Distributed Dynamic Load Balancing in Wireless Networks

Sem Borst, Iraj Saniee, Phil Whiting

Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974

Abstract. Spatial and temporal load variations, e.g. traffic hot spots and sudden load spikes, are intrinsic features of wireless networks, and give rise to potentially huge performance repercussions. Dynamic load balancing strategies provide a natural mechanism for dealing with load fluctuations and alleviating the performance impact. In the present paper we propose a distributed shadow-price-based approach to dynamic load balancing in wireless data networks. We examine two related problem versions: (i) minimizing a convex function of the transmitter loads for given user throughput requirements; and (ii) maximizing a concave function of the user throughputs subject to constraints on the transmitter loads. As conceptual counterparts, these two formulations turn out to be amenable to a common primal-dual decomposition framework. Numerical experiments show that dynamic load balancing yields significant performance gains in terms of user throughputs and delays, even in scenarios where the long-term loads are perfectly balanced.

1 Introduction

Even more so than other communication networks, wireless data networks are characterized by the occurrence of large spatial and temporal load variations. The spatial variations manifest themselves in flash overloads in hot spots in dense urban areas due to mobility, transportation busy hours, accidents, and other unpredictable events. Temporal fluctuations occur in time scales from milliseconds and minutes to hours and days. On the latter time scales, the traffic load varies according to predictable day-of-week and hour-of-day aggregate patterns. On the former time scales, the load fluctuates not only because of the intrinsic randomness in user arrivals and session durations, but also due to variations in the transmission rates that rapidly change due to fast fading. Clearly, the spatial and temporal variation and uncertainty in the traffic will only tend to be more pronounced in ad-hoc deployment environments compared to carefully planned commercial cellular networks.

Third-generation cellular systems aim to provide high-speed data services despite these spatio-temporal variations. In fact, the fast temporal fluctuations are taken advantage of by the base station to allocate resources based on channel feedback and backlog through careful scheduling [1, 3, 7, 13, 17]. In all existing systems, however, including CDMA, UMTS and even IEEE 802.11, each base station, or access point, independently arbitrates among users in its coverage area. That is to say, users simply select the strongest received base station, and each base station allocates resources to users without any coordination with other base stations in its vicinity. As a result, one base station or access point may experience severe overload, while resources might be abundant at surrounding base stations, thus providing scope for performance gains through some form of coordination.

Coordinated resource allocation has recently been considered in several studies, see for instance [5, 6, 18]. The work in [6] shows the gains due to coordination to be significant. Despite these gains, however, coordinated resource allocation among cells remains a challenging task. Centralized coordination requires huge processing capability as well as exchange of vast amounts of information among all users in a geographical area and the coordinating entity.

A possible remedy comes from decentralized or self-organizing schemes where a sufficient degree of coordination is achieved with minimal exchange of state. In previous work [4], the authors considered distributed load balancing as a mechanism for achieving such functionalities for power-controlled services, such as voice connections. In that work it was shown that shadow prices for carefully selected critical resources provide the additional means to dynamically allocate users to cells based on load

considerations, in addition to the standard notion of proximity, and thus achieve a high degree of optimization, without the need for centralization. Dynamic association of users with access points has also been considered in the context of IEEE 802.11 networks, see for instance [2].

In the present paper we combine the utility maximization framework that has been successfully leveraged for scheduling and resource allocation in the uncoordinated case [20, 21] with the distributed optimization approach developed in [4] to achieve a critical level of network optimization. Even though the aim is not to enable full-fledge network-wide scheduling, we show that significant gains result from careful assignment of users to cells and near-optimal allocation of resources from each cell to each user. Additional per-cell scheduling will obviously further improve the performance. The proposed mechanism relies on distributed shadow prices for dynamic load balancing. We examine two related problem versions: (i) minimizing a convex function of the transmitter loads for given user throughput requirements; and (ii) maximizing a concave function of the user throughputs subject to constraints on the transmitter loads. As conceptual counterparts, these two formulations turn out to be amenable to a common primal-dual decomposition framework. Numerical experiments indicate that dynamic load balancing yields significant performance gains in terms of user throughputs and delays, even in scenarios where the long-term loads are perfectly balanced.

The remainder of the paper is organized as follows. In Section 2 we examine the problem of minimizing a convex function of the transmitter loads for given throughput requirements. In Section 3 we turn attention to the problem of maximizing a concave function of the user throughputs subject to constraints on the transmitter loads. We describe how the merits of load balancing schemes can be evaluated in terms of transfer delays and user throughputs in Section 4. In Section 5 we present the numerical experiments that we conducted to benchmark the performance gains from dynamic load balancing.

2 Load minimization

2.1 Model description

We consider a wireless data network with C transmitters. For now we will focus on a static scenario with M users, and address the problem of determining which users should be allocated resources from which transmitters for various optimality criteria. Later we will examine a dynamic setting where users generate random finite-size data transfers and come and go over time. The static scenario considered in the present section may be interpreted as a 'snap shot' in time of such a dynamic situation.

Denote by r_{mc} the feasible transmission rate of user m when served by transmitter c. By feasible rate, we mean the long-term rate that the user would receive if it were allocated all the transmission resources (time slots, power, frequencies) of the transmitter. Let x_{mc} be the actual amount of resources of transmitter c allocated to user m. We assume the transmissions to be (roughly) orthogonal, so that user m receives a total rate of (approximately) $T_m := \sum_{c=1}^{C} x_{mc} r_{mc}$. Let $L_c := \sum_{m=1}^{M} x_{mc}$ be the total load (resource utilization) at transmitter c. Denote by τ_m the rate requirement of user m, and by σ_c the maximum sustainable load on transmitter c, when applicable.

The coefficients r_{mc} only serve as a parsimonious representation of the rate statistics, and likewise the parameters τ_m are only meant to provide a coarse characterization of the traffic demands. We abstract from the specific details of the air-interface structure, and also ignore the burstiness in the traffic processes and the fact that actual transmission rates vary over time because of fast fading. While the latter aspects are clearly crucial for the scheduling at each of the transmitters on a fast time scale, they are less relevant in deciding which users should be served by which transmitters. Also, the quantities x_{mc} will only play the role of decision variables in coordinating the assignment of users to transmitters, with the actual allocation of resources governed by local schedulers residing at the individual transmitters.

