文章编号:1007-5321(2018)06-0110-05 **DOI**:10.13190/j.jbupt.2017-249

一种脉动反馈型两级交换结构

申志军, 高 静, 郭玉波, 白云莉, 李宏慧
(内蒙古农业大学 计算机与信息工程学院, 呼和浩特 010018)

摘要: 为解决反馈型两级交换结构(FTSA)对调度算法的时间限制问题,提出了一种脉动反馈型两级交换结构(PFTSA). PFTSA 将调度算法所需信息以脉动的形式反馈至输入端口, 通过预处理机制使调度算法获得目标缓存的准确信息, 从而避免信元冲突和信元失序. 相对于现有方案, PFTSA 简化了交换结构和交换流程, 同时提高了时延性能.

关 键 词: 包交换; 交换结构; 调度; 反馈机制
中图分类号: TN929.53 **文献标志码**: A

A Pulsating Feedback-Based Two-Stage Switch Architecture

SHEN Zhi-jun, GAO Jing, GUO Yu-bo, BAI Yun-li, LI Hong-hui
(College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China)

Abstract: To solve the time constraint of scheduling algorithm in the feedback-based two-stage switch architecture (FTSA), a new scheme called pulsating feedback-based two-stage switch architecture (PFTSA) is proposed, which transmits the required information back to the input port in a way of pulsating. The accurate data of target buffers can be obtained by the scheduling algorithm with a preprocessing scheme, so as to avoiding the cell conflicting and disordering. As compared to the existing schemes, PFTSA can not only simplify the switch architecture and procedure, but also improve the delay performance.

Key words: packet switching; switch architecture; scheduling; feedback mechanism

在网络视频业务、云计算等新型网络业务和新兴技术的驱动下, 终端处的数据传输压力越来越大. 密集波分复用技术使单根光纤的数据传输带宽达到 10 TB, 这使交换能力相对较弱的中继系统成为当今 Internet 的数据传输瓶颈. 中继设备交换性能关键取决于其交换结构. 在这一领域, 业界的研究重点在于输入缓存(IQ, input queued)[1-2]、输入和交叉点缓存(CICQ, combined input-crosspoint-queued)[3-4]、Clos[5-6]及负载均衡交换结构[7-14]. 其中, 负载均衡交换结构的多级 crossbar 能够将到达输入端的突发数据流均匀散布, 因此能够较好地适应 Internet 自相似数据流. Hu 和 Yeung[9]提出的反馈型两级交换结构(FTSA, feedback-based two-stage switch architecture)巧妙地利用错列对称的 crossbar 连接方式, 使输入端口能通过反馈提前获得目标缓存数据, 从而实现精准的信元调度, 获得极为优异的时延性能. FTSA 是迄今为止理论性能最优的负载均衡交换结构,FTSA 的缺陷在于其输入端口的算法必须在 crossbar 重配置时间内完成调度. crossbar 重配置时间本质上取决于交换

收稿日期: 2017-11-27
基金项目: 内蒙古农业大学优秀青年科学基金项目(2014XYQ-17); 国家自然科学基金项目(61650204, 61462070); 内蒙古自治区自然科学基金项目(2018MS06013)
作者简介: 申志军(1976—), 男, 副教授, 硕士生导师, E-mail: shensljx@sina.com.

芯片元器件的开关速度，当前芯片频率已达到 GHz，即芯片元器件的工作周期为 ns 级．另一方面，现有调度算法如最长队列优先（LQF，longest queue first）、最早离开者优先（EDF，earliest departure first）和轮询算法（RR，round-robin）的算法复杂度均为 $O(N)$．在 ns 级时间之内完成复杂度为 $O(N)$的调度算法是不现实的．实践中因调度耗时超出限定的时间区间而不得不延迟信元传输的开始时间，以等待算法调度的结果，使 FTSA 优异的理论性能无法实现．

