Efficient Parallel Data Processing in the Cloud

Ashley Ingram

BSc Computing

2013/2014

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student)_____

Summary

Fill in the summary here.

Acknowledgements

Most of all I'd like to thank my project supervisor ...

Contents

Chapter 1

Introduction

1.1 Project Aim

Large scale data processing is an emerging trend in Computing, with benefits to both academia and industry. The aim of this project is to investigate the challenges of large scale parallel data processing, discuss the benefits of Cloud Computing and investigate how the efficiency of large scale data processing may be improved.

This project will aim to provide a reasoned and objective evaluation of the efficiency of data processing techniques, whilst comparing current state of the art tools and technologies with newer research in the field.

Specifically, the data processing tools Hadoop and Nephele will be compared, allowing comparison of the individual tools, and a wider discussion of the pros and cons of the MapReduce and PACT programming paradigms.

The project aims to answer the following questions:

- Does PACT overcome the inherent disadvantages of the MapReduce paradigm?
- How well does Nephele perform MapReduce tasks?
- How do Hadoop and Nephele perform in a highly elastic Cloud Computing environment?

1.2 Objectives

The objectives of the project are:

- Background Research The background research will provide context for the project through a summary of Cloud Computing, Big Data, and any other relevant trends. It will discuss current tools and technologies in use for large scale data processing, and identify research which aims to overcome the key deficiencies in existing approaches.

- Experiment Design Experiments will be designed to test the efficiency of tools in relevant situations. This will require identifying relevant areas of interest, and deriving scenarios which will test the efficiency of tools. The experiments should be designed to take into account the various different problems that data processing tools are required to solve. There should also be experiments to test the tools in different environments, in order to evaluate how well they perform in Cloud environments.

  As part of the experiment design, a hypothesis will be produced.

- Experiment Implementation The experiments will be implemented and executed, in order to gather results which indicate the efficiency of tools in different environments and situations.

- Experiment Evaluation The evaluation will compare the efficiency of Hadoop and Nephele using the data obtained in the experiments. It will consider the results and how they are related to existing research (obtained in the background research). It will also discuss how the results compare with the hypothesis and consider the causes of discrepancies (should any exist).

- Project Evaluation The project will be considered as a whole, with a considered evaluation of several aspects. It should consider the relevance of the project in regards to existing literature, how it contributes to the research landscape and whether there is a place for future research in to this area. The project should also be evaluated on whether it has met the aims and objectives.

1.3 Methodology

1.4 Schedule

1.5 Summary

Chapter 2

Background Research

This chapter is intended to give an overview of the current research landscape, and to summarise the core technologies and concepts which form a basis for this project. The chapter will discuss trends toward Cloud Computing, how it is relevant to Data Processing, and key technologies which have been developed in this area, giving a foundation for further investigation into the efficiency of Data Processing techniques and how they are impacted by the Cloud.

2.1 Cloud Computing

Cloud Computing is the latest major infrastructure paradigm which looks to deliver on the promise of Utility Computing. In practice, the term 'Cloud Computing' is ambiguous. Whilst no clear definition exists many experts agree that the cloud exhibits the core benefits of Utility Computing such as elasticity and scalability, whilst making heavy use of virtualisation and pay-per-usage business models [11].

Elasticity in clouds refers to the ability for a user to dynamically select the amount of computing resources they require, allowing them to scale their applications according to demand. The resources that can be acquired are essentially limitless from the user's perspective [7].

Elasticity represents a dramatic shift from the 'traditional' method of building and deploying applications. Rather than purchasing and provisioning hardware and acquiring physical space (such as a data centre), a user can use a Cloud Service Provider. This allows for greater flexibility in business and application development, as users can cope with unpredictable or inconsistent levels of demand. An example of where this flexibility would benefit a company is in the case of an online store. A store may receive fluctuating traffic throughout a year, such as being particularly busy around the Christmas

period. It would be economically inefficient for the store to purchase extra servers to cope with demand over the festive period, as they would be redundant for the majority of the year, but they must improve hardware capability in order to take advantage of the extra business. The inherent flexibility from the elasticity of the cloud would allow the store to simply acquire new computing capacity from their Cloud Service Provider on a temporary basis, giving them the capability of accommodating with the peak traffic but not using (or paying for) the extra resources when they are not needed. Having this scalability is seen as a core benefit of Cloud Computing.

