International Journal of Advanced Technology in Engineering and Science
Vol. No. 09, Issue No. 03, March 2021
www.ijates.com
ijates
ISSN 2348 - 7550

Survey on Efficient Load Balancing Algorithms and Comparative Study

Manjula H Nebagiri [1], Dr. Latha P H[2]

[1]*Information Science, Atria Institute of Technology, (India)*

[2] *Information Science, Sambhram Institute of Technology, (India)*

ABSTRACT

The present era has witnessed tremendous growth of the internet and various applications that are running over it. Cloud computing is the internet-based technology, emphasizing its utility and follows pay-as-you-go model, hence became so popular with high demanding features. Load balancing is one of the interesting and prominent research topics in cloud computing, which has gained a large attention recently. In cloud computing, Load balancing is one of the main challenges which are required to distribute the workload equally across all the nodes. Load balancing uses services offered by many computer network service provider corporations. Load balancing can be different types like network load, storage capacity, memory capacity and CPU load. Load balancing helps to achieve a high user satisfaction and resource utilization ratio by confirming an efficient and fair allocation of every computing resource. Proper load balancing support in implementing failover, enabling scalability, over provisioning, and decreases costs associated with document management systems and maximizes the availability of resources. This paper describes a survey of different dynamic load balancing algorithms in the cloud environment with their comparisons on the bases of different load balancing metric.

keywords*:* *Load balance,SDN, SLA, triangular network*

I.INTRODUCTION

"Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service level agreements (SLA) established through negotiation between the service provider and consumers."[1] In the current and recent works with the maximum improvement of the network, the data/information of the internet is more and more complex issue, and the traffic load become more and more higher. The online data centres may give all varieties of services for people, so they play vital roles in the networks nowadays. Some works proposes very important architectures for the internet data centre. approach to solve the problem, which would use more than 2 or even more to broadcast the internet data flows. The way to achieve the load balancing in the data center is main issues. To minimize the network latency and maximize the throughput is a issue the data centre networks. In this paper discussed about the hardware and software approaches. The hardware(simulation) model proposes more network triangular topologies but basically this is

cost effective. So the software approach apply load balancing methods in the triangular interconnected topologies for higher bandwidth utilization in the current networks. This approach needs to upgrade the current approaches of network traffic flows scheduling. The load balancing approaches can be divided into two classes , static and dynamic load balancing.

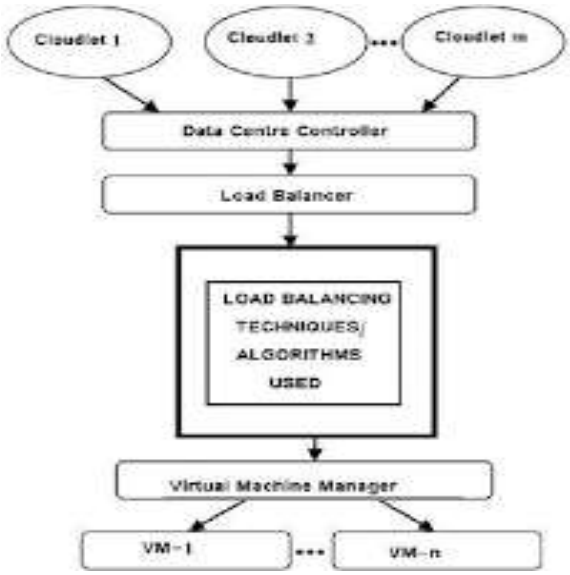

Fig. 1. Load Balancing Process

Static Algorithms: These algorithms do not depend upon the current state of the system and have prior knowledge regarding system resources and details of all tasks in an application. These kinds of algorithms face a major drawback in case of sudden failure of system resource and tasks.

Dynamic Algorithms: These algorithms take decisions concerning load balancing based upon the current state of the system and don't need any prior knowledge about the system. This approach is an improvement over the static approach. The algorithms in this category are considered complex, but have better fault tolerance and overall performance.

This definition describes Cloud Computing using [2], [3]:

Three service models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

Four deployment models: Private Clouds,Community Clouds, Public Clouds, and Hybrid Clouds.

