International Journal of Smart Home
Vol. 10, No. 8 (2016), pp. 241-250
http://dx.doi.org/10.14257/ijsh.2016.10.8.23

Review of the Research on the Optimization of the Energy Consumption of the Cloud Platform

Chun-mao Jiang[1, 2]，Yi-bing Li[1] and Li Zhi-cong[2]

[1]Information and Communication Engineering School,
Harbin Engineering University
Harbin, HeiLongJiang, China 150001
[2]Computer Science and Information Engineering School,
Harbn Normal University
Harbin, HeiLongJiang, China 150025
hsdrose@126.com

Abstract

Cloud platform is a basic platform to support large data, large-scale and high-frequency access computing. The high-energy consumption of the cloud platform and the schedule of multi-constrained combination in cloud applications is challenging issues faced by the cloud computing. We conducted a systematic review in this paper for the energy consumption of a cloud platform, and pointed out the existing problems, and put forward ideas to solve the above problems by constructing the four-level system. Firstly, by conducting virtualization management to the physical resources in the cloud to form the layer of virtual resource. Then, named as presentation layer, to build a formal model of cloud application multi_attribute with effective description, measurement, calculation. Based on this proposed scheduling application layer enables cloud applications to meet the multiple objectives with multi-attribute, heuristic, feedback, iterative scheduling and cloud resources, energy optimization. It laid the foundation for the formation of cloud platform architecture, key technologies and algorithms, which satisfy the multiple objectives: the maximum possible scheduling of the multi-attribute combination constraint application, the optimization for energy consumption and resources in cloud computing.

Keywords*: Cloud platform; Cloud energy consumption; Multi-attribute; Combinational scheduling*

1. Introduction

Grid scheduling technology provides a good basis for resource scheduling of cloud computing[1], however, the more challenging questions in the cloud computing are big data computing, high-frequency access, schedule of the ultra-large-scale computing tasks, making it different from the traditional task scheduling based on distributed and grid computing environment, Specific to these resource scheduling tasks, it is necessary to fully take into account the multidimensional attribute constraint application tasks, data center distribution, data distribution, but also consider the efficient use of physical resources and energy optimization. Existing research sided separates the application-specific properties, while considering task scheduling more neglected the efficient use of the cloud resources and high-energy consumption problem.

ISSN: 1975-4094 IJSH

2. The Research Status

1.1 Research Status of Cloud Task Scheduling

Category 1 simply considers the cost optimization scheduling problem of individual applications, but ignores the time, energy consumption, reliability, transmission cost and other issues.

Chen Hongwei *et al* [2] proposed a scheduling algorithm considering the time and cost, which solved the precedence relation between tasks and achieved the scheduling cost optimization. Under the conditions of time constraints, but the scheduling algorithm makes it difficult to efficient use of resources by Random task, resulting in a waste of energy, Because a lot of time fragments to be produced in order to minimize the cost. According to the market mechanism, starting with the economic interests of supply and demand

It proposed a market equilibrium scheduling algorithm, achieved a maximize revenue to the resource providers, a utility maximization to user and a market equilibrium of a resource between supply and demand under the conditions of the optimal allocation of resources[3].

Ying chun yuan *et al* [4-8]

① Proposed scheduling algorithm DBL based on layered reverse, Computed nodes depth from back to front and assigned the tasks to the same group in which the tasks complete with the synchronization feature,thus reducing the total cost of service and improving the performance, There had been some issues not taken into account such as the balance of the deadline and the total cost in the workflow applications,the failure service failures.

② Proposed a time-optimization algorithm in a cost-constrained workflow and a slicing algorithm from both the forward and reverse, to exchange the minimum completion time for a maximum reduction of the total cost. But the algorithm did not consider the overhead that the transmission required for.

③ Proposed a cost-optimization method based on the serial reduction algorithm, redefined the time in the group of the serial reduction, achieved the cost optimization within groups, thus improving the average performance and saving the cost. However, the algorithm did not consider the time-consuming that the transmission required.

Can-can Liu[9]①proposed the timing characteristics in workflow and a scheduling algorithm using the reverse hierarchical which is based on the task timing-constraints of the same deadline, to ensure the workflow target with the timing-constraints by setting the consistent time point,solved the limited use problems of the reverse layered and optimized the implementing costs.

