Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

Research Paper

Efficient Workload Management in Cloud Computing

Engineering

KEYWORDS : Crumb rubber, utilization, compressive strength, low cost, sustainable

Loveneesh Singla Chandigarh Engineering College, Mohali (Punjab)

Sahil Vashist Assistant Professor, Chandigarh Engineering College, Mohali (Punjab)

ABSTRACT

Cloud Computing faces a huge challenge to deal with the high workload and un precedent requests, therefore in order to see the latency variations in cloud computing and its impact we have propose a strategy which presume the work load on cloud computing and a lot the workload accordingly. By establishing the network connections in real world and simulating the results the effectiveness of our scheme has been shown.

1. Introduction:

In computer science, the Cloud Computing describes a type of outsourcing of computer services which are similar to the way in which the supply of electricity is out sourced. The users can simply use it then they do not need to worry from where the electricity is coming from, how it is made or transported from. Every month, they pay for what they consumed. The idea behind cloud computing is similar that the user can simply use storage computing power or specially crafted development environments without having to worry how these work internally. Cloud computing is a system architecture model for internet based computing. It is the development and use of computer technology on the internet. Task centers and cloud computing services providers hope that the extensive adoption of the cloud will bring them more revenue and they are dynamically promoting the technology. Cloud computing which is a kind of distributed system consisting of a collection of interconnected Task centers which constitutes a numbers of virtual machines obtained by virtualization processes that are dynamically provisioned and accessible as one or more unified computing resources based on service level agreements. A cloud Task center usually deploys many servers, compactly packed to make the most of the space utilization. Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. The 'cloud' in cloud computing can be defined as the set of hardware, network, storage services and interface that combine to deliver aspects of computing as a service. Cloud service includes the delivery of software, infrastructure &storage over the internet based on a user demand. In computer science cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. The popularity of the term Cloud computing can be attributed to its use in marketing to sell hosted services in the sense of Application Service Provisioning that run Client server software on a remote location.

2. Related Work

In **[1]** authors outline previous contributions to the discussion of energy efficiency of cloud computing, provide a working definition of cloud computing and discuss its importance, which will grow as the technology matures and becomes well known. According to the author the assessment of the energy efficiency of cloud computing cannot be based only on task centers due to the importance of the intermediate communication networks that support the overall activity of providing cloud computing services and the devices used to access cloud services. The other components should be taken into account when measuring the energy efficiency of cloud computing. The energy consumption of mobile devices is generally good but the cellular and fixed communication networks that support cloud services have been noted to consume high amount of energy and this consumption is growing. There is the need to improve the energy efficiency of communication networks and the Internet in order to meet the new levels of demand. In this paper they analyzed the energy optimization of the network infrastructure should be paramount if the improved energy efficiency of task centers will result in overall benefits.

Keville et al. [2] examine the use of ARM-based clusters for low-power, high performance computing. This work examines two likely use-modes: (i) a standard dedicated cluster; and (ii) a cluster of pre-configured virtual machines in the cloud. A 40-node department-level cluster based on an ARM Cortex-A9 is compared against a similar cluster based on an Intel Core 2 Duo, in contrast to a recent similar study on just a 4-node cluster. For the NAS benchmarks on 32- node clusters, ARM was found to have a power efficiency ranging from 1.3 to 6.2 times greater than that of Intel. This is despite Intel's approximately five times greater performance. The particular efficiency ratio depends primarily on the size of the working set relative to L2 cache. In addition to energyefficient computing, this study also emphasizes fault tolerance: an important ingredient in high performance computing. In this paper they analyzed that it relies on two recent extensions to the DMTCP check point restart package. DMTCP was extended (i) to support ARM CPUs, and (ii) to support check pointing of the Qemu virtual machine in usermode. DMTCP is used both to checkpoint native distributed applications, and to checkpoint a network of virtual machines. In [3] propose a software and lightweight approach to accurately estimate the power usage of virtual machines and cloud servers. It explores

hypervisor-observable performance metrics to build the power usage model. To configure cloud resources, it considers both the system power usage and the SLA requirements, and leverage learning techniques to achieve autonomic resource allocation and optimal power efficiency. In this paper they analyzed that it implements a prototype of the proposed power management system and test it on a cloud test bed.

In [4] a novel approach to virtual machine consolidation for saving energy is presented. In this paper they analyzed that it is based on energy efficient storage migration and live migration of virtual machines to take advantage of the lacking energy proportionality of commodity hardware.

