Mitigating Signaling Overhead from Multi-Mode Mobile Terminals

Indra Widjaja and Carl Nuzman
Bell Labs, Alcatel-Lucent
Murray Hill, NJ 07974

***Abstract*—Modern cellular networks may be deployed using multiple radio-access-network technologies with multi-mode mobile terminals capable of selecting among different technologies. Such a network typically forms an overlay-underlay architecture where the underlay uses an older technology (e.g., 3G) while the overlay uses a newer technology (e.g., 4G). It has been observed that excessive signaling message updates can arise due to registration ping-pongs, and Idle-mode Signaling Reduction (ISR) has been introduced as a mechanism to reduce this update load. In this paper, we show that while ISR reduces update load, it also has the effect of increasing paging load. We investigate tradeoff between update and paging loads. Our analysis quantifies a threshold that is used to activate or deactivate ISR for each mobile terminal and results in significant signaling load reduction. We also develop a practical threshold-based algorithm that does not rely on knowledge of the structure of an overlay or terminal mobility.**

I. Introduction

It is well known that mobile data/video traffic has one of the highest growth rates among different Internet services as smartphones are becoming more pervasive [1]. Standards bodies, in particular 3GPP [2], have continuously worked on the evolution of wireless network technologies in response to the rapidly increasing demand for bandwidth at air interface. As can be attested in 3GPP work plan, long before 2G technology was upgraded to 3G, it was already recognized that 4G technology would be needed.

In order to protect their investments, wireless service providers usually operate their networks with multiple radio-access-network (RAN) technologies when upgrades are needed. Such a network is said to form an overlay-underlay architecture. Examples include 2G underlay with 3G overlay or 3G underlay with 4G overlay. While more than two technologies operating concurrently are possible, we restrict to the common two-technology case in this paper. Fig. 1 shows an example depicting a wireless network architecture with two RANs: UMTS Terrestrial Radio Access Network (UTRAN) and Evolved UTRAN (EUTRAN) [3][4]. A mobile terminal (MT) is commonly equipped with multi-mode capability so that it can communicate with either RAN seamlessly.

When registered to a cellular network, an MT can be in either of two modes: *connected* or *idle* [3]. When in connected mode, an MT has a connection to the network and is capable of transmitting and receiving data. On the other hand, an MT in idle mode does not maintain a connection to the network and helps to reduce resources consumed in the network. Another advantage of an MT being in idle mode is that most of its circuitry can be turned off and the resulting battery power consumption is substantially reduced. An MT in idle mode, however, needs to be paged (awakened) by the network when there is an incoming session that needs to be established.

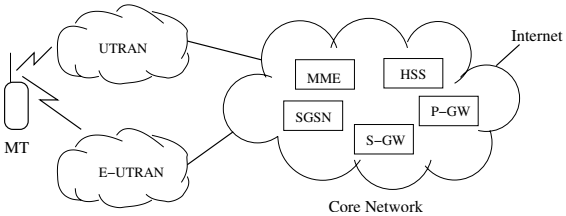

Fig. 1. Wireless network with two RANs (UTRAN and EUTRAN) and a core network (see [3] for details).

A paging area is designed to locate MTs in idle mode. The area consists of a number of base stations, typically on the order of several dozens. When an incoming session arrives, the network can page the recipient MT in idle mode according to the last paging area seen by the network. A paging area is called a tracking area [3] in 4G and a routing area in 3G with data service [5]. An MT that is being paged responds to the network via one of the base stations in the paging area. To ensure that the network knows that a given paging area is up-to-date, each MT needs to inform the network via an *update* when it crosses to a new paging area. In addition, an MT also needs to update the network periodically even if it is stationary. This is necessary for the network to maintain MT registration.

Consider a deployment of an overlay-underlay architecture with two RANs as depicted in Fig. 1. In addition to updating the network when an MT moves to another paging area, the MT also needs to update the network when it moves from one RAN to another so that the MT is registered with the correct RAN. This clearly will have an effect of increasing update rate. Fig. 2 shows an example of representative update rates observed on a network with one RAN (case 1) and two RANs (case 2). As can be seen, the increase in update rate with overlay-underlay architecture can be fairly large. The reason is that deployment of an overlay can have *patchy coverage* caused by the presence of *coverage holes*. Patchy coverage can cause an MT to change registration with underlay and overlay frequently, causing registration "ping-pongs". Update rates have been observed to increase by a factor of 10 or more when an overlay is patchy.

High update rates are undesirable as they can overload network elements, affect network services and reduce MT battery life-times for idle MTs. To combat this problem,

978-0-9836283-0-9

This paper was peer reviewed by subject matter experts for publication in the Proceedings of ITC 2011

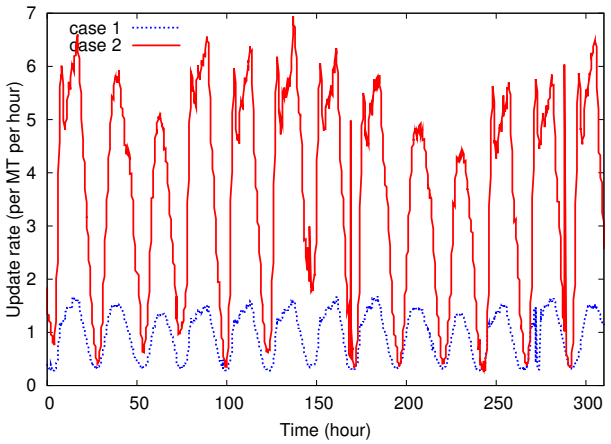

Fig. 2. Update rates in two cases.

