VPLS 中具有时延约束机制的组播问题研究*)

董喜明　余少华
(华中科技大学计算机科学与技术学院　武汉 430074)
(武汉邮电科学研究院　武汉 430074)

摘　要　VPLS 作为一种革新的技术受到了广泛的关注和认可。但是,在用 VPLS 承载数据业务的时候还面临着一个复杂的难题:组播问题。传统的组播问题是具有 NPC 复杂度的 Steiner 问题。本文试图从应用和实现的角度出发,建立具有时延约束机制的组播转发机制。以建立最小时延树和最小开销树作为初始条件,运用循环迭代算法,求解满足时延约束的最小开销树。算法的复杂性为 $O(n^2)$。作为补充,还提出了组播树的剪枝机制。试验结果表明,文中的算法简单可行,易于实现,适合应用于 VPLS 网络中。

关键词　VPLS,组播,最小开销树,最小时延树,时延约束,开销

Study on the Delay-constrained Multicast Issue in VPLS Networks

DONG Xi-Ming　YU Shao-Hua
(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan430074)
(Wuhan Research Institute of Post and Telecommunications, Wuhan430074)

Abstract　VPLS has gained world-wide recognition in recent years. However, deploying VPLS in Metro is confronted with one complicated issue: the multicast problem. Instead of solving the NP-Complete Steiner tree problem, this paper emphasizes more on the experiential and implemental aspect of multicast in VPLS network. It begins with the construction of least cost tree(LCT) and least delay tree(LDT), then an iterative algorithm is proposed to construct all the delay-constrained candidate trees, the one with least cost is chosen. To avoid unwanted traffic sent to PE devices, a pruning mechanism is also suggested. Compared with Steiner solutions, the algorithm is more suitable for implementation as the time complexity is only $O(n^2)$. Simulation result shows that the algorithm is feasible and suitable in building delay-constrained multicast trees over VPLS domain.

Keywords　VPLS, Multicast, LCT, LDT, Delay-constrained, Cost

1　引言

VPLS(虚拟专用局域网业务)是一种通过 IP/MPLS 网络提供虚拟专用以太网桥接域的技术。其原理是在各个 PE 之间建立全网状的 MPLS LSP,将二层以太网帧进行 MPLS 封装,通过 MPLS 交换将用户以太网流量在各个 PE 之间进行转发,从而建立一个点对多点的以太网 VPN[4]。在多媒体业务日益普及的今天,用 VPLS 承载多媒体组播业务成为一种发展趋势,如:用 VPLS 开展视频电话会议,承载 IPTV 业务等等。因此,对时延的控制就成为一个重要的研究课题。本文通过迭代算法,求解 VPLS 域中具有时延约束机制的组播问题。

2　VPLS 网络中的组播时延

同 IP 组播相比,VPLS 中的组播有其自身的特点。网络中共有两种类型的结点:PE 结点和 P 结点。PE 结点需要进行更多的处理,如组播包的复制等等。P 结点则不同,它只根据 MPLS 的标签进行流量的转发。

2.1　VPLS 中的组播

尽管 VPLS 中的发送结点和接收结点在物理上是分离的,但逻辑上仍是一个基于 MAC 转发的逻辑网络[13]。其转发路径的建立主要通过 MPLS 的隧道来完成[5,13]。为了把组播流量传送到不同的 PE 结点,关于 VPLS 的 IETF 草案推荐采用洪泛的方式[6],或将洪泛树作为缺省的组播传送方式[7]。采用洪泛树的优点可以归结为:(1)大大地降低了网络中为了维护组播树而造成的巨大开销;(2)组播树的建立时间很短,能满足实时组播业务的需要。但是,这种处理方式也带来了一个主要的问题,由于流量发送到了并没有接受请求的结点,造成了网络资源的严重浪费[7]。在实现洪泛树的时候,通常只是简单的计算最小生成树,以节省网络中的资源开销。这种方法缺乏时延控制机制。为了更好地实现实时组播流量转发,本文的实现方案采用两个步骤来完成。首先,建立具有时延约束机制的洪泛树,任何一个 PE 结点均可以是洪泛树的根结点。然后,当洪泛树产生的网络资源浪费较大,超过用户设定的参数时,采用剪枝机制建立新的组播传送树。这种方法可以保证网络中的资源浪费被控制在一定的范围内,是一种优化的 VPLS 组播流量传送方法。

