Design and Analysis of a Network-Assisted Fast Handover Scheme for IEEE 802.16e Networks

Lung-Sheng Lee and Kuochen Wang, *Member, IEEE*

***Abstract*—The IEEE 802.16e standard provides quality of service (QoS) for real-time traffic; however, packet transmission disrupted by the handover (HO) process is still a big concern, particularly when frequent HO is performed by mobile stations (MSs) with high mobility. Therefore, an HO scheme that supports frequent HO and provides short service disruption time (SDT) is necessary for providing QoS to real-time traffic. In this paper, we present a novel network architecture, which complies with the IEEE 802.16e standard, to support seamless frequent HO, particularly for MSs with high mobility. Based on this architecture, a *network-assisted fast HO* (NFHO) scheme is proposed to shorten SDT during the HO process. By resolving connection identifier (CID) assignment and uplink (UL) timing adjustment issues, the proposed NFHO scheme can *restart both the UL* and *downlink (DL) packet transmissions before the MS proceeds to the HO ranging*, which is a unique feature of our scheme. In addition, based on the NFHO scheme, an analytic model has been developed to investigate the *expected number of buffered packets, packet loss probability,* and *SDT* during HO. Performance evaluation results show that the NFHO scheme reduces the DL SDT by 75%, compared with the IEEE 802.16e hard HO scheme, and it also reduces the UL SDT by 55.6% and 75% compared with the work of Jiao *et al.* and the IEEE 802.16e hard HO scheme (also the work of Choi *et al.*), respectively. In addition, the proposed NFHO scheme has the best performance in terms of expected number of buffered packets and packet loss probability among existing hard HO schemes for the IEEE 802.16e. Furthermore, our analytic model can be integrated into an admission-control policy to guarantee proper QoS for ongoing HO MSs.**

***Index Terms*—Fast handover (HO), IEEE 802.16e, packet loss probability, service disruption time (SDT).**

I. Introduction

THE IEEE 802.16-2004 standard was designed for rapid worldwide deployment, cost effectiveness, and interoperable multivendor fixed broadband wireless access [1]. To support mobility service, based on the IEEE 802.16-2004, the IEEE 802.16e standard was designed to support mobile stations (MSs) moving at vehicle speeds [2]. The medium access control (MAC) layer handover (HO) process to support mobility between base stations (BSs) is provided. The MAC layer comprises three sublayers. The service-specific convergence sublayer (CS) on top of the sublayers accepts higher layer protocol data units from asynchronous transfer mode (ATM) cell- or packet-based network layers. The common part sublayer (CPS) provides the core MAC functionality of system access, bandwidth allocation, connection establishment, and connection maintenance. The security sublayer (SS) provides authentication, secure key exchange, and encryption [1]. In addition, a service access point (SAP) was defined for interfacing between any two sublayers. The MAC SAP, which is the interface between the CPS and the CS, enables the CPS to support multiple CS specifications, such as ATM CS and packet CS [1]. There are three layer-2 HO modes provided in the IEEE 802.16e. One is a hard HO mode, and the other two are optional soft HO modes, i.e., macro diversity HO (MDHO) and fast BS switching (FBSS) [2]. An MS mandatorily supports the hard HO mode. There are several restrictions and extra hardware and software cost for the BSs to support MDHO or FBSS, such as synchronization on common time source, same frequency assignment, and synchronized frame structure [2]. Therefore, the hard HO mode was adopted in most existing IEEE 802.16e systems. In Section II, we will briefly describe the IEEE 802.16e HO.

The rest of this paper is organized as follows. Section II briefs the HO process in the IEEE 802.16e standard, and Section III reviews some existing IEEE 802.16e HO-enhanced schemes. Section IV describes the proposed network architecture and the proposed network-assisted fast HO (NFHO) scheme. In Section V, an analytic model is described. Performance evaluation results, including analytic and simulation results, are shown in Section VI. Section VII gives some concluding remarks.

II. IEEE 802.16e HO Overview

A. Overview of the IEEE 802.16e HO Stages

The hard HO process consists of the following six stages: 1) cell reselection; 2) HO decision and initiation; 3) synchronization to target BS downlink (DL); 4) ranging and network reentry; 5) termination of MS context; and 6) HO cancellation [2]. These stages can be functionally divided into the following two procedures: 1) HO preparation and 2) HO execution. The HO preparation procedure includes both cell reselection and HO decision and initiation stages. At the cell reselection stage, the MS requests the serving BS an allocation of scanning intervals. After the serving BS grants the scanning intervals, the MS maintains current connections with the serving BS and then scans and synchronized with neighboring (NBR) BSs to evaluate the quality of each channel. An initial ranging

Manuscript received February 24, 2009; revised June 23, 2009 and October 27, 2009. First published December 1, 2009; current version published February 19, 2010. This work was supported in part by the National Science Council under Grant NSC96-2628-E-009-MY3 and Grant NSC97-3114-E-009-001. The review of this paper was coordinated by Prof. J. Li.

The authors are with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: dragonlee@cs.nctu.edu.tw; kwang@cs.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2009.2037508

procedure, which is termed association here, is processed during a scanning interval with one of NBR BSs. The association procedure can obtain service-availability information and ranging parameters, both of which will be used for HO target selection and to expedite the HO-execution procedure [2]. At the end of the cell reselection stage, the MS can obtain an availability list of NBR BSs. At the HO decision and initiation stage, an HO decision is made for the MS to HO from the serving BS to the target BS.

The HO execution procedure includes the following four stages: 1) synchronization to target BS DL; 2) ranging and network reentry; 3) termination of MS context, and 4) HO cancellation. In this procedure, the MS starts the actual HO. The MS synchronizes to the DL of the target BS and obtains uplink (UL) parameters. After completing the ranging process, the MS obtains new basic and primary management connection identifiers (CIDs), which will be used to transport management messages, and adjusts UL parameters to synchronize to the UL. After synchronizing to the UL, the MS is capable of transmitting management messages. The MS then starts the network reentry process that consists of basic capabilities negotiation, authorization, and registration. Note that the ranging process in the ranging and network reentry stage is also termed HO ranging. Moreover, in the registration of the network reentry process, the target BS will reassign transport CIDs, which are used to transport data packets, for active connections. After acquiring the transport CIDs, the MS is capable of starting data packet transmission. Termination of MS context completes the HO process. In this stage, the serving BS releases the MS's context after the resource retain timer expires. In addition, the HO cancellation stage is used to handle the situation of the MS canceling the HO before the resource retain timer expires. In the HO execution procedure, the data-packet transmission is blocked until the MS acquires the transport CIDs in the registration of the network reentry process. To reduce the packet transmission delay in the HO execution procedure, the transport CID assignment and UL synchronization issues must be resolved in advance. Therefore, the motivation of this paper is to propose an HO-enhanced scheme that can resolve the transport CID assignment and UL synchronization issues ahead of the ranging and network reentry stage.

During the HO execution procedure, the serving BS buffers incoming DL packets. Upon completion of network reentry, the target BS, which now is the new serving BS, will forward the data packets received from the old serving BS to the MS. The HO process will prolong the packet transmission delay because data packets are held during the HO preparation and execution procedures. Shortening the service disruption time (SDT) caused by the HO process will minimize the impact on quality of service (QoS). In this paper, we focus on the IEEE 802.16e hard HO mode and propose a novel network architecture for IEEE 802.16e networks. Based on the network architecture, we design a *network-assisted fast HO* (NFHO) scheme to accelerate the HO execution procedure and to reduce the SDT resulted from the ranging and network reentry stage. As mentioned above, the DL packets forwarding from the serving BS to the target BS will prolong the packet transmission delay, particularly in frequent HO situations. The proposed NFHO scheme can reduce the DL packet forwarding delay by bicasting DL packets to both the serving BS and the target BS. It can also handle the problem of frequent HO of ping-pong mobility between BSs. An analytic model is developed to investigate the expected number of buffered packets, packet loss probability, and SDT during the HO execution procedure. The analytic model is also used to analyze the performance among existing IEEE 802.16e HO-enhanced schemes. We will show that the proposed NFHO scheme outperforms the existing IEEE 802.16e HO-enhanced schemes. In addition, we will show that the packet loss probability is affected by the following network parameters: the HO SDT, the packet arrival rate of concurrent HO MSs, and the size of HO packet buffer pool. By the analytic model, we can evaluate the HO packet buffer pool utilization in a BS under different network parameter combinations and obtain a proper network parameter setting that meets the QoS requirement. The analytic model can also be integrated to an admission-control policy to provide proper QoS for incoming HO MSs.

B. IEEE 802.16e HO Scheme

In the IEEE 802.16e network, the BS periodically broadcasts network topology information via the MOB_NBR-ADV message. This message contains channel information of NBR BSs. Note that those message terms with all capital letters were defined in the IEEE 802.16e [2]. When an MS would like to HO, it starts the HO preparation procedure and uses the MOB_SCN-REQ message to request a group of time intervals from the serving BS. Within the time intervals, the MS could seek and monitor a suitable BS from the list of candidate NBR BSs as the HO target. Following the MOB_SCN-RSP message, the MS starts scanning and attempts to synchronize with each NBR BS to evaluate the quality of the channel. During the scanning intervals, all incoming/outgoing packets to/from the MS shall be buffered until exiting the scanning mode and returning to the normal operation mode. The duration of a scanning interval depends on which level (2, 1, or 0) of the association procedure is chosen. With network-assisted association reporting (association level 2) [2], the MS will not wait for an RNG-RSP message from each NBR BS after sending an RNG-REQ message or a code division multiple access (CDMA) ranging code (for orthogonal frequency-division multiple access) to the NBR BS. Instead, each NBR BS will send the serving BS the RNG-RSP message over the backbone. All RNG-RSP messages from each NBR BS are finally collected in an MOB_ASC-REP message, which will be sent to the MS by the serving BS [2]. Therefore, association level 2 needs a shorter scanning interval than the other levels. When the MS decides to HO, an MOB_MSHO-REQ message will be sent to the serving BS. After negotiating with the selected target BS, the serving BS sends the MS an MOB_BSHO-RSP message that may include an action time parameter to specify when the target BS will allocate a fast ranging information element (IE) [2]. The MS could use the fast ranging IE to transmit the RNG-REQ message, which expedites HO ranging. The MOB_HO-IND sent by the MS is used to commit the HO. After committing the HO to the serving BS, the MS enters the HO execution procedure. The

MS proceeds to synchronize with the DL and then performs HO ranging, UL parameter adjustment, basic capability negotiation, authorization, and registration with the target BS. During the HO execution procedure, the serving BS holds data addressed to the MS, and the target BS may use RNG-RSP to notify the MS of pending DL data. Once the MS registers to the target BS successfully, the target BS starts transmitting the retained DL pending data, forwarded from the serving BS, to the MS. After the MS reestablishes Internet protocol (IP) connectivity and completes reception of DL pending data, the target BS then use a backbone message to request the old serving BS to stop forwarding DL data. Note that our proposed NFHO scheme can restart DL/UL data transmission before the HO ranging, which can greatly reduce the SDT. In Section IV, we will detail the proposed NFHO scheme.

