Master thesis proposal
A container based self-adjusted auto-provisioning system at resource level

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

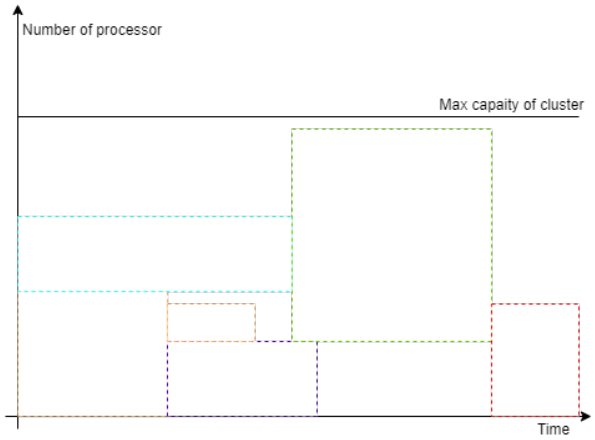
1 INTRODUCTION

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. [**?**] Collectively, there are over 5 PB of data will be stored each year. [**?**] To calibrate the observation by given sky map, SAGECaL is invented and implemented for this purpose.[**?**] By given pre-processed observation data, sky model and parameters, the calibration can be done independently. However, it is a computation consuming application. Currently, eScience Center has developed GPU, MPI and Spark versions for acceleration. All of them have achieved great acceleration compared to the naive version.

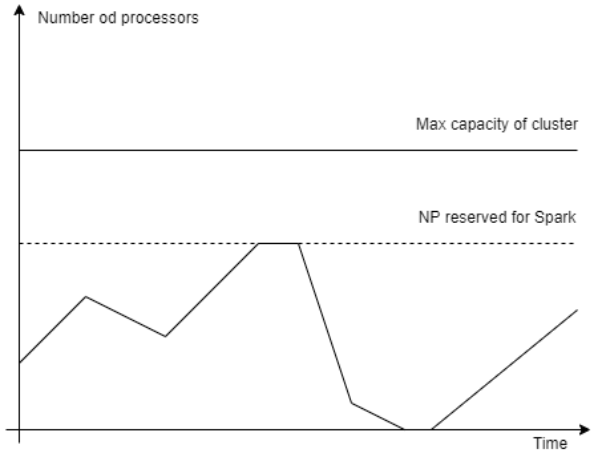
However, the solutions following horizontal scaling idea, MPI, and Spark, will lead to a waste of computation resources in non-dedicated clusters. In this project, we try to build up a system to achieve auto-provisioning at public clusters to drive the computation consuming applications. This solution can be applied to other more applications with the demand for auto-provisioning.

2 THE COMMON ISSUE ON CURRENT VERSIONS

Both MPI version[2] and Spark version[3] take SAGECaL process as a black box, it aims to schedule the task on multiple nodes to achieve acceleration. For the MPI version, the architecture is a master-worker like, but the task scheduling mechanism is very simple. In the Spark version, the tasks are carried by the driver with Java Native Interface. And each task is invoking an encapsulated C++ program. The common issue of MPI and Spark version is that the resources can be not be fully utilized in the non-dedicated cluster. The resource utilization of these two systems 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 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.



(a) Resources utilization of MPI batch jobs on cluster



(b) Resources utilization of Spark on cluster

Figure 1: The waste of compuation on cluster

[1] http://www.lofar.org/
[2] https://github.com/nlesc-dirac/sagecal/tree/master/src/MPI
[3] https://github.com/nlesc-dirac/sagecal-on-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.

3 RESEARCH QUESTION STATEMENT

To tackle on the problems mentioned in previous sections,here, it can be summarized as the research question:

- Is it is possible to build up a system/framework equiped with auto-scaling stratagy at resource level under public clusters?

Moreover, the following topics can be considered together:

- portablility for utilizing remote (heterogeneous) resources
- data locality on cross regions
- fault tolerence

4 RELATED WORKS

The common solution for managing resources is utilizing cloud management platforms like OpenStack or OpenNebula. Tania et al. [**?**] listed multiple methods for auto-scaling on VM based cloud. However, in our case, the application runs in clusters instead of the cloud environment. On the one hand, the VM is very heavy while our task is fixed. On the other hand, the middlewares need to be placed on nodes in advanced like Spark daemon, this leads to the same problem as Spark's. Of course, the strategies of resource management in this report and other papers[**?**][**?**] can be applied in our project.

5 METHODS AND TECHNOLOGIES

In the following subsections, the technologies that we intend to use will be described. And in the last subsection, a rough design will be illustrated.

