基于FPGA实现的10Gbps高速转发引擎设计分析*)

刘勤让[1,2] 邬江兴[1,2]
(解放军信息工程大学信息工程学院 郑州450002)[1]
(国家数字交换系统工程技术研究中心 郑州450002)[2]

摘 要 高速转发引擎的设计是T比特路由器设计中的关键和难点,本文围绕传输带宽需求、查表时间需求和包头处理时间需求以及器件水平等方面对基于FPGA实现的10Gbps高速转发引擎进行了详细的分析,讨论了高速转发引擎各功能模块的设计可行性,并给出了一种可行的实现方案。

关键词 FPGA,10Gbps,转发引擎,并行,流水线

Design Analysis of 10Gbps Forwarding Engine with FPGA

LIU Qin-Rang[1,2] WU Jiang-Xing[1,2]
(Information Engineering College of PLA, Information Engineering Institute, Zhengzhou 450002)[1]
(National Digital Switching System Engineering & Technology R&D Center, Zhengzhou 450002)[2]

Abstract The 10Gbps forwarding engine is one of the most challenge task in terabit router design. In order to design a 10Gbps forwarding engine with FPGA, the transmission bandwidth requirement between each elements in FPGA, IP look-up time, header processing cycle number and the up-to-date development of FPGA are all analyzed in detail. So the number of fiber to provide enough transmission bandwidth is calculated, a fast IP look-up scheme is presented and a header processing module of parallel and pipeline structure is provided. Finally the implementing scheme of the 10Gbps forwarding engine is proposed.

Keywords FPGA, 10Gbps, Forwarding engine, Parallel, Pipeline

1 高速转发引擎设计面对的挑战

随着Internet向下一代网络的演进,对网络的传输带宽、节点处理能力和不同业务的QoS支持能力等都提出了更高的要求。由于光传输技术的飞速进步,无论是单波长载荷速率还是单纤可用波长数量,每年都以超摩尔定律的速度在增长。而作为网络节点设备的路由器则遵循摩尔定律的法则向前发展,导致高速路由器逐渐成为网络发展的瓶颈。图1反映了单纤传输容量和路由器端口速率的发展关系。

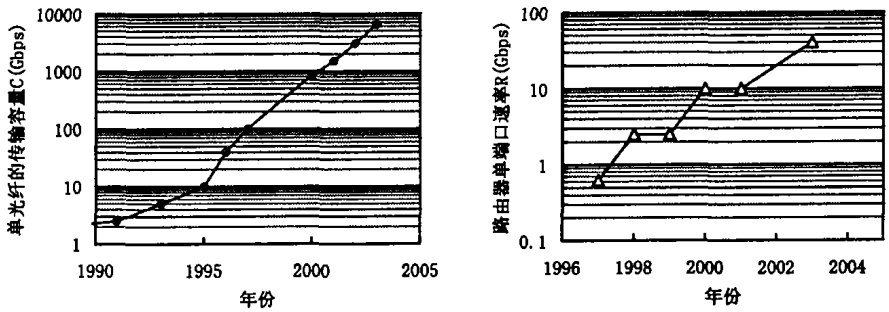

图1 单光纤传输容量和路由器端口速率的增长趋势对比

正是在此背景下业界提出了T比特路由器的设计需求,T比特路由器的基本特点[1]包括:

1. 更大的端口密度;
2. 更高的端口速率;
3. 更大的交换容量。

我们国家为了提高在下一代互联网中的竞争力,由国家十五'863'在信息技术领域中发布重大专项课题'可扩展到T比特的IPv4/v6/MPLS路由器基础平台和实验系统'。本文就是针对T比特路由器中支持QoS的10Gbps高速转发引擎设计中的难点进行了分析和论证,并最终给出一种可行的设计方案。

