# 用 P2P 技术改进网格信息服务*)

黄竞伟　范清风　吴琼莉　何炎祥

(武汉大学计算机学院电信学院　软件工程国家重点实验室　香港城市大学深圳研究院　武汉 430072)

**摘　要**　本文根据网格资源信息的特点阐述了由高度分布式的信息提供者和集合目录组成的网格信息服务基本框架，并分析了它的基础 LDAP (Lightweight Directory Access Protocol)协议，指出 LDAP 目录本质是一种分布式的数据库。由于网格信息系统中 LDAP 目录信息树的动态刷新与复制的频繁发生，我们已提出用环形扩展和线形扩展策略来大幅度提高系统效率；在此基础上，本文进一步提出了文件分块复制法的思想把 LDAP 数据库文件分成若干块，在多个 LDAP 服务器端点间并行复制，最后实践证明，它大幅度提高了以 LDAP 目录分布式数据库为基础与核心的网格信息服务系统的并行效率。

**关键词**　网格信息服务，资源共享，动态复制，分块复制，p2p

## Improved Grid Information Service Using the Technology of P2P

HUANG Jing-Wei　FAN Qing-Feng　WU Qiong-Li　HE Yan-Xiang

(School of Computer, State Key Lab of Software Engineering Wuhan University, Wuhan 430072)

**Abstract**　The infrastructure of grid information service constituted with highly distributed information providers and aggregate directory is brought forward on the basis of the characteristic of grid information resources in the paper. The Grid Information Protocol (GRIP) makes both discovery and enquiry for entity information repeatedly by sorts of flexible measures and the Grid Registration Protocol (GRRP) keeps the resources available by the continuous and periodic updating from the information providers in the infrastructure. The Lightweight Directory Access Protocol (LDAP) that is one of the base protocols is also analyzed. It is put forward that LDAP is a distributed database that has the characteristic of spanning flat and accessing according to the register. The server-to-server communication mode of LDAP defines how to share the LDAP directory and how to update and replicate the information between servers. The dynamic updating and replication of LDAP directory tree would happen frequently for grid information system is distributed widely, highly fault-tolerant, dynamic and diversiform. It has been put forward that the strategy of annular spread and line-form spread can boost the efficiency of grid information service system that regards the distributed database of LDAP as the foundation and core. What's more, we use the viewpoint of file parted replication to divide the LDAP database file into several blocks that are replicated parallel between LDAP sever points then. In such a way, the system efficiency of parallel processing can be boosted by margin. And based on the idea forenamed, we put forward the technique infrastructure and block arithmetic , both of which are proved to be available in improving the system efficiency.

**Keywords**　Grid information services, Enquiry protocol, Registration protocol, Resource sharing, Dynamic replication, P2P

## 1　引言

网格计算是当今计算机界的热门研究方向，而信息服务是网格计算中的重点和难点问题[1]。面对当前的网格环境，Foster[2]提出了由高度分布式的信息提供者和集合目录组成的网格信息服务基本框架，并且分析了它的基础 LDAP 协议，指出 LDAP 目录本质是一种分布式数据库，它是跨平台的，分布式的，按记录存取的。LDAP 协议中的服务器——服务器通讯模式定义了多个服务器之间是如何共享一个 LDAP 目录信息，以及如何更新和复制服务器之间的信息。

由于网格信息系统分布广，容错性强，动态多样性，LDAP 目录信息树的动态刷新与复制将频繁发生，本文提出用环形扩展和线形扩展策略能大幅度提高以 LDAP 目录分布式数据库为基础与核心的网格信息服务系统的效率。不仅如此，在研究了如何在节点级提高复制与刷新效率的策略之后，本文还进一步研究了节点级以下如何提高复制与刷新效率的策略。本文用文件分块复制法把 LDAP 数据库文件分成若干块在多个端点间并行复制，这样进一步提高系统并行效率，提出了技术框架和阻塞算法。最后实践证明它大幅度提高系统效率。

