第33卷 第21期 计 算 机 工 程 2007年11月
Vol.33 No.21 Computer Engineering November 2007
·网络与通信· 文章编号：1000—3428(2007)21—0108—03 文献标识码：A 中图分类号：TP393

一种支持DiffServ模型的CICQ调度策略

李印海，扈红超，郭云飞
（国家数字交换系统工程技术研究中心，郑州 450002）

摘 要：结合大规模接入汇聚路由器需要对不同汇聚业务流进行不同的处理这一实际需求，基于CICQ交换结构，该文给出了一种支持DiffServ模型的调度策略(DS)，该算法以“节点行为”方式对业务流进行调度。和以往算法相比，DS采取了分布式的控制策略，并且具有较低的时间复杂度，工程上更易实现。仿真结果表明，DS不仅能够为EF和AF业务提供带宽保证，而且具有良好的时延性能。

关键词：交换结构；调度策略；CICQ；DiffServ

DiffServ Supporting Scheduling Algorithm for CICQ

LI Yin-hai, HU Hong-chao, GUO Yun-fei
(National Digital Switching System Engineering & Technological Research & Development Center, Zhengzhou 450002)

【Abstract】 Buffered crossbar switches are becoming more and more attractive to high performance routers/switches builders than other schemes, as it can achieve throughput, rate and delay guarantees, and distributing scheduling algorithm can be easily implemented. This paper proposes a distributed scheduling algorithm (shorted by DS) supporting differentiated service model (DiffServ) for CICQ switches, which has lower time complexities than previous algorithms. It evaluates the performances of DS through simulation under burst uniform and non-uniform traffic, and compares it with previous algorithms. Simulation results show that DS can provide minimum bandwidth guarantees for EF and AF traffic and fair bandwidth allocation for BE traffic.

【Key words】 switching fabric; scheduling policy; CICQ; DiffServ

为了满足下一代互联网对QoS的需求，IETF提出了两种QoS解决方案：IntServ[1]和DiffServ[2]。IntServ模型通过资源预留和接纳控制来为每条业务流提供QoS保证，试图将原本面向无连接的Internet改造为面向连接的网络。IntServ模型采用资源预留协议(RSVP)作为它的端到端的信令协议：在发送业务分组前，借助RSVP信令在途中每个节点为该条业务流进行资源预留。由于IntServ要维护每一条流的信息，因此不仅实现起来过于复杂，而且可扩展性差，不适合在骨干核心路由器中使用。与IntServ模型有所不同，在DiffServ模型中路由交换设备不再需要为每条流维护状态信息，网络中每个节点可以基于分组头部的DSCP(diffServ code point field)域以单跳行为(per hop behavior, PHB)独立进行转发。DiffServ定义了3种PHB标准EF(expedited forwarding), AF(assured forwarding)和BE(best effort)。EF业务要求低时延、抖动、端到端的服务和带宽保证；AF业务要求速率和时延的保证，AF又可以分为不同优先级的4类AF1~AF4；BE是尽力而为的业务，不存在质量保证问题。

在DiffServ域中，EF业务具有最高的优先级；为避免EF对其他业务的影响，DiffServ协议为EF业务规定了一个峰值服务速率(peak information rate, PIR)，超过该峰值服务速率的服务请求就会被拒绝；对于AF1~AF4业务而言，DiffServ为每类规定了一个最低服务速率(committed information rate, CIR)；为了避免BE业务产生“饥饿”现象，在满足EF和AF业务时，将网络剩余带宽分配给BE业务。这些节点转发行为PHB是通过网络节点的排队和调度机制来实现的。

