

文章编号: 2095-4980(2014)06-0907-05

一种区分服务模型及其在三层交换机的实现

夏 俊

(深圳市电子学会, 广东 深圳 518057)

摘 要: 服务质量(QoS)是用来定义网络给不同形式的流量提供不同级别的服务保障的能力。当前服务质量的体系结构和模型有多种, 本文参照 RFC2474、RFC2475 标准, 提出了一种新的服务质量——区分服务(Diffserv)应用模型, 该模型包括: 流量的分类器、计量器、标记(或重标记)器、丢弃器和调度器等模块, 其中对于核心模块——调度器, 提出基于最大延迟的加权轮转调度(WRR-MBD)算法, 同时给出区分服务模型在 3 层以太网交换机的实现方法。通过网络测试分析仪对 3 层交换机 QoS 性能进行测试, 得出在 WRR 调度策略下, 不同优先级数据包的丢失率基本按所设置的权重比例进行分配, 从而避免了严格优先级(SP)调度存在的低优先级队列“饥饿”的现象。

关键词: 服务质量; 区分服务; 区分服务编码点; 3 层交换

中图分类号: TN915; TP393

文献标识码: A

doi: 10.11805/TKYDA201406.0907

Implementation of a new Diffserv model based on 3-layer switch

XIA Jun

(Shenzhen Institute of Electronics, Guangdong Shenzhen 518057, China)

Abstract: Quality of Service(QoS) is used to define different forms of network traffic to provide different levels of service assurance capabilities. The models and architectures of QoS fall into several types. A new type of application model of QoS—Differentiated service(Diffserv) application model is proposed according to the standard of RFC2474 and RFC2475. This model is composed of sub-modules such as classifier, meter, marker(remarker), dropper and scheduler etc. Meanwhile, the algorithm of Weighted Round Robin with Max Bounded Delay(WRR-MBD) of scheduler is proposed and analyzed in detail. Finally, the implementation of Diffserv based on 3-layer switch is described. Through a network analyzer to test QoS of 3-layer switch, it is concluded that in WRR scheduling policy, the loss rate of packets with different priorities is basically allocated according to the ratio of weight, which avoids low priority queues “hunger” existing in strict priority scheduling policy.

Key words: Quality of Service; differentiated service; Differentiated Services Code Point; 3-layer switch

随着 IP(Internet Protocol)网络上新应用的不断出现, IP 业务由传统的单一数据业务向同时包含语音、视频等多媒体的综合业务方向发展, 但传统的以最大努力(Best effort)为传输方式的 IP 网络无法满足多媒体应用和各种用户对网络传输质量的不同要求, 因为它将所有的报文都不加区别地等同对待。服务质量(QoS)技术的提出便是致力于解决这个问题, 旨在针对各种应用的不同需求为其提供不同的服务质量, 使传输服务的可靠性、延时、延时抖动、吞吐量、丢包率得到相应的保证^[1-6]。随着语音、视频等实时性的业务的广泛应用, 对此类业务的服务质量的需求变得越来越重要和迫切。本文提出一种新的区分服务模型, 并在基于 TMS(Tornado for Managed Switches)软件体系下的 3 层交换机上得到了整合和实现。

1 Diffserv 模型及其实现

本文用区分服务(Differentiated Services, DS)字段的区分服务编码点(Differentiated Services Code Point, DSCP)

值来标识相似 QoS 要求的流, 将其合并为一类, 以便提供相同的服务, 网络内部节点将每个分组的 DSCP 映射到一组每跳行为(Per-Hop Behavior, PHB), 以便采用合适的转发机制转发, 它不是基于流而是基于类(汇聚流)进行处理的。区分服务按功能可划分为: 分类器、计量器、标记(或重标记)器、丢弃器和调度器等主要模块, 其中分类、计量、标记(重标记)、丢弃是在交换机的入口执行的, 而数据包的调度是在出口执行的, 本区分服务的流程图如图 1 所示。主要模块的功能如下:

1) 分类器(classifier)

将分组报头某些字段的内容, 与系统中制定的规则进行比较, 从而将数据包分到指定的类别中。本文将分类器分为 2 类: BA(Behavior Aggregate)和 MF(Multi Field)分类器, BA 分类器仅根据 DS 字段进行分组分类, MF 分类器可根据分组报头多个字段或其组合选择分组, 如: 源 IP 地址、端口号、协议类型、服务类型(Type Of Service, TOS)字段等来区分不同分组所属的类型并提供相应的 QoS 保证。

2) 计量器(meters)