III. Related Work

The HO of wireless networks can be classified into vertical HO and horizontal HO. An HO is defined as vertical if it occurs between heterogeneous wireless networks. The work in [3] and [4] proposes vertical HO decision algorithms to support HO between heterogeneous wireless networks. In contrast to vertical HO, an HO is defined as horizontal if it occurs between two adjacent cells of the same wireless network. In this paper, we only focus on horizontal HO. Existing IEEE 802.16e horizontal HO-enhanced schemes are reviewed in this section.

A. Existing IEEE 802.16e HO-Enhanced Schemes

During the HO process, real-time services may be disrupted. In the HO preparation procedure, the MS stops normal data transmission for the scanning of NBR BSs. Additionally, in the HO execution procedure, the normal data transmission is blocked until the MS completes the ranging and network reentry stage. The blocking of data transmission disrupts the service of real-time traffic and increases packet transmission delay that impacts the QoS provision.

Existing IEEE 802.16e HO-enhanced schemes focus on either layer 3 or 2. Layer 3 HO schemes, such as [5]–[8], are basically based on the IEEE 802.16e layer 2 hard HO scheme to accelerate layer 3 HO for mobile IPv6, and they did not reduce the packet transmission delay on layer 2. Therefore, these layer 3 HO schemes also have at least the same SDT as the IEEE 802.16e hard HO scheme. Existing layer 2 HO-enhanced schemes can be functionally classified into improvements on either the HO preparation procedure or the HO execution procedure. The work in [9]–[13] focused on the HO preparation procedure, and their algorithms are to predict an HO target BS to reduce ping-pong effects and the number of NBR BS scanning. In addition to HO target BS prediction, Lee *et al.* [9] also proposed a fast synchronization and association scheme that makes the MS be able to do data transmission with the serving BS and to do association with the target BS simultaneously. To realize this scheme, it would costly base on either the MDHO or the FBSS mode. In addition, wrong target BS estimation, due to the exhaustion of target BS resources and so on, may cause huge delays when the MS repeats to evaluate the next target BS candidate.

The studies in [14] and [15] focused on reducing data latency in the HO execution procedure. Choi *et al.* [14] proposed a Fast_DL_MAP_IE message to restart DL data transmission before the MS proceeds to the ranging and network reentry stage. The target BS will use Fast_DL_MAP_IEs and old CIDs for transmitting DL data packets immediately after the MS completes the synchronization to target BS DL stage. After HO ranging is competed and the REG-RSP message, including new transport CID assignment for active connections, is received, the MS will return to use normal DL_MAP_IE with new assigned transport CIDs. Similar to [14], Jiao *et al.* [15] proposed another transport CID-assignment scheme to restart DL data transmission before the MS proceeds to the ranging and network-reentry stage. Through its transport CID assignment scheme, the target BS can use the old transport CIDs, which were used in the serving BS, to transmit DL packets with no CID conflicts between the serving BS and the target BS. The MS uses old transport CIDs until it receives the REG-RSP message, which assigns new transport CIDs to active connections. Both schemes in [14] and [15] restart DL data transmission before the MS proceeds to the HO ranging; however, this feature is not applicable to the UL real-time traffic of these two schemes. This is because the scheme in [14] did not provide mechanisms to preacquire UL synchronization parameters, which are acquired during HO ranging, and new transport CIDs, which are acquired during registration. Concerning the scheme in [15], it can use old transport CIDs for UL data transmission until it receives the REG-RSP message; however, it did not provide mechanisms to preacquire UL synchronization parameters. Thus, the scheme in [15] can advance UL data transmission only after HO ranging.

To restart UL data transmission, the MS should synchronize to the UL first. Since the RNG-RSP message provides UL synchronization parameters, such as frequency corrections, transmission power level corrections, UL timing offset corrections, and basic and primary management CID assignment, to advance UL data transmission, the UL parameters should be obtained ahead of schedule. In this paper, we focus on reducing the SDT during the HO execution procedure and propose an NFHO scheme that can restart both DL and UL data transmissions before the MS proceeds to the HO ranging. We will detail our scheme in Section IV.

B. Qualitative Comparison of Existing IEEE 802.16e HO-Enhanced Schemes

Focusing on the layer 2 IEEE 802.16e HO process, Table I shows a qualitative comparison of the existing IEEE 802.16e HO-enhanced schemes. As mentioned above, existing layer 3 schemes were all based on the IEEE 802.16e layer 2 hard HO scheme. Therefore, from the layer 2 view, we regard the layer 3 schemes as the same as the IEEE 802.16e hard HO scheme. Lee. *et al.* [9] is a representative of those schemes that focus on improving the HO preparation procedure. Choi *et al.* [14] and Jiao *et al.* [15] focus on the HO execution procedure, which is also the focus of the proposed NFHO scheme. Among existing IEEE 802.16e HO-enhanced schemes, [15] is the only paper that dealt with the reduction of both DL and UL data latencies. Our proposed NFHO scheme is also included

TABLE I
COMPARISON AMONG EXISTING IEEE 802.16e HO SCHEMES

Scheme	Improvement on / design focus	Data transmission restarts at	CID_Update required
Association level 2 of 802.16e [2]	HO Preparation procedure/ Reduce NBR BS association time	DL/UL after REG-RSP	Yes
Lee et al. [9]	HO Preparation procedure / 1. Reduce the number of NBR BSs scanning 2. Reduce association time	DL/UL after REG-RSP	Yes
Choi et al. [14]	HO Execution procedure / Reduce DL data latency	1. DL before HO ranging 2. UL after REG-RSP	Yes
Jiao et al. [15]	HO Execution procedure / Momentarily reuse old CIDs to reduce DL and UL data latency	1. DL before HO ranging 2. UL after HO ranging	Yes
NFHO (Proposed)	HO Execution procedure / 1. Fast UL synchronization 2. Reduce DL and UL data latency	DL/UL before HO ranging	Intra-CSSC HO: No Inter-CSSC HO: Yes

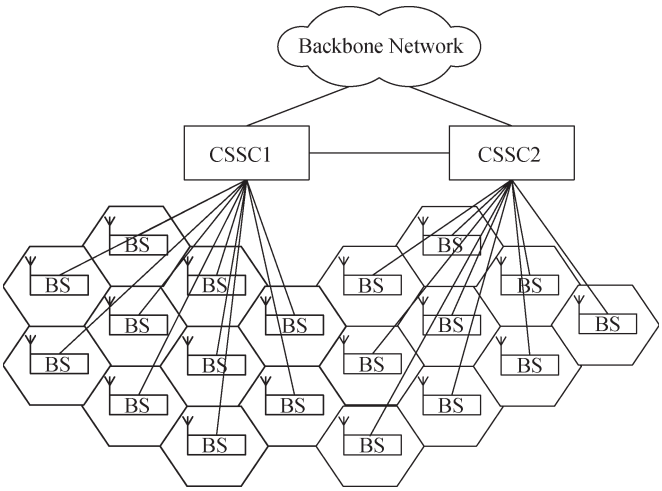

Fig. 1. Proposed network architecture for IEEE 802.16e networks.

in this qualitative comparison. Note that, to the best of our knowledge, our scheme is the only scheme that can restart data transmission at both DL and UL before the MS proceeds to HO ranging.

IV. DESIGN APPROACH

To restart data transmission ahead of the ranging and network reentry stage, two issues need to be resolved. One is UL synchronization, and the other is updating transport CIDs, which are used by active connections for data transmission in the target BS. We first propose a novel network architecture for IEEE 802.16e networks. Based on this architecture, an NFHO scheme is proposed to reduce the execution time of the HO execution procedure. Fig. 1 shows the proposed network architecture. A set of BSs is linked to a CS switch center (CSSC) through the wired network. Here, we design a CSSC that handles two original CS functions: the MAC function of the service-specific CS (either ATM CS or packet CS) and the backbone management message exchange. In addition, the CSSC also handles control messages to bicast data packets to both the serving BS and the selected target BS. A BS handles the functions of the MAC CPS, SS, and PHY layers.

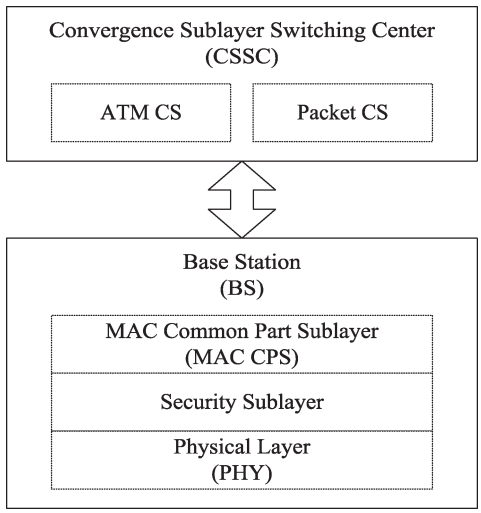

Fig. 2. Protocol layering of the proposed network architecture.

Fig. 2 illustrates the protocol layering of the proposed network architecture. We place ATM CS and packet CS in the CSSC. The HO within a CSSC is termed as *intra-CSSC HO*, and the HO between CSSCs is termed as *inter-CSSC HO*. The CIDs need to be changed only when the MS HOs to a BS that belongs to another CSSC. That is, based on the proposed network architecture, the target BS does not need to reassign CIDs for intra-CSSC HO. To reassign CIDs for inter-CSSC HO, we propose control messages, that are exchanged between the serving BS and the target BS in the HO decision and initialization stage, to preassign CIDs for the MS. Therefore, the MS can obtain new CIDs before entering the HO execution procedure. In this way, the CID reassignment issue can be resolved. The control message exchange for CID preassignment during inter-CSSC HO will be detailed in Section IV-B.

To resolve UL synchronization, we propose to redo association level 2 in the HO decision and initialization stage. After a target BS is selected, the MS may redo scanning and association level 2 for the selected target BS if the current UL synchronization parameters obtained from the cell reselection stage are considered not up-to-date. Therefore, the MS can obtain correct UL synchronization parameters before entering the HO execution procedure. After synchronizing to the DL in the synchronization to target BS DL stage, the MS can adjust the UL synchronization parameters to synchronize to the UL immediately. Furthermore, we also propose an open-loop fine-tuning method that can improve the accuracy of the UL timing-adjustment offset. The open-loop fine-tuning method is detailed in Section IV-A.

