# 基于 SRAM 的核心路由器交换矩阵输入端口设计

陈腾飞，李开航
（厦门大学 物理与机电工程学院，福建 厦门 361005）

**摘 要**：交换矩阵是核心路由器的重要组成部分，为了避免来自不同输入端口的信元同时发往同一个输出端口，需要在输入端口设置缓冲区，即输入排队交换结构。基于静态随机存储器完成了交换矩阵输入端口虚拟输出队列（VOQ）的设计，该设计可以降低核心路由器交换芯片的面积，提高输入端口缓冲区信元的响应速率，并通过 DE-115 开发板完成对设计的验证。

**关键词**：交换矩阵；输入端口；VOQ；SRAM
**中图分类号**：TN911-34 **文献标识码**：A **文章编号**：1004-373X(2012)16-0119-03

## Design of crossbar input port of core router based on SRAM

CHEN Teng-fei，LI Kai-hang
（School of Physics and Mechanical& Electrical Engineering，Xiamen University，Xiamen 361005，China）

**Abstract**：Crossbar is an important part of core router. Some packet buffers need to be set in the input port to avoid the cells coming from different input ports to be sent to a same output port. The design of virtual output queue (VOQ) in crossbar input port was fulfilled based on the static random access memory (SRAM). It can obviously reduce the area of exchange chip in core router and improve the reponding speed of input port buffer. The design was verevied with DE-115 development board.

**Keywords**：crossbar；input port；VOQ；SRAM

## 0 引 言

随着光纤通信技术的飞速发展，路由器的数据处理速度成为网络通信的主要瓶颈，交换矩阵作为核心路由器的重要组成部分则严重制约了路由器的传输速率。目前核心路由器交换结构使用较多的有共享内存和 Crossbar 两种。共享内存结构通过共享输入和输出端口存储器件，减少了对总体存储空间的需求。共享内存结构相对简单，交换效率可根据需求不断优化。共享内存交换结构的交换性能取决于共享内存的存取速率，可扩展性较差，尤其当板卡端口数量较多时，交换效率有所下降。

Crossbar 是一种严格的非阻塞交换结构，输入/输出之间可建立多条通路。Crossbar 采用连接式，即 $N\times N$ 的交叉矩阵。Crossbar 使用调度器，根据各输入点相关的信息，运算调度算法得到输入和输出之间的一个匹配，并配置相应交叉点。调度器的效率非常关键，决定了 Crossbar 的交换速率[1-3]，因此调度算法必须高度完善。但 Crossbar 同样存在扩展性的问题，即交换矩阵的交叉点会随着输入/输出数量的增多呈指数增长。为维持无阻塞交换，需不断完善和改进调度算法，代价是开发的技术成本越来越高，核心交换芯片的面积也越来越大。另外，Crossbar 也同样不能避免排队仲裁，传输效率受到一定影响和限制。但相比共享内存结构，Crossbar 效率和扩展性都比较好[4]，目前大部分高端路由器都使用 Crossbar 交换结构。

基于静态随机存储器（SRAM）的交换矩阵输入端口虚拟输出队列（VOQ）的设计同时结合了共享内存和 Crossbar 两种交换方式的优点，将输入端口中的数据缓冲区移至片外，用高效地调度算法对虚拟输出队列进行调度，可以有效的减小核心交换芯片的面积，并提高数据报文的读取速率。

## 1 系统总体设计

由于核心路由器交换矩阵硬件实现简单，已经在越来越多的 ATM 交换机和高性能路由器中使用。当输入端口使用单一的 FIFO 排队机制时，HOL（Head of Line）阻塞使得开关吞吐率最多只能利用 58%[5]，因此，在目前输入缓冲的交换设备中，输入端口一般采用 VOQ 虚拟输出队列技术，即每个输入端口为到达不同输出端口的信元设置不同的 FIFO 队列。虚拟输出队列技术的采用消除了 HOL 阻塞。

核心路由器交换矩阵主要由三个模块组成，即调度模块，输入模块，输出模块。调度模块主要用来分析输入

收稿日期：2012-03-26

端口的缓存数据报文的目的地址，根据输入端口各个虚拟输出队列的调度请求，使用 iSLIP 调度算法 8 控制输入端口与输出端口之间的连接，防止队列的链头阻塞[6]。输入模块主要是用来将从线卡上接收的数据报文存入不同的基于 SRAM 的虚拟输出队列，同时向调度器发出调度请求，当接收到调度指令后，将报文发往输出端口。输出模块是用来接收输入端口发来的数据报文，并将其重新组合成完整的数据包发送出去，同时给调度器一个反馈指令，交换矩阵的系统框图如图 1 所示。

