Managing connection costs in heterogeneous wireless networks

Abd-Elhamid M. Taha1*, Hossam S. Hassanein2 and Hussein T. Mouftah3

1 Electrical Engineering, Alfaisal University, Saudi Arabia
2 School of Computing, Queen’s University, Canada
3 School of Information Technology and Engineering, University of Ottawa, Canada

ABSTRACT

Common Radio Resource Management techniques have shown great promise in both enhancing network operation and user satisfaction. Such gains are achieved through the joint management of the individual access technologies in a Heterogeneous Wireless Network. The objective of this work is to expand on the existing body of work to accommodate heterogeneity not just at the traditional access-network level but to other connectivity modes such as dynamic spectrum access. Such modes affect operator profitability in both the long and short terms. Specifically, we explore the design of a cost-management model that adapts to the short-term variability in connectivity costs. We also display the operational aspects and effectiveness of this functionality through both simulation and an analytical model. Copyright © 2012 John Wiley & Sons, Ltd.

KEYWORDS

wireless overlay networks; heterogeneous wireless networks; vertical handovers; connectivity cost

1. INTRODUCTION

Both of the recently recognized fourth-generation technologies—Third-generation Partnership Project’s Long Term Evolution (LTE)-Advanced and IEEE’s WirelessMAN-Advanced (IEEE 802.16m)—are a strong support to interwork different access technologies. The resulting composite networks are often labeled Heterogeneous Wireless Networks (HWNs) [1]. Functionally, HWNs enable appropriately equipped terminals—i.e., mobile terminals with multiple radio interface—to initiate sessions through the best access technology, be that in terms of cost, Quality of Service (QoS) (i.e., delay, jitter, reliability, etc.), security, or other possible connectivity parameters [2]. With multihoming facilitated by IPv6 at the terminal level, it will soon become possible that the terminal selects the best combinations of technologies for connectivity—e.g., cellular for voice, Wireless Local Area Network (WLAN) for data, and so on.

A strong feature of HWNs, however, is that they also enable seamless inter-technology handovers, called Vertical Handovers (VHs), whereby a terminal can maintain sessions across the different technologies without having to disconnect and reconnect [3,4]. With VHs, opting for the best technology becomes persistently viable throughout the user’s activity rather than just at session initiation. Much investment is thus made in establishing guidelines for VH decisions, including attribute collection, network selection, and representation of user and network policies [5]. The bulk of interest, however, has been directed at satisfying user requirements with little regard to operator requirements. For instance, it is possible that an operator may direct users from one technology to another to relieve congestion or to enhance overall network utilization. Such a need has been addressed in the literature in various works within the general context of Common Radio Resource Management (CRRM), whereby the resources of the different access technologies are managed jointly and in a manner that improves overall network performance and user satisfaction—see e.g., [6,7]. In a previous work [8], we explored a similar use of handovers as an RRM tool to aid the operator in achieving certain short-term objectives.

A drawback in such previous studies, however, is the focus on access technologies as the only factor of heterogeneity in future networks. At the same time, where
the studies considered connectivity costs, the standing assumption was that these costs did not exhibit short-term variability. Although these assumptions may very well have been valid in traditional settings, the dynamic and user-centric nature of HWNs introduces further considerations that call for more actions at shorter decision time frames. For example, the effect of network load on service cost makes the cost structure—i.e., the structure of costs for individual connectivity modes, highly dependent on demand density. The recent viability of cognitive radios and dynamic spectrum also has more critical effects. Such radios exploit gaps or holes of non-utilized times in different bands across the radio spectrum, which in turn opens the door for spectrum trading and brokerage between operators and between operators and users.

The objective of this paper is thus to investigate the design and operation of a cost-management functionality, which would reside at higher network management levels—i.e., a cluster or several clusters of the network. The functionality would be transparent to the underlying modes of connectivity—i.e., point-to-multipoint, relay, or spectrum. It would also be adaptive to both long-term and short-term variabilities in the network’s cost structure.

The remainder of this work is organized as follows. In the following section, we elaborate on the motivation for designing a dedicated module for Service-Delivery Cost-Reduction (SDCR) and present an overview of related work. In Section 3, we discuss considerations for estimating the cost of a connectivity mode, and in Section 4, we introduce the elements of designing a cost-management module, in addition to the processes involved in user identification and selection for migration. Through simulation and analytical modeling, we elaborate on certain operational aspects of cost-management functionality in Section 5. Finally, in Section 6, we conclude.