2.2 Problem formulation

We first examine the problem of minimizing a convex function $F(L_1, \ldots, L_C)$ of the transmitter loads for given throughput requirements $\tau_1, \ldots, \tau_M$. This formulation is particularly natural when the users have intrinsic rate requirements and conservation of transmission resources (e.g. battery life) is of vital importance, or when the level of congestion is a critical performance measure.

$$\min \; F(L_1, \ldots, L_C) \tag{1}$$

$$\text{sub} \; L_c = \sum_{m=1}^{M} x_{mc} \qquad c = 1, \ldots, C \tag{2}$$

$$T_m = \sum_{c=1}^{C} r_{mc} x_{mc} \geq \tau_m \qquad m = 1, \ldots, M \tag{3}$$

$$x_{mc} \geq 0 \qquad m = 1, \ldots, M, c = 1, \ldots, C.$$

Convex duality implies that the optimal solution to the above problem may be found from the Lagrangian formulation $\max_{\mu \in \mathbb{R}_+^M} F^*(\mu)$, with $F^*(\mu) = \min_{x \in \mathbb{R}_+^{M \times C}} \mathcal{L}(x, \mu)$,

$$\mathcal{L}(x, \mu) := F(L_1, \ldots, L_C) + \sum_{m=1}^{M} \mu_m (\tau_m - \sum_{c=1}^{C} r_{mc} x_{mc}),$$

and $\mu_1, \ldots, \mu_M$ Lagrangian multipliers.
The optimality conditions for the latter formulation read $\frac{\partial F}{\partial L_c} \geq r_{mc} \mu_m^*$, with the complementary slackness conditions $x_{mc}^* \left(\frac{\partial F}{\partial L_c} - r_{mc} \mu_m^* \right) = 0$ for all $c = 1, \ldots, C$, $m = 1, \ldots, M$, and $\mu_m^* (\tau_m - \sum_{c=1}^{C} r_{mc} x_{mc}^*) = 0$ for all $m = 1, \ldots, M$.
Note that the problem (1)-(3) will have a feasible solution (and the Lagrangian will have a finite solution) as long as $\min_{c=1,\ldots,C} r_{mc} > 0$ for all $m = 1, \ldots, M$. Also, there exists an optimal solution with at most $M + C - 1$ non-zero variables x_{mc}^*, which means that there will be at most $C - 1$ additional 'legs' beyond the minimum number that is necessary to connect all the M users.

We now focus on the case where the objective function is of the form $F(L_1, \ldots, L_C) = \sum_{c=1}^{C} K_c(L_c)$, with $K_c(\cdot)$ some strictly convex differentiable function. In that case, x_{mc}^* satisfies $K_c'(L_c^*) = r_{mc} \mu_m^*$ for all $m \in \mathcal{M}_c$, $\mathcal{M}_c = \arg \max_{m=1,\ldots,M} r_{mc} \mu_m^*$, and $x_{mc}^* = 0$ for all $m \notin \mathcal{M}_c$. In the special case where $K_c(x) = x^{1+\beta}/(1+\beta)$ for some $\beta > 0$, we obtain $x_{m_c c} = (r_{m_c c} \mu_{m_c})^{1/\beta}$. The parameter β governs the trade-off between minimizing the total load and the maximum load across all transmitters. As $\beta \downarrow 0$, the objective function becomes $\sum_{c=1}^{C} L_c$, which is minimized by simply assigning each individual user m to the strongest transmitter $c_m = \arg \max_{c=1,\ldots,C} r_{mc}$. In contrast, when $\beta \to \infty$, the problem amounts to minimizing $\max_{c=1,\ldots,C} L_c$, which deserves special treatment and will be examined in further detail below.
The Lagrangian formulation may be interpreted as follows. Each of the users can be allocated resources from each of the transmitters. The cost associated with the load imposed on the transmitter c is specified by the function $K_c(L_c)$, while each unit of throughput obtained by user m carries a reward μ_m. There are two opposing players. Player 1 aims to allocate the resources to users so as to minimize the net cost (or maximize the net revenue) for given rewards μ_m. Player 2 aims to set rewards μ_m, so as to maximize the net cost incurred (minimize the net revenue earned) by the first player.
In principle, the above problems may be readily solved using standard routines. However, these algorithms generally involve centralized computation and require global knowledge of all parameters. Motivated by these issues, we now present an Arrow-Hurwicz type dual-ascent scheme [9] which, while

slower to converge, is mostly distributed in nature, and only involves a limited exchange of information among transmitters and users. The scheme may be interpreted as a repeated game between the two opposing players as described above. The convergence proof is omitted because of page constraints.

Algorithm description for problem (1)–(3)

