

对等网络

林宇¹,程时端¹,李琦²
(1 北京邮电大学,北京100876;
2 北京大学,北京100088)

2

网络资源的变迁,促使网络计算模式发生变化。相应地一种采用对等策略计算模式的网络——对等网络(P2P)得到了广泛关注。P2P是一种分布式网络,网络的参与者共享他们所拥有的硬件资源,这些共享资源能被其他对等结点直接访问而无需经过中间实体。网络中的参与者既是资源提供者,又是资源获取者。为了使读者对P2P有所了解,本讲座分3期对P2P进行介绍:上一期介绍了P2P的拓扑结构、组织模式以及核心查找算法;本期继续介绍P2P研究现状、应用情况以及下一步演进与拓展方向;下一期将讨论P2P给Internet带来的机遇和挑战,并探讨P2P对电信运营商和设备制造商的影响。

中图分类号:TP393.03 文献标识码:A 文章编号:1009-6868 (2006) 02-0057-04

2.3 结构化网络模型

结构化模型与非结构化模型的根本区别在于每个结点所维护的邻居是否能够按照某种全局特定的规则(而不是随机的方式)组织起来。结构化模型这种组织方式决定了结点之间可以方便地快速查找。

结构化对等网络(P2P)模型是一种采用纯分布式的消息传递机制和根据关键字进行查找的定位服务模式,目前的主流方法是采用分布式哈希表(DHT)技术。

分布式哈希表是一个广域范围内大量结点共同维护的巨大散列表。散列表被分割成不连续的块,每个结点被分配给一个属于自己的散列表块,并成为这个散列表块的管理者。在DHT技术中,网络结点按照一定的方式分配一个唯一的结点标识符,资源对象通过散列运算产生一个唯一的资源标识符。(类比下棋应用中,每个下棋人都会被分配一个唯一的标识,通过这个标识,通过某种运算可联系周边结点,这样所有的下棋人就被组织成了一个环)。当需要查找该资源时,通过散列运算可定位到存储该资源的结点。

经典的DHT案例包括Chord、

CAN、Pastry、Tapestry等算法。

Chord算法的主要贡献是提出了一个分布式查找协议,该协议可将指定的关键字映射到对应的结点。在Chord算法中,结点并不需要知道所有其他结点的信息,在由 N 个结点组成的网络中,每个结点只需要维护其他 $O(\log N)$ 个结点的信息,同样,每次查找只需要 $O(\log N)$ 条消息。

CAN算法采用了多维的标识符空间来实现分布式散列算法。CAN将所有结点映射到一个 n 维的笛卡尔空间中,并为每个结点尽可能均匀地分配一块区域。CAN采用的路由算法相当直接和简单,知道目标点的坐标后,就将请求传给当前结点四邻中坐标最接近目标点的结点。

Pastry算法是微软研究院提出的可扩展的分布式对象定位和路由算法,可用于构建大规模的P2P系统。在Pastry算法中,每个结点分配一个128比特的结点标识符,所有的结点标识符形成了一个环形的空间,范围从0到 $2^{128}-1$,结点加入系统时通过散列结点的IP地址在128比特空间中随机分配。

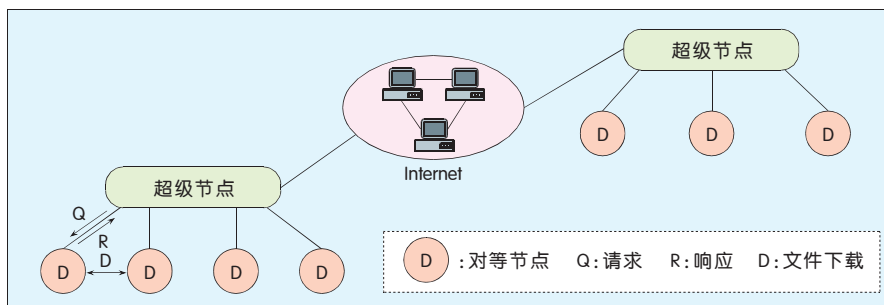
Tapestry算法的思想来源于Plaxton算法,在Plaxton算法中,结点使用自己所知道的邻近结点表,按照目的标识符来逐步传递消息。Tapestry算法在Plaxton算法的基础上,加入了容错机