②Designed algorithm BFTCS, considered the timing characteristics of the workflow, the time flexibility of the task as a priority rule be added to the iterative algorithm,slowed down the growth of the workflow and optimized the implementing costs.

Based on the Amazon EC2, charged based on the amount of the individual consumption, such as CPU, Memory and disk space, The use of cloud resources to meet the user to maximize the quality of service as the starting point. [10, 11]

Chunlin Li[12] designed an algorithm for cloud computing software and services, maximized profits while optimizing the cloud service configuration of resources and meeting the service delivery, but the algorithm ignores the time characteristics to consider.

Above algorithms achieved the cost optimization at the expense of execution time and energy by taking long-distance transmission and other measures, but some problems did not been considered such as Synchronous and parallel of the tasks, time and cost of inter-task communication

Simply consider the time optimization problem.

Optimized time through rational distribution, fuzzy clustering, satisfaction, linear programming, but did not consider the cost of energy consumption, scheduling overhead, reliability, cost, dynamics of tasks, stability, security and so on.

Kliazovich, D. *et al*. [13] focused on the layout method of data between different data centers in the cloud computing environments, reduced time overhead the transmission caused, but it is only discussed from storage resources and network bandwidth, did not save energy though the data layout is more reasonable.

YI Kan *et al* [14] studied and proposed a task distributing algorithm based on load balancing, reduced the response time, but ignored the scheduling overhead, exception handling issues and so on.

Lijun Zhang, proposed an earliest completion time scheduling algorithm which used the Multi-service quality value as a priority constraint. For the multi-QoS needs of each user, making unified measurement and then use the metrics of these needs as task precedence constraints, the algorithm reduced the time span in the process of scheduling.

Xi Li[15], proposed the concept of deadline satisfaction, the time scheduling algorithm- DSESAW was designed based on satisfaction. According to the priority of subtasks, determined the candidate resources of subtasks during the execution of a workflow. The problem of partitioning the global deadline in the workflow been described as a problem of a constrained nonlinear programming, and solved the problem by the existing methods. Improved the performance of adaptability and time guaranteed aspect, but did not consider the reliability, charges and other service indicators.

JianNing Lin① proposed a scheduling algorithm based on genetic algorithm, using the simple coding to get the chromatography relations of the tasks and sorting according to the depth,only considered the execution time of the tasks run on the resources ,and the transmission delay between resources, the dynamic changes of the resource loading and stability are not taken into account.

② proposed a heuristic algorithm based decision path, get the scheduling scheme using the Normal Scheduling Algorithm at first, generating tasks of scheduling decision and the path,then, scheduling the decisive tasks ahead of schedule as soon as possible and running the decisive tasks during the idle periods of resources by using the heuristic algorithm. Achieve the goal to shorten the convergence time of the task, at the same time the same time the method of determining the deadlock been given.

Experimental results show that the new algorithm is superior to other heuristic algorithms. Algorithm saves time overhead, but slightly worse than the genetic algorithm in performance.

Chen Jing, proposed a scheduling algorithm senior who combined with time constraints, network bandwidth, and the prediction mechanism, the shorter length of the scheduling the higher user satisfaction, achieved a double QOS scheduling, but only to verify the validity of the algorithm does not achieve the specific optimization.

Zhi gang chen[11, 16]In order to improve scheduling performance presented a scheduling algorithm multidimensional performance clustering of the resource, constructed the hyper graph of the service's resources at first, clustering around the resources multidimensional, matching the tasks and clustered resources and scheduling, the task completion time shortened, improved performance, but did not consider the aspects of safety and dynamic of the task.

Xiao Li Chen[17] proposed a heterogeneous task scheduling algorithm of fuzzy clustering, divided the network resources reasonably and reduced the time of matching resources.

With the increasing scale of the task, the high superiority is displayed, but the algorithm does not consider the dynamic changes such as the reliability of links, communication ability.

Achieving the time scheduling algorithm basically through the rational layout and the initial linear programming. However, there were lacked of consideration of transmission overhead, dynamic characteristics of the network and multiple performance indicators the in the task scheduling processes.

