多优先级可调速率队列中延迟与速率优化控制研究

杨天明[1,2] 刘景宁[1]

(华中科技大学计算机科学与技术学院 武汉 430074)[1] (黄淮学院 驻马店 463000)[2]

摘 要 在队列网络中,延迟和速率优化控制是一个复杂的问题。针对多优先级、可调服务速率的M/G/1队列,在约束条件为每种优先级业务的平均延迟的情况下,研究了队列的两种凸优化问题,即最小化平均延迟向量的凸函数和最小化平均业务代价的凸函数,并分别提出了一种优化算法。算法使用虚拟队列技术,对这两种具有动态 $c\mu$ 规则变量的优化问题进行了求解。然后算法自适应选择一个严格的优先级政策,以响应在每个忙阶段中观察时刻前的各种业务级别的延迟。利亚普诺夫漂移分析和仿真结果验证了算法的优化性能,并且表明文中所提优先级政策所花费的队列统计资源有限,或者为0。

关键词 多优先级,服务速率,优先级政策,延迟

中图法分类号 TN393 **文献标识码** A

Study on the Control Optimization of Delay and Rate in Multiple-priority Rate-adjustable Queues

YANG Tian-ming[1,2] LIU Jing-ning[1]

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[1]

(Huanghuai University, Zhumadian 463000, China)[2]

Abstract It is a complicated problem in queue networks to optimally control the delay and service rate. For multi-class priority queue and adjustable service rate M/G/1 queues, two convex optimization problems were studied, i. e. , minimizing convex functions of the average delay vector, and minimizing average service cost, both under the constraints of per-class delay, and consiquently an optimization algorithm was proposed for each of them. These algorithms use virtual queue techniques to solve the two problems with variants of dynamic $c\mu$ rules. Then these algorithms adaptively choose a strictly priority policy, in response to past observed delays in all job classes, in every busy period. Lyapunov drift analysis and simulation results validate the optimal performance of these two algorithms, and show that the proposed polices require limited or no statics of the queue.

Keywords Multi-class priority, Service rate, Priority policy, Delay

多类队列系统的动态控制在计算机通信网络和制造系统中广泛应用,因此近年来受到研究者的广泛关注[1,2]。动态控制的一种常用的有力方法是描述某个性能指标的可行域,然后使用优化方法来改进最优控制策略[3,4]。当可行域为多项拟阵时,一个严格的被称为 $c\mu$ 的优先权策略最小化系统的代价,这是一个著名的结论[5]。接下来的问题是在更复杂的系统中这种简单的优先策略是否仍然有效。本文指出在M/G/1队列中的凸优化问题中,如果优化问题的延迟域是多项拟阵,则该队列的在线最优优先策略可与 $c\mu$ 规则同样简单。

本文考虑的是一个为 N 类独立的高斯到达的M/G/1队列系统。控制服务器以非抢占方式一次服务一个任务。完成一个任务后,控制服务器判断下一步服务哪种类型的任务。控制服务器也可以动态调整服务速率。记类型 $n\in\{1,\cdots,N\}$ 的任务的到达速率为 λ_n,该类任务的大小为随机变量 S_n(期望为 $\mathbb{E}[S_n]$)。所有业务的大小是独立同分布的。技术上通常假设所有种类的任务大小 S_n 的分布是任意的,然而其前四阶矩是有限的。任务到达时,控制服务器仅能知道任务的类型而不知其实际的大小。同时假设服务器的服务速率可随着时刻 t 的瞬时代价 $P(t)$ 而改变,记为 $\mu(P(t))$,并假设 $\mu(\cdot)$ 是递增的,$\mu(0)=0$。

记 $\overline{W}_n$ 为类 n 队列的平均队列延迟向量,由于在非抢占队列系统中,平均系统延迟是平均队列延迟与平均服务时间的和,因此本文仅考虑平均队列延迟的优化,而其优化结果可直接推广到平均系统延迟的优化中。考虑下列两种问题:

1)在延迟约束为 $\overline{W}_n\leqslant d_n$ 的情况下,最小化每种任务的平均延迟向量 $(\overline{W}_n)_{n=1}^N$。这里假设服务速率为常数,即我们只考虑分配判决的优化,而不考虑服务时间的控制。

2)在服务时间的动态分配情况下,最小化延迟约束为 $\overline{W}_n\leqslant d_n$ 时的每种任务平均服务代价。

求解第一个问题可使各种任务的延迟具有一种公平性,这与众所周知的速率比例公平性[6]或利用率比例公平性[7]的思想是相同的。另外在第2.1节我们指出延迟比例公平性是平方表达式而非对数表达式。对于第二个问题,其应用之一是利用动态速率缩放[8]来最小化计算机的功率消耗,同时为

