一种区分服务域内的 IP 流量规划方法*

郭国强[1,2], 张尧学[2], 王晓春[2]

[1](常德师范学院 计算机科学与技术系,湖南 常德 415003);
[2](清华大学 计算机科学与技术系,北京 100084)
E-mail: cdggq@263.net; cdggq@21cn.com
http://www.tsinghua.edu.cn

摘要: IP 流量规划与区分服务结合是增强网络服务质量保证的新方法.在分析区分服务对流量规划需求的基础上,设计了区分服务与流量规划结合的功能框架,提出了适合流量规划要求的队列调度方法和流量分割与映射的原理与方法.流分割方法基于表对数据流进行分割.该方法既能体现区分服务数据包的转发优先级的差异,又能保证数据流的延迟与同步要求.

关 键 词: 流量规划;区分服务;服务质量;负载平衡;多路径路由
中图法分类号: TP393 文献标识码: A

由于具有良好的扩展性,区分服务[1,2](differentiated service,简称 DS)成为 Internet 上多媒体应用服务质量(quality of service,简称 QoS)控制研究的重点和热点.

DS 因 PHB(per hop behavior)[3,4]类少,在具有良好扩展性的同时,也存在以下问题:因为 PHB 的确定与用户、ISP 的特殊需求等人为因素有关,标识码相同的不同数据流的流量特性可能有很大的差异,这增加了调度与资源分配的复杂度;因为数据包分类少,类聚集特征很明显,有的路径负载较重,而有的路径空闲或负载较轻,负载不均匀使链路阻塞的可能性增大,也使资源利用率降低.

IP 流量规划[5](traffic engineering,简称 TE)能增强 DS 域的 QoS 保证.TE 在多路径路由[6,7]的基础上能平衡网络负载,提高资源利用率,改善网络流量控制与资源使用性能.基于 DWDM 的骨干网光缆上存在多个并行路径和多路径路由算法构成 TE 应用条件.文献[8,9]提到了要在 DS 域进行流量规划,以改善流量性能.本文根据 TE 与 DS 的特征,找出了多类 PHB 到多输出路径(有限多对多)离散的映射关系.基于该映射关系,我们提出了 PHB 队列调度和流量分割与映射方法,该方法既能把特性有差异的多个流映射到 QoS 参数不同的多路径,又能把特性相似的数据流在多路径上分摊,解决 DS 域存在的问题.

该方法首先按分类优先级把输入缓冲区中的 PHB 安放到多个 FIFO 队列,队列中数据包通过周期性的队列合并、优先级提升动态提升优先级,这样既保证了高优先级 PHB 优先转发,又为满足低优先级 PHB 的延迟要求提供了条件.多个 PHB 队列链接成一个表,借助分类指针指示,表可以提供一个或多个输出点(如图 2 所示).多路径路由算法提供多个备选路径,路径的可用带宽是决定输出路径上能映射流量大小的权重因子.基于 PHB 队列表、输出路径数、输出路径的可用带宽,我们把数据流分割成一个或多个子数据流并映射到输出路径上.

由于对 PHB 队列的调度与数据流的分割进行了统筹考虑,本文提出的方法既保证了数据包转发的优先级,又保证了数据包的延迟与同步需求.

* 收稿日期: 2000-12-28; 修改日期: 2001-06-20
基金项目: 国家重点基础研究发展规划 973 资助项目(G1998030409);湖南省自然科学基金(00JJY2068)

作者简介: 郭国强(1964－),男,湖南常德人,副教授,主要研究领域为多媒体网络服务质量控制,流量规划;张尧学(1956－),男,湖南澧县人,博士,教授,博士生导师,主要研究领域为网络协议与互连,服务质量控制,程序挖掘;王晓春(1968－),安徽合肥人,博士生,主要研究领域为网络服务质量控制.

1 DS 域的 TE 控制功能

图 1 是把 TE 与 DS 结合的功能结构图.图 1 中的序号①~⑩分别为 QoS 代理、分类器、TE 控制中心、TE 计量器、DS 计量器、DS 标识/整形、路由算法、TE 优化器、数据库和调度器,其中③,④,⑧,⑨是 TE 部件,①,②,⑤,⑥,⑩是 DS 部件,E-OSPF[10]是多路径路由模块,它计算备选路径.

