Wireless Networks 9, 103–116, 2003

On Bandwidth Allocation for Data Dissemination in Cellular Mobile Networks

JIANLIANG XU *, DIK L. LEE and BO LI
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Abstract. Wireless bandwidth is a scarce resource in a cellular mobile network. As such, it is important to effectively allocate bandwidth to each cell such that the overall system performance is optimized. Channel allocation strategies have been extensively studied for voice communications in cellular networks. However, for data dissemination applications, studies on bandwidth allocation have thus far been limited to a single-cell environment. This paper investigates the problem of bandwidth allocation for data dissemination in a multi-cell environment, which, to the best of our knowledge, has not been investigated before. The performance objective is to minimize the overall expected access latency given the workload for each cell in a data dissemination system. Two heuristic techniques, called *compact allocation* and *cluster-step allocation*, are proposed to effectively allocate bandwidth for a cellular network. Simulation experiments are conducted to evaluate the performance of the proposed bandwidth allocation schemes. Experimental results show that the proposed schemes substantially outperform the *uniform allocation* and *proportional allocation* schemes.

Keywords: bandwidth allocation, access latency, data dissemination, cellular network, performance analysis

1. Introduction

With continuous advances in wireless communication technologies, we have witnessed the rapid growth of mobile data applications in the commercial market. Examples of mobile data services include local news, entertainments, weather and traffic information, to name but a few. However, a mobile computing environment has many constraints, such as limited wireless bandwidth and battery power [1,8]. Thus, effective data management and resource management techniques are vital to the success of emerging mobile data applications.

There are two fundamental delivery methods for mobile data dissemination: *on-demand access* and *data broadcast* [1,9,11]. In on-demand access, a client sends data requests uplink to the server, and the server returns the results to the client individually. In data broadcast, the server periodically broadcasts information to the entire client population, and the clients monitor the broadcast channel to retrieve the data they are interested in. On-demand access provides fast response for a light-load system but the performance will deteriorate rapidly as workload increases. On the contrary, data broadcast can scale up to a very large client population and fit well to an asymmetric communication environment.

Because wireless bandwidth is a scarce resource in a cellular mobile network, much work has been done on bandwidth allocation for data dissemination in order to achieve a better access performance [1,7,9,11]. For data broadcast, [1,7] proposed methods to assign more bandwidth to frequently accessed items and less to infrequently accessed items such that the overall access latency is optimized. The studies in [2,9] investigated hybrid systems which combine on-demand and broadcast methods so that they complement each other. The bandwidth allocation between on-demand access and data broadcast was analyzed in [9,11]. Unfortunately, these studies are confined to single-cell environments.

In a multi-cell environment, such as a cellular network, bandwidth allocation becomes much more complex than that of a single-cell environment. This is because neighboring cells in a cellular network are not allowed to use the same frequencies for communication simultaneously because of signal interference [14,20]. As such, the available bandwidth for a system must be shared by a group of neighboring cells. On the other hand, if two cells are sufficiently apart, the same frequency can be *reused* in these two cells. As a result, bandwidth allocation cannot be considered on a cell-by-cell basis. Furthermore, the workloads for different cells may differ greatly. Thus, the allocation of frequencies/bandwidth to each cell so as to balance the workload among the cells is a challenging task. Although channel allocation for multi-cell voice communications has been extensively explored (e.g., [10,12,14,17,20]), these studies aim at minimizing call blocking/dropping probabilities or improving the volume of the carried traffic while ensuring certain level of QoS requirement. On the other hand, the characteristics of mobile data applications, which are mostly concerned with access latency, are not considered at all.

This paper investigates the problem of wireless bandwidth allocation for data dissemination in a multi-cell environment. The objective is to obtain the optimal bandwidth allocation scheme which minimizes the overall data access latency for a mobile data dissemination system. We start by formulating the bandwidth allocation problem under a regular cellular model. We then study how to optimally allocate bandwidth in a cell cluster, within which frequencies are not reused. Based on that, two heuristic allocation schemes, called *compact allocation* and *cluster-step allocation*, are proposed for a cellu-

* Corresponding author.

lar network with frequency reuse. The proposed schemes are evaluated by a series of simulation experiments under various system configurations for a three-cell cluster and a 7×7 cellular model. Experiment results show that the proposed schemes substantially outperform the *uniform allocation* and *proportional allocation* schemes.

The rest of the paper is organized as follows. Section 2 provides a motivating example for bandwidth allocation among different cells. In section 3, the bandwidth allocation problem is formulated under a cellular network model. Section 4 describes the proposed bandwidth allocation techniques. Experiment results are presented in section 5. Section 6 reviews related work. Finally, the paper is concluded in section 7.

2. A motivating example

In this section, a simple example is presented to motivate the study on bandwidth allocation for data dissemination in a multi-cell environment. As mentioned before, the performance metric adopted in this study is *access latency,* which is defined as the time elapsed from the moment when a request is submitted to the time when the request is serviced. In the simple example, we assume that the cellular network consists of two cells A and B, each of which is associated with a database containing four data items having the same size of 1 Kb. In both cells, the four data items are broadcast on air based on a *flat broadcast* program, i.e., in a round-robin manner. A mobile client monitors the broadcast in its current cell to retrieve the data of its interest. Assume that the aggregate data access rates are 1 and 4 for cell A and cell B, respectively and that a total wireless bandwidth of 6 Kbps is shared by these two cells. Table 1 shows three different allocation schemes and their corresponding expected access latencies (the calculation of the expected latencies will be described in detail in section 3).

The first method allocates the bandwidth uniformly. In the second method, the bandwidth allocated to each cell is linearly proportional to its aggregate access rate. From table 1, neither of these two solutions yields the best overall access performance. The best overall performance is achieved by the third method which allocates 2 Kbps to cell A and 4 Kbps to cell B. An intuitive explanation for this phenomenon is as follows. For non-uniform traffic loads, a heavy-load cell has more impact on the overall performance than a light-load cell and hence allocating more bandwidth to the heavy-load cell B can improve the overall performance. However, if the bandwidth is over-allocated to cell B, too little bandwidth is left to the light-load cell A. As such, cell A has a very poor latency and thus leads to a worse overall performance. This observation motivates us to quantitatively analyze the impact of skewed bandwidth allocation and find a better strategy in terms of overall access performance.

3. Problem formulation

This section describes a general cellular network model and formulates the bandwidth allocation problem under the model. In this study, we consider a regular cellular network, as adopted in the previous work [4,14,19,20]. The area covered by the cellular network consists of a set of hexagonal cells, each of which has a radius of R. The bandwidth available for the data dissemination system covers certain frequency range of the spectrum. If a certain frequency range is assigned to cell i, it cannot be used in cell i's neighboring cells to avoid interference. The minimum distance at which frequencies can be reused with acceptable interference is called *minimum reuse distance*. For example, figures 1(a) and 1(b) show two cellular networks of minimum reuse distances ($D_{\min}$) of $3R$ and $\sqrt{21}R$, respectively. Cells which can use the same frequency are shown using the same filling pattern. It is assumed that data access requests follow a Poisson process for each individual cell. For data dissemination, each cell employs the data broadcast or the on-demand access method [1,7]. For on-demand access, the uplink cost is ignored because it is generally very small and will not affect the performance results. It is also assumed that the bandwidth allocated to a cell can be aggregated to serve data broadcast or on-demand access [9,11]. To facilitate further discussion, we define some notations in table 2.

Without loss of generality, assume that cells $1, 2, \ldots, N_{\mathrm{b}}$ ($0 < N_{\mathrm{b}} \leqslant N$) use the broadcast method and cells $N_{\mathrm{b}} + 1, \ldots, N$ ($0 \leqslant N_{\mathrm{b}} < N$) employ the on-demand access method. To formulate the problem, we start by deriving the formulae of data access performance for a single cell. Let's first consider cell i where the broadcast method is employed. Suppose that a total of b_i bandwidth is assigned to cell i, and instances of $item_{i,j}$ are spaced with $s_{i,j}$ in the broadcast program. Thus, $item_{i,j}$'s broadcast frequency is $b_i/s_{i,j}$. The expected access latency for $item_{i,j}$ is given by:

$$\frac{s_{i,j}}{2b_i} + \frac{l_{i,j}}{b_i}, \quad 1 \leqslant j \leqslant M_i,\ 1 \leqslant i \leqslant N_{\mathrm{b}},$$

where the first term is the average waiting time and the second is the service time.

Table 1
An example of bandwidth allocation in a two-cell system.

	Bandwidth allocation (Kbps)		Expected access latency (s)		Overall expected access latency (s)
	Cell A	Cell B	Cell A	Cell B	
Uniform	3.0	3.0	1.0	1.0	1.0
Proportional	1.2	4.8	2.5	0.625	1.0
Best	2.0	4.0	1.5	0.75	0.9

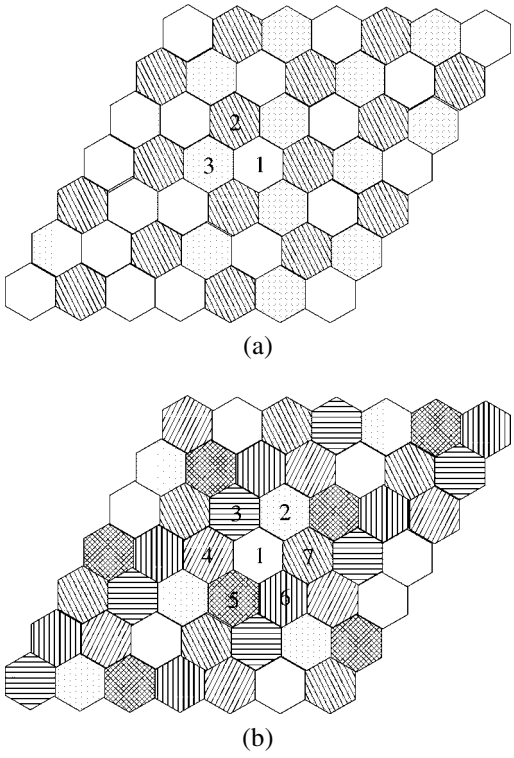

Figure 1. Examples of cellular systems with different minimum reuse distances. (a) $D_{\min} = 3R$, (b) $D_{\min} = \sqrt{21}R$.

