

基于优先级输入排队的调度算法的研究

赵卫星¹ 李秋² 王秀欣²

1.保定供电公司 071051

2.华北电力大学电子与通信工程系 071003

摘要

基于输入排队的调度算法在延时和吞吐率方面有较大优势,特别是极大匹配算法,有复杂度低易于实现等优点。在排队策略中,加入优先级的调度算法可以照顾到优先级较高的信元,可以给予优先调度。本文首先介绍了几种极大匹配调度算法,在此基础上理论分析了 iSLIP^[1]算法的队列延时和加入优先级以后的队列延时。最后通过仿真比较了 iSLIP 算法和 pri_iSLIP(priority iSLIP)算法的信元延时,加入优先级的算法延时较低。

关键词

极大匹配; iSLIP 算法; 优先级

中图分类号: TP393

文献标识码: A

1 引言

路由器和交换机作为网络交换的核心,其结构和性能应满足日前增加的信息数据交换需求。其结构的发展经历了几个阶段,集中式总线结构,分布式共享总线结构,分布式交叉开关(Crossbar)结构以及多机互连的集群结构。高速交换机和路由器一般采用交叉开关(crossbar)以保证交换内核无阻塞,并采用集中式调度器调度定长信元通过交叉开关。基于crossbar的交换机的调度结构有三种,基于输出排队结构,基于输入排队结构和联合输入输出排队结构。基于输出排队的调度算法由于结构优势,可获得高的吞吐量和提供时延保障,但这类算法的主要缺陷是未能充分利用存储器带宽,在高速环境中的应用受很大限制。联合输入输出排队结构是在交换结构的输入输出端口分别进行缓冲排队,但此种结构提出的调度算法复杂度过高,只具有理论意义。

输入排队交换结构存在队头(HOL, Head Of Line)阻塞问题,当端口数目比较多,且在所有输入端口均匀分布的Berboully业务,输入排队算法的吞吐率只有58.6%。在采用虚拟输出排队(VOQ,

Virtual Output Queuing 即在输入缓存中为每个输出端口设一个单独的FIFO缓存队列,可获得高吞吐率。基于输入排队结构的调度算法虽然对QoS的支持不如基于输出排队的调度算法,然而输入排队结构对商用存储器的随机访问速率要求较低,其读写速率仅需等同于线路速率。基于输入排队结构的调度算法主要有最大匹配算法,极大匹配算法,稳定结合算法等。由于复杂度,运行时间以及吞吐率等方面因素,极大匹配占有很大优势。

应用较多的算法是极大匹配算法,SLIP类型的算法,如Cisco公司推出的高端路由器GSR12000系列和斯坦福大学的TinyTera原型机中的交叉开关调度算法分别采用了iSLIP (iterative SLIP)和ESLIP (Extended SLIP)算法^[5],ESLIP是为支持多播传送而在iSLIP算法基础上进行的扩展算法。

2 iSLIP 算法介绍

在实际应用中,一般采用用启发式算法来简化二分图的匹配问题。这些算法具有实现简单、运算速度快等优点。

在讨论调度算法之前,先给出几个假设:

(1)交换网络交换一个信元的时间称为一个时隙。信元的产生、缓存、交换和输出都将在一个时隙内同时进行;

(2)各输入端的信元到达过程相互独立;

(3)输入/输出队列容量足够大。

iSLIP 算法在RRM(round robin matching)^[2]算法的基础上进行了改进,以减少同步现象。三个步骤具体过程如下:

1)请求(request):每个输入向每一个有队列信元的输出发送一个申请。

2)许可(grant):如果一个输出接收很多申请,它选择紧挨着从最高优先级元素开始固定的循环调度的那一个。输出通

知每个输入它的申请是否被允许。只有在这个允许在第一次迭代的第三步被接受的情况下,指向循环调度的最高优先级元素的指针比允许的输入增加1个单元(取模N)。

3)接受(accept):如果一个输入接收到一个允许,它接受紧挨着从最高优先级元素开始固定的循环调度的那一个。指向循环调度的最高优先级元素的指针比接受的输出增加1个单元(取模N)。