Garcia-Arenas, M., *et al* [18] proposed a fill in the blank copies of the data allocation algorithm. According to the frequency of the agent by which the resource is stored, making a reasonable distribution of the copies of the data. Achieved the optimal efficiency with the minimal overhead.

However, the algorithm does not consider computing and throughout capacities of the proxy nodes, and the cost required for transmission, and how to allocate the new data unit. Xin jun Wang, proposed a heuristic algorithm for data distribution, reducing communication costs and improving efficiency, but does not consider the difference of computing power between the nodes.

Akon, M., *et al*. [19, 20] proposed a scheduling algorithm based on P2P phased, the optimal network scheduling scheme, but lack of consideration above service properties. Heidi Liu, proposed a hierarchical genetic algorithm to realize the task scheduling, optimized at the diversity of population, the convergence of the algorithm and the convergence rate, improved the efficiency of genetic algorithm in accuracy and speed, but lack of consideration of the dynamic random characteristics of the tasks.

Proposed the task resource allocation map according to the task dependency, according to the optimal selection from the task map, put forward a scheduling algorithm, the algorithm has advantages in case of large data transfers and a large difference of the task resources.

Yuan Lin[21], proposed a collaborative filtering algorithm and the concepts of satisfaction, to allocate resources by recommending the resources to a user through the use history, improved the efficiency.

However, the algorithm does not consider computing and throughout capacities of the proxy nodes, and cost required for transmission, and without considering how to allocate the new data units, the difference of computing power between nodes, the randomness of the dynamic tasks, the execution of the small amount tasks and so on.

Scheduling optimization for resource utilization, the success rate of the service requests, accuracy, *etc*.

The scheduling research for m * n environment, proposed the Nash equilibrium algorithm in the environment, raising the number of the completed tasks in unit time and the average load in networks and systems, but received only an approximate value and does not prove the convergence of the algorithm[14].

Wu Zhiang *et al*[22] proposed a hierarchical model of QOS, designed a simulation system used to resource management, measuring the QOS parameters on the virtual tissue layer, improved algorithm MIN-MIN, improved the utilization of resources and the success rate of service requests effectively. But the algorithm did not consider other network parameters except the time, the utilization.

Dasgupta, P., *et al*. [23] proposed changed regions activated scheduling algorithm designed for the dynamic scheduling problem in the distributed environments, solved the dynamic scheduling sequence, improved the reuse efficiency and accuracy in designing resources.

The schedule about resource utilization did not consider the convergence of the algorithm, time, efficiency, cost, energy consumption and other service expenses.

Scheduling algorithm considered the work load and load balancing.

Deboosere, L., *et al* [24] proposed random multi-start climbing algorithm for no center scheduling framework, choice the origin by the exponential growth of the repetitions that were generated in selecting a neighbor randomly, adjusted the workload flexibly, optimized the global node choosing effectively, optimizing the execution performance, Even so, it did not take into account the overhead of scheduling and the conditions with very light network load.

Schedule related to faulting - tolerance, reliability.

Literature from Cong Feng Jiang, proposed a safe and fault-tolerant scheduling algorithm, adjusting the number of task backups adaptively based security level of the network systems, rescheduling the failed tasks, improving the success rate of the tasks, within good fault-tolerance and scalability.

Hai Jin and other scholars presented a stochastic Petri net model from the fault-tolerance, through assignment and selection of the tasks, analyzed the fault-tolerant computing performance index, reflecting the impact of the fault on the mission, but the specific measures did not be given.

Deboosere, L., *et al*. [25-28] proposed a scheduling algorithm, using the redundancy and the intelligent agent technology to solve the problem of load balancing and fault-tolerance effectively, but does not consider the cost (such as performance and cost) issues arising from fault-tolerance and dynamic equilibrium.

Peng Xiao, Proposed collaborative scheduling model HPCF based the deadline constraint, providing a service with reliable quality assurance for task scheduling, but it is only a model and does not validate its feasibility.

Some scheduling algorithms rarely consider the scheduling and time overhead, which are related to fault-tolerance and reliability, also did not consider the cost overhead associated with fault-tolerance and dynamic equilibrium.

