计算机科学1998 Vol. 25 No. 4

分组交换网络中提供有界延迟服务的调度策略*)

Scheduling Strategies Providing a Bounded Delay Service in Packet-Switching Network

孙利民　窦文华　周兴铭
(国防科技大学计算机系　长沙410073)

摘　要　Packet scheduling is one of the key components of packet-switching network with a bounded delay service. After a brief introduction to the packet scheduling, several delay-based/rate-based service disciplines are discussed in detail, various issues and tradeoffs in designing service disciplines are presented finally.

关键词　Packet-switching network, Bounded delay service, Packet scheduling, Delay-based/rate-based disciplines, Design requirements

1　引言

将来的高速计算机网络必须支持具有不同通信量特性和不同服务质量要求的各种应用。网络服务质量(QoS)参数主要包括网络端端延迟、延迟抖动、吞吐量和丢失率。网络延迟是实时服务网络最重要的参数，我们称实时服务类型中保证所有包的延迟都不超过给定延迟(以及延迟抖动)的服务为有界延迟服务。提供有界延迟服务的网络必须面向连接，通过资源预约机制分配有限的资源给每个应用。有界延迟网络的关键部件包括确定性通信量模型、包调度和准入控制，确定性通信量模型描述每个连接的最大通信量特性，包调度决定各连接包在输出链路上的发送顺序，准入控制依赖于包调度原理，根据连接的通信量特性，验证网络能否满足已有连接和请求建立连接的网络性能。

在分组交换网络中，不同连接的包在每个交换机内相互作用，如果不进行适当控制，这些相互作用可严重影响网络的服务性能。包调度决定不同连接的作用方式，是网络提供有界延迟服务的关键。目前通常采用 FCFS(先来先服务)调度原理，包在输出链路上的发送顺序完全由它们的到达顺序决定，所有连接具有相同的延迟边界，它虽然实现简单，但灵活性差且网络资源利用率低。本文简要说明包调度的功能，介绍基于速率和基于延迟的各种调度原理，分析它们的性能和实现复杂性，给出了设计包调度的基本原则和目前的发展动向。

2　分组交换网络中包调度的服务模型

在分组交换网络中，假设链路和交换机为任意拓扑结构，包在链路上的传输延迟是有界的，交换机非阻塞，即包到达输入链路时，直接路由到相应的输出链路。包在交换机内没有交换冲突发生，到不同输出链路的包互不干扰，在交换机输出端口排队输出。

有界延迟服务网络采用的服务模型为：在通信前，用户需要向网络说明它的通信量特性和性能要求，网络根据资源分配情况判定是否接收连接，若有足够的资源满足连接的网络性能则接收该用户连接。网络和用户之间存在一种协约关系，网络答应履行它的义务即保证用户的有界延迟服务，前提是用户信守它的约定，发送数据不超过预先说明的通信量。协约在通信前建立，在连接存在期间保持有效。为增加动态性和灵活性，可在连接存在期间修改协约。

在交换机中，包到达流在输入端口分离，根据路由交换到各输出链路端口，存储在输出链路的发送队列中，调度原理决定包的发送顺序。图1显示包通过一个网络交换机的路径，包从左边流向右边，连接A和B的包在同一个输出链路上，包调度器选择首先发送的包。

网络环境下的调度有很多特点，调度分布并独立运行在不同的结点上，它们依次地调度同一个连接的同一个包；调度的最小单位是包，正被调度的包通常不能被剥夺；调度的连接通信量多种多样，有周期性信息，更多的是突发性通信量；调度的应用也有不同的性能要求，如尽最大努力(best-effort)、高速、

*)本课题得到九五国防预研基金资助。孙利民　博士生，主要研究方向为计算机网络、实时通信。窦文华　教授，博士生导师，主要研究领域为高速计算机网络、实时系统。周兴铭　中科院院士，主要研究领域为高性能机体系结构、高速计算机网络、分布式数据库。

