Available online at www.sciencedirect.com
ScienceDirect
Computer Networks 50 (2006) 3466–3484
Computer Networks
www.elsevier.com/locate/comnet

Analytical modeling of CAC in next generation wireless systems ☆

Tuna Tugcu [a,*], H. Birkan Yilmaz [a], Feodor Vainstein [b]

[a] *Computer Engineering, Bogazici University, Bebek, Istanbul 34342, Turkey*
[b] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Savannah, GA 31407, United States*

Received 9 January 2006; accepted 30 January 2006
Available online 24 February 2006

Responsible Editor: I.F. Akyildiz

Abstract

Though *Connection Admission Control* (CAC) in wireless networks has been studied extensively, the heterogeneous structure of *Next Generation Wireless Systems* (NGWS) makes CAC very complex. Accessibility of the subsystems at the time of connection or handoff request, availability of resources in the subsystems, user preferences, and connection class need to be considered in admission control. In this paper, we first give a general CAC algorithm for NGWS. We also propose the first analytical model in the literature for CAC in NGWS. We point out the major challenges in modeling for NGWS and propose a neat solution to calculate state probabilities in a reasonable way even as the state space proliferates.

Keywords: Next generation wireless systems; Call admission control; Analytical modeling

1. Introduction

Aiming to provide high bandwidth access anytime/anywhere, *Next Generation Wireless Systems* (NGWS) merge multiple subsystems with different access technologies. Though similar objectives are pursued in the design of existing PCS, satellite, and WLAN systems, these systems fail to satisfy all of the requirements *simultaneously* due to constraints like global coverage, indoor/outdoor communications, and frequent handoffs. Since it is not likely that emerging new technologies will be able to provide continuous coverage, NGWS will utilize multiple wireless systems (and the new technologies to come) as *subsystems* in order to provide high bandwidth access everywhere. PCS, WLAN, and satellite systems as well as new technologies like 4G Mobile and WiMAX are examples of the subsystems in NGWS. The basic properties of NGWS will be as follows:

- *end-to-end* packet-based service, including the air interface,

☆ A preliminary version of this paper appeared in CONWIN 2005 in Budapest.

* Corresponding author. Tel.: +90 212 359 7611; fax: +90 212 287 2461.
E-mail addresses: tugcu@boun.edu.tr (T. Tugcu), yilmhuse@boun.edu.tr (H. Birkan Yilmaz), fvainste@gtsav.gatech.edu (F. Vainstein).

1389-1286/$ - see front matter
doi:10.1016/j.comnet.2006.01.007

- support for voice, multimedia, and data traffic with QoS provisioning,
- backbone traffic of NGWS carried over the Internet.

In [1,2], a cell design methodologies for the optimal design of multitier cellular systems are proposed. The use of intelligent networks for mobility and internetworking is discussed in [3]. In [4], Zeng et al. discuss emerging technological trends in PCS systems. Key issues about IMT-2000 are discussed from the perspective of the service provider in [5]. Integration of cordless and cellular systems in 3G systems is discussed in [6].

In the literature, there are several proposals for the NGWS architecture under different names including 4G, All-IP, Beyond-3G, and Beyond-UMTS. In [7,8], different NGWS architectures are described with a focus on the possible driving forces for the deployment of these networks. NTT DoCoMo proposes an NGWS architecture, which is an extension of current 3G architecture [9]. A next generation wireless communication architecture that is comprised of old and new wireless communication standards has been presented in [10]. Telefonica's NGWS architecture is composed of WLANs, cellular networks, personal area networks, and distribution networks organized in a layered structure [11]. In the Wine Glass Project, [12], WLANs and UMTS are merged into a next generation wireless network. Furthermore, Siemens [13] adopts 3GPP's IP Based Multimedia System (IMS) specifications [14] and defines its own next generation wireless network architecture. Integrating multiple subsystems into one NGWS brings many challenges ranging from interworking among inherently different wireless subsystems to QoS provisioning [12,15]. The selection of the subsystem for connection establishment and handoff is a key factor in the performance of NGWS. The Global Mobile Broadband System (GMBS) described in [16] aims to include GPRS, UMTS, WLAN, and satellite-based systems. Although certain characteristics of these proposals are common, they are independent of each other in terms of architecture and operation.

In this paper, firstly we give a general *Connection Admission Control (CAC)* scheme for NGWS. We aim to propose a universal scheme in the sense that it does not depend on the number and type of subsystems included in NGWS. Then, we present an analytical model for the admission control scheme. This is the first analytical model for NGWS in the literature to the best of our knowledge. The complexity of NGWS also creates significant challenges, especially in terms of the size of the state space. We demonstrate these challenges, and develop a neat solution that calculates state probabilities for the case where the system is under high load. We present extensive results from our tests. Our aim in this paper is to develop an analytical model for NGWS and demonstrate how the challenges in the size of the state space can be overcome rather than evaluate the performance of a specific algorithm. We hope that this model will serve as a tool for researchers in the field to analytical evaluate their proposed schemes.

The remainder of this paper is organized as follows. In Section 2, the NGWS network architecture is defined. The connection admission control scheme is detailed in Section 3. The analytical model of Connection Admission Control is given in Section 4. The numerical results are explained in Section 6. Finally, Section 7 concludes this paper.

2. Network architecture

NGWS will be composed of multiple subsystems of different types. Each subsystem will serve as an access network to the users. Since the service areas of the subsystems overlap, the mobile terminals, MT, will have access to multiple subsystems simultaneously. WLANs, PCS, satellite systems, and their future variations together with new wireless systems like 4G Mobile are candidates as subsystems. Our NGWS architecture is flexible and allows all or a subset of these subsystems to be a part of NGWS.

We define NGWS as the tuple

$$\mathcal{NGWS} = (\mathcal{S}, \mathrm{HR}), \tag{1}$$

where $\mathcal{S} = \{wl, pcs, sa, 4g, \ldots\}$ is the set of subsystems, and HR is the global home register for all subsystems $s \in \mathcal{S}$. In Eq. 1, *wl*, *pcs*, *sa*, and $4g$ correspond to WLAN, PCS, Satellite, and 4G Mobile subsystems, respectively. Note that the set $\mathcal{S}$ can be expanded as needed to include other types of wireless subsystems. Such additions will not affect our admission control scheme.

Each subsystem has its own cellular infrastructure. For each subsystem, numerous access nodes are deployed to cover the service area, sometimes partially. The access node is the base station in PCS subsystem, access point in WLAN, transponder in satellite subsystem, etc. We denote the ith access

node of subsystems s as b_i^s. For a given access node b_i^s, we define its cell as the set of locations from which it is possible to communicate with that access node. Thus, each subsystem s splits the service area into cells. Clearly, there is a one-to-one correspondence between the cells and access nodes. We use the term access node for the *device* that provides access for MTs and cell for the set of locations served by the same access node. The size, shape, and location of the cells depend on the location and power of the access node, and the terrain. Therefore, the cellular layouts of the subsystems differ.

We state above that NGWS will provide various types of services such as voice, video, and multiple types of data. We denote the set of connection classes as

$$\mathscr{C} = \{\text{voice, video, low bandwidth data, high bandwidth data, } \ldots\}.$$

Each type of connection class has different requirements ranging from bandwidth to end-to-end latency. For the sake of simplicity, we consider only the bandwidth requirement and assume that reasonable values for the other requirements can be achieved if enough bandwidth is provided. We denote the bandwidth requirement of a class k connection as $bw(k)$.

The overlaid structure of NGWS implies that MT has access to multiple subsystems simultaneously. MT knows the set of subsystems it can access by scanning the pilot signals from the access nodes. The selection of the subsystem s for connection establishment depends on several factors:

- *Subsystem accessibility:* The pilot signal of subsystem s must be strong enough for communication.
- *Resource availability:* The load of access node b_i^s, l_i^s, must not exceed the capacity of b_i^s, c_i^s, if the request is admitted.
- *Service class and user preferences:* The user is able to indicate which subsystem he[1] prefers for each connection class. We denote the probability that MT prefers subsystem s for a connection of class k as $p(k,s)$ where $k \in \mathscr{C}$.

In NGWS, there is a *home register* HR that serves all subsystems in mobility and connection management. The home register resides outside the subsystems, as a node in the Internet (Fig. 1). The function of HR is to store static and dynamic information about all registered users. HR has a different interface for each type of subsystem and acts as the home agent for WLAN networks, as the HLR for the PCS networks, etc.

An example configuration of NGWS consisting of satellite, PCS, and WLAN subsystems is depicted in Fig. 1. The service area is covered with overlapping cells of different subsystems. The coverage area of the subsystems may be discontinuous, as in the case of WLAN. Thus, the set of subsystems to which a mobile terminal can access at a given moment varies. Each subsystem has its own local register, LR, and the backbone traffic between the subsystems is carried over the Internet. The global HR serves all subsystems.

In current wireless systems, the central location register is queried for the user information with the user's specific address as the index. The mobile phone number in PCS systems, and the home address in Mobile IPv6 are used for addressing the user in the corresponding systems. In NGWS, the information about a user can be retrieved from the home register by using the user's *next generation user address* (NGUA) as the index. NGUA is the only address visible to the human users, independent of the subsystem in which the mobile terminal resides. It is the responsibility of HR to translate the NGUA specified by the caller party to the address (e.g., phone number, home IP address, etc.) in the relative subsystem.

3. Next generation connection admission control (NGCAC) scheme

In a wireless network, it is the connection admission control scheme that decides whether connection and handoff requests will be accepted. For new connection requests, NGWS must select the appropriate subsystem for connection establishment. Furthermore, the mobility of a user may require changes in the use of wireless resources, resulting in a handoff attempt for the user. It is the duty of the *Next Generation Connection Admission Control* (NGCAC) scheme to manage the connection requests and handoff attempts in a way that maximizes network utilization, minimizes outage, and distributes the load between subsystems.

The connection admission control scheme is triggered in three cases:

[1] Throughout this paper, "he" should be read as "he or she", "his" as "his or her", etc.