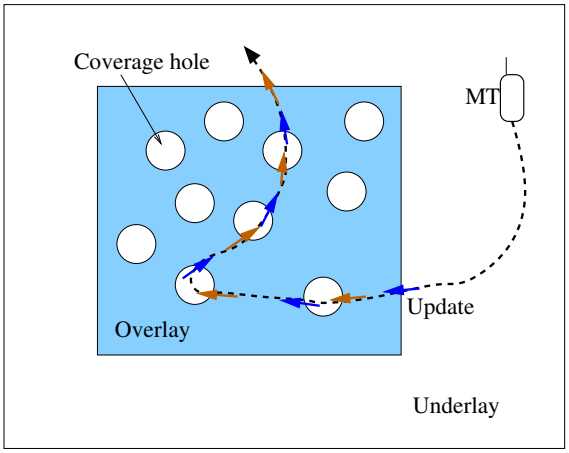

Fig. 3. Overlay-underlay architecture with patchy coverage where registration ping-ponging can occur frequently.

3GPP standards have specified a mechanism, called Idle-mode Signaling Reduction (ISR), that enables an MT to register with *both* RANs simultaneously. We provide a brief overview of ISR in Sec. II. We describe a model for overlay patchiness and quantify how update rate is related to the patchiness. Since ISR requires an MT to register with both RANs simultaneously, paging rate will increase as paging messages are broadcast by both RANs simultaneously. We quantify how paging rate is affected by a dual RAN scenario in Sec. III. Much of previous study on updating and paging focuses only on a single RAN technology without a combination of overlay and underlay [6][7][8]. In an overlay-underlay network, we observe a new phenomenon which we call "ghosting effect" where an MT that is physically located in an underlay can be also registered in an overlay. Such effect tends to further increase paging traffic. We study the tradeoff between paging and update rates in Sec. IV. Given a particular overlay deployment, our main result shows that there exists a threshold λ^* where ISR should be activated for a given MT when its incoming session arrival rate is less than or equal to λ^*. Otherwise, ISR should deactivated. We also provide a stochastic approximation approach to find a good threshold in practice.

II. Update Rate

A. Idle-mode Signaling Reduction (ISR)

ISR is a mechanism to reduce update rate due to frequent MT movement between an overlay and an underlay [3]. Although ISR is currently specified for LTE [9] as an overlay, the principle is also applicable to other technologies. ISR allows an MT to stay registered with two RANs simultaneously. With double registration, updates are avoided when an MT moves between an underlay and an overlay. ISR is activated on a per-MT basis and the network can decide to activate or deactivate ISR on each MT during an update. The underlay is ubiquitous and covers an area larger than the overlay until the overlay becomes a new underlay. We assume that an MT always selects an overlay of newer technology when it is available.

Without ISR (ISR deactivated), an MT only registers with one RAN at any time. This requires the MT to update the network when it moves from an underlay to an overlay and vice versa. As shown in Fig. 3, an update can be originated from an overlay (represented by a blue arrow) or originated from an underlay (represented by a orange arrow). Due to the addition of an overlay, it is clear that overlay-underlay architecture tends to generate higher update rate to a certain degree than a single-RAN architecture. In reality, as can be seen in Fig. 3, the update-rate problem can be more severe when an overlay contains coverage holes. With a single RAN, an MT will be out of coverage when it is in a coverage hole. With dual RANs, an MT encountering a coverage hole will reselect a new RAN by means of an update. We will quantify the effect of coverage holes on update rate. We note that Kalyanasundaram et al.[10] have studied signaling update rate in overlay-underlay architecture. However, their problem setup is completely different from ours. They assume at the outset that the 3GPP ISR mechanism is not used; that is, an MT can only be registered with one technology but not both. The effect of network patchiness is not considered. Also, only update rate but not paging rate is considered. Update rate reduction may arise if an idle-mode MT maintains its registration with an underlay when it moves to an overlay as long as the underlay is available. The MT only changes its registration with an overlay when it is in connected mode. The approach incurs extra delay due to inter-technology handover.

B. Mobility Model

A number of models for the movement of mobile devices are available in the literature, including for example the recent simulation methods of [11] and [12]. We take as a starting point the most analytically tractable approach, which is the well-known fluid-flow model. In this model, we think of MTs as moving randomly over a closed region in such a way that the MT location is uniform, with constant density ρ, and the direction of travel is uniformly distributed over $[0, 2\pi]$. Then the expected rate at which MTs cross a given boundary line is

$$R = \frac{\rho V L}{\pi}, \tag{1}$$

where L is the perimeter of a given region and V is the average MT velocity [13]. The average is over all MTs, which may

have heterogeneous velocities. Since ISR is set on a per-MT basis, it is often more convenient to deal with an individual MT in our setting. As the total number of MTs in the region with area A is ρA, we can also compute the average crossing rate per MT as

$$\hat{R} = \frac{R}{\rho A} = \frac{VL}{\pi A}. \quad (2)$$

C. Analysis of Overlay

We first consider an overlay without coverage holes containing one or more paging areas. Throughout our analysis, we only account for updates originated from an overlay as represented by the blue arrows in Fig. 3. The rates at which overlay-originated and underlay-originated updates occur should be the same under our model. First focus on the signaling update rate due to MTs that move from one paging area to another within an overlay and from an underlay to an overlay. Denote the corresponding update rate per MT by $\hat{R}_o$. Let C_o be the number of equal-sized cells in an overlay and C_a be the number of cells per paging area. There are C_o/C_a equal-sized paging areas in the overlay. If the perimeter of a cell is L_c, then the perimeter of a paging area can be given by $L_a \approx L_c\sqrt{Ca}$. The expression for L_a is exact if the geometry of a cell is a square, and a good approximation for other geometries such as circles or hexagons. From (1), the expected number of updates from MTs entering one paging area is $\rho V L_a/\pi$. Multiplying by the number of paging areas and dividing by the total number of MTs in an overlay ρA_o where A_o is the overlay area, we obtain the average update rate per MT as

$$\hat{R}_o = \left(\frac{C_o}{C_a}\right)\frac{VL_a}{\pi A_o}. \quad (3)$$

To maintain registration to the network, an MT utilizes an update timer with period T so that it can periodically update the network even if it is stationary. A periodic update can be thought off as a keep-alive message and T is typically set to a few hours. The update timer is reset whenever there is an update (triggered by a periodic timer or MT movement). While in overlay, an update timer is used to limit the amount of time between updates for a given MT. It is clear that the rate of timer-triggered updates per MT, $\hat{R}_T$, is bounded above by $1/T$ when the MT is stationary. To further refine the upper bound when an MT is mobile, we assume that an MT stays in an overlay with its dwell time, X, that follows the exponential probability law [14] given by

$$Pr[X > x] = e^{-\mu x},$$

where $\mu = \frac{VL_o}{\pi A_o}$ and L_o is the overlay perimeter.

