光总线交换网络输出排队两级缓冲结构与性能分析

李万林,田 畅,郑少仁
(解放军理工大学通信工程学院交换技术与 ATM 研究中心,江苏南京 210007)

摘 要: 为了解决核心路由器高速无阻塞光总线交换网络体系结构中的高速大容量分组缓冲这一关键技术难题,本文提出了基于 SRAM 技术和 DRAM 技术相结合的输出排队分组两级缓冲结构及相关 LBF - MMA 存储器管理算法,并利用实测的网络流量数据对该缓冲技术的性能进行了仿真分析. 分析表明,两级缓冲结构较好地解决了光总线交换网络中分组缓冲高速度与大容量之间的矛盾,对高速路由器技术的发展也具有一定的指导意义.

关键词: 高速路由器; 分组缓冲; 光总线交换网络
中图分类号: TP393.05 **文献标识码**: A **文章编号**: 0372-2112 (2003) 04-0589-04

Performance Analysis of Output-Queued Two-Stage Packet Buffer Structure for Optical Bus Switching Network

LI Wan-lin, TIAN Chang, ZHANG Shao-ren
(ATM R&D Center, Institute of Communications Engineering, Nanjing, Jiangsu 210007, China)

Abstract: In order to solve the key technical problem that how to efficiently buffer high-speed traffic under the architecture of optical bus switching network, this paper proposes an output-queued two-stage packet buffer structure and related LBF-MMA (Longest beffer first-memory management algorithm) based on the combination of SRAM (Static radom access memory) and DRAM (Dynamic random access memory) technology. Analysis shows that such technique gives a good solution that meets both high speed and large capacity requirements in packet buffering for the optical bus switching network.

Key words: high-speed router; packet buffer; optical bus switching fabric

1 引言

总线交换网络体系结构最初广泛应用于 ATM (Asynchronous Transfer Mode) 交换机[1]. 文献[2]提出了创新的三平面无阻塞光总线交换网络,实现了核心路由器高速分组交换. 图 1 为 8 ×8 2.5Gbps 光总线交换网络,每个输入端口的数据信号调制到光网络上以后,经过 1:8 光分配器,产生 8 路相同的光信号,分别传送到 8 个输出端口上,然后在输出端口执行分组筛选、分组缓冲、拥塞控制、分组调度等功能. 这种光总线交换网络体系结构具有许多鲜明的特点,例如,减少了交换网络的设计复杂度,各个端口之间的通路因没有竞争而能够实现高速交换,易于实现广播和组播,输出排队克服了组合输入输出排队对 QoS 支持的不足[3],易于实现丰富的 QoS 支持等. 同时,所有端口流量都以广播方式涌向一个端口给输出排队分组缓冲带来了挑战.

首先,每个端口都可能以高的突发速率 (2.5Gbps) 涌向某一输出端口,要求输出端口缓冲区具有很高的有效带宽. 8 个 2.5G 端口流量同时涌入某一输出端口,再加上该端口要以 2.5G 速率输出分组,因而仅缓冲这些分组就需要 2.5 ×9 = 22.5 Gbps 有效带宽,再加上 50 - 100 %的诸如 QoS 等高级网络服务开销[4],要求分组缓冲要具有 33.75 - 45Gbps 有效带宽. 目前最新的 DDR333 存储器展宽到 64 位才能够提供 176MHz ×2 × 64 = 21.6Gbps 的峰值带宽,因此使用具有大容量特点的 DRAM (Dynamic Random Access Memory) 技术来实现分组缓冲是非常困难的.

其次,TCP (传输控制协议) 流量对端口分组缓冲容量有很高的要求. 经验上要求端口分组容量大小约为 RTT ×R,其中 R 为端口速率,RTT (Round Trip Time) 为 TCP 流中分组的往返时间[5]. 目前 RTT 一般取为 0.25 秒[5],这样一个 2.5Gbps 端口分组缓冲大小约为 0.25s x 2.5 Gbps = 78MB. 目前快速的 SRAM (Static Radom Access Memory) 的容量都比较小 (目前业界 SRAM 最大容量能够做到 36Mbit),同时成本也比较高,因而使用快速的 SRAM 来实现如此大的分组缓冲目前也是不可取的. 图 1 为光总线网络结构.