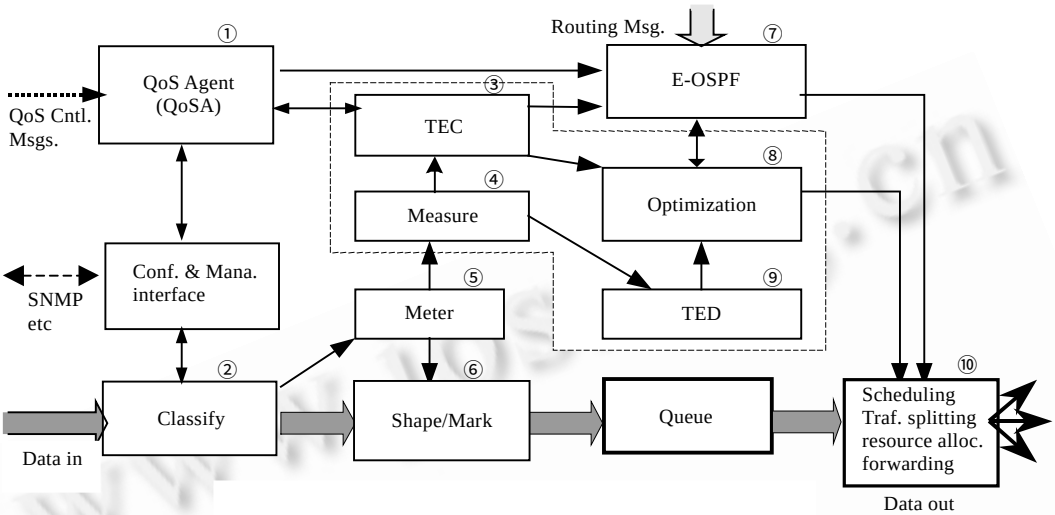


Fig.1 Function framework

图 1 功能结构图

在 DS 部件中,QoSA 是 QoS 控制的关键部件,它向 TEC,E-OSPF 及其他 DS 部件传递 QoS 控制策略及相关参数;分类器按指定的分类条件[1,2]把数据包分成有限类;计量器测算数据包的流量参数值,计量结果传递给整形、标识部件;标识部件根据分类、计量结果及标识策略把标识码写入数据包头的 TOS(type of service)中,确定数据包对应的 PHB.调度器增加了流量分割功能,其工作受 TE,E-OSPF,PHB 队列输出的影响.

在 TE 部件中,TE 控制中心 TEC 控制 TE 各部件工作,负责与非 TE 模块通信;TEC 确定路由算法工作模式(单/多路径路由)、TE 启停策略、流量优化策略;计量部件测算队列的可用空间,输出链路的可用带宽、MTU 等,其数据采集周期由 TEC 决定;TED 存放计量部件的测算结果,存放网络拓扑信息和链路特征信息(资源特性、用户配置管理(带宽需求、管理策略等)信息);优化部件根据配置与计量信息提供数据流调度、分割及映射的控制信息.

DS 结点的 Measure 与 TE 的 METER 结合提供更完备、精确的网络资源和流量信息.由于目的地址相同的 DS 域出口少,DS 域的数据包分类工作量减少,多路径选择和流量分割的计算量减小,在 DS 域实现 TE 的时间代价较低.

DS 域内流量规划原理.当收到在 DS 结点激活 TE 的信息以后,QoSA 使 TEC,E-OSPF 等同时开始工作.主要操作是:E-OSPF 按 QoSA 提供的参数计算出多条备用的路径;分类器按 DS 域出口地址分类数据包;TEC 监测流量计量结果,当满足 TE 启用条件时启动优化器开始工作.优化器根据流量参数、输出路径、来源于 TED 的优化策略控制数据包的调度和数据流的分割与映射.

TE 与 DS 结合要解决 PHB 数据包输出优先级、同步与顺序问题.PHB 优先级的问题主要由调度器解决,解决方法将考虑流量分割的要求;采用 TE 后数据包的平均延迟将明显减小,同步与顺序问题在流分割阶段通过控制同类数据包的映射路径来解决.