2. MOTIVATION AND RELATED WORK

For a wireless network operator, the essential profitability measures may be realized through network management based on long-term observations of user and network characteristics. Through these observations, the cost of various connectivity measures can be discerned, and the selection of the most cost-effective connectivity becomes possible. In HWNs, however, substantial unpredictability is expected to characterize both user and network dynamics. The possible permutations of terminal capabilities, application requirements, and user preferences, both in time and space, render the cost-reduction objective more challenging to achieve and limit the effectiveness of policies such as cost-based admission control. The need therefore exists for mechanisms that enable the persistent management of connectivity costs across the different technologies [3].

The aforementioned relates to the difficulty of keeping users associated with networks on the basis of the operator’s basic network costs. For example, delivery through a WLAN is cheaper than through cellular on the basis of equipment and licensing costs. Towards this end, a cost-based technology assignment has been investigated by Bria in [9]. The author contemplated the issue of technology selection for broadcast messages in a cellular network overlaid with a broadcasting system such as Digital Audio or Video Broadcast. However, the author addressed technology selection only at session initiation and not during an active session. The author’s discussion was also limited to broadcast transfers of known size.

The work in [9] is one of many studies with the general context of CRRM [1,6,7,10–14], all of which advocate and verify that jointly managing the resources of the different access technologies would result in both higher operator profitability and user satisfaction. Such studies further considered various attributes for both networks and users, including static and dynamic characteristics. Our concern is to provide a more expanded view of cost considerations that includes not only networks with heterogeneous access technologies but also those with heterogeneous topologies and dynamic spectrum allocations. We also highlight the short-term variability of connectivity cost that is projected to characterize future wireless networks, as will be elaborated upon in Section 3.

Indeed, several factors are involved in evaluating the cost [15,16]. For example, interest in pricing-based RRM techniques is essentially motivated by the fact that the cost of service delivery actually increases as the load on the utilized network increases [17,18].

The effect of introducing cognitive radios in future wireless networks will also impact connectivity costs in the short-term. With observations that licensed and regulated spectrum is largely underutilized [19], interest is now high in overcoming this underutilization through exploiting spectrum gaps or holes in licensed spectrum. This allows more entities to share the spectrum. The capability to sense and switch between different bands in the spectrum is technically provided by cognitive radios [20]. However, it is expected that, in regulating access to spectrum bands, forms of negotiations and brokerage will exist [21,22]. Among several ramifications, this notably leads to variability in the operator’s cost to access a certain spectrum segment [23,24].

We believe that the notion of Operator Motivated Vertical Handover (OMVH) [8] holds great potential in achieving our cost-reduction objective. In essence, OMVHs were introduced to complement a bulk of work that focused on satisfying only user requirements using VHs, realizing what we labeled as User-motivated Vertical Handovers. We recognized that there are instances when the network may utilize VHs to its advantage such as migrating users out of a congested access technology to another, less-loaded technology in the overlay. In previous work, we focused on such reactive applications of OMVH employed at an overlay level and that responded to and assisted other modules, primarily those concerned with admission and congestion control. The focus of this paper...
is the design of proactive OMVH modules that can be applied at either the cluster or the operator level. We choose SDCR as an encompassing example because it may also include load balancing, a common long-term objective for network operators.†

3. ELEMENTS OF COST EVALUATION

In this section, we elaborate on some of the factors involved in evaluating the cost of service delivery.

Typically, equipment cost and licensing affect an operator’s connectivity costs over the long term and are relatively stable. In performing SDCR, it might be expected that an existing cost structure will dictate a certain flow of user migrations (i.e., from cellular to WLAN). However, when accounting for the expected duration of calls and depending on the type of service, users can be migrated in the opposite direction in minimizing the total service-delivery cost.

Another source of cost variability is induced when a network load approaches or reaches congestion. Whereas some studies have attributed this cost as a marginal cost on the user side, the cost incurred on the operator is realized by dissatisfied users seeking other operators and opportunity costs of serving a certain user in a specific network [26]. To evaluate this cost,‡ a network needs to predict the load within a limited duration appropriate to the decision time frame. Several operationally efficient and online measurement techniques have been proposed for aggregate user traffic; see, for example, the work in [27–30]. At the individual connection level, a connection’s duration can easily be extracted from per-service statistics. As for usage, the per-user effective requirements can also be estimated [31]. Note that evaluating the cost over a certain duration should not imply that the service is delivered over the whole duration.