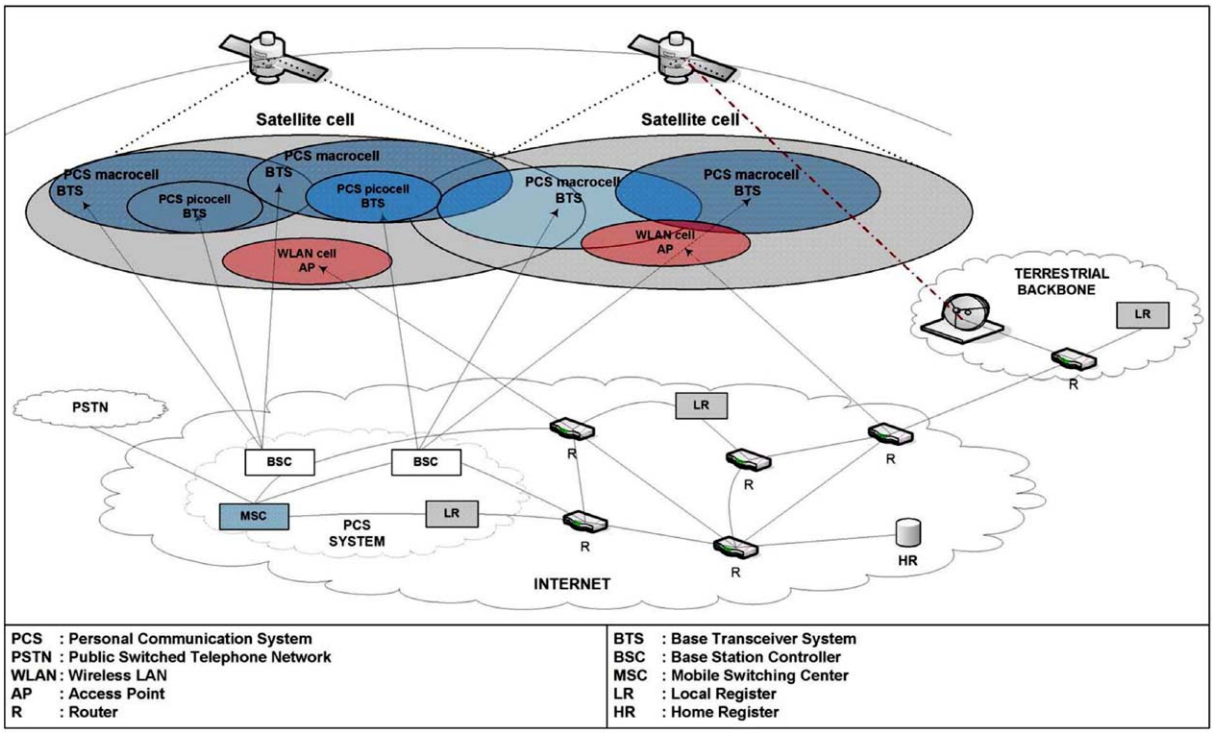

Fig. 1. An example interconnection of subsystems in NGWS.

- Outgoing connection request: When a user initiates a connection request.
- Incoming connection request: When a remote (mobile or fixed) user initiates a connection request destined at the mobile user.
- Handoff request: When a mobile user with an active connection crosses a cell boundary.

3.1. Problem statement and proposed solution

Since MT has access to multiple subsystems simultaneously, NGWS must select one of the subsystems for connection. Among the accessible subsystems, one subsystem that can accommodate the connection request will be selected, subject to connection class and user preferences.

We define the vicinity of b_i^s as

$$\{b_j^t | t \neq s \text{ and } b_i^s \cap b_j^t \neq \emptyset\},$$

and call all access nodes in the vicinity as neighbors. Each access node b_i^s periodically transmits its load information l_i^s to all of its neighbors, and also keeps record of their loads. We denote the recorded value of l_j^t at access node b_i^s as $l_j''^t$. Since the load information is exchanged between only a few access nodes in the vicinity, and the information exchange is performed over abundant wired links, this overhead is negligible. $l_j''^t$ may not be exactly up-to-date, but since load in a cell does not fluctuate wildly, $l_j''^t$ will be reasonably close to l_j^t. We denote the new load of b_i^s if request rq is accepted as $\hat{l}_i^s(rq)$. We also denote the load of b_j^t, based on the recorded value $l_j''^t$, if rq is accepted as $\hat{l}_j''^t(rq)$.

With each connection or handoff request rq, we associate an *ordered list* of accessible access nodes in which ordering criteria is the user's preferences for the connection class of rq. We denote the ordered list of access nodes specified in request rq as $\mathscr{L}_{ac}(rq)$. For outgoing connection setup and handoff requests, MT sends the request to the first access node, b_i^s, in $\mathscr{L}_{ac}(rq)$. However, for incoming connections the initiator (caller) is a remote node that is not aware of the subsystems accessible by MT, availability of the resources in the subsystems, and user preferences for MT. Furthermore, the paging process, which precedes connection establishment, need not be done through the subsystem over which the connection will be established. Therefore, we propose that in the paging reply message, MT specifies $\mathscr{L}_{ac}(rq)$ to be used in

connection admission. Then, connection establishment is performed over the first access node, b_i^s, in $\mathscr{L}_{ac}(rq)$. Since $\mathscr{L}_{ac}(rq)$ contains the identifiers of a few access nodes, its overhead is negligible.

When b_i^s receives request rq, either directly from MT or from a remote node, it checks if the request can be accommodated. If $\hat{l}_i^s(rq)$ remains below capacity c_i^s, b_i^s accepts the request, establishes the connection, and makes the necessary resource allocations. On the other hand, if $\hat{l}_i^s(rq)$ exceeds c_i^s, b_i^s contacts, *on behalf of MT,* all access nodes b_j^t in $\mathscr{L}_{ac}(rq)$ for which $\hat{l}'^t_j(rq) \leqslant c_j^t$. If there exists an access node b_j^t in $\mathscr{L}_{ac}(rq)$ that can accommodate request rq, the request will be transferred to b_j^t, and the connection will be established over that access node. On the other hand, if $\hat{l}'^t_j(rq) > c_j^t$ for all b_j^t in $\mathscr{L}_{ac}(rq)$, request rq will be rejected. In this case, MT may either call off the request or revise the connection class (connection requirements) and resubmit it as a new request. The algorithm for the proposed scheme is presented as a flowchart in Fig. 2.

An example scenario is depicted in Fig. 3 with three subsystems, *sa*, *pcs*, and *wl*. Although the service area is covered by three subsystems, there are *holes* where no signal is received from one or more subsystems. Let us assume MT's subsystem preference decreases in the order *wl*, *pcs*, and *sa*. MT initiates a connection at point *A* where it has access to all subsystems, and selects *wl* according to user preferences. As it moves on its trajectory, it encounters an intra-subsystem handoff between two WLAN cells at point *B*. At point *C*, MT enters a hole in the WLAN cell where access to *wl* is lost. Therefore, MT encounters an inter-subsystem handoff at point *C*, from *wl* to *pcs*, preferring *pcs* over *sa*. MT gets out of the hole at point *D*, but it will keep communicating over *pcs* until it reaches point *F*. At point *F*, MT leaves the PCS cell, so it will encounter another inter-subsystem handoff from *pcs* to *sa*. Finally, at point *G*, MT successfully completes its connection.

4. Analytical model of NGCAC scheme

In the analysis of wireless systems, the service area is typically split into cells since the partitioning criteria is the access node that controls the area. However, in the case of NGWS, cellular granularity is too coarse to define a partition since there are multiple subsystems serving the same service area. Therefore, in our model, the service area is partitioned into smaller regions we call physical areas.

4.1. System definition

Let $\mathscr{B}$ be the set of all access nodes in all subsystems in the service area Ω. We denote the set of all subsets of $\mathscr{B}$ as $\mathscr{A} = 2^{\mathscr{B}}$. We define $\tilde{b} : \Omega \to \mathscr{A}$ such that $\tilde{b}(x)$ is the set of all access nodes reachable from location x. This mapping introduces an equivalence relation $\sim$ on Ω such that $x_1 \sim x_2$ if $\tilde{b}(x_1) = \tilde{b}(x_2)$.

We call the elements of $\mathscr{A}$ as *areas*. For every area $a \in \mathscr{A}$, we define a physical area $\bar{a} \subset \Omega$ as

$$\bar{a} = \{x \in \Omega | \tilde{b}(x) = a\}.$$

It is clear that physical areas are the blocks of the partition induced by "$\sim$". Furthermore, since there exist some $a \in \mathscr{A}$ for which $\bar{a} = \emptyset$, it can be shown that if $\bar{a}_1 = \bar{a}_2 \neq \emptyset$, then $a_1 = a_2$.

We find it more convenient to work with areas rather than with physical areas since it reduces the complexity of the description of the model. We call the areas for which corresponding physical areas are non-empty as *effective areas*. For obvious reasons, we are interested only in the effective areas. The upper bound on the number of effective areas is given by $\binom{|\mathscr{B}|}{K}$ where K is the maximum number of access nodes reachable from one location.[2] The partitioning of the service area and the relationship between areas and cells according to the coverage in Fig. 3 is shown in Fig. 4. The number of area crossings on the trajectory of MT provides a reasonable explanation for not initiating a handoff procedure to the most preferred access node at every area boundary.

For each area a_i, we have the following:

- $n_{a_i}(t)$: Number of users in area a_i at time t.
- $V^k_{a_j,a_i}(t)$: Migration rate of a class k connection from area a_j to a_i at time t. Hence, the number of class k migrating from a_j to a_i during the time period τ is equal to
$$\int_t^{t+\tau} n_{a_j} \cdot V^k_{a_j,a_i}(t)\mathrm{d}t.$$
- $r^k_{a_i}(t)$: Connection generation rate of class k connections in area a_i at time t.

[2] From this point on, we will use the term area in the sense of effective area.

