Research Article

Network Utilization Improvement via a Predictive Load Scheduling Scheme in Heterogeneous Wireless Networks

Bin Fang and Wuyang Zhou

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China

Correspondence should be addressed to Bin Fang; fcb096@mail.ustc.edu.cn

Received 19 January 2015; Accepted 29 March 2015

Academic Editor: Satyajayant Misra

Copyright © 2015 B. Fang and W. Zhou. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In heterogeneous wireless networks (HWNs), the large-coverage cellular networks are usually overlaid with denser small-coverage microcells, wireless local area networks (WLANs), femtocells, and even relays. To prevent unbalanced load distribution in these HWNs from decreasing network utilization and degrading the quality of service (QoS), load scheduling between the overlaid networks is essential and should be reasonably designed. In fact, unbalanced load distribution can be directly reflected by the idle durations of networks and channels. In this paper, a predictive load scheduling scheme is proposed from the point of decreasing the capacity-weighted idle durations of networks and channels. For the proposed scheme, a fairly general HWNs scenario is considered, the scheduling problem is formulated, an iteration-based predictive method is given to obtain the idle duration, and a gradient descent method is derived for the optimal scheduled load. In the simulation, the effectiveness of the idle duration prediction method is validated, and the simulation results show that the proposed load scheduling scheme can greatly improve the network utilization and QoS.

1. Introduction

The integration of heterogeneous wireless networks (HWNs) has become a promising trend to satisfy the fast increasing demand for higher wireless data rate and lower transmission delay. In HWNs, the small-coverage microcells, WLANs, femtocells, and even relays are deployed in hotspots to offload from the large-coverage cellular networks. In general, the deployment of small-coverage cells can increase the bandwidth reusability and global network capacity. Different from the traditional homogeneous cellular networks, the coverage of HWNs with different radio transmission techniques can be overlapped and user equipments (UEs) in the overlapped region can be served by different networks. So, how to reasonably distribute these UEs or their traffic to the overlapped HWNs for load balance is a key problem.

On the other hand, due to the time-variant traffic arrivals and the diversity of traffic classes, the unbalanced load distribution that some networks are overloaded while some other networks are underloaded may occur from time to time. Unbalanced load distribution leads to an irrational result, that under-loaded networks or some of their channels are in idle state when other networks cannot cope with their heavy traffic load and thus decreases the network utilization and degrades the QoS of UE. Therefore, in HWNs with overlapped coverage, when load unbalance occurs, re-distributing these UEs or traffic load in the overlapped regions through load scheduling may improve the load balance.

Meanwhile, with the development of multimode access terminals and multipath technic [1], the emerging mobile wireless terminal would be able to access several networks simultaneously. So a traffic stream in the multimode UE can be divided into several substreams that are transmitted through different networks. This potential traffic stream division method brings breakthrough in load scheduling, that through reasonably dividing traffic streams into substreams and finely distributing traffic data amount in these divided substreams, a more precise load scheduling can be obtained. This paper focuses on traffic streams scheduling and employs this stream division method to implement scheduling.

As an important aspect of load balancing, load scheduling has drawn keen interests of investigators, and its basic idea is
to schedule some traffic load in the overlapped region from the overlaid networks to the underloaded network. These works on load scheduling can be divided into two classes: hard load scheduling [2–5] and soft load scheduling [6–9]. The hard load scheduling schedules the load in units of UEs or traffic streams while the soft scheduling schedules the load in units of substreams or packets by taking multipaths technique [1] into consideration. In the literature, many metrics such as load threshold [2], economic utilities [4, 7–9], balance factors [3], residual resource [5], interference level [6] have been developed to implement load scheduling.

Works in [2] give a hysteresis threshold based load scheduling scheme that the scheduling triggering and scheduled load is decided by the load thresholds. Works in [3] use a load balance factor that considers the global load distribution and UEs' characteristics to implement load scheduling in HWNs and give a fuzzy logic method to decide which UE would be scheduled. Works in [4] define a utility function that considers load states and traffic types for load scheduling and give a Q-learning based method to obtain the scheduled traffic. Works in [5] schedule load based on the residual resources in different networks and give a Markov decision process based method to achieve scheduled load. Works in [6] use the interference level to deploy load scheduling and derive the scheduled load by solving the interference function. Works in [7, 8] formulate utility functions considering traffic load size and transmission rate, and achieve load balance by scheduling substreams. Works in [9] formulate a utility function that considers network capacity and residual resource and obtain the optimal scheduled load by a weighted bargaining game framework.

In this paper, we investigate the soft load scheduling in HWNs and propose a predictive scheduling scheme from the point of decreasing the capacity-weighted idle durations of networks and channels. Compared with the former works, our works have the following novelties and improvements.

(i) Our works consider a fairly general heterogeneous wireless network scenario where the number of networks and their positions are not fixed. Multipaths is considered and the loads that can be and cannot be scheduled between two overlapped networks are differentiated.

(ii) Our works not only consider the current load distribution in the global networks but also consider the arriving traffic loads before the next scheduling. An iteration process based predictive method is proposed to predict the idle durations.

(iii) Our works firstly formulate the objective of minimizing the capacity-weighted idle durations of networks and channels for scheduling and derive a gradient descent method to obtain the optimal solution.

2. Scenario and Problem Formulation

Consider a geographical region with a set $\mathcal{N} = \{1, 2, \ldots, N\}$, of different available wireless base stations (BSs) and access points (APs), as shown in Figure 1. Each BS or AP has its unique coverage and may have overlapped coverage in some regions with other BS/AP. As a result, the geographical region is partitioned to a set, $\mathcal{K} = \{1, 2, \ldots, K\}$, of service regions. Each service region, $k \in \mathcal{K}$, is covered by a unique subset of BSs/APs, as shown in Figure 1. The subset of available wireless BSs/APs at service region $k$ is denoted by $\mathcal{X}_k$, and the subset of service regions that are covered by BS/AP $n$ is denoted by $\mathcal{N}_n$. As shown in Figure 1, the coverage of BS 1 is divided into 5 service regions, so $\mathcal{N}_1 = \{1, 2, 3, 4, 5\}$, and the available wireless BSs/APs at service region 3 include BS 1 and AP 2, so $\mathcal{X}_3 = \{1, 2\}$. It is assumed that all the BSs/APs are connected to a backbone to exchange their roaming signal information, and the proposed scheduling scheme relies on the roaming signal to exchange the traffic load state information and traffic load scheduling information.

