共享缓存式以太网交换机缓存结构分析

郑万立， 涂晓东，田永刚

(电子科技大学 宽带光纤传输与通信网技术教育部重点实验室 成都 610054)

【摘要】针对共享缓存式交换机能提供理想的吞吐量、延时性能和对于一定的丢失率共享缓存交换机需要的内存较其他交换机小的特点，结合现有的共享缓存式以太网交换机芯片，对该交换机的缓存物理结构、数据结构进行分析和总结，并提出了一种共享缓存式以太网交换机缓存结构。

关 键 词 缓存管理单元； 共享缓存池； 分片； 描述符； 输出队列； 分级服务

中图分类号 TN915.05 文献标识码 A

Buffer Structure of Shared-Memory Switch

Zheng Wanli，Tu Xiaodong，Tian Yonggang

(Key Laboratory of Brodband Optical Fiber Transmission and Communication Networks UEST of China, Ministry of Education Chengdu 610054)

Abstract Shared-memory is a common structure of Ethernet switches. It can achieve the optimal throughput and delay performance. Furthermore, for a given loss rate, a shared-memory switch requires less buffers than other switches. This paper analyzes and summarizes the physical and data structure of shared-memory Ethernet switches according to existence shared-memory switch chips, and bring forward our new structure.

Key words memory management unit; common buffer pool; cell; description; egress queue; class of service

共享缓存是交换机中比较常用的一种结构。在这种交换结构中，所有的输入和输出端口都共享一个缓存模块，所有需要经过交换机的数据都在缓存中存储转发。该结构的优点是缓存资源可以得到充分使用，可以获得最佳的时延和吞吐率性能。在交换机芯片的硬件实现上，缓存的集中分布，在芯片的布局和布线上比较方便。如果采用外部存储器作为交换机的共享缓存池，外部缓存可以采用普通的PC内存，这样可以降低交换机的成本，适用于不同的规模设计。但因为缓存的集中管理和缓存的共用，缓存的读写访问时间限制了交换机的规模。如果通过增加缓存的线宽来提高速度，又会增加芯片的设计难度[1]。因此，必须有一种合理的缓存结构来管理有限和缓存空间和缓存带宽。

1 缓存管理涉及的基本问题

1.1 分配和管理缓存的存储空间和带宽

对于传输变长分组的以太网交换机，一般也采用分片存储的方式，而不采用连续存储的方式。

分片存储是将变长的数据包分成定长的分片(Cell)存储在缓存中，一个数据包的各个分片分散在缓存的不同位置。如果一个分片没有被数据填满，则该分片的剩余空间也不能被其他数据包使用。连续存储是将

收稿日期：2004 – 07 – 06

作者简介：郑万立(1981 –)，男，硕士生，主要从事以太网交换机芯片方面的研究；涂晓东(1970 –)，男，博士，副教授，主要从事宽带IP网络、光网络、存储网络方面的研究；田永刚(1981 –)，男，硕士生，主要从事以太网交换机芯片方面的研究.

整个变长数据包直接存储在缓存中，所有的数据的存储地址都是连续的。

虽然连续存储方式可以有效地利用所有的缓存空间，但存在的问题是：对于变长数据包的存储会产生很多无法利用的碎片。如果要整理这些碎片将会涉及到比较复杂的逻辑电路，而且在碎片整理过程中，交换机的性能将降低。缓存分片方式虽然不能有效地利用所有的存储空间，但是缓存的管理比较方便，不会产生碎片。

1.2 缓存带宽的分配

共享缓存式交换机实质上是时分复用的，缓存在一个时钟周期只能允许一个端口读或写。所以，必须最大限度地利用缓存的带宽。在Broadcom等主流交换机中都将缓存的数据总线分成时隙(每个时隙传输一个分片)，各个端口之间以轮循方式使用时间片[2]。这样不仅能做到各个端口之间的公平，还不会造成带宽的浪费。

1.3 流量控制和分组调度

流量控制算法用于处理网络经常出现的队列拥塞。分组调度算法主要是针对每个端口队列的优先级来设置的，决定某个时刻由哪个队列来输出数据。这两种算法在网络设备中有广泛的应用。