1.2 Existing Problems

(1) Lack of the research for the scheduling algorithm based on the combination of the multi_attribute constrained and the resource optimization

Existing studies did not have thoroughly researches in scheduling based the multi-constraint combinations and resource optimization. Lead research has great limitations, cannot be entirely suitable for the actual situation in the cloud applications;

While it only considered its schedulability in the applications, ignored the effective utilization and the energy consumption of the cloud resources, so that reduced the practical and feasible.

(2) Without the sufficiently considering the effect on the scheduling strategy that is from the specific type of the tasks.

In the past, scheduling without the sufficiently considering the effect on the scheduling strategy that is from the specific type of the tasks[29-32].

Because they are the most frequent types in cloud computing, including the big data computing, the high-frequency access and the ultra-large-scale data computation. Its span of the area, time and energy consumption is so wide in all these calculations. It is critical how to layout the initial data based the application type and how to set the storage location and the node position.However, it has not been solved in the existing studies.

3. The Research on the Hierarchy of Task Scheduling in Cloud Considering the Multiple Constraints-Energy Consumption

The Cloud platform maximizes mission satisfaction of the resource requests, and supervises the software and the hardware resources for the cloud tasks, by Gathering a wide range of physical resources.

As it can be seen from Figure 1, the macro indicators of the cloud self-optimization include: the number of the tasks accepted, Energy consumption of cloud platform, available resources of Cloud platforms.

The Energy consumption presents a growing trend with the increased number of the tasks accepted, while the amount of the resources available shows a downward trend[33,34].

So we must schedule effectively to the physical resources, including reservation, distribution, and recycling, to determine the maximum number of the tasks accepted, and the service can achieve in a limited amount of energy and resources.

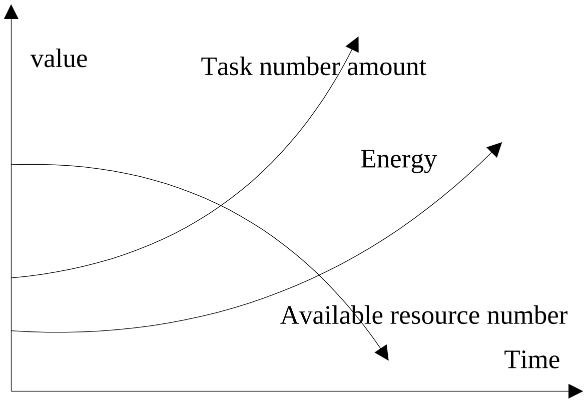

Figure 1. Optimize Goals of Scheduling Services in Cloud Platform

The Executable attribute bounds are more complex that were submitted to the cloud platform from the individual applications, with the different representations, the different calculation methods, different measurement systems, *etc*., And it is more difficult scheduling for resource with multi-combination attributes.

Figure 2, the main task of the resource scheduling is the efficient use and recycling of resources and to optimize energy consumption effectively, which is achieved based the multi-attribute constraint and features of the individual applications, a reasonable allocation of physical resources in the cloud, and guaranteed the Executable Constraint.

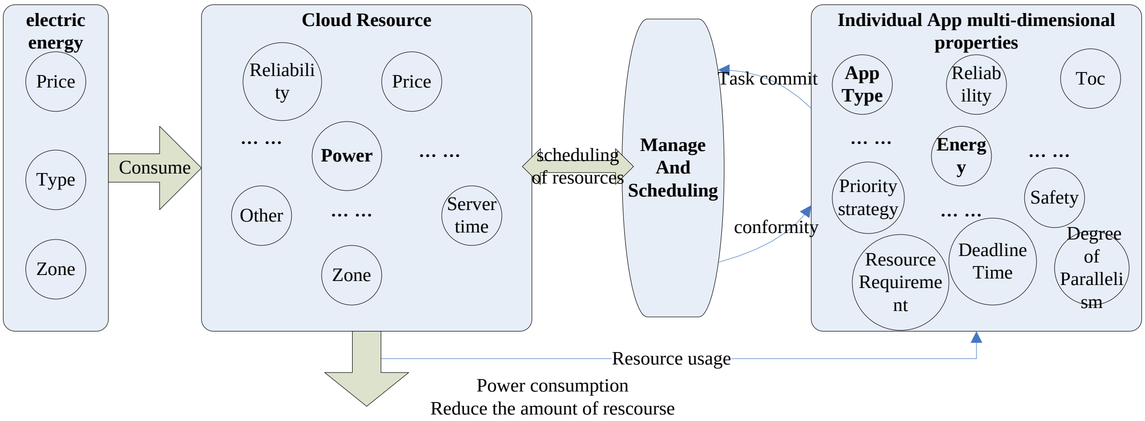

Figure 2. Logic Diagram of Combination Scheduling and Resource Optimization

Since the uncertainty of the multi-attribute constrained combinations from individual applications in the cloud, the amount of available resources gradually reduced with the increase in the number of applications accepted, which will lead to a decrease of the rate of successful scheduling in applications with a high-dimensional constrained composite.

