

Investigation of Network Virtualization and Load Balancing Techniques in LTE Networks

Ming Li¹, Liang Zhao², Xi Li², Xiaona Li¹, Yasir Zaki², Andreas Timm-Giel¹, Carmelita Görg²

¹ Hamburg University of Technology, Schwarzenbergstrasse 95E, 21073 Hamburg, Germany
{ming.li, xiaona.li, timm-giel}@tuhh.de

² University of Bremen, Otto-Hahn-Allee NW1, 28359 Bremen, Germany
{zhaol, xili, yzaki, cg}@comnets.uni-bremen.de

Abstract— Mobile Network Virtualization (NV) is an emerging technique which has drawn increasingly research attention. Network Virtualization enables multiple network operators to share a common infrastructure (including core network, transport network and access network) so as to reduce the investment capital while improving the overall performance at the same time. This is achieved by exploring the multiplexing gain. Similarly, Load Balancing (LB) is a well-known mechanism used in mobile networks to offload excessive traffic from high-load cells (hot spots) to low-load ones within one network operator. This paper aims at investigating the potential gain of applying NV in LTE (Long Term Evolution) networks and compares it with the LB scheme gain. In this paper, we propose an LTE virtualization framework (that enables spectrum sharing) and a dynamic load balancing scheme for multi-eNB and multi-VO (Virtual Operator) systems. We compare the performance gain of both schemes for different applications, e.g. VoIP, video, HTTP and FTP. We also investigate the parameterization of both schemes, e.g. sharing intervals, LB intervals and safety margins, in order to find the optimal parameter settings. The presented results show that the LTE networks can benefit from both NV and LB techniques.

Index Terms—Network Virtualization, Spectrum Sharing, Load Balancing, LTE

I. INTRODUCTION

Network virtualization (NV) is an attractive technique that has recently received the research community attentions([1], [2], [3] and [4]). LTE virtualization is an important use-case of network virtualization. The main advantages of NV include: high resource utilization, improved system performance, lower investment capitals for Virtual Operators (VOs), and better end-user experience, etc. ([5], [6] and [7]). Spectrum sharing is a key technique in LTE virtualization; it can be used at the air interface to adapt to the traffic load variation of different VOs. On the other hand, Load Balancing (LB) is a well-known technique that evenly distributes the traffic load among multiple cells of a cellular system. In this paper, we propose a spectrum sharing scheme for LTE NV and a dynamic LB scheme for multi-eNB and multi-VO systems, and furthermore compare their performance gain of NV with LB for different services.

In [6] and [7], LTE virtualization was investigated as a case study of general wireless virtualization. The main focus was to highlight the advantages and potential gains that can arise from using LTE virtualization, the authors also proposed an initial LTE virtualization framework. In [5], the authors further quantified the multiplexing gain of spectrum sharing in LTE virtualization, from both analytical and simulation perspectives.

A new load estimation mechanism was also proposed in [5] that relies on real time traffic models. However, some other interesting aspects were not addressed, these are: firstly, the load contribution of the nGBR (non-Guaranteed Bit Rate) services (e.g. FTP) has not been directly taken into account in the overall traffic load estimations; secondly, there has been no detailed evaluations and analysis on the impact of NV in a mixed traffic scenarios, e.g. the impact on the end user performance; finally, the multi-VO spectrum sharing results were not compared against other Radio Resource Management (RRM) mechanisms such as LB.

In order to address the aforementioned issues, this paper addresses following aspects: (i) we propose a practical algorithm for spectrum sharing to improve the resource utilization: using safety margins on the load estimation for GBR traffic to guarantee the real-time traffic QoS and applying AMBR (Aggregated Maximum Bit Rate)[8] for load estimation for nGBR traffic; (ii) we analyze in detail the impact of spectrum sharing on the end user performance for different applications, i.e. VoIP, real-time video, HTTP and FTP; (iii) we also compare the gain of spectrum sharing against to one from the LB; (iv) the optimization of the schemes parameters are investigated, e.g. sharing and LB intervals, safety margins and AMBR values.

