一种基于优先级的综合接入系统自适应 I/O 调度方案 [1]

韩国栋 张兴明 邓 勇 邬江兴

(信息工程大学国家数字交换系统工程技术研究中心 郑州 450002)

摘 要: 该文介绍了综合业务接入系统的基本结构，提出了一种新的基于优先级的参数自适应轮询 I/O 调度 (Priority based parameter self adapting round-robin I/O scheduling) 方案，给出了该方案在综合接入系统 I/O 调度中的实现方法，并对几种调度方案的性能进行了实验比较。

关键词: 综合接入系统，基于优先级，参数自适应，I/O 调度

中图分类号: TN919.3 **文献标识码**: A **文章编号**: 1009-5896(2004)01-0131-06

A Kind of Priority-Based Self-Adapting I/O Scheduling Scheme for Integrated-Service Access System

Han Guo-dong Zhang Xing-ming Deng Yong Wu Jiang-xing

(*Nat. Digital Switching System Eng. & Tech. Research Center, Zhengzhou* 450002, *China*)

Abstract On the basis of the introduction of the integrated-service access system, this paper introduces a new kind of scheduling scheme called "priority-based parameter self-adapting round-robin I/O scheduling", and the implementation of this scheme in integrated-service access system is presented. The performances of several scheduling schemes are tested and compared in this paper.

Key words Integrated-service access system, Priority-based, Parameter-self-adapting, I/O scheduling

1 引言

综合接入技术研究的目的在于研究能适应从业务综合走向网络融合发展方向的多业务混合接入系统，同时探索具有实用意义的基于分组化环境和业务分类条件下的接入段带宽分配方案和策略，使其能够共享且高效利用整个接入网带宽 [1−3]，并保证实时业务的 QoS (Quality of Service) 需求。"综合接入"的含义是指接入系统支持多形态的用户业务接入 (如 PSTN, ISDN/BRA/PRA, FR, DDN, ATM, CABLE TV, LAN, IP VPDN, IP PHONE, IP FAX 等)，并按照公平、竞争、合理、高效的原则共享接入段带宽。

一种实用化的综合业务接入系统逻辑结构如图 1 所示。该系统最小配置为：1 局端 +1 远端；也可扩展为：1 局端 $+n$ 远端 ($n \leq 16$)。组网形态支持环形拓扑结构，也支持树形拓扑结构。

图中远端和局端 I/O 调度单元是综合接入系统基于自治域内部标签分组 (Inner Label Packet, ILP) 的输入输出核心处理单元，主要功能是完成系统内部传输带宽的动态分配、ILP 分组的 I/O 调度和系统网管 (NMS) 的通道管理；传输网络是指系统内部光传输通道，速率支持 155Mbps, 622Mbps, 2.5Gbps 等。

[1] 2002-06-20 收到，2002-12-05 改回

国家"863"重大科研项目"实用化综合接入系统研究与开发"(863-317-01-01-01-99) 资助课题

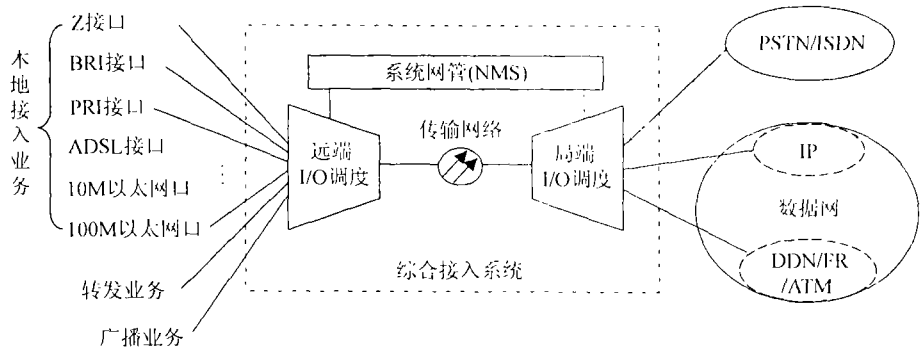

图 1 综合接入系统结构示意图

综合接入系统远端 (接入端) 上行业务包括本地接入业务、转发业务 (环形组网结构)、广播业务等。所有业务数据可分为两类，一是语音或传真类 (包括自治系统网管)，此类需保证 QoS；另一类为数据业务，此类无特别要求。系统的 I/O 调度首先要保证语音业务有足够的传输带宽以满足 QoS，此外须保证最大带宽利用度和高传输效率。

