下一代光核心负载均衡分组交换机研究

夏 羽[1),2)] 曾华燊[1)] 高志江[1)] 申志军[1)]

[1)](西南交通大学信息科学与技术学院 成都 610031)

[2)](纽约大学理工学院电子与计算机工程系 美国 纽约)

摘 要 着重研究了在第五代交换机体系结构中极具竞争力的以阵列波导光栅路由器件为核心的负载均衡交换结构(AWGR-LB).负载均衡结构通常使用严格的时分复用方式,调度简单但性能不够理想;AWGR 的使用能够大大提高交换机整体容量,但由于其波长速率仅为端口速率的 $1/N$ (其中 N 为端口数),其时延性能与传统交换矩阵相比仍有较大差距,因而不能提供良好的服务质量保证.文中提出了一种适应性时分复用调度算法,能够在保持 100%吞吐率优势的同时,极大地改善负载均衡结构的时延性能;同时文中提出的双波导光栅路由器方案,使光交换矩阵能模拟传统交叉开关的工作模式,从而达到以端口线速率交换的效果,与适应性时分复用调度算法相结合,可进一步提高 AWGR-LB 的性能.

关键词 分组交换;负载均衡交换机;Byte-Focal 交换机;时延性能;吞吐率

中图法分类号 TP393 **DOI 号**: 10.3724/SP.J.1016.2011.01332

On Next Generation Optic-Core Load-Balanced Packet Switch

XIA Yu[1),2)] ZENG Hua-Xin[1)] GAO Zhi-Jiang[1)] SHEN Zhi-Jun[1)]

[1)](*School of Information Science & Technology, Southwest Jiaotong University, Chengdu* 610031)

[2)](*Department of Electrical & Computer Engineering, Polytechnic Institute of New York University, New York, United States*)

Abstract This paper focuses on the Arrayed Waveguide Grating Router-Load Balanced (AWGR-LB) switch, a promising candidate for the fifth generation packet switching architecture. The LB switch needs little scheduling but its performance is not very satisfactory for its strict time-division multiplexing manner. AWGR-LB can greatly increase the switching capacity; nevertheless, the transmission delay within the switch fabric is too large compared with traditional ones for its channel rate within the fabric can only be implemented at one N-th of the port rate (N is the switch size), hence makes QoS unsatisfactory. This paper introduces an Adaptive TDM (ATDM) scheduling manner, which dramatically decreases the delay of LB switches while still keeping 100% throughput. By using the authors' dual AWGR design, which can emulate the operation of the traditional crossbar and provide port-rate switching, combined with the ATDM scheme, the performance of AWGR-LB can be further improved.

Keywords packet switch; load-balanced switch; Byte-Focal switch; delay performance; throughput

---

收稿日期:2010-04-24;最终修改稿收到日期:2010-12-08.本课题得到国家自然科学基金(60773102)、"中国工程科技中长期发展战略研究"联合基金(U0970122)和四川大学基金(下一代 Internet 体系结构)资助.**夏 羽**,男,1983 年生,博士研究生,主要研究方向为高速交换结构、交换调度算法、高速交换结构的设计及实现.E-mail: rainsia@gmail.com.**曾华燊**,男,1945 年生,教授,博士生导师,主要研究领域为网络体系结构、高速交换结构、路由器测试技术.**高志江**,男,1985 年生,博士研究生,主要研究方向为高速交换结构的设计及实现.**申志军**,男,1976 年生,博士研究生,主要研究方向为高速交换结构的设计及实现.

1 引　言

随着以 DWDM(Dense Wavelength Division Multiplexing)为代表的光传输技术的飞速发展,光纤中单波长的传输速率已经超过 100Gbps,甚至还高于现阶段骨干网的传输速率.另一方面,由于受商用存储器访问速率的限制,分组交换机(packet switch)/路由器(router)的发展已经远远落后于传输技术.事实上,分组交换机/路由器已经成为现代网络性能的瓶颈.能够适应未来高速数据传输的高性能分组交换机的设计与实现是下一代网络研究的关键问题之一.

商用交换机一共经历了四代.第一代交换机使用和普通计算机相同的架构:用集中式的处理器和存储器来进行 IP 查表、校验和(checksum)计算以及交换等操作.随着计算机网络的广泛应用,集中式的处理器逐渐不能匹配日益提高的传输速率.因此第二代交换机在每一块线卡(网卡)上使用独立的处理器来进行查表、校验和计算等操作,但仍然使用集中式的处理器和存储器来进行交换操作.第一、二代交换机均使用集中式的存储器来存储分组,属于共享存储(Shared-Memory,SM)结构.随着网络传输速率的持续提高和对服务质量的更高要求,共享存储结构已经不能满足高速网络对交换机容量的需求.

第三代交换机引入了交换矩阵(switch fabric),并在线卡上设置独立的缓存以存储暂时不能被转发的分组.最初的第三代交换机使用输出排队(Output-Queued,OQ)方式:分组到达后立刻通过交换矩阵到达对应的输出端口,并进入相应队列中排队等待输出.但是由于交换矩阵和存储器都需要以 N 倍(N 为交换机的端口数)于线卡的速率工作,随着端口速率的提高以及交换规模的扩大,这种交换结构开始变得不现实.而如果分组在到达输入端口时,先在输入端的线卡中排队等待,并且一个时槽内该线卡只通过交换矩阵传输一个分组,则交换矩阵和存储器仅需要以和线卡相同的速率工作,这种交换方式被称为输入排队(Input-Queued,IQ)方式.由于输入排队可以适应更大的交换容量,因而被广泛使用.但是 IQ 结构由于受到队首阻塞(Head-of-Line Blocking,HoL blocking)问题的限制,其吞吐率仅达到 58% 左右[1].虚拟输出队列(Virtual Output Queue,VOQ)[2]作为解决队首阻塞问题的有效方案,被广泛地应用于商用交换机中.但是 VOQ 的引入加剧了输出竞争,需要使用额外的匹配调度算法.典型的调度算法有并行迭代匹配算法 PIM(Parallel Iterative Matching)[3]、带滑动指针的轮询算法 iSLIP(iterative Round-Robin with SLIP)[4]以及双轮询匹配算法 DRRM(Dual Round-Robin Matching)[5]等.实用的 IQ 调度算法大多需要迭代多次才能达到较为理想的时延,且在非均匀流量下吞吐率性能不佳.但随着网络传输速率的持续提高以及交换机规模的持续扩大,留给这些调度算法运行的时间已经越来越紧迫,IQ 交换结构最终将不能适应未来的高速网络.