The rest of this paper is organized as follows: Section II introduces the simulation model and the considered scenarios. Section III presents the detailed algorithms and strategies for the spectrum sharing scheme (including the load estimation for spectrum sharing and resource allocation) and LB scheme. In Section IV the simulation scenarios are defined to evaluate our proposals, and the results are presented and discussed. Section V concludes this work and outlines our future work.

II. SIMULATION MODEL INTRODUCTION

The LTE simulator is implemented using OPNET simulation software (version 15.0) [9]. The LTE virtualization framework in [7] is extended to support multiple-eNB LTE system in order to evaluate the proposed LB scheme. An extended NV scheme (Section III-1) is implemented which takes sharing interval, GBR safety margin, nGBR demand estimation, AMBR into consideration. In addition, handover and LB scheme are integrated into this simulation model.

The designed model follows the 3GPP specifications. It includes all basic E-UTRAN and EPC (Evolved Packet Core) network entities. Figure 1 shows an example scenario with two enhanced NodeBs (eNB), and a number of IP routers

connecting the eNBs and the aGW in the E-UTRAN network. The EPC network entities are represented by the aGW network node. The remote node is represented by an Internet server or any nodes that provide data or voice services.

The radio (Uu) interface includes the radio protocols such as PDCP, RLC, MAC and PHY between the UE entity and the eNB. The PDCP, RLC and MAC (including air interface scheduler) layers are modeled in detail according to the 3GPP specifications in this simulator. The LTE transport network is based on IP technology. The user-plane transport protocols are implemented for both the S1 interface (i.e. the interface between the eNB and the aGW) and the X2 interface (i.e. the interface between the eNBs) according to 3GPP. The S1/X2 interface mainly includes the GTP, UDP, IP and L2 protocols. Ethernet is chosen as the L2 protocol in this simulation model.

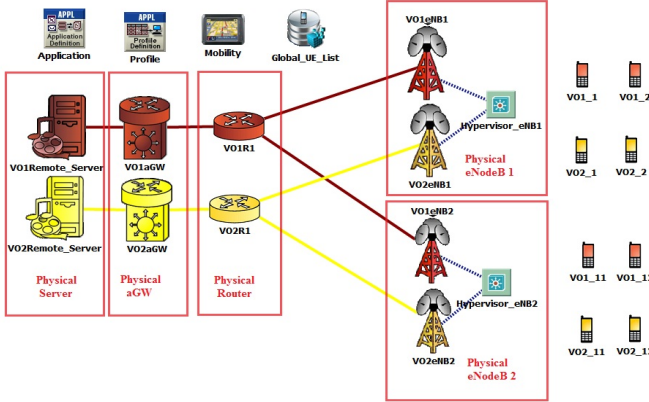


Figure 1. LTE Simulation Model in OPNET

The example scenario contains two Virtual Operators (VOs) sharing the same network infrastructure consisting of the physical eNBs, routers, aGW and servers. This paper focuses only on eNB virtualization, i.e. multiple VOs sharing the spectrum of the same physical eNB. An entity named “Hypervisor” is located at each eNB and is responsible for allocating the radio resources (frequency spectrum in terms of Physical Resource Blocks (PRBs)) to the VOs. In the legacy no-sharing scheme, each VO will get a fixed amount of spectrum resource from the hypervisor regardless of its demand and utilization. In this example scenario, within each physical eNB the hypervisor allocates the spectrum resource to VO1 and VO2 periodically based on their demands.

In addition to virtualization, the LB scheme, which is introduced in Section III-2, is also implemented. By means of LB, it is possible to offload the excessive traffic load from high-load cells to low-load cells. For instance, when the load of VO 1 in eNB 1 is too high, some users can be shifted by means of handover to the neighboring cell, i.e. VO 1 in eNB2.

III. ENB VIRTUALIZATION AND LOAD BALANCING ALGORITHM

1. Algorithm of eNB Virtualization

As explained in Section II, there is a hypervisor in each physical eNB, which is responsible for allocating the spectrum resources to different VOs periodically. Then each VO is able to schedule its available resource to its users. The hypervisor conceptual framework is illustrated in Figure 2. The inputs of

the hypervisor are the total number of PRBs of the physical eNB, the number of VOs, the estimated spectrum demand by different services from different VOs, and the related spectrum sharing parameters. The output of the hypervisor is the allocation unit in terms of number of PRBs for each VO.

