一种新的基于 GPS 的分组公平调度器

邬海涛, 王重钢, 隆克平, 程时端
(北京邮电大学交换技术与通信网国家重点实验室, 北京 100876)

摘 要: GPS(通用处理器共享)是一种调度算法流模型, WFQ(加权公平排队)、WF^2Q(最差情形公平加权公平排队)等调度算法都是基于对 GPS 的模拟. 本文证明了 WFQ、WF^2Q 等算法并不是 P-GPS(基于分组的 GPS), 也就不能保证 P-GPS 的时延及服务特性. 此外, 本文提出了正确的 P-GPS 的分组公平调度器模型.

关键词: 调度算法; GPS; 加权公平排队; 速率比例调度器
中图分类号: TN915.05 文献标识码: A 文章编号: 0372-2112 (2002) 04-0460-04

A Packet Fair Scheduler Based on GPS

WU Hai-tao, WANG Chong-gang, LONG Ke-ping, CHENG Shi-duan
(National Laboratory, Beijing University of Posts & Telecommunications, Beijing 100876, China)

Abstract: GPS(Generalized Processor Sharing) is one of fluid models for scheduling, and some algorithms, such as WFQ (Weighted Fair Queueing) and WF^2Q(Worst-case Fair WFQ), etc., are based on the simulation of GPS. This paper proves that WFQ, WF^2Q, are not the packet-by-packet GPS. Therefore, these algorithms cannot guarantee the delay and services characteristics of P-GPS. A new Packet Fair Scheduler based on the exact simulation of GPS is proposed.

Key words: scheduling algorithm; GPS; weighted fair queuing; rate-proportional server

1 引言

宽带网络的发展要求一个通信网络要能同时支持多种不同的业务特性的流量, 而不同的业务特性也意味着不同的服务质量(QoS), 具体表现为在带宽、时延、时延抖动、分组丢失率等方面有不同的需求. 现在和将来的各种应用可能有着不同的服务质量需求, 因此网络本身必须具有给不同的应用提供相应服务质量的能力. 而队列调度算法正是提供服务质量保证的重要机制之一.

近年来, 基于 GPS[1] 模型的分组公平排队(PFQ: Packet Fair Queuing)[2-10]调度算法得到了较为广泛的研究. 所有的 PFQ 算法皆基于: (1) 系统维持一个全局函数 $P(t)$, 称为系统虚时间函数[1](或 system potential function[8]), 用以记录调度器已提供的服务量; 调度器利用系统虚时间函数为每个分组计算其相应的开始时间标签和完成时间标签, 见式(1); (2) 执行一定的分组选择策略, 有最小完成时间标签优先(SFF[1,2])、最小合格完成时间标签优先(SEFF[3,4])、最小开始时间标签优先(SSF[5]).

$$\begin{cases} S_i^k = \max(F_i^{k-1}, P(a_i^k)) \\ F_i^k = S_i^k + L_i^k / r_i \end{cases} \tag{1}$$

式中: p_i^k 表示连接(或会话) i 的第 k 个分组; a_i^k 表示 p_i^k 的到达时间; S_i^k 表示 p_i^k 的服务开始时间标签; F_i^k 表示 p_i^k 的服务结束时间标签; l_i^k 表示 p_i^k 的分组长度; r_i 表示连接 i 的预约速率; $P(t)$表示在 t 时刻的系统虚时间函数值.

WFQ[1,2]及其在 WFQ 基础上发展起来的一系列分组公平排队算法[3-5]都是基于对 GPS 的模拟. WFQ 被认为就是 P-GPS[1], 包括 GPS 的提出者也将 WFQ 作为 P-GPS 的实现. 本文将证明 WFQ 算法并不是 P-GPS, 也就不具备文献[1, 3]中所证明的时延及服务特性, 以及在此基础上所证明的其它性质. 并且, 本文提出了对 GPS 的正确模拟方法, 能够实现文献[1, 3]中所证明的时延及服务特性.

2 GPS 与 WFQ

GPS 是一个理想化的流模型, 它根据各队列的共享比例对所有的活动队列同时服务. 在任何时间间隔内, 如有 M 个非空的队列, 这个服务器就按照一定的服务比例对这 M 个队列同时进行服务, GPS 对每个队列业务流保证有明确的端到端的时延上限, 而与其他队列业务流无关.

