

Multi-service Load Balancing in a Heterogeneous Network with Vertical Handover

Jie Xu, Yuming Jiang, Andrew Perkis, and Elissar Khloussy

Abstract—In this paper we investigate multi-service load balancing mechanisms in an overlay heterogeneous WiMAX/WLAN network through vertical handover. Considering the service characteristics of the overlay heterogeneous network together with the service requirements of different applications, all streaming applications are served in WiMAX while elastic applications are distributed to WiMAX and WLAN. Two load balancing mechanisms are compared which switch the elastic application with maximum (MAX) and minimum (MIN) remaining size respectively to WLAN. Simulation results indicate that MIN outperforms MAX at the cost of significantly increased number of load balancing actions. Furthermore, it is discovered that both load balancing granularity and proper integration of streaming and elastic applications in WiMAX determine the whole system performance.

Index Terms—vertical handover, wireless heterogeneous networks, load balancing, multi-service

I. INTRODUCTION

After decades of research, it is commonly believed that future wireless network will employ multiple techniques. Especially, for the access part, multiple radio access technologies (RATs) will coexist in terms of both space and time. For example, nowadays there are already lots of WLAN networks which are also covered by other 3G mobile networks at the same time.

The coexistence of heterogeneous networks brings up both challenges and opportunities for providing better wireless service [1]. On the one hand, since wireless communications are intrinsically limited by interference, activities of different networks could interfere with each other and may result in severe service degradation. On the other hand, multiple overlay heterogeneous networks can provide more robust communication guarantee if they could cooperate instead of competition. Therefore, how to integrate coexisted multiple networks is of fundamental importance to the success of future wireless networks.

To take advantage of multiple heterogeneous wireless networks, vertical handover [1] has been proposed as a means for enhancing end users' service quality. Traditionally, due to its operation difficulty and introduced time delay, vertical handover is used as a reactive measure to prevent severe service degradation. Normally vertical handover is only triggered

when the served mobile user are about to move out of the coverage range of current serving network. To this end, various vertical handover mechanisms have been proposed to improve the performance of handover user [2][3].

Recently, proper vertical handover are also used as a proactive means to improve the system performance. In [4][5], vertical handover is adopted as a tool for joint resource management in heterogeneous networks. The objective of vertical handover has been extended to include the whole system performance instead of the performance of handover user. In particular, the main idea is to distribute traffic load among heterogeneous networks in a balanced manner by designing vertical handover protocols. However, only streaming users are considered in these studies although the current wireless network normally serve multi-service applications. In [6], both streaming and elastic applications are considered. The authors preferably distribute streaming applications to cellular network because of its larger coverage and finer QoS guarantee, and the remaining capacities in cellular/WLAN networks are utilized for serving elastic applications. While the scheme in [6] performs well compared to random dispatch, there are still chances that streaming applications are distributed to WLAN and vertical handover is triggered whenever there are enough free capacities in the cellular network.

In this study, we consider system performance of an overlay heterogeneous wireless network where elastic applications share network capacity with prioritized streaming applications. Specifically, we consider the WiMAX/WLAN heterogeneous network and assume all the traffic firstly arrives to the WiMAX network. Streaming applications are given strict preemptive priority over elastic applications in WiMAX. Then according to the comparison result of expected finish time in WiMAX and WLAN, vertical handover of certain elastic applications on their arrivals or during their service to WLAN is conducted. We compare the performance of two different handover mechanisms which selected the file with maximum and minimum remaining size for handover respectively. The results indicate that selection of files with minimum remaining size outperforms the other mechanism at the cost of significant increased number of handovers. Furthermore, based on analysis of simulation results, we conclude that both the load balancing granularity and integration of elastic and streaming applications in WiMAX determine the performance of the whole system.

The remaining of this paper is arranged as follows. The system model is described in the next section. In section II, due to the complexity of exact analysis, theoretical approximations are given. In Section IV, we describe the simulation results,

Jie Xu, Yuming Jiang and Andrew Perkis are with the Centre for Quantifiable Quality of Service in Communication Systems, Norwegian University of Science and Technology in Trondheim.

Elissar Khloussy is with Department of Telematics, Norwegian University of Science and Technology in Trondheim.

Center for Quantifiable Quality of Service in Communication Systems, Center of Excellence, is appointed by The Research Council of Norway, and funded by the Research Council, NTNU and Uninett.