In which chances where the specific constraints are used tolerably, and the Service are able to downgrade. User got the result of the schedule from the user applications based on the current resources and state in the cloud, then giving the results back to cloud platforms after making some adjustments to the constrained combinations, then the cloud platform scheduled recursively.

In order to enhance the success rate of application scheduling and increase the service throughput in the cloud platform. Process is shown in Figure 3.

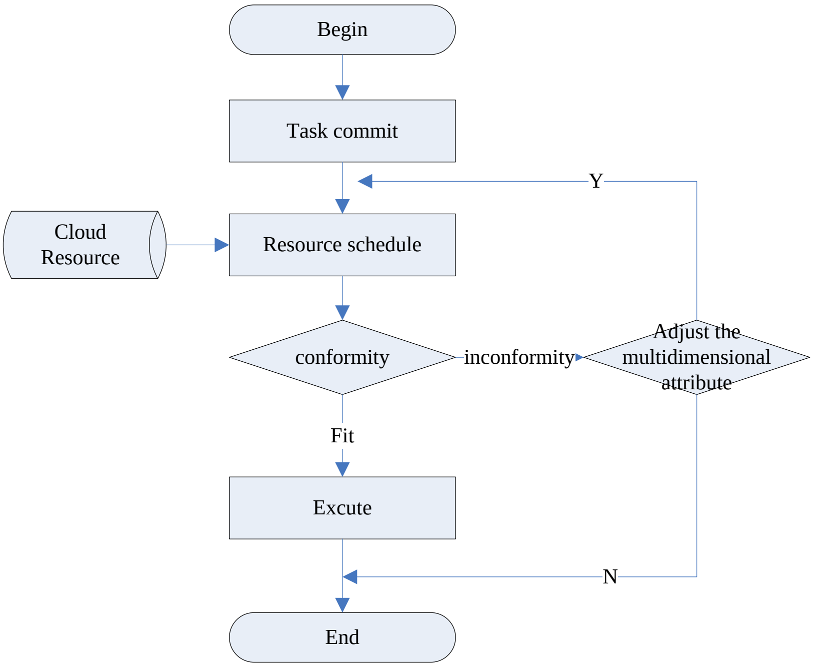

Figure 3. Cloud Platform Inspired, Feedback, Iterative Scheduling Process

3.1 Optimizing the Scheduling Layer

Combined with prior knowledge, statistical analysis and optimization theory, got the specific scheduling policy for different types of tasks and its multi- attributes.

For high-energy applications according to the initial layout of the type of research data, data center set up computed node distribution.

According to the user's multi-attribute constraints, researching how to achieve the maximum scheduling of the tasks using heuristic, feedback, iterative method, while optimizing the energy and the resource effectively.

The Scheduling layer schedule the resources mainly based the multi-attribute constraints from the application requests.

The Scheduling targets-optimizing the energy and the resource after ensuring the greatest possible of success scheduling; and transferring the data generated by scheduling and the service results to the virtual resources layer and the presentation layer.

In order to achieve maximum scheduling of the tasks, giving priority to high-priority attribute constraint within the tolerance threshold that is given by multi- attributes of the application, and estimating the total cost of a service according to the result in each constraint of the scheduling-fit line.

The research has come to a scheme based the multi-attribute threshold adjusting and the heuristic, when the scheduling-fit line with a Reasonable deviation. Beyond the tolerance threshold. Further to maximize services receiving in the mode of feedback and iterative scheduling.