到稿日期:2013-12-12 返修日期:2014-02-27 本文受国家自然科学基金(61303046),河南省教育厅科学技术研究重点项目(14A520020)资助。

杨天明(1968—),男,博士,高级工程师,主要研究方向为数据存储、网络文件系统,E-mail:ytmi196502@126.com;刘景宁(1957—),女,教授,主要研究方向为网络存储、文件系统。

不同业务流提供延迟保证。

当服务速率为常数时，M/G/1 队列系统中的所有可达延迟向量集合为一个多项拟阵，这是一种特殊的多面体，其顶点是通过严格优先策略[9]得到的。这种优先策略根据一种平稳分布在队列忙时随机更新任务的优先级，通过这种方式求出上述第一个优化问题的解。因此我们就需要寻找队列忙时更新任务优先级的最优动态策略。在第二个优化问题中，服务速率控制也是需要考虑的，因此我们需要寻找队列忙时同时更新优先级和服务速率的最优动态策略。

本文的动态优先策略的建立是利用虚拟队列技术完成的，虚拟队列用于监视每种任务中违反延迟约束(存储到虚拟队列备份)的过去观察延迟量。然后在每个判决时刻(忙时的末尾)，为延迟违例越多的任务类型赋予更高的优先级，直至下一个判决时刻，并如此重复。通过利亚普洛夫漂移分析，从技术层面上说本文表明了使虚拟队列优先稳定的策略是解决 M/G/1 队列系统中优化问题的在线最优策略。该策略在每个忙时都会做出最大权重的判决，由于该判决在分配优先级之前对加权后的虚拟队列备份进行排序，因此可将其看作一种 $c\mu$ 策略。本文表明基于该判决的动态 $c\mu$ 规则的性能与最优性能之间的差距为 $O(1/V)$，其中 $V>0$ 为控制参数，可根据最优性选取一个足够大的数值，但较大的值会影响算法的收敛性能，因此需要在最优性和收敛性能之间折中[10,11]。

在有关排队论的文献中，大量的研究集中于对严格优先策略(如 $c\mu$ 规则)最小化线性代价的研究，研究带约束的凸优化问题的文献相对很少[12]。有些文献使用统计近似对单服务器队列的自适应优先策略进行改进[13]，本文的自适应控制方法相比统计近似方法较为简单，并且方法本身就引入了时间平均的约束。

也有文献对单服务器队列中服务速率的状态相关分配问题进行了研究[14]。目前的研究通常使用动态规划(DP)方法来表征最优策略的单调特性。由于本文假设任务是多种类型的，系统的状态空间也是多维的，因此仅使用动态规划方法是不够的。另外，在动态规划方法中引入时间平均约束是比较复杂的，且需要队列的完整统计特性，而本文提出的方法仅需要队列的部分统计特性。本文重点讨论忙时更新服务速率的策略，这种方法可以使得系统分析和控制更加简便。

本文第 1 节说明文中用到的标注法、定义和控制策略；第 2 节对第一个凸延迟优化问题进行求解，并引入延迟比例公平性；第 3 节对第 2 个优化问题即服务速率控制问题进行求解；第 4 节给出仿真结果；最后对全文进行总结和讨论。

1 控制策略模型

考虑一个基于帧结构的 M/G/1 队列系统，其中每个帧由一个闲时和对应的忙时组成。记 $t_k, k\in\mathbb{Z}^+$ 为第 k 个帧的开始时刻。第 k 个帧占据的时间段是 $[t_k, t_{k+1}]$，时间长度为 $T_k\triangleq t_{k+1}-t_k$。定义 $t_0=0$ 并假设系统初始为空。记 $A_{n,k}$ 为第 k 个帧中类 n 任务到达的集合。对于每个任务 $i\in A_{n,k}$，记其延迟为 $W_{n,k}^{(i)}$。

在本文提出的策略下，平均延迟可能不具有显式的极限，这时使用概率论方法[15]定义类 $n\in\{1,\cdots,N\}$ 的平均队列延迟，如式(1)所示。

$$\overline{W}_n\triangleq\limsup_{K\to\infty}\frac{\mathbb{E}\left[\sum_{k=0}^{K-1}\sum_{i\in A_{n,k}}W_{n,k}^{(i)}\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1}|A_{n,k}|\right]}\tag{1}$$

其中，$|A_{n,k}|$ 为第 k 个帧中类 n 任务到达的个数。为了简便，我们仅考虑在帧边缘抽样的平均延迟。

下面两个假设描述了本文所使用的控制策略。

假设 1 调度策略是连续工作的、非抢占式的和非预期的。控制器一次为一个任务服务。每次服务完成时，控制器就判断下一步服务哪个类型的任务。每个类型中的任务的服务顺序是任意的。