1 相关研究与分析

OQ(output queuing)交换结构在QoS保障方面极具优势，而且其调度机制可独立工作于各个输出端口，复杂度较低，易于实现对DiffServ模型的支持。多数支持DiffServ模型的调度算法都是基于输出排队交换结构提出的，典型的如加权轮询服务(WRR)以及联合优先级排队加权轮询(PQWRR)[3]等。然而输出排队交换结构存在N倍加速问题，不具备良好的可扩展特性，高速环境下难以实现，这类支持DiffServ模型的调度算法下应用受限。输入排队交换结构虽然无须加速，并可以在高速环境下实现，但输入排队交换结构必须采用复杂的集中式调度机制才能获得良好的性能，可扩展性差。目前基于输入排队交换结构支持DiffServ模型的调度算法如动态DiffServ调度DDS[4]和分级DiffServ调度HDS[5]，虽然通过采用迭代方式逼近最大匹配可以在一定程度降低算法复杂度，但它们仅能在均匀业务条件下获得较好的性能。对于非均匀业务，业务负载较重时其性能急剧恶化。

近年来随着CMOS技术的进步，在交换单元内部集成一定容量的缓存成为了可能，并已成为该领域的研究热点，其中最具影响力的为联合输入交叉点排队交换结构(CICQ)，一个基本的CICQ交换结构如图1所示。

基金项目：国家“863”计划基金资助项目“大规模接入汇聚路由器(ACR)系统性能与关键技术研究”(2005AA121210)
作者简介：李印海(1964－)，男，副研究员，主研方向：计算机应用技术和交换技术；扈红超，硕士研究生；郭云飞，教授、博士生导师
收稿日期：2007-06-22 **E-mail：**huhongchao@gmail.com

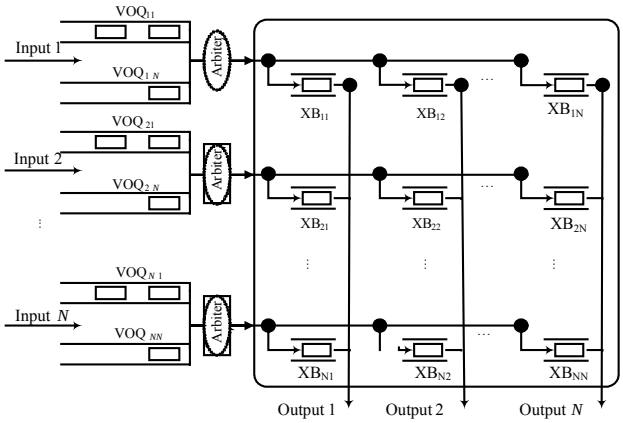

图 1　*N*×*N* CICQ 交换结构

为了避免队头(HOL)阻塞问题，采用了虚拟输出排队机制。可以看出，由于交叉点缓存的引入，将整个交换结构从逻辑上划分为 N 个 $N\times1$ 和 N 个 $1\times N$ 规模的子交换结构，使得分布式调度策略的实现成为了现实。在每个输入端与输出端分别设置一个调度单元，共有 $2N$ 个调度单元。这样每个调度器的设计就大大简化，可扩展性较好。从这一基本结构出发，对其进行了结构扩展以支持 DiffServ 模型。

2　DCICQ 结构及其调度策略 DS

2.1　支持 DiffServ 模型的 DCICQ 交换结构

根据前面的描述，DiffServ模型将互联网中的不同业务流分成了EF, $AF_i(1\le i\le4)$和EF 6 类，并且每种业务具有不同调度优先级。为了在调度时区分不同的业务分组，DCICQ在输入端口将每一VOQ扩展为$P(P=6)$个子虚拟输出排队队列(sub-VOQ)，每一sub-VOQ用来缓存不同类型业务的分组；同时每个交叉点队列相应的被分为P个子交叉点队列，记为sub-XB。拓展后的CICQ交换结构DCICQ如图 2 所示。