As aforementioned, the introduction of cognitive radio introduces a further variability in spectrum-access costs for network operators [21,22]. We note two of the possible ways in which access costs can be evaluated. In the first, an added entity, the spectrum broker, collects information about spectrum demand and supply. Access cost can thus be evaluated through a regulated price or through auctioning. Note that the brokerage entity can also emulate a cooperative setting in which the different operators negotiate access costs. In the second scenario, in which a non-cooperative mode is established, game theoretic approaches can be utilized to evaluate estimated access costs.

†An abridged exposition of the work presented herein was in [25]. This work expands on the arguments presented there and offers a more comprehensive performance evaluation.
‡The choice of methods for evaluating service cost has no bearing on the generality of the proposed Service-Delivery Cost-Reduction.
quality. For selection, SDCR explores the potential cost reduction that can be achieved given the information gathered from SDCR’s first phase and the identification process. Once user selection is made, signals for migrations are initiated so as to change the associations of users’ terminals.

Figure 1 schematizes an operational timeline in which SDCR’s inter-operation time is denoted as $\tau_f$. The valuation of $\tau_f$, which we remark on in Section 5.4, depends largely on the aggregate characteristics of the services delivered. Over relatively longer durations, however, the value of $\tau_f$ can be adjusted according to network conditions.

4.2. Definitions and notations

We consider an operator overseeing a wireless overlay with technologies that are managed by either a single management entity or a group of cooperating management entities. We denote the set of networks in the wireless overlay by $N$ and the set of services provided by the operator in the wireless overlay by $S$. It is assumed that the services are commonly defined in the different networks, so the possibility exists that the services are offered in different networks with different QoS guarantees. Moreover, services need not be available in all of the networks at all times.

Let $U$ be the set of all users associated with all the networks and all services in the wireless overlay. Similarly, let $U^n$ and $U^n_s$, respectively, denote the set of users associated with all services in network $n \in N$ and the set of users associated with service $s \in S$ in network $n$. At times, we will require indicating the temporal dependence of the different sets—e.g., $U^n_s(t)$.

4.3. Identifying candidate users

To identify which users can be migrated from one access technology to another, a measure is required that represents the operator’s perspective of how useful it is to migrate a specific user at a particular time to the SDCR decision process. Such a measure would, for example, process the user’s active interfaces, the user’s applications and mobility profiles, and the user’s location. To facilitate this comparison, we introduced the notion of a migration’s worth in [8], where we discussed the notion at length and what is involved in its valuation. Here, we only extend the notation.

We compute the worth of $u_i$ to be moved from network $i$ to network $j$, denoted by $W^{i,j}(u_i)$ as follows.

$$W^{i,j}(u_i) = \Phi^{i,j}_\Pi(u_i) \cdot \Phi^{i,j}_\Sigma(u_i)$$

where $\Phi^{i,j}_\Pi(u_i)$ is a function of multiplicative factors and $\Phi^{i,j}_\Sigma(u_i)$ is a weighted summation of additive factors. The former strongly dictates the possibility of migration—e.g., whether the service is available in $j$ or whether $u_i$’s terminal does not have the required interface. They can also indicate other attributes that would prevent the network from persistently migrating users in a short duration of time. The additive factors mostly describe the quality of service delivery in network $j$. For example, they can characterize the variation of allocation between the two networks or signal quality. The set of candidate users using class $s$ between network $i$ and network $j$, denoted by $A^{i,j}_s$, can be populated as follows.

$$A^{i,j}_s = \left\{ u_x : u_x \in U^n_s, W^{i,j}(u_x) \geq W_{th} \right\}$$

It is important to note that (i) substantial signaling may be associated with each handover decision [5] and (ii) that a decision to migrate a certain user should ultimately result in benefits that outweigh the handover costs. At the same time, each handover decision puts the user’s connection at the risk of connection loss. Also, although the viability of handovers as a resource management tool makes it a powerful tool, users’ Service Level Agreement (SLA) should be strictly followed.

4.4. User selection

The outcome of engaging the SDCR module is sets of users, each indicating the users using service $s$ to be migrated from a network $i$ to network $j$. Each such set, denoted by $V^{i,j}_s$, will be selected from the respective set of candidate users—i.e., $V^{i,j}_s \subseteq A^{i,j}_s$. Respectively, denote by $V^{O,n}_s$ and $V^{O,n}_s$ service $s$ users to be migrated from
network $n$ to all other networks and those to be migrated from all other networks to network $n$: that is,

$$V_s^{n,O} = \bigcup_{j \neq n, n, j \in N} V_s^{n,j}$$

(3)

$$V_s^{O,n} = \bigcup_{i \neq n, i, n \in N} V_s^{i,n}$$

(4)

If the SDCR module was engaged at a certain time $t$, then the constituents of $U_s^j$ just after the engagement can be computed as follows.

$$U_s^j(t^+) = U_s^j(t) + V_s^{O,j}(t) - V_s^{I,O}(t)$$

(5)

where $t^+$ refers to the instance just after SDCR completes its operation.

