以太网基于优先级的流控仿真设计

龚夏青 窦 军
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 互联网在飞速发展的同时,也给网络系统的正常运行带来了一系列的问题,其中最突出的是由网络流量过大引起的网络拥塞。通过对多种流控方法的研究,最终从队列调度方面着手,利用OPNET网络仿真软件对所设计的三种流控机制进行仿真,初步解决了以太网较高优先级的数据流无法得到及时响应的问题以及公平性问题。

关键词 流量控制,队列调度,严格优先级,加权轮询优先级,OPNET

Ethernet Priority-based Flow Control Simulation Design

GONG Xia-qing DOU Jun
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract With the rapid development of the Internet, there are a series of problems in the normal operation of the network systems, and the most prominent problem is network congestion caused by the overload traffic. The author studied a variety of flow control methods, and by the way of priority-based queuing scheduling, the author used the OPNET network simulation software to simulate the three kinds of designed flow control mechanisms, finally solved the problems of the higher priority data flow not getting a timely response and the fairness issues in Ethernet.

Keywords Flow control, Queuing scheduling, Strict priority(SP), Weighted round robin(WRR), OPNET

1 引言

随着Internet的飞速发展,尤其是视频、语音等多媒体业务的迅猛发展,人们对于在网络上一些关键应用和多媒体应用的需求越来越大,计算机已经不是单纯的处理数据的工具,Internet也由以前的单一的数据网变成了多业务的综合数字网[1]。可见,在Internet的飞速发展的同时,也给网络系统的正常运行带来了一系列的问题,其中最突出的是由网络流量过大引起的网络拥塞。

然而即使有突出性能的网络设备也会受到所联接的网段上的拥塞带来的损害。传统上,通过一个端口的流量必须在只有一个输出队列的缓存中保存,不论它的优先级是多大,也必须按照先进先出的方式被处理。当队列满的时候,任何超出的部分都将被丢弃。此外,当队列变长时,时延也增加了。这个特点使得在传统的以太网上运行实时的事务处理及多媒体应用变得非常困难。由此可见,对网络流量进行智能化的控制显得日益重要。

笔者从队列调度方面着手,通过研究现有的以太网基于优先级的流控标准及建议,分析它们的优劣,最后利用OPNET网络仿真软件对所设计的三种流控机制进行仿真,初步解决了以太网较高优先级的数据流无法得到及时响应的问题以及公平性问题。

2 常用的队列调度

2.1 先进先出队列调度

先进先出(FIFO)调度是最简单的调度算法,就是根据分组到达的先后顺序来调度分组。FIFO调度提供了最基本的存储转发功能,也是目前网络设备中广泛使用的一种方式,它是默认的排队规则。它的优点是实现简单、成本低,缺点是不能区分时间敏感分组和一般分组,并且也不公平,因为这使得排在长分组后面的短分组要等待很长的时间。

2.2 严格优先级队列调度

严格优先级队列调度(SP)就是在先进先出(FIFO)的基础上增加了按优先级排队,这样就能使得优先级较高的分组优先得到服务。

图1是按优先级排队的示意图,图中假定有8个优先级队列。

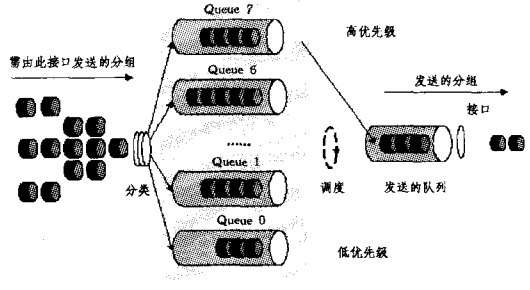


图1 SP示意图

SP队列调度算法,是针对关键业务型应用设计的。在队列调度时,SP严格按照优先级从高到低的顺序优先发送较高优先级队列中的分组,只有当较高优先级队列为空时,才发送较低优先级队列中的分组。