1. Initialize $\mu = (\mu_1, \ldots, \mu_M)$, e.g., $\mu_m^{(0)} = 1$ for all $m = 1, \ldots, M$.
2. For given $\mu = (\mu_1, \ldots, \mu_M)$, find resource allocations $x_{mc}(\mu)$ that minimize $F^*(\mu) := \sum_{c=1}^{C} K_c(L_c) + \sum_{m=1}^{M} \mu_m(\tau_m - \sum_{c=1}^{C} r_{mc}x_{mc})$. This amounts to allocating the resources of each individual transmitter c to the most 'rewarding' user $m_c := \arg\max_{m=1,\ldots,M} r_{mc}\mu_m$; $x_{m_c c}(\mu)$ satisfies $K_c'(x_{m_c c}(\mu)) = r_{m_c c}\mu_{m_c}^*$, and $x_{mc}(\mu) = 0$ for all $m \neq m_c$. In the special case where $K_c(x) = x^{1+\beta}/(1+\beta)$ for some $\beta > 0$, we obtain $x_{m_c c}(\mu) = (r_{m_c c}\mu_{m_c})^{1/\beta}$.
3. Let $T_m(\mu^{(i)})$ be the aggregate throughput of user c for given $\mu^{(i)} = (\mu_1^{(i)}, \ldots, \mu_M^{(i)})$ as determined in step 2. Update $\mu_m^{(i)}$ as
$$\mu_m^{(i+1)} := \mu_m^{(i)} + \varrho_i(\tau_m - T_m(\mu^{(i)})).$$
To guarantee convergence, it is required that $\lim_{i\to\infty} \varrho_i = 0$ and $\sum_{i=0}^{\infty} \varrho_i = \infty$. For example, one may take $\varrho_i = \varrho i^{-1/2+\epsilon}$ for positive constants ϵ, ϱ. To ensure that $\mu_m^{(i+1)} > 0$ for all $m = 1, \ldots, M$, truncate the update step if needed.
4. Let $x_{mc}(\mu^{(i)})$ be the optimal amount of resources allocated by transmitter c to user m for given $\mu^{(i)} = (\mu_1^{(i)}, \ldots, \mu_M^{(i)})$ as determined in step 2. Update $x_{mc}^{(i)}$ as
$$x_{mc}^{(i+1)} = (1 - \varsigma_i)x_{mc}^{(i)} + \varsigma_i x_{mc}(\mu^{(i)}),$$
with $\varsigma_i := \varrho_i / \sum_{j=0}^{i} \varrho_j$. (In particular, $\varsigma_0 = 1$, so that $x_{mc}^{(1)} = x_{mc}(\mu^{(0)})$.)
5. Repeat the above steps until some convergence/stopping criterion is satisfied.

Some observations

i) In view of the complementary slackness conditions, the optimal shadow price vector $\mu^* \equiv (\mu_1^*, \ldots, \mu_M^*)$ is sufficient to determine which users should be assigned to which transmitters. The exact amount of resources allocated by the various transmitters will in practice be dictated by the traffic generated by the users. In that sense step 4 is optional since it is only required in order to obtain the optimal resource allocations x_{mc}^*, and is not needed for finding the optimal shadow price vector.

ii) As mentioned earlier, the dual-ascent scheme is distributed in nature and only involves limited communication among transmitters and users. Specifically, for a given shadow price vector $\mu = (\mu_1, \ldots, \mu_M)$, the throughput vector $T = (T_1, \ldots, T_M)$ in step 2 can be obtained by each of the transmitters autonomously updating its resource allocation. For a given throughput vector $T = (T_1, \ldots, T_M)$, the update in the shadow price vector $\mu = (\mu_1, \ldots, \mu_M)$ can be determined by each of the users separately. We now investigate the case where the objective function is of the form $F(L_1, \ldots, L_C) := \max_{c=1,\ldots,C} w_c L_c$, i.e., the maximum weighted load across all transmitters. In that case, problem (1)–(3) reduces to the following linear program:

$$\min \quad L \tag{4}$$

$$\text{sub} \quad L \geq w_c L_c = w_c \sum_{m=1}^{M} x_{mc} \qquad c = 1, \ldots, C \tag{5}$$

$$T_m = \sum_{c=1}^{C} r_{mc}x_{mc} \geq \tau_m \qquad m = 1, \ldots, M \tag{6}$$

$$x_{mc} \geq 0 \qquad m = 1, \ldots, M, c = 1, \ldots, C.$$

The dual version of the above linear program reads:

$$\max \sum_{m=1}^{M} \tau_m \mu_m \tag{7}$$

$$\text{sub} \sum_{c=1}^{C} \lambda_c \leq 1 \tag{8}$$

$$r_{mc}\mu_m \leq w_c \lambda_c \qquad m = 1, \ldots, M, c = 1, \ldots, C \tag{9}$$

$$\lambda_c, \mu_m \geq 0 \qquad m = 1, \ldots, M, c = 1, \ldots, C,$$

with λ_c and μ_m representing the dual variables or shadow prices associated with the constraints (5) and (6), respectively.

Since optimality demands $\sum_{c=1}^{C} \lambda_c^* = 1$ and $\mu_m^* = \min_{c=1,\ldots,C} w_c \lambda_c^* / r_{mc}$, the latter variables may be eliminated, and the dual problem may be more succinctly cast as:

$$\max \; V(\lambda_1, \ldots, \lambda_C) \tag{10}$$

$$\text{sub} \sum_{c=1}^{C} \lambda_c = 1 \tag{11}$$

$$\lambda_c \geq 0 \qquad c = 1, \ldots, C,$$

with $V(\lambda_1, \ldots, \lambda_c) := \sum_{m=1}^{M} \tau_m \min_{c=1,\ldots,C} w_c \lambda_c / r_{mc}$. The latter quantity represents the minimum value of $\sum_{c=1}^{C} w_c \lambda_c L_c$ for given throughput requirements τ_m.

The latter problem may be interpreted in a similar fashion as above. Each of the users can be allocated resources from each of the transmitters. User m needs to receive throughput τ_m, while each unit of resource allocated by transmitter costs $w_c \lambda_c$, so the cost when transmitter m provides the entire througput required by user m is $w_c \lambda_c \tau_m$. There are two 'opposing' players. Player 1 aims to allocate transmission resources to the users so as to minimize the total cost for given prices λ_c while satisfying the throughput requirements. Player 2 aims to set prices λ_c so as to maximize the total cost incurred by the first player.
The problem (4)–(6) may be solved by a dual-ascent scheme similar to the one that will be described in the following section.

