第40卷第10期 计算机研究与发展 Vol.40,No.10
2003年10月 JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT Oct. 2003

WDM全光网络中实时组播的分布式路由与波长分配算法

黄传河 陈莘萌 贾小华
(武汉大学计算机学院 武汉 430072)
(hwanghe@public.wh.hb.cn)

摘 要 在WDM网络中,由于每条链路上可用波长是动态变化的,在考虑波长转换延迟的条件下,实现实时组播连接的路由与波长分配是十分困难的.假定WDM网络中每条链路有多根光纤,只有部分结点具有波长转换器且波长转换时间是不可忽略的,据此提出了一种用于建立实时组播连接的分布式路由与波长分配算法.该算法以Prim最小生成树算法为基础,生成一棵满足给定延迟时限的最小成本树.当最小成本树不能包括所有目的结点时,对剩余目的结点生成一棵最短延迟树,然后合并两棵树得到一棵组播树.波长分配使用最少波长转换和负载平衡策略.

关键词 WDM网络;路由与波长分配;组播路由;延迟限制路由

中图法分类号 TP393.01

A Distributed Routing and Wavelength Assignment Algorithm for Real-Time Multicast in WDM All-Optical Networks

HUANG Chuan-He, CHEN Xin-Meng, and JIA Xiao-Hua
(*Computer School, Wuhan University, Wuhan* 430072)

Abstract Routing and wavelength assignment for online real-time multicast connection setup is difficult due to the dynamic change of availabilities of wavelengths on links and the consideration of wavelength conversion delay in WDM networks. Assuming that each link has multiple fibres, there are wavelength converters only at part of nodes and the conversion delay is not negligible. A distributed routing and wavelength assignment algorithm for the setup of real-time multicast connections is presented based on the above assumption. The algorithm is based on Prim's MST (minimum spanning tree) algorithm. It generates a sub-minimal cost tree under a given delay bound first. If there are nodes not included in the cost tree, a delay tree is generated to include the rest nodes. The two trees are merged together. The wavelength assignment uses least-conversion and load balancing strategies.

Key words WDM networks; routing and wavelength assignment; multicast routing; delay bound routing

1 引 言

WDM(wavelength division multiplexing)将光纤的带宽分为互不重叠的并行通道,每个通道使用一个波长传输信号.WDM是充分利用光纤带宽的关键技术.WDM网络是面向连接的,在数据传输前,通信双方必须建立连接.在WDM网络中建立连接包括路由选择和波长分配两个过程,简称为RWA(routing and wavelength assignment).

组播是一种组通信机制,其发送者(源结点)将消息同时发送给一组接收者(目的结点).实时组播是一类特定的组播形式,要求在组播请求到达后尽快建立组播连接,同时在所建立的连接中从源结点到任一目的结点的延迟时间不超过给定的时限.实时组播在现代计算机网络中有广泛的应用,例如电视会议、多媒体教学、视频点播、网上拍卖等.

建立实时组播的路由就是找到一棵以源结点为

收稿日期:2002-11-20;修回日期:2003-06-27

树根、包含所有目的结点的路由树,并且从源结点(树根)到任一目的结点(树叶)的传输时间不超过给定的时限,路由树的总成本最小. 寻找这种路由树的问题是 NP-hard 问题,现已有一些启发式方法[1].

在 WDM 网络中建立实时组播连接的主要困难是:

(1) 每个网络结点只知道与其相连的链路上可用的波长,没有任何结点具有全网的拓扑结构或可用波长的信息.

(2) 只有当路由请求到达一个结点时,才知道是否需要进行波长转换,而波长转换时间是不可忽略的,可能会使所选路径超过延迟时限而无效.

(3) 延迟和成本因素是相互独立的,具有最小成本的路径可能具有很长的延迟,反之亦然.

