VM Consolidation or Placement using Utilization Prediction Model and Scheduling Algorithms

Surya Teja Marella, Sai Pedavalli, T Brahma Reddy, A Rama Krishna, Sk. Hasane Ahammad

ABSTRACT: ***Among the challenges faced by most of the business CIOs and IT managers now-a-days, among them, mostly are profitable or efficient utilization of the IT infrastructure, degree of the responsive-ness in urging for new initiatives of business and is much flexible in acclimatizing to changes in the organisation. The continued worries of meeting the IT budget as well as the stringent regulatory requirements are also the main challenges faced by the IT industry. Virtualisation, a technological innovation, helps a lot in deploying creative solutions to these business challenges. Data centre virtualisation is defined as the process of developing, deploying and designing a data centre on virtualization and other computing technologies. It helps to virtualize physical servers in a data centre and does this along with different aspects like storage devices, networking devices and other infrastructure and equipment. [1] Data centre virtualization usually produces a collocated and virtualized cloud data centre. By research, we know that, consolidation of the VMs helps to optimise or in fact reduce the usage of resources thereby reducing unnecessary energy consumption in a cloud data centre. The VM placement has a vital impact in the consolidation of the VMs. The specialists have created many different calculations or algorithms for VM consolidation considering the efficient energy utilization. However, these algorithms lack the use of exploitation mechanism efficiently and most of them focussed on number of physical machines (PMs) minimization and neglected future resource demands. Due to this, there were unnecessary VM migrations done [2]. Moreover, it violated most of the rules of the Service Level Agreement (also abbreviated as SLA) in data centres. In order to resolve this issue, this paper proposes an approach via VM consolidation that considers both, current as well as future utilization of resources. There are two solutions provide where one uses a model based on regression algorithm [3] in order to approximate and predict the CPU and memory utilizations of VMs and PMs and other uses two separate designed algorithms specifically designed to schedule VM in multi-tenant data centers. In order to attain the analysis and performance part, real workload traces like PlanetLab and Google cluster [4] are used. The results achieved via the proposed solutions showed betterment over the heuristic and meta-heuristic algorithms used here in reducing particularly the consumption of energy, the rate of migrations of VM and also the reduction of number of violations of SLA.***

INTRODUCTION:

The term virtualization basically describes the detachment of an asset or essentially a request for a particular service by the client from the delivery of physical resources of that service. Virtualization techniques are applied to various layers of IT like including services in form of server hardware, networks, storage, OS and other applications. This virtual infrastructure serves as a layer between networking, computing and storage and hardware and the applications running on it. This enables the administrators to manage the resources that are pooled of the enterprise thereby, making the IT managers more responsive to dynamic organizational needs. Now-a-days, cloud computing is a growing innovation, which can be used to provide various services in form of resources to perform complicated tasks. These complicated tasks are done with the help of data centres. Data centres help in this case by providing various resources like CPU, storage, network, bandwidth and memory, which gradually has resulted in the increase in the number of data centers in the world. As due to its extensive use, they consume large amount of energy for performing the operations and thereby leading to high operation costs and increase in usage of electricity. Resources are one of the primary causes for the consumption of power [5] in data centers. Energy consumption in data centers can be considered to the usage of resource. Excessive consumption of energy by data centers results in excessive increase in the power bills. Hence, the efficiency of energy of the data centers have to be increased. In past, there were many solutions proposed for this problem, but most of them were primarily focussed on reducing the number of PMs of present, neglecting the future demands of CPU and memory utilizations of VM and PM. To do this, various algorithms can be used which can be used to optimize the usage of the resources and help in reducing the consumption of energy in a data centre. Some of them are:

Solution 1: The VM consolidation or placement of VM in data centres done by developing a prediction algorithm basically a regression algorithm [6] to predict and approximate the CPU and memory utilizations of the future and act and change according to it.

Solution 2: designing algorithms for example, VM scheduling algorithm for minimum energy (MIN ES) and

VM scheduling algorithm for minimum communication (MIN CS) which can be used to compare and yield scheduling that can be used by the data centers to reduce the consumption of energy.