定义 1 一个 GPS 调度器是工作保持的(work-conserving), 对 N 个队列进行服务, 并且可以用 N 个正实数 $r_1, r_2, \cdots, r_n$ 来表示, 且 $r = \sum_i r_i$. 设 $W_i(\tau, t)$ 为连接 i 在 $(\tau, t]$ 期间所接受的服务. 若某个连接在时间 t 排队的流量为正数, 则称这个连接在时间 t 阻塞. 那么 GPS 可定义为对在时间段 $(\tau, t]$ 内持续阻塞的任一连接 i, 都有下式成立:

收稿日期: 2000-11-27; 修回日期: 2001-04-09
基金项目: 国家自然科学基金(No. 69972008); 高等学校博士点专项科研基金(No. 20010013003)

$$\frac{W_i(\tau, t)}{W_j(\tau, t)} \geqslant \frac{r_i}{r_j} \quad j = 1, 2, \cdots N \tag{2}$$

GPS 模型是流系统, 它可以对队列同时进行服务, 不存在非抢占(non preemption)的单元. 但是实际的系统都是分组系统: 在任何给定的时刻只能有一个分组可以得到服务, 分组的传输是不能被先占的. 因此出现了一类采用逼近基于流的 GPS 模型的分组算法: 分组公平排队算法.

P-GPS 逼近 GPS 模型的方法是: 在假设无后续分组到达的前提下, 选择在 GPS 中最先接受服务或最先完成服务的分组并对之进行调度. 其具体方法是: (1) 引入了虚时间的概念, 系统虚时间函数表示系统在当前活动期间已提供给所有业务流的归一化的服务量; (2) 当分组到达时就对其赋予一个虚开始或完成时间标签(队列的虚时间为队首分组的虚时间标签); (3) 系统根据时间标签选择分组进行服务. 所有的 PFQ 算法都是在某个虚时间函数的基础上按照一定的分组选择策略对队列进行调度.

WFQ 的提出先于 GPS, 其分组选择策略属于 SFF; 而 WF^2Q 的分组选择策略属于 SEFF. WFQ 等系列算法的时间标签都是按式(1)计算, 其系统虚时间都按下式计算:

$$\begin{cases} P(0) = 0 \\ P(t_{j-1} + \tau) = P(t_{j-1}) + \dfrac{\tau \times r}{\sum\limits_{i \in B_j} r_i} \\ \tau \leqslant t_j - t_{j-1}, j = 2, 3, \cdots \end{cases} \tag{3}$$

这里假定在时间段(t_{j-1}, t_j)系统内阻塞的连接没有变化, 且这些阻塞的连接的集合用 B_j 来表示.

定义 2　假定在时间 τ 以后系统无分组到达, 调度器选择在相应的 GPS 系统里最先完成服务的分组进行服务, 则该调度器为 P-GPS(Packet-by-Packet, GPS).

Parekh 等人提出 GPS, 并证明了 GPS 与 P-GPS 之间的关系. 根据文献[1], GPS 系统的分组服务完成的顺序与将来的分组到达无关. GPS 与其对应的 P-GPS 间存在下列性质[1]:

$$d^k_{i,\,\text{P-GPS}} - d^k_{i,\,\text{GPS}} \leq L_{\max}/r \quad \forall i, k \tag{4}$$

$$W_{i,\,\text{GPS}}(0, \tau) - W_{i,\,\text{P-GPS}}(0, \tau) \leq L_{\max} \quad \forall i, k \tag{5}$$

其中 $d^k_{i,\,\text{P-GPS}}$和 $d^k_{i,\,\text{GPS}}$分别表示连接 i 的第 k 个分组在 P-GPS 和 GPS 中的离开时间, 而连接 i 在 P-GPS 和 GPS 下所接受的服务分别用 $W_{i,\,\text{P-GPS}}(0, \tau)$和 $W_{i,\,\text{GPS}}(0, \tau)$来表示.

因为 GPS 的提出者就认为 WFQ 就是 P-GPS, 使得人们认为 WFQ 满足式(4)、(5), 在此基础上推导出 WFQ 等算法的时延和业务特性.