Table 2
Notation definitions.

Notation	Definition
N	number of cells in the network
N_b	number of cells that employ the broadcast for data dissemination, $0 \leqslant N_b \leqslant N$
N_c	size of an interference cluster
B	total amount of bandwidth available for the system
M_i	number of data items for cell i's database
$item_{i,j}$	data item j for cell i's database
λ_i	data access rate for cell i
$p_{i,j}$	access probability for $item_{i,j}$ in cell i, $\sum_{j=1}^{M_i} p_{i,j} = 1$
$\lambda_{i,j}$	mean data access rate for $item_{i,j}$, i.e., $\lambda_{i,j} = p_{i,j}\lambda_i$
λ_c	aggregate data access rate for an interference cluster
$l_{i,j}$	length of data item $item_{i,j}$
l_i^s	sum of item sizes for cell i, $l_i^s = \sum_{j=1}^{M_i} l_{i,j}$
l_i	average size of accessed items for cell i, $l_i = \sum_{j=1}^{M_i} p_{i,j} l_{i,j}$
$s_{i,j}$	space distance between instances of $item_{i,j}$ in cell i's broadcast program
b_i	the amount of bandwidth allocated to cell i, $b_i > 0$
c_i	coordination of the center of cell i
$D_{\min}$	minimum frequency reuse distance
$\bar{t}(i, b_i)$	expected access latency for cell i with b_i bandwidth
$\bar{t}^b(i, b_i)$	expected access latency for cell i with b_i bandwidth for data broadcast
$\bar{t}^o(i, b_i)$	expected access latency for cell i with b_i bandwidth for on-demand access

The expected access latency for cell i is obtained:

$$\bar{t}^b(i, b_i) = \frac{l_i + 1/2 \sum_{j=1}^{M_i} p_{i,j} s_{i,j}}{b_i}, \quad 1 \leqslant i \leqslant N_b. \tag{1}$$

It can be observed that $\bar{t}^b(i, b_i)$ is inversely proportional to b_i, and $\bar{t}^b(i, b_i)$ can also be expressed with $(1/b_i)\bar{t}^b(i, 1)$. Now consider cell i where the on-demand access method is employed. The system can be approximately modeled as an $M/M/1$ queueing model. Let l_i denote the average size of accessed items for cell i and b_i the bandwidth allocated to cell i. The service rate is b_i/l_i, and we obtain the expected access latency for cell i:

$$\bar{t}^o(i, b_i) = \frac{l_i}{b_i - \lambda_i l_i}, \quad N_b + 1 \leqslant i \leqslant N. \tag{2}$$

Thus, we have the overall expected access latency for the system as follows ($0 \leqslant N_b \leqslant N$):[1]

$$\bar{t}(b_1, b_2, \ldots, b_N) = \frac{1}{\sum_{j=1}^{N} \lambda_j} \left(\sum_{i=1}^{N_b} \frac{\lambda_i \bar{t}^b(i, 1)}{b_i} + \sum_{i=N_b+1}^{N} \frac{\lambda_i l_i}{b_i - \lambda_i l_i} \right). \tag{3}$$

Our objective is to find out the optimal allocation of a total of B bandwidth to each cell which minimizes the overall expected latency. For ease of presentation, we define a new name *interference cluster* Q as a *maximal* subset of cells which are within the distance of mutual interference, i.e., $|c_i - c_j| < D_{\min}$, for all $i, j \in Q$, where $|c_i - c_j|$ is the distance between cell i and cell j. An interference cluster is a maximal subset in the sense that the union of this cell cluster and any other cell is not a cluster within which all the cells mutually interfere. For example, the set of cells numbered one through three in figure 1(a) and the set of cells numbered one through seven in figure 1(b) are two interference clusters, but the set of cells numbered one and two is not an interference cluster in either figure because it does not satisfy the maximality criterion. Thus, we obtain the mathematical representation of the bandwidth allocation problem given by

$$\text{Minimize } \bar{t}(b_1, b_2, \ldots, b_N), \tag{4}$$

$$\text{(BA)} \quad \text{subject to} \quad \sum_{k \in Q} b_k \leqslant B, \text{ for any interference cluster } Q \subseteq \{1, 2, \ldots, N\}, \tag{5}$$

where $\bar{t}(b_1, b_2, \ldots, b_N)$ is given by (3). Note that the constraints defined in (5) require that the bandwidth shared by any interference cluster does not exceed B. It is well known that the total bandwidth required for a regular cellular network equals the maximum bandwidth needed by an interference cluster [5,6]. Therefore, (5) guarantees that the total bandwidth for the data dissemination system required by any feasible solution to the BA problem will not exceed the available bandwidth B. Once a bandwidth allocation solution to the BA problem is obtained, the problem of assigning frequencies to

[1] Throughout this paper, it is assumed that the result is 0 for a sum function when its superscript is less than its subscript.

each cell matching the allocation solution can be solved by a *frequency assignment scheme,* which is outside the scope of this paper. Interested readers are referred to [5,6] for details.

4. The proposed heuristic allocation techniques

The optimization problem BA is complicated by not only bandwidth allocation among an interference cluster but also the factors of frequency interference and reuse. One naive solution for this problem is to enumerate all possible bandwidth allocation solutions which satisfy the constraints defined in (5) and find out the optimal solution. As can be seen, this scheme has a very high complexity, which is in the order of B^N. Moreover, if the amount of bandwidth allocated to a cell can be infinitely small, this scheme is actually infeasible. In this section, some sophisticated allocation techniques for the BA problem are proposed. According to (5), an interference cluster can share a maximum of B bandwidth. Thus, we first analyze the optimal bandwidth allocation for an interference cluster. When the network size is equal to the cluster size, this is an exact solution to the BA problem. For the case where the network size is greater than the cluster size, we further propose two heuristics, namely *compact allocation* and *cluster-step allocation*, based on the result for an interference cluster.

4.1. Optimal bandwidth allocation for an interference cluster

4.1.1. The analysis

This subsection presents the analysis for the optimal bandwidth allocation in an interference cluster. Since no frequency reuse is allowed within an interference cluster, the problem is greatly simplified. Let N_c denote the number of cells in an interference cluster and λ_c the aggregate data access rate for the cluster, i.e., $\lambda_c = \sum_{i=1}^{N_c} \lambda_i$. The optimization problem of minimizing the overall expected latency for an interference cluster is defined by

$$(\mathrm{BA_c}) \quad \text{Minimize } \bar{t}(b_1, b_2, \ldots, b_{N_c}), \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^{N_c} b_i = B, \tag{7}$$

where $\bar{t}(b_1, b_2, \ldots, b_{N_c})$ is given by (3).

For the $\mathrm{BA_c}$ problem, it is obvious that no feasible solution can be found if $B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i \leqslant 0$. In this case, the system has to increase the available bandwidth to offer a good access performance. When $B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i > 0$, the following theorem is obtained.

Theorem 1. For the optimization problem $\mathrm{BA_c}$, if $B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i > 0$, the minimum overall expected access latency is achieved when the bandwidth allocated to cell i, b_i, is given by

$$b_i = \begin{cases} \dfrac{\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}}{\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}} \\ \quad \times \left(B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i \right), \\ \quad \text{if } 1 \leqslant i \leqslant N_b; \\ \dfrac{\sqrt{\lambda_i l_i}}{\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}} \\ \quad \times \left(B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i \right) + \lambda_i l_i, \\ \quad \text{if } N_b + 1 \leqslant i \leqslant N_c. \end{cases} \tag{8}$$

The minimum overall expected latency is

$$\bar{t}^{\,*}(b_1, b_2, \ldots, b_{N_c}) = \frac{1}{\lambda_c (B - \sum_{i=N_b+1}^{N_b} \lambda_i l_i)} \times \left(\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i} \right)^2. \tag{9}$$

Proof. This is a constrained-minimum problem. It can be solved using the Lagrange multiplier theorem. Detailed proof is shown in appendix A. □

In the $\mathrm{BA_c}$ problem, $N_b = 0$ refers to the case where all the cells in the cluster employ the on-demand access method. In the other extreme, $N_b = N_c$ refers to the case where all the cells employ the broadcast method. In the following, for $N_b = N_c$ we consider two kinds of broadcast schedules of particular interest to us.

Flat broadcast. Under this model, each cell broadcasts all its data items in a round-robin manner. The space distance between instances of each $item_{i,j}$ is the sum of item sizes for cell i, i.e., $s_{i,j} = \sum_{j=1}^{M_i} l_{i,j} = l_i^s$. Thus, we have the following corollary.

