文章编号:1000 - 5080(2003)04 - 0073 - 04

Qos 调度算法综述

许宪成
(广东外语外贸大学 计算机系,广东 广州 510420)

摘要:对常见的主要 Qos 调度算法进行了评述,着重介绍了 WFQ 系列、DRR 和 CBQ 的工作机理,并对 WFQ 系列进行了理论分析。最后给出比较结果和适用范围,展望了发展方向。

关键词:服务质量;调度算法;服务程序
中图分类号:TP393　　文献标识码:A

0 前言

IP Qos[1,2] 目前已成为业界关注和研究的热点,不论在 IntServ 还是在 DiffServ 里,都涉及到队列调度机制。队列调度机制对 Qos 保证至为重要,其主要目的是为每个业务流提供一定的服务质量保证(如带宽、时延及时延抖动、丢包率等),调度器的基本功能就是网络中继节点或路由器如何从一个或多个队列中选择下一个待转发的分组。

一个有效的队列调度机制取决于它采用的调度算法,衡量调度算法的性能指标主要有:公平度、时延特性、对恶意流的隔离能力、链路带宽利用率、复杂度等,前 4 个指标与 Qos 密切相关[1]。目前常见的队列调度算法主要有基于轮转(Round Robin)和基于广义处理机共享 GPS(Generalize Processor Sharing)的算法两大类[2]。前者是轮流对每个队列进行服务,其实现简单,但不能对业务提供时延保证,目前主要有 Weighed RR、Deficit RR 等;后者目前主要有:Packetized GPS(即 WFQ)、自时钟公平排队(SCFQ)、VC(Virtual Clock)、Leap Forward VC 等,尤其是 WFQ 及其派生算法能提供较好的公平度、时延特性以及对恶意流的隔离能力,但当队列数较多时,其实现复杂度较高。

本文重点讨论已被广泛接受并实现的关键调度算法:WFQ(Weighted Fair Queueing)及其派生算法、DRR(Deficit Round Robin)算法和 CBQ(Class Based Queueing)算法。为讨论方便,假定有 N 个流被基于 GPS 的算法调度,$W_i(t_1,t_2)$ 是时间(t_1,t_2)内流 i 所接受的服务,调度器服务速率为 r,公平度(Fairness)为$|W_i(t_1,t_2)/r_i - W_j(t_1,t_2)/r_i|$,等待调度的流 i 的队列首分组 p_i^k 为该数据流的第 k 个分组,其长度为 L_i^k(所有流中最长分组长度为 $L_{\max}$),到达时刻为 a_i^k,离开时刻为 d_i^k。

1 WFQ 及其派生算法

1.1 WFQ 调度算法

WFQ 的基础是 GPS 算法[3]。GPS 流体模型隐含分组无限可分及所有的流可以同时接受服务,还假定只要其中任一队列有分组进入,调度器就立即处于工作状态,并保证链路容量不会因调度算法本身而受到任何额外损失。GPS 公平度最高(0),按流的权重提供带宽,时延 $D_{i,GPS} \leq \sigma_i / r_i$,所以 GPS 具有良好的隔离性和扩展性,是一个理想的调度算法。

在实际系统中,一个调度器一次只能为一个流服务,而且服务的最小单元为 1 个分组。GPS 算法的理想化流体模型无法实现,但通过模拟 GPS 参考模型,就可以定义一些能够近似实现 GPS 的调度算法,如 PGPS(即 WFQ)算法,Parekh 和 Gallager 对此进行了深入的分析。另一相似的调度算法是基于 BR(Bit by Bit Round Robin)的公平排队 FQ(Fair Queueing),由 Damers 等提出[4]。

WFQ 将一个分组模型化为相应的 GPS 参考系统的一定流体量,当 GPS 服务了与分组相对应的流体量时就视同该分组被发送。为此 WFQ 引入虚拟服务时间函数 $V(\tau)$ 表示调度器在时刻 τ 已经提供的服务,

