INTERNATIONAL JOURNAL OF RESEARCH
In Advanced Engineering Technologies

Volume 5, Issue 2 NOV 2015

DYNAMIC HETEROGENEITY-AWARE RESOURCE MACHINERY FOR CLOUD STORAGE

[1] M. V. SUDHA RANI, [2] MR. B. BHARATH KUMAR

[1]M.Tech Student, Department of CSE.

sudhatharak@gmail.com

[2] Assistant Professor, Department of CSE.

bandlabharathkumar@gmail.com

Abstract*— *Data centers consume tremendous amounts of energy in terms of power distribution and cooling. ynamiccapacity provisioning could be a promising approach for reducing energy consumption by dynamically adjusting the number of active machines to match resource demands. However, despite in depth studies of the matter, existing solutions haven't totally thought of the non uniformity of each employment and machine hardware found in production environments. particularly, production information centers usually comprise heterogeneous machines with different capacities and energy consumption characteristics. Meanwhile, the assembly cloud workloads typically incorporates numerous applications with completely different priorities, performance and resource needs. Failure to consider the non uniformity of each machines and workloads can result in each sub-optimal energy-savings and long planning delays, attributable to incompatibility between employment needs and also the resources offered by the provisioned machines. to deal with this limitation, we have a tendency to gift Harmony, a Heterogeneity-Aware dynamic capacity provisioning theme for cloud information centers. Specifically, we have a tendency to initial use the K-means cluster formula to divide employment into distinct task categories with similar characteristics in terms of resource and performance requirements. Then we have a tendency to gift a method that dynamically adjusting the quantity of machines to attenuate total energy consumption and planning delay. Simulations mistreatment traces from a Google's cipher cluster demonstrate Harmony will cut back energy by twenty eight % compared to heterogeneity-oblivious solutions.

I. INTRODUCTION

DATA centers have recently gained significant quality as an economical platform for hosting large-scale service applications. whereas massive data centers get pleasure from economies of scale by amortizing long-term capital investments over an oversized variety of machines, they conjointly incur tremendous energy prices in terms of power distribution and cooling. as an example it has been reportable that energy-related prices account

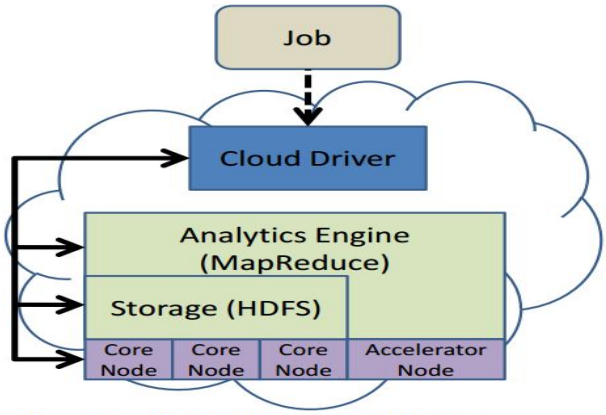
for approximately twelve p.c of overall information center expenditures. for big corporations like Google, a 3 p.c reduction in energy price will translate to over 1,000,000 greenbacks in price savings . At the same time, governmental agencies still implement and rules to push energy efficient computing . As a result, reducing energy consumption has become a primary concern for today's information center operators. In recent years, there has been in depth analysis on up information center

energy potency . One promising technique that has received important attention is dynamic capability provisioning (DCP). The goal of this technique is to dynamically change the quantity of active machines in an exceedingly information center so as to cut back energy consumption whereas meeting the service level

objectives (SLOs) of workloads within the context of workload planning in information centers, a metric of particular importance is planning delay, that is that the time missive of invitation waits within the scheduling queue before it's scheduled on a machine. Task planning

delay could be a primary concern in information center environments for many reasons: (1) A user may need to right away proportion Associate in Nursing application to accommodate a surge in demand and thence needs the resource request to be glad as presently as
possible.

(2) Even for lower-priority requests (e.g., background applications), long planning delay will lead to starvation, which might considerably hurt the performance of those applications. In apply, however, there\'s typically a trade-off between energy
savings and planning delay. despite the fact that turning off an oversized variety of machines are able to do high energy savings, at a similar time, it reduces service capacity and thence results in high planning delay.



