并行交换中支持包保序的缓存结构及调度算法

兰巨龙[1], 董雨果[2], 陈 越[1], 温建华[1]
(1. 郑州信息工程大学国家数字交换系统工程技术研究中心, 河南郑州 450002;
2. 空军工程大学电讯工程学院, 陕西西安 710077)

摘 要: 并行交换结构的并行工作原理和负载平衡特性使得属于同一流的分组或者信元被分散到多个交换模块进行处理, 在输出端口它们的先后顺序无法得到保证. 为了解决该问题, 本文提出一种支持包保序的新技术, 包括两级虚拟输入排队(VIQ)结构和包保序轮询(SKRR)调度算法, 并且从理论上分析了这种新技术的吞吐率和时延性能.

关键词: 并行交换; 负载平衡; 时延; 调度
中图分类号: TN919.21 文献标识码: A 文章编号: 0372-2112 (2004) 12A-035-04

The Buffer Structure and Scheduling Algorithm for Maintaining Packet Order in the Parallel Switch

LAN Jur-long[1], DONG Yur-guo[2], CHEN Yue[1], WEN Jiarr-hua[1]
(1. *NDSC, Information Engineering University, Zhengzhou, Henan* 450002, *China*;
2. *Telecommunications Engineering Institute, Air Force Engineering University, Xi' an, Shaanxi* 710077, *China*)

Abstract: Due to the parallelism and load-balancing of parallel switches, the packets (or cells) within the same flow will be spread into several low speed switching fabrics for processing. When these packets are sent to the output, however, their sequence can not be guaranteed. For keeping packet order, this paper proposes a novel technique that includes the buffer structure of two-stage Virtural Input Queues (VIQ) and the scheduling algorithm named sequence keeping round-robin (SKRR). The throughput and average delay performance of the technique are also analyzed.

Key words: parallel switch; load balancing; delay; scheduling

1 引言

随着高速宽带通信网的日益发展, 网络设备的交换能力正成为一个制约现代网络发展的主要瓶颈. 并行交换(PS-parallel switching)结构[1~4]能够将多个 G 比特级的交换结构组建成 T 比特级(甚至更高级)的交换系统, 从而突破器件技术的限制, 极大提升网络设备的交换能力. 包保序对于 PS 结构是十分重要的, 因为, 如果到达的包通过不定长分组的形式进行交换, 分组包乱序将会导致当前版本 TCP 出错[5,6]; 而如果到达的包首先通过切片成为很多 cell[7,8], cell 失序将为以后 cell 重组造成严重的问题[3,9].

在一个 PS 结构中, 根据并行原理(parallelism), 目的地为同一输出端口的高速分组包将被分散到多个低速的交换模块中进行交换处理. 由于各个交换模块的时延状态各异, 当这些包被送到输出端口时, 它们的先后顺序很可能被破坏. 目前公开的文献中至少有两类措施来纠正包失序问题: 第一类措施是在低速交换模块中建立整形机制(reshaping mechanism)[3], 但这种办法不能解决包在输出端口失序的问题; 第二类措施是在输出端口设置大容量的缓存以便重新恢复包的先后顺序, 例如输出排队 OQ[2] 和单级虚拟输入排队 VIQ[10,11], 但是该方法可能导致 HOL 阻塞[12]. 此外还存在其他的包保序技术, 例如两级交换的 3DQ 缓存技术[12] 和 IPv6 路由器的 POKD 技术[13], 但是它们并不适用于 PS 结构.

本文为 PS 结构提出一种支持包保序的新技术, 包括两级虚拟输入排队(VIQ)结构和包保序轮询(SKRR)算法: 包在两级 VIQ 中按照它们的先后顺序重新排序, 然后由 SKRR 算法进行调度, 分配到目的输出端口. 理论分析表明, PS 结构采用这种新技术后, 拥有与输出排队(OQ-output queue)结构相同的吞吐率, 并且其平均时延不大于 OQ 的平均时延加上一个常数.