本文受国家自然科学基金(60773102)资助。

龚夏青(1989—),女,硕士生,主要研究方向为网络与通信技术;窦 军(1963—),男,副教授,主要研究方向为网络体系结构。

SP 的缺点是：如果较高优先级队列中总有分组，低优先级队列中的分组就会长期得不到服务，出现“饿死”现象。

2.3 公平队列调度

公平队列调度(FQ)可以解决 SP 里出现的“饿死”现象。FQ 就是对每种类别的分组流设置一个队列，然后轮流使每一个队列一次只能发送一个分组，对于空的队列就跳过去。但是公平队列也有不公平的地方，即长分组得到的服务时间长，这时短分组就比较吃亏，而且 FQ 并没有区分时间敏感分组和一般分组。

2.4 加权轮询队列调度

图 2 所示为加权轮询队列调度的示意图。

加权轮询队列调度(WRR)是 SP 和 FQ 的综合体。WRR 是在队列之间轮流调度，保证每个队列都能得到一定的服务时间。以端口有 8 个优先级队列为例，WRR 可为每个队列配置一个加权值(依次为 w7、w6、w5、w4、w3、w2、w1、w0)，加权值表示获取资源的比重。如一个 100M 的端口，配置它的 WRR 队列调度算法的加权值为 50、50、30、30、10、10、10、10(依次对应 w7、w6、w5、w4、w3、w2、w1、w0)，这样可以保证最低优先级队列至少获得 5Mbit/s 带宽，避免了采用 SP 调度时低优先级队列中的报文可能长时间得不到服务的缺点。WRR 队列还有一个优点是，虽然多个队列的调度是轮循进行的，但对每个队列不是固定地分配服务时间片——如果某个队列为空，那么马上换到下一个队列调度，这样带宽资源可以得到充分的利用。

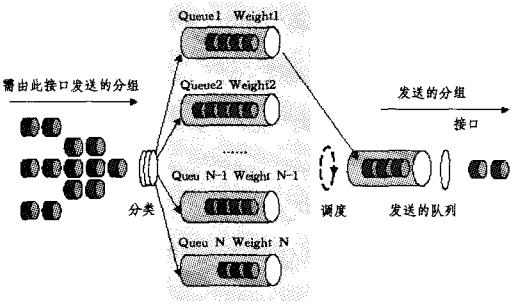


图 2 WRR 示意图

3 仿真

3.1 仿真模型设计

网络域拓扑结构如图 3 所示，这是为本次仿真设计的一个包交换网络，它由四个周边节点和一个中心节点组成。hub 为中心节点，用于处理包的存储转发。node0、node1、node2、node3 为周边节点，用于发送或接收包。各节点之间采用全双工链路进行连接。

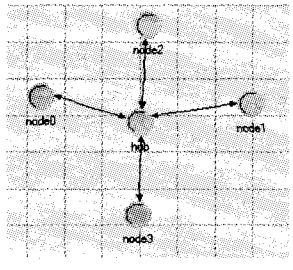


图 3 网络域拓扑结构

图 3 中，节点 node0、node1、node2、node3 即上述的四个周围节点，为了仿真方便，使 node0 只发送优先级为 0 的数据包，node1 只发送优先级为 1 的数据包，node2 只发送优先级为 2 的数据包，node3 只发送优先级为 3 的数据包。优先级 0、1、2、3 的优先级别依次降低。周围节点都是采用的 simple_source模型来产生数据包，simple_source 模型支持用户设置的发送包的间隔和发送包的大小符合的概率分布函数。

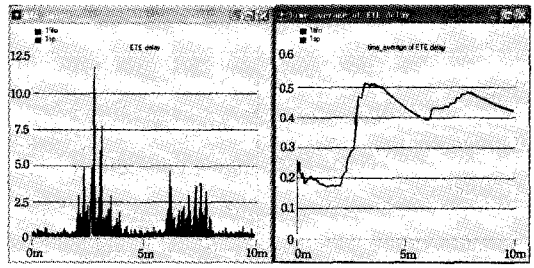
3.2 仿真结果分析

表 1 所列为产生包的相关设置，该场景的目的在于观测在极端情况下，所设计的三种流控方案的时延。以 node0 介绍表中各参数的意义，node0 节点，从第 120s 开始到 240s 结束，每 1 秒平均产生 100 个数据包，每个数据包的大小恒为 500 bits。这里 node0 产生的数据包为最大优先级的，其他节点的参数设置含义与此类似，其中 infinity 表示包产生结束时间为无穷大，当然仿真时，会随着仿真的结束而结束。这里需要说明一下，node1、node2 节点主要是为了便于观测在较为极端的情况下各种方案下的仿真结果的差异而设计的。

