EWFQ:一种新的高速网络分组调度算法*)

EWFQ:A Novel Packet Scheduling Algorithms in High Speed Networks

任立勇　卢显良

(电子科技大学计算机学院　成都610054)

Abstract　Packet scheduling algorithm is one of crucial technologies of routers in high speed networks. In this paper we first discuss the limitation of some existing packet scheduling algorithms, then show the quantitative relationships between the GPS system and its corresponding packet WFQ system. A novel packet scheduling algorithm is proposed. It is proven to have following properties: (1) it ensures fair allocation of bandwidth among all sessions; (2) it provides deterministic delay upper bounds to a session whose traffic is constrained by a leaky bucket; (3) it has a relatively low asymptotic complexity of O(logN); (4)it has a relatively low Worst-case Fair Index(WFI). So it can be deployed in routers of high speed networks.

Keywords　Fair queueing, Scheduling algorithms, Quality of service, High speed networks

1　引言

宽带综合业务网要求能给不同的应用提供不同的服务质量(QoS),其中分组调度算法作为网络路由器中的一个重要组件起着相当关键的作用,传统的Internet是基于尽力而为(best-effort)模型实现的,该模型采取先来先服务(FCFS)的分组调度算法,这种模型具有实现简单的特点,它在假定所有应用互相协作的情况下工作得非常好,但当网络发生拥塞时,实时应用的服务质量往往得不到保证。同时,连接间的隔离性能也非常差,吞吐量大的连接得到更多的服务,某些不良行为的连接可能造成其他连接的服务质量急剧下降。

A.K.Parekh 等提出的广义处理器共享(GPS)[1]能较好地解决上述问题:1)当在数据源端实施漏桶算法的流量整形时,GPS能提供端到端的延迟界限,2)GPS能为有分组积压的连接提供公平的带宽分配。由于GPS是基于流体模型的理想化调度算法,不能用于实际系统,于是各种各样的基于分组调度的近似GPS算法相继提出[2~5],这些算法均是在连接带宽公平性分配与计算复杂度间作出平衡,如WFQ尽管提供了可与GPS相当的性能,但由于其计算复杂度较高(O(N))而不能在高速网络中运行,SCFQ算法重新定义了系统虚拟时间计算函数,将计算复杂度降低为O(1),但不足之处在于当网络连接增多时,连接的隔离和保护性能变差,同时SCFQ提供的端到端延迟较大。

为此,本文在研究WFQ与SCFQ的工作机制与不足的基础上,提出了一种新的分组调度算法EWFQ(Extended Weighted Fair Queueing),理论分析与实验证明EWFQ不仅具有计算复杂度低的特点,同时也更好地接近了GPS性能水平。

*)本文得到国家九五重点攻关项目基金资助,信息产业部生产发展基金资助。任立勇　博士研究生,主要研究方向为网络资源管理、网络协议等。卢显良　教授,博士生导师,主要研究方向为操作系统与网络应用技术。

由器的体系结构正朝着速度更快、服务质量更好和更易于综合化管理三个方向发展。

参考文献

1　Asthana A, Delhp S, Jagdish H, et al. Towards a gigabit IP router. Journal of High-Speed Networks, 1992, 1(4): 281～288

2　Partidge C, Carvey P, Burgess E, et al. A 50Gb/s IP router. IEEE Trans. on Networking, 1998, 6(3): 237～248

3　Chen J S, Guerin R. Performance study of an input queueing packet switch wiich with two priority classes. IEEE Trans. Commun[J], 1991, 39(1): 117～126