Hence, the traffic load that can be scheduled between two BSs is the unserved traffic load in the overlapped regions of the two BSs. The traffic load in each BS $n$, $n \in \mathcal{N}$, can be divided into a set $\mathcal{R}_n = \{r_{n,k}, k \in \mathcal{N}_n\}$, of traffic loads in the regions of $\mathcal{N}_n$, and $r_{n,k}$ denotes the traffic load of BS $n$ in region $k$. The traffic load that would be scheduled from BS $n$ to BS $m$ in region $k$, which is denoted by $x_{n,m}^{k,n}$, should be less than the traffic load of BS $n$ in region $k$, that is $r_{n,k}$, so

$$0 \leq x_{n,m}^{k,n} \leq r_{n,k}, \quad k \in \mathcal{N}_n, k \in \mathcal{X}_m. \quad (1)$$

And the total scheduled load from BS $n$ in region $k$ should be less than $r_{n,k}$; that is

$$\sum_{m \in \mathcal{X}_k} x_{n,m}^{k,n} \leq r_{n,k}. \quad (2)$$

Denote the downlink transmission capacity of BS $n$ as $C_n$. So when BS $n$ stays in idle state for a duration of $t$, the network utilization loss is $C_n \cdot t$, and when a channel of BS $n$ stays in idle duration of $t$, the network utilization loss is $C_n \cdot y_n \cdot t$, where $y_n$ is the ratio of a channel capacity to the total capacity $C_n$. For the network utilization loss, the idle duration $t$ of a channel is equivalent to an idle duration $y_n \cdot t$ of BS $n$. Thus we only consider the idle durations of BSs in the following by assuming that all the idle durations of channels have been converted into BS idle durations.

An example of idle state durations and active state durations for BS $n$ between two scheduling is shown in Figure 2. When the residual load equals 0, the BS $n$ transfers to idle state, and when new traffic streams arrive, the BS $n$ in idle state transfers to active state. The sum of idle durations in BS $n$ between the current scheduling and the next scheduling is denoted by $\mathcal{D}_n(r_n, x_n, T)$, where $r_n$ is the traffic load in BS $n$ before the current scheduling, $r_n = \sum_{k \in \mathcal{X}_k} r_{n,k} \cdot x_n = \sum_{m \in \mathcal{N}_n, k \in \mathcal{X}_k} x_n^{m,n-k} - \sum_{m \in \mathcal{N}_n, k \in \mathcal{X}_k} x_n^{m,n}$, and $T$ is the duration between the current scheduling and the next scheduling. The network utilization loss between the current scheduling and the next scheduling can be given as $h(r_{nk}, x_n^{k,n}, T) = \sum_{m \in \mathcal{N}_n} C_n \mathcal{D}_n(r_n, x_n, T)$. Therefore,
3. System Model

3.1. Networks Distribution. The HWNs may consist of a series of networks with different spectra, cell size levels, and radio transmission technics, such as cellular network, WLAN, WiMax, microcell, and femtocell. In this paper, a general scenario of HWNs is considered, as shown in Figure 1, that some networks may be regularly distributed while the other networks are distributed irregularly. The networks using different spectra can be overlapped, while networks using the same spectra cannot be overlapped and the border is determined by the UE-side received power. So any region \( k \in \mathcal{K} \) cannot be covered with two BSs using same spectra. The downlink capacity of each BS \( n \in \mathcal{N} \) (denoted by \( C_n \)) is assumed to be constant.

3.2. Traffic Model. The traffic arrival in each BS \( n \in \mathcal{N} \) from each region \( k \in \mathcal{K}_n \) is assumed to be time-heterogeneous Poisson arrival with arrival rate of \( \lambda_{n,k}(t) \), where \( t \) denotes the time, so the traffic arrival rate in each region \( k \) is \( \lambda_k(t) = \sum_{n \in \mathcal{N}_k} \lambda_{n,k}(t) \). Traffic data size are assumed to be negative exponential distributed with mean value \( f_d \), and its probability density function (PDF) is given by

\[
f(x) = \frac{1}{f_d} e^{-x/f_d}, \tag{4}
\]
3.3. Service Model. No bandwidth or channel reservation is considered for each network, and a process-sharing scheme in which all the admitted user equipments (UEs) equally share the available bandwidth is adopted by all the networks. That means all the bandwidth would be in use when the BS is in active state. Each UE is assumed to be equipped with multiple network interfaces so that can access all types of wireless networks. Multipaths technic [1] is considered for the UE that each UE can be admitted to one or more available networks simultaneously and the traffic in the UE can be divided into multiple substreams.

