J. Parallel Distrib. Comput. 71 (2011) 316–332

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

On-demand minimum cost benchmarking for intermediate dataset storage in scientific cloud workflow systems

Dong Yuan *, Yun Yang, Xiao Liu, Jinjun Chen

Faculty of Information and Communication Technologies, Swinburne University of Technology, Hawthorn, Melbourne 3122, Victoria, Australia

ARTICLE INFO

Article history:
Received 23 January 2010
Received in revised form
13 August 2010
Accepted 2 September 2010
Available online 16 September 2010

Keywords:
Dataset storage
Scientific workflow
Cloud computing
Cost benchmarking

ABSTRACT

Many scientific workflows are data intensive: large volumes of intermediate datasets are generated during their execution. Some valuable intermediate datasets need to be stored for sharing or reuse. Traditionally, they are selectively stored according to the system storage capacity, determined manually. As doing science on clouds has become popular nowadays, more intermediate datasets in scientific cloud workflows can be stored by different storage strategies based on a pay-as-you-go model. In this paper, we build an intermediate data dependency graph (IDG) from the data provenances in scientific workflows. With the IDG, deleted intermediate datasets can be regenerated, and as such we develop a novel algorithm that can find a minimum cost storage strategy for the intermediate datasets in scientific cloud workflow systems. The strategy achieves the best trade-off of computation cost and storage cost by automatically storing the most appropriate intermediate datasets in the cloud storage. This strategy can be utilised on demand as a minimum cost benchmark for all other intermediate dataset storage strategies in the cloud. We utilise Amazon clouds' cost model and apply the algorithm to general random as well as specific astrophysics pulsar searching scientific workflows for evaluation. The results show that benchmarking effectively demonstrates the cost effectiveness over other representative storage strategies.

1. Introduction

Scientific applications are usually complex and data intensive. In many fields, such as astronomy [14], high-energy physics [24] and bioinformatics [27], scientists need to analyse terabytes of data either from existing data resources or collected from physical devices. The scientific analyses are usually computation intensive, hence taking a long time for execution. Workflow technologies can be facilitated to automate these scientific applications. Accordingly, scientific workflows are typically very complex. They usually have a large number of tasks and need a long time for execution. During the execution, a large volume of new intermediate datasets will be generated [15]. They could be even larger than the original dataset(s) and contain some important intermediate results. After the execution of a scientific workflow, some intermediate datasets may need to be stored for future use because: (1) scientists may need to re-analyse the results or apply new analyses on the intermediate datasets; (2) for collaboration, the intermediate results may need to be shared among scientists from different institutions and the intermediate datasets may need to be reused. Storing valuable intermediate datasets can save their regeneration cost when they are reused, not to mention the waiting time saved by avoiding regeneration. Given the large sizes of the datasets, running scientific workflow applications usually need not only high-performance computing resources but also massive storage [15].

Nowadays, popular scientific workflows are often deployed in grid systems [24] because they have high performance and massive storage. However, building a grid system is extremely expensive and it is normally not an option for scientists all over the world. The emergence of cloud computing technologies offers a new way to develop scientific workflow systems, in which one research topic is cost-effective strategies for storing intermediate datasets.

In late 2007, the concept of cloud computing was proposed [32] and it is deemed the next generation of IT platforms that can deliver computing as a kind of utility [11]. Foster et al. made a comprehensive comparison of grid computing and cloud computing [17]. Cloud computing systems provide high performance and massive storage required for scientific applications in the same way as grid systems, but with a lower infrastructure construction cost among many other features, because cloud computing systems are composed of data centres which can be clusters of commodity hardware [32]. Research into doing science and data-intensive applications on the cloud has already commenced [25], such as early experiences like the Nimbus [21] and Cumulus [31] projects. The work by Deelman et al. [16] shows that cloud computing offers a cost-effective solution for data-intensive applications, such as scientific workflows [20]. Furthermore, cloud computing systems offer a new model: namely, that scientists from all over the world can

* Corresponding author.
E-mail addresses: dyuan@swin.edu.au (D. Yuan), yyang@swin.edu.au (Y. Yang), xliu@swin.edu.au (X. Liu), jchen@swin.edu.au (J. Chen).

0743-7315/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jpdc.2010.09.003

collaborate and conduct their research together. Cloud computing systems are based on the Internet, and so are the scientific workflow systems deployed in the cloud. Scientists can upload their data and launch their applications on the scientific cloud workflow systems from everywhere in the world via the Internet, and they only need to pay for the resources that they use for their applications. As all the data are managed in the cloud, it is easy to share data among scientists.

Scientific cloud workflows are deployed in a cloud computing environment, where use of all the resources need to be paid for. For a scientific cloud workflow system, storing all the intermediated datasets generated during workflow executions may cause a high storage cost. In contrast, if we delete all the intermediate datasets and regenerate them every time they are needed, the computation cost of the system may well be very high too. The intermediate dataset storage strategy is to reduce the total cost of the whole system. The best way is to find a balance that selectively stores some popular datasets and regenerates the rest of them when needed [1,36,38]. Some strategies have already been proposed to cost-effectively store the intermediate data in scientific cloud workflow systems [36,38].