假设 2 当服务速率可调时，第 k 个忙时的服务速率固定为 $\mu(P_k)$，$P_k\in[P_{\min},P_{\max}]$，控制器的判断过程也可能是随机的。闲时的服务速率为 0。假设最大代价 $P_{\max}$ 是有限的，但其值足够大以保证延迟约束的可行性。最小代价 $P_{\min}$ 也假设足够大使得即使队列的代价始终为 $P_{\min}$ 时队列也是稳定的。即我们需要 $\sum_{n=1}^{N}\lambda_n\mathbb{E}[S_n]/\mu(P_{\min})<1$ 或 $\mu(P_{\min})>\sum_{n=1}^{N}\lambda_n\mathbb{E}[S_n]$。

2 凸延迟优化

凸延迟优化的描述如式(2)所示。

$$\begin{aligned}&\text{minimize }\sum_{n=1}^{N}f_n(\overline{W}_n)\\&\text{subject to }\overline{W}_n\leqslant d_n,n=1,\cdots,N\end{aligned}\tag{2}$$

使用的调度策略如假设 1 描述。函数 f_n 对所有的 n 都是连续的、凸的、非降的和非负的。这里，假设控制器的服务速率为常数，且所有的延迟约束为可行的。

2.1 延迟比例公平性

若延迟向量 $(\overline{W}_n^*)_{n=1}^N$ 在平方惩罚函数 $f_n(\overline{W}_n)=\frac{1}{2}c_n(\overline{W}_n)^2$ 下对所有 n 都是最优的，则称该延迟为加权延迟比例公平的，这里 c_n 是给定的正常数。在这种情况下，任意可行的延迟向量 $(\overline{W}_n)_{n=1}^N$ 需要满足式(3)。

$$\sum_{n=1}^{N}f_n'(\overline{W}_n^*)(\overline{W}_n-\overline{W}_n^*)=\sum_{n=1}^{N}c_n(\overline{W}_n-\overline{W}_n^*)\overline{W}_n^*\geqslant 0\tag{3}$$

这与速率比例公平性的思想是相同的，速率比例公平性式(4)所示。

$$\sum_{n=1}^{N}c_n\frac{x_n-x_n^*}{x_n^*}\leqslant 0\tag{4}$$

其中，$(x_n)_{n=1}^N$ 为可行速率向量，$(x_n^*)_{n=1}^N$ 为最优速率向量。直观上看，由于对延迟的要求尽可能小，而对速率要求尽可能大，因此延迟比例公平性具有式(3)所示的乘积形式，而不是如式(4)所示的比例形式。为了进一步阐明这个问题，以两个用户为例，说明式(3)和式(4)具有相同的比例折中。首先令 $c_1=c_2=1$。在速率比例公平性中，考虑两个可行速率 $x_1=100, x_2=10$，用户 2 的速率是用户 1 的 1/10。然后如果用户 1 的速率增加 x 个单位，用户 2 的速率并不会因此减少 $x/10$ 个单位。在延迟比例公平性中，假设 $\overline{W}_1=10, \overline{W}_2=100$，用户 2 的延迟是用户 1 的 10 倍。然后根据式(3)，如果用户 1 的延迟减少 x 单位，由于用户 2 的延迟为用户 1 的 10 倍，用户 2 的延迟最多仅能增加 $x/10$ 个单位。

2.2 延迟公平性策略

对于每个用户 $n\in\{1,\cdots,N\}$，定义两个离散时间的虚拟延迟队列 $\{Z_{n,k}\}_{k=0}^{\infty}$ 和 $\{Y_{n,k}\}_{k=0}^{\infty}$，其中 t_{k+1} 时刻 $Z_{n,k+1}$ 和 $Y_{n,k+1}$ 根据式(5)和式(6)计算。

$$Z_{n,k+1}=\max\left[Z_{n,k}+\sum_{i\in A_{n,k}}(W_{n,k}^{(i)}-d_n),0\right]\tag{5}$$

$$Y_{n,k+1}=\max\left[Y_{n,k}+\sum_{i\in A_{n,k}}(W_{n,k}^{(i)}-r_{n,k}),0\right]\tag{6}$$