2.2　VPLS 网络中的组播时延

从源到目的地,组播时延由三部分组成,Δ_1 是从源到相连的 PE 结点的时延,Δ 是在服务提供商网络中的时延,Δ_2 是

*)本文得到国家"863"项目新型城域网关键技术及以太网交换机核心芯片开发(项目编号分别为 2003AA121110 和 2003AA1Z1180)资助。
董喜明　博士研究生,主要研究方向为城域网。**余少华**　博士,教授,博士生导师。

从 PE 到接收目的地的时延。总的时延可表示为：$\Delta_A=\Delta_1+\Delta+\Delta_2$，其中，时延 Δ_1，Δ_2 是在接入的用户网络中产生的，由用户控制。难控制的时延部分为通过服务提供商网络产生的时延，是本文的研究对象。为了讨论问题的方便，假设将组播源上移至与其相连的 PE 结点，且任何一个 PE 结点均可以成为组播源。所有的发送源共享同一棵洪泛树，以进行最初的组播流量的转发。

2.3 网络模型及问题的定义

一个 VPLS 网络可以用一个无向图 $G=(V,E)$ 来表示。其中，V 是结点的集合，E 是边的集合。任何一条边均通过一个二元组 (Cw,Dw) 来描述，Cw 是对应边的开销，Dw 为对应边的时延。树 $T(s)$ 是以源结点 $s\in V$ 为根，并包含从源 s 到其它结点 $v\in(V-\{s\})$ 的路径。树 $T(s)$ 的开销可以表示为：

$$Cost(T(s))=\sum_{t\in T(s)}C(t) \tag{1}$$

如果用 $P(T(s),v)\in T(s)$ 表示树中从组播源到接收目的结点的路径，则该路径的开销为：

$$Cost(P(T(s),v))=\sum_{t\in P(T(s),v)}C(t) \tag{2}$$

路径上端到端的时延表示为：

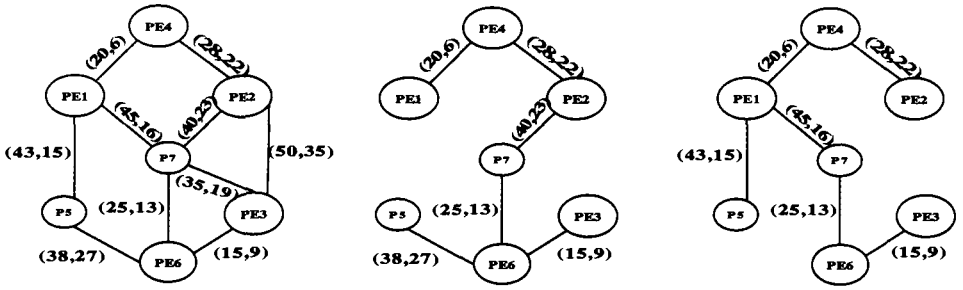
$$Delay(P(T(s),v))=\sum_{t\in P(T(s),v)}d(t) \tag{3}$$

根据(1)～(3)，研究的问题形式化为：$\forall v\in(V-\{s\})$，$Delay(P(T(s),v))\leqslant\Delta$ 且 minimal$(Cost(T(s)))$。即：对于任意的目的结点，在满足时延约束机制的同时，使组播树的开销尽可能地小。

3 VPLS 中求解具有时延约束机制的洪泛树

对 VPLS 中组播的处理分为两个步骤。本节主要讨论具有时延约束机制的洪泛树的建立过程，它是 3.2 节中剪枝机制的基础。基于消息的剪枝是在洪泛树的副本上进行的。

3.1 迭代算法



(a) 一个VPLS网络的实例　(b) 最小开销树(LCT)　(c)最小时延树(LDT)

图 1　迭代算法的初始过程

图 1(a)是一个 VPLS 网络的实际例子。Cw 和 Dw 作为权值，通过调用最小生成树的贪婪算法 prim[1,9]，求出最小开销树(LCT)和最小时延树(LDT)，见图 1(b)和图 1(c)。建立 LDT 的主要目的是检验是否存在满足时延约束机制的洪泛树。如果 LDT 不能满足时延约束，则需要重新进行时延参数的协商。否则，网络中就一定存在一棵能够满足时延约束机制的洪泛树。算法 It_Dcmt 中，行 3～4 将 LDT 作为时延约束的检查条件。行 3～7 是对 LCT 的时延特性进行测试，如果 LCT 能满足时延约束条件，则这是最佳解，此时网络的开销最小。

