Applications And Challenges Of Efficient Energy Optimization In Cloud Computing

Ghanshyam Parmar

N. V. Patel College of Pure and Applied Sciences, Vallabh Vidyanagar, Gujarat, India.

Abstract — Cloud computing is a new field of study that benefits all types of users, from average consumers to major corporations. The present research interests of researchers have been on several cloud computing study fields, such as load balancing, cost management, workflow scheduling, etc. Some conventional approaches have been devised to address these issues; however, they are not very successful. Since the past ten years, a significant area of concern has been the application of energy optimisation in cloud computing. This chapter emphasises the applicability of nature-inspired algorithms to a variety of cloud computing challenges in a detailed (yet condensed) survey report. The chapter seeks to give a thorough understanding of nature-inspired optimisation methods and their application to the challenges of cloud technology. There is also discussion of some potential future possibilities for cloud computing research.

***Keywords* –** Cloud computing, Efficient Energy optimization, Load balancing, Work flow scheduling, Cost optimization

I. INTRODUCTION

Optimisation is the process of selecting the best solution from those that are available to solve a problem, subject to the satisfaction of certain criteria. Depending on the type of problem, every optimisation problem is either a minimization problem or a maximisation problem. Optimisation is virtually always present in our everyday lives. The basic goal of any optimisation technique used to solve engineering issues is to strike a balance between exploration and exploitation. Constraints (barriers), design variables, and the objective function (the core of optimisation) are the essential components of every optimisation [1]. Depending on the number of objective functions, the kind of objective functions, the kind of restrictions, the kind of variables, and the nature of optimisation, there are several kinds of optimisation algorithms. These factors will determine whether the optimisation is single- or multi-objective / Pareto-optimal multi-objective solutions that can be local or global, limited, or unconstrained, smooth, or non-smooth, stochastic, or deterministic, continuous, or discrete, etc [2]. There are many different types of optimisations, however it can be challenging to select the precise optimisation technique that would work best for a given complex engineering problem. The complexity of real-world optimisation problems is increasing as science and technology advance in tandem. The classic optimisation methods that were previously created are unable to explain the precise and actual solution to nonlinear and non-differential problems in huge search spaces.

These algorithms' primary drawbacks include early convergence, the usage of intricate stochastic functions, and the employment of higher order derivatives while resolving the equations. Some well-known optimisation algorithms have previously demonstrated their efficacy in tackling various real-world situations during the past few decades [3]. The genetic algorithm (GA), created in 1992 by John Holland and Goldberg (1989), is the most well-known evolutionary algorithm. This technique works better at local optima and has a very low possibility of trapping at local minima places as compared to gradient search-based algorithms. Then, in 1995, Kennedy and Eberhart created a method called PSO that was influenced by stochastic swarms. It is currently one of the most well-liked stochastic and heuristic-based search techniques. In the past ten years, several variations of this category of physical, chemical, and nature-based algorithms have been introduced [4]. Even though they have been used to several complicated situations, they still have significant flaws, such as convergence criteria, when utilised in isolation. This is because regulating factors like population size, environmental factors, iteration count, etc. are used extensively. As a result, in order to test the algorithms' capacity to solve difficult issues, these variations have been created by incorporating certain alterations to the parameters or by conducting any type of hybridization of the algorithms. Hybridization is not the ideal technique to handle these complicated problems because any significant change in the parameter selection may alter the functional aspects of the entire algorithm.

Researchers from diverse research communities have recently become interested in the practicality of nature-inspired optimisation methods [5]. The capacity to tackle NP-hard problems is the primary factor in the success rate of nature-inspired and swarm-based algorithms. Many real-life problems have been solved using heat and trail approaches because some of the earlier proposed optimisation strategies fail to tackle real-life difficulties. This serves as the foundation for researchers to concentrate on creating some efficient optimisation algorithms that are competitive and effective at solving complicated issues.

Since the last two decades [6], internet usage has grown significantly in popularity across all user levels, from business to the automation industry at a lower cost. A new era known as cloud computing has emerged as a result of the expansion of data storage, access, and processing with a wide range of applications. Basically, cloud computing is concerned with hosting and providing services across a network. Due to their flexibility in allowing for the usage of resources in accordance with customer preferences, cloud services are more popular and in demand. The majority of the main IT sectors today are successfully utilising cloud services to meet their needs for organisational change.