Dharwar et al. [5] outlines the state-of-the-art in power-management technology on server hardware and describes how these raw features can be abstracted into a set of energy policies. In this paper they analyzed to explain how these policies or energy-profiles can be used to run a cloud taskcenter energy efficiently. Further, this work also highlights some of the challenges involved in running cloud infrastructures in the emerging markets optimally despite some unique energy constraints.

Ye et al. [6] focuses on the live migration strategy of multiple virtual machines with different resource reservation methods. It first describe the live migration framework of multiple virtual machines with resource reservation technology. Then it perform a series of experiments to investigate the impacts of different resource reservation methods on the performance of live migration in both source machine and target machine. Additionally, it analyze the efficiency of parallel migration strategy and workload-aware migration strategy. In this paper they ana-

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

lyzed the metrics such as downtime, total migration time, and workload performance overheads are measured.

Kejiang et al. [7] present a virtual machine based energy-efficient task center architecture for cloud computing. It investigates the potential performance overheads caused by server consolidation and live migration of virtual machine technology. In this paper they analyzed the experimental results show that both the two technologies can effectively implement energy-saving goals with little performance overheads. Efficient consolidation and migration strategies can improve the energy efficiency.

Yamini et al. [8] propose workflow consolidation particularly in clouds has become an important approach to streamline resource usage and in turn improve energy efficiency. Based on the fact that resource utilization directly relates to energy consumption, it has modeled their relationship and developed two energy-conscious workflow consolidation heuristics. The cost functions incorporated into these heuristics effectively capture energy-saving possibilities and their capability has been verified by my evaluation study. In this paper they analyzed the results in this study should not have only a direct impact on the reduction of electricity bills of cloud infrastructure providers, but also imply possible savings(with better resource provisioning) in other operational costs (e.g., rent for floor space).

Karthikeyan et al. [9] focuses heuristics energy efficiency approach to reduce the energy consumption and carbon emission by using efficient VM migration algorithm. To enhance the quality of services, improved version of auto scaling technique as scalable frame work can be used. In this technique cloud resources can be allotted and booted quickly to meet response time requirements depends up on incoming load. In this paper they analyzed in addition to that we will also observe what are the challenges faces during the implementation and what are the performance metrics have to be taken.

Zhou et al. [10] This paper focuses propose a random dynamic scheduling scheme to deal with the demand uncertainty based on Monto-Carlo sampling. Moreover, it show that the proposed scheduling scheme possesses a desirable property since it inherits from the problem of known user demand. Subsequently, it present exact steps to implement the proposed scheduling. Finally, in this paper they analyzed the numerical simulation results are provided to validate its efficiency.

Anton et al. [11] define an architectural framework and principles for energy-efficient Cloud computing. Based on this architecture, we present our vision, open research challenges, and resource provisioning and allocation algorithms for energy-efficient management of Cloud computing environments. The proposed energy-aware allocation heuristics provision task center resources to client applications in a way that improves energy efficiency of the task center, while delivering the negotiated Quality of Service (QoS). In this paper they analyzed that it conducts a survey of research in energy-efficient computing and propose: (a) architectural principles for energy-efficient management of Clouds; (b) energy-efficient resource allocation policies and scheduling algorithms considering QoS expectations and power usage characteristics of the devices; and (c) a number of open research challenges, addressing which can bring substantial benefits to both resource providers and consumers.

Zhang al. [12] present a survey of cloud computing, highlighting its key concepts, architectural principles, state-of-the-art implementation as well as research challenges. The aim of this paper is to provide a better understanding of the design challenges of cloud computing and identify important research directions in this increasingly important area.

Berl et al. [13] the context of cloud computing, reviews the usage of methods and technologies currently used for energy-efficient operation of computer hardware and network infrastructure. After surveying some of the current best practice and relevant literature in this area, this paper identifies some of the remaining key research challenges that arise when such energy-saving techniques are extended for use in cloud computing environments.

Anton et al. [14] It discuss causes and problems of high power / energy consumption, and present a taxonomy of energy-efficient design of computing systems covering the hardware, operating system, virtualization and task center levels. In this paper they analyzed that the survey of various key works in the area and map them to our taxonomy to guide future design and development efforts.

3. System Model:

In the following, we start by refining the problem definition and then present the cost and time objective functions that we consider in this work. The main objective of this work is to deal with the heavy flow of workflow workflow and to stream line it with the user's process to provide better scheduling strategy. Since the main aim of the work is to view services of cloud computing as the source of workload processing tool as shown in Figure 1.