Both solutions serve the same purpose and yield results that lead to less power consumption. The working is mentioned below;

Revised Manuscript Received on April 12, 2019.
Surya Teja Marella, K L University, Guntur, AP, India
Sai Pedavalli, K L University, Guntur, AP, India
T Brahma Reddy, K L University, Guntur, AP, India
A Rama Krishna, K L University, Guntur, AP, India
Sk. Hasane Ahammad, K L University, Guntur, AP, India

A) The first solution i.e, using a regression algorithm to predict and basically approximate the CPU and memory utilizations. Here the model named, Utilization Prediction Model used, is an important component for allocation of resource as well as scheduling of job problems in virtualized servers and clouds. It is primarily used by a resource manager in order to balance the excess load of the work among the working nodes of the system, thereby, increasing the resource usage efficiency. In virtual machine distribution task here in our case, the accurate prediction of CPU utilization as well as memory utilization, allows to relocate at least one virtual machines, maintaining a strategic distance from the abundance physical machines' over-burden.

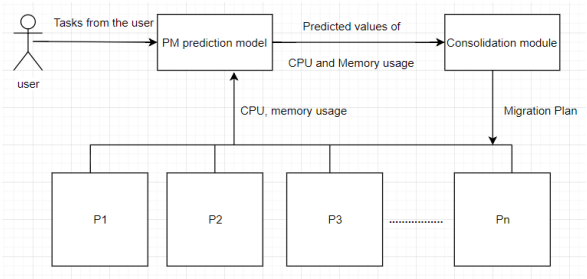

Figure 1 showing the architecture of the solution presented

In this approach, a particular data centre is considered that comprises of n heterogeneous PMs, P = hp1,..,pni, as observed from the architecture diagram given above, termed figure1.

Here, each PM consists of D type or D number of resources, for example, CPU, memory, arrange I/O and capacity limit. The user can allocate multiple VMs to each PM via Virtual Machine Monitor (VMM). In the proposed model, the allocation is done based on the Best Fit Decreasing (BFD) algorithm. At a time, the user submits his/her request for allocating of m VMs, V = hv1,...,vmi, which are allocated to the respective PMs. Multiple physical machines as given in the figure, form a cluster. Each physical machine consists of VMM and assigned number of VMs. To optimize the VM placement according to a given work load quantity, an efficient algorithm must be used to the proposed algorithm which can be applied periodically as utilization of VMs and PMs vary overtime.

The UP-VMC algorithm here to attain this result. The proposed system consists of 2 agents namely,

(1) Local Agents that are fully distributed (LAs) in PMs,

(2) a master node residing Global Agent (GA).

figure 2, below shows the architecture and working under each PM,

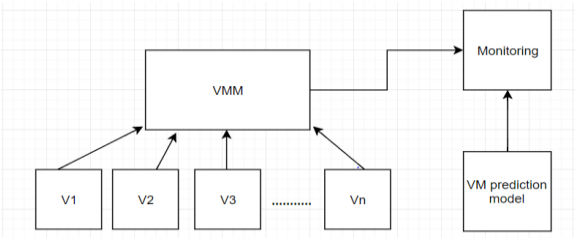

Figure 2 showing the architecture of each PM

The task is performed as following,

a) Each LA keeps a check or tracks the current asset usage of all VMs in a PM in a periodical sense. Then it computes the utilizations of all VMs in a PM of future using the prediction model based on regression algorithm.

b) The GA gathers all the required data from nearby agents local to system in order to keep up the general perspective on present and future asset use of the respective VMs.

c) Then, the GA checks and clears the VM placement by using the UP-VMC algorithm and then, it sends necessary required commands to VMMs to perform placement of VM. The commands given indicate which Virtual Machines from a source PM migrate to which VM.

d) The VMMs carries out the actual migration process from the source to destination after receiving the commands from the GA.

In brief the working is, UPVMC constructs a process for consolidation of VM as a multiple-objective well known process of vector bin packing where it takes utilization of resources both current and future in order to consolidate the VMs into the least number of physical machines in active state (PMs) and the utilization of resource of future is predicted by using a model based on regression prediction algorithm. Here, in order to get more better results out of the model, and to achieve QoS and to reduce the migrations, a VM allocating procedure primarily focussed on prediction models is used.

B) The second solution is designing algorithms i.e, algorithm for scheduling VM for reduced energy (MIN ES) and algorithm for scheduling VM for reduced communication (MIN CS) and comparing them with various predefined algorithms. Here, the VM scheduling problem for minimum energy is modelled as a much-used programming problem for integer and is shown to be NP-hard, the algorithms are designed. MIN ES takes care of extra powering up of unwanted servers and gadgets by setting VMs which have comparable contract end times. MIN CS is dependable to designate VMs of same occupant on same server so as to diminish the general vitality utilization on both system and servers. After developing the algorithms, they are compared to the integer programming optimal solution in a smaller scale level cluster and then compared to algorithms designed before, like mPP and native greedy placement algorithm and analyse the results leading to predicting the effectiveness of the designed algorithms.