表 1 产生包的相关设置

节点名	产生包的间隔	包产生开始时间	包产生结束时间	产生包的大小
node0	possion(0.01)	120 s	240 s	constant(500)
node1	possion(0.1)	0.0 s	infinity	constant(1000)
node2	possion(0.1)	0.0 s	infinity	constant(1000)
node3	possion(0.01)	360 s	480 s	constant(500)

图 4(b)表明 FIFO 方案和 SP 方案下的平均 ETE delay 几乎完全一致，而观察图(a)中 2min～4min 区间可以看出，SP 的最大时延明显比 FIFO 的大许多，6min～8min 区间时延的差异不是很大。这是因为 2min～4min 间 node0 在以平均 100 Packet/s 的速度产生优先级为 0 即最高优先级的数据包，对于 SP 来说，其他优先级的数据包都不能被服务，必须等到最高优先级发完之后才可被服务，所以会出现 ETE delay 增大许多的现象。而 6min～8min 区间，node3 在以平均 100Packet/s 的速度产生优先级为 3 即最低优先级的数据包，SP 方案对它并没有特别地照顾，反而有抑制，但是总的来说不是很明显，这是因为此时较高优先级的数据包的产生速率要远远小于最低优先级的数据包的产生速率。

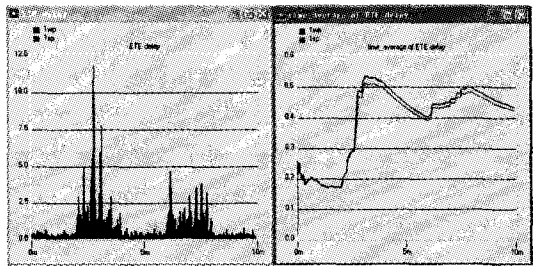


(a)ETE delay (b)平均 ETE delay

图 4 FIFO 和 SP 的 ETE delay

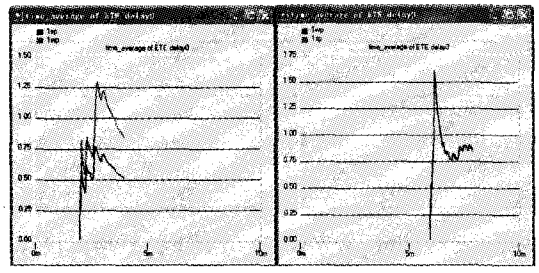
图 5 所示为 WRR 和 SP 方案下的端到端时延，从图 5(a)可以看出，WRR 方案要比 SP 方案的 ETE delay 的最大值小许多，虽然前者的平均 ETE delay 大于后者。这是因为 WRR 方案里有轮询机制，这样就可以有效地防止高优先级的数据包一直被服务，而低优先级的数据包只能等待的状况。

图 6 所示为 SP 和 WRR 方案中优先级为 0、3 的数据包的平均端到端时延情况，从图 6(a)可以看出对于最高优先级的数据包，WRR 方案下的平均时延明显大于 SP 方案的，这是因为 WRR 是基于公平性的，所以这里的时延增加了，一定会有其他优先级的数据包的时延减少。从图 6(b)可以看出对于最低优先级的数据包，WRR 方案和 SP 方案的统计结果很接近，这是由于特定的数据源所致。