2 PS 结构模型

一种典型的 PS 结构模型如图 1 所示, 该 PS 结构为 $N \times N$, 端口线速率为 R, 每个输入端口包含一个分路器(Demultiplexer), 输入分路器中的缓存分为 K 段 FIFO, 每段 FIFO 存储对应的到中间各交换模块的业务[3,14,15], 每个输出端口包含

收稿日期: 2004-08-18; 修回日期: 2004-11-06
基金项目: 国家 863 重大课题基金(No. 2001AA121011)

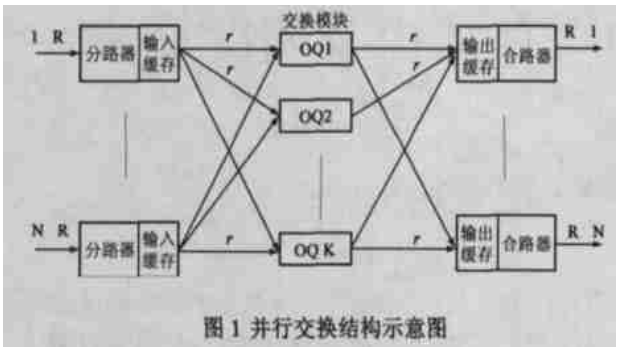

图 1 并行交换结构示意图

一个合路器(Multiplexer); 交换部分由K 个OQ 交换模块组成, 每个OQ 交换模块为 $N \times N$, 端口线速率为 r, 内部加速比(Core speedup) S 定义为Kr/R. 从工程实现角度出发, 本文假定 S 为 1, PS 结构处理的单元为 cell, 每个输入分路器的调度判决都是独立执行的, 各交换模块之间没有共享信息, PS 内部亦无反馈通信, 并且 PS 结构满足尽职工作(Work-conserving)[1~3]的条件.

我们在本文中约定, Cell: 由 IP 分组包切片后形成的等长数据包, 长度一般为 512 比特或 64 字节; 时钟: 在线速率为 R 的条件下, 发送或接收一个 Cell 的时间; 层: 指中间的 OQ 交换模块, 例如, 第 k 层即指第k 个 OQ 交换模块.

PS 模型中包保序的要求是: 对于任意输出端口, 来自相同输入端口的 cell 必须依照其进入 PS 结构的先后顺序来离开 PS 结构[2,3,9]. 例如, 任意一段时间内, 目的输出端口为 j 的所有 cell 按照先后次序 Q 抵达输入端口 i, 在 PS 结构中进行并行交换后, 它们从输出端口 j 发出, 必须按照次序 Q 离开 PS 结构. 然而, 如下原因将导致 cell 在 PS 结构中失序.

® cell 在输入缓存中失序——因为负载平衡的需要, 到同一输出端口的 cell 将被均分到各个交换层对应的 FIFO. 由于各个FIFO 的拥塞状况不尽相同, 因而这些 cell 在FIFO 中的排队时延也不一样, 它们抵达 PS 结构的先后次序可能因为排队时延不相同而被扰乱. 当然, 同一 FIFO 中的 cell 是不会失序的;

® cell 在 OQ 交换模块中失序——与上面的分析同理, cell 因为在各个OQ 交换模块中面临不同的排队时延而导致失序.

3 包保序技术

我们解决 PS 结构中 cell 失序问题的基本思路为: 首先, 在每一层 OQ 的输出端口根据 PS 结构的输入端口对 cell 进行整形排序, 纠正之前的失序问题, 然后采用保序调度算法把重新排序后的 cell 分配到 PS 结构输出端口的虚拟输入排队缓存中, 从而确保 cell 按照先后顺序离开输出端口. 以下分别介绍两级虚拟输入排队缓存结构和保序调度算法的原理.

3.1 两级虚拟输入排队

如图 2 所示, 在每一层的输出端口, cell 根据输入端口在 VIQ_1 中排队, VIQ_1 包含 N 个 FIFO, 对应 N 个输入端口. 例如, $FIFO_1$ 缓存来自输入端口 1 的 cell, $FIFO_2$ 缓存来自输入端口 2 的 cell, …, $FIFO_N$ 缓存来自输入端口 N 的 cell. 我们可以看出 VIQ_1 中任意 FIFO 里缓存的 cell 具备如下性质: 它们的输入端口和输出端口都是相同的, 例如, 从输入端口 2 来的、通过第 k 层、目的地为输出端口 j 的所有 cell 都会在第 k 层输出端口 j 的 VIQ_1 中的 $FIFO_2$ 中缓存排队. 如我们在第 2 节分析的那样, 同一 FIFO 内的 cell 不会失序, 所以 VIQ_1 能够确保 cell 的先后顺序不被扰乱. 以下我们用 $VIQ_1(i,k,j)$ 表示第 k 层的第 j 输出端口的 $FIFO_i$.

