



International Journal of Engineering, Business and Enterprise Applications (IJEBA)

www.iasir.net

A Policy Driven Architecture for Effective Service Allocation in Cloud Environment

¹Mansi Goyal, ²Richa Chhabra
¹Student, ²Faculty
ITM University, Gurgaon, Haryana, India

Abstract: A Cloud Environment provides the integration of multiple clients and server in a distributed environment. But in this environment the cloud servers are limited and there are number of cloud clients. To perform the effective cloud service allocation, some rule oriented model is required that can perform the analysis on the cloud server features as well as client characteristic analysis. In this work, a policy based architecture is shown that covers the cloud service allocation along with location identification and migration assistance.

Keywords: Load Balancing, Cloud Scheduling, Request Scheduling

I. INTRODUCTION

A cloud computing is distribution system that provides the integrated virtual environment. The presented work is defined as the integrated system that combines the cloud service, network system and the application software in an integrated environment. The cloud system is the shared system in which the resources and the services are shared in the effective service environment [1][2].

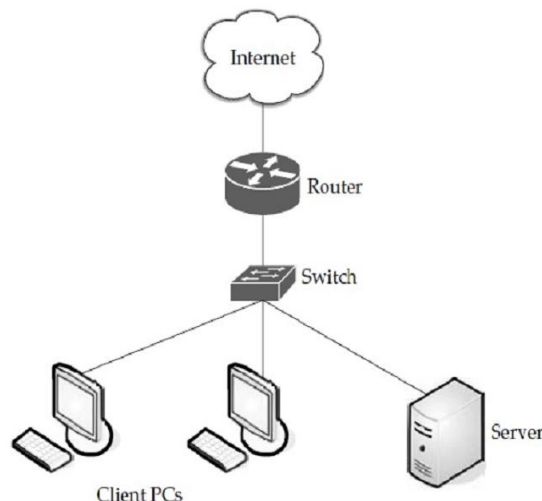


Figure 1: Basic Client Service Interconnection Model

Here figure 1 is showing the basic integration model. As we can see, the client and server both are connected to the web based system in a generic integrated environment. In this environment, the service provider avail the services to the clients under the characterization analysis so that the distribution of the services to the client will be effective. In this environment, different kinds of cloud servers are available under the characteristics specification such as public availability, private restricted access and the limited secure access. As the user enter to the system, it basically connected to the intermediate layer where it get the information about all the available services along with cloud server specifications. But as the number of clients over the system increases, the challenges associated with the cloud system also increases. These challenges include the scheduling of the client requests, client service allocation, load balancing, security etc. To perform he effective cloud service allocation, there is the requirement of some effective mechanism that can perform the effective identification of the cloud and client characteristics. To handle these client requests, there is the requirement of some reliable and efficient service allocation is required. The cloud computing is one of the most effective architecture available over the web and mobile system to provide the sharing of services and the resources. It also improves the cloud system efficiency and the throughput. The distributed cloud system is capable to handle the multiple requests in an integrated environment along with independent resource specifications or the shared resources. These resources include the memory specification, storage area definition etc. The effectiveness of the cloud system can be

achieved to gain the effective turnaround time, wait time etc. The cloud system is able to handle the multiple requests in the cloud environment as well as provide the integrated distributed cloud environment so that the processes present in the job queue will be processed effectively. This cloud system is having the different service allocation architecture to provide the effective distribution of the services to the clients. These allocation processes are also defined under the scheduling mechanism. In this section the exploration to the scheduling system is defined.

A. Scheduling in Distributed Cloud

When a distributed cloud system is generated, one of the challenges is to decide the order of client request processing. There are number of scheduling approaches that are either handled by the centralized cloud controller or some independent cloud system controller. These requests or jobs will be handled under the cloud system specification. The centralized controller will manage the allocation of these services in an effective way so that the effective generation of the cloud system will be performed. The objective of the scheduling processes is to manage the relation between the cloud system and the clients so that the resource allocation and the process execution will be done effectively. The distributed cloud system is also defined under the cost estimation so that the resource management in such system will be effective and adaptive. It will also explore the fault tolerance, scalability, reliability to the system. In this paper, the cloud environment exploration is been defined under the cloud service allocation process. In this section, the cloud system is defined with basic model specification. This section also defined the scheduling approach and service allocation system in cloud environment. In section II, the work defined by the earlier researchers is explored. In section III, the service allocation model is explained. In section IV, the conclusion derived from the work is discussed and presented.

