P2P应用层数据流量优化

任立勇，雷 明，张 磊

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】为减少对主干网络带宽的消耗，提出了一套数据流量优化方法。首先通过把Peer间的邻居关系明确划分为物理邻居关系和逻辑邻居关系，并用探路者算法来发现Peer的物理邻居，实现拓扑匹配；然后在数据调度算法中，引入通报/退避机制，把大部分数据传输控制在城域网络内部。通过建立模拟仿真环境进行实验分析，证实了该方法可以减少90%以上的网络数据流量。

关 键 词 邻居； 优化； P2P； 调度算法； 拓扑

中图分类号 TP311 **文献标识码** A doi:10.3969/j.issn.1001-0548.2011.01.021

Data Traffic Optimization in P2P Application Layer

REN Li-yong, LEI Ming, and ZHANG Lei

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract At present, too much network bandwidth has been occupied by a variety of P2P applications that greatly affect the backbone of the network. A scheme for data flow optimization is presented to reduce the consumption of the network bandwidth. In the scheme, the peers relationship is divided into physical and logical neighbors, the pathfinder algorithm is used to determine the relationship between peers and achieve topology matching, and the notice/drawback mechanism is introduced in the data scheduling algorithms to make most data transmission under the control of metropolitan area network. The experiment results of a simulation environment prove that the scheme mentioned above is an effect way of reducing the network traffic by 90%.

Key words neighbor; optimization; P2P; scheduling algorithms; topology

P2P应用的迅速发展，大量消耗着网络带宽资源。有关资料显示，中国电信现有骨干网流量有很多都属于P2P流量，在东部地区某些省份白天可高达60%～70%，夜间更高。这些应用之所以会占用如此之大的网络带宽，一方面是由于P2P系统中的节点(peer)在选择邻居时的随机性造成系统逻辑拓扑与物理拓扑失配；另一方面是由于系统中的peer在下载数据时的随机性和无序性。BitTorrent[1]、CoolStreaming/DONet[2]、SplitStream[3]、ZIGZAG[4]、PeerStreaming[5]这些典型的P2P都没有或很少涉及拓扑匹配问题，也没有基于拓扑性的数据调度。

减少P2P应用对网络带宽消耗的关键是：做到拓扑匹配，使P2P网络的节点的逻辑关系与其在物理网络中的关系相适合，并在此基础上合理安排数据调度，减少网间数据交互。但是，优化不应以破坏P2P应用的原有特性为代价而获得网络流量的降低，否则容易造成系统鲁棒性、稳定性、效率等降低，以及形成数据孤岛节点群等问题。

目前在拓扑匹配方面的研究可分为：1) 依靠特定服务器静态返回具有拓扑邻近性节点表的方式[6]；2) P2P网络中的节点自己进行动态探测节点间物理拓扑邻近关系，并自动调整节点的逻辑位置的方式[7-10]。在拓扑匹配基础上的数据调度优化可分为集中调度管理和自组织调度管理方式。集中方式如建立超节点[11-12]；自组织方式多以提高获取数据的效率为目的，较少考虑减少网络流量问题。

本文建立了一套减少网络总体流量的通用优化方法，该方法以节点间关系是邻居关系(或伙伴)，数据调度以拉方式工作的网状拓扑P2P应用系统架构为基础，将邻居关系进一步划分成物理邻居和逻辑邻居。逻辑邻居关系的建立过程就是所要优化的P2P应用系统的原有邻居关系的建立过程，本文不作详述。物理邻居关系以探路者算法建立，邻居之间的数据调度方式不作大的改变，仅在数据下载时引入

收稿日期：2009－07－14； 修回日期：2009－11－16
基金项目：国家自然科学基金(61073181)
作者简介：任立勇(1971－)，博士，副教授，主要从事分布式计算、高级操作系统方面的研究.