作者简介:许宪成(1965 -),男,河南开封人,讲师.
收稿日期:2003 - 04 - 15

并选择参考系统中具有最小结束服务时间 F_i^k 的分组进行发送。与相应的 GPS 参考系统相比,在 WFQ 情况下,对各业务流 i 有:

$$d_{i,WFQ}^k - d_{i,GPS}^k \leq L_{\max}/r$$

$$W_{i,GPS}(0,\tau) - W_{i,WFQ}(0,\tau) \leq L_{\max}$$

若业务流由令牌桶(σ_i, r_i)规约定义和限定,则时延 D_i 保证上限为:

$$D_{i,WFQ} \leq \sigma_i/r_i + L_{i,\max}/r$$

可见与相应的 GPS 参考系统相比,WFQ 的分组结束服务的时间最多滞后发送该业务流的一个最长分组的时间,$D_{i,WFQ}$最后一项反映了实际系统逐分组传输的特点。所以在一个实际的系统中,仅从结束服务时间而言,WFQ 已经具备了最佳的性能参数。另外,由于 WFQ 总是选择最早离开的分组进行发送,因而可以采用一个优先级队列予以实现。

1.2 WF^2Q 调度算法

WFQ 仅以分组在 GPS 参考系统中的 F_i^k 为依据选择分组进行发送,这就会造成一些尚未在 GPS 参考系统中得到服务的分组仅因其 F_i^k 较小而被选中并发送,导致超前 GPS 参考系统很多。所以 WFQ 的分组离开次序与 GPS 参考系统的离开次序间可能会存在明显差异,导致时延抖动累计效应增大和算法效率的下降。WFI(Worst - case Fair Index)[5] 公平因子 C_i 被引入用来衡量这个差异,显然 $C_{i,GPS}=0$。

为了降低 WFQ 算法的 WFI 因子 C_i,WF^2Q(Worst - case Fair Weighted Fair Queueing)[5] 采用 SEFF (Smallest Eligible virtual Finish time First) 分级选择策略来更准确地模拟 GPS:首先筛选在 GPS 参考系统中已经得到服务的分组,然后再选择其中具有最小结束服务时间的分组进行发送。以此避免分组离开 WF^2Q 的时间过于偏离参考系统的离开时间。在从 WFQ 派生的各种算法中,WF^2Q 算法具有最小的 WFI 因子:

$$C_{i,WF^2Q} = L_{i,\max}/r_i + (L_{\max} - L_{i,\max})/r$$

而 WF^2Q 的时延上限依然与 WFQ 相同,所以,WF^2Q 被认为是模拟 GPS 的优化算法。

1.3 WF^2Q+ 算法

WF^2Q+ 算法[6] 是 WF^2Q 的改进版,它在保持与 WF^2Q 相同时延上限的同时,降低了算法复杂度,用以支持 H-GPS(Hierarchical-GPS)。为此,WF^2Q+ 算法仍采用 WF^2Q 的 SEFF 分组选择策略,通过重新定义 $V_{WF^2Q+}(t+\tau)$ 消除了模拟相应 GPS 参考系统的耗费:

$$V_{WF^2Q+}(t+\tau) = \max[V_{WF^2Q+}(t) + W(t,t+\tau)], \qquad \min_{i\in B(t+\tau)}(S_i^{h_i(t+\tau)})$$

式中 $W(t,t+\tau)$为$(t,t+\tau)$期间的服务总量,$h_i(t+\tau)$为业务流 i 的队列首分组序列号,而 $S_i^{h_i(t+\tau)}$ 为该分组的虚拟开始服务时间。另一方面,WF^2Q+ 又通过重新定义 S_i 简化了计算量[6]:只有当分组到达队首时才需计算其开始和结束服务时间,而不必象原方案那样,只要有分组到达或发送系统都必须重新计算相应参数。

通过上述改进,WF^2Q+ 将复杂度由 WFQ 和 WF^2Q 的 $O(n)$,降至 $O(\log n)$,同时仍保持 WF^2Q 算法具有的最小 WFI 因子。

