效用及实时业务 QoS 联合保证的资源分配方案

青 亮 方旭明* 曾 斌
(西南交通大学信息编码与传输省重点实验室 成都 610031)

摘 要：为了有效利用和公平分配有限的网络资源，从而优化网络性能，提高社会福利，该文提出了一种效用及实时业务 QoS 联合保证的资源分配方案，将效用模型与具有不同服务质量需求的业务相结合，引入需求带宽和期望带宽，建立优化模型。方案不仅体现了资源分配效率的要求，达到系统效用的最优，同时优先保证实时业务的 QoS 要求，并达到 QoS 保证业务和尽力而为(Best Effort, BE)业务之间的公平分配。仿真结果表明，提出的方案较传统方案有效地提高了系统总效用，实现了面向服务的资源分配目标，优化了网络系统的整体性能。

关键词：效用函数；实时业务；服务质量；资源分配；公平性
中图分类号：TN92 **文献标识码**： A **文章编号**：1009-5896(2013)05-1257-05
DOI: 10.3724/SP.J.1146.2012.01174

Utility and Real Time Traffic QoS Jointly Guaranteed Resource Allocation Scheme

Qing Liang Fang Xu-ming Zeng Bin
(*Provincial Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: Resource allocation plays a significant role in current wireless communications. In order to improve the performance of the whole wireless network and social welfare, the limited network resources should be allocated efficiently and fairly. In this paper, a jointly guaranteed resource allocation scheme is proposed, which firstly introduces the required bandwidth and the desired bandwidth, and then combines the utility model with services that have different QoS (Quality of Service) requirements. An optimized model is built and the corresponding resource allocation scheme is proposed. The scheme could guarantee different QoS requirements, while the utility of the system could be improved. Furthermore, it can realize the fairness allocation between QoS and BE (Best Effort) traffics. The simulation results show that the proposed scheme can effectively improve the system utility compared with the traditional method, meanwhile, the goal of service oriented allocation is achieved, and the performance of the whole system is optimized.

Key words: Utility function; Real time traffic; Quality of Service (QoS); Resource allocation; Fairness

1 引言

近年来网络的快速发展，各种多媒体业务和高速率数据业务层出不穷，如视频业务、在线游戏等，与此同时，用户数也急剧增加。这些都直接导致了全球范围的网络资源紧张[1]，也使网络的资源分配问题面临着巨大的挑战。传统的资源分配策略更多地关注资源的高效利用，其主要目的为提高资源利用率。但互联网学术研究和协议发展的进程表明，网络传输是面向用户服务的，仅仅依靠工程技术手段来解决资源分配问题，实际达到的效用很低。因而资源分配的优化目标不能仅仅只考虑资源利用率，还应该考虑到用户满意度、分配的公平性、社会福利及不同业务的 QoS 需求等因素。

文献[2]首次提出了网络效用最大化(Network Utility Maximization, NUM)模型，该模型将经济学中的效用函数引入到资源分配中，通过求解网络效用最大化的优化目标，较好地实现了网络资源的合理分配。文献[3]从对偶问题的角度对文献[2]的模型做了进一步研究，并得到了相应的资源分配机制。文献[4]考虑到无线网络资源的不稳定性，提出一种基于效用函数的带宽自适应策略，根据网络负载的情况，调节分配到带宽的动态升降。文献[5]对具有 QoS 保证的 OFDMA 网络的资源分配进行了研究。文献[6,7]研究了效用函数的非凸优化问题。其中文献[6]针对 OFDMA 网络中不同业务类型的资源分配

2012-09-10 收到，2012-12-28 改回
科技部中芬国际合作项目(2010DFB10570)资助课题
*通信作者：方旭明 xmfang@swjtu.edu.cn

问题，提出一种连续优化技术，较好地解决了基于效用资源分配的非凸优化问题。文献[7]研究了非弹性业务效用函数的非凸优化问题，应用粒子群优化算法来解决聚合效用的最大化。文献[8]针对分配变化敏感的用户提出一种新的激励分配架构，该方案允许通过调节效用函数的参数，实现用户资源分配均值和方差的折中。文献[9]针对无线环境下的信道容量的不可预知性，提出一种自适应的拥塞控制算法，并验证了在不预测网络容量特性的情形下，将网络效用最大化。但是，这些研究仅仅只从效用最大化的角度考虑，没有考虑到实时业务的 QoS 需求问题。文献[10]虽然考虑了多业务下基于效用的资源分配，但并未对不同 QoS 需求进行区分，而这种区分采用纯粹的数学方法是难以做到的，并且也没有考虑到不同业务间资源分配的公平性问题。