II. RESEARCH METHODOLOGY

Lot of work is already done in the area of resource allocation and the process scheduling. Some of the earlier work done in same area is presented here. In year 2006, Vikki Tang has defined a work to reduce the instruction scheduling under the dynamic compilers. Author defined a scheduling approach under the feedback analysis so that effective allocation will be done. The presented framework is defined to benefit the instruction scheduling under multi threaded server applications [1]. In year 2013, Lichen Weng has defined a work on multithreaded Distributed Cloud system to perform the dynamic modelling. The paper describes the design under three steps. At first, author converts a scheduling policy to dynamic to evaluate the runtime of pattern mapping. The another step is to define the regression model to achieve the scheduling policy to identify the changing behavior of the threading system. The main objective of author was to define a scalable heuristic approach for estimating the growth of the system count[2]. Hsiang-Yun Cheng is defined as an analytical model to achieve the task scheduling under the analytical modelling. Author estimated the potential aspects under the memory and bandwidth analysis to restrict the number of task. Author implemented the scheduling under the real hardware [3]. In year 2013, Vishakha gupta has performed the performance analysis for the functionality analysis under asymmetric platforms. Author has performed the analysis under the heterogeneity under the utility and applicability analysis. Author has defined the work under the workload analysis and defined it under different processes and different configuration for the resource analysis [4].

Morris A. Jette defined the characteristics analysis under the scheduling process for multi programmed environments. Author defined a time and space slicing mechanism for the parallel programming and defined the concurrent job execution under single Distributed Cloud environment. Author has defined a performance analysis system under the utilization and responsiveness under different computing platforms[5]. Another work for the heterogeneous scheduling policies for real time multi Distributed Cloud system is considered for the multimedia mapping for design space. Author has defined a suitable scheduling policy so that system energy can be minimized. The presented framework includes the analysis on energy reduction approaches for dynamic power management [6]. Another work on power management for multi-core architecture for the process scheduling is defined for the process estimation under platform evaluation. Author defined the effectiveness and scalability of the system. Author highlighted the scalability limitations for the thread scheduling algorithm for small scale multi Distributed Cloud system. Author has defined the scheduling overhead without loss of accuracy [7]. In Year 2005, Rony Ghattas presented some approach to improve the functionality of the micro Distributed Cloud system under the energy and power constraints. This system was defined under low bit system and to enhance the system performance. The main advantage of the system is to reduce the cost and complexity of this new micro Distributed Cloud system along with the reduction of power consumption [8].

In Year 2003, Andrei Terechko defined the scheduling under the high level language with some variable definition with global values. Author defined the long range and large impact schedule for the compiler optimization for local values under the scheduling units. The paper has defined three main algorithms for assigning the values to different cluster under the multi pass scheduling approach under the variable definition. Author also defined the performance measures for optimizing the algorithm [9]. In Year 2004, Andrew Riffel also defined a multi pass partitioning problem with recursive denominator split along with heuristic algorithm so

that the robustness over the approach will be achieved. This paper redefines the MPP as a scheduling problem and uses scheduling algorithms that allow incremental resource estimation and pass computation in effective time[10]. Another work on improvement over the energy efficiency was presented by Hiroshi Sasaki. The proposed method groups several instructions as a single issue unit and reduces the required number of ports and the size of the structure for dispatch, wakeup, select, and issue. The present paper describes the micro architecture mechanisms and shows evaluation results for energy savings and performance [11]. Flavius Gruian presented an addresses scheduling approach for reduced energy of hard real-time tasks with fixed priorities assigned in a rate monotonic or deadline monotonic manner. The approach Author describes can be exclusively implemented in the RTOS. It targets energy consumption reduction by using both on-line and off-line decisions, taken both at task level and at task-set level [12].

Martin Schoeberl performed the investigation on the overhead analysis on object oriented operations. Author also presented the work so that the overhead over the system will be reduced as well as the dispatch and field access will be done effectively. Author presented this work for a real time embedded system. The main objective presented by the author to reduce the hardware cost and to optimize the application output [13]. In Year 2000, Jared Stark presented work on instruction scheduling for pipelined processing. Author defined the work to improve the pipelined scheduling. Author has defined the technique to eliminate the ability to improve the execution of dependent instruction under the consecutive cycles. The presented approach by the author has defined the frequency check with the sacrifice of IPC [14].