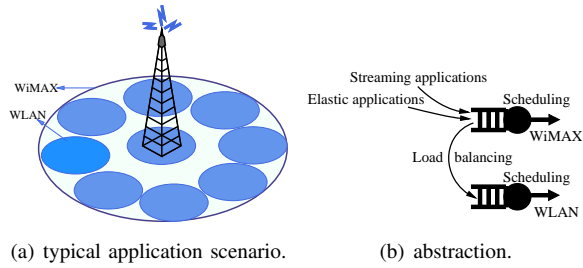


Fig. 1. System model.

followed by discussions in Section V. Finally, we conclude the paper in Section VI.

II. SYSTEM MODEL

A. WiMAX/WLAN Heterogeneous Network

The network scenario of this study is shown in Fig. 1(a). Within the coverage area of the WiMAX network, there are multiple WLAN networks as well. WiMAX is infrastructure-based and can provide guaranteed service with centralized control. In addition, as one option for the future mobile networks, WiMAX has a large coverage area but limited rate to each end user. On the other hand, WLAN is contention based and incapable of providing service guarantee. However, since WLAN uses free frequency it can provide higher data rate to end users if the network is not congested. Therefore, WiMAX and WLAN networks are complementary in terms of service characteristics. It is beneficial to design the overlay heterogeneous network in a cooperative way.

One possible way of cooperation is to distribute different types of applications to specific network by taking account of both the service requirement of applications and the network service characteristic. In this study, the cooperation scheme is shown in Fig. 1(b). Specially, all the streaming applications are served in WiMAX due to their stringent service requirements. Then the remaining service capacity of WiMAX and the total service capacity of WLAN are devoted to elastic applications. Furthermore, to prevent the disturbance of elastic applications, streaming applications are served with higher preemptive priority. Namely the streaming applications arrives and leaves without any disturbance of elastic applications.

For streaming applications, the system can be modeled as an $G/G/K$ loss-queue system. Specifically, if we assume the streaming applications arrive according to Poisson process and the duration follows minus exponential distribution, then the system can be seen as $M/M/K$ Erlang system for which lots results have been obtained. Suppose the capacity of WiMAX is C_{wimax} , then the number of channels K can be calculated as $\lfloor C_{wimax}/B_s \rfloor$ where B_s is the bandwidth requirement of each streaming application. In this study, when there are already K streaming applications in the system, the new streaming arrivals will be simply rejected and discarded.

For elastic applications, both the remaining capacity of WiMAX and the total capacity of WLAN can be utilized. Due to the dynamic state of streaming applications, the remaining capacity of WiMAX varies. For WLAN, we assume the total capacity is fixed to simplify the analysis. Therefore, there are

actually two servers for elastic applications. In each server, residing elastic applications share the capacity in processor-sharing (PS) manner. In particular, no requirements on the maximum or minimum bandwidth for each elastic application is specified.

B. Load Balancing Mechanisms

Real-time load balancing is performed on application arrivals and departures as the network state only changes on these occasions. Once the network state changes, whether load balancing action should be conducted is checked in order to improve the performance of elastic applications. According to the application arrival assumption, in this study we perform unidirectional handover check only from WiMAX to WLAN.

Two types of load balancing actions could be triggered depending on the system states. One kind of action is vertical handover which switches elastic applications that have already been served partly by WiMAX to WLAN. To proceed handover, both criteria for handover check and handover candidate selection need to be clearly defined. In fact, lots of efforts have been devoted to propose efficient algorithms since they actually determine the handover performance [3][2]. We conduct handover check based on the expected finish time of elastic applications. Since existing elastic applications share the service in PS way, the expected finish time can be linearly represented by service rate. Then these mechanisms actually belong to the bandwidth-based handover decision mechanisms. The service rates in two networks under current condition are compared and a handover decision is made if the service rate could be increased after handover. Handover candidate is selected from all the elastic applications in WiMAX including the newly arrival. Specifically, we select handover application based on remaining size. Two mechanisms which select the file with maximum and minimum remaining size respectively are tested. Later we refer the two mechanisms as MAX and MIN for convenience. In real applications the remaining size is difficult to get and may introduce lots of complexity. However, since we determine remaining size on flow level, the complexity introduced can be seen as affordable.

The other kind of action is dispatching which switches the elastic application on its arrival to WLAN. Because we assume all traffic initially arrives to WiMAX, dispatching can be seen as one special case of vertical handover. However, it should be noted that the actual cost for dispatching is much lighter compared with vertical handover since the application has not been served yet.

