Multi-Layer Mobility Load Balancing in a Heterogeneous LTE Network

Panagiotis Fotiadis[1], Michele Polignano[1], Daniela Laselva[2], Benny Vejlgaard[2], Preben Mogensen[2,1], Ralf Irmer[3], Neil Scully[3]

[1]*Aalborg University*
Department of Electronic Systems
paf@es.aau.dk, mpo@es.aau.dk

[2]*Nokia Siemens Networks*
Research Center, Aalborg (DK)
{name}.{surname}@nsn.com

[3]*Vodafone Group*
Research & Development, Newbury (UK)
{Name}.{Surname}@vodafone.com

Abstract—**This paper analyzes the behavior of a distributed Mobility Load Balancing (MLB) scheme in a multi-layer 3GPP (3rd Generation Partnership Project) Long Term Evolution (LTE) deployment with different User Equipment (UE) densities in certain network areas covered with pico cells. Target of the study is to evaluate MLB in terms of efficient pico cell utilization and macro layer load balancing (LB). The analysis focuses on video streaming traffic due to specific service characteristics (e.g. play-out buffer delay/ jitter protection) that might make any mobility performance degradation transparent to the end user performance. Results have shown that the proposed MLB scheme can significantly improve the overall network resources utilization by eliminating potential load imbalances amongst the deployment layers and consequently enhance user experience. However this occurs at the cost of increased Radio Link Failures (RLF), a fact that might be critical for further applying MLB in real-time conversational services without additional mobility optimization and interference management techniques.**

Keywords-LTE, Load Balancing, Mobility; Heterogeneous Networks; Self Organizing Networks (SON)

I. INTRODUCTION

Mobile data traffic is growing extensively and it is expected that, compared to 2011, an 18-fold increase in total network traffic will occur by 2016 [1]. Obviously, this growth of mobile broadband poses new challenges to operators in terms of meeting the future coverage and capacity requirements. Although LTE is an emerging technology which is expected to provide enhanced spectral efficiency, macro-only deployment will not be sufficient. The installation of additional low power small cells (pico/femto) seems to be a promising solution for tackling the above mentioned requirements and operators are planning to complement their macrocell systems in such a manner, creating multi-layer topologies, referred to as Heterogeneous Networks (HetNet). However, interference co-ordination in case of co-channel deployment, mobility management and efficient network utilization are major challenges for HetNets, and they should be tackled in an automated manner due to additional system complexity. Self-Organizing Networks (SON) [2] target towards this direction and 3GPP standardization has already defined specific features for achieving an autonomous network management.

Mobility Load Balancing (MLB) is included within the SON LTE framework and its responsibility is to optimally distribute traffic among the different layers [3] by exploiting mobility management and load knowledge of the neighboring cells. In such a manner, overloaded cells can identify potential under-utilized nearby eNBs and attempt to shift part of their traffic towards them by adjusting handover parameters. Nevertheless, such an approach might disrupt mobility management as users are forced to stay connected longer to low loaded cells at a cost of lower spectral efficiency. This results in a higher risk of Radio Link Failures (RLF), if coverage degrades significantly. However the impact of RLFs on end user performance may not be the same for different type of services. Streaming applications typically do not demand as strict constraints as conversational services do. In fact, playout buffering at the receiver makes them, to a certain degree, immune to delay jitter [4]. Thus, even experiencing a RLF might be totally transparent to the UE, unless buffer underflow occurs during the connection re-establishment procedure.

Our objective is to investigate the potentials of load balancing in a HetNet LTE network. Compared to prior studies [5-6], which have only focused on single layer (Intra-Frequency) MLB scenarios, multi-layer HetNet deployments provide an additional level of freedom for load balancing to act, as it can operate between different carrier frequencies and base station technologies, given that the sufficient overlapping coverage is provided. For that reason, a threshold-based multi-layer MLB algorithm is proposed and its impact on video streaming applications is investigated under macro/pico co-channel interference conditions. Moreover, an additional macro carrier at a different lower frequency (escape carrier) is provided for coverage purposes.