Data Analytic Cloud Architecture

Finally, the heterogeneity- aware DCP theme ought to also take into consideration the reconfiguration prices associated with change on and off individual machines. this is often as a result of oftentimes turning on and off a machine will cause the "wear-and-tear" result that reduces the machine life. Despite the very fact that an oversized variety of DCP schemes have been planned within the literature in recent years, a key challenge that usually has been unmarked or considered tough to handle is heterogeneousness, which is current in production cloud information centers. we have a tendency to summarize the categories of heterogeneousness found in production environments as follows: Machine heterogeneity. Production information centers typically comprise several varieties of machines from multiple generations. they need heterogeneous processor architectures and speeds, hardware options, memory and disk capacities. Consequently, they have different runtime energy consumption rates.

(A) Resource Allocation

In this section, we'll examine the resource allocation strategy to form information analytics within the cloud economical. By "efficient", we tend to mean minizing the dimensions of the cluster whereas meeting the performance req uirements. In this way, we are able to improve overall information center utilization by letting the remaining machines used by alternative applications, or scale back price by not paying for redundant machines. to it finish, we tend to propose Associate in Nursing design to create a data analytics system within the cloud as seen in Figure one. In this design, collaborating nodes square measure sorted into one among 2 pools:

(1) long-living core nodes to host both information and computations, and

(2) accelerator nodes that square measure value-added to the cluster quickly once extra computing power is required. Associate in Nursing analytic engine (e.g., Hadoop MapReduce) runs on nodes in each pools whereas a storage system (e.g., HDFS) is deployed solely on core nodes. This approach is comparable to the previous work of Chohan et al. [6], within which spot instances (cheaper however could also be terminated while not notice) of Amazon EC2 square measure value-added to process nodes to hurry up Hadoop jobs. The cloud driver manages nodes allotted to the analytic cloud and decides once to add/remove what form of nodes to/from that pool, and the way several. Users submit employment to the cloud driver with many hints about the duty characteristics, as well as memory demand, ability to use special options like GPUs, and the deadline, if out there. several production queries square measure habitually processed, therefore the cloud driver keeps the history of query executions to estimate the submission rate

of those queries and update the hints provided. It additionally monitors the storage system to estimate the incoming rate. In this way, the cloud driver predict the resource needs to method queries and to store information. The cloud driver is answerable for allocating resources to the cluster. the amount of core nodes is decided primarily based on the desired storage size. additionally, the cloud driver refers to the history to examine if additional nodes should be value-added to the core pool to accommodate the production queries. once several production queries with tight deadlines square measure anticipated or an oversized ad-hoc question is submitted, the cloud driver can add nodes to the accelerator pool quickly to handle them instead of allocating too several core nodes which will be underutilized. When adding nodes, the cloud driver additionally makes a decision on that resource instrumentality (e.g., virtual machine) to use. As Associate in Nursing illustration, we tend to examine the case when we use Amazon EC2 for the cloud. If we tend to contemplate solely the price for the storage, using m1.large instances is that the most cost-effective. However, these instances even have the smallest amount computing power among instance sorts out there , thus we would would like additional nodes to accommodate the assembly queries. during this case, using other instance sorts like c1.medium may be additional economical in terms of the total expense. Moreover, some jobs could run considerably quicker on nodes of a selected instance sort (e.g., cg1.4xlarge instances with GPUs).

Hence, it's vital to understand the job/instance sort relationship (which we tend to call *job affinity*) to find a good mixof different instances that minimize the cost to maintain the cluster.

II. RELATED WORK

WORKLOAD ANALYSIS:

To understand the heterogeneousness in production cloud knowledge centers, we've conducted AN analysis of work traces for one in every of Google's production reckon clusters consisting of approximately 12000 machines. The work traces

contain programming events, resource demand and usage records for a complete of 672,003 jobs and 25;462;157 tasks over a time span of twenty nine days. Specifically, employment is AN application that consists of one or a lot of tasks. every task is scheduled on one physical machine. once employment is submitted, the user can specify the utmost allowed resource demand for each task in terms of needed processor and memory

size.