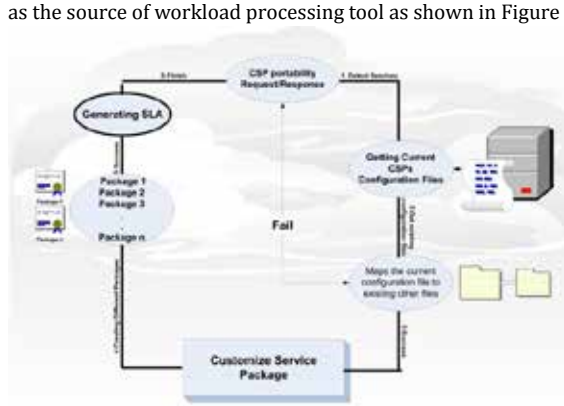

Figure 1: Workload on different CSPs

Formally, a workflow application is represented by G = (T,E), where T = {t1, ..., tn} which is a set of a number of job workflow. Workflow ti is called the immediate parent workflow of tj, i.e. tj generates workload ti. In a given graph, a workflow without any precedents is called an input workflow, denoted tinput and a workflow without successors is called an exit workflow, denoted texit. Let Task be a n × n matrix of communication task, where task[i, j] is the amount of task required to be transmitted from workflow ti to workflow tj . Since workflow workflow allocation and scheduling algorithms depends upon one input and one output. So, if there are more than one output they are associated to a non valued cost and non time cost. This modification allows obtaining a DAG with only one input and one exit workflow. Moreover, a workflow is an indivisible unit of work and is non-preemptive.

To manage a given workload all virtual machines are called as per demand and if we create a graph we can show that the edges which represents parent-child graph denoted RG. Formally, a resources graph is represented by RG = (VM,V), where VM = {VM1, ..., VMm} is a finite set of virtual machines types that define a virtual machine images and a task center locations. However, we consider that there is enough virtual machines for each type. Thus, a costumer can request and obtain sufficient virtual machines at any time. This assumption is reasonable in rather Cloud computing environment because it gives for user an "illusion of infinite resources". When there is no ambiguity, we omit term type and use virtual machine instead virtual machine type. V represents the set of directed edges. Each edge is denoted (VMi, VMj) corresponding to the link between these virtual machines. Let B be a m×m matrix, in which B[i, j] is the bandwidth between virtual machines VMi and VMj , where B[i, i] $\longrightarrow \infty$ means that there is no transfer task. Figure 1 (right side) shows an example of resources graph with real-life measurement of task transfer speeds (bandwidth) between different task centers of the Cloud provider Amazon.

3.1 Function focusing Time

Since focusing time is necessary to define the ST and FT attributes, which are generated from the virtual machines workflow. The partial allocation and scheduling refers to the fact that for each workflow the ST and FT values are computed using only the workflow that must be performed before it as shown in the following. More precisely, SM(tj) and FM(tj) are the latest start execution time and the latest finish execution time of workflow tj . For the input workflow tinput:

SM(tinput) = 0,

FM(tinput) = ST(Tinput) + ET(tinput, VM(tinput))…..(1)

For the other workflow in the graph, the *ST* and the *FT* values are computed recursively, starting from the initial workflow, as shown in Equation 2 and Equation 3. In order to compute the *FM* of a workflow *tj*, all immediate predecessor workflow of *tj* must have been assigned and scheduled with the consideration of the transfer time:

FT(tj) = ST(tj) + ET(tj, VM(tj)) …….(2)

ST(tj) = max

tp∈pred(tj)

{FT(tp) + TT(VM(tp), VM(tj))} ……..(3)

where pred(tj) is the set of immediate predecessors of workflow tj.

After all workflow in a graph are scheduled, the schedule length (i.e., the overall completion time) will be the finish time of the exit workflow. The schedule length, also called *makespan*, is defined as:

makespan = *FW*(*texit*)…….. (4)

Therefore, the time objective function is to determine the assignment of workflow of a given workflow application to virtual machines such that its schedule length is minimized.

3.2 Function focusing cost

In the following, we focus on the cost objective function which is the total expense for workflow execution including i) the workflow execution cost and ii) the task transfer cost between the used virtual machines. The cost function is a structure independent criterion defined as the sum of the costs

of executing all workflow. Thus, the cost objective function is to determinate the assignment of workflow of a given workflow application such that its overall execution cost is minimized.

