DiffServ 中动态优先级调度算法的延迟分析

杜慧军
DU Hui-jun

广东技术师范学院 电信学院,广州 510665
College of Electronics and Information,Guangdong Polytechnical Normal University,Guangzhou 510665,China
E-mail:hdu287@163.com

DU Hui-jun.Delay analysis in dynamic priority scheduling algorithm based on DiffServ.Computer Engineering and Applications,2009,45(17):99-101.

Abstract: Through the analysis of four kinds of priority queues and the priority scheduling algorithm in DiffServ system,the dynamic priority scheduling algorithm is obtained which can be used to solve the fairness problem that occurs when IP blocks are being forwarded.The following problem,however,is the delay phenomenon at the same time.After determining the more precise criterion of delay threshold and the concrete method of realization,find that with the dynamic priority scheduling algorithm,forwarding IP blocks dose not exceed the delay threshold.From the simulations,with the algorithm,it can provide the QoS guarantee that four kinds of priority queues are forwarded fairly.

Key words: Quality of Service(QoS);DiffServ;scheduling algorithm;fairness;network delay

摘 要:通过对 DiffServ 体系的 4 种优先级队列和优先级调度算法的分析,得出了动态优先级调度算法可以解决 IP 分组转发时的公平性问题。但随之而来的问题是 IP 分组转发时的超延迟现象。在确定出较精确的延迟门限标准和具体的实现方法后,得出动态优先级调度算法使 IP 分组的转发不会超出延迟门限。从仿真实验表明,动态优先级调度算法在一般的网络环境和条件下,4 种优先级队列分组的公平性转发能够提供 QoS 保证。

关键词:服务质量;区分服务;调度算法;公平性;网络延迟
DOI:10.3778/j.issn.1002-8331.2009.17.030 **文章编号**:1002-8331(2009)17-0099-03 **文献标识码**:A **中图分类号**:TP393

1 引言

当前,在区分服务(DiffServ)[1]体系的 QoS 保证机制的研究中,业界把主要研究的重点集中在 DiffServ 体系的核心交换机设备上。核心交换机的研究重点是第三层交换与路由相结合的方法。由于核心交换机的体系结构大多趋于采用 Crossbar 交换开关[2]为核心的分布式结构,因此,调度算法也就成为核心交换机中的一个重要研究内容。调度算法的主要目的是为了解决 IP 分组转发时的公平性和延迟性。DiffServ 体系的路由节点缓存区由 4 种不同级别的独立缓存队列构成,它是一种支持优先级的缓存队列,与此相对应的调度算法是优先级调度算法。优先级调度算法分为两种:第一种是静态调度算法;第二种是动态调度算法。静态优先级调度算法实际上是一种单播调度算法。其中,iSLIP 算法[3]是目前较为成功的一种单播调度算法。该算法通过 RR 指针保证了多个端口之间 IP 分组转发时的公平性。该文将在 iSLIP 算法的基础上讨论动态优先级调度算法对不同优先级的 IP 分组转发时的公平性和延迟性。

2 基于优先级调度算法的缓存队列

DiffServ 体系定义了四种聚集业务类型:即:(1)迅速型业务[4](PHB-EF)。它是一种“三低一保证”的业务,“三低一保证”是指低延迟、低抖动、低丢失率和带宽保证等四项指标。它的 DSCP 值(DSCP 是 DS 标记域中的具体值)是“101110”,该业务是一种类似于高速的专线服务;(2)确保型业务[5](PHB-AF)。它是一种“预约带宽”的业务,在网络拥塞的情况下仍能保证该类业务流拥有一定量的预约带宽。它的 DSCP 值是“100XXX、011XXX、010XXX、001XXX”;(3)选择型业务[6](PHB-CS)。它是一种按“分级服务”的业务,它的 DSCP 定义值类型是“XXX000”,它在 DiffServ 域中是按原 IntServ/RSVP 定义的不同级别进行转发处理;(4)缺省型业务[7](PHB-BE)。它是一种“尽力而为”的业务,它的 DSCP 值是“000000”,任何一种路由器都支持缺省型业务流。按照这四种聚集业务的定义,DiffServ 体系在所有的路由设备中构建了 4 种缓冲队列,以致于把进入 DiffServ 体系的所有信息流划分归类为 4 种聚集流,分别缓存到不同的队列中,并进行维护和管理。如图 1 所示。Queue1 表示为第 1 种迅速型 EF 聚集流;Queue2 表示为第 2 种确保型 AF 聚集流;Queue3 表示为第 3 种选择型 CS 聚集流;Queue4 表示为第 4 种缺省型 BE 聚集流。见图 1 所示。显然,这种队列为严格的优先级提供了实现 QoS 的基础。

