一种区分服务综合方案的模型与性能分析

盛立杰,林 闯,吴建平
(清华大学计算机科学与技术系,北京 100084)

摘 要: 本文提出了一种将 Internet 网络分组传输延时和丢失控制的区分服务要求相结合的综合方案,它具有较强的通用性和灵活性. 文章给出了 PHB 实现机制的一种基于随机 Petri 网模型的分析框架,并给出了性能指标的分析和数值结果. 模型求解采用了一种分解、迭代的近似方法,可以有效降低求解复杂度. 近似分析结果和模拟结果的比较证明,这种近似求解方法是可行的.

关键词: 区分服务; 随机 Petri 网模型; 性能分析
中图分类号: TP393 **文献标识码**: A **文章编号**: 0372-2112 (2000) 11A-0032-04

Modeling and Performance Analysis of a Differentiated Services Scheme

SHENG Li-jie, LIN Chuang, WU Jian-ping
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: This paper proposes an integrative scheme of Differentiated Services (DiffServ) for the Internet. In our scheme, qualities of service are described with the scales of importance and urgency integratively, which makes the service negotiation and implementing mechanism more flexible and universal. PHB implementing mechanism is modeled with Stochastic Petri Net, and its performance is analyzed, with an approximate analysis technique to reduce the complexity.

Key words: differentiated services; stochastic Petri nets; performance analysis

1 引言

现有的 Internet 网络仅提供单一的尽量做好(best effort)服务,不能满足不同客户、不同应用对服务质量(QoS)的不同要求. 因此服务质量控制问题已成为近年的研究热点[1]. 区分服务[2](Differentiated Service, DiffServ)是 IETF 近年来提出的一套 QoS 提供框架. 它将复杂性推到网络边界,而在网络内部节点仅实现简单的面向少量流聚集的调度算法,因此具有良好的可扩展性. 按照服务质量的定义方法,区分服务方案可以分为"绝对的"区分服务和"相对的"区分服务两类[3]. 绝对区分服务以 QoS 指标的绝对数值度量服务质量;而相对区分服务只强调不同服务级别间的相对关系. 奖赏服务(Premium Service, PS)可以看作是一种绝对的区分服务,而确信服务(Assured Service, AS)则只在丢失率方面刻画了服务质量的相对性,延迟方面的相对关系目前没有规范. 本文提出的综合方案属于相对的区分服务,服务质量以延迟和丢失率两个方面的相对级别来刻画.

本文的另一部分重要工作是给出了一种针对区分服务实现方案的较为通用的性能分析框架. 在区分服务研究领域,目前获得的性能评价结果大多利用模拟方法获得,基于数学模型的分析结果较少[4]. 本文采用随机 Petri 网来模型综合方案,并给出了分析求解方法. 这种模型分析方法在输入流为泊松和自相似两种信源模型时都可以应用.

2 建议方案的描述

网络传输的服务质量要求主要可以归结为信息传输的实时性与可靠性两个方面,因此本方案中不同服务水准的划分以"实时性"与"可靠性"为标准. "实时性"在具体 QoS 指标上体现为延迟、抖动,"可靠性"体现为丢失率. 在总体结构上,综合服务方案分别实现于传输路径中的三个部分:客户端、网络边界与网络内部节点. 在客户端,客户应用流依据实时性与可靠性的不同需要选择不同的服务水准,在分组包头的区分服务标记域(DS field)做预标记. 在网络边界节点,根据服务协议(Service Level Agreement, SLA)制定方式的不同,有两种处理方式. 一种是在 SLA 中指定客户流在不同服务级别中的流量大小与特性;边界节点根据 SLA 调节客户流,通过整形、降级、丢弃等手段强制其符合 SLA 中定义的流规范(profile),这是传统的区分服务框架中的处理方式. 另一种是 SLA 中只说明不同服务级别的收费方式,但对客户流的流量不做任何假设和限定;边界节点既不改变客户流中分组的服务级别,也不专门丢弃任何分组,而只是记录客户在不同服务级别中的流量大小,

收稿日期:2000-06-14;修回日期:2000-10-08
基金项目:国家高技术研究发展计划(No. 863-306-ZT05-01-02;No. 863-300-05-04-02-00);国家重点基础研究发展规划项目(No. G1999032707);国家自然科学基金项目(No. 69873012)

作为收费依据.后一种方式的好处是处理、控制简单,而且为用户动态选择服务质量提供了便利.无论采用哪种方式,经边界节点最终进入区分服务网络的是实时性与可靠性各不相同的流聚集(Traffic Aggregate).在区分服务网络内部,内部节点根据分组包头标记,通过实时优先级调度与多级别丢弃控制相结合的综合算法,为不同流聚集提供不同质量的转发服务,其外特性就是区分服务框架中所谓的逐点行为(PHB Per-Hop-Behavior).