第四代交换机以多机柜架构的出现为标志:由于交换机的规模不断扩大,多级多平面交换机开始出现,使用单一的机柜已经不能满足交换机在体积、功耗及散热等方面的要求.现代交换机多将交换背板放置于一独立的机柜中,而将线卡分别放置于其它的机柜中,这些线卡和背板之间使用光纤连接,如图 1 所示.多机柜架构交换机的典型代表有:Alcatel 7670 RSP、Juniper TX、Avici TSR 和 Cisco CRS-1 等.

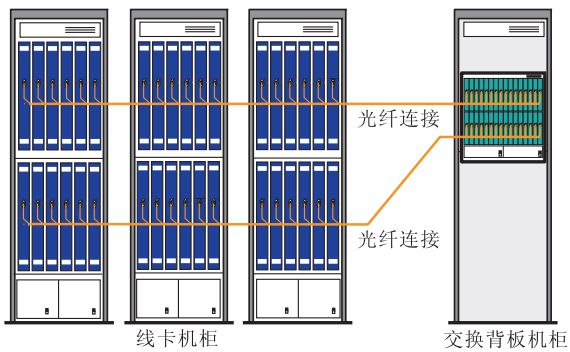


图 1　多机柜架构的交换机

本文的主要研究对象是面向未来高速网络的第五代交换机,第五代交换机以光交换核心的使用为标志.典型代表是斯坦福大学项目[6],该项目首先提出了使用阵列波导光栅路由器(Arrayed Waveguide Grating Router,AWGR)作为核心交换矩阵,结合负载均衡(Load-Balanced,LB)的交换架构[7],可以构建容量高达 100Tbps 的交换机,但是,由于使用的调度算法性能不够理想,导致其时延性能非常差.香港大学项目[8]也采用 AWGR 扩展 LB 交换机,正像文献[8]所说,由于工艺原因,AWGR 中每个信道速率仅为传统交换矩阵的 $\frac{1}{N}$,因此需要使用传统交换矩阵中 N 倍的时间来传送一个分组,这也将导致时延性能非常差.

针对以上两个问题,本文中,我们首先提出了适应

性时分复用(Adaptive Time-Division Multiplexing, ATDM)的调度方式,从而较大地改善了 LB 结构时延性能,同时可以证明 ATDM 方式能够保持 LB 交换机 100%吞吐率的优势.此外,为了配合 ATDM,我们提出了一种使用双 AWGR 串联方式来模拟传统交叉开关的可行方案,我们将其称为通过双 AWGR 模拟的交叉开关(Dual AWGR Emulated Crossbar,DAWGREC)模块.DAWGREC 的主要优势是可以以端口速率传输分组,从而避免 N 倍传输时间的问题,而这也将进一步改善 AWGR-LB 交换机的时延性能.仿真实验表明新的 LB 交换机在各种流量下均有良好的时延性能.

本文第 2 节主要讨论 LB 交换机的研究现状;第 3 节介绍 ATDM 调度方式以及采用 ATDM 调度的交换机体系结构;第 4 节证明使用 ATDM 的负载均衡交换机可以在任意许可流量下达到 100%的吞吐率;第 5 节介绍如何扩大 AWGR 交换机容量;第 6 节用仿真实验说明新 LB 交换机在各种流量下均有很好的性能;第 7 节总结全文.

2 负载均衡的分组交换机研究现状

由于 LB 交换机调度简单,可以适应高速网络,近 10 年来,它一直是交换机研究的热点.LB 交换结构是一种特殊的两级交换结构,最简单的 LB 结构如图 2 所示.它由两个 $N\times N$ 的交叉开关组成,之间设置 N 个缓存用于暂存从第一级交叉开关转发过来的信元(cell,即定长分组,若分组为变长,则在进入到交换机之前将其切分),这些缓存被组织成为 VOQ 的形式,每一组 VOQ 连接一个第一级交叉开关的输出端口和一个第二级交叉开关的输入端口,因此共有 N^2 个 VOQ.

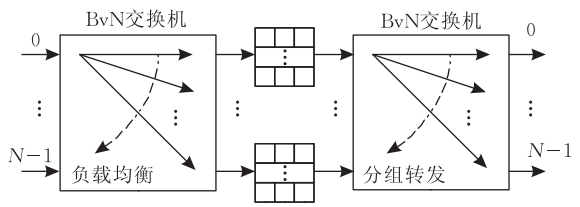


图 2 负载均衡交换结构

和 IQ 交换结构不同,LB 交换结构中的交叉开关不需要使用调度算法来配置,而是周期性地,在预定的 N 种连接模式之间顺序切换.通过选择恰当的连接序列,N 个时槽内,交叉开关的各输入端和输出端将各连接一次.如果用 i 表示交叉开关的输入端,j 表示交叉开关的输出端,而用 t 表示当前时槽,则一种可行的连接模式为 $j=(i+t) \bmod N$.这种连接模式称为时分复用(Time-Division Multiplexing,TDM)方式,而使用 TDM 方式的交换机在文献[7]中被称为 BvN(Birkhoff-von Neumann)交换机.

在 LB 交换结构中,第一级交换矩阵作为一个负载均衡器,将到达的信元均匀地发送到各个中间级,而第二级交换矩阵才真正将信元转发到正确的端口.由于第一级负载均衡器的作用,第二级 BvN 交换机仅需要处理均匀业务流,而 BvN 交换机可以均匀地服务各个端口.文献[7]证明了 LB 交换结构可以在弱混合流量下达到 100%的吞吐率.

虽然 LB 结构有以上优势,但是它存在两个严重的缺陷:(1)如果将“流(flow)”定义为从同一端口进入,而指向同一输出端口的分组序列,则在 LB 结构中,属于同一个流的分组离开的顺序会和其到达顺序不同,这被称为分组失序问题;(2)其时延性能和传统交换机相比有较大差距.

文献中有多种方法解决 LB 结构的分组失序问题,这些方法主要可以分为 3 类:(1)在分组到达输出端口之前就阻止分组失序,代表算法及结构有:UFS(Uniform Frame Spreading)算法[9]、满帧填充 PF(Padded Frame)算法[10]、Mailbox 结构[11]和基于反馈的(Feedback-Based)结构[12];(2)在分组到达输出端口时将失序度控制在一定范围内,然后再在输出端设置一定大小的维序缓存(resequencing buffer)将失序的分组重新排序输出.其代表有 FOFF(First Ordered Full Frame First)[6]算法及 Byte-Focal 结构[13];(3)组合方法,即将几种方法组合使用,典型的代表是 CR(Contention and Reservation)交换机[14],它结合了 Mailbox 和 UFS 两种方法.