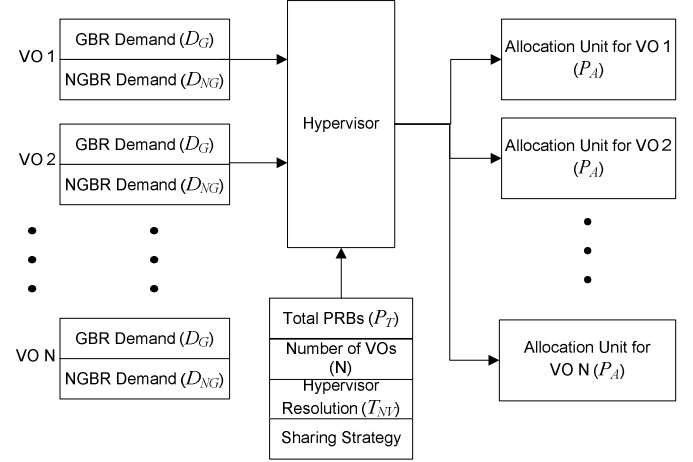


Figure 2. Hypervisor overview

TABLE I. NOTATIONS I

Parameter	Descriptions
P_T	Total number of PRBs per physical eNB, e.g. 50PRBs (10MHz spectrum bandwidth) shared by multiple VOs.
P_A	Number of allocated PRBs from hypervisor that the VO will get in the next hypervisor resolution
T_{NV}	Hypervisor resolution, i.e. the time interval in seconds that the hypervisor is triggered (the time between two consecutive spectrum resource reallocations)
f	GBR Margin factor, in terms of percentage which is used for the GBR demand estimation
N	Total number of VOs within the physical eNB
i	The index of VO
R_{AMBR}	Aggregate maximum bit rate (AMBR)
D_G	Spectrum demand (in PRBs) for GBR services
D_{nG}	Spectrum demand (in PRBs) for nGBR services
R_G	Estimation of PRBs usage for GBR services
R_{nG}	Estimation of PRBs usage for nGBR services
U_G	Instantaneous number of PRBs used for GBR services
$U_{G,u}$	Instantaneous number of PRBs used per user for GBR services
$U_{nG,u}$	Instantaneous number of PRBs used per user for nGBR services
r	The user's instantaneous data rate in one TTI
t	The time in term of the index of current TTI

The sharing strategy includes two options: *no sharing* and *dynamic sharing*. In case of no sharing, each VO will get a constant allocation unit all the time:

$$P_A(i) = \frac{P_T}{N} \quad (1)$$

If the dynamic scheme is in use, then the proposed dynamic spectrum allocation algorithm is enabled. Based on the following algorithms (Eq. (2) – (5)), the hypervisor will periodically allocate the available resources to VOs according to their estimated demands. The demands for GBR and nGBR traffic are estimated separately. For the GBR, an additional safety margin is added to its requirement with the purpose of guarantying the QoS. In order to estimate the nGBR demand, an AMBR which limits the total bit rate of all of the non-GBR bearers for a user is implemented and is enabled in our simulator. The reason behind the use of the AMBR limit is that the TCP based applications are greedy and try to use as much as resources which makes the estimation of the nGBR demand not

possible. The detailed algorithms of the demand estimations are explained as follows.

GBR Demand (D_G): It is estimated based on the number of PRBs requirement for GBR services during the last hypervisor resolution time.

$$D_G(i)(t) = R_G(i)(t) * f \quad (2)$$

The R_G is updated at the MAC scheduler in every TTI (1ms) based on the exponential moving average (EMA) formula (the unit of T_{NV} is in terms of TTIs), Here $U_G(i)$ is the sum of PRB requirements by GBR services of all users of the VO i .

$$R_G(i)(t) = (1 - \frac{1}{T_{NV}}) * R_G(i)(t-1) + \frac{1}{T_{NV}} * U_G(i)(t) \quad (3)$$

nGBR Demand (D_{nG}): The nGBR demand for each VO is updated at the MAC scheduler in every TTI based on the EMA.