Suppose that the timer expires T hours after the last update. Then the expected number of timer-triggered periodic updates per MT while in an overlay is

$$\hat{N}_T = \sum_{k=1}^{\infty} Pr\{X > kT\} = \sum_{k=1}^{\infty} e^{-\mu kT} = \frac{1}{e^{\mu T} - 1}.$$

Dividing by the expected dwell time $1/\mu$, we obtain the average rate of timer-triggered updates per MT

$$\hat{R}_T = \frac{\mu}{e^{\mu T} - 1}. \quad (4)$$

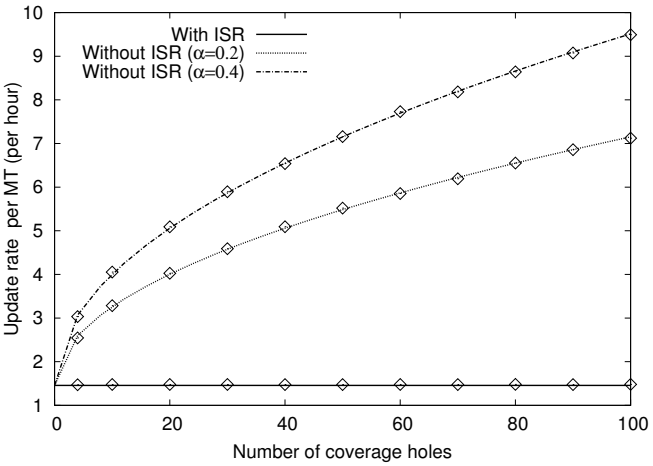

Fig. 4. Update rate per MT vs number of coverage holes, N_h, for different values of α. Corresponding simulation results indicated by "$\diamond$" verify the accuracy of the analysis.

Simulation results show that the impact of different dwell time distributions on $\hat{R}_T$ is small. Moreover, the contribution of $\hat{R}_T$ on the overall update rate is also generally very small.

D. Analysis of Coverage Holes

We now consider the signaling update rate due to MTs leaving coverage holes. We assume that the total area of all holes is a fraction αA_o of the total overlay area, and that there are N_h equal-sized holes in an overlay, each of area $A_h = \alpha A_o/N_h$. Modeled as a circle, each hole has radius $r_h = \sqrt{(\alpha A_o)/(\pi N_h)}$ and perimeter $L_h = 2\sqrt{\alpha\pi A_o/N_h}$. We adopt equal-sized holes in order to limit the number of parameters. Different hole sizes can be incorporated with increasing number of parameters. Our analysis is also applicable to other common geometries.

A coverage hole that intersects the boundary of a paging area may change the effective perimeter of the area used for deriving the update rate. However, this effect is expected to be small since a coverage hole is significantly smaller than a paging area, and thus we ignore this effect. A coverage hole may also delay a periodic timer update when an MT happens to be in a coverage hole when its timer expires. It is expected that this effect is small when α is small. We also ignore this effect in our analysis.

With ISR (ISR activated), updates are not generated by MTs leaving a coverage hole. Without ISR, under the fluid model, updates by MTs leaving one of N_h holes are generated at the rate of $\rho V L_h/\pi$. Multiplying this value by N_h and dividing by ρA_o, we obtain the per-MT update rate due to MT leaving coverage holes as

$$\hat{R}_h = \begin{cases} 2V\sqrt{\frac{\alpha N_h}{\pi A_o}}, & \text{without ISR} \\ 0, & \text{with ISR.} \end{cases} \quad (5)$$

E. Results

Fig. 4 plots the overall average update rate per MT. The rate is due to three components: (1) an MT entering an overlay, (2) periodic timer updates while on overlay, and (3) an MT

leaving coverage holes. In this scenario, the parameter values are $V = 10$ km/hr, $T = 3$ hr, $L_c = 3.5$ km and $C_a = C_o = 100$ (in urban area). Note that when ISR is activated on the MT, the average update rate is flat at around 1.5 updates per hour as the third component is zero. However, when ISR is deactivated, the average update rate increases as a square root of N_h while the total area of coverage holes, αA_o, is fixed. Indeed, a large number of tiny holes can have very detrimental effect on the update rate. Thus, to reduce update rate without ISR in early deployment of base stations of new technology as overlay, it is often better to have concentrated deployments of these base stations so that there are not too many coverage holes. Referring to Fig. 4 again, the update rate without ISR is about 7.2 when $N_h = 100$ and $\alpha = 0.2$. If N_h is reduced to 50, the rate will be lower than 7.2 even if α (total hole area) is increased by up to two times.

To verify our analysis, we simulate a random-trip mobility model [15]. MT movement consists of a sequence of flights in an overlay of a square region. At the beginning of each flight, MT's flight duration is chosen from some distribution and its direction is chosen uniformly over $[0, 2\pi]$. The MT then travels in a straight line in the chosen direction with a given speed, which is also chosen from some distribution. If the MT hits the boundary of the overlay, it continues its flight in the same direction at the opposite side. This transforms the region to a torus and ensures that MT location at the end of each flight is uniformly distributed over the region. The MT may optionally pause for a given duration at the end of a flight. The numbers of times the MT crosses the holes and the overlay are recorded to keep track of the update rate during the entire sequence of flights. The intersections, if any, between a given MT's fight path and a hole can be solved through a quadratic equation. As can be seen from Fig. 4, the simulation results, marked as $\diamond$'s, validate the analytical expression. Extensive experiments also show that the average update rate is insensitive to the distribution of flight duration, pause duration or velocity.