PROCESS INTRODUCTION FOR EXPERIMENT OR ANALYSIS:

Solution 1:

The main idea of the solution is to present a approach for utilization prediction-awareness for the consolidation of VM problem. UP-VMC is used here because it performs consolidation of VM as a bi-dimension vector packing problem and considers memory utilization in addition to CPU utilization. With this feature, it can identify the causes

of SLA violations more easily and help in preventing them in the future. To consider current and future resource utilization, two regression-based prediction models i.e, linear [8] and k-nearest neighbour are proposed. Since the main focus is on small term prediction of the load due to the dynamic change in workload, these prediction models are used. Here, LR utilizes the authentic asset usage information and finds out a linear function and K-NNR finds the future utilization by selecting a data set having a local average. Then, it is given in form of k which is the number of nearest neighbour samples near to the new obtained sample. The k-value best, is computed by the method called leave-one-out cross-validation, a fundamental incentive among one and the quantity of tests. It estimates the accuracy of prediction by summing of square of residuals for each k and then chooses the value best for k having the minimum total residual. These are the reasons due to which the prediction models are selected. The evaluation of UP-VMC is done on a simulated large-scale data centre using the real traces systems PlanetLab and Google workloads. The main reason of using utilization prediction model for VM consolidation is shown by comparison with Sercon [6] and a modified version of FF and BF. In terms of experimental analysis, the analysis of UP-VMC is done by checking the impact of n different VM selection policies and then the prediction performance is evaluated and in comparison, the experimental results with the algorithms for benchmark is done. The main criteria of this approach are to provide a guarantee that basic SLAs are not violated, minimizing the physical servers usage and minimizing the number of VM migrations. So, these criteria are to be fulfilled primarily in any given scenario of the analysis. Here the VM selection policies [7] are Minimum Migration Time (MMT), Maximum Load (MaxL- VM with max load) and Minimum Load (MinL- VM with min load) and the baseline algorithms include Modified Best Fit Decreasing (MBFD), Modified First Fit Decreasing (MFFD), PM Utilization Prediction-aware VM Consolidation (PUP-VMC), VM Utilization Prediction-aware VM Consolidation (VUP-VMC), Sercon and ACS-based VM Consolidation (ACS-VMC).

Solution 2:

To take on this problem, we need to know about the energy models and how much energy is consumed by the data centres. Mostly, in data centres, emerges essentially from perspectives like network devices, framework and servers (physical) and utility costs of electricals where, servers take up the greater part of the vitality. In this way, lessening the vitality disseminated in type of heat consumption by servers and devices for systems administration will likewise reduce the vitality devoured by different frameworks like cooling frameworks and power draw. In this arrangement, since the issue is characterized NP hard and arrangements given by programming solvers take quite a while to implement and provide result, algorithms are designed to save energy done by powering down the servers. In brief, 2 models are to be considered, one being the Physical server energy consumption model and the Switch consumption of energy model as the other. For the physical server, most of the energy consumed is by the CPU, memory, stockpiling frameworks, the quantity of dynamic system cards and so forth. We also know that energy consumption fluctuates depending on the server or CPU load. So, considering the CPU to be in normalised state, the energy consumed by the server can be given by, if p(idle) is energy consumed by the server when idle and p(busy) is the consumption of energy by the server when busy,

$$p(u)= p(idle)+(p(busy)-p(idle)) *u \text{ where } u \in [0,1]$$

Whereas, in Switch energy consumption model, configuration of equipment and flowing of traffic through the gadgets both influences the vitality utilization. As we probably are aware, the majority of the systems administration gadgets utilized in current server farms are fundamentally product switches, here vitality utilization model of switches is just considered. As far as sparing vitality, it is sensible to kill unused ports and as most product switches utilized now-a-days are ToR switches outfitted with a solitary line-card, the vitality utilization of a switch can be given as,

$$p(s)= p(i) + p(p) *s \quad \text{where } p(i)=p(idle\_s) \text{ and } p(p)=p(port)$$