通报/退避机制，以减少主干网上的总数据流量。该方法可适用于P2P流媒体应用和P2P文件下载。

1 划分物理邻居和逻辑邻居

现有P2P系统很多采用随机选择邻居的方式，随机性可以带来很好的系统鲁棒性[13]，并且不会使系统太过复杂，因而具有很强的实用性，但不可避免地会造成系统逻辑拓扑与物理拓扑失配。

将节点间邻居关系明确划分为物理邻居关系和逻辑邻居关系，逻辑邻居关系是指两个物理拓扑上相隔较远(往返时延较大)的节点之间结成的邻居关系。物理邻居关系是指两个在物理拓扑上相隔很近(往返时延较小)的节点之间结成的邻居关系。这种划分既可以通过逻辑邻居保持原有系统的特性，又可以通过建立物理邻居关系实现节点关系与物理拓扑相匹配。在做数据调度时可以因邻居关系的不同而采用不同的优化措施。两种邻居的关系如图1所示。

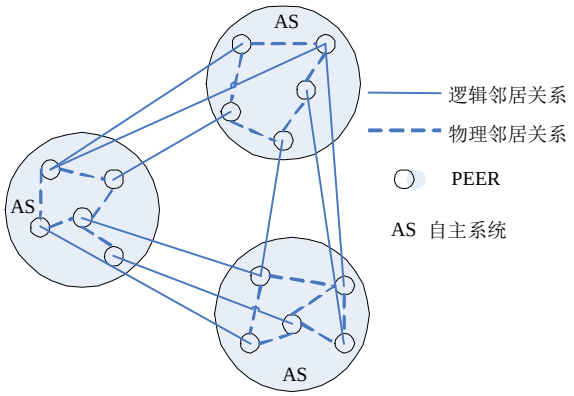

图 1 物理邻居关系与逻辑邻居关系

物理邻居关系主要体现节点间的物理距离，因此该关系相对逻辑邻居关系更加稳定，不会因为暂时数据交互量变少或网络拥塞而被破坏。逻辑邻居关系主要以交互数据为主，一旦相互之间交互的数据变少就很容易切断该邻居关系，并与其他节点建立新的邻居关系。

2 物理邻居关系建立与探路者算法

建立物理邻居关系需要做两方面的工作：首先让节点掌握足够的其他节点的信息，然后从这些信息中判断出离其很近的节点作为自己的物理邻居。探路者算法可以完成这两方面的工作。

探路者指由某节点发出的数据包，该数据包被收到它的其他节点转发，数据包里记录了其所经过路径上一定数量的节点信息。每个节点都周期性地发出探路者，探路者经过*N*跳后被撤消。

下面用伪代码对探路者算法进行描述：

```
Input:
Packet. Type; /* Pathfinder, Pathfinder_answer,
RTT_probe, RTT_answer */
LogNeiTable; //Logical neighbor table
PhyNeiTable; //Physical neighbor table
MinLog; //minimum logical neighbors
MinPhy; //minimum physical neighbor
LimitRTT: //maximum Phyneighbor RTT
Algorithm:
Switch (Packet. Type)
Case Pathfinder:
    Send(Pathfinder_answer) to source Neighbor;
    If (LogNeiTable.size<MinLog) then
        RequestLogNei(selectpeer(Pathfinder.
    Peertable, MinLog–LogNeiTable. size);
    If (PhyNeiTable. size<MinPhy) then
        Foreach (peer in Pathfinder. peertable)
            If (peer is not neighbor)
                Send(RTT_prober) to peer;
    If(Pathfinder.TTL >1) then {
        Add self peer into Pathfinder;
        Pathfinder. TTL－－;
        Sleep(some time);
        Send (Pathfinder) to a Neighbor;
        }
    Break;
Case RTT_probe:
    Send (RTT_answer) to source neighbor;
    Break;
Case Pathfinder_answer:
    RTT=this time–send time;
    If (RTT<LimitRTT) then
        Change LogNeighbor to PhyNeighbor;
    Break;
Case RTT_answer:
    RTT=this time–send time;
    If (RTT<LimitRTT) then
     RequestPhyNei (this peer);
    Break;
Case other: /*其他部份*/
```