2 Broadcom公司BCM5690交换机芯片的缓存结构

BCM5690是一个千兆以太网交换机芯片，该芯片支持12个千兆以太网端口和一个10 Gbps的HiGig堆叠端口。

2.1 交换机的整体结构

BCM5690的整体结构如图1所示，图中共享缓存管理单元(Memory Management Unit，MMU)是整个系统的核心，MMU除了管理缓存外还有一个总线仲裁器的作用，所有的数据都会由MMU存储到共享缓存池(Common Buffer Pool，CBP)中，输出时也是由MMU来调度输出顺序，MMU为每个千兆端口接口控制器(Gigibit Port Interface Control，GPIC)的输入和输出分配总线带宽[3]。

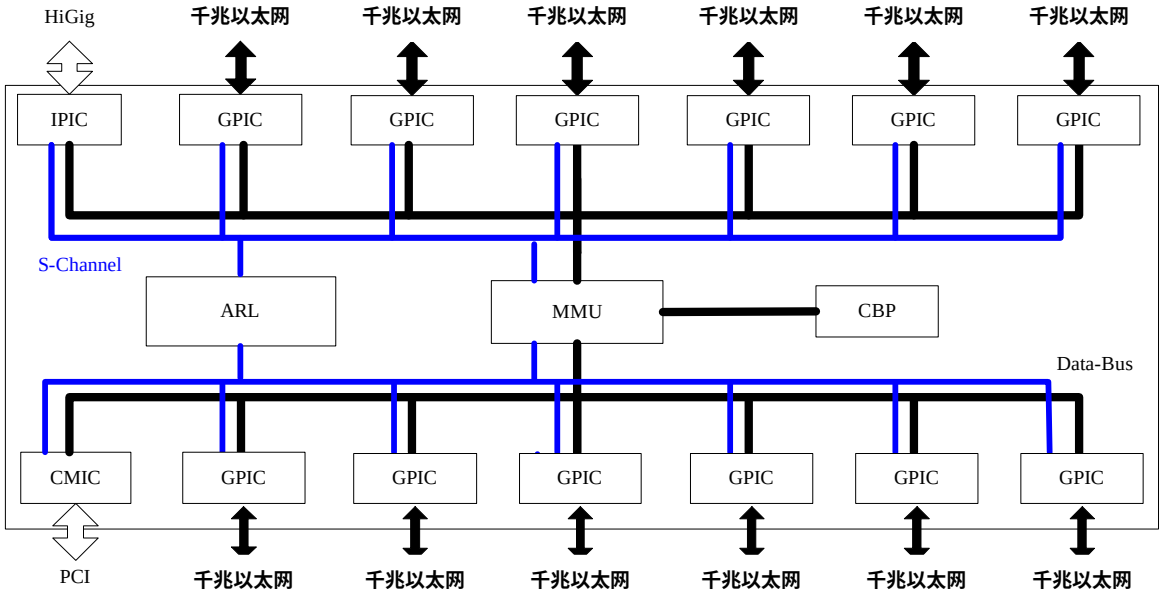

图1　BCM5690 的整体结构

MMU和GPIC之间是通过32 Gbps的片内总线相连。在BCM5690 中，MMU和其他模块之间的控制信息是通过S-Channel总线传输，在硬件实现上，也可以通过控制线传输。

GPIC是物理层与缓存之间的过渡模块，并且在GPIC中完成所有查找过程(包括二层查找，路由查找，流分类的一些功能)，然后等待MMU分配总线时隙给GPIC，将数据包连同查找结果一起传输给MMU；MMU输出的数据包也是在GPIC中组装再传输给物理层，

另外，交换机的流量控制功能也是在MMU和GPIC之间协同完成的。MMU有专门的寄存器统计和判断

相关的量(如队列长度)，当这些量到达一定门限的时候，MMU通过S-Channel向GPIC发出信息，GPIC根据信息决定相关数据包(如到达某个队列的数据包)是接收还是丢弃。这样作可以节约缓存的带宽和空间，避免需要丢弃的数据包进入缓存[3]。