从技术水平现状和未来几年发展趋势来看,RAM 技术难于达到核心路由器高速端口分组缓冲的要求. 因此,为了解决光总线交换网络架构下高速大容量分组缓冲这一关键技术难题,必须突破存储技术水平的限制,采用新的体系结构.

收稿日期:2002-01-04;修回日期:2002-07-11
基金项目:国家 863 计划资助项目 (No. 2001AA121063)

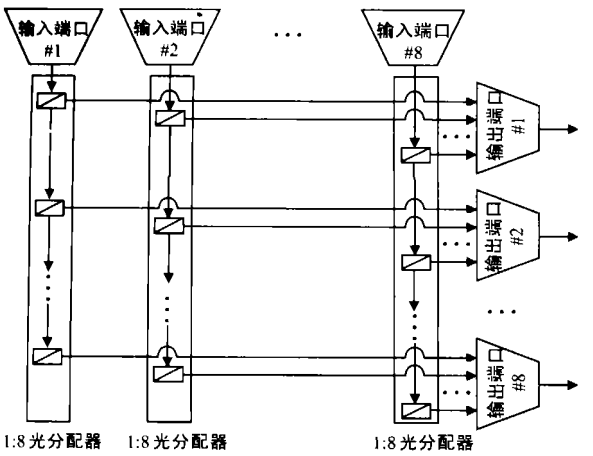

图 1 OmniRouter880 光总线交换网络结构

本文结合 SRAM 技术的高速度优势和 DRAM 技术的大容量特点,基于这种光总线交换网络体系结构,提出了一种两级分组高速缓冲结构,并利用实测的网络流量对这种分组缓冲技术的性能进行了仿真分析. 分析表明,两级缓冲结构大大降低了分组缓冲对存储器的苛刻要求,具有很好的可实现性和可行性,是光总线交换网络架构下分组缓冲良好的解决方案.

2 两级缓冲结构

本文提出的基于光总线交换网络的两级缓冲结构,如图 2 所示,主要包括四个组成部分:信元筛选器、一级 SRAM 快速缓冲、存储器管理算法和二级并行 DRAM 阵列缓冲等. 为了提高存储器管理效率,并降低交换网络的实现复杂度,采用固定长度分组(即信元),大小通常为 64 字节[6]. 信元筛选器过滤掉相对应的输入端口广播过来的目的端口地址非本地的所有信元,只允许目的地址为本端口的信元写入一级缓冲,因而它以线路速率 R 工作.

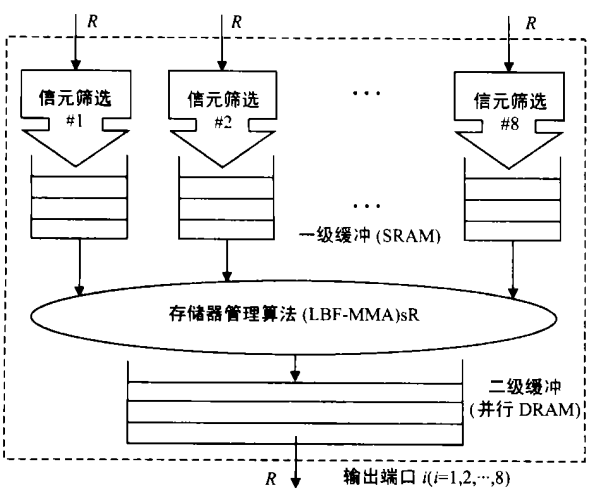

图 2 基于光总线交换网络两级分组缓冲结构

一级缓冲是采用 SRAM 技术的高速小容量缓冲,它用来暂时缓存来自相应输入端口、目的端口地址为本端口的所有信元. 二级缓冲为大容量的并行 DRAM 阵列,用来为所有目的地址为本端口的信元提供大容量的缓冲.