The paper is organized as follows. Section II describes the basic system model, while the proposed MLB framework is presented in Section III. Section IV outlines the simulation parameters. Finally, numerical results are provided in Section V, whereas Section VI concludes the paper.

II. SYSTEM MODEL

A. Mobility Management

The studied LTE HetNet scenario consists of two co-sited macro layers (Macro 800 MHz & 2600 MHz respectively) and a pico layer sharing the 2600 MHz frequency (2 pico cells /macro sector area). Regarding mobility management, both Intra and Inter Frequency handovers are triggered by the A3 event [7], which is reported to the serving eNode (eNB) if the

978-1-4673-1881-5/12/$31.00 ©2012 IEEE

UE detects a better neighboring cell. The corresponding mathematical expression is presented in Eq (1). $M_{S,eNB}$, $M_{T,eNB}$ correspond to the UE measurements for the serving and target eNB respectively in terms of Reference Signal Received Quality (RSRQ) [7], and H_m is the cell-pair defined handover offset, also referred to as Cell Individual Offset (CIO).

$$M_{s,eNB} + H_m < M_{T,eNB} \tag{1}$$

However, since Inter-Frequency handovers cost more in terms of signaling overhead a higher CIO value is applied compared to the equivalent Intra-Frequency A3 trigger. Note that Inter-Frequency handovers can further be limited, if the corresponding measurements are not always performed. In our study, the A2 event [7] ('*serving quality becomes worse that threshold* ') is utilized for that purpose and the choice of the measurement triggering threshold is a compromise between Inter-Frequency handover signaling cost and user-perceived performance.

B. Load Definition and Composite Available Capacity

LTE MLB functionality is governed by standardized signaling that takes places over the X2 interface. Cells contiguously monitor their load conditions and report their Composite Available Capacity (CAC) to neighboring eNBs, as specified in [8]. CAC represents the overall available resource level that can be offered for LB, given that an operational load target is defined. The above described framework is modeled below.

If $n_i(t_k)$ is the amount of occupied resources at the measurement interval t_k, and, $N_{i,PRB}$, the total bandwidth of the cell i in terms of PRBs, then the instantaneous load sample $\rho_i(t_k)$ and the corresponding cell load estimation $\tilde{\rho}_i\,(t_k)$ can be expressed as:

$$\rho_i(t_k) = \frac{n_i(t_k)}{N_{PRB}} \tag{2}$$

$$\tilde{\rho}_i\,(t_k) = (1-a)\cdot\tilde{\rho}_i\,(t_{k-1}) + a\cdot\rho_i(t_k) \tag{3}$$

Note that the cell load estimate is based on an Infinite Impulse Response (IIR) filter, where α represents the memory of the filter. Given that ρ_{Target} is the target operational load in terms of resources occupancy, CAC is modeled as follows:

$$CAC = 100\cdot\left(1-\frac{\tilde{\rho}_i\,(t_k)}{\rho_{Target}}\right) \tag{4}$$

III. MLB ALGORITHM

The proposed distributed multi-layer MLB scheme adjusts CIOs in favor of under-utilized cells, relying on the exchanged load information. The term multi-layer implies that MLB can operate between cells that belong to different base station technology (macro/ pico cell) or carrier frequency, as long as the X2 interface is present. Based on $\tilde{\rho}_i\,(t_k)$ and the two vendor-specific load thresholds Thr_{High} and Thr_{Low}, a cell is either tagged as active, passive or neutral, specifying certain actions depending on its state (Table I). The hysteresis range between the thresholds is defined according to Eq. (5).