The problem addressed in this work deals with the workflow application allocation and scheduling while simultaneously minimizing the makespan and the overall cost execution. As mentioned previously, this problem can be approached in several ways. In our work, we have opted for the effective solutions (Pareto solutions) computation. To achieve this objective, we propose in the following three multi-objective approaches. The first one is based on the overall cost execution function while the second one is conducted by the makespan function. Finally, we propose an approach taking into account the two functions together.

3.3 Resource allocation phase

In this phase the selection of an "optimal" virtual machine for each workflow is decided. In other words, the virtual machine which gives minimum execution and communication costs for a workflow is selected and the workflow is assigned to that virtual machine. More precisely, given the labeling of workflow in the graph levels, the allocation process explores the graph by starting the allocation workflow of level k, where the value of k is given by the following strategies:

i) topdown,
ii) bottom-up and
iii) mixed exploration and allocation strategy.

The top-down strategy consists of starting by the allocation of the initial workflow (level l1) to the virtual machine which gives minimum execution cost. After this assignment, the graph is traversed in a top-down fashion from level 2 to level L.

3.4 Time-based approach

While the previous approach is based on minimizing the *cost* function, the time-based approach, detailed in the following, is based on the *makespan* criterion. More precisely, the time-based approach attempts to minimize the overall completion time (i.e. execution and communication time). As the cost-based approach, the time-based approach is an application matching and scheduling algorithm for an "unbounded" number of virtual machines, which has three major phases, namely: i) a workflow sorting phase ii) an allocation phase and iii) Pareto selection phase.

3.5 Workflow sorting phase

This phase is the same as for the cost based algorithm. Recall that this phase allows grouping the workflow application workflow that are independent of each other. The cost-based and the time-based approaches differ mainly at resource allocation phase. Indeed, the first one approach focus on minimizing the cost function

Algorithm 1 Evaluation-Scheduling algorithm

Input: free slot n

Output: workflow's scheduling results

Step1: node n sends heartbeat to JobTracker to request workflow distribution

Step2: each workflowTracker tune workflow proportion factor

according to each workflow's real processing condition

Step3: JobTracker compute workflow's completing time TimeRemain and its executing rate , on i workflowtype Excute Rate nodes, among which Workflowtype means map or reduce.

Step4: be sure that if there have failure workflow, if have, tune the failure workflow, if not, tune the workflow which have never been tuned. And if there have no workflow which have never been tuned, then sort the running workflow according to their estimating executing time, and select the longest workflow i to tune.

Step5: calculation of the free slot n, *n workflowtype Excute Rate* and estimate the slow workflow i's executing time on the node n. if *n* < , then retune this slow workflow on *T* R*emain Time*node n, or else do nothing.

4. Simulation and Results:

We have implemented a custom simulator to model the proposed system. In our study we consider a maximum of 5000 task centers belonging to different CSPs. For simplicity we assume that one CSP has only one task center which participates in the federation. The simulator considers 6 different geographical locations and randomly assigns task centers to different locations, such that , the simulator keeps the number of resources available at different task centers as constant.

In Figure 1 we can evaluate that the latency of different locations have different impact on the performance of workload further if we move at the Probability distributive function (PDF) at Figure 2 we can see that the values at which the different strategies are working and the output is coming.

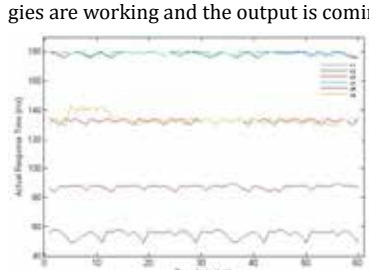

Figure 1: Latency at different Locations (ms)