制,从而可适应P2P动态变化的特点。

由于采用了确定性拓扑结构,DHT类结构能够自适应结点的动态加入和退出,有着良好的可扩展性、鲁棒性、结点标识分配的均匀性和自组织能力。DHT可以提供精确的资源发现,只要目的结点存在于网络中,DHT总能发现它。总的说来,DHT算法比较适合大规模对等网络应用,目前这种技术的应用主要集中在数据和文件共享系统上。

DHT网络结构最大的问题是DHT的维护机制较为复杂,尤其是结点频繁加入和退出造成的网络波动会极大地增加DHT的维护代价(类比下棋应用中,每次下棋人进入系统和离开系统,都需要重新调整P2P结点的拓扑)。DHT所面临的另外一个问题是由于DHT建立在精确哈希散列的基础上,因此仅支持精确关键词匹配查询,无法支持内容/语义等复杂查询(比如下棋应用中,只能通过标识来逐步发现下棋人,不能通过一些模糊信息,比如下棋人的段位信息、下棋人和自己之间的网络性能信息进行查询)。基于分布式哈希表的路由机制也有无法解决的问题,比如经过哈希运算之后,结点的位置信息被破坏了,来自同一个子网的站点很可能结点号相距甚远,不利于查询性能的优化(通过DHT建立起来的拓扑可能与

基金项目:国家“973”计划项目
(2003CB314806);国家自然科学基金项目
(90204003)



▲图5 半分布式结构(含有超级结点)

实际的网络拓扑不相符合,比如湖北和广州的两个棋友可能被分配在相邻的位置,而两者在网络通信上时延很大)。

2.4 混合式网络模型

Kazaa模型是P2P混合式模型(半分布式结构)的典型代表(如图5所示,eDonkey、eMule等也可以划分为这个类型),它在纯P2P分布式模型基础上引入了超级节点的概念,综合了集中式P2P快速查找和纯P2P去中心化的优势。Kazaa模型将结点按能力不同(计算能力、内存大小、连接带宽、网络滞留时间等)区分为普通结点和搜索结点两类。其中搜索结点与其临近的若干普通结点之间构成一个自治的簇,簇内采用基于集中目录式的P2P模式,而整个P2P网络中各个不同的簇之间再通过纯P2P的模式将搜索结点相连起来,甚至也可以在各个搜索结点之间再次选取性能最优的结点,或者另外引入一个新的性能最优的结点作为索引结点来保存整个网络中可以利用的搜索结点信息,并且负责维护整个网络的结构。

由于普通结点的文件搜索先在本地所属的簇内进行,只有查询结果不充分的时候,再通过搜索结点之间进行有限的“泛洪”。这样就极为有效地消除了纯P2P结构中使用“泛洪”算法带来的网络拥塞、搜索迟缓等不利影响。同时,由于每个簇中的搜索结点监控着所有普通结点的行为,这也能确保一些恶意的攻击行为能在网络局部得到控制,并且超级结

点的存在能在一定程度上提高整个网络的负载平衡。总的来说,基于超级结点的混合式P2P网络结构比以往有较大程度的改进。然而,由于超级结点本身的脆弱性也可能导致其簇内的结点处于孤立状态,因此这种局部索引的方法仍然存在着一一定的局限性。

半分布式结构的优点是性能、可扩展性较好,较容易管理,但对超级结点依赖性大,易于受到攻击,容错性也受到影响。

2.5 各种结构模型的性能比较

表2对中心化拓扑、全分布式非结构化拓扑、全分布式结构化拓扑、半分布式拓扑4种结构模型的综合性能进行了比较。可以看出不同结构模型在系统复杂性、可扩展性、功能上有不同的均衡性能。

3 P2P网络的典型应用

Internet最初产生和发展的主要动力之一就是资源共享。而文件交换的需求直接导致了P2P技术的兴起,这是P2P最初的应用,也是最成功的应用之一。针对这类应用的共享软件Napster使得人们在客户/服务器模式

▼表2 不同结构模型的性能比较

拓扑结构	可扩展性	可靠性	可维护性	发现算法效率	复杂查询
中心化拓扑	差	差	最好	最高	支持
全分布式非结构化拓扑	差	好	最好	中	支持
全分布式结构化拓扑	好	好	好	高	不支持
半分布式拓扑	中	中	中	中	支持