分组失序的根本原因是属于同一个流的不同分组在经过不同的中间级时所经历的时延不同.第 1 类算法保证分组通过不同的中间级时能经历相同的时延.UFS 算法使用逐帧(frame-by-frame)转发的概念,一个帧由 N 个连续的属于同一个流的信元组成.如果每个输入端口的信元都按帧的方式,在 N 个连续的时槽内被顺序转发到各个对应的 VOQ 中,则每个中间级对应的 VOQ 的队长都相等,从而同一个帧的分组在中间级所经历的时延也相同.但是在中低负载下,一个 VOQ 可能需要等待较长时间才能构成一个帧,因此 UFS 算法的时延性能较差.

满帧填充(PF)算法在 VOQ 没有满帧时,通过

填充“空分组”凑成满帧,这样虽然可以减少等待成帧的时间,但是其代价是使用过多的带宽来转发空分组,从而造成带宽浪费,时延性能虽较 UFS 有所提高,但是改进仍然不明显.

Mailbox 交换结构使用竞争中间级队列的方式,并采用了对称式 TDM 连接模式巧妙地提供了竞争成功与否的反馈路径,使得在低负载下时延性能得到较大改善.但竞争失败的信元将被阻塞,因此不能达到 100%的吞吐率.当 $\delta=0$ 时,该交换结构只能达到 58%左右的吞吐率[11].

基于反馈的(feedback-based)交换结构类似于 Mailbox,它提供了同时具有 staggered symmetry 和 in-order packet delivery 两种属性的连接模式来交换中间级队列的占用信息和解决分组失序的问题.FB 是目前时延性能最好的 LB 交换机,但由于其每个中间级 VOQ 仅缓存一个分组,因此各个输入端存在竞争冲突,从而不能达到 100%的吞吐率,事实上,它被证明在 2 倍加速比下才能达到 100%的吞吐率[12].

FOFF 算法允许非满帧 VOQ 发送非满帧分组.虽然可以在很大程度上改进 UFS 算法的性能,但是在中低负载下,发送的分组大部分属于非满帧分组,失序仍然可能发生,因此在输出端需要维序队列来重组失序的分组,维序不但需要额外的通信量,而且也造成 FOFF 的时延性能依旧较差.

CR 交换机巧妙地结合了 Mailbox 和 UFS 两种方法,在低负载下使用 Contention 模式(即 Mailbox 结构)来获得较好的时延性能,而在高负载下时使用 Reservation 模式(即 UFS 算法)来保证 100%的吞吐率.CR 交换机在中高负载下,时延性能仍然非常差.

在所有解决方案当中,Byte-Focal 是唯一一种不需要在端口和中间级之间交换任何信息,且时延性能较好,同时也能保持 100%吞吐率的结构.它使用逐分组(packet-by-packet)的方式来解决失序问题.Byte-Focal 结构在第一级和中间级均使用 VOQ,而在输出端采用虚拟输入队列(Virtual Input Queue,VIQ)集合的结构,并加入流首(Head-of-Flow)指针,巧妙地利用了分组经过中间级的顺序,解决了分组失序问题.有关 Byte-Focal 的详细描述可以参考文献[13].

3　使用轮询调度加强的负载均衡结构

以上这些方案虽然都解决了 LB 交换结构分组失序的问题,且时延性能都有所提高,但并没有做出根本性的改善.

我们深入分析后发现,LB 结构时延性能低下的主要原因是 TDM 调度存在严重的“连接浪费”问题.所谓“连接浪费”是指由于 TDM 完全按照预定的连接模式建立连接,当某端口对之间建立连接之后,可能并没有信元传输.从而在其它队列均为空的情况下,一个等待传输的信元,最坏情况下将被延迟 $N-1$ 个时槽.若两级 BvN 都存在连接浪费,情况将更糟.

基于以上分析,我们认为完全抛弃匹配算法的方式虽然调度简单但会导致极差的性能,并不可取,因此我们提出了适应性的 TDM(ATDM)调度.ATDM 在传统 TDM 的基础上使用复杂度仅为 $O(1)$的带滑动指针的轮询(Round-Robin)调度来消除 BvN 结构的“连接浪费”,改进后的 BvN 结构称为使用轮询调度加强的 BvN(Round-Robin Enhanced BvN,RRE-BvN)结构.RRE-BvN 结构的工作过程如下:时槽开始时,各输入、输出端口先根据 TDM 方式结合实际的队列情况建立匹配,对应队列为空的端口不匹配,随后进行经典的 1-SLIP 三步调度[4],即

(1)请求阶段.所有未匹配的输入端口,向所有不为空的 VOQ 所对应的输出端发送请求.

(2)授权阶段.所有未匹配的输出端口,收到请求后,从中选择一个在顺时针方向离授权指针最近的输入端口的请求进行授权.如果该授权在下一步被接受,则授权指针指向被授权的输入端口的下一个端口;否则授权指针不移动.

(3)接受阶段.输入端收到授权之后,从中选择一个在顺时针方向离接受指针最近的输出端口接受其授权.然后将接受指针移动到该输出端口的下一个端口.

我们将用 RRE-BvN 替换传统的 BvN 后所建立的新 LB 结构称为使用轮询调度加强的负载均衡(Round-Robin Enhanced LB,RRE-LB)交换结构,如图 3 所示.一方面,TDM 方式使得在调度开始就存在一些匹配,从而无需迭代即可建立比 1-SLIP 更多的匹配;另一方面,轮询调度所建立的匹配可以有效消除 BvN 结构的“连接浪费”.且轮询算法的复杂度仅为 $O(1)$且实现简单[15].

除替换 BvN 结构以外,RRE-LB 和 Byte-Focal 类似,在中间级队列采用 VOQ,而在输出端使用 VIQ 集合结构来保证分组以正确顺序离开.VIQ 集合结构如图 4 所示,其工作方式是将属于同一个流

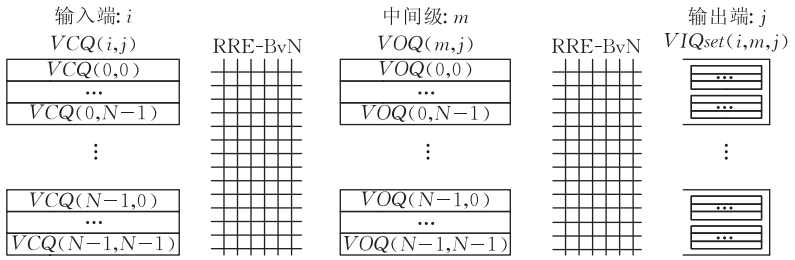


图 3 轮询调度加强的负载均衡结构

(即从相同输入端口到达且指向同一个输出端口)但是经过不同中间级的分组入队到不同的 VIQ 中,然后为每一个流维护一个流首指针(每个端口各 N 个),该指针从第一个 VIQ 开始,每次 VIQ 中属于该流的分组离开后,对应指针移向下一个 VIQ. VIQ 集合的细节可以参看文献[13].

