INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1459)
Available online at: https://www.ijariit.com

Resource demand prediction for minimizing power consumption at data centers

Gaurav Thakur
gauravthakur573@gmail.com
Thapar Institute of Engineering and Technology, Patiala, Punjab

ABSTRACT

Technical progresses in servers, systems and capacity virtualization are empowering the production of resource pools of servers that allows various application workloads to share each server in the pool. This approach proposes and assesses the parts of a capacity management process for creating an efficient utilization of such pools and also facilitate huge quantity of business administrations. The objective of our approach is to give a capacity management procedure to resource pools that let capacity organizers coordinate free market activity for resource limit in an given interval of time. In this approachwe will describe the workloads of big business applications to pick up experiences for their conduct. Inthis paper we will follow a trace-based approach for capacity management that relies upon definition forrequired capacity and portrayal of workload requestdesigns. The exactness of scope quantificationexpectations relies upon our capacity to describe workload request examples, to perceive patterns for expected changes in future requests and to reflect business conjectures for any of the sudden changes in future requests. A contextual analysis with 6 months of information which speaks to the asset utilization of 159 workloads in a venture server farm shows the adequacy of the proposed limit administration handle. Our results and conclusion will show that whenever we will use 8 processor systems, we will predict the exact future per- server required capacity to 1 processor every 98 percentof the time. This approach will enable 38 percent reduction inprocessor utilization as compared to today's current best practice for workload placement. This knowledge will helpresourcepool operators to decide the best capacity for their server pools.

Keywords: *Workload, Capacity Management, Sliding window, Resource Pool.*

1. INTRODUCTION

In the earlier times data centers were made up of small numbers of very large mainframe computers in which eachcomputer was hosting a several number of application workloads with many users. Capacity planning workershelped to ensure that sufficient capacity was available every time, as it was needed. With the advancements in distributed computing new application workloads were typically assigned to their own small servers. The single sizefor all philosophies used for permanently allocating data path resources in today's superscalar CPUs to maximizeperformance across a wide range of applications resulted in the over commitment of resources in general. To reduce power consumption in the data path, the resource allocationscan be dynamically adjusted based on the demands of applications [1]. The incremental cost of capacity from small servers was quite less in price than that of the increased cost of capacity on mainframes. Capacity plannerswould often predict an application's workload demands two years in advance and pre-provision a new server with sufficient capacity so that the workload could be managed efficiently. However, the explosive growth in both hardwarecomputing and internet computing has led to server sprawl in data centers. Hardware data centers are typically full of large numbers of lightly utilized servers that incur high price of ownership including facilities price, as rent andpower for computing and cooling, high software licensing price and high price for human management activities. An online energy saving framework for data center by using reinforcement learning and intelligent feedback were also proposed in this referenced paper in order to realize energy saving from the view of software optimization. And theresults of the simulation experiment on CloudSim platform showed the proposed intelligent framework and online mapping methodology can effectively improve the average resource utilization of data centers and reduce the energy consumption while guaranteeing real-time performance [2]. Many enterprises have currently started to endeavor the resource pools of servers bolstered by virtualization components that empower numerous application workloads to be facilitated on each and every server. The essential inspiration for enterprises to embrace such innovations is expanded adaptability, the capacity to rapidly re-purposeserver ability to better address the issues of use workload proprietors, and to decrease general expenses of possession. Tragically, the intricacy of these conditions introduces extra

administration challenges. There are numerous workloads,a limited number can be facilitated by every server andevery workload has some capacity requirements that muchof the time changes in view of the required business needs. Capacity management techniques are not yet accessible to oversee such pools in a financially way. Researchers also proposed the multiplexing benefits achievable in on- demand data centers. They used real web workloads to study the effect of various factors on these benefits. The results demonstrated that fine-grained multiplexing at short time scales of the order of seconds to a few minutes combined with fractional server allocation leads tosubstantial multiplexing gains over coarse-grained reallocation [3].