We use the same CIDs and the preassigned CIDs for intra-CSSC HO and inter-CSSC HO, respectively. Furthermore, the

proposed scheme enables the MS to synchronize to the UL immediately after the DL is synchronized. Since DL packets are bicast to the selected target BS as well as the serving BS at the HO decision and initialization stage, the transmission of both DL and UL packets can be immediately restarted after the MS completes the synchronization to target BS DL stage. Therefore, we can shorten the packet transmission delay resulted from waiting for the completion of the ranging and network reentry stage during the HO execution procedure. The detailed intra- and inter-CSSC HO message sequence charts (MSCs) are described in Section IV-B.

Note that the proposed network architecture complies with the IEEE 802.16e standard. Since the CSSC still handles its original CS functions, it can follow MAC SAP specifications to access the MAC CPS. Concerning the backbone management messages, exchanging between CSSC and MAC CPS, they can be handled by an add-on software component for supporting the proposed NFHO scheme. That is, only an add-on software component with a small cost is required to support the proposed NFHO scheme. Note that even without the add-on software component, the CSSC and BS can still function through the MAC SAP to support the original IEEE 802.16e hard HO scheme. Therefore, the proposed network architecture complies with the IEEE 802.16e standard. Note that in the proposed network architecture, the CSSC is logically separated as an individual component; however, from an implementation perspective, the CSSC can be still located in one of the BSs, and the other BSs have wired links to the CSSC.

A. Acquiring UL Synchronization Parameters

Association in the cell reselection stage is used to acquire ranging parameters and service-availability information that could expedite the HO execution procedure. In the IEEE 802.16e, association level 2 provides better efficiency than the other two association levels (0 and 1). Here, we assume association level 2 is adopted. As mentioned above, after an HO target BS is selected, we redo association to acquire fresh UL synchronization parameters in the HO decision and initiation stage. Fig. 3 shows the flowchart of reassociation to the selected target BS for acquiring UL synchronization parameters. If the UL parameters of the selected target BS are considered to be out-of-date, the MS should renew the UL parameters by doing association with the selected target BS. The decision to renew UL parameters depends on some factors, such as the difference of DL arrival times since the last association, the decay of mean carrier-to-interference-and-noise ratio since the last association, and a refresh timer. Here, we simply assume that the MS decides to renew the UL parameters when a refresh timer has expired.

Based on the difference of DL signal delays between the synchronization at the association and the synchronization at the synchronization to target BS DL stage, we propose an open-loop fine-tuning method to acquire a correct UL timing adjustment offset. Fig. 4 shows the difference of DL signal delays between locations "a" and "b." The MS did association with the target BS at location "a," and a frame was synchronized at γ_a, which was the clock time measured by the MS. During the synchronization to target BS DL stage, the MS moves to

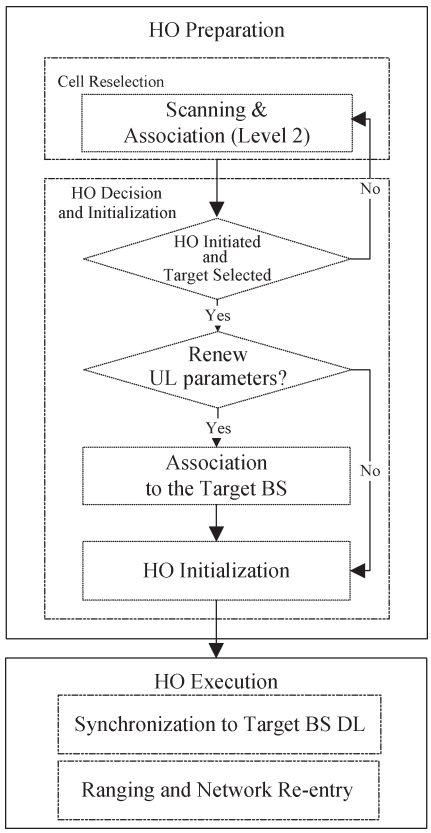

Fig. 3. Reassociation to renew UL parameters.

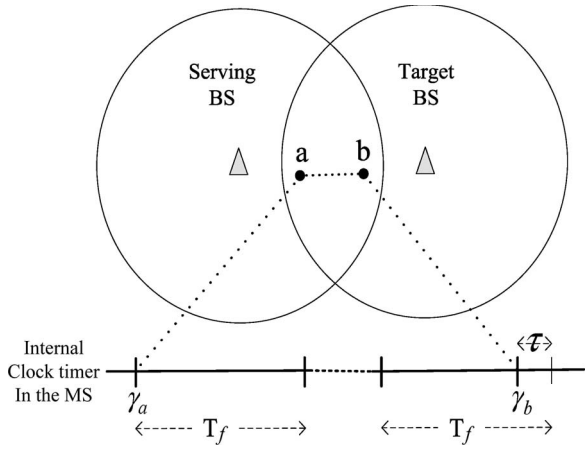

Fig. 4. Difference of DL signal delays between synchronization at the association and synchronization at the synchronization to target BS DL stage.

location "b," and a frame is synchronized at γ_b, which is the clock time measured by the MS. Therefore, we have

$$\gamma_b - \gamma_a = n \cdot T_f + \tau, \qquad |\tau| \ll T_f$$

where T_f is the frame duration, n is the number of frames in the time period between "a" and "b," and τ is the difference of DL signal delays between locations "a" and "b." Since the UL has also the same difference of the UL signal delays between "a" and "b" as that of the DL, we add the difference τ to fine-tune the UL timing adjustment offset. Thus, we have

$$\eta_b = \eta_a + \tau$$

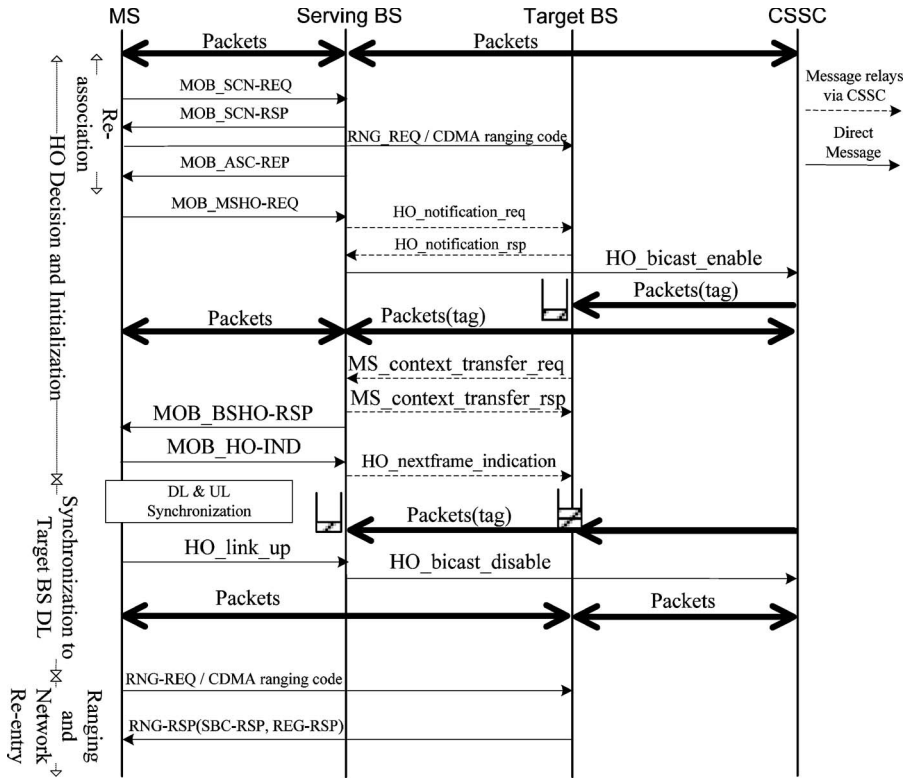

Fig. 5. Intra-CSSC HO MSC.

where η_a is the UL timing-adjustment offset obtained from the last association, and η_b is the fine-tuned UL timing-adjustment offset that will be used at the synchronization to target BS DL stage. By acquiring a correct UL timing adjustment offset, the MS can also synchronize to the UL immediately after synchronizing to the DL of the target BS. As a result, the MS can restart the DL and UL data transmissions before the ranging and network reentry stage.

B. Fast HO Execution Procedure With QoS Support

To reduce the packet transmission delay due to the HO execution procedure, based on our proposed network architecture, we bicast DL traffic to both the serving BS and the selected target BS to avoid the latency caused by data forwarding. Note that bicasting DL traffic will incur extra bandwidth overhead between CSSC and BS. However, since the link between CSSC and BS is a wired link, the extra bandwidth overhead is not a concern. In the HO execution procedure of the IEEE 802.16e, the UL/DL data packets can only be transmitted after the ranging and network reentry stage. In the ranging and network reentry stage, the IEEE 802.16e provides HO optimization options that could omit authorization by sharing the MS's context between the serving BS and target BS and grouping the basic capabilities negotiation and registration into the HO ranging. Note that the configuration of the HO optimization options in each BS shall be announced in the MOB_NBR-ADV message. Therefore, by configuring the HO optimization options, data transmission could be restarted immediately after the HO ranging. In the HO ranging, the target BS provides the MS necessary parameters, including UL timing-adjustment offset, frequency corrections, transmission power-level corrections, and basic and primary management CIDs. The proposed NFHO scheme shortens the data-transmission delay by restarting the UL and DL data transmissions ahead of the HO ranging. Be reminded that these parameters, except for the basic and primary management CIDs, can be obtained in the last association, and the UL timing-adjustment offset can be further fine-tuned for UL synchronization in the synchronization to target BS DL stage. In addition, based on the proposed network architecture, the intra-CSSC HO does not need to update transport CIDs. Therefore, normal data-packet transmission can be restarted before the HO ranging, and the SDT resulted from the HO ranging could be eliminated.

Fig. 5 shows the MSC of the proposed intra-CSSC NFHO scheme. The bold lines show packet flows, and the others are management messages. The management messages with dashed lines exchanged between BSs are relayed via a CSSC. The management messages with solid lines are direct messages from source to destination. The management message terms with all capital letters are the MAC management messages defined in the IEEE 802.16e, and the other management messages terms with leading capital letters and followed by lower case letters are added messages for the NFHO scheme. The proposed intra-CSSC NFHO scheme that shortens SDT is detailed as follows.

1) After the HO target is selected in the HO decision and initialization stage, the MS starts a reassociation to renew UL parameters for the target BS if the MS considers the current UL parameters of the target BS to be out-of-date. The renewed UL parameters will be used for UL synchronization in the synchronization to target BS DL stage.