通常情况下，LCT 难以满足时延约束的要求。为此，我们采用迭代算法进行问题的求解。在迭代的开始，对 LCT 和 LDT 的边集进行升序排列，分别记为 E_c 和 E_d。然后求出两个边集的差值 E_s(算法 1 中行 9)。显然，E_s 中的边也是按照时延的升序方式进行排列的。

迭代法的基本思想是：以提高网络的开销为代价，将 E_s 中时延较小的边加入到 LCT 中，以期求出满足时延约束条件的洪泛树。随着迭代的进行，所有的候选树均被计算出来，但只有具有最小开销的候选树才能被选中，作为下一轮迭代的输入，对这样的候选树我们称之为迭代树。

定理 1　设 T_i 为一棵迭代树，边 $e\in E$。当 e 加入到 T_i 中时，T_i 中就会形成环路[1,9]。

证明：任何一棵迭代树均为生成树，具有 $n-1$ 条边。最开始的迭代树是 LCT。当加入一条边以后，迭代树 T_i 中就存在 n 条边，这与生成树的性质(任何一棵生成树有且只有 $n-1$ 条边)相矛盾，所以 T_i 中存在环路。

在 VPLS 域中，如果存在环路，容易在服务提供商的网络中造成广播风暴，因此环路必须被破除。算法 1 中，行 18～27 就是不断地加入时延较小的边，去掉时延较大的边，以破除树中的环路。如果生成的新树不能满足时延约束条件，则丢弃。否则，加入到候选树的集合中。在每一轮迭代的最后，满足时延约束且开销最小的树被选为迭代树，作为下一轮迭代的输入。

定理 2　设迭代算法每轮最后生成的树为 T，则一定有 $Delay(T)\leqslant Delay(T_s)$。

证明：在迭代的过程中，有两种结果产生：(1)一条时延较大的边被时延较小的边所替代，因此生成的候选树 T 满足 $Delay(T)<Delay(T_s)$。(2)生成的候选树虽然满足时延约束，但是，当 $Cost(T)>Cost(T_s)$，算法的返回值依然是 T_s，也就是说，新加入的边在该轮迭代中并不产生候选树，T_s 仍然作为返回值进入下一轮的迭代。此时，$Delay(T)=Delay(T_s)$。综合(1)(2)，定理 2 成立。

定理 3　算法 1 的返回结果是一棵满足时延约束的最小开销树。

证明：从定理 2 中可以看出，每次迭代的返回值均是此次迭代产生的满足时延约束的具有最小开销的候选树。另一方面，如果生成的树的开销大于该轮的迭代树，则返回进入下一轮迭代的仍是上一轮的迭代树。这实际上是一个边迭代边求最小值的过程。定理 3 成立。

Algorithm 1　*It_Dcmt* (G,V,s,D,Δ)

/* D 是接收结点的集合，Δ 是时延约束条件的集合，Δ_v 是结点 v 的时延约束条件，$\Delta_v\in\Delta$ */
1 $T_d\leftarrow$minimal-delay tree
2 $E_d\leftarrow$all edges in sorted order of Dw in T_d
3　if($\exists v\in D$, delay$(P(T_d(s),v))>\Delta_v$) then
4　　return FAIL /* 不存在满足时延约束条件的洪泛树 */

```
5   T_c← minimal-cost tree
6 if (∀v∈D,delay(P(T_c(s),v))<Δv)then
7       returnT_c/*LCT满足时延约束,这是最佳解*/
8 E_c← all edges in sorted order of Dw in T_c
9 E_s=E_d−E_c
10 T_s← T_c/* T_c(LCT)作为迭代的初始值 */
11 for each edge e∈E_s in ascending order of Dw /* 循环迭代 */
12   T'← Iteration(T_s,V,Δ,e)
13   T_s← T'
14 endfor
15 return T_s/*返回迭代树作为下一轮迭代的输入*/

   Iteration(T_s, V,Δ,e)
16 t_r←{}
17 T_s' ← add e to T_s
18 for any edge k in the loop in T_s'
19   if d(k)>d(e)/* 时延较大的边被替换*/
20       T_s''← remove k from T_s'
21       if(∃v∈D,d(P(T_s''(s),v))>Δ_v) then
22         continue /*新生成的树不满足时延约束条件,丢弃
*/
23       else
24         t_r=t_r∪ T_s''
25       endif
26   endif
27 endfor
28 T←the tree with the minimal cost in t_r/*取本轮迭代中开销最小
   的候选树*/
29 if (cost(T)<cost(T_s)) and (T≠LCT) /* if t_r=NULL,then
   cost (T)=+∞ */
30 return T
31 else
32   return T_s
33 endif
```