下开始重新认识P2P思想对人们使用网络习惯的影响。

随着人们对P2P思想的理解和技术的发展,作为一种软件架构,P2P还可以被开发出种类繁多的应用模式,除了最初的文件交换之外,还出现了一些分布式存储、深度搜索、分布式计算、个人即时通信和协同工作等新颖应用。其中最著名的例子是基于分布式计算的搜索外星文明的科学实验SETI@home,每个志愿参加者只需下载并运行相应软件,就可以贡献自己闲置的计算能力,参与分析Arecibo射电望远镜的无线电磁波数据并回送计算数据。另外,随着Sun公司将其JXTA协议扩展到诸如个人数字助理(PDA)和移动电话等手持终端上,并允许人们屏蔽具体的物理平台进行资料共享和文件交换等,P2P技术在移动通信和智能网领域也开始呈现出较大应用前景。

近年来,Internet上各种P2P应用软件层出不穷,P2P计算技术正不断应用到军事、商业、政府信息、通信等领域。根据具体应用不同,P2P可以分为如下类型:

(1)提供文件和其他内容共享的P2P网络,例如Napster、Gnutella、eDonkey、emule、BitTorrent等。

(2)挖掘P2P对等计算能力和存储共享能力的应用,例如Xenoservers^[25]、SETI@home、Avaki、Popular Power等。

(3)实现基于P2P方式的协同处理与服务共享平台,例如JXTA、Magi、Groove、NET My Service等。

(4)提供即时通信交流,例如ICQ、OICQ、Yahoo Messenger等。

(5)实现安全的P2P通信与信息共

享,例如 Skype、Crowds、Onion Routing 等等。

(6)实现P2P应用层组播。就是在应用层实现组播功能而不需要网络层的支持,这样就可以避免由于网络层迟迟不能部署对组播的支持而使组播应用难以进行的情况出现。应用层组播需要在参加的应用结点之间实现一个可扩展的,支持容错能力的重叠网络,而基于DHT的发现机制正好为应用层组播的实现提供了良好的基础平台。例如天天在线、QQ视频、深圳蓝波网络、PPLive等。

(7)实现Internet间接访问基础结构。为了使Internet更好地支持组播、单播和移动等特性,Internet间接访问基础结构提出了基于汇聚点的通信抽象。在这一结构中,并不把分组直接发向目的结点,而是给每个分组分配一个标识符,目的结点根据标识符接收相应的分组。标识符实际上表示的是信息的汇聚点。目的结点把自己想接收的分组的标识符预先通过一个触发器告诉汇聚点,当汇聚点收到分组时,将会根据触发器把分组转发给相应的目的结点。Internet间接访问基础结构实际上在Internet上构成了一个重叠网络,它需要对等网络的路由系统提供相应的支持。

4 P2P网络的研究和现状

4.1 中国学术机构P2P研究现状

(1)北京大学的Maze系统

Maze是北京大学网络实验室开发的一个中心控制与对等连接相融合的对等计算文件共享系统,结构类似于共享软件Napster,对等计算搜索方法类似于共享软件Gnutella。网络上的一台计算机,不论是在内网还是外网,可以通过安装运行Maze的客户端软件自由加入和退出Maze系统。每个结点可以将自己的一个或多个目录下的文件共享给系统的其他成员,也可以分享其他成员的资源。Maze系统

支持基于关键字的资源检索,也可以通过好友关系直接获得。

(2)华中科技大学的AnySee系统

AnySee是华中科大设计研发的视频直播系统。它采用了一对多的服务模式,支持部分网络地址转换(NAT)和防火墙的穿越,提高了视频直播系统的可扩展性。

4.2 企业研发产品

(1)广州数联软件技术有限公司的POCO平台

POCO是中国最大的P2P用户分享平台,是有安全、流量控制力的无中心服务器的第三代P2P资源交换平台,也是世界范围内少有的盈利的P2P平台。目前已经形成了可以支持2600万用户的能力,平均在线用户58.5万,在线用户峰值突破71万,并且全部是宽带用户。