Table I. MLB CELL STATE CHARACTERIZATION

Cell Status	Condition	Action
Passive	$\tilde{\rho}_i(t_k) < Thr_{Low}$	Performs cell range extension
Neutral	$Thr_{Low} \le \tilde{\rho}_i(t_k) \le Thr_{High}$	Does not participate in MLB
Active	$\tilde{\rho}_i(t_k) > Thr_{High}$	Requests cell shrinking

$$Thr_c = \begin{cases} \rho_{Target} + \rho_{hyst}\,, & c = High \\ \rho_{Target} - \rho_{hyst}\,, & c = Low \end{cases} \tag{5}$$

In order to avoid system instability due to rapid load fluctuations around the ρ_{Target} region, neutral cells do not participate in any MLB activity. Hence, CIO negotiations/ adjustments are only allowed between active-passive (*high - low loaded* respectively). Note that the algorithm is only triggered by active cells, as such an approach, minimizes the signaling overhead over the X2 interface. Furthermore, by setting ρ_{Target} sufficiently high we ensure that MLB will not be triggered at low load conditions, where load balancing is not of vital importance.

Similarly to CAC, the Composite Missing Capacity (CMC) can be defined for active cells according to Eq. (6).

$$CMC = 100\cdot\left(\frac{\tilde{\rho}_i\,(t_k) - \rho_{Target}}{\rho_{Target}}\right) \tag{6}$$

In case of $\tilde{\rho}_i\,(t_k)$ exceeding Thr_{High}, the active cell calculates its CMC and initiates the Resource Status Update [8] requesting CAC information from its neighbor passive cells. Since CMC/ CAC represent the percentage of occupied resources below and above ρ_{Target} respectively, the ideal load shift ratio (LSR) for a negotiating cell can be estimated by Eq. (7).

$$LSR = \begin{cases} 1 - \dfrac{CMC}{100+CMC} < 1, for\ active\ cells \\ \\ \dfrac{100}{100-CAC} > 1, for\ passive\ cells \end{cases} \tag{7}$$

Regarding the CIO negotiation (Mobility Setting Change Procedure [8]), LSR is mapped to a maximum ΔCIO coverage adjustment as follows:

$$\Delta\mathrm{CIO_{max}} = 10\log_{10}\left[\left(\sqrt{\mathrm{LSR}}\right)^{\beta}\right] \tag{8}$$

where β is a vendor specific parameter that can be configured per cell/ layer pair (e.g. higher β for macro-pico cell pairs can be used for offloading). Note that ΔCIO_{max} values might differ in an active-passive cell pair, as they depend on the corresponding CMC and CAC respectively. In order to avoid asymmetrical cell shrinking/ range extension, the minimum value is selected according to Eq. 9.

$$\Delta CIO_{Neg} = \min\left(\left|\Delta CIO_{max,active}\right|, \Delta CIO_{max,passive}\right) \qquad (9)$$

Thus, if neighbor cells i, j are active and passive respectively, $\mathrm{CIO}_{\mathrm{Initial}}$ is the initial A3 offset value and, $CIO_{relax,max}$, the maximum allowed CIO relaxation, the updated CIO values in each direction are given by Eq. (10) & Eq. (11). Similarly to [6], ΔCIO_{Neg} is applied symmetrically such as to minimize the chance of ping pong handovers occurrence.

$$CIO_{New,i \to j} = \max\left(CIO_{old,i \to j} - \Delta CIO_{Neg}, CIO_{Initial} - CIO_{relax,max}\right) \qquad (10)$$

$$CIO_{New,j \to i} = \min\left(CIO_{old,j \to i} + \Delta CIO_{Neg}, CIO_{Initial} + CIO_{relax,max}\right) \qquad (11)$$

Note that $CIO_{relax,max}$ determines the maximum range in which CIO values can be adjusted. In principle, SON Mobility Robustness Optimization (MRO) [2] should dynamically control this parameter relying on periodic mobility-related KPIs collection. However, since MRO is out of the scope of this study, a fixed value is considered. The proposed algorithm has a predictive behavior and therefore, it might be possible that no traffic will be shifted to a passive cell unless a handover is triggered. In that case, the active cell will remain overloaded and MLB will again be triggered. In order to minimize the negotiation attempts and also for the sake of system stability, whenever a CIO is adjusted, further negotiations between that specific cell-pair are frozen for a certain time duration, referred to as Wait Period (WP).