4　Mckeown N. Fast Switched Backplane for a Gigabit Switched Router. White Paper, http://www.cosco.com

5　Semeria C. Internet Backbone Routers and Evolving Internet Design. White Paper, http//www.juniper.com

6　李津生,等编著.下一代Internet网络技术.人民邮电出版社,2001

2 公平排队调度算法

公平排队调度算法是近似 GPS(Generalized Processor Sharing)服务器[1]的分组调度算法。GPS 服务器持续工作(Work-Conserving)并以固定的速率 r 发送数据，它是一个基于流体模型的理想化调度算法，即 GPS 服务器假定分组无限可分并能同时为多个连接服务。一个有 N 个队列的 GPS 系统表示为 N 个正实数 $\varphi_1,\varphi_2,\cdots,\varphi_N$，其中 φ_i 为 i 连接分配的带宽权值，设 $W_i(t_1,t_2)$ 为连接 i 在时间间隔 (t_1,t_2) 得到的服务量，$W(t_1,t_2)$ 为服务器总的服务量，则 GPS 服务器可定义为：

$$\frac{W_i(t_1,t_2)}{\varphi_i}=\frac{W(t_1,t_2)}{\sum_{j\in B(t_1,t_2)}\varphi_j}\geqslant\frac{W(t_1,t_2)}{\sum_{j=1}^{N}\varphi_j} \quad (1)$$

其中 $B(t_1,t_2)$ 为时间间隔 (t_1,t_2) 内所有有分组积压的连接的集合。令 $g_i=r*\varphi_i$，由(1)可推导出 $W_i(t_1,t_2)\geqslant g_i*(t_2-t_1)$(假定 $\sum_i\varphi_i=1$)，则连接 i 在有分组积压时的实际发送速率 $r_i\geqslant g_i$。可见，GPS 服务器可保证给连接 i 分配最小的服务速率 $g_i=r*\varphi_i$。因此，通过调整连接权值 φ_i，GPS 服务器可灵活地为连接分配带宽。

GPS 良好的工作特性促使许多研究者研究各种近似 GPS 的分组调度算法，其中加权公平排队(WFQ)被认为是最理想的近似 GPS 算法。

定理1 GPS 调度算法与相应的分组调度算法 WFQ(或 PGRS)有如下关系：

(1)设 $d_{i,WFQ}^k,d_{i,GPS}^k$ 分别为 WFQ 与 GPS 系统中连接 i 的第 k 个分组的离开时间，则

$$d_{i,WFQ}^k-d_{i,GPS}^k\leqslant\frac{L_{max}}{r} \quad (2)$$

(2)设 $W_{i,GPS}(0,t),W_{i,WFQ}(0,t)$ 分别为 GPS 与 WFQ 系统中第 i 个连接在时间间隔 $[0,t)$ 内的服务量，则

$$W_{i,GPS}(0,t)-W_{i,WFQ}(0,t)\leqslant L_{max} \quad (3)$$

(3)当在信源端实施参数为 (σ_i,r_i) 的漏桶控制时，WFQ 系统中连接 i 的延迟界限为：

$$\frac{\sigma_i}{r_i}+\frac{L_{max}}{r} \quad (4)$$

(4)令 WFQ 系统中有 N 个连接，连接 i 的速率为 r_i，则它的最坏情况公平指数为：

$$C_{i,WFQ}=N*\frac{r_i}{r}*\frac{L_{max}}{r} \quad (5)$$

文[1]给出了定理1的证明。

WFQ 系统在实现时，设置了一个系统虚拟时间函数 $V(t)$，每个分组到达时计算其虚拟发送时间 S_i^k 与虚拟完成时间 F_i^k，在选择下一个发送分组时，WFQ 系统采用的是 SFF(Smallest Finish First)服务机制。尽管 WFQ 调度算法提供了与 GPS 相当的特性，但它存在两个致命的弱点：(1)在最坏情况下，服务器的 N 个连接全部活跃，此时该算法的计算复杂度为 O(N)，因此，几乎不可能在高速主干网中的路由器中配置 WFQ 调度算法；(2)WFQ 服务器的最坏情况公平指数(WFI，Worst-case Fair Index)与系统中的连接数 N 成正比，因此造成时延抖动增加。