2 PHB 队列调度

区分服务有 3 个服务类[2],服务类由 PHB 实现.服务类是 P 服务(premium service,简称 PS)、A 服务(assured service,简称 AS),尽力转发服务(best effort,简称 BE).PS 仅由 1 个 EF[3](expedited forwarding) PHB 实现;AS 由 12

个 AF[4](assured forwarding) PHB 实现,AF PHB 分为 4 个子类,每子类 3 个优先级.BE 类数据包认为是特殊的 PHB.

调度目标是,按 PHB 到达时间和优先级顺序转发数据包,同时使低优先级的数据包能及时被转发.由于 TE 是有条件启用,要求 PHB 队列既能单路输出,又能路多路输出,并能在两种状态之间平滑地切换.

队列设置.队列分类数及队列总数影响调度器复杂度及时空开销.AF PHB 按优先级分成高、中、低 3 大类(原为 12 类),EF PHB 和 BE 各 1 类,共 5 类 PHB.每类 PHB 使用两个队列,共使用 10 个队列.由于调度器动态提升队列优先级,在优先级提升前$(t-\varepsilon)$后$(t+\varepsilon)$到达的同一类 PHB 数据包会进入不同的队列,因此,设 2 个队列(P_i , P_i^+)存放它们,又通过队列合并使它们进入同一队列,并保持先后顺序不变.

调度原理.在循环周期 Δ 内,调度器的工作分为输入、合并和提升 3 个主要过程.队列的合并、优先级提升是通过调整队列首指针、重标队列号实现,处理后的数据包存放在各类队列的 $P+$队列中,以备流分割与包转发.算法描述如下:

Input(P_i**)**; //输入数据包 P_i,分类累计数据包数量

{

input P_i; //接纳数据包

$N_i=N_i+1$ (i=1,2,…,10) //累计数据包数量

}

Merg(P_i,P_i^+**)** //合并 P_i,P_i^+

{

$P_i=P_i+P_i^+$ //p_i^+串联在 p_i 的尾部

}

Upgrade(P**)** //提升 p_i 的优先级

{

$(P_i-1)^+=P_i$ //p_i 升高一级

New(P_i**)** //新创建 P_i

}

数据包按分类优先级存放实现了服务质量的区分,优先级提升体现了数据延迟性能的改善,队列的合并还实现了数据包的同步.

图 2 是 3 类 PHB 队列的调度实例.其中 0+和 3+分别是优先级最高、最低的队列,且设每次每类有 1 个数据包从输入缓冲区进入 PHB 队列.其中①,④,⑦为包输入,②,⑤,⑧是队列合并,③,⑥,⑨是队列优先级升级.经过几个周期,0+中包含各类 PHB,这既体现了调度器对各类 PHB 的优先级差异的考虑,又体现了优先级提升对低优先级 PHB 延迟性能的考虑.图中可提供 3 个输出 0+、1+、2+,也可以只提供 1 个(多队列串联)输出.

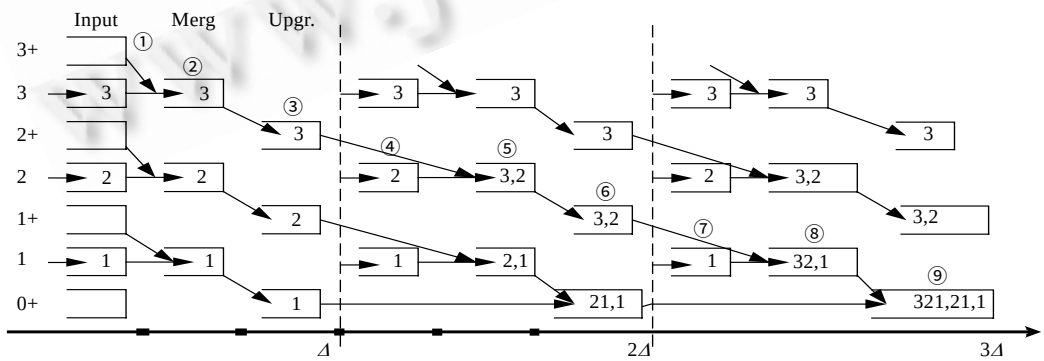


Fig.2 Example of scheduling operation

图 2 调度器工作示例图

3 数据流分割与映射

完成流量分割功能的部件称为分割器.分割器是在 TE 控制下完成多输入、多输出联接的多路开关,从多个 PHB 队列输入,向 1 个或多个端口队列输出.它应在低开销条件下,既按优化比例分割数据流,又通过分割算法保证数据包的顺序和同步要求.