在网络内部的每一个节点处,PHB 的实现机制可抽象为一个发送调度器和多个缓冲队列(如图 2 所示)构成的模型.不同实时要求的流聚集使用不同的缓冲队列;同一实时级中不同丢弃优先级的流聚集共享同一缓冲队列.设每个节点提供 n 个实时优先级,并假设流实时调度优先级按序号反向排列,亦即,实时级为 i 的流的实时调度优先级高于 $i+1$ 的流.各实时级中又分多个丢弃优先级,实时级 i 中的丢弃优先级数的最大值记为 ND_i,丢弃优先级为 1 的流聚集丢弃概率最大、可靠性最低,丢弃级为 ND_i 的丢弃概率最小、可靠性最高.记实时级为 $i(1 \le i \le n)$ 丢失级为 $j(1 \le j \le ND_i)$ 的流聚集为 $Flow_{ij}$,实时级为 i 的流进入的缓冲队列记为 Q_i,其最大空间记为 B_i,队列是头向尾排序,它是一个 FIFO 队列见图 1.

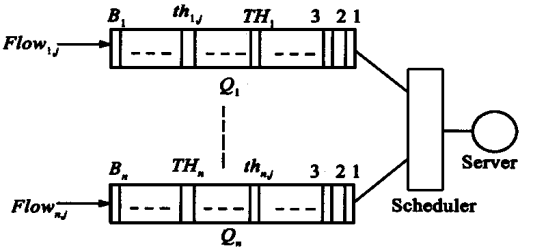


图 1　调度方案的抽象模型

在模型中用两类参数分别描述实时调度与丢弃控制算法.实时调度与丢弃控制有多种算法可以选择,不同的算法在可控性和效率等方面各有千秋.这里限于篇幅,只以最简单的队列长度阈值(queue length threshold,QLT)调度方案[5]和阈值控制(threshold control,TC)丢弃算法为例说明.

QLT 的调度依据是缓冲队列长度与调度阈值的相对大小关系.以 TH_i 表示 Q_i 的调度缓冲阈值.当 Q_i 中的缓冲占有值 $M(Q_i)$ 达到或超过 TH_i,即 $M(Q_i) \ge TH_i$,并且更高实时优先级队列的缓冲占有值都小于它们自己的调度缓冲阈值(即 $\forall 1 \le j < i, M(Q_j) < TH_j$)时,发送权将交给 Q_i.若所有队列的缓冲占有值都小于其调度阈值,则非空队列中优先级最高的将获得发送权.

TC 的丢弃策略也根据队长和丢弃阈值的大小关系确定.令 th_{ij} 表示 $Flow_{ij}$ 的丢弃阈值,满足 $1 \le th_{ij} \le B_i$, $ND_i = B_i$.并且对同一实时级 i,丢弃级 j 越小其阈值 th_{ij} 也越小,即 $1 \le j < k \le ND_i = > 1 \le th_{ij} \le th_{ik} \le B_i$.当 $Flow_{ij}$ 中的分组到达时,若 $M(Q_i) < th_{ij}$,则分组进入 Q_i,否则此分组被丢弃.

3　方案的随机 Petri 网模型

假设读者对随机 Petri 网(SPN)的理论和应用有一些基本的了解.有关 SPN 的详细描述可参阅文献[6,7].在 SPN 模型中,缓冲队列可由位置来表示,缓冲占有程度可由位置的标识(marking)表示.在模型中,允许变迁的可实施条件可用变迁的谓词规定.对变迁的实施阻止概率或服务竞争条件概率可由变迁实施概率来规定.

图 2 给出了网络内部节点的队列调度方案的 SPN 模型,此模型已经过精化设计[8].模型中的变迁和位置的描述含意如下:

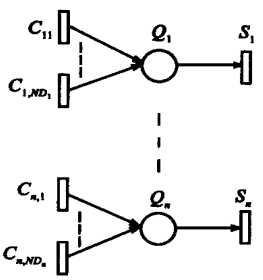


图 2　调度方案的 SPN 模型

c_{ij}:表示 $Flow_{ij}$ 中分组的到达,到达过程为泊松到达,平均速率为 λ_{ij}.

s_i:表示 Q_i 中分组的发送服务,服务时间是负指数分布的,服务速率为 i.这里假设同一实时优先级的分组的服务速率相同.

