文章编号:1001 - 2486(2005)05 - 0025 - 05

TCP 友好的路由器设计*

吕高锋,胡晓峰
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:TCP 流是网络中主要数据流,TCP 协议处理机制的研究有助于我们认知网络内部复杂特性,合理利用网络资源,这些对于路由器设计具有重要意义。根据 TCP 协议处理机制,在分析 TCP 协议加增倍减的拥塞控制模式对报文缓冲的影响和 TCP 流保序性要求对报文处理的影响的基础上,对路由器的设计中报文缓冲和报文分配等关键问题提出了 TCP 友好的解决方案。TCP 友好的路由器设计有利于提高路由器处理能力。

关键词:路由器体系结构;TCP 拥塞控制;报文乱序
中图分类号:TP393 文献标识码:A

TCP Friendly Design of Router

LU Gao-feng, HU Xiao-feng
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Researching on the characteristic of TCP flow helps to cognize the complex structure of network and to make use of the resource of network effectively. Referring to the characteristic of TCP flow and analyzing TCP congestion control of additive increase multiplicative decrease and packet out-ordering, we present TCP friendly design of router which accelerates the processing of router.

Key words: router; TCP congestion control; packet out-ordering

核心路由器是 Internet 的基础,路由器主要用于连接各子网络并完成报文转发,由查找路由表、交换、输出缓存、输出调度等模块实现主要功能。路由器的结构如图 1(a)所示,主要由三大部分组成,线卡(Line Card)用于连接与之相连的各子网;交换机构(Switch Fabric)用于在路由器内部连接各线卡,实现数据交换;主控卡(Main Control Card)主要完成建立路由表等管理功能。

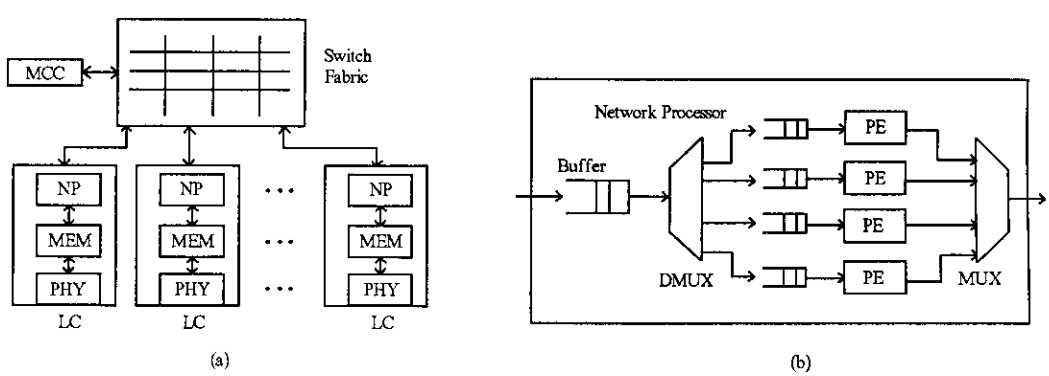

图 1 路由器体系结构

Fig.1 The architecture of router

下一代 Internet 中,40Gps 的光通信技术将极大提高链路速率。根据缓冲区计算的经验公式,路由器

* 收稿日期:2005 - 05 - 20
基金项目:国家重点基础研究发展计划项目(2003CB314802);国家自然科学基金项目(90104001);国家 863 高技术研究发展计划基金项目(2003AA115130)
作者简介:吕高锋(1980—),男,博士生。

需要10Gbits的存储器,如此巨大的缓冲区,实现难度很大。同时需要更高处理速率的电子器件,根据摩尔定律可知,微电子器件处理能力每18个月翻一番,仍然滞后于链路速率的增长,不能满足线速处理的性能要求。为了弥补电子器件处理能力的不足,路由器体系结构不断进行创新。目前路由器普遍采用并行处理方式,各线卡中多个报文处理引擎(Process Engine)并行处理,如图1(b)所示,以提高处理能力。