本文提出了一个在 WDM 网络中建立实时组播连接的分布式路由与波长分配算法. 该算法首先构造一棵最小成本树连接满足延迟限制的目的结点. 对没有连入最小成本树的结点,利用构造一棵最短延迟树,使其包含所有其余目的结点. 然后,将两棵树合并成一棵统一的树. 该算法的优点是:

① 它是完全分布式的,路由与波长分配只根据本结点的信息完成;

② 在建立组播连接时尽量避免波长转换;

③ 所构造的树在满足延迟时限的条件下具有接近最优的成本.

2　问题定义

网络用无向图 $G(V,E)$ 表示,其中 $V=\{1,2,\cdots,N\}$ 为结点集,$E=\{(i,j)\}$ 是光纤链路集,链路 (i,j) 可能包括多条光纤,编号为 $1,2,\cdots,k$. 每条链路 (i,j) 有 3 个参数:

① $\sigma_{ij}^k \subseteq \{1,2,\cdots,W\}$ 表示链路 (i,j) 的光纤 k 上当前可用波长的集合;

② c_{ij} 表示使用链路 (i,j) 的成本;

③ d_{ij} 表示链路 (i,j) 的延迟时间.

链路 (i,j) 上的可用波长是动态变化的,只有与此链路相连的两个结点确切知道当前 σ_{ij}^k 的值. 在建立连接时,为该连接的每一链路分配一根光纤及该光纤上一个当前未被使用的波长. 被分配的波长一直被占用,直到通信结束连接被终止.

网络中结点具有光纤交换功能,即从一根光纤上输入的光信号可以从输出链路的任一光纤上输出. 只有部分结点具有波长转换器. 同时假定每个波长转换器可以将任一波长转换为任一其他波长. $R[i]=1$ 表示结点 i 有波长转换器.

在一条路径上的通信延迟包括链路延迟和波长转换时间两部分. 链路延迟 d_{ij} 表示信号从结点 i 经链路 (i,j) 到达结点 j 所需的时间. 波长转换时间 $d_i^c(\lambda_x,\lambda_y)$ 表示结点 i 将波长 λ_x(输入波长)转换为波长 λ_y(输出波长)所需的时间. 假定所有波长转换器完成任何两个波长之间的转换所需时间是相同的. 如果没有进行波长转换,即 $\lambda_x=\lambda_y$,则 $d_i^c(\lambda_x,\lambda_x)=0$. 如果结点 i 没有波长转换器,且 $\lambda_x \neq \lambda_y$,则 $d_i^c(\lambda_x,\lambda_x)=\infty$.

考虑建立组播连接的实时请求 $R=(s,D,\Delta)$,其中 s 是源结点,D 是目的结点集合,Δ 是延迟时限值. 组播连接是一棵树 T,T 的总成本定义为

$$COST(T)=\sum_{(i,j)\in T} c_{ij}. \tag{1}$$

令 $P(u,v)$ 表示树 T 中从结点 u 到结点 v 的路径,则从树中结点 u 到 v 的延迟 $DELAY(u,v)$ 定义为

$$DELAY(u,v)=\sum_{(i,j)\in P(u,v)} d_{ij}+\sum_{i\in P(u,v)} d_i^c(\lambda_x,\lambda_y). \tag{2}$$

从树根 s 到任意结点 v 的延迟记为 $DELAY(s,v)$,树 T 的延迟定义为

$$DELAY(T)=\max\{DELAY(s,d),\forall d\in D\}. \tag{3}$$

延迟时限条件可以表述为

$$DELAY(T)\leqslant\Delta. \tag{4}$$

本文所考虑的问题是设计一个分布式的路由和波长分配算法来构造一棵波长路由树,使得该树在满足式(4)定义的条件下成本尽可能小.

3　相关研究成果

最小成本树是 Steiner 树,加上延迟限制后,寻找带延迟限制的 Steiner 树是 NP-hard 问题[1]. 关于构造最优组播树的研究已有一些结果[1,2],但这些结果并不能直接用于 WDM 网络.