Denote the cost of user $u_x$, where $u_x \in U$, in the interval between $t$ and $t + \tau_f$, by $C(u_x,t)$. Accordingly, the cost of service delivery to $U$ in the same duration can be expressed as

$$C(U(t_j)) = \sum_{u_x \in U} C(u_x,t).$$

(6)

The objective of the SDCR module can be stated as

$$\min_{V_s^{i,j} \in A_s^{i,j}} C \left(U(t^+)\right)$$

(7)

In seeking this objective, SDCR needs to consider certain constraints. For example, a user migration can be limited to a single technology instead of allowing for multihoming:

$$V_s^{i,j} \cap V_s^{i,k} = \emptyset \quad \forall i \neq j \neq k$$

(8)

Additionally, the operation of SDCR is bounded by the capabilities of the technologies within the overlay. Denote by $Q(U^n)$ the QoS allocations (e.g., bandwidth, provided for $U^n$) and by $Q^n$ the maximum QoS allocations available in network $n$, either absolutely or for the purpose of cost reduction. The constraints can be described as follows:

$$Q \left(U^n(t^+)\right) \leq Q^n$$

(9)

Equally important, SDCR operation must guaranteed that a user’s QoS is guaranteed, that is,

$$Q(u_x,t) \geq Q_{SLA}(u_x) \quad \forall u_x \in U$$

(10)

These are the basic constraints for the operation of any SDCR module. Other constraints can certainly be added. Several mechanisms for the solutions are possible, including linear programming and other assignment algorithms—e.g., online bin-packing heuristics.

Once the users to be migrated are selected, the selection is passed on to the relevant network management entities to perform the OMVHs.

5. CONSIDERATIONS FOR IMPLEMENTATION

As pointed out in Section 4.4, the core of the user-selection element reduces to an assignment problem that can be tractably solved—e.g., using offline bin-packing heuristics. There are other aspects of implementing SDCR, however; that must be taken into consideration. Our intent in this chapter is to explore operational aspects such as the possible transient response to an SDCR implementation, in addition to its response to variation in load and cost structure. We also look at the valuation of $\tau_f$, which dictates the frequency of operation, and we offer a Markov model than can be used in selecting an appropriate value.

5.1. Simulation environment

Simulation experiments were carried out in an event-driven simulation that was built utilizing C++ and MATLAB. The SDCR core was implemented through a Mixed Integer Linear Programming formulation solved using the GLPK package of the GNU project [32].

An overlay involving two networks, schematized in Figure 2, is used. We will refer to the network with the persistently larger coverage as network 1 and the other network as network 2. For purposes of evaluation, the coverage of the two networks are concentric. A fixed number of users was uniformly distributed over the area of the larger coverage. Users make connection requests with an aggregate inter-arrival time that is exponential with a controllable mean. The capacities of networks 1 and 2 are, respectively, 80 and 40 effective bandwidth units (ebus). In both networks, two classes of services are defined. The first, service 1, is allocated six ebus in network 1 and four ebus in network 2. The second service requires four and two ebus, respectively, in networks 1 and 2. The connection holding time of the first service is exponentially distributed with a mean of 150 s, regardless of the network choice. Connection holding times for the second service are fixed, regardless of the network choice, for a duration of

![Figure 2. An overlay of cellular and Wireless Local Area Network (WLAN) coverages.](image-url)
All users are assumed to be dual-mode users—i.e., can request and receive services in both networks. Experiments ran for a simulated time of 3600 s, and each result represents the outcome of ten experiments.

Note that the values used here are arbitrary and that other values were used in the intensive investigation performed, which displayed similar trends to the ones presented in the following text.

5.2. Observing the transient response

Figure 3 shows the temporal response for the two networks, with and without SDCR being employed. In the figure, the instantaneous service-delivery cost per user is shown. The cost structure was fixed for classes 1 and 2 (six and four in network 1, and two and two in network 2, all per ebu per second.) The aggregate arrival rate was set at 10 calls per minute with 60% of the calls seeking network 1 and 70% requesting service 1. Coverage ratio was set at 1:0.6, with users uniformly distributed over the coverage area of the larger coverage—i.e., network 1. This roughly means that 60% of the users will reside in the overlap area. At each decision instant, the migration of each user within the overlap area is randomly assigned. The worth threshold, \( W_{th} \), was set at 0.3, and SDCR was employed every 20 s. The instantaneous cost was also sampled every 20 s.