Five characteristics: on-demand self-service, broadnetwork access, resource pooling, rapid elasticity, and measured service.

Any cloud computing system consists of three major components which are:

Client: Client is end users, which interact with the clouds to manage information related to the cloud. Clients can be Mobile client, Thin client and Thick client.

Data centre: Datacentre is the collection of servers hosting different applications and it may exist at a large distance from the clients.

Distributed Servers: Distributed servers are the part of a cloud which actively checks services of their hosts and available throughout the internet hosting different applications.

In data centre the network flows may change all the time , so the link costs of the network have to be changed and machine should learn and cost should change automatically to learn.. The static load balancing approaches have no ability to get the information on real-time loads and to decrease time of dispatching data flows among nodes. So dynamic load balancing can overcome the issue, they will bring more work for monitoring network statistics and scheduling data flows.

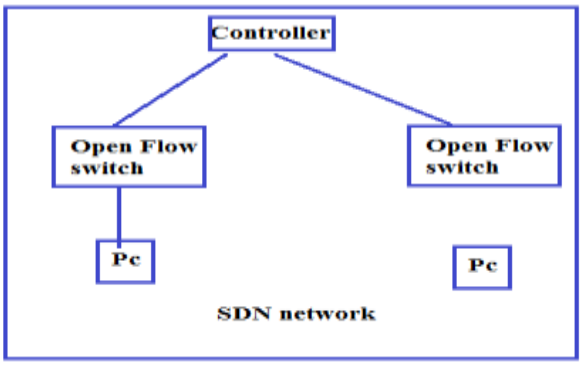

Fig. 2. SDN Network

Achieving the load balancing at the datacenter level is the key issue now a days. To reduce the network latency and to increase the thoughput is a big problem over the datacenter topology. There were two models[2 proposed , they are hardware and software have been proposed. The model of hardware and and some models of the network topologies, but generally the cost of hardware over the datacenter is over cost than the regular hardware. The software model will apply load balancing models for peak bandwidth usage in current topologies. This infrastrucute requied to modelize the existing models with network data traffic scheduling with respect to flows. The load balancing approaches can be classified into 2 classes[6]dynamic and static load balancing. Over the datacenter, the network topologies models of flows might change all the time, so the link costs of the topology changes with accordingly. The static model of load balancing have not much ability to fetch the information over th real-time loads and to decrease time of delivering data broadcast between nodes. But dynamic load balancing models can overcome the issue, they will fetch more working models to monitor the topology statistics with proper scheduling the model of data flows.

This paper mainly focuses on dynamic load balancing algorithms. The rest of the paper is organized as follows: Section 2 contains an introduction of load balancing. Section 3 describes the existing load balancing algorithms. Comparison of all the algorithms areanalysed in section 4 using different load balancing metrics and final conclusion of the work is given in section 5.

II.CLOUD LOAD BALANCING

In cloud environment, Cloud load balancing is a type of load balancing that is performed in cloud computing. Cloud load balancing is the process of distributing workloads across multiple computing resources. Cloud load balancing reduces costs associated with document management systems and maximizes availability of resources. Where in general Load balancing is a technique that distributes the excess dynamic local workload evenly across all the nodes. Load balancing is used for achieving a better service provisioning, resource utilization and improving the overall performance of the system. For the proper load distribution, a load balancer is used which received tasks from different location and then distributed to the data centre. A load balancer is a device that acts as a reverse proxy and distributes network or application load across a number of servers [4][5].

III.EXISTING LOAD BALANCING ALGORITHMS IN CLOUDCOMPUTING

In cloud computing environment, there are various Load Balancing Algorithms which are closely analysed and compared on the bases of some predefined metrics, including throughput, response time, Overhead, performance, fault tolerance, migration time, resource utilization, and scalability. Some of the commonly known Load Balancing Algorithms are;

Active Clustering: Active Clustering algorithm [11][12] works on the principle of grouping the similar nodes and work together on the available groups. A set of processes is iteratively executed by each node on the network. Initially any node can become an initiator and selects another node from its neighbours to be the matchmaker node satisfying the criteria of being a different type than the former one [17]. The matchmaker node then forms a connection between neighbours of it which are similar to the initiator. The matchmaker node then removes the connection between itself and the initiator [18].