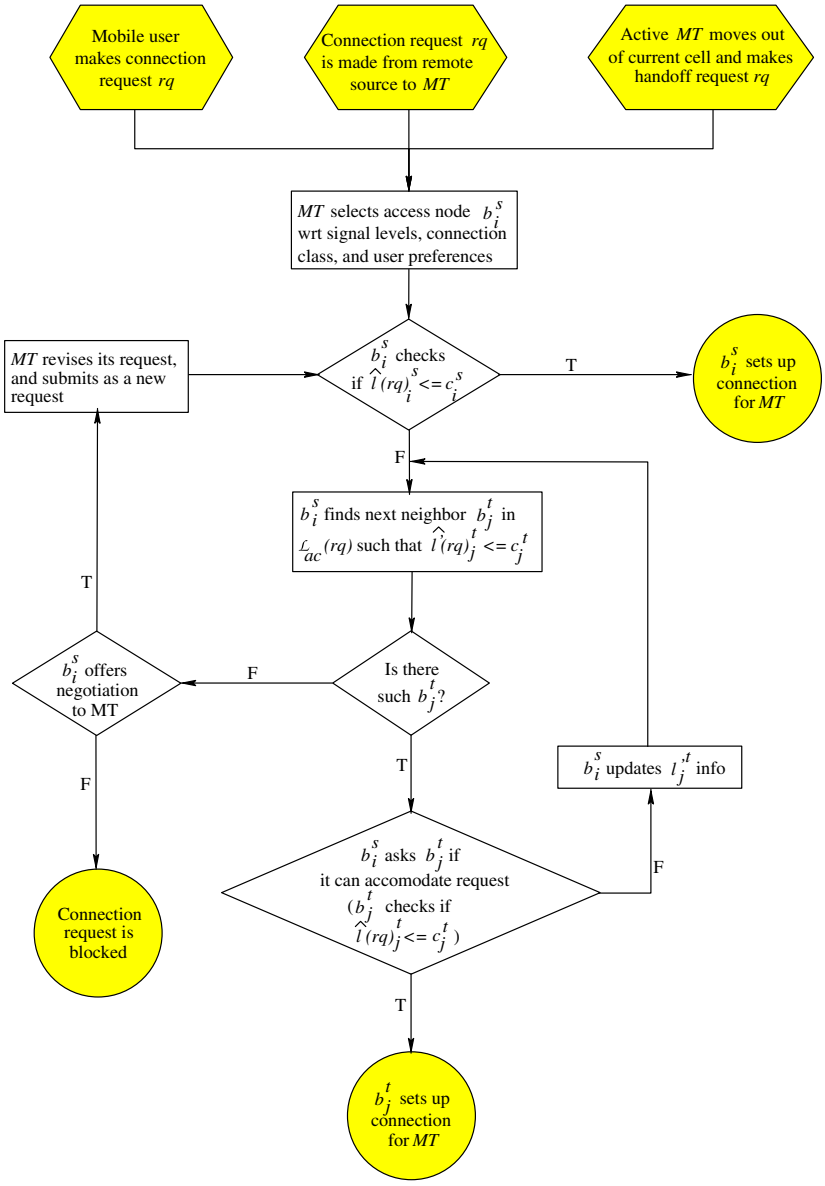

Fig. 2. The algorithm of the connection admission control scheme.

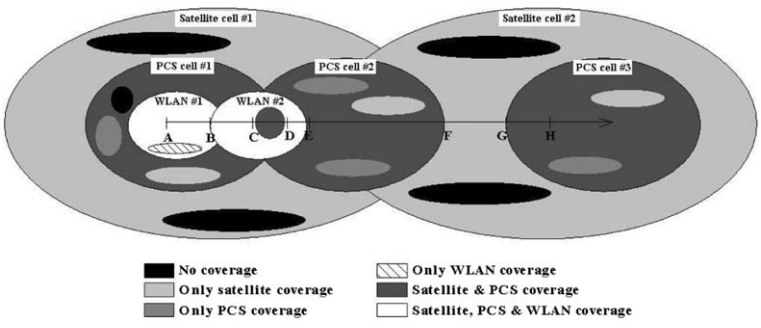

Fig. 3. Connection admission scenario.

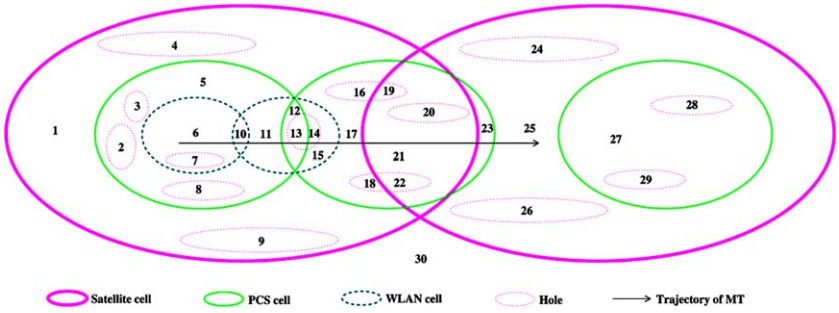

Fig. 4. An example partitioning of the service area.

- $f(a_i,k,t)$: Connection profile of class k connections in area a_i at time t, i.e., the distribution of class k such as 0.60 for voice, 0.10 for multimedia, 0.30 for data.
- $p(a_i,k,s,t)$: Probability that a user in area a_i with a class k connection prefers subsystem SS^s at time t.

We denote the number of active connections of class k communication with access node b in area a as $x_a^k(b)$. The state of the system is any particular distribution of values taken by the variables $x_a^k(b)$. Thus, the state of the system is the tuple

$$g=\begin{bmatrix} x_1^1(1), & x_1^1(2), & \cdots, & x_1^1(|\mathscr{B}|)\\ x_1^2(1), & x_1^2(2), & \cdots, & x_1^2(|\mathscr{B}|)\\ \cdots & & & \\ x_1^{|\mathscr{C}|}(1), & x_1^{|\mathscr{C}|}(2), & \cdots, & x_1^{|\mathscr{C}|}(|\mathscr{B}|)\\ x_2^1(1), & x_2^1(2), & \cdots, & x_2^1(|\mathscr{B}|)\\ \cdots & & & \\ x_2^{|\mathscr{C}|}(1), & x_2^{|\mathscr{C}|}(2), & \cdots, & x_2^{|\mathscr{C}|}(|\mathscr{B}|)\\ \vdots & & & \\ \vdots & & & \\ x_{|\mathscr{A}|}^1(1), & x_{|\mathscr{A}|}^1(2), & \cdots, & x_{|\mathscr{A}|}^1(|\mathscr{B}|)\\ \cdots & & & \\ x_{|\mathscr{A}|}^{|\mathscr{C}|}(1), & x_{|\mathscr{A}|}^{|\mathscr{C}|}(2), & \cdots, & x_{|\mathscr{A}|}^{|\mathscr{C}|}(|\mathscr{B}|) \end{bmatrix}. \quad (2)$$

The state can also be represented as a mapping

$$g:\mathscr{X}\to\mathbb{N}\cup\{0\}, \quad (3)$$

where $\mathscr{X}=\{x_a^k(b)\}$. Thus, a state is any particular distribution of values taken by individual $x_a^k(b)$. We denote the set of all states as $\mathscr{G}$.

Although each variable $x_a^k(b_i^s)$ can assume values up to the capacity of corresponding access node b_i^s, we have

$$\sum_{a\in\mathscr{A}}\sum_{k\in\mathscr{C}} w(k)\cdot x_a^k \leqslant c_i^s$$

since the same wireless resources are shared by all areas covered by the same cell. Therefore, many states in the state space are never visited.

4.2. Elementary events

There are various elementary events that cause the system to switch from one state to another. Keeping in mind that a state of the system is a mapping as given by Eq. 3, each elementary event causes a change in the mapping at one or two points. For example, let us assume that the current state of the system is defined by the mapping g_i, and access node b accepts a new connection request of class k from area a. This new connection causes $g_i(x_a^k(b))$ to be incremented by one, defining a new mapping g_j. In Fig. 5, an example in which $x_5^1(2)$ is incremented due to a new connection. Two mappings, g_i and g_j are the same except at the point marked with an arrow.

The elementary events are classified under three groups as follows.

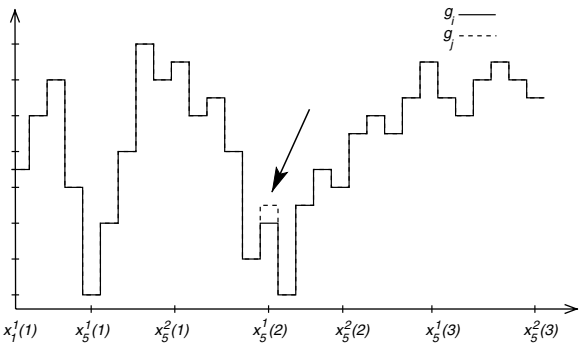

Fig. 5. State transition due to a new connection of class 1 from area a_5 over access node 2.

4.2.1. New connection events

4.2.1.1. Outgoing new connection. MT initiates a new connection request of class k in area a, and sends request rq directly to b_i^s, the first access node in $\mathscr{L}_{ac}(rq)$.

- Direct outgoing new connection:
  In this case, b_i^s accepts rq since $\hat{l}_i^s(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_i^s)$:
  $$g_j(x_a^k(b_i^s)) = g_i(x_a^k(b_i^s)) + 1.$$
- Indirect outgoing new connection:
  In this case, b_i^s cannot accommodate rq since $\hat{l}_i^s(rq)$ exceeds c_i^s. b_i^s checks $\mathscr{L}_{ac}(rq)$ for the first access node b_j^t such that $\hat{l'}_j^t(rq) \leqslant c_j^t$. Note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b_j^t for connection setup. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_j^t)$:
  $$g_j(x_a^k(b_j^t)) = g_i(x_a^k(b_j^t)) + 1.$$

4.2.1.2. Incoming new connection. MT receives a paging request from a remote source for a class k connection while in area a. MT specifies $\mathscr{L}_{ac}(rq)$ in the paging reply, and the source initiates the connection setup procedure over b_i^s, the first access node in $\mathscr{L}_{ac}(rq)$.

- Direct incoming new connection:
  In this case, b_i^s accepts rq since $\hat{l}_i^s(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_i^s)$:
  $$g_j(x_a^k(b_i^s)) = g_i(x_a^k(b_i^s)) + 1.$$
- Indirect incoming new connection:
  In this case, b_i^s cannot accommodate rq since $\hat{l}_i^s(rq)$ exceeds c_i^s. b_i^s checks $\mathscr{L}_{ac}(rq)$ for the first access node b_j^t such that $\hat{l'}_j^t(rq) \leqslant c_j^t$. Note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b_j^t for connection setup. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_j^t)$:
  $$g_j(x_a^k(b_j^t)) = g_i(x_a^k(b_j^t)) + 1.$$

4.2.2. Migration events

4.2.2.1. Intra-cell movement. In this case, MT with an active connection of class k over access node b_i^s, in area a_u moves to area a_v, which is covered by the same access node. Therefore, handoff will not occur, but the system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_i^s)$ and $x_{a_v}^k(b_i^s)$:

$$g_j(x_{a_u}^k(b_i^s)) = g_i(x_{a_u}^k(b_i^s)) - 1,$$
$$g_j(x_{a_v}^k(b_i^s)) = g_i(x_{a_v}^k(b_i^s)) + 1.$$

4.2.2.2. Intra-subsystem handoff. In this case, MT with an active connection of class k over access node b_i^s, in area a_u moves to area a_v, which is covered by b_j^s but not by b_i^s. Therefore, MT encounters a handoff from b_i^s to b_j^s, within subsystem s. The system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_j^s)$ and $x_{a_v}^k(b_j^s)$:

$$g_j(x_{a_u}^k(b_i^s)) = g_i(x_{a_u}^k(b_i^s)) - 1,$$
$$g_j(x_{a_v}^k(b_j^s)) = g_i(x_{a_v}^k(b_j^s)) + 1.$$

4.2.2.3. Inter-subsystem handoff. In this case, MT with an active connection of class k over access node b_z^w, in area a_u moves to area a_v, which is not covered by any access node of subsystem w. To continue uninterrupted service, MT sends a handoff request directly to b_i^s, the first access node in $\mathscr{L}_{ac}(rq)$.