Understanding Machine Heterogeneity:

The traces conjointly give data regarding the types of machines utilized in the cluster. A machine is characterized by its capability in terms of central processing unit, memory and disk size additionally as a platform ID, which identifies the micro-architecture (e.g., trafficker name and chipset version) and memory technology (e.g., DDR or DDR2) of the machine. just like tasks, machine capacities ar normalized specified the

largest machine encompasses a capability up to one. shows the various sorts of machines and their characteristics (capacity and platform ID (PFID)). We found ten sorts of machines wherever over fifty and 30 % of the machines belong to machine types one and a couple of, severally. On the opposite hand, machine sorts three and four have around 1;000 machines each. The remaining machine sorts (5 to 10) constitute but a hundred machines. sadly, the traces don't give careful data regarding hardware specifications, however, it's possible that this heterogeneity interprets into completely different energy consumption models.

Understanding Task Heterogeneity

In order to investigate the work heterogeneity, we have a tendency to premeditated tasks' needs and their durations for the 3 priority teams. the CPU and memory size of tasks happiness to every priority group. The coordinates of every

purpose in these figures correspond to a mixture of C.P.U. and memory requirements. Radius of every circle is power in number of tasks at intervals its proximity. It may be seen that most of the tasks have low resource requirements.

III. FRAME WORK

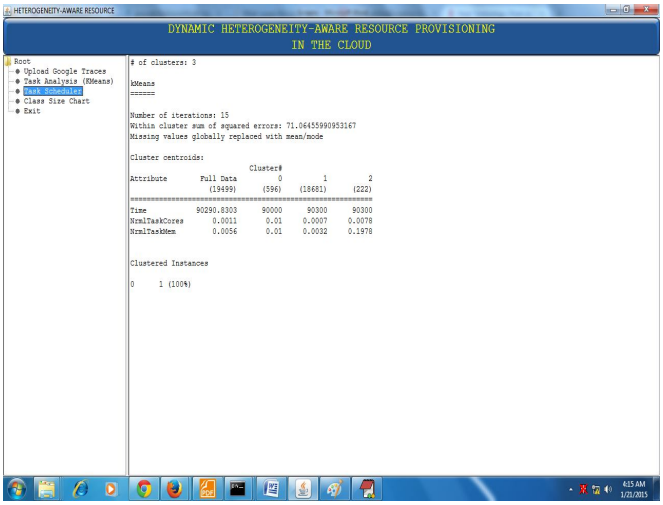
Despite intensive studies of the difficulty, existing solutions haven't absolutely thought of the heterogeneity of each work and machine hardware found in production environments. it can be easily integrated with existing planning algorithms, variants of first-fit and best-fit algorithms and Open source platforms like Eucalyptus will adopt this mechanism by ever-changing the programming policy to weight round-robin 1st job and weight round-robin best fit, severally. the most advantage of CBP is its simplicity and usefulness for preparation in existing systems . current work on this subject has not addressed a key challenge, that is that the non uniformity of workloads and physical machines.

The fact that an outsized variety of DCP schemes are planned within the literature in recent years, a key challenge that usually has been overlooked or thought of tough to deal with is heterogeneity, that is prevailing in production cloud data centers. gift the formulation of the heterogeneity-aware DCP, and supply 2 technical solutions . Discusses the readying of Harmony in practice. Finally, we have a tendency to appraise our planned system using Google employment traces. as each hardware design and capability upgrade need a careful understanding of the employment characteristics in terms of arrival rate, needs, and period .We have analyzed the employment of a Google cipher cluster, associated planned an approach to task classification mistreatment k-means bunch .