The main aim of this paper is to provide a capacity management process for resource pools which will help the capacity planners to match the supply and demand for the resource capacity for a given interval of time. When overseeing gatherings of resources, there are a considerable measure of capacity management issues that should be routed to guarantee that the resources are utilized successfully. For example: what amount of limit will it take to support the current workloads, what are the different workloads that are assigned out to every resource,what is the effect on the scheduler workload execution arrangement or the strategies representing the trade, inwhat capacity the workloads should be apportioned to make working hours or strategy settings more powerful, what should be done when a resource does not requireenough ability to address the issues of the workload, what number of resources should be required in case of an arranging skyline. For all these issues, we have outlined a capacity management approach. The outcomes demonstrate that during the combination of 8 processor frameworks, we anticipated the capacity required to keep up 1 processor time of 98 percent of the time when the gauge capacity was 6 weeks required. The server afterwards was permitting a 38 percent reduction in CPU usage contrasted with current accepted procedures for the position of workloads.

2. PROBLEM STATEMENT

Capacity management and demand estimating capacities inserver farms have turned out to be one of the key viewpoints in limiting power and power utilization in server farms. It is depicted as a procedure of capacitymanagement of asset pools. The emphasis is on the meanings of required limit and the workload of specializedforecast. proposed the design and implementation of the PRESS system, a novel predictive resource scaling system for cloud systems. PRESS employs light-weight signal processing and statistical methods to predict dynamic resource demands of black box applications and performs elastic VM resource scaling based on the prediction results [4]. When overseeing gatherings of resources, thereare several measures of capacity management issues that should be routed to confirmthat the resources are utilized successfully. For example: what amount of limit will it take to support the current workloads, what are the different workloads that are assigned out to every resource,what is the effect on the scheduler workload execution arrangement or the strategies representing the trade, in what capacity the workloads should be apportioned to make working hours or strategy settings more powerful, what should be done when a resource does not require enough ability to address the issues of the workload, what number of resources should be required in case of an arrangingskyline.

We took note of Z imperatives of the frame [5](Sn, Un, Pn), n = 1 ... Z, where:

- Sn be a fixed window with intervals of one hour sothat Sn = Cn.d
- Un be a point of confinement in the rate ofcapacity use for window.
- Pn be the rate of windows that are permitted tohave applications past the UI (user interface).

Using these variables we calculate the value of Y andstudy the trends.

3. SOLUTION METHODOLOGY

It is must to define the required capacity. The required capacity is the minimum capacity needed to meet the resource needs for the workloads on a server resource. Service workload placement: the workload collocation service we use recommends the placement of application workloads between servers in the pool to reduce the numberof servers that are used to balance workloads or servers. Theservice implements a trace-based approach to characterize demand for resources and recommend solutions. Basically, each workload is characterized using a trace-time-varying demands for key capacity attributes such as CPU usage and memory usage. The workload colocation service includes greedy algorithms to consolidate workloads into a small set of servers to balance workloads on a fixed number of servers. It also includes a search optimization based on the genetic algorithm designed to improve the intensive solutions. In each case, the algorithms simulate multi-task scenarios. Each scenario considers zero location or multiple workloads on each server**.** The aggregate demand workloads assigned to a server are characterized by using a trace which is the sum of the demand vary depending on theworkload per unit time. Service recommends the best workload location we can find on all servers or for consolidation or load leveling. Finally, the service accepts the limitations and the workload sites that include affinity between workloads, for example, workloads should or should not be placed on the same physical server and affinity between workloads and a list of one or more specific servers to three objectives:

- Helps in determining if requests for a workload changes significantly over time;
- It supports the production of synthetic traces representing the future demands of each workload to support capacity planning activities;
- provides a practical model that can be used to support foresight exercises. The service implements model discovery of the techniques described in our implementation part.