3 Throughput maximization

We now turn attention to the problem of maximizing a concave function $G(T_1, \ldots, T_M)$ of the user throughputs for given load (resource utilization) constraints $\sigma_1, \ldots, \sigma_C$. This formulation is appropriate when users have elastic traffic demands and the consumption of transmission resources (e.g. power) is constrained by hard limits, but not a crucial criterion otherwise.

$$\max \; G(T_1, \ldots, T_M) \tag{12}$$

$$\text{sub} \; T_m = \sum_{c=1}^{C} r_{mc} x_{mc} \qquad m = 1, \ldots, M \tag{13}$$

$$L_c = \sum_{m=1}^{M} x_{mc} \leq \sigma_c \qquad c = 1, \ldots, C \tag{14}$$

$$x_{mc} \geq 0 \qquad m = 1, \ldots, M, c = 1, \ldots, C.$$

The above formulation is conceptually similar to the multi-path extension of the basic utility maximization problem in [11], i.e., joint routing and rate control, see also for instance [10, 12, 14, 15, 19, 22–24].
Convex duality implies that the optimal solution to the above problem may be found from the Lagrangian formulation $\min_{\lambda \in \mathbb{R}_+^C} G^*(\lambda)$, with $G^*(\lambda) = \max_{x \in \mathbb{R}_+^{M \times C}} \mathcal{L}(x, \lambda)$,

$$\mathcal{L}(x, \lambda) := G(T_1, \ldots, T_M) + \sum_{c=1}^{C} \lambda_c (\sigma_c - \sum_{m=1}^{M} x_{mc}),$$

and $\lambda_1, \ldots, \lambda_C$ Lagrangian multipliers.
The optimality conditions for the latter formulation read $\frac{\partial G}{\partial T_m} \leq \lambda_c^*/r_{mc}$, with the complementary slackness conditions $x_{mc}^* \left(\frac{\partial G}{\partial T_m} - \lambda_c^*/r_{mc} \right) = 0$ for all $c = 1, \ldots, C$, $m = 1, \ldots, M$, and $\lambda_c^* (\sigma_c - \sum_{m=1}^{M} x_{mc}^*) = 0$ for all $m = 1, \ldots, M$.
Note that the problem (12)-(14) will have a finite solution as long as $\min_{c=1,\ldots,C} r_{mc} > 0$ for all $m = 1, \ldots, M$. Also, there exists an optimal solution with at most $M + C - 1$ non-zero variables x_{mc}^*, which means that there will be at most $C - 1$ additional 'legs' beyond the minimum number that is necessary to connect all the M users.

We now focus on the case where the objective function is of the form $G(T_1, \ldots, T_M) = \sum_{m=1}^{M} U_m(T_m)$, with $U_m(\cdot)$ some strictly concave differentiable function. In that case, x_{mc}^* satisfies $U_m'(T_m^*) = r_{mc} \mu_m^*$ for all $c \in \mathcal{C}_m$, $\mathcal{C}_m := \arg \min_{c=1,\ldots,C} \lambda_c^*/r_{mc}$, and $x_{mc}^* = 0$ for all $c \notin \mathcal{C}_m$.

We will in particular consider the family of α-fair utility functions defined by $U_\alpha(T) = \frac{T^{1-\alpha}}{1-\alpha}$ for some $\alpha > 0$, yielding $x_{m_c c}^* = r_{m_c c}^{1/\alpha - 1} (\lambda_c^*)^{-1/\alpha}$. The parameter α represents a fairness coefficient which characterizes the trade-off between the total throughput and the minimum throughput across all users [16]. In particular, the cases $\alpha \downarrow 0$, $\alpha \to 1$ and $\alpha \to \infty$ correspond to maximum throughput, Propertional Fairness and max-min fairness, respectively. As $\alpha \downarrow 0$, optimality is achieved by simply allocating all the resources of each individual transmitter c to the strongest received user $m_c = \arg \max_{c=1,\ldots,C} r_{mc}$. In contrast, when $\alpha \to \infty$, the problem merits special treatment and will be revisited below.

Algorithm description for problem (12)–(14)

1. Initialize $\lambda = (\lambda_1, \ldots, \lambda_C)$, e.g., $\lambda_c^{(0)} = M/C$ for all $c = 1, \ldots, C$.
2. For given $\lambda = (\lambda_1, \ldots, \lambda_C)$, find resource allocations x_{mc} that maximize $\mathcal{L}(x, \lambda) := \sum_{m=1}^{M} U(T_m) + \sum_{c=1}^{C} \lambda_c (\sigma_c - \sum_{m=1}^{M} x_{mc})$. This amounts to allocating each individual user m resources from the most attractive transmitter $c_m := \arg \min_{c=1,\ldots,C} \lambda_c/r_{mc}$; x_{mc_m} satisfies $U_m'(x_{mc_m} r_{mc_m}) = \lambda_{c_m}/r_{mc_m}$, and $x_{mc} = 0$ for all $c \neq c_m$. In the special case where $U_m(x) = x^{1-\alpha}/(1-\alpha)$ for some $\alpha > 0$, we obtain $x_{mc_m} = (\lambda_{mc_m}/r_{c_m})^{-1/\alpha}/r_{mc_m} = r_{mc_m}^{1/\alpha - 1} \lambda_{c_m}^{-1/\alpha}$.
3. Let $L_c(\lambda^{(i)})$ be the optimal load at transmitter c for given $\lambda^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_C^{(i)})$ as determined in step 2. Update $\lambda_c^{(i)}$ as
$$\lambda_c^{(i+1)} := \lambda_c^{(i)} + \varrho_i (L_c(\lambda^{(i)}) - \sigma_c).$$
To guarantee convergence, it is required that $\lim_{i \to \infty} \varrho_i = 0$ and $\sum_{i=0}^{\infty} \varrho_i = \infty$. For example, one may take $\varrho_i = \varrho i^{-1/2+\epsilon}$ for positive constants ϵ, ϱ. To ensure that $\lambda_c^{(i+1)} > 0$ for all $c = 1, \ldots, C$, truncate the update step if needed.
4. Let $x_{mc}(\lambda^{(i)})$ be the optimal amount of resources allocated by transmitter c to user m for given $\lambda^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_C^{(i)})$ as determined in step 2. Update $x_{mc}^{(i)}$ as
$$x_{mc}^{(i+1)} = (1 - \varsigma_i) x_{mc}^{(i)} + \varsigma_i x_{mc}(\lambda^{(i)}),$$

with $\varsigma_i := \varrho_i / \sum_{j=0}^{i} \varrho_j$.

5. Repeat the above steps until some convergence/stopping criterion is satisfied.

The convergence proof is skipped because of page limitations.