2.2 缓存管理模块的结构

缓存管理模块的结构如图2所示[3]。

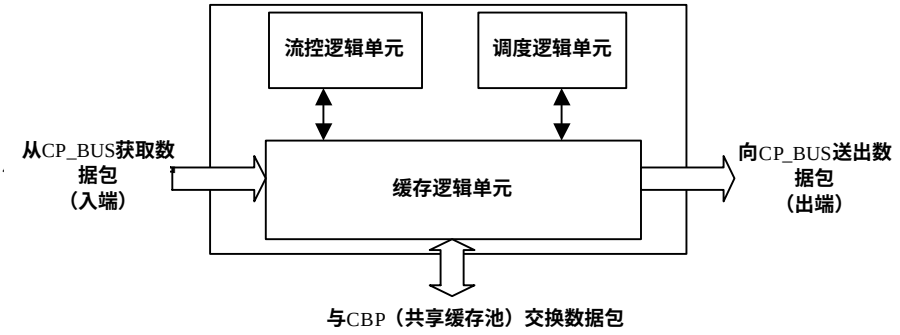

图2 MMU的物理结构

1) 缓存逻辑单元(Buffer logic)

缓存逻辑单元接收来自片内总线的数据包，再把它们存在CBP中，此时MMU可以同时以线速接收来自其他模块的数据包。缓存逻辑单元也可以重新获得CBP中的数据包，再把它们传输到片内总线上，此时MMU也能以线速把数据包同时发往其他模块。

2) 流控逻辑单元(Flow Logic)

流量逻辑单元可以对设备的流量控制选项进行管理。它包含多个用来追踪CBP中的数据包的计数寄存器，这些寄存器在完成流量控制中起了重要的作用。每个端口的寄存器都设置了激活门限，当达到任一个门限时，MMU会采取适当的流量控制策略。

3) 调度逻辑单元(Schedule Logic)

根据缓存调度算法(如Weighted Round Robin)决定对该数据包的处理方式(放入哪个端口的哪个优先级队列)。该模块包含缓存队列的所有信息。

缓存逻辑单元的内部结构如图3所示，缓存逻辑单元的功能如表1所示。

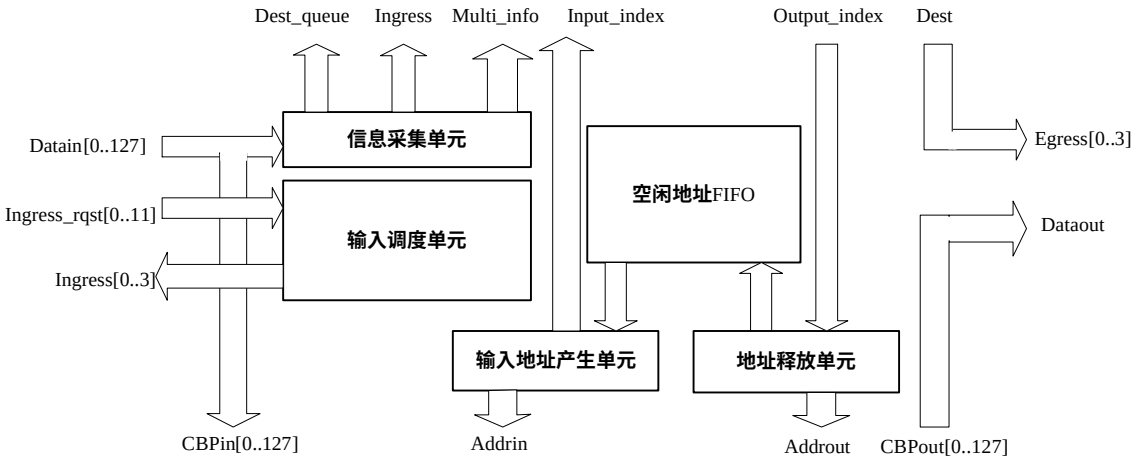

图3 缓存逻辑单元的结构

BCM5690将1 MB的片内缓存(CBP)划分为8 K个128 B的分片。数据包到达MMU后，MMU读取控制的信息，将数据包分成128 B的分片，存入相应的队列。在实际的芯片中，数据包的分割过程是在接收数据的同时完成的，因为可以认为片内的数据总线是与CBP的数据线直接相连的。每8个时钟周期改变一次分片的地址索引(由地址索引可以直接映射出CBP的物理地址)，就可以做到数据的分片存储。