本文只以RTT为判断依据，如果结合其他指标进行，物理邻居关系将更加精确。

3　通报/退避机制

拓扑匹配可以使物理相邻的节点汇聚在一起，但不能防止在同一群物理邻居节点内的多个节点同时向各自的逻辑邻居处下载数据，因此在数据调度方面，需要在物理邻居之间建立一种协商机制，使同一个数据只有一个或少数几个副本进入同一个物理邻居群。通报/退避机制可以以少量的通信和时间代价完成这种协商。该机制不涉及数据的搜索，仅在数据下载时进行优化。

该机制的基本思想是，节点向逻辑邻居请求某片数据前，先向所有物理邻居发出通告并进入等待期，通告中含有往返时延代价，确定没有冲突后再向逻辑邻居请求该片数据；如果有冲突就比较代价，由代价较小的节点向逻辑邻居请求数据，另外一个节点就进入退避期，在退避期内的节点收到通报立即回送一个拒绝消息。

通报/退避机制的状态转换图如图2所示。图中，等待期的超时时间设定为物理邻居间的最大往返时延；退避期的超时时间设定为逻辑邻居的最大往返时延。

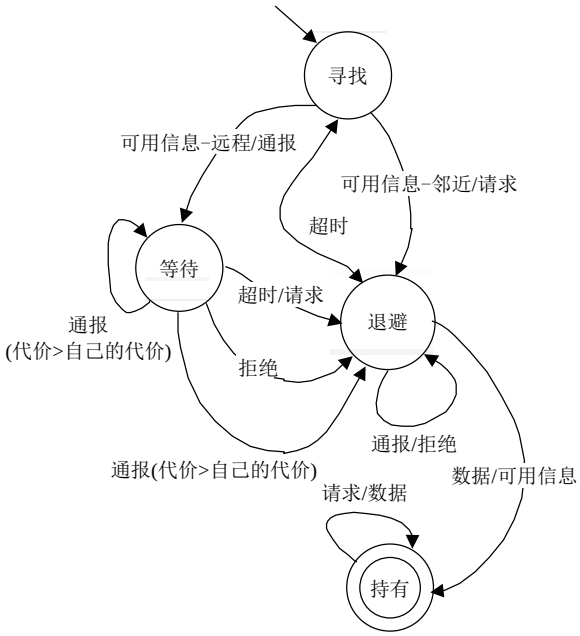

图2　通报/退避机制

通报/退避机制减少网间数据流量的过程如图3所示。图3中的节点之间都存在物理邻居关系，并且都没有数据片A。1、2号节点首先试图向自己的逻辑邻居下载数据片A，1号节点分别向2、3、4、5、6、7号节点发送请求通报信息；2号节点也同时向1号节点发送请求通报信息，如果1号节点的时间代价小于2号节点的时间代价，2号节点进入退避期，3、4、5、6、7号节点也进入退避期，1号节点向其逻辑邻居发出数据请求。在此期间，13、18号节点也试图向自己的逻辑邻居下载数据片，它们分别向5、6、17号节点发送请求通报，5、6号节点正处于退避期内，5、6号节点分别向13、18号节点发送拒绝信息，13、18号节点也进入退避期。正常情况下，退避期到期前，1号节点已经下载到了数据片A，并已将数据通过物理邻居传遍了所有18个节点。如果1号节点下载失败，又会重复上述过程。

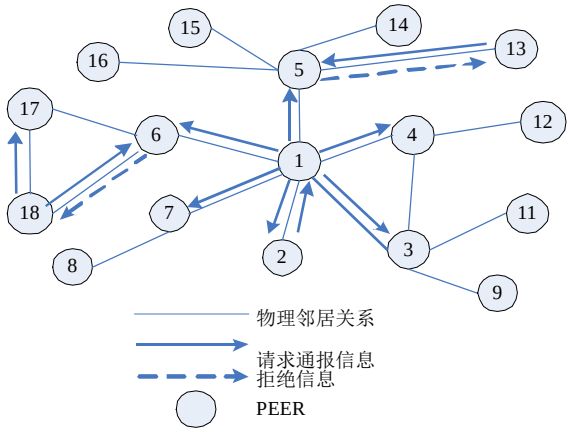

图3　通报/退避机制示例

通过通报/退避机制可以保证当一个节点下载某数据片时，其物理邻居及物理邻居的物理邻居都不会从逻辑邻居处下载同一片数据，从而能减少网络数据流量。

4　仿真分析

本文以CERNET为背景设计仿真环境，该仿真环境为3层网络模型。首先生成一个国内主干网络(以下简称一级网络)，在国内主干网上生成20个省内主干网络(以下简称二级网络)，在省内主干网上共生成200个城域网络(以下简称三级网络)，其中城域网络包含了局域网的功能。在该3层网络中分别引入10～15、1～10、0.3～1 ms的时延，并实现按IP转发的功能；在城域网络中还实现地址转换功能。每个网络可单独计算进入本网络的数据流量。