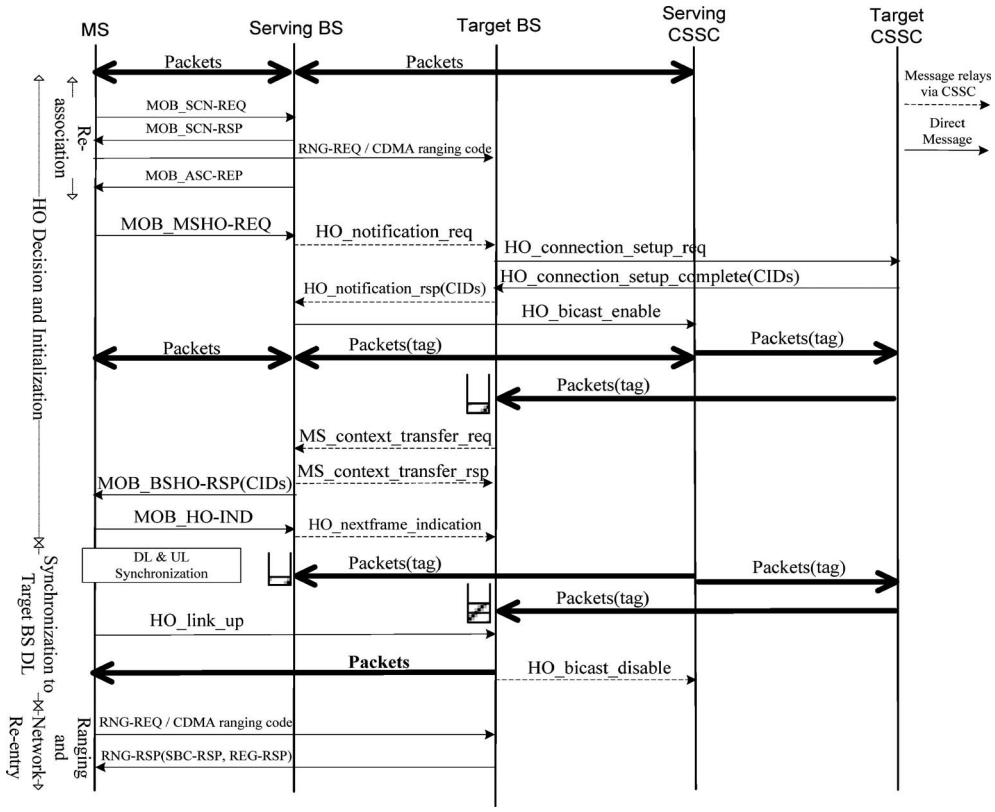

Fig. 6. Inter-CSSC HO MSC.

2) When an MS issues MOB_MSHO-REQ to the serving BS, the serving BS should negotiate the QoS requirement of the MS with the target BS via the added backbone messages: HO_notification_req and HO_notification_rsp.
3) The serving BS issues an HO_bicast_enable message to enable bicasting DL packets to the target BS. After that, each DL packet bicast to both the serving and target BSs is tagged with a sequence number. In Fig. 5, we denote a tagged packet flow as Packets(tag).
4) Because the HO process optimization options are enabled to omit authorization, the serving and target BSs must use MS_context_transfer_req and MS_context_transfer_rsp to transfer the context of the MS. After that, the serving BS sends MOB_BSHO-RSP to the MS.
5) The MS sends MOB_HO-IND to inform the serving BS that the HO is started. Then, the MS enters the synchronization to target BS DL stage and starts synchronizing to the target BS. Meanwhile, the serving BS sends HO_nextframe_indication to notify the target BS of the tagged sequence number of the next expected DL packet. After synchronizing to the DL of the target BS, the MS also synchronizes to the UL by the acquired UL parameters described in Section IV-A. As mentioned above, there is no need to update transport CIDs for the active connections of the MS during intra-CSSC HO. Therefore, the target BS can start to allocate DL_MAP_IE and UL_MAP_IE for DL and UL data packets after receiving HO_link_up. The MS sends HO_link_up to inform the target BS of UL synchronization. Note that the HO_link_up could be a bandwidth-request packet instead. After that, the target BS issues HO_bicast_disable to disable bicasting and let the DL data path switch to the target BS only.
6) The HO ranging, which exchanges RNG-REQ and RNG-RSP, is started after normal data packet transmission. Since the HO optimization options are enabled, the unsolicited SBC-RSP (basic capabilities negotiation) and REG-RSP (registration) are appended to the RNG-RSP.

Fig. 6 shows the MSC for the proposed inter-CSSC NFHO scheme. The MSC is similar to that of the intra-CSSC HO, except that the target CSSC updates transport CIDs for the active connections of the MS, and these updated transport CIDs are sent back to the MS through MOB_BSHO-RSP. The message followed by "(CIDs)" denotes that these updated transport CIDs are attached. The following details the inter-CSSC HO MSC.

1) The MS performs association level 2 to the selected target BS and then obtains UL parameters from MOB_ASC-REP.
2) The MS starts an HO by issuing MOB_MSHO-REQ to the serving BS. The serving BS uses HO_notification_req to negotiate MS's QoS requirements with the target BS. Since this HO is an inter-CSSC HO, the target CSSC needs to update transport CIDs for the active connections of the MS. The target BS requests a connection setup to the target CSSC by sending HO_connection_setup_req. Then, the target CSSC assigns transport CIDs and initializes a classifier which will classify upcoming bicasting packets by the assigned transport CIDs. The target CSSC sends HO_connection_setup_complete with assigned

transport CIDs to the target BS. Finally, the assigned transport CIDs are delivered to the serving BS through HO_notification_rsp.

3) The HO_bicast_enable is issued by the serving BS to enable bicasting DL data packets to the target CSSC. After that, each DL packet tagged with a sequence number is bicast to both the serving and target BSs. The target CSSC and target BS will use the assigned transport CIDs to transmit packets.
4) MS_context_transfer_req and MS_context_transfer_rsp are used to transfer the context of the MS from the serving BS to the target BS. The serving BS sends an MOB_BSHO-RSP message with the assigned transport CIDs appended to respond to the received MOB_MSHO-REQ.
5) After sending MOB_HO-IND, the MS enters the Synchronization to the target BS DL stage. After synchronizing to the DL, the MS also immediately synchronizes to the UL by the acquired UL parameters described in Section IV-A. The serving BS sends the target BS HO_nextframe_indication to notify the next expected sequence number of the tagged DL packet. Since the connection setup between the target BS and the target CSSC have been done and the bicast DL packets in the target BS have been constructed with assigned transport CIDs, the target BS can go ahead to schedule for UL/DL data transmission without waiting for the completion of the HO ranging. The HO_link_up issued by the MS is used to indicate that the MS has synchronized to the UL of the target BS; therefore, the HO_link_up could be a bandwidth request packet instead of a specific management message. After that, the target BS sends HO_bicast_disable to switch the DL data path to the target BS only.
6) The HO ranging is started after normal data packet transmission. Since the target BS enables the HO optimization options, the unsolicited SBC-RSP and REG-RSP are appended to the RNG-RSP.

V. Analytic Model

Since the proposed NFHO scheme bicasts DL packets to the target BS in advance in addition to the serving BS, the expected number of buffered packets in the target BS is a performance metric. The exhaustion of the packet buffer pool may result in the packet loss of subsequent incoming DL packets, and the packet loss probability is another performance metric to evaluate the QoS for the MS. Therefore, including the packet loss probability to the policy of admission control is necessary to guarantee QoS to the ongoing HO MSs.

In the following, we focus on the inter-CSSC HO and analyze the expected number of buffered packets in Section V-A. In Section V-B, with a limited packet buffers constraint, we analyze the number of concurrent HO MSs and the packet-loss probability from a system viewpoint. In Section V-C, we evaluate existing HO schemes and analyze the SDT and the expected number of buffered packets during an HO. By applying the expected number of buffered packets obtained in Section V-C to the derivations in Section V-B, we can obtain the packet-loss probability for each existing HO scheme.

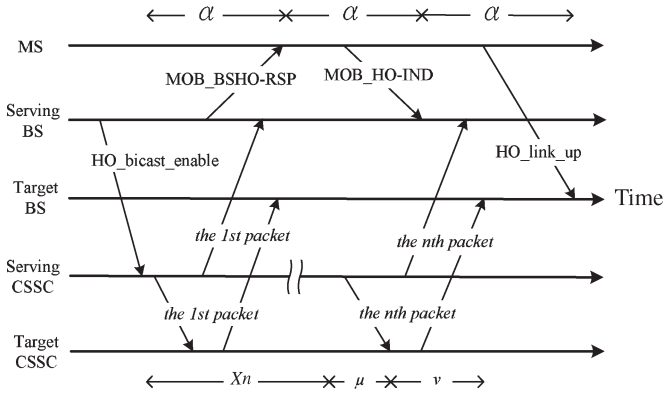

Fig. 7. Timing diagram extracted from Fig. 6.

A. Expected Number of Buffered Packets

Fig. 7 shows the timing diagram extracted from Fig. 6. After HO_bicast_enable arrives at the serving CSSC, the serving CSSC starts to bicast subsequent DL packets to the serving BS and the target BS. The packets bicast to the target BS are buffered in the buffer pool until the MS synchronizes to the DL of the target BS and informs the target BS of UL synchronization by issuing an HO_link_up. After the target BS receives HO_link_up, the buffered packets start to be consumed by sending them to the MS. Therefore, in the target BS, the number of packets buffered for the MS will reach a maximum at the moment that HO_link_up arrives at the target BS. Assume that the arrival of packets to the serving CSSC is a Poisson arrival process with arrival rate λ [16], [18]. Thus, the interarrival times are exponentially distributed. Assume that an HO_bicast_enable message arrives at the serving CSSC at time t_0 and that t_n is the time that the nth packet arrives at the serving CSSC. Let $X_n = t_n - t_0$ denote the nth bicast packet arriving at the serving CSSC in the period [0, t]. Thus, X_n has an Erlang distribution with density function [18], [20] given by

$$f_{X_n}(t) = \frac{(\lambda t)^{n-1}}{(n-1)!}\lambda e^{-\lambda t}. \tag{1}$$

The Laplace transform for X_n is

$$f^*_{X_n}(s) = \left(\frac{\lambda}{s+\lambda}\right)^n. \tag{2}$$

Without loss of generality, we assume that the processing delays of HO_BSHO-RSP, MOB_HO-IND, and HO_link_up have the same distribution with the following mixed-Erlang density function [18], [19], [21]:

$$f_\alpha(t) = \sum_{i=1}^{I} \alpha_i \frac{(\sigma_i t)^{p_i-1}}{(p_i-1)!}\sigma_i e^{-\sigma_i t} \tag{3}$$

where

$$\sum_{i=1}^{I} \alpha_i = 1$$

and α is a random variable that represents the processing delay. The Laplace transform for α is expressed as

$$f_{\alpha}^{*}(s) = \sum_{i=1}^{I} \alpha_i \left(\frac{\sigma_i}{s+\sigma_i} \right)^{p_i}. \tag{4}$$

We select the mixed-Erlang distribution because it has been proven that it can well approximate many other distributions [17], [18], [21].