与PIM(Parallel Iterative Matching)^[4]算法相比,iSLIP算法的主要优点是实现简单且具有高吞吐率。在均匀的独立Bernoulli到达下,即使是单次迭代,iSLIP算法也能达到100%的吞吐率^[3],明显高于PIM算法。下面我们对iSLIP算法的性能进行分析。

3 iSLIP 算法的延时分析

首先我们定义如下关于延迟的参数^[6]:

信元到达时间:信元到达输入端口的时间;

信元离开时间:信元到达输出端口的时间;

信元延迟:信元离开时间减去信元到达时间;

信元服务时间:交换网络用于传送一个信元的时间;

信元等待时间:信元延迟减去信元服务时间,也就是信元在输入队列中的等待时间。

3.1 无优先级算法延时分析

对于固定的负载,队列延时为一个不变的值。在低或高负载时表现的不同性能,原因在于仲裁器同步的程度。在低负载时,到达的信元在随机位置找到仲裁器,iSLIP算法表现出来的性能类似于PIM一次迭代的情况。信元立即发送的概率正比与无其他信元等待发送到同一个输出端口。忽略低负载时队列的延时,每个

输出端的信元大约为 $\lambda(1 - (N-1/N)^{N-1})$, 随着 N 的增加到 $\lambda(1 - (1/e)^{N-1})^7$. 因此, 对于小的常量 λ 队列的延时是随着常量 N 增加的, 在高负载时, 算法每 N 个时隙调度一次 FIFO 队列, FIFO 队列将类似一个具有到达概率 λ/N 服务时间为 N 时隙的 M/D/1 队列。因为一个具有随机服务时间为 S , 到达率为 λ , 服务率为 μ 的 M/G/1 队列的队列延时为:

$$d = \frac{\lambda E(S^2)}{2(1 - \frac{\lambda}{\mu})} \quad (1)$$

所以对于 iSLIP 算法在高负载的 Bernoulli 业务源时, 延时接近于

$$d = \frac{\lambda N}{2(1 - \lambda)}, \text{ 它正比于 } N。$$

3.2 具有优先级算法延时分析

在一般的排队问题里, 先到先占是一种最自然的排队规则, 但是在实际问题里有时我们也会利用其他的规律以符合自己的需要, 对于较为紧急的事件, 延时费用较高的事件, 或者仅仅是比较容易做的事件, 人们往往愿意给予较高的优先权^[7]。在每个输入端口都有 N 个 VOQ 队列, 加入优先级以后, 每个 VOQ 队列又分成了 m 个子队列, 在输入端就有 mN 个 VOQ 队列。每 m 个子队列分别代表 m 个不同优先级。输入排队优先级策略图如图 1 所示。

(1) 信元到达率: λ_i 和 λ 分别表示不同优先级的信元和所有信元的到达率, 它们有如下关系:

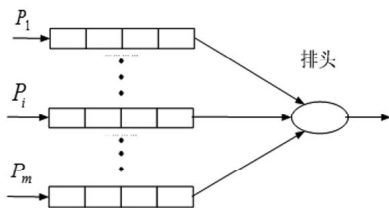


图1 输入排队优先级队列

$$\lambda = \sum \lambda_i \quad (2)$$

(2) 信元服务率: μ_i 和 μ 分别表示不同优先级的信元和所有信元的服务率, 则:

$$\frac{1}{\mu} = \sum \frac{\lambda_i}{\lambda} \times \frac{1}{\mu_i} \quad (3)$$