已有的路由器设计方法注重提高路由器处理能力,对路由器处理对象网络数据流等因素研究不足。路由器作为一种特殊的信息处理的计算机,可以根据其处理对象的特征,进行优化和设计。分析和研究网络协议处理机制,可以提供一条有效地探索网络内部运行机制的途径。TCP流是骨干网中主要的数据流,因此重点分析TCP协议处理机制。通过研究TCP协议处理机制对路由器设计中报文缓冲、报文处理等关键问题的影响,提出了TCP友好的路由器设计方法。

1 TCP

1.1 TCP流

美国NLANR(National Laboratory for Applied Network Research)研究中心对不同的学术和研究机构接入骨干网的数据流进行了长期的监测。对采样数据用四个度量标准进行统计,分别是字节数、报文数、数据流数和源目的对的数量。前两个标准是主要的,可直接测量得到;后两个标准是通过对数据处理和汇聚得到的。再对采集的数据流根据报文协议进行分类,主要有TCP、UDP、ICMP、IGMP、GRE、ESP、SKIP等协议。TCP流占据总流量的60%~90%,UDP流占据总流量的10%~40%,其他协议小于5%[1]。统计结果如表1所示,表明TCP是Internet中主要的传输协议。

表1 网络流量统计

Tab.1 Statistic of the internet flow

Metric	TCP(%)	UDP(%)	Other(%)
Bytes	(83±11)	(16±11)	(1±1)
Packets	(75±12)	(22±11)	(3±2)
Flows	(56±15)	(33±10)	(11±7)

TCP流作为网络中主要的数据流,因此重点对TCP流特性进行分析,研究TCP协议处理机制。

1.2 TCP拥塞控制机制

如果链路承载的数据量超过了实际运载能力,就会引起网络拥塞。如果对网络流控制不好,将会出现网络崩溃,导致吞吐量下降,网络性能降低。TCP Reno是一个高效的拥塞控制算法,采用滑窗机制控制源端发送速率,同时实现了快速重传和快速恢复算法,较好地解决了网络拥塞可能造成的吞吐量急剧下降的网络崩溃问题。拥塞窗口(cwnd)更新算法如下所示。

(1)收到一个非重复的确认ACK

慢启动阶段:if cwnd < ssthresh,then cwnd = cwnd + 1;

拥塞避免阶段:cwnd = cwnd + 1/cwnd。

(2)收到3个重复的ACK

ssthresh = cwnd/2,ndup = 3;

重传ACK指示丢失的报文;

cwnd = ssthresh + ndup;

当新的报文的ACK到达时,cwnd = ssthresh。

(3)重传计时器超时

ssthresh = W,cwnd = 1。

其中,ssthresh是拥塞窗口慢启动阈值。ndup是重复的ACK的数。

当收到同一个报文段的3个重复的ACK,TCP Reno触发快速重传机制,发送ACK指示丢失的报文,而不用等待超时计时器超时才重传,这样会提高网络的利用率和吞吐量。

从TCP拥塞控制机制出发,对TCP协议处理机制进行研究,探讨TCP处理方式对路由器设计中若干关键问题的影响。

2 TCP流对报文缓冲的影响

从排队论的角度来看,网络中是队列的集合,每个队列有一个缓冲区(Buffer)临时保存到达的报文。

报文到达后,缓存转发,则会产生延迟。若达到的报文大小超过缓冲区容量,则会丢弃报文,需要发送者对丢弃的报文进行重传,导致网络吞吐量降低[2]。因此路由器设计中缓冲区大小的设置非常重要。

2.1 缓冲区大小计算的经验公式

Villamizar 通过实验得到经验公式 $B = \overline{RTT} \times C$,计算 TCP 流所需要的缓冲区大小[3]。如果缓冲区设置满足经验公式,在网络瓶颈处的路由器的缓冲区不会被读空,吞吐量不会下降。TCP 流通过瓶颈带宽需要的缓冲区大小等于带宽和延迟两者的乘积,能够有效防止链路空闲,损失吞吐量。参考网络流模型,如图 2 所示。TCP 源连续地发送报文,报文大小是固定的。这条流通过路由器缓冲区 B,到达接收者。若发送者的接入链路速率 C' 大于接收者瓶颈链路 C,报文必须在路由器中进行缓存。发送者到接收者的传播延迟记作 T_P。

