An IoT-Oriented Data Storage Framework in Cloud Computing Platform," *IEEE transactions on Industrial Informatics*, vol. 10, no. 2, May 2014.
[4] Saima Abdullah and Kun Yang, "A QoS Aware Message Scheduling Algorithm in Internet of Things Environment," *IEEE Online Conference on Green Communications (OnlineGreenComm)* 2013.
[5] Qi Zhang, Mohamed Faten Zhani, Raouf Boutaba and Joseph L.Hellerstein, "Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud," *IEEE transactions on Industrial Informatics*, vol. 2, no. 1, January-March 2014.

[6] Abirami S.P. and Shalini Ramanathan, "Linear Scheduling Strategy for Resource Allocation in Cloud Environment," *International Journal on Cloud Computing: Services and Architecture (IJCCSA)*,vol.2, no.1,February 2012.

[7] R.Buyya, C.S.Yeo, S.Venugopal, J.Broberg, and I.Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. Future Generation Computer Systems," 25(6):599-616, *Elsevier*, June 2009.

[8] N.B.Mabrouk, S.Beauche, E.Kuznetsova, N.Georgantas and V.Issarny, "Qos-aware service composition in dynamic service oriented environments," *In Proceedings of Middleware, Middleware* '09, pages 7:1-7:20, 2009.

[9] U.Lublin and D.Feitelson, "The workload on parallel supercomputers: modeling the characteristics of rigid jobs," *Journal of Parallel and Distributed Computing*, 63(11):1105–1122, 2003.

[10] J.Dean and S Ghemawat, " MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, 51(1):107–113, 2008.

[11] Yuhui Shi, "Particle Swarm Optimization," *IEEE Neural Network Society*, pp.8-13, 2004.

[12] Tzu-Chiang Chiang, Po-Yin Chang, and Yueh-Min Huang, "Multi-Processor Tasks with Resource and Timing Constraints Using Particle Swarm Optimization," *IJCSNS International Journal of Computer Science and Network Security*, vol.6 no.4, pp. 71-77, 2006.

[13] Shih-Tang Lo, Ruey-Maw Chen, Der-Fang Shiau and Chung-Lun Wu, "Using Particle Swarm Optimization to Solve Resource-constrained Scheduling Problems," *IEEE Conference on Soft Computing in Industrial Applications*, 25-27 June, 2008.

[14] Rui Mendes, James Kennedy and José Neves, "The Fully Informed Particle Swarm: Simpler, Maybe Better," *IEEE Transactions of Evolutionary Computation*, vol. 8, no. 3, pp. 204- 210, 2004.

[15] Parsopoulos K.E, and Vrahatis M.N, "Recent approaches to global optimization problems through particle swarm optimization," *Natural Computing* , vol.1, pp. 235 – 306, 2002.