IV. SIMULATION ASSUMPTIONS

Two hotspot areas are randomly generated per macro sector and pico cells are placed in the center of the hotspot (Fig. 1). The hotspot radius is set to 80m. Hotspot UEs represent the 66% of total users, while their movement is confined within the circular area defined by the hotspot center and the respective radius (*bouncing back when they reach the border*). The remaining UEs are free moving users moving in straight lines at either 3 km/h or 50 Km/h. The ratio between free moving UEs at 3 Km/h and 50 Km/h is set to 0.5. A detailed summary of all simulation assumptions is provided by Table II.

A. Load Balancing Index

Similarly to [9], in order to better visualize the MLB performance, the load balancing index ξ is defined in Eq. (12), where $\rho_l(t)$ is the average load of layer l and $\overline{\rho(t)}$ the average load of the whole network at time t. Better load distribution amongst the different layers is indicated as $\xi(t) \to 0$.

$$\xi(t) = \sum_l \left(\rho_l(t) - \overline{\rho(t)}\right)^2 \qquad (12)$$

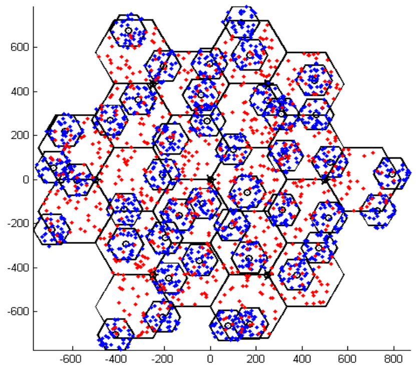

Figure 1: Network Layout Example (Blue is used for hotspot UEs and red for free moving ones)

B. Scheduling Assumptions

Users with an average experienced bit rate below the source rate requirement r_{min} (1 Mbps) are ranked into descending wideband SINR order and are allocated the necessary number of PRBs in order to meet r_{min}. The remaining resources are then shared equally among the UEs with an average experienced bit rate above r_{min}.

Table II. SIMULATION PARAMETERS

Network Layout	Hexagonal grid, 7 cell sites, 3 sectors per site
Carrier Frequency (Bandwidth)	Macro 800 MHz (10MHz) Macro 2600 MHz (20MHz) Pico 2600 MHz (20MHz)
ISD	500 m
Propagation Model	Hata COST 231 (Macro), 3GPP (Pico)
Transmission Power	Macro: 43 dBm, Pico: 36 dBm
Shadowing Std. Deviation	Macro: 8 dB , Pico: 10 dB
Shadowing Correlation Length	Macro: 50 m, Pico: 13 m
Simulation Length	20 min
Initial CIO (Intra / Inter)	2 dB (Intra-HO) / 4dB (Inter-HO)
A2 event threshold	-12 dB
UE measurement rate	100 msec
Time-to-Trigger Window	0.4 sec (Intra-HO) / 0.5 sec (Inter-HO)
HO execution timer	0.15sec (Intra-HO) / 0.25 sec (Inter-HO)
Cell load measurement rate	500 msec
MLB Wait Period	10 sec
MLB ρ_{Target}	0.7
MLB ρ_{hyst}	0.1
MLB $CIO_{relax,max}$	5 dB
RLF modelling	Based on T310 [10], re-connection based on channel quality