4. Predictive Method for Idle Duration

4.1. Idle Duration Division. Consider a BS having constant traffic arrival rate of \( \lambda \) at the period \( T \), that is the duration between the current scheduling and next scheduling. And the BS has a traffic load of \( L_0 \) after the current scheduling and has constant capacity of \( C \). The expected sum of idle durations (denoted by \( D \)), of the considered BS at the period \( T \) can be given as

\[
D = \sum_{i=0}^{\infty} D_i P_i, \tag{5}
\]

where \( D_i \) denotes the average idle duration when \( i \) traffic streams arrive in the period \( T \), and \( P_i \) denotes the probability that \( i \) traffic streams arrive in the period \( T \). According to the Poisson process, the expression of \( P_i \) can be given as

\[
P_i = \frac{(\lambda T)^i}{i!} \exp(-\lambda T). \tag{6}
\]

Indeed, the average idle duration when no traffic stream arrives in the period \( T \) can be given by

\[
D_0 = T - \frac{L_0}{C}. \tag{7}
\]

4.2. Idle Duration Expectation with One Traffic Stream. When one traffic stream arrives in the period \( T \) (\( i = 1 \)), the idle duration distribution in the period \( T \) is shown in Figure 3. In Figure 3, \( t_a \) denotes the traffic stream arrival time, \( a \) denotes the data size of the arrived traffic stream. When \( t_a \leq \frac{L_0}{C} \), the idle duration distribution is shown in Figure 3(a), based on which the expected idle duration \( (D_1) \) can be derived as

\[
D_1 \left( 0 \leq t_a \leq \frac{L_0}{C} \right) = \mathbb{E} \left[ T - \frac{L_0 + a}{C} \right] \bigg\vert 0 \leq t_a \leq \frac{L_0}{C} \bigg\] \[
= \int_0^{\frac{L_0}{C}} \int_0^{CT-L_a} \left( T - \frac{L_0 + a}{C} \right) \frac{1}{f_d} \exp \left( -a/f_d \right) \, da \, dt_a \]
\[
= \frac{L_0}{CT} \left( T - \frac{L_0}{C} \right) \frac{f_d L_0}{C} + \frac{f_d}{C} \exp \left( \frac{L_0 - C T}{f_d} \right), \tag{8}
\]

where \([x]^+\) equals \( x \) when \( x > 0 \), equals 0 otherwise, \( f(t_a) \) is the PDF of the arrival time \( t_a \) and is given by the following referred to [10]

\[
f(t_a) = \begin{cases} \frac{1}{T}, & 0 \leq t_a \leq T, \\ 0, & \text{else} \end{cases} \tag{9}
\]

Besides, when \( L_0/C < t_a \leq T \) holds, the idle duration distribution is shown Figure 3(b), based on which the idle duration expectation can be derived as

\[
D_1 \left( \frac{L_0}{C} < t_a \leq T \right) = \mathbb{E} \left[ \left( t_a - \frac{L_0}{C} \right) + \left[ T - t_a - \frac{a}{C} \right]^+ \right] \bigg\vert \frac{L_0}{C} < t_a \leq T \bigg\] \[
= \int_0^T \int_{\frac{L_0}{C}}^{CT-t_a} \left( \left( t_a - \frac{L_0}{C} \right) + \left( T - t_a - \frac{a}{C} \right) \right) \frac{1}{f_d} \exp \left( -a/f_d \right) \, da \, dt_a \]
\[
= \frac{1}{T} \left( \frac{2L_0 T}{C^2} + \frac{T^2}{C^2} + \frac{t_a^2 - f_d T}{C^2} + \frac{T^2}{C^2} + \frac{f_d^2}{C^2} \right) - \frac{f_d^2}{C^2} \exp \left( \frac{L_0 - C T}{f_d} \right). \tag{10}
\]

Based on the expressions of \( D_1(0 \leq t_a \leq L_0/C) \) and \( D_1(L_0/C < t_a \leq T) \), the idle duration expectation of \( D_1 \) is derived as

\[
D_1 = D_1 \left( 0 \leq t_a \leq \frac{L_0}{C} \right) + D_1 \left( \frac{L_0}{C} < t_a \leq T \right) \]
\[
= T - \frac{L_0}{C} \frac{f_d}{C} + \frac{f_d^2}{C^2 T} + \left( \frac{f_d}{C^2 T} \frac{L_0 - C T}{f_d} \right) \exp \left( \frac{L_0 - C T}{f_d} \right). \tag{11}
\]
Based on (7) and (11), the following iteration expression for $D_1$ could be obtained:

$D_1 = D_0 - \frac{f_d}{C} \left[ \frac{CT-f_d}{CT} + \frac{f_d-CT+CD_0}{CT} \exp\left(-\frac{CD_0}{f_d}\right) \right].$

(12)

4.3. The Two-Stage Predictive Method. Similar to the above analysis, when there are $i$ traffic streams arriving at the period $T$ and $i \geq 2$ holds, there are two cases of the last traffic stream's arrival time: the former $i-1$ traffic streams have finished service or have not finished service before the last traffic stream's arrival. Thus when there are $i$ traffic streams arriving during the period $T$ ($i \geq 0$), there are $2^i$ cases of idle duration distributions.

Consequently, it is practically impossible to derive the formulation of $D_i$ when $i$ is large. Indeed, we have found that the result obtained for $D_i$ increases rapidly with $i$. Nevertheless, based on the relationship of $D_i$ and $D_{i+1}$ and through approximation, a two-stage predictive method is given to approximate $D_i$ in

$$D_i = \begin{cases} D_{i-1}^{(1)}, & 1 \leq i < \frac{CT-L_0}{f_d} - 1, \\ D_{i-1}^{(2)}, & i \geq \frac{CT-L_0}{f_d} - 1, \end{cases}$$

(13)

where

$$D_{i-1}^{(1)} = D_1 - \frac{f_d}{C} \left[ \frac{CT-f_d}{CT} + \frac{f_d-CT+CD_i}{CT} \exp\left(-\frac{CD_i}{f_d}\right) \right],$$

$$D_{i-1}^{(2)} = D_1 - \frac{f_d}{C} \left[ \frac{CT-f_d}{CT} + \frac{f_d-CT+CD_i}{CT} \exp\left(-\frac{CD_i}{f_d}\right) \right] + \frac{L_0 + CD_i}{CT}.$$

(14)