Corollary 1. When the flat broadcast schedule is employed in each cell, the minimum overall expected latency, $1/(\lambda_c B) \times (\sum_{i=1}^{N_c} \sqrt{\lambda_i (l_i + 1/2 l_i^s)})^2$, is achieved when the bandwidth allocated to cell i is proportional to $\sqrt{\lambda_i (l_i + 1/2 l_i^s)}$.

The proof of this corollary is simply by following theorem 1 and substituting the parameters of $s_{i,j}$ and $\bar{t}^{\,b}(i,1)$. Reconsider the example presented in section 2. According to the above corollary, the minimum latency is achieved when the bandwidth allocated to cell A and cell B follows the ratio of $\sqrt{1 \cdot (1 + 4/2)} / \sqrt{4 \cdot (1 + 4/2)} = 1/2$, and the optimal latency is $1/((1+4) \cdot 6))(\sqrt{1 \cdot (1 + 4/2)} + \sqrt{4 \cdot (1 + 4/2)})^2 = 0.9$. The result is consistent with the example.

(1) 1st-round allocation
remain_band := B;
sum_para := $\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}$;
$L := \sum_{i=N_{b+1}}^{N_c} \lambda_i l_i$;
for $i := 1$ **to** N_c **do**
  if $i \leqslant N_b$ **then** $b_i := \lfloor \sqrt{\lambda_i \bar{t}^{\,b}(i,1)}/sum\_para \cdot (B - L)/B_u \rfloor \cdot B_u$;
  else $b_i := \lfloor (\sqrt{\lambda_i l_i}/sum\_para \cdot (B - L) + \lambda_i l_i)(1/B_u) \rfloor \cdot B_u$;
  if $b_i == 0$ **then** $b_i := B_u$;
  remain_band := *remain_band* $- b_i$;
end for
(2) 2nd-round allocation
while *remain_band* > 0 **do**
  select i for which $\lambda_i \bar{t}(i, b_i) - \lambda_i \bar{t}(i, b_i + B_u)$ is the largest;
  $b_i := b_i + B_u$;
  remain_band := *remain_band* $- B_u$;
end while
while *remain_band* < 0 **do**
  select i for which $b_i > B_u$ and
  $\lambda_i \bar{t}(i, b_i - B_u) - \lambda_i \bar{t}(i, b_i)$ is the least;
  $b_i := b_i - B_u$;
  remain_band := *remain_band* $+ B_u$;
end while

Algorithm 1. Bandwidth allocation with allocation granularity constraint.

Optimal broadcast. Flat broadcast is simple to implement. However, its average access performance is poor when the client access patterns are skewed. For a single cell i, the minimum expected latency is achieved when spacing $s_{i,j}$ of $item_{i,j}$ is proportional to square root of its length and inversely proportional to square root of its access probability,[2] i.e., $s_{i,j} = \sqrt{l_{i,j}/p_{i,j}} \sum_{j=1}^{M_i} \sqrt{p_{i,j} l_{i,j}}$. Such a schedule minimizing the expected latency is termed as the *optimal broadcast schedule*. Therefore, following theorem 1 we have the corollary for the optimal broadcast schedule as follows:

Corollary 2. When the optimal broadcast schedule is employed in each cell, let $q_i = \sum_{j=1}^{M_i} \sqrt{p_{i,j} l_{i,j}}$, the minimum overall expected latency, $1/(\lambda_c B)(\sum_{i=1}^{N_c} \sqrt{\lambda_i (l_i + 1/2q_i^2)})^2$, is achieved when the bandwidth allocated to cell i is proportional to $\sqrt{\lambda_i (l_i + 1/2q_i^2)}$.

4.1.2. Constraint of bandwidth allocation granularity
So far we have quantitatively analyzed the optimal bandwidth allocation for different cells in an interference cluster. However, in some situations, there is a constraint of bandwidth allocation granularity. Suppose that the minimum bandwidth allocation unit is B_u, then the bandwidth allocated to a cell must be nB_u, where n is a positive integer. In this case, the bandwidth allocation needs to approximate the optimal results obtained in the previous subsection. We propose a heuristic for such situations. The pseudo-code of the algorithm is described in algorithm 1. It is a two-round bandwidth allocation scheme. In the first round, the algorithm assigns bandwidth to a cell according to the optimal solution but truncates its fractional bandwidth. If the bandwidth is 0 after truncation, we set it to B_u to prevent the expected latency becoming infinite. In the second round, if the remaining bandwidth is larger than 0, B_u bandwidth is repeatedly assigned to a cell which achieves the greatest incremental decrease[3] in the objective function $\bar{t}(1, 2, \ldots, N_c)$ until the remaining bandwidth is exhausted. On the other hand, if the remaining bandwidth is less than 0,[4] B_u bandwidth is repeatedly extracted from a cell which has more than B_u bandwidth and introduces the least incremental increase (i.e., $\lambda_i \bar{t}(i, b_i - B_u) - \lambda_i \bar{t}(i, b_i)$) in the objective function. The performance of this heuristic is investigated in the simulation experiments.

[2] This result can be easily obtained by extending the study performed in [7,16], where waiting time was employed as the performance metric.

4.2. Bandwidth allocation for the cellular network

Due to factors such as frequency reuse and non-uniform traffic loads, allocating bandwidth for a cellular network to achieve good access performance is not an easy task. In the following, we present two heuristics to do bandwidth allocation, namely *compact allocation* and *cluster-step allocation*. In theorem 1, there is an allocation factor associated with each cell i; hereinafter we call it *deserved allocation factor* of cell i. For example, for data broadcast, the deserved allocation factor for cell i is $\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}$; for on-demand access, the deserved allocation factor can be $\sqrt{\lambda_i l_i}$ (when $B - \sum_{i=N_{b+1}}^{N_c} \lambda_i l_i$ dominates) or $\lambda_i l_i$ (when $\sum_{i=N_{b+1}}^{N_c} \lambda_i l_i$ dominates). We take a pessimistic estimation and use $\lambda_i l_i$ as the allocation factor for on-demand access.

4.2.1. Compact allocation
Under this scheme, the *compact* frequency allocation approach [20] is used. In assigning certain frequencies to a cell, the same frequencies are reused in other cells such that the average distance between them is minimal. To illustrate the idea, we borrow the *equivalence relation* Γ definition from [4].

Definition 1. For any two cells i, j in the cellular network, define $(i, j) \in \Gamma$ if and only if one of the following holds true: (i) $i = j$; (ii) $|c_i - c_j| = D_{min}$; (iii) there exists a cell k such that $(i, k) \in \Gamma$ and $(j, k) \in \Gamma$.

This is an equivalence relation, which, for example, partitions the network into three and seven equivalence classes in figures 1(a) and 1(b), respectively. The implication of this definition is that for all the cells in an equivalence class we can assign the same frequencies to them such that the optimal reuse distance is achieved. Regarding our bandwidth allocation problem, the same bandwidth can be allocated to all the cells in an equivalence class. Thus, the BA problem is reduced to the BA_c problem of bandwidth allocation in an interference cluster, which consists of a set of cells from all different equivalence classes with exactly one cell from each class. Examples of such cell clusters are the set of cells

[3] That is, $\lambda_i \bar{t}(i, b_i) - \lambda_i \bar{t}(i, b_i + B_u)$, where $\bar{t}(i, b_i)$ is given in (1) and (2) for broadcast and on-demand access, respectively.

[4] This occurs when there are too many cells that have bandwidth of 0 after truncation and are assigned B_u bandwidth.

numbered one through three in figure 1(a) and the set of cells numbered one through seven in figure 1(b). In such a cluster, for each cell we take the average parameter values (i.e., average $\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}$, $\sqrt{\lambda_i l_i}$, and $\lambda_i l_i$) of the cells that belong to its equivalence class. Then, the techniques presented in section 4.1 are used to allocate bandwidth to the cells according to their average parameter values. It is easy to see that under this scheme the bandwidth utilization is maximized and the constraints in (5) are satisfied. Obviously, this heuristic is good for the case where the loads for the cells in an equivalence class is homogeneous, but may not perform well for a heterogeneous case.

4.2.2. Cluster-step allocation

If the loads for the cells in an equivalence class differ greatly, it might be better to allocate different amount of bandwidth to them. In the cluster-step allocation approach, for all possible interference clusters,[5] the bandwidth allocation is carried out one interference cluster after one depending on their importance. The importance factor for a cluster is determined by its *aggregate deserved allocation factor*. The larger the aggregate deserved allocation factor, the more important the cluster. For bandwidth allocation in each interference cluster, the techniques presented in section 4.1 are applied. If some cells in a cluster Q have been assigned certain bandwidth before Q is picked only the unassigned cells are considered, i.e., the problem becomes how to assign bandwidth B' to an interference cluster which consists of the unassigned cells in Q. To avoid over-allocating bandwidth to the unassigned cells, B' is determined by taking the minimum of B_1 and B_2, where B_1 is the remaining available bandwidth in Q and B_2 is the deserved allocated bandwidth for the unassigned cells if no cells have been assigned bandwidth for Q.

Theorem 2. If data broadcast is employed in all the cells of the system and there is no constraint of bandwidth allocation granularity, under the cluster-step allocation scheme, no related interference cluster will exceed the available bandwidth B after each allocation for an interference cluster.