C. *Video Streaming Traffic & Satisfaction Model*

Constant Bit Rate (CBR) streaming traffic is assumed with a source rate of 1 Mbps. Session arrivals are exponentially distributed and the video duration time is set to 2 min. The UE satisfaction per session is evaluated based on the buffer at the terminal's side. The video starts playing only when the buffering threshold is reached and frames are read from the buffer at the play-out rate. If the downlink bit rate is the same with the play-out rate, the buffer size shall remain constant and no re-buffering will occur. On the other hand, decreasing amount of buffered data implies that the downlink bit rate is lower than the play-out one and re-buffering will occur if the buffer eventually empties. Our model records both the initial and total re-buffering time and assesses Quality of Experience (QoE) based on the thresholds presented in Table III. The required initial buffering threshold is set to 5sec. Finally, the satisfaction ratio is defined as the number of happy streaming sessions over the total number of initiated ones.

Table III. STREAMING SATISFACTION MODEL

Criterion	**QoE Assessment**		
	Happy	***Unhappy***	***Dropped***
Initial Buffering Time (t_{ib})	$t_{ib} \leq 8$ sec	$8 < t_{ib} \leq 10$ sec	$t_{ib} > 10$sec
Total Re-buffering Time (t_{rb})	$T_{rb} = 0$ sec	$0 < t_{rb} \leq 10$ sec	$t_{rb} > 10$sec

V. NUMERICAL RESULTS

The assessment of MLB performance is conducted based on a sensitivity analysis for different offered load conditions. The simulated traffic levels (*Mbps per macro sector area*) correspond to an average total network utilization of ~40% to ~80%.

Fig. 2 shows the load balancing index with and without MLB. The balanced load distribution among the different layers is depicted by the significantly lower value of ξ compared to the reference case. In specific, MLB decreases the index by a factor of 8 at high load, whereas smaller gains are observed at lower offered traffic conditions. Moreover, the continuous increase of ξ when MLB is not applied indicates that larger load imbalances are created. Increasing offered load implies more severe co-channel interference at the 2600 MHz layers. Hence, more users are handing over to the escape carrier resulting in overload; a condition which eventually is resolved when MLB is applied.

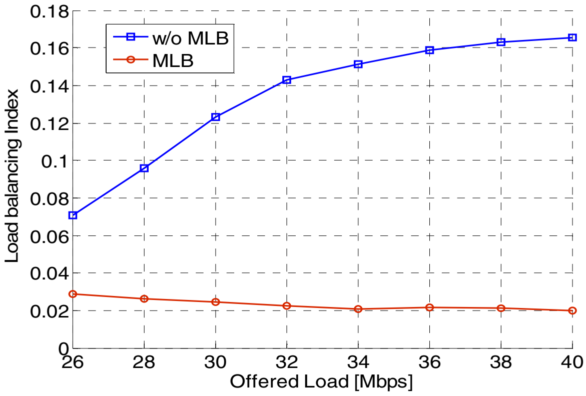

Figure 2: Load Balancing Index

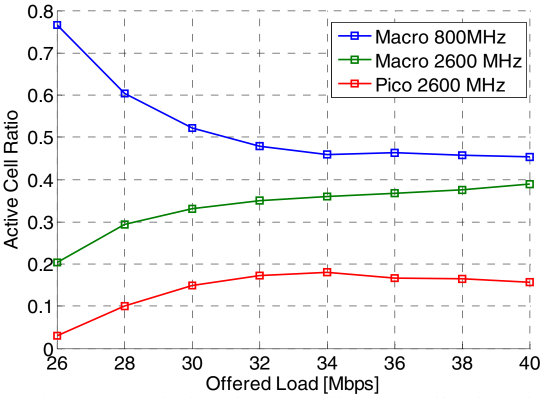

Figure 3: Distribution of Active Cells versus Offered Load

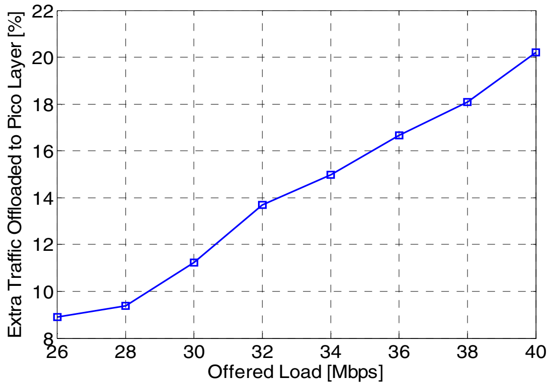

Figure 4: Additional amount of traffic offloaded to picos (MLB enabled)