高速转发引擎主要完成对报文的第三层查表和报头处理等操作,具体包括提取IP报头,根据报文类型分别进行单播查表、组播查表、安全过滤、优先级映射以及报头本身的有效性检查(包括版本号、TTL超时和地址范围等),综合上述的查表和检查结果,生成对报文的处理方式,包括正常转发报文、产生记录报文、转发并记录报文和丢弃等四种操作方式。对转发报文和记录报文还需要根据上述的查表和检查结果生成相应的内部标签,贴于报文的头部,供交换或主控查看。高速转发引擎的最大挑战就是必须对最短包支持线速处理。高

*)基金项目:国家十五'863'信息技术领域重大专项课师(NO 2003AA103510)。刘勤让 博士生,研究方向为IPQoS和高速网络节点中的QoS实现;邬江兴 工程院院士,教授,博士生导师,国家数字交换系统工程技术研究中心主任,国家863计划高性能宽带信息网总体组组长。

速转发引擎的实现可以采用网络处理机、ASIC 专用芯片和可编程 FPGA 等实现方式，本文讨论的实现方案是基于 FPGA 的。

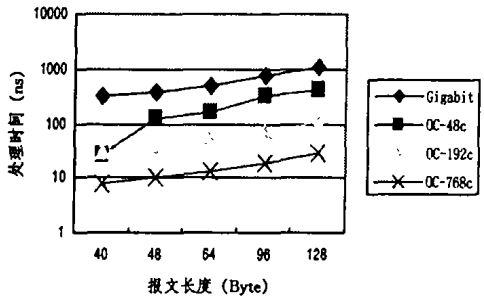

图2　不同端口速率线速包转发的处理时间

由图2可以看出，要完成对10Gbps 端口速率的最短40字节报文线速转发，高速转发引擎的最长处理时间只有32ns，若采用125Mhz 内部时钟设计，也只有短短的4个周期。所以在高速转发引擎的设计中，都无一例外地引入了并行机制和流水线操作的思想[2]，这便是高速转发引擎的典型结构，如图3所示。

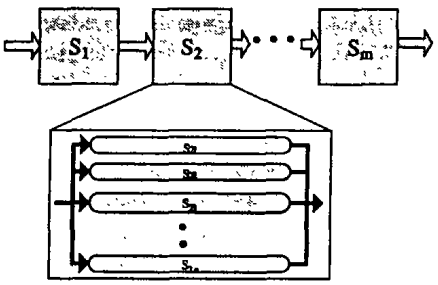

图3　高速转发引擎的并行流水线结构

并行机制就是对任务 S 划分为空间上‘相互独立’的 N 个任务子单元 $s_i(i=1,2,\cdots,N)$，从而实现由 N 个简单的任务子单元 s_i 在相同的时间内经过并行操作完成一个复杂的任务 S。所以通过并行机制可以降低实现复杂度。

流水线操作就是将一个重复的时序过程分解为若干个时间上‘相互独立’的子过程（每个子过程称为流水线的“段”或“级”），而每个子过程都可以有效地在其专用功能段上与其它子过程同时执行。“段”的数目称为流水线的“深度”。流水线的各段中，时间最长的功能段会造成流水线的“堵塞”和“断流”，成为流水线的瓶颈，同时也决定整个流水线操作的最长时间。所以通过流水线操作可以减少时序过程的重复时间间隔。

2　10Gbps 高速转发引擎的设计分析

本节将围绕10Gbps 高速转发引擎的设计从传输带宽需求、线速查表、报头处理和器件水平等方面进行分析。

2.1　传输带宽约束分析

传输带宽约束包括输入传输带宽约束、内部传输带宽约束、输出传输带宽约束。如图4所示。

为了描述方便，首先对符号定义如下：

R－端口速率；ξ－链路承载效率；L－报文长度；M－端口添加输入标签的字节数；N－转发引擎添加路由标签的字节数；K－转发引擎的数据送往交换网络的平面数。C－FPGA 内部的时钟频率；W－FPGA 内部数据传输的总线宽度；对于转发 FPGA 和端口以及交换网络的互连通常采用多路光纤并行传输，设 B－光纤的互连容量；$n1$－端口与转发之间的并行光纤数；$n2$－转发与交换网络的单平面并行光纤数；η－光纤互连效率。