- Direct inter-subsystem handoff:
  In this case, b_i^s accepts rq since $\hat{l}_i^s(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_z^w)$ and $x_{a_v}^k(b_i^s)$:
  $$g_j(x_{a_u}^k(b_z^w)) = g_i(x_{a_u}^k(b_z^w)) - 1,$$
  $$g_j(x_{a_v}^k(b_i^s)) = g_i(x_{a_v}^k(b_i^s)) + 1.$$
- Indirect inter-subsystem handoff:
  In this case, b_i^s cannot accommodate rq since $\hat{l}_i^s(rq)$ exceeds c_i^s. b_i^s checks $\mathscr{L}_{ac}(rq)$ for the first access node b_j^t such that $\hat{l'}_j^t(rq) \leqslant c_j^t$. Note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b_j^t for handoff. The system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_z^w)$ and $x_{a_v}^k(b_j^t)$:
  $$g_j(x_{a_u}^k(b_z^w)) = g_i(x_{a_u}^k(b_z^w)) - 1,$$
  $$g_j(x_{a_v}^k(b_j^t)) = g_i(x_{a_v}^k(b_j^t)) + 1.$$

4.2.3. Hangup event

In this case, MT with an active connection of class k in area a over access node b_i^s terminates the

connection voluntarily. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_i^s)$: $g_j(x_a^k(b_i^s)) = g_i(x_a^k(b_i^s)) - 1$.

4.3. Transition graphs

We introduce *transition graphs* to explain how the next state is found if the present state is given. There is an individual graph $\Gamma(e,s)$ for every elementary event e and every subsystem s. All transition graphs have the same vertex set $\mathscr{G}$, the set of all states. The set of arcs, $\mathscr{V}$, depends on e and s. We define the set of graphs associated with event e as $\Gamma_e : \mathscr{G}, \mathscr{V}, e, \mathscr{S}$. The number of states adjacent to a given state is of the order of $|\mathscr{X}|$.

We explain the transition graphs with an example in which subsystems s_1, s_2, and s_3 are accessible in area a. The received signal levels from the access nodes of these subsystems will be different at every point in a. We denote the probability that the received signal from the access node of subsystem s_n is the strongest for any randomly selected user in area a as q_n. We examine the transition graph for the *outgoing new connection* event. Depending on the state, we have the $\{s_1,s_2,s_3\}$, $\{s_1,s_2\}$, $\{s_1,s_3\}$, $\{s_2,s_3\}$, $\{s_1\}$, $\{s_2\}$, $\{s_3\}$, and $\emptyset$ cases where $\{x,y,z\}$ represents the case that corresponding cells of the subsystems x, y, and z have enough resources, and $\emptyset$ represents the case that none of the subsystems has enough resources. The outgoing arcs in the transition diagrams that correspond to the $\{s_1,s_2,s_3\}$, $\{s_1,s_3\}$, and $\emptyset$ cases are depicted in Fig. 6(a)–(c), respectively. In Fig. 6(a) for the $\{s_1,s_2,s_3\}$ case, the flow out of g_i due to outgoing new connections is split into three according to the ratios q_1, q_2, and q_3 toward the states g_j, g_k, and g_l. However, in Fig. 6(b) for the $\{s_1,s_3\}$ case, the flow out of g_i is split into two according to the ratios $\frac{q_1}{q_1+q_3}$ and $\frac{q_3}{q_1+q_3}$ toward the states g_j and g_l.

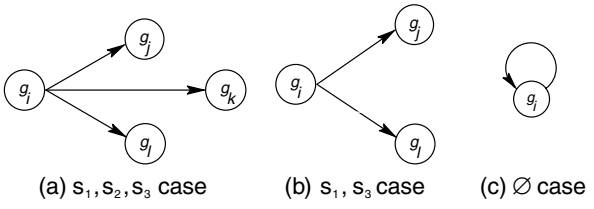

Fig. 6. Outgoing arcs for state g_i in the transition graph.

4.4. Transition probabilities

Transitions between states occur when MT initiates connections, moves in the service area, and hangups. Due to the complexity of the NGCAC scheme, we examine each case separately.

4.4.1. New connection events

4.4.1.1. Outgoing new connection

- Direct outgoing new connection:
  The probability that a direct outgoing new connection attempt of class k occurs for access node b is

$$P_{no}(b,k,\Delta t) = \Delta t \cdot \sum_{a\in\mathscr{A}(b)} \{r_a(t)\cdot f(a,k,t) \cdot n_a(t)\cdot p(a,k,s(b),t)\} + o(\Delta t), \tag{4}$$

  where $\mathscr{A}(b)$ is the set of areas constituting the cell of access node b, $r_a(t)$ is the connection generation rate of all connection classes in area a at time t, and $s(b)$ is the subsystem to which access node b belongs.
- Indirect outgoing new connection:
  The probability that an indirect outgoing new connection attempt of class k occurs for access node b is

$$\widetilde{P}_{no}(b,k,\Delta t) = \Delta t \cdot \sum_{a\in\mathscr{A}(b)} \{r_a(t)\cdot f(a,k,t) \cdot n_a(t)\cdot \alpha(b)\} + o(\Delta t), \tag{5}$$

  where

$$\alpha(b) = \sum_{\substack{u=1\\u\neq b}}^{|\mathscr{B}|} \left(\sum_{\substack{v=1\\v\neq b}}^{u} R^k(v)\cdot p_v \right) \cdot \frac{p_b}{\sum_{w=1}^{|\mathscr{B}|} p_w - \sum_{v=1}^{u} p_v} \tag{6}$$

  represents the total probability that MT prefers other access nodes serving the same area, but those access nodes cannot accommodate the request due to lack of resources, and $R^k(v)$ is the probability that access node v rejects a request of class k, i.e.,

$$R^k(v) = P\left(\left\{ \sum_{k\in\mathscr{C}} \sum_{a\in\mathscr{A}(b)} x_a^k(v) \right\} + bw(k) > C(v) \right),$$

where $C(v)$ is the capacity allocated to access node v.

4.4.1.2. Incoming new connection

- Direct incoming new connection:
  The probability that a direct outgoing new connection attempt of class k occurs for access node b is

$$P_{ni}(b,k,\Delta t) = \Delta t \cdot \sum_{a\in\mathscr{A}(b)} \{r_a^k(t)\cdot n_a(t) \cdot p(a,k,s(b),t)\} + o(\Delta t). \quad (7)$$

- Indirect incoming new connection:
  The probability that a direct outgoing new connection attempt of class k occurs for access node b is

$$\widetilde{P}_{ni}(b,k,\Delta t) = \Delta t \cdot \sum_{a\in\mathscr{A}(b)} \{r_a^k(t)\cdot n_a(t) \cdot \alpha(b)\} + o(\Delta t). \quad (8)$$

Thus, the probability that a new connection attempt of class k occurs for access node b is

$$P_{\text{new}}(b,k,\Delta t) = P_{no}(b,k,\Delta t) + \widetilde{P}_{no}(b,k,\Delta t) + P_{ni}(b,k,\Delta t) + \widetilde{P}_{ni}(b,k,\Delta t) + o(\Delta t). \quad (9)$$

4.4.2. Migration events

4.4.2.1. Intra-cell movement. The probability that a mobile that has an active connection of class k over access node b moves from one area to another in the same cell without changing its access node is

$$P_{\text{intra}}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j\in\mathscr{A}(b)} \sum_{\substack{a_i\in\mathscr{A}(b)\\ a_i\neq a_j}} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(b) + o(\Delta t). \quad (10)$$

4.4.2.2. Intra-subsystem handoff. The probability that access node b receives a handoff request of class k from another cell in the same subsystem, $s(b)$, is

$$P_{\text{intra}}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j\in\mathscr{A}} \sum_{a_i\in\mathscr{A}(b)} \sum_{\substack{c\in\mathscr{B}\\ s(c)=s(b)\\ c\neq b}} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(c) + o(\Delta t). \quad (11)$$

4.4.2.3. Inter-subsystem handoff

- Direct inter-subsystem handoff:
  The probability that access node b receives a direct handoff request of class k from another cell of another subsystem, $s(b)$, is

$$P_{\text{inter}}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j\in\mathscr{A}} \sum_{a_i\in\mathscr{A}(b)} \sum_{\substack{c\in\mathscr{B}\\ s(c)\neq s(b)}} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(c)\cdot p(a_j,k,s(b)) + o(\Delta t). \quad (12)$$

- Indirect inter-subsystem handoff:
  The probability that access node b receives an indirect handoff request of class k from another cell of another subsystem, $s(b)$, is

$$\widetilde{P}_{\text{inter}}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j\in\mathscr{A}} \sum_{a_i\in\mathscr{A}(b)} \sum_{\substack{c\in\mathscr{B}\\ s(c)\neq s(b)}} V^k_{a_j,a_i}(t) \cdot x^k_{a_i}(c)\cdot \alpha(b) + o(\Delta t). \quad (13)$$

Thus, the probability that a handoff attempt of class k occurs for access node b is

$$P_{\text{handoff}}(b,k,\Delta t) = P_{\text{intra}}(b,k,\Delta t) + P_{\text{inter}}(b,k,\Delta t) + \widetilde{P}_{\text{inter}}(b,k,\Delta t) + o(\Delta t). \quad (14)$$

4.4.3. Hangup event

The probability that MT terminates the connection voluntarily is

$$P_{\text{hangup}}(b,k,\Delta t) = \Delta t \cdot \sum_{a\in\mathscr{A}(b)} h_a^k(t)\cdot x_a^k(t) \cdot p(a,k,s(b),t) + o(\Delta t), \quad (15)$$

where $h_a^k(t)$ is the rate at which class k connections in area a hangup at time t.

5. Calculating state probabilities

5.1. Analytical approach

Analytically calculating the state probabilities for such a complex system is a challenge on its own. The state probabilities in our dynamic system can be calculated by solving the eigenvalue problem of the transition matrix.

We denote the probability that the system is in state g_i at time t as $\mathbf{P}_i(t)$. Initially, the system starts in state g_0 in which there are no ongoing connections. Thus, $\mathbf{P}_0(0)$, the probability of being in state g_0 at time 0, is equal to 1.0 while $\mathbf{P}_i(0) = 0.0$ for all other states g_i. As time elapses, the elementary events mentioned in Section 4.2 will occur, carrying the system from state to state. This will cause the state probabilities to change in time.

