A Survey on Resource Allocation in a Cloud

Santhosh Pawar
Department of CS
KSWU Vijayapura

Abstract-*Cloud computing is essential in the field of modern computing systems, where cloud providers have to provide effective resource for the users to increase the quality of service. A data centre consists of a large number of servers or hosts which are the key factor of cloud environment. The use of the enormous amount of computing and data centres generally leads to consumption of large amount of energy Therefore, the datacentre resources have to be distributed such that energy efficiency is maximized. The present paper surveys various resource allocation strategies and methods that are efficient in terms of energy. These strategies have been compared on the basis of their techniques. We analyse the most promising existing research in resource management and examine monitored values, supported application classes and the most important criteria for evaluating the effectiveness of the approach. It discusses methods to evaluate and model the energy consumed by these resources, and describes techniques that operate at a distributed system level, trying to improve aspects such as resource allocation.*

Keywords-Cloud computing, resource allocation.Saas,PaaS, IaaS

I. INTRODUCTION

Cloud computing is increasingly being used for what was known as „on-demand‟ and „utility computing‟. It appears that „Cloud‟ has become the label of choice for pay-per-use access to a wide variety of third-party applications and computational resources on a massive scale. Clouds are now supporting patterns of less-predictable resource use for applications and services across the IT spectrum, from online office applications to high-throughput transactional services and high-performance computations involving substantial quantities of processing cycles and storage. Clouds may be considered as an extension of service-oriented computing that covers computational hardware-based resources as well as software, with concomitant business benefits in cost reduction where such services scale efficiently. There are currently three cloud-based service models: Software as a Service (SaaS), where the consumer uses an application, but does not control the operating system, hardware or network infrastructure. In this situation, the user steers applications over the network. Next is Platform as a Service (PaaS), where the users host an environment for their applications. The users control the applications, but do not control the operating system, hardware or network infrastructure, which they are using. Finally, there is Infrastructure as a Service (IaaS), where the user accesses fundamental computing resources such as CPU, memory, middleware and storage. The consumer controls the resources, but not the cloud infrastructure beneath them. Optimizations for energy conservation can be made at hardware and software levels. Hardware level energy optimizations are achieved through circuit design by implementing smaller silicon process geometries, auto idle detection circuits and active well biasing techniques [3]. Software level energy optimizations are implemented in operating system through Green Scheduling techniques that analyze active processes for energy requirements. The performance of distributed computing systems at large scale has been significantly constrained by their excessive power requirements. This ever increasing energy consumption for powering and cooling is a major limiting factor in the running and expansion of the data centres. Besides, the energy consumption issue in these systems raises both business and environmental concerns. Actually clouds deal with dynamic and heterogeneous resources and applications, scheduling and resource allocation play an important role in optimization of resource usage: hence, better energy efficiency. The gains that we get in energy efficiency from moving business and personal software on the cloud is achieved from the fact that data centers which hosts these cloud servers are far more energy efficient than the IT infrastructure that the small companies deploy and also the fact that the servers of data centers are much more efficient.

II. RESOURCE ALLOCATION

Resource allocation is process of assigning the resources that are available to the third parities by the cloud providers in the cloud environment when needed. The resource allocations are based on the pricing schema and the time of allocation.

Allocation of resources includes two factors to be avoided, • Over provisioning of resources- this complexity arises when the resources are sold more than the resources that are available. • Under provisioning of resources – this problem arise only when less number of resources are assigned to the people than the demand.

A. Types of Resources

Computing Resource - Which includes collection of memory, network, processor, input/output devices in the cloud environment .these are collectively called as the physical machines (PM). According to the user needs the computing resources should be allocated or purchased. The concept of VM comes under PM where PM creates virtual software for user to run on virtual machine in different OS, application and platform. Networking Resources - Bandwidth, storage, communication, challenges, traffics such problems arrives in networking side which can be taken care by working on protocols to enhance the QOS of the cloud. Storage Resources - When the scalability comes to storage it must be achieved by considering the ACID property. Now a day the cloud storage is based on the "NO SQL" data storage technologies under some functional conditional which has been includes for storing documents, key-value. Power Resources - Usage of power per day by the system deals with power resource. The energy consumed by the system for providing and allocating the resource is much less than energy consumed by the system that is idle, waiting for a resource to be allocated. This lead to another technology namely green-cloud computing.