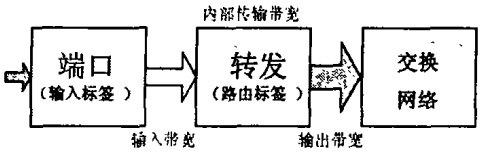

图4　高速转发引擎的带宽需求分析

对于10Gbps POS 接口的实际链路速率为 $R=2.488\times4=9.952$Gbps，同时由于采用 VC-4容器的 OC-192 SDH 帧[3]的承载效率为261/270，类 HDLC 帧[4]的承载效率为 $L/(L+9)$，所以对于长度为 L 的 IP 包，链路层承载效率 ξ 为：

$$\xi=\frac{261}{270}\times\frac{L}{L+9} \tag{1}$$

若光纤传输线路码型采用8B/10B 编码，用于分组定界的开销为（分组头—4字节 K27.7码，分组尾—4字节 K29.7码）8字节，所以对于长度为 L 的 IP 包，互连效率 η 为：

$$\eta=\frac{8}{10}\times\frac{L}{L+8} \tag{2}$$

所以转发引擎的输入和输出带宽以及内部传输带宽需求可以表达为：

$$B\times\eta\times n1\geqslant R\times\xi\times(1+\frac{M}{L}) \tag{3}$$

$$B\times\eta\times n2\geqslant R\times\xi\times(1+\frac{M}{L})\times(1+\frac{N}{L+M})\times\frac{1}{K} \tag{4}$$

$$C\times W\geqslant R\times\xi\times(1+\frac{M}{L})\times(1+\frac{N}{L+M}) \tag{5}$$

在报文长度 $L=40$字节时，根据我们的设计，取 $K=2$，$B=3.0$Gbps，$M=16$字节，$N=16$字节，可以求得 $n1=6$，$n2=4$，同时参考 Xilinx[5]和 Altera[6]的 FPGA 器件水平，内部可以通过采用125Mhz 时钟进行128比特的数据传输，可以提供16Gbps 的内部传输带宽，以满足内部传输带宽需求。

所以要实现10G POS 接口40字节 IP 包线速转发，线路接口模块到转发引擎需要6路光纤并行作为输入，转发引擎到交换网络需要2×4=8路光纤并行作为输出，而转发引擎 FPGA 内部可以采用125Mhz 时钟128位数据总线的设计。

2.2　线速查表分析

在转发引擎的线速查表设计中，通常用 TCAM[7,8]来存储表项，表项结构为查表关键字和查表结果两个部分。当一个 IP 报文到来时，转发引擎 FPGA 根据输入 IP 报文的报头信息提取出查表关键字，并将该关键字由 TCAM 的数据总线 DBUS 送给 TCAM，TCAM 开始查找包含有该关键字的表项。若查表命中，TCAM 则通过结果总线 RBUS 给出该匹配表项在 TCAM 中的保存地址；FPGA 再执行一次 TCAM 读表项操作，读出该匹配表项的完整表项，通过 DBUS 输出，查表结果即被读出。以上是传统的基于单个 TCAM 的查表结构，查表过程由 TCAM 查表搜索和 TCAM 读表项两个操作串行进行，且无法流水操作，因而整个查表过程需要的时间为二者的和，通常这个过程需要十几个时钟周期，显然无法满足10Gbps 最短包的线速转发要求。

为此我们设计了支持流水线操作的 TCAM＋SRAM 查表结构[9~11]，将表项的关键字存储于 TCAM 中，查表的结果存放于 SRAM 中，具体查表过程如图5所示。

在该结构下，将查表操作分为送查表关键字、输出匹配地址和读取查表结果等三个流水子过程，且支持流水线操作。整个查表时间由最长的查表功能段决定，在实际的设计中组播查表关键字的生成决定最长的查表时间，为4个时钟周期。在 TCAM 支持的125Mhz 工作频率下，可以满足10Gbps 转发引