If no resources were capable to support the resulting capacity requirements, we may either try to balance the larger workloads, adjust the quality of the workload of the service requirements or combine these two approaches."Add workload" indicates whether a location can be found for a new Workload.We set out to the required capacity as the most secure confinement is fulfilled. For instance, either: S0 = 30 minutes, U0 and P0 = 100% = 100% and s1 = 5 minutes = 100% U1 and P1 = 98%. The main limitation permits instinctive necessity that limit request ought not surpass control for a very long time, for instance 60 minutes. The second limitation confines the recurrence at which the demand is approved to surpass the offer on a shorter time scale, example 2 minutes. This restrains the effect of overabundance saves the workload applications ina shorter time. The count of Y happens with the concentration by asking for the unit amid the workload situation works out. Give N, a chance to be the workload number follows to the review. Each track has W weeks of perceptions with T intervals every day, measured eachmoment. For intervals of 5 minutes

we have T = 288intervals for every day. We deal out every interval by a record $1 \leq t \leq T$. Every day x of the seven days of the week has a remark for every interim t. Every perception has a deliberate an incentive for every capacity trait considered in the examination. To build up Y, consider a property thathas a capacity point of confinement of the units asking for R. Let D, x, t be the entirety of the quality necessities by the workloads for one week w, x the day and the interim t .We characterize the deliberate an incentive for Y as taken after [5].

$$Y = W \min w = 1 \; T \min t = 1 \; \Sigma 7x = 1 \min (D, x, t, R) \; \Sigma 7x = 1 \; D, x, t.$$

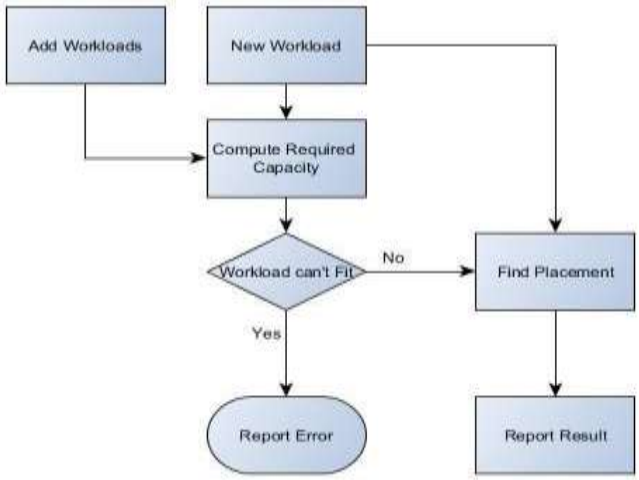

Figure - 1: Flowchart

Thus, Y is accounted for as the base likelihood of access to resources got on week one of the T intervals every day.

The primary favorable position of this approach by the sliding window approach is that the unsatisfied request in a hole is displayed as repeated for the following interval. Contrasted with the straightforward sliding windowapproach, the over- burden conditions are characterized in connection to the application units that are not fulfilled rather than various adjacent over-burden periods. It has been shown that such a value of Y can be used to determineWorkload Manager Scheduler parameters for workload managers that support two service priorities.

Simple Sliding Window

Our simple sliding window definition for required capacity defines the minimum needed capacity for a capacity attributeas the minimum number of units of capacity needed so that demand for capacity doesn't exceed the supply of capacity for more than an overload S as expressed in minutes. The current energy and environmental cost trends of datacenters are unsustainable. It is critically important to develop datacenter-wide power and thermal management (PTM) solutions that improve the energy efficiency of the datacenters [6]. Researchers also proposed the data traffic for wireless metering in smart grid is considered. A dynamicspectrum allocation scheme is proposed for the power load prediction [7]. If a unit of demand is not satisfied because demand exceeds supply, i.e., an overload, then that unit of demand propagates forward in time until there is available capacity to satisfy the demand. For a performance critical environment S may be chosen as 0, which means all demandmust always be satisfied.

Per Unit of Demand Sliding Window:

Researchers also proposed ACES, an automated server provisioning system that aims to meet workload demandwhile minimizing energy consumption in data centers. To perform energy-aware server provisioning, ACES faces three key tradeoffs between cost, performance, and reliability [8]. We now consider our preferred definition for required capacity that also uses the sliding window. With this approach,we find the smallest capacity such that no unit of demand is propagated from one interval to the next for more than S minutes. We specify a resource accessprobability Y that a unit of demand will be satisfied upon demand, and hence not propagated.

4. IMPLEMENTATION

With the use of data set: To evaluate the effectiveness of ourmethods and processes we obtained six months of workload trace data for 139 workloads from a data center. The data center specializes in hosting enterprise applications such as customer relationship management services and supply chain management services for small and medium sized businesses. Cloud computing allows business customers to scale up and down their resource usage based on needs.Many of the touted gains in the cloud model come from resource multiplexing through virtualization technology [9].