以图1中的拓扑为例，假设组播源位于PE4之后，时延约束条件由 $\Delta=\{30,30,70,30,70,70\}$ 给定，求得 $E_s=\{(43,15),(45,16)\}$。首轮迭代中，LCT作为迭代树，运行算法1后的结果如图2所示。

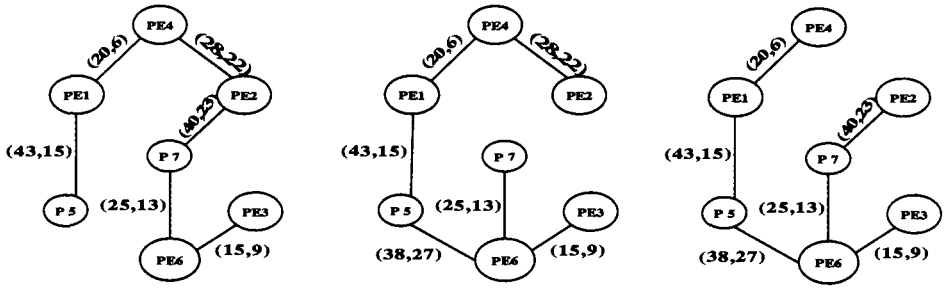


图2 迭代过程

首轮迭代中，共产生3棵树，(c)因为无法满足时延约束条件，故删除。(a)和(b)满足约束条件，成为候选树。但(b)的开销更小，因此选作迭代树，进入下一轮迭代。第二轮迭代产生的树和LDT相同，由于LDT的开销较大，所以(b)作为洪泛树返回，算法结束。

3.2 剪枝[8,13]

同具有NP-Complete复杂度的Steiner问题相比，算法1在实现上要简单得多。但是，这种解决办法也有自身的不足。在VPLS网络中，虽然将组播流量发送到所有的PE结点是一种可行的解决方案，但这种做法可能将流量发送到并不需要接受该组播流量的结点处，造成网络资源的浪费。为了量化这种方式所造成的资源浪费，我们引入式(4)：

$$\delta(t,g)=\frac{C(T_{ter})-C(tp(g))}{C(tp(g))} \tag{4}$$

式(4)中 T_{ter} 是算法1求得的洪泛树，$t_p(g)$ 是完全匹配的组播树[14]，$\delta(t,g)$ 用以计算造成网络资源浪费的百分比。很显然，采用洪泛树大大地降低了由于管理组播转发状态而带来的巨大开销，但网络资源的浪费也必须进行抑制。为了弥补这种算法所带来的不足，可以采取一种折中的处理方案，即：预定义一个门限值 bth，当 $\delta(t,g)>bth$ 时，就触发基于消息的剪枝机制生成新的组播树。剪枝是在洪泛树的副本上进行的，洪泛树依然保存在各个PE结点中作为满足式(4)条件下的组播转发树。剪枝后，只有发出接受请求的结点才能收到组播流量。

定理4 剪枝生成的新组播树依然满足时延约束机制。

证明：算法1中的迭代保证到树中的任何结点都满足时延约束，剪枝生成的树是洪泛树的子树，显然定理4成立。

3.3 动态加入和离开

当VPLS域中的某个成员动态地发出组播组加入请求时，必须进行两方面的工作，(1)采用算法1求得的洪泛树开始组播流量的转发，以减少请求建立时间；(2)计算 $\delta(t,g)$，如果 $\delta(t,g)>bth$ 成立，则触发剪枝机制，计算新的组播树。经过一个合理的时间段以后，组播流量的发送切换到新建立的组播树上[8]。当一个组播组离开的时候，必须检查组播转发所采用的组播树，如果是洪泛树，则停止组播源继续发送数据；否则，在停止组播源发送数据的同时，还必须删除对应的组播树，以避免系统由于维护大量的组播树而带来的巨大开销。