2 带宽的动态分配

基于 ILP 的综合接入系统的 I/O 调度模块是系统内部数据传输和 I/O 调度，以及系统内全带宽利用度动态分配的核心单元。I/O 调度的数据流向主要有 3 个端口 4 个方向，如图 2 所示 (以远端模块为例)。其中 LACS-up/dn(Local Access Services-up/down) 为本地业务 (上行/下行) 缓冲队列，BCAS(Broadcast Services) 为广播业务缓冲队列，TRAS(Transmit Services) 为广播业务缓冲队列，MUXR-up/dn 为混合输入输出缓冲队列。

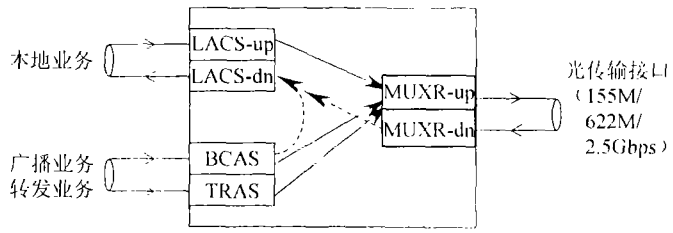

图 2 ILP I/O 数据调度流向示意图

每个端口的 ILP 数据均按业务类型和不同的优先级形成 $n(n \leq 6)$ 个缓存队列阵，同一队列按 FIFO(First In First Out) 原则占用传输带宽。队列阵带宽分配结构如图 3 所示。

按照基于优先级的时间片轮询式调度策略 (Priority-based time-slice round-robin scheduling strategy) 本地业务接入模块从用户接口板读入 ILP 包，并根据 ILP 的包类别指示将 Type-0 和 Type-1 类包分别存入 Queue1 和 Queue2，Type-2 类包则按包长的不同分别存入 Queue3, Queue4, Queue5, Queue6 中，形成本地上行业务调度缓存区 LACS-up；对于转发业务和广播业务，同样依据 ILP 包类别、包长指示，将分组包分别写入不同的队列，形成转发业务缓存区 TRAS 和广播业务缓存区 BCAS；输出调度器 (Output scheduler) 按照基于优先级的参数自

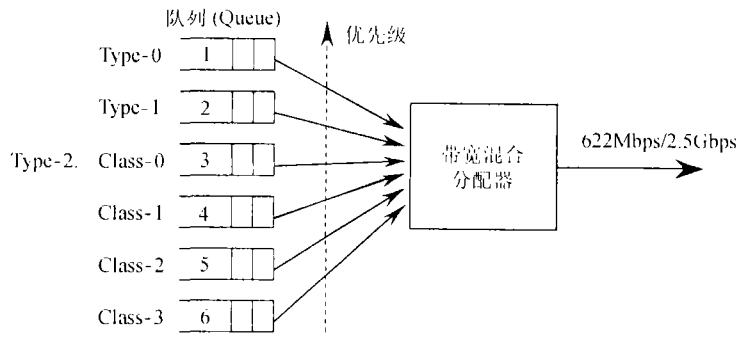

图 3　多队列带宽分配结构

适应轮询调度算法 (Priority-based parameter auto-adaptive round-robin scheduling) 从 LACS-up, TRAS, BCAS 中形成基于优先级的多队列输出缓冲存储区 MUXR-up, 并按加权轮询算法 (WRR[4]) 从 MUXR-up 向输出端口调度输出 ILP 包。