Despite the hoopla surrounding cloud computing research, several of its earlier problems are still open for discussion, and new ones are constantly emerging. It is necessary to address issues like automated service delivery, virtual machine migration, server consolidation, efficient energy management, and traffic control, among others [7]. However, several academics are using optimisation methods to handle specific issues including task scheduling, load balancing job/flow scheduling, resource optimisation, resource allocation, etc. For effectively overcoming those difficulties, most of them have adopted optimisation methodologies inspired by nature. In this chapter, a thorough analysis of the application of nature-inspired optimisation algorithms to various cloud computing challenges has been provided, along with observations, analytical debates, and recommendations for resource-use improvement.

II. WHAT IS CLOUD COMPUTING

Now-a-day's Cloud computing is a new technology that is used to host and provide cloud services globally over the internet [9]. It offers a solution for providing on-demand services to satisfy service provider and end user requirements based on service level agreements (SLAs). Both commercial and academic applications use it. Users of cloud computing can access powerful computer resources in a virtualized setting. It offers pay-as-you-use, dynamic, scalable, and adaptable services to the end users. Cloud service providers [8] as Google, Amazon, Microsoft, and others set up their data centres to host cloud computing applications all around the world. IaaS (infrastructure as a service), PaaS (platform as a service), and SaaS (software as a service) are the names of these services [12]. Service level Agreements (SLAs) are used to deliver these cloud services to the end user for a variety of capabilities, including compute, network, and storage services.

Numerous researchers have provided their own definitions of cloud computing. Cloud computing is described as "a distributed and parallel computing environment that consists of a sizable collection of virtualized and interconnected computing resources that are dynamically presented and provisioned as one or more unified computing resources based on SLA" by Buyya et al. [8]. The National Institute of Standards and Technology (NIST) defined cloud computing as "it is a pay-per-use model for enabling convenient, on-demand computing resource access to a common pool of computing resources that can be dynamically released and provisioned with service provider effort or negligible management effort" [12].

III. CLOUD COMPUTING ARCHITECTURE

Based on the capabilities and service level given by the service provider and the level of abstraction used, cloud computing services are broadly categorised. In a cloud context, layered architecture can also serve as a representation of this abstraction level. Four layers, including the data centre layer, platform layer, infrastructure layer, and application layer, can be used to describe the cloud computing layered architecture [10].

Datacenter Layer: In the cloud environment's data centre, this layer oversees controlling the physical resources including servers, switches, routers, power supplies, cooling systems, etc. To provide services to the end user, all resources are available and maintained in numerous data centres. There are many physical servers in the data centre, and they are all connected by high-speed equipment like switches and routers.

Infrastructure Layer: Through several virtualization technologies, including Xen, KVM, and VMware, physical resources are divided into a collection of virtual resources at this virtualization layer. The cloud environment's central layer is where resources are dynamically provisioned utilising a variety of virtualization techniques [13].

Platform Layer: The operating system and application software make up this layer. This layer's goal is to directly deploy apps on virtual machines.

Application Layer: This layer comprises of several real cloud services that customers use. These programmes offer services to the user in accordance with their needs.

Cloud Business Model: To meet client needs, the cloud computing environment uses a service-oriented business model. It indicates that end users are provided with physical resources and services as needed. According to the abstraction level and service level, the services offered by cloud providers are divided into three categories in this section [14].

IV. Service Models

Infrastructure as a Service (IaaS): In terms of virtualized resources like virtual machines, IaaS offers on-demand physical resources like CPU, storage, memory, etc. Each virtual machine has its own computing power and can-do specific tasks based on the needs of the user. On-demand service provisioning of servers in the form of virtual machines is a feature of cloud architecture. IaaS suppliers include, for instance, EC2, GoGrid, etc [15].

Platform as a Service (PaaS): It offers a platform for the development and deployment of apps. This provides a high degree of abstraction to help programmers access the cloud environment more easily. Microsoft Azure, Google AppEngine, and other PaaS providers are examples [16].

Software as a Service (SaaS): It offers end consumers on-demand services or apps via the Internet. Software as a Service, or SaaS, is the name of this form of service delivery. It will do away with the expense of maintaining software and streamline end user development. PaaS suppliers include, for instance, RackSpace and salesfource.c.

V. Challenges in Cloud Computing Environment

Cloud computing has adopted many of the grid computing works features, including virtualization, resource allocation, scheduling, service discovery, scalability, and on-demand service. Currently, job scheduling [18], resource allocation [37], and energy efficient scheduling [17], among other things, in large scale contexts [19], provide the major obstacles for these concepts. The above notions and its current research directions are the main emphasis of this section.