Let μ be a random variable to denote the transmission delay from the serving CSSC to the target CSSC and ν be a random variable to denote the transmission delay from the target CSSC to the target BS. Thus, the transmission delay from the serving CSSC to the target BS is $\mu + \nu$. Assume that both μ and ν also have the mixed-Erlang distributions with the following density functions [18], [19], [21]:

$$f_{\mu}(t) = \sum_{j=1}^{J} \mu_j \frac{(\rho_j t)^{q_j - 1}}{(q_j - 1)!} \rho_j e^{-\rho_j t} \tag{5}$$

$$f_{\nu}(t) = \sum_{k=1}^{K} \nu_k \frac{(\delta_k t)^{r_k - 1}}{(r_k - 1)!} \delta_k e^{-\delta_k t} \tag{6}$$

where

$$\sum_{j=1}^{J} \mu_j = 1, \qquad \sum_{k=1}^{K} \nu_k = 1.$$

The corresponding Laplace transforms are expressed as follows:

$$f_{\mu}^{*}(s) = \sum_{j=1}^{J} \mu_j \left(\frac{\rho_j}{s+\rho_j} \right)^{q_j} \tag{7}$$

$$f_{\nu}^{*}(s) = \sum_{k=1}^{K} \nu_k \left(\frac{\delta_k}{s+\delta_k} \right)^{r_k}. \tag{8}$$

Let Φ_n represent the nth packet that arrives at the target BS. From Fig. 7, we have

$$\Phi_n = \begin{cases} 1, & X_n + \mu + \nu < 3\alpha \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Therefore, the expected number of buffered packets, which is denoted as $Q_{\max}$, is expressed as [18]

$$Q_{\max} = \sum_{n=1}^{\infty} E(\Phi_n) \tag{10}$$

$$= \sum_{n=1}^{\infty} \Pr(X_n + \mu + \nu < 3\alpha). \tag{11}$$

Let $T_n = X_n + \mu + \nu$. Then, (10) can be rewritten as

$$Q_{\max} = \sum_{n=1}^{\infty} \Pr(T_n < 3\alpha) \tag{12}$$

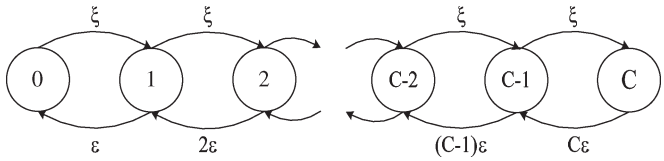

Fig. 8. State transition diagram for HO processing.

where $\Pr(T_n < 3\alpha)$ is calculated as

$$= \sum_{n=1}^{\infty} \left\{ \int_{T=0}^{\infty} f_{T_n}(T) \left[\int_{t=T}^{\infty} f_{3\alpha}(t) dt \right] dT \right\}$$

$$= \sum_{n=1}^{\infty} \left\{ \int_{T=0}^{\infty} f_{T_n}(T) \left[\sum_{i=1}^{I} \frac{\alpha_i \sigma_i^{P_i}}{3^{P_i}(P_i - 1)!} \left(\int_{t=T}^{\infty} t^{P_i - 1} e^{-\frac{\sigma_i}{3} t} dt \right) \right] dT \right\}$$

$$= \sum_{n=1}^{\infty} \left\{ \int_{T=0}^{\infty} f_{T_n}(T) \left[\sum_{i=1}^{I} \alpha_i \left(\sum_{h=0}^{P_i - 1} \frac{(\sigma_i T)^h}{3^h h!} e^{-\frac{\sigma_i}{3} T} \right) \right] dT \right\}$$

$$= \sum_{n=1}^{\infty} \sum_{i=1}^{I} \alpha_i \sum_{h=0}^{P_i - 1} \frac{\sigma_i^h}{3^h h!} \left(\int_{T=0}^{\infty} f_{T_n}(T) T^h e^{-\frac{\sigma_i}{3} T} dT \right)$$

$$= \sum_{n=1}^{\infty} \sum_{i=1}^{I} \alpha_i \sum_{h=0}^{P_i - 1} \frac{\sigma_i^h}{3^h h!} \left((-1)^h \frac{d^h}{ds^h} f_{T_n}^{*}(s) \Big|_{s=\frac{\sigma_i}{3}} \right). \tag{13}$$

Since $T_n = X_n + \mu + \nu$ and $f_{T_n}(t)$ denotes the density function of T_n, by applying the convolution property, its Laplace transform $f_{T_n}^{*}(s)$ is given as follows:

$$f_{T_n}^{*} = f_{X_n}^{*} f_{\mu}^{*} f_{\nu}^{*}$$

$$= \left(\frac{\lambda}{s+\lambda} \right)^n \left[\sum_{j=1}^{J} \mu_j \left(\frac{\rho_j}{s+\rho_j} \right)^{q_j} \right] \left[\sum_{k=1}^{k} \nu_k \left(\frac{\delta_k}{s+\delta_k} \right)^{r_k} \right]. \tag{14}$$

B. Packet-Loss Probability During HO

Since the target BS only has limited packet buffers to queue bicast packets during HO, bicast packets arrived at the target BS are dropped if there is no available packet buffer. Suppose the BS could provide a common HO packet buffer pool to hold bicast packets for all incoming HO MSs. After HO_link_up is received, the BS allocates the MS a dedicated connection buffer pool. Then, the buffered packets for the MS are moved from the HO packet buffer pool to the MS's dedicated connection buffer pool, and the packet buffers are released to the HO packet buffer pool immediately.

Consider a BS that can handle at most C MSs that HO into the BS concurrently. Assume that the arrival of an HO request is a Poisson process with arrival rate ξ. The HO processing time is exponentially distributed with mean $1/\varepsilon$. We model the HO processing as M/M/m/m queuing model. Fig. 8 shows the state

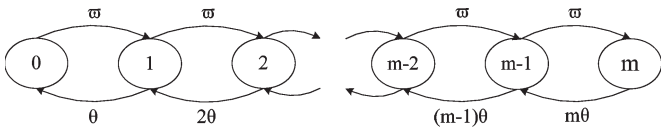

Fig. 9. State transition diagram for packet-buffer utilization.

transition of the Markov process for HO processing. The arrival rate of state k in this model is

$$\xi_k = \begin{cases} \xi, & k < C \\ 0, & k \geq C. \end{cases} \tag{15}$$

The departure rate in state k is given by

$$\varepsilon_k = k\varepsilon. \tag{16}$$

Let the probability in state k be π_k. Therefore, we have

$$\sum_{k=0}^{C} \pi_k = 1. \tag{17}$$

From [20], we have

$$\pi_k = \begin{cases} \pi_0 \prod_{i=0}^{k-1} \frac{\xi}{(i+1)\varepsilon}, & k \leq C \\ 0, & k > C. \end{cases} \tag{18}$$

By substituting π_k into (17), we get

$$\pi_0 = \left[\sum_{k=0}^{C} \left(\frac{\xi}{\varepsilon} \right)^k \frac{1}{k!} \right]^{-1}. \tag{19}$$

To analyze the usage of the HO packet buffer pool, we model it as an M/M/m/m queuing model. Assume that the HO packet buffer pool can provide at most m packet buffers. Fig. 9 shows the state transition of the Markov process for the utilization of HO packet buffers. The buffer requests to the HO packet buffer pool are a Poisson arrival process with rate ϖ. The buffer request rate can be obtained from the expected number of ongoing HO MSs and the packet-arrival rate of each ongoing HO MS. Therefore, ϖ can be expressed as

$$\varpi = \lambda \sum_{k=1}^{C} k\pi_k. \tag{20}$$

Let θ denote the rate of the HO packet buffers that are released to the HO packet buffer pool. In addition, an HO packet buffer holding time Θ can be expressed as $\Theta = 1/\theta$. Let Th_n denote the duration of the nth bicast packet that is buffered at the target BS. Thus, the expected time of a bicast packet held at the target BS can be obtained as

$$\Theta = Q_{\max}^{-1} \sum_{n=1}^{\infty} E(Th_n) \tag{21}$$

where $E(Th_n)$ is calculated as follows:

$$\begin{aligned} &E(Th_n) \\ &= \int_{T=0}^{\infty} \int_{t=T}^{\infty} f_{T_n}(T) f_{3\alpha}(t)(t-T)dtdT \\ &= \int_{T=0}^{\infty} \int_{t=T}^{\infty} f_{T_n}(T) f_{3\alpha}(t) t dt dT \\ &\quad - \int_{T=0}^{\infty} \int_{t=T}^{\infty} f_{T_n}(T) f_{3\alpha}(t) T dt dT \\ &= \sum_{i=1}^{I} \alpha_i \left[\sum_{h=0}^{P_i} \frac{p_i \sigma_i^{h-1}}{3^{h-1} h!} \left(\int_{T=0}^{\infty} f_{T_n}(T) T^h e^{-\frac{\sigma_i}{3}T} dT \right) \right] \\ &\quad - \sum_{i=1}^{I} \alpha_i \left[\sum_{h=0}^{P_i-1} \frac{\sigma_i^h}{3^h h!} \left(\int_{T=0}^{\infty} f_{T_n}(T) T^{h+1} e^{-\frac{\sigma_i}{3}T} dT \right) \right] \\ &= \sum_{i=1}^{I} \alpha_i \left\{ \left[\sum_{h=0}^{P_i} \frac{p_i \sigma_i^{h-1}}{3^{h-1} h!} \left((-1)^h \frac{d^h}{ds^h} f_{T_n}^*(s) \Big|_{s=\frac{\sigma_i}{3}} \right) \right] \right. \\ &\quad \left. - \left[\sum_{h=0}^{P_i-1} \frac{\sigma_i^h}{3^h h!} \left((-1)^{h+1} \frac{d^{h+1}}{ds^{h+1}} f_{T_n}^*(s) \Big|_{s=\frac{\sigma_i}{3}} \right) \right] \right\}. \end{aligned} \tag{22}$$

In this model, the packet buffer request rate in state k is

$$\varpi_k = \begin{cases} \varpi, & k < m \\ 0, & k \geq m. \end{cases} \tag{23}$$

The packet buffer release rate in state k is given by

$$\theta_k = k\theta. \tag{24}$$

Assuming the probability in state k denoted as ψ_k, we have

$$\sum_{k=0}^{m} \psi_k = 1. \tag{25}$$

Thus, from [20], the probability in state k is

$$\psi_k = \psi_0 \prod_{i=0}^{k-1} \frac{\varpi}{(i+1)\theta}, \qquad k \leq m. \tag{26}$$

By substituting ψ_k into (25), we get

$$\psi_0 = \left[\sum_{k=0}^{m} \left(\frac{\varpi}{\theta} \right)^k \frac{1}{k!} \right]^{-1}. \tag{27}$$

C. SDT During the HO Execution Procedure

The SDT is defined as the duration that the MS and BS stop data transmission during the HO execution procedure. In this section, we evaluate existing HO schemes, including the proposed NFHO, and analyze the SDT for the HO execution

TABLE II
SDT OF THE EXISTING HO SCHEMES

SCHEME	DL	UL
802.16e [2]	$T_{Sync} + T_{Rng} + T_{Reg}$	$T_{Sync} + T_{Rng} + T_{Reg}$
Choi et al. [14]	T_{Sync}	$T_{Sync} + T_{Rng} + T_{Reg}$
Jiao et al. [15]	T_{Sync}	$T_{Sync} + T_{Rng}$
Proposed NFHO	T_{Sync}	T_{Sync}

procedure of each scheme. We assume that the HO process-optimization options are enabled to omit the basic capabilities negotiation and authorization in the HO execution procedure. First, the following parameters are defined:

T_{Sync} mean time of the DL synchronization;
T_{Rng} mean time of the HO ranging;
T_{Reg} mean time of the registration.