可靠、高速并可靠等。另外，包调度必须高速实现，否则会成为网络的瓶颈，这要求包调度应简单。本文主要讨论有界延迟服务的包调度。

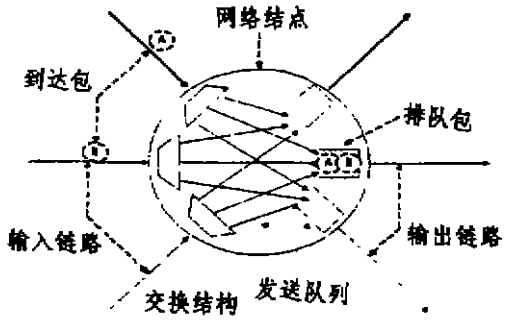

图1 通过网络交换机的包路径

包调度可分类为连续型调度和非连续型调度。如果等待发送队列中存在包，包调度器就不能空闲，称为连续型调度。非连续型调度是即使等待发送队列中存在包，如果包不是合法的可调度包，调度器就可能空闲。考虑到延迟是有界延迟服务的核心参数，本文将包调度分为基于延迟的调度和基于速率的调度。基于延迟的调度使用延迟上界，如静态优先级调度给每个连接赋予延迟上界，根据延迟约束优先级排序到达的包，保证连接的延迟上界。基于速率的调度提供连接最小吞吐率，给每个连接分配整个带宽的一部份，根据连接的通信量特性计算延迟上界。

3 基于速率的调度原理

所有基于速率的调度模拟以下两个系统：(1)时分多路复用系统(TDM)：把时间分为固定大小的帧，每个帧又分成时间槽，分配时间槽给每个连接；(2)通用处理机共享系统(GPS)：每个连接赋予共享因子，提供给连接的服务正比于它的共享因子。

停走(Stop-and-Go)调度、层次轮询 HRR(Hierarchical-Round-Robin)调度和虚时钟 VC(Virtual Clock)调度模拟 TDM 系统，连接间相互隔离，每个连接分配固定的带宽。Stop-and-Go 和 HRR 调度采用分帧机制，在一帧期间到达的包排队可等待在下一帧时间内发送，不同连接的包在一帧内的发送顺序为任意的。分帧策略的缺点类似于 TDM，若连接没有用完分配的带宽，剩余的带宽会浪费掉。VC 调度采用统计复用技术，每个包赋予虚拟完成时间，虚拟完成时间是包若在 TDM 系统中的完成时间，按虚拟完成时间增加的顺序发送包。加权公平排队 WFQ(Weighted-Fair-Queuing)调度类似于 VC 调度，但它模拟的是 GPS 系统。下面分别说明 VC 类调度和 WFQ 类调度的原理。

3.1 VC 类调度

1)VC 调度原理。假设连接 i 分配的发送速率为 r_i，第 k 个包长度为 $L_{k,i}$，到达时间为 $t_{k,i}$，它的虚拟完成时间为 $F_{k,i}$，那么：

$$F_{1,i} = t_{1,i} + L_{1,i}/r_i \quad (1)$$

$$F_{k,i} = \max\{t_{k,i}, F_{k-1,i}\} + L_{k,i}/r_i, \quad k > 1 \quad (2)$$

包的虚拟完成时间 $F_{k,i}$ 等同于连接在速率为 r_i 的专用参考服务器上的完成时间。到达的包根据它的虚拟完成时间插入排序队列中。排序复杂性为 O(Log N)，N 为队列中的包个数。VC 调度没有分帧调度的带宽浪费，但存在惩罚利用剩余带宽连接的问题。VC 调度的两个扩展是及时离开(Leave-in-Time)调度和突发调度。

2)Leave-in-Time 调度。其原理类似于 VC 调度，不同之处是在到达包插入发送队列前，速率控制机制保存并延迟到达的包。延迟包的目的是为了包到达流的突发性不至于增加到下个结点，通过控制通信量破坏，保证端端的延迟抖动上界。

3)突发调度。很多应用级数据单元太大不能封装在单个包内，必须分段才能在网络内传输。数据单元的整组包的性能比单个包的性能更重要，如压缩的动态视频图像在 ATM 网络上传输。突发调度将信息分为连续的 Bursts，每个 Burst 是一组封装一个应用级数据单元的包，赋予同一组 Burst 的所有包相同的虚拟完成时间，提高了调度实现的效率。突发调度又扩展为组优先级调度。

3.2 WFQ 类调度

1)WFQ 调度原理。WFQ 调度又称为逐包通用处理机共享 PGPS 调度(Packet-by--packet Generalized Processor Sharing)，它模拟 GPS 系统。假设 GPS 系统有 N 个连接，输出链路的发送速率为 1，连接 i 赋予服务共享因子 φ_i，服务器在时间 t 提供给连接 i 的服务速率为 $g_i = \varphi_i/\Sigma_{j\in B(t)}\varphi_i$，其中 $B(t) \subseteq N$ 表示在时间 t 有等待包发送的连接集合。因此，GPS 系统给有包发送的连接 i 分配的带宽正比于它们的服务共享因子 φ_i。在最坏情况下 $B(t) = N$，每个连接收到最小的保证速率为 $\varphi_i/\Sigma_{j\in N}\varphi_i$。然而，GPS 实际上不可能实现，因为它同时服务所有非空连接，实际的包作为整体发送而非逐位的复用。