其中，$W_{n,k}^{(i)}$代表在帧$[t_k,t_{k+1})$内，第n类任务排队延迟；$r_{n,k}\in[0,d_n]$是时刻t_k选取的辅助变量，它与帧大小T_k和类型n的帧到达$A_{n,k}$是独立的。为队列$Z_{n,k}$和$Y_{n,k}$中的每个虚拟到达$W_{n,k}^{(i)}$分别匹配一个虚拟服务d_n和$r_{n,k}$。假设对所有n，初始化$Z_{n,0}=Y_{n,0}=0$。队列$Y_{n,k}$的稳定有利于最小化式(2)中的凸延迟惩罚。$Z_{n,k}$的值反映了第n类任务过去时间内超过指定延迟域d_n的延迟量。接下来将说明如果$Z_{n,k}$长时间内保持在一个较小的值上，则平均延迟约束$\overline{W}_n\leqslant d_n$将得到满足。这可由引理1来描述，该引理说明了$Z_{n,k}$队列的稳定性使得延迟约束$\overline{W}_n\leqslant d_n$得到满足。

定义1 如果队列$Z_{n,k}$满足$\lim_{K\to\infty}\mathbb{E}[Z_{n,K}]/K=0$，则称该队列是期望速率稳定的。

引理1 如果队列$Z_{n,k}$是期望速率稳定的，则$\overline{W}_n\leqslant d_n$。

下面这个策略将求解式(2)所示的凸优化问题。

延迟公平性策略(延迟公平)

1)在每个帧k中，按照比例$(Z_{n,k}+Y_{n,k})/\mathbb{E}[S_n]$的降序对优先任务进行分类，其中$\mathbb{E}[S_n]$为第$n$类任务的平均大小，任务与服务器的连接可任意断开。

2)在帧k的末尾，分别使用式(5)和式(6)，对所有n，计算$Z_{n,k+1}$和$Y_{n,k+1}$，其中$r_{n,k}$是如式(7)所示的凸规划问题的解。

$$\begin{aligned}&\text{minimize } Vf_n(r_{n,k})-Y_{n,k}\lambda_n r_{n,k}\\&\text{subject to } 0\leqslant r_{n,k}\leqslant d_n\end{aligned}\tag{7}$$

其中，$V>0$是一个预定义的控制参数。

延迟公平策略仅需要任务的到达率λ_n和平均任务大小$\mathbb{E}[S_n]$，而不需要高阶统计量。当$f_n(\cdot)$可微时，式(7)所示问题是易于求解的。

对于纯粹的以寻求达到所有延迟约束$\overline{W}_n\leqslant d_n$(这时对式(2)中的所有$n$，都有$f_n=0$)为目的的可行性问题，可以使用被称为延迟可行策略的另一不同算法来解决。该策略不适用队列$Y_{n,k}$，且不需要任何有关任务到达和大小的统计信息。

延迟可行性策略(延迟可行)

- 在每个忙时，按照$Z_{n,k}$的降序对优先任务进行分类。
- 在每个忙时的末尾，对所有n，更新$Z_{n,k}$的值。

2.3 延迟公平和延迟可行策略的性能

定理1 在假设1中定义的调度策略类下，给定任意一组可行的延迟界$\{d_1,\cdots,d_N\}$，延迟公平和延迟可行策略都可以使所有$Z_{n,k}$队列的期望速率稳定，因此对所有n都满足延迟约束$\overline{W}_n\leqslant d_n$(引理1)。另外，延迟公平策略使得凸延迟代价满足下式。

$$\limsup_{K\to\infty}\sum_{n=1}^{N}f_n\left(\frac{\mathbb{E}\left[\sum_{k=0}^{K-1}\sum_{i\in A_{n,k}}W_{n,k}^{(i)}\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1}|A_{n,k}|\right]}\right)\leqslant\frac{C\sum_{n=1}^{N}\lambda_n}{V}+\sum_{n=1}^{N}f_n(\overline{W}_n^*)$$

其中，$V>0$是预定义的控制参数，$C>0$为有限常数，向量$(\overline{W}_n^*)_{n=1}^N$是问题的最优解。当参数$V$足够大时，凸延迟代价可任意接近其最优值$\sum_{n=1}^{N}f_n(\overline{W}_n^*)$。

3 延迟约束最优速率控制

本部分引入了服务速率的动态分配机制。如假设2所述，重点关注在第k个忙时分配一个固定服务速率$\mu(P_k)$的基于帧的策略。这里，帧的大小T_k、忙时时间长B_k、各帧类型n的到达$A_{n,k}$以及队列延迟$W_{n,k}^{(i)}$都依赖于$\mu(P_k)$或代价P_k。与式(1)类似，定义平均服务速率如式(8)所示。

$$\overline{W}_n\triangleq\limsup_{K\to\infty}\frac{\mathbb{E}\left[\sum_{k=0}^{K-1}P_kB_k(P_k)\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1}T_k(P_k)\right]}\tag{8}$$