当 ILP I/O 业务总量超过系统容限，丢包事态严重时，系统将启用丢包控制和数据用户接口流量控制——反馈式流量控制机制 (Feedback traffic control)。控制量的产生由网管系统考虑了全局总流量、局部流量、以及各缓冲 Buffer 的充盈程度等因素后，以网管指令的形式动态下达至数据端口，限制相应的用户端口业务发生速率，从而实现全局的带宽动态管理。丢包和端口流量抑制原则是 [5,6]：

(1) 保证 Type-0 业务，Type-1 业务有足够的传送带宽 (系统配置时此两类业务不超出系统承载能力)；

(2) 保证数据接口 (Type-2 业务) 的预约或租用带宽，控制新增长预约业务；

(3) 其余 Type-2 业务按 Best-of-Effort 原则允许接入。

对于下行业务，首先按 ILP 类型和数据分组的长度信息形成 MUXR-dn 缓冲队列，再分别按目的节点号和广播状态指示判断落地业务、转发业务或广播业务，分别形成 LACS-dn, TRAS, BCAS 多队列缓冲区，然后对落地业务按板号、端口号分发，转发和广播业务同上行处理。此外，下行业务的分发、分类等处理的轮询和调度算法同上行处理。

此外，对于 2 类 ILP 的短包，在实时流量和丢包统计的基础上，在保证 Type-0, Type-1 无丢包的情况下，可以通过软件动态增加带宽，以充分提高带宽利用率，同时保证话音业务的 QoS。

3 参数自适应 I/O 调度算法

最基本的调度算法是时间片轮询调度 (Time-slice-based round-robin scheduling)，它的实现逻辑和 FPGA 制作简单。但当轮询到长数据包 (Class-2, 3) 时，会造成话音数据传输延迟，影响话音质量。其次是采用基于优先级的时间片轮询调度，这种调度算法尽管能充分保证语音业务不会丢包并且传输延迟小，但是 IP 数据业务得到的带宽小，特别是当语音业务的流量不是很大，主要是数据业务时，系统会在话音队列的轮询上浪费时间，造成带宽的不必要浪费。

因此笔者对上述算法进行了修正，提出了一种新的基于优先级的参数自适应轮询调度算法。该算法的基本原理是：

(1) 将缓冲队列的充盈程度分为“忙”和“闲”两种状态 (对于数据业务，当缓冲队列的剩余容量不够容纳两个 1500byte 长 IP 分组的状态定义为“忙”状态，否则为“闲”状态；对于 Type-0 和 Type-1 类业务缓冲队列非空即为“忙”状态，否则为“闲”状态)；

(2) 当被轮询的缓冲队列处于“忙”状态时， Type-0 和 Type-1 队列立即读出 ILP 分组，直至队列状态变为“闲”；其它数据缓冲队列则结合优先级从该队列中读出 $S=[N/p]$ 个 ILP 分组 (N 是系统在各个处理点上根据系统总流量、实时流量和丢包统计由软件计算并设置的包数量， p 是优先级系数， $[\bullet]$ 表示取整)；

(3) 当被轮询的缓冲队列处于“闲”状态时， Type-0 和 Type-1 队列通过软件设置轮询等待次数 k，再进入数据队列轮询调度。数据队列每轮询一次， k 值减 1，直至 $k=0$ 时再巡检 Type-0 和 Type-1 队列状态 (考虑到 Type-0 和 Type-1 空闲时必为话音接入间隙，这种等待是合理的)；

(4) 对于数据缓冲队列 Type-2，按优先级加权系数 p 为 Class-0, Class-1, Class-2, Class-3 分别设定最终输出轮询等待次数 $C(i)$，每次该队列被轮询到时， $C(i)$ 值减 1，直至轮询等待次数为零时才输出 S 个该队列分组。同时中断 k 值的等待，进入下一轮循环。