生成网络后，在三级网络上生成一定数量的节点，这些节点实现了3种协议：1) 类CoolStreaming协议(简称协议1)；2) 协议1基础上引入的探路者算法(简称协议2)；3) 协议2基础上引入的通报/退避机制(简称协议3)。

本文分别在仿真系统中生成1 000、2 000、5 000和10 000个节点进行仿真，测试分析以上3种协议下的网络流量、系统建立物理邻居的速度以及物理邻居数与网络流量的关系等问题。

为便于比较，采用网络流量度*d*衡量进入网络的流量。*d*定义为：单位时间内某层网络中的所有网络从外部收到的数据流量之和与所有节点收到的有效数据之和的比值。

仿真系统动态生成节点，起始过程中，每10 ms生成一个节点。从开始生成节点到运行650 s，系统中节点平均物理邻居数及网络流量度都基本达到稳定(以下简称稳定过程)，系统记录该过程中每个节点和每个网络每秒收到的数据量。第640 s时各级网络的网络流量如图4所示。

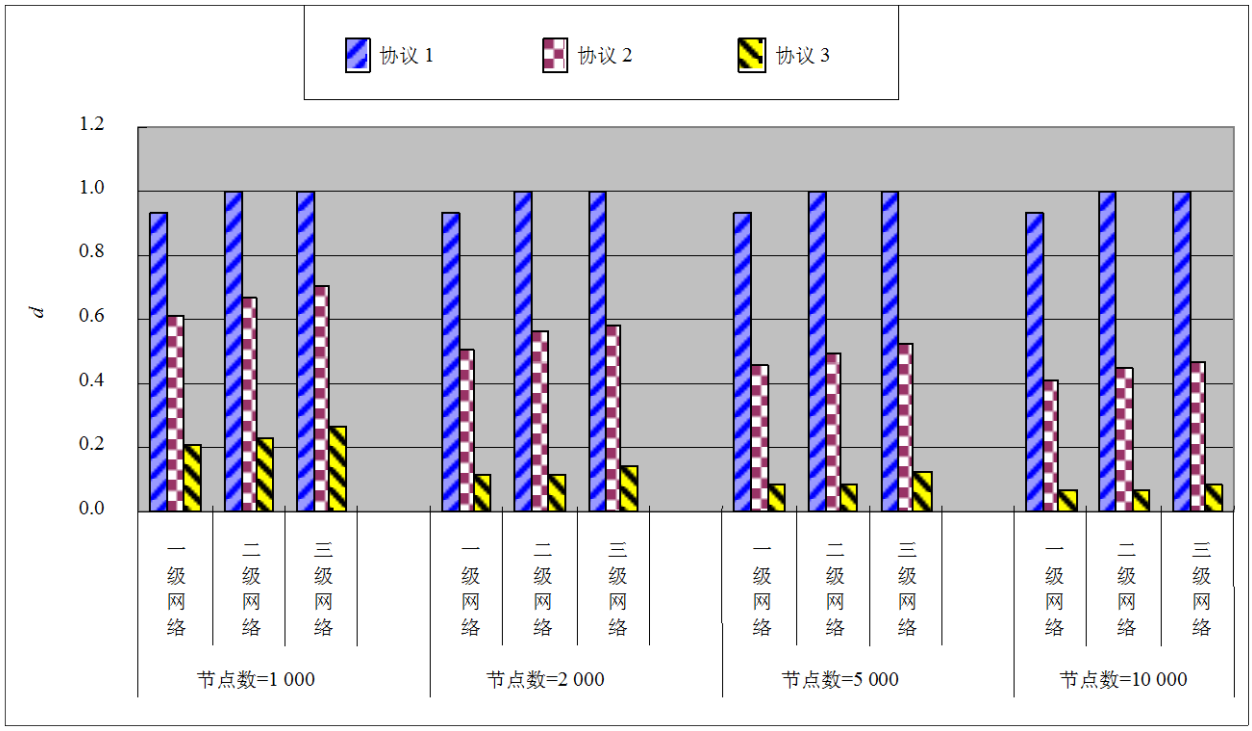

图4 3种协议下的网络流量度

由图4可看出，没有优化前，节点所收到的数据绝大部分(93%以上)来自其他网络的节点，几乎都穿过了一级和二级网络，并且各级网络的网络流量度也不会因为系统中节点的变化而明显变化。当节点建立了物理邻居之后，网络流量度降低了40%～60%(与文献[6,12]所述相吻合)，此时有一半左右的数据没经过其他网络，传输被限制在三级网络内部。在采用了通报/退避机制之后，只要系统中的节点达到一定规模(5 000个以上)，网络流量度可降低90%以上，尤其是当节点数在10 000以上时，进入二级网络以上的数据只占全部节点收到的数据的7.2%，节点得到的数据有93%来自于同一个三级网络内部。

