DOI : https://doi.org/10.56452/7-2-561

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Harmony: A Dynamic Heterogeneity−Aware Resource Provisioning In the Cloud Environment

Kanchan Naithani

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

HARMONY, a resource management system for heterogeneity-aware settings that dynamically provide capacity. The workload is divided into different task classes with comparable properties in terms of resource and performance needs using the K-means clustering technique. The DCP is an optimization problem that takes into account the heterogeneity of machines and workloads. It is more adaptable and may offer a very effective DCP solution for any workload composition. The heterogeneous characteristics will produce insufficient energy savings and protracted scheduling delays as a result of a mismatch between workload requirements and the resources offered by the supplied machines. a special technique for dynamically adjusting the number of machines of each kind to lower total energy consumption and performance costs with regard to scheduling delays. These findings imply the need for more effective provisioning and scheduling techniques to shorten the scheduling lag for these challenging jobs.

1. INTRODUCTION

Large-scale service applications may now be hosted on a cost-effective platform in data centres, which has lately seen tremendous growth in popularity. While big data centres benefit from economies of scale by spreading out long-term capital expenditures across a large number of machines, they also have very high energy costs for cooling and electricity delivery. In instance, it has been stated that the cost of electricity makes up around 12% of total data centre expenses. A 3% decrease in energy costs may result in cost savings of more than $1 million for big businesses like Google.

There has been a lot of research done recently to increase data centre energy efficiency, especially with the aim of Dynamic Capacity Provisioning (DCP). This method aims to minimise energy usage while maintaining SLOs for workloads by adjusting the number of active computers in a data centre. In data centre environments, scheduling delays are a serious problem because of the requirement to swiftly scale up an application to handle a spike in demand and the risk of hunger. Energy savings and scheduling delay, however, may have to be traded off because shutting down a lot of equipment might result in significant energy savings but also lower service capacity and increase scheduling delay.

However even though several DCP methods have been put out in recent years, heterogeneity, which is common in production cloud data centres, remains a significant issue that is frequently disregarded or regarded as being impossible to handle.

Dynamic capacity provisioning in cloud computing settings is made possible by HARMONY, a heterogeneity-aware resource management system. Using the K-means clustering technique, the workload is divided into groups of tasks with comparable resource and performance needs. Machine and workload heterogeneity are taken into account by the DCP as an optimization challenge. For

arbitrary workload compositions, it can offer a very effective DCP solution that is more adaptable. Because workload demands and the resources provided by the supplied machines are incompatible, the heterogeneous features will result in both sub-optimal energy savings and lengthy scheduling delays. a unique method for dynamically altering the number of machines of each kind to reduce overall energy use and the performance cost in terms of scheduling delay. These findings imply the need for more effective provisioning and scheduling techniques to shorten the scheduling lag for these challenging jobs.

The proposed system is capable of handling the workload and machine heterogeneity present in one of Google's production compute clusters. Our aim is to provide effective task categorization at run time while ensuring high characterization accuracy. By carefully managing the number of work classes, it may be characterised with a high degree of precision. The number of machines is dynamically adjusted between energy savings and scheduling delay. Up to a 20% increase in energy savings may be achieved through harmony, which also considerably reduces job scheduling delay. The Harmony method uses k-means clustering to enhance task categorization.

2. LITERATURE SURVEY

The requirement for representative workload benchmarks is driven by the growing size and complexity of big compute clusters, which makes it necessary to assess the performance effect of system modifications in order to improve scheduling algorithms and manage management tasks. To do this, workload characterizations must be built, from which realistic performance benchmarks may be derived [1].

In a setting with diverse virtualized server clusters, the authors look at the design, implementation, and assessment of a power-aware application placement controller. The application management middleware's placement component considers the power and migration expenses as well as the performance advantage when deciding where to put the application containers on the physical servers [2].

By shutting off unused equipment, data centre capacity may be dynamically adjusted, which is an efficient method for conserving energy. To determine the best control strategy, use Model Predictive Control (MPC). We demonstrate that our suggested approach may achieve a considerable decrease in energy cost through in-depth analysis and simulation utilising actual workload traces from Google's compute clusters [3].

A key method for increasing resource utilisation and power efficiency in cloud infrastructures is server consolidation based on virtualization. Yet, dynamic resource pool management through online adaption is essential to guarantee adequate performance of shared resources while adjusting the workload of applications [4].

