Future Generation Computer Systems 78 (2018) 353–368

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

pipsCloud: High performance cloud computing for remote sensing big data management and processing

Lizhe Wang[a,b], Yan Ma[b,*], Jining Yan[b], Victor Chang[c], Albert Y. Zomaya[d]

[a] *School of Computer Science, China University of Geoscience, Wuhan 430074, PR China*
[b] *Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, PR China*
[c] *Xi'an Jiaotong-Liverpool University, Suzhou, PR China*
[d] *School of Information Technologies, University of Sydney, Australia*

HIGHLIGHTS

- A Cloud-enabled HPC platform for large-scale RS applications.
- Hilbert-R^+ Tree based data indexing for optimal RS big data indexing.
- Collaborative large-scale RS workflow processing across data centers.
- Cloud-enabled virtual HPC environment with VMs and bare-metal provisioning.

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form
25 May 2016
Accepted 12 June 2016
Available online 23 August 2016

Keywords:
Big data
Cloud computing
Data-intensive computing
High performance computing
Remote sensing

ABSTRACT

Massive, large-region coverage, multi-temporal, multi-spectral remote sensing (RS) datasets are employed widely due to the increasing requirements for accurate and up-to-date information about resources and the environment for regional and global monitoring. In general, RS data processing involves a complex multi-stage processing sequence, which comprises several independent processing steps according to the type of RS application. RS data processing for regional environmental and disaster monitoring is recognized as being computationally intensive and data intensive.

We propose pipsCloud to address these issues in an efficient manner, which combines recent cloud computing and HPC techniques to obtain a large-scale RS data processing system that is suitable for on-demand real-time services. Due to the ubiquity, elasticity, and high-level transparency of the cloud computing model, massive RS data management and data processing for dynamic environmental monitoring can all be performed on the cloud via Web interfaces. A Hilbert-R^+-based data indexing method is employed for the optimal querying and access of RS images, RS data products, and interim data. In the core platform beneath the cloud services, we provide a parallel file system for massive high-dimensional RS data, as well as interfaces for accessing irregular RS data to improve data locality and optimize the I/O performance. Moreover, we use an adaptive RS data analysis workflow management system for on-demand workflow construction and the collaborative processing of a distributed complex chain of RS data, e.g., for forest fire detection, mineral resources detection, and coastline monitoring. Our experimental analysis demonstrated the efficiency of the pipsCloud platform.

1. Introduction

Remarkable advances in high-resolution Earth observation techniques have led to explosive increases in the volume and rate of remote sensing (RS) data [1]. The latest generation space-borne sensors are capable of generating continuous streams of observation data at a rate of several gigabytes per second [2] almost every hour, every day, and every year. The current volume of globally archived observation data probably exceeds one exabyte according to statistics in an OGC report [3]. The volume of RS data acquired by a regular satellite data center has increased by several terabytes per day, especially with high-resolution missions [4]. The use of high-resolution satellites with higher spatial, spectral, and temporal resolution of data inevitably leads to higher pixel

[*] Corresponding author.
E-mail address: mayan@radi.ac.cn (Y. Ma).

http://dx.doi.org/10.1016/j.future.2016.06.009
0167-739X/© 2016 Elsevier B.V. All rights reserved.

dimensionality. The current and future sensors are highly diverse, and RS data are commonly regarded as "*big RS data*" or "*big earth observation data*" in terms of the data volume as well as the complexity of data.

The proliferation of RS big data is revolutionizing the way that RS data are processed, analyzed, and interpreted to obtain knowledge [5]. In large-scale RS applications, regional or even global coverage multi-spectral and multi-temporal RS datasets are used for processing to meet the growing demands for more accurate and up-to-date information. Continental scale forest mapping normally involves the processing of terabytes of multi-dimensional RS datasets containing the available forest information [6]. Moreover, large-scale applications are also exploiting multi-source RS datasets to compensate for the limitations of a single sensor. Thus, massive data volumes as well as the increasing complexity of data are major issues. In particular, many time-critical RS applications demand real-time or near real-time processing capacities [7,8], such as in large-scale debris flow investigation ([9], flood hazard management [10]), and the surveillance of large ocean oil spills [11,12]. In general, the large-scale data processing problems in RS applications [4,13,14] with high quality of service (QoS) requirements are typically regarded as both computationally intensive and data intensive. Similarly, innovative analyses and high QoS requirements are driving improvements to traditional RS data processing systems. The timely processing of vast volumes of multi-dimensional RS data incurs unprecedented computational requirements, which far exceed the capabilities of conventional instruments. Employing the cluster-based high-performance computing (HPC) paradigm in RS applications is the most widespread and effective approach [15–19]. Both NASA's NEX system [5] for global processing and InforTerra's "Pixel Factory" [20] for massive imagery auto-processing use cluster-based platforms for QoS optimization.