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

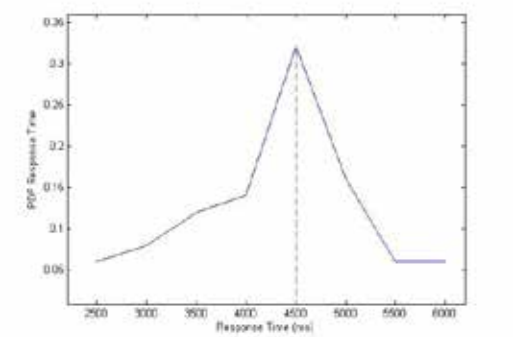

Figure 2: Probability distributive Function of Response Time

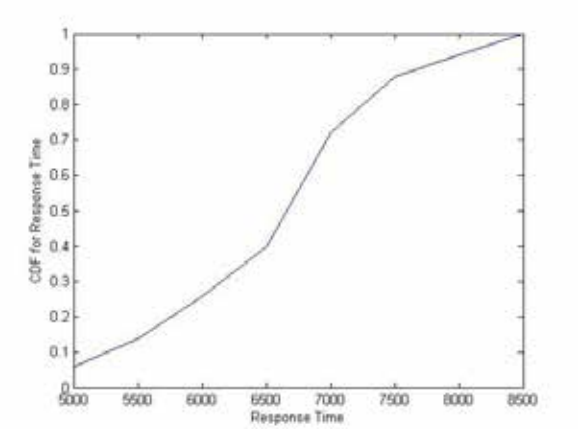

Figure 3: Cumulative Distributive Function of Response Time

In Figure 3 we can observe that our strategy process all the queries under 8500 ms which is more significant to deal with 500 queries/second.

5. Conclusion and Future Work

In this work we have evaluated that workload pre assumption plays a significant role in the management of requests from users end. Further the processing time and hypervisor monitoring needs to be done in order to get more detailed and effective pre-assumption of workload of a virtual machine.

REFERENCE

[1] Andreas Berl, ErolGelenbe, Marco di Girolamo, Giovanni Giuliani, Hermann de Meer, Minh Quan Dang and Kostas Pentikousis ," Energy-Efficient Cloud Computing" in Advance Access publication on August 19, 2009, Published by Oxford University Press on behalf of The British Computer Society. | | [2] Anton Beloglazov, RajkumarBuyya, Young Choon Lee, and Albert Zomaya " A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems" in Proc. of the ACM/IEEE Conf. on Supercomputing (SC), Seattle, WA, 2005, pp. 1–9. | | [3] Anton Beloglazov, JemalAbawajy, RajkumarBuyya ," Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing" Proc. of the ACM/IEEE Conf. on Supercomputing (SC), Pittsburgh, PA, 2004, pp. 47–58. | | [4] DeepthiDharwar, Srivatsa S. Bhat, VaidyanathanSrinivasan, DipankarSarma, Pradipta Kumar Banerjee, "Approaches towards energy-efficiency in the cloud for emerging markets"inProceedings of twenty-first ACM SIGOPS symposium on Operating systems principles. ACM, 2007, p. 278. | | [5] Francis Owusu, Colin Pattinson, "The current state of understanding of the energy efficiency of cloud computing" on 11th International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2012, pp. 1948 –1953. | | [6] Kurt L. Keville, RohanGarg, David J. Yates, KapilArya, Gene Cooperman, "Towards Fault-Tolerant Energy-Efficient High Performance Computing in the Cloud", International Conference on Cluster Computing, 2012 IEEE, pp. 622-626. | | [7] Kejiang Ye, Xiaohong Jiang, Dawei Huang, Jianhai Chen, Bei Wang "Live Migration of Multiple Virtual Machines with Resource Reservation in Cloud Computing Environments" IEEE, vol. 4, pp. 422–431, 2011. | | [8] Kejiang Ye, Dawei Huang, Xiaohong Jiang, Huajun Chen, Shuang Wu, "Virtual Machine Based Energy-Efficient Data Center Architecture for Cloud Computing: A performance Perspective" in 2010 IEEE/ACM International Conference on Green Computing and Communications & 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing, vol. 6161, pp. 172–182. | | [9] Liang Zhou, BaoyuZheng, Jingwu Cui, and Sulan Tang, "Toward Green Service in Cloud: From the Perspective of Scheduling" in Proc. of the International Conference on Computing, Networking and Communications, Big Sky, MT, 2009, pp. 1–14. | | | | [10] Pablo Graubner, Matthias Schmidt and Bernd Freisleben, "Energy-efficient Virtual Machine Consolidation for Cloud Computing"inInformation and Communication on Technology Springer, 2011, pp. 1–9. | | [11] Qi Zhang, Lu Cheng, RaoufBoutaba, "Cloud computing: state-of-the-art and research challenges" J Internet ServAppl (2010) 1: 7–18 | | [12] R.Karthikeyan , P.C hitra, "Novel Heuristics Energy Efficiency Approach for Cloud Data Center" in 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies, pp. 210-216 | | [13] Yamini, "Power Management in Cloud Computing Using Green Algorithm", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), pp.120-127. | | [14] Ziming Zhang, Qiang Guan and Song Fu, "An Adaptive Power Management Framework for Autonomic Resource Configuration in Cloud Computing Infrastructures Performance Computing and Communications Conference (IPCCC), 2012 IEEE 31st International conference, pp. 51-60. | |