Q_i:实时级为 i 的分组进入的队列,容量为 B_i,其中 token 数即实际队列长度为 $M(Q_i)$.

在模型中,变迁 s_i 和 c_{ij} 的相关联的实施概率分别由 x_i 和 p_{ij} 表示,因此这两个变迁的实际实施速率分别为 $x_i \times \mu_i$ 和 $p_{ij} \times \lambda_{ij}$. s_i 的谓词描述队列实时调度方案.对于 QLT 方案,s_i 的谓词为:

$$\left\{\neg(M(Q_i) \ge TH_i) \wedge (\forall j, 1 \le j < i, M(Q_j) < TH_j)\right\} \vee \left[\begin{array}{l}(\forall j, 1 \le j < i, M(Q_j) = 0) \wedge (M(Q_i) > 0) \wedge \\ (\forall j, i < j \le n, M(Q_j) < TH_j)\end{array}\right] \quad (1)$$

也可以使用 s_i 的实施概率 x_i 来描述 QLT 调度方案:

$$x_i(M(Q_i)) = \begin{cases} 1, \text{in the condition of (1)} \\ 0, \text{otherwise} \end{cases} \quad (2)$$

c_{ij} 的实施概率 p_{ij} 描述分组丢弃控制方案,对于 TC 可写为

$$p_{ij}(M(Q_i)) = \begin{cases} 1, \text{when } 0 \le M(Q_i) < th_{ij} \\ 0, \text{otherwise} \end{cases} \quad (3)$$

4　性能分析

对于图 2 中的 SPN 模型,可以直接构造与之同构的马尔可夫链并写出其状态转换矩阵,进而可以获得所有状态的稳定状态概率.这是通常解 SPN 模型最直接的方法.用 $P[M]$ 表示状态 M 的稳定状态概率.则 $Flow_{ij}$ 的丢失率 L_{ij} 可以表达为:

$$L_{ij} = 1 - \sum_{k=0}^{B_i} p_{ij}(M(Q_i) = k) \times P[M(Q_i) = k] \quad (4)$$

$Flow_{ij}$ 的吞吐量为

$$T(s_{ij}) = \lambda_{ij} \times (1 - L_{ij}) \quad (5)$$

实时级为 i 的流的吞吐量 $T(s_i)$ 为:

$$T(s_i) = \sum_{1 \le j \le ND_i} T(s_{ij}) \quad (6)$$

当分组到达缓冲队列 Q_i 时,如果发现有 k 个分组在队列中,那么系统的响应时间就是 $k+1$ 个分组的服务时间.则 $Flow_{ij}$ 的分组在缓冲队列 Q_I 中的平均延时 D_{ij} 可以表达为:

$$D_{ij}=\left(\sum_{k=0}^{B_i-1}\frac{(k+1)}{T(s_i)}\times P[M(Q_i)=k]\times p_{ij}(M(Q_i)=k)\right)/(1-L_{ij}) \quad (7)$$

上式分母中的$(1-L_{ij})$表示平均延时的计算仅考虑未丢失分组,不包括丢弃分组.

当模型规模较小时,可以考虑采用上述直接求解方法.但在一般情况下,上述模型是一个 n 维的马尔可夫链,随着 B_i 和 n 的增大,将出现状态爆炸问题.一种可能的解决方法是基于模型分解和迭代求解子模型的近似求解方法.

由于图 2 的模型已经过精化设计,因此每个结构独立部分就是一个子模型.子模型之间的相互影响,亦即子模型之间的输入输出参数关系,描述在变迁 s_i 的实施条件谓词中.每个变迁 s_i 能否实施,不但与本子模型的队列状态相关,而且与其他子模型的队列状态相关.在模型中,队列状态(标识)不能直接做子模型间的输入、输出参数,因为它们是时间函数;但标识在稳定状态下的概率可做为输入、输出参数.每个子模型 A_i 以所有其它子模型 $A_k(1\le k\le n, k\ne i)$ 的稳定状态下的队列占有分布概率做为输入参数.在子模型 A_i 中,记缓冲占有量的稳定状态概率为

$$\forall r, 0\le r\le B_i, G_i(r)=P[M(Q_i)=r] \quad (8)$$

则子模型 A_i 中,缓冲队列占有量小于 TH_i 的概率可表达为:

$$F_i(TH_i)=P[M(Q_i)<TH_i]=\sum_{r=0}^{TH_i-1}G_i(r) \quad (9)$$

子系统的缓冲队列竞争发送的影响可以表示为每个缓冲队列发送概率的降低.对于子系统 $A_i(1\le i\le n)$,变迁 s_i 的实施概率 x_i 可以表达为

$$x(M(Q_i))=\begin{cases}0, & M(Q_i)=0\\ \left(\prod_{k=1}^{i-1}G_k(0)\right)\times\left(\prod_{k=i+1}^{n}F_k(TH_k)\right), & M(Q_i)<TH_i\\ \prod_{k=1}^{i-1}F_k(TH_k), & M(Q_i)\ge TH_i\end{cases} \quad (10)$$

式(10)表达了来自其它子系统的输入参数(函数 G, F)对子系统 A_i 行为的影响.计算时,可以按照 A_1 到 A_n 的次序循环迭代求解.迭代过程中,每个子系统求解时都要利用最新得到的输入参数,并在求解后立即更新输出参数.初始迭代时,$F_j(TH_j)$ 的初值可设置为 0 到 1 之间的任意值.到迭代收敛时,将得到系统的近似解.

现在回到每个子系统模型的求解.考虑到变迁实施时间负指数分布的假定,在子系统 i 中,队列 Q_i 中分组个数变化的随机过程是一个马尔可夫链,更具体的,是一个简单的生死过程,因此很容易写出其乘积形式的解.限于篇幅,这里不再列出.

5 数值结果

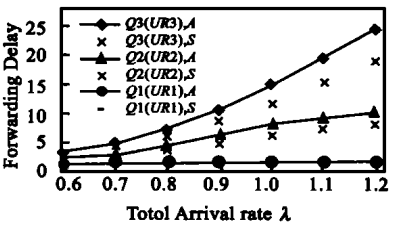
以上近似分析方法可以直接编程实现.同时,为验证近似求解方法的有效性,还编制了模拟求解程序.限于篇幅,下面仅给出一个典型模型的分析结果,此模型可以同时实现 PS 与 AS.

模型中共有 3 个队列,对应三个实时优先级.实时优先级 1 中有一个丢弃级,实时级 2 和 3 中各有 2 个丢弃级.整个模型中共实现了 5 个 PHB,在适当的参数配置下,最高实时级中的 PHB 可视为 EF,另两个实时级对应两个 AF 组,组内各有两个 AF.具体参数设置为:(1)5 个 PHB 对应的流聚集的入速率相同,若设总速率为 λ,则每个流的入速率为 $\lambda_{ij}=\lambda/5$;(2)各队列的服务速率相同,均为 $\mu=1.0$;(3)各队列的实时调度阈值为 $TH_1=1$, $TH_2=3$, $TH_3=3$;(4)各流聚集的丢弃阈值为 $th_{11}=3$, $th_{21}=5$, $th_{22}=6$, $th_{31}=8$, $th_{32}=10$.

令总入速率 λ 从 0.1 增长到 1.5,可以观察各 PHB 对应流聚集的性能指标的变化.图 3 显示了三个实时优先级的平均转发延迟随入速率的变化,图 4 显示了 5 个 PHB 对应流聚集的丢失率随总入速率的变化.分别比较每个指标的模拟与近似计算结果可见,当负载小于 0.8 时模拟与近似计算结果极其接近;而在负载大于 0.8 时模拟与计算结果有一定差别,但两者的发展趋势是相似的.这表明近似分析方法是有效的.

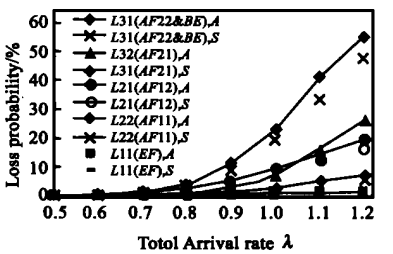
由图 3 中 Q_1、Q_2、与 Q_3 的延迟曲线可见,不同实时级分组的平均延迟差别极其显著,三条曲线的发展明显不同.而图 4 中的曲线则表明,在同一实时级中,高丢弃优先级的流聚集将有更大的丢失率.从数值结果看,综合方案有效地实现了不同级别流聚集服务质量的区分.