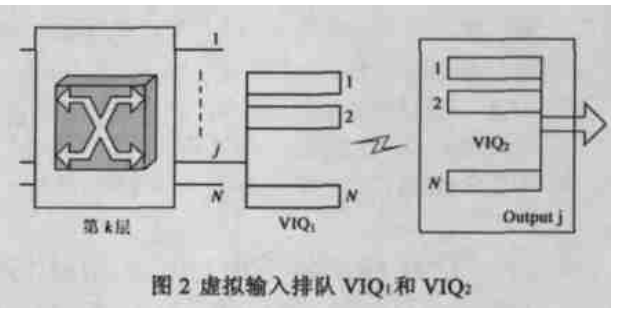

图 2 虚拟输入排队 VIQ_1 和 VIQ_2

VIQ_2 也包含 N 个FIFO, 到达同一输出端口的 cell 根据输入端口在 VIQ_2 中排队. $VIQ_2(i,j)$ 表示第 j 输出端口的 $FIFO_i$. 这样, VIQ_2 中的每个 FIFO 实际上是缓存了来自相同输入端口、经过不同层的 cell. 基于 VIQ_1 和 VIQ_2, 我们提出的问题是: 要实施 cell 保序, VIQ_1 中的 cell 应该怎样分配给 VIQ_2? 显然, 我们需要合适的负载平衡调度算法(输入端口分路器执行的)和保序调度算法(输出端口合路器执行的).

3.2 保序轮询调度算法

文献[9, 16] 分析得出, 一种简单的负载平衡调度算法即为把目的地为同一输出端口的 cell 按照轮询方式均匀的分配给每一层. 本文采纳这种负载平衡算法, 例如, 我们考虑如何调度输入端口 i 上的要到输出端口 j 的所有 cell, 第 1 个 cell 分配给第 1 层, 并在 $VIQ_1(i,1,j)$ 中排队; 第 2 个 cell 分配给第 2 层, 并在 $VIQ_1(i,2,j)$ 中排队 …第 N 个 cell 分配给第 N 层, 在 $VIQ_1(i,N,j)$ 中排队; 第 $N+1$ 个 cell 分配给第 1 层, 在 $VIQ_1(i,1,j)$ 中排队. 显然, 为了保序, $VIQ_2(i,j)$ 应该从 VIQ_1 中按照如下方式收集 cell: 第 1 个 cell 从 $VIQ_1(i,1,j)$ 收集, 第 2 个 cell 从 $VIQ_1(i,2,j)$ 收集…第 N 个 cell 从 $VIQ_1(i,N,j)$ 收集, 然后, 又开始从 $VIQ_1(i,1,j)$ 收集…

在 SKRR(sequencer keeping round robin) 算法中, $VIQ_2(i,j)$ 用轮询指针 $p1(i,j)$ 来记住前一个 cell 来自于哪个层或 $VIQ_1(i,k,j)$, 输出端口 j 用指针 $p2(j)$ 以轮询方式从 VIQ_2 中收集 cell. 可以看出, 指针 $p1(i,j)$ 用来支持 cell 保序, 而指针 $p2(j)$ 用来保证尽职工作以及调度的公平性. 这两种指针的主要区别是: 如果 $p1(i,j)$ 指向一个空的 VIQ_1, 那么它将一直指向该 VIQ_1 项直至该 VIQ_1 不为空; 但是, 如果 $p2(j)$ 指向的 VIQ_2 为空, 它将跳过空的 VIQ_2 项直到指向一个非空的 VIQ_2 项. SKRR 算法的形式化描述如下:

每一个 VIQ_2 执行如下操作:

1. 初始化, $p1(i,j) \leftarrow VIQ_1(i,1,j)$;
2. $p1(i,j) \leftarrow VIQ_1(i,(k+1) \bmod K, j)$.

并且, 每一个输出端口 j 执行如下操作:

If $VIQ_2(i,j) \neq 0$, then $p2(j) \leftarrow VIQ_2(i,j)$; else $p2(j) \leftarrow VIQ_2((i$

+ 1) mod N, j).

从以上分析描述可以看出,SKRR 算法简单且便于实现,不过该算法需要 N^2K 个缓存,而相同交换能力的 OQ 仅仅需要 N^2 个缓存,这还意味着 SKRR 算法需要复杂的缓存管理机制来处理大量的指针.