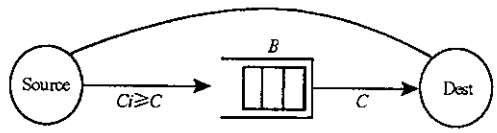

图 2 网络模型

Fig.2 Model of the flow

从时刻 t_0 开始,发送者逐渐增加窗口大小,填充缓冲区,直到缓冲区开始丢弃第一个报文。在一个往返传输延迟之后,因为发送者等待被已经丢弃报文的 ACK,在 t_1 时刻超时。它立即把拥塞控制窗口减半,从 $W_{\max}$ 减小到 $W_{\max}/2$。发送方等待接收方的应答报文,当接收到 $W_{\max}/2$ 应答报文后,发送方可以继续发送报文。发送者拥塞窗口变化如图 3 所示。

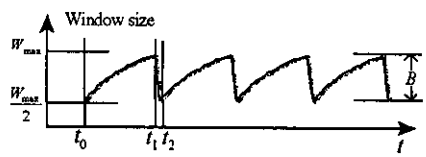

图 3 TCP 流量的‘锯齿’特性

Fig.3 Characteristic of the TCP flow

如果缓冲区被读空,则路由器以恒定的速率 C 向瓶颈链路发送报文,这也意味着 ACK 到达速率为 C。发送者等待的时间是 $(W_{\max}/2)/C$,然后它才可以恢复传输,再次增加窗口大小。而读空缓冲区内报文时间是 B/C。在 t_2 时刻缓冲区不会被读空,仅当发送者重新发送的第一个报文在缓冲区被读空之前到达路由器,即 $(W_{\max}/2)/C \leqslant B/C$,故 $B \geqslant W_{\max}/2$。我们知道 TCP 发送报文的速率是 $R = W/RTT$,即 $C = W/RTT = W_{\max}/2T_P$。因此 $W_{\max}/2 = 2T_P \times C$,可推导出经验公式 $B \geqslant 2T_P \times C = \overline{RTT} \times C$。

2.2 缓冲区大小计算

经验公式由来的基础是认为骨干网中数据流是同步的。而目前骨干网路由器中处理成千上万的流,这些流有不同的 RTT 值,是不完全同步的。RTT 或者处理时间微小的变化已经足以防止流同步。在网络流量特性已经改变的情况下,我们有必要重新审视经验公式。

如果‘锯齿’不是同步的,更多流的加入,它们窗口的汇聚效果就更不像‘锯齿’。它们彼此平滑了峰值,汇聚以后的拥塞窗口峰值到槽值的距离更小,如图 4 所示。2004 年,Appenzeller 研究了新的网络环境下报文缓冲与 TCP 流的关系[4]。

假设一组流有随机(互相独立)的起始时间和传播延迟,可以认为它们是非同步的,窗口 $W_i(t)$ 的变化过程也是互相独立的。总的窗口 W 大小变化可以看作一个随机过程,由多个独立的锯齿组成。由中心极限定理得知,窗口变化过程汇聚以后可以转化成一个高斯过程。

从窗口变化过程,我们知道 t 时刻队列占用情况:$Q(t) = \sum_{i=1}^{n} W_i(t) - (2T_P \times C) - \varepsilon$,所有未应答的报文在队列 $Q(t)$ 中,或者在链路 $(2T_P \times C)$ 中,或者被丢弃,ε 表示被丢弃的报文。如果缓冲区足够大,TCP 正常处理,ε 与 $(2T_P \times C)$ 相比,可以忽略不计。因此,$Q(t)$ 的分布可以记作 $Q^d = W - 2T_P \times C$。

因为 W 是正态分布,则 Q 也是正态分布,有一个偏移量。如果知道 $(Q < b)$ 的概率,即可得到链路利用率的上限。因为 Q 是正态分布,根据缓冲区设置值 b 利用 error-function 估算该概率,计算链路利用率的下界(参考文献[4]对此进行了详细的说明)。