为了降低一级缓冲的信元丢失率,本文提出了 LBF-MMA (Longest Buffer First-Memory Management Algorithm,LBF-MMA) 存储器管理算法,完成一级缓冲中的信元到二级缓冲的转储. 从统计角度来看,某一输出端口的信元到达速率为 R,考虑到流量的突发性,MMA 以速率 sR 从一级缓冲转储到二级缓冲,s 为加速因子. 采用 LBF-MMA 最长缓冲优先存储器管理算法,优先从当前信元最多的缓冲中转储信元,而不采用简单的轮询算法.

这种两级缓冲结构,由于采用 SRAM 和 DRAM 相结合的技术,能够支持高速交换网络的分组缓冲. 对一级缓冲来说,流入和流出的峰值速率分别为 R 和 sR,因此要求其有效带宽大于 $(1+s)R$;对二级缓冲来说,具有固定的平均流出速率 R 和峰值流入速率 sR,因此要求并行 DRAM 阵列能够提供 $(1+s)R$ 有效带宽. 在转储加速因子 s 不太大的条件下,对 SRAM 和 DRAM 的要求相对都不苛刻.

3 仿真分析

3.1 仿真模型

由以上分析可知,影响两级缓冲结构实用性的主要因素是:信元丢失率、一级缓冲容量大小和 MMA 转储加速因子. 一级缓冲大小决定了使用高速 *Cache* 的可行性,加速因子的大小决定了使用 DRAM 存储的复杂性和可行性. 二级缓冲所引起的信元丢失率主要是由于到达本端口总的流量速率和流量突发方式以及突发时间长度的分布有关. 一级缓冲整形作用降低了涌入二级缓冲流量的突发度,在一定程度上反而降低了仅使用单级缓冲系统的信元丢失率. 一级缓冲性能成为两级缓冲系统性能的关键因素,因此本文在分析中忽略二级缓冲所引入的信元丢失率. 围绕分析目的,抛弃一些次要因素,抓住问题本质,简化模型,做三点假定:所有端口都以相同的速率 r 到达目的端口;来自某一输入端口的流量等概率地(以 1/8 概率)到达各个目的端口;所有一级缓冲区的容量相同. 仿真模型如图 3 所示.

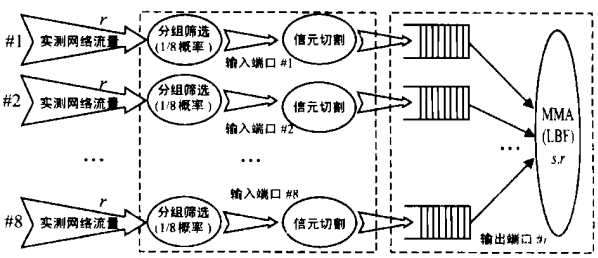

图 3 仿真模型

由于信元丢失率除了与缓冲区大小 m 和 MMA 转储加速因子 s 有关外,还可能受 Internet 流量特点等因素影响. 目前还没有公认的可供使用的网络流量模型[7],本文采用 Internet 骨干节点网流量实测数据[8]来仿真. 由于 MMA 采用 LBF 方式转储一级缓冲中的信元,因而对于每个一级缓冲来说其转储的峰值速率可以达到 sr,而一级缓冲的信元到达速率峰值能够达到 r.

设第 i 个缓冲的信元丢失率为 d_i,在仿真过程中总共到达个 a_i 信元,丢失 b_i 个信元,则两级缓冲系统的信元丢失率 d 为

$$d = \sum_{i=1}^{8} b_i / \sum_{i=1}^{8} a_i \qquad (1)$$

如果 $a_1 = a_2 = \cdots = a_8$,则

$$d = (1/8) \sum_{i=1}^{8} b_i / a_i = (1/8) \sum_{i=1}^{8} d_i \qquad (2)$$

而不是直观感觉的各个信元丢失率的和. 本文使用式(1)来计算两级缓冲结构的信元丢失率.