擎设计需求。

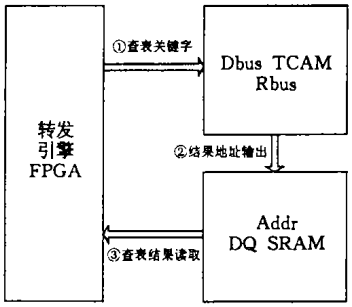

图5 TCAM+SRAM 的流水线查表操作

2.3 报头处理分析

在转发引擎的设计中，报头处理的逻辑关系最为复杂，它要完成对输入报文报头的有效性检查(包括版本号、TTL 超时和地址范围等)，对于组播报文，还要完成组播直连和组播 RPF 检查，结合报头处理对的有效性检查结果和查表结果，生成对当前报文的处理方式，包括转发、丢弃、上报和转发同时上报四种处理方式，根据报文的处理方式生成新的路由标签，对报文进行重新封装后分别送往交换网络和主控模块。同时所有的这些任务都必须支持线速处理，为此我们设计了有效性检查和查表操作的并行结构以及报头处理本身的流水线结构来支持对10Gbps 端口速率的40字节报文线速转发，如图6所示。

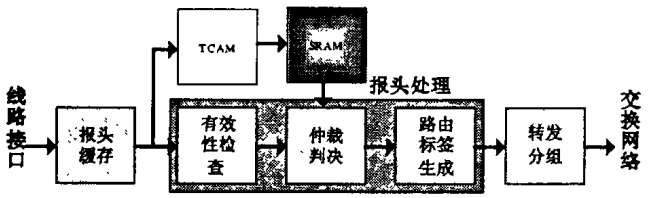

图6 报头处理的流水线实现结构

2.4 FPGA 器件水平分析

随着 FPGA 器件水平的飞速发展，也为基于 FPGA 实现 10Gbps 高速转发引擎提供了良好的技术开发平台，具体表现在以下几个方面：

(1)高速 MGT Xilinx 和 Altera 两个主要 FPGA 器件供应商都针对高速 FPGA 的开发在器件内部集成了多达24个 MGT(Multi-Gigabit Transciver)模块。MGT 可以实现高达10.3125Gbps 的高速数据传输，同时抗噪声能力更高、功耗更低并且可减少信号数量，可降低电路板复杂性。

(2)封装和电平 提供大量封装类型以及大量 IO 引脚数(最大1200)，从而可满足接口所需要的吞吐量要求。同时 FPGA 的每个引脚都支持数字控制阻抗匹配(DCI)技术，可减少成百/上千的片外端接匹配电阻，因此可以简化电路板布局布线工作。表现为可以减少电路板的层数，缩短布线长度，从而可获得更高的系统可靠性。FPGA 支持多种单端和差分电平标准，如 HSTL、SSTL、LVCOMS、LVTTL、PECL、LVDS 等。

(3)时钟 FPGA 需要连接多种外部器件，因此必须面对具有不同频率的多个时钟域。为此 FPGA 提供了多达12个 DCM 的支持，DCM 可以补偿由于时钟传输延迟以及电路板布局限制所产生的信号畸变，12个 DCM 提供了相位移动和频率合成能力，特别适合具有多个时钟域和关键时序要求的系统。

DCM 支持超过 400MHz 的时钟输出，从而可支持领先的总线接口标准，如 RapidIO 和 SPI-4。DCM 的数字化特点使其可不受系统温度和电压波动的影响。DCM 提供了一个可保证精确50/50占空比的零延迟时钟缓冲。DCM 可精确控制一个时钟周期内的相移，精度达到时钟周期的1%，这对建立和保持时间的调整非常关键。DCM 支持精确生成24MHz 至 420MHz 范围间的频率。