III. Paging Rate

In the preceding section, ISR was shown to reduce update rate significantly when the overlay is patchy. However, since ISR maintains double registration, it also has the potential to increase paging rate. In this section, we quantify the effect of ISR on paging rate.

Paging is triggered when the network receives the first downlink packet for an idle-mode MT. If ISR is not activated, the network sends paging messages to the paging area that the MT was last seen. Note that the paging area may belong to an underlay or overlay depending on which technology the MT is registered with. If ISR is activated, the network sends paging messages to both paging areas of underlay and overlay that the MT was last seen. In either case, the MT will receive a paging message from a base station that the MT is associated with and respond to the network.

A. Ghosting effect for overlay paging with ISR

When an MT with its ISR activated first enters an overlay, it will register with the overlay while simultaneously maintaining its registration with the underlay. The MT deregisters from the overlay if, after its periodic timer expires, it is no longer in the overlay. If the MT is in the overlay when the periodic timer expires, the MT performs an update to remain in the overlay and the timer resets. Thus, the deregistration occurs at a time of the form $t_{\text{reg}} + kT$, where t_{reg} is the registration time, k is an integer, and T is the timer expiration period. In particular, deregistration occurs at the earliest time instant of this form such that the MT is not in the overlay.

Because of the timeout mechanisms, the set of MTs registered with the overlay includes those physically in the overlay region as well as a group of MTs no longer in the overlay region whose timers have not yet expired. We call an MT that is not physically in the overlay but is still registered with the overlay a "ghost MT". The consequent increase in the number of MTs registered with the overlay is called *ghosting*.

Ghosting can be useful; indeed, ISR takes advantage of the ghosting effect to reduce update messages by ghost MTs that wander into overlay coverage holes. On the other hand, the cost incurred by ghost MTs is an increase in the number of MTs registered with the overlay and the associated increase in paging traffic.

Using a density model for MT population, the number of MTs physically in the overlay is proportional to the overlay area, $N_o = \rho A_o$. The increase in the number of MTs due to ghosting may be thought of as increasing the effective area of the overlay, from the point of view of paging. In other words, if N_s MTs are registered with the overlay, we say the effective area is $A_s = N_s/\rho$. We are interested in the ghosting ratio $A_s/A_o = N_s/N_o$, which is always greater than or equal to one for a convex region.

The ghosting ratio in a given scenario depends primarily on the MT mobility patterns and the timer expiration period. If the MTs are moving in and out of the overlay at high speed, without returning, and the timer period is long, then the number of ghost MTs will be large. On the other hand, if the MTs move very little during the timer period, then there will only be very few ghost MTs. To arrive at a tractable model, we assume that a given MT moves at a constant speed v and in a straight line while crossing an overlay. Different MTs may have different velocities, according to a specified distribution.

Suppose that the trajectory of a given MT takes it through an overlay at a point where the overlay has width w. Assuming that the overlay consists of a single tracking area, then the distance that the MT will travel before deregistering will be of the form kvT, for k an integer. In particular, deregistration will occur at the smallest value of k such that $kvT > w$, namely $k = \lceil w/vT \rceil$. Thus ghosting increases the effective cross-section of the overlay at that point from w to $f(w, vT) = \lceil w/vT \rceil \, vT$. When the distance vT traveled during the timer period is much smaller than the overlay width w, the effect of ghosting is negligible. On the other hand, if vT is much larger than w, there is a large ghosting effect.

Although the expression for the effective overlay cross-section is simple for a given width w and speed v, one needs to take into account a variety of widths and speeds in order

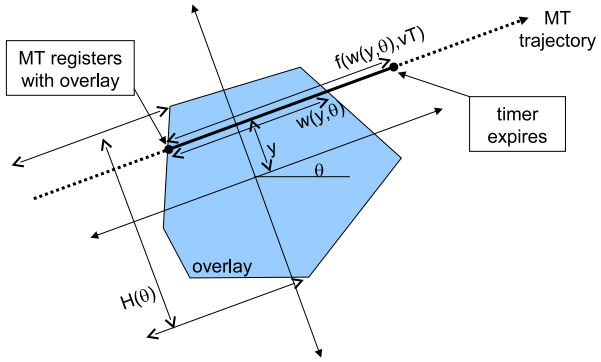

Fig. 5. Ghosting of an MT and associated notation. The MT is traveling at direction θ and at minimum distance y from the origin. The MT registers with the overlay upon entering it, and is deregistered at the first timer expiration that occurs outside the overlay.

to estimate the total effect. Consider Fig. 5 where we define a class of MTs to be a set of MTs having a given speed v and a given direction of travel, θ. Denote a central point of the overlay as an origin. The trajectory of a given MT traveling in direction θ can be characterized by the perpendicular distance y from the trajectory to the origin. We may denote the width of the overlay under this trajectory as $w(y,\theta)$. The distance traveled by the MT before deregistering is then $f(w(y,\theta), vT)$. To calculate the area of the region in which MTs moving in direction θ at speed v are registered with the overlay, we integrate the effective cross-section over y to obtain

$$A_s(\theta, v) = \int f\left(w(y,\theta), vT\right)\, dy.$$

Next, given a distribution function $p_\theta(\theta)$ for the direction of travel, for $0 \le \theta \le 2\pi$, we can integrate over direction to obtain the ghosting area as a function of velocity,

$$A_s(v) = \int A_s(\theta, v) p_\theta(\theta)\, d\theta.$$

Finally, the effective area is obtained by integrating over a velocity distribution, $p_v(v)$, so that

$$A_s = \int A_s(v) p_v(v)\, dv.$$

In Fig. 6, the solid curve shows the effective area $A_s(v)$ as a function of velocity, for a circular overlay of diameter 15 km, when the timeout period is 3 hours. The plot is normalized to the actual overlay area, $A_o = \pi(15/2)^2 \text{km}^2$. Overlays of different shapes lead to slightly different area functions.