3　WFQ 不是 P-GPS

通过一个反例可以说明 WFQ 等都不能满足式(4)、(5). 与文献[1]中的假设相同, 认为分组的到达时间是该分组的最后一个比特到达调度器的时间, 分组的服务完成时间是该分组的最后一个比特离开调度器的时间. 假定调度器的服务速率为单位速率 $r = 1$, 所有连接的分组都是单位长度, 即 $L_{\max} = L_i = 1$. 在该调度器中存在 6 个连接, 连接 1 到 6. 其中连接 1 的确保速率为 0.5, 其余 5 个连接的确保速率为 0.1. 各个连接的到达方式如图 1 所示. 连接 1 的到达速率与其确保速率相同; 连接 2 到 6 的到达速率大于其确保速率, 使得连接 2 到 6 持续阻塞. 开始时间设为 0.

在 WFQ 里, 分组的到达和离去被视为事件. 当事件的发生使得阻塞队列的集合 B_j 发生了变化时, 虚时间和计算虚时间的导数都要更新. 分组的到达时间如图 1 所示: $a^k_1 = 2(k-1)$, $k = 1, 2, \cdots$, 而 $a^1_i = 0$, $i = 1, 2, \cdots, 6$. $P(0) = 0$. 各个连接的分组标签按式(3)计算, 可得: $S^1_i = 0$, $i = 1 \cdots 6$; $F^1_1 = 2$, $F^1_i = 10$, $i = 2 \cdots 6$. 因此在时间 $t = 0$ 时分组 p^1_1 首先被选中服务. 在时间段$(0, 1)$内阻塞的连接为 1 到 6, 因此 $\sum_{i \in B_j} r_i = 1$, 可得, $P(1) = 1$. 由于连接 1 的第一个分组在 $t = 1$ 时被服务完, 在时间段$(1, 2)$内阻塞的连接为 2 到 6, 因此 $\sum_{i \in B_j} r_i = 0.5$, 可得, $P(2) = P(1) + 1/0.5 = 3$. 依次类推, 可得 $P(10) = 15$, $P(12) = 18$, 分组的服务顺序如图 1(分组符号位置的不同是为了清楚的表现连接 1 分组的服务时间). 连接 2 到 6 的第二个分组的时间标签分别为 $F^2_i = 20$, $i = 2 \cdots 6$. 对连接 1 的第 6 个到第 10 个分组, 其完成时间标签分别为: $F^6_1 = 15$, $F^7_1 = 18$, $F^8_1 = 21$, $F^9_1 = 23$, $F^{10}_1 = 25$. 从图 1 中可以看到, p^8_1 的分组服务完成时间在 GPS 下和在 WFQ(或 WF^2Q)下相差的时间为 2, 即 $2L_{\max}/r$, 显然式(4)不成立. 并且 $W_{1,\,GPS}(0, 17) = 8.5$, 而 $W_{1,\,P-GPS}(0, 17) = 7$, 显然式(5)也不成立.

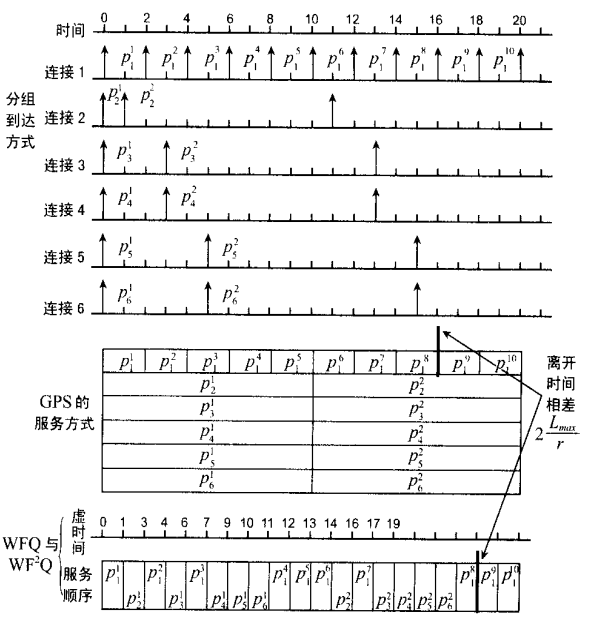

图 1　GPS 与 WFQ(WF^2Q)服务的差异

WFQ(WF^2Q)算法不能保证式(4)、(5)成立, 也就不是 P-GPS. 因此, 在式(4)、(5)基础上所得出的 WFQ 等能保证流量受漏桶算法限制的连接的时延特性的结论是不对的.

