

网格 Cache 若干问题分析

陈 梅 都志辉

(清华大学计算机科学与技术系 北京100084)

摘 要 网格中资源能力的不均衡和异构特性会给系统造成瓶颈,引起客户端延迟。网格 cache 系统能够有效地缓解这一问题。本文分析了网格 cache 的研究目的、网格 cache 和传统 cache 的关系,介绍了目前与网格 cache 相关的研究和网格 cache 系统设计的若干问题。

关键词 caching, 网格, cache 系统

An Analysis of Some Problems in Grid Cache

CHEN Mei DU Zhi-Hui

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract Disproportions between the capabilities of the resources in the grid and its property of heterogeneity cause bottlenecks in the system and high latency seen by the users. Grid cache system plays a crucial role in relaxing this problem. In this paper we analyze the purpose of the Grid cache system and the relations between grid cache and conventional cache. We also introduce the researches about the grid cache and some problems with need to be solve.

Keywords Cache, Grid, Cache system

随着网络基础设施、网络协议标准和技术的飞速发展,在互联网更进一步的资源共享的要求使网格技术产生和发展起来。从20世纪90年代产生网格计算的概念至今,全球范围的政府、工商业界和学术界独立或联合进行的网格研究越来越深入,网格应用的领域也日渐广泛,形成了世界性的网格研究组织^[1],正式的网格标准和协议也将出现。网格 cache 技术是网格系统的一项重要技术,在解决通信瓶颈、提高网格系统性能方面起着重要的作用。本文分析了网格 cache 的研究目的、网格 cache 和传统 cache 的关系,介绍了目前与网格 cache 相关的研究和网格 cache 系统设计的若干问题。

1 网格 cache 的研究目的

网格应用之所以有强大的吸引力,在于它具有的几个独特优势^[2]:1. 网格能够提供超级计算存储能力,它可以协同非集中控制的分布异构资源完成现有系统难以完成的复杂任务;2. 网格能够用较低的成本完成系统目标,因为它可以充分发挥加入网格的各种资源的能力;3. 网格能够提供简单统一的操作平台,它有一套标准、开放、通用的协议和接口,以使大量分布设备和数据资源可以无缝远程访问。

网格具有下面的一些特点:1. 网格资源是动态分布的,网格中的资源集合可以看作是一个虚拟组织(Virtual Organization, VO)^[3],地理上分布的成员(服务提供者和使用)可以自由加入和退出;2. 资源是异构的,除了计算机系统之外还可以是其它设备、它们的规模不同、体系结构和管理策略不同,而且资源间的能力不均衡;3. 网格成员是自治的。

网格管理虚拟组织成员的加入和退出、为异构资源提供统一的接口、保证资源间通信的顺畅和安全,要在拓扑结构变化、不安全可靠的环境中给用户架设一个安全可靠的服务平台,需要有可靠的网格 I/O 系统的支撑。此外,由于网格起源于完成超级复杂的计算和存储任务,通常需要处理的任务量非常大,例如欧洲在研的 DataGrid 项目^[4]目标就是处理以 P 为单位的数据,对 I/O 系统的性能也提出了很高的要求。网格系统 I/O 性能的提高一个很重要的方面是运用软件技术

解决系统通信瓶颈问题。通信瓶颈的产生通常归因于系统部件的能力差异,在网格中主要体现在以下方面:带宽需求和实际可用带宽间、不同网络的传输能力间、不同存储系统的 I/O 速度之间以及各种资源协同工作时处理能力的差异。一部分使这种差异恶化的原因在于异构系统间传输的额外开销。

采用合适的 cache 技术可以有效地缓解上面提到的问题,屏蔽客户端延迟。这里 cache 是指存储能力强于本地磁盘弱于主存结点的暂时性的存储^[5]。用 cache 存储处理需要的远程数据、保存计算的中间结果,可以减少多个任务使用相同文件时的重复传输,一方面减少了带宽占用、缓解中央服务器的负载,一方面也减少了异构系统间传输造成的处理开销。cache 还可以在不同 I/O 速度存储系统间起到缓冲的作用。