In order to simplify the analysis and reduce the number of parameters, we next develop approximations using upper and lower bounds on A_s. The following bounds are easily verified:

$$f(0, vT) = 0 \quad (6)$$
$$f(w, vT) \ge vT, \quad \text{for } w > 0 \quad (7)$$
$$f(w, vT) \ge w \quad (8)$$
$$f(w, vT) \le w + vT \quad (9)$$

Using the upper bound (9), we may write

$$A_s(\theta, v) \le \int_{w>0} w(y,\theta)\, dy + vT \int_{w>0} dy = A_o + vT\, H(\theta)$$

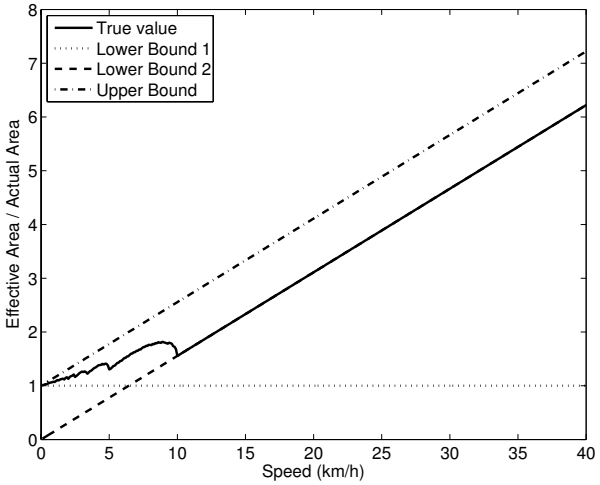

Fig. 6. Relative increase in effective area due to ghosting, $A_s(v)/A_o$, as a function of MT speed. Scenario depicted is a circular overlay with diameter 15 km, with periodic timer set at $T = 3$ hours.

where

$$H(\theta) = \int_{w>0} dy$$

is the "height" of the overlay in the direction perpendicular to angle (see Fig. 5). Averaging over angle and speed, we obtain the upper bound

$$A_s \le A_o + VT\bar{H}$$

where V is average velocity and $\bar{H}$ is average overlay height. Similarly, by applying the lower bounds (7) and (8), we obtain

$$A_s \ge \max\left(A_o, VT\bar{H}\right).$$

Thus the increase in MT registration due to ghosting satisfies the following bounds:

$$\max\left(1, \frac{VT}{A_o/\bar{H}}\right) \le \frac{N_s}{N_o} \le 1 + \frac{VT}{A_o/\bar{H}}, \quad (10)$$

which only require knowledge of the ratio between the average linear distance traveled by an MT during a timeout period, VT, and the average overlay width, $A_o/\bar{H}$. These upper and lower bounds are depicted in Fig. 6.

B. Ghosting When Coverage Holes are Present

The preceding analysis has been for convex overlay areas, for which an MT moving in a straight trajectory enters and exits the overlay exactly once. For an overlay with coverage holes, the situation is more complicated. However, the upper and lower bounds can be easily modified to handle the case with coverage holes.

Consider an overlay region with coverage holes, and assume that the region is convex when the coverage holes are filled in. Call this convex region the filled region. Define the function $w(y,\theta)$ to be the width of a given trajectory passing through the filled region, as in the previous section. Define $\tilde{w}(y,\theta) \le w(y,\theta)$ to be the total length of the intersection of the trajectory with the true region; that is, $\tilde{w}$ is obtained by subtracting from w the intervals where the trajectory passes through coverage holes.

For a given trajectory, the effective cross-section of the region is the total length of the sub-intervals of the trajectory in which an MT would be registered with the overlay. For a convex region, the effective cross-section function f only depends on y and θ through the width $w(y,\theta)$. For a region with coverage holes, the effective cross-section function is much more complicated, and will be denoted by $\tilde{f}(y,\theta,vT)$. However, it is easy to see that $\tilde{f}$ satisfies bounds similar to (6)-(9):

$$\tilde{f}(y,\theta,vT) = 0 \quad \text{for } \tilde{w}(y,\theta)=0 \tag{11}$$
$$\tilde{f}(y,\theta,vT) \geq vT, \quad \text{for } \tilde{w}(y,\theta)>0 \tag{12}$$
$$\tilde{f}(y,\theta,vT) \geq \tilde{w}(y,\theta) \tag{13}$$
$$\tilde{f}(y,\theta,vT) \leq w(y,\theta)+vT \tag{14}$$

Eqs. (11)-(14) reflect that: (a) there is no registration on a trajectory that does not intersect the overlay, (b) for a trajectory that intersects the overlay, the MT is registered for at least time T, (c) the MT is registered while at least it is physically in the overlay, and (d) the MT is registered for at most time T after it physically leaves the filled overlay. Note that integrating $\tilde{w}(y,\theta)$ over y yields $(1-\alpha)A_0$, which is the area of the overlay with coverage holes removed. After averaging the bounds on $\tilde{f}$ with respect to y, θ, and v, as in the previous section, we obtain the following bounds:

$$\max\left(1-\alpha, \frac{VT}{A_o/\bar{H}}\right) \leq \frac{N_s}{N_o} \leq 1+\frac{VT}{A_o/\bar{H}}. \tag{15}$$

for the fraction of MTs in the filled overlay region of area A_0 that are registered with an overlay containing coverage holes of total area αA_0.

IV. ISR or No ISR

A. Tradeoff Between Paging and Updating

We first consider tradeoff between the update rate and paging rate with and without ISR. Although updating and paging contribute most significantly in the overall network signaling load, there are other events that also induce load on the network. Since these events induce the same amount of load independent of ISR setting, they can be ignored in our tradeoff study. To simplify exposition, we also only account the signaling load handled by a network element in the overlay. The load handled by the underlay can be similarly derived.

