Literature study
Increasing resource utilization of cloud with limited resources

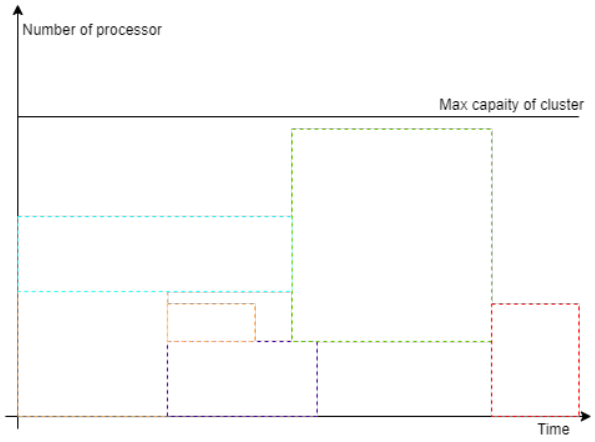
You Hu
adolphus.hu@student.vu.nl
VU Amsterdam

1 INTRODUCTION

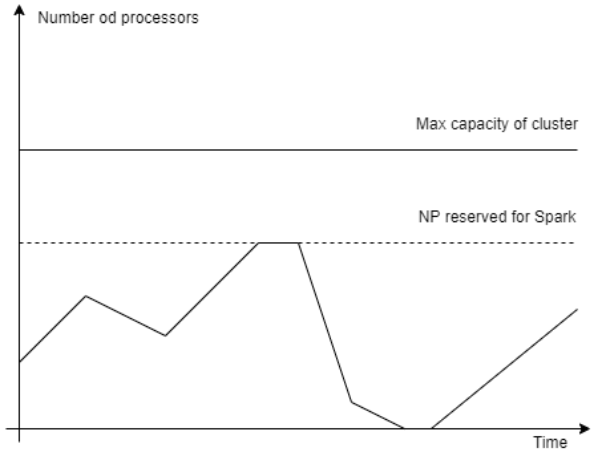
The Netherlands eScience Center has developed solutions for calibrating imaged observation collected by LOw Frequency Array(LOFAR) telescope[1]. The LOFAR consists of 51 stations cross Europe and a typical LOFAR observation has the size of 100TB, after frequency averaging, the size can be reduced to 16TB. [6] Collectively, there are over 5 PB of data will be stored each year. [1] This large volume of data requires vast computation availability to calibrate, on the other hand, since the process availability is not able to catch up with the data producetion, the observation will be processed when it is needed.

The current solutions , MPI, and Spark, will lead to a waste of computation resources in non-dedicated clusters. It can be visualized as Fig. 1. As it is shown in Fig. 1a, MPI jobs are scheduled as fixed batch jobs which are colored boxes in the figure. It is very common to meet the situation that all jobs in the queue are too large and the number of idle resources is not enough for the jobs in the queue. For the Spark version, the possible situation can be visualized as Fig. b. The nodes should be reserved for Spark in advance and Spark task manager handles the task scheduling. For typical Spark applications relying on RDD, the number of required executors can be up and down dynamically. Part of computation resources may be wasted since this computation power is exclusive for Spark. However, of course, the current Spark implementation for calibration is based on Driver mode and the granularity is one executor per task. In this case, the waste of resources won't happen within the Spark cluster. But when there are a lot of idle nodes in the cluster that not reserved for Spark at the beginning, they can not be used for Spark.

In this project, we try to tackle on the issue of resource utilization. The LOFAR owns computation facility their own. Therefore, we take both perspectives of cloud provider in Section 2 and cloud user in Section 3 into discussion.



(a) Resources utilization of MPI batch jobs on cluster



(b) Resources utilization of Spark on cluster

Figure 1: The waste of compuation on cluster

2 CLOUD PROVIDER PERSPECTIVE

The cloud providers own data centers with large quantity of and kinds of facilities. The resource management is the key point that the cloud provider will concern. In 2013, Jennings and Stadler made a survey and listed challenges on Cloud resource management [4]. They firstly lists actors in the cloud economy and then the fields of cloud technologies. Manvi and Shyam also made a survey on the same topic but it mainly focuses on IaaS cloud related aspects [5]. As it

[1] http://www.lofar.org/

provides more observation from the technology side, the issues and definition of resource in clouds are given firstly. In this section, the metrics around cloud are listed, as they are usually as the counterpart of resource utilization for trade off making. And then, few related topics in the resource management field will be discussed.

