云计算及其关键技术

● 邓倩妮 上海交通大学计算机系 上海 200040
● 陈 全 上海交通大学计算机系 上海 200040

摘要：

论文对新兴的计算模型——云计算进行了简要的介绍。论文给出了云计算的定义，介绍了云计算的发展背景和应用场景，分析了云计算和网格计算以及传统超级计算的区别，总结了云计算的关键技术：存储技术、数据管理技术以及编程模型。

关键词：云计算; 数据存储; 数据管理; 编程模型

1. 云计算产生背景及定义

1.1 云计算的定义

云计算(Cloud Computing)是一种新近提出的计算模式。维基百科给云计算下的定义：云计算将IT相关的能力以服务的方式提供给用户，允许用户在不了解提供服务的技术、没有相关知识以及设备操作能力的情况下，通过Internet获取需要服务[1]。

中国云计算网将云定义为：云计算是分布式计算（Distributed Computing）、并行计算（Parallel Computing）和网格计算（Grid Computing）的发展，或者说是这些科学概念的商业实现[2]。

Forrester Research 的分析师 James Staten 定义云为:“云计算是一个具备高度扩展性和管理性并能够胜任终端用户应用软件计算基础架构的系统池”。

虽然目前云计算没有统一的定义，结合上述定义，可以总结出云计算的一些本质特征，即分布式计算和存储特性，高扩展性，用户友好性，良好的管理性。云计算技术具有以下特点：

(1) 云计算系统提供的是服务。服务的实现机制对用户透明，用户无需了解云计算的具体机制，就可以获得需要的服务。

(2) 用冗余方式提供可靠性。云计算系统由大量商用计算机组成机群向用户提供数据处理服务。随着计算机数量的增加，系统出现错误的概率大大增加。在没有专用的硬件可靠性部件的支持下，采用软件的方式，即数据冗余和分布式存储来保证数据的可靠性。

(3) 高可用性。通过集成海量存储和高性能的计算能力，云能提供一定满意度的服务质量。云计算系统可以自动检测失效节点，并将失效节点排除，不影响系统的正常运行。

(4) 高层次的编程模型。云计算系统提供高级别的编程模型。用户通过简单学习，就可以编写自己的云计算程序，在“云”系统上执行，满足自己的需求。现在云计算系统主要采用Map-Reduce模型。

(5) 经济性。组建一个采用大量的商业机组成的机群相对于同样性能的超级计算机花费的资金要少很多。

1.2 云计算的应用场景

云计算有着广泛的应用前景。如表1所示：

表1 云计算的应用领域

领域	应用场景
科研	地震监测
	海洋信息监控
	天文信息计算处理
医学	DNA信息分析
	海量病历存储分析
	医疗影像处理
网络安全	病毒库存储
	垃圾邮件屏蔽
图形和图像处理	动画素材存储分析
	高仿真动画制作
	海量图片检索
互联网	Email服务
	在线实时翻译
	网络检索服务

云计算在天文学[7]、医学等各个领域有着广泛的应用前景。

趋势科技和瑞星等安全厂商纷纷提出了“安全云”计划。如今，每天有2万多种新的病毒和木马产生，传统的通过更新用户病毒库的防毒模式，受到了严峻的挑战，用户端的病毒库将过于庞大。趋势科技和瑞星的“安全云”将病毒资料库放在“云”端，与客户端通过网络相连，当“云”在网络上发现不安全链接时，可以直接形成判断，阻止其进入用户机器，从根本上保护机器的安全。

据趋势科技大中华区执行总裁张伟钦介绍，趋势科技已投入了大量资金，在全球数个地方建设了新型数据中心。同时，趋势科技还花费了1000多万美元，租借了34000多台服务器，构建了一个服务遍及全球的“安全云”。目前趋势科技已将公司中低端的部分产品线放到“云安全”计划中，而高端的大部分产品线，仍在准备过程中。

谷歌提供的Gmail、Google Earth、Google Analytics等服务都基于其云计算服务器运行[8]。谷歌基于云计算提供的翻译服务具有现今最好的性能[9]。对互联网和美国人生活的一项研究显示，大约70%的在线用户使用以上“云计算”服务。