协议3下系统由初始化进入稳定过程中的网络流量度及节点平均物理邻居数分别如图5和图6所示，清楚地反映了物理邻居数据与网络流量度之间的关系。

系统初始过程中不断地生成节点，因此需要一定的时间达到稳定。在稳定前，由于节点没有或只有很少的物理邻居，因而网络流量度较高。由图5和图6可看出，在生成完所有节点100 s后，系统趋于稳定，并且随着物理邻居的增加，网络流量度也成反比地下降。

从图6还可看出，当节点数量为1 000时，每个三级网络的节点数较少(平均为5)，因此平均物理邻居数较少(平均为4.1)。随着节点的增多，物理邻居数也会有一定程度的增加，但当节点数量大于5 000后，节点的平均物理邻居数稳定在7.5左右，不会再因节点的增加而明显增加。这是因为探路者只能记录20个节点造成的，节点能感知的范围有限。若想增加节点的平均物理邻居数，可以通过加大探路者记录的节点个数来实现，但也不能无限制地增加节点数，因为这样会加大探测开销，也会使系统收敛时间增长。

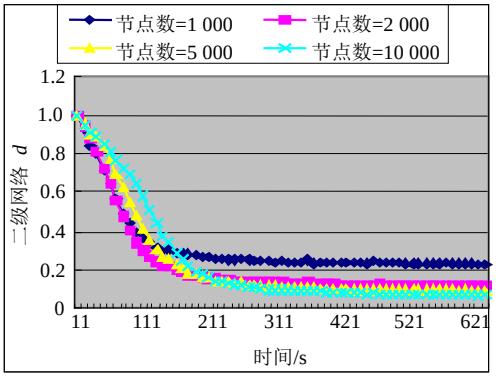

图5 系统初始过程中进入二级网络的网络流量度

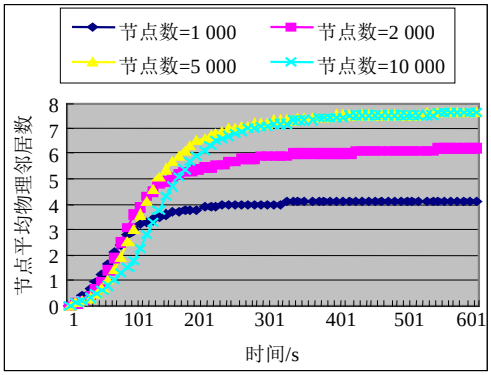

图6　系统初始过程中节点的平均物理邻居数

5　结 束 语

P2P系统中节点组织与数据调度是两大核心问题，以前的P2P系统大都主要考虑系统本身的适用性、鲁棒性、稳定性和效率等问题，占用了太多的网络资源。

本文通过划分节点邻居关系为物理邻居与逻辑邻居关系，用探路者算法建立物理邻居并引入通报/退避机制，形成一套优化方案，在不破坏原有P2P系统性能的前提下，实现节点组织的拓扑匹配及物理邻居之间的有效协作，将大量数据传输控制在城域网络内部。