As updating and paging events typically generate different numbers of messages, we use *message* rate rather than *event* rate. Let $\tilde{R}^U$ and $\tilde{R}^P$ denote the average update message rate per MT and the average paging message rate per MT, respectively. Let N^U and N^P denote the number of messages generated per update event and the number of messages generated per paging event, respectively. From (3), (4), (5) and (15), we can write

$$\tilde{R}^U = \begin{cases} N^U(\hat{R}_o+\hat{R}_T+\hat{R}_h), & \text{without ISR} \\ N^U(\hat{R}_o+\hat{R}_T), & \text{with ISR} \end{cases} \tag{16}$$

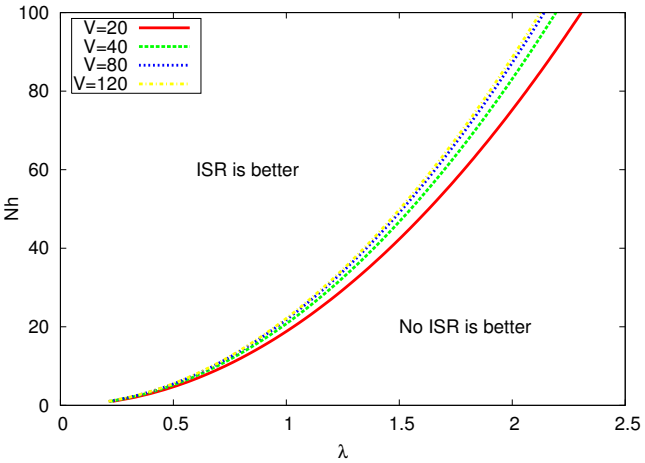

Fig. 7. (N_h,λ) for which ISR and no-ISR induce the same load.

and

$$\tilde{R}^P = \begin{cases} N^P\lambda(1-\alpha), & \text{without ISR} \\ N^P\lambda\max\{1-\alpha, \frac{VT}{A_o/H}\}, & \text{with ISR (LB)} \\ N^P\lambda\,(1+\frac{VT}{A_o/H}), & \text{with ISR (UB)} \end{cases} \tag{17}$$

where λ is the incoming (network-originating) mean session arrival rate per MT (sessions per hour). It is clear from (16) that ISR has the advantage at reducing update rate when there are coverage holes. On the other hand, (17) shows that ISR is subjected to additional paging due to ghost MTs in coverage holes and outside an overlay.

Let $g(V,N_h,\lambda) = \tilde{R}^{U,I} + \tilde{R}^{P,I} - \tilde{R}^{U,\bar{I}} - \tilde{R}^{P,\bar{I}}$, where superscript I represents the case with ISR while $\bar{I}$ is for the case without ISR in (16) and (17). It follows that ISR is beneficial when $g<0$ and harmful when $g>0$. The break-even point between ISR and no-ISR occurs at $\lambda^*(V,N_h) = \{\lambda : g(V,N_h,\lambda)=0\}$.

Fig. 7 plots several curves for different MT velocities where break-even-point, $g=0$, occurs for different pair of values (N_h,λ). It is assumed that $\alpha=0.2$, $N^U=19$, $N^P=12$ and other parameter values are the same as those used in the scenario depicted in Fig. 4. The figure uses the lower bound for paging with ISR in (17) as extensive verification shows that the lower bound is significantly more accurate than the upper bound (for example, see Fig. 6). For a given MT velocity, the area below the curve defines a region where deactivating ISR on the MT is better, while the area above the curve is where activating ISR is better. Observe also that λ at a break-even point is very sensitive to overlay patchiness, N_h.

In Fig. 8, we plot the overall average message rate per MT (messages per hour), $\tilde{R}(\lambda) = \tilde{R}^U(\lambda) + \tilde{R}^P(\lambda)$, as a function of λ. There are six curves corresponding to three different velocities, with and without ISR. N_h is set to 60 and other parameter values are the same as those in Fig. 7. For a given MT velocity, observe that with ISR activated, the message rate depends strongly on λ while with ISR deactivated, the message rate is insensitive to λ. These two curves with and without ISR intersect at a break-even point (indicated by a bullet) when $\lambda=\lambda^*$. As can be seen from the figure, the

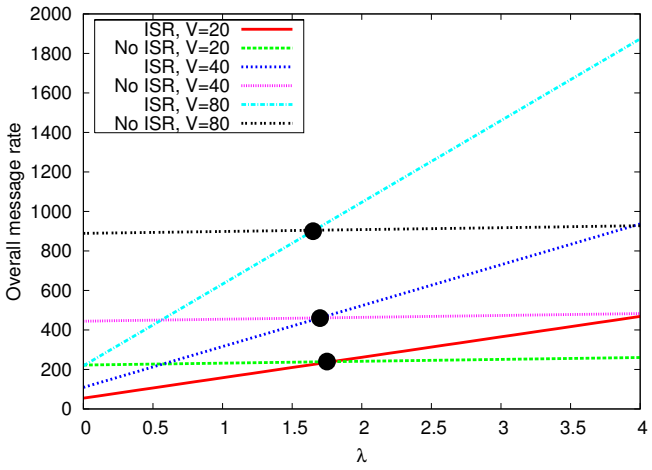

Fig. 8. Overall message rate vs. λ

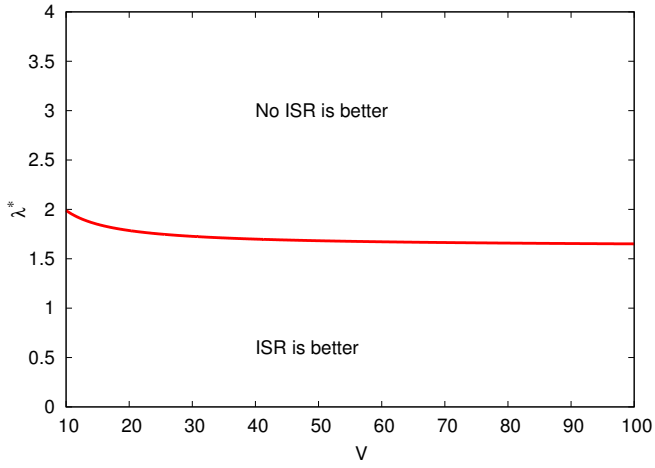

Fig. 9. λ^* vs V at break-event point.

values of λ^* are rather insensitive to MT velocity although the message rate depends strongly on MT velocity.