(4)块 RAM 超过10Mb 嵌入式 BlockRAM 可以实现对报文的大容量、高速缓存，以及对 QoS、流量监管等多种复杂控制的支持。

3 一种基于 FPGA 实现的10Gbps 高速转发引擎实现方案

综合文中的上述分析，本节给出了一种基于 Xilinx Virtex-II Pro 系列 FPGA 实现的10Gbps 高速转发引擎实现方案。本方案中用8对2.5Gbps 数据速率的 RocketIO 和10对2.5Gbps 数据速率的 RocketIO 来满足10Gbps 转发引擎的输入和输出带宽需求，内部采用125Mhz 时钟的128位数据总线传输来满足内部传输带宽需求，查表采用我们设计的 TCAM+SRAM 的结构来支持10Gbps 端口的线速查表，如图7所示。

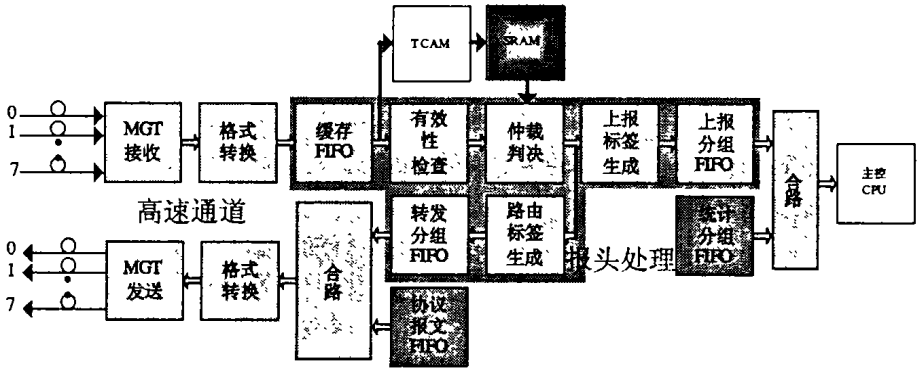

图7 一种基于 FPGA 实现的10Gbps 转发引擎结构

结论 目前国内外在下一代网络中的技术竞争已经全面展开，围绕大容量、高端口速率的 T 比特路由器开发引起了国内外的广泛关注，我国为了提高在下一代网络中的竞争力，在国家八六三计划中资助了重大专项课题"可扩展到 T 比特的 IPv4/IPv6路由器基础平台及实验系统"研究。本文针对该项目中10Gbps 高速转发引擎的实现难点和可行性进行了详细的分析，并基于自主知识产权考虑，设计了一种基于 FPGA 实现的10Gbps 高速转发引擎结构。

基于多代理的网格任务调度研究*)

曾万聃[1] 周绪波[2] 戴 勃[1] 游新冬[1] 常桂然[1]

(东北大学信息科学与工程学院 沈阳110004)[1](清华大学软件学院 北京100081)[2]

摘 要 随着网格技术的发展,代理技术近年来被用在网格的实现当中。多代理技术用分布式自主结构代替集中式的非自主性结构,具有更强的实时性,特别适合于动态调度,本文提出的网格任务调度系统就采用了多代理的体系结构。由于Globus已经提供了网格操作系统的大部分功能,本系统构建在Globus之上。在充分利用底层Globus提供的功能并与之结合的基础上,在代理中加入人工智能、知识学习的方法和服务质量实现策略,对网格任务进行灵活智能的调度,实现更好的负载平衡并达到一定的服务质量。采用JATLite来创建该系统,利用该系统可以根据应用背景需求来快速定制动态任务调度平台。

关键词 网格,智能代理,多代理,JATLite

Research on Grid Task Scheduling Based on Multi-Agent

ZENG Wan-Dan[1] ZHOU Xu-Bo[2] DAI Bo[1] YOU Xin-Dong[1] CHANG Gui-Ran[1]

(School of Information Science and Engineering, Northeastern University, Shenyang 110004)[1]

(Software Institute, Tinshua University, Beijing 100081)[2]