Honey Bee Foraging: Honey Bee Foraging algorithm [13] is derived from the behaviour of honey bees for finding and reaping food. In order to check for fluctuation in demand of services, servers are grouped under virtual servers, having its own virtual service queues. Each server processing a request from its queue calculates a profit or reward on basis of CPU utilization, which is corresponds to the quality that the bees show in their waggle dance and advertise on the advert board. Each of the servers takes the role of either a forager or a scout. A server serving a request, calculates its profit and compare it with the colony profit, if profit was high, then the server stays at the current virtual server and if it was low, then the server returns to the forager or scout behaviour, thus balancing the load with the server [17].

Join Idle Queue: Join Idle Queue load balancing algorithm [14] is applied for dynamically scalable web services. This technique involves a dispatcher to whom processors informs at the time of their idleness, without interfering with job arrivals. Thus removing the load balancing work from the critical path of request processing, system load is reduced; no communication overhead at job arrivals and no increment in actual response time [17].

Load Balance Min-Min (LBMM): LBMM scheduling algorithm [15] and new optimized Load Balancing Max-MinMax (LB3M) [16] had main objective to minimize execution time of each task, also avoid unnecessary replication of task on the node thereby minimizing overall completion time. Opportunistic Load balancing algorithm when combined with LBMM (OLB + LBMM) [15] keeps every node in working state to achieve load balance. Similar to LBMM, LB3M [16] also calculate average completion time for each task for all nodes. Then mark the task with maximum average completion time. After that it dispatches the task of marked node to the unassigned node with minimum completion task, thus balancing the workload evenly among all nodes [17].

Ant Colony Optimization (ACO): ACO algorithm [17] is mainly proposed for load balancing of nodes and aims efficient distribution of workload among the nodes [3]. The ant will start to move towards the source of the food from the head node when the request is initialized. Ant records their data for future decision making and it keeps records for every node and it makes a visit to the record. Every ant is built with their own individual result set and further built for giving the complete solution. It makes to update continuously with a single result set rather than own result set is updating. This ant works in searching of new sources food with the use of existing food sources to shift the food back to the nest. This mainly aims that efficient distribution of the load among the nodes. It does not encounter the dead end of the movement to the node for building an optimum solution set. In ACO two types of pheromones are used Foraging Pheromone (FP) used to explore overloaded node by forward movement of ants while Trailing Pheromone (TP) used to discover its path back to the under loaded node. In order to limit the number of ants in the network, they would commit suicide once it finds the target node [18].

ACCLB (Load Balancing mechanism based on Ant Colony and Complex network theory): In [19] author proposed a load balancing mechanism based on ant colony and complex network theory in an open cloud computing federation. It uses small world and scale-free characteristics of a complex network to achieve better load balancing. This technique overcomes heterogeneity, is adaptive to dynamic environments, is excellent in fault tolerance and has good scalability hence helps in improving the performance of the system.

Exponential Smooth Forecast based on Weighted Least Connection (ESWLC): ESWLC algorithm [20] is improved form of Weighted Least-Connection (WLC) along with its features; it also takes into account time series and trials. WLC counts the connections of each server and reports the appropriate server based on the multiplication of a server weight and its count of connections. ESWLC algorithm concludes assigning a certain task to a node only after getting to know about the node capabilities. ESWLC builds the decision based on the experience of the node's CPU power, memory, number of connections and the amount of disk space currently being used. ESWLC then predicts which node is to be selected based on exponential smoothing [18].

Honey Bee Behaviour inspired Load Balancing (HBB-LB): According to [21][22], HBB-LB is a technique, which helps to achieve even load balancing across virtual machine to maximize throughput. It considers the priority task waiting in queue for execution in virtual machines. After that, the work load on VM calculated decides whether the system is overloaded, under load or balanced and based on these VMs are grouped [17].