通过建立仿真模型分析证实，探路者算法可以较快地实现拓扑匹配，平均在一分钟时间内，节点就可以找到5个以上的物理邻居，而通报/退避机制能实现将主干网上的数据流量减少90%～94%，实现了最初的设计目的。

本文的优化目前只能将数据传输控制在一个城域网的范围内，如何进一步缩小范围，比如将数据传输控制在局域网范围内，只用往返时延作为选择物理邻居的依据不足以满足要求，必须结合其他指标进行。另外，如何改进物理邻居组织，以扩大通报影响范围也需要进一步地思考论证。

本文的研究工作得到电子科技大学校青年基金重点项目(L0801060JX0808)的资助，在此表示感谢。

参　考　文　献

[1] COHEN B. Incentives build robustness in BitTorrent [C]//Proceedings of 1st Workshop on Economics of Peer-to-Peer Systems. [S.l.]: [s.n.], 2003.

[2] ZHANG Xin-yan, LIU Jiang-chuan, LI Bo, et al. CoolStreaming/DONet: a data-driven overlay network for efficient live media streaming[C]//IEEE INFOCOM 2005. New York: IEEE Press, 2005: 2102-2111.

[3] CASTRO M, DRUSCHEL P. SplitStream: high-bandwidth content distribution in a cooperative environmen[C]// Proceedings of 2nd International Workshop on Peer-to-Peer Systems. [S.l.]: [s.n.], 2003: 292-303.

[4] TRAN D, HUA K, SHEU S. Zigzag: an efficient Peer-to-Peer scheme for media streaming[C]//Proceedings of IEEE INFOCOM. San Francisco: [s.n.], 2003: 1283-1292.

[5] LI J. Peer Streaming: a practical receiver-driven Peer-to-Peer media streaming system[R]. MSR-TR-2004-101, Microsoft Research, 2004.

[6] 谢勇均, 闫涛, 郑婕, 等. Tracker中一种具有拓扑意识的结点选择算法(TAPS)[J]. 微电子学与计算机, 2007, 24(1): 34-37.
XIE Yong-jun, YAN Tao, ZHENG Jie, et al. A topology awareness peer selection algorithm in tracker (TAPS)[J]. Microeletronics & Computer, 2007, 24(1): 34-37.

[7] LIU Yun-hao, LIU Xiao-mei, XIAO Li, et al. Location-aware topology matching in P2P systems[C]//Proceedings of IEEE INFOCOM 2004. [S. l.]: IEEE Press, 2004.

[8] 邱同庆, 陈贵海. 一种令P2P覆盖网络拓扑相关的通用方法[J]. 软件学报, 2007, 18(2): 381-390.
QIU Tong-qing, CHEN Gui-hai. A generic approach to making P2P overlay network topology-aware[J]. Journal of Software, 2007, 18(2): 381-390.

[9] REN S, GUO L, JIANG S, et al. SAT-Match: a self-adaptive topology matching method to achieve low lookup latency in structured P2P overlay networks[C]//Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS04). [S.l.]: [s.n.], 2004.

[10] QIU Tong-qing, CHEN Gui-hai, CHAN E, et al. Towards location-aware topology in both unstructured and structured P2P in structured P2P overlays[C]//IEEE ICPP 2007. [S.l.]: [s.n.], 2007.

[11] HOONG P K, MATSUO H. Push-pull two-layer super-peer based P2P live media streaming[J]. Journal of Applied Sciences, 2008, 8(4): 585-593.

[12] 杨寿保, 许通, 胡云. 用户需求适应的P2P超级节点选取机制[J]. 电子科技大学学报, 2009, 38(3): 385-388.
YANG Shou-bao, XU Tong, HU Yun. User-demand adaptive P2P super node selection strategy[J]. Journal of University of Electronic Science and Technology of China, 2009, 38(3): 385-388.

[13] BARAB´ASI A L. Linked: the new science of net-works[M]. [S.l.]: Perseus Publishing, 2002.

编　辑　税　红