(2)深圳市点石软件有限公司的OP平台

OP又称为Openext Media Desktop,一个网络娱乐内容平台,是共享软件Napster的后继者,它可以最直接的方式找到用户想要的音乐、影视、软件、游戏、图片、书籍以及各种文档,随时在线共享文件,容量数以亿计,号称“十万影视、百万音乐、千万图片”。OP整合了Internet Explorer、Windows Media Player、RealOne Player和ACDSee,是中国的网络娱乐内容平台。

(3)基于P2P的在线电视直播共享软件PPLive

PPLive是一款用于互联网上大规模视频直播的共享软件。它使用网状模型,有效地解决了当前网络视频点播服务的带宽和负载有限问题,实现用户越多,播放越流畅的特性,整体服务质量大大提高(2005年的超级女声决赛期间,这款软件非常地火爆,同时通过它观看湖南卫视的观众达到上万人)。

4.3 国际P2P技术应用情况

国际开展P2P研究的学术团体主

要包括P2P工作组(P2PWG)、全球网格论坛(Global Grid Forum,GGF)。目前P2PWG已经和GGF合并,由该论坛管理P2P计算相关的工作。GGF负责网格计算和P2P计算等相关内容的标准化工作。

从国外公司对P2P计算的支持力度来看,Sun公司、Microsoft公司和Intel公司投入较大。

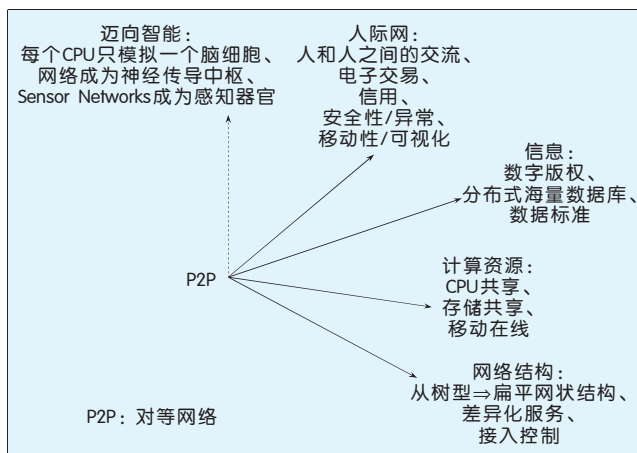
Sun公司以Java技术为背景,开展了JXTA项目。JXTA是基于Java的开放P2P平台,任何个人和组织均可以加入该项目。因此,该项目不仅吸引了大批P2P研究人员和开发人员,而且已经发布了基于JXTA的即时聊天软件包。JXTA定义了一组核心业务:认证、资源发现和管理。在安全方面,JXTA加入了加密软件包,允许使用该加密包进行数据加密,从而保证消息的隐私、可认证性和完整性。在JXTA核心之上,还定义了包括内容管理、信息搜索以及服务管理在内的各种其他可选JXTA服务。在核心服务和可选服务基础上,用户可以开发各种JXTA平台上的P2P应用。

Microsoft公司成立了Pastry项目组,主要负责P2P计算技术的研究和开发工作。目前Microsoft公司已经发布了基于Pastry的软件包SimPastry/VisPastry。美国休斯顿莱斯(Rice)大学也在Pastry的基础之上发布了FreePastry软件包。

Intel公司成立了P2P工作组开展P2P的研究。工作组成立以后,Intel公司积极与应用开发商合作,开发P2P应用平台。2002年Intel公司发布了Net基础架构之上的P2P加速工具包(Accelerator Kit)和P2P安全应用编程接口(API)软件包,从而使得微软NET开发人员能够迅速地建立P2P安全Web应用程序。