It examines a recent Google public release of a large-scale production workload trace. We provide a statistical profile of the data, which includes some intriguing findings on task durations, CPU and memory usage, and job arrival trends. In addition, they use k-means clustering to uncover common groupings of employment, making a number of methodological changes and arriving at different conclusions from earlier research on comparable data [5].

3. PROPOSED SYSTEM

Because of the variety of workloads and physical computers, the current system functions. The current systems are capable of carrying out these tasks, including analysing workload along with machine characteristics in production clouds and dynamic capacity provisioning for achieving a balance between energy-saving goals and application performance goals. In accordance with machine availability as well as workload distribution at any given time, a heterogeneity aware DCP controller is intended to change the number of active machines. Accurate task classification during runtime is more challenging since the system typically doesn't know how long a job will take to complete. Regarding task scheduling delay, CBP makes no performance guarantees.

DOI : https://doi.org/10.56452/7-2-561

Nevertheless, the current technique has a number of drawbacks, including scheduling during run time and postponing energy savings. DCP is an optimization problem that takes into account realtime electricity prices as well as machine and workload heterogeneity. Rising energy consumptions and dynamically altering the number of active devices.

One of Google's production compute clusters has a workload and machine heterogeneity, and the proposed solution works well under both conditions. Our aim is to provide effective task categorization at run time while ensuring high characterization accuracy. By carefully managing the number of work classes, it may be characterised with a high degree of precision. The number of machines is dynamically adjusted between energy savings and scheduling delay. Up to a 20% increase in energy savings may be achieved through harmony, which also considerably reduces job scheduling delay. The Harmony method uses k-means clustering to enhance task categorization. The following are some of the proposed approach's benefits:

- By keeping the bottleneck resource at 80% utilisation, scheduling delays and energy savings are avoided.
- A Rise in Diversity
- Services for Stricter Access Control and Security
- Minimize re-labeling process errors.

The steps that must be taken to implement the suggested strategy are described in the section that follows:

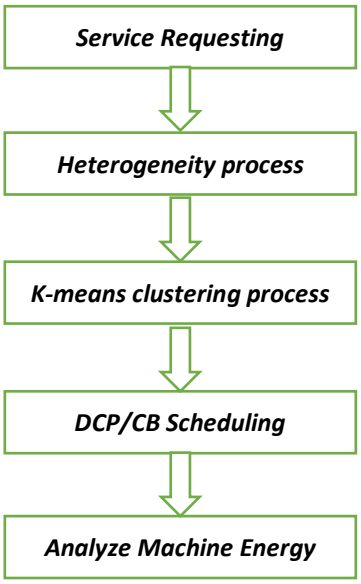

Fig 1: System Architecture

1. Service Requesting

Cloud storage services offer a specialised setting that is secured by a company's firewall. Those that need customization and more control over their data should use private clouds. A minimum of one private cloud infrastructure and one public cloud infrastructure are included in the cloud storage, which is a hybrid of the other two types. An organisation may, for instance, store structured data that is being utilised right away in a private cloud and unstructured data that is being archived in a public cloud.

2. Heterogeneity process

Workload and machine heterogeneity are two of the two processes that make up heterogeneity. It uses energy at varying rates when in use. Increase or decrease the number of active computers in a data centre dynamically. Several separate task classes with a comparable set of resource and performance goals can be created from a heterogeneous workload.

3. K-means Clustering process

The job of arranging a set of items into a single group so that they are more similar to one another than to those in other groups is known as cluster analysis or clustering (clusters). With a collection of unlabeled data, clustering seeks to identify the intrinsic grouping. Assessed the calibre of the K means algorithm's resultant clusters.

4. DCP/CB Scheduling

In production cloud settings, the machine and workload have a significant impact on the creation of DCP systems. The DCP is an issue that requires optimisation takes run-time electricity prices into account along with machine and workload heterogeneity. DCP framework that can increase application performance while also being more effective in terms of energy conservation.

5. Analyze Machine energy

The devices that could house the containers while also consuming the least amount of energy. The monitoring module is in charge of gathering a variety of data on tasks as well as machines, such as CPU and memory use, available resources, along with current task durations. Moreover, it alerts the management framework to any errors and irregularities.