针对这一问题，利用脉动反馈机制为调度算法提供接近一个时隙的执行时间，有效解决了其对调度算法的时间限制问题，提高了实践可行性．

1　相关工作

解决 FTSA 的时间限制问题有两种思路：前置反馈交换结构（FFTS，front-feedback-based two-stage switch architecture）[13] 和使用 2-错列对称的改进方案（FTSA-2-SS，FTSA using 2-staggered symmetry connection pattern）[14]．

FFTS 和 FTSA-2-SS 的交换结构相同，如图 1 所示，其两级 crossbar 分别记为 X_1 和 X_2，位于 X_1 之前的缓存记为 VOQ1，任意 VOQ1(i,k)用于缓存到达输入端口 i 且目标端口为 k 的信元；位于 X_1 和 X_2 之间的缓存记为 VOQ2，任意 VOQ2(j,k)用于缓存到达中间端口 j 且目标端口为 k 的信元，任意 VOQ2(j,k)设置 2 个信元的缓存空间．输出端口设置重排序缓存来纠正信元离开交换机的顺序．

FFTS 和 FTSA-2-SS 均能在一定程度上解决算法执行时间不足的问题，但付出了巨大的代价．

1）交换结构复杂化．二者都只能简单地依据不完整的信息进行调度．可能导致“信元冲突”[13-14] 问题，故二者均需为任意 VOQ2(j,k)设置“第二信元空间”．第二信元空间虽然可临时缓存发生冲突的信元，但也丧失了原本的信元保序机制，故 FFTS 和 FTSA-2-SS 均需在输出端设置重排序缓存来纠正信元离开交换机的顺序．

2）时延性能恶化．信元在第二信元空间和重排序缓存中的等待时间也恶化了时延性能，仿真实验表明，在相同的条件下，二者的时延性能相对于 FTSA 均有明显的下降．

提出的脉动反馈型两级交换结构（PFTSA，pulsating feedback-based two-stage switch architecture）采用一种前置的脉动反馈机制使得输入端口能够提前获得准确的目标缓存状态数据，既能扩展算法的执行时间区间，又能避免信元冲突和信元失序，无需设置第二信元空间和重排序缓存．

2　脉动反馈型两级交换结构

PFTSA 采用如图 1 所示的交换结构，其两级 crossbar 采用错列对称方式[9]，其任意 VOQ2(j,k)仅设置一个信元的缓存空间．

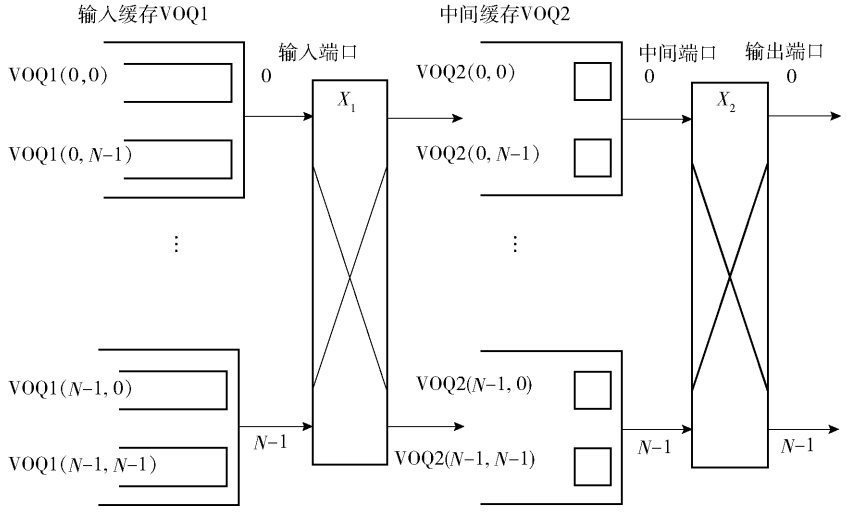

图 1　PFTSA 交换结构

为便于表述，做如下约定：

1）交换结构的输入/输出端口数记为 N，端口号的加减操作实际都要对 N 取模，即 $i-1$ 实质上是$(i-1) \bmod N$.

2）输入端口 i 记为 I_i，中间端口 j 记为 M_j，输出端口 k 记为 O_k.