The iteration expressions of $D_{i-1}^{(1)}$ and $D_{i-1}^{(2)}$ are derived through approximation in Appendix A, and the effectiveness of the two-stage iteration method has been validated through extensive simulation. In the simulation, we found that the effectiveness of the two-stage iteration method is mainly determined by the ratio of $f_d/CT$ and the method has a good performance when $f_d/CT < 0.1$ holds. Some simulation results are given as Figures 4 and 5. In the two figures, the $Y$ axis represents the normalized idle duration (i.e., $D/T$), the estimate value of idle-duration is calculated by (5) and (13), and the real value of idle duration is obtained by simulating the traffic arrival process and the traffic serving process. Figure 4 shows the averaged normalized idle-durations with the variation of the arrived traffic numbers in the period $T$, and Figure 5 shows the averaged normalized idle-durations with the variation of the $\lambda \cdot T$. For Figure 5, the arrived traffic number in the period $T$ is determined by the Poisson with arrival rate $\lambda$. From the two figures, it can be seen that the effectiveness of the predictive method is almost determined by the ratio of $f_d/CT$, and the estimate value by the two-stage iteration method matches well with the real value when $f_d/CT \leq 0.1$. Indeed, the condition of $f_d/CT \leq 0.1$ can be fulfilled by the current and future wireless networks like 4G cellular network, WiMAX and WLAN, and so forth.

5. Method for Optimal Solution

With the idle duration of $D(R, x, T)$ approximated by (5) and (13), the load scheduling problem (3) would be proved to be a convex optimization problem by proving $\frac{\partial^2 h(R, x, T)}{\partial x_k^{n-m} \partial x_k^{m-n}} \geq 0$ for all $n, m \in \mathcal{N}$, and $k \in \mathcal{K}$ in the following. Then the gradient expression of $\frac{\partial h(R, x, T)}{\partial x_k^{n-m}}$ is derived to obtain the gradient descent method for optimal solution.

According to the expression of $h(R, x, T)$ and (5) and (13), $\frac{\partial^2 h(R, x, T)}{\partial x_k^{n-m} \partial x_k^{m-n}}$ is derived as

$$\frac{\partial^2 h(R, x, T)}{\partial x_k^{n-m} \partial x_k^{m-n}} = \frac{\partial^2}{\partial x_k^{n-m} \partial x_k^{m-n}} \left[ \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \frac{\partial^2 h(R, x, T)}{\partial x_k^{n-m} \partial x_k^{m-n}} \right].$$

(15)

Thus, $\frac{\partial^2 h(R, x, T)}{\partial x_k^{n-m} \partial x_k^{m-n}} \geq 0$ can be proved by proving $\frac{\partial^2 D_i}{\partial T^2} \geq 0$. According to (13), $\frac{\partial^2 D_i}{\partial T^2} \geq 0$ can be proved by proving $\frac{\partial^2 D_i}{\partial T^2} \geq 0$ and $\frac{\partial^2 D_i}{\partial T^2} \geq 0$. For the clarity of proof, define $f(D_i)$ and $g(D_i)$ as follows:

$$f(D_i) = \frac{f_d}{C} \left[ \frac{CT-f_d}{CT} + \frac{f_d-CT+CD_i}{CT} \exp\left(-\frac{CD_i}{f_d}\right) \right],$$

$$g(D_i) = \frac{CD_i + L_0}{CT}.$$

(16)
Based on (A.6) and (A.8), the following relationships hold: \( D_{i+1}^{(1)} = D_i - f(D_i) \) and \( D_{i+1}^{(1)} = D_i - f(D_i)g(D_i) \). So, when \( d^2D_i^{(1)}/dL_0^2 \geq 0 \), the following relationships can be derived:

\[
\frac{d^2D_i^{(1)}}{dL_0^2} = \frac{d^2D_i^{(1)}}{dL_0^2} - \frac{d^2f(D_i^{(1)})}{dL_0^2} = \frac{d^2D_i^{(1)}}{dL_0^2} - \frac{d^2D_i^{(1)}}{dL_0^2} \left( 1 \frac{D_i^{(1)}}{T} \right) \exp \left( -\frac{CD_i^{(1)}}{f_d} \right)
\]

\[
\geq \frac{d^2D_i^{(1)}}{dL_0^2} \frac{D_i^{(1)}}{T}
\]

Hence, \( d^2D_i^{(1)}/dL_0^2 \geq 0 \) only when \( d^2D_i^{(1)}/dL_0^2 \geq 0 \) holds. Since \( d^2D_i^{(1)}/dL_0^2 = 0 \) holds, \( d^2D_i^{(1)}/dL_0^2 \geq 0 \) is proved.
The expectation of normalized idle duration

\[
\lambda^* T
\]

\[
L_0/(CT) = 0.01
\]

\[
L_0/(CT) = 0.3
\]

\[
L_0/(CT) = 0.8
\]

Figure 5: Estimate and real value of the average \(D/T\) with variations of \(f_d/CT\), \(L_0/CT\), and \(\lambda T\).

When \(d^2 D^{(2)}_i/dL^2_0 \geq 0\) holds, \(d^2 D^{(2)}_i/dL^2_0 \geq 0\) can be proved as the following equation based on (22), (23), and Lemma 2 in Appendix B:

\[
\frac{d^2 D^{(2)}_i}{dL^2_0} \geq \frac{d^2 f(D^{(2)}_i)}{dL^2_0} g(D^{(2)}_i) - f(D^{(2)}_i) \frac{d^2 g(D^{(2)}_i)}{dL^2_0} \geq 0.
\]