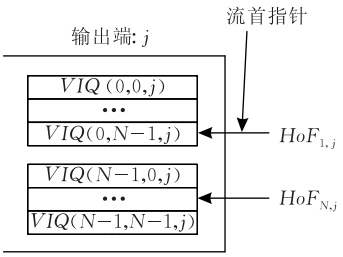


图 4 VIQ 集合结构

为了配合 ATDM 调度方式,同时也降低调度复杂度,我们抛弃了 Byte-Focal 的第一级需要复杂调度的 VOQ 结构[13],而在 RRE-LB 第一级使用虚拟中间级队列(Virtual Central Queue,VCQ). VCQ 结构如图 5 所示,其工作方式如下:从输入端口 i 中到达而指向输出端口 j 的流的信元被以 $VCQ_{i,0}$, $VCQ_{i,1}$,…,$VCQ_{i,N-1}$,$VCQ_{i,0}$,…的顺序均匀地分配到 N 个 VCQ 中,而在 $VCQ_{i,m}$ 中的分组一定被转发到交换矩阵的第 m 个出口.

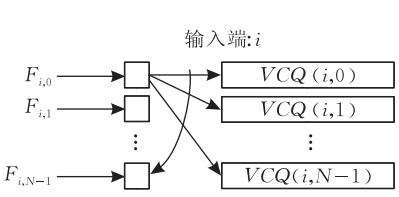


图 5 VCQ 结构

性质 1. 在 RRE-LB 结构中,属于同一个流的分组离开的顺序和到达顺序一致.

证明. 根据 VCQ 的工作原理,属于同一个流的分组被分配到中间级的顺序一定满足:$m=0,1,2,\cdots,N-1,0,1,2,\cdots,N-1,\cdots$. 虽然这些分组通过中间级的排队时延可能不同,但是在输出端,由于有 VIQ 集合及流首指针的限制,同一个流中离开的分组所经过的中间级顺序如下:$0,1,2,\cdots,N-1,0,1,2,\cdots,N-1,\cdots$. 因此,属于同一个流的分组离开 RRE-LB 结构的顺序一定与其到达顺序相同. 证毕.

4 稳定性分析

定义 1(稳定性,stability). 对于一个排队系统,如果其中的队列的平均长度均有上界($<\infty$),则称该排队系统稳定. 而一个稳定的交换系统可以达到 100%的吞吐率,即所有进入该系统的分组都可以在有限时间内被转发离开.

定义 2(许可流量,admissible traffic). 一个 $N\times N$ 的交换机中,若用 $\lambda_{i,j}$ 表示从输入端口 i 到达而指向输出端口 j 的流的平均到达速率,如果该流量满足如下两式:

$$\sum_{i=0}^{N-1}\lambda_{i,j}\leq 1,\quad \forall j=0,1,\cdots,N-1;$$

$$\sum_{j=0}^{N-1}\lambda_{i,j}\leq 1,\quad \forall i=0,1,\cdots,N-1,$$

则该流量称为许可流量.

引理 1. 使用 VOQ 的 RRE-BvN 交换结构可以对任意 $\lambda_{i,j}\leq 1/N$ 的流量提供 100%的吞吐率.

证明(反证法). 对 RRE-BvN 中任意一个 $VOQ_{i,j}$,假设其不稳定,则其队长会无限增长. 但是由 RRE-BvN 结构的特点,一个队列不为空时,TDM 调度在 N 个时槽的周期内一定服务该队列一次;而 1-SLIP 调度在其它队列为空的情况下还可能服务该队列一次或以上,则在 t 个时槽内,该队列会被服务至少$\lfloor t/N\rfloor$次,则从长时间来看,服务速率

$$d_{i,j}\geq\lim_{t\to\infty}\frac{\lfloor t/N\rfloor}{t}=1/N.$$

由排队论可知到达速率小于服务速率的队列应该稳定,这与假设矛盾. 证毕.

引理 2. 在许可流量下,RRE-LB 交换结构的第一级可以达到 100%的吞吐率.

证明. 如图 6 所示,将从输入端口 i 进入而指

向输出端口 j 的流的平均到达速率用 $\lambda_{i,j}$ 来表示,而实际进入队列 $VCQ_{i,m}$ 的流的平均速率用 $\lambda'_{i,m}$ 来表示. 根据 VCQ 的工作原理有

$$\lambda'_{i,m}=\frac{1}{N}\lambda_{i,0}+\frac{1}{N}\lambda_{i,1}+\cdots+\frac{1}{N}\lambda_{i,N-1}=\frac{1}{N}\sum_{k=0}^{N-1}\lambda_{i,k},$$

根据许可流量定义 2 有 $\lambda'_{i,m}\leq\frac{1}{N}$.

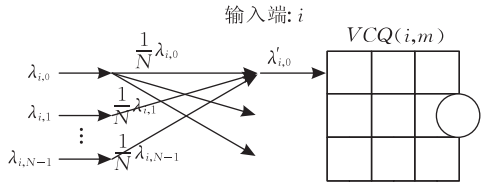


图 6　到达第一级 VCQ 的流量速率

由引理 1 可知,RRE-LB 的第一级可以达到 100%的吞吐率.　　证毕.

引理 3.　在任意许可流量下,RRE-LB 交换结构的第二级可以达到 100%的吞吐率.

证明.　根据引理 2 可知,所有到达的流量都会被完全转发到第二级. 我们将进入中间级 $VOQ_{m,j}$ 的流量的平均到达速率用 $\lambda''_{m,j}$ 表示. 如图 7 所示,由于从输入端口 i 到达而指向输出端口 j 的流量在第一级被均匀地分配到 N 个 VCQ 中,因此到达第二级输入端口 m,而指向输出端口 j 的流量的平均到达率为 $\frac{1}{N}\lambda_{i,j}$. 又由于每个输入端口都有指向某输出端口 j 的流量,因此总体 $VOQ_{m,j}$ 的平均到达速率:

$$\lambda''_{m,j}=\sum_{l=0}^{N-1}\frac{1}{N}\lambda_{l,j}.$$

由许可流量的定义 2 有 $\lambda''_{m,j}\leq 1/N$. 由引理 1 可知,RRE-LB 的第二级可以达到 100%的吞吐率.

