一种基于云计算模型的遥感处理服务模式研究与实现 *

刘 异, 吴 维, 江万寿, 龚健雅
(武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430074)

摘 要: 随着空间遥感技术、对地观测技术的不断发展,一个以多时相、多分辨率、多传感器、多波段为特征的多层、立体、多角度、全方位和全天候遥感对地观测数据获取与处理体系正在形成。该体系必然会带来海量、多源的遥感数据。提出了采用目前商业上成功的云计算模型来实现一个高性能、高可扩展性、高可用的遥感处理服务,并结合原型系统,详细阐述了该处理系统的组成与关键技术。

关键词: 分布式计算; 遥感; MapReduce; 并行; 服务模型; 云计算
中图分类号: TP208 **文献标志码**: A **文章编号**: 1001-3695(2009)09-3428-04
doi:10.3969/j.issn.1001-3695.2009.09.064

Research of remote sensing service based on cloud computing mode

LIU Yi, GUO Wei, JIANG Wan-shou, GONG Jian-ya
(*State Key Laboratory for Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan* 430074, *China*)

Abstract: As the development of technologies of earth observation, receiving and processing systems for earth-observed data are emerging. Managing the large volume and multi-source data is a key problem for the earth observation system. This paper analyzed several modes of remote sensing service and proposed adopting the cloud computing mode to implement remote sensing service for better scalability, performance and stability. Showed the prototype system named OpenRS-Cloud to illustrate the system framework and key technologies.

Key words: distributed computing; remote sensing; MapReduce; parallel computing; service mode; cloud computing

0 引言

2006—2020 年的《国家中长期科技发展规划纲要》中明确指出:发展基于卫星、飞机和平流层飞艇的高分辨率先进对地观测系统,发射一系列的高分辨率遥感对地观测卫星,建立覆盖可见光、红外、多光谱、超光谱、微波、激光等观测谱段的、高中低轨道结合的、具有全天时、全天候、全球观测能力的大气、陆地、海洋先进观测体系。到 2020 年,建成稳定的运行系统,提高我国空间数据的自给率,形成空间信息产业链条[1]。

在这样的环境和要求下,建立一个能快速、高效、并行、并功能可扩展的遥感处理服务是必需的,也是必要的。而之前的各种遥感处理系统有些虽然具备了海量数据并行处理能力,但是其系统并不开放,并且并行实现手段多基于高耦合性的 MPI[2~4],算法本身和并行模型强耦合并没有从框架上考虑第三方算法的集成[5,6]。而目前主流 IT 界提出的云计算模型[7],是一种具备处理规模化、管理集中化、功能开放化、存储海量化和客户端轻量化等特点的新型服务计算模型,最初由 Google 发起[8],目前已经得到微软、雅虎、亚马逊等 IT 各大主流公司的支持,在通用 IT 领域取得了极大的成功。本文将详细讨论在云计算模型下的遥感处理服务的必要性、系统的组成和实现技术,为大规模遥感处理服务的实现提供一种新的思路。

1 遥感处理云计算模式的必要性

虽然云计算在通用 IT 领域已经取得了极大的成功,但是在探讨遥感领域的云计算模式前,需要详细地讨论遥感处理领域是否需要云计算。可以将遥感处理领域的相关人员分为四类,即平台架构者、算法开发者、数据提供者和最终用户。平台架构者侧重于提供基础的遥感数据读写、通用处理算法和用户界面等;算法开发者侧重于实现算法逻辑,实现过程中可采用平台提供的通用方法;数据提供者侧重于提供原始数据;最终用户在平台架构者提供的界面上,通过算法提供者的算法来处理数据提供者所提供的数据,形成最终所需要的数据文件。图 1 阐述了四者之间的关系。

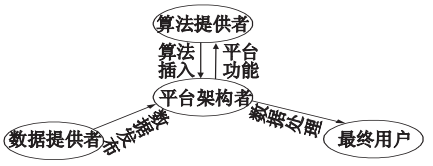

图1 遥感处理领域角色关系

虽然现在的遥感平台有很多,但是都符合上述模型,只是

收稿日期: 2009-01-08; **修回日期**: 2009-02-16 **基金项目**: 国家"863"计划资助项目(2007AA120203)
作者简介:刘异(1983-),女,江西人,博士研究生,主要研究方向为分布式地理信息系统、P2P、遥感服务框架、GeoSimulation 等(liuyiwhu@gmail.com).

