DHT网络中基于测量的QoS监控系统

李莹峰，邓晓衡
（湖南工业大学 计算机与通信学院，湖南 株洲 412008）

摘 要：为对P2P系统提供QoS支持，深入研究了现有的分布式哈希表、网络测量技术和网络测量系统架构，并在此基础上设计了一个架构在DHT网络上的网络测量系统。该测量系统对DHT网络中的Peer之间的网络性能与Peer自身资源进行测量、分析、存储与发布，为DHT网络的有效利用提供决策支撑。实验表明该系统具有较好的可扩展性、自治性、多种测量工具协同能力。

关键词：对等网络；分布式哈希表；全分布式结构化P2P网络；网络测量；服务质量
中图分类号：TP393 **文献标识码**：A **文章编号**：1673-629X(2009)05-0188-04

Measurement-Based QoS Monitor System in DHT Network

LI Ying-feng, DENG Xiao-heng
(School of Computer and Communication, Hunan University of Technology, Zhuzhou 412008, China)

Abstract: A network measurement system is designed, which is built on DHT network, to provide QoS service for P2P systems, through in-depth study of the existing distributed Hash table, network measurement technology and network measurement system architecture. The measurement system is used to measure, analyse, storage and publish network performance from peer to peer and peer's resources and it is able to afford decision support to DHT system. The experiments show that the system has better scalability, autonomy, capacity of collaboration in a variety of measurement tools.

Key words: P2P; distributed Hash table; DHT network; network measurement; QoS

0 引 言

最近几年，Peer-to-peer技术快速发展，P2P系统的拓扑结构也不断发生变化，在已有的拓扑结构中，采用分布式哈希表(Distributed Hash Table，DHT)技术的全分布式结构化拓扑(Decentralized Structured Topology)也被称为DHT网络[1]。当前P2P领域的最新成果集中在DHT技术方面，其中产生了四种典型的DHT机制：Chord[2]、CAN[3]、Pastry[4]和Tapestry[5]。可以有效地完成节点组织和搜索定位，并具有较好的搜索效率和扩展能力，所以现在也产生了一系列的后继改进和应用项目，并引导了P2P领域的发展趋势。

DHT技术为DHT网络提供节点自我管理、关键字查询定位等服务，通常被认为是一种Overlay层，它不考虑网络底层的物理连接关系，仅仅假设应用层之下的传输层、网络层和数据链路层能够正常连接，并提供足够的带宽等底层网络性能参数，保证上层应用的正常进行[6]。然而在某些P2P应用，特别是一些网格运算和Web服务应用中，仅仅假设底层网络性能参数能保证系统正常运行是不够的，这些系统中需要了解网络和节点性能的具体参数，以保证系统的优化运作和对服务质量(Quality of Service，QoS)的支持，所以有必要在DHT层中提供一定的网络和节点性能监控。

文中的主要工作是在DHT网络中引入网络测量思想，通过对DHT网络系统的节点间网络性能和节点性能测量，提取网络中的基础物理数据进行存储与分析，为P2P应用项目提供网络性能监控与评估，并根据DHT技术的层次结构特点提出整个测量系统的设计结构和实现方式。

收稿日期：2008-09-01
基金项目：中国博士后科学基金(20060400879)；湖南省自然科学基金(06JJ30032)；湖南省高等学校科学研究项目(07C231)
作者简介：李莹峰(1980-)，男，硕士研究生，研究方向为计算机网络、网络测量；邓晓衡，博士后，副教授，硕士生导师，研究方向为网络计算、网络拥塞控制与网络测量。

1 研究现状

QoS是指网络提供更高优先服务的一种能力，同时也是用来解决网络延迟和阻塞问题的技术，它通过提供专用带宽、控制网络抖动和延时、改进网络丢包率等方式进行控制，但其控制的基础仍然是网络底层的

性能。网络测量是指遵照一定的方法和技术，利用软、硬件工具测量或验证表征网络性能指标的一系列活动。它的主要任务是收集网络中可直接观察到的物理现象与参数，为网络行为的分析与 QoS 技术提供基础数据。