According to the load on VM the task is scheduled on VMs, which is removed earlier. To find the correct low loaded VM for the current task, tasks which are removed earlier from over loaded VM are helpful. Forager bee is used as a Scout bee in the next steps [18].

Equally Spread Current Execution (ESCE): According to [24], ESCE is a dynamic load balancing algorithm, which handles the process with priority. It determines the priority by checking the size of the process. This algorithm distributes the load randomly by first checking the size of the process and then transferring the load to a virtual machine, which is lightly loaded. The load balancer spreads the load on different nodes, and hence, it is known as spread spectrum technique.

Throttled Load Balancer (TLB): Throttled load balancer is a dynamic load balancing algorithm [24] in which the client first requests the load balancer to find a suitable virtual machine to perform the required operation. In Cloud computing, there may be multiple instances of virtual machine. These virtual machines can be grouped based on the type of requests they can handle. Whenever a client sends a request, the load balancer will first look for that group, which can handle this request and allocate the process to the lightly loaded instance of that group.

Genetic Algorithm (GA): According to [25], Genetic Algorithm has been used as a soft computing approach, which uses the mechanism of natural selection strategy. A ***simple Genetic Algorithm is composed of three operations:*** genetic operation, selection, and replacement operation. The advantage of this technique is that it can handle a vast search space applicable to complex objective function and can avoid being trapped in locally optimal solution. A generation is a collection of artificial creatures (strings). In every new generation, a set of strings is created using information from the previous ones. Occasionally, a new part is effort for good measure. According to Genetic Algorithms are randomized, but they are not simple random walks. They adept exploit historical information to speculate on new search points with expected improvement. The effectiveness of the GA depends in appropriate mix of exploration and exploitation.

Particle Swarm Optimization (PSO) Algorithm: Particle Swarm Optimization (PSO) as a meta-heuristic's method is a self-adaptive global search-based optimization technique introduced by Kennedy and Eberhart [27]. The PSO algorithm is alike to other population-based algorithms like Genetic algorithms (GA) but, there is no direct recombination of individuals of the population. The PSO algorithm focuses on minimizing the total cost of computation of an application workflow. The objective is to minimize the total cost of execution of application workflows on Cloud computing environments. Results show that PSO based task-resource mapping can achieve at least three times cost savings as compared to Best Resource Selection (BRS) based mapping for

application workflow. In addition, PSO balances the load on compute resources by distributing tasks to available resources.

IV. LOAD BALANCING METRICS AND COMPARISON OF THE ENTIRE ALGORITHMS

After studying the dynamic load balancing algorithms, we have compared all the algorithms on the bases of some predefined metrics. These metrics are as follows [17]: Throughput: Throughput is used to calculate the number of jobs whose execution has been completed. It should be high to improve the performance of the system. Overhead: It determines the amount of overhead involved while implementing a load balancing algorithm. Overhead should be minimized so that a load balancing technique can work efficiently. Fault Tolerance: Fault tolerance system is a system in which the processing does not get affected because of the failure of any particular processing device in the system. The load balancing should be fault tolerant.

Migration time: Migration is the time of movement of job of the master system to the slave system and vice versa in case of results. Migration time is the overhead, which cannot be removed but should be minimized.

Response Time: It is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized. Resource Utilization: It is used to check the utilization of resources. It should be optimized for an efficient load balancing.

Scalability: It is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved. Performance: It is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays. Table1 shows the comparisons of the dynamic load balancing algorithm, which were discussed in section 3.

Table 1

Comparisons of existing Load Balancing Algorithms

→) Algorithms (↓)	Through put	Overhead	Fault Tolerance	Migration Time	Resp onse Time	Resource Utilization	Scala bility	Performa nce
Active Clustering	N	Y	N	Y	N	Y	N	N
Honeybee Foraging	N	N	N	N	N	Y	N	N
Join Idle queue	N	Y	N	N	Y	N	N	Y
LBMM	N	N	N	N	N	Y	N	Y
ACO	Y	N	N	N	N	Y	Y	Y
ACCLB	N	N	Y	N	N	Y	Y	Y
ESWLC	Y	N	Y	N	N	Y	N	Y
HBB-LB	Y	N	N	N	N	N	Y	Y
ESCE	Y	N	N	N	Y	Y	N	Y
TLB	N	N	Y	Y	Y	Y	Y	Y
Genetic Algorithm	N	N	N	N	N	Y	N	Y
PSO	Y	N	N	N	Y	Y	N	Y