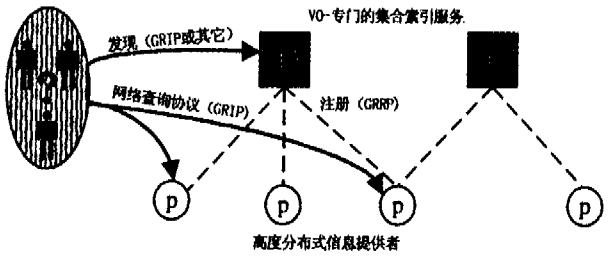
## 2　系统结构概观

网格信息系统中，信息服务组件越来越分布式，而且个体来源常常失效，信息提供者的总数要求很大，而且种类越来越多样性，这是网格信息服务的基本要求。

我们的网格信息服务系统结构(图 1)包括两个基本的实

*)"网上信息收集和分析的基础问题和模型研究"，国家自然科学基金重大研究计划，2002，1-2005，12，项目编号：90104005。黄竞伟　教授，博导，研究方向：分布并行处理，演化计算，网格计算等。范清风　博士生，研究方向：分布并行处理，网格计算。吴琼莉　博士生，研究方向：网络通信。何炎祥　教授，博导，研究方向：分布并行处理，移动计算，网格计算等。

体:高度分布式信息提供者和专门的集合目录服务。

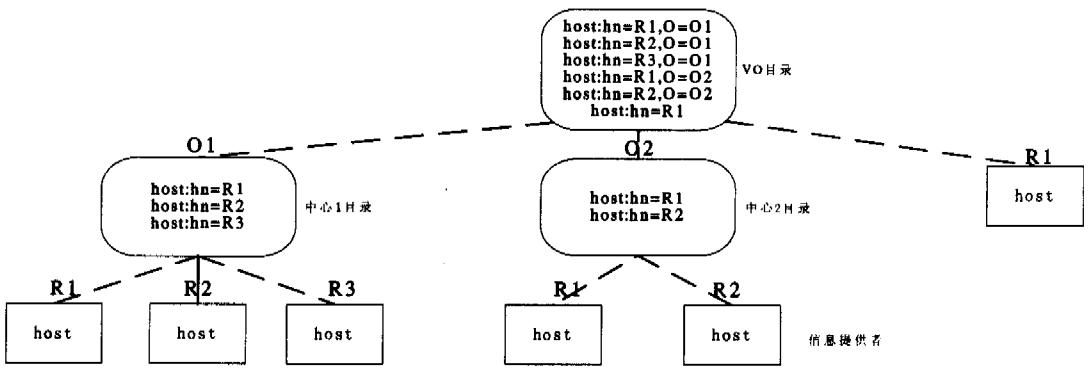



用网格信息协议 GRIP 通过各种灵活的方法反复地发现和查询实体信息,用网格注册协议 GRRP,通过信息提供者定期地连续地提供刷新,来保持资源可用性。

图 1　系统结构概观

信息提供者可定义为是一种表达两项基本协议的服务。网格信息协议 GRIP(Grid Information Protocol)通过各种灵活的方法反复地发现和查询实体信息。而网格注册协议 GRRP(Grid Registration Protocol)通过信息提供者定期的连续的提供刷新,来保持资源可用性。

集合目录是一项服务,它用 GRRP 和 GRIP 从一组信息提供者处获得关于一组实体的信息,然后对有关这些实体的查询做出答复。另一方面集合目录本身就可以作为信息提供者,如图 2 所示。每个目录使用 GRIP 数据模式,查询语言和协议,收集其他实体信息;并且每个目录自己充当信息提供者,目录使用 GRRP 来向更高级目录注册,以建构这个等级。




两个资源中心和一个体正给一虚拟组织提供资源(底部,方形盒)。形成有关等级发现服务的这三个集合目录(上部圆形盒)以匹配此逻辑结构的方式组织起来。请注意资源名是如何用于对特殊组织的全范围搜索,如果要求是如此;换句话说,搜索可以面向根目录而不用涉及到范围。

图 2　等级发现

由此可见,这个系统结构体现了许多与环球网(World-Wide Web)相同的结构原则。GRIP 与 HTTP(服务程序所用的协议)相一致,集合目录与搜索引擎相一致。