为克服 WFQ 的上述缺点，S. Golestani 提出了另一种算法：SCFQ(Self-Clock Fair Queueing)[3]，SCFQ 算法取当前发送分组的完成时间为系统虚拟时间 $V(t)$，这使算法的计算复杂度降为 $O(1)$，但 SCFQ 算法不足之处在于当网络连接增多时，连接的隔离和保护特性将变差，同时 SCFQ 算法的延迟(Delay)与延迟抖动(Delay-jitter)要比 WFQ 大得多[4]，这主要是因为 SCFQ 在计算系统虚拟时间时的不准确性产生的。J. C. R. Bennett 提出了一种叫 WF²Q(Worst-case Fair Weighted Fair Queueing)的算法[5]，由于 WF²Q 在选择下一个发送分组时采用的是 SEFF(Smallest Eligible Finish First)服务机制，因此其最坏情况公平指数与服务器中的连接数无关，同时，在任一时间段 $[0,t)$ 内，WF²Q 服务器提供给连接 i 的服务量与相应 GPS 服务器提供给连接 i 的服务量相比不会超过该连接的最大分组长度的一个百分比，即 $W_{i,WF^2Q}(0,t)-W_{i,GPS}(0,t)\leqslant(1-\frac{r_i}{r})L_{i,max}$，但遗憾的是，WF²Q 与 WFQ 具有相同的时间复杂度($O(N)$)。

3 改进的分组调度算法

如第2节所述，WFQ 与 WF²Q 算法虽然具有较为满意的分组调度特性：公平性、确定的分组延迟上界、小的 WFI(只针对 WF²Q)，但由于它们采用相同的系统虚拟时间计算函数 $V_{GPS}(t)$，因此具有相同计算复杂度 $O(N)$[1]，很难在高速网络中实现。为此，我们设计并实现了一种新的分组调度算法 EWFQ(Extended Weight Fair Queueing)，EWFQ 采用与 WF²Q 相同的调度机制 SEFF，但重新定义了系统虚拟时间函数，将计算复杂度降为 $O(\log N)$，该函数具体如下：

$$V(t)=\max\{V(\tau)+W(\tau,t),\min_{i\in B,\tau}\{S_i(\tau)\}\} \quad (6)$$

①当连接 i 在信源实施漏桶控制(σ_i,r_i)时，WFQ 提供的端到端延迟及延迟抖动分别为$\frac{\sigma_i+nL_{max}}{r_i}+\sum_{j=1}^{n}\frac{L_{max}}{C_j}$，$\frac{\sigma_i+nL_{max}}{r_i}$，而 SCFQ 的端到端延迟与延迟抖动分别为$\frac{\sigma_i+nL_{max}}{r_i}+\sum_{j=1}^{n}K_j\frac{L_{max}}{C_i}$，$\frac{\sigma_i+nL_{max}}{r_i}+\sum_{j=1}^{n}(K_j-1)\frac{L_{max}}{C_j}$。其中 n 为从源到目的跳数，C_j、K_j 分别为第 j 个路由器的链路速率与共享该链路的连接数。

式中 $\tau \leqslant t$，$W(\tau,t)$为时间间隔(τ,t)内系统的服务量，$B(\tau)$为时刻τ时有分组积压的连接的集合，$S_i(\tau)$为连接i的第一个分组的虚拟开始时间。

如图1所示，EWFQ 算法的具体实现可分为两部分：1）当有新的分组到达时，$ARRIVE(i,k,P)$函数计算该分组的虚拟开始时间与虚拟完成时间，并给分组打上时间标记，然后将该分组送入相应连接的队列；2）分组发送完成后，$FINISH(i,k,P)$函数重新计算系统虚拟时间，然后选择下一个合格的具有最小完成时间(SEFF)分组进行发送。