(3) 负载: 信元到达率与信元服务率之比, ρ_i 和 ρ 分别表示不同优先级的信元和所有信元的负载, 则:

$$\rho = \sum \rho_i \quad (4)$$

(4) 平均包服务时间: $E(S_i)$ 和 $E(S)$ 分别表示不同优先级的信元和所有信元的平均包服务时间。

(5) 平均包等待时间: $E(W_i)$ 和 $E(W)$ 分别表示不同优先级的信元和所有信元的平均包等待时间。

我们这里采用抢占优先占用调度方式, 为了简化起见, 这里的优先级定为两级, 即信元的优先级分为高优先级和低优先级, 当高优先级信元到达输入端口时, 我们可以得到下列关系。

$$E[W_1] = E[V_1] = E[M_1]E[S_1] + \frac{\lambda_1 E[S_1^2]}{2} \quad (5)$$

由于 $E[M_1] = \lambda_1 D_1$, 因此高优先级信元的平均等待时间为:

$$E[W_1] = \frac{\lambda_1 E[S_1^2]}{2(1 - \rho_1)} \quad (6)$$

为了计算 $E[W_2]$, 在低优先级时总的业务如同一个 M/G/1 队列中以 λ_2 为到达率, S_2 为服务时间的 FCFS 状态下的业务一样。由 M/G/1 队列的结果, 我们已知这项值的平均数为

$$E[V_2] = \frac{a_2 E[Y_2^2]}{2(1 - \theta_2)} \quad (7)$$

以 V_2 作为第一次服务时间, S_1 为以后逐次服务时间, 那么当到达率为 λ_1 时, 平均需要

$$E[W_2] = \frac{E[V_2]}{(1 - \rho_1)} \quad (8)$$

单位时间服务系统上所有工作 V_2 一级由 V_2 引起的高优先级服务才会全部完成, 这时就是这个低优先级到达者首次占用服务所等待的平均延误时间。

$$E[W_2] = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} \quad (9)$$

4 性能仿真与性能分析比较

仿真针对 8×8 的 crossbar, 使用与实际情况相符的 Bernoulli 独立同分布业务源模拟均匀业务, 用 Bursty 产生连续发往同一端口的业务源模拟突发业务。分别对 iSLIP 算法和具有优先级的 pri_iSLIP 算法进行仿真。我们这里选取的迭代次数 i 为 4, 优先级的级数为 2。

4.1 Bernoulli 独立同分布业务源

首先针对所有输入端口采用 Bernoulli-iid-uniform 业务源, 随机选择目的端口, 且概率相等。对 iSLIP 和 pri_iSLIP 算法的信元平均延时进行比较, 如图 2 所示。

在这种业务源下两种算法的性能相当, 随着负载增高, 信元平均延时增加。在负载大于 80% 时, pri_iSLIP 算法比 iSLIP 算法延时稍低一些。所以在 Bernoulli 独立同分布业务条件下 lp_iSLIP 算法表现出的性能略优于 iSLIP 算法。