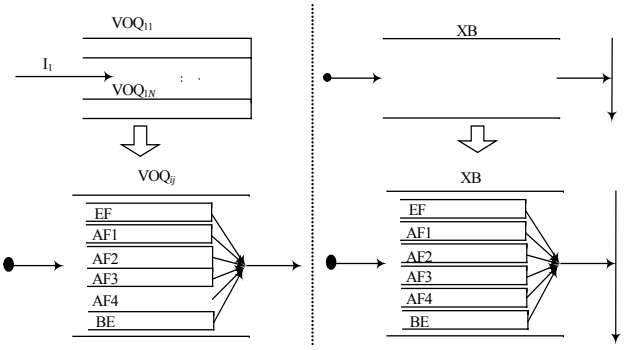

图 2　支持 DiffServ 模型的 DCICQ 交换结构

2.2　支持 DiffServ 模型的 DS 调度策略

在给出DS算法之前，首先给出DS算法中的带宽分配准则——DBA (dynamically bandwidth allocation)[4]。在DBA带宽分配准则中，每个输出端口的带宽B_w被划分为两部分：预留带宽与剩余带宽。其中预留带宽在EF与AF业务之间进行分配，EF和$AF_i(\le i\le4)$5 类业务都具有预定分配的带宽份额；剩余带宽在AF与BE业务之间进行分配。令B_p代表任一输出端口p类业务流的预留带宽份额，则B_p满足如下约束条件：

$$\sum_{p=1}^{P}B_p=1,\qquad \sum_{p=1}^{P-1}B_p<1 \tag{1}$$

令$B_{jp}(n)$表示在n时隙时，在输出端口j的p类业务流的获得的统计带宽。对于p=1 来说，$B_{j1}(n)=c_{j1}(n)/n$；对于 $1<p\le5$，$B_{jp}(n)=c_{jp}(n)/(n \bmod F)$。其中，$c_{j1}(n)$表示截止到$n$时隙时调度的EF业务分组数；$c_{jp}(n)$ $(1<p\le5)$表示从第k“帧”开始时刻调度的p类业务分组的数目，帧序数$k=\lceil n/F\rceil$，F为一帧所占用的时隙数。为了描述DS调度方案的方便性，给出描述DCICQ结构的若干符号定义：

定义 1 VOQ_{ijp}：缓存从i端口去往j端口的p类业务分组的虚拟输出队列；

定义 2 $v_{ijp}(n)$：标示在时隙n时VOQ_{ijp}队列是否为空的Boolean类形变量：“0”，空；“1”：非空；

定义 3 $g_{ijp}(n)$：标示在时隙n内VOQ_{ijp} 队列HOL信元的年龄，若$v_{ijp}(n)=0$，则$g_{ijp}(n)=-1$。i端口HOL信元的年龄向量$G_{ip}(n)=[g_{i1p}(n), g_{i2p}(n),\cdots,g_{iNp}(n)]_N$；

定义 4 XB_{ijp}：和VOQ_{ijp}一一对应，缓存从i端口去往j端口p类业务分组的交叉点队列；

定义 5 $f_{ijp}(n)$：在时隙n内XB_{ijp}中的信元数目，满足$0\le f_{ijp}(n)\le S$，S为交叉点队列容量。

定义 6 $w_{ijp}(n)$：在时隙n内XB_{ijp}队列HOL信元的年龄，若$f_{ijp}(n)=0$，则$w_{ijp}(n)=-1$。j端口HOL信元的年龄向量$W_{jp}(n)=[w_{1jp}(n),w_{2jp}(n),\cdots,w_{Njp}(n)]_N$。

在任意输入端口i'，若$VOQ_{i'jp}$满足：$v_{i'jp}(n)>0$，$0<f_{i'jp}(n)<S$，则$VOQ_{i'jp}$是“候选”的，输入端口i'的p类候选VOQs集合记为E_{ip}；对应输出端口j'，若交叉点队列$XB_{ij'p}$满足：$0<f_{ij'p}(n)$，$B_{j'p}(n)<B_{wp}$，输出端口j' 对应p类XBs的集合为$E_{j'p}$。

支持DiffServ的DS调度策略分为两个过程来实现：输入调度阶段(IS)和交叉点调度阶段(CS)。在输入调度阶段，每个输入调度单元独立的进行仲裁，选出当前要调度的VOQ_{ijp}，并将其队头分组送到相应的交叉点队列；在交叉点调度阶段，每个交叉点调度单元选出一交叉点队列，并将队头分组发送到外部链路。所有的IS和CS调度单元都可以分布式与并行工作。分别描述如下：

(1)IS：对于某一输入端口i，从p=1 类业务开始至p=6 直到找到某一虚拟输出队列VOQ_{ijp}满足条件：$VOQ_{ijp}\in E_{ip}$，且其对应的队头信元年龄$g_{ijp}(n)$在可选的虚拟输出队列集合中是最大的。