## 3　网格信息服务的 LDAP 信息模型

网格信息服务结构的基础是遵从 LDAP 模型的,主要由目录信息树 DIT(Directory Information Tree)层次和对象组成。我们采用标准的轻量目录存储协议(LDAP),作为 GRIP 和 GRRP 的基础协议,LDAP 明确了:数据模式、查询语言和线路协议[3]。

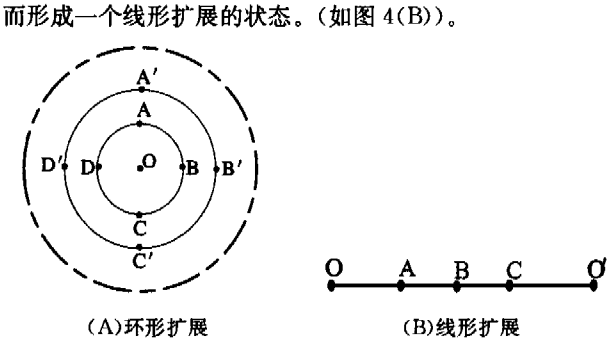
LDAP 目录也是一种类型的分布式数据库,但不是关系型数据库。LDAP 是跨平台的协议,它适合用于那种需要频繁读取场合,LDAP 服务器可以是分布的,它以一条条表项(entry)存储的,各表项属性可变。在 LDAP 协议中存在两种通信模式:客户-服务器通信和服务器-服务器通信。基本的客户-服务器通信允许用户程序连接 LDAP 服务器进行创建、检索、修改、删除数据等操作。服务器-服务器通信定义了多个服务器如何共享一个 LDAP 目录信息,以及如何更新和复制服务器之间的信息。

## 4　用动态复制策略提高效率

由于网格信息服务系统分布广,容错性强,动态多样,LDAP 目录分布式数据库的服务器与服务器间的更新和复制将相当频繁。为了提高系统效率,我们根据网格不同的组成基础,在比较多种复制策略之后,最终发现环形扩展和线形扩展的策略能大幅度提高以 LDAP 目录分布式数据库为基础与核心的网格信息服务系统的效率[4]。

**策略 1**　环形扩展。可以把这个策略比作一个三环喷泉,水源于顶部;当它从顶部的边缘落下来,将进入下一层;当这一层的水又要溢出来时,将落到再下一层。数据流量是相似的道理:一旦文件的阈值超过了根节点,将在下一层产生复制,最佳位置是需求量最高的客户的根节点;进而文件的需求数超过第二层,它将在下一层中复制;一个经常使用的文件可能最终复制到客户机本身。需求文件的客户机存贮一个逻辑拷贝,当文件很大(2G)时,客户机只能在同一时刻存贮一个拷贝,该文件将很快被替换。服务器周期性地标记流行文件而且把它们繁殖到层次的下层中去,从而形成一个环形扩展的状态。(如图 4(A))

**策略 2**　线形扩展。在该方式中,复制文件存贮在从服务器到客户机路径上的每一个节点中。就是说,当一个客户需要某个文件时,该文件的拷贝将存贮在路径的每一层上,从而形成一个线形扩展的状态。(如图 4(B))。




(A)环形扩展　　(B)线形扩展

图 4

通过些实验比较，依据低响应时间和低带宽占用的原则，在环形扩展与线形扩展中的交易容易实现。如果主要目的是加快系统的响应则环形扩展策略将工作的较好。如果保存带宽是首要目的，线形扩展将是一个较好的复制策略。

## 5 文件分块复制法

在研究了如何在节点级提高复制与刷新效率的策略之后，本文还进一步研究了节点级以下如何提高复制与刷新效率的策略。文件分块复制法 FPR(Files Parted Replicating)[4] 用 P2P 作为一种寻求并行效率的方法，它能够进一步大幅度提高系统效率。