Task Scheduling: One of the burgeoning study directions for the research community is the scheduling of tasks [18] with the available resources. Applications used in the cloud can be either standalone tasks known as Bag-of-tasks or connected standalone tasks known as workflows. There are two types of workflow scheduling: deterministic and non-deterministic. A directed acyclic graph (DAG) can predict the execution path in advance for deterministic algorithms, but the execution path for non-deterministic algorithms is decided dynamically. The workflow scheduling problem, which is NP-Complete, deals with a variety of issues, including dynamicity, heterogeneity, elasticity, service quality, analysis of vast volumes of data, etc. It is challenging to identify the overall ideal solution in this scheduling. The most crucial factors in determining the quality of services in a cloud computing environment are things like make span, cost, elasticity, and energy usage, among others. Applications in the commercial cloud are divided into those that are workflow-driven or single service oriented. Task to resource mapping is one of the most significant issues with cloud computing. This issue includes three components: choosing the optimum resource provisioning strategy for virtual machines, scheduling tasks on virtual machines, and choosing the virtual machines.

Resource Allocation: Users of the cloud computing environment have access to a limitless amount of computing resources, allowing them to dynamically increase or decrease resource utilisation in response to demand. In a resource allocation model, there are two main goals: the cloud provider wants to maximise income by achieving high resource utilisation, and cloud users want to keep costs to a minimum while still getting what they need. In a cloud context, the major goal of resource allocation [20] is to dynamically distribute VM resources among users based on demand while not going over their resource limits and cost budgets. Users of the cloud computing environment have access to a limitless amount of computing resources, allowing them to dynamically increase or decrease resource utilisation in response to demand. In a resource allocation model, there are two main goals: the cloud provider wants to maximise income by achieving high resource utilisation, and cloud users want to keep costs to a minimum while still getting what they need. In a cloud context, the major goal of resource allocation [20] is to dynamically distribute VM resources among users based on demand while not going over their resource limits and cost budgets.

Cloud Federation: When a cloud provider distributes resources among many users, the cost goes up and the capacity of the cloud provider is proportionately occupied. When a cloud provider distributes resources among many users, the cost goes up and the capacity of the cloud provider is proportionately occupied [21]. In order to maximise resource utilisation and reduce overall costs, it is important to assign each user to a resource that they will use.

VM Consolidation: By dynamically reallocating VMs utilising live migration technologies, VM consolidation [22] is utilised to increase the dynamic utilisation of physical resources and decrease energy consumption. In order to conserve energy and maximise server resource utilisation, VM consolidation seeks to fit as few physical servers as possible with active VMs. Virtualization technologies are utilised in cloud computing architectures to carry out the VM consolidation concept. By separating virtual machines from physical servers and allowing them to be placed on those servers where energy consumption can be increased, virtualization is the newest technology that lowers the energy consumption of datacenters. By expanding their client base, adding more virtual machines (VMs) while using fewer physical servers, and lowering data centre power usage, cloud providers attempt to lower the operational costs of their data centres. In order to maximise each physical server's utilisation, VMs should be spread among the fewest possible physical servers. Therefore, consolidation offers greater efficiencies with fewer computers that are turned on, which results in decreased energy usage in data centres.

Energy Efficient Scheduling: In a cloud computing context, scheduling is the issue of allocating jobs to a specific machine to perform their work in accordance with Service Level Agreements. Energy-conscious task scheduling involves giving tasks to a computer in a way that uses the least amount of energy possible [23]. The problem can be implemented in many scheduling algorithms in order to acquire the optimal outcome in order to achieve this minimization. With cloud computing, users of the service can access the necessary resources without physically being at their workplace since they pay for the resources, they use by signing SLAs with the cloud service provider.

Energy expenses are rising today as a result of the rapid rise in data centre resource management costs while yet maintaining high service level performance. A server's energy usage in datacenters varies dynamically according to the workloads it is now handling [24]. The number of active processors on a computing server determines how much power the server consumes, and the power consumption of each processor is mostly determined by its CPU utilisation. The immense carbon footprints created by the enormous amounts of electricity needed for the power supply and cooling of the numerous servers housed in those data centres cause carbon dioxide emissions in data centres to increase day by day.

By using energy-aware job scheduling, the major goal is to reduce energy usage in cloud data centres [17]. Because of the extensive requirements for dynamic computing applications, reducing the energy consumption of cloud data centres is a difficult problem. Therefore, a suitable green cloud computing environment is required, one that not only minimises energy usage in cloud data centres but also lowers operational costs without going against SLA policy.