WFQ 用 VC 模拟 TDM 相同的方法模拟 GPS 系统：每个包都赋予虚拟完成时间，虚拟完成时间对应于它们在 GPS 系统中的完成时间，按虚拟完成时间递增的顺序发送包。如果分配连接 i 的服务共享因子 φ_i，它的第 k 个包在时间 $t_{k,i}$ 到达，其发送时间为 s，它的虚拟完成时间为：

$$F_{0,i} = 0 \quad (3)$$

$$F_{k,i} = \max\{t_{k,i}, F_{k-1,i}\} + [s\Sigma_{j\in B(t)}\varphi_i]/\varphi_i, \quad k > 0 \quad (4)$$

比较公式(4)和(2)，WFQ 调度的虚拟完成时间依赖于有等待包的连接，因此需要跟踪连接的活动

性，实现上比 VC 调度复杂，但它提供良好的公平性。对于 WFQ 调度，若包在时间 t 离开 GPS，它离开 WFQ 不迟于 $t + s^{max}$，s^{max} 是系统中最大包的发送时间。因此，WFQ 调度的包不会落后于相应 GPS 系统大于单个包的发送时间。假设有两个连接，所有包的大小为 1，链路的发送速率为 1，连接的保证速率均为0.5。从 t = 0 开始，有 1000 个包从连接 1 按 1 个包/秒的速度达到，连接 2 在(0,900)期间无包达到，从 t = 900起，有 450 个包从连接 2 以 1 包/秒速度达到。那么，在(0,900)期间，连接 1 的前 900 个包发送出去，等待队列内无包。如果使用 VC 调度，t = 900 时，连接 1 的虚拟完成时间为 1800，连接 2 的虚拟完成时间为 900。当连接 2 的包达到时，它们的虚拟完成时间为 900,902,…,1798，连接 1 达到包的虚拟完成时间为 1800,1802,…,1998，连接 2 达到的所有包都将在连接 1 的包之前服务。如果使用 WFQ 调度，两个连接预留的带宽相同，t = 900 之后，交互调度连接 1 和连接 2 的包。造成 VC 调度惩罚利用剩余带宽连接的原因在于：VC 虚拟完成时间的定义参考静态 TDM 系统，连接的虚拟完成时间独立于其它连接的活动，完全依赖本连接的达到历史和预留速度，一旦利用剩余带宽发送过多包，即使不影响其它连接的性能，也要受到惩罚。WFQ 参考了动态 GPS 系统，虚拟完成时间的计算依赖整个系统连接的活动性。

2) 最坏情况公平加权公平排队 WF^2Q 调度 (Worst-case Fair Weighted Fair Queuing)。WFQ 仅利用了 GPS 系统的虚拟完成时间，WF^2Q 同时利用了 GPS 系统的虚拟开始时间和虚拟完成时间，达到更加精确地模拟 GPS 系统。在 WF^2Q 中，当在时间 τ 选择下一个要发送的包时，并不是从所有在发送队列等待的包中选取(WFQ 是这样)，而是从发送队列虚拟开始时间大于时间 τ 的包中选取，虚拟开始时间是包在相应 GPS 系统开始接收服务的时间。

对于 WF^2Q 调度，若包在时间 t 离开 GPS 系统，它离开 WF^2Q 调度器的时间不迟于 $t + s^{max}$，也不早于 $t-g_i$，s^{max}，s^{max} 是系统中最大包的发送时间。因此，WF^2Q 比 WFQ 具有更好的公平性，实现的复杂性也增加了。

3) 自计时公平排队 SCFQ 调度 (Self-Clocked Fair Queuing)。与 WF^2Q 相反，SCFQ 调度是为减少实现复杂性而提出的。WFQ 和 WF^2Q 因计算精确的虚拟时间而需维持参考的 GPS 系统，花费的代价很大。SCFQ 调度基于这样的观察：任何时间的系统虚拟时间约为正在接收服务包的虚拟服务时间。估计的虚拟时间定义 $\hat{V}(t)$ 为 $F^p, S^p < t < f^p, s^p$ 和 f^p 表示包在 SCFQ 系统开始服务和完成服务的时间。假设连接 i 分配服务共享因子 φ_i，连接 i 的第 k 个包在时间 $t_{k,i}$ 到达，发送时间为 s，SCFQ 的虚拟完成时间为：

$$F_{0,i} = \hat{V}(t_{1,i}) \tag{5}$$

$$F_{k,i} = \max\{\hat{V}(t_{k,i}), F_{k-1,i}\} + [s\Sigma_{j\in N}\varphi_j]/\varphi_i \quad k > 0 \tag{6}$$