C. Performance Metrics

For the aforementioned system model, we are only interested in the performance of elastic applications since the performance of streaming applications does not change with the adopted load balancing mechanism. Specifically, we consider three common performance metrics for elastic applications as follows.

- Average sojourn time: The sojourn time defines the time interval of elastic application from its arrival to its departure. As users are very sensitive to duration of

elastic application, this metric is highly related to the user experience of service quality.

- Time-average throughput: The time-average throughput defines the ratio of total service amount to total service time. This metric can be seen as an indicator of system performance in terms of service capability.
- Call-average throughput: The call-average throughput defines the mean of individual ratios of service size to its service time. This metric integrates the influence of service size on users' expectation of service time. Therefore, it is believed to be the best metric representing the users' quality of experience (QoE) [7].

In addition, another important metric for load balancing mechanisms is the number of load balancing actions. Although no specific cost is considered in this paper, it is still beneficial to compare the number of load balancing actions as normally costs of these actions do not vary with different applications. In addition, due to different costs of vertical handover and dispatch, we record both of them in simulations respectively.

III. THEORETICAL APPROXIMATION

It is usually fairly complicated to exactly analyze system performance when integrated services are involved [8][9][10]. Therefore, in this section, we provide a simple approximation for theoretical analysis of our system model. The simplified analysis results provide a reference for further comparisons with our simulation results.

First, the two servers are approximated as one server with capacity $C = C_{wimax} + C_{wlan}$. With this approximation, the system becomes the traditional integrated service system where elastic applications share the server capacity with prioritized streaming applications. However, even for this simplified system, the calculation of stationary results is still very difficult since no closed formulation can be derived [7].

However, approximate results can be calculated with either of the two quasi-stationary assumptions. One quasi-stationary condition assumes elastic applications evolve much faster than streaming applications. This assumption is reasonable since the streaming applications usually last longer than elastic applications. However, theoretical derivation based on this assumption can only be obtained for rather light elastic traffic since it requires uniform stationarity [8].

In our analysis, we take the other quasi-stationary condition which assumes streaming applications evolve much faster than elastic applications. While this assumption is not quite reasonable, it can provide upper performance bounds for elastic applications [8]. In this case, the service devoted to elastic application can be approximated as the total capacity minus the average aggregated service rates of streaming applications.

Based on the former simplification, we could get approximate results for elastic applications. Specifically, since elastic applications share the capacity with processor-sharing policy, insensitivity of PS policy to service distribution can be applied. The average throughput can be expressed as $C_e * (1 - \rho_e)$, and the mean sojourn time can be expressed as

$$E[T] = \frac{E[x]}{C_e * (1 - \rho_e)} \quad (1)$$

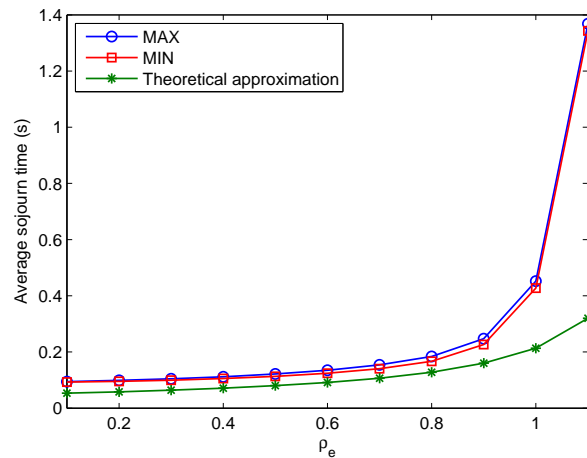


Fig. 2. Average sojourn time with minus exponential distributed file size.

where ρ_s and ρ_e represent the load of streaming and elastic applications respectively with respect to the capacity of WiMAX. In addition, $C_e = C(1 - (1 - P_b) * \rho_s)$, $E[x]$ denotes the average size of elastic applications and P_b is the blocking probability of streaming applications.

IV. NUMERICAL RESULTS

To evaluate the performance of the two handover mechanisms, we have conducted extensive simulations with varying parameters. Each simulation result is obtained based on the average of 30 runs with different seeds. In each run, we simulate 10^6 files and remove the initial stage of the first 5×10^3 files. The results of 30 runs are checked with Skewness and Kurtosis which ensure these runs follow a normal distribution. This guarantees the validation of our simulation results.