每条路径的可用带宽决定路径权重.我们依据路径权重分割、映射数据流.当某个路径权重较大时,意味着路径的可用带宽大,该路径的输出端口队列便可与多个 PHB 队列建立关联,多个子数据流会聚后映射在此输出路径上.

3.1 流量分割的描述

设 $V=(v_1,\ldots,v_k)$,$W=(w_1,\ldots,w_k)$是数据流分割向量,S_i,S_j 是向 i, j 路径转发的数据流量,$S_{\max}$ 是最大数据包的体积,K 是输出路径数.

(1) 当 $w_1,\ldots,w_k$ 是流量分割系数时,$w_1+\ldots+w_k=1$ 且 $w_i\geq 0$;

(2) 当 $v_1,\ldots,v_k$ 是输出路径的可用带宽时,$S_i/S_j=v_i/v_j$ (比例分割);

因为 Internet 是包传递网络,对“流”并不是任意分割,因此有

$$\left|\frac{S_i}{v_i}-\frac{S_j}{v_j}\right|\leq\frac{S_{\max}}{\min(v_i,v_j)};$$

(3) V 与 W 的关系为 $w_i=v_i/(v_1+\ldots+v_k)$.

3.2 流分割的控制

设 q 是到达目标结点的最优路径(延迟最小),$\delta=(\delta_1,\ldots,\delta_k)$是延迟裕量向量,则

(1) $q+\delta_k$ 是第 k 条路径的延迟;

(2) $\delta_1,\ldots,\delta_k\geq 0$ 且 $\mathrm{Min}\{\delta_1,\ldots,\delta_k\}=0$.

设 D 是数据包经多条路径到达目标的最小延迟抖动,则

(1) $\mathrm{Max}\{\delta_1,\ldots,\delta_k\}\geq D$ 是分割成功的基本要求(同步要求);

(2) $\delta_1=,\ldots,=\delta_k=0$ 是最优分割(使用多条完全并行的路径)

3.3 分割与映射方法

TE 在每个循环开始时被启用,输出路径可在循环中随机减少直到停止 TE.

按使用优先级从高到低,输出端口队列为 $Q_1,\ldots,Q_k$,按转发优先级排列,PHB 分类队列是 $P_1,\ldots,P_N$.$Q_1,\ldots,Q_i,\ldots,Q_k$ 与 P 中的 $P_{1'},\ldots,P_{i'},\ldots,P_{K'},\ldots,P_N$ 关联,流分割与映射原理是

$$Q_i=P_{i'}+\ldots+P_{k'-1},\quad Q_k=P_{k'}+\ldots+P_N\ (P_{k'-1}\text{ 是 }P_{k'}\text{的前一个队列},N\text{ 是队列总数}).$$

当 $K=1$ 时,$Q_1=P_1+\ldots+P_N$,是所有 PHB 队列串联的单输出路径情形.

当 $K=3$ 时,$N=5$ 是 3 输出路径的情形,可能的分割方法是

$$Q_1=P_1;\ Q_2=P_2+P_3+P_4;\ Q_3=P_5.$$

把 $P_1,(P_2,P_3,P_4),P_5$ 视为 EF PHB,AF PHB,BE 队列时,这便构成 DS 流的一种最简单的分割方案.

3.4 其他情况说明

(1) $K>N$.当输出路径比 PHB 队列多($K>N$)时,出现一个 PHB 队列对应多个输出队列的情形.一方面,通过优化器的控制,把 PHB 队列分得更多;另一方面,设 L_i,L_j 表示 P_i 队列中相邻数据包优先级,并且 $L_i<L_j$,则 L_j 只能映射到 L_i 对应的 Q_i 或优先级更低的 $Q_j(j\geq i)$中,只有在新的周期开始时才按优先级整体调整输出路径.