Though we do not know the system state exactly at a given time, we can calculate the probability of being in each state g_i at time t, literally $\mathbf{P}_i(t)$. During an infinitesimally short time interval Δt, the system may switch to/from g_i due to an elementary event. Therefore, $\mathbf{P}_i(t+\Delta t)$ will be different from $\mathbf{P}_i(t)$. The rate of change in state probability of g_i during Δt, $\mathbf{P}'_i(t)$, is given by

$$\mathbf{P}'_i(t) = \mathbf{P}_i(t) \cdot \left(-\sum_{j \neq i} \lambda_{ij} \right) + \sum_{j \neq i} \mathbf{P}_j(t) \cdot \lambda_{ji}, \tag{16}$$

where λ_{ij} is the rate of flow from state g_i to g_j. λ_{ij} values are calculated using Eqs. (4)–(15) depending on the state transitions triggered by each elementary event.

In theory, when the system reaches equilibrium we have

$$\forall g_i \in \mathcal{G} \quad \mathbf{P}'_i(t) = 0. \tag{17}$$

From Eq. 17, we derive a method to solve the state probabilities with the assumption that the system is in equilibrium. We use the power method, which is an iterative process, for solving state probabilities. In each iteration, probabilities converge to equilibrium state probabilities that correspond to the eigenvector of largest eigenvalue of the transition matrix.

In the power method, we start with an initial probability vector that sums up to one. We represent the vector of state probabilities, i.e., the vector composed of state probabilities of all states at time t, as $[\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}$. Then we start the iterations, correcting the state probabilities with the assistance of the transition matrix in each step. In each iteration we evaluate new probabilities for the next iteration according to Eq. 18. After sufficient iterations, state probabilities converge to the steady state probabilities. We use the metric d in Eq. 19, to represent the difference between two probability vectors:

$$\mathbf{P}_i(t) = \frac{\sum_{j \neq i} \mathbf{P}_j(t) \cdot \lambda_{ji}}{\sum_{j \neq i} \lambda_{ij}}, \tag{18}$$

$$d([\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}, [\mathbf{P}_i(t+\Delta t)]_{g_i \in \mathcal{G}}) = \sum_{g_i \in \mathcal{G}} |\mathbf{P}_i(t) - \mathbf{P}_i(t+\Delta t)|. \tag{19}$$

We utilize metric d to analyze the convergence of the state probability vector. While solving the system numerically, iteration is stopped when d evaluated between successive iterations is less than a threshold. In Eq. 19, d is the metric derived from ℓ_1-norm. Here, we have two probability vectors, $[\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}$ and $[\mathbf{P}_i(t+\Delta t)]_{g_i \in \mathcal{G}}$. We sum the absolute probability difference of all states in these probability vectors. Calculating d between identical vectors gives zero due to non-degeneracy of the metric. The final probability vector in the last iteration is close to the probability vector in equilibrium situation according to the metric in Eq. 19.

5.2. Challenges in analytical approach

The analytical model presented above provides a tool to calculate state probabilities using the power method. However, the size of the state space constitutes a major challenge. Any combination of values for the tuple given in Eq. 2, as long as the access node capacities are not exceeded constitutes a different state. To make the problem more tangible, we consider the simple cellular layout in Fig. 7, where each access node is assigned only 4 channels. We also assume only one connection class for the sake of simplicity. Below, we show that even this example is a challenging scenario.

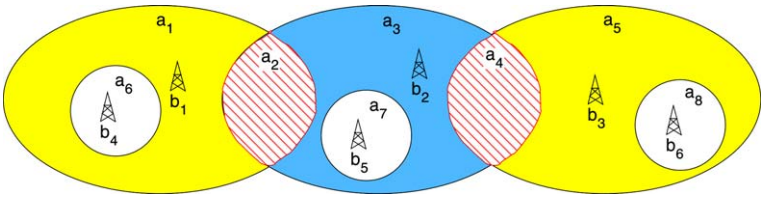

Fig. 7. Cellular layout.

The reader should note that some of the users in a_2 may communicate over b_1 while others communicate over b_2. The sum of the variables $x^1_{a_1}(b_1)$, $x^1_{a_2}(b_1)$, and $x^1_{a_6}(b_1)$ should not exceed 4, the capacity of b_1. Considering such constraints for all variables, we find that there are 201 different combinations for each access node. Since we have 6 access nodes, the number of states reaches $201^6 = 65\,944\,160\,601\,201$. However, some of these states are not effectively possible. For example, the value of the variable $x^1_{a_3}(b_1)$ should always be 0. When we exclude such states and consider only *effective states,* the number of states reduces down to 8 962 362 486 784. Since it is practically impossible to evaluate a model with so many states, we need to find an intelligent way to consider only states that are relevant for modeling purposes.

5.3. Practical approach

Due to challenges mentioned in the previous subsection, we propose another method, which is practical in the number of states considered, subject to some assumptions.

The states that represent the cases where the system is far from high load are of no interest for research. Therefore, we analyze the state probabilities given that the system has high load. We evaluate the state probabilities for *Fr0 states* (no free channel in any cell), *Fr1 states* (only one free channel available in one of the cells), and *Fr2 states* (only two free channels available in all network). The reader should note that *Fr0* is not a single state; since there are multiple combinations for $x^k_a(b)$ variables under full load, all channels can be in use in different ways. Similar idea applies for *Fr1* and *Fr2 states*. Assuming that the system has high load means

$$\sum_{i\in Fr0}\mathbf{P}_i(t)+\sum_{i\in Fr1}\mathbf{P}_i(t)+\sum_{i\in Fr2}\mathbf{P}_i(t)\simeq 1, \tag{20}$$

which implies the sum of the probabilities of the remaining states is very close to zero. This assumption is expressed by Eq. 21.

$$\sum_{i\notin(Fr0\cup Fr1\cup Fr2)}\mathbf{P}_i(t)\simeq 0. \tag{21}$$

By considering only the states in *Fr0*, *Fr1*, and *Fr2*, we reduce the number of states down to 156 800, which is reasonable. Using this approach, we are able to find the probability that the system goes to one of the critical states in *Fr0*, *Fr1*, or *Fr2*, given that it is operating in the range close to full capacity. Results in the next section are produced using this approach.

6. Numerical results

In this section, we analyze the effects of several parameters on system performance. First, we discuss the convergence of the iterations and dropping rate, then argue about the effect of migration rate on the system, and finally discuss the effect of connection generation rate on the system using the practical approach in Section 5.3.

In the rest of this section, we use the state definitions in Eqs. 2 and 3. We do not consider all states, but we consider only the states in *Fr0*, *Fr1*, and *Fr2*. Hence we do not find the exact state probabilities, but calculate the probabilities of states in *Fr0*, *Fr1*, and *Fr2* subject to Eq. 20. We denote these *conditional probabilities* by $\mathbf{P}_c$.

6.1. Iteration parameters

In our tests, we use the following parameters:

Δt	Time interval
ε	Threshold for convergence
MR	Migration rate
CGR	Connection generation rate
HUP	Hangup rate
NUSR	Number of users

To allow the system load to stabilize, we keep HUP equal to CGR in the tests.

6.2. Convergence with respect to d

We first analyze how the system converges to steady state. For the system to converge with respect to metric d, the change in state probabilities should decrease as the number of iterations increase.

In Fig. 8, we analyze convergence with different values of Δt. The x-axis represents iterations and the y-axis represents the metric evaluated for two probability vectors formed in successive iterations. For larger time increments, the state probabilities change faster in the beginning. However, the system converges around 3500–4000 iterations for all cases.

With different time increments the system converges to the same $\mathbf{P}_c$ values. Table 1 illustrates that the system converges to approximately the same $\mathbf{P}_c$ value for different Δt. We have used a time increment of 0.05 s in the rest of the tests.

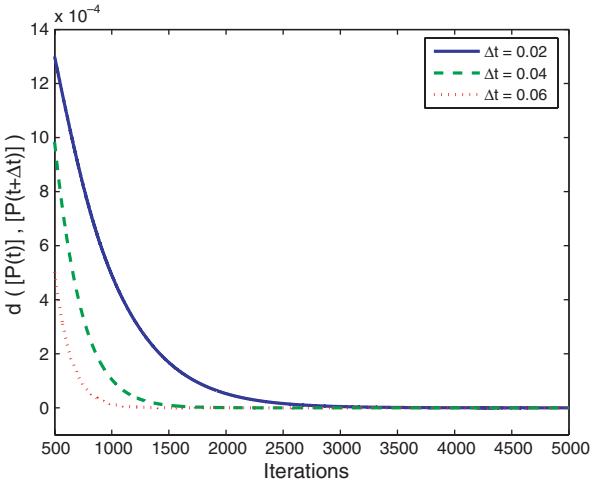

Fig. 8. Effect of Δt on convergence with respect to d (MR = 0.1, CGR = 0.25, NUSR = 50, HUP = 0.25).

Table 1
$\mathbf{P}_c$ values for different Δt

	$\mathbf{P}_c[Fr0]$	$\mathbf{P}_c[Fr1]$	$\mathbf{P}_c[Fr2]$
$\Delta t = 0.02$	0.03559912	0.22212804	0.74231886
$\Delta t = 0.04$	0.03559909	0.22212781	0.74231946
$\Delta t = 0.06$	0.03559913	0.22212803	0.74231922

6.3. Convergence of $\mathbf{P}_c$(*dropping*)

In addition to convergence with respect to metric d, we also evaluate the conditional probability of dropping active connections, $\mathbf{P}_c$(*dropping*). In Fig. 9, the x-axis represents the number of iterations and the y-axis represents the conditional probability of dropping. The tests are performed using different values for migration rate, MR. It is apparent from the figure that increasing the migration rate also increases $\mathbf{P}_c$(*dropping*). Furthermore, we observe that the tests converge around *1000* iterations for all values of MR.

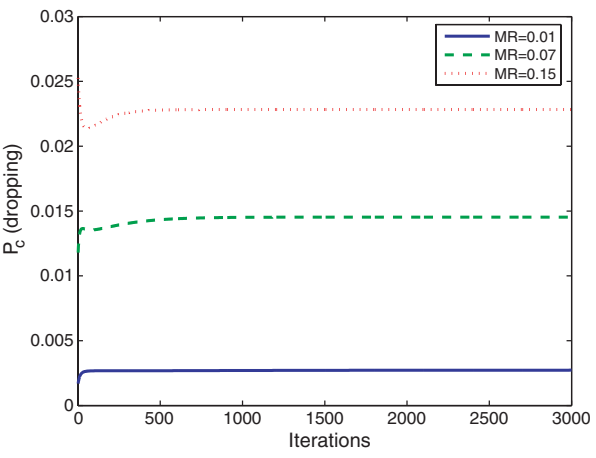

Fig. 9. Effect of MR on convergence of $\mathbf{P}_c$(*dropping*) ($\Delta t = 0.05$, CGR = 0.3125, NUSR = 50, HUP = 0.3125).