2 DRR 算法

DRR[7] 最初是作为 STFQ(Stochastic Fair queueing)的功能扩展提出的,用以弥补 WRR 没有考虑分组长度变化的缺憾,它也可看作是另一类为克服 WFQ 复杂性、可扩展性问题的算法。DRR 采用 Hash 法将业务流的分组映射到某个队列中,通过限定队列数来克服可扩展性问题;为实现与 WFQ 相近的公平性,DRR 首先为每个队列 i 分配一个服务额度(Quantum) Q_i,同时,用一个状态变量 Deficit Counter(亦称赤字计数器) D_i 与每个队列相联系,表示上一轮调度时未用完的服务额度。当某个队列的一个分组因长度 L_i 过大而未获得服务时,作为补偿,服务额度 Q_i 结转供下一轮使用,这就是 DRR 的基本思想。

具体实现时,系统还使用一个 ActiveList 记录所有的非空 Active 队列,用一个指针标记当前的发送队列。如果一个队列由非空变成为空,则将其从 Active 队列中移出。

显然,DRR 算法支持长度变化的分组且实现简单,所以很多路由器实现了该调度算法。CISCO 的

MDRR(Modified DRR)实际上就是 DRR 和基于 IP TOS 字段的优先级相结合而成。

DRR 的缺点是对时延的保证不够有力。为此 Shreedhar 和 Varghese 提出了它的一种变体 DRR +[7], DRR + 算法把业务流分成尽力服务(best-effort)时延关键(latency critical)两类。若后者在接入时(admission time)没有超额发送分组,则相对每一个其它同类流而言,系统将提供至多不超过一个分组的时延保证。

3　CBQ 算法

上述讨论中 WFQ 或 DRR 调度器都需要为每个单流维持一个队列,在实际网络中,业务流数量往往极其巨大,实施每流排队的代价很高。如果把若干业务流聚集(aggregated)组成"类"(Class),再用这些类构成树形的层次链路共享结构,则可基于类提供带宽,从而在一定时间内,保证属于该类的业务流获得相应的链路容量,在此前提下遵循一定的原则,将剩余带宽动态调整分配给其它类,达到合理有效地使用带宽的目的,保证公平:这正是层次链路共享模型 HLS(Hierarchical Link Sharing)的基本思想。Floyd 和 Jackbson 在文献[8]阐述了如何用 CBQ(Class Based Queueing)构建层次链路共享结构以支持实时服务,实现上述目的。

CBQ 就是基于"类"的 HLS 算法,每个类均与一定的带宽相联系。这里的"类"是一个广义的概念,它可以表示一个组织,也可以代表一类业务,还可以是一个具体的端到端业务。由这些各种各样的"类"组成了一个 CBQ 树。可以通过图 1 来说明其原理,这是一个实例的层次化链路共享结构。图中的每个节点,都称为"类"(Class)。其中根节点代表物理资源(总带宽)。内部节点代表业务流聚集,而叶节点则代表具体业务流。

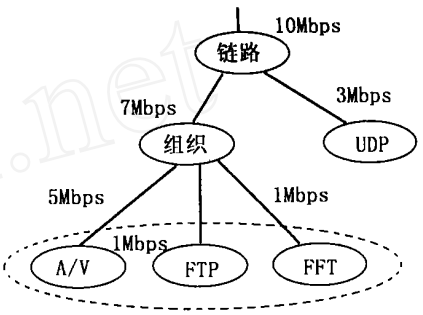

图 1　层次链路共享结构

这种结构能够实现不同类间隔离,而且并不要求在所有的链路层次采用相同的调度算法。文献[8]中 CBQ 方案就先将业务划分为尽力服务和实时服务,加以隔离后再对后者实施基于优先权的调度算法:根据需要,实时服务内的不同类再赋予相应的优先权值,以减少时延。值得一提的是,如果 CBQ 树中每个节点均采用 DRR 或 W^2FQ+ 调度器,则可构成两个著名的层次链路共享模型 Hierarchical-DRR 和 Hierarchical-W^2FQ+[6],其中 H-DRR 提供带宽保证,比较容易实现;而 H-W^2FQ+ 则模拟 H-GPS,既支持链路共享、又支持实时服务和尽力服务,在 H-WFQ 系列中具有上佳的性能参数。