2 网格 cache 与传统分布式 cache

2.1 分布式 cache 系统

基于网络的分布式 cache 系统大致包括 web cache、p2p cache、cache 集群以及各种分布式文件系统(AFS、NFS、DFS、DPFS)cache 等,尽管各种 cache 系统有各自的特点,要解决的问题不尽相同,然而基本目的都是想找到一个最佳的方案缓存用户请求的数据备份,以便在最短的时间以最小的代价满足客户的数据请求。

2.1.1 分布式 cache 系统结构 cache 系统结构决定了 cache 的组织方式、cache 间的相互关系和互操作的方式。常用的有层次结构(hierarchy)、分布式结构(distributed)和混合结构(hybrid)^[6~10]。采用什么样的结构对算法的选择具有很大的影响。下面分析前两种结构,混合结构是一种集成的结构,主要是利用了前两种结构的优点。

层次结构是广泛采用的一种结构,它的分层缩短了客户请求的距离,中央结点的负载被分散出去,每层内部结点有一定的相关性,有利于提高 cache 命中率,缩短查找时间。为了减少管理开销,一般分层的 cache 结构最多为三至四层,最下层是客户端的 cache,最上层则直接连接中央服务器,为一个树型结构。在层次结构中,下层 cache 中的数据也存于其上一

层,每层 cache 都存有数据备份,因而占用空间较大,并且 cache 缺失时的逐层查找会造成额外层间延迟。层次结构基本上是依网络拓扑分布,高层 cache 一般存在于网络交换中心,容易形成瓶颈,一旦出现问题恢复困难。

分布式结构减少了对中间层的空间要求,只在底层缓存数据,对 cache 条目分层或分区管理。分布式结构中结点 cache 间的网状关系与层次结构的父子结点的单一关系不同,有较好的容错性,也不存在因为层间分布不平衡导致的负载不均问题。不过在大规模的分布式 cache 系统中会占用过多的带宽,连接时间过长,系统的管理也是个问题。

2.1.2 分布式 cache 算法和目录 根据问题的性质可以把 cache 算法分成基本算法和高级算法。基本算法是 cache 系统里必不可少的维持系统正常工作的一类算法,包括以下内容:复制,解决是否需要复制和复制到哪的问题^[11];查找,发现请求对象;决策,解决从哪一资源上获取的问题;替换^[12];统一界面,解决异构系统互操作问题。在基本算法的基础上可以配置更高级的算法以改进 cache 系统性能,高级算法包括:一致性,维护 cache 内副本的有效性^[13];预取,以提高访问命中率^[14,15];压缩,减小存储空间和带宽需求^[16]等。

衡量一个 cache 算法的性能要考虑两个方面,一是算法本身的效率,二是算法需要的成本。算法的效率由延迟、命中率、空间利用率等反映。查找速度和调页速度是影响延迟的关键因素,目录基础设施和系统结构的设计直接影响到查找和调页的速度。命中率除了跟算法有关,在一定程度上还跟 cache 空间大小和缓存粒度有关系。要提高系统性能,选择与系统环境相匹配的算法很重要,算法在不同的环境中会有不同的表现。考虑到系统扩展的问题,好的算法应该是轻量级的。成本则主要反映在算法占用的带宽和 CPU 时间上。

不论是何种 cache 系统,在每个 cache 中都需要维护一个或若干个目录^[17],主要用于请求资源的查找,还可在替换和预取等算法中提供参考。其中 cache 数据目录是必备的,它记录了 cache 内容在本地和远程文件之间的映射情况包括文件名、用户列表,文件的更新、修改情况,一致性维护及其他算法需要的一些记录比如锁记录、时间戳等。根据需要还可能配置摘要目录、相邻 cache 结点目录等。

2.2 网络 cache 与传统分布式 cache 系统的不同

网络 cache 与传统 cache 尽管有着密切的联系,但是仍然存在一些差别:

首先,在 OGSA 架构中,网络中的一切都是服务^[3,18],cache 功能可以看作是能够在网络中得到的一种服务,它既可以透明地作为其他服务的支撑,也可以作为单独的服务提供,通过提供标准的服务接口,实现异构用户间的资源共享。可以在底层的 cache 服务和其他网络服务(比如安全服务)的基础上构造更高级的 cache 服务,网络 cache 系统能够根据用户的不同需要提供不同 QoS 的服务,而传统 cache 一般是中央控制、执行统一策略的系统。