定理2 EWFQ 的计算复杂度为 $O(\log N)$。

证明：由图1可知，EWFQ 的计算时间主要花费在两部分：计算系统虚拟时间 V(t)与选择下一个最早完成的分组进行发送。其中第一部分需要对所有有分组积压连接的队头分组的虚拟开始时间进行排序以选择在 GPS 系统中最早开始服务的分组(即合格分组)，第二部分需要对所有合格分组按虚拟完成时间排序以选择在实际系统中最早服务分组。最坏情况下，N 个连接全部活跃，此时两部分的计算复杂度均为 $O(\log N)$[5]，因此 EWFQ 的计算复杂度为上述两部分计算复杂度之和，即 $O(\log N)$。 □

```
ARRIVE(i,k,p)
1 IF Q_i(t)=0
2  THEN
3   S_i^k=max{F_i^{k-1},V(t)}
4  ELSE
5   S_i^k=F_i^{k-1}
6 F_i^k=S_i^k+L_i^k/r_i
7 Stamp(P,S_i^k,F_i^k)
8 enqueue(Q_i,P)
FINISH(i,k,p)
1 dequeue(Q_i,P)
2 temp←min_{i∈B(t)}{S_i}
3 V(t)←V(a_i^k)-L_i
4 V(t)←max{V(t),temp}
5 select(p)
```

图1 改进的加权公平排队(EWFQ)

定理3 EWFQ是持续工作的(Work-Conserving)。

证明：由(5)式可以看出，系统虚拟时间至少等于所有队头分组中有最小虚拟开始时间分组的开始时间，反过来讲就是在任意时刻 t 时，只要系统中有分组积压，就至少有一个分组，其虚拟开始时间小于或等于系统虚拟时间，即至少有一个合格分组，这就保证了以 SEFF 为调度机制的 EWFQ 是持续工作的。 □

定理4 当在信源端实施参数为(σ_i,r_i)的漏桶控制时，EWFQ 系统中连接 i 的延迟界限为：

$$\frac{\sigma_i}{r_i}+\frac{L_{max}}{r} \tag{7}$$

由于本定理的证明相当复杂，因此这里只给出相关说明，EWFQ 采用了 SEFF 服务机制，运用文[7]中相同的方法，可以证明 EWFQ 实现的系统为分组速率比例服务器，同时，文[7]证明了对于具有 SEFF 调度机制的分组速率比例服务器，如果在连接 i 的信源端实施参数为(σ_i,r_i)的漏桶控制时，连接 i 的延迟界限为$\frac{\sigma_i}{r_i}+\frac{L_{max}}{r}$。 □

为精确量化分组调度算法与流体 GPS 的差别，文[5]中给出了一个叫最坏情况公平指数(WFI，Worst-case Fair Index)的定义。

定义1 任意时间 t 时，若连接 i 的分组在服务器 s 中的延迟小于等于$\frac{1}{r_i}Q_{i,s}(t)+C_{i,s}$，则称服务器 s 对连接 i 是最坏情况公平的，即

$$d_{i,s}^k \leqslant a_i^k+\frac{Q_{i,s}(a_i^k)}{r_i}+C_{i,s} \tag{8}$$

式中，a_i^k，$d_{i,s}^k$分别表示连接 i 的第 k 个分组到达时间与离开时间，$Q_{i,s}(a_i^k)$表示连接 i 的第 k 个到达时，连接 i 在服务器中分组的队列长度，r_i 为连接 i 的服务速率，$C_{i,s}$为连接 i 的最坏情况公平指数 WFI。

从定义中可以看出为保证服务器对每个连接都是最坏情况公平的，其 WFI 必须尽量小，并与服务器中连接数无关。文[5]中证明了 $C_{i,PS}=0$，同时举例证明了 $C_{i,WFQ}=N\frac{L_{max}}{r}$。

定理5 EWFQ 系统中连接 i 的最坏情况公平指数 WFI 为 $c_{i,EWFQ}=\frac{L_{i,max}}{r_i}-\frac{L_{i,max}}{r}+\frac{L_{max}}{r}$，即：

$$d_{i,EWFQ}^k-a_i^k \leqslant \frac{Q_{i,EWFQ}(a_i^k)}{r_i}+\frac{L_{i,max}}{r_i}-\frac{L_{i,max}}{r}-\frac{L_{max}}{r} \quad \forall i,k \tag{9}$$