4 实验仿真

4.1 时间复杂度分析

算法1中，行1和行5调用了prim贪婪算法来生成LCT和LDT，时间复杂度为 $O(n^2+n^2)$。行11－14是一个循环迭代过程，在最好的情况下，LCT和LDT的边集合完全相同，时间复杂度为0。最坏的情况下，二者的边集合完全不同，将进行 $n-1$ 次迭代。这种情况极少见。因此迭代过程的时间复杂度为 $O(n)$。基于消息的剪枝机制是一个线性的过程，时间复杂度可以忽略不计。由此可见，文中算法的时间复杂度为 $O(n^2+n^2+n)\approx O(n^2)$，比求解Steiner树[12]的时间复杂度要小得多。表1是本文中的算法与Steiner算法的对照表。

表1 同Steiner树的比较

	It_Dcmt	Steiner树
时间复杂度	$O(n^2)$	NP-complete
开销处理	较少	很大
维护的状态数	少	很多
洪泛开销	存在	不存在
组的加入和离开	剪枝	重计算组播树
实现难易度	易	难
组播建立时间	短	长

(下转第85页)

的支持，ICE作为一种综合的解决方案将有着非常广阔的应用前景。

参考文献

1 Rosenberg J. Interactive Connectivity Establishment (ICE): A Methodology for Nettwork Address Translator(NAT) Traversal for the Session Initiation Protocol(SIP). draft-rosenberg-sipping-ice-01(work in progress), July 2003

2 Rosenberg J, Schulzrinne H, Camarillo G, et al. SIP: Session Initiation Protocol. RFC 3261, June 2002

3 Rosenberg J, Schulzrinne H. An Extension to the H. 323 for Symmetric Response Routing, RFC 3581, August 2003

4 Rosenberg J, Weinberger J, Huitema C, Mahy R. STUN - Simple Traversal of User Datagram Protocol(UDP) Through Network Address Translators(NATs). RFC 3489, March 2003

5 Rosenberg J. Traversal Using Relay NAT(TURN), draft-rosenberg-midcom-turn-02(work in progress), October 2003

6 Rosenberg J. Examples of Network Address Translation(NAT) and Firewall Traversal for the Session Initiation Protocol(SIP), draft-rosenberg-sipping-nat-scenarios-01

7 刘杨，姜琳颖，王中. 基于ICE方式的SIPNAT解决方案研究. 见：中科院第八届计算机科学与技术学术 研讨会论文集[C]，2004

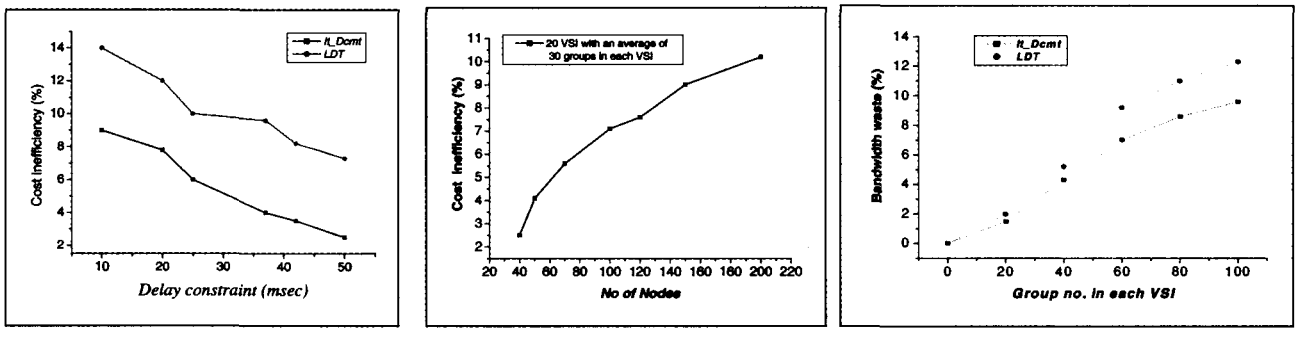
(上接第27页)