Some observations

i) In view of the complementary slackness conditions, the optimal shadow price vector $\lambda^* \equiv (\lambda_1^*, \ldots, \lambda_C^*)$ suffices to determine which users should be served by which transmitters. The exact amount of resources allocated to the various users will in practice be governed by local schedulers at each of the transmitters. In that sense step 4 is again optional, as it only serves to obtain the optimal resource allocations x_{mc}^*, and plays no role in finding the optimal shadow price vector.

ii) As before the dual-ascent scheme allows a distributed implementation, with both users and transmitters making mostly autonomous decisions and only parsimoniously exchanging information.

We now investigate the case where the objective function is of the form $G(T_1, \ldots, T_M) := \min_{m=1,\ldots,M} T_m$, i.e., the minimum throughput across all users. In that case, problem (12)–(14) reduces to the following linear program: (with $v_m \equiv 1$ for all $m = 1, \ldots, M$):

$$\begin{aligned} \max \ & T \\ \text{sub} \ & T \leq v_m T_m = v_m \sum_{c=1}^{C} r_{mc} y_{mc} \qquad c = 1, \ldots, C \qquad (15) \\ & L_c = \sum_{m=1}^{M} y_{mc} \leq \sigma_c \qquad m = 1, \ldots, M \qquad (16) \\ & y_{mc} \geq 0 \qquad m = 1, \ldots, M, c = 1, \ldots, C. \end{aligned}$$

The above problem is equivalent to (4)–(6) with $v_m = 1/\tau_m$ and $\sigma_c = 1/w_c$ in the sense that the optimal solutions are related and hence will not be discussed further.

4 Dynamic setting

In the previous section we addressed the problem of maximizing a throughput utility function for a given static user population. While utility maximization provides a useful guiding principle for fair and efficient resource sharing among competing users, the utility function does not necessarily have any physical meaning in terms of actual perceived performance. In particular, the exact numerical value of the utility function or the fact that the aggregate system utility has been maximized may not be of any direct relevance to a data user. What a data user does perceive, is the performance experienced in terms of delays or actual received throughputs for example, and hence we will evaluate the merits of the load balancing schemes in terms of these metrics. In order to do so, we will consider a dynamic setting where users generate random finite-size data transfers over time. For convenience, we assume that users belong to one of K classes, with transmission rates taking values in a discrete set of values, but the results easily extend to scenarios with a continuum of rates. Class-k users arrive as a stationary ergodic process of rate ν_k, and have generally distributed service requirements with mean β_k (bits). Define $\rho := (\rho_1, \ldots, \rho_K)$, with $\rho_k := \nu_k \beta_k$ the traffic intensity associated with class-k users. Denote by R_{kc} the feasible transmission rate of class-k users when served by transmitter c. As before, we assume the transmissions to be (roughly) orthogonal, so that class k receives an aggregate rate of (approximately) $\sum_{k=1}^{K} x_{kc} R_{kc}$, when x_{kc} is the total amount of resources of transmitter c allocated to class k.

In order for delays and user throughputs to be meaningful, a first prerequisite is that the system is stable. Define the rate region of the system by

$$\mathcal{R} := \{r \in \mathbb{R}_+^K : \exists x \in \mathcal{X} : r_k \leq \sum_{c=1}^{C} x_{kc} R_{kc} \text{ for all } k = 1, \ldots, K\},$$

with $\mathcal{X} := \{x \in \mathbb{R}_+^{K \times C} : \sum_{k=1}^{K} x_{kc} \leq \sigma_c \text{ for all } c = 1, \ldots, C\}$ representing the set of feasible resource allocations. Clearly, $\rho \in \mathcal{R}$ is a necessary condition for the existence of a resource allocation strategy that achieves stability, while $\rho \in \text{interior}(\mathcal{R})$ is a sufficient condition.

We will consider four different scenarios.

(i) The 'greedy' scheme simply assigns users to the strongest received transmitter. Thus, class-k users are statically assigned to transmitter $c_k := \arg \max_{c=1,\ldots,C} R_{kc}$. Let $\mathcal{K}_c := \{k : c_k = c\}$ be the set of user classes assigned to transmitter c, and define $\bar{\sigma}_c := \sum_{k \in \mathcal{K}_c} \rho_k / R_{kc}$ as the resulting load on transmitter c. It is easily seen that the greedy assignment achieves stability if and only if $\bar{\sigma}_c < \sigma_c$ for all $c = 1, \ldots, C$.

(ii) The 'fractional' α-fair strategy assigns users to transmitters so as to maximize the aggregate α-fair utility. Specifically, suppose that there are N_k class-k users at some point in time. Then the fractional α-fair strategy solves the following optimization problem:

$$\begin{aligned}
\max \ & \sum_{k=1}^{K} \sum_{n=1}^{N_k} U_\alpha(T_{kn}) \\
\text{sub} \ & T_{kn} = \sum_{c=1}^{C} R_{kc} x_{knc} \qquad n = 1, \ldots, N_k, k = 1, \ldots, K \qquad (17) \\
& L_c = \sum_{k=1}^{K} \sum_{n=1}^{N_k} x_{knc} \leq \sigma_c \qquad c = 1, \ldots, C \qquad (18) \\
& x_{knc} \geq 0.
\end{aligned}$$

It is easily verified that all the class-k users will receive the same throughput, and thus the above problem can alternatively be phrased as maximizing $\sum_{k=1}^{K} U_\alpha(T_k/N_k)$ subject to $(T_1, \ldots, T_K) \in \mathcal{R}$. Since $\mathcal{R}$ is a convex set, it then follows that the fractional α-fair strategy achieves stability for any $\rho \in \text{interior}(\mathcal{R})$, provided $\alpha > 0$.