6.4. Effect of migration rate

Migration rate changes the dynamics of the system. To analyze the effect of migration rate, we apply the practical approach with different MR values. We sum $\mathbf{P}_c$ of states in *Fr0*, *Fr1*, and *Fr2* separately to analyze where the system inclines to. For example, higher $\mathbf{P}_c$ value for *Fr2* means that the system mostly operates in *Fr2* states, when it is highly loaded. In other words, having two channels available in the whole network is more probable than having no channel under high load assumption. We examine the effects of MR using single and multiple CGR values in the following sections.

6.4.1. Effect of MR for single connection generation rate

In Fig. 10, the x-axis represents MR and the y-axis represents the sum of $\mathbf{P}_c$ for *Fr0*, *Fr1*, and *Fr2* states. For these tests, we fix the value of CGR and vary MR. For each MR value, probability vector converges to the equilibrium vector. Summing up $\mathbf{P}_c$ values for *Fr0*, *Fr1*, and *Fr2* allows us to grasp where the system operates. Desired operation mode is where $\mathbf{P}_c[Fr2]$ is higher than $\mathbf{P}_c[Fr1]$ and $\mathbf{P}_c[Fr0]$. If $\mathbf{P}_c[Fr2]$ value is the highest, then it means probability of being in one of the states in *Fr2* is higher than being in one of the states in *Fr1* or *Fr0*.

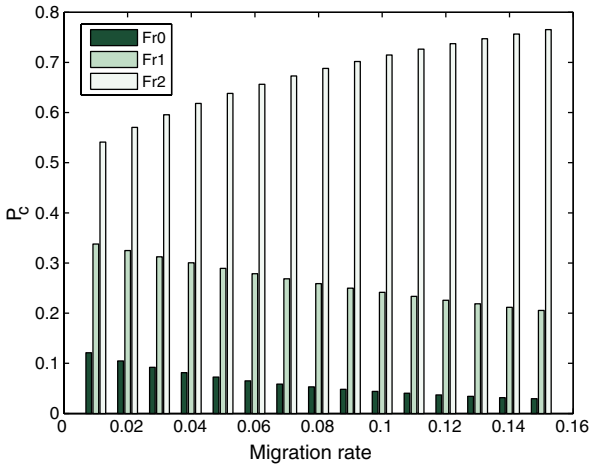

Fig. 10. Effect of migration rate on $\mathbf{P}_c$ ($\Delta t = 0.05$, CGR = 0.3125, NUSR = 50, HUP = 0.3125).

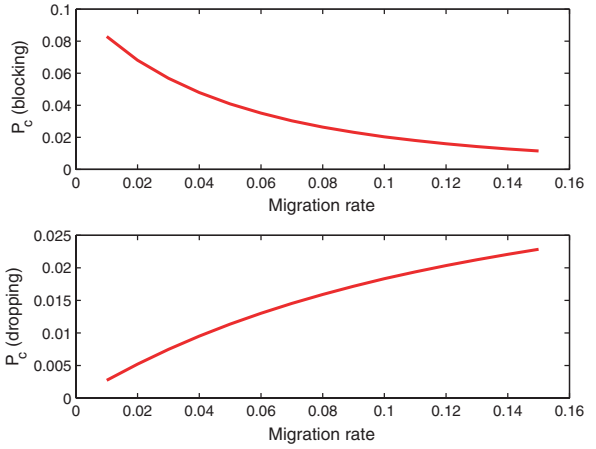

Fig. 11. Effect of migration rate on $\mathbf{P}_c(blocking)$ and $\mathbf{P}_c(dropping)$ ($\Delta t = 0.05$, CGR = 0.3125, NUSR = 50, HUP = 0.3125).

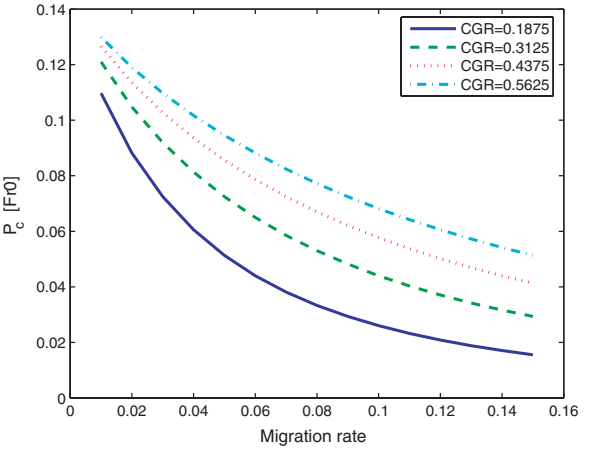

Fig. 12. Effect of migration rate on $\mathbf{P}_c[Fr0]$ with multiple CGR values ($\Delta t = 0.05$, NUSR = 50).

Inclination of the system towards *Fr2* can be analyzed by Figs. 10 and 11. When MR increases, $\mathbf{P}_c$ value of *Fr2* increases and $\mathbf{P}_c$ values of *Fr1* and *Fr0* decrease. This means the system moves towards lower load as MR values increases. Though this behavior seems strange at first, we can explain it by analyzing $\mathbf{P}_c(dropping)$. Fig. 11 depicts dropping and blocking with respect to migration rate. As the migration rate increases, many calls will be dropped due to high load. Thus, $\mathbf{P}_c(dropping)$ increases with increasing MR. With every dropped connection, a channel is released in the system. Therefore, the system is inclined towards *Fr2* states as MR increases. Hence, there are more channels than *Fr0*, therefore system has room for new connection competitors, resulting in decreasing $\mathbf{P}_c(blocking)$.

6.4.2. Effect of MR for multiple connection generation rates

We also analyze effect of MR for different CGR values on $\mathbf{P}_c[Fr0]$, $\mathbf{P}_c[Fr1]$, and $\mathbf{P}_c[Fr2]$ individually. First we analyze $\mathbf{P}_c[Fr0]$ and $\mathbf{P}_c(blocking)$ together. In Fig. 12, the *x*-axis represents MR values, the *y*-axis represents sum of $\mathbf{P}_c$ values of states in *Fr0*, and each curve corresponds to a different CGR value. Similar to Fig. 10, as MR value increases, all the curves corresponding to different CGR values decrease. We also observe that tests with higher CGR values result in higher $\mathbf{P}_c[Fr0]$ since there are more connection attempts with higher CGR values. Thus, the system leans toward *Fr0* more than *Fr1* and *Fr2*. Hence we have higher $\mathbf{P}_c(blocking)$ for higher CGR values as seen in Fig. 13.

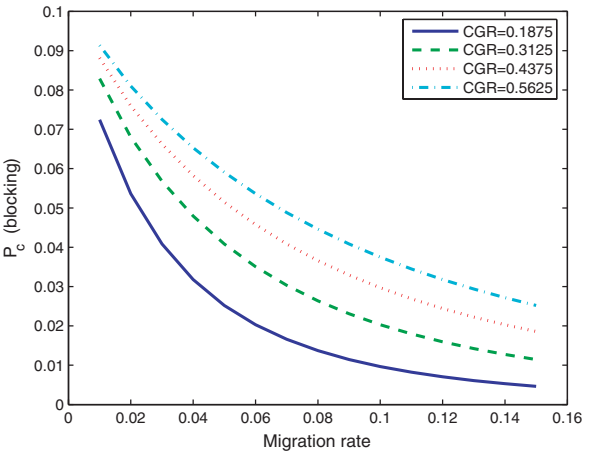

Fig. 13. Effect of migration rate on $\mathbf{P}_c(blocking)$ with multiple CGR values ($\Delta t = 0.05$, NUSR = 50).

Similar behavior is observed for *Fr1* in Fig. 14. However, for *Fr2* the situation changes. In Fig. 15, the *x*-axis represents MR values, the *y*-axis represents $\mathbf{P}_c[Fr2]$, and each curve corresponds to different CGR value. As MR value increases, all the curves corresponding to different CGR values increase. The increase in $\mathbf{P}_c[Fr2]$ can be explained by the increase in $\mathbf{P}_c(dropping)$, as shown in Fig. 16. Dropping events occur more if we increase MR value, so available channels increase frequently in the system. In Fig. 16, the *x*-axis represents MR values, the *y*-axis represents $\mathbf{P}_c(dropping)$, and each curve corresponds to different CGR value. For a fixed MR value we observe that lower CGR values imply, less attempts to fill empty channels in system. Hence, we expect that lower CGR values result in

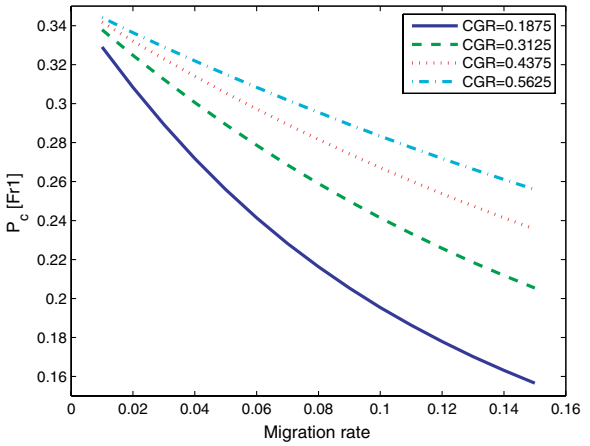

Fig. 14. Effect of migration rate on $\mathbf{P}_c[Fr1]$ with multiple CGR values ($\Delta t = 0.05$, NUSR = 50).

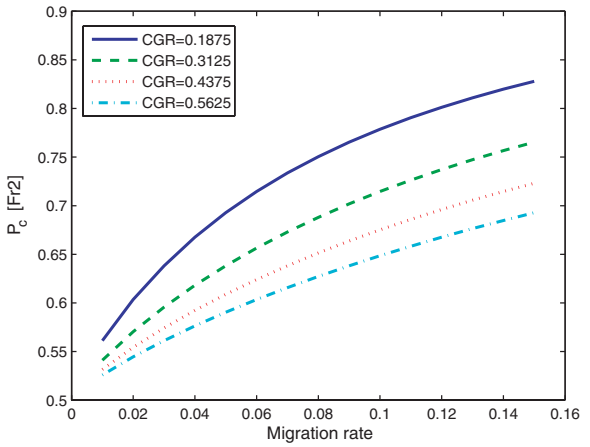

Fig. 15. Effect of migration rate on $\mathbf{P}_c[Fr2]$ with multiple CGR values ($\Delta t = 0.05$, NUSR = 50).