Ibis Portability Layer(IPL)

Ibis[**?**] is a programming environment that combines Java's "run everywhere" portability both with flexible treatment of dynamically available networks and processor pools, and with highly efficient, object-based communication.[4] IPL can be used for communication between executors. The advantages of IPL are the cross-domain router, builtin leader election mechanism, and user-friendly interface. All of these help to implement a parallel application.

Docker container

A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries, and settings.[5] In our case, the executors are based on containers which encapsulate all related environment. It is more lightweight than VMs and easier to manage.

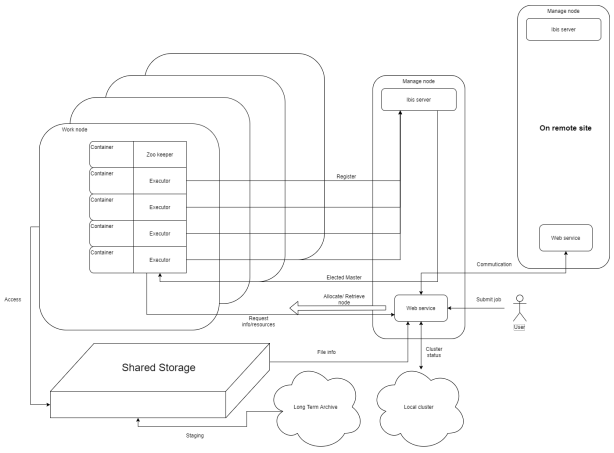


Figure 2: An over view design for the system

Xenon

Xenon is a middleware developed by eScience Center that tends to provide a single programming interface to many different types of remote resources, allowing applications to switch without changing a single line of code.[6] This tool enables the program to communicate with the cluster and to manage resources automatically.

Rough design

Here, a rough design is purposed and shown in Fig. 2. The idea of our solution is that for each cluster, we assume there is one node never crash conveys ibis-server and service for: job submitting; computation node allocation and release; communication with other clusters. Once a node is assigned, it will join/create the docker swarm or Kubernetes cluster. The containers will execute the same code, and elect ibis master. The master receives a job list and allocates tasks to workers. It also monitors the job list to send scaling up or down demand to the server(this part can be discussed in detail later). The job list or task arrangement can be persisted by something like Zookeeper to reduce the cost of the crash of the master.

6 EVALUATION

There are a few kinds of features that need to be evaluated. The core is the resource utilization of the cluster. The average usage rate and overall running time for given tasks can

[4]https://www.cs.vu.nl/ibis/javadoc/ipl/index.html

[5]https://www.docker.com/resources/what-container

[6]https://github.com/xenon-middleware/xenon

be measure for old versions(MPI and Spark) and the new version. Since here we would use JNI to invoke SAGECaL function, we can replace it by any computation consuming and data-intensive application. For data locality features, we can compare the time spent between the remote access version and the version with locality optimization. For fault tolerance, we can randomly force kill nodes to test this feature.

7 PLAN

- Phrase 1 - literature study March 30th- April 27th
  - Week1-3: literature study on resources management algorithm and container-based auto-scaling systems
  - Week4: summarizing and reporting
- Phrase2 - implement prototype : April 27th - May 1st
  - Week1-3: implement a prototype with simple scaling mechanism in one cluster
  - Week4-5: design a benchmarking framework for different scaling mechanism
- Phrase3 - add&test features : May 1st - May 29th
  - Week1-3: test different(2-3) mechanisms and add more features
  - Week4: enhance testing system for cross region requirement
- Phrase4 - final evaluation : May 29th - June 26th
  - Week1-2: evaluate the performance and try to compare to existed versions
  - Week3-4: debugging and optimizing
- Phrase5 - organize thesis paper: June 29th - July 24th
  - In previous stages: build up skeleton and fill in content at any time
  - Week1-2: make the draft
  - Week3-4: review and improve

8 SUPERVISION

This project is initialed by the Netherlands eScience Center. The supervision is provided by:

- Dr. Jason Maassen[7], Technical lead efficient computing at Netherlands eScience Center
- Prof. Adam Belloum[8], Full Professor at UvA and Technical lead optimized data handling

There will a regular weekly meeting with Prof.Belloum, and continuously reporting to Dr. Maassen about the progress.

This proposal has been agreed by You Hu, Jason Maassen and Adam Belloum. For convenience and due to the COIVD-19 situation, there is no signature in this version. If needed, please contact to You Hu for making signed version.

[7] e-mail:J.Maassen@esciencecenter.nl
[8] e-mail:A.S.Z.Belloum@uva.nl