SCFQ 的虚拟时间计算比较简单，精确度比 WFQ 差。考虑图 2 的例子。假设所有包的大小为 1，链路的发送速率为 1，连接 1 的保证速率为 0.5，其余 10 个连接的保证速率为 0.05。在 GPS 中，连接 1 的第 k 个包的完成时间为 2k，k = 1 ～ 10，第 11 个包的完成时间为 21，连接 2 ～ 11 的第 1 个包的完成时间为 20，WFQ 的服务顺序如图第 4 行所示。如果使用 SCFQ，t = 0，连接 1 的第 1 个包因有最小的虚拟完成时间而首先服务；t = 1，连接 2 ～ 11 的包虚拟完成时间均为 20，按 tie-breaking 规则，调度连接 2 的包，由于 SCFQ 将正在服务包的虚拟完成时间作为目前的虚拟完成时间，因此，$\hat{V}(1) = \hat{V}(2) = 20$；t = 2，连接 1 的第 2 个包的虚拟完成时间为：$F_{2,1} = \max(2, 20) + 2 = 22$，它的虚拟完成时间最大而排在最后。图例说明，虽然连接 1 按它的说明发送包，它的包却显著延迟。

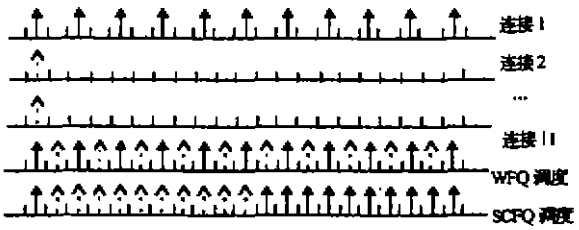

图 2 SCFQ 调度和 WFQ 调度

4 基于延迟的调度原理

基于延迟的调度给每个连接的包赋予延迟上界，延迟边界测试基于所有连接的通信量和调度原理的特性，计算连接的最大延迟。延迟边界测试是连接准入控制的主要内容，如果计算出连接的最大延迟超过连接要求的延迟上界，网络就不接收该连接。基于速率的调度没有延迟边界测试。下面说明几种基于延迟的调度原理，并给出它们的延迟边界测试的充分必要条件。

4.1 最小期限优先 EDF (Earliest- Deadline-First) 调度

EDF 调度给每个到达的包赋予调度期限 (deadline)，即到达时间 t 与延迟上界 d_i 之和，选择最小期限的包发送，正在发送的包不能被中断。EDF 需要维持根据包期限递增顺序排列的发送包队列，当新的包到达时，首先计算包的期限，然后查找其在队列中的正确位置并插入队列。EDF 调度是优化的调度：如果任何包调度原理可调度一组延迟约束的连接，EDF 调度原理也可调度该组连接。

若连接 i 的通信量为 A^*，延迟边界为 d_i，EDF 调度的延迟边界测试的充分必要条件参见表 1。EDF 调

度有两个扩展：Delay-EDD 调度和 Jitter-EDD 调度。

1)Delay-EDD(Delay-Earliest-Due-Date) 调度。其连接采用(X_{min},X_{ave},I,s^{max})通信量模型，X_{min} 是相邻两个包到达时间间隔的最小值，X_{ave} 是在任意 I 时间长度期间最大的平均包到达间隔，s^{max} 是最大包的发送时间。调度器与通信量源之间有服务约定，若每个源的通信量满足它说明的通信量模型，调度器提供有界延迟服务。Delay-EDD 调度的关键问题是到达包的期限赋值，假设连接 i 的第 k 个包在时间 $t_{k,i}$ 到达，延迟边界为 $d_{k,i}$，第 k 个包的期限为：

$$D_{k,i} = \max\{t_{k,i} + d_{k,i}, D_{k-1,i} + X_{min,i}\} \quad (7)$$

2)Jitter-EDD(Jitter-Earliest-Due-Date) 调度。它在 Delay-EDD 调度的基础上增加了延迟抖动控制，提供连接确定的延迟抖动上界(连接的任意两个包延迟之差的最大值)。当包离开调度器时，在它的包头记录着其期限与实际离开调度器时间之差(PreAhead)，下一个调度器的调整器延迟 PreAhead 时间后，该包才成为合法的可调度包(eligible)。图 3 说明包通过相邻两个调度器的过程。由于在两个调度器之间包的可调度时间差是固定的，等于包在前一个调度器的延迟上界与两个调度器之间链路延迟之和，若目的主机不控制延迟抖动，则端端延迟抖动上界为最后一个调度器的局部延迟。