Fig. 9 plots λ^* as a function of MT velocity V for a given deployment using parameter values as in Fig. 8. To minimize overall message rate, the network should deactivate ISR for a given MT when $\lambda > \lambda^*$ and activate it when $\lambda \leq \lambda^*$. Notice that except at very low velocities, λ^* is insensitive to MT velocity. This observation suggests that a global threshold value that is *independent* of MT velocity, $\tilde{\lambda}$, can be used for deciding whether ISR is to be activated or not. This allows us to avoid measuring MT velocity, which is non-trivial in practice. Since the problem of deciding ISR/no-ISR becomes crucial at high message rate which occurs at high velocity (see Fig. 8), the global threshold should be tuned to conditions at high velocities (e.g., above 20 km/hr).

Deciding whether or not to activate ISR has the largest impact when session arrival rate λ is very low or very high. Therefore, using a threshold $\tilde{\lambda}$ to activate or deactivate ISR gives the largest benefit when the distribution of session arrival rates among MTs has high variability. Such distributions, with a relatively small percentage of chatty users (high λ values) and high percentage of quiet users (low λ values), are common

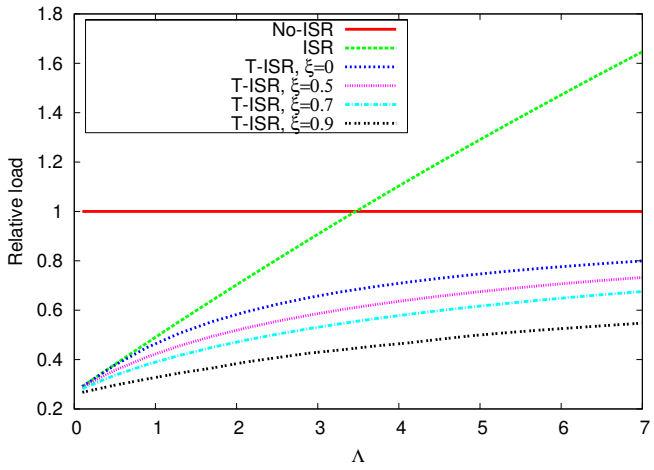

Fig. 10. Comparison of no-ISR, ISR and T-ISR.

in practice.

To evaluate our threshold-based ISR approach (T-ISR), we perform a Monte-Carlo simulation experiment. We assume that there are $K = 500,000$ MTs and denote them by $U_k, k = 1, \cdots, K$. Let Λ denote the mean session arrival rate over the entire MT population. We can model the chattiness of different MTs from measurement by assigning a random session arrival rate λ_k for each U_k from a Generalized Pareto cumulative distribution function defined by

$$F(x;\xi,\sigma) = \begin{cases} 1-(1+\xi x/\sigma)^{-1/\xi}, & \xi \neq 0, \\ 1-\exp(-x/\sigma), & \xi = 0, \end{cases}$$

for $x \geq 0$ and $\sigma = \Lambda(1-\xi)$. Based on all session arrival rates, we compute the overall message load from all MTs, $M = \sum_{k=1}^{K} \tilde{R}(\lambda_k)$. We evaluate three cases: (1) ISR (ISR is activated for each MT), (2) no-ISR (ISR is deactivated for each MT), and (3) T- ISR (ISR of the MT is activated if its $\lambda_k \leq \tilde{\lambda}$ and is deactivated otherwise).

For comparison purpose, we normalize the message load to the no-ISR case such that the message load for the no-ISR case is always 1. Fig. 10 compares the three cases with $\tilde{\lambda}$ set to 1.7 (see Fig. 9). We assume $V = 20$ km/hr and other parameter values are as before. As can be seen, ISR and T-ISR reduce the load by 60% or so relative to no-ISR when MTs are quiet (Λ is very low). As Λ increases, the advantage of ISR diminishes and eventually ISR becomes worse than no-ISR. On the other hand, T-ISR always outperforms the other two cases. Note that T-ISR becomes more advantageous as the shape parameter ξ increases. The advantage of T-ISR also becomes more pronounced with increased MT velocity or more patchy overlay.

B. Practical Setting of ISR

In principle, the preceding analysis allows us to compute an optimal value for the global threshold $\tilde{\lambda}$ that minimizes signaling load. In practice, such an open-loop approach is unlikely to work well because model parameters such as number of holes and hole area are not available.

Instead, a closed-loop approach can be used to iteratively optimize the threshold.The problem fits into the framework

of a one-dimensional stochastic approximation: we wish to minimize an unknown function $M(\tilde{\lambda})$ of a control variable $\tilde{\lambda}$, and we have access to noisy observations $\hat{M}(\tilde{\lambda}, t, \tau)$. In our case, $M(\tilde{\lambda})$ is the overall message load as a function of the global threshold $\tilde{\lambda}$, and $\hat{M}(\tilde{\lambda}, t, \tau)$ is an empirical estimate of the message load obtained by taking the (random) number of messages that arrive in a given time interval $(t, t+\tau)$, and dividing by τ. A number of different stochastic approximation algorithms could be used. A canonical example is the Kiefer-Wolfowitz algorithm which converges to the minimum under certain technical conditions [16].