Researchers also proposed the novel EHEROS scheduler which is an expansion of the state-of-art network and energy aware schedulers. EHEROS is particularly intended to operate in heterogeneous systems. It is based on the aggregation of utilization and network links [10]. Each workload was hosted on its own server so that we coulduse the resource demand measurements for any server to characterize the workload's demand trace. Each trace describes resource usage, processor demands, as measured during time interval of 5 minutes. The fixed window approach typically results in higher required capacity values, i.e. more over provisioning, compared to sliding window technique. This is because the fixed window is more constraining than the sliding

window, when P = 100% all demands must be satisfied within each fixed window. Neither the percentile nor the fixed window approach characterizes the impact of over-load epochs on units of demand. However, such a characterization is helpful and necessary when deciding workload manager scheduler settings.

Table - 1: Implementation Table

Workload	**Simple Sliding Window (Requirement capacity values)**	**100-p**	**99-p**	**95-p**	**Fixed window (Requirement capacity values)**	**Per unit demand and sliding window**
w1	850	950	770	625	860	840
w2	540	590	415	300	560	540
w3	80	80	80	75	80	80
w4	35	55	45	12	40	25
w1-w4	1200	1400	1100	850	1300	1200
w5	470	540	410	320	455	450
w6	190	190	190	185	185	190
w7	45	50	40	30	45	45
w8	80	130	65	55	85	75
w5-w8	590	680	530	435	610	560

As expected, the per demand unit window approach typically ends up in a rather lower required capacity value, i.e., more efficient capacity usage, than the easy window definition because it permits longer periods of overload as long as no unit of demand remains unsatisfied after s = half-hour. We conducted a walk- forward test over the six months of knowledge to emulate how well our capacity management processwould have served the info center for the six months.

- Starting with the primary week, a window with w weeks of knowledge is employed to recommend a consolidated configuration C1 i.e., each workload is assigned to a selected server, for the system. The configuration reportsexpected required capacity values for every server within the configuration.
- The next y weeks of information are then simulated with relation to C1. This simulation gives the particular required capacity for thefollowing y weeks. The difference between a server's estimated and actual required capacity gives absolutely the error for the estimate of required capacity. The negative errors reflect"under-estimated" capacity while the positiveerrors correspond to "over-estimated" capacity. We use a special CDF that reflects both formsof errors for the walk-forward test.
- The steps within the walk-forward test are repeated iteratively with w weeks of knowledgebut now starting with weeks 2, 3, and so on.
- Let i be the step number within the walk- forward test. Step i computes a brand new configuration C i and a brand new set of differences between estimated and actual required capacity values for every server. error(y axis)

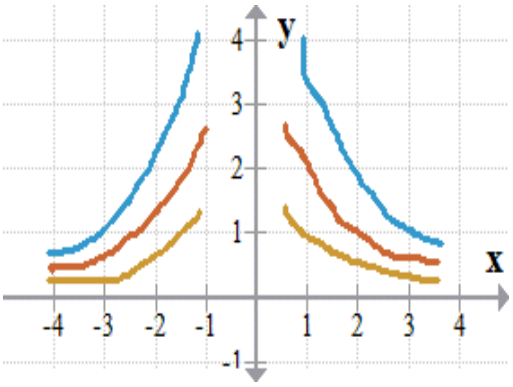

Figure - 2: Variation from actual value

30 Hours dataset
1 week dataset
6 weeks dataset

By exploring through the graph, we will presume that after we compute workload seeable of 30 hours information, there's more errors. As we move towards more weeks then errors come less. this best practice forserver consolidation in industry decides the height request of each application workload and consolidates the applications to few servers with the top goal thatthe overall of peak requests for applications at every server isn't the maximum amount because the limit of the server. On the off chance that we recommend placements supported peak per-application demands then we require 238 processor servers. By using the trace-based method described during this paper we require only 15 servers. The approach enables a 38% reduction in processor usage as

compared to workload placement that stacks applications basedonly on their peak demands. The precision ofexpectations for required capacity recommends that such asset investment may be accomplished with less risk.