Abstract With the development of Grid technology, agent technology has been utilized in the implementation of Grids. The distributed autonomous infrastructure of multi-agent can achieve better real-time response than the centralized non-autonomous ones, and it is very suitable to dynamic scheduling. Because Globus has provided most of the functions of a Grid operating system, our system is based on Globus. This system not only makes full use of and integrates with the functions of Globus, but also utilizes artificial intelligence and machine learning methods and adopts the strategies of QoS, to achieve efficient scheduling of Grid tasks with certain levels of QoS assurance. JATLite is used to build the system. This system provides a platform with which a dynamic task scheduling system satisfying the application requirements can be created quickly and efficiently.

Keywords Grid, Agent, Multi-agent, JATLite

1 引言

网格彻底地改变了计算机和数据的访问方式,将成为下一代分布式计算的体系结构标准,它提供从孤立的系统到紧密结合的簇、企业范围内聚簇及地理上分散的计算机环境之间联系的途径[1]。网格技术使用户无论在何时何地都能透明地访问计算和存储资源,并保证一定的服务质量成为可能[2]。由于网格环境的异构性和复杂性,为网格资源提供透明的无缝的并且稳定可靠的服务成为一大难题。代理技术多年前就已经被用于解决计算机负载均衡的问题,近几年代理技术也开始被应用在网格的实现当中[3]。

在网格体系结构中,由下至上的四个管理层分别是:构造层,连接层,资源层,汇聚层。资源层包括计算资源、存储资源、网络资源、代码库等等;连接层处理简单的安全通讯,提供单点登录、代理,与不同的本地安全方案的结合以及基于用户的信任关系;资源层关注于单个资源,资源层协议的两个基本类就是信息协议和管理协议;汇聚层提供目录服务、协同分配、服务发现和调度、监控和诊断服务、数据复制服务、基于网格的编程系统,等等。

在广域分布的系统中实现汇聚资源的协调是复杂的高层任务,服务的发现、协同分配,及其监控、诊断、授权等等这些过程需要智能的、自治的和社会性的能力,智能代理都满足这些特性[4]。服务的可扩展性问题是网格的一个关键问题,资源管理、协同分配和调度等这些服务都必须考虑到系统的可扩展性问题[5]。基于代理的技术是能够提供可扩展性和自适应性服务的最有前景的一种方法。本文将多代理的体系结构用于组织和协调网格环境中的分布式资源调度,讨论了基于多代理的网格任务调度的思想、设计和实现方法。

*)基金项目:高等学校博士学科点专项科研基金(20030145017)。曾万聃 博士生。

参 考 文 献

1 汪斌强,王建东.高性能路由器技术体系、关键问题及发展趋势.电信科学,2003(10)

2 Wang Jun, Klara Nahrstedt Parallel IP Packet Forwarding for Tomorrow's IP Routers. In: Proc. of 2001 IEEE Workshop on High Performance Switching and Routing (HPSR'01), Dallas, TX, May 2001. 353～357

3 肖萍萍,周芳.SDH原理与技术.北邮出版社,2002

4 Simpson W. PPP in HDLC-like Framing. STD 51, RFC 1662, July 1994

5 Xilinx. Virtex-II Pro Platform FPGA Handbook. 2003

6 Altera. Stratix Device Handbook. 2003

7 Cypress. Ayama 10000 Network Search Engine. April 2003

8 Netlogic. Nse5000 Network Search Engine Device. Nov. 2002

9 Waldvogel M, Varghese G, et al. Scalable High-Speed IP Routing Lookups. Procedures ACM SIGCOMM, 1997. 25～36

10 徐恪,吴建平,吴剑.基于TCAM的高速路由查找.小型微型计算机系统增刊(CERNET 2001学术年会)

11 Gupta P. Algorithms for routing lookups and packet classification: [PHD thesis]. http://klamath.stanford.edu/～pankaj/thesis/thesis 1sided.pdf, Dec. 2000