证毕.

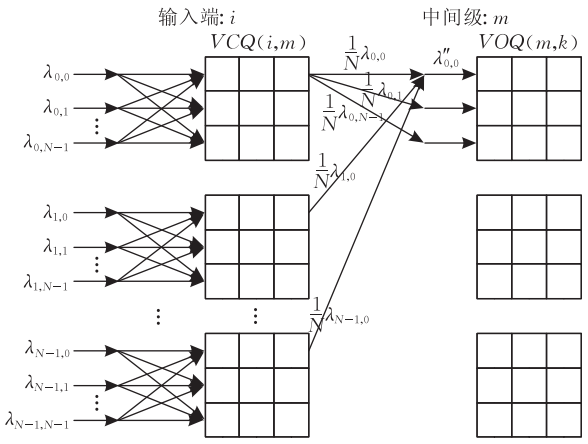


图 7　到达第二级 VOQ 的流量速率

定理 1.　RRE-LB 可以在任意许可流量下达到 100%的吞吐率.

证明.　由引理 2 和引理 3 可知 RRE-LB 结构的第一级和第二级均可达到 100%的吞吐率,因此最终 RRE-LB 可以在任意许可流量下达到 100%的吞吐率.　　证毕.

值得注意的是,在文献[7]中,为了证明 100%的吞吐率,作者假设流量符合弱混合(weak mixing)特性;而在我们的证明过程中,对流量的唯一要求仅仅是符合许可流量,而这正是所有稳定的交换结构对流量的必要约束. 因此 RRE-LB 解除了传统 LB 结构对流量弱混合的要求.

5　使用 AWGR 扩展 RRE-LB 结构

5.1　AWGR 概述

阵列波导光栅路由器(Arrayed-Waveguide Grating Router, AWGR)[16]是一种被动光交换器件. 如图 8 所示,一个输入端的光信号如果通过不同的波长传播,则会被转发到 AWGR 的不同输出端口. 具体来说,对于一个 $N\times N$ 的 AWGR,输入端口 i 传输 N 种波长的光信号,分别用 $\lambda_0^i,\lambda_1^i,\cdots,\lambda_{N-1}^i$ 表示,则输入端口 i 通过波长 λ_j^i 传输的光信号,会被转发到输出端口 $(i+j)\bmod N$. 简单来说,$N\times N$ 的 AWGR,每个端口使用 N 种波长,可以提供输入端和输出端之间的全连接,共 N^2 条信道.

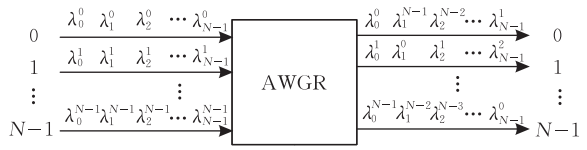


图 8　阵列波导光栅路由器

AWGR 有如下优势:(1)造价低;(2)零功耗;(3)端口间的全连接;(4)无需配置. 但是它也存在如下缺点:(1)每个信道传输速率仅为 $\frac{1}{N}R$,而传统交叉开关的信道速率为 R,因此 AWGR 需要用原来 N 倍的时间来传输一个信元;(2)由于 AWGR 本身的特点,其规模不能过大,一般来说,超过一百个端口的 AWGR 是非常少见的[16].

5.2　提高交换速率

虽然 AWGR 可以提供输入端口和输出端口之间的全连接,但是由于技术和工艺原因,目前实验室内部也只实现了最高 40Gbps 的 AWGR,市场上可以使用的 AWGR 速率将更低. 对于 100Gbps 以上的网络,如果使用高速 AWGR,肯定会造价过高,但如果每一个信道只能以 $\frac{1}{N}$ 端口速率的低速工作,则

又会造成 N 倍传输时延.本小节我们提出一种使用两个低速 AWGR 串联,能够以线速率 R 模拟任意交叉开关连接模式的交换矩阵,我们将其称为通过双 AWGR 模拟的交叉开关(DAWGREC)模块,以低廉的实现成本达到高性能.

如图 9 所示,DAWGREC 由两个 AWGR 串联而成,两个 AWGR 之间用可变换波长的激光发射器件(Tunable-Wavelength Laser Transmitter,TWLT)连接.当长度为 L 的信元到达输入端口时,将被平均分为 N 个长度为 L/N 的分片(segment),这些分片被均匀地发送到每一个中间级,而通过配置 TWLT,即可在这些分片到达第二个 AWGR 的入口时变换为正确的波长,从而被转发到正确的目的地,并在第二个 AWGR 的出口被重新组合成信元.这样转发一个信元的时间仍然为一个时槽.值得注意的是,由于从同一个输入端口被均匀发送到各个 TWLT 的分组最终的目的端口都一样,因此其波长映射都一样,从而需要的波长映射实际上只有 N 种而不是 N^2 种.

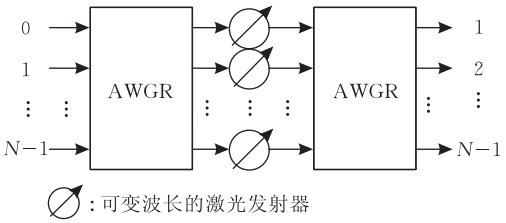


图 9 通过两个 AWGR 模拟交叉开关示例

为了便于理解 DAWGREC 的原理,我们在图 10 中显示了两组通过 DAWGREC 模拟 4×4 的交叉开关连接的示例.

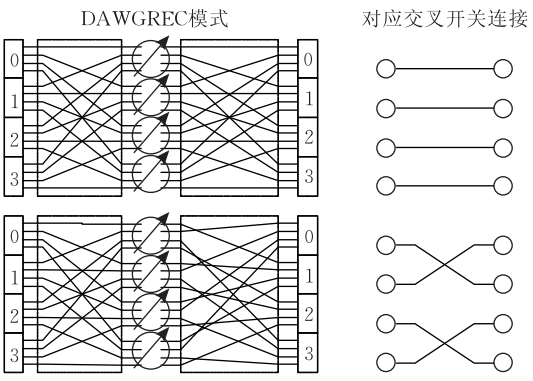


图 10 DAWGREC 模拟交叉开关示例

5.3 扩大交换规模

由于 AWGR 要达到一百个以上的端口非常困难,因此通过 DAWGREC 来实现的 RRE-BvN 也不可能支持过大的交换规模.本节我们使用多个较小规模的 DAWGREC 模块互联从而扩展为较大的规模.