由于高优先级数据分组轮询等待次数比低优先级分组少，故能保证高优先级数据分组时延减少。新的基于优先级的参数自适应轮询调度算法流程如图 4 所示。

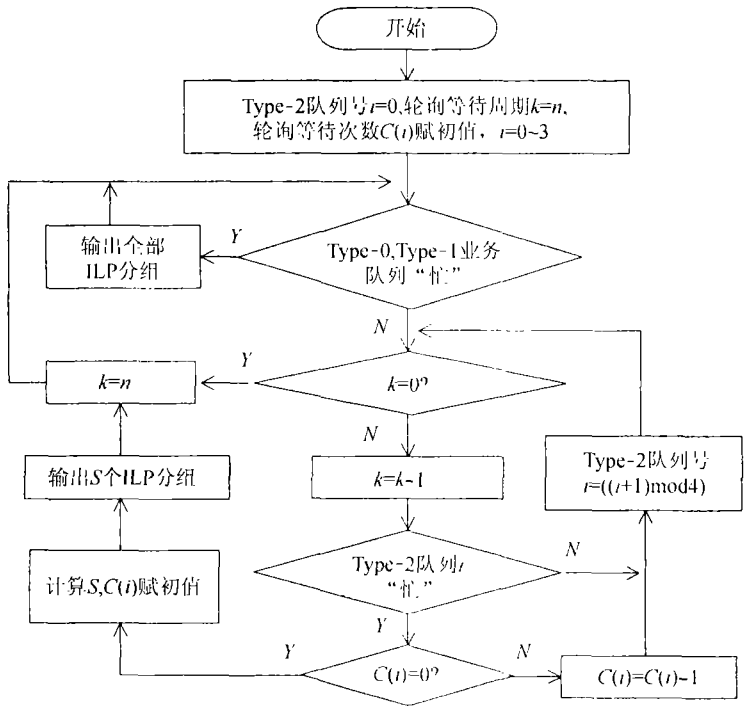

图 4 一种新的基于优先级的参数自适应轮询调度算法流程图

4 测试结果分析

在理论分析的基础上 [7]，作者在研发过程中对不同的 I/O 调度算法进行了电路分组业务传送时延、数据分组业务丢包统计等参数的对比实验和测试，结果如表 1 、表 2 所示 (数据采用 IP 包方式， 30%×1Gbps 表示按总流量 (1Gbps) 的 30% 速率发送数据，测试设备： Smartbits2000 等)。

从表 1 的结果可见，采用基于优先级的参数自适应轮询调度算法的话音业务时延特性明显优于采用时间片轮询调度算法，时延与包长成正比例增长。从表 2 的结果可见，采用基于优先

级的参数自适应轮询调度算法的丢包性能明显优于采用基于优先级的时间片轮询调度算法，丢包率随着流量的增加而增大，随着包长的增大而减小 (同一流量下)。这是因为流量增大对系统的处理能力要求加大，超负荷时必然丢包；包越长包间隔越大，给系统的包处理时间相对延长，不至于使内部缓冲器拥塞，从而丢包率降低。

表 1　电路业务传送时延测试结果对照表

时间片轮询调度算法①			基于优先级的参数自适应轮询调度算法②		
IP 包长 (byte)	流量 (Gbps)	时延 (μs)	IP 包长 (byte)	流量 (Gbps)	时延 (μs)
64	30%	14.65	64	30%	9.45
65	30%	15.10	65	30%	9.90
128	30%	39.85	128	30%	24.50
512	30%	263.55	512	30%	183.65
1024		未测	1024		未测

表 2　数据丢包统计测试结果对照表 (测试时长：40s)