B. Resource Allocation problem

The cloud service providers deployed a large number of datacenters across the world to full fill the cloud consumer request and to provide the cloud service. Datacenters usually contain a huge number of server and these serving machines also fulfill the client request. In datacenters many physical machines are grouped into the form of units, and these units are called clusters [7]. Resource Allocation in cloud can be termed as the allocation of datacenter‟s tools to the to the client request along to the needs and use. The resource assigning problem must be rooted on such criteria as, service needs, Quality of Service (QoS), efficient and green utilization of resources [2]. The cloud service producer allocates these datacenters resources in a mobile way for their users and charge them pay per use-based guaranteed and reliable services. The cloud computing offers a multitenant environment, where multiple clients generate multiple requests. Numerous service users can request number of cloud services along with QoS simultaneously at any time with specified amount of resources that they need. Upon receiving a consumer request, the cloud mediator check out the matching of task demand to extant computation tools and determine its presumption, whether consumer task demand can be served on available computation tools rooted on QOS needs or not. Then the mediator refers the task demand to resource scheduler for scheduling through well-tuned rationing of resources. On the behalf of QoS needs the resources are allocated from cloud resource pool for user‟s workload and the mapping between user workload and available resources is done by resource scheduler [8]. The resource requirements of consumer request in cloud despotic and period species. Hence, it is prerequisite to organize the presence of resources and scheduling of all the resources instantly. Afterward here is a necessity to supply the sufficient resources to demanded service user in a well administered path for completing their needs. Efficient allocation of resources is a very tough job in cloud data centers. At present Allocating resources efficiently for cloud consumer in cloud data centers has been an affinity for researcher.

C. Energy Consumption Model

The models described are based on power consumption measurements and published specifications of representative equipment [15], [16]. Those models include descriptions of the common energy-saving techniques employed by cloud computing service providers. The models are used to calculate the energy consumption per bit for transport and processing, and the power consumption per bit for storage. The energy per bit and power per bit are fundamental measures of energy consumption, and the energy efficiency of cloud computing is the energy consumed per bit of data processed through cloud computing. Performing calculations in terms of energy per bit also allows the results to be easily scaled to any usage level. i. User Equipment

A user may use a range of devices to access a cloud computing service, including a mobile phone (cell phone), desktop computer, or a laptop computer. These computers typically comprise a central processing unit (CPU), random access memory (RAM), hard disk drive (HDD), graphical processing unit (GPU), motherboard, and a power supply unit. Peripheral devices including speakers, printers, and visual display devices are often connected to PCs. These peripheral devices do not influence the comparison between conventional computing and clouds computing and so are not included in the model. In our analysis, we assume that when user equipment is not being used it is either switched off or in a deep sleep state (negligible power consumption) [17]. ii. Data Centers A modern state-of-the-art datacenter has three main components -data storage, servers, and a local area network (LAN) [15]. The functionality of this equipment as well as some of the efficiency improvements in cloud computing data centers over traditional data centers [17]. Long-term storage of data in a datacenter is provided by hard disk arrays, together with associated equipment. Hard disk arrays include supporting functionality such as cache memories, disk array controllers, disk enclosures, and redundant power supplies. In a cloud computing datacenter, all the storage space in the datacenter is consolidated and hard disk usage is centrally coordinated [19], [20]. Consolidation and central coordination

minimizes the total number of hard disks used, greatly increasing the overall energy efficiency of storage. In addition, files that are not accessed regularly are stored in a different set of capacity optimized hard disks [21]. These hard disks enter a low-power mode when not in use and consume negligible energy. While infrequently used data files are stored on a disk, the data rate and latency of disk read operations is generally inadequate for services such as file hosting, which entail frequent accesses to the file. Data for these services are cached in RAM on one or more servers. Additional servers perform datacenter management and, in a high performance computing facility, provide on-demand computing. The server performance depends on the computational characteristics of the task being performed, including the number of floating point operations, memory accesses, and suitability for parallel processing. Through server virtualization/consolidation, a very large number of users can share a single server, which increases utilization and in turn reduces the total number of servers. Users do not have or need any knowledge of the tasks being performed by other users and utilize the server as though they are the only user on the server. During periods of low demand, some of the servers enter a sleep mode which reduces energy consumption. To reflect the efficiency gains from sleeping, virtualization, and consolidation, in our analysis, the computation servers and file hosting servers are fully utilized.