The figure shows the general effectiveness of the SDCR module. Note that in the graphs, there are two horizontal lines indicating the median cost per user, and in both instances, the median cost per user is lower by about 16% when the SDCR is employed.

A subtle aspect of the SDCR module can be observed in Figure 4 and that is the effect of the value of \( T_I \). In the figure, instead of employing SDCR every 20 s, it is employed every 200 s—i.e., a 10x factor. We note that the value of the median cost when SDCR is employed is higher than in the case of \( T_I = 20 \) s, resulting in a cost reduction of 5% compared with the 16% observed in Figure 3. More importantly, there are times when the instantaneous cost under SDCR is higher than when SDCR is not employed. This can be explained as follows. First, we previously mentioned that the choice of \( T_I \) depends, in part, on the characteristics of services employed. For exponentially distributed service 1 and fixed-duration service, the call holding times are 150 (mean) and 300, respectively. Hence, at \( T_I = 200 \) and minding the arrival rate, the SDCR module is bound to miss reduction opportunities. Second, the bound on worst-case cost occurs not when SDCR is disengaged but when each user is persistently assigned to most expensive network. Equivalently, optimal cost occurs if cost reduction is persistently applied. With respect to the overhead in terms of processing and signaling resulting from such a setting, the conscious design choice of making SDCR operate in a periodic manner should be noted.

In Figure 5, the same settings are applied with the exception of a demand surge between 1000 and 1500 s. During the surge, all new users request network 1, whereas the choice of service is evenly divided between the two services. It should be observed that the performance of the SDCR module is maintained during the surge, resulting in a reduction of 16%.
Figure 6. Instantaneous service-delivery cost per user, with and without Service-Delivery Cost-Reduction (SDCR) being engaged, cost structure changed randomly.

Figure 6 shows an extreme evaluation of the SDCR module. Service costs in network 1 were varied between 50% and 150% of their original values, whereas service costs in network 2 were varied between 100% and 300% of their original values. The changes were applied at every change in the simulation state—i.e., arrival, departure, engaging SDCR, or measurement. To make the figure more legible, we slightly reduced the sampling to a measurement every 40 s. Again, despite instantaneous cost variations, whether or not the SDCR is employed, employing SDCR results in a lower median cost. In this case, the achieved reduction is 18%.

Before proceeding, it is worth noting that the actual reduction in cost is highly dependent on the outcome of the cost-valuation scheme. If there is no significant variation in the cost of service delivery between technologies in the wireless overlay, the actual reduction caused by SDCR will accordingly decrease.

5.3. Effect of network load

Figures 7–10 show some operational aspects with and without employing SDCR and under different aggregate loads. The aggregate arrival rate was changed from 10 to 25 requests per minute, in steps of 2.5 requests per minute. To accommodate the increasing load, SDCR frequency was increased to once every 5 simulated s. In all, requests were evenly distributed between the two services. However, the portion of requests choosing a certain network was varied. Otherwise, all of the aforementioned settings were maintained, except for the surge and the variable cost settings.

In observing subplots (a) in Figures 7 and 8, we note that employing SDCR results in lowering the total cost (an average $4 \times 10^5$ in Figure 7(a) and up to $4 \times 10^5$) in Figure 8(a)). Note that in the figures, the percentage of requests made to network 1 was, respectively, 20% and 50%. Understanding the subplots (b) to (c) in the same two figures can be eased by looking at the cost structure, which basically has services less expensive to deliver in the network 2 than in network 1. This naturally results in blocking probability generally enhancing in network 1 and worsening in network 2.

This particular facet results in an important effect to be observed in Figure 9, where the percentage of requests made to network 1 is raised to 80%. Specifically, in Figure 9(a), where employing SDCR actually results in an increase in the total cost of service delivery. However, observing subplots (b) and (c) in Figure 9 indicates a substantial drop in the blocking probability when employing SDCR (unlike the effect in Figures 7 and 9) This indicates that employing SDCR with a given cost structure results in more users being accommodated in the system. As more users are accommodated in the system, the total delivery cost is bound to increase. Moreover, observing subplot (b) in Figure 10, where the per-user median instantaneous cost is plotted against the load, it becomes readily apparent that, despite the increase in the total cost of service delivery, the actual per-user cost is still reduced.