基于优先级的时间片轮询调度算法①			基于优先级的参数自适应轮询调度算法②		
IP 包长 (byte)	流量 (Gbps)	丢包率 (%)	IP 包长 (byte)	流量 (Gbps)	丢包率 (%)
64	30%	0.000	64	30%	0.000
64	40%	15.421	64	40%	14.496
128	30%	0.000	128	30%	0.000
128	40%	1.474	128	40%	0.671
256	30%	0.000	256	30%	0.000
256	40%	0.912	256	40%	0.127
512	30%	0.000	512	30%	0.000
512	40%	0.374	512	40%	0.087
1024	30%	0.000	1024	30%	0.000
1024	40%	0.098	1024	40%	0.024
1280	30%	0.000	1280	30%	0.000
1280	40%	0.002	1280	40%	0.000
1518	40%	0.000	1518	40%	0.000

注：①国家数字交换技术系统工程研究中心 (NDSC) 自测；② 863-317 主题专家测试组测试。

5 结束语

从前面的讨论可以得出，基于 ILP 的综合接入系统的 I/O 调度和动态带宽分配的实现是从 4 个方面来得到保障的：丢包统计和控制机制、包类别分拣、业务优先级别和 I/O 调度算法。调度的目的是实现对各端口缓冲队列的 I/O 管理，资源公平共享，减少包丢失概率，保证带宽的最大利用度。本文提出的“基于优先级的参数自适应轮询调度算法”在 1999 – 2000 年度国家“863 计划通信技术主题的重大课题“实用化综合接入系统研究与开发”(863-317-01-01-01-99) 项目中得到了应用和验证。结果表明，在多业务、宽频带 (2.5Gbps 吞吐率)、全动态的要求和前提下，该调度算法能有效地保证系统时延、丢包率、抖动等 QoS 指标。

此外，基于系统内部标签包 (ILP) 的接入和传输方式，使得接入网的馈线段、配线段、以及自治系统内部 (局端至远端) 的用户载荷数据无须进行其它任何形式的信元转换，直接以 Payload 的形式进入骨干网，这对于加快作为网络通信“最后一公里”的接入网技术向宽带 IP 化转移的进程无疑具有十分重要的意义。

参　考　文　献

[1] McKeown N, Izzard M, Mekkittikul A. The tiny tera: a packet switch core [J]. *IEEE Micro Magazine*, 1997, 17(1): 26–33.

[2] Lakshman T V, Stiliadis D. High speed policy-based packet forwarding using efficient multi-dimensional range matching[C]. Proc. ACM SIGCOMM'98, Vancouver, Canada, Sept. 1998: 203–214.

[3] Duffield N G, Lakshman T V, Stiliadis D. On adaptive bandwidth sharing with rate guarantees[C]. Proc IEEE INFOCOM, San Francisco, CA, April 1998: 1122–1130.

[4] Katevenis M, Sidiropoulos S, Courcoubetis C. Weighted round-robin cell multiplexing in a general-purpose ATM switch chip[J]. *IEEE J. on Selected Areas in Communications*, 1991, 9(3): 1265-1279.

[5] Kar K. Scheduling of variable size packets in input queued switches. [Masters' Thesis], University of Maryland at College Park, December 1999.

[6] Kim J H. Bandwidth and latency guarantees in low-cost high performance networks[D]. [Ph. D. Thesis], Department of Computer Sciences, University of Illinois, Urbana-Champaign, January 1997.

[7] Casetti G, Kurose J, Towsley D. A new algorithm for measurement-based admission control in integrated services packet networks[R]. Proc. of the Fifth International Workshop on Protocols for High-Speed Networks, France, October 1996: 13–28.

韩国栋： 男，1964 年生，副教授，博士，从事数据通信及接入技术研究，承担国家 863 信息网技术主题“实用化综合接入系统”项目.

张兴明： 男，1963 年生，副教授，硕士，从事数据通信及接入技术研究，承担国家 863 信息领域跨主题重大课题“核心路由器”.

邓 勇： 女，1961 年生，讲师，从事数据通信研究.

邬江兴： 男，1953 年生，教授、博士生导师、国家数字交换系统工程技术研究中心主任、国家 863 计划通信技术主题专家组副组长兼信息网专业专家组组长、中国高速信息示范网总体组组长.