作者简介:杜慧军(1953-),男,教授,主要研究方向:下一代高智能互联网技术、广域网通信协议。
收稿日期:2008-04-08 **修回日期**:2008-08-13

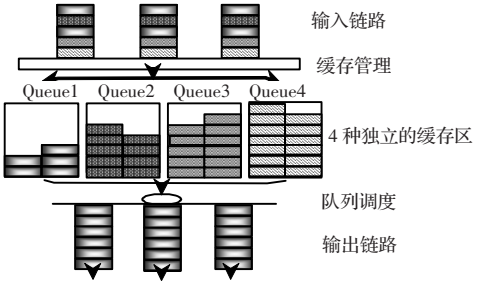

图 1 Diffserv 体系的优先级缓存队列

3 优先级调度算法简介

3.1 静态优先级调度算法简介

静态优先级调度算法是指 IP 分组的优先级别在传输之前便根据某种要求已确定,而且该优先级别在整个传输过程中保持不变。该算法实现简单,可为最高级别的聚集流提供 QoS 保证。iSLIP 算法便是一种静态优先级调度算法。iSLIP 算法的目的是为了公平、有效、快速地匹配一个输入队列从输入端口到输出端口。它是一种迭代算法。每次迭代由三个步骤组成:分别为请求、准许和接受。采用 RR 选择指针进行匹配来保证多个端口之间 IP 分组转发时的公平性。iSLIP 算法的优点是多个端口之间 IP 分组转发的公平性好,容易扩展,对 N 个端口的交换开关只需 lb N 次迭代,易于硬件实现。但是,iLISP 算法每次迭代请求都是为最高级别的 IP 分组由输入端口向输出端口发出的一个信息。包括准许和接受都是针对最高级别 IP 分组的。所以,在 DiffServ 体系中,其他几种低级别聚集流有可能一直得不到处理,这是因为只有在高一级的聚集流队列为空时,才能服务给定较低一级的优先级队列,尤其在高优先级聚集流不断进入缓存队列的情况下,就会导致低级别的聚集流无限期地等待。显然这种 iSLIP 调度算法对 4 种不同优先级聚集流的转发是不公平的。

3.2 动态优先级调度算法简介

动态优先级调度算法是指 IP 分组的优先级别在传输中根据某种要求随着时间而变化,而且随着时间的增大,优先级别也随之提高。这样,等待时间较长的 IP 分组,总会因其优先级别不断提高而被调度转发,该算法实现也相对简单,它在为最高级别的聚集流提供转发的同时,而且还能为其他几种低级别聚集流提供了 QoS 保证,显然它是一种较好的调度算法。这种调度算法的思想为不同优先级别的 IP 分组转发提供了公平性保证。

这里规定,静态与动态优先级调度算法均设定相同的缓存区最大利用率。两者区别仅限于优先级别队列在转发传输过程中优先级别的不变和可变。

4 IP 分组的转发延迟门限

在静态优先级调度算法中,由以上讨论可知,最高级别聚集流的 QoS 是有保证的。在动态优先级调度算法中,由于在最高级别聚集流队列中加入了较低级的 IP 分组,这样是否会影响到最高级别 IP 分组的按时转发,为此作以下延迟门限分析:

现以端到端的传输延迟不大于 0.025 s 为准[8]。由于局域网(校园网内)路过的节点数很有限,所以,端到端的传输延迟主要取决于广域网中核心交换机的节点个数,现推算端到端之间经过的最大节点个数 n,因为全球骨干网的路由节点数已超过十万台[9],但是任一对端到端之间的路由节点数远远没有那么多,由最新资料统计,全球共有 224 个国家和地区[10]。因此得到每个国家平均含有约 446 台骨干网的路由节点。以我国为例进一步推算,有 32 个省直辖市,每个省平均含有约 14 台骨干网的路由节点。设省内骨干网边缘节点到省内骨干网中心节点之间平均有可能经过的最大节点数是 7 台(按两条路计算),而国家级骨干网边缘节点到国家级骨干网中心节点之间平均有可能经过的最大节点数是 8 台(按 4 条路计算)。全球骨干网则是 224 个国家和地区的路由节点之间的连接,所以任意两个国家之间有可能经过的路由最大节点数是 112 台(按 2 条路计算)。因此 n=7+8+112+8+7=142。在全球骨干网中,一般使用 IGRP(专有路由协议),它的最大跳数是 255。显然,142 满足条件。考虑到园区网、校园网的情况,取一个保守值 200,由此得 t=0.025 s/n=0.025 s/200=125 000 ns。现以此延迟门限标准作为 IP 分组的转发延迟门限讨论。

5 动态优先级调度算法的实现方法

按照动态优先级调度算法的思想,较低级 IP 分组随着时间的增大,其优先级别也随之提高。设 $Q_1^i(t)$、$Q_2^i(t)$、$Q_3^i(t)$、$Q_4^i(t)$ 分别代表 4 种不同级别的聚集流。$Q_1^i(t)$ 为最高级别聚集流,$Q_4^i(t)$ 为最低级别聚集流,上标 i 表示分组顺序号。以 $Q_j(t)$ 表示 t 时刻第 j 个队列的 IP 分组个数;以 B 表示整个缓存区的IP 分组个数大小;以 B/4 表示每个队列所能容纳的 IP 分组个数;用 $L_j(t)$ 表示代表 t 时刻第 j 个队列的 IP 分组个数的上限,用 $\triangle t$ 表示传输一个 IP 分组所需时间。

当一个分组在 t 时刻到达缓存区入口时,首先记录这个 IP 分组到达的时间,以保证缓存区每个分组都有唯一的时间,并按照优先级别把该分组归入不同的优先级队列(设为第 j 列),然后计算 t 时刻第 j 队列的 IP 分组个数 $Q_j(t)$,即,

$$Q_j(t)=Q_j^1(t)+Q_j^2(t)+\cdots+Q_j^n(t)$$

根据 RED-DT 算法[11]可知,RED-DT 算法总是预留出一小部分缓存空间,留作更好地处理信息流的突发性。它的缓存区利用率为 $\rho=\beta B/(1+\beta K)$。从 ρ 的表达式可得,当 K=4 时,设定 β=1,那么,整个缓存区的最大利用率为 ρ=80%。这个最大利用率是 IP 分组丢弃的门限。因此晋升门限应小于丢弃门限。又因为缓存区容量一般大于 100 个 IP 分组,所以,选取(B/4)×70%作为晋升门限是合理的。

以此晋升门限为准,对每个进入缓存区的 IP 分组,均要对某一种缓存队列中每个 IP 分组的等待时间进行计算以及 $Q_j(t)$ 计算和 $Q_j(t)/(B/4)$ 判断,所以,只要网络不产生严重拥塞,就不会出现因 IP 分组丢弃而得不到晋升的情况。

按照上述的选取,动态优先级调度算法晋升门限为:

$L_j(t)=(B/4)\times 70\%$

若 $Q_j(t)/(B/4)>L_j(t)$,则取 $F_j(t)=\mathrm{MAX}_{\mathrm{time}}\{Q_j^1(t),Q_j^2(t),\cdots,Q_j^n(t)\}$;作以下计算:

$$F_j(t)=\begin{cases}DS\cap(111111) & j=1\\ DS\cup(101110)\cap(101110) & j=2\\ DS\cup(100111)\cap(100111) & j=3\\ DS\cup(111000)\cap(111000) & j=4\end{cases}$$

其中,DS 是 IP 分组头中的优先级字段,通过这样的方法可使较

低级的 IP 分组向上晋升一级优先级别。这样,由于分组流不断进入缓存队列,因此,在某较低级的队列中可能存在着低级队列和高级队列之分,这时,如果该低级的队列中个数超出丢弃的门限,则丢弃没有晋升的分组,而晋升过的分组受到保护。这种晋升分组方法保证了不同级别聚集流的公平性。

具体的不同优先级队列转发步骤如下:

(1)从 t 时刻开始,对最高优先级队列进行转发,当转发完时刻或到 $t+(B/4)\times70\%\times\triangle t$ 时刻暂停;

(2)从转发完时刻或 $t+(B/4)\times70\%\times\triangle t$ 时刻开始,对所有小于 t 时刻得 3 种较低级别队列均晋升优先级别一个档次,然后选择晋升中最高优先级队列进行转发,当转发完时刻或到$[t+(B/4)\times70\%\times\triangle t]\times2$ 时刻暂停;

(3)重新赋值 $t=[t+(B/4)\times70\%\times\triangle t]\times2$ 或转发完时刻,从 t 时刻开始,回到最高优先级队列再进行转发,当转发完时刻或到 $t+(B/4)\times70\%\times\triangle t$ 时刻暂停;

(4)从转发完时刻或 $t+(B/4)\times70\%\times\triangle t$ 时刻开始,对所有小于 t 时刻的 3 种较低级别队列均晋升优先级别一个档次,然后选择晋升中等待时间最长的最高优先级队列进行转发,当转发完时刻或到$[t+(B/4)\times70\%\times\triangle t]\times2$ 时刻暂停;

(5)重复(3)和(4),直到结束为止。

现以 1 000 个字节长度的 IP 分组为例,取广域网骨干网带宽为 2.5 Gb/s,这时,每个时钟数按每秒钟 2.5 Gb/s 计算,得每个时钟周期为 0.4 ns。以现行的接口数据线 64 位为准,每个总线周期能传送 8 个字节,每个总线周期为 4 个时钟周期,所以解出带宽为 2.5 Gb/s 时,1.6 ns 传送 8 个字节,也就是 1 ns 传输 5 个字节。这种条件下,转发一个分组需要 200 ns,由此得出在一个延迟门限内可转发 625 个 IP 分组。按照上述的具体转发步骤可描述为:Queue1→Queue2→Queue1→Queue3→Queue1→Queue4→Queue1→Queue2→……,如此进行下去,直到转发完 625 个分组。

6　两种调度算法延迟比较

这里用 MATLAB 做了一个仿真实验。这个实验是假设网络没有发生拥塞环境下的仿真,主要目的是只观察 4 种优先级别分组的转发个数情况。并设缓存区容量可容纳 100 个 IP 分组,最大利用率为 ρ=80%晋升门限是 70%,并以 1 000 个字节长度的 IP 分组为例。仿真实验首先随机地发送了 625 个 IP 分组,讨论在一个可允许的延迟门限内(125 000 ns),4 种优先级别分组的转发延迟情况。先做静态优先级调度算法的仿真实验,结果见图 2 所示。后做动态优先级调度算法的仿真实验,结果见图 3 所示。