目前关于 WDM 网络的路由与波长分配的研究主要是针对点到点通信的. 如 Spath 提出了几种路由策略[3],Jia 等人提出了一种针对静态请求、旨在使波长转换最少的 k-cut 方法[4].

在 WDM 网络中,路由算法应以链路上当前可用波长为基础. Sahasrabuddhe 等人提出的光树概

念[5],Li[6],Pankaj[7]等人分别提出的计算需要最少波长数的路由树的方法,Jia 等人提出的一种为一组静态组播请求分配路由的算法[8],都为组播路由与波长分配问题提供了有益的借鉴.

但所有这些关于 WDM 网络的研究都是将路由(构造树)与波长分配作为独立的过程对待,或者假定网络状态、参数是静态不变的或具有特定的拓扑结构. 对一般的 WDM 网络上的组播它们并不适用.

4 分布式算法

4.1 算法基本思想

本文所提出的分布式算法由两部分组成:GenCtree 和 GenDtree. GenCtree 以 Prim 的 MST 算法为基础,构造一棵最小成本树 Ctree. 其工作过程是:每次选取一个离树最近(按成本计算)的结点,如果其延迟满足约束条件,则将其加入到树中. 如果 GenCtree 能将所有目的结点加入到 Ctree 中,则算法终止. 否则,调用过程 GenDtree 将那些不能满足延迟约束的目的结点构成一棵最短延迟树 Dtree. GenDtree 以延迟为指标,不考虑成本,采用并行搜索方法建立最短延迟树.

构造 Ctree 和 Dtree 后,将其合并为一棵树. 在合并时可能会产生回路. 因此必须消除 Ctree 中的某些边以保证树的性质,同时满足延迟条件.

4.2 数据结构

路由表 CRoutab 和 DRoutab:每个结点都保存有成本路由表 CRoutab 和延迟路由表 DRoutab. 路由表中的项 CRoutab[d]/DRoutab[d]表示到达目的结点 d 的最小成本/最小延迟及其可能的输出链路,其中第 1 个链路为主链路,其余链路为候选链路.路由表可以使用距离向量路由算法计算得到.

目的结点到树的距离:在构造树时,为每个目的结点记录一个三元组〈*treenode*, *dest*, *dist*〉,用于跟踪每个目的结点到已经构造的树的最短距离. 〈*treenode*, *dest*, *dist*〉表示到目的结点 *dest* 的最短距离是 *dist*,它通过树上结点 *treenode* 连接到树中.

可用波长链路数 $NL(\lambda)$:每个结点记录与其关联的链路上的可用波长. $NL(\lambda)$表示与本结点相关联的链路中波长 λ 可用的链路总数,每根光纤计算为 1.

4.3 最小成本树 Ctree 的构造

(1) Ctree 的构造算法 GenCtree

① 源结点 s 启动 GenCtree,将 s 加入到 Ctree 中.

② s 利用 CRoutab 选择一个最近的目的结点,选择通往该目的结点、具有可用波长的一条输出链路. 如果主链路没有可用波长就选择一条候选链路. 然后通过所选择的链路向下一个结点发送一个 CFIND 消息.

③ 下一结点收到 CFIND 消息后,使用相同的方法选取通往选定的目的结点的输出链路,直到到达目的结点.

④ CFIND 消息在传递过程中收集从所构造的部分树到其他目的结点的最短路径的信息. CFIND 消息所到达的目的结点通过所收集的三元组信息负责选取下一个加入到树 Ctree 中的目的结点 *dest* 及其相应的在树上的连入结点 *treenode*,然后向该树结点发送一个消息,由该树结点启动路径〈*treenode*, *dest*〉的建立过程.

⑤ 每当 CFIND 消息到达一个目的结点,该目的结点就负责选择下一个连入树 Ctree 中的目的结点. 如果一个目的结点因为违反延迟条件而不能加入到树中,则首次发现这一条件的结点负责选择一个新的目的结点,并通知其进行相应的操作.

⑥ 重复上述过程直到没有目的结点可以加入到树中,这时,调用 GenDtree 过程.