4. RESULTS

In systems utilizing cloud computing, Harmony is a heterogeneity-aware resource management system for dynamic capacity provisioning. The workload is broken down into separate task classes exhibiting equivalent resource as well as performance characteristics needs using K-means clustering methods. This method considerably improves task scheduling latency while reducing the scheduling delay for certain hard-to-schedule jobs by up to 20%. It can also result in considerable energy savings. The most difficult part of utilizing containers for scheduling is choosing the right container size, which guarantees that every activity can be scheduled without going against machine capacity restrictions. These findings imply the need for more effective provisioning and scheduling techniques to shorten the scheduling lag for these challenging jobs. This system can achieve high accuracy in characterization and facilitate effective task categorization at run time by carefully managing the number of job classes.

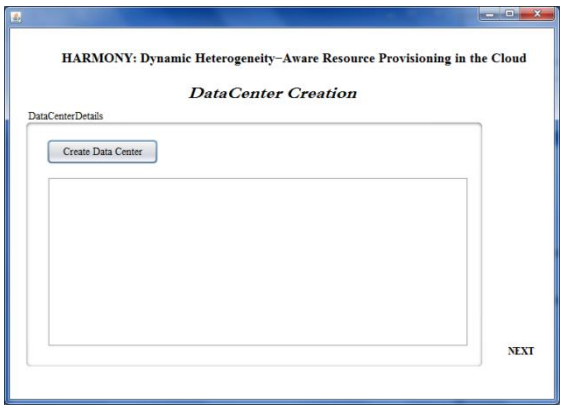

Fig 2: Data Center Creation

DOI : https://doi.org/10.56452/7-2-561

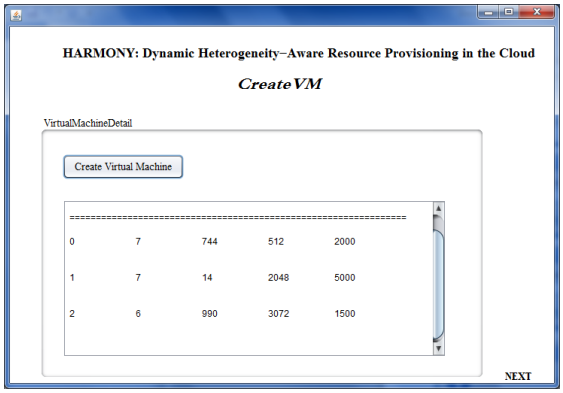

Fig 3: Virtual Machine Creation

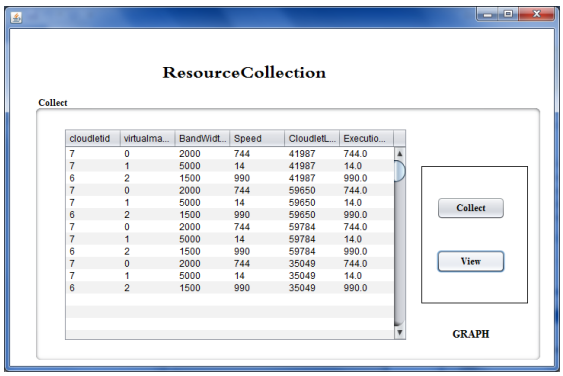

Fig 4: Resource Collection

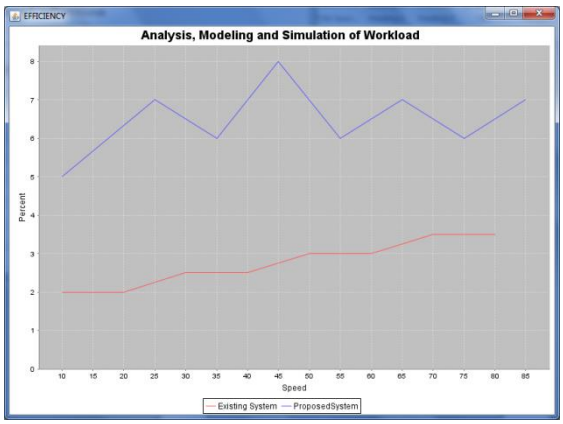

Fig 5: Performance Analysis

5. CONCLUSION

The jobs are categorised utilizing the k-means algorithm based on the static attributes, priority, CPU, and memory size supplied in the job request. The task running time is used to further categorise each task class into subclasses in the second phase. The benefit of this strategy is that it not only makes relabeling easier, but also lessens the amount of mistake that is created during the process. By carefully regulating the number of task classes, a high degree of characterisation accuracy may be achieved. The most difficult part of utilising containers for scheduling is choosing the right container size. On the one hand, scheduling every job while abiding by machine capacity restrictions is ensured by adjusting the container size to the highest permitted value for the class.