Metrics related to resource utilization

In the cloud environment, there are many kinds of resources and a set of aspects around the cloud economy. Both Jennings [4] and Manvi [5] starts with the definition of resources.

Jennings et al. categorize the resources into compute, networking, storage, and power. Manvi at al. summarizes that there are physical resources(CPU,memory, storage, network elements and sensors) and logical resources(OS, engergy, Network throughput/bandwidth,Load balancing mechanisms and so on). There are overlap between two of them especially in the physical part, while Manvi adds API, OS, load balancing to logical resource concept. However, it is understandable that API , OS and some protocols and mechanisms can be viewed as some sort of asset of cloud owner, but they are more acceptable to be considered as part of Quality of Service(QoS) which needs resources to fulfill. Therefore, in this section, we mainly focus on the utilization of physical resources like CPU, memory, storage; and consider the trade-off between utilization rate and QoS.

QoS (Quality of Service) metrics are import to both cloud provider and consumer they are good for optimizing resource utilization efficiency Bardsiri and Hashemi listed detailed metrics from four kinds of features:performance, economic, security and general. [2] Their coverage is comprehensive. There are plenty of features and corresponding metrics that cloud users would put concern.

Given the background of researching utilization under limited resources, there are few metrics we consider important. For performance features, the CPU load rate and packet loss frequency are what users may concern while the cloud provides needs to make a compromise for utilization of resources. In the economic aspect, the price per resource unit is the key point that cloud providers and users wrestle on. However, from the technical view, the time for VM booting/deleting/suspending/provison attract more attantion. Besides, the availability and reliability are very important. The response time is the key metric for auto-scaling mechanism. Cloud providers need to pay effort on fault tolerance to make sure the safety of the cloud.

In the following sections, we will explore how cloud providers face resource management issues and the metrics shown above play important roles in those researches.

Resource demand profiling

There is no doubt that before allocation and provision, it is vital to estimate the demanded resource of each workload. In this section, the workloads can be classified as two types: batch and interactive. Different workload has different nature of their requirement of resources.

For the batch applications, it is easier to estimate the resource demand. Given the parameters of each pre-identified application, the required resource over time can be calculated by a pre-trained model. As an example, Becerra et al. purposed a methodology to profile the batch jobs. [3] The workload profile is updated according to the deviation of CPU, memory, network usage by the time.

It is clear that the workload of web applications varies dynamically over multiple time scales. One approach is that all the workloads are

REFERENCES

[1] 2020. PROviding Computing solutions for ExaScale ChallengeS. 777533 (2020), 1–163.

[2] Amid Khatibi Bardsiri and Seyyed Mohsen Hashemi. 2014. QoS Metrics for Cloud Computing Services Evaluation. *International Journal of Intelligent Systems and Applications* 6, 12 (2014), 27–33. https://doi.org/10.5815/ijisa.2014.12.04

[3] Yolanda Becerra, David Carrera, and Eduard Ayguadé. 2009. Batch job profiling and adaptive profile enforcement for virtualized environments. In *2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing*. IEEE, 414–418.

[4] Brendan Jennings and Rolf Stadler. 2015. Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management* 23, 3 (2015), 567–619. https://doi.org/10.1007/s10922-014-9307-7

[5] Sunilkumar S. Manvi and Gopal Krishna Shyam. 2014. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. *Journal of Network and Computer Applications* 41, 1 (2014), 424–440. https://doi.org/10.1016/j.jnca.2013.10.004

[6] Hanno Spreeuw, Souley Madougou, Ronald Van Haren, Berend Weel, Adam Belloum, and Jason Maassen. 2019. Unlocking the LOFAR LTA. (2019), 467–470. https://doi.org/10.1109/eScience.2019.00061