where, p(idle_s) is basically the vitality devoured as the switch is on with every one of the ports in debilitated state, P(port) is the vitality devoured by a solitary port, Also here, s is the quantity of dynamic ports present on the switch. By this, we become acquainted with how much vitality is being devoured by the server farms so as to utilize it to make an enhanced answer for diminish the utilization of vitality. For the advancement of improved arrangement, we have to initially ensure that the base vitality VM booking issue is a whole number programming issue and turn out to be NP-hard. As we realize that the general vitality utilization can be decreased by limiting the quantity of dynamic physical servers consequently demonstrating it as a bin packing issue, which is basically NP- hard. Since, the problem being NP-hard is proved, the first case test is done i.e, trying to get the optimal solution using software solvers for example, Gurobi. Gurobi software solver is basically known for providing of optimised results in integer linear programming problems. After getting the required values like time required to linearize the problem etc are taken out, the two algorithms are designed namely, MIN ES and MIN CS.

In MIN ES algorithm, the virtual machines are scheduled one after other in order to reduce the amount of energy consumed per each virtual machine. In every iterative step, allocation of virtual machine is done according to the performance and stability of the system. Due to it being time slotted, the allocation process of the requested VMs carries out at the initial step of the next immediate slot. But, the problem of MinES is that the scheduling of the virtual machines are done one by one, it requires more time if the problem's dimension is considered. Hence, MinCS in this scenario is developed which schedules the set of VMs that have a place with a similar process on those close assets close to it physically, i.e., which are on a similar server, or in a similar rack, etc.

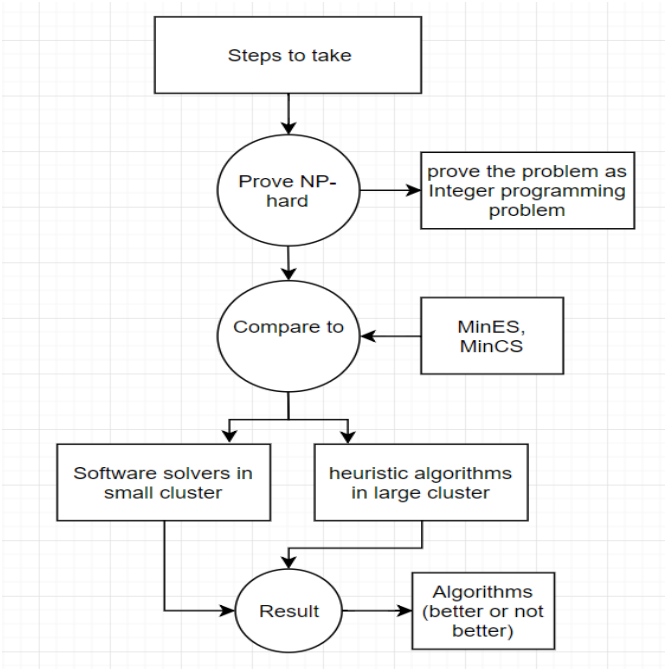

Figure 4 representing the architecture of the solution 2

From the figure 4 above, it is stated that after developing the algorithms, they are evaluated in terms of performance under different scenarios and are compared to those obtained by the software solvers like Gurobi in small clusters and then, compared to mPP and best first fit approach algorithms in large clusters. At last, the sensitivity measure of MIN ES for inaccuracy of contact end times is also tested.

CONCLUSION OR EXPERIMENTAL RESULTS OBSERVATION:

Solution 1:

In terms of VM selection policies, the UP-VMC helped in reducing the consumption of energy, violations of SLA and also reduced the number of migrations significantly when it used MMT policy and can perform even better when K-NNR prediction model is used due to its accuracy in measuring the utilization of resources particularly of PMs and VMs than LR. In terms of performance of prediction models which include KNN and Linear Regression models, the RMSE of KNN is lower than LR and the Mean Absolute Percentage Error (MAPE) is comparatively lower as compared to LR indicating that K-NN can be used to compute the utilization more accurately by giving out less residuals. In the comparison of Baseline algorithms with UP-VMC, it was inferred that, in case of CPU threshold values, in terms of SLA violations, UP-VMC compared to other algorithms reduced the SLA violations percentage rate more efficiently. The VM consolidation approach with UP-VMC thereby reduces the energy consumption compared to MFFD and other algorithms. This is due to the proposed model here, minimises the number of PMs that are active by packing the virtual machines into the most-loaded PMs. When VM migration case considered, the total number of migrations of VM in the UP-VMC performs much better than the benchmark algorithms because of the utilization prediction done beforehand, and therefore decrease in the number of migrations in VM. Hence, its outright to say that The VM consolidation approach with UP-VMC is much more stable and satisfies all the three criteria mentioned above and leads to less energy consumption.