具体来说,我们使用 4 个 $N/2\times N/2$ 的 DAWGREC 模块构造一个 $N\times N$ 的 DAWGREC 模块.将要到达 N 个输出端口的信元分为两部分,一部分指向前 $N/2$ 个输出端口,另一部分指向后 $N/2$ 个输出端口.这 4 个 DAWGREC 模块通过图 11 的方式连接.通过配置这 4 个 DAWGREC 模块中的 TWLT,即可以模拟 $N\times N$ 的交叉开关的任意连接模式.

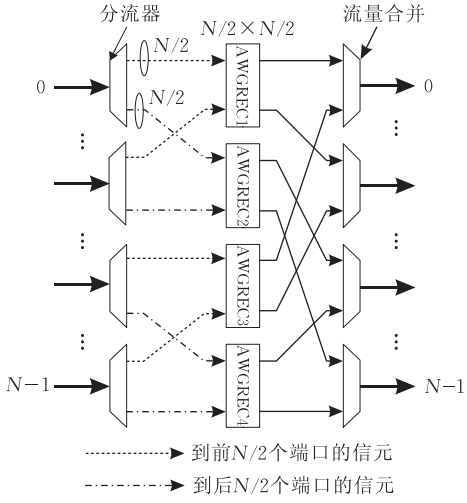


图 11 DAWGREC 规模的扩展

一般的,可以使用 k^2 个 $n\times n$ 的 DAWGREC 矩阵扩展成为一个 $kn\times kn$ 的 DAWGREC 模块.而又可以将合成的 DAWGREC 模块作为基础模块再次进行扩展.

因此,如果需要更大的交换规模,则可以使用扩展的 DAWGREC 模块再次通过上面的方法组合.通过多次扩展,理论上可以支持任意规模,以适应下一代网络对更大交换规模的需求.

6 仿真实验

吞吐率和时延是评价交换机好坏的两个重要指标,前面我们已经证明了 RRE-LB 交换机可以保证 100%的吞吐率.本节我们在为 NS-3(Network Simulator 3)设计的交换机仿真平台[1]上对 RRE-LB 交换机进行仿真,并对比其它负载均衡方案的时延性能.在仿真中,我们设置端口数为 32,对于其它端口数,也有类似结果,限于篇幅,我们不再一一列举.

为了便于对比 RRE-LB 交换机的时延性能,我们在仿真结果中加入了如下交换机和调度算法:

(1)输出排队(OQ).交换机时延性能的最优界,在高速网络中无法实际实现.

(2)5-SLIP.iSLIP 算法,32 端口交换机,文

献[4]推荐 5 次迭代,由于迭代次数较多从而无法用于高速或大规模交换.

(3) 1-SLIP. 仅一次迭代的 iSLIP 算法,复杂度仅为 $O(1)$,和 RRE-LB 具有类似的复杂度.

(4) CR 交换机. 100%吞吐率的 LB 交换机,作为时延性能参考.

(5) 基于反馈的 LB 交换机(FB). 使用 2 倍加速比才能达到 100%的吞吐率,但具有较好的时延.

(6) Byte-Focal 交换机(BF). 目前最实际的 LB 交换机,可以达到 100%的吞吐率,其中性能最好的 LQF 调度算法的时间复杂度为 $O(\log N)$.

在第一组仿真中,我们比较各种交换机在 Bernoulli 和突发均匀流量下的时延性能. 突发流量中停留在 ON 和 OFF 状态的时槽数均服从几何分布. 在 ON 状态下,每个时槽均产生一个分组. 而停留在 ON 状态的时槽数称为突发长度(burst length),用 b 表示. 同一个突发内的分组的目的端口相同,而不同突发之间的指向从 N 个输出端口中均匀地随机选择.

仿真结果如图 12 所示. 可见,随着突发长度的增大,各种算法之间的性能渐渐接近. RRE-LB 由于有第一级负载均衡器的作用,其在高负载下的时延性能远远好于 1-SLIP 算法. 而 CR 交换机虽然在低负载下时延稍小,但是仍然高于 RRE-LB. Feedback-Based 交换机在低负载下与 Byte-Focal 性能相当,而在高负载下,性能较好. 总体来说 RRE-LB 无论在高负载还是低负载下,都能达到很好的性能.

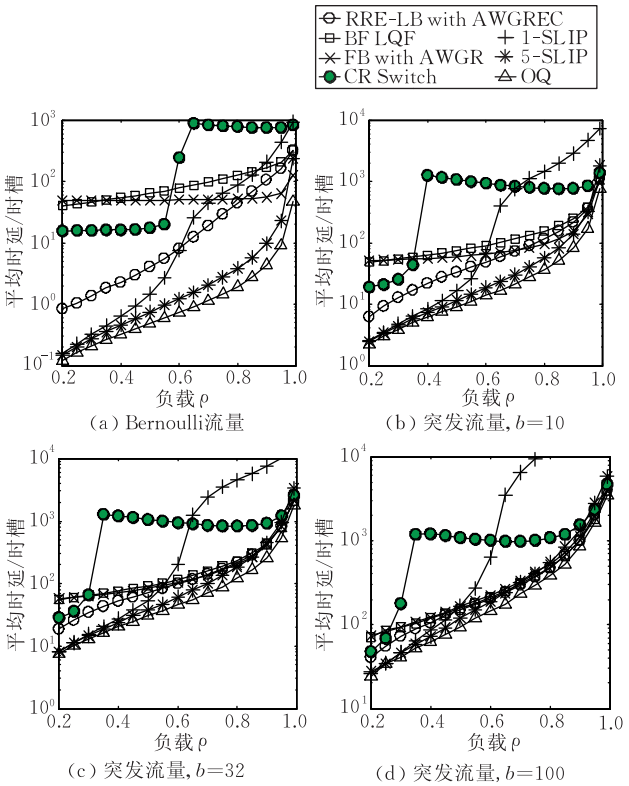


图 12　均匀流量下的时延性能

在下一组仿真中,我们主要考察各种交换结构在 Bernoulli 非均匀流量下的时延性能. 我们使用了 4 种在交换机仿真中广泛使用的非均匀模型:非对称(asymmetric)、对角(diagonal)、热点(hotspot)以及非平衡(unblance). 假设输入端口 i 的流量平均到达率为 λ_i,而用 $\lambda_{i,j}^{A}$、$\lambda_{i,j}^{D}$、$\lambda_{i,j}^{H}$ 和 $\lambda_{i,j}^{U}$ 分别表示从输入端口 i 到达而指向输出端口 j 的流量分别在非对称、对角、热点以及非平衡模型中的速率,则有如下关系成立:

$$\lambda_{i,j}^{A}=a_k\lambda_i,\quad j=(i+k)\bmod N,$$

其中,

$$a_k=\begin{cases}0, & k=0\\ \dfrac{(r-1)}{(r^N-1)}, & k=1,\ \text{而}\ r=f^{-\frac{1}{N-2}}\\ a_1 r^{k-1}, & \text{其它}\end{cases},$$

$$\lambda_{i,j}^{D}=\begin{cases}d\lambda_i, & i=j\\ (1-d)\lambda_i, & j=(i+1)\bmod N\\ 0, & \text{其它}\end{cases},$$

$$\lambda_{i,j}^{H}=\begin{cases}h\lambda_i, & i=j\\ \dfrac{h}{N-1}\lambda_i, & \text{其它}\end{cases},$$