Fig. 3 displays the distribution of active cells in the layers. Observe that in all cases, the higher probability for triggering MLB is at the Macro 800 MHz layer, a fact that further validates our previous statement. On the other hand, higher load conditions are needed in order to trigger MLB at the 2600 MHz carrier as the allocated bandwidth for these layers is double compared to the escape carrier one.

Macro-to-pico offloading is shown in Fig. 4. We observe that the additional traffic carried by the picos (*reference is the non MLB case*) is higher as the offered load increases. This observation is primarily due to the inband 2600 MHz Macro-to-Pico MLB operation. At high load conditions, the only potential passive state candidates for a Macro 2600 MHz are its neighboring pico eNBs and consequently offloading occurs.

The end user performance in terms of concluded happy sessions is presented in Fig. 5. The benefit of multi-layer MLB is quite evident as the overall session satisfaction ratio is kept significantly high, even at the 40 Mbps case. However, this can happen only up to a certain traffic level, as heavier load conditions will dramatically decrease the probability of finding a passive neighbor for negotiating a new CIO adjustment. An even more important aspect is definitely the capacity gains that the algorithm provides for a fixed target

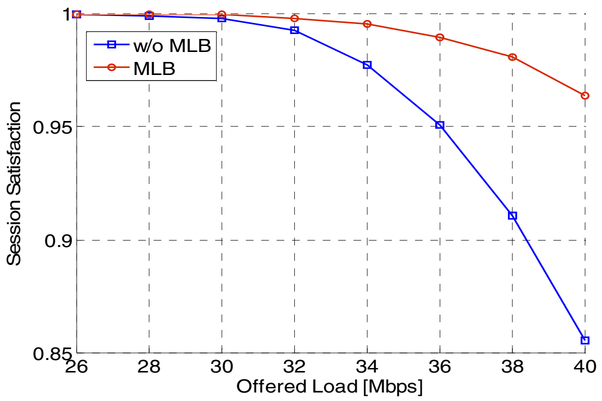

Figure 5: Session Satisfaction Ratio

performance network requirement. By assuming a minimum acceptable session satisfaction ratio of 95%, the network is capable of carrying more than 10% of additional traffic when MLB is applied while maintaining the desirable end user performance.

As far as mobility performance is concerned, Fig. 6 depicts the average RLF rate for the different offered traffic cases, expressed as the number of RLFs per user per hour. Compared to the non MLB case, MLB increases significantly the RLF rates, by a factor of ~5 at high load conditions. This effect is totally expected due to fact that MLB tries to utilize more the 2600 MHz layers, a behavior that inevitably leads into higher levels of co-channel interference. Note that the major cause of RLFs in this scenario is late pico-to-macro handovers. Combining the results from Fig. 5 and Fig. 6, the RLF cost on UE satisfaction is minimal due to the inherent delay jitter robustness of video steaming applications provided by the play-out buffer functionality. By exploiting the capacity of the 2600 MHz layers, users are better served as resources availability compensates for the degraded spectral efficiency. Thus, it is possible to guarantee that the amount of buffered data is sufficient for avoiding any buffer underflow during the connection re-establishment that would otherwise trigger undesirable video rebuffering. Although the RLF rates in the investigated scenario are quite small due to the coverage provided by the 800 MHz carrier, there is a clear indication that MLB causes a significant RLF increase in the presence of strong co-channel interference, while degradation can be even worse in scenarios where higher mobility is assumed and no escape carrier is deployed. Hence, the traffic steering impact on real-time services, such as typical conversational applications, in a similar network deployment would have been much more severe as their requirements in terms of delay and data loss are significantly stricter.