(2) 断路问题.某备选路径停止使用时,相关的 PHB 队列在下一个周期开始时,重新对应输出端口或合并到其他队列.

(3) 同步、顺序问题.由于 PHB 队列构成了数据表,并且基于 PHB 标识码和输出路径权重(分割系数或链路速率)确定数据包转发路径,因此,我们的分割方法类似基于表的 HASH 分割方法[11],数据包的同步、顺序要求能

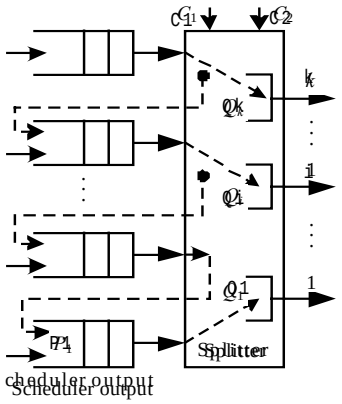


Fig.3 Example of splitter operation

图 3 分割器工作示例图

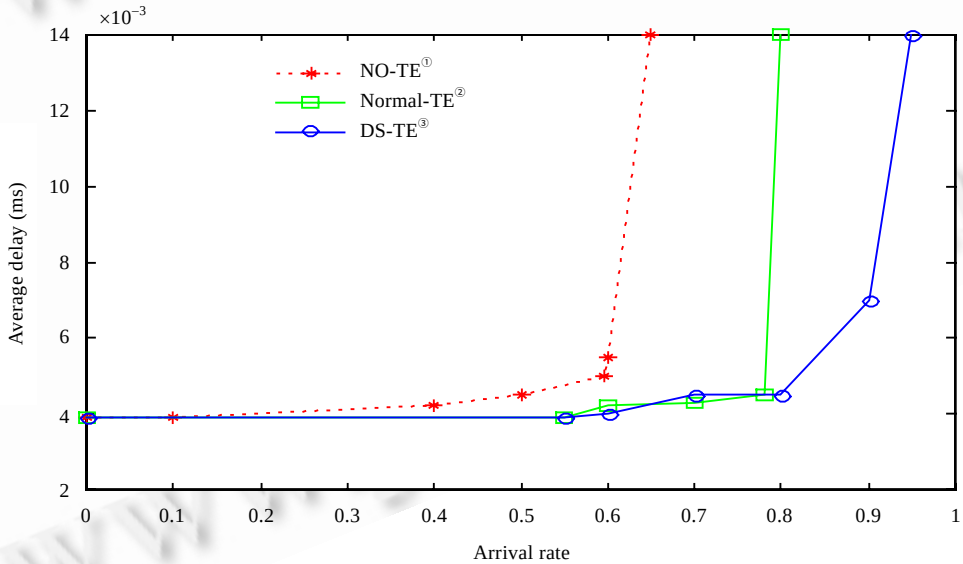
得到保证.具体地说,D 决定了 E-OSPF 选定的路径满足数据包同步要求.由于输出路径与 PHB 队列都是按优先级排列、对应,分割与映射连续流不会出现数据包现顺序颠倒的问题.当 $K>N$ 或某路径停止使用时,低优先级数据包可能要在高优先级路径上转发,但由于采取了短期(当前处理周期内,不许使用更高优先级路径)控制和周期性的整体性对应关系调整,也不会有数据包接收顺序颠倒的问题.

数据流映射可以使用硬件、软件两种实现方法.在采用硬件方法时,Q 与 P 对应关系确定输出指针初始位置,K 个部件按指针值并行转发数据包.硬件方法转发效率高,但分割精确度不高.软件方法采用轮询方法建立 PHB 数据包到输出端口队列的映射关系.

图 3 说明了 TE 在 DS 中实施流分割与映射的情况.其中,C_1,C_2 分别是图 1 中的多路径路由、优化器的信息.图中 $Q_1=P_1+P_2+\ldots$(多 PHB 队列串联),$Q_i=P_i$,$Q_k=P_k$(一类 PHB 单独使用一条输出路径).

4 模拟与分析

文献[12]肯定了多路径路由对网络流量性能及资源利用率的改善.以第 3.1 节中的分割系数(路径可用带宽)作为输出路径权重,我们分别模拟了无 TE 作用(NO-TE)、TE 应用于非 DS 环境(NORMAL-TE)、TE 应用于 DS 域(DS-TE)时网络延迟性能的变化(如图 4 所示).