表1　缓存逻辑单元功能

逻辑单元	功　　能
信息采集单元	当数据包到来时采集数据包的Header信息，传给流控和调度模块
输入调度单元	根据输入端的空闲情况，用轮循方式为输入端分配带宽
输入地址产生单元	从空闲地址先进先出(First in First out, FIFO)队列中取出空闲分片的索引，将该索引传给Schedule单元，并将索引转化为物理地址作为接收数据包的输入地址。
空闲地址FIFO	存储空闲的分片的索引号
地址释放单元	接收需要输出的分片的索引，将索引放回空闲地址FIFO中，并将索引转化为物理地址读取缓存中的数据。

2.3　缓存中的队列

缓存队列是由调度逻辑单元建立和更新的。缓存队列中存储的业务队列(Transaction Queue，XQ)表项如图4所示，XQ表项记录每个Cell的地址索引号及相关的控制信息。在XQ队列中同一个数据包的XQ表项是连续的，而在共享缓存中，同一数据包的分片是离散存储的。

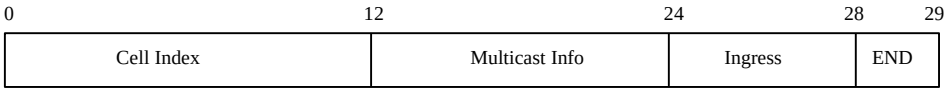

图4　XQ表项的数据结构

当数据离开MMU时，MMU会根据XQ表项中的Cell地址索引依次读取数据，并发送到片内总线上，然后MMU会回收释放的分片地址，将这些地址重新记入空闲地址FIFO。离散存储虽然在一定程度上会导致存储空间的浪费——数据包的最后一个分片不一定会被数据完全填满，但是这样能够方便缓存的管理。

3　分片存储、描述符作为队列单元的缓存管理方案

BCM5690芯片能够有效地完成整个交换过程，但是芯片设计的某些部分硬件实现上有一定难度。1) BCM5690的XQ队列是针对每个分片，需要有8 K个XQ表项，这样出端服务等级队列(Class Of Service，COS)队列会占用较大的存储空间，而片内存储空间相当有限；2) BCM5690数据包的查找操作是在每个GPIC上完成的，这涉及到二层表，三层表，VLAN表，多播表的镜像和查找引擎的复用，因此查找表的更新比较困难[1]。

本文结合BCM5690的结构提出了一种基于分片存储和描述符队列的缓存结构，这种结构的处理过程和数据结构与BCM5690有很大不同。

3.1　缓存管理方案的数据结构

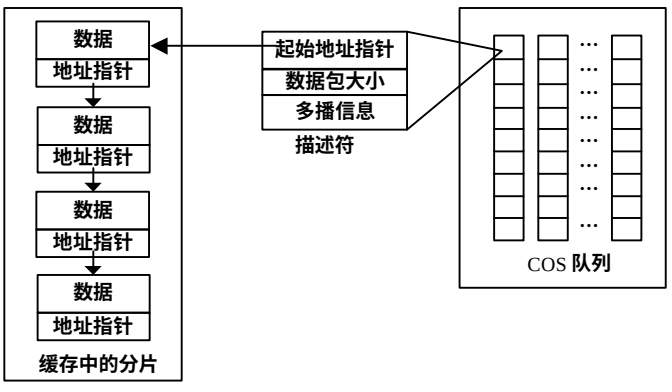

图5　缓存数据结构

缓存管理方案的数据结构如图5所示。每一个队列由数据包描述符组成。每个描述符代表一个数据包(而不是一个分片)。描述符中记录了该数据包的第一个分片的地址，每个分片除了记录数据还记录了下一个Cell

的地址，一个数据包的所有分片形成一个链表。该数据结构的优势在于：1) COS队列的结构很简单，不需要记录太多的信息，相对来说比BCM5690的COS队列占用的空间少。2) 能够支持较小的分片，提高缓存的利用率。3) 在COS队列中自然而然地实现数据包之间的分界，不需要专门的定界符(如图4中的报尾标识字段)。