GenCtree 用于选择目的结点的标准是最小成本. 树 Ctree 与目的结点 d 之间的最小成本距离定义为

$$DIST(Ctree,d) = \min\{COST(t,d), \forall t \in Ctree,\ d \in D\text{-}Ctree\}. \quad (5)$$

设所选择的输出链路为(v,w),则从源结点 s 经 v 到达 v 的邻结点 w 的延迟 $DELAY(s,w)$定义为

$$DELAY(s,w) = DELAY(s,v) + d_{vw} + d_v^c(\lambda_v,\lambda_w), \quad (6)$$

该延迟在结点 v 进行计算并测试.

如果试探失败,发现失败的结点负责通知源结点,然后另外选取一个目的结点开始其连入树的操作,同时释放为失败的路径所预留的波长.

(2) GenCtree 的波长分配策略

选择输出链路的结点负责进行分配波长. 波长分配策略为:

① 输出链路任一光纤上与输入链路上相同的波长可用时,优先分配该光纤及波长;

② 如果该波长在所有光纤上不可用,则选择具有最大 $NL(\lambda)$值的波长及具有该波长的第 1 根光纤.

4.4 最短延迟树 Dtree 的构造

(1) Dtree 的构造算法 GenDtree

GenDtree 以源结点 s 为树根,将所有没有包含在 Ctree 中的目的结点连接起来构造一棵最短延迟树 Dtree. 在构造 Dtree 时,GenDtree 需考虑 3 个因素:

① 当最短延迟树存在时,GenDtree 必须能够以很高的概率构造成功;

② GenDtree 必须简单、执行速度快,以便能快速完成组播连接的建立;

③ 所构造的延迟树 Dtree 应该易于与成本树 Ctree 合并. Dtree 和 Ctree 可能有公共链路和结点,合并可能会产生回路.

GenDtree 以并行方式运行,构造算法为:

① 对剩余的每个目的结点 d,源结点 s 从 DRoutab[d]中的每个输出链路上发送 DFIND 消息.

② 在每个中间结点,采用受限扩散方式,最多选取 K 条输出链路转发 DFIND 消息,直到选定的目的结点被连入 Dtree 树中,或者是因不能满足延迟条件或无可用波长而终止.

在中间结点 v,选取到达目的结点 d 的输出链路的标准是到达 d 的最小延迟. v 选取下一个结点 w 的评价函数为

$$f(w)=\begin{cases} DELAY(v-w,d), & DELAY(s,v)+ \\ \quad DELAY(v,d)+d_v^c(\lambda_v,\lambda_w)\leqslant\Delta, \\ \infty, \quad \text{otherwise}. \end{cases} \tag{7}$$

③ 当 DFIND 消息第 2 次到达一个结点时,GenDtree 采取下述方法处理重复的消息:

· 如果本结点先前收到过 DFIND 消息,并且当前收到的 DFIND 消息与先前的 DFIND 消息来自不同的输入链路,则在两条输入链路中选择具有较长输出延迟(离开延迟)的链路从树中删除.

· 如果本结点的一个或多个子结点及其链路已经由先前的 DFIND 消息加入到树 Dtree 中,但当前的 DFIND 消息没有选择它们加入到 Dtree 中,则 GenDtree 计算到这些邻结点的延迟以决定是否将它们保留在 Dtree 中. 决定取舍的标准是新的延迟(当前 DFIND 消息计算结果)是否比原延迟小. 当 Dtree 的一条旧链路被确定删除时,该链路以后的子树一同被删除.

④ 每个成功加入到 Dtree 的目的结点向源结点发送一个 DFOUND 消息,以通知源结点当前的进展状况.

(2) GenDtree 的波长分配策略

① 如果所选择的链路已经是 Ctree 的一条链路,则使用在 Ctree 中已分配的波长;

② 如果所选择的链路已经是 Dtree 的一条链路,则可以根据需要重新分配波长;

③ 如果所选择的链路是一条新链路,则按 GenCtree 的分配策略进行分配.