为了实现 8 个相同平均速率的实测网络流量来灌入本文的仿真系统,首先对从网上下载的网络实测流量预处理. 为此,创建一个庞大的循环链表 Ω,Ω 中的每个节点包含如下信息:(1)分组到达间隔时间;(2)分组长度,以便执行分组分割;(3)下一节点地址. 创建 8 个指针均匀地指向这个循环链表,作为 8 个实测流量数据. 为了能够较精确地仿真 Internet 实时流量(无限长时间网络实测流量),Ω 的长度不能太短,在仿真实验中,本文使用的所有 Ω 的长度都大于 7300000(个分组). 为此,我们从 NLANR 官方网站下载了约 3G 字节的 Internet 流量实测数据. 在一定意义下,仿真实验中采用的实时网络流量非常精确的模拟了 Internet 网络实时流量.

由于网络实测数据中没有路由器动态路由表信息,本文采用经典文献[1,9]采用的输入流量等概率分布到目的端口方法,进行建模.

3.2 仿真结果

首先,仿真分析 MMA 转储加速因子固定条件下,信元丢失率与一级缓冲区容量大小之间的关系. 缓冲区容量大小以信元个数为单位,仿真结果如图 4 所示.

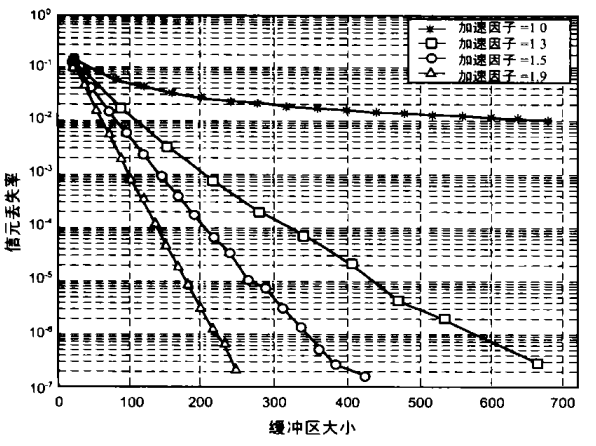

图 4　信元丢失率与缓冲区大小之间关系

图中横轴为一级缓冲容量大小,单位是信元个数,线性坐标;纵轴为信元丢失率,对数坐标. 图 4 表明,在相同加速因子条件下,随着缓冲区大小的增加,在一定的范围内,信元丢失率呈指数下降的趋势.

其次,在一级缓冲区容量固定条件下,信元丢失率与 MMA 转储加速因子之间关系的仿真结果如图 5 所示. 图中横轴为 MMA 转储加速因子,线性坐标;纵轴为信元丢失率,对数坐标. 图 5 表明,在缓冲区固定条件下,信元丢失率随着加速因子的增加呈指数下降的趋势.