III. SCHEDULING APPROACHES

In the distributed cloud environment, the scheduling approach is having the importance to resolve the load balancing problem. To perform the distributed load balancing, the parallel queue handling on the the intermediate layer is performed. While performing work on distributed queuing, cooperative and non-cooperative process analysis will be performed. The analysis of the queue elements will be performed under the different parameters. These parameters include the response time analysis, wait time analysis, resource availability, resource requirement etc. In the second scheduling mechanism, the processes input by the users is maintained in a single global queue and scheduling is performed on this global queue initially and later on the process allocation to different clouds will be performed. In the centralized cloud computing environment, the different considerations are taken while performing the scheduling. These considerations are shown in figure 1.

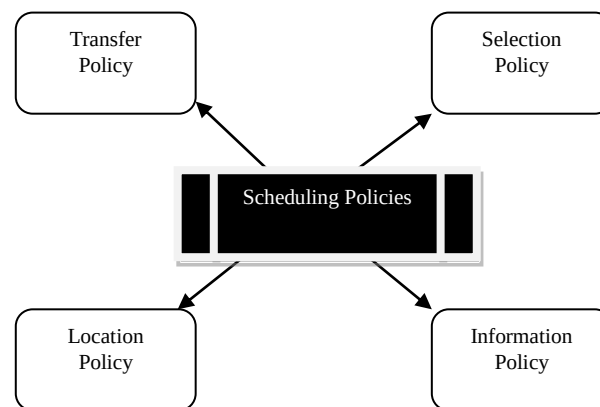


Figure 1: Policies Under Scheduling Consideration

A. *Transfer Policy*

One of the most effective considerations of the scheduling scheme is the transfer policy. According to this policy, the job transfer can be performed from one cloud server to other. This approach is also called cloud migration policy. According to this policy, the cloud server analysis is performed under the client request. If the particular cloud server is not able to handle the request in such case, the cloud migration will be performed. To perform the migration, the transfer policy is used. According to this policy, the analysis of the cloud system is performed under the current acceptability of the client request on the server. If the server availability parameters are adapted to the request will be migrated.

B. *Selection Policy*

The selection policy is about the selection of the cloud server based on the user request parameters. If the user request parameters are adapted to the cloud server availability. This analysis will be performed on all the available cloud servers. The cloud server that is feasible to the user request parameters will be consider effective to the selection policy. The selection criteria is based on the scheduling algorithm such as the wait time analysis

based will perform the cloud server allocation to the critical request first. The adaptive selection parameters are shown in figure 2. These parameters can be used individually or in group to take the effective scheduling decision so that the

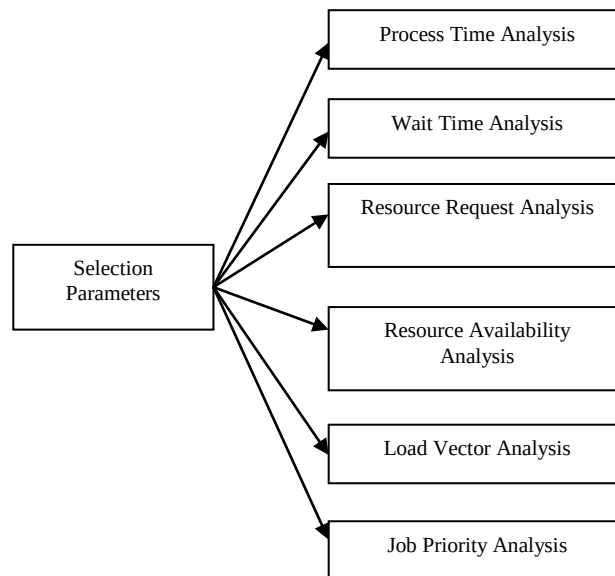


Figure 2: Selection Policy Parameters

C. Location Policy

Location policy is about the decision of the process execution server. Some specialized user request requires the availability of some specialized servers such as database server. The location policy also depends on the specialized attributes such as the physical location of the physical location of the server, the language domain of the processing etc. Sometimes, the utilization of the server cannot be performed even if the server is available because of the location boundation specified by the client. Generally the location policy is either user specific or adaptive.