(iii) The 'integral' α-fair strategy also assigns users to transmitters so as to maximize the aggregate utility, but subject to the additional constraint that users can only be assigned to a single transmitter. It is readily checked that in this case all the class-k users assigned to the same transmitter will receive the same throughput. Also, the above problem can be equivalently stated as maximizing $\sum_{k=1}^{K} U_\alpha(T_k/N_k)$ subject to $(T_1, \ldots, T_K) \in \mathcal{R}(N_1, \ldots, N_K)$. Here $\mathcal{R}(N_1, \ldots, N_K) \subseteq \mathcal{R}$ is some subset with the property that $\lim_{N_1,\ldots,N_K \to \infty} \mathcal{R}(N_1, \ldots, N_K) = \mathcal{R}$. This implies that the integral α-fair strategy also achieves stability for any $\rho \in \text{interior}(\mathcal{R})$, provided $\alpha > 0$.

(iv) The 'ideal' scenario is where the resources of all the transmitters can be pooled into a single transmitter which offers a transmission rate $R_k^{\max} = \max_{c=1,\ldots,C} R_{kc}$ to class-k users. This is a hypothetical scenario typical propagation conditions, and is only meant to provide an absolute bound on the achievable performance. It is easily seen that stability occurs in the ideal scenario if and only if $\sum_{c=1}^{C} \bar{\sigma}_c < \sum_{c=1}^{C} \sigma_c$.

In conclusion, both the fractional and the integral α-fair strategies achieve stability whenever feasible. The greedy assignment may generally fail to do so, while the stability region for the ideal scenario will typically be strictly larger than $\mathcal{R}$, except in the rather special circumstance that $\bar{\sigma}_c \equiv \bar{\sigma}$ for all

$c = 1, \ldots, C$. In that case the stability regions for both the greedy assignment and the ideal scenario coincide with interior($\mathcal{R}$).

We now compare the performance in terms of delays and perceived throughghputs in the various scenarios. Clearly, if $\rho \in$ interior($\mathcal{R}$), but $\bar{\sigma}_c > \sigma_c$ for some $c = 1, \ldots, C$, i.e., the loads are imbalanced, but the total load is sustainable, then the delay under the greedy assignment will be infinite, whereas it is finite under both the fractional and integral α-fair schemes. Now consider the case $\bar{\sigma}_c < \sigma_c \equiv 1$ for all $c = 1, \ldots, C$. If we assume Poisson arrivals and fair sharing of resources among competing users, then the mean number of active users under the greedy assignment is

$$EN^{greedy} = \sum_{c=1}^{C} \frac{\bar{\sigma}_c}{1 - \bar{\sigma}_c},$$

whereas the mean number of users in the ideal scenario is

$$EN^{ideal} = \left(\sum_{c=1}^{C} \bar{\sigma}_c \right) / \left(C - \sum_{c=1}^{C} \bar{\sigma}_c \right).$$

It is easily verified that for a given value of $\sum_{c=1}^{C} \bar{\sigma}_c$, EN^{greedy} is minimal if $\bar{\sigma}_c \equiv \bar{\sigma} = \sum_{c=1}^{C} \bar{\sigma}_c / C$, and then equal to $C\bar{\sigma}/(1-\bar{\sigma}) = \sum_{c=1}^{C} \bar{\sigma}_c / (1 - \sum_{c=1}^{C} \bar{\sigma}_c / C) = CEN^{ideal}$. Thus, even in the best-case scenario where the loads are perfectly balanced, the mean number of active users under the greedy assignment is C times as large as in the ideal scenario. Because of Little's law, this implies that the mean delays will be C times as large as well, and thus the throughputs (defined as the ratio of service requirement and delay) will be C times higher. Although the delays for the fractional and integral α-fair strategies are expected to be "somewhere in between", this appears difficult to prove, let alone quantify where exactly they fall relative to the greedy scheme and the ideal scenario. Hence, we will examine the delay performance in the various scenarios in the next section through numerical means.

5 Numerical experiments

We now discuss the numerical experiments that we conducted to benchmark the performance gains from dynamic load balancing. We consider a dynamic setting where users generate random finite-size data transfers as described in the previous section, and evaluate the performance in terms of transfer delays and blocking rates.

We first examine a linear network with just two transmitters. While admittedly simple, a two-transmitter scenario is likely to provide conservative estimates for the potential gains, since the scope for load sharing increases with the number of neighboring transmitters, as will in fact be confirmed later.

The two transmitters cover an interval $[0, D]$ and are located at positions $D/6$ and $5D/6$, respectively. The path loss q behaves as a function $q = d^{-\gamma}$ of distance d, with $\gamma = 3.5$. The feasible transmission rate r at transmitter c (in bits/second behaves as a function $r_c = \zeta \log(1 + snr_c)$ of the Signal-to-Noise Ratio (SNR), with $snr_c = q_c/(\eta + \theta q_c + q_{3-c})$, $c = 1, 2$, with ζ, θ, η system-specific parameters. Throughout we take $\eta = 0.01$, $\theta = 0.1$, $\zeta = 800$.

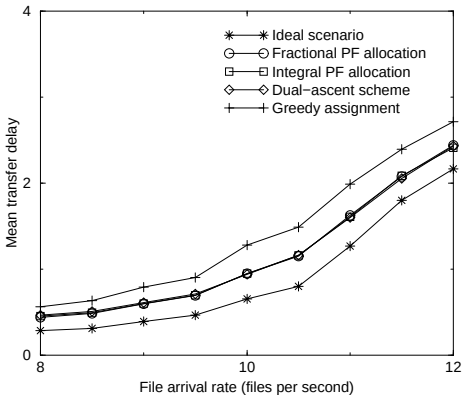
Users arrive as a Poisson process of rate ν (per second), and have service requirements with mean β (in bits). Throughout we take $\beta = 250$ Kbits (31.25 Kbytes). At most $L = 40$ users are admitted into the system simultaneously. Users that generate a transfer request when there are already L transfers in progress, are blocked and lost. Let (R_1, R_2) be the rate pair of an arbitrary user. The nominal average load on transmitter c under the greedy assignment may then be derived as

$$\nu\beta\mathbb{E}\{\frac{1}{R_c}\mathrm{I}_{\{R_c > R_{3-c}\}}\} = \nu\beta\mathbb{E}\{\frac{1}{\max\{R_1, R_2\}}\}\mathbb{P}\{R_c > R_{3-c}\}.$$

We compare the four scenarios described in Section 4 as well as the dual-ascent scheme. We take $U(x) = \log(x)$, i.e., $\alpha \to 1$, which corresponds to Proportional Fair (PF) scheduling. In the dual-ascent scheme, we only executed 30 iterations for every change in the user population, with $\varrho_i = 0.5/\sqrt{i}$.