1.3 云计算的发展

目前，亚马逊，微软，谷歌，IBM，Intel等公司纷纷提出了“云计划”。例如亚马逊的AWS（Amazon Web Services)[3]、IBM和谷歌联合进行的“蓝云”计划等。这对云计算的商业价值给予了巨大的肯定。同时学术界也纷纷对云计算进行深层次的研究。例如谷歌同华盛顿大学以及清华大学合作，启动云计算学术合作计划(Academic Cloud Computing Initiative)，推动云计算的普及，加紧对云计算的研究。美国卡耐基梅陇大学等提出对数据密集型的超级计算（DISC：Data Intensive SuperComputing）进行研究，本质上也是对云计算相关技术开展研究。

IDC的调查显示，未来五年云计算服务将急速增长，预期2012年市场规模可达420亿美元。目前企业导入云计算已逐渐普及，并且有逐年成长趋势。估计在2012年，企业投入在云计算服务的支出将占整体IT成本的25%，甚至在2013年提高至IT总支出的三分之一。

由此可见，在各大公司以及学术界的共同推动下，云计算技术将会持续发展。

1.4 云计算与其他超级计算的区别

1.4.1 云计算与网格计算的区别

Ian Foster 将网格定义为：支持在动态变化的分布式虚拟组织（Virtual Organizations）间共享资源，协同解决问题的系统[4]。所谓虚拟组织就是一些个人、组织或资源的动态组合。

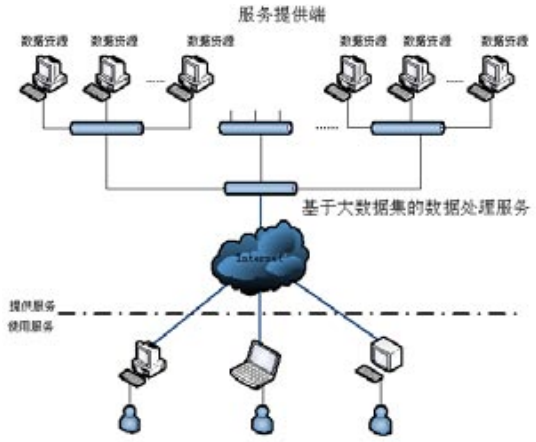

图1“云”系统的结构

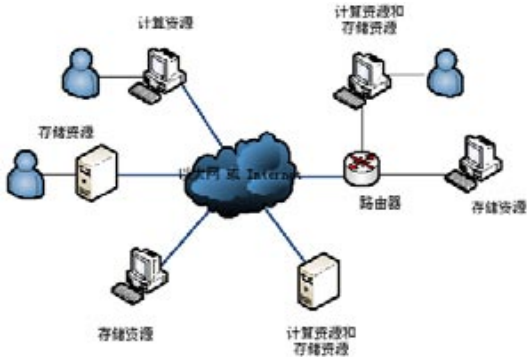

图2 网格的结构

图1和图2分别为云及网格的结构示意图。图1 显示，云计算是一种生产者—消费者模型，云计算系统采用以太网等快速网络将若干机群连接在一起，用户通过因特网获取云计算系统提供的各种数据处理服务。图2显示，网格系统是一种资源共享模型，资源提供者亦可以成为资源消费者，网格侧重研究的是如何将分散的资源组合成动态虚拟组织。

云计算和网格计算的一个重要区别在于资源调度模式。云计算采用机群来存储和管理数据资源，运行的任务以数据为中心。即调度计算任务到数据存储节点运行。而网格计算，则以计算为中心。计算资源和存储资源分布在因特网的各个角落，不强调任务所需的计算和存储资源同处一地。由于网络带宽的限制，网格计算中的数据传输时间占总运行时间的很大一部分。

1.4.2 云计算系统与传统超级计算机的区别

超级计算机拥有强大的处理能力，特别是计算能力。2008年11月17日，最新一期的Top500[6]榜单发布。冠军“RoadRunner”是IBM为美国 Los Alamos 国家实验室建造的计算机系统。它的运算速度达到了

1.026 Petaflop/s。RoadRuner超级计算机包含12960个IBM PowerXcell 8i处理器以及6948个分布于刀片服务器上的AMD Opteron芯片刀片服务器安装在288个IBM BladCener机架上。RoadRuner拥有80TB的内存，外存使用1.5PB容量的Panasas存储，外存通过10Gb/秒以太网进行连接。耗资超过1亿美元。

TOP500对超级计算机的排名方式可以看出，传统的超级计算机注重运算速度和任务的吞吐率。以运算速度为核心进行计算机的研究和开发。而云计算则以数据为中心，同时兼顾系统的运算速度。传统的超级计算机耗资巨大，远超云计算系统。例如，趋势科技花费1000多万美元租用34000多台服务器，构建自身的“安全云”系统。