4.2 实验仿真

为了评估算法1的性能，VPLS网络采用Waxman所建议的拓扑随即生成模型[10,11]。确定两结点(u,v)间是否存在链路的概率函数由 $P(u,v)=\beta\exp(\frac{-d(u,v)}{aL})$ 定义，$d(u,v)$为结点u和v之间的距离。我们评估三种类型的VPLS网络，其结点数分别为50，100和200。同时为了评估采用了时延约束机制后给网络带来的额外开销，我们还定义了开销失效度(Cost inefficiency)：

$$inefficiency=\frac{Cost(T_{iter})-Cost(T_{LCT})}{Cost(T_{LCT})} \quad (5)$$

其中，$Cost(T_{LCT})$是用最小生成树作为洪泛树的开销，$Cost(T_{iter})$是算法1生成的洪泛树的开销。该式用作评估此条件下网络开销增加的百分比。时延约束参数则是由$\Delta v=d(P(s,v))\times\delta, v\in V$定义。我们发现，时延约束机制越严格，开销失效度就增大，意味着网络中将需要分配更多的资源来满足时延的约束。但是，算法1所生成的洪泛树造成的开销失效度要优于直接使用LDT，见图3(a)。另外，开销失效度还与网络的规模有一定的关系，在50个结点的网络中为2.7%，100个结点的网络中为4.5%，150个结点的网络中则达到了6.8%，见图3(b)。图3(c)则是为了证实剪枝机制的必要性。在50个结点、20个VPLS虚拟转发实例的网络中，设定bth为10%。图中表明，用算法1生成的洪泛树要优于直接使用LDT。实验过程中，剪枝机制并没有触发。因此，各个结点中维护的组播转发状态很少，因而开销也大大地减少。可以肯定，通过洪泛树和剪枝机制的结合，可以在城域以太网中取得较好的性能。



(a) (b) (c)

图3 实验

结束语 VPLS域中的组播是近期的研究热点。本文通过迭代法生成具有时延约束机制的组播洪泛树，并结合基于消息的剪枝机制完成VPLS域中组播的处理。这种处理机制充分的结合了VPLS网络的特点，从实际应用和实现的角度出发，建立具有时延约束机制的组播转发机制。同时为了抑制组播带来的网络开销，提出了一种较好的折中处理机制。文中算法的时间复杂度为$O(n^2)$，因此易于实现，适合城域以太网中VPLS的组播处理。

参考文献

1 Cormen T H, et al. Introduction to Algorithms. MIT Press, Cambridge, MA: 1990

2 Kabada B K, Jale J M. Routing to multiple destinations in computer networks. IEEE trans. Commun, 1983. 31～3

3 Estrin, D, et al. Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification, RFC 2362. June 1998

4 Aggarwal R, Morin T, Fang L. Multicast in BGP/MPLS VPNs and VPLS. work in progress, draft-raggarwa-l3vpn-mvpn-vpls-mcast-01. txt, 2004

5 Lasserre M, Kompella V. Virtual Private LAN Services over MPLS. work in progress, draft-ietf-l2vpn-vpls-ldp-06. txt, 2005

6 Serbest Y, Qiu R, Hemige V, Nath R. Supporting IP Multicast over VPLS. work in progress, draft-serbest-l2vpn-vpls-mcast-01. txt, 2004

7 Sajassi A, Salama H. VPLS based on IP Multicast. work in progress, draft-sajassi-mvpls-00. txt. 2002

8 Williamson B. Developing IP Multicast Networks, Volume I. Cisco press, 2000

9 Knuth D E. The Art of Computer Programming, Vol 3. London: Addison-Wesley Publishing Company, 1973

10 Waxman B M. Routing of multipoint connections. IEEE Journal of Selected Area in Communications, 1998. 1617～1622

11 Salama H F, Reeves D S, Viniotis Y. Evaluation of multicast routing algorithms for real-time communication on high-speed networks. IEEE JSAC, 1997, 15-3: 332～345

12 Kuipers F, Van M P. MAMCRA: A constrained-based multicast routing algorithm. Computer communications 25, 2002

13 Dong Ximing, Yu Shaohua . VPLS: An Effective Technology for Building Scalable Transparent LAN Services. In: Proceedings of SPIE, Network architectures, management, and applications II. 2004. 137～147

14 Fei Aiguo, Cui Junhong, Gerla M, Faloutsos M. Aggregated Multicast: an Approach to Reduce Multicast State. In: Proc. IEEE Global Internet (GLOBECOM'01), 2001