4.2 突发(bursty)业务源

在所有输入端口均采用突发业务源条件下, 这里采用的业务源是 Bursty, 即可产生连续发往同一端口的信元。突发业务时信元平均延时比较如图 3 所示。

从图 3 中可以看出, 在负载低时, 两种算法的信元平均延时都较小, 随着负载的增加, pri_iSLIP 算法比 iSLIP 算法延时

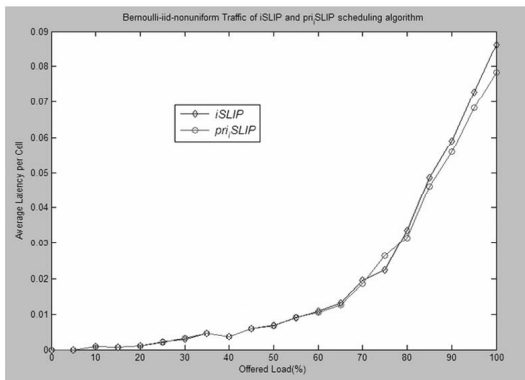


图2 Bernoulli独立同分布业务源信元平均延时比较

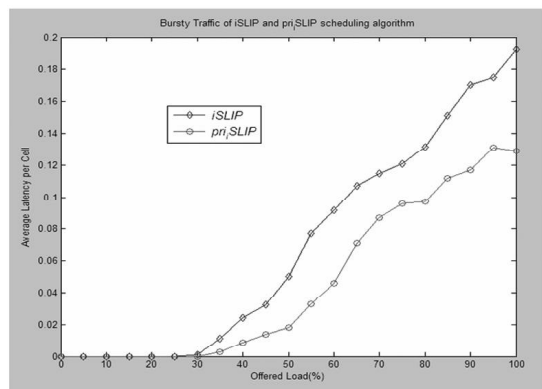


图3 突发(bursty)业务源信元平均延时比较

下转第 38 页 ▶▶

能, 极其广泛的应用范围和极具潜力的应用价值, 吸引了全世界科学家的关注。有关碳纳米管的研究也愈加深入, 成为当前重要的国际前沿课题之一, 我国科学家在这一领域已经取得了重要进展。碳纳米管具有非常光明的前景, 在未来几年内, 随着碳纳米管应用技术及产品的开发, 碳纳米管将会对众多领域产生重大而深刻的影响, 并给人类带来巨大的利益。

参考文献

- [1] Lijima S. Helical microtubes of graphitic carbon[J]. Nature.1991, 354: 56-58.
- [2] 王敏, 赵红, 张相育, 等. 电弧法制备单臂碳纳米管的机理及进展[J]. 化学工程师. 2006, 135 (12): 38-40.
- [3] Guo T. Catalytic growth of single-walled nanotubes by laser vaporization[J]. Chem Phys Lett. 1995, 243: 49-51.
- [4] Ivanov V, Nagy J B, Lambin P, et al. The study of carbon nanotubes produced by catalytic method[J]. Chem Phys Lett. 1994, 223: 329-331.
- [5] Nelly M R, Myung S K, Terry K, et al. Carbon nanofibers: a unique catalyst support medium[J]. J Phys Chem. 1994, 89: 13108-13110.
- [6] Hata K, Lijima S. Water-assisted highly efficient synthesis of impurity-free single-walled carbon nanotubes[J]. Science, 2004, 306: 1362-1364.
- [7] 朱宏伟, 慈立杰, 梁吉等. 浮游催化法半连续制取碳纳米管的研究[J]. 新型炭材料. 2000, 15 (1): 48-51
- [8] Cheng H M, Li F, Su G, et al. Large-scale and low-cost synthesis of single-walled carbon nanotubes by the catalytic pyrolysis of hydrocarbons[J]. Appl Phys Lett, 1998, 72: 3282-3285.
- [9] 侯鹏翔, 白塑, 成会明等. 气体流动状态对纳米碳纤维制备的影响[J]. 新型炭材料. 2000, 15 (4): 17-20.
- [10] Ebbesen T W, Lezec H J, Hiura H, et al. Electrical conductivity of individual carbon nanotubes[J]. Nature, 1996, 382: 54-55.

作者简介

张艳荣, 本科生。从事无机非金属材料的研究。

上接第 33 页

显然 f 在 $x = \pm 1$ 点不可导, 但 $x=0$ 时, $f'(0)=0$, 由此知这样变化条件后定理的应用范围缩小了。

(5) 满足罗尔定理条件的中间值 ζ 可以有无穷多吗? 可能。

例如, 函数 $f(x) = \begin{cases} x^2 \sin^2 \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases}$ 在 $[-1, 1]$ 上满足罗尔定理的条件, 当 $x \neq 0$ 时,

$$f'(x) = 2x \sin^2 \frac{1}{x} - 2 \sin \frac{1}{x} \cos \frac{1}{x} = 2x \sin^2 \frac{1}{x} - \sin \frac{2}{x},$$

当 $\xi = \frac{1}{2n\pi}$ 时 ($n \in \mathbb{Z}$), $f'(\xi) = 0$, 说明定理中的中间值 ζ 可以有无穷多。

(6)^[3] 罗尔定理的条件可以减弱。函数 $f(x)$ 在闭区间 $[\alpha, b]$ 的两端点 α, b 处即使不连续, 只要在两端点处的单侧极限存在且相等, 定理结论仍然成立; 定理中的区间也可以是无穷区间, 只要满足一定的条件, 也能保证定理结论的成立。