其中，$B_k(P_k)$和$T_k(P_k)$表征了B_k和T_k对P_k的依赖性。易证$B_k(P_k)$和$T_k(P_k)$随着P_k增加而递减。本部分的目的是求解如式(9)所示的延迟约束速率控制最优化问题。

$$\begin{aligned}&\text{minimize } \overline{P}\\&\text{subject to } \overline{W}_n\leqslant d_n,n=1,\cdots,N\end{aligned}\tag{9}$$

所用的控制策略如假设1和假设2所述。

3.1 动态速率控制策略

首先建立如式(5)所示的虚拟队列$Z_{n,k}$。假设初始时，对所有n，都有$Z_{n,0}=0$。

动态速率控制(动态速率)策略

1)在帧$k\in\mathbf{Z}^+$中，使用严格优先策略π_k，即按照比例$Z_{n,k}/\mathbb{E}[S_n]$的降序对优先任务进行分类，其中$\mathbb{E}[S_n]$为第n类任务的平均大小，任务与服务器的连接可任意断开。

2)帧$k\in\mathbf{Z}^+$在忙时，分配服务速率$\mu(P_k)$，其中P_k是式(10)所示问题的解。

$$\begin{aligned}&\text{minimize } \left(V\sum_{n=1}^{N}\lambda_n\mathbb{E}[S_n]\right)\frac{P_k}{\mu(P_k)}+\sum_{n=1}^{N}Z_{n,k}\lambda_n\overline{W}_n(\pi_k,P_k)\\&\text{subject to } P_k\in[P_{\min},P_{\max}]\end{aligned}\tag{10}$$

其中，$V>0$是预定义的控制参数。$\overline{W}_n(\pi_k,P_k)$是服务速率为$\mu(P_k)$并使用严格优先策略$\pi_k$情况下，在所有忙时，类$n$的平均队列延迟。如果任务类型按照优先策略$\pi_k$正确地重排序，使得

$$\frac{Z_{1,k}}{\mathbb{E}[S_1]}\geqslant\cdots\geqslant\frac{Z_{N,k}}{\mathbb{E}[S_N]}$$

则平均延迟$\overline{W}_n(\pi_k,P_k)$可写为式(11)。

$$\overline{W}_n(\pi_k,P_k)=\frac{\frac{1}{2}\sum_{i=1}^{N}\lambda_i\mathbb{E}[X_i^2]}{(1-\sum_{i=0}^{n-1}\rho_i)(1-\sum_{i=0}^{n}\rho_i)}\tag{11}$$

其中，随机变量$X_i\triangleq S_i/\mu(P_k)$表示帧$k$中类型为$n$的任务的服务时间，$\rho_i\triangleq\lambda_i\mathbb{E}[X_i^2]$，$i=1,\cdots,N$，$\rho_0\triangleq 0$。

3)在每个帧的边界，根据式(5)更新所有$Z_{n,k}$队列。

动态速率策略需要到达速率和任务大小的前两阶矩。为了移除该策略对二阶统计量的依赖，将式(10)中的目标函数除以一个未知常数$\tilde{R}\triangleq\frac{1}{2}\sum_{n=1}^{N}\lambda_n\mathbb{E}[S_n^2]$，并重新定义变量$\tilde{V}=V/\tilde{R}$。修改后的策略与动态速率策略具有相同的性能(这可由定理2说明)，且仅依赖于一阶统计特性。

3.2 动态策略的性能

定理2 动态策略对所有类型n都满足延迟约束$\overline{W}_n\leqslant d_n$，且平均代价$\overline{P}$满足式(12)。

$$\overline{P}\leqslant\frac{C\sum_{n=1}^{N}\lambda_n}{V}+P^*\tag{12}$$

其中，P^*为式(9)中的最优平均代价，$C>0$为有限常数，$V>0$为预定义的控制参数。当V足够大时，$\overline{P}$与P^*之间的差距可无限小。

4 仿真结果及分析

本部分在两类任务非抢占式M/G/1队列中仿真延迟公平策略、延迟可行策略和动态速率策略的性能。记$W(P)$为

服务速率 $\mu(P)$ 恒定时的队列延迟区域，定义 $\rho_n \triangleq \lambda_n \mathbb{E}[X_n]$，$R \triangleq \frac{1}{2}\sum_{n=1}^{2}\lambda_n \mathbb{E}[X_n^2]$，其中 $X_n = S_n/\mu(P)$ 为类型为 n 的任务的服务时间。根据文献[9]，有式(13)。

$$\boldsymbol{W}(P)=\left\{(\overline{W}_1,\overline{W}_2)\left|\begin{array}{l}\overline{W}_1\geqslant\frac{R}{1-\rho_1},\overline{W}_2\geqslant\frac{R}{1-\rho_2}\\ \rho_1\overline{W}_1+\rho_2\overline{W}_2=\frac{(\rho_1+\rho_2)R}{1-\rho_1-\rho_2}\end{array}\right.\right\} \tag{13}$$