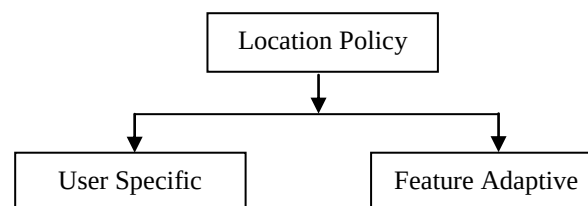


Figure 3: Location Policy Parameters

The user specific parameters includes the requirement specification in terms of server country specification, language specification etc. The feature adaptive specification is identified by the model itself based on the requirement and the availability analysis. The load balancing machismo is also the parameter for the location adaptive assignment.

D. Information Policy

The information policy is about the extraction of the information related to the cloud server as well as the relative environmental vector. This policy deals on two main ends i.e. client side and the server side adaptation as shown in figure 4. The client side adaption will capture the request related information such as dead line criticality evaluation, process sharing policy analysis etc.

IV. CONCLUSION

In this paper, an exploration to the cloud service scheduling mechanism is explored. The complete scheduling approach is defined under some policy specifications. These policies define the rules for the process generation and its execution on the cloud server. The paper has explored the policy parameters as well as its inclusion as the effective stage in the process execution mode.

REFERENCES

- [1] Vikki Tang, "A Framework for Reducing Instruction Scheduling Overhead in Dynamic Compilers".
- [2] Lichen Weng, "Scheduling Optimization in Multicore Multithreaded MicroDistributed Clouds through Dynamic Modeling", CF'13, May 14–16, 2013, Ischia, Italy. ACM 978-1-4503-2053-5
- [3] Hsiang-Yun Cheng, "An Analytical Model to Exploit Memory Task Scheduling", INTERACT-14, March 13, 2010, Pittsburgh, PA, USA ACM 978-1-60558-921-3/10/03
- [4] Vishakha Gupta, "Kinship: Efficient Resource Management for Performance and Functionally Asymmetric Platforms", CF'13, May 14–16, 2013, Ischia, Italy. ACM 978-1-4503-2053-5
- [5] Morris A. Jette, "Performance Characteristics of Gang Scheduling in Multiprogrammed Environments", 1997 ACM 0-89791-985-8/97/0011
- [6] Minyoung Kim, "Design Space Exploration of Real-time Multi-media MPSoCs with Heterogeneous Scheduling Policies", CODES+ISSS'06, October 22–25, 2006, Seoul, Korea. ACM 1-59593-370-0/06/0010
- [7] Jonathan A. Winter, "Scalable Thread Scheduling and Global Power Management for Heterogeneous Many-Core Architectures", PACT'10, September 11–15, 2010, Vienna, Austria. ACM 978-1-4503-0178-7/10/09
- [8] Rony Ghattas, "Energy Management for Commodity Short-Bit-Width Microcontrollers", CASES'05, September 24–27, 2005, San Francisco, California, USA. ACM 1-59593-149-X/05/0009
- [9] Andrei Terechko, "Cluster Assignment of Global Values for Clustered VLIW Distributed Clouds", CASES'03, Oct. 30 – Nov. 1, 2003, San Jose, California, USA. ACM 1-58113-676-5/03/0010
- [10] Andrew Riffel, "Mio: Fast Multipass Partitioning via Priority-Based Instruction Scheduling".
- [11] Hiroshi Sasaki, "Energy-Efficient Dynamic Instruction Scheduling Logic through Instruction Grouping", ISLPED'06, October 4–6, 2006, Tegernsee, Germany. ACM 1-59593-462-6/06/0010
- [12] Flavius Gruian, "Hard Real-Time Scheduling for Low-Energy Using Stochastic Data and DVS Distributed Clouds", ISLPED'01, August 6-7, 2001, Huntington Beach, California, USA. ACM 1-58113-371-5/01/0008
- [13] Martin Schoeberl, "Architecture for Object-Oriented Programming Languages", JTRES '07 September 26-28, 2007 Vienna, Austria ACM 978-59593-813-8/07/09
- [14] Jared Stark, "On Pipelining Dynamic Instruction Scheduling Logic", 0-7695-0924-X/2000© 2000 IEEE