本文对效用函数的性质、效用函数的确定以及资源分配的策略等问题进行了研究。将效用模型与具有不同 QoS 需求的业务结合起来，提出一种效用及实时业务 QoS 联合保证的资源分配方案。该方案不再单纯以效用最优为资源分配目标，而是定义了期望带宽与需求带宽，结合效用的变化趋势确定排序准则，以排序准则依次满足 QoS 业务的需求带宽，在此基础上将剩余带宽在 QoS 保证业务与尽力而为(BE)业务间以效用最大化的目标进行资源分配。仿真验证表明本方案不仅有效地提高了系统总效用，同时优先保证实时业务的 QoS 要求，减少 QoS 保证业务与 BE 业务间的不公平度。

2 系统描述

近年来，基于效用的资源分配算法的研究成果层出不穷，大部分算法都可归结于凸优化问题，利用诸如 KKT(Karush-Kuhn-Tucker)条件的经典理论和算法就可以得到很好的解决。然而，许多研究表明，凸效用函数只能描述弹性业务，对于具有 QoS 需求的非弹性业务则无法适用[11]。虽然一些学者对非凸函数的效用优化算法也进行了研究，但多是从对偶的角度，采用集中式或分布式的算法进行求解。事实上，采用此类算法，其复杂度较高，且分析较困难。通常将产生不可行的或次优的速率分配，其解并不一定是全局最优解[12]。

基于以上分析，本文提出一种新的解决方案，试图将复杂的非凸优化问题进行分解，主要达到以下 3 个目的：(1)系统总效用最优(社会福利最大化)；(2)考虑 QoS 保证业务的期望带宽并优先保证实时用户的 QoS 要求；(3)减少 QoS 保证业务与 BE 业务间的不公平度。

2.1 效用函数的确定

在经济学中，效用用于表征消费者对产品使用的主观满意程度。定义 $u(r)$ 为用户使用某类业务的效用函数，其物理意义为用户使用某类业务所分配到 r 的网络资源后对网络服务的满意程度。

本文将考虑 BE 业务与 QoS 保证业务的混合场景。根据文献[10]提出的效用函数确定方法，分别推导出 QoS 保证业务与 BE 业务的效用函数形式。对于 QoS 保证业务，其效用函数的形式为

$$U_1(r)=1\Big/\left[1+e^{-A_1(r-r_0)}\right] \tag{1}$$

其中 r_0 为资源最低需求。A_1 为斜度参数，为用户对某类业务的服务质量的要求程度，A_1 值越大，表示用户对服务质量要求越苛刻。

对于 BE 业务，其效用函数的形式为

$$\left.\begin{aligned} &U_2(r)=\frac{1}{1+me^{-A_2 r}}-d\\ &1/(1+m)-d=0 \end{aligned}\right\} \tag{2}$$

其中 QoS 属性为 $r_0=0$, A_2 为斜度参数，为系统资源利用率与用户公平性的表征。A_2 越大，表示侧重用户公平性；A_2 越小，表示侧重资源利用率。m, d 为满足条件 $U'(r)>0$, $U''(r)\le 0$ 的常数。

值得注意的是，效用函数的具体形式不影响本文的资源分配机制及算法性能。

2.2 期望带宽与期望效用

本文将服务等级中的服务品质以数量化进行描述。研究表明，带宽为效用函数中的必要因素[2]，因而本文中的网络资源以带宽为主。考虑 k 类业务的资源需求，定义用户对第 k 类业务的请求带宽为期望带宽(desired bandwidth)，即表示为 $r_{k,d}$。同理，定义用户对第 k 类业务的最低需求带宽为必须带宽(required bandwidth)，即表示为 $r_{k,r}$，那么带宽参数的设定区间将形成用户效用的变化区间。

设第 k 类业务根据系统决策情况分配到的带宽资源在$[r_{k,r}, r_{k,d}]$范围内波动；每一类业务于用户都有一个相应的效用函数 $U(r_k)$与之对应，代表用户对第 k 类业务所分配资源(本文为带宽)的满意程度。其中必须带宽 $r_{k,r}$可根据不同 QoS 保证业务的最低需求来确定，期望带宽可根据用户的期望效用值并结合式(1)来获得。

2.3 系统效用最大化模型

考虑一个单小区的场景，小区间干扰可以忽略。假设小区接入的业务种类数为 K，其中包含了 K_1 类 QoS 保证业务及 K-K_1 类 BE 业务，则 QoS 保证业务和 BE 业务总数分别为

$$\left.\begin{aligned} M=\sum_{k=1}^{K_1} n_k \\ N=\sum_{k=K_1+1}^{K} n_k \end{aligned}\right\} \tag{3}$$