Framework for working of Dynamic Load Balancing Load balancing is a technique of distributing the total load to the individual nodes of the collective system to the facilitate networks and resources to improve the response time of the job with maximum throughput in the system [6]. The important things which said about load balancing are estimation of load, load comparison, different system stability, system performance, interaction between the nodes, nature of work to be transferred, selecting of nodes and many other ones to consider while developing such algorithm [7]. In the area of cloud computing, the main objective of load balancing techniques is to improve performance of computing in the cloud, backup plan in case of system failure, maintain stability and scalability for accommodating an increase in large scale computing, reduces associated costs and response time for working in the cloud and also maximizes the availability of resources.

V.CONCLUSION

In the cloud computing environment, load balancing is one of the main issues, which is required to distribute dynamic local workload to all the nodes in the cloud to improve the performance and maximize resource utilization. This paper explains cloud computing, load balancing, types of load balancing algorithms, components of dynamic load balancing algorithms and load balancing metrics. This paper primarily focuses on dynamic load balancing algorithm in cloud environment. For this, various existing dynamic load balancing algorithms are surveyed. By comparing the algorithms on different metrics tried to find the scope for improving fault tolerance, throughput, performance, resource utilization and minimizing response time, migration time, overhead in the load balancing algorithm. Future work is related to designing a new dynamic load balancing algorithm with fault tolerance for better resource utilization, minimum response time and fast throughput of the cloud computing environment.

REFERENCES

[1] 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, 25:599_616, 2009.*

[2] P. Mell and T. Grance, *The NIST Definition of Cloud Computing,* National Institute of Standards and technology, Information Technology Laboratory, *Technical Report Version 15*, 2009.

[3] Rimal, Bhaskar Prasad, Eunmi Choi, and Ian Lumb. *"A taxonomy and survey of cloud computing systems."* INC, IMS and IDC, 2009. *NCM'09*. Fifth International Joint Conference on. IEEE, 2009.

[4] L. M. Vaquero, L. Rodero-Merino, J. Caceres and M. Lindner, *"A break in the clouds: towards a cloud definition,"* SIGCOMM ACM Computer Communication Review,*vol. 39,* pp.

[5] December Rahman, Mazedur, Samira Iqbal, and Jerry Gao. *"Load Balancer as a Service in Cloud Computing."* In Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on, *pp. 204-211*. IEEE, 2014.

[6] Tong R, Zhu X. *A load balancing strategy based on the combination of static and dynamic*[C]//Database Technology and Applications (DBTA), 2010 2nd International Workshop on. IEEE, 2010: *1-4*

[6] R. Shimonski. *"Windows 2000 & Windows Server 2003 Clustering and Load Balancing"*, Emeryville. McGrawHill Professional Publishing, CA, USA (2003), *p 2, 2003*.

[7] Ali M. Alakeel, *"A Guide to Dynamic Load Balancing in Distributed Computer Systems"*, IJCSNS International Journal of Computer Science and Network Security, *VOL.10 No.6, June 2010*.

[8] M.Armbrust, A.Fox, R. Griffit,et al., *"A view of cloud computing"*, Communications of the ACM, *vol. 53, no.4, pp. 50–58, 2010.*

[9] M. Amar, K. Anurag, K. Rakesh, K. Rupesh, Y. Prashant (2011). *SLA Driven Load Balancing For Web Applications in Cloud Computing Environment,* Information and Knowledge Management, 1(1), *pp. 5-13, 2011*.

[10] O. Abu- Rahmeh, P. Johnson and A. Taleb-Bendiab, *"A Dynamic Biased Random Sampling Scheme for Scalable and Reliable Grid Networks"*, INFOCOMP - Journal of Computer Science, *ISSN 1807-4545, 2008, VOL.7, N.4, December, 2008, pp. 01-10.*

[11] F. Saffre, R. Tateson, J. Halloy, M. Shackleton and J.L. Deneubourg, *"Aggregation Dynamics in Overlay Networks and Their Implications for Self-Organized Distributed Applications."* The Computer Journal, March 31st, 2008.