1.5 云计算的关键技术

云计算是一种新型的超级计算方式，以数据为中心，是一种数据密集型的超级计算[10]。在数据存储、数据管理、编程模式等方面具有自身独特的技术。

1.5.1 数据存储技术

为保证高可用、高可靠和经济性，云计算采用分布式存储的方式来存储数据，采用冗余存储的方式来保证存储数据的可靠性，即为同一份数据存储多个副本。

另外，云计算系统需要同时满足大量用户的需求，并行地为大量用户提供服务。因此，云计算的数据存储技术必须具有高吞吐率和高传输率的特点。

云计算的数据存储技术主要有谷歌的非开源的GFS（Google File System）[11]和 Hadoop 开发团队开发的GFS的开源实现HDFS（Hadoop Distributed File System）[12][13]。大部分IT厂商，包括yahoo、Intel的“云”计划采用的都是HDFS的数据存储技术。

未来的发展将集中在超大规模的数据存储、数据加密和安全性保证、以及继续提高I/O速率等方面。

1.5.2 数据管理技术

云计算系统对大数据集进行处理、分析向用户提供高效的服务。因此，数据管理技术必须能够高效的管理大数据集。其次，如何在规模巨大的数据中找到特定的数据，也是云计算数据管理技术所必须解决的问题。

云计算的特点是对海量的数据存储、读取后进行大量的分析，数据的读操作频率远大于数据的更新频率，云中的数据管理是一种读优化的数据管理。因此，云系统的数据管理往往采用数据库领域中列存储的数据管理模式。将表按列划分后存储。

云计算的数据管理技术最著名的是谷歌的BigTable[14]数据管理技术，同时Hadoop开发团队正在开发类似BigTable的开源数据管理模块。

由于采用列存储的方式管理数据，如何提高数据的更新速率以及进一步提高随机读速率是未来的数据管理技术必须解决的问题。

1.5.3 编程模式

为了使用户能更轻松的享受云计算带来的服务，让用户能利用该编程模型编写简单的程序来实现特定的目的，云计算上的编程模型必须十分简单。必须保证后台复杂的并行执行和任务调度向用户和编程人员透明。

云计算采用类似MAP-Reduce[15]的编程模式。现在所有IT厂商提出的“云”计划中采用的编程模型，都是基于MAP-Reduce的思想开发的编程工具。

MAP-Reduce不仅仅是一种编程模型，同时也是一种高效的任务调度模型。Map-Reduce这种编程模型并不仅适用于云计算，在多核和多处理器、cell processor、以及异构机群上同样有良好的性能[16, 17, 18]。

该编程模式仅适用于编写任务内部松耦合、能够高度并行化的程序。如何改进该编程模式，使程序员得能够轻松的编写紧耦合的程序，运行时能高效的调度和执行任务，是Map-Reduce编程模型未来的发展方向。

2. 数据存储技术

为了满足云计算的分布式存储方式、同时保证数据可靠性和高吞吐率以及高传输率的需求。目前各IT厂商多采用GFS或HDFS的数据存储技术。

以GFS为例。GFS是一个管理大型分布式数据密集型计算的可扩展的分布式文件系统。它使用廉价的商用硬件搭建系统并向大量用户提供容错的高性能的服务。

GFS和普通的分布式文件系统有以下区别，如表2所示：

表2 GFS与传统分布式文件系统的区别

	GFS	传统分布式文件系统
组件失败管理	不作为Exception处理	作为Exception处理
文件大小	少量大文件	大量小文件
数据写方式	在文件末尾附加数据	修改现存数据
数据流和控制流	数据流和控制流分开	数据流和控制流结合

GFS系统由一个Master和大量块服务器构成。Master存放文件系统的所有的元数据，包括名字空

间、存取控制、文件分块信息、文件块的位置信息等。GFS中的文件切分为64MB的块进行存储。

在GFS文件系统中，采用冗余存储的方式来保证数据的可靠性。每份数据在系统中保存3个以上的备份。为了保证数据的一致性，对于数据的所有修改需要在所有的备份上进行，并用版本号的方式来确保所有备份处于一致的状态。