4　GPS 的正确模拟方法

对 GPS 的模拟需要引入系统虚时间的概念, 而定义 1 是一个理想的流模型的定义, 我们需要另一个等效定义来提供

系统的虚时间, 即 RPS[8-10](速率比例调度器)模型. 在 RPS 里, 每个正接受服务的连接都能即刻得到与其预约速率成正比的服务量, 所有正接受服务的连接的虚时间函数值以相同的速率增加, 这提供了公平调度的保证.

根据文献[8], GPS 是 RPS 的特例, 在任何时刻, GPS 的系统虚时间都和所有阻塞的连接的虚时间相等, 即

$$P(t)=P_i(t)\quad \forall i\in B(t) \tag{6}$$

因此, 可根据下式来计算 GPS 的系统虚时间

$$P(t_2)-P(t_1)=P_i(t_2)-P_i(t_1)=\frac{W_i(t_1,t_2)}{r_i}$$

$$=\frac{\sum_i W_i(t_1,t_2)}{\sum_i r_i}=\frac{W(t_1,t_2)}{\sum_i r_i}=\frac{r\times(t_2-t_1)}{\sum_i r_i}$$

$$\forall i\in B(t_1,t_2) \tag{7}$$

这里假定在时间段(t_1,t_2)内阻塞的连接的集合$B(t_1,t_2)$没有发生变化. 从式(7)可以看出, 对 GPS 模拟的关键在于正确的找到 GPS 系统中正阻塞的连接的集合. 而 WFQ(WF^{2}Q)是采用式(3)对 GPS 系统的虚时间进行模拟, 由于分组系统都是非抢占的, 因此二者阻塞的连接集合并不一定是相同的, 从而导致 WFQ(WF^{2}Q)不能保证式(4)和(5)成立.

在 GPS 系统中, 对某个分组开始服务时的系统虚时间是该分组达到时的系统虚时间和该分组所在连接的前一分组的服务完成时的系统虚时间的最大值. 因此, 按式(1)来对分组打标签的策略是可行的, 分组的选择策略可以变化, 关键问题是系统虚时间的计算. 假定对 GPS 系统里的分组也按式(1)加上标签(GPS 仍按 RPS 的规则服务, 该标签只是为了便于理解). 那么, 当某个连接的分组的完成标签的最大值(即该连接的队尾分组的完成标签)大于系统虚时间时, 该连接有分组在 GPS 系统中接受服务, 即该连接在 GPS 中是阻塞的. 因此, 判断一个连接是否在 GPS 中阻塞应根据其队尾分组的完成标签. 根据以上分析, 有以下定理:

定理 1 GPS 系统中, 若某个连接的分组按式(1)计算所得的完成标签的最大值大于系统虚时间, 是该连接处于阻塞状态的充要条件.

证明 充分性. 因为 GPS 下的分组完成标签是分组服务完成时的系统虚时间, 所以如果某个队列存在分组标签大于系统虚时间, 则该分组必然仍在服务或尚未服务, 则它所在的队列处于阻塞状态.

必要性. 若某个连接阻塞, 则必定有分组未服务完, 该分组的完成时间标签必大于系统虚时间. 则该连接的分组完成时间标签的最大值大于系统虚时间. 证毕.

在分组系统中仍然可以计算分组的时间标签. 因此, 可以通过判断某个队列的队尾分组的完成标签是否大于系统虚时间来判断该连接是否在对应的 GPS 系统中处于阻塞状态. 根据文献[1, 3], 有

定义 3 系统忙期是系统处于不空闲状态的最大时间段.

引理 1 若两个服务系统都是工作保持的, 则两个系统的系统忙期相同.

本文中$F_{i,h}$表示连接i的队首分组的完成标签, $F_{i,t}$表示连接i的分组的完成标签的最大值, 即连接i的队尾分组的完成标签; 设$F_{min,h}$表示所有连接的队首分组的完成标签的最小值(也是所有分组的完成标签的最小值), 设$F_{min,t}$表示所有连接的队尾分组的完成标签的最小值.

定义 4 分组系统当时间$t=\tau$时的虚阻塞集合为$VB(\tau)=\{j: F_{j,t}>P(\tau)\}$.

定义 5 分组系统的阻塞更新事件为下列事件的任一事件:

(1)τ时刻某个空闲的连接有分组p_i^k到达, 并且满足$P(a_i^k)\geqslant F_i^{k-1}$, 设$F_i^0=0$, 将连接$i$加入虚阻塞集;

(2)τ时刻某个队列的$F_{i,t}=P(\tau)$, 将连接i从虚阻塞集中删除.