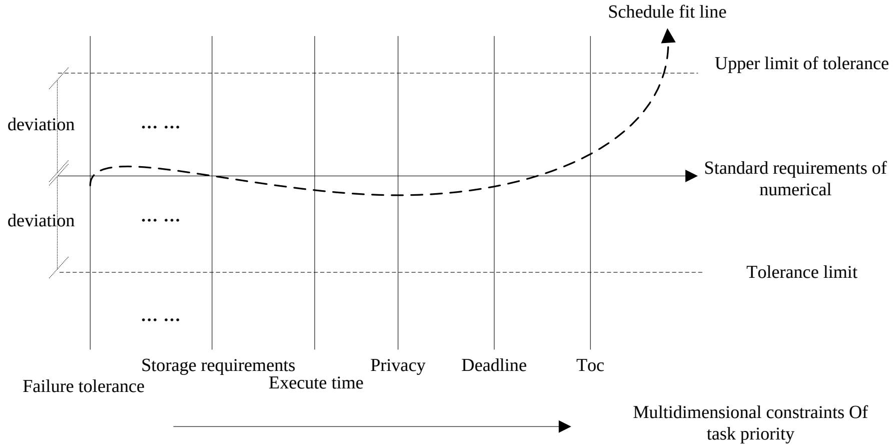

Figure 4. Logic Diagram of Multidimensional Constraint Combination Scheduling

4. Summary

In order to solve the problems of energy consumption, multi-attributes combination scheduling and resource optimization, this paper presents Four-Tier Architecture, the presentation layer, resource scheduling layer, resource virtualization layer and the physical resources hyper esthetically, the presentation layer and its formal description model analyzed the issues, including the Value Composition of multi-attribute, measurement methods, computing theory. The scheduling theory proposed in the paper based on virtual resources layer and the presentation layer, unlike the traditional research, which best meet the multi- attribute constraint of the tasks using the heuristics, feedback, iterative mechanism, and optimizing energy consumption and resources so that improved the overall service acceptance and service quality. And laid the foundation for the formation of cloud platform architecture, key technologies and algorithms, which satisfy the multiple objectives: the maximum possible scheduling of the multi-attribute combination constraint application, the optimization for energy consumption and resources in cloud computing.

Acknowledgement

This work is supported by the Harbin technology bureau youth talented project (2014RFQXJ073) and China postdoctoral fund projects(2014M561330).

References

[1] Foster, I., *et al*. Cloud computing and grid computing 360-degree compared. in Grid Computing Environments Workshop, 2008. GCE'08. 2008. Ieee.

[2] Chen Hongwei and Wang Ruchuan, A Grid DAG Scheduling Algorithm for Cost-Time Optimization [J]. Acta Electronica Sinica, 2005. 33(8): p. 1375-1380.

[3] Li Zhijie., A Sequential Game-Based Resource Allocation Strategy in Grid Environment [J]. Journal of Software, 2006. 17(11): p. 2373-2383.

[4] Yuan Yingchun, Li Xiaoping, and Wang Xi, Cost Optimization Heuristics for Grid Workflows Scheduling Based on Serial Reduction [J]. Journal of Computer Research and Development, 2008. 45(2): p. 246-253.

[5] Yuan Yingchun, Time Optimization Heuristics for Scheduling Budget-Constrained Grid Workflows [J]. Journal of Computer Research and Development, 2009(2): p. 194-201.

[6] Yuan Yingchun, Grid Workflows Schedule Based on Priority Rules [J]. Acta Electronica Sinica, 2009(007): p. 1457-1464.

[7] Yuan Yingchun, Bottom Level Based Heuristic for Workflow Scheduling in Grids [J]. Chinese Journal of Computers, 2008. 31(2): p. 282-290.

[8] an Yiming, Zeng Guosun, and Wang Wei, Policy of Energy Optimal Management for Cloud Computing Platform with Stochastic Tasks [J]. Journal of Software, 2012. 23(2): p. 266-278.

[9] Liu Chanchan, Temporal Consistency Based Heuristics for Cost Optimization in Workflow Scheduling [J]. Journal of Computer Research and Development, 2012. 49(6): p. 1323-1331.

[10] Randles, M., D. Lamb, and A. Taleb-Bendiab. A comparative study into distributed load balancing algorithms for cloud computing. in Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on. 2010. IEEE.