Proof. See appendix B. □

The above theorem shows that the cluster-step allocation scheme can prevent the system from over-using the available bandwidth B, i.e., (5) is satisfied, for the exclusive broadcast case. For other cases, we use a detection approach to obtain a feasible solution. After allocating bandwidth for an interference cluster, if it causes a related interference cluster Q' to over-use bandwidth B, the following operation is performed. For each cell i in Q', the bandwidth is adjusted to the minimum of b_i and b'_i, where b_i is the currently assigned bandwidth and b'_i is cell i's deserved allocated bandwidth if no cells have been assigned bandwidth for Q'. This method leads to a feasible solution.

[5] An efficient algorithm for finding all interference clusters for a cellular network can be found in [3].

The rationale behind the cluster-step scheme is that it is more important to optimally allocate bandwidth in an interference cluster which has more impact on the overall system performance. Compared with the compact allocation scheme, this scheme can allocate more bandwidth to a more important cell. On the other hand, however, the frequency reuse distances produced by this scheme are not optimized and thus may lead to a poor performance for a homogeneous case.

5. Performance evaluation

5.1. Basic experiment configuration

The basic experiment configuration is as follows. It is assumed that there is a total of 672 Kbps bandwidth available for the data dissemination system. The database associated with each cell consists of 1,000 data items. Data item sizes follow an exponential distribution with a mean of 10 Kb. The cellular network has a number of cells. Data access rates for the cells are set as follows. Each cell i is assigned a relative access rate λ_i^{r}. If cell i uses data broadcast, we set cell i's access rate to λ_i^{r} since the cell's access performance depends on access probabilities over items and is independent to its data access rate. If cell i uses on-demand access, we introduce a factor of utilization rate ρ. Unless explicitly specified, cell i's access rate is set to $(B \cdot \rho/(\overline{\lambda^{\mathrm{r}}} \cdot \bar{l} \cdot N_{\mathrm{c}}))\lambda_i^{\mathrm{r}}$, where B is the total available bandwidth, $\overline{\lambda^{\mathrm{r}}}$ is the average of λ_i^{r}'s for the system, $\bar{l}$ is the average item length and N_{c} is the size of an interference cluster. In the experiments, ρ is set to 0.5 by default. The settings for the relative access rates, λ_i^{r}'s, will be specified for each set of experiments in the following subsections. In each cell i, data requests over the items follow the *Zipf* distribution [21] with a skewness parameter of θ_i, and θ_i is randomly chosen between 0 and 1. The experiment results are obtained after the system has reached the stable state, i.e., after 5,000 queries have been made in each cell. A total of 1,000,000 queries are evaluated.

In the experiments, besides the proposed allocation algorithms, two additional allocation schemes are evaluated:

- *Uniform allocation.* The bandwidth is uniformly allocated to each cell regardless of their workloads.
- *Proportional allocation.* This scheme allocates bandwidth to the cells skewly in an intuitive way. The bandwidth allocated to a cell is linearly proportional to its data access rate.

5.2. Performance evaluation for an interference cluster

We first investigate the performance of the optimal bandwidth allocation approach for an interference cluster. To take a close look at the effects of non-uniform workloads, the size of an interference cluster, N_{c}, is set to three. For the three cells, the relative access rates are set to $base\_rate^i$ $(i = 0, 1, 2)$, and *base_rate* is set to 3 by default; access skewness parameters θ_i's $(i = 0, 1, 2)$ are randomly set to 0.20, 0.59, 0.38,

respectively. In order to evaluate the performance under different data dissemination methods, all the cells employ either the broadcast or the on-demand access method. For data broadcast, two kinds of schedules, flat broadcast and optimal broadcast [7], are considered.

5.2.1. Effect of the workload skewness

In this subsection, the bandwidth allocation schemes are evaluated under various levels of workload skewness for the cluster. In the experiments, *base_rate* is varied from 1 to 6. The larger the *base_rate* value, the more skewed the workload of the system. The results are presented in figures 2(a) and 2(b) for optimal broadcast and on-demand access, respectively. The performance trends for flat broadcast are similar to the optimal broadcast case. In figure 2(a), theoretical lower bounds are calculated according to (9). In figure 2(b), theoretical lower bounds are not available because they require solving the $M/G/1$ model which is infeasible analytically. In the figures, the optimal allocation denotes the scheme that allocates bandwidth for the cluster according to (8).

From figures 2(a) and 2(b), two main observations are made. First, among the three allocation schemes, the proposed optimal allocation scheme performs the best. For data broadcast, in all cases, the performance of the optimal allocation approaches the theoretical lower bounds very closely. When the traffic loads are uniform in the cluster (i.e., *base_rate* $=$ 1), all the schemes have the same performance. With increasing workload skewness, the performance improvement of the optimal allocation scheme over the uniform and proportional allocation schemes becomes much greater. For example, for on-demand access, when *base_rate* is larger than 3, the uniform allocation scheme makes the system saturated because the heavy-load cells have too little bandwidth, whereas the optimal allocation scheme can adjust bandwidth allocation among the cells and achieve a very good performance; and when *base_rate* $= 6$ the optimal allocation scheme is 32% better than the proportional allocation scheme. Second, although more bandwidth is allocated to the heavy-load cells in the proportional allocation scheme, its performance is not good. On average, it is 24.2% and 23.4% worse than the optimal allocation scheme for optimal broadcast and on-demand access, respectively. It is observed in the experiments that, in terms of variance, the optimal allocation scheme performs similarly to the uniform allocation scheme but the proportional allocation scheme has a very bad performance. This verifies our intuition that the need of allocating more bandwidth to the heavy-load cells is over-estimated in the proportional allocation scheme.

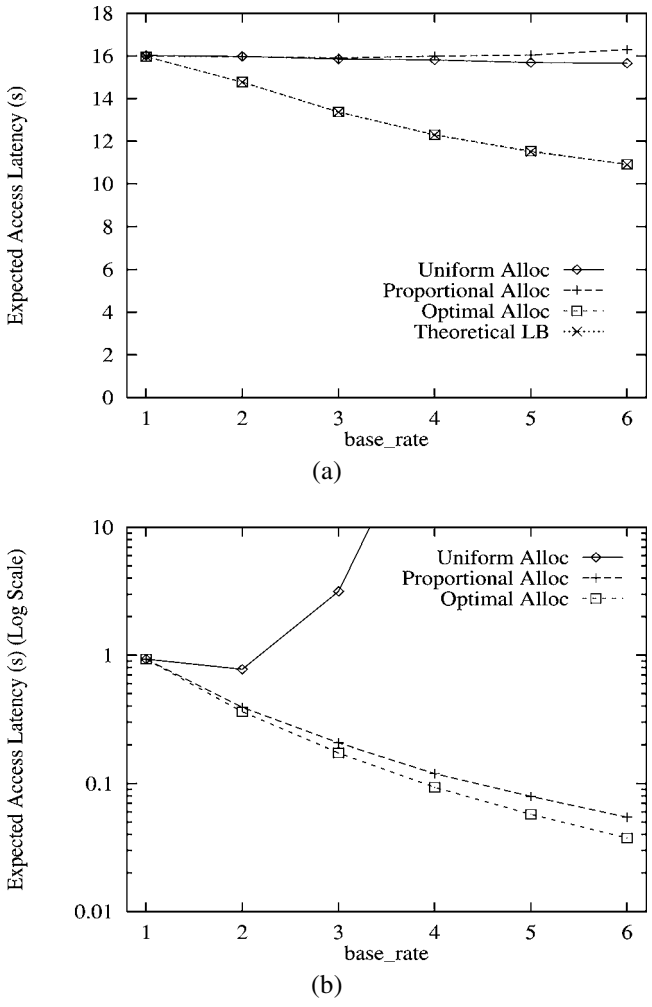

Figure 2. Access latency performance under skewed workloads. (a) Optimal broadcast. (b) On-demand-access.

5.2.2. Effect of the bandwidth allocation granularity

This subsection investigates the performance under a constraint of bandwidth allocation granularity. The performance of the proposed scheme is compared to the theoretical lower bounds, which are calculated without the constraint. Since on-demand access has no available lower bounds, only data broadcast is evaluated. The uniform and proportional allocation schemes are also included for comparison. For the uniform and proportional allocation, if the number of allocated bandwidth units for a cell is not an integer, the aggregate fractional bandwidth is re-assigned unit by unit in a way such that the overall performance is the best. The minimum bandwidth allocation unit is varied from 224 Kbps to ∞^{-1}, where ∞^{-1} denotes the case where there is no constraint on bandwidth allocation granularity. In other words, the number of bandwidth allocation units is varied from 3 to ∞. In the figures, the optimal allocation denotes the scheme that allocates bandwidth for the cluster using algorithm 1.