4　性能分析

本节用一个 OQ 结构作为参考交换结构来对比分析 PS 结构的性能.方便起见,以下用 SKRR 来表示应用两级 VIQ 和 SKRR 的 PS 结构.我们将比较 SKRR 和 OQ 的平均时延性能和吞吐率.一些符号说明如下,$IQ_i^k(t)$:系统时间[0, t] 内第 i 输入端口中上与第 k 层对应 FIFO 的队列长度;$A_{ij}(t)$:系统时间[0, t] 内抵达第 i 输入端口、目的地为第 j 输出端口的 cell 总和;$A_{ij}^k(t)$:$A_{ij}(t)$ 中被分配到第 k 层的 cell 总量;类似地,$B_{ij}(t)$和 $B_{ij}^k(t)$ 分别表示离开输入端口抵达交换层和抵达第 k 交换层的 cell;$C_{ij}(t)$ 和 $C_{ij}^k(t)$ 分别表示离开交换层和离开第 k 交换层抵达输出端口的 *cell*;$D_{ij}(t)$ 表示离开输出端口的保序 *cell* 总量.

4.1　平均排队时延

本文的分析采用了与文献[12, 17] 相同的思想,即为:在外来输入流量为 PS 结构输入端口流量的条件下,比较 SKRR 与 delayed-OQ(见定理 1);然后在外来输入流量为 PS 结构交换层流量的条件下,比较 delayed-OQ 与 OQ(见定理 2).

定理 1　*SKRR* 的平均时延不大于一个 delayed-OQ 的平均时延加上一个常数 NK.

证明　根据 OQ 的性质[4],delayed-OQ 具备性质

$$D_{ij}^{DOQ}(t+\lambda) \geqslant A_{ij}(t) \tag{1}$$

其中 λ 是平均时延,$D_{ij}^{DOQ}(t)$ 是输出端口的 *cell* 数量.类似地,对于任意输入端口,我们可以有 $B_{ij}^k(t+K\cdot IQ_i^k(t)) \geqslant A_{ij}^k(t)$.故得到

$$\sum_k B_{ij}^k(t+K\cdot IQ_i^k(t)) \geqslant \sum_k A_{ij}^k(t) \tag{2}$$

以输入端口的视角来看,交换层和输出端口可以被认为是一个 delayed-OQ.由式(1) 我们得到

$$D_{ij}(t+\lambda) \geqslant \sum_k B_{ij}^k(t) \tag{3}$$

由式(2) 和式(3) 我们能够得到.因为 $D_{ij}(t+K\cdot IQ_i^k(t)+\lambda \geqslant \sum_k A_{ij}^k(t)$.因为 $IQ_i^k(t) \leqslant N$[3] 以及 $D(t)$ 是非递减的函数,我们最后得到下式

$$D_{ij}(t+NK+\lambda) \geqslant D_{ij}^{DOQ}(t+\lambda) \tag{4}$$

结论得证.

定理 2　SKRR 的平均时延不大于一个标准 OQ 的平均时延加上一个常数 2NK.

证明　首先我们比较标准 OQ 和 delayed-OQ 的平均时延.根据其性质,OQ 结构满足 $D_{ij}^{OQ}(t+\tau) \geqslant A_{ij}(t)$,其中 τ 为平均时延,$D_{ij}(t)$ 为输出端口的 cell 数量.我们先计算 VIQ_1 的平均时延,因为如第 2 节分析,离开 VIQ_1 的 cell 要保持它们抵达输入端口的先后顺序,所以 VIQ_1 必须补偿输入端口排队导致的时延差异[3,9].又因为输入端口的最大时延为 NK 系统时钟[9],所以 VIQ_1 的最大时延为 *NK* 系统时钟,故我们可得

$$C_{ij}^k(t+NK) \geqslant B_{ij}^k(t) \tag{5}$$

另外,SKRR 的所有输出端口可被当作一个 OQ 结构,因此我们有

$$D_{ij}(t+\tau) \geqslant \sum_k C_{ij}^k(t) \tag{6}$$

由式(5) 和式(6),并且把 VIQ_1 和 VIQ_2 当作 delayed-OQ,我们得到 $D_{ij}^{DOQ}(t+NK+\tau) \geqslant \sum_k B_{ij}^k(t)$.所以 OQ 和 delayed-OQ 的关系为

$$D_{ij}^{DOQ}(t+NK+\tau) \geqslant D_{ij}^{OQ}(t+\tau) \tag{7}$$

由式(4) 和式(7),我们可以得出 SKRR 和 OQ 的平均时延差异的最大值为 2NK 系统时钟,结论得证.