有些平台比较封闭,并不支持第三方算法插件。目前绝大部分遥感平台有一个共同点,就是并不提供网络服务,都是以桌面软件形式提供给用户使用,而网络服务的形式已经在 GIS 领域被广泛研究和使用,如著名的 ArcGIS 软件已经成功地实现将原桌面版本的所有功能移植成为 Web 服务。众所周知,网络服务比普通单机桌面版具有更强大的处理能力,更易集成和更开放,但是目前遥感处理为什么没有服务的大量涌现,笔者认为主要有如下几个原因:

a)资源紧缺。遥感处理涉及到海量数据和庞大的计算能力,普通的单服务器计算模式并不能满足要求,需要采用集群环境,而普通开发者并没有适合的硬件环境。

b)开发困难。缺乏适合遥感处理的 Web 开发平台。目前 GIS 服务多数基于 Web service 相关平台。由于微软、Sun 等公司提供了一个简单利用的 Web service 开发环境,开发比较简单。但是普通的 Web service 只针对单个请求的响应时间比较短的情况,而单个遥感处理请求可能需要 10 min 或更长的处理时间,普通的 Web service 开发环境并不适合。

c)缺乏统一平台级商业模式。商业模式对于注重算法产权的遥感领域是很重要的。而长期以来,如何保护普通算法开发者权益的问题一直没得到重视和解决。只有统一的平台级商业模式才能引导遥感处理领域的算法发展。

上述问题是阻碍遥感处理服务发展的根本原因。针对上述三个问题,如果采用云计算模式,可以得到良好的解决。

1)资源紧缺　云计算提供统一并强大的计算环境和硬件资源,普通算法开发者可以将自己的算法按照接口规范提交到平台即可,而无须自己架构网站。

2)开发困难　云计算提供的统一接口可使算法开发变得简单,算法开发者无须知道网络通信相关知识即可被最终用户所调用。针对这点,后面将详细阐述 OpenRS-Cloud 的解决方案。

3)缺乏统一平台级商业模式　云计算的商业模式已经得到成功,亚马逊公司已经盈利上亿美金。在云计算中,数据提供者、算法提供者和平台提供者将从最终用户的每次请求中获利,进行合理的分成,最终用户也拥有了更多的算法和数据选择权利。

结合前面对云计算特点的描述,有理由相信将云计算应用于遥感处理是可行的。下面将详细阐述遥感处理云计算的系统构成与关键技术。

2　遥感处理云计算系统构成

根据上面的分析,本章提出一种遥感处理的云计算模式,并在后面利用原型系统加以验证。图 2 是设想的遥感云计算模式。

从图 2 中可以看出该模式分为下面几个系统:

a)Web 入口系统。该系统主要面向最终用户,利用 Ajax、Flash 或 Slivelight 技术,为用户提供一个直接在 Web 浏览器上就能使用的用户界面。

b)多任务并行作业系统。该系统采用 Google 云计算的并行模式 MapReduce,通过集成大量的服务器或普通 PC 机提供一个针对海量用户的分布式任务调度和管理环境。

c)WebGIS 数据发布与管理系统。由于待处理的影像是通过浏览器展示给用户的,需要 WebGIS 系统来进行数据的发布和元数据查询。

d)算法插件系统。为算法提供者制订统一的接口方式和插件开发方法。

e)分布式数据存储系统。由于云计算涉及的数据庞大,需要类似 Google 的分布式文件系统来进行数据的存储和管理。数据提供者调用该系统来发布自己的数据,而 WebGIS 数据发布系统会根据分布式数据存储系统的变化定时发布新的数据到网络上。

f)计费系统。当最终用户享用数据下载或数据计算服务后,需要付一定的费用,而这些费用将按照一定的比例分配给算法提供者、数据提供者和平台系统。这种商业模式已经被 Google、亚马逊成功应用。