[12] Dhurandher, Sanjay K., Mohammad S. Obaidat, Isaac Woungang, Pragya Agarwal, Abhishek Gupta, and Prateek Gupta. *"A cluster-based load balancing algorithm in cloud computing."* In Communications (ICC), 2014 IEEE International Conference on, pp. 2921-2925. IEEE, 2014.

[13] Randles, M., D. Lamb and A. Taleb-Bendiab, *"A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing,"* in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Perth, Australia, April 2010.

[14] Yi Lua, QiaominXiea, Gabriel Kliotb, Alan Gellerb, James R. Larusb, Albert Greenbergc,*" Join-Idle-Queue: A Novel Load Balancing Algorithm for Dynamically Scalable Web Services"* Volume 68 Issue 11, November, 2011, pp:1056-1071, Elsevier Science Publishers, 2011.

[15] S. Wang, K. Yan, W. Liao, and S. Wang, *"Towards a Load Balancing in a Three-level Cloud Computing Network",* Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, China, September 2010, *pages 108-113.*

[16] Che-Lun Hung, Hsiao-hsi Wang and Yu-Chen Hu *"Efficient Load Balancing Algorithm for Cloud Computing Network"*, International Conference on Information Science and Technology (IST 2012), April 28-30, pp; 251-253.

[17] Sushil Kumar, Deepak Singh Rana and Sushil Chandra Dimri, *"Fault Tolerance and Load Balancing algorithm in Cloud Computing: A survey",* International Journal of Advanced Research in Computer and Communication Engineering, July 2015.

[18] Dharmesh Kashyap, Jaydeep Viradiya, *"A Survey of Various Load Balancing Algorithms In Cloud Computing",* International Journal of Scientific & Technology Research, *Vol. 3, Issue 11, November 2014.*

[19] Zhang, Z. and X. Zhang, *"A load balancing mechanism based on Ant Colony and Complex Network Theory in Open Cloud Computing federation."* In proc. 2nd International Conference on. Industrial Mechatronics and Automation (ICIMA), IEEE, *Vol. 2, pp:240-243, May 2010.*

[20] Ren, X., R. Lin and H. Zou, *"A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast"* in proc. International Conference on. Cloud Computing and Intelligent Systems (CCIS), IEEE, *pp: 220-224, September 2011*.

[21] Dhinesh B. L.D, P. V. Krishna, *"Honey bee behaviour inspired load balancing of tasks in cloud computing environments"*, in proc. Applied Soft Computing, *volume 13, Issue 5, May 2013.*

[22] Ganesh, Amal, M. Sandhya, and Sharmila Shankar. *"A study on fault tolerance methods in Cloud Computing."* In Advance Computing Conference (IACC), 2014 IEEE International, *pp. 844-849. IEEE, 2014.*

[23] Galloway, Jeffrey M., Karl L. Smith, and Susan S. Vrbsky. *"Power aware load balancing for cloud computing."* Proceedings of the World Congress on Engineering and Computer Science. *Vol. 1. 2011.*

[24] Domanal, Shridhar G., and G. Ram Mohana Reddy. *"Load Balancing in Cloud Computingusing Modified Throttled Algorithm."* Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on. IEEE, 2013.

[25] Ye, Zhen, Xiaofang Zhou, and AthmanBouguettaya. *"Genetic algorithm based QoS-aware service compositions in cloud computing."* Database systems for advanced applications. Springer Berlin Heidelberg,

[26] Dam, Scintami, et al. *"Genetic algorithm and gravitational emulationbased hybrid load balancing strategy in cloud computing."* Computer, Communication, Control and Information Technology (C3IT), 2015 Third International Conference on. IEEE, 2015.

[27] Pandey, Suraj, Linlin Wu, SiddeswaraMayura Guru, and Rajkumar Buyya. *"A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments."* In Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, pp. 400-407. IEEE, 2010.