In this paper, we propose a novel algorithm that can calculate the minimum cost for intermediate dataset storage in scientific cloud workflow systems. The intermediate datasets in scientific cloud workflows often have dependencies. During workflow execution, they are generated by the tasks. A task can operate on one or more datasets and generate new one(s). These generation relationships are a kind of data provenance. Based on the data provenance, we create an intermediate data dependency graph (IDG), which records the information of all the intermediate datasets that have ever existed in the cloud workflow system, no matter whether they are stored or deleted. With the IDG, we know how the intermediate datasets are generated and can further calculate their generation cost. Given an intermediate dataset, we divide its generation cost by its usage rate, so that this cost (the generation cost per unit time) can be compared with its storage cost per time unit, where a dataset's usage rate is the time between every usage of this dataset that can be obtained from the system logs. Then we can decide whether an intermediate dataset should be stored or deleted in order to reduce the system cost. However, the cloud computing environment is very dynamic, and the usages of intermediate datasets may change from time to time. Given the historic usages of the datasets in an IDG, we propose a cost transitive tournament shortest path (CTT-SP) based algorithm that can find the minimum cost storage strategy of the intermediate datasets on demand in scientific cloud workflow systems. This minimum cost can be utilised as a benchmark to evaluate the cost effectiveness of other intermediate dataset storage strategies.

The remainder of this paper is organised as follows. Section 2 gives a motivating example of a scientific workflow and analyses the research problems. Section 3 introduces some important related concepts and the cost model of intermediate dataset storage in the cloud. Section 4 presents the detailed minimum cost algorithms. Section 5 demonstrates the simulation results and the evaluation. Section 6 discusses related work. Section 7 is a discussion about the data transfer cost among cloud service providers. Section 8 addresses our conclusions and future work.

2. Motivating example and problem analysis

2.1. Motivating example

Scientific applications often need to process a large amount of data. For example, the Swinburne Astrophysics group has been conducting a pulsar searching survey using the observation data from the Parkes Radio Telescope, which is one of the most famous radio telescopes in the world [8]. Pulsar searching is a typical scientific application. It involves complex and time-consuming tasks and needs to process terabytes of data. Fig. 1 depicts the high-level structure of a pulsar searching workflow, which is currently running on the Swinburne high-performance supercomputing facility [30].

First, raw signal data from the Parkes Radio Telescope are recorded at a rate of one gigabyte per second by the ATNF [7] Parkes Swinburne Recorder (APSR) [6]. Depending on the different areas in the universe in which the scientists want to conduct the pulsar searching survey, the observation time is normally from 4 min to 1 h. Recording from the telescope in real time, these raw data files have data from multiple beams interleaved. For initial preparation, different beam files are extracted from the raw data files and compressed. They are 1–20 GB each in size, depending on the observation time. The beam files contain the pulsar signals which are dispersed by the interstellar medium. De-dispersion is used to counteract this effect. Since the potential dispersion source is unknown, a large number of de-dispersion files needs to be generated with different dispersion trials. In the current pulsar searching survey, 1200 is the minimum number of the dispersion trials. Based on the size of the input beam file, this de-dispersion step takes 1–13 h to finish, and it generates up to 90 GB of de-dispersion files. Furthermore, for binary pulsar searching, every de-dispersion file needs another step of processing named accelerate. This step generates accelerated de-dispersion files with a similar size in the last de-dispersion step. Based on the generated de-dispersion files, different seeking algorithms can be applied to search pulsar candidates, such as FFT Seeking, FFA Seeking, and Single Pulse Seeking. For a large input beam file, it takes more than one hour to seek the 1200 de-dispersion files. A candidate list of pulsars is generated after the seeking step, which is saved in a text file. Furthermore, by comparing the candidates generated from different beam files in the same time session, some interferences may be detected and some candidates may be eliminated. With the final pulsar candidates, we need to go back to the de-dispersion files to find their feature signals and fold them to XML files. Finally, the XML files are visually displayed to the scientists, for making decisions on whether a pulsar has been found or not.

As described above, we can see that this pulsar searching workflow is both computation and data intensive. It needs a long execution time, and large datasets are generated. At present, all the generated datasets are deleted after having been used, and the scientists only store the raw beam data extracted from the raw telescope data. Whenever there are needs to use the deleted datasets, the scientists will regenerate them based on the raw beam files. The generated datasets are not stored, mainly because the supercomputer is a shared facility that cannot offer unlimited storage capacity to hold the accumulated terabytes of data. However, it would be better if some datasets were to be stored, for example, the de-dispersion files, which are frequently used. Based on them, the scientists can apply different seeking algorithms to find potential pulsar candidates. Furthermore, some datasets are derived from the de-dispersion files, such as the results of the seek algorithms and the pulsar candidate list. If these datasets need to be regenerated, the de-dispersion files will also be reused. For large input beam files, the regeneration of the de-dispersion files will take more than 10 h. This not only delays the scientists from conducting their experiments, but also requires a lot of computation resources. On the other hand, some datasets need not be stored, for example, the accelerated de-dispersion files, which are generated by the accelerate step. The accelerate step is an optional step that is only used for binary pulsar searching. Not all pulsar searching processes need to accelerate the de-dispersion files, so the accelerated de-dispersion files are not that often used. In light of this, and given the large size of these datasets, they are not worth storing, as it would be more cost effective to regenerate them from the de-dispersion files whenever they are needed.