III. RESOURCE ALLOCATION STRATEGIES FOR CLOUD COMPUTING

In cloud computing resource allocation plays an important role specifically in pay-per-use dispersion where the counting of tools is changed to service allocator. Cloud service providers established large-scale virtualized datacenters for fulfilling the request of the clients. Put running these data centers permanently for continue service allocation, commonly they postulate a huge mass of power. Therefore, the resources have to be allocated such that energy efficiency is maximized. Hence, energy consumption due to improper resource allocation is a critical issue. The issue here is to allocate proper resources to perform the computation with minimal energy consumption. To identify a technique that can QoS as well as reduce energy consumption in cloud environment is one of the challenging task. In this section, we present significant research carried out in resource allocation for cloud computing. An Energy Efficient Computing on Green Cloud Computing architecture of resource management for a virtualized cloud datacenter has been proposed by Nitin Jain. In [10] Moving towards Cloud Computing, high performance computing usage of huge datacenter (DC) and huge cluster is increasing day by day and energy consumption by these DC and energy dissipation in environment by these DC is also rising day by day. The large amount of CO2 dissipation in environment has generated the necessity of Green computing (saving energy by recycling it and reusing it over a period of time and minimizing the wastage in terms of usage of resources). More processor chips generates more heat, more heat requires more cooling and cooling again generates heats and thus we come to a stage where we want to balance the system by getting the same computing speed at decreased energy consumption. Dynamic Resource Allocation using Virtual Machines proposed by Zhen Xiao and Weijia Song. In [1] present a system that uses virtualization technology to allocate datacenter resources dynamically based on application demands and support green computing by optimizing the number of servers in use. We introduce the concept of "skewness" to measure the unevenness in the multi-dimensional resource utilization of a server. By minimizing skewness, we can combine different types of workloads nicely and improve the overall utilization of server resources. We develop a set of heuristics that prevent overload in the system effectively while saving energy used. Trace driven simulation and experiment results demonstrate that our algorithm achieves good performance. Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing proposed by Daniel Warneke. In [8] discuss the opportunities and challenges for efficient parallel data processing in clouds and present our research project Nephele. Nephele is the first data processing framework to explicitly exploit the dynamic resource allocation offered by today‟s IaaS clouds for both, task scheduling and execution. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution. Based on this new framework, we perform extended evaluations of MapReduce-inspired processing jobs on an IaaS cloud system and compare the results to the popular data processing framework Hadoop.

Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds proposed by Xiaomin Zhu. In [3] firstly propose in this paper a novel rolling-horizon scheduling architecture for real-time task scheduling in virtualized clouds. Then a task-oriented energy consumption model is given and analysed. Based on our scheduling architecture, we develop a novel energy-aware scheduling algorithm named EARH for real-time, aperiodic, independent tasks. The EARH employs a rolling-horizon optimization policy and can also be extended to integrate other energy-aware scheduling algorithms. Furthermore, we propose two strategies in terms of resource scaling up and scaling down to make a good trade-off between task‟s schedulability and energy conservation. Extensive simulation experiments injecting random synthetic tasks as well as tasks following the last version of the Google cloud tracelogs are conducted to validate the superiority of our

EARH by comparing it with some baselines. The experimental results show that EARH significantly improves the scheduling quality of others and it is suitable for real-time task scheduling in virtualized clouds. Towards Pay-As-You-Consume Cloud computing process proposed by Shadi Ibrahim, Bingsheng. In [2] studies reveal the reason for such variations is interference among concurrent virtual machines. The amount of interference cost depends on various factors, including workload characteristics, the number of concurrent VMs, and scheduling in the cloud. In this work, we adopt the concept of pricing fairness from micro economics, and quantitatively analyze the impact of interference on the pricing fairness. To solve the unfairness caused by interference, we propose a pay-as-you-consume pricing scheme, which charges users according to their effective resource consumption excluding interference. The key idea behind the pay-as-you-consume pricing scheme is a machine learning based prediction model of the relative cost of interference. Our preliminary results with Xen demonstrate the accuracy of the prediction model, and the fairness of the pay-as-you-consume pricing scheme. Dataflow-Based Scientific Workflow Composition Framework proposed by Xubo Fei and Shiyong Lu. In [7] Scientific workflow has recently become an enabling technology to automate and speed up the scientific discovery process. Although several scientific workflow management systems (SWFMSs) have been developed, a formal scientific workflow composition model in which workflow constructs are fully compositional one with another is still missing. In this work, we used a dataflow-based scientific workflow composition framework consisting of 1) a dataflow-based scientific workflow model that separates the declaration of the workflow interface from the definition of its functional body; 2) a set of workflow constructs, including Map, Reduce, Tree, Loop, Conditional, and Curry, which are fully compositional one with another; 3) a dataflow-based exception handling approach to support hierarchical exception propagation and user-defined exception handling. Our workflow composition framework is unique in that workflows are the only operands for composition; in this way, our approach elegantly solves the two-world problem in existing composition frameworks, in which composition needs to deal with both the world of tasks and the world of workflows.