Therefore, \(\partial^2 h(r_{nk}, |x_{k}^{m-m}|, T)/\partial x_{k}^{m} \geq 0\) is proved, so the load scheduling problem (3) is a convex optimization problem when it adopts the two-stage predictive method. There are many methods proposed in the literature such as descent method, gradient descent method, and Newton's
(1) Given a starting point of $x_{k}^{n-m} \in [0, r_{k}]$ for all $n, m \in N$ and $k \in K$.
(2) Calculated $\partial h (r_{nk}, \{x_{k}^{n-m}\}, T) / \partial x_{k}^{n-m}$ for each $(n, m, k)$ by (19)–(21).
(3) Initialized $j = 0$, step size $\delta_{nk}^{n-m}$ for each $(n, m, k)$.
(4) while existing $(n, m, k)$ satisfy $|\partial h (r_{nk}, \{x_{k}^{n-m}\}, T) / \partial x_{k}^{n-m}| > \epsilon$ do
(5) Calculated $x_{k,j}^{n-m} = x_{k,j}^{n-m} - \delta_{nk}^{n-m}(\partial h (r_{nk}, \{x_{k}^{n-m}\}, T) / \partial x_{k}^{n-m})$.
(6) Updated $j = j + 1$, and step size $\delta_{nk}^{n-m}$.
(7) Calculated $\partial h (r_{nk}, \{x_{k}^{n-m}\}, T) / \partial x_{k,j}^{n-m}$ for each $(n, m, k)$ by (19)–(21).
(8) end while
(9) Obtain the optimal solution $x_{k,j}^{n-m} = x_{k,j}^{n-m}$.

**Algorithm 1: Gradient descent method for the optimization problem (3).**

The method [11] for convex optimization problem. Here, we give a gradient descent method to solve problem (3) and derive the gradient expression of $\partial h (r_{nk}, \{x_{k}^{n-m}\}, T) / \partial x_{k}^{n-m}$ in the following:

$$
\frac{\partial h (r_{nk}, \{x_{k}^{n-m}\}, T)}{\partial x_{k}^{n-m}} = \sum_{n \in N, \forall} \frac{C_{n}}{\sum_{i=0}^{\infty} P_{i}} \frac{\partial D_{1} (L_{0} = r_{n1} + x_{n2})}{\partial L_{0}} \left( \frac{\partial x_{n1}}{\partial (x_{k}^{n-m})} \right),
$$

where

$$
\frac{\partial x_{n1}}{\partial (x_{k}^{n-m})} = \begin{cases} 
1, & k \in M, k \in M, m = n1 \\
-1, & k \in M, k \in M, n = n1 \\
0, & \text{else}.
\end{cases}
$$

And $\partial D_{1} (L_{0} = r_{n1} + x_{n2}) / \partial L_{0}$ is given by

$$
\frac{dD_{i+1}}{dL_{0}} = \begin{cases} 
\frac{dD_{i}}{dL_{0}} - \frac{df (D_{i})}{dL_{0}}, & k < C_{n1}T - L_{0} - 1 \\
\frac{dD_{i}}{dL_{0}} - \frac{df (D_{i})}{dL_{0}} - g (D_{i}) - f (D_{i}) \frac{dg (D_{i})}{dL_{0}}, & \text{else},
\end{cases}
$$

where

$$
\frac{df (D_{i})}{dL_{0}} = \frac{dD_{i}}{dL_{0}} \left( 1 - \frac{D_{i}}{T} \right) \exp \left( - \frac{C_{n1}D_{i}}{f_{d}} \right),
$$

$$
\frac{dg (D_{i})}{dL_{0}} = \frac{1}{T} \frac{dD_{i}}{dL_{0}} + \frac{1}{C_{n1}T}, \quad \frac{dD_{0}}{dL_{0}} = \frac{1}{C_{n1}}.
$$

And according to Lemma 2 in Appendix B, the following relationships can be derived:

$$
\frac{df (D_{i})}{dL_{0}} \leq 0, \quad \frac{dg (D_{i})}{dL_{0}} \geq 0.
$$

All the results above can be combined to obtain the gradient descent method [11] for problem (3) as Algorithm 1. By implementing the iteration process in the gradient descent method, the optimal solution of scheduled loads ($x_{k}^{n-m}$, $n, m \in N$, $k \in K$) can be obtained.

### 6. Simulation Results and Discussion

In this section, the effectiveness of the proposed load scheduling scheme is validated, the performance of the proposed scheme and other schemes are compared, and the impacts of system parameters are evaluated.

In the simulation, the considered scenario is shown in Figure 6: three BSs of $M = \{0, 1, 2\}$ and three regions of $K = \{0, 1, 2\}$.

#### 6.1. Validation of the Proposed Load Scheduling Scheme

In the literature, a load balance index based method has been extensively investigated [12–14], and the general expression of the balance index is given by

$$
B (y) = \frac{\left( \sum y_{i} \right)^{2}}{\left( \sum y_{i}^{2} \right)},
$$

(24)

The reference load scheduling scheme, denoted by "ref" in the simulation results, is obtained by maximizing $B (y)$ with $y_{i} = (r_{i} + x_{i}) / C_{i}$. In the simulation results, the scheme without load scheduling is denoted by non, the proposed load scheduling scheme is denoted by eff, and the throughput-optimal load scheduling scheme is denoted by opt. The opt scheme obtain its optimal scheduling loads by exhaustive search and trials.

In the simulation for effectiveness validation of the proposed scheme, the sums of capacity-weighted idle durations are collected by fixing the residual load of $r_{nk}$ ($n \in N$, $k \in K$) before each scheduling, and the results are averaged by 10000 trials. The simulation parameters are configured as $r_{0,1} = 15$ Mbits, $r_{0,2} = 10$ Mbits, $r_{1,1} = 5$ Mbits, $r_{2,2} = 10$ Mbits, $C_{0} = 20$ Mbps, $C_{1} = 20$ Mbps, $C_{2} = 20$ Mbps, $\lambda_{0,0} = 10$ calls/s, $\lambda_{0,1} = 2.4$ calls/s, $\lambda_{1,1} = 9.6$ calls/s, $\lambda_{0,2} = 2.4$ calls/s, $\lambda_{1,2} = 2.4$ calls/s, $\lambda_{2,2} = 9.6$ calls/s, $f_{d} = 1$ Mbits, and the maximum acceptable UE number is 40 for each BS.

