Computers & Geosciences 60 (2013) 109–116

Contents lists available at ScienceDirect

Computers & Geosciences

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

Process virtualization of large-scale lidar data in a cloud computing environment

Haiyan Guan [a], Jonathan Li [b,a,*], Liang Zhong [c], Yongtao Yu[b], Michael Chapman [d]

[a] *Department of Geography & Environmental Management, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1*
[b] *Key Laboratory for Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian 3610005, China*
[c] *Changjiang Spatial Information Technology Engineering Co., Changjiang Institute of Survey Planning Design and Research, Wuhan, Hubei 430010, China*
[d] *Department of Civil Engineering, Ryerson University, Toronto, Ontario, Canada, M5B 2K3*

ARTICLE INFO

Article history:
Received 14 December 2012
Received in revised form
12 July 2013
Accepted 15 July 2013
Available online 23 July 2013

Keywords:
Process virtualization
Lidar
Condor
Computing environment

ABSTRACT

Light detection and ranging (lidar) technologies have proven to be the most powerful tools to collect, within a short time, three-dimensional (3-D) point clouds with high-density, high-accuracy and significantly detailed surface information pertaining to terrain and objects. However, in terms of feature extraction and 3-D reconstruction in a computer-aided drawing (CAD) format, most of the existing stand-alone lidar data processing software packages are unable to process a large volume of lidar data in an effective and efficient fashion. To break this technical bottleneck, through the design of a Condor-based process virtualization platform, we presented in this paper a novel strategy that uses network-related computational resources to process, manage, and distribute vast quantities of lidar data in a cloud computing environment. Three extensive experiments with and without a cloud computing environment were compared. The experiment results demonstrated that the proposed process virtualization approach is promisingly applicable and effective in the management of large-scale lidar point clouds.

1. Introduction

Light detection and ranging (lidar) technologies, including airborne, mobile and terrestrial laser scanning (ALS, MLS, and TLS), have gradually become common mapping practices in three-dimensional (3-D) data acquisition. Existing lidar systems can provide point clouds with a point density of up to thousands of points per square metre. For example, an Optech® Lynx Mobile Mapper captures a total of 144 million points of five blocks in 20 min (Conforti and Zampa, 2011). This revolutionary data acquisition technology allows for various applications, such as transportation, utility, forestry, mining, and urban planning. In addition to the huge quantity of point clouds acquired by a laser scanner, multiple high-spatial-resolution cameras (as common components) provide a significantly large quantity of image data. For example, a Trimble® MX-8 system, integrating two RIEGL® VQ-250 laser scanners and four CCD cameras, collects a total of 35 Gigabytes in 20 min (Guan et al. , 2013a).

Furthermore, it is intricate and sophisticated to efficiently store, manage, and process lidar data and images for customized products, such as object identification (Secord and Zakhor, 2007; Guan et al., 2013b), 3D building models (Habib et al., 2005; Zhang et al., 2005; Mitishita et al., 2008; Nebiker et al., 2010) and Digital Orthophoto Maps (DOMs) (Liu et al., 2007) because creating these lidar-derived products are costly and intensive computation due to a variety of methods and a diverse selection of parameters. Current lidar data processing software packages (e.g. QT Modeler, LasTools, InPho, LiDAR Explorer for ArcGIS, Terrasolid, and TopPIT) are with stand-alone and task-oriented features that considerably limit lidar data applications (GIM, 2012); thus, much attention has been paid to research on organizing lidar data more effectively and efficiently.

Although a group of level-of-detail (LOD) variants have proven to be suitable for adaptive visualization of a notably large volume of lidar data, these variants are time-consuming in pre-processing and progressively inefficient in query performance because of their unbalanced tree structures (Pfister et al., 2000; Rusinkiewicz and Levoy, 2000). Moreover, a variety of tree structures, ranging from binary-tree, quad-tree, octree to their combinations, have been presented to accelerate the lidar data processing procedure,

* Corresponding author at: Key Laboratory for Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian 3610005, China. Tel.: +86 059 225 80003.
E-mail addresses: junli@xmu.edu.cn, junli@uwaterloo.ca (J. Li).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.07.013

(Lu and He, 2008; Liu et al., 2008; Kreylos et al., 2008; Elseberg et al. 2011). For example, the binary-tree is employed for building extraction (Sohn et al., 2008); the octree is used for split-and-merge segmentation (Wang and Tseng, 2010). Those two-dimensional (2-D) tree-based algorithms are limited by their unbalanced index structures and one-dimensional (1-D) space partitioning.

To overcome the above two limitations, 3-D R-trees, the most well-known index structure for spatial data, have been applied to the real data by adaptively adjusting index structures (Zhu et al., 2007). In Gong et al. (2012), a hybrid spatial index method (3DOR-Tree) that integrates R-Tree and Octree structures is used to overcome the unbalanced data distribution. Similar to B-tree, R-tree is a height balanced tree that hierarchically splits spaces into possibly overlapping subspaces. However, the tree-node overlapping and complexity of 3-D R-trees cause multipath queries, resulting in lower query efficiencies in lidar data processing.