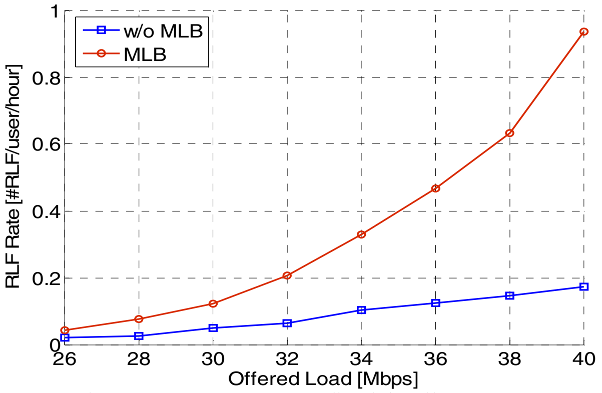

Figure 6: MLB Impact on Radio Link Failures Rate

VI. Conclusions

In this paper, we have studied a multi-layer MLB scheme that dynamically adjusts the cell-pair CIOs based on load information signaling over the X2 interface. The algorithm has been evaluated in a LTE heterogeneous network consisting of a macro/ pico co-channel deployment at 2600 MHz, supplemented by an escape carrier at 800 MHz. We have shown that MLB can efficiently distribute the load across the different layers, leveraging as much as possible the macro-to-pico offloading. The threshold-based trigger ensures that MLB does not disrupt the system at low load conditions, given that the operational load target is set sufficiently high. However, at high load conditions and strong interference levels, MLB can significantly increase the RLF rates. The mobility degradation has minimal effect on video steaming applications due to the robustness provided by the play-out buffer. Hence, in such a deployment, MLB should operate on a service-aware manner, unless further interference management/ mobility optimization techniques are additionally applied; otherwise, the impact on real-time conversational applications might be critical. For that purpose, future work includes mobility enhancements by providing MLB/MRO integration and performance evaluation in an even more challenging HetNet deployment (e.g. escape carrier removal, 4 picocells/ macro sector area).

VI. References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016", Cisco Public Information, February 14, 2012.

[2] 3GPP standardization, "Self-configuring and self-optimizing network use cases and solutions (Release 9)", TR 36.902 v1.2.0, May 2009.

[3] S. Hamalainen, H. Sanneck, and C. Sartori, "LTE Self-Organizing Networks (SON): Network Management Automation for Operational Efficiency", Wiley 2012.

[4] V. Vukadinovic, and G. Karlsson, "Video Streaming Perfomance under Proportional Fair Scheduling", IEEE Journal on Selected Area In Communications, Vol. 28, No. 3, April 2010.

[5] R. Kwan, R. Arnott, R. Patterson, R. Trivisonno, and M. Kubota "On Mobility Load Balancing for LTE Systems", IEEE Vehicular Technology Society, Sep. 2010.

[6] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load Balancing in Downlink LTE Self-Optimizing Networks",IEEE Vehicular Technology Conference, May 2010.

[7] 3GPP TS 36.331,Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA);Radio resource Control (RRC); Protocol Specification (Release 9).

[8] 3GPP TS 36.423,Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); X2 Application Protocol (Release 10).

[9] H. Wang, L. Ding, P. Wu, Z. Pan, N. Liu, and X. You, "Dynamic Load Balancing and Throughput Optimazation in 3GPP LTE Networks", International Wireless Communications &Mobile Computing, July 2010

[10] 3GPP TS 36.133, "Evolved Universal Terrestrial Radio Access (EUTRA),"Requirements for support of radio resource management", V10.3.0 (2011-06).