Some results of the four scheduling schemes are given by Figures 7 and 8, where the normalized utilization loss is calculated by $h (r_{nk}, \{x_{k}^{n-m}\}, T) / (C_{0} + C_{1} + C_{2})T$. Figure 7 shows...
the normalized utilization loss values of the four schemes with the variation of the initial residual load in BS 0 from region 0 (i.e., \( r_{0,0} \)), and Figure 8 shows the normalized utilization loss values with the variation of the scheduling period \( T \). In the two figures, the throughput-optimal load scheduling scheme (denoted by “opt”) always has the best performance and the scheme without load scheduling always has the worst performance. The performance of the proposed scheme is very close to the performance of the throughput-optimal scheme, and the performance gap is smaller when the initial residual load \( r_{0,0} \) is larger or when the scheduling period \( T \) is smaller. In Figure 7, the normalized utilization loss values of the four schemes decrease with the initial residual load \( r_{0,0} \) because larger initial residual load means smaller idle-duration expectation. In Figure 8, the normalized utilization loss values of the four schemes increases with the period \( T \) because the total capacity \( \sum_{n \in N} C_n \) is larger than the sum of arriving traffic loads \( \sum_{n \in N, k \in K} \lambda_{nk} \int_0^T d \). Furthermore, similar results have been obtained with variations of initial residual loads in the other two BSs and traffic arrival rates in the other two regions, but they are not given here due to space limitation. All the results show that the proposed scheduling scheme can closely approach the optimal solution.
6.2. Throughput and Delay Improvement. We evaluate the performance of the proposed load scheduling scheme with the periodically changed traffic arrival rates as Figure 9, and the common fluctuation cycle of traffic arrival rates is 300 s. And in Figure 9, $\lambda_i (i = 0, 1, 2)$ denotes the traffic arrival rates in region $i$, $\lambda_{iH}$ and $\lambda_{iL}$ ($i = 1, 2$) denote the peak value and minimal value of arrival rates in region $i$, and $\lambda_{iH} = 2\lambda_{iL}$. We collect results by running the emulator of load serving process and traffic arrival process for 10000 seconds for each trial and average the results through 100 trials.

Figures 10 and 11 show the blocking probabilities and delays of the three load scheduling schemes when traffic load is light, respectively. The blocking probability is the probability that a new arrival UE is blocked, and the blocking happens when the numbers of admitted UEs in all the available BSs reach their upper limits. The traffic arrival rates are configured as $\lambda_0 = 10$ calls/s, $\lambda_{1H} = \lambda_{2H} = 5$ calls/s, $\lambda_{0,0} = 10$ calls/s, $\lambda_{0,1} = 0.3\lambda_1$, $\lambda_{0,2} = 0.7\lambda_2$, $\lambda_{1,1} = 0.7\lambda_1$, $\lambda_{2,2} = 0.3\lambda_2$, and other parameters are the same as those in the former subsection.

As shown in Figures 10 and 11, the proposed load scheduling scheme always has the best performance of decreasing average delays and blocking probabilities. Compared with the “non” scheme, the proposed load scheduling scheme can reduce almost 20% blocking probabilities and 30% delays when the scheduling period $T$ is small. In the two figures, the blocking probabilities and average delays of the proposed scheme and the reference scheme increase with the scheduling period, because the scheduling frequency decreases and the gain brought by scheduling decreases accordingly when $T$ increases.

Figures 12 and 13 show the performance of these schemes when traffic load is relatively heavy, which is configured as $\lambda_0 = 10$ calls/s and $\lambda_{1H} = \lambda_{2H} = 9$ calls/s. As shown in Figures 12 and 13, the proposed load scheduling scheme can almost reduce 60% blocking probability and reduce about 50% delay compared with the scheme without load scheduling when the scheduling period $T$ is small. The proposed load scheduling scheme always has the best performance. Comparing these results with the results when the traffic load is light, we can see that the proposed load scheduling scheme has better performance improvements of blocking probability and average delay when the traffic load is heavier.
Eff
Scheduling period $T$ (/s)
4
7
10
13
16
19
22
25
28
31 ...
̂$D_{(u)}$
$i+1$ and is given by
̂$D_{(u)}$
$i+1$ = $E \left[ [t_{i+1} - L_0]_C \right] + [L_0 + \sum_{j=1}^{i} a_j]_C + [T - a_j - (CT - L_0 - if_d)/C].$ (A.4)

with a variety of system parameters, the effectiveness and is formulated, a objective of minimizing the capacity-weighted idle durations scheme, a fairly general HWNs scenario is considered, the idle durations of networks and channels. For the proposed posed from the point of decreasing the capacity-weighted performance of the proposed load scheduling scheme is demonstrated, and the impacts of system parameters are evaluated. This work may shed some light for future extension and study.

Appendices

A. Derivation of the Two-Stage Predictive Method

Lemma 1. The idle duration expectation of $D_i$ satisfies

$$D_i > \frac{CT - L_0 - if_d}{C}.$$ (A.1)

Proof. Denote $E_i$ as the idle duration expectation when all the $i$ traffic streams arrive at time 0, and $E_i$ is given by

$$E_i = E \left[ [T - \sum_{j=1}^{i} a_j]_C \right] = \frac{CT - L_0 + if_d - if_d \exp \left(-\frac{(CT - L_0 - if_d)}{if_d}\right)}{C},$$ (A.2)

where $a_j$ is the data size of the $j$th traffic stream.

Thus $D_i > E_i > (CT - L_0 - if_d)/C$ holds.

When there have been $i$ arrived traffic streams in the period $T$, the period $T$ would be divided into a series of idle durations and active durations, as shown in the top of Figure 14. In Figure 14, $[t^n_m]$ ($n \in \{1, \ldots, N\}$) denotes the idle durations set, $[t^m_a]$ ($m \in \{1, \ldots, M\}$) denotes the active durations set without containing the active duration of $[0, L_0/C]$. So $D_i$ is the sum of all the idle durations. When the $(i+1)$th traffic stream arrives in the period $T$ (primary period $T$), some idle durations may disappear and $D_{i+1}$ is the sum of the residual idle durations.