### 5.1 基本原理

当 LDAP 目录数据库文件被平凡复制与刷新时，所有的复制费用是由主 LDAP 服务器负担的。而使用文件分块复制法，把文件分成若干块，在多个客户机同时复制相同文件的情况下，不同的客户机首先复制不同的文件块，然后再相互复制它们没有的文件块。这种上传和下载间的重新分配，极大地扩展了网格信息服务系统的复制与刷新能力[5]。如图 5 所示。

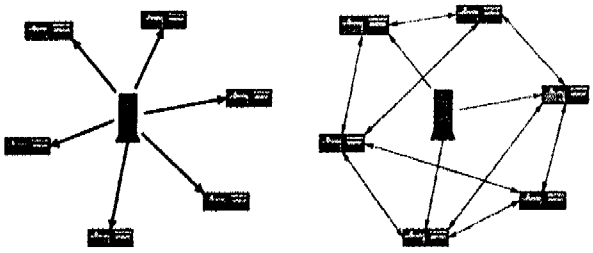


图 5 文件分块复制法原理图

对文件分块复制法的部署是由 LDAP 主服务器决定的。需复制的下层服务器根据需求使用文件分块复制法来尽可能高效地得到所需的目录文件。

### 5.2 技术框架

5.2.1 *发布内容* LDAP 分布式数据库中设一张轨迹表(种子)记录了数据库文件的长度、名字，碎片信息和轨迹的 url。这些轨迹将帮助 LDAP 服务器彼此发现。LDAP 服务器通过这些协议传输复制的文件名，端口和同一文件的不同端点，并把相关信息组成一个列表，放于轨迹之中，LDAP 服务器用这些信息去彼此联系。

5.2.2 *端点分布* 所有复制的逻辑问题是如何处理 LDAP 服务器之间的交互。关于 LDAP 服务器间彼此复制的信息是发送给轨迹的，轨迹的责任是严格地限制 LDAP 服务器端点间的彼此发现。轨迹是 LDAP 服务器端点间彼此发现的唯一方式，标准的轨迹算法是返回一个服务器端点的高强壮性的随机图表，很多端点选择算法产生强大的规则图表，它能够在少量的搅动后重新获得碎片。

为了保持轨迹上 LDAP 服务器端点间的有效连接，文件分块复制法把 LDAP 文件分成固定大小的块，典型的为四分之一兆。每个复制服务器报告它的所有端点的所有块，为了保证数据的完整性，碎块的所有信息都包括在 .frr 文件中。每个 LDAP 服务器端点不断地从其所有能到达的端点下载，当然不能从那些它们连接不上的，没有需求块的，或当前不允许下载得端点下载。

5.2.3 *流水线作业* 文件分块复制法在传输数据时，为了避免块传送间的超时，让几个请求马上挂起是很有必要的。它的措施是把块分成更小的子块——16k 大小，总是保持 5 个，要求及时的流水线作业。子块一旦到达，新的要求就要发出。在流水线上的数据量是连接饱和度的重要参数。

### 5.3 下载块选择法

在块下载时选择一个好的顺序对提高性能是非常必要的。一个环的块选择，将导致所有的块同时提供，或同时等待，而不能把任何块下载到所需要的任何端点上。

1)严格的优先级：文件分块复制法选择块的首要策略是：一旦某一子块被要求，则这些相关的子块将先于其它未要求的子块。这种策略对尽快地收集完全块非常有利。

2)最少块优先：当选择那一块复制时，端点服务器总是复制那些其它端点服务器拥有最少的块，该技术我们称之为最少块优先。该技术可以确保端点可以得到所有其它端点缺少的块，所以当再次需要该块时，可以反向复制。这同时也保证了那些更普通的块留在后面。这样也有利于保证反向复制的端点的持久兴趣。

3)随机的第一块：最少者优先的一个例外是在复制开始时。这时，端点服务器没有什么去上传，所以尽可能快地得到一个完全的块是非常重要的。比起那些在多个端点上，可以从不同的端点上下载的子块，稀少块总是仅存在于一个端点上，因而它往往下载得比较慢。因此，只到第一块完全下载之前，下载块是随机选择的，之后才是最少者优先。