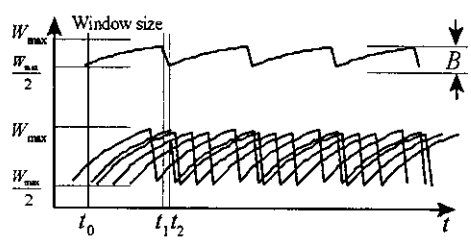

图 4 非同步的 TCP 汇聚

Fig.4 Aggregate of desynchronized TCP flows

以上分析过程中,只考虑了长的 TCP 流。网络中还存在

一定量的短流。在慢启动阶段,缓冲队列吸收了短的突发数据,而在拥塞避免阶段,它适应窗口大小的缓慢改变。研究发现,短流对队列的影响很小,缓冲区的大小主要由长流的流量特性决定。

2.3 TCP友好的报文缓冲区设计

根据分析得到的链路利用率估计公式,就可以计算得到一定链路利用率下所需要的缓冲区大小。比如我们假设网络中有10 000条数据流,计算缓冲区大小,如表2所示,其中,$B_0 = \frac{2T_P \times C}{\sqrt{n}}$。

新的缓冲区大小计算公式从TCP的处理机制出发,重新考虑了TCP流"锯齿"变化的汇聚效果,以网络中TCP流为模型。根据新的计算公式设置缓冲区,融入了TCP的因素,从而形成TCP友好的设计方法。从G.Appenzeller的推导过程可以看出,这种设置报文缓冲区大小的方法,不仅可以满足报文缓冲的要求,不降低链路的利用率。同时报文通过较小的缓冲区,经历较小排队延迟,还能够提高报文的处理速率。

表2 不同链路利用率要求下缓冲区大小

Tab.2 Buffer size with differentiated utilization

Utilization	Buffer
$Util \geqslant 98.99\%$	$B = 1 \times B_0$
$Util \geqslant 99.99988\%$	$B = 1.5 \times B_0$
$Util \geqslant 99.99997\%$	$B = 2 \times B_0$

对于40Gb/s的线卡,假设同时处理40 000条流,每条流的大小是1Mb/s,根据经验公式,需要的缓冲区是10Gbits。实现时需要用到片外存储器,如DRAM,然而存取速度比较慢。根据新的计算公式,所需要的缓冲区为50Mbits,可以用片上存储器实现,如SRAM,而且速度比较快。可以看出,新的缓冲区大小计算公式计算得到缓冲远远小于经验公式计算值,大大减小路由器中报文缓冲实现的复杂度,为Internet中新型路由器的设计提供了理论指导。

我们知道TCP"锯齿"形的拥塞控制算法通过报文丢失来推测网络拥塞情况,因此发送者可能会填满网络中间节点处任意大的缓冲区。不管我们把瓶颈处的路由器的缓冲区做得如何大,TCP都可能将缓冲区填满。如果路由器的缓存设置很大以避免报文丢失,维持高的吞吐量,那么将会导致更大的延时。

3 TCP流对报文处理的影响

TCP提供可靠的字节流服务。发送者将字节流发送到TCP连接上,接收者窗口必须出现同样顺序的字节流。当TCP出现报文乱序时,接收者必须对数据进行重新排序,因此TCP是一个对报文乱序敏感的协议。目前并行处理是提高路由器性能的重要方法,并行处理可能导致报文乱序,从而影响处理TCP报文速率。

3.1 报文乱序

如果TCP报文迟于其正常顺序3个以上报文才到达,称其为深乱序报文,否则为浅乱序报文。已有的研究表明,浅乱序只会减缓发送者拥塞窗口的增长,因为发送者会丢弃重复的确认[5]。根据TCP Reno协议可知,深乱序报文使发送者触发不必要的快速重传和快速恢复。在拥塞避免阶段,如果发生报文丢失,发送者将进入快速恢复阶段,拥塞窗口降为原来的一半,快速恢复阶段将持续 n 个 RTT,n 表示丢弃的报文数。如果存在深乱序报文,发送者也将拥塞窗口大小减为原来的一半。

浅乱序和深乱序对TCP将产生完全不同的影响。浅乱序仅仅使发送者拥塞窗口的增长速率降低,报文的发送速率没有急剧变化;而深乱序使拥塞窗口的大小突然减半。下面主要分析存在深度乱序时TCP的发送速率。