The recreated periods of $T_{r,1}$ and $T_{r,2}$ are created by the following steps. Firstly, aggregate the idle durations ($[t^n_m]$) together as a unified idle duration, and aggregate the active durations ($[t^m_a]$) together as a unified active duration. Then if put the unified idle duration to the end of the unified active duration, the $T_{r,1}$ is created, as shown in the middle of Figure 14. And if put the unified active duration to the end of the unified active duration, the $T_{r,2}$ is created, as shown in Figure 14.

So, if we calculate $D_{i+1}$ based on the recreated period $T$ of $T_{r,1}$, the result is denoted by $\bar{D}_{i+1}^{(1)}$ and is given by (A.3), where $t_{i+1}$ is the arrival time of the $(i+1)$th traffic stream. Consider

$$\bar{D}_{i+1}^{(1)} = E \left[ (t_{i+1} - (T - D_i))_C \right] + (T - a_{i+1})_C.$$ (A.3)

If we calculate $D_{i+1}$ based on the recreated period $T$ of $T_{r,2}$, the result is denoted by $\bar{D}_{i+1}^{(2)}$ and is given by

$$\bar{D}_{i+1}^{(2)} = E \left[ (t_{i+1} - L_0)_C \right] + \left[ \frac{L_0}{C} + D_i - t_{i+1} - \frac{a_{i+1}}{C} \right]_C.$$ (A.4)

7. Conclusion

In this paper, a predictive load scheduling scheme is proposed from the point of decreasing the capacity-weighted idle durations of networks and channels. For the proposed scheme, a fairly general HWNs scenario is considered, the objective of minimizing the capacity-weighted idle durations is formulated, a two-stage predictive method is given to predict the idle duration, and a gradient descent method is given for the optimal scheduled loads. By simulation, the effectiveness of the two-stage predictive method is validated with a variety of system parameters, the effectiveness and

Figure 12: Blocking probability of traffic when load is relatively heavy.

Figure 13: Average delay of traffic when load is relatively heavy.
Thus, \( D_{i+1}^{(1)} \leq D_{i+1} \leq D_{i+1}^{(u)} \) holds. So when \( CT - L_0 - if_d \gg f_d \) holds, \( D_i \gg f_d/C \) can be obtained due to \( D_i > (CT - L_0 - if_d)/C \), and the following relationship can be obtained:

\[
D_{i+1} - D_{i+1}^{(1)} \leq D_{i+1}^{(u)} - D_{i+1}^{(1)} = \frac{D_i C F_d + f_d^2}{C^2 T} \exp \left( -\frac{D_i C}{f_d} \right) = 0. \tag{A.5}
\]

Therefore, when the condition of \( CT - L_0 - if_d \gg f_d \) holds, \( D_{i+1} \) can be approximated by \( D_{i+1}^{(1)} \); that is, given by

\[
D_{i+1} = D_{i+1}^{(1)} = D_i - \frac{f_d}{C} \left[ \frac{CT - f_d}{CT} + \frac{f_d - CT + CD_i}{CT} \exp \left( -\frac{CD_i}{f_d} \right) \right]. \tag{A.6}
\]

On the other hand, when the condition of \( (L_0 + if_d)/CT \gg 1 \) holds, \( D_i \gg f_d/C \) can be obtained. Thus, when \( t_{i+1} > L_0/C \) and the BS is on active state at the time of \( t_{i+1} \), the probability that the served load of the \((i + 1)\)th traffic stream equals zero in the period \( T \) is given by

\[
\mathbb{P}\left[ L_s = 0 \right] > \mathbb{P}\left( R_{all} > D_i C \right) \times \mathbb{P}\left( R_m > D_i C \right) = \exp \left( -\frac{D_i C}{f_d} \right) = 1, \tag{A.7}
\]

where \( L_s \) is the served load of the \((i + 1)\)th traffic stream in the period \( T \), \( R_{all} \) is the total residual load of the BS at time \( t_{i+1} \), \( R_m \) is the residual load of the \(m\)th traffic stream, here the \(m\)th traffic stream is the traffic stream with the arrival time immediately before the \((i + 1)\)th traffic stream, \( R_m \) conforms to negative-exponential distribution due to the \(a_m \) (the data size of the \(m\)th traffic stream) being negative-exponential distribution with memoryless property.

So \( D_{i+1} \) decreases only when \( t_{i+1} < L_0/C \) or the BS is in idle duration at \( t_{i+1} \), and the probability is \((L_0 + CD_i)/CT\).

Therefore, when the condition of \((L_0 + if_d)/CT \gg 1 \) holds, \( D_{i+1} \) can be approximated as \( D_{i+1}^{(1)} \) given by

\[
D_{i+1} = \frac{f_d}{C} \left[ \frac{CT - f_d}{CT} + \frac{f_d - CT + CD_i}{CT} \exp \left( -\frac{CD_i}{f_d} \right) \right]. \tag{A.8}
\]

**B. Boundaries of \( dD_i/dL_0 \)**

**Lemma 2.** For any \( i \geq 0 \), \( -1/C \leq d\bar{D}_i/dL_0 < 0 \) holds, where \( \bar{D}_i \) is the approximation of \( D_i \) according to (13).

**Proof.** An inductive method is adopted to prove this lemma. Firstly, \( d\bar{D}_0/dL_0 = -1/C \), the lemma is true when \( i = 0 \).