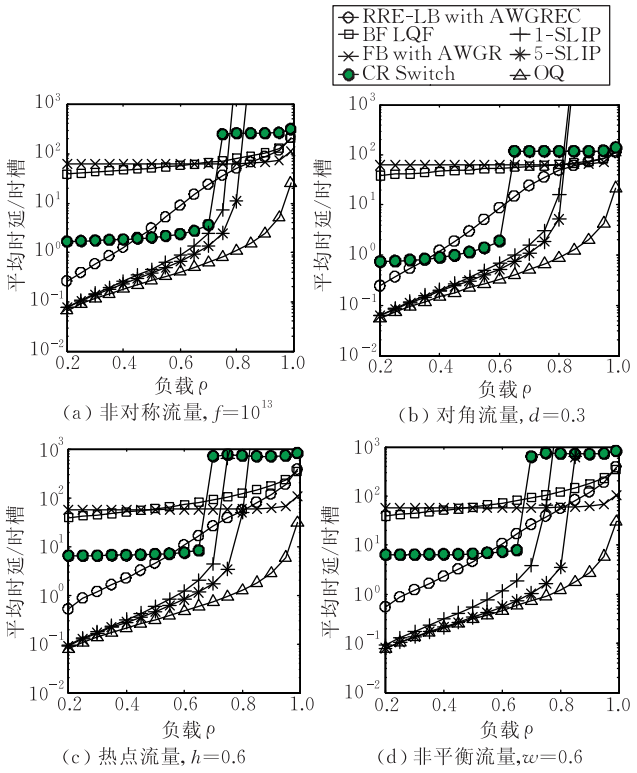
$$\lambda_{i,j}^{U}=\begin{cases}\lambda_i\left[w+\dfrac{1-w}{N}\right], & i=j\\ \lambda_i\dfrac{1-w}{N}, & \text{其它}\end{cases}.$$

其中 $f\in(1,+\infty]$称为非对称系数,$d\in[0,0.5]$称为对角系数,$h\in[0,1]$称为热点系数而 $w\in[0,1]$称为非平衡系数. 在仿真中我们取文献中的典型值 $f=10^{13}$、$d=0.3$、$h=0.6$ 以及 $w=0.6$.

仿真结果如图 13 所示. 可见,由于第一级负载均衡器的作用,RRE-LB 无论在高负载还是低负载下都稳定,且时延性能良好,而具有相同调度复杂度的 1-SLIP 算法在高负载下不稳定. 和其它 LB 结构相比,RRE-LB 的时延性能在非均匀流量下具有较大优势.

文献[18]及其相应的后续研究指出网络中的流量大多为突发流量. 为了更接近实际流量,我们结合突发和非均匀两种流量模型,在下一组仿真中用突发长度为 10 的非均匀流量评估各种交换机的时延性能. 其中突发流量的设置和第一组仿真相同,而非均匀流量的配置和第二组仿真相同.

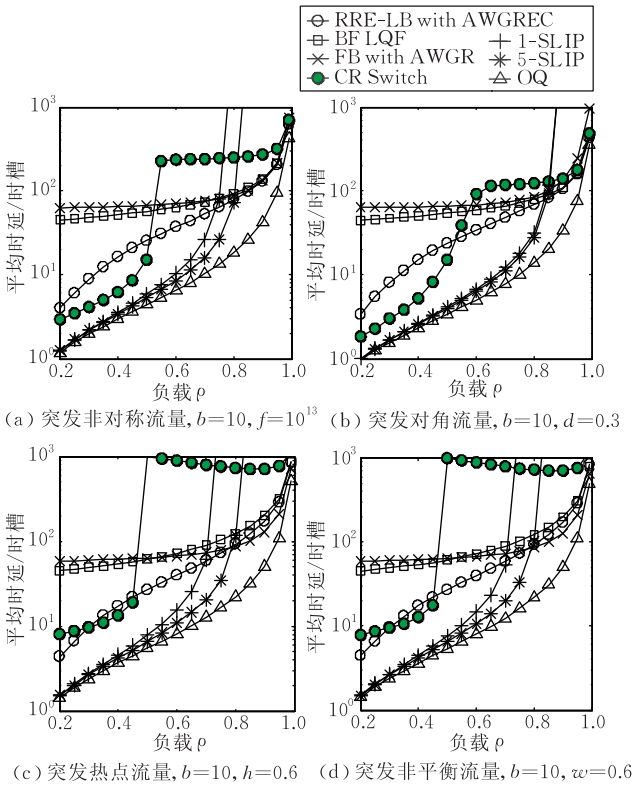
仿真结果如图 14 所示. 在各种 LB 结构中,RRE-LB 的时延性能仍然有较大优势.



(a) 非对称流量, $f=10^{13}$ (b) 对角流量, $d=0.3$

(c) 热点流量, $h=0.6$ (d) 非平衡流量, $w=0.6$

图 13 Bernoulli 非均匀流量下的时延性能



(a) 突发非对称流量, $b=10$, $f=10^{13}$ (b) 突发对角流量, $b=10$, $d=0.3$

(c) 突发热点流量, $b=10$, $h=0.6$ (d) 突发非平衡流量, $b=10$, $w=0.6$

图 14 突发非均匀流量下的时延性能

7 总 结

本文提出了一种时分复用(TDM)和简单轮询调度结合的适应性 TDM(ATDM)调度以替换传统负载均衡(LB)交换机中的 TDM 调度. 新的 LB 交换机称为轮询加强的负载均衡(RRE-LB)交换机,该交换机有效地改善传统 LB 交换机由于 BvN 结构使用严格的 TDM 调度而导致的"连接浪费"问题,从根本上解决困扰 LB 交换机的时延性能问题.