3.2 报文乱序对TCP性能的影响

假设TCP进入了稳定状态,当拥塞窗口增长到 b_n 时,网络发生拥塞,TCP将拥塞窗口减半设为 $b_n/2$。然后拥塞窗口以一个 RTT 时间增加1个报文段大小进行线性增长,增加到 b_n 时,再次发生拥塞。我们将这样一个过程称为拥塞周期,如图5所示。TCP的发送速率可以用一个拥塞周期内的报文发送速率表示。

用 T 表示报文发送时间,以 RTT 为单位,S 表示发送的报文数,则TCP的发送速率为:$R_1 = \frac{3}{4} b_n$。

如果在拥塞窗口为 c_n 时发生了深乱序,则拥塞窗口将减半为$\frac{c_n}{2}$,然后再线性增加到 b_n,则 TCP 的报文发送速率为:$R_2=\frac{3(b_n^2+c_n^2)}{4(b_n+c_n)}$。性能下降比例为:$\eta=\frac{R_2}{R_1}=\frac{b_n^2+c_n^2}{b_n^2+b_nc_n}$,其中 $a=\frac{b_n}{c_n}(0.5<a<1)$。

假设深乱序是平均分布,则 a 的平均值为 0.75,此时 $\eta=0.89$。即在一个拥塞周期内存在 1 个深乱序时,TCP 的传输性能下降了 10%。存在 2 个深乱序时,得到 $\eta=0.80$,性能下降更多。因此,深乱序严重降低 TCP 的报文发送速率,降低 TCP 的性能。

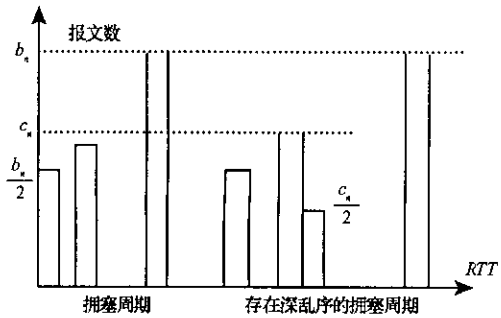

图 5 拥塞周期内报文发送数

Fig.5 The number of outstand packets

3.3 TCP 友好的报文分配

在如图 1(b)所示的路由器结构中,线卡中存在多个并行的报文处理引擎。多个报文并行处理时,如果先到达的报文经历较长的处理延时,而后续报文的处理延时较小,那么它将有可能先于前面的报文离开系统,从而在多个并行处理的报文之间形成乱序,可能降低网络性能。因此在并行报文处理方式下必须在一定程度上保证分组顺序,特别是严格保证不增加 TCP 中的深乱序的出现概率。

在报文并行处理方式下,报文分配以及处理引擎负载均衡必须针对并行度和报文保序做出折衷,即不能使 TCP 处理性能下降,又要保证报文处理引擎发挥较好的并行性,同时具有较高的利用率。在网络负载较轻或突发性较弱时,报文丢失概率很小。在重负载或强突发条件下,路由器队列变长,报文的丢失概率增加。因此在轻负载或弱突发情况下需要优先考虑报文的保序问题,而重负载或强突发情况下可以优先考虑负载均衡。结合 TCP 报文保序要求,形成 TCP 友好的报文分配方式,如图 6 所示。假设有 N 个独立的并行处理报文处理引擎,第 i 个引擎记为 $N_i(0\leqslant i<N)$。利用 Hash 函数根据流标识产生哈希值。哈希值作为表索引,通过查表最终确定报文处理引擎。

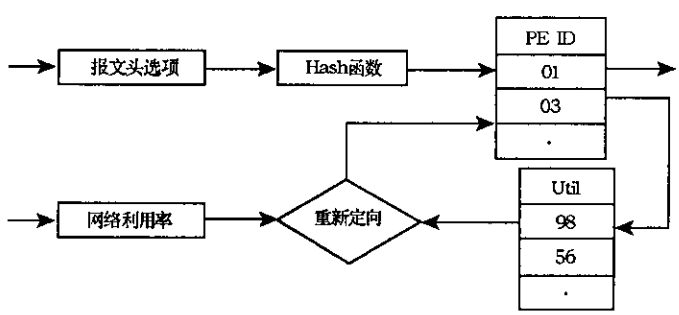

图 6 TCP 友好的报文分派方式

Fig.6 TCP friendly distribute of packets

TCP 友好的报文分派方式通过映射表建立报文和处理引擎的映射关系,能够根据网络负载和各处理引擎负载动态调整二者的映射关系,能够实现负载均衡。同时综合考虑网络负载和 TCP 报文处理保序性要求,对报文保序性提供了有力保证。