In Table II, the SDT of some classical HO schemes and the proposed NFHO scheme are compared. Since the schemes in [14] and [15] and the proposed NFHO can immediately restart the DL transmission after the synchronization to target BS DL stage, these schemes have the same DL SDT and shorter DL disruption time than the IEEE 802.16e hard HO scheme. The scheme in [15] can temporarily use transport CIDs until these CIDs are updated in the registration. Thus, it can restart the UL transmission immediately after the HO ranging. In [14], it only improves on DL transmission, and therefore, it has the same UL SDT with the IEEE 802.16e hard HO scheme. By resolving transport CID assignment and UL synchronization issues, the proposed NFHO can also restart the UL transmission immediately after the synchronization to target BS DL stage.

We assume that both the DL and UL have the same traffic model and define Y_n as the nth packet arrived at the target BS in the DL part or generating from the MS in the UL part during the SDT. Assume that Y_n has an Erlang distribution with the following density function [18], [20]:

$$f_{Y_n}(t) = \frac{(\lambda t)^{n-1}}{(n-1)!}\lambda e^{-\lambda t}. \tag{28}$$

The Laplace transform for the density function of Y_n is given by

$$f^*_{Y_n}(s) = \left(\frac{\lambda}{s+\lambda}\right)^n. \tag{29}$$

Let $R1$, $R2$, and $R3$ be random variables that represent the processing times of DL synchronization, HO ranging, and registration, respectively. Assume that $R1$, $R2$, and $R3$ also have mixed-Erlang distributions with the following density functions [18], [19], [21]:

$$f_{R1}(t) = \sum_{i=1}^{I} a_i \frac{(\beta_i t)^{p_i-1}}{(p_i-1)!}\beta_i e^{-\beta_i t} \tag{30}$$

$$f_{R2}(t) = \sum_{j=1}^{J} b_j \frac{(\phi_j t)^{q_j-1}}{(q_j-1)!}\phi_j e^{-\phi_j t} \tag{31}$$

$$f_{R3}(t) = \sum_{k=1}^{K} c_k \frac{(\varphi_k t)^{r_k-1}}{(r_k-1)!}\varphi_k e^{-\varphi_k t} \tag{32}$$

where

$$\sum_{i=1}^{I} a_i = 1 \quad \sum_{j=1}^{J} b_j = 1 \quad \sum_{k=1}^{K} c_k = 1.$$

The corresponding Laplace transforms are shown as follows:

$$f^*_{R1}(s) = \sum_{i=1}^{I} a_i \left(\frac{\beta_i}{s+\beta_i}\right)^{p_i} \tag{33}$$

$$f^*_{R2}(s) = \sum_{j=1}^{J} b_j \left(\frac{\phi_j}{s+\phi_j}\right)^{q_j} \tag{34}$$

$$f^*_{R3}(s) = \sum_{k=1}^{K} c_k \left(\frac{\varphi_k}{s+\varphi_k}\right)^{r_k}. \tag{35}$$

Let Γ_n denote the nth packet that is buffered during the SDT. In the DL part, Γ_n denotes the nth packet that is buffered at the BS, and in the UL part, Γ_n denotes the nth packet that is buffered at the MS during the SDT. Then, we have

$$\Gamma_n = \begin{cases} 1, & Y_n \le R1 + R2 + R3 \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

Let Q_{SDT} denote the number of buffered packets during the SDT. Therefore, Q_{SDT} can be expressed as

$$Q_{\mathrm{SDT}} = \sum_{n=1}^{\infty} E(\Gamma_n) \tag{37}$$

$$= \sum_{n=1}^{\infty} \Pr(Y_n \le R1 + R2 + R3). \tag{38}$$

Let $T_e = R1 + R2 + R3$. Then, (38) can be rewritten as

$$= \sum_{n=1}^{\infty} \Pr(Y_n \le T_e)$$

$$= \sum_{n=1}^{\infty} \int_{T=0}^{\infty} f_{T_e}(T) \left[\int_{t=0}^{T} \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} dt\right] dT$$

$$= \sum_{n=1}^{\infty} \int_{T=0}^{\infty} f_{T_e}(T) \left[1 - \sum_{k=0}^{n-1} \left(\frac{\lambda^k}{k!}\right) T^k e^{-\lambda T}\right] dT$$

$$= \sum_{n=1}^{\infty} \left[\int_{T=0}^{\infty} f_{T_e}(T) dT - \sum_{k=0}^{n-1} \left(\frac{\lambda^k}{k!}\right) \int_{T=0}^{\infty} f_{T_e}(T) T^k e^{-\lambda T} dT\right]$$

$$= \sum_{n=1}^{\infty} \left[f^*_{T_e}(s)|_{s=0} + \sum_{k=0}^{n-1} (-1)^{k+1} \left(\frac{\lambda^k}{k!}\right) \frac{d^k}{ds^k} f^*_{T_e}(s)|_{s=\lambda}\right]. \tag{39}$$

Since $T_e = R1 + R2 + R3$ and $f_{T_e}(t)$ denote the density function of T_e, by applying the convolution property, its Laplace transform $f^*_{T_e}(s)$ is given as follows:

$$\begin{aligned} f^*_{T_e} &= f^*_{R1} f^*_{R2} f^*_{R3} \\ &= \left[\sum_{i=1}^{I} a_i \left(\frac{\beta_i}{s+\beta_i}\right)^{p_i}\right]\left[\sum_{j=1}^{J} b_j \left(\frac{\phi_j}{s+\phi_j}\right)^{q_j}\right] \\ &\quad \times \left[\sum_{k=1}^{K} c_k \left(\frac{\varphi_k}{s+\varphi_k}\right)^{r_k}\right]. \end{aligned} \tag{40}$$

Let Tb_n denote the duration of the nth packet that is buffered at the target BS (DL part) or the MS (UL part). Thus, the mean time of a packet buffered in the target BS (DL part) or in the MS (UL part) can be expressed as

$$\begin{aligned} \Theta &= Q^{-1}_{\mathrm{SDT}} \sum_{n=1}^{\infty} E(Tb_n) \qquad (41) \\ &= Q^{-1}_{\mathrm{SDT}} \sum_{n=1}^{\infty} \int_{T=0}^{\infty} f_{T_e}(T) \int_{t=0}^{T} f_{Y_n}(t)(T-t)dtdT \\ &= Q^{-1}_{\mathrm{SDT}} \sum_{n=1}^{\infty} \left[\int_{T=0}^{\infty} f_{T_e}(T) T \int_{t=0}^{T} f_{Y_n}(t)dtdT \right. \\ &\qquad \left. - \int_{T=0}^{\infty} f_{T_e}(T) \int_{t=0}^{T} f_{Y_n}(t)tdtdT\right] \\ &= Q^{-1}_{\mathrm{SDT}} \sum_{n=1}^{\infty} \left\{\left[\int_{T=0}^{\infty} f_{T_e}(T)TdT \right.\right. \\ &\qquad \left. - \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} \int_{T=0}^{\infty} T^{k+1} e^{-\lambda T} dT\right] \\ &\qquad - \left(\frac{n}{\lambda}\right)\left[\int_{T=0}^{\infty} f_{T_e}(T)dT \right. \\ &\qquad \left.\left. - \sum_{k=0}^{n} \frac{\lambda^k}{k!} \int_{T=0}^{\infty} f_{T_e}(T) T^k e^{-\lambda T} dT\right]\right\} \\ &= Q^{-1}_{\mathrm{SDT}} \sum_{n=1}^{\infty} \left\{(-1)\left(\frac{d}{ds} f^*_{T_e}(s)|_{s=0} + \frac{n}{\lambda} f^*_{T_e}(s)|_{s=0}\right)\right. \\ &\qquad + \frac{n}{\lambda} f^*_{T_e}(s)|_{s=\lambda} + \sum_{k=0}^{n-1} (-1)^k \frac{\lambda^k}{k!} \\ &\qquad \left. \times \left[\frac{k+1-n}{k+1}\right] \frac{d^{k+1}}{ds^{k+1}} f^*_{T_e}(s)|_{s=\lambda}\right\}. \qquad (42) \end{aligned}$$

VI. Performance Evaluation

Based on the analytic model derived in Section V, we first present analytic results of various IEEE 802.16e HO-enhanced schemes, including the proposed NFHO scheme. A simulation model using NS-2 was also developed to validate the analytic model. The simulation model is based on the analytic model described in Section V-C. The same parameter settings shown in Table III [18] were used for simulation as well as for analysis. In the simulation, the packet-interarrival time and the processing times for DL synchronization, HO ranging, and registration were generated by the random functions of their corresponding distributions. If a BS accepts an incoming HO MS, the SDT and the packet arrival time of each packet during the SDT can be computed. Thus, we can obtain the expected number of buffered packets. Furthermore, the packet loss probability can be obtained by monitoring the usage of the HO packet buffer pool. In Section VI-A, the performance evaluation results, including analytic and simulation results, are compared among the existing HO schemes and the proposed NFHO scheme. In addition, based on the analytic model of the proposed inter-CSSC HO procedure, Section VI-B presents the expected number of buffered packets and the packet loss probability of the proposed NFHO.

TABLE III
Parameter Settings for Performance Evaluation

Variable	Parameters
R1	$a_1 = a_2 = 0.5$ $\beta_1 = 200$, $\beta_2 = 66.67$ (packets/sec)
R2	$b_1 = b_2 = 0.5$ $\phi_1 = 100$, $\phi_2 = 66.67$ (packets/sec)
R3	$c_1 = c_2 = 0.5$ $\varphi_1 = 100$, $\varphi_2 = 40$ (packets/sec)
M	200 (packets)
ξ/ε	20

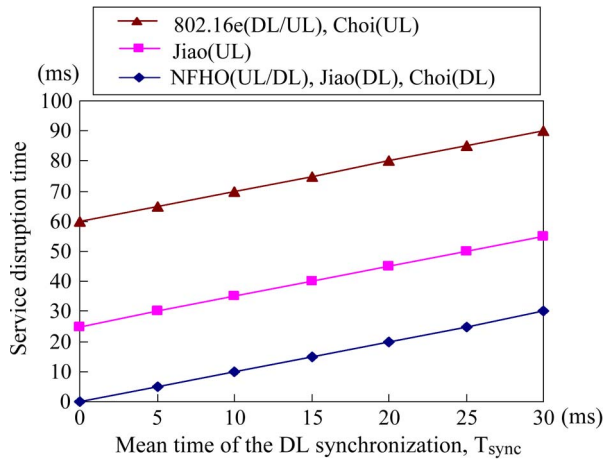

Fig. 10. HO service-disruption time versus T_{Sync}.