5.4. Frequency of operation

As shown earlier, the value of $\tau_f$ affects the efficiency of the SDCR module. The underlying intuition is that a small $\tau_f$ realizes persistent optimality at the cost of high overhead, whereas a large $\tau_f$ may render the SDCR module ineffective. Such considerations are relative to distributions of user mobility and behavior—e.g., connection durations—in addition to the rate at which the operator’s cost structure varies. In this section, we provide an analytical model for a network employing SDCR, and relating $\tau_f$ to the model’s rate of change. In doing so, we offer a formalism with which the designer can reason about the value of the value of $\tau_f$. Without loss of generality, we will maintain our considerations for an overlay of two technologies, whereby all users can connect to either access technology. For illustration, we limit the scope of considerations to interactions between the two networks.

We detail the terminal’s state transitions as follows:

1. **Arrival to overlay**: users initiate connection to either network 1 or 2 at rates $\lambda_{h1}$ and $\lambda_{h2}$, respectively.
2. **User-solicited handovers**: given user preferences and behavior, users migrate from network $i$ to network $j$ (i.e. $i \neq j$ and $j = \{1, 2\}$) at a rate of $\lambda_{u,i,j}$. This rate does not include the outcome of SDCR or any other OMVH module.
3. **SDCR migrations**: these take place at each $\tau_f$, with $\lambda_{s,i,j}$ as the rate for SDCR-caused handovers. Denote $\lambda_s (\lambda_{s,12} + \lambda_{s,21})$ as the rate at which SDCR is engaged.
4. **Departure from overlay**: users depart from network $i$ at a rate $\mu_i$. Deriving the rates $\mu_1$ and $\mu_2$ depends on the interaction between the arrival process, the users’ behavioral model, and the variation in the cost structure.
We can now consider the state transitions for the overlay as a whole. We indicate the state of the overlay by the tuple \((|U_1|, |U_2|)\), where \(|\cdot|\) operator computes the cardinality of the set within. Although this definition easily accommodates instances of multihoming, we assume that a user in the overlay can only be associated with a single network at a time—i.e., \(U_1 \cap U_2 = \emptyset\). Denote by the \(Q^n\) the maximum QoS allocation available for network \(n\) and by \(Q(U^n)\) the QoS allocation required for \(U^n\), with the conditions \(Q(U^n) \leq Q^n\) and \(n = 1\) or \(2\), bounding all states in the overlay—i.e., both user requirements and network capabilities are observed at all possible states. Denote the resulting state space by \(S\).

Ignoring rudimentary states, we readily distinguish three individual processes contributing to the overlay state transitions. In the following, \((a, b)\) is an arbitrary state.
(1) **Arrivals and departures:** this process is governed by arrivals to and departures from either network. The overlay transitions between the following states at the respective rates as follows:

- $(a, b)$ to $(a + 1, b)$ at rate $\lambda_{n1}$;
- $(a, b)$ to $(a, b + 1)$ at rate $\lambda_{n2}$;
- $(a, b)$ to $(a - 1, b)$ at rate $\mu_{n1}$; and
- $(a, b)$ to $(a, b - 1)$ at rate $\mu_{n2}$.

(2) **User-solicited handovers:** this process is governed by the user-initiated handovers between the networks. Herein, we will not expand on the considerations for this process because it has been
adequately addressed elsewhere (see, e.g., Hasib and Fapojuwo in [33]). We nevertheless indicate the resulting state transitions for the overlay:

- \((a, b)\) to \((a - 1, b + 1)\) at rate \(\lambda_{a,12}\)
- \((a, b)\) to \((a + 1, b - 1)\) at rate \(\lambda_{a,21}\).

(3) SDCR migrations: this process occurs at a rate of \(\lambda_s\). If the overlay is at state \((a, b)\), it can potentially move to any state \((i, j)\), given that \(a + b = i + j\) (crudely speaking, this means any state on the positive diagonal shown in Figure 11 on which \((a, b)\) lies). The rate at which the overlay changes between \((a, b)\) and a given \((i, j)\) depends on the probability that state \((i, j)\) is the one that yields the minimum aggregate service-delivery cost. Denote by \(D_{(a,b)}\) the set of states \((i, j)\) for which \(i + j = a + b\), including state \((a, b)\), and define the function \(C_{(i,j)}\) as the aggregate cost for supporting users comprising state \((i, j)\). The transition rate between states \((a, b)\) and \((i, j)\) in \(D_{(a,b)}\) is \(\lambda_s \cdot p_{ab,ij}\), where \(p_{ab,ij} = P(C_{(i,j)}) \leq C_{(k,l)}, \forall (k,l) \in D_{(a,b)}\). State transition for SDCR can then be described as follows:

- \((a, b)\) to \((i, j)\) in \(D_{(a,b)}\) at rate \(\lambda_s \cdot p_{ab,ij}\).