However, despite their enormous computational capacities, cluster platforms that are not optimized for data-intensive uses still struggle to cope with huge data analysis and intensive data I/O. The mainstream multi-core clusters are characterized by a multilevel hierarchy and increasing scale. These HPC systems are almost out of reach for those without expertise in HPC because programming based on message passing interface (MPI)-enabled HPC platforms is not a simple task. Moreover, the main online processing needs are seldom satisfied. Thus, there is a lack of simple methods to help end users to exploit massive RS data processing capabilities in ubiquitous large-scale HPC environments. As a consequence, RS data processing typically involves multiple stages of data flow processes. On-demand processing also requires the ability to customize and serve dynamic processing workflows, instead of predefined static flows. In addition, the on-demand provision of resources will lead to unpredictable and volatile requirements for large-scale computing resources, and thus substantial investments will be essential to ensure that system upgrades and scale-out are maintained. In addition, the build and maintenance of these platforms is highly complex and expensive.

Cloud computing [21] provides scientists with a revolutionary paradigm for utilizing computing infrastructure and applications. Based on virtualization, computing resources and various algorithms can be accommodated, and delivered as ubiquitous on-demand services according to an application's requirements. The cloud paradigm has also been implemented widely in large-scale RS applications, such as the Matsu project [22] for cloud-based flood assessment. At present, clouds are rapidly being adopted in HPC systems such as clusters as variable scientific platforms [23]. Scientists can readily customize their HPC environment and access huge computing infrastructures in the cloud. However, compared with conventional HPC systems or even supercomputers, the clouds are not QoS-optimized large-scale platforms. Moreover, they differ from the traditional cloud because the data center clouds deployed in data-intensive RS applications need to facilitate massive RS data processing and intensive data I/O.

To address the issues mentioned above, we propose *pipsCloud*, which is a cloud-enabled high-performance RS data processing system for large-scale RS applications. In particular, we incorporate the cloud computing paradigm in cluster-based HPC systems to address various issues from a system architecture perspective. First, by employing an application-aware data layout optimized for data management and Hilbert R^+ tree-based data indexing, RS big data including imagery, interim data, and products can be managed and accessed by users in an efficient manner. Based on virtualization and bare-metal (BM) provisioning [24], virtual machines (VMs) and BM machines with lower performance penalties are deployed on-demand to allow easy scale up and out. Moreover, generic parallel programming skeletons are employed to simplify the programming of efficient MPI-enabled RS applications. In this manner, the cloud-enabled virtual HPC environment for RS big data processing can be dynamically encapsulated and delivered as online services. According to the dynamic scientific workflow technique, *pipsCloud* gives the ability to readily customize collaborative processing workflows for large-scale RS applications.

The remainder of this paper is organized as follows. In Section 2, we review some related research. In Section 3, we discuss the challenges related to building and enabling a high performance cloud system for data-intensive RS data processing. In Section 4, we explain the design and implementation of *pipsCloud* from a system-level perspective. In Section 5, we describe the experimental validation and analysis of *pipsCloud*. Finally, we give our conclusions in Section 6.

2. HPC for RS big data: State of the art

Each solution has its advantages and disadvantages. In this section, we compare the current dominant system architectures that are used frequently for RS data processing, including cluster-based HPC platforms and clouds. In Section 2.1, we explain the incorporation of a multi-core cluster HPC structure in RS data processing systems and applications. In Section 2.2, we introduce some new methods for enabling large-scale RS applications by exploiting the cloud computing paradigm.

2.1. Cluster computing for RS data processing

Growing numbers of improved sensor instruments are being incorporated in satellites for Earth observation and we are now entering an era of "RS big data." In addition, the urgent demands of large-scale RS problems with boosted computational requirements [5] have led to the widespread application of multi-core clusters. First, the NEX system [5] for global RS applications was built by NASA on a cluster platform with 16 computers during the middle of the 1990s. The "Pixel Factory" system [20] developed by InforTerra employed a cluster-based HPC platform for massive RS data auto-processing, especially ortho-rectification. These HPC platforms have been employed to accelerate hyperspectral imagery analysis [25]. It should be noted that the 10,240-CPU Columbia supercomputer[1] equipped with an InfiniBand network has been employed for RS applications by NASA.

Several traditional parallel paradigms are widely employed in multi-level hierarchy-featured cluster systems, i.e., the OpenMP

[1] Columbia Supercomputer at NASA Ames Research Center, http://www.nas.nasa.gov/Resources/Systems/columbia.html.