TABLE I
SIMULATION PARAMETERS

Parameters	Meaning	Values
C_{wimax}	WiMAX capacity	1000 kbps
C_{wlan}	WLAN capacity	600 kbps
ρ_s	Load of streaming applications	0.3
ρ_e	Load of elastic applications	0.1-1.1
$1/\mu_s$	Average length of streaming applications	140 s
$1/\mu_e$	Average size of elastic applications	64 kbits

The values of simulation parameters are listed in Tab. I. Specifically, we calculate ρ_e based on the capacity of WiMAX.

A. Exponential Distribution

First we assume the size of elastic applications follows exponential distribution, which has been assumed and analyzed extensively for traffic modeling.

In Fig. 2, the average sojourn time of finished files is shown. It can be seen that for exponential distributed files, the average sojourn time with MAX is slightly longer than that with MIN. However, compared with the theoretical approximation, the performance of MAX and MIN are worse. In particular, when the traffic load is heavy, the average sojourn time of finished files with MAX and MIN increases very rapidly.

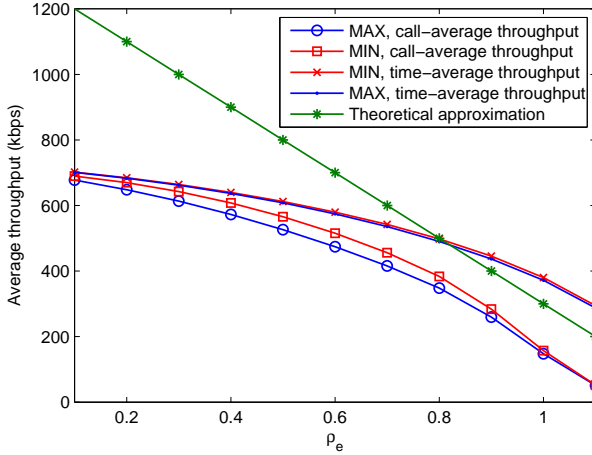


Fig. 3. Average throughput with minus exponential distributed file size.

In Fig. 3, the average throughput performance is shown. It can be seen that the throughput performance of MIN is better than MAX as well. The time-average and call-average throughputs are similar for low-load elastic traffic and the gap increases as the load of elastic applications becomes heavier. The main reason for the fast decay of call-average throughput of heavy-loaded elastic applications is the fast increase of average sojourn time as shown in Fig. 2. However, the average-time throughput is not heavily influenced by sojourn time as it only depends on the system throughput. In addition, when the load of elastic applications is low, the simulation results of both MAX and MIN are fairly worse than the theoretical approximation. However, with the increase of elastic traffic load, the performance of MAX and MIN first gets closer to and then the time-average throughput outperforms the theoretical approximation. The reason for this performance alternation is due to the use of the service capacity of WLAN for heavy-loaded elastic applications. When the load of elastic applications is low, almost all the elastic applications are served in WiMAX according to the handover criteria. However, with increasing load of elastic applications, more and more files are dispatched or handed over to WLAN. Thus all the service capacities of WiMAX and WLAN are utilized when the load of elastic applications are heavy enough.

In Fig. 4 the dispatch and handover times are shown to illustrate the cost of each load balancing mechanism. It can be seen that MIN introduces much more load balancing actions (dispatch/handover) than MAX and even two times more when the load of elastic applications is heavy. Moreover, MIN adopts much more handovers than MAX while the numbers of dispatch for two mechanisms are comparable.

In summary, for exponential distributed files, the performance results of MAX and MIN are similar. However, MIN introduces much more operation costs with a large number of load balancing actions. Therefore, we prefer MAX to MIN with consideration of overall performance in this context.

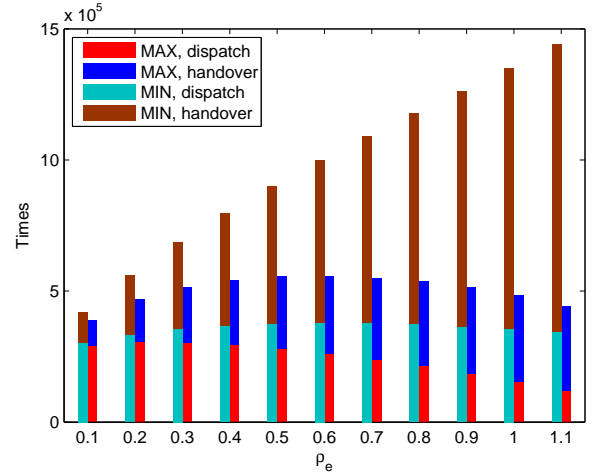


Fig. 4. Dispatch and handover times with minus exponential distributed file size.