3) I_i与M_j相连记为$I_i \rightarrow M_j$, M_j与O_k相连记为$M_j \rightarrow O_k$, I_i通过M_j与O_k相连记为$I_i \rightarrow M_j \rightarrow O_k$.

4) M_j在t时隙开始时刻的缓存状态数据记为$Q_j(t^b)$, M_j在t时隙结束时刻的缓存状态数据记为$Q_j(t^e)$, $Q_j(t^b)$和$Q_j(t^e)$都仅有Nbit, 若其第v位为"1"表示VOQ2(j,v)非空, 反之表示VOQ2(j,v)为空.

5) I_i在t时隙之初向中间端口传输Nbit的$P_i(t)$, $P_i(t)$最多只有1 bit为"1", 其第v位为"1"表示I_i将在t时隙向中间端口传输VOQ1(i,v)中的信元, $P_i(t)=0$表示I_i在t时隙不向中间端口传输信元.

6) crossbar重配置时间记为T_R, 沿X_1或X_2传送Nbit的发送和传播时延之和记为T_N;输出端口将Nbit的数据反馈至位于同一线卡的输入端口的时间记为T_F, 输入端口进行一次预处理的处理时延记为T_P;信元在X_1或X_2上的传输时延和传播时延之和记为T_C. 因T_R、T_N、T_F、T_P等均耗时极短, 故记$T_{STD}=\max(T_R, T_N, T_F, T_P)$.

2.1 脉动反馈机制

不失一般性, 不妨设t时隙有$I_i \rightarrow M_j \rightarrow O_k$, 则由错列对称连接方式的特性可知:$t+1$时隙必有$I_k \rightarrow M_j$. 以$t$时隙$I_k$的算法所需信息的传输过程为例, 说明脉动反馈机制的工作方法.

1) 0时刻, I_i向M_j传输$P_i(t)$;M_j向O_k传输本地缓存状态数据$Q_j(t^b)$,如图2(a)所示.

2) T_{STD}时刻, $Q_j(t^b)$到达O_k;$P_i(t)$到达M_j;O_k向I_k反馈$Q_j(t^b)$;M_j向O_k传输$P_i(t)$;I_i开始向M_j发送信元,如图2(b)所示.

3) $2T_{STD}$时刻, $Q_j(t^b)$到达I_k;$P_i(t)$到达O_k;I_k对$Q_j(t^b)$进行预处理(具体处理方法见2.2节);O_k向I_k传输$P_i(t)$;M_j向O_k发送信元,如图2(c)所示.

4) $3T_{STD}$时刻, $Q_j(t^b)$预处理完成, 结果记为$Q_j(t^{be})$;$P_i(t)$到达I_k;I_k开始对$Q_j(t^{be})$进行预处理(具体处理方法见2.2节),如图2(d)所示.

5) $4T_{STD}$时刻, $Q_j(t^{be})$预处理完成, 结果即为$Q_j(t^e)$;I_k开始进行调度,如图2(e)所示.

6) T_C+T_{STD}时刻, 信元到达M_j, 如图2(f)所示.

7) T_C+2T_{STD}时刻, 信元到达O_k,如图2(g)所示.

8) T_C+3T_{STD}时刻, crossbar重配置完成, 算法调度结束, $t+1$时隙开始,如图2(h)所示.

脉动反馈机制的核心在于t时隙之初,$P_i(t)$依次经M_j和O_k反馈至I_k. $P_i(t)$是PFTSA能够避免信元冲突和信元失序问题的关键. 因为通过$P_i(t)$, PFTSA可得到M_j在t时隙结束时刻其缓存状态的准确数据, 即$Q_j(t^e)$. 基于准确的$Q_j(t^e)$进行"有的放矢"的调度, 不会出现"信元冲突"问题, 自然无