To elaborate, the SDCR migrations process mimics the SDCR’s operation as described in Section 4.4. In terms of operational complexity, a worst-case scenario is implied as the model assumes that the migration of all users is always possible. Denote the cost \(C_{(i,j)}\) as a random variable \(X_k\), where \(k = (i, j) \in D_{(a,b)}\), and assume that the number of states in \(D_{(a,b)}\) is \(n_d\). The costs of these \(n_d\) states can thus be considered independent random variables, each with identical pdf \(f(x)\). This is justified as the cost is evaluated on a per-user basis. Let \(Y = \min(X_1, X_2, \ldots, X_{n_d})\). To find \(p_{(i,j)}\), we have to find the Cumulative Distribution Function (CDF) of \(Y\) as it details the probability that state \((i, j)\) is the state with the minimum cost at the instant of engaging the SDCR module. Let \(P(Y < y < Y + dy)\) be the probability that one of the random variables falls in \(y, y + dy\) and that all others are greater than \(y\). The probability that a certain \(X_i, i = 1 \ldots n_d\) falls in this range is

\[
f(y)dy = \left(\int_y^\infty f(x)dx\right)^{n_d-1}\]

(11)

Because there are \(n\) ways to choose the state with the minimum cost, then

\[
f_Y(y) = n_d \cdot f(y)dy \left(\int_y^\infty f(x)dx\right)^{n_d-1}\]

(12)

Accordingly,

\[
f_Y(y) = n_d \cdot f(y)dy [1 - F(y)]^{n_d-1}\]

(13)

and the actual probability can be calculated by

\[
F_Y(y) = \int_0^y f_Y(y')dy'.
\]

The aforementioned description readily applies to a homogeneous and ergodic Markov chain, as can be directly inferred from the finiteness of the state space, the irreducibility of the chain, and an independence from \(\tau_f\). Furthermore, the description implies an \(N\)-dimensional \(M/G/m/m\) loss system, where \(N\) is the number of networks in the overlay. The basic performance measures for such a system can be directly computed [34]. Our interest here, however, is identifying an effective value for \(\tau_f\). This is viable through identifying the mean recurrence times for the states in the chain, which will help in realizing a sensible frame of reference in terms of how often the SDCR module should be engaged. To derive the mean recurrence rate, we require the chains’ steady-state probabilities,
denoted by \( \pi_{(i,j)} \) for state \((i, j)\). For a given state \((a, b)\), we have the following balance equations:

\[
\sum \text{Rate out} = \pi_{(a,b)} \cdot (\mu_{n1} + \mu_{n2} + \lambda_{n1} + \lambda_{n2} + \lambda_{u,12} + \lambda_{u,21}) \cdot \pi_{(a,b)} \tag{15}
\]

\[
\sum \text{Rate in} = \mu_{n1} \cdot \pi_{(a+1,b)} + \mu_{n2} \cdot \pi_{(a,b+1)} + \lambda_{n1} \cdot \pi_{(a-1,b)} + \lambda_{n2} \cdot \pi_{(a,b-1)} + \lambda_{u,12} \cdot \pi_{(a+1,b-1)} + \lambda_{u,21} \cdot \pi_{(a-1,b+1)}
\]

\[+ \left( \sum_{(i,j) \in D_{(a,b)}} p_{ij,ab} \right) \cdot \lambda_{S} \cdot \pi_{(i,j)} \tag{16} \]

Using \( \sum \text{Rate in} = \sum \text{Rate out} \) and adding the constraint that \( \sum_{(i,j)} \pi_{(i,j)} = 1 \), the steady-state probabilities can be identified. Given ergodicity, the mean recurrence time for state \((i, j)\), denoted \( M_{(i,j)} \), is \( 1/(\pi_{(i,j)} \cdot q_{(i,j)}) \). The mean recurrence rate of state \((i, j)\), denoted by \( q_{(i,j)} \), is the negative of the state’s Rate out at the steady state. Define \( M = \{M_{(i,j)} \mid (i, j) \in S\} \). A proper valuation of \( q_{(i,j)} \) is thus essentially a function of \( M \). Possibilities include the median of \( M \) or, more effectively, a percentile value dictating the module’s aggressiveness.