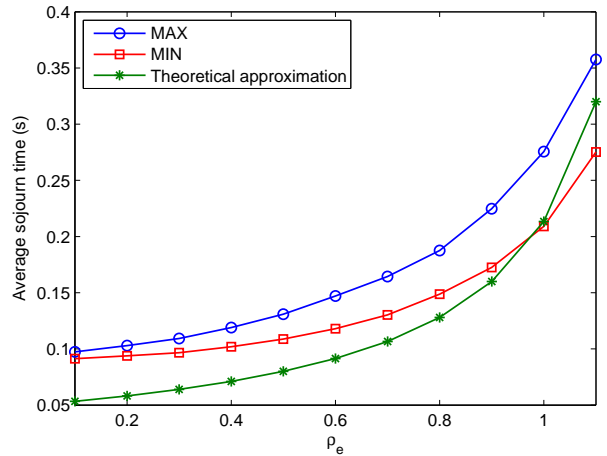


Fig. 5. Average sojourn time with Pareto distributed file size.

B. Pareto Distribution

For realistic applications, it is believed that the size of elastic applications follows heavy-tailed distribution. Normally Pareto distribution is chosen as the representative heavy-tailed distribution especially for file sizes. Therefore, we also present results with Pareto distributed file sizes. Specifically, we take the shape parameter as 1.2 which is a typical value for Pareto distribution.

In Fig. 5, the average sojourn time is shown. Compared with Fig. 2, there are two major differences. First, the average sojourn time with Pareto distributed file sizes is shorter than that with exponential distribution. This phenomenon has been stated in [11] and the reason is that Pareto distribution has higher variability than exponential distribution. Second, the difference between MAX and MIN is more visible with Pareto distributed file sizes. Moreover, for heavy-loaded elastic applications, MIN provides shorter average sojourn time than the theoretical approximation. In fact, since MAX and MIN actually take advantage of the size information of files, there are chances that the average sojourn time is shorter than the

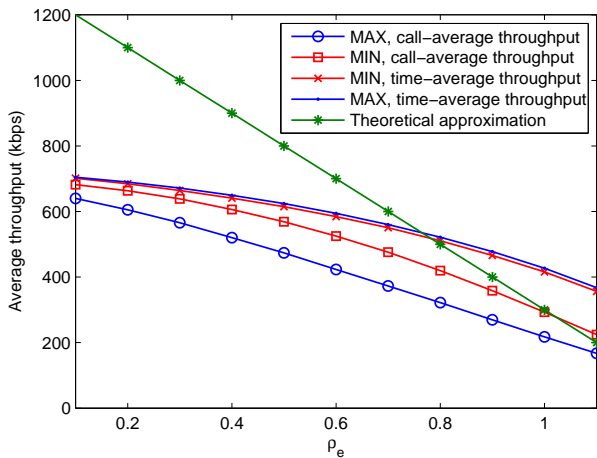


Fig. 6. Average throughput with Pareto distributed file size.

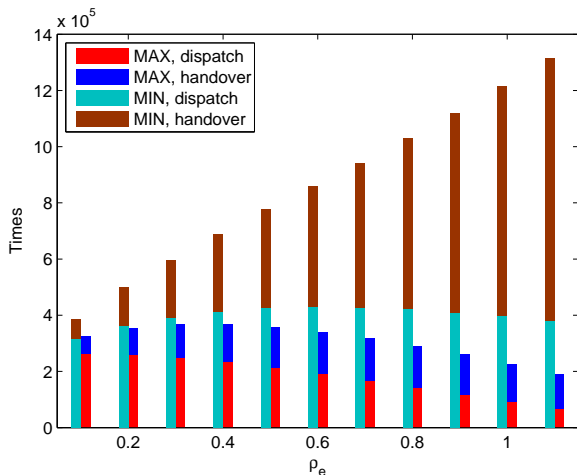


Fig. 7. Dispatch and handover times with Pareto distributed file size.

theoretical approximation based on PS.

In Fig. 6, the throughput results are shown. It can be seen that the difference of call-average throughputs with MAX and MIN is much larger compared with Fig. 3. The difference complies with the visible difference of average sojourn time in Fig. 5. These results indicate that the performance difference of MAX and MIN is much larger with Pareto distributed files. In addition, there is a crossover between call-average throughput with MIN and the theoretical approximation. This is consistent with the result in Fig. 5.