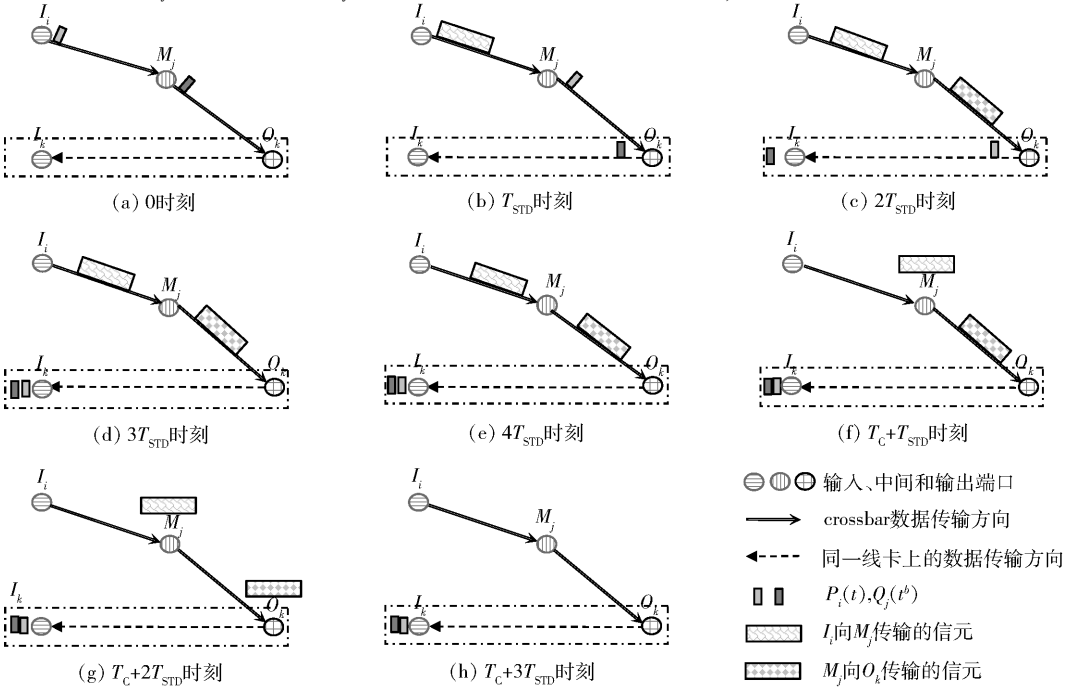

图2 PFTSA的脉动反馈机制

需设置第二信元空间，没有第二信元空间自然不会导致信元失序问题(文献[9]已证明)，因此 PFTSA 不再需要第二信元空间和重排序缓存.

2.2　信息预处理过程

M_j在 t 时隙开始和结束时刻的缓存状态数据至多只可能有 2 个 bit 的不同.

首先考虑 t 时隙离开 M_j的信元，因 t 时隙有 $M_j \rightarrow O_k$，故若有信元离开，则必属于 VOQ2(j,k). 因此预处理方法如下(处理结果记为 $Q_j(t^{be})$)：

$$Q_j(t^{be}) \leftarrow Q_j(t^{b}) \& (\sim(1 \ll (k)))$$

其次考虑 t 时隙到达 M_j的信元，因 $P_i(t)$ 中已经包含了这一信息，故只需：

$$Q_j(t^{e}) \leftarrow Q_j(t^{be}) \mid P_i(t)$$

处理得到的 $Q_j(t^e)$ 正是 PFTSA 中 I_k在 t 时隙的调度算法所需要的准确目标缓存状态数据.

2.3　调度算法

PFTSA 中调度算法依然采用 Hu 和 Yeung[9] 提出的 LQF、EDF 和 RR 算法进行调度.

3　相关分析

3.1　算法执行时间的扩展效果

基于 2.1 节分析可知，PFTSA 中调度算法的执行时间为 $T_C - T_{STD}$，运用相同的分析方法，FTSA 中调度算法的执行时间不足一个 $T_{STD}(T_R - T_F)$，FFTS 中调度算法的执行时间为 $T_C + T_{STD}$，FTSA-2-SS 中调度算法的执行时间为 $T_C + 2T_{STD}$.

考虑到 $T_C \gg T_{STD}$，相对于 FTSA，PFTSA 能够大幅度地改善调度算法的执行时间限制问题.