3.2 缓存的物理结构及流程

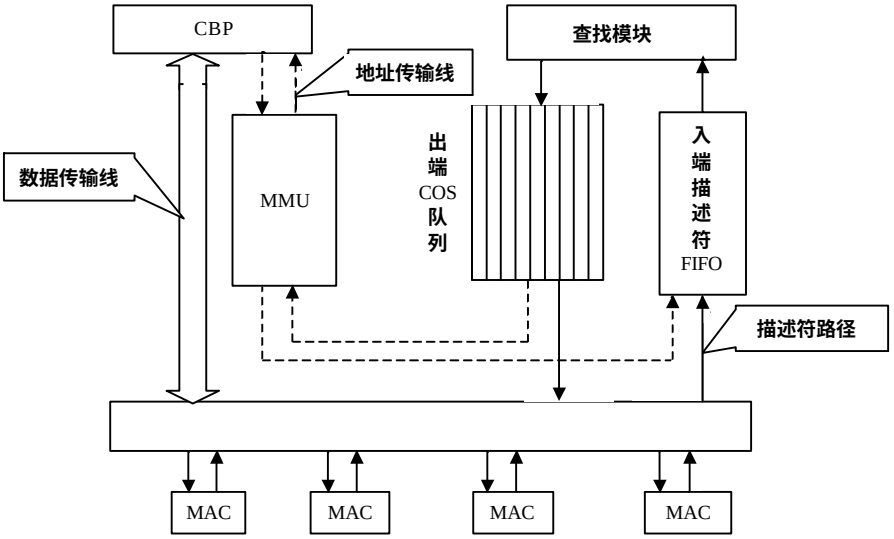

图6 整体物理结构

缓存的整体物理结构如图6所示。方案中的数据包流程为：1) 当数据包由数据传输单元传入的时候，数据包本身存入CBP中，MMU负责分配空闲空间，并返回存储地址，而数据包的包头会进入入端描述符FIFO中，产生一个数据包描述符。2) 在以后的操作中，交换机都是对数据包描述符的处理，数据包本身只有在输出的时候才会根据描述符和输出队列的信息调出。3) 数据包描述符依次从入端描述符FIFO输出，在查找模块中进行各种查找操作，将查找结果记录在描述符中。4) 然后将描述符放入相应的出端COS队列中。对描述符进行输出排队。5) 最后由调度算法调出描述符，根据描述符找到数据包的物理地址，从缓存中取出数据传送给数据传输单元。同时，MMU回收已释放的分片地址，这些分片以后可以分配给其他输入数据包使用。

该方案的运作过程与BCM5690有很大差别。它是先将数据包存入缓存，建立数据包描述符，再根据描述符提供的控制信息(这些信息包括从数据包帧头复制的信息和数据包在缓存中的地址)进行以一系列查找操作，然后将描述符入队。这样真正做到了数据包的“零拷贝”[2]，在数据包输入和输出过程中，只有数据包描述符的移动。描述符是数据包的唯一标识，如果描述符被删除(比如由于加权早期随机丢弃标识算法产生的丢弃)，缓存中的数据包就没有意义，很快就会被新的数据包覆盖掉。其优势在于：1) 数据包本身的传输被简化。2) 数据和控制部分的功能划分明确，而且相对独立。3) 所有的查找功能都集中在查找模块中，物理实现的时候，L2、L3等地址表的复用，更新以及查找模块的复用结构清晰，容易实现。

本文研究工作得到中兴通信科研基金资助，在此表示感谢。

参 考 文 献

[1] Chao H J, Cheuk H L, Eiji O. Broadband packet switching technlogies[M].New York: A Wiley-Interscience Publication John Wiley & Sons, Inc., 2001. 83-96

[2] Garcia-Haro J, Jajszczyk A. ATM shared-memory switching architectures[J]. IEEE Network Magazine, 1994,8: 18-26

[3] Lau M V, Shieh S, Wang Peifeng, *et al*. Gigabit ethernet switches using a shared buffer architecture[J]. IEEE Communications Magazine, 2003, 41: 76-84

编 辑 漆 蓉