(a)ETE delay　　(b)平均 ETE delay

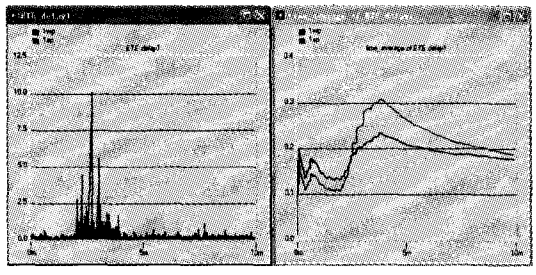
图 5　WRR 和 SP 的 ETE delay



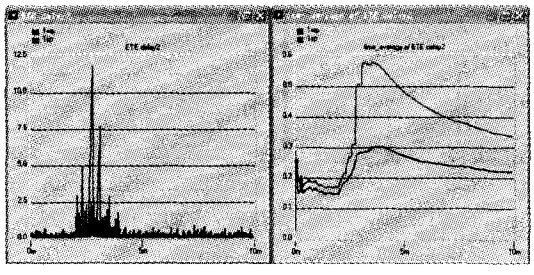
(a)平均 ETE delay0　　(b)平均 ETE delay3

图 6　优先级为 0、3 的包的平均 ETE delay

通过对下面一组图的进一步分析，可以更为清晰地看到 WRR 的公平性的体现。



(a)ETE delay1　　(b)平均 ETE delay1



(c)ETE delay2　　(d)平均 ETE delay2

图 7　优先级为 1、2 的包的 ETE delay

图 7 所示为 SP 和 WRR 方案中优先级为 1、2 的数据包的端到端时延情况，从图 7(a)、(c)可以看出 WRR 方案在处理较低优先级的数据包时明显优于 SP 方案，图 7(a)、(c)中的最大时延的差异是由优先级别的大小所致。现对图(b)作详细分析，从图中可以看出，图中的红色曲线和蓝色曲线的交点对应的时刻是 2min，最高点对应的时刻是 4min，这两个时刻恰好是最高优先级数据包产生时间区间，在 0min～2min 区间，优先级为 1 的数据包就是此时的最高优先级别的，SP 方案的平均时延小于 WRR 方案的平均时延，这是因为 SP 是绝对优先调度较高优先级的数据包的，而 WRR 是支持公平性的，这里公平性的实现需要较高优先级的队列付出一定的“代价”，即时延的增加。而在 2min～4min 区间，产生了更高优先级别的数据包 0，SP 方案的平均时延就明显大于 WRR 方案的，这也侧面反映出了 SP 方案的不稳定性，而 WRR 方案则有着较好的适应性。在 6min～8min 区间，有很多的最低优先级的数据包产生，从图中可以看到 WRR 方案的平均时延小于 SP 方案的平均时延，这也体现了 WRR 的公平性。

结束语　本次设计中，笔者设计了两种基于优先级的流控机制，即基于严格优先级的流控机制和基于加权循环的流控机制。通过设置合理的数据源，对所设计的流控机制的仿真结果进行分析，验证了所设计方案的可行性和适用性。

但是还存在以下不足，需要进一步研究：

1. 中心节点处理机的队列缓冲区的设置，需要更为可靠的理论支持。

2. 数据源的设置应更接近现实。

3. 算法有待于进一步优化，或是需要设计出适用性更强的算法。

4. 在 OPNET 中所进行的模拟都是通过简化的网络模型来实现的，这样只能定性的说明问题，模型有待于进一步的完善。

参 考 文 献

[1] 刘军，雷振明. 以太网流量工程研究[J]. 计算机工程与应用，2002,15(3):8-11

[2] 李彦东，马宏斌，许生旺. 基于 OPNET 的试验通信网 Qos 性能分析与仿真[J]. 工程实践及应用技术，2008,34(6):53-55

[3] 王沛. 基于 IP 网络队列调度算法的研究[J]. 大众科技，2009,10(114):66-68

[4] 孟坤，朱翼隽. 动态优先级队列的离散时间排队分析[J]. 科学技术与工程，2008,8(24):6491-6495

[5] Lestas M, Plisillides A, Icannou P, et al. Adaptive congestion protocol: a congestion control protocol with learning capablity [J]. Computer Networks: The International Journal of Computer and Telecommunications Networking, 2007, 51(13): 3773-3798