再假设系统总资源为 R,r_k 为分配给业务 k 的资源数量。以最大化系统总效用为目标，在一定约束条件下建立优化模型如下：

$$\left.\begin{aligned} &\max\ U=\sum_{k=1}^{K_1} n_k U_1(r_k)+\sum_{k=K_1+1}^{K} n_k U_2(r_k) \\ &\text{s.t.}\ \sum_{k=1}^{K} n_k r_k \le R \\ &\quad r_k \ge 0, \qquad k=1,2,\cdots,K \\ &\quad r_k \in \left[r_{k,r}, r_{k,d}\right],\ k=1,\cdots,K_1 \\ &\quad r_k \in [0,R], \qquad k=K_1+1,\cdots,K \end{aligned}\right\} \tag{4}$$

其中 $k=1,\cdots,K_1$ 时表示 QoS 保证业务，此时 $U_1(r_k)$ 表示用户使用不同 QoS 保证业务相应的效用函数；$k=K_1+1$ 时表示 BE 业务，$U_2(r_k)$ 表示用户使用不同 BE 业务相应的效用函数。

2.4 资源分配机制

根据 2.1 节描述的效用函数形式，确定效用函数中的参数 A_1, A_2 及业务的需求带宽 $r_{k,r}$，以及期望带宽 $r_{k,d}$。此模型是一个非凸优化问题，很难通过解析方式得到最优解。本文采用如下策略解决：首先分配通过当前资源情况以及排序原则确定的部分或全部 QoS 保证业务的需求带宽，然后将剩余资源采用迭代算法进行以系统总效用最大为目标的分配。通过该算法，可以优先保障 QoS 保证业务的要求，同时在满足系统效用最优的目标下，尽量减少 QoS 保证业务和 BE 业务之间的不公平度，体现一定的公平性。具体资源分配步骤如下：

步骤 1　前期准备，估算 QoS 保证业务数 M 与 BE 业务数 N。

步骤 2　对 M 个 QoS 业务进行排序，在前文的描述中，对于业务的效用函数，A 值越大，则效用值相对资源的增益越大，即相同资源下，A 值越大的业务将获得较高的效用增益，从而使得系统总效用最大。因而在分配资源时，资源应优先分配给 A 值越大的业务。以此制定 QoS 保证业务的排序原则为：按照效用函数中 A 值从小到大的顺序排列，即：$[r_1, r_2, \cdots, r_M]$。排序原则的具体分析见下文。

步骤 3　计算 $m=\lfloor R/r_0 \rfloor$，估算当前资源下最多能支持的 QoS 保证业务数 m，根据步骤 2 的排列顺序 $[r_{M-m+1}, r_{M-m+2}, \cdots, r_M]$ 分配资源，使得 $[r_{M-m+1}, r_{M-m+2}, \cdots, r_M]=r_0$，表示让 m 个 QoS 业务均能分配到资源，此时剩余资源为 $R'=R-mr_0$。

步骤 4　资源 R' 在 m 个 QoS 保证业务与 BE 业务间进行再分配。再分配的原则为：通过迭代使得系统效用最大，求出此时的效用 U_1 及分配方案。由于当前的资源分配方案不一定为系统效用最大下的资源分配方案，因而必须探讨已经分得资源的 QoS 保证业务与 BE 业务间效用增益的情形，即让部分 QoS 保证业务的资源分配给 BE 业务，以换取效用的增益，且根据公平性原则，减少的 QoS 保证业务数须小于系统中的 BE 业务数。

步骤 5　分别令步骤 3 中的 m 减为 $m-1$, $m-2$, $\cdots$, $m-N+1$，重复步骤 4，分别得到效用值 $U_2, U_3, \cdots, U_N$ 及相应的资源分配方案。

步骤 6　比较 $U_1, U_2, U_3, \cdots, U_N$ 的值，得到最优的效用值及资源分配方案。

排序原则的分析及验证：

从前文的式(1)，式(2)中可以得到：对于 QoS 保证业务及 BE 业务，参数 A_1, A_2 取值不同，其效用函数的变化趋势也不同。当在同样的分配资源下，参数 A_1 取值越大，其效用函数值将越快接近于期望值。当在达到同样的期望效用值的条件下，参数 A_1 取值越小，需要分配的资源数越大。当资源不充足时，为了获得较大的系统效用增益，可以优先满足边际效用较大的业务，即是参数 A_1 取值较大的业务。总之，A_1 值越大的业务分配资源的优先级越高。