Figures 3(a) and 3(b) show the experiment results for flat broadcast and optimal broadcast, respectively. As can be seen, the optimal allocation scheme performs pretty well. When the number of allocation units is larger than 15, it has a very close performance to the theoretical lower bounds. When the number of allocation units is smaller than 15, its performance is only a little worse. For example, when there are only 3 allocation units available, the optimal allocation scheme is 19% worse than the lower bound for both flat broadcast and optimal broadcast. However, since in this case the only reasonable solution is to allocate one unit for one cell, in fact this is the best result that a practical scheme could ob-

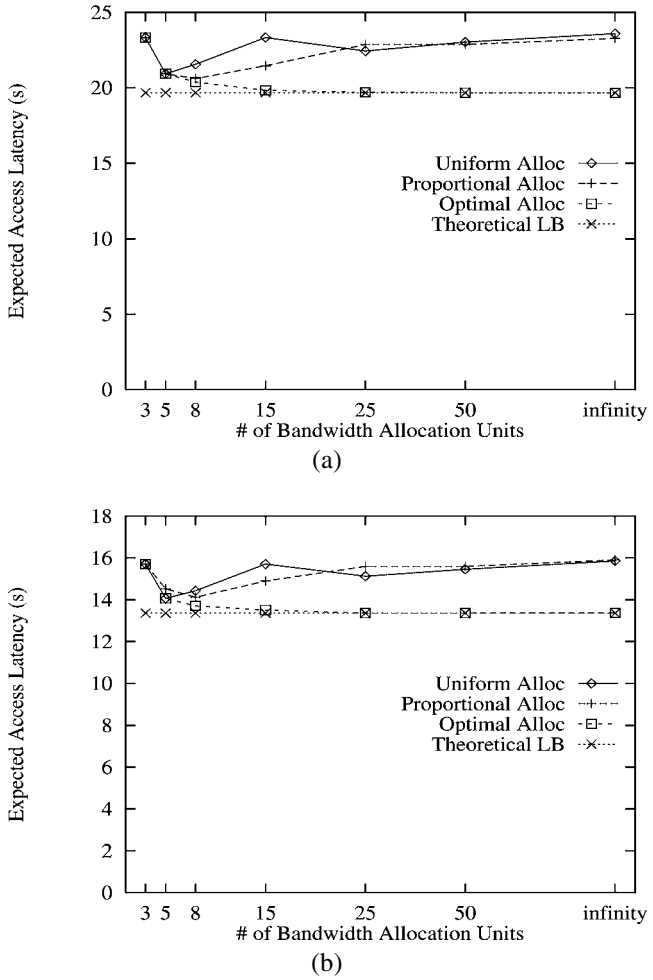

Figure 3. Effect of different bandwidth allocation granularity. (a) Flat broadcast. (b) Optimal broadcast.

tain. In all cases, the optimal allocation scheme performs no worse than the uniform and proportional allocation schemes.

5.3. Performance evaluation for the cellular network

This section evaluates the performance of the proposed allocation schemes for a cellular network. It is assumed that the network consists of 7×7 cells as shown in figure 4. A "wrapped-around" cellular network model is further assumed, in which the left-most column of cells is regarded as adjacent to the right-most column of cells and the top-row is adjacent to the bottom-row. The minimum reuse distance is assumed to be $\sqrt{21}R$ [14,20]. In other words, the size of an interference cluster is seven. Thus, it is easy to see that there are totally 49 interference clusters for this cellular network. The minimum bandwidth allocation unit is set to 9.6 Kbps, i.e., there are totally 70 bandwidth allocation units. Three data sets with different features are tested. In data set #1, the relative access rate for each cell is randomly chosen between 1 and 50. In data set #2, the relative access rate for the cells in the same equivalence class i is randomly chosen between 1 and 2^{i+1}. In data set #3, the relative access rates for most cells are uniformly distributed between 1 and 10 except

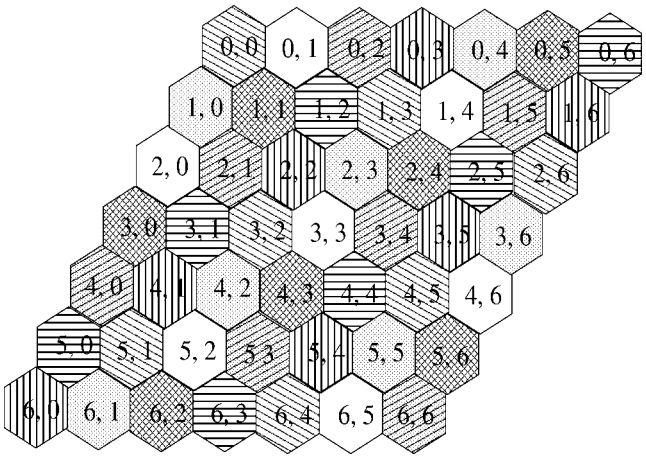

Figure 4. The cellular system used in the experiments.

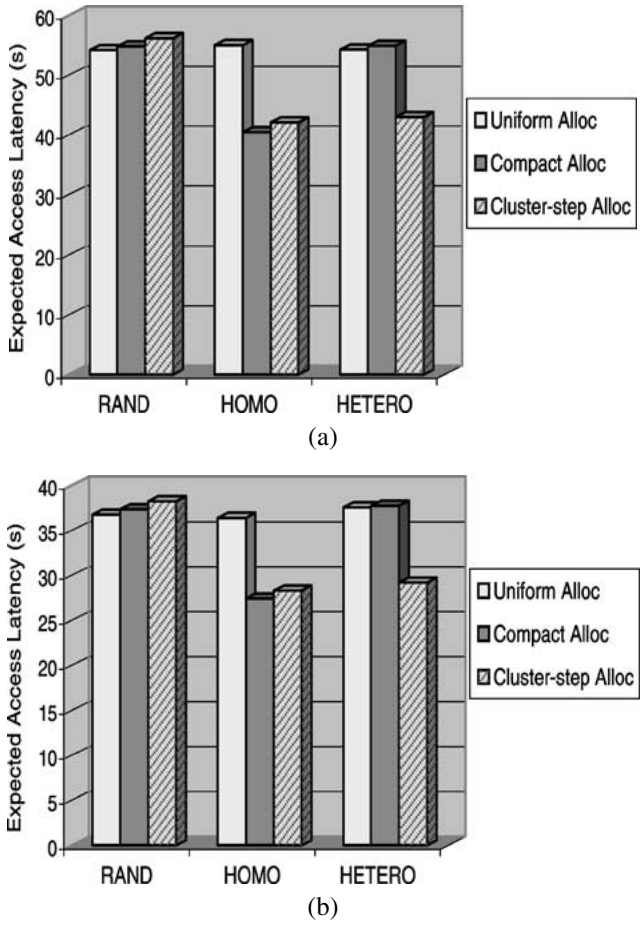

Figure 5. Performance of data broadcast for the cellular network. (a) Flat broadcast. (b) Optimal broadcast.

that some cells have a much higher rate of 100. These three sets are labeled as *random access* (*RAND*), *homogeneous access* (*HOMO*), and *heterogeneous access* (*HETERO*) cases, respectively. The real settings for the relative access rates λ_i^r's and access skewness parameters θ_i's are shown in tables 3–5. Since the proportional allocation demonstrates a similar performance to the uniform allocation scheme in the previous subsection, we do not include it for comparison here.

Table 3
Data set #2: random access.

$(\lambda^{r}_{(x,y)}, \theta_{(x,y)})$	$x=0$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$
$y=0$	(1, 0.03)	(33, 0.70)	(16, 0.31)	(34, 0.39)	(6, 0.79)	(26, 0.83)	(25, 0.91)
$y=1$	(31, 0.85)	(19, 0.62)	(13, 0.99)	(19, 0.79)	(42, 0.95)	(9, 0.38)	(15, 0.44)
$y=2$	(33, 0.32)	(40, 0.28)	(50, 0.54)	(41, 0.49)	(24, 0.07)	(27, 0.02)	(2, 0.02)
$y=3$	(13, 0.05)	(36, 0.55)	(36, 0.91)	(49, 0.71)	(17, 0.51)	(23, 0.59)	(36, 0.09)
$y=4$	(37, 0.47)	(9, 0.40)	(1, 0.85)	(40, 0.96)	(3, 0.69)	(30, 0.75)	(13, 0.27)
$y=5$	(28, 0.61)	(26, 0.26)	(20, 0.31)	(10, 0.22)	(33, 0.01)	(36, 0.12)	(16, 0.62)
$y=6$	(49, 0.99)	(42, 0.42)	(20, 0.04)	(36, 0.21)	(7, 0.35)	(31, 0.21)	(28, 0.01)

Table 4
Data set #2: homogeneous access.

$(\lambda^{r}_{(x,y)}, \theta_{(x,y)})$	$x=0$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$
$y=0$	(1, 0.03)	(3, 0.70)	(3, 0.31)	(11, 0.39)	(4, 0.79)	(34, 0.83)	(63, 0.91)
$y=1$	(20, 0.85)	(24, 0.62)	(33, 0.99)	(1, 0.79)	(4, 0.95)	(2, 0.38)	(5, 0.44)
$y=2$	(3, 0.32)	(7, 0.28)	(16, 0.54)	(26, 0.49)	(30, 0.07)	(69, 0.02)	(2, 0.02)
$y=3$	(16, 0.05)	(91, 0.55)	(2, 0.91)	(4, 0.71)	(3, 0.51)	(8, 0.59)	(23, 0.09)
$y=4$	(6, 0.47)	(3, 0.40)	(1, 0.85)	(51, 0.96)	(6, 0.69)	(2, 0.75)	(1, 0.27)
$y=5$	(72, 0.61)	(2, 0.26)	(2, 0.31)	(2, 0.22)	(11, 0.01)	(23, 0.12)	(20, 0.62)
$y=6$	(16, 0.99)	(27, 0.42)	(25, 0.04)	(91, 0.21)	(1, 0.35)	(3, 0.21)	(5, 0.01)

Table 5
Data set #3: heterogeneous access.