5. CONCLUSION

We describe a capability management method for resourcepools. the tactic depends on services that change and alter management for resource pool operators. we've a bent to targeted on definitions for needed capability and on a work- load demand prediction technique. A case studyexploited six months of information for 139 enterpriseapplications to measure the effectiveness of our strategies. The machine-controlled strategies expected (the needed) the desired (the desired) capability of servers hosting the workloads to among one processor out of eight ninety fifthof the time once predicting required capability 5 weeksinto the long run whereas reducing mixture processor necessities by thirty fifth while not vital risks. Such advance information will facilitate resource pool operators to return to a choice whether or to not order further capability for his or her pools.

The work demand prediction service depends on patternand trend recognition strategies. Our case study resultsshow that trend prediction are useful as long as we've aninclination to don't exaggerate however way into the longrun we've a bent to expect trends to continue. we've got abent to believe that job demand prediction strategies aradvantageous for a capability management method. Theyacknowledge odd patterns which might act within thelonger term, e.g., three day and five day patterns might act each fifteen days, and might facilitate to report once aworkload's demands seem to deviate from past behavior. Our future work includes: developing on-line auto-matedstrategies for watching and news unplanned changes tofigure characteristics; higher exploiting notions ofconfidence and risk regarding predictions for futureneeded capacity; and better act business fore-casts forgrow to be our approach. Additional work is additionallyrequired to characterize work demand patterns forenterprise services exposed as confidence and riskregarding predictions for future needed capacity; andbetter integration business fore-casts for turn into ourapproach. Additional work is additionally required tocharacterize work demand patterns for enterprise servicesexposed as internet services.

6. REFERENCES

[1] Dmitry Ponomarev, Gurhan Kucuk, and Kanad Ghose. 2006. Dynamic Resizing of Superscalar Datapath Components for Energy Efficiency. IEEE Trans. Comput. 55, 2 (February 2006), 199–213.

[2] Yuan, Jingling & Miao, Xuyang & Li, Lin & Jiang, Xing. (2013). An Online Energy Saving Resource Optimization Methodology for Data Center. Journal of Software. 8. 10.4304/jsw.8.8.1875-1880.

[3] Chandra, Abhishek & Goyal, Pawan & Shenoy,Prashant. (2003). Quantifying the benefits of resource multiplexing in on-demand data centers.

[4] Gong, Zhenhuan & Gu, Xiaohui & Wilkes, John. (2010). PRESS: PRedictive Elastic ReSource Scaling for cloud systems. 9 - 16. 10.1109/CNSM.2010.5691343.

[5] Gmach, D., Rolia, J.A., Cherkasova, L., & Kemper, A. (2007). Capacity Management and Demand Prediction for Next Generation Data Centers. IEEE International Conference on Web Services (ICWS 2007), 43-50.

[6] Pakbaznia, Ehsan & Ghasemazar, Mohammad & Pedram, Massoud. (2010). Temperature-aware dynamic resource provisioning in a power-optimized datacenter. Proceedings -Design, Automation and Test in Europe, DATE. 124-129. 10.1109/DATE.2010.5457223.

[7] Habault, Guillaume & Lefrançois, Maxime & Lemercier, François & Montavont, Nicolas & Chatzimisios, Periklis & Papadopoulos, Georgios Z.. (2017). Monitoring Traffic Optimization in Smart Grid. IEEE Transactions on Industrial Informatics. PP. 1-1. 10.1109/TII.2017.2742584.

[8] Guenter, Brian & Jain, Navendu & Williams, Clyde. (2011). Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning. Proceedings- IEEE INFOCOM. 1332 - 1340. 10.1109/INFCOM.2011.5934917.

[9] Xiao, Zhen & Song, Weijia & Chen, Qi. (2013). Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment. Parallel and Distributed Systems, IEEE Transactions on. 24. 1107-1117.10.1109/TPDS.2012.283.

[10] Yassa S, Chelouah R, Kadima H, Granado B. Multi- objective approach for energy-aware workflow schedulingin cloud computing environments. ScientificWorldJournal. 2013 Nov 4;2013:350934. doi: 10.1155/2013/350934. PMID: 24319361; PMCID: PMC3835373.