(2)CS：对于任一输出端口j，从p=1~p=5 类业务直到找到某一交叉点缓存队列CB_{ijp}满足条件：$XB_{ijp}\in E_{jp}$，且对应的队头信元的年龄$w_{ijp}(n)$在候选的交叉点队列集合中是最大的。若找到，则返回；否则，从p=2~p=6 类业务中，选择队头信元年龄最大的交叉点队列。

3　仿真性能评估

有效性通过平均时延来衡量；而公平性通过不同优先级的DiffServ业务的归一化带宽来验证。同时为了方便比较，给出了相同仿真条件下PQWRR和DDS算法的性能仿真结果。仿真参数的选取与文献[4]类似，即业务到达过程采用ON-OFF模型；突发长度为 32；每个输入端口EF业务与AF业务的比例依次为 18%, 24%, 20%, 16%, 12%。R_{j1}到R_{jp}依次设置为0.198 ,0.24 ,0.20, 0.16, 0.12。一帧设置为 1 000 个时隙。

带宽性能评估采用 4×4 交换结构，DDS算法的迭代次数为 4。为了制造过载环境，所有输入端口产生的业务分组具有相同的目的端口，每一输入端口业务流量强度以 0.1 为间隔从 0.1 增长到 1.0。图 3 给出了DDS和DS的实际获得带宽曲线，可以看出DS在复杂度较低的情况下获得了与DDS相似的带宽性能。当业务流强度达到 0.3 时，DDS和DS的EF与AF_i $(1\le i\le4)$业务获得带宽趋于稳定，而BE业务获得带宽比例有所下降。

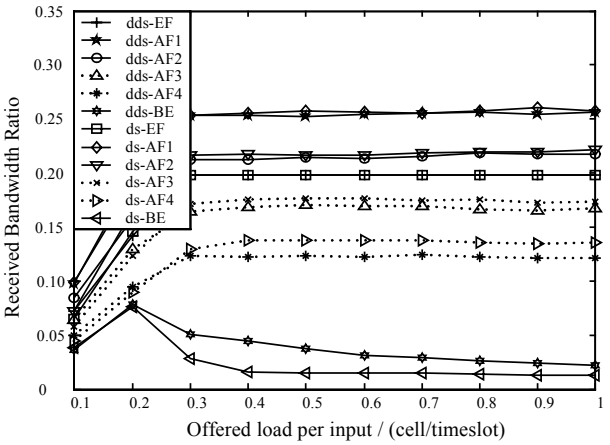

图 3　DDS 和 DS 的实际带宽分配

图 4 给出了PQWRR,DDS以及DS算法在均匀与非均匀条件下EF业务时延性能，交换规模均为 16×16。对于EF业务，3 种方案均能够提供具有低时延的服务，DDS和DS获得了相似的性能曲线，而PQWRR算法性能较差[4]。

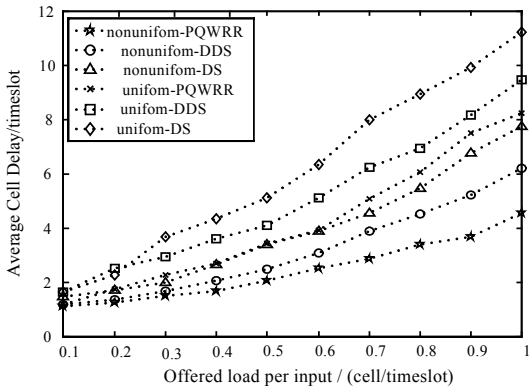

图 4　均匀与非均匀下的 EF 业务性能

图 5 给出了 AF1, AF3 业务在非均匀模型下时延性能，可以看出：在低负载情况下，DS 和 DDS 算法性能类似，当负载强度超过 0.7 后，DDS 算法性能急剧恶化；PQWRR 和 DDS 变化较为缓慢。