客户端不通过Master读取数据，避免了大量读操作使Master成为系统瓶颈。客户端从Master获取目标数据块的位置信息后，直接和块服务器交互进行读操作。

GFS的写操作将写操作控制信号和数据流分开，如图3[11]所示：

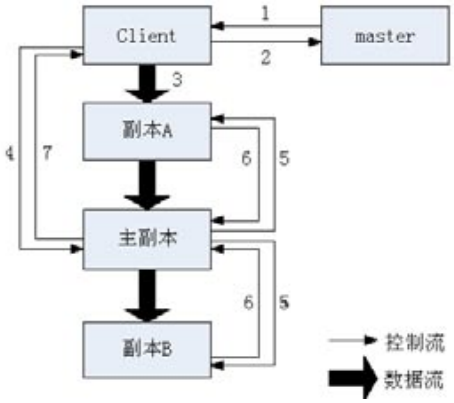

图3 写控制信号和写数据流

即，客户端在获取Master的写授权后，将数据传输给所有的数据副本，在所有的数据副本都收到修改的数据后，客户端才发出写请求控制信号。在所有的数据副本更新完数据后，由主副本向客户端发出写操作完成控制信号。具体请见[11]。

3. 数据管理技术

为了满足云计算的大规模数据集管理，高效的数据定位需求。谷歌采用BigTable的数据管理技术。在各大IT厂商的支持下，Hadoop开发团队正在开发其开源版本。

以BigTable为例。BigTable数据管理方式设计者——Google给出了如下定义：“ BigTable是一种为了管理结构化数据而设计的分布式存储系统，这些数据可以扩展到非常大的规模，例如在数千台商用服务器上的达到PB(Petabytes)规模的数据。 ”

BigTable对数据读操作进行优化，采用列存储的方式，提高数据读取效率。BigTable管理的数据的存储结构为：<row: string, column: string, time: int64> ->string。BigTable的基本元素是：行，列，记录板和时间戳。其中，记录板是一段行的集合体。

BigTable中的数据项按照行关键字的字典序排列，每行动态地划分到记录板中。每个节点管理大约100个记录板。时间戳是一个64位的整数，表示数据的不同版本。

BigTable在执行时需要三个主要的组件： 链接到每个客户端的库，一个主服务器，多个记录板服务器。主服务器用于分配记录板到记录板服务器以及负载平衡，垃圾回收等。记录板服务器用于直接管理一组记录板，处理读写请求等。

为保证数据结构的高可扩展性，BigTable采用三级的层次化的方式来存储位置信息，如图4[14]所示。

其中第一级的Chubby file中包含Root Tablet的位置，Root Tablet包含所有METADATA tablets的位置信息，每个METADATA tablets包含许多User Table的位置信息。具体见[14]。

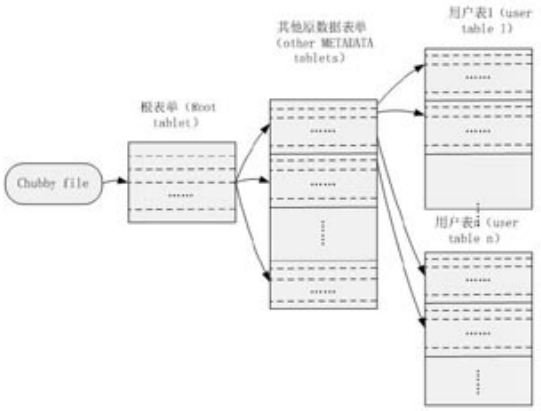

图4 BigTable中存储记录板位置信息的结构

4. 编程模型技术

当前各IT厂商提出的“ 云 ”计划的编程工具均基于Map-Reduce的编程模型。Map-Reduce是一种处理和产生大规模数据集的编程模型，程序员在Map函数中指定对各分块数据的处理过程，在Reduce函数中指定如何对分块数据处理的中间结果进行归约。用户只需要指定map和reduce函数来编写分布式的并行程序。当在机群上运行Map-Reduce程序时，程序员不需要关心如何将输入的数据分块、分配和调度，同时系统还将处理机群内节点失败以及节点见通信的管理等。图5给出了一个Map-Reduce程序的具体执行过程。

从图5可以看出，执行一个Map-Reduce程序需要五个步骤：输入文件、将文件分配给多个worker并行地执行、写中间文件（本地写）、多个Reduce workers同时运行、输出最终结果。本地写中间文件在减少了对网络带宽的压力同时减少了写中间文件的时间耗费。执行Reduce时，根据从Master获得的中间文件位置信息，将Reduce命令发送给中间文件所在节点执行，进一步减少了传送中间文件对带宽的需求。