Because our problem aims at optimizing a "live" communications system, we are more concerned with controlling the impact of our steps, and with robustly tracking changes in system state, than with convergence to a fixed ideal. For these reasons, in our experiments we chose to use an algorithm with fixed step size, on a transformed control variable. Let $y = F(\tilde{\lambda})$ be the fraction of MTs having arrival rate lower than $\tilde{\lambda}$, and let $q(y) = F^{-1}(y)$ be the inverse transformation. We seek an optimal parameter $y^* \in [0, 1]$ that minimizes $M(q(y))$. We first set $y_0 = 0$, and then iteratively evaluate

$$y_{n+1} = y_n + \beta \ \mathrm{sgn}(\hat{M}(q(y_n), t_n, \tau) - \hat{M}(q(y_n + \delta), t_n + \tau, \tau))$$

where δ is perturbation step size and β is increment step size. We also limit y_{n+1} to the interval $[0, 1]$. We estimate the message load during adjacent time periods of length τ, turning on ISR for an additional fraction of MTs, δ, during the second time period. Depending on which of the two periods experiences a higher load, we update our threshold to increase or decrease the fraction of MTs with ISR by a fixed fraction β.

Fig. 11 illustrates numerical results of using the stochastic approximation algorithm in three scenarios with aggregate arrival rates $\Lambda = \{1, 3, 7\}$ calls per MT per hour. The system consisted of 10,000 MTs with individual arrival rates $\{\lambda_k\}$ drawn from a Generalized Pareto distribution with shape parameter $\xi = 0.7$, and all other system parameters as used for Fig. 10. The algorithm used parameters $\delta = 0.1$, $\beta = 0.05$, with an observation time of $\tau = 30$ minutes. With these settings, the algorithm converges to and remains close to nearly optimal settings within about 20 iterations. Signaling load per MT is reduced significantly relative to the initial state $y_0 = 0$ in which no MTs use ISR.

Taking into account of the diurnal load variations in Fig. 2, it may make most sense to optimize the threshold $\tilde{\lambda} = q(y)$ for busy-period conditions. After initial convergence, periodic updates should be sufficient since key parameters such as network patchiness and aggregate MT behavior change slowly. For example, an operator could choose to run a single iteration of the stochastic approximation algorithm once each day, during the busy period.

V. Conclusion

The proliferation of multi-mode MTs and deployment of overlay-underlay wireless network architecture has the potential of stressing network signaling load and creating service disruption. 3GPP has developed a mechanism, called ISR, to allow an MT to be registered in both overlay and underlay to reduce update load at the cost of increasing paging load. We have analyzed the tradeoff between updating and paging with and without ISR and quantified a single threshold value that can be used to activate or deactivate ISR on each MT to minimize signaling load. To deal with more realistic scenarios, we have further developed a stochastic approximation algorithm to activate or deactivate ISR without relying on knowledge of overlay deployment or terminal mobility.

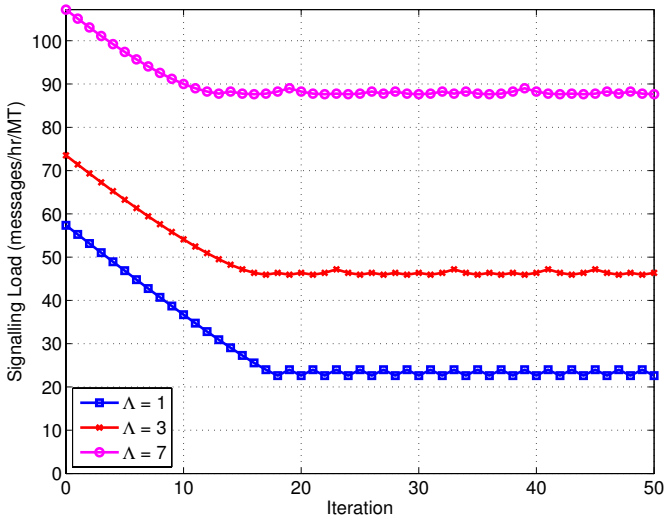

Fig. 11. Progress of the stochastic approximation algorithm, for three values of session arrival rate

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update," White paper, Feb. 2011.

[2] http:/www.3gpp.org.

[3] 3GPP TS 23.401, "GPRS Enhancement for E-UTRAN Access", Dec. 2010.

[4] "Flat IP Architectures in Mobile Networks: From 3G to LTE," Heavy Reading, Apr. 2008.

[5] 3GPP TS-23-060, "GPRS Service Description," Dec. 2008.

[6] A. Bar-Noy, I. Kessler and M. Sidi, "To update or not to Update", IEEE INFOCOM, 1994.

[7] I. Akyildiz, J. Ho and Y. Liu, "Movement-Based Location Update and Selective Paging for PCS Networks," IEEE/ACM Transactions on Networking, pp. 629-638, Aug. 1996.

[8] V. Wong and V. Leung, "Location Management for Next Generation Personal Communication Networks," IEEE Network Magazine, pp. 18-24, Sep. 2002.

[9] H. Ekstrom et al., "Technical Solution for the 3G Long Term Evolution," IEEE Communications Magazine, pp. 38-45, Mar. 2006.

[10] S. Kalyanasundaram et al., "Signaling Reduction in Idle Mode for Inter-Technology Mobility," IEEE VTC 2007, 2007.

[11] K. Lee et al., "SLAW: A Mobility Model for Human Walks," INFOCOM 2009, Rio, Brazil, May 2009.

[12] A. Mei and J. Stefa, "A Simple Model to Generate Small Mobile World," INFOCOM 2009, Rio, Brazil, May 2009.

[13] D. Lam, D.C. Cox and J. Widom, "Teletraffic Modeling for Personal Communications Services," IEEE Communications Magazine, pp. 79-87, Feb. 1997.

[14] R.A. Guerin, "Channel Occupancy Time Distribution in a Cellular Radio System," IEEE Trans. Vehicular Technology, vol. 35, no. 3, pp. 89-99, Aug. 1997.

[15] J. Le Boudec and M. Vojnović, "Perfect Simulation and Stationarity of a Class of Mobility Models," INFOCOM 2005, Apr. 2005.

[16] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," Annals of Mathematical Statistics 23, pp. 462-466, Sep. 1952.