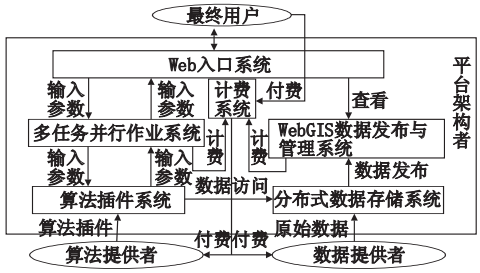

图2　遥感处理服务云计算模式

上述几大系统是遥感处理云计算平台的基础系统,缺一不可,通过这些系统提供的数据发布、数据搜索、多用户计算、海量存储和 Web 浏览器访问来为其他三类用户提供良好的生态环境。最终用户得到了一个不断扩展功能,并且拥有海量处理能力和存储能力的系统;算法提供者不用关心计算能力和硬件条件的缺乏,而只需关注本身算法的有效性,并且底层的数据读取、通用算法、矩阵等操作均由平台提供;数据提供者利用平台提供的工具上传数据,平台将自动进行元数据录入和数据发布;当最终用户付给相应的费用后,三方均能得到一定的收益。遥感处理云计算平台并不以建立算法完备的系统为目标,而是希望建立一个良好的具有计算能力、存储能力和发布能力的生态环境来吸引用户群。除了上面的基础系统外,一个上线运行的系统还需配置其他辅助子系统,如用户权限管理、数据上传下载系统等。后面将结合 OpenRS-Cloud 原型系统来详细介绍各个系统的实现方法、采用的技术手段和界面设计。

3　符合遥感处理云计算模式的原型系统设计与实现

上面阐述了遥感处理云计算模式的基本组成,本章将结合名为 OpenRS-Cloud 原型系统作更深入的阐述。OpenRS(开放式遥感处理平台)系统是国家“863”十一重点项目多源遥感数据综合处理与服务系统第二课题的研究目标。OpenRS-Cloud 是 OpenRS 的子项目,希望可以利用廉价 PC 机群进行高性能

遥感数据处理的云计算系统,主要提供任务并行化处理、Web 监控、分布式存储、算法插件化和多语言接口等功能。OpenRS-Cloud 的系统架构是符合遥感处理云计算模式的,拥有五大核心系统(目前未实现计费系统)。下面将阐述 OpenRS-Cloud 五大核心系统的实现方法和技术手段。

3.1　Web 入口系统

Web 入口系统是最终用户与其他子系统的桥梁。OpenRS-Cloud 提供了一个基于 Web 浏览器、利用 Google Web Toolkit 开发的 Ajax 用户界面。从图 3 中可以看出,整个界面分为任务管理、插件管理、数据文件选择和元数据搜索四个部分。

任务管理主要负责计算任务的算法选择、参数输入、任务状态监控和相关信息查看,如图 4 所示。

插件管理主要负责管理和注册用户的插件系统,用户可以通过该界面上传的第三方插件到服务器集群系统上来进行相关处理,如图 5 所示。

图3　OpenRS-Cloud界面布局

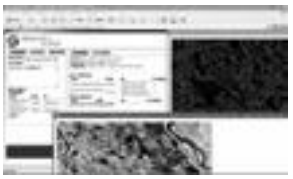

图4　任务提交界面布局

图5　插件管理界面布局

数据文件选择的主体是一张全球地图,而图上的每个红点代表一份原始影像数据,每个影像数据具有覆盖范围、地点、描述、缩略图和提供者等信息,这些数据全部位于后台的分布式存储系统中,通过后台的 WebGIS 系统(目前采用 Google Map)发布在地图上(目前采用 KML),每个数据文件可通过点击来作为算法输入参数参与计算,这种方式取代了传统遥感软件的本地目录查找与选择,结合地图搜索和元数据搜索能让用户更方便地查找所需数据,如图 3 所示。

元数据搜索采用全文搜索技术,在通过针对各个元数据的描述信息进行倒排索引后,可提供全文搜索能力。该项功能可以与地图位置搜索联合使用,实现空间与属性的联合查询。

OpenRS-Cloud 提供了一个简单有效的界面模式,目前没有考虑算法本身(如交互式地物提取)更加复杂的交互功能,但是对于遥感处理的大多数算法已经适合。针对某些交互性强的特定算法,可采用 Flash 或 Slivelight 等富客户端技术来实现。

3.2　算法插件系统

OpenRS-Cloud 的底层算法插件系统采用 OpenRS 系统。OpenRS 系统是一个具备可扩展能力的插件系统。该系统分为如图 6 所示的四个层次,即内核插件系统、服务插件、对象插件和应用程序。详细描述参见文献[9]。