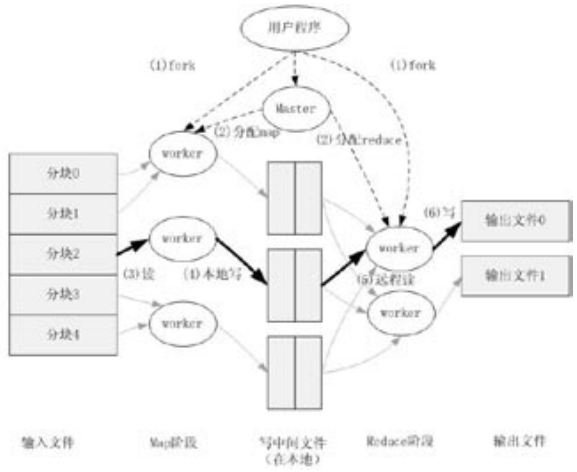

图5 Map-Reduce程序的具体执行过程

Map-Reduce模型具有很强的容错性，当worker节点出现错误时，只需要将该worker节点屏蔽在系统外等待修复，并将该worker上执行的程序迁移到其他worker上重新执行同时将该迁移信息通过Master发送给需要该节点处理结果的节点。Map-Reduce使用检查点的方式来处理Master出错失败的问题，当Master出现错误时，可以根据最近的一个检查点重新选择一个节点作为Master并由此检查点位置继续运行。

5. 结语

综上所述，云计算是一种新型的计算模式。它的最主要特征是系统拥有大规模数据集、基于该数据集，向用户提供服务。它使用大量的普通商用机来构建系统，通过冗余存储的方式确保整个系统的可靠性和可用性。与传统超级计算机在底层编程不同，数据密集计算的云系统上使用基于Map-Reduce的高级编程模式。这使得编程人员可以不用考虑底层的并行化方式，专心与程序的逻辑实现。普通用户经过简单的学习，可以编写出满足自身需要的简单程序。

越来越多的IT厂商提出了自己的“云”计划，并投入大量资金推动云计算的发展。这恰恰为云计算提供了良好的发展机遇。虽然现在的云计算并不能完美地解决所有的问题，但是在不久的将来，一定会有越来越多的云计算系统投入实用，云计算系统也会不断地被完善，并推动其他科学技术的发展。

参考文献：

[1] 维基百科http://en.wikipedia.org/wiki/Cloud_computing

[2] 中国云计算网。 http://www.cloudcomputing-china.cn/Article/ShowArticle.asp?ArticleID=1

[3] Jinesh Varia. Cloud architectures- Amazon web services [EB/OL]. ACM Monthly Tech Talk , http://acmbangalore.org/events/monthly-talk/may-2008--cloud-architectures---amazon-web-services.html, May, 2008

[4] IAN FOSTER; CARL KESSELMAN; STEVEN TUECKE. The anatomy of the grid enabling scalable virtual organizations. International Journal of High Performance Computing Applications. August 2001, 15(3): 200-222

[5] FRAN Berman, GEOFFREY Fox, TONY Hey. The grid: past, present, and future [A]. Grid Computing: Making the Global Infrastructure a Reality [C]. John Wiley & Sons, Ltd, 2003. 9-50.

[6] Top 500 supercomputing sites. http://www.top500.org/

[7] ALEXANDER S. Szalay, PETER Kunszt, ANI Thakar, JIM Gray, DON Slutz, ROBERT J. Brunner. Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey [A]. SIGMOD International Conference on Management of Data Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM, 2000. 29(2): 451-462

[8] Luiz Andr é Barroso, Jeffrey Dean, Urs H-lzle. Web search for a planet: The Google cluster architecture [J]. IEEE Micro, Mar/Apr, 2003, 23(2): 22 - 28.

[9] Google tops translation ranking[N]. News@Nature, http://www.nature.com/news/2006/061106/full/news061106-6.html, Nov. 6, 2006.

[10] RANDAL E.Bryant. Data-Intensive supercomputing: the case for DISC[R]. CMU Technical Report CMU-CS-07-128. May 10, 2007.

[11] SANJAY GHEMAWAT; HOWARD GOBIOFF; PSHUN-TAK LEUNG. The Google file system. Proceedings of the nineteenth ACM symposium on Operating systems principles. Oct. 2003

[12] Hadoop. http://hadoop.apache.org/

[13] Yahoo! Hadoop Tutorial. http://public.yahoo.com/gogate/hadoop-tutorial/start-tutorial.html

[14] Fay Chang, Jeffrey Dean, Sanjay Ghemawat et al. BigTable: a distributed storage system for structured data [A]. Operating Systems Design and Implementation, 2006.