计量器是用来检测由分类器选出的业务流的临时特性(如速率), 并依此作为是否对包丢弃或重标记的尺度^[7]。为了监控复杂的流量, 实施更灵活的调控规则, 设计了三色标记桶, 定义 2 个突发尺寸: Large Burst Size(LBS)和 Middle Burst Size(MBS), 从而将包流分为 3 个符合等级, 当进入交换机的包流大于设定的 LBS 时被认为不符合包流, 介于 LBS 和 MBS 的包流被认为部分符合, 小于 MBS 的包流为符合流量要求, 即将 3 种包流用 3 种颜色来划分: 绿色代表符合, 黄色代表部分符合, 红色代表不符合。

3) 标记器(marker)

将分组的 DS 字段设为特定的 DSCP, 从而将这个分组加到与编码点对应的 DS 行为聚合体中, 它反映用户对一定业务性能等级的要求。标记值直接映射为网络节点将要执行的每一跳行为(PHB), PHB 只对聚合流操作^[8]。

4) 丢弃器(dropper)

由于交换机工作于存储转发方式, 则可能会出现交换机接收的分组要比它能发送出去的分组多的现象, 这导致交换机内部的缓冲区产生溢出, 因此交换机必然要丢弃分组。本丢弃器采用了随机早期检测(Random Early Detection, RED)分组丢弃算法, 它是在链路出现拥塞的早期征兆时, 在拥塞发生前就限制数据流。RED 算法的机理是: 网络转发节点通过一个低通滤波函数计算缓冲区的平均占用量, 并依此作为网络拥塞程度的度量, 缓冲区的占用量是通过队列的平均长度来反映的, 队列长度一般以分组的个数为单位。当缓冲的平均占用量达到一定的程度时, 就对到达的分组按照一定的概率进行丢弃, 当占用量达到门限值时, 丢弃所有到达的分组, 从而避免网络拥塞。

5) 调度器(scheduler)

业务流被分类并按着给定的优先级输入到相应的输出队列中, 并送往输出接口。3 层交换机通过调度器从多个输出队列缓冲区中选择下一个待转发的分组, 实现了对实时业务的转发。调度器通过采用不同的分组调度算法来实现多种调度策略的实施。

本交换机的调度器支持的调度算法有: 先进先出(First In First Out, FIFO)队列调度、严格优先级(Strict Priority, SP)调度、加权轮转(WRR)调度、基于最大延迟的加权轮转调度(WRR-MBD)。

本文提出的 WRR-MBD 调度算法是为了支持实时数据流转发应用而设计的, 它在 WRR 算法的基础上, 增加了对某一个(或几个)有实时性要求的队列的最大转发时间的限制, 当服务于其他队列的时间超过了设定的阈值, 则调度器必须终止对其他队列的服务, 而转到对此有实时性要求的队列中进行数据包的转发。此调度算法特别对延迟敏感、实时性要求较高的数据流如语音、视频业务优势明显。当交换机产生拥塞时, 交换机将不同的业务加以区别对待, 确保语音视频等数据流在最大延迟范围之内。

2 WRR-MBD 调度算法机理

本文提出的 WRR-MBD 调度算法的原理如下:

假设某一出口的队列数量为 $N=\{N_0, N_1, \dots, N_{n-1}\}$, $W(N_i)$ 表示队列 N_i 的权值, 变量 i 表示所服务的队列的序号,

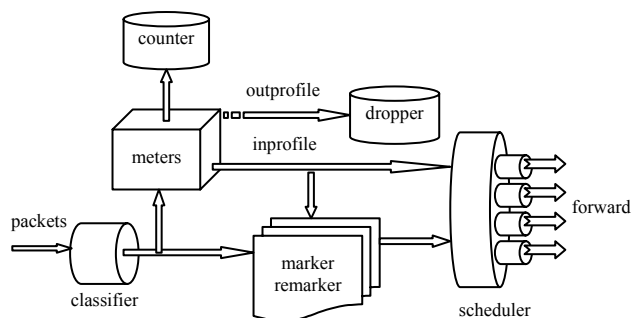


Fig.1 Flow chart of Diffserv
图 1 区分服务流程框图

$T(N_i)$ 表示队列 N_i 所处理的分组数量。则 $\sum T(N_i)$ 表示一个轮转周期所需要处理的分组总量。 $\sum W(N_i)$ 表示各队列权值总和。则： $T(N_i)/\sum T(N_i)=W(N_i)/\sum W(N_i)$ ，表明对各队列的分组转发处理数量是按照各权值占权值总数的比例进行分配。则每个队列分配的带宽为：(端口的总带宽) $\times W(N_i)/\sum W(N_i)$ 。