A Cloud Service Provider is an organisation that provides access to Cloud Computing resources. They manage the underlying hardware, and typically provide APIs and other methods for a user to manage their resources. Some of the largest Cloud Service Providers are Amazon through AWS (Amazon Web Services), Microsoft through Windows Azure and Google through Google Apps.

A Cloud Service Provider does not have to be an external organisation, but when they are they typically use pay-per-usage business models. Rather than pay a fixed monthly cost, or have a one-off license fee, customers will pay the Cloud Service Provider for the resources they use (usually on a per-hour basis). For example, a 'Medium' size Virtual Machine costs 0.077 an hour from Windows Azure [8]. This allows businesses to only pay for the resources that they need.

2.1.1 Related Ideas

2.1.1.1 Utility Computing

Utility Computing is the idea that households and businesses could outsource their demand to external companies, who provide the relevant amount of service on a pay-per-usage basis. Customers would access computing resources over a network, and would pay for the length of computing time that they use.

This is analogous to other utilities such as Gas or Electricity. In the case of Electricity, power is provided from the National Grid and the customer pays for how much they use. This allows a customer change the amount of power they require without having to pay a fixed cost (for example, using less electricity when they are on holiday).

Utility Computing is an established concept, with leading thinkers such as Leonard Klienrock (part of the original ARPANET project) referencing it as early as 1969 [6]. Various technologies have emerged which offer some attributes associated with utility computing, with Grids and Clouds appearing to be the most promising [3].

2.1.1.2 Virtualisation

Virtualisation is one of the key enabling technologies behind Cloud Computing. It abstracts away the details of the physical hardware and allows Cloud Service Operators to run several virtual machines on one physical machine [12], completely independently of one another. This allows for customers applications and the physical hardware to be consolidated, utilising resources more efficiently and

making it financially feasible to run a cloud [4].

The configurability afforded by virtualisation is another property essential to Cloud Computing. It allows Cloud Service Providers to support a diverse range of applications, which may have different requirements (high compute, high memory, etc). Virtualisation allows this to be achieved, as it would be prohibitively costly at hardware level [4].

The reliability of a cloud can also be improved through the use of virtualisation techniques. Virtual Machines can be backed up, migrated and replicated simply, allowing applications to recover from hardware failure.

2.2 Cloud Computing Service Models

Depending on the scenario, the Cloud offers several different service models. These models allow for clients to provision services in a different manner, depending on what they need from the cloud.

The different service models provide different levels of abstraction for the user. In Infrastructure as a Service, the user has full control over the machines that they acquire from the Cloud Service Provider, where in Software as a Service they are given less control, and need not worry about the underlying hardware whatsoever.

2.2.1 Infrastructure as a Service

Infrastructure as a Service (IaaS) provides an abstraction on top of a virtualisation platform, so that the client does not need to worry about what method of virtualisation is being used, and does not have to learn about the underlying technologies [2].

Clients can request Virtual Machines in varying configurations, and a Virtual Infrastructure Manager will provision an appropriate Virtual Machine on a physical machine which has capacity. In addition to allowing users to provision virtual machines, IaaS systems may allow a user to configure other infrastructure elements, such as virtual networks.

This provides a great deal of control to the user, as they are essentially renting a machine of a requested specification for a short period of time. They are free to install whatever Operating System and software on the machine as required, and can configure it in essentially any way.