图 6 非均匀模型下 PQWRR、DDS 和 DS 时延抖动性能仿真曲线图。可以看出 DS 算法表现出最好的时延抖动性能。

综合以上公平性和有效性的仿真结果可以得出：只有 DS 方案在公平性和有效性两个方面均表现出良好的性能，并且比其它两种方案更易于在高速环境下通过硬件实现，能够更好地支持 DiffServ 模型。

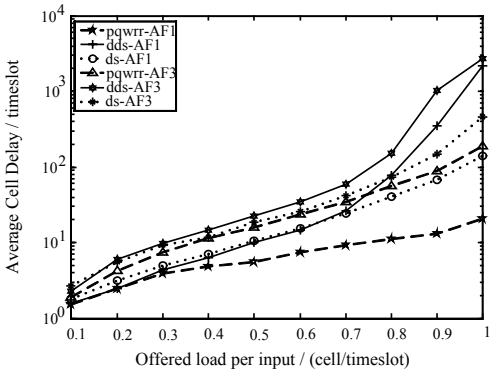

图 5　非均匀下 AF1 和 AF3 业务性能

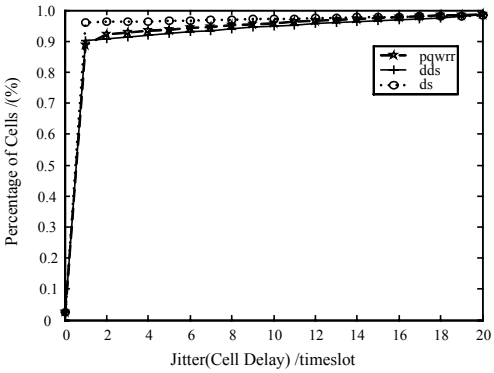

图 6　非均匀下 AF1 和 AF3 业务性能

4　结束语

本文对基本的 CICQ 结构进行了扩展，并提出了一种支持 DiffServ 模型的分布式调度策略 DS。仿真结果表明：与 PQWRR 和 DDS 相比，DS 在带宽保证、时延和时延抖动方面都表现出良好的性能。同时，DS 算法采用分布式与并行策略来实现的，在硬件上更易实现，具有较低的硬件复杂度。

参考文献

1 Braden R, Clark D, Shenker S. Integrated Services in the Internet Architecture: An Overview[S]. RFC 1633, 1994-06.

2 Carlon M, Wesis W, Blake S, et al. An Architecture for Differentiated Services[S]. RFC 2475, 1998-12.

3 Mao J, Moh W M, Wei B. PQWRR Scheduling Algorithm in Supporting of DiffServ[C]//Proc. of ICC'01. 2001: 679-684.

4 Yang M, Lu E, Zheng S Q. Scheduling with Dynamic Bandwidth Share for DiffServ Classes[C]//Proc. of ICCCN'03. 2003: 319-324.

5 Yang Mei, Wang J, Lu E. Hierarchical Scheduling for DiffServ Classes[C]//Proc. of IEEE Globecom. 2004: 707-712.

（上接第 107 页）

4　结束语

本文提出了对等实体间的网络用户漫游机制，构造了一个基于信任度的对等实体间的网络用户漫游模型。从模型的安全性分析和验证可以看出，所提出的漫游模型具有可行性。

参考文献

1 Chamberlin N. A Brief Overview of Single Sign-on Technology[EB/OL]. (2005-03-10). http://www.gitec.org/asserts/pdfs.

2 Damiani E, Vimercati D C, Paraboschi S, et al. A Reputation-based Approach for Choosing Reliable Resources in Peer-to-peer Networks[C]//Proc. of the 9th ACM Conference on Computer and Communications Security. 2002.

3 Prashant D, Partha D. PRIDE: Peer-to-Peer Reputation Infrastructure for Decentralized Environments[C]//Proc. of the 13th Wide Web Conference, New York. 2004.

4 Ratnasamy S. A Scalable Content-addressable Network[D]. USA: University of Berkeley, 2002.

5 Yu B, Singh M Y. Detecting Deception in Reputation Management[C]//Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multi-agent Systems. 2003: 73-80.

6 Zhong Y, Bhargava B. Authorization Based on Evidence and Trust[C]//Proc. of Data Warehouse and Knowledge Management Conference, France. 2002.