其次,网络 cache 存在的环境比传统结构复杂,具体表现在:1. 组成 cache 系统的成员类型复杂,现存的所有类型的 cache 系统甚至网络 cache 都有可能加入到网络中,而且成员是异构的、能力有高低,这是由网络成员属于分散控制域的特性决定的;2. 网络需要缓存的数据类型比较复杂,除了常规服务数据,还有注册和控制信息、服务信息、元数据以及网络拓扑,cache 系统的设计要考虑到各种类型的缓存要求;3. 传统 cache 系统的成员和数据来源都是静态分布的,而网络系统中数据的使用者、数据的提供者,甚至可能 cache 系统本身都具有动态性质;4. 网络 cache 系统服务对象的复杂关系对

cache 决策很有影响,cache 的使用者既可以是网络服务的提供者也可以是使用者,或者在提供某一服务的同时使用另一个服务,网络中的一个服务请求可能会解析成不同地点的多个服务,这种复杂性必然导致 cache 管理和算法设计的复杂化。

最后,网络 cache 是对现有 cache 资源的整合和优化配置,目前分立的各网络 cache 比如 AFS、DPSS、NFS 等可以整合成一个大的 cache 池,客户通过 cache 管理服务和监控服务使用这个 cache 池^[19]。

3 网络 cache 及与之相关的研究

3.1 OGSA 中的 cache management API 与复制

网络里的一切都被抽象成服务,开放网络服务体系结构 OGSA(Open Grid Services Architecture)解决了两个重要的问题,即标准服务接口的定义和协议的识别,Globus Toolkit^[20]用于构建 OGSA 中的服务(持久的或瞬时的)。在 Globus Toolkit 中定义了 cache management API,这是一个可选的底层 API,用于在特定的数据传输和访问策略中实现基本的 cache 功能。它定义在 Global Access to Secondary Storage(GASS)协议^[21]中,GASS 定义了一个全局命名空间,为远程文件访问提供标准 I/O 接口,它利用了 Globus Toolkit 的安全(GSI)和通信(gridFTP)等服务。cache management API 的功能包括操作本地 cache,由用户直接控制插入、锁定、删除和参考计数,控制什么时候缓存文件以及缓存到什么地方,协调多个任务对 cache 的同步操作。cache management API 为 GASS 中的 file access API 提供服务。GASS 采取的四种 I/O 模式在一致性问题上都比较简单,不用考虑全局一致性问题。

在传统 cache 中,复制是 cache 技术的一部分,缓存数据本身就是复制的数据备份,cache 要决定是否复制一个文件,复制到哪,如何维护复制数据和主存数据的一致性,如何替换等问题。由于一般网络分布式系统规模较大、布局分散,一些系统倾向于采取 cache 和复制分工的策略,如文[22,23],以利用复制增强数据的可达性。在 Globus Toolkit 中也提供了专门的复制工具:replica catalog 和 replica management services^[5]。cache 和复制都属于 Globus Toolkit 部件中的资源管理部分,其中 cache 定义为客户端现象,为 I/O 服务,复制是服务器端的现象,由服务器决定何时何地建立文件备份^[24]。复制的对象是相对稳定、修改不多的数据,粒度通常比 cache 数据大。在实现上,复制需要利用 cache API 提供的缓存服务,复制对 cache 来说则是透明的,cache 不需要了解数据的来源是原始提供者还是复制点。

3.2 IBM 的 caching grid

针对网络的商业应用,IBM 把 Web caching 和网络结合起来研究了 caching grid^[25],它的功能是在客户请求率升高或请求延迟的 QoS 降低时提供 cache 服务,并提供在线负载测量、实时预测和事件产生机制来提醒用户的增加。caching grid 中的 cache 来源于集群中的一部分结点(另一部分做服务器)和在集群管理服务注册成 cache 代理的客户机,当 caching grid 预测到将有超载时,就分配给服务器一些代理以平衡负载,如果服务器代理不够,就联结注册过的客户机代理。可以看出 caching grid 的核心功能是在线建模和预测,它由一个 eModel 工具来完成。用户需要建立一个 SLA(service level agreements)定义需要的服务,例如在一定时间周期内以特定的平均反映时间满足一定吞吐量,服务可达等,自治系统元件遵照 SLA 执行监控和自我配置。如果 SLA 不能根据原