$(\lambda^{r}_{(x,y)}, \theta_{(x,y)})$	$x=0$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$
$y=0$	(1, 0.03)	(100, 0.70)	(7, 0.31)	(14, 0.39)	(3, 0.79)	(11, 0.83)	(10, 0.91)
$y=1$	(13, 0.85)	(8, 0.62)	(6, 0.99)	(8, 0.79)	(17, 0.95)	(4, 0.38)	(100, 0.44)
$y=2$	(13, 0.32)	(16, 0.28)	(20, 0.54)	(17, 0.49)	(100, 0.07)	(11, 0.02)	(13, 0.02)
$y=3$	(5, 0.05)	(15, 0.55)	(100, 0.91)	(20, 0.71)	(7, 0.51)	(9, 0.59)	(15, 0.09)
$y=4$	(100, 0.47)	(4, 0.40)	(1, 0.85)	(16, 0.96)	(1, 0.69)	(12, 0.75)	(5, 0.27)
$y=5$	(12, 0.61)	(11, 0.26)	(8, 0.31)	(4, 0.22)	(13, 0.01)	(100, 0.12)	(7, 0.62)
$y=6$	(20, 0.99)	(17, 0.42)	(8, 0.04)	(100, 0.21)	(3, 0.35)	(13, 0.21)	(11, 0.01)

The allocation schemes are first evaluated for the case where all the cells employ the broadcast method for data dissemination. The results are shown in figure 5. Examine the optimal broadcast case (figure 5(b)). For *RAND*, the proposed compact allocation scheme performs similarly to the uniform allocation scheme. For *HOMO*, the compact allocation and cluster-step allocation schemes have much better performance than the uniform allocation scheme, producing the improvement of 24.7% and 22.2%, respectively. For both *RAND* and *HOMO*, the cluster-step allocation scheme is a little bit worse than the compact allocation scheme. This is mainly because the frequency reuse distances are not optimized in the cluster-step allocation scheme. For *HETERO*, the compact allocation scheme has the performance as bad as the uniform allocation scheme, whereas the cluster-step allocation scheme improves the performance 22% over them. The compact allocation scheme is not good, because it only takes into consideration the average load for each equivalence class. Similar results are observed for flat broadcast.

In the following set of experiments, we compare the allocation schemes for the case where all of the cells employ the on-demand access method. The results are shown in figures 6(a) and 6(b) for the utilization rates of 0.2 and 0.5, respectively. When the system has a light utilization (see figure 6(a)), results similar to the exclusive broadcast case are obtained. However, for *HOMO* and *HETERO* the improvement of the proposed schemes over the uniform allocation scheme is much greater. The reason is that under the on-demand access model, if the data access rate exceeds a cell's capacity, its performance will become extremely poor. Thus on-demand access is more sensitive to bandwidth allocation. When the system reaches a medium utilization (see figure 6(b)), the proposed schemes can adjust bandwidth allocation among the cells to achieve a good overall performance. However, some of the cells in the uniform allocation scheme have a very poor performance, which leads to a worse overall performance for all three cases. Moreover, in figure 6(b), the cluster-step allocation scheme performs better than the compact allocation scheme for both *RAND* and *HOMO*. We also conducted experiments for a heavy load case ($\rho = 0.8$), and the observed results are similar to the case of $\rho = 0.5$.

In the last set of experiments, the allocation schemes are investigated for a hybrid system. In a hybrid system, the cells of relative access rates $\lambda^{r}_{i} < T$ employ the on-demand access and the rest use the optimal broadcast. Figures 7(a) and 7(b) illustrate the experiment results for $T = 5$ and $T = 10$, respectively. From the figures, the first observation is that the relative performance of the allocation schemes is simi-

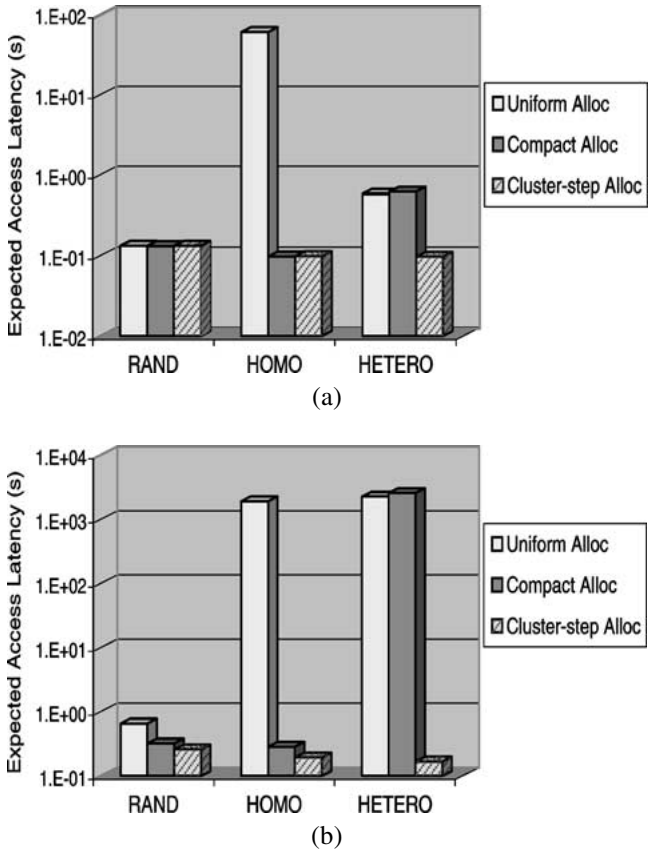

Figure 6. Performance of on-demand access for the cellular network. (a) $\rho = 0.2$. (b) $\rho = 0.5$.

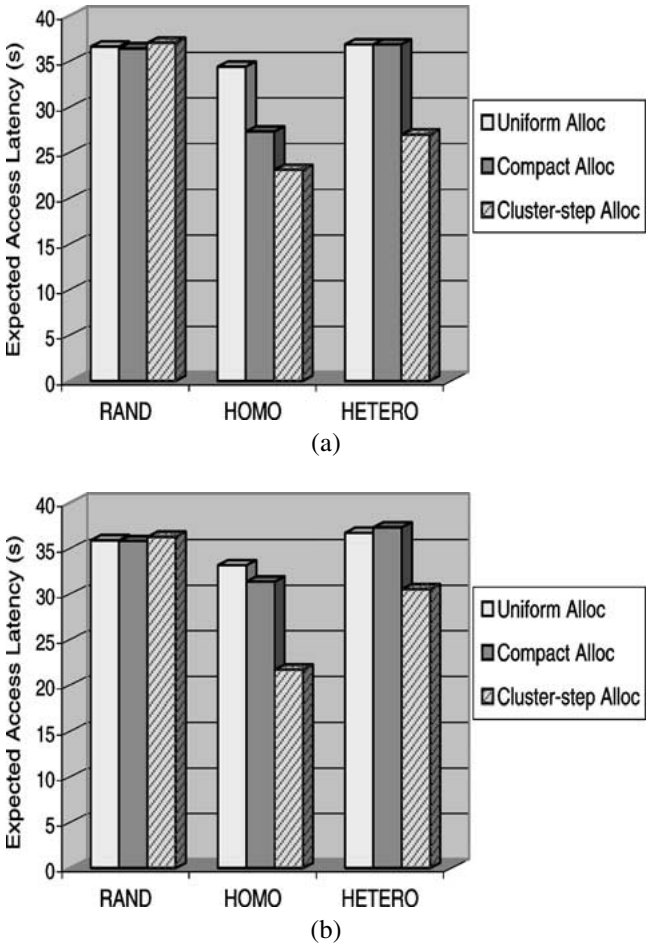

Figure 7. Performance of hybrid access for the cellular network. (a) $T = 5$. (b) $T = 10$.

lar to the exclusive broadcast case except that for *HOMO* the compact allocation scheme has a worse performance than the cluster-step allocation scheme. This can be explained as follows. The compact allocation scheme assigns less bandwidth to cells which have a lighter average load. However, if the loads for some cells in this class are heavy, as observed before for on-demand access, their performance is very poor, resulting in a worse overall performance. That is also the reason that the performance for the compact allocation becomes even worse when $T = 10$. The second observation is that a hybrid system can improve the overall system performance if a reasonable number of light-load cells use the on-demand access method instead of the broadcast method. We illustrate this observation using the *HETERO* data set as an example. Compare figures 5(b) and 7(a); for *HETERO* the performance of the cluster-step allocation scheme is improved by 7.7% when the cells of $\lambda_i^r < 5$ employ the on-demand access. However, as T is increased to 10, i.e., the cells of $\lambda_i^r < 10$ employ the on-demand access, its performance degrades 13% (see figure 7(b)). The main reason is that under a light load on-demand access can achieve a better performance than data broadcast but under a heavy load data broadcast wins.

6. Related work

6.1. Bandwidth allocation for data dissemination

Bandwidth allocation strategies for data dissemination so far have focused on two aspects. One aspect is bandwidth allocation among different data items for broadcast schedules [1,7,18]; the other is bandwidth allocation between data broadcast and on-demand access for hybrid data dissemination systems [9,11]. However, both of them are limited to a single-cell environment.

A *broadcast schedule* consists of two subproblems: (i) bandwidth allocation among different items, and (ii) order of items to be broadcast. In [1], a periodic dissemination architecture, called *Broadcast Disk* (*Bdisk*), was proposed. The Bdisks superimpose multiple (logical) disks, each spinning at a different speed, on a single broadcast channel. Using the *Bdisk*, one can construct a fine-grained memory hierarchy such that the items of higher popularities are assigned more bandwidth and are closer to the clients. Optimal broadcast schedules were investigated in [7,18], where a *square-root rule* was discovered for minimizing the overall waiting time.