VI. Emerging Research Areas of Cloud Computing

There are several older and more recent cloud computing study fields. To address the issues that arise in the real world when using clouds, researchers are constantly creating new, extremely effective solutions. The usage of nature-inspired optimisation algorithms can be realised in the study fields that this section addresses.

Energy Optimization: Energy optimisation is a difficult task in any cloud situation because the data centres must use the resources effectively [25]. Knowing the list of power providers and how to control usage are vital for managing power consumption. Any computing facility that requires energy to function uses it. Therefore, poor use may contribute to the creation of power waste. To address this issue, most researchers employ the dynamic reduction strategy in the number of clusters. Additionally, by migrating VMs, the goal may be attained via the virtualization technique. Numerous methods exist for achieving energy optimisation.

First, the dynamic voltage and frequency scaling technique can be used to automatically change the frequency and voltage in order to optimise the operation of various physical machines. The primary goals of such a technology are to maximise power and reduce heat produced by processors. Second, the data centres employ the power capping technique, which aids in system-level power budgeting [26]. In addition, this method, known as power shifting, is helpful for individually allocating power to each server in the specified cluster. In this method, systems with higher priorities are provided more power rather than those with lower priorities.

C-states systems [23], which are equipped with CPUs, are another significant recent advancement for power conservation. This method allows for a better use of the power by turning off the unused parts of idle systems. Although this method is ideal for devices that are used for extended periods of time, such personal laptops, it has limitations like deep sleep. The state of deep sleep is one in which the brain's processing units are stuck at one spot and take a long time to awaken.

Load Balancing: For better performance in any cloud situation, the overall execution time of the jobs must be decreased. The computing elements of clouds are referred to as virtual machines (VM). Tasks executed by virtual machines (VMs) for business purposes perform faster and in parallel. As a result, scheduling issues with respect to the available resources arise. For effective scheduling, the scheduler must use every available resource. Multiple tasks are allocated to one or more virtual machines (VMs) for concurrent completion. In this case, the scheduler needs to make sure that all the virtual machines are given the same tasks and loads, rather than some of them being under heavier load and others being idle.

Thus, the scheduler oversees distributing/assigning all loads to all VMs evenly. The major goal of load balancing algorithms is to make sure that resources are used as efficiently as possible while also improving the application that the client is offering in terms of response time. The techniques are also helpful for quick application execution amid runtime workload variations [27]. Both static and dynamic load balancing strategies can be used in homogeneous and heterogeneous situations. Static methods are appropriate when there is little fluctuation at the node.

These strategies are ineffective if the load fluctuates often, such as in cloud systems. Therefore, dynamic algorithms are preferred to static approaches that have an additional cost overhead. Dynamic algorithms are preferred, particularly in heterogeneous situations. Several nature-inspired based solutions are proposed for dynamic load balancing and continue to work well. Numerous authors have researched various aspects of load balancing algorithms in the literature.

Task Scheduling: Since its introduction, cloud computing has significantly decreased the financial and upkeep costs for deploying different applications. Clients are not concerned about any resource or corresponding income loss because of the excellent scalability of the system [28, 29]. Multiple systems connected via the internet can readily access the resources at any distant location by using virtual machines (VMs). One of the main goals of cloud computing is to maximise income generation on both the cloud provider's and the clients' sides. job scheduling has been a crucial component of cloud computing because poor job scheduling can result in significant revenue loss, performance deterioration, etc.

In order to effectively address issues like reaction time, resource usage, make span, and data communication costs, scheduling algorithms must be effective. A multidimensional cloud computing framework with a genetic algorithm for effective task scheduling has been created by Zhu and Liu [30]. To increase job scheduling efficiency, they considered parameters like task completion time and client financial needs. Their performance result is good in terms of efficiency when compared to conventional genetic algorithms. Wu et al. [31] compare the performance of two distinct optimisation strategies, such as GA and CRO, and energy-aware job scheduling. They argue that the CRO algorithm uses the make span more effectively than the GA method based on the results of their simulations. Their suggested approach beats previous methods like FCFS and can handle all types of work scheduling problems, regardless of size. In addition to these efforts, various nature-inspired solutions [32] have been developed to address the job scheduling issue in the cloud environment.