1)插件系统　提供四大基础服务,这里的服务概念是指在整个 OpenRS 运行生命周期内惟一的对象,可创建其他对象。

2)服务插件　提供通用服务功能,如序列化、空间参考、矩阵计算等。

3)对象插件　提供可创建式对象,实现文件读取、分类等各种算法。插件对象管理采用插件层次树的形式,如图 7 所示,每个对象都根据其接口继承关系位于插件树上。OpenRS 系统采用纯接口开发,算法提供者能很方便地实现符合 OpenRS 接口规范的各种算法插件,并且每个算法提供者还可以利用平台提供或其他用户提供的算法插件。

4)应用程序　桌面级应用,提供桌面级的支持可扩展性点的 GUI 界面。

图6　OpenRS插件系统构成图

图7　OpenRS插件接口层次树

OpenRS-Cloud 能直接调用实现了 orsIExecute 或 orsIParallelExecute 接口的 OpenRS 算法插件。orsIExecute 是指非并行算法,算法的输入和输出为具有 XML 类似层次接口的属性。在计算过程中,算法可以输出进度和日志信息。这些信息将直接反馈给 Web 用户。orsIParallelExecute 是指那些可以进行子任务分解和合并的并行算法,这些算法分解的子任务将会并行运行在由多任务并行作业系统所分配的节点机上,每个子任务依然可以输出进度和日志信息,orsIExecute 可以看做是 orsIParallelExecute 的特例。

在 OpenRS 系统中,算法参数的输入和输出采用 XML 描述。因为采用 XML 参数描述,可以让 OpenRS-Cloud 系统自动地生成参数界面系统,如图 3 所示。目前 OpenRS 插件系统内置支持整数、浮点、字符串、文件名、数组和矩阵几种类型,并且可方便地扩展到空间点、线、面和用户自定义类型的支持,Web 入口系统可根据参数类型生成相应的界面元素。

从上述描述可以看出,算法实现者并不需要关心网络通信与资源分布情况,只需实现算法逻辑即可。实际上实现了 orsIExecute或 orsIParallelExecute 的算法不仅可以在 OpenRS-Cloud 中运行,也可以在桌面版的 OpenRS 系统中运行,这正是本文采用的纯接口的方式而带来的网络化与桌面化运行的无缝集成。

3.3　多任务并行作业系统

多任务并行作业系统是 OpenRS-Cloud 系统的计算核心,设计模型采用 MapReduce 模型。MapReduce 模型最早被 Google 资深科学家 Jeff Dean 用于分布式计算,他在处理 Google 的日常工作时,认识到 Google 所需的绝大部分数据处理都可以归结为一个简单的并行算法:MapReduce。这个算法能够在

很多种计算中达到相当高的效率,而且是可扩展的。MapReduce 的两个主要概念是 map(映射)和 reduce(化简),主要思想是从函数式编程语言里借来的。简单而言,map 函数是用来把一组键值对映射成一组新的键值对,用于子任务执行;而 reduce 函数是用来保证所有映射的键值对中的每一个共享相同的键组,用于子任务处理结果的合并。MapReduce 会生成大量的临时文件,为了提高效率,一般采用分布式文件系统来管理和访问这些文件。

多任务并行作业系统通过结合 OpenRS 插件系统与 MapReduce并行计算模型,为各种遥感算法提供一个通用的并行算法开发环境。系统的任务执行流程分为三个阶段,即任务分解、子任务执行(map 过程)和子任务合并(reduce 过程),如图 8 所示。

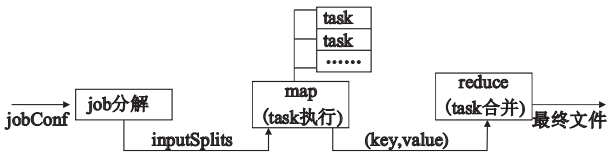

图8　任务抽象执行流程

在多任务并行作业系统中,笔者将任务称为 job,而子任务称为 task。每个 job 在经过任务分解过程后,形成多个 task,每个 task 具有分离输入(inputSpilit),系统根据优先级将子任务发送到各个处理服务器进行并行处理,最终结果汇集在一台服务器来做最终的合并工作。在上述过程中,map 所产生的中间结果存储在分布式文件系统中,用于最终 reduce 操作的数据读取和处理。图 9 是任务运行界面图。从图中可以看出,每个用户提交的任务执行情况非常清楚,如同桌面系统一样,惟一不同的是各个任务同时运行在后台多个节点机上。