近期对实际网络的大量测量与分析表明,网络传输在多个时间量级上表现出二阶自相似性[9],也称为长相关性.在文[10]中,Anderson 和 Niekson 提出一种通过叠加若干个两态马尔可夫调制的泊松过程(2-state MMPP)来获得自相似流的模型方法.它可以在多个时间量级上模型流的相关结构,从而可以刻画传输流在各个时间量级上的突发性.应用文[10]中的方法,对于每个输入流 $Flow_{ij}$,用 4 个两态 MMPP 的叠加来近似其长相关性,之后利用前面的近似分析方法做了分析求解.求解时唯一的不同是,由于分解后得到的每个子模型没有简单的乘积形式解,因此需要采用 SPNP[7] 对子模型求解.限于篇幅,这里不再详述.



("A" for analytical results; "S" for simulation results)

图 3 各队列平均转发延迟的对比



("A" for analytical results; "S" for simulation results)

图 4 各流聚集丢失率的对比

5 结论与进一步的工作

本文描述的 Internet 区分服务综合方案将"实时性"与"可靠性"的组合作为服务区分的标准,易于在客户端、网络边界与内部节点上实现.文中给出了 PHB 实现机制的 SPN 模型,并给出了一种通用的性能分析方法.分析过程中所采用的分解与迭代的近似求解方法,可以将一个复杂、大规模的非乘积解的 SPN 模型分解为相对简单的子模型进行求解,可应用到多种复杂系统模型的分析中.

文中作为例子所分析的 PHB 实现机制是相对简单的 QLT + TC 组合,其优点是实现简单,但缺点也很明显,即虽然可以区分不同的服务级别,但区分的程度不易控制,而且各种阈值均为静态设定,缺乏动态性,导致队列系统的总体效率受输入的影响较大.针对这样的问题,我们已经并将继续尝试设计一些新的队列调度和缓冲管理方案,并将比较各种方案的优劣.

参考文献:

[1] 林闯. 多媒体信息网络 QoS 的控制 [J]. 软件学报,1999,10(10):1016 - 1024.

[2] 林闯,单志广,盛立杰,吴建平. 区分服务及其几个热点问题的研究 [J]. 计算机学报,Apr. 2000,23(4):1 - 15.

[3] Dovrolis C,Stiliadis D,and Ramanathan P. Proportional differentiated services:delay differentiation and packet scheduling [A]. ACM SIGCOMM [C],Sept. 1999.

[4] May M,Bolot J,Marie A,and Diot C. Simple performance models of differentiated services schemes for the internet [A]. IEEE INFOCOM'99 [C],March 1999,New York City:1385 - 1394.

[5] Lin C and Lam E C M. Dynamic queue length thresholds for scheduling real-time traffic in ATM networks [A]. Proc. of 1999 Inter. Conf. on Communications [C]. IEEE Computer Society,June 6 - 10,1999,Vancouver,BC Canada:869 - 874.

[6] Lin C and Marinescu D C. Stochastic high-level Petri nets and applications [J]. IEEE Trans. on Computers,1988,37(7):815 - 825.

[7] Ciaodo G,Muppala J and Trivedi K S. SPNP:stochastic Petri net package [A]. Proc. of the Petri nets and performance models [C],Kyoto,Japan,December,1989:142 - 151.

[8] 林闯. 随机 Petri 网模型的精化设计 [J]. 软件学报 2000,11(1):104 - 109.

[9] W. E. Leland,M. S. Taqqu,W. Willinger and D. V. Wilson. On the self-similar nature of Ethernet traffic (Extended Version) [J]. IEEE/ACM Transactions on Networking,Feb 1994,2(1):1 - 15.

[10] A. T. Andersen and B. F. Nielsen. A Markovian approach for modeling packet traffic with long-range dependence [J]. IEEE Journal on Selected Areas in Communications,June 1998,16(5):719 - 732.

作者简介:

盛立杰　1976 年生,清华大学计算机系博士研究生,主要研究方向为计算机网络体系结构,分布式系统性能评价,网络资源管理与服务质量控制.

林　闯　1948 年生,清华大学计算机系教授,博士导师,同时为《计算机学报》编委,中科院网络中心和北京科技大学兼职教授.主要研究领域为计算机网络,系统性能评价,随机 Petri 网,逻辑推演和推理系统.已在 IEEE Transactions on Computers,IEEE Transactions on Knowledge and Data Engineering,ACM Journal of Wireless Networks,International Journal of Intelligent Systems,IEICE Transactions of Fundanmentals,计算机学报,软件学报,电子学报,通信学报等国内外核心期刊上和 IEEE Computer Society 的学术年会上发表论文 60 多篇.