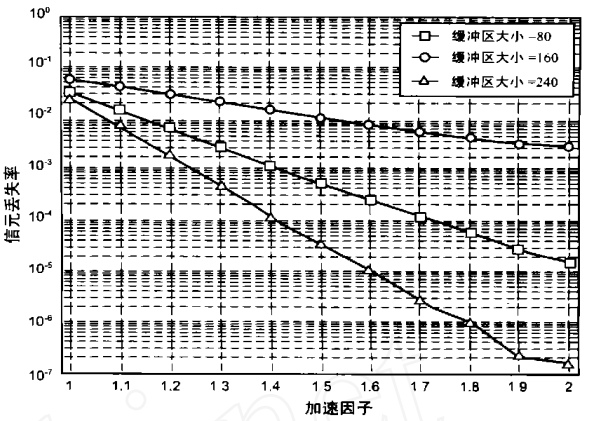

图 5　信元丢失率与加速因子之间的关系

仿真结果表明,当一级缓冲较小,MMA 转储加速因子较小时候,两级缓冲系统具有较高的信元丢失率. 然而可以通过增加一级缓冲的大小和增大 MMA 转储加速因子来明显的降低信元丢失率. 例如,当加速因子为 1.5 时,只要一级缓冲区容量超过 440 个信元(即 27.5K 字节),信元丢失率就可以降到 10^{-7} 以下. 这基本上不影响整个分组缓冲系统的信元丢失率.

4　进一步讨论

本文在对模型的仿真中做了三点假设:端口速率均为 r,一级缓冲大小均为 m,输入端口以相同的概率 $p = 1/8$ 分布其流量到各个输出端口. 下面本文放宽这些限制,进一步讨论两级缓冲在一般条件下的性能.

设端口 i 的瞬时速率为 $R_i(t)$,端口 i 向端口 j 瞬时转发概率为 $p_{i,j}(t)$,端口 j 第 i 个一级缓冲的容量大小为 l_{ij},其中 $i, j = 1, 2, \cdots, 8$. 只要涌向端口 j 的平均流量速率与端口 j 的平均输出流量速率 $\bar{R}_j$ 满足式(3),

$$\sum_{i=1}^{8} \bar{R}_i \cdot \bar{p}_{ij} \leq \bar{R}_j \qquad (3)$$

就不会由于第二级缓冲排队急速膨胀而导致信元丢失率迅速增加,一级缓冲信元丢失率仍然是两级缓冲系统性能的关键因素. 本文定义归一化的缓冲区长度 $l'_{ij}(t)$ 为,

$$l'_{ij}(t) = c_{ij}(t) / l_{ij} \qquad (4)$$

式中 $c_{ij}(t)$ 为端口 j 的第 i 个一级缓冲中在时刻 t 时的信元总数. 本文 LBF-MMA 算法为总是优先转储归一化缓冲区长度 $l'_{ij}(t)$ 最大的那个一级缓冲中的信元. 由于 MMA 以 LBF 方式转储一级缓冲中的信元,当某个一级缓冲的信元到达速率增大时,其对应的服务速率也会线性增加,峰值能够达到 $s\bar{R}_j$. 极端情况是,某个端口 k 以 $R_j(t)$ 速率全速涌向端口 j,其他端口都不向端口 j 发送信元. 这时候端口 j 的第 k 个一级缓冲的的信元到达速率为 $R_j(t)$,服务速率同样也增加并达到 $s\bar{R}_j$,信元丢失率的影响不大. 因此,在一般情况下,两级缓冲结构具有良好的性能.

其次,讨论一下两级缓冲的可实现性问题. 对于一级缓

冲,采用 SRAM 技术实现是可行的. 一级缓冲以有效带宽 $(1+s)R$ 工作,对于 2.5G 端口,加速因子 $s=1.5$ 条件下,这个数值为 6.25Gbps,SRAM 技术完全能够提供足够的带宽. 另外,一级缓冲的容量也较小,一般为几十 K 字节,使用 SRAM 技术容量上也没有问题. 二级缓冲采用并行 DRAM 阵列,只要线路卡物理尺寸允许,能够实现很高的总带宽和庞大的容量. DRAM 阵列要能够提供 $(1+s)R$ 有效带宽,由于 MMA 转储加速因子不大,对 DRAM 的要求也不高. 对于 2.5G 端口,如果采用 $s=1.5$,则这个数值为 6.25Gbps 带宽. 再加上诸如 QoS 等高级网络服务 50 - 100 %额外开销,缓冲区需要 9.375Gbps - 12.5Gbps 带宽. 目前最新的 DDR333 存储器能够提供 21.6Gbps 的峰值带宽,完全能够满足二级缓冲带宽要求.