网络测量技术伴随着计算机网络的发展而发展，现已形成一些较成熟的测量方式与测量架构。根据测量方式的不同，网络测量方法可以分为主动测量和被动测量两种，文献[7]就 P2P 网络对这两种方法进行了详细说明，并将现有的 P2P 测量按照目的进行了分类，详细介绍了国内外的主要研究内容、研究项目等。然而一个测量系统在实现测量方式时不但包括了数据采集部分，还包括数据存储与分析部分等，因此测量系统的体系架构和实现也是网络测量的主要研究内容。国内外也提出了多种网络测量架构方案和系统，如：NIMI[8]、GMA[9]、Advisor[10]等。虽然在网络测量方式和测量系统架构上已经存在了一些研究成果，其中包括 P2P 系统内的测量理论与系统，但是缺少针对 DHT 网络特点的测量系统，伴随着 DHT 技术与网格技术的发展，在 DHT 网络中同样需要进行测量研究。

笔者在参考上述文章的基础上提出一种架构在 P2P 系统上的，专门针对 DHT 网络的网络性能测量架构与系统，对处于互联网中的 DHT 节点性能以及网络端到端性能进行测量、存储与分析，并将分析所得到的数据进行存储，为 P2P 网络 QoS 提供依据。

2　系统设计

由于 P2P 系统和 DHT 网络存在一些特性，所以在 DHT 网络中引入性能监控时也需要考虑这些特性的影响。P2P 系统取消了 C/S 结构中的中央服务器节点，那么在进行测量时，由某一个节点测量网络中的所有节点的性能，然后向所有节点发布信息显然不合理，也违背了 P2P 系统的设计初衷。根据这一特点，提出在网络性能的测量中也需要相应地引入由各节点自主测量，并为自身或其他节点提供服务的思想，并将测量系统集成到 DHT 节点中。如图 1 所示，测量系统应该作为 DHT 网络节点中的一个模块，独立进行网络测量，获取、存储测量结果，为 DHT 节点和 P2P 应用程序提供数据操作接口，减少与 P2P 系统的耦合度。其中独立测量表示不需要依赖 DHT 搜索机制，同时也与 P2P 应用程序无关。

测量系统采用分层架构，将测量系统的数据采集层、数据存储与分析层、数据发布层进行分离。数据采集层主要解决获得网络物理性能和 P2P 节点计算机性能数据；数据存储与分析层主要按照一定格式将取得的数据分类存储，并通过性能分析对网络性能进行评估；数据发布层主要是定义取得性能数据和分析结果数据的函数，并按照某种方式在 DHT 网络中进行传输。三个层次又分别可以划分成几个小模块，如图 2 所示，网络性能数据采集层包括网络性能测量工具和测量工具管理程序两个模块；测量数据存储分析层包括测量数据管理程序、测量数据分析程序和数据存储三个模块；性能数据发布层主要是性能数据发布接口。