### 6. CONCLUSION

In this paper, we studied how proactive and periodic implementations can be used to attain certain operational objectives. Projecting variability in the cost of service delivery in future HWNs, we chose the objective of reducing the operator’s cost for service delivery to the mobile end user, showing the elements involved in the design and operation of SDCR. SDCR holds potential gains and stands robustly independent of the methods by which an operator’s cost are evaluated and the variability that these cost structures undergo. And although it is possible, on the basis of load distribution and cost structure, that SDCR results in a higher total cost because of increased admission, SDCR was found to persistently reduce the per-user cost.

Notwithstanding, in implementing SDCR, the designer should be aware of the variation of costs in the operator’s cost structure. This is because the magnitude of SDCR’s gains are dependent on this variation. Considerations should also be made to signaling required for handovers, as well as the risks associated with each handover decision. More generally, a more concrete evaluation of payoffs—i.e., effective impact of the decisions—should be made [35].

### ACKNOWLEDGEMENT

This work was made possible by the National Priorities Research Program grant from the Qatar National Research Fund (Member of Qatar Foundation).

### REFERENCES


1186
Managing connection costs in heterogeneous wireless networks


AUTHORS’ BIOGRAPHIES

Abd-Elhamid M. Taha has written and lecture extensively on broadband wireless networks, focusing on radio resource management techniques. He is the co-author of the recently published book ‘LTE, LTE-Advanced and WiMAX: Toward IMT-Advanced Networks’ by Wiley & Sons and a presenter for several tutorials at flagship IEEE Communication Society events. Dr. Taha’s other interests include scalability issues in the Internet of Things and modeling challenges in networked cyber physical systems. Over the last 2 years, he has co-founded and organized the workshop on the design, modeling and evaluation of Cyber Physical systems, CyPhy (www.cyphy.org).

Hossam S. Hassanein is a leading authority in the areas of broadband, wireless and mobile networks architecture, protocols, control, and performance evaluation. His record spans more than 400 publications in journals, conferences, and book chapters, in addition to numerous keynotes and plenary talks. Dr. Hassanein’s leadership and editorial activities include his role as the Editor-in-Chief of the IEEE/ACM Transactions on Networking and his previous term as the Editor-in-Chief of IEEE Transactions on Mobile Computing. He sits on the editorial boards of several journals, including the IEEE Journal on Selected Areas in Communications, IEEE Journal on Select Areas in Networking, and IEEE Transactions on Wireless Communications.


talks in flagship venues. Dr. Hassanein has received several recognition and best papers awards at top international conferences. He is also the founder and director of the Telecommunications Research (TR) Lab at Queen’s University School of Computing, with extensive international academic and industrial collaborations. Dr. Hassanein is a senior member of the IEEE and is the former chair of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). Dr. Hassanein is an IEEE Communications Society Distinguished Speaker (Distinguished Lecturer 2008-2010).

Hussein T. Mouftah joined the School of Electrical Engineering and Computer Science of the University of Ottawa in 2002 as a Tier 1 Canada Research Chair Professor, where he became a University Distinguished Professor in 2006. He has been with the ECE Department at Queen’s University (1979-2002), where he was prior to his departure a Full Professor and the Department Associate Head. He has 6 years of industrial experience mainly at Bell Northern Research of Ottawa (now Nortel Networks). He served as Editor-in-Chief of the IEEE Communications Magazine (1995-1997) and IEEE ComSoc Director of Magazines (1998-1999), Chair of the Awards Committee (2002-2003), Director of Education (2006-2007), and Member of the Board of Governors (1997-1999 and 2006-2007). He has been a Distinguished Speaker of the IEEE Communications Society (2000-2007). He is the author or coauthor of 7 books, 53 book chapters, and more than 1100 technical papers, 12 patents and 140 industrial reports. He is the joint holder of 12 Best Paper and/or Outstanding Paper Awards. He has received numerous prestigious awards, such as the 2007 Royal Society of Canada Thomas W. Eadie Medal, the 2007-2008 University of Ottawa Award for Excellence in Research, the 2008 ORION Leadership Award of Merit, the 2006 IEEE Canada McNaughton Gold Medal, the 2006 EIC Julian Smith Medal, the 2004 IEEE ComSoc Edwin Howard Armstrong Achievement Award, the 2004 George S. Glinski Award for Excellence in Research of the U of O Faculty of Engineering, the 1989 Engineering Medal for Research and Development of the Association of Professional Engineers of Ontario, and the Ontario Distinguished Researcher Award of the Ontario Innovation Trust. Dr. Mouftah is a Fellow of the IEEE (1990), the Canadian Academy of Engineering (2003), the Engineering Institute of Canada (2005) and the Royal Society of Canada RSC Academy of Science (2008).