(7) 微分中值定理中只指出 ζ 的存在性, 没有指出它的具体位置, 某种意义上说微分中值定理就是适合特定等式的某区间内的“中值点”的存在定理。从文献^[4]中得到如下定理:

定理 3 设函数 $f(t), g(t)$ 在 $[\alpha, x]$ 上连续, 在 (α, x) 内可导, $f_{+}^{(n+1)}(a), g_{+}^{(n+1)}(a)$ 存在, 且 $g'(t) \neq 0$, 记 $F(t) = f'(t)/g'(t)$, $F^*(a) = F''(a) = \dots = F^{(n-1)}(a) = 0, F_{+}^{(n)}(a) \neq 0$, 则对于柯西中值公式 (2) 中的 ζ 有

$$\lim_{x \rightarrow a^+} \frac{\xi - a}{x - a} = \frac{1}{\sqrt[n]{n+1}}.$$

参考文献

- [1] 华东师范大学数学系. 数学分析[M]. 北京: 高等教育出版社. 2001.
- [2] 盛祥耀. 高等数学[M]. 北京: 高等教育出版社. 1987
- [3] 齐春玲, 李晓培. 关于罗尔(Rolle)中值定理条件的研究[J]. 河南科技大学学报(自然科学版). 2007(5): 96~97.
- [4] 陈新一, 唐文玲. 关于 Cauchy 中值定理“中值点”的一个注记[J]. 甘肃教育学院学报(自然科学版). 2003(7): 4~6.

作者简介

常敏慧 (1981 -), 女, 山西原平人, 本科, 山西运城学院助教, 从事高等数学教学工作。

上接第 35 页

明显减小, 表现出良好的性能。

5 结论

具有优先级策略的调度算法性能有一定的提高, 通过仿真也可以看出。优先策略可用于多种排队模型, 比如输入输出排队, 两级调度结构。通过理论分析, 它可以减少队列的延时, 特别是对于优先级高的队列, 通过仿真也得到了相似的结果。随着高速路由器和交换机的发展, 传统的调度算法方案在速度及可靠性等方面表现不能满足人们的需要, 成为系统性能的瓶颈。我们要通过各种方法开阔思路来提高调度算法的性能。本文总结了一些前人的经验, 从对信元加入优先级的角度对极大匹配算法进行了研究, 较未加优先级的算法性能有一定提高, 对实际使用的高速路由器和交换机调度算法的改进有一定的参考价值。

参考文献

- [1] McKeown N. The iSLIP scheduling algorithm for input-queued switches[J]. IEEE/ACM Trans Networking, 1999, 7(4): 188-201.
- [2] Richard, Dimitrio. Two dimensional round-robin schedulers for packet switches with multiple input queues[J]. IEEE/ACM Trans. Networking, 1994, 2(5): 471-482.
- [3] McKeown N, Anantharan V, Walrand J. Achieving 100% throughput in an input-queued switch[A]. IEEE Infocom '96[C]. San Francisco: IEEE, 1996. 296-302.
- [4] Anderson T, Owicki S, Saxes J, Thacker C. High speed switch scheduling for local area networks. ACM Transactions in Computer Systems, Dec. 1993, 319-352.
- [5] 王斌. 高性能路由器的交换技术研究[D]. 北京邮电大学, 20050420: 5-16.
- [6] 李文杰, 刘斌. 输入排队中抢占式的短包优先调度算法[J]. 电子学报, 2005, 4(33): 577-583.

[7] 华兴. 排队论与随机服务系统[M]. 上海翻译出版公司, 1987: 180-185.

作者简介

赵卫星 (1958 -), 男, 河北保定人, 工程师, 从事电力系统通信工程设计; 李秋 (1983 -), 女, 山东兖州人, 硕士研究生, 研究方向为信息系统及信息安全; 王秀欣 (1983 -), 硕士研究生, 研究方向为信息系统及信息安全。