其中，两不等式说明了当某一类任务的优先级高于另一类时，该类任务的平均延迟就达到了最小化。该式中的等式是 M/G/1 队列的守恒法则。

本部分所有仿真结果都是在 10^6 个帧场景下运行 10 次后取平均的结果。

4.1 延迟公平和延迟可行策略的仿真结果

考虑具有两类任务的 M/M/1 队列，到达率为 $(\lambda_1,\lambda_2)=(1,2)$，平均服务时间为 $(\mathbb{E}[X_1],\mathbb{E}[X_2])=(0.4,0.2)$。假设服务速率是恒定的，根据式(13)，平均队列延迟的性能域为式(14)。

$$\boldsymbol{W}=\{(\overline{W}_1,\overline{W}_2)\,|\,\overline{W}_1+\overline{W}_2=2.4,\overline{W}_1\geqslant 0.4,\overline{W}_2\geqslant 0.4\} \tag{14}$$

延迟公平策略：考虑如式(15)所示的延迟比例公平性问题。

$$\begin{array}{l}\text{minimize } f(\overline{W}_1,\overline{W}_2)=0.5(\overline{W}_1)^2+2(\overline{W}_2)^2\\ \text{subject to } (\overline{W}_1,\overline{W}_2)\in\boldsymbol{W},\overline{W}_1\leqslant 1.95,\overline{W}_2\leqslant 1\end{array} \tag{15}$$

该问题的最优解是 $(\overline{W}_1,\overline{W}_2)$。不同控制参数 V 下延迟公平策略的性能仿真结果如表 1 所列。

表 1 不同 V 下延迟公平策略的延迟比例公平性

V	$\overline{W}_1$	$\overline{W}_2$	$f(\overline{W}_1,\overline{W}_2)$
100	1.6607(0.0055)	0.7424(0.0052)	2.4814(0.0239)
1000	1.7977(0.0057)	0.5984(0.0043)	2.3321(0.0199)
2000	1.8339(0.0056)	0.5639(0.0053)	2.3176(0.0217)
5000	1.8679(0.0073)	0.5276(0.0050)	2.3014(0.0222)
最优	1.92	0.48	2.304

延迟可行策略：仿真中使用了 5 种不同的延迟约束集合 $\overline{W}_n\leqslant d_n, n\in\{1,2\}$，即 $(d_1,d_2)=(0.45,2.05),(0.85,1.65),(1.25,1.25),(1.65,0.85)$ 和 $(2.05,0.45)$，其中每个分量都比 $\boldsymbol{W}$ 中的可行点大于 0.05。延迟可行策略的仿真结果如图 1 所示。

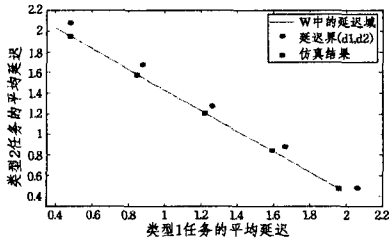

图 1 不同延迟约束集合下的延迟可行策略的结果

每种延迟约束集下都进行了 10 次仿真，每种任务类型的延迟样本标准差近似为 0.017，所有仿真都是连续进行的，产生的是连续的结果。由图 1 可知，对不同的约束集合，延迟可行策略都可以自适应输出可行的平均延迟。

4.2 动态速率策略的仿真结果

考虑具有两种类型任务的 M/G/1 队列，到达率为 $(\lambda_1,\lambda_2)=(1,2)$。类型 1 任务的大小为 0.5 的概率为 0.8，大小为 3 的概率为 0.2；类型 2 任务的大小始终为 1。考虑服务速率 $\mu(P)=\sqrt{P}$，其中 P 的取值集合为 $\{16,25\}$。

在动态速率控制策略下，完全队列延迟域记为 $\boldsymbol{W}$，它是两个独立的延迟域 $\boldsymbol{W}(16)$ 和 $\boldsymbol{W}(25)$ 的凸包，这两个延迟域由式(13)定义，其中 $\boldsymbol{W}(16)$ 和 $\boldsymbol{W}(25)$ 分别为恒定服务速率，P 分别为 16 和 25 时的队列延迟域。

使用动态速率策略求解如式(16)所示的优化问题

$$\begin{array}{l}\text{minimize } \overline{P}\\ \text{subject to } (\overline{W}_1,\overline{W}_2)\in\boldsymbol{W}\\ \overline{W}_1\leqslant 0.4,\overline{W}_2\leqslant 0.325\end{array} \tag{16}$$