In addition, by comparing Fig. 2 with Fig. 5 and Fig. 3 with Fig. 6, we obtain following observations. For the light-loaded area, the performance results with Pareto or exponential distributions are quite similar. However, for the heavy-loaded area, namely when the load of elastic applications is larger than 0.9, the performance results with Pareto distribution are much better than those with exponential distribution.

Fig. 7 presents the numbers of load balancing actions. Compared with Fig. 4, the difference between MAX and MIN is even larger. Specifically, for MIN, the total number of load balancing actions is comparable for both Pareto and

exponential distribution while more handovers are adopted with Pareto distribution. For MAX, much less load balancing actions are taken with Pareto distribution especially in the heavy-loaded area.

V. DISCUSSION

Based on the simulation results and theoretical approximations, we present the following discussions on the impact of size of elastic applications and further thinking on load balancing.

The simulation results suggest that MIN outperforms MAX in terms of all the performance metrics for elastic applications. However, this performance overwhelm comes with cost of significantly increased number of load balancing actions. Besides the increased number of handover times, another possible reason for the better performance of MIN is that MIN lets large files stay in WiMAX. This moves the system towards the quasi-stationary assumption that streaming applications evolves much faster than elastic applications in WiMAX. However, MAX drives the system towards another quasi-stationary assumption that elastic applications evolve much faster. As stated in [8], the former quasi-stationary assumption leads to better average performance. However, this statement applies with condition of uniform stationarity where the load of elastic applications needs to be fairly low.

Another possible reason for the better performance of MIN over MAX is the better load balancing granularity. Much more files need to be distributed to WLAN with MIN to achieve load balancing since those files are relatively small. Then both the frequency and size of load balancing actions ensure better load balancing granularity with MIN. However, as shown in Figs. 4 and 7, this also introduces a large proportion of handover which is much costly than dispatching.

The dilemma between MAX and MIN inspires us to think how to keep the advantage of MIN while in the meanwhile reducing the number of load balancing actions. To this aim, we point out that two aspects should be taken into consideration. First, sufficient load balancing granularity needs to be provided. Second, the characteristics of integrated services in WiMAX need to be explored and utilized.

It is worth highlighting that in this study we have made several assumptions in order to investigate the fundamental effect of different load balancing mechanisms. Specifically, we assume the capacity of WLAN is fixed and does not depend on the number of users in the network. While the assumptions may not always hold, we believe that the trends discovered in this study will remain under released assumptions. Moreover, the results in the paper could also be helpful for designing load balancing mechanisms for other overlay heterogeneous networks.

VI. CONCLUSION

In this paper we study load balancing for multi-service in an overlay heterogeneous network. Based on the analysis of simulation results of two load balancing mechanisms, we draw the conclusion that both load balancing granularity and integration of elastic and streaming applications in WiMAX

affect the whole system performance. This knowledge could provide guidance for further developing better load balancing mechanisms.

REFERENCES

- [1] N. Nasser, et al., "Handoffs in fourth generation heterogeneous networks," *IEEE Communications Magazine*, vol. 44, pp. 96-103, 2006.
- [2] M. Kassar, et al., "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Computer Communications*, vol. 31, pp. 2607-2620, 2008.
- [3] X. Yan, et al., "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks," *Computer Networks*, vol. In Press, Corrected Proof, 2010.
- [4] X. Liu, et al., "Joint Radio Resource Management through Vertical Handoffs in 4G Networks," in *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*, 2006, pp. 1-5.
- [5] A.-E. M. Taha, et al., "Vertical handoffs as a radio resource management tool," *Computer Communications*, vol. 31, pp. 950-961, 2008.
- [6] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *Wireless Communications, IEEE Transactions on*, vol. 8, pp. 725-735, 2009.
- [7] R. Litjens, et al., "Throughputs in processor sharing models for integrated stream and elastic traffic," *Performance Evaluation*, vol. 65, pp. 152-180, 2008.
- [8] F. Delcoigne, et al., "Modeling integration of streaming and data traffic," *Performance Evaluation*, vol. 55, pp. 185-209, 2004.
- [9] R. Malhotra and J. L. v. d. Berg, "Flow level performance approximations for elastic traffic integrated with prioritized stream traffic," in *Telecommunications Network Strategy and Planning Symposium, 2006. NETWORKS 2006. 12th International, 2006*, pp. 1-9.
- [10] S. Borst and N. Hegde, "Integration of Streaming and Elastic Traffic in Wireless Networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, 2007, pp. 1884-1892.
- [11] R. Litjens and R. J. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the QoS," *Performance Evaluation*, vol. 52, pp. 193-220, 2003.