In most cases, to achieve final lidar-derived products, existing established lidar processing algorithms and software tools must divide substantial amounts of lidar points into a number of data blocks (Pu et al., 2011) and thin out or rasterize the lidar data for a series of post-processing procedures (Van Gosliga et al., 2006; Mongus and Zalik, 2012). Besides software limitations, computer hardware, such as the amount of computer memory ranging from a few hundred megabytes to gigabytes, is usually unable to support intensive calculation requirements if couples of threads are simultaneously implemented for lidar data processing on a multi-processor computer. Thus, a lidar data processing system requires an open, shared and interoperated environment, where data processing, management, and distribution are automatic, intelligent, and real-time. To this end, advanced techniques like parallel processing are used to improve large-scale lidar data processing efficiency by distributing them among multiple shared-servers (Wand et al., 2008; Ma and Wang, 2011). First, through data index structures, such as grid, quadtree, and octree, those techniques, regarding spatial relationship of the datasets, uniformly divide and distribute mass remotely sensed data into data servicers,. A client retrieves the data from the data servers directly and efficiently according to data spatial locations and boundaries, which enables the client to maximize the capability of parallel computing. With advances in Grid Computing, Cloud Computing is a promising choice of massive remotely sensed data processing (Xue et al., 2011). Therefore, besides a high throughput computation and grid workflow for remote sensing quantitative retrieval applications, a cloud computing environment is motivated by the requirement for customized remote sensing products, especially lidar-derived products.

In this paper, we design a Condor-based virtual platform of massive remotely sensed data processing for lidar-derived products in a cloud computing environment, and analyze the platform's performance on three common uses of lidar processing techniques: filtering, DEM interpolation, and DOM generation. Specifically, we take advantage of cloud computing, one of the newest internet-based paradigms in computation in the field of lidar data processing to accomplish the following: (1) solve the expanding data and task-intensive computation problems encountered within customized lidar-derived products as the demand for these products increases; (2) develop a prototype of a Condor-based process virtualization platform for large-scale remotely sensed data processing, analysis, and intensive computation; (3) provide an efficient method for users to make full use of various idle internet resources and established lidar-relevant data processing algorithms; (4) explore the potential for improving the parallel efficiency of our Condor-based process virtualization platform by synchronizing computation and communication procedures.

The remainder of the paper is organized as follows: Section 2 introduces cloud computing. Section 3 presents a Condor-based middleware design for lidar data processing. Section 4 discusses extensive experimental results and evaluates the performance of the proposed process virtualization platform in the cloud computing environment. Finally, Section 5 states the concluding remarks.

2. Cloud computing

The “cloud”, a natural evolution of distributed computing, is of the Web 2.0 protocol and is a particularly widespread virtualization technology. Associated with a new paradigm for the provision of computing infrastructure, cloud computing shifts infrastructure locations from the desktop to the network (Boss et al., 2007; Vaquero et al., 2008). Those network-related capabilities and resources are provided as services, via the on-demand and accessible internet without knowing the detailed knowledge of the underlying technology (Bolze and Deelman, 2010). Cloud computing in the early development period was called “Grid Computing” – a term that originated in the 1990s. The main research of Grid Computing ranged from Giga Ethernet testbed to Metacomputing. In Smarr and Catlett (1992), Metacomputing, focusing on managing and harnessing heterogeneous computational resources, is considered as the prototype of Grid Computing. Typical representative projects of Metacomputing included FAFNER (an internet-based sieving effort from Cooperating Systems Corporation), I-WAY, and Information Wide-Area Year (Foster et al., 1996). FAFNER was followed by distributed projects such as SETI@home (Korpela et al., 2001) and Distributed.Net; whereas, Globus (a toolkit of middleware components for Grid Computing infrastructure) (Foster et al., 2001) and Lehion (an object-based approach to Grid Computing) projects were based on I-WAY. The 1990s period of Grid Computing was characterised by the use of distributed interconnected computers and resources collectively to achieve high performance computational capabilities and resource sharing (Wilkinson, 2010). Grid computing technology subsequently evolved into a wider range of science and engineering disciplines, including biomedical research, industrial research, high-energy physics, bioinformatics, chemistry, earth science, and geometric modelling. In 2001, the Open Grid Services Architecture (OGSA), a new generation of grid structure as a standard of Grid Computing, was originally proposed by Foster et al. (2003) to integrate web services supported by industrial communities with computation services (Foster and Kesselman, 2003).

Network-based remotely sensed data management and distribution systems have achieved significant breakthroughs in commercial software (e.g., Lockheed Martin's Intelligent Library System, the Microsoft Terraserver, Z/I Imaging corporation's TerraShare) and scientific research platforms (e.g., Graz Distributed Server System(GDSS), Data and Information Access Link (DIAL)). However, to the best of our knowledge, most systems and commercial products concentrate on data retrieving, distribution, and storage. Little attention has been paid to fast and effective data processing and analysis services for such a huge volume of remotely sensed data, such as lidar point clouds in particular.

Virtualization is defined as creating something virtually or non-existent rather than having an actual physical version. The process virtualization of remotely sensed data in cloud computing environments is a system that connects internet-oriented service technology with a remote-sensing database for data retrieval, processing and feedback. There are two ways to virtualize lidar data in a cloud computing environment. One is a tightly coupled model for parallel computation; the other is a loosely coupled model in distributed computation. Physical and software interactions are highly inter-dependent with stable and robust communications in the tightly coupled model; while in a loosely coupled architecture, a significantly large number of interactions operate independently in different geographical positions, leading to costly and unreliable communications. Meanwhile, a successful remotely sensed data