定义 6 若某分组调度系统按式(1)为分组加标签, 按定义 5 的阻塞更新事件来更新系统的虚阻塞集合, 分组选择策略为 SFF, 即选择最小完成时间标签的分组服务, 按下式计算系统虚时间,

$$P(t_2)-P(t_1)=r\times(t_2-t_1)/\sum_i r_i$$

$$\forall i\in VB(t_1,t_2), \text{若}(t_1,t_2)\text{内虚阻塞集不变} \tag{8}$$

则称该系统为新加权公平排队系统, 简称 N-WFQ.

引理 2 N-WFQ 的系统虚时间和与它对应的 GPS 的系统虚时间相等.

证明 因为 N-WFQ 按 SFF 选择分组, 是工作保持的, 其对应的 GPS 也是工作保持的, 根据引理 1, 二者的系统忙期相同. 因此, 只需证明一个系统忙期内二者的虚时间相同. 如果系统虚时间在每个阻塞更新事件发生时都和对应的 GPS 的虚时间相等, 则 NWFQ 的虚阻塞集合与 GPS 的阻塞集合相同, 则该系统能保证虚时间始终相等. 下面用数学归纳法证明这一假设.

设τ_k表示第k个阻塞更新事件发生的时间. 当第一个分组到达时, 系统从空闲变为忙, 设该时间为 0, 则$P(0)=0$. 而$t=0$也正是第一个阻塞更新事件发生的时间, 因此虚时间相等对第一个阻塞更新事件的时间成立, 并且此时 GPS 的阻塞集合与 N-WFQ 的虚阻塞集合相同. 假设虚时间相等对第k个阻塞更新事件的时刻成立, 即$P_{NWFQ}(\tau_k)=P_{GPS}(\tau_k)$, 到下一次阻塞更新事件发生之前, GPS 系统的阻塞集合没有变化, 根据虚阻塞集合以及阻塞更新事件的定义, N-WFQ 的虚阻塞集合也不变, 即二者保持相同, 因此第$k+1$个阻塞更新事件的系统虚时间相等也成立. 证毕.

定理 2 N-WFQ 是 P-GPS 系统.

证明 根据定义 2, 6 以及引理 2, 可得出该结论. 证毕.

推论 1 N-WFQ 能满足式(4)和(5), 能保证 P-GPS 的时间及服务特性.

定义 7 若某分组调度系统与定义 7 中系统的唯一区别是按 SEFF 选择分组, 即选择最小合格完成时间标签的分组服务, 则该调度系统为新 WF^{2}Q, 简称 N-WF^{2}Q.

推论 2 N-WF^{2}Q 能满足式(4)和(5), 且能保证文献[3]中定理 1 证明的 WF^{2}Q 的时间及服务特性(但 WF^{2}Q 实际上不能

实现这些特性).

推论3　将式(3)阻塞连接集合改为定义4中的虚阻塞集合, 能真正实现文献[1]中定理1, 2证明的(但实际上WFQ等所不能保证的)性质, 即式(4)、(5).

为便于理解N-WFQ系列算法与WFQ系列算法的不同, 将第3节反例中分组到达模式下N-WFQ的分组服务顺序画在图2中.

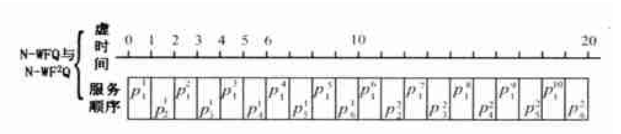

图2　N-WFQ(N-WF²Q)的分组服务顺序

5　算法的实现

我们考虑N-WFQ的实现, 对N-W²FQ的实现可采用文献[3]中实现WF²Q的方法, 由调节器(regulaor)加N-WFQ的方法来等效实现N-W²FQ.

N-WFQ选择分组服务时需要找出完成时间标签最小的分组, 可将所有队列的队首分组按完成时间标签排序, 新来的分组加入时的复杂度为 $O(\log_2 N)$, 这一点与WFQ以及其它分组公平排队的算法排序的复杂度相同.