现在增加另一个约束条件，即对某一个或几个队列设置最大转发延迟时间(MBD)的限制。在多个约束条件下，给定系统执行操作的优先级顺序如下：

1) MBD 的优先级高于 WRR 调度策略；

2) 若对多个队列分别设置了 MBD，则对应于高优先级队列的 MBD 优先级更高。

假设对 x 队列和 y 队列所设置的最大允许延迟时间分别为 D_x 和 D_y ， x 队列的优先级高于 y 队列，则 D_x 优先级高于 D_y 。这说明如果对 x 队列的计时时间达到所设定的 D_x 值，并且当前正服务于 y 队列时，调度器将暂停对 y 队列的服务，而转向对 x 队列的处理。

假设系统从一个队列中每执行一次分组调度所需时间为 T_s ，调度单位长度的分组所需时间为 T_p ，则调度器每执行一个轮回所需时间为： $N \times T_s + \sum W(N_i) \times T_p$ 。则从第 x 队列执行完调度之后到下一轮回再对此 x 队列调度之前的间隔时间为 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_x) \times T_p$ 。同理，对 y 队列进行调度的间隔时间为 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_y) \times T_p$ 。

根据网络的拥塞情况和 WRR 权重和 MBD 参数的设置情况，可以将调度顺序分为以下几种：

1) 若 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_x) \times T_p \leq D_x$ 且 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_y) \times T_p \leq D_y$ ，说明为 x 队列和 y 队列分配的计时器的计时时间还未达到 D_x 和 D_y 时，调度器已经又回到对此队列的服务，这表明调度器采用的是正常的 WRR 调度顺序，所设置的 MBD 不起作用。

2) 若 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_x) \times T_p \geq D_x$ 且 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_y) \times T_p \leq D_y$ ，说明为 x 队列分配的计时器的计时时间已达到 D_x ，此时调度器正在服务于其他队列，还未回到对 x 队列处理，这时调度器终止对其他队列的服务而回到 x 队列中。

3) 若 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_x) \times T_p \leq D_x$ 且 $(N-1) \times T_s + \sum W(N_i) \times T_p - W(N_y) \times T_p \geq D_y$ ，说明为 x 队列分配的计时器的计时时间已达到 D_y ，此时调度器正在服务于其他队列，同样调度器终止对其他队列的服务回到对 y 队列的操作。

WRR-MBD 调度算法对延迟敏感的业务做到了实时转发，同时也满足了对其他不同优先级业务的带宽的合理分配，非常适合于语音、视频等实时性要求较高的业务与其他业务同时争用同一端口带宽资源的情况。

3 3 层交换机软件框架结构下的区分服务模型

本 3 层交换机软件系统采用 Wind River 公司的 VxWorks 6.8 操作系统，本文在分析和论证了 TMS 体系结构的基础上，实现了 Diffserv 模块的添加和功能实现。结构框图如图 2 所示，其中红色方框为增加的 Diffserv 部分，红色椭圆圈定的为在原 TMS 基础上修改和添加的模块。

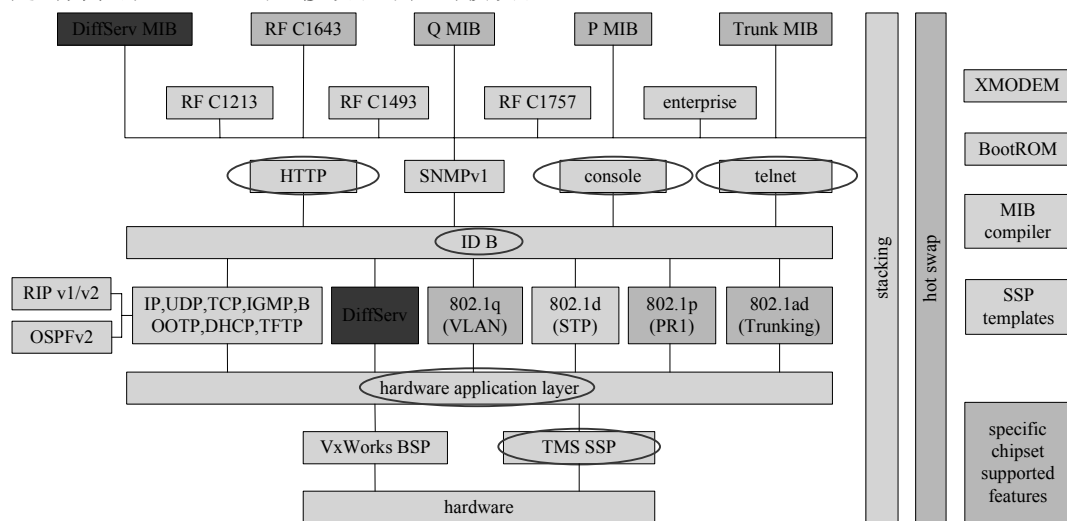


Fig.2 Diffserv module of TMS architecture
图 2 TMS 体系结构下的 Diffserv 模块

从图 2 可以看出, TMS 主要由以下几部分组成:

1) 管理应用层: 包括超文本传输协议(Hyper Text Transport Protocol, HTTP)网络服务器, 控制台(通过串口的超级终端方式访问), 远程登录访问(telnet), 使用 Envoy9.1/EnvoyGlue(TMS 支持 SNMP v1, v2 和 v3 协议加 SNMPv2/v3 MIBs), 标准和企业的管理信息库(Management Information Database, MIB)等。

2) 协议层: 包括 IP/TCP/UDP 协议栈, 路由信息协议(Routing Information Protocol, RIP)、开放式最短路径优先(Open Shortest Path First, OSPF)路由协议, 组播管理协议(Internet Group Management Protocol, IGMP), 引导程序协议(BOOTstrap Protocol, BOOTP)和动态主机配置协议(Dynamic Host Configuration Protocol, DHCP), 简单文件传送协议(Trivial File Transfer Protocol, TFTP)等。

3) 链路层: 如 STP(Spanning Tree Protocol, IEEE 802.1d 生成树协议), GARP(Generic Attribute Registration Protocol, 通用属性注册协议, 支持 GVRP 和 GMRP), IGMP 监听, LACP(Link Aggregation Control Protocol, 支持 IEEE 802.3ad 的链路聚合协议, 即 Trunking)等。

4) 硬件适配层: 主要包含了系统抽象层(System Abstraction Layer, SAL), BCM(Broadcom Monitor)监视层和硬件驱动程序层(driver)。

5) 核心层: 包含了实时操作系统 VXWORKS, BSP 和 SSP。BSP 为板集支持包(Board Support Package), 提供了软件和硬件芯片的接口函数, BSP 是针对 CPU(Central Processing Unit)及一些外围设备的, 完成对诸如串口、RAM(Random Access Memory)、Flash、实时时钟、非易失存储器等的操作。SSP 为交换支持包(Switch Support Package), 用于驱动 BROADCOM 公司的交换芯片, 完成了 TMS 与 SDK(Software Development Kit)的接口功能。

用户可通过 Web、console 或 telnet 实现对 Diffserv 的策略规则的设定和配置, 整个系统使用 IDB+(Information Database Plus)作为管理引擎, 应用层模块通过与 IDB+交互, 实现对 Diffserv MIB 库信息的存取, 通过 Diffserv MIB 库, 将配置的信息传递给协议层的 Diffserv 处理模块, 并经硬件适配层及 SSP 完成对 BCM5615 交换芯片的配置和操作。Diffserv 模块的添加需在 IDB+引擎注册, 在系统启动时 IDB+引擎就自动实现了对 Diffserv MIB(Management Information Base)模块的注册。

对于 Diffserv 模块, IDB+主要完成 Diffserv Module 的注册、Diffserv 参数的写入和读取以及配置信息的存储, Diffserv 功能的具体实现是在 IDB+的下层(包括协议层、链路层、适配层、核心层)来完成。

4 实验测试

采用 SmartBits 网络测试分析仪对交换机 QoS 性能进行测试, 用 SmartBits 的 ML7710 卡的 4 个 100 M 端口发送数据包, 用 1 个 100M 端口接收数据, 在 SmartBits 测试仪的发送端分别采用 80 byte, 512 byte 和 1 518 byte 长度的数据包发送数据。对交换机的出口配置为 4 个优先级队列, 即 COS(Class Of Service)q0~COSq3, 并建立 802.1p 优先级与优先级队列的映射关系。

当 SmartBits 每个发送端同时以 50 M 的带宽速率发送数据时, 则在交换机的出口端产生 $4 \times 50 \text{ M} = 200 \text{ M}$ 的带宽, 即产生 200% 的负载率, 在采用 SP 调度方式对出口的服务类别(COS)队列进行调度时, 在 SmartBits 接收测试得到了 4 个不同优先级的数据包的损失率, 见表 1。

从表 1 中可以看到, 当产生拥塞时, 交换机将完全丢弃优先级低的数据包而转发高优先级的数据包。当采用 WRR-MBD 调度时, 设置各优先级队列所分配的权重如表 2 所示。