其中，$\boldsymbol{W}$ 为上述完全延迟域。当式(16)约束条件中的等式成立时，延迟代价就达到了最小。$(\overline{W}_1,\overline{W}_2)=(0.4,0.325)$ 对应的平均延迟代价为 13.5，此时的平稳随机策略是最优的。随着控制参数 V 逐渐增大，动态速率策略达到最优时各类型任务的平均代价和平均延迟结果如表 2 所列。

表 2 不同控制参数下的动态速率策略的结果

V	$\overline{W}_1$	$\overline{W}_2$	$\overline{P}$
1	0.3562(0.00078)	0.3929(0.00032)	13.802(0.018)
10	0.3984(0.00022)	0.3274(0.00005)	13.510(0.026)
100	0.4003(0.00013)	0.3252(0.00010)	13.504(0.022)
最优	1.92	0.48	2.304

结束语 本文提出的自适应方法说明了多类任务类型的队列中的凸优化问题可化简为一系列线性代价最小化问题，每一个更新阶段就对应一个线性代价最小化问题。在M/G/1队列中，由于延迟域是多项拟阵，因此这些线性代价最小化问题可通过 $c\mu$ 规则求解。其他一些排队系统的性能域也具有多项拟阵的性质，因此也可以利用严格优先策略来最小化线性代价。本文的方法可用于求解某些凸优化问题并改进其中的在线动态优先策略。

参 考 文 献

[1] Shimada T, Liyama N, Kimura H, et al. Dynamic Control Method of Queuing Delay with/without OEO Conversion in a Multi-Stage Access Network [C]// World Telecommunications Congress (WTC). Miyazaki, Mar. 2012: 1-6

[2] 陈雪莲，杨智应. 桥吊可动态分配的连续泊位分配问题算法[J]. 计算机应用，2012(5)：1453-1456

[3] Le L B, Modiano E, Shroff N B. Optimal Control of Wireless Networks with Finite Buffers [J]. IEEE/ACM Transactions on Networking, 2012, 20(4): 1316-1329

[4] 严黎明，牛玉刚. 基于队列敏感性的无线接入网络用赛控制算法[J]. 计算机应用，2012(1)：123-126

[5] 朱红雷，彭元喜，尹亚明，等. 一种动态分配虚拟输出队列结构的片上路由器[J]. 计算机研究与发展，2012，530(1)：183-192

[6] Leith A, Alouini M-S, Dong K, et al. Flexible Proportional-Rate Scheduling for OFDMA System [J]. IEEE Transactions on Mobile Computing, 2013, 12(10): 1907-1919

[7] 侯华，李亘煊. 加权比例公平群智能跨层资源分配算法[J]. 计算机应用研究，2012(3)：1038-1043

[8] Prabhu B J, Tugui A E, Verloop I M. Steady-state Approximations of Dynamic Speed-scaling in Data Centers [C]// 2012 6th International Conference on Network Games, Control and Optimization (NetGCoop). Avignon, Nov. 2012: 135-138

[9] Gelenbe E, Mitrani I. Analysis and Synthesis of Computer Systems(2nd ed)[M]. Imperial College Press, 2010

[10] Hariharan S, Shroff N B. On Sample-Path Optimal Dynamic Scheduling for Sum-Queue Minimization in Forests [J]. IEEE/ACM Transactions on Networking, 2013(99):1

[11] 汪浩,黄明和,龙浩. 基于 G/G/1-FCFS、M/G/1-PS 和 M/G/∞ 排队网络的 Web 服务组合性能分析[J]. 计算机学报,2013(01):22-38

[12] 郑建国,王翔,刘荣辉. 求解约束优化问题的 ε-DE 算法[J]. 软件学报,2012(09):2374-2387

[13] 柯鹏,金姗姗,李文翔. 面向多业务通信调度的优先级排队模型研究[J]. 计算机科学,2013(3):159-162

[14] Prado S M, Louzada F, Rinaldi J G, et al. A New Distribution for Service Model with State Dependent Service Rate [C]// 2013 Second International Conference on Informatics and Applications (ICIA). Lodz, Poland, Sept., 2013:294-299

[15] Neely M J. Dynamic Optimization and Learning for Renewal Systems [C]// Asilomar Conference on Signals, Systems and Computers. Pacific Grove U. S. A., Nov. 2010:681-688

[16] 唐宏,李敏,周到. TD-LTE 集群通信系统基于优先级划分的随机接入算法[J]. 重庆邮电大学学报:自然科学版,2013,25(6):711-715

(上接第 110 页)