Work Flow Scheduling: In order to manage the interdependent tasks successfully and utilise the resources, work flow must be well scheduled. The completion of all the activities necessary to finish a task constitutes work flow. Workflow may include a variety of elements, including data, programmes, reporting, sequencing, etc. Any process, data aggregation, data segmentation, and distribution/redistribution are the structures to operate with such components [33]. Work flow scheduling is more widely used than job scheduling because it is more effective at identifying the best course of action for handling complex applications by considering various task-related limitations.

For the most part, researchers model workflow scheduling using directed acyclic graphs. In any cloud environment, it is crucial to take the computation cost and completion time into account while planning the work flow. Any IT staff member who must manually calculate the work flow must have the necessary background knowledge to do it correctly [34]. In the cloud context, it's crucial to satisfy the work flow scheduling optimisation requirements in order to implement an effective work flow management system. The work flow management system finds, controls, and executes the work flows using the available computing resources. Computational logic representation handles the sequence of work flows' execution. The work flow management system may be built to address issues like process optimisation, effective process management, system integration, rescheduling, improvement in maintainability, etc. [35]. Several nature-inspired based or soft computing-based methodologies are developed in the cloud environment for efficient cost optimisation. Cost optimisation can be handled by single objective and multi objective algorithms. The optimisation criteria could include things like wait times, service level agreements, availability, budget allocation, and dependability control.

VII. CONCLUSION

The potential of nature-inspired optimisations to generate promising solutions for a variety of applications has made them extremely popular over the past 20 years. However, it does not seem that because they are infants, there is no immediate need for treatment. This chapter explores the applicability of nature-inspired optimisation methods for several cloud computing research fields. The focus of this survey has remained on applications like load balancing, task scheduling, process scheduling, cost optimisation, etc. Additionally, numerous perspectives are used to analyse the type and character of distinct nature-inspired algorithms. It has been discovered that PSO and ACO have been used frequently to address practically all types of issues in cloud computing.

In contrast to some of the traditional cloud computing techniques, nature-inspired algorithms are reliable, scalable, and efficient to utilise. Realising the situation, there is still more work to be done on such algorithms. Although these algorithms are quite effective at providing the best results, there may still be a considerable gap between theory and practise. There are still problems to be solved in the areas of thorough mathematical analysis, convergence analysis to determine the optimality condition, appropriate trade-offs between exploration and exploitation, precise tuning of algorithmic parameters, etc. In addition to these, recently created algorithms based on natural phenomena have been developed, including the multi verse optimisation, lion optimisation, whale optimisation, dragonfly optimisation, virus colony search optimisation, elephant herding optimisation, social spider optimisation, social emotional optimisation, moth search algorithm, intelligent water drop algorithm, krill herd algorithm, wind driven optimisation, kidney inspired algorithm, bird swarm algorithm, and ant lion optimisation.

References

[1] J. Kennedy, Russell Eberhart, Particle swarm optimization. Proc. IEEE Int. Conf. Neural Networks 4, 1942–1948 (1995)

[2] X.S. Yang, S. Deb, Cuckoo search via Lvy flights, in World Congress on Nature and Biologically Inspired Computing (NaBIC 2009) (IEEE, 2009), pp. 210–214

[3] X.S. Yang, A new meta-heuristic bat-inspired algorithm, Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), p. 6574

[4] K.M. Passino, Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst. 22(3), 52–67 (2002)

[5] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Global Optim. 39(3), 459–471 (2007)

[6] D. Teodorovi, M. DellOrco, Bee colony optimization cooperative learning approach to complex transportation problems, inAdvanced OR and AI Methods in Transportation: Proceedings of 16th MiniEURO Conference and 10th Meeting of EWGT (13–16 September 2005) Poznan: Publishing House of the Polish Operational and System Research (2005), pp. 51–60

[7] R. Tang, S. Fong, X.S. Yang, S. Deb, Wolf search algorithm with ephemeral memory, in IEEE Seventh International Conference on Digital Information Management (ICDIM) (2012), pp. 165–172

[8] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Gener. Comput. Syst. 25(6), 599–616 (2009)

[9] I. Foster, Y. Zhao, I. Raicu, S. Lu. Cloud computing and grid computing 360-degree compared. in Grid Computing Environments Workshop, IEEE, GCE'08 (2008), pp. 1–10

[10] Q. Zhang, L. Cheng, R. Boutaba, Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. 1(1), 7–18 (2010)

[11] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Kaminski, G. Lee et al., A view of cloud computing. Commun. ACM 53(4), 50–58 (2010)

[12] P. Mell, T. Grance, The NIST definition of cloud computing. National Institute of Standards and Technology, Information Technology Laboratory, Technical Report Version 15, (2009)