4.5 Ctree 和 Dtree 的合并

Ctree 和 Dtree 建立之后,需要将它们合并为一棵单一的树. 源结点 s 沿 Ctree 和 Dtree 的所有链路发送一个 MERGE 消息开始合并操作,该消息被逐结点地进行处理,直到到达目的结点. 合并算法可简述为:

① 如果 Ctree 和 Dtree 在结点 x 相交,并且分属于 Ctree 和 Dtree 的输入链路不同,则去掉 Ctree 的输入链路.Ctree 中 x 的所有后继结点重新计算延迟时间;

② 如果沿合并的新树的路径不满足延迟条件,则算法失败终止;

③ 如果 Ctree 的一条链路被删除,则 Ctree 中的前驱链路沿通向树根的方向逐一被删除,直到到达一个目的结点或 Ctree 与 Dtree 的相交结点.

4.6 算法分析

结论 1. 所构造的树包含源结点和所有目的结点,并且满足约束条件式(4),同时具有接近最小的成本.

证明. 显然,所构造的树包含源结点和所有目的结点.

算法的每一步往树中增加一条链路及对应的对端结点,从源结点 s 到新增加结点的路径延迟不超过延迟时限. 所以从源结点到达每个目的结点的路径延迟都满足延迟约束条件式(4).

由于 GenCtree 每次增加一条最小成本的路径到树中,因此每次加入到树中的是从树到目的结点的最小成本路径. 只有 GenDtree 增加的路径不具有最小成本,根据 Prim 算法可知,总的成本接近最优.

如果初始波长或者波长分配策略选择不当,可能会导致大量的波长转换,从而可能导致路径延迟超过时限. 尽管本算法在可能时试图替换部分链路或路径,但因为它并不为已经构造的部分树重新分配波长,因此仍然有可能最终不能构造出一棵满足要求的树. 当进行 Ctree 和 Dtree 的合并时,Ctree 的一些链路可能会被删除,这可能导致合并后的树无效,而算法并没有试图为删除的目的结点寻找候选路径. 因此存在着虽然有解但找不到解的可能. 但由于算法使用的是启发式方法,即选取最少转换和最小负载的方法为链路分配波长,因此上述情况通

常可以避免. 证毕.

结论 2. 算法的通信复杂性是 $O(n)$.

证明. 在 GenCtree 的路由选择过程中,CFIND, CFOUND 等消息的总数分别不超过 $O(n)$,最坏复杂性为 $O(n)$. 在 GenDtree 中,DFIND 消息是以并行扩散方式传输的,最大消息数也是 $O(n)$. 这样消息总数为 $O(n)$,所以通信复杂性不超过 $O(n)$.

证毕.

5 模拟结果

本文在模拟时,网络规模固定为 200 个结点,网络拓扑随机生成并测试,直到生成一个连通的网络为止. 每条链路假定只有一根光纤,具有波长转换器的结点数为 50% 且随机分布. 链路成本为 1～15 之间的随机数,链路延迟为 1～10 之间的随机数,光纤链路上的波长数为 8. 除非特别情况,假定 $|D|$ 为网络规模的 20%,链路上波长的平均可用性定为 50%,$K=1$. 网络拓扑在模拟过程中保持不变.

本文模拟 3 个算法进行比较:SPT, MST 及本文提出的算法 dRWA. SPT 和 MST 用本文的方法进行计算,但不考虑延迟限制. 模拟对象为组播树成本,它们针对 4 个参数即 Δ、$|D|$、波长可用性及相对波长转换时间进行模拟;组播连接建立时间,针对 $|D|$;相对波长转换数,针对波长可用性.