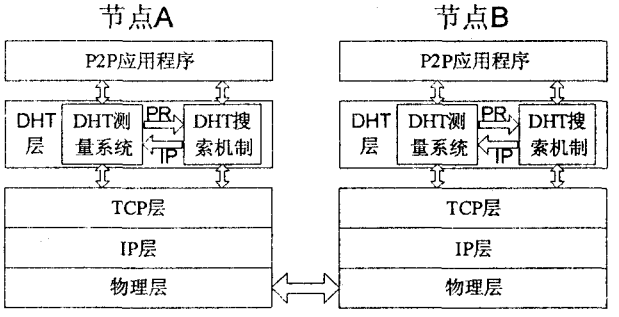

图 1　测量监控系统网络层次关系图

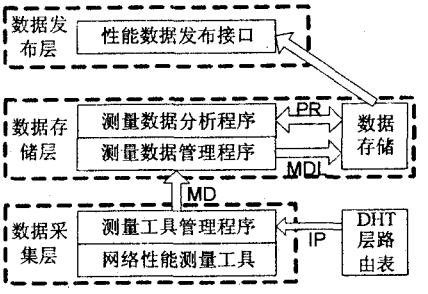

图 2　测量监控系统层次架构图

2.1　数据采集层

测量系统不直接对网络性能进行测量，通过调用现有测量工具来实现测量数据的获得，从而提高测量系统的可扩展性。所测量的网络性能参数包括节点间端到端的网络带宽、传输延时、丢包率、被测节点的 CPU 和内存利用率、硬盘容量等。测量工具管理程序用来控制和管理网络性能测量工具，其主体是实现从现有 DHT 网络系统节点的分布式哈希表所维护的路由表中提取网络节点的 ID 和对应的物理 IP 地址、端口号，然后将物理 IP 和端口号以参数形式传送给测量工具，调用测量工具进行网络物理性能数据采集。主体以外部分是对多种测量工具进行调用过程的实现。另外该部分还需要实现数据采集的测量采样间隔和触发机制，这是由于 P2P 网络中的所有节点是可以动态加入与退出的，同时大量节点存在于边缘网络中网络性能并不稳定。

2.2　数据存储层

数据存储层与数据采集层之间通过接口联系，接口由数据存储层提供，由数据采集层进行调用，将每次

测量所获得的性能数据传送给测量数据管理程序。测量数据管理程序控制传送过来的数据的存储,它将不同测量工具所取得的测量数据分类,并按照一定的数据结构进行封装,并将封装后的数据进行存储。数据存储模块作为该层的中心,其工作是负责将所分类好的测量数据进行存储,提供测量数据分析程序接口,由测量数据分析程序提取数据分析,并将分析后的结果返回存储到数据存储部分。测量数据分析程序主要负责对测量数据进行分析,部分测量工具在测量过程中会对测量数据进行初步计算,得到网络性能的一些参数,这些参数对于一些应用程序已经可以作为性能衡量标准,而一些测量工具所得到的只是原始数据,必须由用户做分析,针对这类测量数据,系统引入测量数据分析模块。由于测量数据的采集是有一定的采样间隔和触发机制,因此数据分析过程也需要按照一定的时间间隔,提取数据存储部分中所存储的最近采集数据进行分析。

2.3 数据发布层

数据发布层采用请求 - 回应机制,它与数据存储层之间通过数据存储层提供的接口进行联系,按照一定条件取得测量数据或性能分析数据,为 DHT 网络中其他节点提供服务。当 DHT 网络中节点需要进行性能优化的服务时,调用数据发布层的性能查询函数,性能查询函数发起性能查询请求,被查询的节点在接收到性能查询请求后,按照请求查询的性能类型从数据存储部分取得对应的性能分析数据或测量数据,将数据返回主动查询节点。最终优化组合方式由 P2P 应用程序按照自身要求独立制定。

3 系统实现

根据上节对 DHT 网络特点和测量系统架构分析,测量系统调用测量工具,取得网络基础设备上的物理数据和 DHT 网络节点性能数据,并提交分析、存储与发布的整个过程中,需要对测量工具、测量数据、性能分析数据等进行控制。因此,在实现系统设计过程中定义了四种数据结构,用来管理与控制这些基本信息。TS(ToolSets),测量工具集合,用来预设可以进行网络测量的测量工具,它包括预设的测量工具标识符、测量工具名称、测量工具类型说明等数据元素;PT(PerformanceTypes),性能类型,DHT 网络中可以测量得到的基本性能,它包括可测量的节点性能标识符,可测量的节点性能类型等数据元素,其中性能类型包括带宽、延时、丢包率、CPU 利用率等网络和计算机节点性能;MDL(MeasureDataList),测量数据封装列表,将用测量工具所得到的直接测量数据 MD(MeasureData),按照 PT 分类并按照一定格式封装,方便存储与系统内的分析调用,它包括测量节点 DHT 标识符,被测节点 DHT 标识符,测量时间,测量工具类型,测量结果等数据元素;PR(PerformanceResult),性能分析数据,对 MDL 数据进行分析后得到的性能分析数据,按照 PT 分类。