同WFQ相比, N-WFQ实现的关键区别是对虚阻塞集合的维护. 可采用下面的方法: 按定义4维护一个队列, 该队列中的元素不是分组, 而是每个队列的队尾分组的完成时间标签, 所有元素所代表的队列构成虚阻塞集合. 根据定义5的阻塞更新事件, 该队列中元素的加入或更新是在新分组到达时: 某个队列的队尾分组的完成时间标签发生变化, 则该队列也需进行相应的调整, 复杂度为 $O(\log_2 N)$. 按照定义5的第2条, 某个队尾分组服务完离开队列时, 并不将其对应的元素从该队列中删除, 因此需考虑元素的删除时间, 即某个队尾时间标签值与系统虚时间相等的时刻, 规则如下: 设 t 为当前时间, $Next(t)$为下一次删除元素的时间, 假定 t 以后没有分组到达, $Next(t)$的值可按下式计算

$$F_{min,t} = P(t) + (Next(t) - t) / \sum_{i \in VB(t)} r_i \qquad (9)$$

$$\Rightarrow Next(t) = t + (F_{min,t} - P(t)) \sum_{i \in VB(t)} r_i \qquad (10)$$

如果由于新分组的到达使得队列中的元素只是顺序发生变化, 则虚阻塞集合不变, 但需更新 $Next(t)$. 因为需要取队尾分组标签的最小值, 因此该队列需要排序, 将新的队尾标签插入的操作的复杂度为 $O(\log_2 N)$, 这样, 系统排序复杂度的数量级没有增加. 而且, 因为只是对固定大小的元素(时间标签)排序, 可采用硬件固化操作, 以提高速率.

6　结束语

作为提供服务质量的一种重要的手段, 分组调度算法近年来得到了较充分的研究. 本文的贡献主要如下: (1)证明了WFQ, WF²Q及其系列算法不能保证其相应的文献中证明的性质, 以及由这些性质所推出的结论; (2)提出并证明了如何对GPS的系统虚时间进行正确模拟的方法; (3)应用这一方法, 可得到实现P-GPS的正确算法N-WFQ, 能保证文献[1]中定理1, 2所证明的性质(即式(4)、(5)).

参考文献:

[1] A K Parekh, R G Gallager. A generalized processor sharing approach to flow control in integrated service networks: the single-node case [J]. IEEE/ACM Trans. Networking, 1993, 1(3): 344-357.

[2] A Demers, S Keshav, S Shenker. Analysis and simulation of fair queueing algorithm [J]. J Internetworking Res Experience, 1990, 1(10): 3-26.

[3] J C R Bennett, H Zhang. WF²Q: Worst-case fair weighted fair queuing [C]. IEEE INFOCOM'96. San Francisco: CA, 1996.

[4] J C R Bennett, H Zhang. Hierarchical packet fair queuing algorithms [J]. IEEE/ACM Trans On Networking, 1997, 5(5): 675-689.

[5] P Goyal, H M Vin. Start-time fair queuing: A scheduling algorithm for integrated services packet switching networks [J]. IEEE/ACM Trans on Networking, 1997, 5(5): 690-703.

[6] H Zhang. Service disciplines for guaranteed performance service in packet switching networks [J]. Proceedings of the IEEE, 1995, 83(10): 1374-1396.

[7] 王重钢, 隆克平, 龚向阳, 程时端. 分组交换网络中队列调度算法的研究及其展望[J]. 电子学报, 2001, 29(4): 553-559.

[8] D Stiliadis, A Varma. Rate proportional servers: A design methodology for fair queuing algorithms [J]. IEEE/ACM Trans on Networking, 1998, 6(2): 164-173.

[9] D Stiliadis, A Varma. Efficient fair queueing algorithms for packet-switched networks [J]. IEEE/ACM Trans on Networking, 1998, 6(2): 175-185.

[10] 王重钢, 隆克平, 龚向阳, 程时端. ERPS: 一种增强的速率成比例调度器[J]. 电子学报, 2001, 29(7): 912-915.

作者简介:

邬海涛　男, 1976年9月生于江西省南昌市, 北京邮电大学交换技术与通信网国家重点实验室博士研究生, 1998年毕业于北京邮电大学电信工程工程系, 2000年9月攻博, 目前研究方向为宽带网络服务质量, TCP/IP改进, 区分服务, 流控和拥塞控制, 及无线分组网络性能.

王重钢　男, 1974年5月生于四川省乐山市, 北京邮电大学交换技术与通信网国家重点实验室博士研究生. 1999年9月入北京邮电大学攻读博士学位, 目前研究方向为宽带网络资源管理和业务控制, 路由算法及服务质量.