图9　多任务并行作业系统任务监视

3.4　WebGIS 数据发布与管理系统

WebGIS 数据发布与管理系统主要提供数据底图的显示、元数据的查找和数据发布。目前数据底图采用 Google map。后台的元数据服务按照相关地图浏览或查询指令自动生成 KML 显示在 Google map 上。本系统为成熟的 WebGIS 系统,这里不再详细阐述。

3.5　分布式数据存储系统

分布式数据存储系统采用 Hadoop 的 HDFS 分布式文件存储系统。由于目前遥感处理算法开发包只针对本地文件系统,而不能直接访问 HDFS 分布式文件存储系统,需要将 hdfs 映射为本地文件系统。该技术可采用 WebDAV(Web-based distributed authoring and versioning)。该协议是基于 HTTP 1.1 的一个通信协议。它为 HTTP 1.1 添加了一些扩展,使得应用程序可以直接将文件写到 Web server 上,并且在写文件时可以对文件加锁,写完后对文件解锁,还可以支持对文件所做的版本控制。由于该技术与 IT 领域的 Cloud 计算完全一致,不作更多阐述。

4　遥感云计算与网格计算

虽然 OpenRS-Cloud 是一整套云计算系统,与网格计算在模式上有所不同,但是从技术上而言,是完全可以与之共存的。因为网格计算以集成 Web service 服务见长,所以 OpenRS-Cloud 可通过提供一套以 WSDL 定义的任务提交与状态查询接口来作为网格计算的一个节点参与到网格计算中。具体的实现模式如图 10 所示。

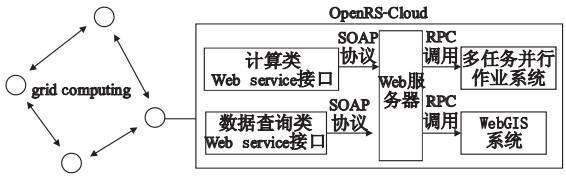

图10　OpenRS-Cloud 与 grid computing的关系

5　结束语

云计算作为目前 IT 领域的新兴技术,极大地吸引了各种 IT 厂商、普通开发者和使用者,在短短的时间内形成一种崭新的商业模式,并取得了极大的成功,这是之前单纯的 Web service或者网格计算技术所没有达到的。作为遥感处理专业领域,如何进行海量数据存储与处理、系统的扩展与开放等是该领域长期的瓶颈,云计算模式的出现给解决这些问题带来了希望。本文详细探讨遥感云计算的系统构成和实现方法,并以一个具体的原型系统展现了遥感云计算模式的用户界面、技术手段与运行流程。

参考文献:

[1]　国家中长期科学和技术发展规划纲要(2006—2020)[EB/OL]. http://www.gov.cn/jrzg/2006-02/09/content_183787.htm.

[2]　黄国满,郭建峰.分布式并行遥感影像处理中的数据划分[J].遥感信息,2001(2):9-12.

[3]　卢丽君,廖明生.分布式并行计算技术在遥感数据处理中的应用[J].测绘信息与工程,2005,30(3):1-3.

[4]　周海芳.遥感图像并行处理算法的研究和应用[D].长沙:国防科学技术大学,2003.

[5]　蒋利顺,刘定生.遥感图像 K-means 并行算法研究[J].遥感信息,2008(1):27-30,115.

[6]　王攀峰,杜云飞,周海芳,等.基于复小波变换的遥感图像并行融合算法[J].计算机工程与科学,2008,30(3):35-39.

[7]　云计算[EB/OL]. http://en.wikipedia.org/wiki/Cloud_computing.

[8]　JEFFREY D, SANJAY G. MapReduce: simplified data processing on large clusters[C]//Proc of the 6th Symposium on Operating System Design and Implementation. San Francisco, CA:[s.n.], 2004.

[9]　OpenRS 插件系统[EB/OL]. http://www.openrs.org/.

[10]　沈占峰.高分辨率遥感影像并行处理数据分配策略研究[J].哈尔滨工业大学学报,2006,38(11):1968-1971,1976.