3.1 数据采集层

在 DHT 节点中系统维护 DHT 路由表,同时将维护<key, value>数组,value 中存储的信息包括对应 key 的 DHT 节点 ID、物理 IP 地址和通信端口号,因此,测量工具管理程序在 DHT 节点运行时读取节点所维护的 DHT 路由表,获得节点中所存储的 value 中的信息,主要是被测量节点的物理 IP 和端口信息,并将其作为参数传递给测量工具,由测量工具独立进行测量并返回测量结果。在对测量工具的管理中,测量系统参考每种测量工具所提供的操作函数,实现调用方式,并根据所有的测量工具和测量的性能定义 TS 和 PT 数据结构中对应数据元素。对测量的时间间隔与触发机制,测量系统针对不同情况分开进行:DHT 节点在稳定运行时测量工具管理程序采用定时测量,其采样时间间隔同样定义为周期时间 T;DHT 节点在发现有节点动态加入和退出系统时,首先向测量系统发送禁止测量消息,暂停 DHT 路由表更新这段时间内的测量系统自动进行的网络测量,然后对其所维护的 DHT 路由表进行更新,更新后进入新的稳定运行状态,当重新达到 DHT 路由表的稳定状态后,DHT 节点向测量系统发送重新测量消息,测量系统从 DHT 路由表中重新取得更新后的内容,独立进行网络测量。在这一过程中 DHT 节点对测量系统进行了干预,目的是减少无用网络测量,一定程度减少测量过程对网络性能的影响。

3.2 数据存储层

数据采集层在采集到网络性能数据后,调用数据存储层的接口,将测量数据与所用测量工具类型 TS 传送给测量数据管理程序。后者接收数据后开始工作,按照 TS 类型和 PT 类型将测量数据封装成 MDL 类型,然后存入数据存储模块中。数据存储模块使用现有的微型数据库,便于采用统一接口对数据进行存储与获取。将测量数据按照 TS 和 PT 类型分类存储之后,测量数据分析程序按照一定的时间间隔和触发机制,调用数据存储模块中需要进行分析的原始测量数据进行分析。在分析程序所采用的时间间隔和触发机制中,除了调用测量工具所采用的时间间隔与触发机制外,数据分析过程所用时间也是需要考虑的问题。由于当前所用测量工具的原因,在实现系统时暂没有考虑数据分析过程所用时间,因此该部分采用等待时

间周期 T 的时间间隔机制。测量数据分析程序在数据存储模块中,按照采集时间大于前次分析时间与采集时间小于或等于当前分析时间作为查询条件,确定是否在两个时间点中测量系统取得了新的测量数据。如果没有取得新数据则表明 DHT 路由表中信息可能在进行更新,系统测量过程已经被禁用,因此,对应的分析过程将停止,如果取得新数据则表明系统稳定的按照周期 T 在对网络状况进行测量,因此,对应的分析过程启动。测量数据分析程序在完成分析后将分析结果回存入数据存储模块,用于数据发布层进行查询与发布。

3.3 数据发布层

测量系统的性能数据发布接口主要是实现请求-回应机制,其主要是作为与 DHT 网络中其他节点查询其中一个节点所维护的性能数据的接口。在具体实现时使用消息机制在节点间进行通讯,消息定义出所进行的操作以及操作所针对的 PT 类型,这里的操作主要是查询数据,包括查询性能数据、测量数据以及数据条数。其运作过程为:请求查询的 DHT 节点向网络中被查询节点发送一个包含 PT 类型的查询消息;被查询节点接收到上述消息后,根据查询的性能类别与数据条数,调用数据查询函数对性能进行查询;查询函数返回最近 n 次的数据,其中 n 为查询请求节点发送的数据条数;被查询节点最后将查询所得数据返回给请求节点。