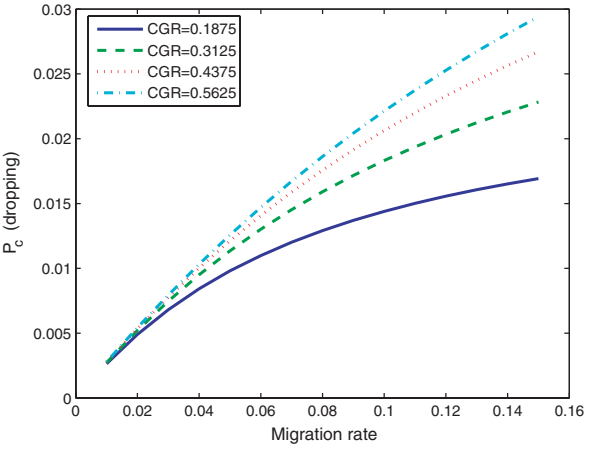

Fig. 16. Effect of migration rate on $\mathbf{P}_c(dropping)$ with multiple CGR values ($\Delta t = 0.05$, NUSR = 50).

lower $\mathbf{P}_c(dropping)$ values. We realize the expectations in Fig. 16.

We analyze the effect of MR for different CGR values, by considering the order of the curves in Figs. 12–14 observe that the conditional probability decreases with increasing MR and lower CGR. In figures, curves corresponding to smaller CGR values are below the others and decreasing. However, in Fig. 15, the conditional probability increases with increasing MR and lower CGR. In figure, the curves corresponding to smaller CGR values are above the others and increasing. When CGR value is lower we expect that $\mathbf{P}_c$ value of being in *Fr0* is lower. In Figs. 12 and 14, curve corresponding to the smallest CGR value is below the others. Due to lower $\mathbf{P}_c$ value of being in *Fr0* and *Fr1* we expect that $\mathbf{P}_c$ value of being in *Fr2* is higher than the others. We observe the positive effect of higher MR value on the system since *Fr2* states are better states than *Fr0* and *Fr1* states. However, we analyze the effect of MR in Fig. 16 on $\mathbf{P}_c(dropping)$ we observe that increasing MR value results in higher $\mathbf{P}_c(dropping)$.

6.5. Effect of connection generation rate

Connection generation rate is another factor that changes the dynamics of the system. To analyze the effect of connection generation rate, we apply the practical approach with different CGR parameter values. We examine the effects of CGR using single and multiple MR values.

6.5.1. Effect of CGR for single migration rate

In Fig. 17, we analyze the effect of connection generation rate on $\mathbf{P}_c$. We vary connection generation rate from 0.05 to 0.75 conn/min, and fix MR. As the connection generation rate increases, we observe that the system shifts from *Fr2* states to *Fr1* and *Fr0* states. Thus, the analytical model provides a tool for the system designer to understand at which load the system goes to *Fr0*.

Fig. 18 demonstrates system behavior under different loads. $\mathbf{P}_c$ values for *Fr0*, *Fr1*, and *Fr2* are shown on the *x*-axis using separate bars for each CGR value. We observe from the figure that the system is mostly in *Fr2* states. As the load increases, the system spends more time in *Fr1* and *Fr0* states. Analyzing the change in $\mathbf{P}_c$ individually for each type of state shows that for *Fr0* and *Fr1* states, $\mathbf{P}_c$ increases as the load increases. However, for *Fr2* states, $\mathbf{P}_c$ decreases as the load increases since there is a shift towards *Fr0* states.

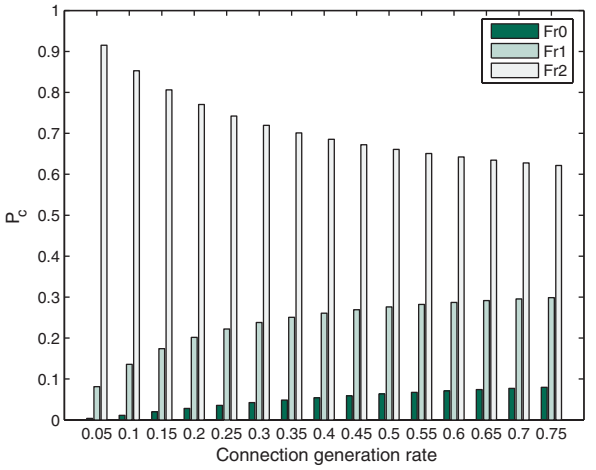

Fig. 17. Effect of connection generation rate on $\mathbf{P}_c$ ($\Delta t = 0.05$, MR = 0.1, NUSR = 50).

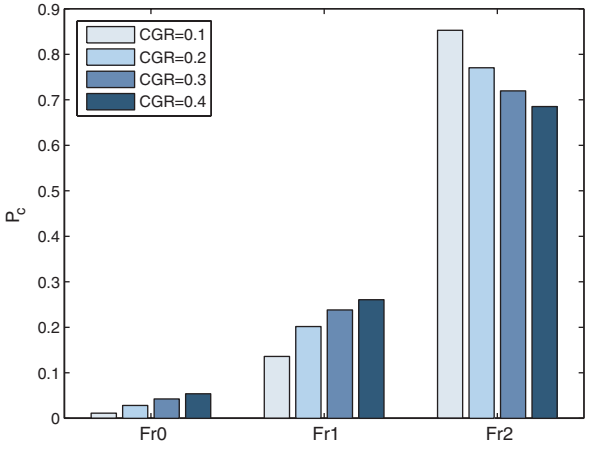

Fig. 18. State type versus $\mathbf{P}_c$ ($\Delta t = 0.05$, MR = 0.1, NUSR = 50).

In Fig. 19, we analyze the effect of CGR on $\mathbf{P}_c(dropping)$ and $\mathbf{P}_c(blocking)$. The x-axis represents CGR, the y-axis represents $\mathbf{P}_c(blocking)$ and $\mathbf{P}_c(dropping)$, respectively. The figure demonstrates the relationship between $\mathbf{P}_c(blocking)$ and $\mathbf{P}_c(dropping)$ while MR is fixed. We observe that the increase the increase in CGR causes an almost linear increase in $\mathbf{P}_c(blocking)$. However, the increase rate of $\mathbf{P}_c(dropping)$ slows down for higher CGR values. This behavior can be explained by the fact that we are operating close to full capacity. For higher CGR values, we know from Fig. 17, that the probability of the system being in *Fr0* and *Fr1* states is higher. Hence, increasing CGR further does not add many more connections into the system

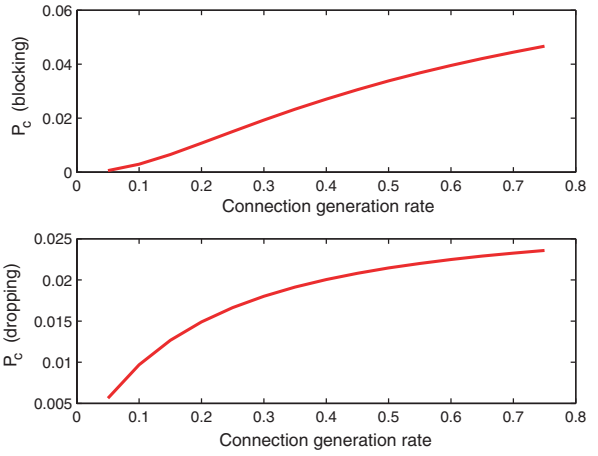

Fig. 19. Effect of connection generation rate on $\mathbf{P}_c(blocking)$ and $\mathbf{P}_c(dropping)$ ($\Delta t = 0.05$, MR = 0.1, NUSR = 50).

since most of these attempts are blocked. Therefore, the increase rate of $\mathbf{P}_c(dropping)$ slows down as CGR increases further.

6.5.2. Effect of CGR for multiple migration rate

We also analyze effect of CGR for different MR values on $\mathbf{P}_c[Fr0]$, $\mathbf{P}_c[Fr1]$, and $\mathbf{P}_c[Fr2]$ individually. In Fig. 20, the x-axis represents CGR values, the y-axis represents sum of $\mathbf{P}_c$ values of states in *Fr0*, and each curve corresponds to a different MR value. The increase in CGR value results in higher $\mathbf{P}_c[Fr0]$ since there are more connection attempts. We also observe that tests with higher MR values result in lower $\mathbf{P}_c[Fr0]$, as explained previously in Section 6.4.

In Fig. 21, the x-axis represents CGR values, the y-axis represents $\mathbf{P}_c(blocking)$, and each curve corresponds to a different MR value. The increase in

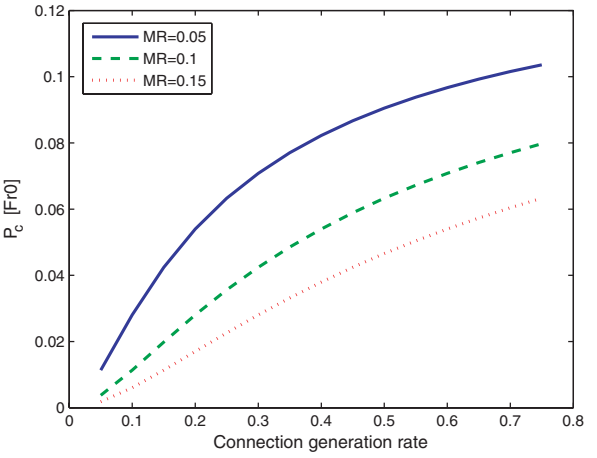

Fig. 20. Effect of connection generation rate on $\mathbf{P}_c[Fr0]$ for multiple MR values ($\Delta t = 0.05$, NUSR = 50).

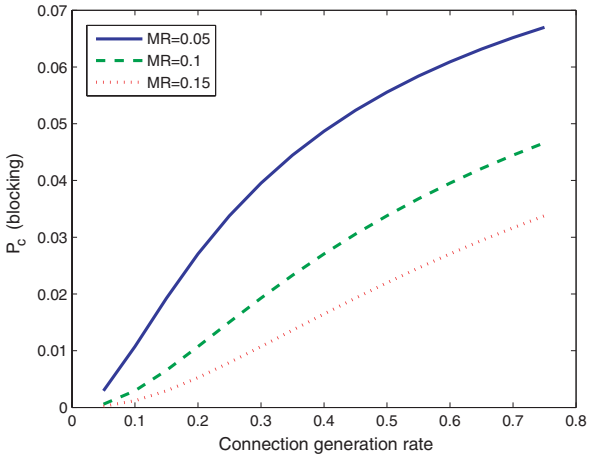

Fig. 21. Effect of connection generation rate on $\mathbf{P}_c(blocking)$ for multiple MR values ($\Delta t = 0.05$, NUSR = 50).