始数据得到或者需求变化需要更新,eModel 可用于新 SLA 的建立;服务提供者在配置及承诺客户 SLA 时,需要考虑它的可用容量,eModel 可用于分析接受这个 SLA 的风险;eModel 在系统运行中还可以测量和监控性能,预测完成服务目标的潜在危险。网格中服务请求到达模式和服务机制比较复杂,需要应用复杂的统计和建模技术来分析系统,预测请求到达怎样随时间改变,在不同平台上建立系统模型预测请求服务过程。

3.3 cache 集群

cache 集群是一种紧密耦合的 web caching 结构,网格 cache 可以借鉴其调度和负载均衡的方法。基于 LVS 的 cache 集群的体系结构^[26]:第一层是负载调度器,一般采用 IP 负载均衡技术,可以使得整个系统有较高的吞吐率;第二层是 cache 服务器池,一般 cache 服务器放置在接近主干 Internet 连接处,它们可以分布在不同的网络中。调度器可以有多个,放在离客户接近的地方。负载调度器采用了一种 IP 隧道方法(即 VS/TUN 方法)或者 VS/DR 方法,调度器只调度 Web cache 请求,而 cache 服务器将响应数据直接返回给客户。cache 服务器采用本地硬盘来存储可缓存的对象。cache 服务器间有专用的多播通道(Multicast Channel),通过 ICP 协议(Internet Cache Protocol)来交互信息。

4 网格 cache 系统设计

4.1 网格 cache 系统的主要管理目标

一、实现良好的可扩展性和适应性。网格 cache API 的设计满足轻量级的要求,有利于实现这一目标。适应系统扩展造成的服务对象数量和类型的增加和对 I/O 性能的更高要求;协调配合网格内的其他服务,减少系统内的冲突和瓶颈;满足不同组织的访问权限和安全要求。

二、有效。cache 系统本身给网格造成的负担要小,要实现 cache 系统资源效益的最大化。

三、提供可靠服务。cache 系统应该具有健壮性,它是正常访问的基础,要避免由于 cache 本身的低效和错误引起性能下降。在 cache 故障时需要有恢复策略。cache 系统的负载平衡也是对可靠性的良好支持。

四、简单、易于使用。cache 是一种底层的技术,而且作为软件实现的底层,需要尽量减少复杂性以提高效率。易于使用不仅指用户接口,而且要易于系统开发和部署。

4.2 网格 cache 系统需研究的问题

目前网格 cache 系统有以下问题需要解决:

异构系统间的接口问题,已经开发出的 Storage Resource Broker(SRB)^[27]界面可以处理异构存储系统,Globus 项目正在开发 GASS 和 SRB 的接口。

什么时候对用户启用远地 cache,如何选择使用合适的 cache 结点,采用何种备份方式;网格可以灵活地把程序传送到数据端,如何决定是传送数据到处理端还是传送程序到数据端。

向用户提供不同 QoS 服务的问题,例如预约 cache 空间,并监控 cache 服务情况。

如何减少 cache 开销;如何提供全球统一命名空间;尽管网格中多数数据是只读的,一致性维护的问题也还存在。

总结 网格系统 I/O 性能的改善离不开网格 cache 系统的支持,本文尝试分析了导致 cache 需求的网格问题,网格 cache 与传统 cache 有着密切的联系,因此又综合介绍了传统 cache 系统的主要技术,以及在应用传统 cache 技术时要注意的网格 cache 与传统 cache 系统的一些差别,并在对网格