本文假设期望的效用值为 0.9，设置此时的效用函数的参数分别如下：

$A_1=0.1, 0.2, 0.4, 0.8, 3.2, 12.8$; $A_2=0.2, 0.8$; $r_0=10$

则不同的 A_1, A_2 值对应的期望带宽如图 1 所示。

3　数值仿真及分析

3.1 仿真参数设置

系统总资源 $R=(15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100, 120, 150)$；用户的资源请求 $r_r=r_0$=10；用户的期望带宽对应的效用均为 0.9; QoS 保证业务

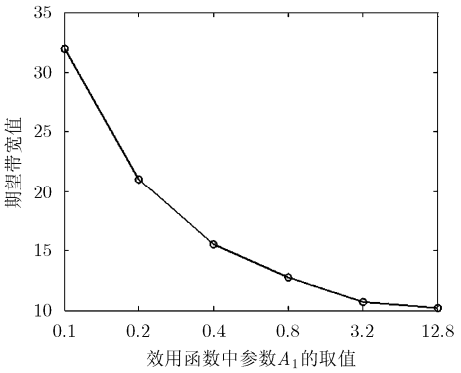

图 1 效用函数在不同参数下分配的资源情况

数 M=6, BE 业务数 N=2; QoS 保证业务效用函数形式为：$U_1(r)=\dfrac{1}{1+e^{-A_1(r-r_0)}}$；BE 尽力而为业务效用函数形式为：$U_2(r)=\dfrac{1}{1+1.5e^{-A_2 r}}-0.4$。

3.2 仿真结果及分析

本文通过仿真两种资源分配方案来评估系统性能。方案 1 为本文提出的资源分配策略。方案 2 为传统的分配方案，即优先满足 QoS 保证业务的最低需求后，剩余资源全部分配给 BE 业务。

图 2，图 3 分别对这两种方案的资源分配结果进行了仿真。其中，业务编号 1–6 为 QoS 保证业务，编号 7，编号 8 为 BE 业务。从图 2 中可以看到，本文提出的方案在系统的资源不足或者说当前 QoS 保证业务过多时，优先满足效用增益大的业务，也即是编号较大的业务，且剩余资源的再分配满足效用增益最大的原则。即在 BE 业务的效用增益大于 QoS 保证业务的效用增益时，将资源优先分配给 BE 业务，以获得系统的效用最大化。当系统的资源充足时，QoS 保证业务分配的资源满足排列原则。从图 3 中可以看出，在资源不充足时，优先满足 QoS 保证业务的最低需求后，剩余资源全部分配给 BE 业务，这样造成 QoS 保证业务只能分配到其最低需求的资源。

图 4 中，“*”线代表方案 1，“o”线代表方案 2。根据传统方案的描述，需要对资源不充足情况进行分段讨论，即图中[15,20), [20,30), [30,40), [40,50), [50,60)这 5 个区间，以及资源充足时即[60,150]这一区间。经过本文算法以及传统算法得到了每一区间所分别对应的两种方案的系统累积效用。从图中可以看出，随着总资源数的增加，方案 1 的系统累积效用上升且远高于方案 2 的系统累积效用。值得注意的是，当资源分别为 20, 30, 40, 50, 60 时，方案 2 资源刚好全部分配给对应的 2, 3, 4, 5, 6 个 QoS 保证业务，没有剩余资源，此时 BE 业务效用为零，总效用也最低，因而效用值出现如图所示的波动。

根据经济学知识[13]，有如下结论：当且仅当系统总效用在取得最大值时，系统利用率及用户公平性达到最优。从图 2，图 3，图 4 的对比分析可以看到，在资源不充足时，本文提出的分配机制能够体现一定的公平思想。如当资源分别为 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 时，通过减少少部分 QoS 保证业务的资源，让更多的 BE 业务分配到资源或分配到更多的资源，减少了这两种业务之间的不公平度，并不过度满足 QoS 保证业务需求，兼顾效用最大化情形下的 BE 业务需求，达到效率和公平性的最优，使得系统总效用提升。