图 1 说明了组播树成本与延迟时限 Δ 之间的关系.SPT 和 MST 的曲线为常数,因为二者都不受 Δ 影响. 本文算法的曲线介于二者之间,其高端接近于 SPT 曲线,低端接近于 MST 曲线. 当 Δ 越小,越多的 MST 路径违反 Δ 限制而用 SPT 路径代替. 使得构造的树更宽,组播树成本就更高. 随着 Δ 的增加,更多的目的结点通过 MST 路径连入树中,导致树的成本下降. 当 Δ 足够大时,不会影响路由选择,最终的组播树变成 MST.

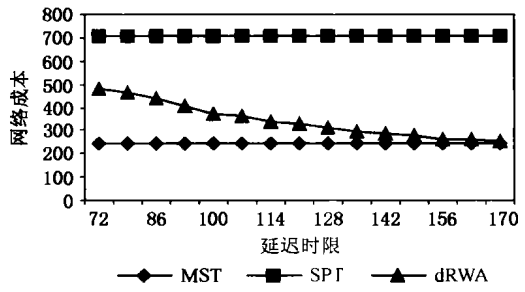

图 1 不同 Δ 时的成本

图 2 说明树的成本与目的结点集大小间的关系.当目的结点增加时,组播树包括更多的目的结点,导致树的成本增加.SPT 曲线在其他两个曲线之上并上升得更快. 这是因为 SPT 不考虑路径共享. 本文算法的性能接近 MST,二者的曲线随目的结点集的增大而增加得非常缓慢,因为目的结点集大,共享路径的可能性就大.

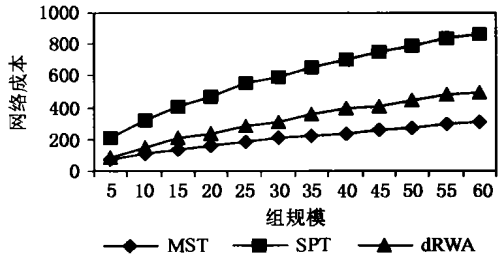

图 2 不同 $|D|$ 时的成本

图 3 说明树的成本与波长可用性之间的关系. 当波长可用性增大时,每条链路上有更多的可用波长,在构造组播树时有更多的链路可供选用,进行波长转换的机会也会更少,从而导致组播树的成本下降. 模拟结果显示,当每条链路上波长可用性大约大于 30% 时(平均有 2.4 个可用波长),组播树的成本受波长可用性的影响变得较小. 也就是说,当网络的负载不超过 70% 时,路由选择与波长分配受负载影响较小.

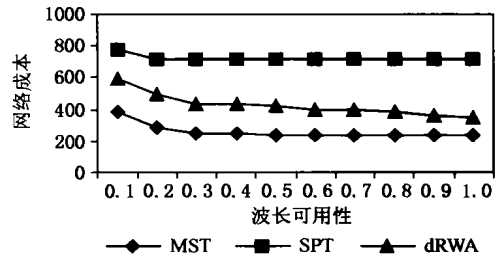

图 3 不同波长可用性的成本

图 4 说明树的成本与相对转换延迟之间的关系.相对转换延迟是指波长转换时间与全网的平均链路延迟的比值. 结果显示,用 SPT 及本文算法构造树的成本随相对转换延迟的增加而非常缓慢地增加,MST 树的成本是常数. 本文算法的曲线几乎与

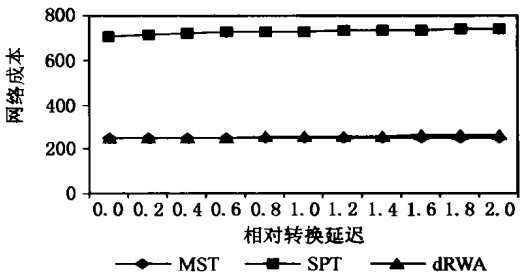

图 4 不同转换延迟的成本

MST 曲线重合,说明当延迟时限给定之后,转换延迟时间主要影响建立连接的成功率,而对组播树成本的影响较小. 增大相对转换延迟的效果类似于减小延迟时限 Δ.