A. Performance Evaluation Results Among the Existing HO Schemes and the Proposed NFHO Scheme

To evaluate the proposed NFHO scheme with respect to the existing HO schemes, we first compare the SDT. Based on Table II, Fig. 10 shows the analytic results of the SDT. The values of T_{Rng} and T_{Reg} were set to 25 and 35 ms [14], respectively. We found that in the DL part, the existing HO schemes and the proposed NFHO scheme have the same SDT, and all outperform the IEEE 802.16e HO scheme. In the UL part, the NFHO scheme has a shorter SDT than the other schemes because the NFHO scheme allows the UL data transmission to immediately restart after the DL synchronization. Note that

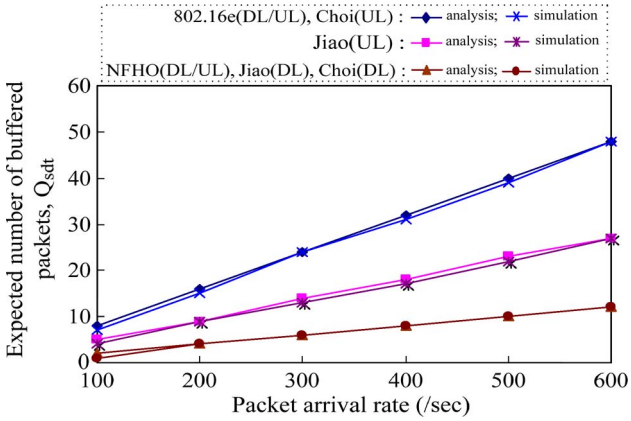

Fig. 11. Q_{SDT} versus packet-arrival rate.

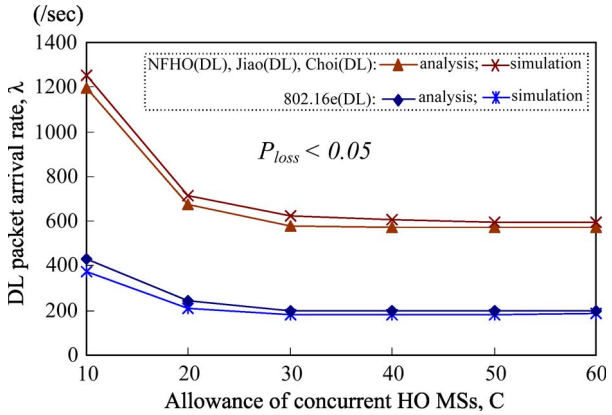

Fig. 12. Supported DL packet-arrival rate versus the allowance of concurrent HO MSs under a given packet loss probability constraint ($P_{\mathrm{loss}} < 0.05$).

T_{Sync} is set to 20 ms [15]. From Fig. 10, it is shown that the proposed NFHO scheme reduces the DL SDT by 75% compared with the IEEE 802.16e HO scheme, and it also reduces the UL SDT by 55.6% compared with the Jiao *et al.* [15] scheme and by 75% compared with both the Choi *et al.* [14] and IEEE 802.16e HO schemes. As mentioned in Section IV, only an add-on software component with a small cost is required to support the proposed NFHO scheme. Therefore, our NFHO scheme is cost effective.

In the following performance evaluation (see Figs. 11 and 12), including analysis and simulation, the parameters listed in Table III [18] were used, unless there were individual parameter assignments in each evaluation. Note that even with other parameter settings, we can still obtain similar performance evaluation results. Assume that $R1$, $R2$, and $R3$ all have mixed-Erlang-2 distributions. By (37), we investigate the expected number of buffered packets (Q_{SDT}) among the proposed NFHO and the existing HO schemes. Fig. 11 shows the expected number of buffered packets during the HO execution procedure. It can be seen that Q_{SDT} linearly increases with the increase of packet-arrival rate λ. This is because the packets are buffered in the BS (for DL packets) or the MS (for UL packets), and MS and BS stop transmitting during the SDT. The DL part of the proposed NFHO, the Jiao *et al.* [15], and the Choi *et al.* [14] schemes all have the minimum Q_{SDT} because these schemes can restart DL transmission immediately after the DL is synchronized ($R2 = R3 = 0$). In the UL part, the Q_{SDT} of the proposed NFHO ($R2 = R3 = 0$) is the lowest, followed

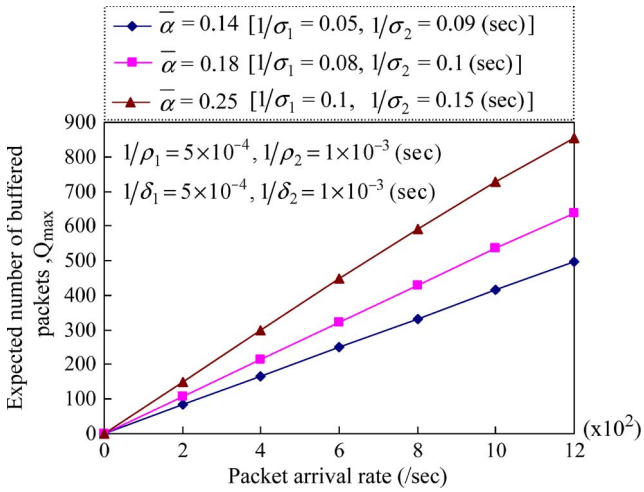

Fig. 13. $Q_{\max}$ versus packet arrival rate.

by the Q_{SDT} of Jiao *et al.* ($R3 = 0$), and the Choi *et al.* and IEEE 802.16e HO schemes have larger Q_{SDT} than the others. In Fig. 11, the analytic and simulation results are very close, which confirms the validity of our analytic model.

With a given packet loss probability constraint, by (18), (26), (37), and (41), we analyze the relationship between the allowance of concurrent HO MSs (i.e., C) and the supported DL packet-arrival rate (i.e., λ) per MS. We assume that each MS has the same UL/DL packet rate. As shown in Fig. 12, by restricting the packet loss probability to be less than 0.05 during HO, in the DL part, the NFHO, the Jiao *et al.*, and Choi *et al.* schemes can support a higher packet-arrival rate than the IEEE 802.16e HO scheme. Fig. 12 also shows that λ becomes stable, even when C keeps increasing. It is because that a given ξ/ε ratio implies a limit (threshold) on the number of concurrent HO MSs in a BS. λ gradually decreases until C exceeds by the threshold. Again, in Fig. 12, the analytic and simulation results are very close, which confirms the validity of our analytic model. In addition, in the UL part, the supported UL data rate depends on the number of available packet buffers locally allocated in the MS. The UL performance evaluation results among the proposed NFHO and the existing HO schemes has been shown in Fig. 11. In the UL part of Fig. 11, with a given Q_{SDT}, we found that the NFHO scheme can support a larger packet-arrival rate than the other schemes. Note that a fixed Q_{SDT} is equivalent to assigning a P_{loss} because the available packet buffers are allocated in the MS.

B. Additional Performance-Evaluation Results of the Proposed NFHO Scheme

With different transmission delays α, we first investigate the relationship between the expected number of buffered packets in the target BS and the packet-arrival rate. We assume that the random variables μ, ν, and α have mixed-Erlang-2 distributions [18], [19], [21]. The coefficients for μ, ν, and α were set to $\mu_1 = \mu_2 = 0.5$, $\nu_1 = \nu_2 = 0.5$, and $\alpha_1 = \alpha_2 = 0.5$, respectively [18]. Other parameter settings are shown in each figure. In addition, the mean time of HO message-processing delay is computed as $\overline{\alpha} = 2(\alpha_1/\sigma_1 + \alpha_2/\sigma_2)$. By (10), Fig. 13 shows the relationship between the expected number of buffered packets in the target BS and the packet-arrival rate under

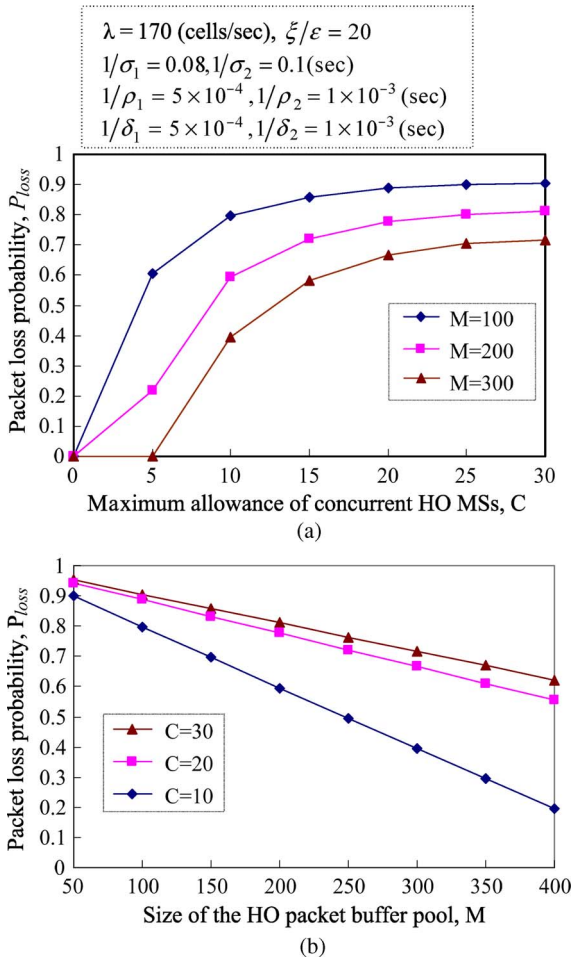

Fig. 14. (a) Packet loss probability versus C. (b) Packet loss probability versus M.

three different $\overline{\alpha}$. It was observed that the expected number of buffered packets $(Q_{\max})$ is positively correlated to the packet-arrival rate (λ) of the MS. An increase in λ leads to a larger $Q_{\max}$, and a larger $\overline{\alpha}$ also results in a larger $Q_{\max}$. Note that even with other parameter settings, we still can obtain similar results.