We first consider a scenario where the user locations are uniformly distributed across the coverage area. In this case, the nominal load on each of the two transmitters is $\rho/2$, with $\rho := \nu\beta\mathbb{E}\{\frac{1}{\max\{R_1,R_2\}}\}$. Since the long-term loads are perfectly balanced, this provides a lower bound for the potential gains from load balancing.

Figure 1 shows the mean transfer delay as function of the arrival rate. Note that the delay in the ideal scenario is roughly half of that under the greedy assignment, as indicated by the analysis in the previous section At high load, the relative difference reduces though. This may be explained from the fact that a significant fraction of the users are blocked under the greedy assignment (not shown in the figure), effectively reducing the load on the system, while the blocking in the ideal scenario remains negligible throughout. Further observe that the performance of the dual-ascent scheme is virtually indistinguishable from that of the globally optimal integral or fractional PF allocation. Given the small number of iterations, this indirectly demonstrates that the dual-ascent scheme converges rapidly enough to achieve similar performance as a globally optimal allocation. It also suggests that the dynamic load balancing is hardly hindered by refraining from soft hand-off and assigning users to just a single transmitter. The mean transfer delay in each of these three cases is approximately 15–25% lower than under the greedy assignment, even though the long-term loads are perfectly balanced.



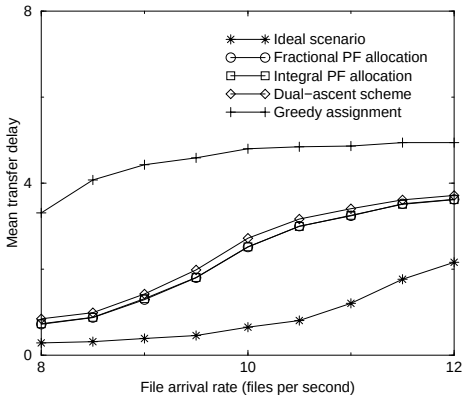
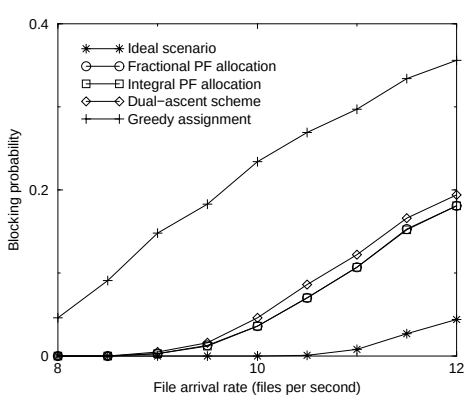


Figure 1 (left): Mean transfer delay as function of arrival rate; uniform traffic. Figure 2 (right): Mean transfer delay as function of arrival rate; non-uniform traffic.

We now look at a scenario where the density of users is three times higher in one half of the coverage area than the other. In this case, the nominal loads on the two transmitters are $\rho/4$ and $3\rho/4$, respectively.

Figure 2 plots the mean transfer delay as function of the arrival rate. As before, the performance of the dual-ascent scheme is practically identical to that of the globally optimal integral or fractional PF allocation. The reduction in mean transfer delay in each of these three cases varies from 30% to 80%, which means that the improvement in perceived user throughput can be as large as 500%. Further note that at high load, the relative improvement diminishes. This reflects the fact that a substantial fraction of the users are blocked under the greedy assignment, as shown in Figure 3, essentially lowering the load on the system, while the blocking with load balancing remains moderate. Observe that even with load balancing the system will become overloaded at some point, but load balancing helps to move that point significantly further out.



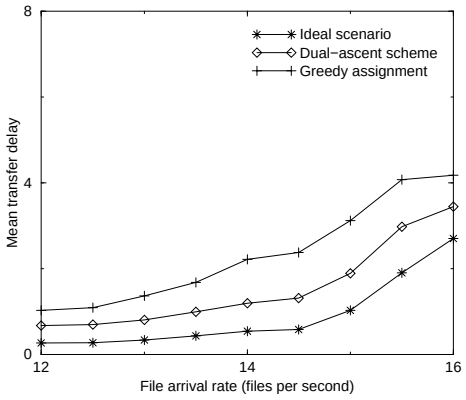


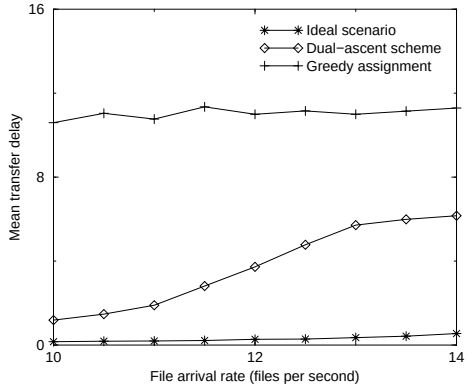
Figure 3 (left): Blocking probability as function of arrival rate; non-uniform traffic. Figure 4 (right): Mean transfer delay as function of arrival rate; uniform traffic.

We now investigate a network with a square coverage area $[0, D] \times [0, D]$, with four transmitters located at positions $(D/6, D/6)$, $(D/6, 5D/6)$, $(5D/6, D/6)$, $(5D/6, 5D/6)$. In this case the SNR at transmitter c is determined by $snr_c = q_c/(\eta + \sum_{d=1}^{4} q_d - (1-\theta)q_c)$, with q_d the path loss to transmitter d governed by a power-law as function of distance with exponent $\gamma = -3.5$ as specified earlier. At most $L = 80$ users are admitted into the system simultaneously. We only present the results for the dual ascent scheme, and not for the globally optimal integral and fractional PF allocation.