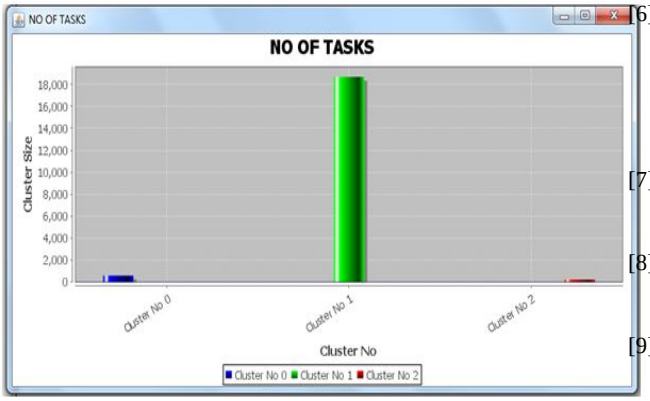
Realizing that directly determination DCP isn't possible, during this section we have a tendency to gift 2 quick heuristics for determination DCP. each techniques think about determination the integer-relaxation of DCP (i.e., quiet the constraints that variables should take number values) called DCP nine RELAX, that is far easier to unravel than DCP. Once the answer for DCP nine RELAX is obtained, one amongst our resolution techniques referred to as containerbased provisioning (CBP) directly roundsthe numbers of machines to the closest number values and use these values for capability provisioning. On the other hand, the containerbased planning (CBS) technique tries to search out a possible placement of containers in physical machines, and use containers for run-task planning. In each cases, the capability provisioning module 1st adjusts the amount of active machines, and informs the computer hardware concerning how tasks ought to be appointed to every form of machines.

IV. EXPERIMENTAL RESULTS

The following picture shows that input request given by the user who was belong to Cluster 0.



Below graph displaying the number of records each cluster contains in the form of graph as follows



V. CONCLUSION

Dynamic capability provisioning has become a promising answer for reducing energy consumption in knowledge centers in recent years. However, existing work on this subject has not addressed a key challenge, that is that the non uniformity of workloads and physical machines. during this paper, we 1st give a characterization of each work and machine non uniformity found in one among Google's production figure clusters. Then we have a tendency to gift Harmony, a heterogeneity-aware framework that dynamically adjusts the quantity of machines to strike a balance between energy savings and planning delay, whereas considering the reconfiguration price. Through experiments exploitation Google work traces, we found Harmony yields giant energy savings whereas significantly rising task planning delay.

REFERENCES

[1] Energy Star Computer Server Qualified Product List, energystar.gov/ia/products/prod_lists/enterprise_servers_prod_list.xls, 2014.

[2] Googleclusterdata - traces of google workloads, http://code.google.com/p/googleclusterdata/, 2014.

[3] U.S. Energy Information Adminstration, ttp://www.eia.gov/, 2014.

[4] R. Boutaba, L. Cheng, and Q. Zhang, "On Cloud Computational Models and the Heterogeneity Challenge," J. Internet Services and Applications, vol. 3, pp. 77-86, 2012.

[5] S. Boyd et al., Convex Optimization. Cambridge Univ. Press, 2004.

[6] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing Server Energy and Operational Costs in Hosting Centers," ACM SIGMETRICS Performance Evaluation Rev., vol. 33, pp. 303-314, 2005.

[7] J. Diaz et al., "A Guide to Concentration Bounds," Handbook on Randomized Computing. Springer, 2001.

[8] D. Gross and C. Harris, Fundamentals of Queueing Theory, pp. 244- 247, John Wiley & Sons, 1998.

[9] S. Lee, R. Panigrahy, V. Prabhakaran, V. amasubrahmanian, K. Talwar, L. Uyeda, and U. Wieder, "Validating Heuristics for Virtual Machines Consolidation," Microsoft Research, MSR-TR-2011-9, 2011.

[10] A.K. Mishra, J.L. Hellerstein, W. Cirne, and C.R. Das, "Towards Characterizing Cloud Backend Workloads: Insights from Google Compute Clusters," ACM IGMETRICS Performance Evaluation Rev., vol. 37, pp. 34-41, Mar. 2010.

[11] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the Electric Bill for Internet-Scale Systems," Proc. ACM SIGCOMM, 2009.

[12] S. Ren et al., "Provably-Efficient Job Scheduling for Energy and Fairness in Geographically Distributed Data Centers," Proc. IEEE 32nd Int'l Conf. Distributed Computing Systems (ICDCS), 2012.

[13] A. Verma et al., "Pmapper: Power and Migration Cost Aware
Application Placement in Virtualized Systems," Proc. Ninth ACM/ IFIP/USENIX Int'l Conf. Middleware (Middleware), 2008.

[14] Q. Zhang, M.F. Zhani, R. Boutaba, and J.L. Hellerstein, "HARMONY: Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud," Proc. IEEE Int'l Conf. Distributed Computing Systems (ICDCS), 2013.