现在我们来估算 SKRR 在高速路由器中的时延代价.考虑一个 16 端口的 PS 交换机,端口线速率为 OC768(40Gbps),内部加速比为 1,并且 N= K,cell 的长度取为 64 字节.根据定理 2 的结论,该 PS 交换机的平均时延比一个标准的 OQ 交换机最大高出 $2NK = 2\cdot 16^2$ 个时钟,每个时钟约为 13ns,则 2NK 个时钟为 6.7μs,该时延性能满足商业标准.

4.2　吞吐率

本文中,如果交换结构 A 的总排队队长 $Q^A(t)$ 和交换结构 B 的队长 $Q^B(t)$ 满足 $Q^B(t) \leqslant Q^A(t) \leqslant Q^B(t)+C$($C$ 是常数),那么我们认为这两个交换结构的吞吐率相同[12,18].与文献[12] 采用复杂计算从而得出严格上确界不同,我们在定理 3 中给出一个宽松的界值.

定理 3　SKRR 和 OQ 具有相同的吞吐率.

证明　由定理 2 的证明中我们知道 $VIQ_1(i,k,j) \leqslant N$,再加之 $IQ_i^k(t) \leqslant N$[3],因此我们得到 $Q^{SKRR}(t) = \sum_i K\cdot IQ_i^k(t) + \sum_i\sum_k\sum_j VIQ_1(i,k,j) + Q^{OQ}(t) < KN^2 + KN^3 + Q^{OQ}(t)$

又因为定理 2 有 $Q^{OQ}(t) \leqslant Q^{SKRR}(t)$,根据吞吐率的定义,结论得证.

因为 OQ 交换结构的吞吐率具备很多重要的性质,定理 3 非常有用,例如,OQ 在 admissible Bernoulli i.i.d. 流量条件下吞吐率为 100%,所以由定理 3,SKRR 也有该性质.

4.3　性能对比

我们比较了 OQ-PPS[3]、两级交换[12] 和本文提出的 VIQ&SKRR 三种交换结构的性能,性能指标选取为是否支持包保序、系统所需缓存数量、相对时延(relative delay)[3] 和吞吐率.简便起见,我们默认 $K = N$,且加速比为 1.

如表 1 所示,VIQ&SKRR 比 OQ-PPS 多需要 N^3-N^2 个缓存,这些额外需要的缓存主要用来支持 cell 保序,不过,缓存的增加并没有加大 VIQ&SKRR 的时延,它们具有相同的时延能力.此外,与两级交换相比,虽然它们都支持保序,但是 VIQ&SKRR 多需要 N^2 个缓存,这些额外多出的缓存主要用来增加并行处理能力,所以 VIQ&SKRR 的时延比两级交换少 2N2 系统时钟.

表 1　交换结构的性能比较

交换结构	支持保序	缓存数量	相对时延	吞吐率
OQ-PPS	否	$3N^2$	$2N^2$	与 OQ 相同
两级交换	是	N^2+N^3	$4N^2-2$	与 OQ 相同
VIQ & SKRR	是	$2N^2+N^3$	$2N^2$	与 OQ 相同

5 结论

PS 结构充分利用了计算机技术中的并行处理原理, 具备良好的可扩展性和负载平衡能力, 但是包保序问题严重阻碍着 PS 结构的工程实现, 是一个未被很好解决的公开难题. 本文提出一种新技术来解决这个问题, 该新技术包含: 对失序 cell 进行重组的两级 VIQ 缓存结构, 以及保证 cell 保序和公平性的高效调度算法 SKRR. 理论分析表明, 应用该技术的 PS 结构和 OQ 结构具有相同的吞吐率, 并且其平均时延最多比 OQ 结构高出 2NK 系统时钟. 这种新技术的代价是需要较大数量的缓存并且缓存管理机制较复杂, 我们在下一步研究工作中将考虑采用共享缓存, 从而降低复杂度和提高此技术的工程应用性.

参考文献:

[1] S Iyer, A Awadallah, N McKeown. Analysis of a packet switch with the memories running slower than the line rate[A]. Proc of IEEE INFOCOM 2000, vol 2[C]. Tel Aviv, Israel: IEEE, 2000. 529- 537.

[2] Wang W, Dong L, Wolf W. A distributed switch architecture with dynamic load-balancing and parallel input-queued crossbars for terabit switch fabrics[A]. Proc of IEEE INFOCOM 2002[C]. New York, USA: IEEE, 2002. 352- 361.

[3] S Iyer, N McKeown. Making parallel switches practical[A]. Proc of IEEE INFOCOM 2001, vol 3[C]. Anchorage, Alaska: IEEE, 2001. 1680- 1687.