4 结束语

在深入研究 P2P 技术和网络测量领域的已有研究成果的基础上,设计与实现了架构在 DHT 层上的针对 DHT 网络的三层架构网络测量系统。该系统能对 DHT 网络中节点间网络性能以及节点自身性能进行测量、存储、分析和发布,为 DHT 网络的优化利用提供决策支持。测量系统参考 GMA 系统架构,具体设计与实现了如下内容:设计提取 DHT 节点所维护的 DHT 路由表中节点信息的方式;对测量工具的调用与管理方式;测量采样过程的触发与控制机制;测量数据的存储系统;测量性能的发布机制。

在现有的 P2P 系统特别是 P2P 文件分享系统中存在多对多的通讯与服务情况,单纯考虑 DHT 节点间一对一的通讯与测量还存在不足。基于小世界理论的 DHT 思想保证了查找的可达性问题,但长链路上的性能不是其关注重点,如何综合考虑保证整条长链路上的性能也将是需要进一步考虑的问题。同时,伴随着测量工具的引入,测量分析程序、系统内的时间控制与触发机制都有改进空间。

参考文献:

[1] 罗杰文. Peer-to-Peer 综述[EB/OL]. 2006-08-25. http://www.intsci.ac.cn/users/luojw/P2P/index.html.

[2] Stoica I, Morris R, Karger D. Chord: a scalable peer-to-peer lookup service for internet applications[C]//Proceedings of ACM SIGCOMM 2001. San Deigo, CA:[s.n.], 2001: 149-160.

[3] Ratansamy S, Francis P, Handley M, et al. A scalable content-addressable network[C]// Proceedings of ACM SIGCOMM 2001. San Deigo,CA:[s.n.], 2001:161-172.

[4] Rowstron A, Druschel P. Pastry: scalable, distributed object location and routing for large-scale peer-to-peer systems [C]//IFIP/ACM International Conference on Distirbuted Systems Platforms (Middleware). Heidelberg:[s.n.], 2001: 329-350.

[5] Zhao Y B, Huang L, Stribling J, et al. Tapestry: a resilient global-scale overlay for service deployment[J]. IEEE Journal on Selected Areas in Communications,2004,22(1):41-53.

[6] Gribble S D, Brewer E A, Hellerstein J M, et al. Scalable, Distributed Data Structures for Internet Service Construction [C]//In Proceedings of OSDI 2000. San Diego:[s.n.], 2000:68-76.

[7] 刘　琼,徐　鹏,杨海涛,等. Peer-to-Peer 文件共享系统的测量研究[J]. 软件学报, 2006,17(10):2131-2140.

[8] Paxson V, Mahdavi J, Adams A, et al. An Architecture for Large-Scale Internet Measurement[J]. In IEEE Communications,1998,36(8):48-54.

[9] Tierney, Aydt R, Gunter D, et al. A Grid Monitoring Architecture[EB/OL]. 2002-01-16. http://www-didc.lbl.gov/GGFPERF/GMA-WG/papers/GWD-GP-16-2.

[10] Latter T, Kutzko M, Engelhardt S, et al. Network Performance Advisor[EB/OL]. 2005-09. http://dast.nlanr.net/projects/advisor/.

(上接第 153 页)

研究[J]. 计算机工程,2007(5):58-60.

[5] Kunz T. The Influence of Different Workload Descriptions on a Heuristic Load Balancing Scheme[J]. IEEE Trans. on Software Eng,1991,17(7):725-730.

[6] Balter M H, Downey A B. Exploiting Process Lifetime Distributions for Dynamic Load Balancing[J]. ACM Transactions on Computer Systems,1997,15(3):253-285.

[7] 陆克中,林晓辉. MPI 并行程序设计的负载平衡实现方法[J]. 微计算机信息,2007(3):226-227.

[8] Zambonelli F. Exploiting Biased Load Information in Direct-neighbour Load Balancing Policies[J]. Parallel Computing, 1999,25(6):745-766.