式中：a_i^k，$d_{i,EWFQ}^k$分别表示连接 i 的第 k 个分组到达时间和离开时间，$Q_{i,EWFQ}$为时刻 a_i^k 时连接 i 的队列长度，$L_{i,max}$，L_{max}分别表示连接 i 与系统中最大分组长度。

证明 由文[5]可知，WF²Q 系统具有以下特性：

$$d_{i,WF^2Q}^k-d_{i,GPS}^k \leqslant \frac{L_{max}}{r} \tag{10}$$

$$W_{i,WF^2Q}(0,t)-W_{i,GPS}(0,t) \leqslant (1-\frac{r_i}{r})L_{i,max} \tag{11}$$

由于 EWFQ 采用了 SEFF(Smallest Eligible Finish First)调度机制，因此可采用与文[5]中定理1相同的方法证明 EWFQ 也具有上述特性，即：

$$d_{i,EWFQ}^k-d_{i,GPS}^k \leqslant \frac{L_{max}}{r} \tag{12}$$

$$W_{i,EWFQ}(0,t)-W_{i,GPS}(0,t) \leqslant (1-\frac{r_i}{r})L_{i,max} \tag{13}$$

由定理2可知，EWFQ 是持续工作的，因此由(13)式立即可得

$$Q_{i,GPS}(\tau)-Q_{i,EWFQ}(\tau) \leqslant (1-\frac{r_i}{r})L_{i,max} \tag{14}$$

同时由 WFI 的定义可知，GPS 的 WFI 为0，即

$$d_{i,GPS}^k-a_i^k \leqslant \frac{Q_{i,GPS}(a_i^k)}{r_i} \tag{15}$$

由(12)式可得

$$(d_{i,EWFQ}^{k}-a_i^{k})-(d_{i,GPS}^{k}-a_i^{k})\leqslant\frac{L_{max}}{r} \tag{16}$$

$$d_{i,EWFQ}^{k}-a_i^{k}\leqslant(d_{i,GPS}^{k}-a_i^{k})+\frac{L_{max}}{r}\leqslant\frac{Q_{i,GPS}(a_i^{k})}{r_i}+\frac{L_{max}}{r}$$

$$\leqslant\frac{Q_{i,EWFQ}(a_i^{k})+(1-\frac{r_i}{r})L_{i,max}}{r_i}-\frac{L_{max}}{r}$$

$$=\frac{Q_{i,EWFQ}(a_i^{k})}{r_i}+\frac{L_{i,max}}{r_i}-\frac{L_{max}}{r}+\frac{L_{max}}{r} \tag{17}$$

即 $c_{i,EWFQ}=\frac{L_{i,max}}{r_i}-\frac{L_{i,max}}{r}+\frac{L_{max}}{r}$ □

由于EWFQ是在WF²Q算法的基础上设计而成，因此它除了具有上述特性之外，同时还可参照文[5]相同的方法证明EWFQ在带宽分配方面具有公平特性，各连接间的隔离特性也相当好。