图 5 说明建立连接时间与目的结点集大小之间的关系. 建立时间是指最长路径上各链路的链路延迟时间的总和,没有计算波长转换时间. 当 $|D|$ 增大时,建立时间随之增大. 因为 D 是随机产生的,可能分布在全网范围,因此建立时间的增长速度远小于 $|D|$ 的增长速度.

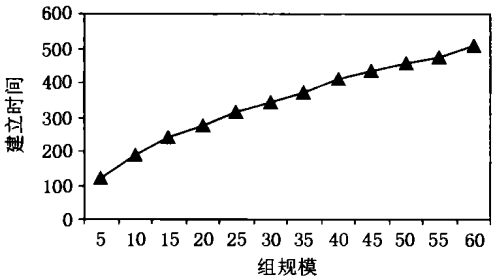

图 5　不同 $|D|$ 的建立时间

图 6 说明相对转换次数与波长可用性之间的关系. 相对转换次数是指总的波长转换次数与树中链路总数(即所用波长的总数)的比值,也就是平均链路(即波长)的转换次数. 当波长可用性增大时,每条链路上的可用波长增加,波长转换就会减少.结果显示,本文算法所进行的波长转换比 MST 少,但比 SPT 多. 当波长可用性很低时,需要进行波长转换的概率非常高,波长转换的累计时间会显著增加,因此会导致建立连接失败的概率增大.

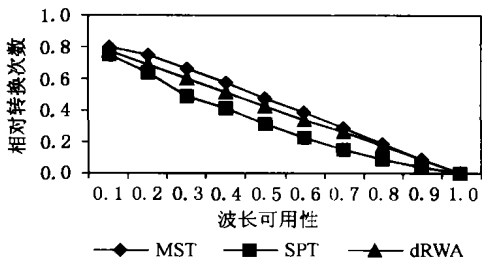

图 6　不同波长可用性的波长转换次数

6 结　论

本文提出了一种在 WDM 网络中建立实时组播连接的方法,该方法构造一棵满足延迟时限并具有接近最小成本的组播树,同时为树中的每条链路分配光纤与波长.该方法考虑了波长转换时间对路由选择与波长分配的影响,并且是完全分布式的.

对于 WDM 网络中波长转换器只有部分波长转换能力的情况本文没有考虑,对此需要做进一步的研究.

参 考 文 献

1　Bin Wang, J C Hou. Multicast routing and its QoS extension: Problems, algorithms and protocols. IEEE Network, 2000, 14 (1/2): 22～36

2　C P Low, Y J Lee. Distributed multicast routing with end-to-end delay and delay variation constraints. Computer Communications, 2000, 23(9): 848～862

3　J Spath. Dynamic routing and resource allocation in WDM transport networks. Computer Networks, 2000, 32(4): 519～538

4　X Jia, Ding-zhu Du, Xiao-dong Hu *et al*. A new wavelength assignment method for minimal wavelength conversions in WDM networks. In: Proc of the 9th ICCCN, 2000. 621～624

5　L H Sahasrabuddhe, B Mukherjee. Light trees: Optical multicasting for improved performance in wavelength-routed networks. IEEE Communications Magazine, 1999, 37(2): 67～73

6　Deying Li, Xiufeng Du, Xiaodong Hu *et al*. Minimizing number of wavelengths in multicast routing trees in WDM networks. Networks, 2000, 4: 260～265

7　R K Pankaj. Wavelength requirements for multicasting in all-optical networks. IEEE/ACM Trans on Networking, 1999, 7 (3): 414～424

8　X Jia, D Du, X Hu *et al*. Optimization of wavelength assignment for QoS multicast in WDM networks. IEEE Trans on Communications, 2001, 49(2): 341～350

黄传河　男,1963 年生,教授,博士生导师,主要研究方向为计算机网络、分布并行处理、量子计算.

陈莘萌　男,1939 年生,教授,博士生导师,主要研究方向为分布并行处理、新型计算机理论.

贾小华　男,1962 年生,教授,博士生导师,主要研究方向为计算机网络、分布式系统.