Secondly, assume that \( -1/C \leq d\bar{D}_i/dL_0 < 0 \) holds, then the following relationships can be derived:

\[
\frac{df(\bar{D}_i)}{dL_0} = \frac{d\bar{D}_i}{dL_0} \left( 1 - \frac{\bar{D}_i}{T} \right) \exp \left( -\frac{\bar{D}_i}{f_d} \right), \tag{B.1}
\]

\[
\frac{dg(\bar{D}_i)}{dL_0} = \frac{d\bar{D}_i}{dL_0} \left( 1 + \frac{\bar{D}_i}{T} \right) - \frac{1}{CT},
\]

so

\[
\frac{d\bar{D}_i}{dL_0} \left( 1 - \frac{\bar{D}_i}{T} \right) < \frac{df(\bar{D}_i)}{dL_0} < 1, \tag{B.2}
\]

\[
0 < f(\bar{D}_i) < \bar{D}_i, \tag{B.3}
\]

\[
0 < \frac{dg(\bar{D}_i)}{dL_0} < \frac{1}{CT},
\]

\[
\frac{d\bar{D}_i}{dL_0} \left( 1 - \frac{\bar{D}_i}{T} \right) < \frac{df(\bar{D}_i)}{dL_0} < 0.
\]
Then, the following equations can be derived:
\[
\frac{d\hat{D}_{i+1}}{dL_0} = \frac{d\hat{D}_i}{dL_0} - \frac{df(\hat{D}_i)}{dL_0} < \frac{d\hat{D}_i}{dL_0} \frac{1}{T} < 0,
\]
\[
\frac{d\hat{D}_{i+1}}{dL_0} = \frac{d\hat{D}_i}{dL_0} - \frac{df(\hat{D}_i)}{dL_0} > \frac{d\hat{D}_i}{dL_0} \geq -\frac{1}{C},
\]
\[
\frac{d\hat{D}_{i+1}}{dL_0} = \frac{d\hat{D}_i}{dL_0} - \frac{df(\hat{D}_i)}{dL_0} \theta(\hat{D}_i) - f(\hat{D}_i) \frac{dg(\hat{D}_i)}{dL_0}
\]
\[
< \frac{d\hat{D}_i}{dL_0} \frac{d\hat{D}_i}{dL_0} \left(1 - \frac{d\hat{D}_i}{dL_0} \frac{1}{T}\right) < 0,
\]
\[
\frac{d\hat{D}_{i+1}}{dL_0} > \frac{d\hat{D}_i}{dL_0} \left(\frac{1}{CT} + \frac{1}{T} \frac{d\hat{D}_i}{dL_0}\right)
\]
\[
> \frac{d\hat{D}_i}{dL_0} - T \left(\frac{1}{CT} + \frac{1}{T} \frac{d\hat{D}_i}{dL_0}\right) = -\frac{1}{C}.
\]
Thus, the lemma is true for \(\hat{D}_{i+1}\).

Thirdly, based on the above two steps, \(-1/C \leq d\hat{D}_i/dL_0 < 0, \forall i \geq 0\) is proved. \(\square\)

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**Acknowledgments**

The authors wish to thank the anonymous reviewers for their hard work and constructive comments. This paper is supported by National Programs for High Technology Research and Development (2014AA01A707) and Natural Science Foundation of China under Grant (6146136002).

**References**

[1] J. R. Iyengar, P. D. Amer, and R. Stewart, “Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths,” *IEEE/ACM Transactions on Networking*, vol. 14, no. 5, pp. 951–964, 2006.

[2] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, “Analytical model of the single threshold mechanism with hysteresis for multi-service networks,” *IEICE Transactions on Communications*, vol. 95, no. 1, pp. 120–132, 2012.

[3] S. Jie, Y. Zheng, T. Liang-Rui, and H. Jian-Hong, “A novel load balancing algorithm based on utility functions and fuzzy logic in heterogeneous wireless networks,” in *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD’12)*, pp. 414–418, IEEE, May 2012.

[4] R. Nasri, A. Samhat, and Z. Altman, “A new approach of umts-wlan load balancing: algorithm and its dynamic optimization,” in *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM ’07)*, pp. 1–6, Espoo, Finland, June 2007.

[5] B. Fang, W. Zhang, S. Zhang, and W. Zhou, “Distributed algorithm for ergodic global network utility maximization based-on joint RRA and VHD,” in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC ’13)*, pp. 1440–1445, April 2013.

[6] H. Son, S. Lee, S.-C. Kim, and Y.-S. Shin, “Soft load balancing over heterogeneous wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2632–2638, 2008.

[7] B. Li, W. Shi, and Y. Zhao, “A load balancing algorithm based on dividing IP flow for high-speed traffic over heterogeneous wireless networks,” in *Proceedings of the 3rd International Congress on Image and Signal Processing (CISP ’10)*, pp. 4294–4298, October 2010.

[8] N. Wang, W. Shi, S. Fan, and Y. Liu, “Flow diversion-based vertical handoff algorithm for heterogeneous wireless networks,” *Journal of Information and Computational Science*, vol. 7, no. 13, pp. 4863–4870, 2011.

[9] J. J. Liu, G. Wei, and Y. G. Wang, “Game-theoretic rate allocation with balanced traffic in collaborative transmission over heterogeneous wireless access networks,” *IET Communications*, vol. 6, no. 10, pp. 1245–1251, 2012.

[10] G. Gross, *Stochastic Processes*, 1996.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[12] I. Jabri, N. Krommenacker, T. Divoux, and A. Soudani, “IEEE 802.11 load balancing: an approach for qos enhancement,” *International Journal of Wireless Information Networks*, vol. 15, no. 1, pp. 16–30, 2008.

[13] Q.-T. Nguyen-Vuong and N. Agoulmine, “Efficient load balancing algorithm over heterogeneous wireless packet networks,” *IEEE Journal on Electronics and Communications*, vol. 1, no. 1, pp. 53–61, 2011.

[14] G. Bianchi and I. Tinnirello, “Improving load balancing mechanisms in wireless packet networks,” in *Proceedings of the International Conference on Communications (ICC ’02)*, vol. 2, pp. 891–895, May 2002.