Since the size of the HO packet buffer pool is limited, the packet loss probability during HO is affected by the packet-arrival rates of ongoing HO MSs and the available packet buffers in the pool. We consider three different sizes of the HO packet-buffer pool, and the BS can only support a limited number of concurrent HO requests. By (18), (21), and (26), we evaluate the relationship among C (the allowance of concurrent HO MSs), M (the size of the HO packet-buffer pool), and P_{loss} (packet loss probability). Fig. 14(a) shows that the packet loss probability (P_{loss}) of an HO MS increases with the increase of the allowance of concurrent HO MSs (C). The BS that provides a larger size of HO packet-buffer pool (M) has a lower packet loss probability during HO. Note that P_{loss} approximates zero when C and M are set to 5 and 300, respectively. This is because M is large enough to accommodate all arrival packets from C HO MSs. With the same traffic parameters, Fig. 14(b) also shows the relationship among P_{loss}, C, and M. The parameter settings, which are listed on top of Fig. 14(a), will also be used in Figs. 15–17, unless there are individual settings listed in each figure.

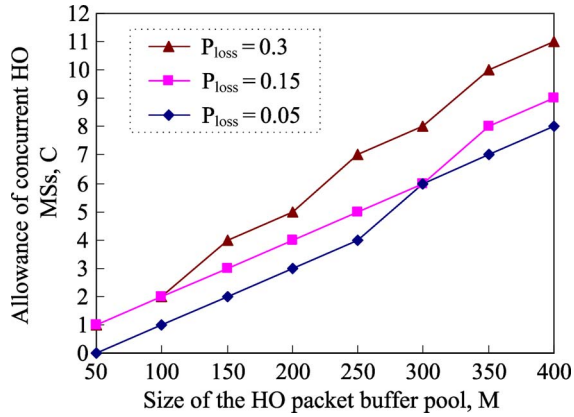

Fig. 15. Relationship between M (the size of the HO packet buffer pool) and C (the maximum allowance of concurrent HO MSs) under a given P_{loss} (packet loss probability) constraint.

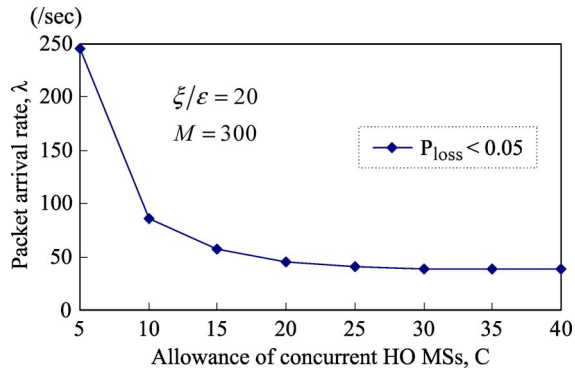

Fig. 16. Packet-arrival rate (λ) of each MS versus the allowance of concurrent HO MSs (C) under a given packet loss probability (P_{loss}).

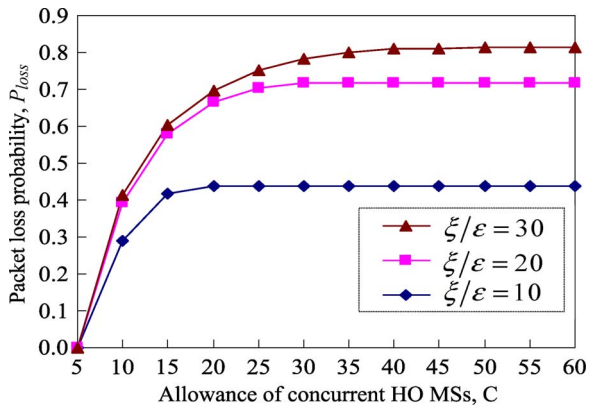

Fig. 17. P_{loss} versus C under a given ξ/ε.

Packet loss probability is an important QoS parameter for HO. According to the size of the HO packet buffer pool allocated in a BS, the BS can control the allowance of concurrent HO MSs to meet the packet loss probability requirement. With a given packet loss probability, by (18), (21), and (26), we evaluate the relationship between C (the allowance of concurrent HO MSs) and M (the size of the HO packet-buffer pool). Fig. 15 shows that when allocating a larger M, the BS can accept a higher C while still meeting the packet loss probability constraint. With a given P_{loss}, Fig. 16 also shows that the higher C that the BS allows, the lower λ (packet-arrival rate) that the BS can support to each MS. Note that without loss of generality,

the ξ/ε ratio was set to 20, and M was set to 300. It can be observed that when we increase C, the supported λ of each MS decreases. The decreasing of λ is sharp until C is greater than 20. It is because the expected value of packet-holding time grows until C exceeds the ξ/ε ratio. Similar results are also shown in Fig. 17. Given three different ξ/ε ratios, it can be seen that when we increase C, P_{loss} grows sharply until C is greater than the corresponding ξ/ε ratio. Figs. 16 and 17 also show that the BS can guarantee the packet loss probability requirement for ongoing HO MSs by controlling the allowance of concurrent HO MSs (C) and the ξ/ε ratio. Therefore, to provide proper QoS (e.g., the packet loss probability, P_{loss}) to ongoing HO MSs, (18), (21), and (26), which were the basis of Figs. 16 and 17, can assist in setting up an appropriate admission-control policy to grant or reject incoming HO requests.

VII. Conclusion

In this paper, we have presented a novel IEEE 802.16e network architecture. Based on this architecture, an NFHO scheme has been proposed to shorten the SDT during HO. Since the proposed NFHO scheme resolves the transport CID assignment and UL-synchronization issues before HO ranging, the UL/DL data transmission can be restarted before the MS proceeds to the HO ranging. In this scheme, DL packets are bicast to the target BS as well as the serving BS before breaking the connection with the serving BS. In addition, we have also developed an analytic model to evaluate the expected number of buffered packets, the packet loss probability, and the SDT. Based on the analytic model, we can also evaluate the relationship among the packet loss probability, packet-arrival rate, number of concurrent HO MSs, and the size of the HO packet-buffer pool. Performance evaluation results have shown that the proposed NFHO scheme reduces the DL SDT by 75% compared with the IEEE 802.16e hard HO scheme. It also reduces the UL SDT by 55.6% compared with the Jiao *et al.* scheme and by 75% compared with both the Choi *et al.* scheme and the IEEE 802.16e hard HO scheme. In summary, the NFHO scheme has the best performance in terms of the expected number of buffered packets, packet loss probability, and SDT among existing hard HO schemes for the IEEE 802.16e. The derived analytic model can also be integrated to an admission-control policy to provide proper QoS to ongoing HO MSs.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments that improved the quality of this paper.

References

[1] IEEE Standard for Local and Metropolitan Area Network—Part 16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE Std. 802.16-2004, Oct. 2004.
[2] IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE Std. 802.16e-2005, Feb. 2006.
[3] S. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie, "Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 865–881, Feb. 2009.
[4] B.-J. Chang and J.-F. Chen, "Cross-layer-based adaptive vertical handoff with predictive RSS in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3679–3692, Nov. 2008.
[5] Y.-H. Han, H. Jang, J. Choi, B. Park, and J. McNair, "A cross-layering design for IPv6 fast handover support in an IEEE 802.16e wireless MAN," *IEEE Netw.*, vol. 21, no. 6, pp. 54–62, Nov./Dec. 2007.
[6] J. Park, D.-H. Kwon, and Y.-J. Suh, "An integrated handover scheme for fast mobile IPv6 over IEEE 802.16e systems," in *Proc. IEEE VTC*, Sep. 2006, pp. 1–5.
[7] Y.-W. Chen and F.-Y. Hsieh, "A cross layer design for handoff in 802.16e network with IPv6 mobility," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2007, pp. 3847–3851.
[8] Y.-S. Chen, K.-L. Chiu, K.-L. Wu, and T.-Y. Juang, "A cross-layer partner-assisted handoff scheme for hierarchical mobile IPv6 in IEEE 802.16e," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2008, pp. 2669–2674.
[9] D. H. Lee, K. Kyamakya, and J. P. Umondi, "Fast handover algorithm for IEEE 802.16e broadband wireless access system," in *Proc. 1st Int. Symp. Wireless Pervasive Comput.*, Jan. 2006, pp. 1–6.
[10] P.-S. Tseng and K.-T. Feng, "A predictive movement based handover algorithm for broadband wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2008, pp. 2834–2839.
[11] J. Chen, C.-C. Wang, and J.-D. Lee, "Pre-coordination mechanism for fast handover in WIMAX networks," in *Proc. 2nd Conf. Wireless Broadband Ultra Wideband Commun.*, Aug. 2007, p. 15.
[12] O. C. Ozdural and H. Liu, "Mobile direction assisted predictive base station switching for broadband wireless systems," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 5570–5574.
[13] S. Cho, J. Kwun, C. Park, J.-H. Cheon, O.-S. Lee, and K. Kim, "Hard handoff scheme exploiting uplink and downlink signals in IEEE 802.16e systems," in *Proc. IEEE VTC*, May 2006, pp. 1236–1240.
[14] S. Choi, G.-H. Hwang, T. Kwon, A.-R. Lim, and D.-H. Cho, "Fast handover scheme for real-time downlink services in IEEE 802.16e BWA system," in *Proc. IEEE VTC*, May 2005, pp. 2028–2032.
[15] W. Jiao, P. Jiang, and Y. Ma, "Fast handover scheme for real-time application in mobile WiMAX," in *Proc. IEEE ICC*, Jun. 2007, pp. 6038–6042.
[16] Y. Xiao, H. Li, Y. Pan, K. Wu, and J. Li, "On optimizing energy consumption for mobile handsets," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1927–1941, Nov. 2004.
[17] S. Asmussen, *Applied Probability and Queues*. New York: Wiley, 1987.
[18] A.-C. Pang, Y.-B. Lin, H.-M. Tsai, and P. Agrawal, "Serving radio network controller relocation for UMTS all-IP network," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 4, pp. 617–629, May 2004.
[19] X. Wang, P. Fan, J. Li, and Y. Pan, "Modeling and cost analysis of movement-based location management for PCS networks with HLR/VLR architecture, general location area and cell residence time distributions," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3815–3831, Nov. 2008.
[20] L. Kleinrock, *Queuing Systems*, vol. 1, *Theory*. New York: Wiley, 1975.
[21] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1062–1072, Jul. 1999.

Lung-Sheng Lee received the M.S. degree in computer and information science in 1997 from the National Chiao Tung University, Hsinchu, Taiwan, where he is currently working toward the Ph.D. degree in computer science and information engineering with the Department of Computer Science.

His research interests include personal communications service networks, WiMAX networks, and real-time operating systems.

Kuochen Wang (M'86) received the B.S. degree in control engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1978 and the M.S. and Ph.D. degrees in electrical engineering from the University of Arizona, Tucson, in 1986 and 1991, respectively.

He is currently a Professor with the Department of Computer Science and the Director of the Institute of Computer Science and Engineering, National Chiao Tung University. From 1980 to 1984, he was a Senior Engineer with the Directorate General of Telecommunications in Taiwan. He served in the army as a second lieutenant communication platoon leader from 1978 to 1980. His research interests include wireless (ad hoc/sensor) networks, mobile (cloud) computing, and power management for multimedia portable devices.