Meeting Deadlines of Scientific Workflows in Public Clouds with Tasks Replication proposed by Rodrigo N. Calheiros and Rajkumar Buyya. In [6] studies, previous research in execution of scientific workflows in Clouds either try to minimize the workflow execution time ignoring deadlines and budgets or focus on the minimization of cost while trying to meet the application deadline. However, they implement limited contingency strategies to correct delays caused by underestimation of tasks execution time or fluctuations in the delivered performance of leased public Cloud resources. To mitigate effects of performance variation of resources on soft deadlines of workflow applications, we propose an algorithm that uses idle time of provisioned resources and budget surplus to replicate tasks. Simulation experiments with four well-known scientific workflows show that the proposed algorithm increases the likelihood of deadlines being met and reduces the total execution time of applications as the budget available for replication increases. Balancing Energy in Processing, Storage, and Transport in cloud proposed by Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker. In [9] present an analysis of energy consumption in cloud computing. The analysis considers both public and private clouds, and includes energy consumption in switching and transmission as well as data processing and data storage. We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing can enable more energy-efficient use of computing power, especially when the computing tasks are of low intensity or infrequent. However, under some circumstances cloud computing can consume more energy than conventional computing where each user performs all computing on their own personal computer (PC).

IV. CONCLUSION AND FUTURE WORK

In recent time, energy efficient allocation of data centres resources has evolved as one the critical research issue. In any cloud computing environment if we want to renovate the power proficiency of the cloud datacenters, we must look into improving the resource allocation made to the datacenters. This paper surveys various resource allocation models and techniques used to improve energy consumption by cloud data centers supporting the cloud computing. Since the resource allocation is the core concept of cloud computing and various mechanisms has been dealt for allocation in cloud environment, the allocation mechanisms changes according to the evolution methods at different levels of cloud. Hence the conclusion is done by setting a fundamental research technique. Finally, we highlighted the potential benefits of applying application classification in combination with a general application model in this area. We believe that this combination can have a significant impact, especially regarding the development and automation of general resource management solutions. Our future research involves the evaluation of these possibilities.

REFERENCES

[1] Z. Xiao, W. Song, and Q. Chen, “Dynamic resource allocation using virtual machines for cloud computing

environment", IEEE Trans.Parallel Distrib. Syst., vol. 24, no. 6, pp. 1107-1117, Jun. 2013.

[2] S. Ibrahim, B. He, and H. Jin, "Towards pay-as-you-consume cloud computing", Proc. 8th IEEE Int‟l. Conf. Service Computing (SCC „11), pp. 370-377, Jul. 2011.

[3] X. Zhu, L. Yang, H. Chen, J. Wang, S. Yin, and X. Liu, "Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds", IEEE trans. Cloud Computing, vol. 2, no. 2, pp. 168-180, Jun. 2014.

[4] X. Liu, Y. Yang, Y. Jiang, and J. Chen, "Preventing temporal violations in scientific workflows: where and how", IEEE Trans. Softw. Eng., vol. 37, no. 6, pp. 805-825, Nov. 2011.

[5] G. Juve, A. Chervenak, E. Deelman, S. Bharathi , G. Mehta, and K. Vahi, "Characterizing and profiling scientific workflows", Future Gener. Comput. Syst., vol. 29, no. 3, pp. 682-692, Mar. 2013.

[6] R.N. Calheiros and R. Buyya, "Meeting Deadlines of Scientific Workflows in Public Clouds with Tasks Replication", IEEE Trans. Parallel Distrib. Syst., vol.25, no.7, pp. 1787-1796, Jul. 2014.