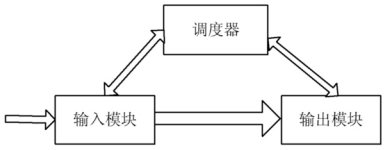


图 1 交换矩阵整体结构

## 2 VOQ 虚拟输出队列设计

影响 Crossbar 交换效率的因素主要是输入排除链头阻塞问题和调度算法的选择。输入排队链头阻塞问题的解决方案就是采用给每个输入到输出建立一个虚拟缓冲队列的输入排队交换内核的体系结构，基本思想是每一个输入端口在其输入缓冲器中为每一个输出端口保存一个先进先出(FIFO)队列。对于 8×8 的交换结构，共有 8×8 个 VOQ。到达输入端口的信元按照它的输出端口，置入相应的 VOQ 队列中。在每个交换时隙，调度器调度所有 VOQ，使得每一个输出端口只有一个 VOQ 接受服务，然后发送其最前端的分组，不仅消除了由 FIFO 队列造成的链头阻塞，更不用考虑设置加速比问题，VOQ 的具体结构如图 2 所示。

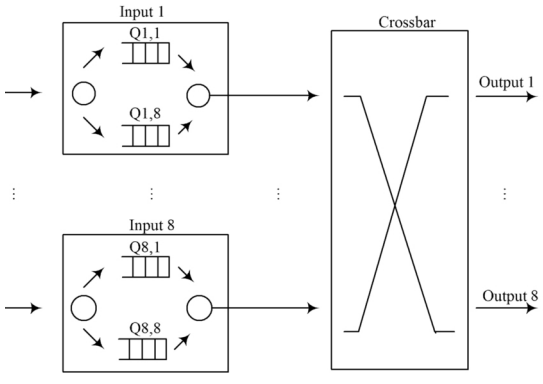


图 2 VOQ 虚拟输出队列设计

VOQ 方式将目的输出端口不同的信元放在不同的队列中缓存，因此发往不同输出端口的信元相互不存在 HOL 阻塞。在某些调度算法下，VOQ 方式可 100% 获得交换开关的利用率。目前 Cisco GSR12000，BBN-MGR 等路由器都采用 VOQ 方式组织输入队列。消除 HOL 阻塞后，交换开关仍存在另外两种阻塞，即输入端口阻塞和输出端口阻塞。由同一输入端口不同 VOQ 队列中的信元竞争输入端口而产生的阻塞称为输入端口阻塞，由不同输入端口的信元竞争同一输出端口而产生的阻塞称为输出端口阻塞。调度器根据各输入端口 VOQ 队列的状态决定 Crossbar 内部的拓扑关系，从而解决上述两种阻塞[7]。系统主要由交换阵列、调度器、输入控制器、输出控制器和 SRAM 组成。输入控制器从线卡接收信元，根据其目的端口号将其存入双端口 SRAM 中，每个输入端口共 8 个队列，分别存放发往不同输出端口的信元。输入端口控制器根据队列的空满情况向调度器发出请求[8]。调度器根据各输入端口的请求公平地分配输出端口，并将调度结果传送到 Crossbar 交换阵列和各输入/输出控制器。输入端口控制器接收到调度结果后，从相应的 VOQ 队列取出一个信元送交换阵列交换。同时输出端口控制器根据调度结果，将接收的信元放入相应的输出端口寄存器中。若输出接口控制器检测到寄存器中有重组完毕的报文，将报文发往相应的线卡中。

## 3 输入端口设计

调度算法的选择和输入排队链头阻塞问题是影响交换矩阵交换速率的关键因素。i-SLIP 调度算法的硬件实现比较简单，并且支持优先级调度，可以很好地满足调度的要求。输入端口 VOQ 队列的设计则可以很好的解决链头阻塞问题，由于输入端口在交换芯片中占据了很大的面积，所以将报文缓冲区移到片外可以显著地降低交换芯片的面积，输入端口的设计如图 3 所示。

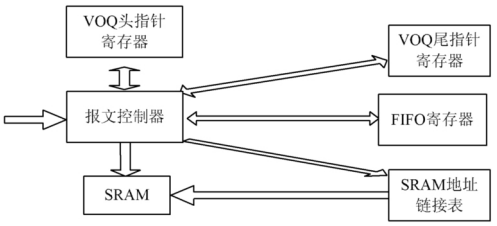


图 3 输入端口控制器设计

从线卡传输到交换网络输入端口的数据包有着固定的长度，它的长度共有72 位，包括 6 位的包头和66 位的包数据，其中包头的前 3 位是源地址，后 3 位是目的地址[9]。当报文控制器接收到从线卡传输来的 72 位的数据包时，便将其存入 SRAM 中的空地址中，FIFO 寄存器是专门用来存放 SRAM 中的空地址，报文控制器根据 FIFO 寄存器的空地址将数据包存入到 SRAM 中，同时更新 SRAM 地址链接表和 VOQ 尾指针寄存器，以便接收下一个数据包。当需要从 SRAM 中读取数据包时，首先根据 VOQ 头指针寄存器找到 SRAM 地址链接表，SRAM 地址链接表中存放的是数据包在