An example of an Infrastructure as a Service provider would be Amazon EC2 [1]. Amazon provide a variety of different Virtual Machine types, including those specialising in High Performance Computing or applications requiring a large amount of memory. Virtual Machines can use a range of images provided by Amazon (including Windows and various distributions of Linux), or users can create and upload their own custom Virtual Machine images.

2.2.2 Platform as a Service

Platform as a Service (PaaS) is a higher level abstraction which allows applications to be built and deployed without worrying about the underlying Operating System or runtime environment [5]. The

user still specifies the resources required, but no longer has to manually manage the virtual machines. The Cloud Service Provider will maintain the machines, providing the necessary software (Operating Systems, Web Servers, etc) and updating them frequently.

The advantage of PaaS is that it allows users to deploy their own applications, without having to worry about maintaining the underlying infrastructure. Whilst this decreases the control the user has over the deployment environment, it reduces the complexity of managing the infrastructure themselves.

PaaS offerings may also provide supplementary services to users, such as health and availability monitoring, or auto-scaling.

Windows Azure is an example of a Platform as a Service provider [9]. Whilst they provide Infrastructure as a Service offerings, they also provide Platform as a Service capabilities through Windows Azure Web Sites. Windows Azure Web Sites allow users to upload applications written in a variety of web technologies (ASP.NET, Python, PHP, Node.js) and have them hosted in the Windows Azure runtime environment. This means the client does not manually have to manage web servers, frameworks and other necessary technologies.

2.2.3 Software as a Service

Software as a Service (SaaS) refers to providing access to applications over the internet on-demand [12]. Software is centrally hosted by the Cloud Service Provider, and clients can access the application through a web browser or other form of client. As the software is centrally hosted, Cloud Service Providers can handle updating the software for all users, ensuring all users benefit from bug fixes or additional features.

Software as a Service applications can reduce the cost of deploying and using software for an organisation as they don't have to purchase their own hardware, install and configure software, and can avoid having technical support staff. An example of a successful Software as a Service application is Salesforce [10]. Salesforce is a Customer Relationship Management tool which charges organisations per user, making it a viable choice for small businesses. Salesforce can be accessed through a web browser, enabling customers to use their software regardless of location or device.

2.3 Cloud Computing Deployment Models

2.4 Big Data

2.5 MapReduce

2.6 PACT

2.7 Summary

Chapter 3

A Long but Boring Chapter

Bibliography

[1] Amazon. Aws — amazon elastic compute cloud (ec2) - scalable cloud services, 2014.

[2] Alex Amies, Harm Sluimen, Qiang Guo Tong, and Guo Ning Liu. *Developing and Hosting Applications on the Cloud*. IBM Press/Pearson, 2012.

[3] Rajkumar Buyya, Chee Shin Yeo, and Srikumar Venugopal. Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In *High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on*, pages 5–13. Ieee, 2008.

[4] Ian Foster, Yong Zhao, Ioan Raicu, and Shiyong Lu. Cloud computing and grid computing 360-degree compared. In *Grid Computing Environments Workshop, 2008. GCE'08*, pages 1–10. Ieee, 2008.

[5] Platform as a service (paas) drives cloud demand. Technical report, Intel, August 2013.

[6] L. Kleinrock. UCLA to be first station in nationwide computer network. UCLA Press Release, July 1969.

[7] Peter Mell and Timothy Grance. The nist definition of cloud computing. 2011.

[8] Microsoft. Pricing calculator — windows azure, 2014.

[9] Microsoft. Windows azure, 2014.

[10] salesforce. Salesforce crm, 2014.

[11] Luis M Vaquero, Luis Rodero-Merino, Juan Caceres, and Maik Lindner. A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1):50–55, 2008.

[12] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications*, 1(1):7–18, 2010.

Appendix A

Personal Reflection

I think my project went well.

Appendix B

Record of External Materials Used

I developed all of the materials presented in this project myself.

Appendix C

How Ethical Issues were Dealt With

No ethical issues arose during my project.