4)结束模式：有时一块将以非常慢的速度传送出来，在复制中这不是问题，但它可能延误复制的完成。为了防止该现象的发生，一旦某个端点服务器没有的子块被请求，则它向所有端点发送这些子块的请求。删除那些浪费在冗余数据传输上而占有大量带宽的子块。在实践中，并没有很多带宽是以这种方式浪费的。因为结束期很短，而且文件的结尾部分总是复制的很快。

### 5.4 上传控制算法

文件分块复制法没有中心的资源分配，每个端点服务器自我管理，寻求最高复制率。服务器端点为了完成这一使命，尽可能地从任何它所能达到的端点复制，而且经由一个 P2P 变量来决定向那些服务器端点上传。为了合作，端点上传；为了不合作，端点阻塞。阻塞是对上传的一个暂时拒绝，它停止上传，但复制仍能够发生，而且当重新上传时，不必重新协商连接。

上传控制算法对保持良好特性是很有必要的。一个好的上传控制算法可以充分利用所有可利用的资源，为每个端点服务器提供合理的下载组成率，而且防止那些只下载不上传的端点服务器。

1)并行效率：上传控制算法把端点文件复制给向它们上传的端点，从而在任何时候都有几个端点服务器双向传送。这样系统也会更充分地上传，从而获得更好的传输率。

2)上传控制：算法规定，每个服务器端点总是保证上传固定数量的其它端点服务器(默认为 4)。对那些端点上传的控制是严格基于当前复制率的：目前主要是用一个滚动的 20 秒平均法。为避免由于端点迅速的上传和阻塞而引发的资源浪费状态，服务器端点每 10 秒钟重新计算谁将上传谁将阻塞，然后保留该状态，只到下一个 10 秒期到来。

3)优化上传：那些提供最佳复制率的端点服务器，会因为没有方法发现而受到损失。为了解决这个问题，任何时候，每

(下转第 96 页)

2004

5 Ferraiolo D F, Sandhu R, Gavrila S, et al. Proposed NIST standard for role-based access Control. ACM Transactions on Information and System Security (TISSEC), 2001, 4(3)

6 Bhatti R, Joshi J B D, Bertino E. Access Control in Dynamic XML-based Web-Services with X-RBAC. In: Proceedings of the First International Conference on Web Services, Las Vegas, USA, 2003

7 Wonohoesodo R, Tari Z. Role Based Access Control System for Web Services. In: Proceedings of the 2004 IEEE International Conference on Services Computing (SCC'04), Shanghai, China, 2004. 49～56

8 Bhatti R, Bertino E, Ghafoor A. A Trust-based Context-Aware Access Control Model for Web Services. In: Proceedings of the IEEE International Conference on Web Services (ICWS'04), San Diego, California, USA, 2004

9 OASIS Standard. Security Assertion Markup Language (SAML) V1. 1, October, 2003. http://www. oasis-open. org/committees/security/docs/cs-sstc-core-01. pdf

10 Simple Object Access Protocol (SOAP) V1. 1. May, 2000. http://www. w3. org/TR/2000/NOTE-SOAP-20000508

11 Farrel S, Housley R. An Internet Attribute Certificate Profile Authorization. http://www. itef. org/ rtf/rfc3281. txt, April 2002

12 Winsborough W H, Jacobs J. Automated Trust Negotiation in Attribute-based Access Control. In: Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX), Washington, D C, April 2003

13 OASIS Standard. eXtensible Access Control Markup Language (XACML) Version 1. 0. February, 2003. http://www. oasis-open. org/committees/xacml

(上接第 70 页)

个端点有一个单独的优化上传，不管现在的下载率如何处，它都不会阻塞。优化上传的端点，是以每三个阻塞检查周期(每周期 30 秒)循环的。30 秒对上传和下载操作足够了。

4)反冷落上传：有时某个端点服务器会被以前所有它能够下载的端点服务器阻塞。这种情况下，它只能得到较差的端点下载率，除非优化上传发现更好的端点。为了解决这个问题，如果来自某个特殊的端点单独块，在一分钟内没有响应，文件分块复制法认为它被该端点冷落了，而且除非作为优化上传，否则不会对之上传，因而对该端点作一个反冷落上传。这就导致多个同时的特殊上传(前面所提出的一个优化上传原则的特殊的反例)，将导致在系统颠簸时复制率尽快恢复。