3.2　交换结构和流程的简化效果

相对 FFTS 和 FTSA-2-SS 而言，PFTSA 无需设置第二信元空间和重排序缓存.

1) PFTSA 在每个中间端口节约 N 个信元空间，全部 N 个中间端口共节约 N^2个信元空间.

2) 申志军等[13]已证明 FFTS(FTSA-2-SS 类似)需为每个输出端口设置 N 个信元的重排序缓存，亦即 PFTSA 共节约 N^2个信元空间.

3) 考虑 $N=32$ 且每时隙转发 256 Byte，PFTSA 可节约 2^{19} Byte 的存储空间.

3.3　时延性能

因为信元无需在第二信元空间和重排序缓存中等待，PFTSA 的时延性能优于 FFTS 和 FTSA-2-SS. 但 PFTSA 的每次调度都是提前进行，调度开始时尚无法得知本时隙到达本地输入端口的信元信息，亦即无法转发本时隙到达输入端口的信元，故相对于 FTSA 而言，其时延性能必有所损失.

不失一般性，以 I_k在 t 时隙的调度算法为例，当且仅当同时满足下列 3 个条件时，对 PFTSA 的时延性能产生显著负面影响的事件才会发生.

1) t 时隙有信元 p 到达 VOQ1(i,v).

2) t 时隙 p 是 VOQ1(i,v) 中唯一的信元.

3) t 时隙 p 是唯一满足算法需求的信元.

从发生的条件来看，导致 PFTSA 时延性能明显损失的概率较低，故其时延性能损失较为有限.

3.4　代价

PFTSA 的优势本质上在于以下两方面的折中：

1) PFTSA 将调度算法执行时间扩展到 $T_C - T_{STD}$，大幅度缓解了算法的时间限制问题.

2) 简化了交换流程和结构，提高了时延性能.

PFTSA 的代价在于如下 2 个方面：

1) PFTSA 对调度算法执行时间的扩展效果稍逊于 FFTS 及 FTSA-2-SS. 但考虑到 $T_C \gg T_{STD}$，故其对算法执行时间的扩展效果仍然是较为可观的.

2) PFTSA 每时隙都会比 FTSA 多传输 N bit，传输效率会降低. 但若以 $N=32$，每时隙传输 256 Byte 为例，效率损失为 1.5%，其代价仍可承受.

4　仿真结果

理论上 FFTS 和 FTSA-2-SS 在相同的交换环境中具有相同的时延性能，故仅对 FTSA、FFTS 和 PFTSA 三种交换结构进行仿真.

基于 Opnet14.5，使用 32×32 的交换模型分别仿真上述 3 种结构在均匀数据流环境、突发数据流环境和 hot-spot 数据流环境中的平均时延.

均匀数据流环境预设信元独立且以伯努利分布到达输入端口，信元的目的输出端口服从均匀分布. 均匀数据流常用于模拟传统的交换环境.

突发数据流环境预设信元以突发的形式到达输入端口，每个突发块中的信元具有相同的目的输出端口. 突发数据流常用于模拟自相似交换环境.

hot-spot 数据流环境预设信元独立且以伯努利分布到达输入端口，其中 1/3 的信元具有相同的目的输出端口，2/3 的信元具有相同的另一个目的输出端口. hot-spot 数据流常用于模拟不规则的数据流交换环境.

3 种环境中的仿真结果分别如图 3～图 5 所示.