最后,比较一下两级缓冲技术和单级缓冲技术的性能. 对于单级缓冲系统,输出排队要求缓冲区能够提供 $(1+N)R$ 带宽. 对于两级缓冲结构,只要求缓冲区能够提供 $(1+s)R$ 带宽,其中 s 远小于 N. 至于两级缓冲系统的信元丢失率,虽然增加了一级缓冲所带来的信元丢失因素,但同时由于一级缓冲的整形作用,二级缓冲的信元丢失率肯定有所降低,所以总的信元丢失率性能也与单级缓冲差不了多少. 两级缓冲结构能够支持更高速率端口,并且具有更好的可实现性能.

5 结语

为了解决核心路由器光总线交换网络体系结构下高速大容量分组缓冲这一关键技术难题,必须突破存储器技术水平限制,采用创新的存储体系结构. 结合 SRAM 技术的高速度优势和 DRAM 技术的大容量特点,本文提出了输出排队两级缓冲结构及相关的 LBF-MMA 存储器管理算法. 分析表明,两级缓冲结构的信元丢失率主要取决于一级缓冲的大小和 LBF-MMA 转储加速因子,并随着缓冲区大小的增加或者 LBF-MMA 加速因子的增加呈指数下降的趋势. 当 LBF-MMA 加速因子为 1.5 时,一级缓冲的大小只要有数十 K 字节,两级缓冲系统的信元丢失率就降低到很低,因而具有可行性和良好的可实现性.

两级缓冲结构不仅较好地解决了光总线交换网络架构下分组高速大容量缓冲这一关键技术难题,而且对于高速路由器技术也具有一定的指导意义. 本文的分析和讨论是基于光总线交换网络和 2.5G 速率端口的,对于其他类型交换网络和更高速率的端口,本文的思想也具有一定的借鉴意义.

参考文献:

[1] M D Prycker. Asynchronous Transfer Mode-Solution for Broadband ISDN Second Edition [M]. Ellis Horwood, 1993. 177 - 187.

[2] 汪斌强,戚文芽,兰巨龙等,基于光总线的无阻塞交换网络的工程实现 [J]. 电信科学,2001,17(7):40 - 43.

[3] S Chuang, A Goel, N McKeown, et al. Matching output queuing with a combined input/output-queued switch [J]. IEEE J Sel Areas in Communications, 1999, 17(6):1030 - 1039.

[4] Rambus Inc, Rambus DRAM for OC-192 Data Rate Line Card Applications [Z]. http://www.rambus.com, 2000.

[5] S Iyer, R R Kompella, N McKeown. Analysis of a memory architecture for fast packet puffers [J]. IEEE Workshop on High Performance Switching and Routing, 2001.

[6] A Birman, H R Gail, An optimal service policy for buffer systems [J]. Journal of the Association for Computing Machinery, 1995, 42(3):641 - 657.

[7] 田畅,王海,郑少仁,基于用户行为的网络流量模型及自相似性分析[J]. 通信学报,2000,21(9):19 - 25.

[8] NLANR, http://pma.nlanr.net/Traces/Traces/daily/20011012/ [Z], 2001.

[9] N McKeown. Scheduling algorithms for input-queued cell switch [D]. PhD thesis, university of California at Berkeley, 1995.

作者简介:

李万林 男,1976 年 4 月出生于河南省虞城县,在读博士研究生,1997 年在解放军理工大学获得电信交换工程专业学士学位,2000 年在解放军理工大学获得程控交换专业硕士学位,目前在解放军理工大学攻读通信与电子系统专业博士学位,主要研究方向为宽带网络技术.

田 畅 男,1963 年 2 月出生于山东省青岛市,2001 年于解放军理工大学获通信与电子系统专业博士学位,现为解放军理工大学全军交换技术与 ATM 交换中心副教授,中国电子学会高级会员,主要从事宽带交换技术、网络安全和无线分组网的研究,在国内外有关刊物、会议发表论文 30 余篇.