Solution 2:

After getting the qualities from the product solver [10] (ex. Gurobi), it is seen that, the un ravelling progresses toward becoming tedious if extensive number of bilinear imperatives with exceptionally expansive number of new factors and direct requirements are considered, state, a group of 70 servers with an occupant solicitation of 200 VMs. This much measure of time utilization isn't impressive on account of server farms. In this way, the calculations structured i.e, MIN CS, MIN ES when contrasted with the estimations of the as of now created estimations of the product solver, would be wise to results as far as vitality utilization and time it required to perform. This is demonstrated by considering n sorts of VMs with n distinctive VM arrangements for the test, by utilizing Google Job Trace and grabbing n timespan from the follow records, in MIN ES, the VMs were booked to a particular host server which had comparatively less or in fact negligible vitality utilization leading to the most modest number of dynamic servers stamping less utilization of vitality and for MIN CS, as mentioned VMs are distributed under same switch, it turns on the most modest number of switches. On the off chance that the standard deviation is watched for both the cases, it is of a lot littler incentive than contrasted with the normal and choices suggesting better outcomes from these calculations. Next, the correlation in little bunch with the result of solver within a group with state, 50 servers(physical) are finished. When utilizing Gurobi programming solver, it was seen that, the execution of MIN CS and MIN ES is similarly much stable when VM qualities are liable to change. On account of vast bunches, with state, 500 physical servers, the calculations are contrasted and the produced consequences of past heuristic calculations like mPP and best initially fit. As far as performance [11], on a normal, MIN ES and MIN CS accomplish much better outcomes with section level arrangement of the servers and conversely, gives much better execution in a few different situations. The calculations, MIN ES and MIN CS fared well because of utilizing of transmission capacity assets adequately when contrasted with different calculations which saved transfer speed in a guileless manner. Going to the terms of affectability, the calculations are generally delicate to the blunder in foreseeing contract end dates for instance, bigger variety of VM spans. In this way, the sending occasion of the calculations, a great component is required to estimated or foresee and deal with inhabitants' agreements end dates.

REFERENCES:

1. K. Thirupathi Rao and Sai Kiran2 L.S.S.Reddy, “Energy Efficiency in Datacenters through Virtualization: A Case Study “, pp. Vol. 10 Issue 3 (Ver 1.0), April 2010.
2. W.Voorsluys, J.Broberg, S.Venugopal, and R.Buyya,“Cost of virtual machine live migration in clouds : A performance evaluation,” in 1st International Conference on Cloud

Computing (CloudCom). Springer-Verlag, 2009, pp. 254–265
3. F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, and H. Tenhunen, “Utilization prediction aware vm consolidation approach for green cloud computing,” in Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on, 2015, pp. 381–388.
4. K. Park and V. S. Pai, “Comon: A mostly-scalable monitoring system for planetlab,” vol. 40, 2006, pp. 65–74 and “Traces of google workloads,” 2015, http://code.google.com/p/ googleclusterdata/.
5. F.Farahnakian, P.Liljeberg,andJ.Plosila,“Energy-efficientvirtual machines consolidation in cloud data centers using reinforcement learning,” in 22nd Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), 2014, pp. 500–507.
6. A. Murtazaev and S. Oh, “Sercon: Server consolidation algorithm using live migration of virtual machines for green computing,” vol. 28, 2011, pp. 212–231.
7. M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, “Resource allocation algorithms for virtualized service hosting platforms,” J. Parallel Distrib. Comput., vol. 70, no. 9, pp. 962–974, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc. 2010.05.006
8. F. Farahnakian, P. Liljeberg, and J. Plosila, “LiRCUP: Linear regression based CPU usage prediction algorithm for live migration of virtual machines in data centers,” in Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on, 2013, pp. 357–364.
9. J. Han, Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.
10. Gurobi optimizer reference manual (accessed on 29/05/2014). [Online]. Available: http://www.gurobi.com/documentation/5. 6/reference-manual/
11. R. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. 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 (SPE), vol.41, pp. 23 – 50, 2011.