4 仿真实验

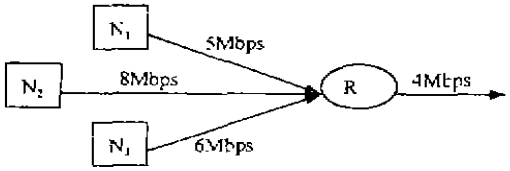

图2 模拟实验网络模型

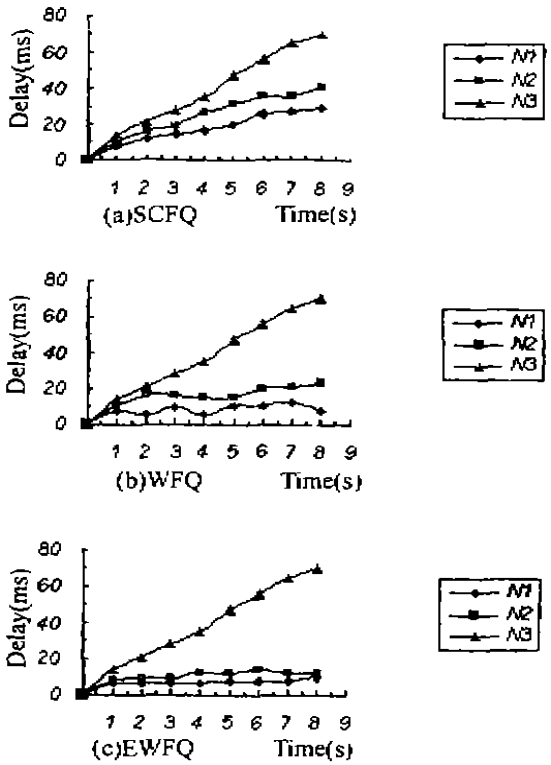

图3 三种算法的延迟比较

下面通过仿真实验进一步说明EWFQ算法的性能。为简单起见，我们采用图2所示网络模型，设该仿真系统有3个连接共享输出链路，其中N_1、N_3发出的信源符合ON-OFF模型，两种状态持续时间分别服从负指数分布，这两个连接信源流都经过漏桶算法整形，整形后流量符合(σ,ρ)模型，假定N_1发出的信源为不良行为流，实验中，该流持续以超过其分配带宽的速率发送数据。如图3、图4分别为路由器采用SCFQ、WFQ与EWFQ算法时各连接的时延及实际占用带宽比较。

如图3所示，仿真实验表明，由于SCFQ算法的隔离性和保护性较差，尽管N_1、N_2发出的流经过漏桶算法整形，但由于N_3发出的流未经任何限制，持续以超过其分配带宽的速率发送数据，从而导致N_1、N_2流延迟特性变差。同时，尽管WFQ系统中漏桶算法整形后的连接有确定的时延上界(如(4)式)，但由于WFQ算法计算复杂度较高，WFI也比较大，因此其时延特性虽比SCFQ算法好，但时延抖动相对较大，比EWFQ算法提供的时延特性差。同样道理，如图4所示，随着时间的变化，N_3发出的流将占用SCFQ系统的绝大部分带宽，而WFQ算法与EWFQ算法由于采用相同的带宽分配策略，因此在带宽分配方面就公平得多，平滑得多。

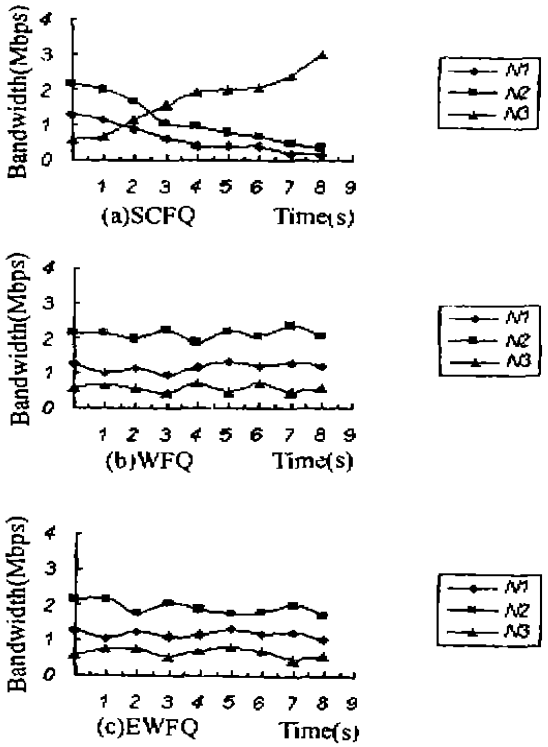

图4 三种算法的带宽比较

结束语 本文提出了一种新的分组调度算法

基于宽带的无线接入技术 LMDS

Wireless Access Technology Based on Broadband

马 燕

(重庆师范学院物理学与信息技术系 重庆400047)

Abstract LMDS(Local Multipoint Distribution Service)is a broadband wireless point-to-multipoint communication system operating above 20 GHz. It occupies the spectrum frequency at around 28 GHz and 31 GHz. It can handle telephony and video programming as well as data services such as Internet access. A big advantage of LMDS is that cable or copper lines for access to the home are not needed. LMDS provides an effective last-mile solution for the incumbent service provider and can be used by competitive service providers to deliver services directly to end users. In this paper, The author analyses the performance, working principle, structure of system of LMDS and its key technology. Also, the paper discusses its foreground.