4 结束语

TCP 拥塞控制机制以及报文乱序的研究对路由器的设计具有重要指导价值。根据已有的 TCP 协议处理机制研究成果,在分析 TCP 流加增倍减特性对报文缓冲影响和 TCP 流保序性要求对报文处理影响的基础上,针对路由器的设计中报文缓冲和报文分配等关键问题提出了 TCP 友好的解决方案,TCP 友好的路由器设计方法,综合了 TCP 对路由器的要求,有利于提高处理 TCP 报文的能力,提高网络利用率。

(下转第 94 页)

4 结 论

求解了液体火箭尾焰辐射在吸收、散射及自身辐射性介质中传输的简化微积分方程，主要考虑高温气体分子的红外辐射，略去散射在所研究方向引起的辐射增量，计算结果与实测值的比较说明方法是可行的。

参 考 文 献：

[1] Ludwig C B. Handbook of infrared radiation from combustion gases[R]. NASA SP-3080,1973.

[2] Ludwig C B, Malkmus W, Walker J, et al. A Theoretical model for absorbing, emitting, and scattering plume radiation. Hortn T E(Ed). Spacecraft Radiative Transfer and Temperature Control[C]. New York: AIAA Progress Series in Astronautics and Aeronautics, 1982,83: 111-127.

[3] 徐南荣.喷气流红外辐射场的数值计算[J].航空动力学报.1995,16(6):647-653.

[4] 聂万胜,杨军辉,何浩波,庄逢辰.液体火箭发动机尾焰流场及其燃烧组分的谱带模型参数计算[A].装备指挥技术学院学术交流会论文集[C].北京:军事科学出版社,2005.

(上接第29页)

参 考 文 献：

[1] 李枝清,梁阿磊,彭路. 高速路由器设计方案研究[J]. 计算机工程,2001,27(8):94-96.

[2] 林闯,周文江,李寅,等. 基于Intel网络处理器的路由器队列管理:设计、实现与分析[J].计算机学报,2003,26(9):1069-1077.

[3] Villamizar C,Song C. High Performance TCP in ANSNET[J]. ACM Computer Communications Review, 1994, 24(5):45-60.

[4] Appenzeller G,Keslassy I. Sizing Router Buffers[A]. SIGCOM'04[C], 2004.

[5] 胡晓峰,孙志刚,苏金树,等. 高速路由器并行交换技术研究[J]. 计算机研究与发展,2004,41(1):60-64.

(上接第66页)

[2] Sfakiotakis M, Lane D M, Davies J B C. Review of Fish Swimming Modes for Aquatic Locomotion [J]. IEEE J of Oceanic Engineering, 1999, 24(2):237-252.

[3] 王光明,胡天江,等. 长背鳍波动推进游动研究[J]. 机械工程学报, 2005,(10).

[4] Consi T R, Seifert P A, Triantafyllou M S, et al. The Dorsal Fin Engine of the seahorse (Hippocampus sp.) [J]. Journal of Morphology, 2001, 248:80-97.

[5] 施法中. 计算机辅助几何设计与非均匀有理B样条(CAGD&NURBS)[M]. 北京:北京航空航天大学出版社,1994:22-25.

[6] 孙世贤,黄圳圭,等. 理论力学教程[M]. 长沙:国防科技大学出版社.1997:99-102.

[7] Hu T J, Li F, Wang G M, et al. Morphological Measurement and Analysis of Gymnarchus Niloticus [J]. Journal of Bionics Engineering, 2005, 2(1):25-31.

[8] 胡天江,李非,沈林成."尼罗河魔鬼"长背鳍波动包络线的提取算法[J]. 国防科技大学学报, 2005, 27(6).