①不使用 TE,②普通 TE,③DS 域 TE.

Fig.4 Comparison of the delay performance of traffic

图 4 流延迟性能比较

从图 4 可以看出,在负载较小时,几种情形的延迟几乎相同,在重负载时效果明显,说明必须有条件使用 TE;在启用 TE 以后,由于使用多条路径降低了连接的失败概率和减小了数据包在输出队列中的等待时间,数据包的延迟明显减小;由于区分服务的分类特性明确(3 大类或 5 小类)和满足要求的输出路径足够(大于或等于 3 或 5),流分割与映射的效果明显,3 种情况下网络发生阻塞时的数据包到达率分别是 0.6,0.78 和 0.9,TE 在 DS 域发挥的作用最好;发生阻塞时的数据到达率高,说明 DS-TE 的资源利用率明显高于非区分服务的情形.

5 结 语

我们研究了 IP 流量规划在区分服务域的应用,文中的队列度策略既保证了 DS 域内 PHB 数据包的转发优先级和延迟性能,又能适合数据流在单路径和多路径两种情况下的流量分割与映射.模拟结果说明了我们的方法能更好地改善网络流量性能,避免阻塞,确保网络的服务质量,提高网络资源利用率.在今后的研究中,我们将利用多路径路由仿真、测试环境,深入研究 IP 流量规划对网络流量性能控制等的影响.

References:

[1] Blake, S., *et al.* An Architecture for Differentiated Service. RFC 2475, 1998.

[2] Nichols, K., *et al.* A Two-Bit Differentiated Services Architecture for Internet. RFC 2638, 1999.

[3] Jacobson, V., *et al*. An Expedited Forwarding PHB. RFC 2598, 1999.

[4] Heinanen, J., *et al*. Assured Forwarding PHB Group. RFC 2597,1999.

[5] Daniel, O., *et al*. A Framework for Internet Traffic Engineering. IETF Draft, 2000.

[6] Thaler, D., *et al*. Multipath Issues in Unicast and Multicast Next-Hop Selection. RFC 2991, 2000.

[7] Chen, J., *et al*. An Efficient Multipath Forwarding Method. INFOCOM'98, 1998.

[8] Rekhter, Y., *et al.* A Provider Architecture for Differentiated Services and Traffic Engineering. RFC 2430, 1998.

[9] Rabbat, R., *et al*. Traffic Engineering Algorithms Using MPLS for Service Differentiation. In: International Conference on Communication (ICC 2000). 2000.

[10] Katz, D., Yeung, D. Traffic Engineering Extensions to OSPF. IETF Draft, 1999.

[11] Cao, Zhi-ruo, *et al*. Performance of hashing-based schemes for Internet load balancing. 2000. http://www.cs.berkeley.edu/~shelleyz/.

[12] Cidon, Israel, Rom, Raphael. Analysis of multi-path routing. IEEE/ACM Transactions on Networking, 1999,7(6):1003~1015.

An IP Traffic Engineering Method for Differentiated Services*

GUO Guo-qiang[1,2], ZHANG Yao-xue[2], WANG Xiao-chun[2]

[1](*Department of Computer Science and Technology*, *Changde Normal University*, *Changde* 415003, *China*);

[2](*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084, *China*)

E-mail: cdggq@263.net; cdggq@21cn.com

http://www.tsinghua.edu.cn

Abstract: The integration of IP TE (traffic engineering) and DS (differentiated service) is a new method to enhance QoS (quality of service) guarantee in network. In this paper, based on analyzing the requirement of DS for TE, the integrating framework is constructed, a schedule method suitable to TE is put forward, as well as the principle and the method of partitioning and mapping. This method carries out traffic splitting according to table. It not only distinguishes the priorities of packets forwarded in DS domain, but also guarantees the requirements for delay and synchronization of data flow.

Key words: traffic engineering; differentiated service; quality of service; load balancing; multipath routing

* Received December 28, 2000; accepted June 20, 2001

Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030409; the Natural Science Foundation of Hu'nan Province of China under Grant No.00JJY2068