Keywords LMDS, Base station, User station, Protocol

随着信息化社会的到来,作为连接全球信息的 Internet 的业务量日益巨增,正呈现出爆炸性的增长,因此近年各国都高度重视 Internet 骨干网络的建设,核心网络实现了光纤化,其带宽基本可满足当前对信息获取的需求。而网络的瓶颈越来越集中体现在接入网方面,即用户与核心网络的连接部分。

从理论上说,全光纤接入网络将是比较完美的解决方案,但实现上面临很多的困难,即使在发达国家也还远远未能实现,因此,人们在网络的接入技术方面做了大量的研究,现已提出了多种解决方案,如 cable-modem,即广电的宽带多媒体传送系统,它利用了有线电视系统可用频谱中的一小部分来传送数字信号;DSL,即"数字用户线路",其最具代表性的 ADSL 技术是利用现有的电话双绞铜线对用户提供高速 Internet 接入等业务;ISDN 技术,是传送话音、数据、图像等综合业务的主要手段之一;还有其它如 HFC、PON(无源光网络)、APONC(基于 ATNK 的无源光网络)、SDH 等方式的光纤宽带接入等,这些接入技术都有各自的特点和针对性。

另一方面,无线接入技术特别是宽带无线技术正迅速发展。运用该技术可以将数据、Internet、话音、视频和多媒体应用传送到商业和家庭用户。无线网的组建方便快捷、对人群和环境影响最少,而且无需巨额基础设施和场地投资,作为具有较强竞争性的宽带无线接入技术具有更美好的前景。特别是近年来发展起来的 LMDS 技术,是无线宽带技术的一个典范。

一、LMDS 综述

LMDS(Local Multipoint Distribution Service)是近年来逐渐发展起来的一种工作于24GHz～38GHz 频

EWFQ,理论分析与实验证明该算法不仅具有连接独立的特性,能更公平、灵活地为连接分配带宽,而且它的计算复杂度比 WFQ 算法低得多。同时,EWFQ 还具有较好的时延特性和较低的 WFI,可以为连接提供确定的时延上界和稳定的时延抖动。因此,该算法能更好地满足高速网络的要求。

参 考 文 献

1 Parekh A K, Gallager R G. A generalized processor sharing approach to flow control in integrated services networks: the single-node. IEEE/ACM Trans. Networking, 1993, 1(3): 344～357
2 Demers A, Keshav S, Shenker S. Analysis and simulation of a fair queueing algorithm. In: Proc. ACM SIGCOMM'89: 3～12
3 Golestani S. A self-clock fair queueing scheme for broadband applications. Proc. IEEE INFOCOM'94, Toronto, CA, 1994, 4: 636～646
4 王宏宇,顾冠群. 集成服务网络中的分组调度算法研究综述. 计算机学报,1999,22(10):1090～1099
5 Bennett J C R, Zhang H. WF²Q: Worst-case fair weighted fair queueing. IEEE INFOCOM'96, San Francisco, CA, 1996, 3: 120～128
6 Stoica I, Wahab H A. Earliest eligible virtual deadline first: A flexible and accurate mechanism for proportional share resource allocation. [Tech. Rep TR-95-22]. Old Dominion Univ., 1995, 11
7 Stiliadis D, Varma A. Rate-proportional server: a design methodology for fair queueing algorithms. IEEE/ACM Trans. Networking, 1998, 6(2): 164～174