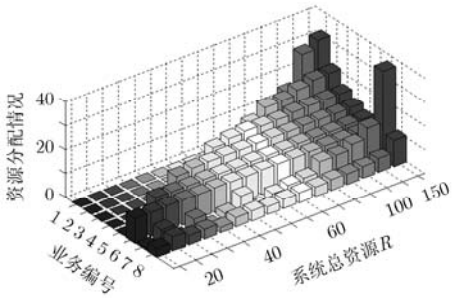

图 2 本方案下资源分配结果图

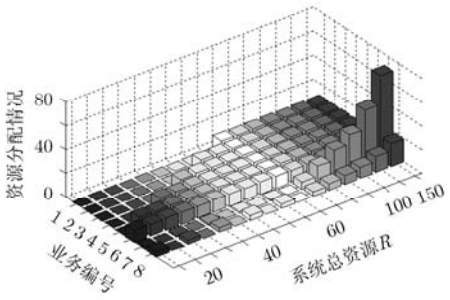

图 3 方案 2 资源分配方法结果图

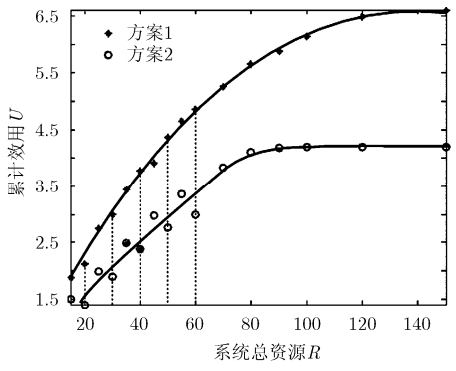

图 4 两种方案下系统总效用变化趋势对比图

4 结论

本文提出一种效用及实时业务 QoS 联合保证的资源分配方案，不再单纯以效用最优为资源分配目标，通过定义期望带宽与需求带宽，结合效用的变化，建立排序准则，在该准则下依次首先满足 QoS 保证业务的需求带宽。在此基础上，再将剩余带宽在 QoS 保证业务与 BE 业务间以效用最大化的优化目标进行资源分配。仿真结果表明，本文提出的方案不仅较传统方案有效地提高了系统总效用，同时优先保证实时业务的 QoS 要求，并减少 QoS 保证业

务与 BE 业务之间的不公平度，且其算法复杂度较低，易于实现，为优化网络系统的整体性能提供了研究思路。

参 考 文 献

[1] China Internet Network Information Center: 30th China Internet Development Report [OL]. http://news.sciencenet.cn/htmlnews/2012/7/267248.shtm. 2012.08.

[2] Kelly F P, Maulloo A K, and Tan D K H. Rate control for communication networks: shadow prices, proportional fairness and stability[J]. *The Journal of the Operational Research Society*, 1998, 49(3): 237–252.

[3] Low S H and Lapsley D E. Optimization flow control, I: basic algorithm and convergence[J]. *IEEE/ACM Transactions on Networking*, 1999, 7(6): 861–874.

[4] Lu Ning and Bigham J. On utility-fair bandwidth adaptation for multi-class traffic QoS provisioning in wireless networks[J]. *Computer Networks*, 2007, 51(10): 2554–2564.

[5] Pischella M and Belfiore J C. Resource allocation for QoS-aware OFDMA using distributed network coordination [J]. *IEEE Transactions on Vehicular Technology*, 2009, 58(4): 1766–1775.

[6] Mehrjoo M, Moazeni S, and Shen Xue-min. An interior point penalty method for utility maximization problems in OFDMA networks[C]. IEEE International Conference on Communications, Dresden, Germany, 2009: 1–5.

[7] Tang Mei-qin, Long Cheng-nian, and Guan Xin-ping. Non-convex maximization for communication systems based on particle swarm optimization[J]. *Computer Communications*, 2010, 33(7): 841–847.

[8] Joseph V and De Veciana G. Variability aware network utility maximization[OL]. http://arxiv.org/abs/1111.3728.

[9] Herzen J, Aziz A, Merz R, *et al*.. A measurement-based algorithm to maximize the utility of wireless networks[C]. Proceedings of the 3rd ACM Workshop on Wireless of the Students, Las Vegas, 2011, DOI:10.1145/2030686.2030691.

[10] Chen Li and Wang Bin. Utility-based resource allocation for mixed traffic in wireless networks[C]. IEEE INFOCOM 2011 International Workshop on Future Media Networks and IP-based TV, Shanghai, 2011: 91–96.

[11] Lee J W, Mazumdar R R, and Shroff N B. Non-convex optimization and rate control for multi-class services in the internet[J]. *IEEE/ACM Transactions on Networking*, 2006, 13(4): 827–840.

[12] Fazel M and Chiang M. Network utility maximization with nonconcave utilities using sum-of-squares method[C]. IEEE Conference on Decision and Control, Seville, Spain, 2005: 1867–1874.

[13] Varian H R. Intermediate Microeconomics: A Modern Approach[M]. New York: WW Norton & Co, 2010: 500–700.

青　亮：　男，1987 年生，硕士生，研究方向为话务量理论、用户网络行为建模研究.

方旭明：　男，1962 年生，教授，博士生导师，研究方向为无线资源管理、无线多跳中继网络.

曾　斌：　男，1987 年生，博士生，研究方向为用户网络行为建模研究、无线资源管理.