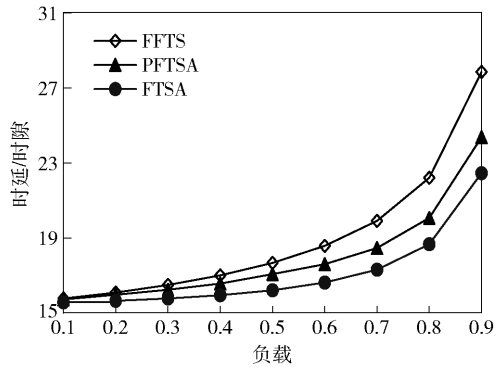

图3 均匀数据流环境中的时延比较

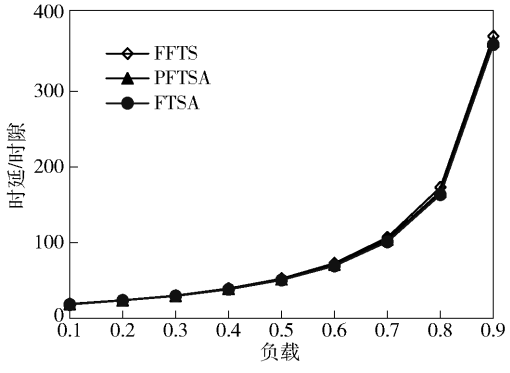

图4 突发数据流环境中的时延比较

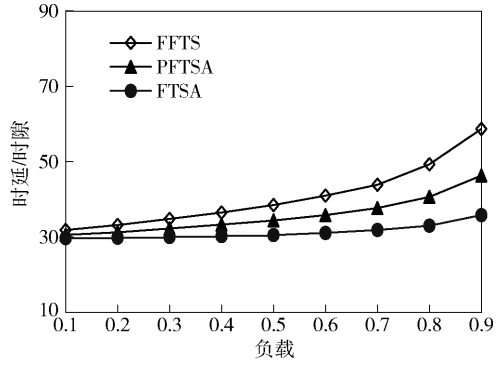

图5 Hot-Spot 数据流环境中的时延比较

图3表明,均匀数据流环境下 FTSA、FFTS 和 PFTSA 的时延性能均较为优异,其中 FTSA 的时延性能最好,PFTSA 的时延性能优于 FFTS,这是因为信元无需第二信元空间和重排序缓存中等待. PFTSA 的时延性能略逊于 FTSA,原因是其提前调度的策略无法转发当前时隙到达的信元.

图4表明,在突发数据流环境中,FTSA、FFTS 和 PFTSA 的时延均大幅度上升,但 PFTSA 的时延性能仍然符合3.3节的性能预期.

图5表明,3种结构均能较好地适应 hot-spot 这种不规则的数据流交换环境,PFTSA 的平均时延介于 FTSA 和 FFTS 之间,依然符合预期.

5 结束语

PFTSA 通过脉动的形式使得输入端口可提前获得调度算法所必需的准确数据. 基于这种机制,PFTSA 得以简化交换结构和交换流程,同时提高了系统的时延性能. 下一步研究将从降低算法耗时的角度来进一步缓解反馈型两级交换结构对调度算法的时间限制问题,使之能够更好地适应未来的高速交换环境.

参考文献:

[1] Xiao Jie, Yeung K L, Jamin S. Pipelined scheduler for unicast and multicast traffic in input-queued switches[C]// 2016 IEEE Global Communications Conference. New York: IEEE Press, 2016: 1-6.

[2] Hu Bing, Yeung K L, Zhou Qian, et al. On iterative scheduling for input-queued switches with a speedup of 2-1/*N*[J]. IEEE/ACM Transactions on Networking, 2016, 24(6): 3565-3577.

[3] Jin Hao, Pan Deng, Liu Jason, et al. OpenFlow based flow level bandwidth provisioning for CICQ switches[J]. IEEE Transactions on Computers, 2013, 62(9): 1799-1812.

[4] 王斌, 王文鼐. 高性能组合输入交叉点排队交换机[J]. 北京邮电大学学报, 2012, 35(1): 95-98.
Wang Bin, Wang Wennai. High performance CICQ fabric [J]. Journal of Beijing University of Posts and Telecommunications, 2012, 35(1): 95- 98.

[5] 高雅, 邱智亮, 张茂森, 等. Clos 交换网络中随机化的加权匹配调度算法[J]. 北京邮电大学学报, 2013, 36(4): 90-94.
Gao Ya, Qiu Zhiliang, Zhang Maosen, et al. Randomized weight matching dispatching scheme for Clos-network switches[J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(4): 90-94.

[6] Xia Yu, Hamdi M, Chao H J. A practical large capacity three stage buffered Clos network switch architecture[J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(2): 317-328.