cache 研究现状了解的基础上分析了网格 cache 系统需要解决的几个问题。

参考文献

- GGF homepage. <http://www.gridforum.org>
- Foster I. What is the grid? A three point checklist. <http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf>
- Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, 2001, 15(3)
- Data Grid homepage. <http://eu-datagrid.web.cern.ch/eu-datagrid/>
- Lamehamed H, Szymanski B, Shentu Z, Deelman E. Data replication strategies in grid environments. In: *Proc. of Fifth Intl. Conf. on Algorithms and Architectures for Parallel Processing*, 2002. 378~383
- Wang J. A survey of web caching schemes for the internet. *ACM Computer Communication Review*, 1999, 29(5): 36~46
- Rodriguez P, Spanner C, Biersack E W. Analysis of Web caching architectures: hierarchical and distributed caching. *IEEE/ACM Transactions on Networking*, 2001, 9(4): 404~418
- Wei S, Schenfeld E. Hierarchical interconnection cache networks. In: *Proc. of Seventh Intl. Parallel Processing Symposium*, 1993. 135~141
- Mahanti A, Williamson C, Eager D. Traffic analysis of a Web proxy caching hierarchy. *IEEE Network*, 2000, 14(3): 16~23
- Dykes S, Jeffery C, Das S. Taxonomy and design analysis for distributed Web caching. In: *Proc. of the 32nd Annual Hawaii Intl. Conf. on System Sciences*, 1999. HICSS-32, Volume: Track8, 1999. 10
- Kangasharju J, Ross K W. A replicated architecture for the Domain Name System. In: *Proc. of IEEE INFOCOM 2000. Nineteenth Annual Joint Conf. of the IEEE Computer and Communications Societies. Volume: 2*, 2000. 660~669
- Willick D L, Eager D L, Bunt R B. Disk cache replacement policies for network filesystems. In: *Proc. the 13th Intl. Conf. on Distributed Computing Systems*. 1993. 2~11
- Liu Chenjie, Cao Pei. Maintaining strong cache consistency in the World-Wide Web. In: *Proc. of the 17th Intl. Conf. on Distributed Computing Systems*. 1997. 12~21
- Shih F W, Lee T-C, Ong S. A file-based adaptive prefetch caching design. In: *Proc. of IEEE Intl. Conf. on Computer Design: VLSI in Computers and Processors*, 1990. 463~466
- Foygel D, Strelow D. Reducing Web latency with hierarchical cache-based prefetching. In: *Proc. of 2000 Intl. Workshops on Parallel Processing*, 2000. 103~108
- Cecchet E. Whoops!: a clustered Web cache for DSM systems using memory mapped networks. In: *Distributed Computing Systems Workshops. Proc. 22nd Intl. Conf. on*, 2002. 806~811
- Wu Yuguang, Muntz R R. A new cache directory scheme. In: *Proc. of Second Intl. Symposium on Parallel Architectures, Algorithms, and Networks*. 1996. 466~472
- 都志辉, 陈渝, 刘鹏, 王小鹤, 杜江, 周念生. 以服务为中心的网格体系结构 OGSA. <http://hpclab.cs.tsinghua.edu.cn/~duzh/publications.htm>. 计算机科学, 2003, 30(7): 26~29
- Tierney B L. Data grids and data grid performance issues outline. <http://www.didc.lbl.gov/>
- Globus Toolkit 1.1.3 System Administration Guide. <http://www.globus.org/>
- Bester J, Foster I, Kesselman C, Tedesco J, Tuecke S. GASS: A Data Movement and Access Service for Wide Area Computing Systems. In: *Sixth Workshop on I/O in Parallel and Distributed Systems*, May 1999
- Han Hua, Guo Chaoyang, Dai Yafei, Li Xiaoming. Global file store: a massive storage system on the internet concept and design. In: *Proc. of 2001 Intl. Conf. on Computer Networks and Mobile Computing*, 2001. 165~175
- James B L, Hurson A R. Transaction processing in mobile, heterogeneous database systems. *IEEE transactions on knowledge and data engineering*, 2002, 14(6)
- Ranganathan K, Foster I T. Identifying Dynamic Replication Strategies for a High-Performance Data Grid. In: *Proc. of the Second Intl. Workshop on Grid Computing*. Nov. 2001
- Crawford C H, et al. Commercial Applications of Grid Computing. IBM Research Division Thomas J. Watson Research Center. 2003. <http://www.ibm.com>
- Linux 服务器集群系统(二)LVS 集群的体系结构. <http://www-900.cn.ibm.com/developerWorks/cn/linux/cluster/lvs/part2/index.shtml#5>
- Wan M, Rajasekar A, Moore R, Andrews P. A simple mass storage system for the SRB data grid. In: *Proc. of 20th IEEE / 11th NASA Goddard Conf. on Mass Storage Systems and Technologies*, 2003. 20~25