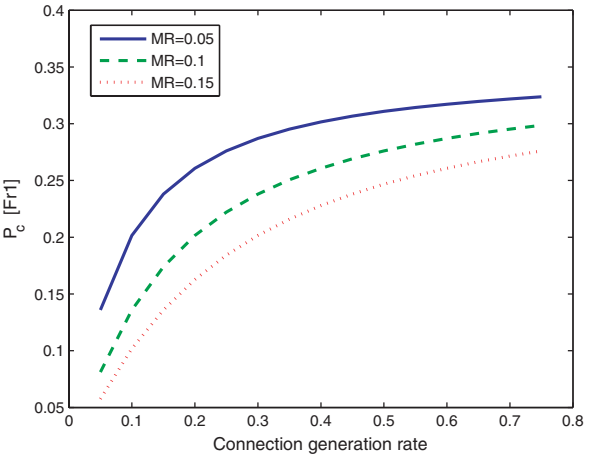

Fig. 22. Effect of connection generation rate on $\mathbf{P}_c[Fr1]$ for multiple MR values ($\Delta t = 0.05$, NUSR = 50).

CGR value results in higher $\mathbf{P}_c(blocking)$ since more connection attempts are blocked. We also observe that tests with higher MR values result in lower $\mathbf{P}_c(blocking)$, as explained previously in Section 6.4. There is a close relation between $\mathbf{P}_c[Fr0]$ and $\mathbf{P}_c(blocking)$ since blocking events occur more frequently when the system is in *Fr0* states. So $\mathbf{P}_c(blocking)$ increases as CGR value increases due to increase in probability of being in *Fr0*. For higher MR values $\mathbf{P}_c(dropping)$ is higher which implies lower $\mathbf{P}_c[Fr0]$ value. Due to lower $\mathbf{P}_c[Fr0]$ value, $\mathbf{P}_c(blocking)$ is lower for higher MR values.

Similar behavior is observed for *Fr1* in Fig. 22. However, for *Fr2* the situation changes. In Fig. 23, the *x*-axis represents CGR values, the *y*-axis represents $\mathbf{P}_c[Fr2]$, and each curve corresponds to different MR value. As CGR value increases, $\mathbf{P}_c[Fr2]$ decreases for all curves. We also observe that tests with higher MR values result in higher $\mathbf{P}_c[Fr2]$ since higher MR values imply higher $\mathbf{P}_c(dropping)$, as explained previously in Section 6.4. We observe the negative effect of higher CGR value on the system, since *Fr2* states are better states than *Fr0* and *Fr1* states.

In Fig. 24, the *x*-axis represents CGR values, the *y*-axis represents $\mathbf{P}_c(dropping)$, and each curve corresponds to different MR value. Dropping events occur more frequently as CGR increases, because channels released due to dropping events are occupied faster. Since the channels occupied faster new dropping events are more probable to happen. However, at high CGR values, the increase rate of $\mathbf{P}_c(dropping)$ slows down as explained in Fig. 19. For a fixed CGR value we observe that higher

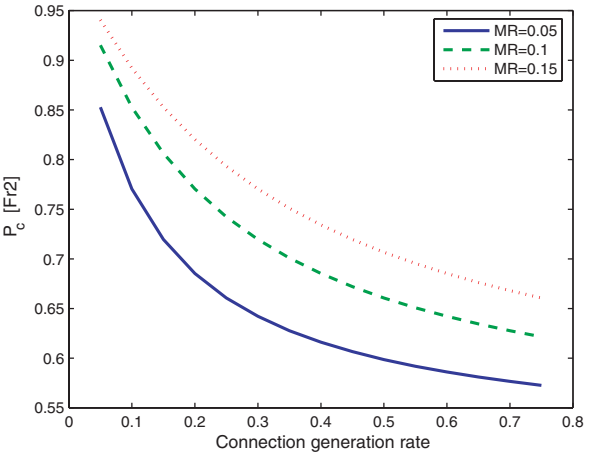

Fig. 23. Effect of connection generation rate on $\mathbf{P}_c[Fr2]$ for multiple MR values ($\Delta t = 0.05$, NUSR = 50).

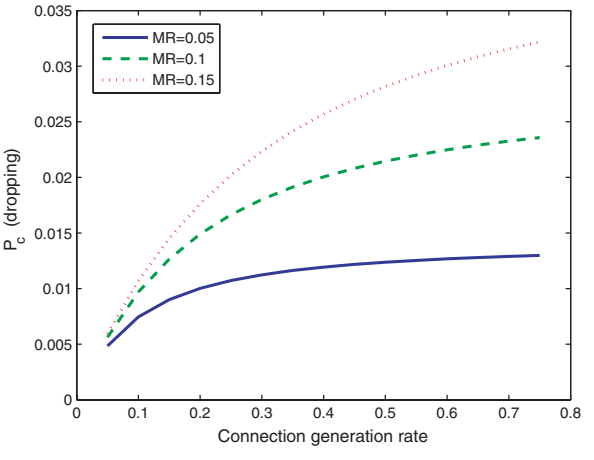

Fig. 24. Effect of connection generation rate on $\mathbf{P}_c(dropping)$ for multiple MR values ($\Delta t = 0.05$, NUSR = 50).

MR values imply, higher $\mathbf{P}_c(dropping)$ as explained previously in Section 6.4.

7. Conclusions and future work

NGWS will be composed of multiple subsystems. The selection of the appropriate subsystem for connection setup is a crucial issue for overall system performance. In this paper, after defining the network architecture, we proposed a novel connection admission control scheme. The proposed NGCAC scheme considers the accessibility of the subsystems and the availability of the resources in those subsystems in addition to connection class of the connection request and the user's preferences. We also provided an analytical model of the proposed scheme. We pointed out major challenges in analytically modeling NGWS and provided a workaround. As future work, we will consider more complex scenarios and evaluate the effects of more parameters on the performance of NGWS.

Acknowledgment

This research is partially supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under grant number 104*E*032 and Bogazici University Research Fund under grant number *BAP*04*S*104.

References

[1] A. Ganz, C.M. Krishna, D. Tang, Z.J. Haas, On optimal design of multitier wireless cellular systems, IEEE Communications Magazine 35 (February) (1997) 88–93.
[2] X. Lagrange, Multitier cell design, IEEE Communications Magazine 35 (August) (1997) 60–64.
[3] N. Faggion, T. Hua, Personal communications services through the evolution of fixed and mobile communications and the intelligent network concept, IEEE Network Magazine (July) (1998) 11–18.
[4] M. Zeng, A. Annamalai, V.K. Bhargava, Recent advances in cellular wireless communications, IEEE Communications Magazine (September) (1999) 128–138.
[5] K. Buchanan, R. Fudge, D. McFarlane, T. Philips, A. Sasaki, H. Xia, IMT-2000: service provider's perspective, IEEE Personal Communications Magazine (August) (1997) 8–13.
[6] S.G. Niri, R. Tafazolli, Cordless-cellular Network Integration for the 3rd Generation Personal Communication Systems, in: Proceedings of the IEEE Vehicular Technology Conference 1, 1998, 402–408.
[7] A. Bria, F. Gessler, O. Queseth, R. Stridh, M. Unbehaun, J. Wu, J. Zander, M. Flament, 4th-Generation wireless infrastructures: scenarios and research challenges, IEEE Personal Communications Magazine (2001).
[8] M. Flament, F. Gessler, F. Lagergren, O. Queseth, R. Stridh, M. Unbehaun, J. Wu, J. Zander, Telecom scenarios for the 4th generation wireless infrastructures, in Proceedings of the PCC Workshop, Stockholm, 1998.
[9] NTT-DoCoMo, Outline of Fourth-Generation Mobile Communications, 2002. Available from: <http://www.nttdocomo.co.jp/corporate/rd/new_e/4gen01_e.html>.
[10] R. Berezdivin, R. Breinig, R. Topp, Next generation wireless communications concepts and technologies, IEEE Communications Magazine 40 (March) (2002) 108–116.
[11] J. Jimenez, Towards the 4G. Available from: <http://research.ac.upc.es/conferencies/ITCSS/jimenez.pdf>.
[12] WINEGLASS—Wireless IP NEtwork as a Generic platform for Location Aware Service Support, 2002. Available from: <http://domo-bili.cselt.it/WineGlass>.
[13] W.W. Lu, Compact multidimensional broadband wireless: the convergence of wireless mobile and access, IEEE Communications Magazine (November) (2000) 119–123.
[14] 3rd Generation Partnership Project, IP Based Multimedia Services Framework Report, 3GPP Technical Specification 23.228 V5.3.0.
[15] T. Robles, A. Kadelka, H. Velayos, A. Lappetelainen, A. Kassler, H. Li, D. Mandato, J. Ojala, B. Wegmann, QoS support for an All-IP system Beyond 3G, IEEE Communications Magazine 39 (August) (2001) 64–72.
[16] P. Conforto, C. Tocci, G. Losquadro, A. Spazio, R. Sheriff, M. Chan, Y. Hu, Ubiquitous Internet in an integrated satellite-terrestrial environment: The SUITED solution, IEEE Communications Magazine 40 (January) (2002) 98–107.

Tuna Tugcu received his BS and PhD degrees in Computer Engineering from Bogazici University in 1993 and 2001, respectively, MS degree in Computer and Information Science from New Jersey Institute of Technology in 1994. He pursued post-doctorate study in Broadband and Wireless Networking Lab at Georgia Institute of Technology until July 2002. He also worked as a visiting assistant professor at Georgia Institute of Technology (Savannah Campus) for two years. He is currently an assistant professor in Computer Engineering at Bogazici University. His research interests include real-time systems, communication networks, and wireless communications.

H. Birkan Yilmaz received his BS degree in Mathematics from Bogazici University in 2002. He is pursuing his MS study in Computer Engineering at Bogazici University since 2002. He also works as a teaching assistant in Mathematics Department. His research interests include next generation wireless systems, mathematical modeling, Markov models, routing in ad-hoc networks, cryptography, algebraic structures.

Feodor Vainstein received his MS degrees in Electronics and Electrical Engineering and in Applied Mathematics from Moscow Institute of Electrical and Computer Engineering in 1971 and 1974, respectively. In 1987, he immigrated to the USA, and received his Ph.D. in Electrical Engineering from Boston University in 1992. Following this, he became an associate professor at North Carolina A&T State University, and later served as the Director of Computer Engineering. Dr. Vainstein is currently a Professor of Computer Engineering at Georgia Institute of Technology, Savannah Campus.