In a hybrid data dissemination system, the data broadcast and on-demand access are combined to complement each other [9,11]. Imielinski and Viswanathan [9] considered one

channel in a cell and the channel is divided into two subchannels: broadcast and on-demand access. Based on this architecture, they proposed an algorithm that achieves the minimum number of uplink requests while making sure that the optimal mean access latency does not exceed a certain predefined threshold. There were no simulation studies in [9]. Lee et al. [11] assumed that there are a number of channels in a cell. Based on an $M/M/m/B$ queuing model, a dynamic channel allocation scheme was proposed to obtain the minimum overall access latency. Basically, the proposed algorithms in [9,11] aim at identifying the optimal set of broadcast items and determining the amount of bandwidth used for data broadcast and on-demand access.

The authors in [2,13,15] studied the schemes for allocating data on broadcast and on-demand channels. However, their studies assumed a fixed ratio of broadcast to on-demand bandwidth, as opposed to dynamic assignment based on system workloads.

6.2. Channel allocation for voice communication

Channel allocation schemes have been extensively studied for multi-cell voice communication systems (see [10] for a survey and [12,17] for some more recent work). Different from data dissemination, the main concerns for voice communications are *blocking probability* for new calls, *dropping probability* for hand-off calls, and delay in channel assignment etc.

Basically, the existing channel allocation schemes for voice communications can be divided into three categories [10]: fixed channel allocation, dynamic channel allocation, and hybrid channel allocation. In fixed channel allocation, a number of channels are permanently allocated to each cell depending on long-term traffic loads. In dynamic channel allocation, all channels are placed in a pool and they are assigned to new calls on a call-by-call basis. Hybrid channel allocation strategies combine aspects of both fixed and dynamic channel allocation schemes. Each cell is assigned a set of permanent channels and when the permanent channels for a cell become inadequate it requests channels as needed from the shared channel pool.

Among these three categories, closest to our work is fixed channel allocation [14,20]. In [20], the non-uniform compact pattern allocation scheme was proposed, which attempts to allocate channels to cells in a way such that the overall call blocking probability is minimized. A regular hexagonal cellular network was considered. *Cochannel cells* of a channel refer to the cells that can use this channel simultaneously without causing signal interference. The allocation of a channel to the set of cochannel cells forms an *allocation pattern*. The allocation pattern with minimum average distance between cochannel cells is referred to as the *compact allocation pattern*. Given the traffic load for each cell, the proposed heuristic finds the optimal compact allocation pattern for one channel at a time until all available channels are allocated.

In [14], new calls and hand-off calls are distinguished. It considered prioritized channel assignment, where higher priorities are given to hand-off calls, i.e., only hand-off calls can be admitted after the number of remaining free channels drops below some threshold. The objective considered is to minimize the overall hand-off call dropping probability while ensuring sufficient level of QoS for new call attempts. As such, a marginal capacity allocation technique was proposed for channel allocation in a cell cluster. On determining the assignment of each channel, the cell which can achieve the largest incremental decrease in the objective function is selected. It is not clear how to extend the result to a cellular network in [14].

7. Conclusion

Studies on bandwidth allocation for data dissemination in a multi-cell mobile network have been limited in the literature. To the best of our knowledge, this is the first study that attempts to address the multi-cell bandwidth allocation problem for data dissemination. With the boom of mobile data applications, it is believed that the bandwidth allocation problem will become more and more important.

In this paper, we have formulated the bandwidth allocation problem formally under a regular cellular model, with the objective of minimizing the overall expected data access latency. We have analyzed how to optimally allocate bandwidth for an interference cluster without frequency reuse. Two heuristic bandwidth allocation schemes have also been proposed for a cellular network with frequency reuse. The compact allocation scheme achieves the optimal frequency reuse distance, whereas the cluster-step allocation scheme assigns more bandwidth to important cells.

The proposed schemes have been evaluated by a series of experiments under various system configurations. For a three-cell interference cluster, the proposed optimal allocation method shows a substantially better performance than the uniform and proportional allocation schemes for non-uniform workloads. It is also demonstrated that the proposed allocation approach has a very good performance when there is a constraint on the bandwidth allocation granularity. For a 7×7 cellular network, the proposed compact allocation and cluster-step allocation schemes were investigated under three kinds of data sets. For the random access case, the proposed schemes perform better than the uniform allocation scheme when the average system utilization is medium or high for on-demand access. For the homogeneous access case, both of the proposed schemes outperform the uniform allocation scheme significantly. The compact allocation scheme has a better performance than the cluster-step allocation scheme for data broadcast and light-load on-demand access, but the cluster-step allocation scheme dominates in the other cases. For the heterogeneous access case, the cluster-step allocation scheme performs significantly better than the uniform allocation scheme, whereas the compact allocation scheme cannot improve the performance.

As for future work, we are going to extend this study to irregular cellular network models and other optimization objectives. As a starting point, it is assumed in this paper that

each cell employs either broadcast or on-demand access for data dissemination. A third model is hybrid data dissemination, in which one needs to consider how much bandwidth is assigned to a cell and how to allocate the assigned bandwidth between broadcast and on-demand access. Thus, this hierarchical bandwidth allocation problem becomes more complicated, which deserves more work. Furthermore, the bandwidth allocation problem can be considered in an integrated service model where voice and data services are combined. In addition, this study considered allocating bandwidth based on long-term traffic loads for the cells. It is interesting to investigate how hand-offs and/or dynamic loads would affect the system performance.

Acknowledgements

The research was supported by the Research Grant Council, Hong Kong SAR, China under grant numbers DAG99/00.EG09, HKUST6163/00E, and HKUST6241/00E.

Appendix A. Proof of theorem 1

Since $\sum_{i=1}^{N_b} b_i = B$, we have $\sum_{i=1}^{N_c} b_i - B = 0$. Let

$$L(b_1, b_2, \ldots, b_{N_c}, \gamma) = \bar{t}(b_1, b_2, \ldots, b_{N_c}) + \gamma\left(\sum_{i=1}^{N_c} b_i - B\right). \quad (A.1)$$

It is obvious that the optimization problem defined by the above equation and (7) is equivalent to the BA_c problem.

Substituting (3) for $\bar{t}(b_1, b_2, \ldots, b_{N_c})$ we rewrite (A.1) as follows:

$$L(b_1, b_2, \ldots, b_{N_c}, \gamma) = \frac{1}{\lambda_c}\left(\sum_{i=1}^{N_b} \frac{\lambda_i \bar{t}^{\,b}(i,1)}{b_i} + \sum_{i=N_{b+1}}^{N_c} \frac{\lambda_i l_i}{b_i - \lambda_i l_i}\right) + \gamma\left(\sum_{i=1}^{N_c} b_i - B\right). \quad (A.2)$$

The differential technique is applied to solve this optimization problem. Differentiate (A.2) by b_i, we obtain

$$\frac{\partial L(b_1, \ldots, b_{N_c}, \gamma)}{\partial b_i} = -\frac{1}{\lambda_c}\frac{\lambda_i \bar{t}^{\,b}(i,1)}{b_i^2} + \gamma, \quad \text{if } 1 \leqslant i \leqslant N_b, \quad (A.3)$$

and

$$\frac{\partial L(b_1, \ldots, b_{N_c}, \gamma)}{\partial b_i} = -\frac{1}{\lambda_c}\frac{\lambda_i l_i}{(b_i - \lambda_i l_i)^2} + \gamma, \quad \text{if } N_b + 1 \leqslant i \leqslant N_c. \quad (A.4)$$

Equal them to zero, after some transformation we have:

$$b_i = \begin{cases} \dfrac{\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}}{\sqrt{\lambda_c \gamma}}, & \text{if } 1 \leqslant i \leqslant N_b; \\ \dfrac{\sqrt{\lambda_i l_i}}{\sqrt{\lambda_c \gamma}} + \lambda_i l_i, & \text{if } N_b + 1 \leqslant i \leqslant N_c. \end{cases} \quad (A.5)$$

As $\sum_{i=1}^{N_c} b_i = B$, we obtain

$$\sum_{i=1}^{N_b} \frac{\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}}{\sqrt{\lambda_c \gamma}} + \sum_{i=N_b+1}^{N_c}\left(\frac{\sqrt{\lambda_i l_i}}{\sqrt{\lambda_c \gamma}} + \lambda_i l_i\right) = B,$$

$$\frac{1}{\sqrt{\lambda_c \gamma}}\left(\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}\right) + \sum_{i=N_b+1}^{N_c} \lambda_i l_i = B,$$

$$\sqrt{\lambda_c \gamma} = \frac{\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_{b+1}}^{N_c} \sqrt{\lambda_i l_i}}{B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i}. \quad (A.6)$$

Substituting $\sqrt{\lambda_c \gamma}$ in (A.5), we get b_i as follows:

$$b_i = \begin{cases} \dfrac{\sqrt{\lambda_i \bar{t}^{\,b}(i,1)}}{\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}} \times \left(B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i\right), \\ \quad \text{if } 1 \leqslant i \leqslant N_b; \\ \dfrac{\sqrt{\lambda_i l_i}}{\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}} \times \left(B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i\right) + \lambda_i l_i, \\ \quad \text{if } N_b + 1 \leqslant i \leqslant N_c. \end{cases} \quad (A.7)$$