驶的可能方向,因此所提出的算法获得了较高的效率和成功率。

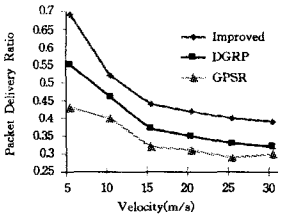

图 10 包递交率

图 11 是路由开销和车辆速度之间的关系图。由于该算法没有进行 RNG 和 GG 图的计算,使得报文递交率增加,重传率减少,所提出的算法开销明显少于 GPSR 算法。图 11 与图 10 对应,从图 11 可以看出速度低于 15m/s 的情况下,DGRP 是很准确的,但是随着速度的升高错误率也会增加。相反,所提出的改进算法是趋于稳定的。

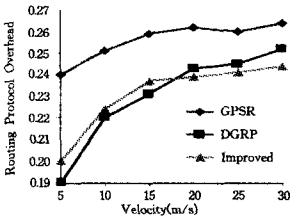

图 11 路由协议开销

图 12 是车辆行驶速度与吞吐量之间的关系图。数据包的递交率得到改善,使得吞吐量也明显增加。

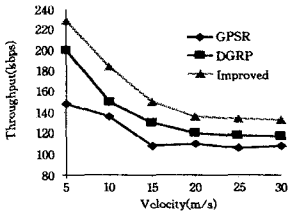

图 12 吞吐量

结束语 该策略通过信标的广播信息确定节点的类型,引入向量和一种新的判定十字路口节点的信标方式,对于直行道的普通节点、预测节点或者十字路口节点,根据不同的节点类型采取不同的转发方式:普通节点采取贪婪转发的方式,预测节点采取相应的计算判定方法来转发数据包,而十字路口节点决定路由策略的方向。由于引入了基于向量的判定方法,其转发效率要高于 GPSR 协议。本文比较了 GPSR、DGRP 和新算法的包的递交率、路由协议的开销、吞吐量,对得到的结果进行了分析。我们不仅改进了 GPSR 路由协议,而且修改了算法使其适合城市场景。模拟结果显示利用向量决定下一跳转发节点和在交叉路口采用预测模式确实有效地改善了算法的性能,增加了报文递交率。对于路由开销,新算法明显比 GPSR 和 DGRP 稳定。今后我们将测试所提出的算法在不同场景下的适应性,并开发出适合不同场景的地理位置路由协议,努力解决局部最大问题并修改恢复策略。

参 考 文 献

[1] Shafiee K, Leung V C M. Connectivity-aware minimum-delay geographic routing with vehicle trackingin VANETs[J]. Ad hoc Networks, 2011, 9(2):131-141

[2] Cha Si-ho. Comparison of greedy routing protocols for vehicular ad hoc networks[C]// 2012 International Conference on ICT Convergence (ICTC). Oct 2012:565-566

[3] 杨成恩,徐家品. 稀疏车辆 Ad hoc 网络移动模型研究[J]. 通信技术,2011,44(5):88-91

[4] 陈潜,刘云. 动态高速环境下 Ad hoc 路由协议研究[J]. 中北大学学报:自然科学版,2011,32(5):579-582

[5] 杜昱宏,秦华标. 城市非连通车载自组网中低时延路由协议[J]. 计算机工程,2010,6(15):111-113

[6] 罗涛,王昊. 车载无线通信网及其应用[J]. 中兴通信技术,2011,17(3):1-7

[7] 黄振旺,郭达. 基于地理位置的车载网络路由协议的研究[J]. 移动通信,2012,36(5):1006-1010

[8] 胡淼,李剑锋. 车载自组织网络中基于贪婪算法的地理位置路由[J]. 中兴通信技术,2011,17(3):24-28

[9] 刘杰,唐伦,龚璞,等. 一种基于连接性的 VENETS 地理机会路由协议[J]. 计算机应用研究,2013,30(4):1116-1119

[10] 郑新旺,杨光松,黄联芬,等. 车载自组织网路由协议连通性能仿真[J]. 重庆理工大学学报:自然科学版,2010,24(2):6-10

[11] Karp B, Kung T H. GPSR: Greedy perimeter stateless routing for wireless networks [C]// Proceedings of the 6th annual International Conference on Mobile Computing and Networking. Aug 2000:243-254

[12] Brahmi N, Boussedjra M, Mouzna J. Mobility Support and Improving GPSR Routing Approach in Vehicular Ad hoc Networks [M]// New Technologies, Mobility and Security. Nov 2008:1-6

[13] Kumar R, Rao S V. Directional Greedy Routing Protocol (DGRP) in Mobile Ad-Hoc Networks[C]// Information Technology, 2008(ICIT'08). Dec 2008:183-188

[14] 李元振,廖建新,李彤红,等. 地理和交通信息感知的车载 Ad hoc 路由[J]. 北京邮电大学学报,2009,32(5):56-61