SRAM 中的地址，然后根据 SRAM 地址链接表找到需要从 SRAM 中读取的数据包的地址，从而读取所需要的数据，同时更新 VOQ 头指针寄存器和 SRAM 地址链接表[10]。

由于报文的头尾标志用 2 b 定义，因此具有很好的故障恢复能力。例如因此硬件传输时受到外界干扰，10 标志变成 $n$，这时不需任何例外处理，带来的危害仅仅影响连续的两个报文(两个报文合并成一个)。

## 4　SRAM 读写测试

交换矩阵输入端口的设计取决于能否根据输入端口中 FIFO 寄存器中的空的 SRAM 的地址和 SRAM 地址链接表准确地读取 SRAM 中的数据报文。该输入端口设计以 Atera DE-115 开发板上的 SRAM 芯片为基础，编写 SRAM 的仿真模型，该芯片的存储容量为 2 MB，并在 Modelsim 中完成了对设计的验证。仿真结果如图 4 所示。

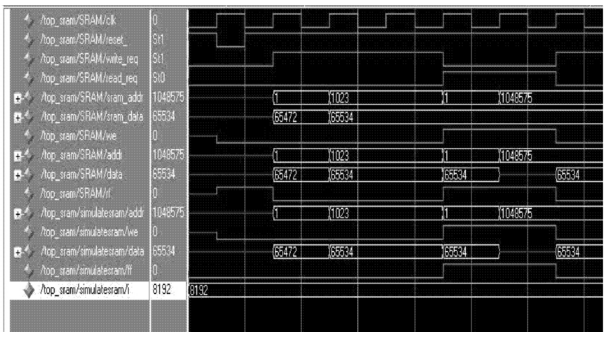


图 4　SRAM 仿真模型测试

## 5　结　语

本文设计了一个基于 SRAM 的交换矩阵的输入端口，该设计有效的消除了输入排队链头阻塞的问题，极大地提高交换开关的利用率，将输入端口数据报文存放在片外 SRAM 中，可以显著降低交换芯片的面积，提高虚拟队列中数据报文的读取速度，并在 Altera 开发板上完成了验证，系统性能稳定，具有很好的应用前景与研究意义。

## 参　考　文　献


[1] 过锋. 高性能计算机互连网路由器研究与实现[D]. 长沙：国防科技大学，2004.

[2] 余鑫，黄本雄，阮加勇. LOARD：分层的开放分布式路由器结构[J]. 计算机科学，2005，32(8)：30-33.

[3] 马丽红，蔡祥宝. 带 VOQ 的输入队列交换网络中的分组调度算法研究[J]. 江西通信科技，2006(2)：12-14.

[4] 左忠卫. 基于 FPGA 的大容量 ATM 交换模块的设计与实现[D]. 西安：西安电子科技大学，2008.

[5] 孙志刚，苏金城，卢锡城. 高效的 Crossbar 仲裁算法[J]. 计算机学报，2000，23(10)：1078-1082.

[6] MEKEOWN Nick，ALLANTHARAM Venkat，WALRAND Jean. Aehieving 100% throughput in an inputqueued switch [C]// Proeeedings of IEEE Infoeom '96. San Francisco，Usa：IEEE，1996：18-25.

[7] 张福刚. 基于高端路由系统中逻辑控制 ASIC 芯片的研究[D]. 上海：复旦大学，2008.

[8] MCKEOWN Nicholas William. Scheduling algorithms for input-queued cell switches [D]. CA，USA：University of California，1995.

[9] 郑燕峰. 基于输入排队的可扩展交换结构调度算法的研究[D]. 北京：中国科学院计算机研究所，2006.

[10] 刘化君，刘斌. iSLIP 调度算法研究与实现[J]. 小型微型计算机系统，2003，24(9)：1593-1596.


作者简介：陈腾飞　男，1985 年出生，河南人，硕士研究生。主要研究方向为数字集成电路设计。
李开航　男，1967 年出生，福建人，副教授。主要研究方向为模拟集成电路设计。

(上接第 111 页)


[6] SHALOM Y B，BIRMIWAL K. Variable dimension filter for maneuvering target tracking [J]. IEEE Trans. on AES，1982，18(3)：621-629.

[7] 刘望生，李亚安. 闪烁噪声下目标跟踪的改进粒子滤波算法[J]. 兵工学报，2011，32(1)：91-95.

[8] YANG Y，HE H，XU G. Adaptively robust filtering for kinematic geodetic positioning [J]. Journal of Geodesy，2001，75：109-116.

[9] 巴欣宏，赵宗贵，杨飞，等. 一种新的机动目标跟踪的多模型算法[J]. 电子与信息学报，2005，27(1)：13-16.


作者简介：闫常浩　男，1986 年出生，辽宁辽阳人，硕士研究生。研究方向为传感器。