This means that when b_i is given by (A.7), (A.1) is minimized, and hence $\bar{t}(b_1, b_2, \ldots, b_{N_c})$ is optimized as well. Substituting b_i in (3), we obtain the minimum overall expected latency as follows:

$$\bar{t}^{\,*}(b_1, b_2, \ldots, b_{N_c}) = \frac{1}{\lambda_c(B - \sum_{i=N_b+1}^{N_c} \lambda_i l_i)} \times \left(\sum_{i=1}^{N_b} \sqrt{\lambda_i \bar{t}^{\,b}(i,1)} + \sum_{i=N_b+1}^{N_c} \sqrt{\lambda_i l_i}\right)^2.$$

Appendix B. Proof of theorem 2

Under the cluster-step allocation scheme, at each allocation step for an interference cluster, only the unassigned cells are considered. Thus, each allocation affects the total assigned

bandwidth only for the interference clusters that have not been picked. Assume the current cluster considered is Q. Let's consider a related interference Q' which has not been picked yet and intersects with Q. Without loss of generality, assume that cells $1, 2, \ldots, N_a$ $(1 \leqslant N_a \leqslant N_c)$ in Q' have been assigned certain bandwidth after the current allocation. Further assume that the bandwidth of cell i $(1 \leqslant i \leqslant N_a)$ is assigned when cluster Q_i is picked. When the exclusive broadcast method is employed, the deserved allocation factor is $f_i = \sqrt{\lambda_i \bar{t}^b(i, 1)}$ for cell i. Let f_i^q be the aggregate allocation factor for cluster Q_i $(1 \leqslant i \leqslant N_a)$, and $f_{Q'}^q$ be the aggregate allocation factor for cluster Q', $f_{Q'}^q \geqslant \sum_{i=1}^{N_a} f_i$. Now consider cell i. If no cells have been assigned bandwidth when Q_i is picked, then the bandwidth allocated to cell i is $b_i = (f_i/f_i^q)B$. If some cells have been assigned when Q_i is picked, let f_i^u denote the aggregate allocation factor for the unassigned cells; because the algorithm limits the available bandwidth B' for the unassigned cells up to the amount of bandwidth that they are deserved to have, then $b_i = (f_i/f_i^u)B' \leqslant (f_i/f_i^u)((f_i^u/f_i^q)B) = (f_i/f_i^q)B$. Thus, the bandwidth allocated to cluster Q' so far is $TB \leqslant \sum_{i=1}^{N_a}(f_i/f_i^q)B$. Since Q_i $(1 \leqslant i \leqslant N_a)$ has been picked but Q' has not, this means $f_i^q \geqslant f_{Q'}^q$ $(1 \leqslant i \leqslant N_a)$ according to the algorithm. Therefore, $TB \leqslant \sum_{i=1}^{N_a}(f_i/f_{Q'}^q)B \leqslant B$, and the theorem follows.

References

[1] S. Acharya, R. Alonso, M. Franklin and S. Zdonik, Broadcast disks: Data management for asymmetric communications environments, in: *Proceedings of ACM SIGMOD Conference on Management of Data*, San Jose, CA, USA (May 1995) pp. 199–210.

[2] S. Acharya, M. Franklin and S. Zdonik, Balancing push and pull for data broadcast, in: *Proceedings of ACM SIGMOD Conference on Management of Data*, Tucson, AZ, USA (May 1997) pp. 183–194.

[3] C. Bron and J. Kerbosch, Finding all cliques of an undirected graph, Communications of the ACM 16(9) (1973) 575–577.

[4] X. Dong and T.H. Lai, An efficient priority-based dynamic channel allocation strategy for mobile cellular networks, in: *Proceedings of IEEE INFOCOM'97*, Kobe, Japan (April 1997) pp. 892–899.

[5] A. Gamst, Some lower bounds for a class of frequency assignment problems, IEEE Transactions on Vehicular Technology 35(1) (1986) 8–14.

[6] A. Gamst and W. Rave, On frequency assignment in mobile automatic telephone systems, in: *Proceedings of IEEE Global Telecommunications Conferences (GLOBECOM)*, Miami, FL, USA (November 1982) pp. 309–315.

[7] S. Hameed and N.H. Vaidya, Efficient algorithms for scheduling data broadcast, Wireless Networks (WINET) 5(3) (1999) 183–193.

[8] T. Imielinski and B.R. Badrinath, Wireless mobile computing: Challenges in data management, Communications of the ACM 37(10) (1994) 18–28.

[9] T. Imielinski and S. Viswanathan, Adaptive wireless information systems, in: *Proceedings of the Special Interest Group on DataBase Systems (SIGDBS) Conference*, Tokyo, Japan (October 1994) pp. 19–41.

[10] I. Katzela and M. Naghshineh, Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey, IEEE Personal Communications 3(3) (1996) 10–31.

[11] W.-C. Lee, Q.L. Hu and D.L. Lee, A study of channel allocation methods for data dissemination in mobile computing environments, Mobile Networking and Applications (MONET) 4(2) (1999) 117–129.

[12] B. Li, L. Yin, K.Y. Michael Wong and S. Wu, An efficient and adaptive bandwidth allocation scheme for mobile wireless networks based on on-line local parameter estimations, Wireless Networks (WINET) 7(2) (2001) 107–116.

[13] C.W. Lin and D.L. Lee, Adaptive data delivery in wireless communication environments, in: *Proceedings of 20th IEEE International Conference on Distributed Computing Systems (ICDCS'OO)*, Taipei, Taiwan (April 2000) pp. 444–452.

[14] S.-H. Oh and D.-W. Tcha, Prioritized channel assignment in a cellular radio network, IEEE Transactions on Communications 40(7) (1992) 1259–1269.

[15] K. Stathatos, N. Roussopoulos and J.S. Baras, Adaptive data broadcast in hybrid networks, in: *Proceedings of the 23rd VLDB Conference*, Athens, Greece (August 1997) pp. 326–335.

[16] N.H. Vaidya and S. Hameed, Scheduling data broadcast in asymmetric communication environments, Wireless Networks (WINET) 5(3) (1999) 171–182.

[17] A.L. Wijesinha, S.P. Kumar and D.P. Sidhu, Handover and new call blocking performance with dynamic single-channel assignment in linear cellular arrays, Wireless Networks (WINET) 6(2) (2000) 121–129.

[18] J.W. Wong, Broadcast delivery, Proceedings of the IEEE 76(12) (1988) 1566–1577.

[19] S. Wu, K.Y. Michael Wong and B. Li, A dynamic call admission policy with precision qos guarantee using stochastic control for mobile wireless networks, IEEE Transaction on Networking (2001) to appear.

[20] M. Zhang and T.P. Yum, The nonuniform compact pattern allocation algorithm for cellular mobile system, IEEE Transactions on Vehicular Technology 40(2) (1991) 387–391.

[21] G.K. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Reading, MA, 1949).

Jianliang Xu is an Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. He received the BEng degree from Zhejiang University, Hangzhou, China, in 1998, and the Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2002. His research interests include mobile computing, wireless networks, Internet technologies, and distributed systems.
E-mail: xujl@cs.ust.hk

Dik Lun Lee received the M.S. and Ph.D. degrees in computer science from the University of Toronto in 1981 and 1985, respectively. He is a Professor in the Department of Computer Science at the Hong Kong University of Science and Technology, and was an Associate Professor in the Department of Computer and Information Science at the Ohio State University, Columbus, OH, USA. He has served as a guest editor for several special issues on database-related topics, and as a program committee member and chair for numerous international conferences. He was the founding conference chair for the International Conference on Mobile Data Management. His research interests include document retrieval and management, discovery, management and integration of information resources on Internet, mobile and pervasive computing. He was the chairman of the ACM Hong Kong Chapter.
E-mail: dlee@cs.ust.hk

Bo Li received the B.S. (summa cum laude) and M.S. degrees in computer science from Tsinghua University, Beijing, P.R. China, in 1987 and 1989, respectively, and the Ph.D. degree in the computer engineering from University of Massachusetts at Amherst in 1993. Between 1994 and 1996, he worked on high performance routers and ATM switches in IBM Networking System Division, Research Triangle Park, North Carolina. Since then, he has been with the Computer Science Department, the Hong Kong University of Science and Technology. His current research interests include wireless mobile networking supporting multimedia, video multicast and all optical networks using WDM. He has been on editorial board for ACM Mobile Computing and Communications Review (MC2R), ACM/Kluwer Journal of Wireless Networks (WINET), IEEE Journal of Selected Areas in Communications (JSAC) – Wireless Communication Series (to be named IEEE Transactions on Wireless Communications), SPIE/Kluwer Optical Networking Magazine (ONM), KICS/IEEE Journal of Communications and Networks (JCN). He served as a guest editor for IEEE Communications Magazine Special Issue on Active, Programmable, and Mobile Code Networking (April 2000), IEEE Journalof Selected Areas in Communications Special Issue on Protocols for Next Generation Optical WDM Networks (October 2000), SPIE/Kluwer Optical Networks Magazine Special Issue on Wavelength Routed Networks: Architecture, Protocols and Experiments (January/February 2002), and an ACM Performance Evaluation Review Special Issue on Mobile Computing (2002). In addition, He has been involved in organizing over two dozens of conferences, was the vice-chair for IEEE Infocom'2001.

E-mail: bli@cs.ust.hk