We first consider a scenario where the user locations are uniformly distributed across the coverage area. In this case, the nominal load on each of the four transmitters is $\rho/4$, with $\rho := \nu\beta\mathbb{E}\{\frac{1}{\max\{R_1, R_2, R_3, R_4\}}\}$. Figure 4 shows the mean transfer delay as function of the arrival rate. Note that the delay in the ideal scenario is now roughly a quarter of that under the greedy assignment, as predicted by the analysis in the previous section. At high load, the relative difference decreases again because a significant fraction of the users are blocked under the greedy assignment, effectively shedding load from the system, while the blocking in the ideal scenario remains negligible throughout. The mean transfer delay for the dual ascent-scheme is now 30–50% lower than under the greedy assignment, which corroborates the earlier assertion that the gains from load balancing tend to increase with the number of transmitters.

We now look at a scenario where the density of users is three times higher in one square corner of the coverage area than elsewhere. In this case, the nominal load on one transmitter is $\rho/2$ and $\rho/6$ on the other three.

Figure 5 plots the mean transfer delay as function of the arrival rate. The reduction in mean transfer delay achieved by the dual-ascent scheme ranges from 50% to 90%, which means that the improvement in perceived user throughput can be as large as 1000%. Further note that at high load, the relative improvement diminishes again because a substantial fraction of the users are blocked under the greedy assignment, as shown in Figure 6.



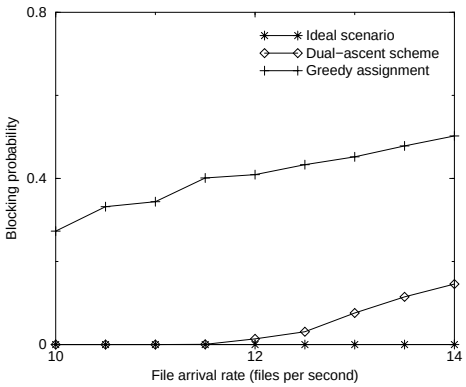


Figure 5 (left): Mean transfer delay as function of arrival rate; non-uniform traffic. Figure 6 (right): Blocking probability as function of arrival rate; non-uniform traffic.

References

1. D.M. Andrews, L. Qian, A.L. Stolyar (2005). Optimal utility-based multi-user throughput allocation subject to throughput constraints. In: *Proc. Infocom 2005.*
2. Y. Bejerano, S.-J. Han, L. Li (2004). Fairness and load balancing in wireless LAN's using association control. In: *Proc. ACM MobiCom 2004*, 315–329.
3. Chaponniere, E.F., Black, P.J., Holtzman, J.M., Tse, D.N.C. (2002). Transmitter directed code division multiple access system using path diversity to equitably maximize throughput. US Patent 6,449,490.
4. S.C. Borst, G. Hampel, I. Saniee, P.A. Whiting (2006). Load balancing in cellular wireless networks.
5. T. Bu, L. Li, R. Ramjee (2006). Generalized Proportional Fair scheduling in third-generation wireless networks. In: *Proc. Infocom 2006*
6. S. Das, H. Viswanathan, G. Rittenhouse (2003). Dynamic load balancing through coordinated scheduling in packet data systems. In: *Proc. Infocom 2003.*
7. E. Eryilmaz, R. Srikant (2005). Fair resource allocation in wireless networks using queue length based scheduling and congestion control. In: *Proc. Infocom 2005.*
8. H. Han, S. Shakkottai, C.V. Hollot, R. Srikant, D. Towsley (2006). Multi-path TCP: a joint congestion control and routing scheme to exploit path diversity in the Internet. *IEEE/ACM Trans. Netw.*, to appear.
9. L. Hurwicz, K. Arrow, H. Uzawa (1958). *Studies in Linear and Non-Linear Programming.* Stanford University Press.
10. K. Kar, S. Sarkar, L. Tassiulas (2001). Optimization based rate control for multi-path sessions. Technical Report 2001-1, Institute for Systems Research, University of Maryland.
11. F.P. Kelly, A. Maulloo, D. Tan (1998). Rate control in communication networks: shadow prices, proportional fairness, and stability. *J. Oper. Res. Soc.* **49**, 237–252.
12. F.P. Kelly, T. Voice (2005). Stability of end-to-end algorithms for joint routing and rate control. *Comp. Commun. Rev.* **35**, 5–12.
13. Liu, X., Chong, E.K.P., Shroff, N.B. (2003). A framework for opportunistic scheduling in wireless networks. *Comp. Netw.* **41**, 451–474.
14. X. Lin, N.B. Shroff (2003). Utility maximization for communication networks with multi-path routing. *IEEE Trans. Aut. Control* **51**, 766–781.
15. S.H. Low (1999). Optimization flow control with on-line measurements or multiple paths. In: *Proc. ITC-16.*
16. J. Mo, J. Walrand (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.* **8**, 556–567.
17. M.J. Neely, E. Modiano, C. Li (2005). Fairness and optimal stochastic control for heterogeneous networks. In: *Proc. Infocom 2005.*
18. A. Sang, X. Wang, M. Madihian, R.D. Gitlin. Coordinated load balancing / cell-site selection and scheduling in multi-cell packet data systems. In: *Proc. ACM Mobicom 2004.*
19. V. Srinivasan, C. Chiasserini, P. Nuggehalli, R. Rao (2005). Optimal rate allocation for energy-efficient multi-path routing in wireless ad hoc networks. *IEEE Trans. Wireless Commun.* **3**, 891–899.
20. A.L. Stolyar (2005). On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation. *Oper. Res.* **53**, 12–25.
21. A.L. Stolyar (2005). Maximizing queueing network utility subject to stability: greedy primal-dual algorithm. *Queueing Systems* **50**, 401–457.
22. T. Voice (2006). Stability of multi-path dual congestion control algorithms. In: *Proc. ValueTools 2006*, to appear.
23. J. Wang, L. Li, S.H. Low, J.C. Doyle (2005). Cross-layer optimization in TCP/ICP networks. *IEEE/ACM Trans. Netw.* **13**, 582–595.
24. W.H. Wang, M. Palaniswami, S.H. Low (2002). Optimal flow control and routing in multi-path networks. *Perf. Eval.* **52**, 119–132.