[7] Chang C S, Lee D S, Jou Y S. Load balanced Birkhoff-von Neumann switches[C]// 2001 IEEE Workshop on High Performance Switching and Routing. New York: IEEE Press, 2001: 276-280.

(下转第136页)

energy efficiency tracking circuits for converter-less energy harvesting sensor nodes[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2017, 64(6): 670-674.

[2] Huang Liang, Bi Suzhi, Qian Liping. Optimal threshold-based transmission scheduling policy for energy harvesting sensor nodes[C]// 2016 IEEE Global Communications Conference. New York: IEEE Press, 2016: 1-6.

[3] Sunny A. Joint scheduling and sensing allocation in energy harvesting sensor networks with fusion centers[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(12): 3577-3589.

[4] Yang Shusen, Tahir Y, Chen P Y, et al. Distributed optimization in energy harvesting sensor networks with dynamic in-network data processing[C]//The 35th Annual IEEE International Conference on Computer Communications. New York: IEEE Press, 2016: 1-9.

[5] Cai Songfu, Lau V K N. MIMO precoding for networked control systems with energy harvesting sensors[J]. IEEE Transactions on Signal Processing, 2016, 64(17): 4469-4478.

[6] Gong Jie, Zhou Sheng, Niu Zhisheng. Optimal power allocation for energy harvesting and power grid coexisting wireless communication systems[J]. IEEE Transactions on Communications, 2013, 61(7): 3040-3049.

[7] Yang Jian, Yang Qinghai, Kwak K S, et al. Power-delay tradeoff in wireless powered communication networks[J]. IEEE Transactions on Vehicular Technology, 2017, 66(4): 3280-3292.

[8] Zhou Xun, Ho C K, Zhang Rui. Wireless power meets energy harvesting: a joint energy allocation approach in OFDM-based system[J]. IEEE Transactions on Wireless Communications, 2016, 15(5): 3481-3491.

[9] Mandjes M. Large deviations for Gaussian queues: modeling communication networks [M]. Hoboken: Wiley, 2007: 25.

[10] Gardner E S. Exponential smoothing: the state of the art-Part II[J]. International Journal of Forecasting, 2006, 22(4): 637-666.

[11] Blasco P, Gunduz D, Dohler M. A learning theoretic approach to energy harvesting communication system optimization[J]. IEEE Transactions on Wireless Communications, 2013, 12(4): 1872-1882.

(上接第114页)

[8] Shen Yanming, Panwar S S, Chao H J. Design and performance analysis of a practical load-balanced switch[J]. IEEE Transactions on Communications, 2009, 57(8): 2420-2429.

[9] Hu Bing, Yeung K L. Feedback-based scheduling for load-balanced two-stage switches [J]. IEEE/ACM Transactions on Networking, 2010, 18(4): 1077-1090.

[10] Cai Yan, Wang Xiaolin, Gong Weibo, et al. A study on the performance of a three-stage load-balancing switch [J]. IEEE/ACM Transactions on Networking, 2014, 22(1): 52-65.

[11] Durkovic S, Cica Z. Birkhoff-von Neumann switch based on greedy scheduling[J]. IEEE Computer Architecture Letters, 2018, 17(1): 13-16.

[12] Huang An, Hu Bing. The optimal joint sequence design in the feedback-based two-stage switch[J]. Journal of Network and Computer Applications, 2014, 45(4): 27-34.

[13] 申志军, 曾华燊, 夏羽. 基于前置反馈的两级交换结构[J]. 通信学报, 2011, 32(5): 56-62.
Shen Zhijun, Zeng Huashen, Xia Yu. Front feedback-based two-stage switch architecture [J]. Journal on Communications, 2011, 32(5): 56-62.

[14] 申志军, 曾华燊, 高志江. 一种改进的反馈制两级交换结构 FTSA-2-SS[J]. 电子与信息学报, 2011, 33(6): 1319-1325.
Shen Zhijun, Zeng Huashen, Gao Zhijiang. An improved feedback-based two-stage switch architecture using 2-staggered symmetry connection pattern [J]. Journal of Electronics & Information Technology, 2011, 33(6): 1319-1325.