CBQ 一般只提供对带宽的保证,相对非层次化结构而言,在 CBQ 树中同时实施带宽、时延、时延抖动等 Qos 策略比较困难,对节点的调度算法有更严格的性能要求。事实上,W^2FQ+ 算法的提出正是基于这种背景。

4　结束语

本文对 IP Qos 机制中常见关键调度算法进行了分析讨论,表 1 概括给出它们之间的性能比较。

表 1　常见关键调度算法比较

调度算法	虚拟服务时间	带宽时延结合	公平度	实现难度	复杂度	附加说明
GPS	- -	是	0	∞	∞	理想算法
DRR	- -	或然性	中	中	低	考虑分组长度
WFQ	GPS ref、选最小 F_i^k	是	∞	高	高	考虑逐分组传输
WF^2Q	GPS ref、采用 SEFF	是	中	高	高	考虑筛选合格分组
WF^2Q+	GPS ref、、简化 S_i 计算	是	中	高	中	理想模拟 GPS 算法

由此可见,WFQ 系列算法提供带宽时延保证,调度较公平,但实现复杂目前多用于较低速链路,DRR 容易实现,可用于高速链路但一般仅提供带宽保证,CBQ 性能取决于树中节点采用的具体算法。总之,较新的算法都是在先前的基础上发展而来,具体实现往往要结合使用并进行功能扩展。如何对算法进行优化,降低复杂度以便实际使用则始终是努力方向。当然,采用何种调度算法还与 Qos 的体系结构 IntServ 和

DiffServ有关。调度算法也需要结合接入控制、整形(令牌桶)与冲突检测(如RED)等机制,才能提供完整的端到端Qos保证[1,2]。以上讨论仅局限于有缆环境下带宽固定的情况,至于如何在突发网络流量情况下仍保持调度公平,如何在无线环境下支持移动IP、如何使调度算法能同时支持组播通信等领域[9],还有待进一步的研究。

参考文献:

[1] Zheng Wang. Internet Qos: Architectures and Mechanisms for Quality of Service[M]. Morgan Kaufmann Publishiers, 2002.

[2] William Stallings. High-Speed Networks And Internets: Performance and Quality of Service[M], Second Edition, Prentice Hall Press, 2002.

[3] Abhay K Parekh, Robert Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single Node Case[J]. IEEE/ACM Trans on Networking, 1993, 1(3): 344 - 357.

[4] Alan Demers, Srinivasan Keshav, Scott Shenker. Analysis and Simulation of a Fair Queueing Algorithm[A]. Proceedings of SIGCOMM '89[C], 1989.

[5] Jon C R Bennett, H Zhang. WF^{2}Q: Worst-Case Fair Weighted Fair Queueing[A]. IEEE INFOCOM '96[C], 1996.

[6] J C R Bennett, H Zhang. Hierarchical Packet Fair Queueing Algorithms[A]. IEEE/ACM Transactions on Networking[C], 1996.

[7] M Shreedhar, G Varghese. Efficient Fair Queueing Using Deficit Round robin[A]. Proceedings for SIGCOMM '95[C], 1995.

[8] Floyd S, Jacobson V. Link-Sharing and Resource Management Models for Packet Networks[J]. IEEE/ACM Transactions on Networking, 1995, 3(4): 365 - 386.

[9] Lars Wischhof, John W Lockwood. Packet Scheduling for Link-Sharing and Quality of Service Support in Wireless Local Area Networks[R]. WUCS-01-35, 2001.

Survey of Qos Scheduling Algorithms

XU Xian-Cheng

(Dep. of Comput. Sci. & Tech., Guangdong Univ. of Foreign Stud., Guangzhou 510420, China)

Abstract: Scheduling algorithms is very important for IP Qos mechanism. Most of researches on scheduling algorithms are round robin or GPS based. The recent developments of scheduling algorithms are introduced. The emphasis of the paper is to introduce the classification of scheduling algorithm and analyse the operating mechanisms of three typical implements of scheduling algorifhm: WFQ series, DRR and CBQ.

Key words: Qos; Scheduling algorithms; Service routines