Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 90 – 98

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Performance of a Low Cost Hadoop Cluster for Image Analysis in Cloud Robotics Environment

Basit Qureshi[a,*], Yasir Javed[a,e], Anis Koubâa[a,b], Mohamed-Foued Sriti[c], Maram Alajlan[d]

[a]*Prince Megrin Data Mining Center, Prince Sultan University, Saudi Arabia*
[b]*CISTER/INESC-TEC, ISEP, Polytechnic Institute of Porto, Porto, Portugal*
[c]*Al-Imam Mohammad Bin Saud University, Saudi Arabia*
[d]*King Saud University, Saudi Arabia*
[e]*eFIT, UNIMAS, Sarawak, Malaysia*

Abstract

With the emergence of cloud robotics, the cloud computing paradigm becomes increasingly attractive to robotics, where the cloud acts as the remote brain of low-cost robots, such as commodity drones. The idea is to offload heavy computations, like image processing, from the robot to the cloud; process it in short time (near real-time) and send back commands to the robot. This paper investigates the performance of a back-end cloud computing framework in deploying robotics-like applications (i.e. image analysis and processing) using low-cost Hadoop clusters. The design of a low-cost mini-data center built with readily available commodity 32-bit ARM boards, i.e. Raspberry Pi 2 Model B, is presented. Furthermore, the performance of RPi-based clusters is extensively tested with different types of data including text, text/image and image, and a comparative analysis against Hadoop cluster running on virtual machines is presented. The Hadoop Image Processing Interface (HIPI) Library was used and also configured to optimally utilize the Pi Cluster resources for improved performance. Results show that the RPi Hadoop cluster lags in performance when compared to Hadoop cluster running on virtual machines, the low cost and small form factor makes it ideal for remote Image analysis in surveillance / disaster recovery scenarios where UAVs can transmit image streams to the Cluster for remote processing.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: UAVs; HIPI; Hadoop cluster; Image analysis; Distributed processing; Embedded systems

1. Introduction

Nowadays, rapid development in Cloud Robotics research has enabled manufacturers to mass market Unmanned Aerial Vehicles (UAV) often known as drones. These UAVs are capable of flying with a remote controlled device and can capture image and video streams of data. Recently there has been a trend in use of UAVs in surveillance applications where real-time images can be captured. In such applications, detection of suspicious items and objects is a fundamental task that requires intensive computation and processing of objects features and parameters against database of suspicious objects. Since UAVs have limited onboard processing and storage capabilities that restrict their

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: qureshi@psu.edu.sa

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.013

abilities in processing and handling such tasks, these could be offloaded to a cloud computing environment capable of storing petabytes of data and processing intensive image processing and analysis applications.

In the light of these limitations, we investigate the use of a low cost cloud infrastructure built with readily available commodity ARM processors that can be used to build clusters. Raspberry Pi (RPi)[1] are a class of embedded computers build with 32-bit ARM processors with on-board memory, Network I/O and Storage. These computers have desirable characteristics such as low-cost, low-power yet high performance boards capable of intensive computation and storage. Michael Laurenzano in[2] characterize the performance and energy of HPC computations drawn from a number of problem domains on current ARM and x86 processors. Pleiter and Richter in[3] study the energy efficiency of embedded system with ARM Cortex-9 processors for scientific numerical applications. Michael Cloutier in[4] investigate the currently available 32-bit low-cost Embedded Computing platforms including Raspberry Pi[1], Beaglebone Black[5] as well as other systems. Abrahamsson in[6] investigate the use of Raspberry Pi to build an affordable and energy-efficient cloud computing cluster with 300 Raspberry Pi computers.

Apache Hadoop[7] is an open-source software framework derived from Google's MapReduce and Google File System (GFS)[8]. Recently, Hadoop became a de-facto standard in the area of BigData analytics. A growing number of Big Data applications are being run on the public cloud service providers. Manikandan et. al.[9] investigate use of Big Data Analytics techniques for classification and clustering of Big Data using MapReduce. Loewen in[10] test performance of virtualization within cloud with variable dataset of different sizes. Mathiya[11] notice that configuring YARN parameters in Hadoop 2.0 greatly improves optimal performance for various applications. Xhafa et.al. in[12] evaluate Data Mining Frameworks under a Hadoop cluster and show that improvements in time efficiency to a certain scale for most mining functions. Unfortunately, there are many limitations to Hadoop supporting image/video analysis applications for the following reasons:

- Lack of image read/write interface: Hadoop provides many interfaces for reading and writing text data but no read/write interface for image data. In particular, object recognition applications may require the information from preceding and subsequent images or frames of videos.
- Existing Image and video processing applications are not compatible with Hadoop framework. Much work needs to be done in allowing popular object recognition libraries such as OpenCV which are coded in C/C++ to execute in Hadoop environment.
- The HIPI Library[13] introduced in 2011 provides an interface for storing images in HDFS using Hib file format. Since Hadoop uses Input Split format for splitting large files based on HDFS block size, the performance of image analysis algorithms running on HIPI desire much improvement.

The contributions of this paper are in three-folds: First, we design and setup a low-cost mini-data center with commodity RPi computers organized into four clusters each with five machines. We discuss the challenges faced to build the cloud infrastructure and to install and configure the cloud applications for big data processing. Second, we evaluate the performance of the RPI clusters for computation and storage intensive applications, in terms of execution time. In particular, we contrast the performance of Hadoop applications running on RPI Cluster with Hadoop running on regular PCs in virtualized environments. Finally, we extend the HIPI library to optimally utilize the computation and storage capabilities of the RPI Clusters. Furthermore, we evaluate the performance of RPI cluster using the extended HIPI Library and compare the results of the library for large scale compute intensive applications for Image analysis-as-a-service.

2. Related Work

Manikandan et. al.[9] investigate use of Big Data Analytics techniques such as Joins, Indexing, graph search and applied these for classification and clustering of Big Data using Map reduce. Loewen in[10] test performance of virtualization within cloud with variable dataset of different sizes. Their results show improved performance when using distributed processing. Mathiya[11] notice that configuring YARN parameters in Hadoop 2.0 greatly improves optimal performance for various applications. They study and analyze customizing parameter configurations setting for the performance tuning of Apache Hadoop jobs and notice improvement in utilization of available hardware resources. Xhafa et.al. in[12] evaluate Data Mining Frameworks under a Hadoop cluster and show that improvements in time effi-