5 P2P网络的继续演进

图6表明了P2P应用进化的几个方向(网络结构、计算资源、信息、人际网、智能化)及其对网络的促动。基



▲图6 P2P应用的演化及其对网络的促进作用

于这些要素的组合，P2P应用在不断进化之中。

5.1 对计算能力的拓展

“计算拓展”意味着分散的计算机和设备的剩余处理能力的有效应用。因特网就是一个拥有巨大剩余能力的计算机。在这个虚拟的拥有数量巨大的分散CPU的超级计算机上所能做的事情超乎想象。在这方面已经有很多成功的应用，比如搜索外星文明实验和癌症基因分析等，都利用了分散的计算能力。现在，通常意义上的计算机的CPU处理能力还在不断提高，内存容量也在不断扩大。另一方面，PDA、移动电话等等也可以看成多样化的计算机。宽带通信的出现使得CPU的处理能力发生了巨大的变化。在有限的网络环境里长时接续变得平常普遍，从IPv4到IPv6的演变是不可避免的发展趋势。

5.2 对信息/数据应用的拓展

“信息/数据拓展”意味着分散数据库访问技术的发展应用。在同步检索、文件传输、备份文件保管方面，基于分散形式的数字化产品的开发体现了分散数据的活力。信息通常被保存在各种数据库、文件库、图书馆、博物馆，或者以数字化商品的形式存在。通常在分散的环境里保存保管数据的方案并不十分有效。但是在人们

离开保存信息的固定场所以后，希望通过网络检索获取信息的时候，分散环境保存保管数据的益处就体现出来了。

在因特网上，信息检索的工具从开始的Fetch、Gopher到通用的Browser，再到最新的P2P文件共享。可扩展标记语言(XML)在这种数据交换的背景下得到了

广泛应用。当然这种发展也带来数字版权 (Digital Copyright) 等亟待解决的问题。如何使P2P技术更有效合理地得到应用需要进一步研究。

5.3 人际网的拓展

“人际网拓展”意味着要确保信息发送和接收的主体同未知或已知的同伴之间可以容易地进行交流，特别是同商业伙伴之间能容易地经常地交流。像ICQ这样的即时交流工具就是P2P应用的一个典型代表。利用这些交流工具人们可以在网络上方便地同具有相同爱好的个人、组织进行交流。

人在本质上都是“移动地”存在的，对各种“位置”关心或厌恶的同时，维持着“总体”的关系，这导致人们总是有欲望同他人交流信息和资源，无论是真实的还是虚拟的。也就是说人的本性之一就是：在不安于现状，拒绝千篇一律的同时，对新奇和与众不同的事物有挑战的欲望。这种矛盾的存在是P2P进化的基本动力。

电子商务的本质使人们能够在市场交易中获得有效的竞争力。因为商业的目的，依据人们的兴趣取向的不同，各种各样的网络交流社区出现了。比如几乎同e-mail和Web一样，聊天(Chat)、空间信息网格(SIG)等越来越引人注目。现在网上拍卖也成为P2P的商业形式。进一步发展，虚拟公

司、虚拟组织会渐渐成为人们生活的一部分。

5.4 智能化方面的拓展

智能到底能否被创造？这是似乎是个遥远的问题。现有的神经系统系统的主要局限在于：

(1) 单个神经元及其传导模型仿真过于简单。

(2) 神经元系统只能模拟数量在几百万量级的神经元系统，而人脑有几十亿甚至几百亿的神经元。

(3) 对人脑的工作机理了解不足。

在P2P分布式网络下，每一个结点可以模拟足够复杂的神经元及其传递模型，只要网络足够大，就可能达到几百亿结点的规模，再加上有线/无线传感器网络作为系统的“感知器官”，已经有了创造智能的条件。

(待续)

收稿日期：2005-11-22

作者简介



林宇，北京邮电大学网络与交换国家实验室副教授，博士。主要研究方向包括QoS控制、测量和管理、P2P技术、移动应用等。曾作为主持研究人和项目负责人参与国家“973”、国家自然科学基金、国家“863”项目10余项。已出版专著3部，发表论文60余篇，论文被SCI收录8篇，EI收录近30篇。



程时端，北京邮电大学教授、博士生导师，北京邮电大学学术委员会副主任。目前研究方向包括ISDN、ATM、TCP/IP、ATM和IP网的话音通信技术、协议工程、流量工程、宽带网络性能和服务质量等。



李琦，北京大学教授、博士生导师，中国图形图像学会技术委员会主任，全国地理信息标准化技术委员会委员，国家电子标签标准工作组成员，国家海洋局信息中心特聘教授，国务院信息化工作办公室顾问，北京市政府专家顾问，中国数字地球发展战略研究专家。