[7] X. Fei and S. Lu, A dataflow-based scientific workflow composition framework, IEEE Trans. Services Computing, vol. 5, no. 1, pp. 45-58, Mar. 2012.

[8] D. Warneke and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud", IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 6, pp. 985-997, Jun. 2011.

[9] J. Baliga, R.W. Ayre, K. Hinton and R.S. Tucker. "Green cloud computing: Balancing energy in processing, storage, and transport", Proceedings of the IEEE, vol. 99, no. 1, pp. 149-167, Dec. 2011.

[10] A. Jain, M. Mishra, S.K. Peddoju, Energy efficient computing-Green cloud computing, Proc. 1st IEEE Int‟l. Conf. Energy Efficient Technologies for Sustainability (ICEETS„13), pp. 978-982, Apr. 2013.

[11] L. Qi, Y. Tang, W. Dou, and J. Chen, "Combining local optimization and enumeration for qos-aware web service composition", IEEE Int‟l. Conf. Web Services (ICWS), pp. 34-41, Jul. 2010.

[12] X. Li, J. Wu, S. Tang, and S. Lu. Let‟s Stay Together: Towards Traffic Aware Virtual Machine Placement in Data Centers. Proc. 33rd IEEE Int‟l. Conf. Computer Communications (INFOCOM ‟14), pp. 1842-1850, Apr. 2014.

[13] Y. Guo, A.L. Stolyar, and A.Walid, "Shadow-routing based dynamic algorithms for virtual machine placement in a network cloud", Proc.32nd IEEE Int‟l. Conf. Computer Communications (INFOCOM ‟13), pp. 620-628, Apr. 2013.

[14] G. Terzopoulos and H. Karatza, "Performance evaluation and energy consumption of a real-time heterogeneous grid system using DVS and DPM", Simulation Modelling Practice and Theory, vol. 36, pp. 33-43, May. 2013.

[15] Y.H. Du, P.C. Xiong, Y.S. Fan, and X. Li, "Dynamic checking and solution to temporal violations in concurrent workflow processes", IEEE Trans. Man and Cybernetics, Part A: Systems and Humans, vol. 41, no. 6, pp. 1166-1181, Nov. 2011.

[16] J. Zheng, T.S.E. Ng, K. Sripanidkulchai, and Z. Liu. "Pacer: A progress management system for live virtual machine migration in cloud computing", IEEE Transactions on Network and Service Management, vol. 10, no. 4, pp. 369-382, Dec. 2013.

[17] D. Meisner, B.T. Gold, and T.F. Wenisch, "PowerNap: eliminating server idle power", ACM SIGARCH Computer Architecture News, vol. 37, no. 1, pp. 205-216, Mar. 2009.

[18] Y. C. Lee and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems", The Journal of Supercomputing, vol. 60, no. 2, pp. 268-280, May 2012.

[19] C. Isci, J. Liu, B. Abali, J.O. Kephart, and J. Kouloheris, "Improving server utilization using fast virtual machine migration", IBM J. Research and Development, vol. 55, no. 6, pp. 1-12, Nov. 2011.

[20] S. He, L. Guo and Y. Guo, "Real time elastic cloud management for limited resources", Proc. 3rd IEEE Int‟l. Conf. Cloud Computing (CLOUD „11), pp. 622-629, Jul. 2011.

[21] Designing Your Cloud Infrastructure.

[22] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement", Proc.29th IEEE Int‟l. Conf. Computer Communications (INFOCOM ‟10), pp. 1-9, Mar. 2010.

[23] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", Software: Practice and Experience, vol. 41, no. 1, pp. 23-50, Jan. 2011.

[24] N. Liu, Z. Dong and R. Rojas-Cessa, "Task scheduling and server provisioning for energy-efficient cloud-computing data centers", 33rd IEEE Int‟l. Conf. Distributed Computing Systems Workshops (ICDCSW‟13), pp. 226-231, Jul. 2013. [25] S. Zhang, B. Wang, B. Zhao and J. Tao, "An Energy-Aware Task Scheduling Algorithm for a Heterogeneous Data Center", 12nd IEEE Int‟l. Conf. Trust, Security and Privacy in Computing and Communications (TrustCom‟13), pp. 1471-1477, Jul. 2013.