4.2 静态优先级 SP(Static-Priority) 调度

ED 调度因查找排序队列而使操作复杂性高，为降低复杂度提出了 SP 调度。它将连接集 C 分成 P 个子集 $\{C_p\}_{1\le p\le P}$，子集 C_p 中的所有连接具有相同的延迟上界 d_p，$d_p < d_q$，$p < q$。SP 调度维持 P 个具有优先级的 FIFO 队列，FIFO1，FIFO2，…，FIFO P，FIFO 1 的优先级最高，FIFO P 的优先级最低，优先级高的延迟上界小。子集 C_p 中所有连接的到达包排到 FIFO p 队列中，SP 调度选择优先级最高的非空队列的包发送。SP 调度只有少量的固定的操作，性能介于 FCFS 调度和 EDF 调度之间。SP 调度的延迟边界测试的充分必要条件参见表 1。

4.3 轮询优先级队列RPQ (Rotating-Priority-Queues) 调度

RPQ 调度象 SP 调度将连接集分成 P 个连接子集 C_1，C_2，…，C_p，同一子集中所有连接具有相同的延迟上界。有一个称为轮询时间的系统参数 $\Delta > 0$，所有子集的延迟上界均为 $d_p = n_p\Delta$，$n_p < n_q$，$p < q$，且 $n_1 > 0$。RPQ 调度维持 $(n_p + 1)$ 个有序 FIFO 队列，每个 FIFO 队列对应一个索引整数 n，$0 \le n \le n_p$，索引为 n 的 FIFO 称为 n-队列。子集 C_p 中所有连接包到达时，都插入当前的 n_p-队列中。由于所有 $n_p > 0$，没有包插入当前 0-队列。0-队列的包具有最高优先级。RPQ 调度从索引数最小的非空队列中选择包发送。在 Δ 周期结束时，RPQ 调度重新标识 FIFO 队列，对 $n \ge 1$ 队列，当前的 n-队列改为 (n − 1)-队列，当前的 0-队列变为 n_p-队列，FIFO 队列完成了一次轮询。

RPQ 调度吸取了 SP 调度和 EDF 调度两者的优点，实现上采用多个 FIFO 队列，比 SP 增加了少量的轮询操作，性能上接近 EDF 调度。它的延迟边界测试的充分必要条件参见表 1。

表 1　基于延迟调度的延迟边界测试

调度原理	延迟边界测试充分必要条件
EDF	对所有的 $t \ge d_1$，有 $t \ge \Sigma_{i\in N}A_i^*(t - d_i) + \max_{k,d_k>t} S_k^{max}$
SP	对所有的 $p, t \ge 0$，$\exists\ \tau \le d_p - S_p^{max}$ 使得：$t + \tau \ge \sum_{i\in C_p} A_i^*(t) + \sum_{q=1}^{p-1}\sum_{j\in C_q} A_j^*(t+\tau) - S_p^{min} + \max_{r>p} S_r$
RPQ	对所有的 $t \ge d_1$，有 $t \ge \Sigma_{i\in C1}A_i^*(t - d_1) + \sum_{q=2}^{pp}\sum_{j\in C_q} A_q^*(t + \Delta - d_q) + \max_{r,d_r>i+\Delta} S_r^{max}$

结束语　为提供有界延迟服务，提出了大量的包调度原理，不少调度原理没有充分利用网络资源。好的包调度原理应满足下列标准：

(1) 高效：为保证确定的性能，因网络资源有限而需要通过准入控制机制限制连接的个数和通信量。高效的调度原理充分利用网络资源，支持更多的有界延迟的连接。

(2) 灵活：很多应用有着显著不同的通信量特性和不同的服务要求。灵活的调度原理要能满足各种应用的不同通信量的不同性能要求。

(3) 防火特性：防火特性指调度原理保护行为良好的用户(按通信量说明发送信息的用户)获得应有的服务，不会受到网络负载波动、不良行为用户及“尽最大努力”的通信量的影响。