[11] Vouk, M.A., Cloud computing–issues, research and implementations. Journal of Computing and Information Technology, 2004. 16(4): p. 235-246.

[12] Li, C. and L.Y. Li, Optimal resource provisioning for cloud computing environment. The Journal of Supercomputing, 2012: p. 1-34.

[13] Kliazovich, D., GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. in Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE. 2010. IEEE.

[14] Yi Kan and Wang Ruchuan, Decentralized integration of task scheduling with replica placement strategy [J]. Journal on Communications, 2010(009): p. 94-101.

[15] Li Xi, Grid Workflow Scheduling Algorithm Based on Deadline Satisfaction [J]. Journal of Computer Research and Development, 2011. 48(5): p. 877-884.

[16] Chen Zhigang and Yang Bo, Task Scheduling Based on Multidimensional Performance Clustering of Grid Service Resources [J].

[17] Du Xiaoli, A Grid DAG Scheduling Algorithm Based on Fuzzy Clustering [J]. Journal of Software, 2006. 17(11): p. 2277-2288.

[18] Garcia-Arenas, M., Assessing Speed-ups In Commodity Cloud Storage Services For Distributed Evolutionary Algorithms. 2011 Ieee Congress on Evolutionary Computation. 2011, New York: Ieee. 304-311.

[19] Akon, M, A cross-layered peer-to-peer architecture for wireless mobile networks. 2006 IEEE International Conference on Multimedia and Expo - ICME 2006, Vols 1-5, Proceedings, 2006: p. 813-816.

[20] Arrifano, A, JOINT SOURCE-CHANNEL DECODING OF MOTION-INFORMATION USING MAXIMUM-A-POSTERIORI, in 2011 18th Ieee International Conference on Image Processing. 2011.

[21] Lin Yuan, Luo Siwei, and Yang Liner, Recommendation-Based Grid Resource Matching Algorithm [J]. Journal of Computer Research and Development, 2009(011): p. 1814-1820.

[22] WU Zhiang, Luo Junzhou, and Song Aibo, QoS-Based Grid Resource Management [J]. Journal of Software, 2006. 17(11): p. 2264-2276.

[23] Dasgupta, P., Improving peer-to-peer resource discovery using mobile agent based referrals. Agents and Peer-To-Peer Computing, 2004. 2872: p. 186-197.

[24] Deboosere, L., Efficient resource management for virtual desktop cloud computing. The Journal of Supercomputing, 2012: p. 1-27.

[25] Lefèvre, L. and A.-C. Orgerie, Designing and evaluating an energy efficient cloud. The Journal of Supercomputing, 2010. 51(3): p. 352-373.

[26] Baliga, J., Green cloud computing: Balancing energy in processing, storage, and transport. Proceedings of the IEEE, 2011. 99(1): p. 149-167.

[27] Beloglazov, A., A taxonomy and survey of energy-efficient data centers and cloud computing systems. Advances in Computers, 2011. 82(2): p. 47-111.

[28] Berl, A., Energy-efficient cloud computing. The Computer Journal, 2010. 53(7): p. 1045-1051.

[29] Nikolopoulos, V., Web-based decision-support system methodology for smart provision of adaptive digital energy services over cloud technologies. Software, IET, 2011. 5(5): p. 454-465.

[30] Wang, X., Y. Wang, and H. Zhu, Energy-Efficient Multi-Job Scheduling Model for Cloud Computing and Its Genetic Algorithm. Mathematical Problems in Engineering, 2012. 2012.

[31] Mezmaz, M., A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems. Journal of Parallel and Distributed Computing, 2011. 71(11): p. 1497-1508.

[32] Aggarwal, B., Stratus: energy-efficient mobile communication using cloud support. ACM SIGCOMM Computer Communication Review, 2011. 41(4): p. 477-478.

[33] XIAO Lu-xin RESEARCH ON THE OPTIMIZATION OF ENROLLMENT DATA RESOURCES 2015.Vol.2, No.2, 2015, pp.9-12

[34] MingMing Guo, ShengLong Yang, Huan Yan, LiFeng Kan and Bo Yang. MOBILE VIDEO ALARM SYSTEM BASED ON CLOUD COMPUTING. REVIEW OF COMPUTER ENGINEERING STUDIES. 2014 Vol.1, No.2, pp.5-10