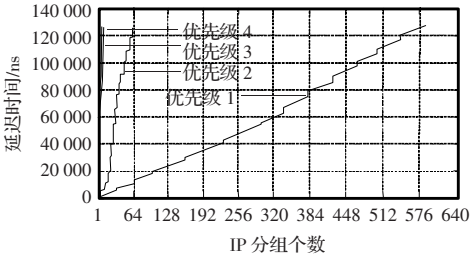

图 2　静态优先级调度算法

由图 2 中可看出,因 IP 分组优先级别不同而转发延迟不同。优先级别最高的分组(优先级 1)在一个延迟门限内可发送 570 个左右 IP 分组,优先级 2 可转发 60 个左右 IP 分组;优先级 3 和优先级 4 的 IP 分组在这段时间内一直没有被转发。这个结果表明,静态优先级调度算法只能为优先级别最高的分组提供QOS 保证。其他 3 种均不能得到保证,所以公平性较差。

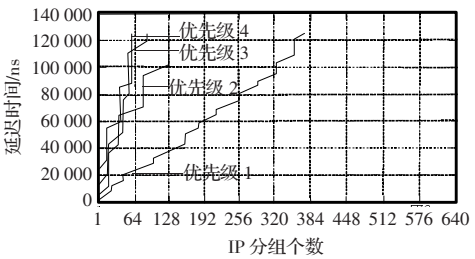

图 3　动态优先级调度算法

从图 3 中可看出,优先级别最高的分组在一个延迟门限内可发送 370 个左右 IP 分组,同时还能可转发 120 多个优先级2的 IP 分组;转发 80 多个优先级 3 的 IP 分组;转发 60 多个优先级 4 的 IP 分组。这个结果表明,动态优先级调度算法为优先级别最高的分组提供 QOS 保证的同时,也能为其他 3 种优先级别的分组提供相应保证。即保证了 4 种优先级别的分组在一个延迟门限内都能得到相应的转发,所以公平性较好。

7　结论

在 DiffServ 体系的 4 种独立缓存队列中,通过引入动态优先级调度算法可以改变过去静态优先级调度算法无法解决的 4 种优先级别 IP 分组转发时的公平性问题。解决公平性问题的具体方法是依据缓存队列容量的 70%来确定晋升门限和依据 IP 分组进入缓存区的时间,当某一种缓存队列的 IP 分组个数超过这个门限时,则晋升一组等待时间最长的 IP 分组级别,除 Queue1 最高优先级队列外,其他晋升优先级队列按照等待时间最长的最高优先级队列优先转发的原则。总之,整个不同优先级队列的转发过程均由这两个参数来控制。这一具体方法解决了不同优先级别 IP 分组转发时的公平性问题。动态优先级调度算法在一般网络环境下使用,无论在公平性上和延迟性上都有很好的性能表现。结果表明这种动态优先级调度算法是 DiffServ 体系中一种能够满足 QoS 保证的调度算法。

参考文献:

[1] Rahbar A G P,Yang O.OCGRR:A new scheduling algorithm for differentiated services networks[J].IEEE Transactions on Parallel and Distributed Systems,2007,18(5):697-710.

[2] Abel F,Minkenberg C,Iliadis I,et al.Design issues in next-generation merchant switch fabrics[J].IEEE/ACM Transactions on Networking,2007,15(6):1603-1615.

[3] W Jing-cun,W Qin,X Xin-ai,et al.TA-iSLIP:A traffic adaptive iSLIP scheduling algorithm[C]//Communications and Networking in China,ChinaCom'06,25-27 Oct.2006:1-5.

[4] Jiang Y M.Delay bound and packet scale rate guarantee for some expedited forwarding networks[J].Computer Networks,2006,50(1):15-28.

[5] Pakdeepinit P,Yeophantong T,Chen P,et al.Balancing secondary traffic metering for DiffServ assured forwarding classes [C]//15th IEEE International Conference on Networks,19-21 Nov 2007:406-411.

(下转 128 页)