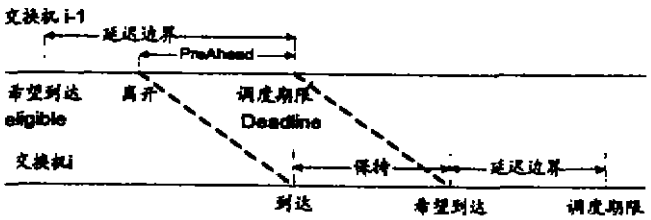

图 3　Jitter-EDD 中的包调度

计算机科学1998 Vol. 25 №. 4

Internet 上分布式虚拟现实系统的研究

Researches on Distributed Virtual Reality Systems over Internet

蒋遂平 周明天

（电子科技大学计算机科学与工程学院 成都610054）

摘 要 Due to the limitations of bandwidth and delay of Internet and the performance of Personal Computer connected to Internet, It is difficult to develop distributed, multiuser virtual reality systems over Internet. This paper discussed some problems in the development and the architecture of distributed virtual reality systems.

关键词 Virtual reality, Virtual environments, Distributed virtual environments, Distributed interactive simulation, Internet

虚拟现实方面的研究工作最先集中于单用户系统，单用户系统强调系统的沉浸特征，追求多种感知的逼真性，随着 Internet 技术的发展，特别是 WWW 应用所取得的成功，以及军用分布式交互仿真网络 SimNet 的成功，人们开始将单机或局域网上的单用户虚拟现实系统扩展到 Internet 上的多用户分布式虚拟现实系统。在这方面的一个重要进展是虚拟现实建模语言（Virtual Reality Modeling Language，VRML）的出现，VRML 促进了以页面为中心的第一代 Web 正在向以交互、动态和逼真的三维世界为特征的第二代 Web 的发展。第一代 Web 主要用于信息发布，第二代 Web 则主要用于通信，两代 Web 都可以采用图1所示的模型来描述，尽管各自的组成部分内部很不相同。

第二代 Web 浏览器比第一代 Web 浏览器复杂得多，其组成如图2所示。其中，跟踪包括用户头、手、眼、体等位置和方向的检测，而绘制也是多感知的，包括视觉、听觉、力觉、触觉等的反馈。对象模拟包括各虚拟对象行为模拟，以及一些高层交互行为，如碰撞检测和地形跟随等。第二代 Web 中其它成分将在后面讨论。

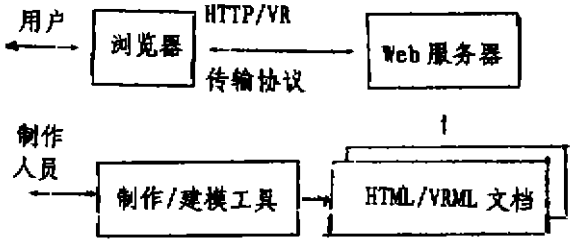

图1 Web 模型

Internet 上的分布虚拟现实系统与以前强调沉浸的单用户系统不同，它更强调系统的交互性和用户间的通信，能让一组分布于不同地点的用户同时在相同的虚拟环境中相遇并交互，以支持协同工作。它的基本思想是：构造一个由用户"居住"的空间，用户可以用适合于各种社交情况的自然和非正式的方式相互通信。

Internet 上分布式虚拟现实系统的主要特征有：(1)地理上分散的用户可在同一时刻，进入同一虚拟环境，进行实时交互，用户数目可达数千人；(2)用户

---

(4)可分析性强：准入控制判定在不影响已有连接网络性能的前提下，能否满足新连接的网络性能，这需要包调度条件。调度条件验证包的最大延迟是否超过连接的延迟上界。如果调度条件不精确，准入控制将没有必要地限制网络可支持的连接数。

(5)低复杂度：包调度必须在发送一个包时间内完成，否则链路会空闲，调度器成为网络的瓶颈，这就要求调度的复杂度低。

一个包调度不可能优化上述所有性能，特别是调度的高效与低复杂度相互矛盾，每个调度根据特定的条件，在满足上述要求时采用了折衷方案。目前，有界延迟服务调度的发展动向为：

·在增加少量复杂度的前提下，改进已有调度原理，如调度 WF^2Q^+，RPQ^+；

·根据应用的性能要求和通信量特点，提高调度的效率和性能，如突发和组调度；

·对 internet 异构交换机不同调度环境下，研究通用化的调度算法。

（参考文献共20篇略）