其次,为了适应下一代网络对高速交换的需要,我们在文中通过结合两个阵列波导光栅路由器(AWGR)的方式,使得核心交换器件可以以线速率模拟传统交叉开关的运行模式,这种器件称为通过双 AWGR 模拟的交叉开关(DAWGREC)模块. 通过使用这种模块,可以进一步提高光核心交换机的交换效率,改善时延性能.

最后,由于 AWGR 器件的规模不能过大,而为了适应下一代网络对大规模交换的需要,我们提出了通过多个 DAWGREC 模块互联的方式模拟更大规模的交叉开关的方案. 理论上,通过多次扩展,DAWGREC 模块可以模拟任意规模的交叉开关.

我们相信,不管未来使用何种网络体系结构,分组交换技术仍然是计算机网络的基础. 而能够保持 100% 吞吐率,同时解决了时延性能以及交换规模问题的新 LB 结构一定可以为未来网络提供更好的服务质量保证.

参 考 文 献

[1] Karol M, Hluchyj M, Morgan S. Input versus output queueing on a space-division packet switch. IEEE Transactions on Communications, 1987, 35(12): 1347-1356

[2] Tamir Y, Frazier G. High-performance multiqueue buffers for VLSI communication switches//Proceedings of the 15th Annual International Symposium on Computer Architecture. Honolulu, Hawaii, USA, 1988: 343-354

[3] Anderson T E, Owicki S S, Saxe J B, Thacker C P. High-speed switch scheduling for local-area networks. ACM Transactions on Computer System, 1993, 11(4): 319-352

[4] McKeown N. The iSLIP scheduling algorithm for input-queued switches. IEEE/ACM Transactions on Networking, 1999, 7(2): 188-201

[5] Li Yi-Han, Panwar S, Chao H. On the performance of a dual round-robin switch//Proceedings of IEEE Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2001). Anchorage, Alaska, USA, 2001, 3: 1688-1697

[6] Keslassy I, Chuang S, Yu K, Miller D, Horowitz M, Solgaard O, McKeown N. Scaling internet routers using optics//Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'03). Karlsruhe, Germany, 2003: 189-200

[7] Chang C, Lee D, Jou Y. Load balanced Birkhoff-von Neu-

mann switches, Part I: One-stage buffering. Computer Communications, 2002, 25(6): 611-622

[8] Wang Xin, Yeung K L. Load balanced two-stage switches using arrayed waveguide grating routers//Proceedings of the IEEE Workshop on High Performance Switching and Routing (HPSR 2007). Brooklyn, New York, USA, 2007: 1-6

[9] Keslassy I. The load-balanced router[Ph. D. dissertation]. Stanford University, Stanford, California, 2004

[10] Jaramillo J, Milan F, Srikant R. Padded frames: A novel algorithm for stable scheduling in load-balanced switches. IEEE/ACM Transactions on Networking, 2008, 16(5): 1212-1225

[11] Chang C-S, Lee D-S, Shih Y-J, Yu C-L. Mailbox switch: A scalable two-stage switch architecture for conflict resolution of ordered packets. IEEE Transactions on Communications, 2008, 56(1): 136-149

[12] Hu Bin, Yeung K L. Feedback-based scheduling for load-balanced two-stage switches. IEEE/ACM Transactions on Networking, 2009, 18(4): 1077-1090

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

[14] Yu C, Chang C, Lee D. CR switch: A load-balanced switch with contention and reservation. IEEE/ACM Transactions on Networking, 2009, 17(5): 1659-1671

[15] Chao J. Saturn: A terabit packet switch using dual round robin. IEEE Communications Magazine, 2000, 38(12): 78-84

[16] Ngo H, Pan D, Qiao C. Constructions and analyses of non-blocking WDM switches based on arrayed waveguide grating and limited wavelength conversion. IEEE/ACM Transactions Networking, 2006, 14(1): 205-217

[17] Xia Yu, Zeng Hua-Xin, Shen Zhi-Jun. Design and implementation of switch module for NS-3//Proceedings of the 4th International ICST Conference on Performance Evaluation Methodologies and Tools (ICST ValueTools 2009). Pisa, Italy, 2009

[18] Leland W E, Taqqu M S, Willinger W, Wilson D V. On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Transactions on Networking, 1994, 2(1): 1-15

XIA Yu, born in 1983, Ph. D. candidate. His research interests include computer network architectures, high-performance packet switches and switch scheduling algorithms.

ZENG Hua-Xin, born in 1945, Ph. D., professor, Ph. D. supervisor. His research interests include computer network architectures and communications technology, high-performance packet switch architectures and router test system.

GAO Zhi-Jiang, born in 1985, Ph. D. candidate. His main research interests include design and implementation of high-performance packet switches.

SHEN Zhi-Jun, born in 1976, Ph. D. candidate. His main research interests include design and implementation of high-performance packet switches.

Background

The main background of this research is the Single User-data transfer Platform Architecture Network (SUPANET) platform proposed by the Sichuan Network & Communications Technology Key Lab. The research of this platform is supported in part by National Natural Science Foundation of China project "Next Generation Internet Architecture and Its Key Technologies" under grant 60773102 and "Long-term Strategy of Engineering and Technology Development in China" united foundation project "Future Network Architecture and Network Technology Strategies" under grant U090122, and in part by Sichuan University Foundation (SCUF) project "Next Generation Internet Architecture".

As early as in 2003, the research group has deeply analyzed the problems in current Internet architecture, and found that the 30 years old Internet architecture, which provides the best effort service, has been the root of the inability of current Internet to provide QoS guarantee, thus an "clean-slate" architecture must be used in order to provide QoS guarantee in future Internet. This idea is the same as what GENI (Global Environment for Network Innovations) announced in 2006. SUPANET is one answer to this "clean-slate" architecture. By using physical frame, the authors put the QoS guarantee directly on physical layer, and this leads to the concept of Ethernet-oriented Physical Frame Timeslot Switching (EPFTS). Meanwhile, SUPANET is a connection-oriented network, i. e. a connection must be established by using QoSNP (QoS Negotiation Protocol) before the data can be transferred on the network. By doing so, the QoS of each connection can be easily reserved and then conserved. The authors have a lot research results on SUPANET platform, and their many papers have been indexed by EI and SCI.

The development of packet switching technologies has been lagging far behind the progress of transmission technologies. Current switching technology is unable to meet the requirement of high-speed data transmission in future networks. No matter what kind of network architecture is used in future network, packet switching technology must be a core problem to solve in the future computer network, which is the very reason why the authors do the project presented in this paper.