[13] H. Das, G.S. Panda, B. Muduli, P.K. Rath. The complex network analysis of power grid: a case study of the West Bengal power network. in Intelligent Computing, Networking, and Informatics (Springer, New Delhi, 2014), pp. 17–29

[14] Das, Himansu, S.K. Mishra, D.S. Roy, The topological structure of the Odisha power grid: a complex network analysis. IJMCA 1(1), 012–016 (2013)

[15] Das, Himansu, D.S. Roy, A grid computing service for power system monitoring. Int. J. Comput. Appl. 62(20) (2013)

[16] H. Das, A.K. Jena, P.K. Rath, B. Muduli, S.R. Das. Grid computing-based performance analysis of power system: a graph theoretic approach. in Intelligent Computing, Communication and Devices (Springer, New Delhi, 2015), pp. 259–266

[17] I. Kar, R.N Ramakant Parida, H. Das, Energy aware scheduling using genetic algorithm in cloud data centers. in IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016), pp. 3545–3550

[18] Sarkhel, Preeta, H. Das, L.K. Vashishtha. Task-Scheduling Algorithms in Cloud Environment. in Computational Intelligence in Data Mining (Springer, Singapore, 2017), pp. 553–562

[19] C.R. Panigrahi, M. Tiwary, B. Pati, H. Das, Big Data and Cyber Foraging: Future Scope and Challenges. in Techniques and Environments for Big Data Analysis (Springer International Publishing, 2016), pp. 75–100

[20] X. Xin, H. Yu, A game theory approach to fair and efficient resource allocation in cloud computing. Math. Probl. Eng. 2014, 1–14 (2014)

[21] L. Mashayekhy, M.M. Nejad, D. Grosu, Cloud federations in the sky: Formation game and mechanism. IEEE Trans. Cloud Comput. 3(1), 14–27 (2015)

[22] A. Corradi, M. Fanelli, L. Foschini, VM consolidation: A real case based on OpenStack cloud. Future Gener. Comput. Syst. 32, 118–127 (2014)

[23] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, D. Wetherall, reducing network energy consumption via sleeping and rate-adaptation, in NSDI, Vol. 8 (2008), p. 323336

[24] P. Arroba et al., Server power modeling for run-time energy optimization of cloud computing facilities. Energy Procedia 62, 401–410 (2014)

[25] M.A. Sharkh, A. Shami, An evergreen cloud: Optimizing energy efficiency in heterogeneous cloud computing architectures. Veh. Commun. 9(11), 199–210 (2017)

[26] D.-M. Bui et al., Energy efficiency for cloud computing system based on predictive optimization. J. Parallel Distrib. Comput. 102, 103–114 (2017)

[27] D.L. Eager, E.D. Lazowska, J. Zahorjan, Adaptive load sharing in homogeneous 647 distributed systems. IEEE Trans. Softw. Eng. 12(5), 662675

[28] M.G. Avram, Advantages and challenges of adopting cloud computing from an enterprise perspective. Procedia Technol. 12, 529534 (2014)

[29] M.D. Assuno, A. Costanzo, R. Buyya, A cost-benefit analysis of using cloud computing to extend the capacity of clusters. Clust. Comput. 13(3), 335347 (2010)

[30] Y. Zhu, P. Liu, Multi-dimensional constrained cloud computing task scheduling mechanism based on genetic algorithm. Int. J. Online Eng. 9, 1518 (2013)

[31] L. Wu, Y.J. Wang, C.K. Yan, Performance comparison of energy-aware task scheduling with GA and CRO algorithms in cloud environment, in Applied Mechanics and Materials (2014), p. 204208

[32] S. Xue et al., An ACO-LB algorithm for task scheduling in the cloud environment. J. Softw. 9(2), 466473 (2014)

[33] S. Xue, J. Zhang, X. Xu, An improved algorithm based on ACO for cloud service PDTs scheduling. Adv. Inf. Sci. Serv. Sci. 4(18), 340348 (2012)

[34] W.-J. Wang, Y.-S. Chang, W.-T. Lo, Y.K. Lee, Adaptive scheduling for parallel tasks with QoS satisfaction for hybrid cloud environments. J. Supercomput. 66(2), 129 (2013)

[35] K. Li, G. Xu, G. Zhao, Y. Dong, D. Wang, Cloud task scheduling based on load balancing ant colony optimization. in 2011 IEEE Sixth Annual Chinagrid Conference (ChinaGrid) (2011), pp. 3–9