[4] Indra Widjaja, Anwar I Elwalid. Exploiting parallelism to boost datapath rate in high-speed IP/MPLS networking[A]. Proc of IEEE INFOCOM 2003, vol 1[C]. San Francisco California, USA: IEEE, 2003. 566 - 575.

[5] J C R Bennett, C Partridge, N Shectman. Packet reordering is not pathological network ehavior[J]. IEEE/ACM Transactions on Networking, 1999, 7(6): 789- 798.

[6] E Blanton, M Allman. On making TCP more robust to packet reordering [J]. ACM Computer Communication Review, 2002, 32(1): 20- 30.

[7] ITU-T Recommendation I. 363. 5. B-ISDN ATM Adaptation Layer specification: Type 5 AAL[S]. Aug. 1996.

[8] H Jonathan Chao, Kung-Li Deng, Zhi-gang Jing. A petabit photonic packet switch (P3S)[A]. Proc of IEEE INFOCOM 2003, vol 1[C]. San Francisco California, USA: IEEE, 2003. 775- 785.

[9] S Iyer, N McKeown. Analysis of the Parallel Packet Switch Architecture [J]. IEEE/ACM Transactions on Networking, 2003, 11(2): 314- 324.

[10] A Aslam, K Christensen. Parallel packet switching using multiplexors with virtual input queues[A]. 2002 Proc of IEEE LCN[C]. Tampa, Florida, USA: IEEE, 2002. 270- 277.

[11] A Aslam, K Christensen. A parallel packet switch with multiplexors containing virtual input queues[J]. Computer Communications, 2004, 27(3): 1248- 1263.

[12] I Keslassy, N McKeown. Maintaining packet order in two-stage switches [A]. Proc of IEEE Infocom 2002, vol 2[C]. New York, USA: IEEE, 2002. 1032- 1041.

[13] Yue Chen, Yuguo Dong et al. A packet-order-keeping demultiplexer in parallel-structure router based on flow classification[A]. The 2003 International Conference on Computer Networks and Mobile Computing [C]. Shanghai, China: IEEE 2003. 415- 418.

[14] Yuguo Dong, Peng Yi, Yunfei Guo. Analysis of the stable working for the buffered PPS[A]. AINA 2003[C], Xi' an, China: IEEE, 2003. 252 - 256.

[15] Yuguo Dong, Peng Yi, Yunfei Guo. Analysis and designing of the stable parallel packet switch[A]. PDCAT 2003[C]. Chengdu, China: IEEE 2003. 296- 300.

[16] Yuguo Dong, Zupeng Li, Yufei Guo. On the load balancing of a parallel switch with input queues[A]. PDCAT 2003[C]. Chengdu, China: IEEE 2003. 301- 305.

[17] Cheng-Shang Chang, Duarr-Shin Lee, Ching-Ming Lien. Load Balanced Birkhoff-von Neumann Switches, Part II: Multi-stage buffering[J]. Computer Communications, 2002, 25: 623- 634.

[18] E Leonardi, M Mellia, F Neri, M A Marsan. On the stability of input-queued switches with speed-up[J]. IEEE/ACM Transactions on Networking, 2001, 9(1): 104- 118.

作者简介:

兰巨龙 男, 1962 年出生于河北张北, 解放军信息工程大学国家数字交换系统工程技术研究中心教授, 博士生导师, 1988 年获西安电子科技大学通信与电子系统专业工学硕士学位, 2001 年获解放军信息工程大学通信与信息系统专业博士学位. 主要研究方向为网络路由理论与技术、并行交换结构和 IPv6 技术等. E-mail: ljl@mail.ndsc.com.cn.

董雨果 男, 1976 年出生于四川雅安, 空军工程大学电讯工程学院讲师, 2001 年获空军工程大学电讯工程学院信号与信息处理专业硕士学位, 2004 年获解放军信息工程大学信息工程学院通信与电子专业博士学位. 主要研究方向为并行交换结构和网络安全等.

陈 越 男, 1965 年出生于河南开封, 解放军信息工程大学副教授, 硕士生导师, 1990 年获国防科技大学计算机软件专业工学硕士学位, 解放军信息工程大学通信与信息系统专业在读博士研究生, 主要研究方向为网络路由理论与技术、数据库技术等.

温建华 男, 1970 年出生于重庆江津, 解放军信息工程大学国家数字交换系统工程技术研究中心讲师, 1992 年获解放军信息工程学院计算机工程学士学位, 主要研究方向为软件工程质量管理和 IPv6 技术等.