表 1 在 SP 调度 200%负载率下的数据包损失率
Table1 Packet loss rate of 200% load ratio under SP scheduling

packet size/bytes	priority 0/%loss	priority 2/%loss	priority 4/%loss	priority 7/%loss
80	99.99	92.08	0	0
512	99.97	98.53	0	0
1 518	99.91	99.42	0	0

表 2 不同 COS 队列的 WRR 权重设置
Table2 WRR weight of different COS queue

weight of WRR	COS queue
10	COSq0 (low priority)
20	COSq1
30	COSq2
40	COSq3(high priority)

经系统测试, 得出在接收端数据包丢失的百分率如表 3 所示。可以看出, 在 WRR-MBD 调度策略下, 不同优先级数据包的丢失率基本上是按所设置的权重比例分配的, 避免了 SP 调度中存在的低优先级队列出现的“饥饿”现象。

表 3 在 WRR-MBD 调度 200%负载率下数据包损失率

Table3 Packet loss rate of 200% load ratio under WRR-MBD scheduling				
packet size/bytes	priority 0/%loss	priority 2/%loss	priority 4/%loss	priority 7/%loss
80	82.25	63.79	46.04	0
512	80.14	59.53	39.70	19.08
1 518	77.28	53.75	31.12	37.13

5 结论

针对语音、视频等实时性要求较高的业务普遍需求,本文提出一种新的区分服务模型,着重对其中的核心调度算法——最大延迟加权轮转调度(WRR-MBD)算法进行了详细阐述,同时给出了区分服务模型在3层交换机的具体实现方法。测试表明,当网络发生拥塞时,WRR-MBD调度可根据权重实现各队列的带宽的合理分配,在确保高优先级的业务得到优质服务的同时,也使低优先级业务得到基本保障。

参考文献:

- [1] Andelman N, Mansour Y, ZHU A. Competitive queueing policies for QoS switches[C]// Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms. Baltimore, USA: [s.n.], 2003: 761–770.
- [2] Sudha S, Ammasaigounden N. An aggregate marker for bandwidth fairness in DiffServ[J]. Journal of Network and Computer Applications, 2012, 35(6): 1973–1978.
- [3] WU Chin-Chi, WU Hsien-Ming, LIN Woei. High-performance packet scheduling to provide relative delay differentiation in future high-speed networks[J]. Computer Communications, 2008, 31(10): 1865–1876.
- [4] Yusuf Sani, Mohamed Othman. A provision-aware fair bandwidth distribution marker algorithm for DiffServ networks[J]. Journal of Network and Computer Applications, 2011(2): 715–721.
- [5] Mary Looney, Oliver Gough. A provision aware proportional fair sharing three colour marker[J]. Journal of Network and Computer Applications, 2013, 36(1): 476–483.
- [6] Jung E S, Ranka S, Sahni S. Topology aggregation for E-science networks[C]// 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. Melbourne, Australia: [s.n.], 2010: 530–533.
- [7] Szymanski T H. Future internet video multicasting with essentially perfect resource utilization and QoS guarantees[C]// 2011 IEEE 19th International Workshop on Quality of Service (IWQoS). San Jose, CA: IEEE, 2011: 1–3.
- [8] Barabas T. A traffic engineering algorithm for differentiated multicast services over MPLS network[C]// 7th IEEE International Symposium on Applied Computational Intelligence and Informatics. Timisoara, Romania: IEEE, 2012: 185–190.

作者简介:



夏俊(1971–),男,河南省信阳市人,硕士研究生,主要研究方向为通信芯片设计、无线传输技术及两化融合过程中底层设备与管理信息系统通信处理. email: sharisxia@163.com.

(上接第906页)

- [10] 张月雷. 航空维修 IETM 关键技术研究 and 应用框架设计[D]. 南京: 南京航空航天大学, 2007. (ZHANG Yue-lei. Key technologies research and applied frames design on the IETM of aeronautic maintenance[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2007.)

作者简介:



余蓉(1969–),女,四川省荣县人,本科,工程师,主要从事计算机软件产品的设计、开发、IETM技术研究. email: yrfine@sina.com.

胥永康(1973–),男,四川省遂宁市人,本科,研究员,主要从事计算机软工工程、网络信息安全方面的研究.

吴家菊(1978–),女,四川省资阳市人,硕士,高级工程师,主要从事计算机软工工程、IETM方面的研究.

王峰(1988–),男,四川省遂宁市人,硕士,工程师,主要从事计算机软件设计开发、可靠性研究.