5)仅仅上传：一旦某个端点完成了复制，它不再有可用的复制率决定那个端点去上传，现在它将变成更好的上传端点，专门用作上传，这对提高系统效率非常有益。

### 5.5 现实世界的经验

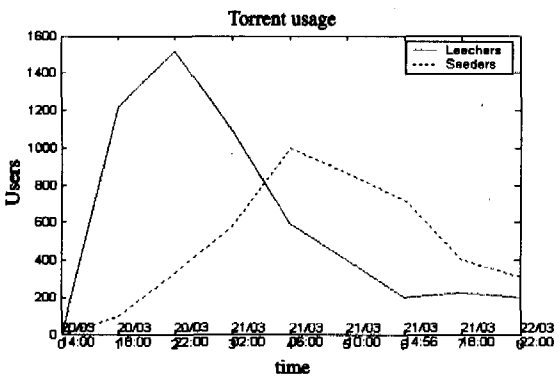



图 6 复制情况分析

这是一个大于 400 兆的文件在网格系统中提供复制的全过程，如图 6 所示，完成复制的数量(seeders 蓝线)和未完成复制的数量(leechers 红线)。其间，未完成复制的数量在文件可利用之后增加的很快，一旦达到顶点后将以指数的速度落下。相对而言，已完成复制的数量增长较慢，其峰值晚于未完成的峰值，之后缓慢下降，积分值大。未完成复制的数量成指数的剧烈增减和已完成复制的数量的稳定说明 LDAP 目录服务器的动态刷新与复制过程迅速扩展并完成，这证明文件分块复制法能进一步大幅度提高以 LDAP 目录分布式数据库为基础与核心的网格信息服务系统的效率。

本文中的文件分块复制法是源于因特网上流行的下载工具 Bittorrent。Bittorrent 不仅已经被使用，而且流传得很广，它为上百兆的文件下载服务，可面向上千个同时的下载者。

**结论和以后的工作** 网格技术使得广泛的大规模共享成为可能。网格资源信息服务是网格项目中的基础部分。文中我们根据网格资源信息的特点提出了由高度分布式的信息提供者和集合目录组成的网格信息服务基本框架，并且分析了它的基础 LDAP 协议，指出 LDAP 目录本质是一种分布式的数据库。由于网格信息系统分布广，容错性强，动态多样性，LDAP 目录信息树的动态刷新与复制将频繁发生。本文经过试验比较了多种复制策略，最终证明环形扩展和线形扩展的策略可大幅度提高系统效率；不仅如此，还用文件分块复制法，提出了技术框架和上传算法，把 LDAP 数据库文件分成若干块在多个端点间并行复制，这样进一步提高以 LDAP 目录分布式数据库为基础与核心的网格信息服务系统的并行效率。

目前我们研究了如何在节点级提高复制与刷新效率的策略，及节点级以下如何提高复制与刷新效率的策略；此外还应考虑 LDAP 目录的负载平衡问题，在什么地方复制将带来最佳的系统效果，等等。

## 参 考 文 献

1 Czajkowskiy K, Fitzgeraldz S, Fosterx I, et al. Grid Information Services for Distributed Resource Sharing. In: Proc. 10th IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), IEEE Press, 2002

2 Foster I, Kesselman C, Tuecke S. The anatomy of the Grid: Enabling scalable virtual organizations. Intl. Journal of Supercomputing Applications, (to appear) 2002. http://www. globus. org/research/papers/anatomy. pdf

3 Howes T A. Lightweight Directory Access Protocol. http://www. kingsmountain. com/directory/doc/ldap/ldap. html

4 He Yanxiang, Fan Qianfeng, Zhang Lifei. Design of dynamic replication strategies for a grid computing . Computer Engineering, 2004(2)

5 Cohen B. Incentives Build Robustness in BitTorrent 2003. 5. http://www. bittorrent. com/bittorrentecon. pdf