$$D_{nG}(i)(t) = (1 - \frac{1}{T_{NV}}) * D_{nG}(i)(t-1) + \frac{1}{T_{NV}} * R_{nG}(i)(t) \quad (4)$$

The R_{nG} is the sum of nGBR users' requirements within a VO. For each user, the requirement equals $U_{nG,u}$ plus one deviation which is updated based on r , and R_{AMBR} . If the user's instantaneous data rate r is bigger than R_{AMBR} , a negative deviation is applied, otherwise a positive one. The purpose is to estimate the demand for nGBR users based on the AMBR.

The allocation unit (P_A) is calculated based on the Eq. (5). The hypervisor firstly considers allocating P_A based on the GBR demand. If the GBR demand (D_G) in one eNB is larger than the total available resources (P_T), then P_A for each VO is proportional to its GBR demand over total GBR demand. If not, each VO will get its GBR demand. Besides, the rest PRBs will be allocated to VOs proportional to their nGBR demands (D_{nG}).

$$\begin{aligned} \text{if } \sum_{i=1}^N D_G(i)(t) \geq P_T \quad P_A(i) &= \frac{D_G(i)(t)}{\sum_{i=1}^N D_G(i)(t)} * P_T \\ \text{else} \quad P_A(i) &= D_G(i)(t) + \frac{D_{nG}(i)(t)}{\sum_{i=1}^N D_{nG}(i)(t)} * (P_T - \sum_{i=1}^N D_G(i)(t)) \end{aligned} \quad (5)$$

2. Algorithm for Load Balancing

The eNB virtualization aims to distribute the eNB spectrum resources to different VOs so as to efficiently utilize the resources of the eNB. In contrast, load balancing aims to balance the overall spectrum resources utilization among multiple eNBs belonging to the same VO. Each VO executes the LB independently of the others VOs. The notations used for LB are listed in Table II.

TABLE II: NOTATIONS II

Parameter	Descriptions
k	The index of eNB
T_{LB}	LB resolution, i.e. the time interval in seconds that the LB is triggered
M	Total number of eNBs
P_{LB}	The total available PRBs for the VO in one eNB, e.g. 25 PRBs
P_{offset}	Safety margin for LB with the default setting 3PRBs to avoid the Ping-Pong effect
D_T	Total demand for GBR and nGBR (same algorithm with LB, but using T_{LB} instead of T_{NV} in formula 2 to 4)
P_s	The spare resource in the eNB = $P_{LB} - D_T$ (can be negative in case of congestion)
I_s	The index of source eNB for handover
I_t	The index of target eNB for handover

Similar to virtualization, the LB is also periodically triggered (T_{LB}). When LB is triggered, the following steps are performed:

1) Decide whether the LB condition is fulfilled. Only if there is an eNB that is overloaded (Eq. (6)) and an eNB that has enough spare resources (Eq. (7)), one user can be selected for handover.

$$\forall k \text{ in } M, \text{ if any } P_s(k) < 0 \quad (6)$$

$$\forall k \text{ in } M, \text{ if any } P_s(k) > P_{offset} \quad (7)$$

2) Find the user for handover. One user will be selected from the highest congested eNB (I_s , with smallest P_s). The user with the lowest priority QCI class (based on Table III, MAC priority mapping) and lowest throughput (r) is selected, e.g. the cell edge FTP user.

3) The eNB with most spare resource (P_s) is selected as the target eNB (I_t) for the handover.

IV. SCENARIOS AND SIMULATION RESULTS

In order to compare the impact and potential gains of NV and LB, a set of simulations are performed. The simulation scenarios include GBR (VoIP, Video) and nGBR (HTTP, FTP) traffic. The effect of NV as well as LB on individual applications is analyzed. Some NV and LB parameters, e.g. NV and LB interval (resolution), GBR safety margin, are deeply analyzed. The detailed scenario settings are shown in Table III. The traffic is set in a way that all VOs have the same amount of average GBR traffic during the simulation time. Nevertheless, the two VOs within the same eNB alternatively have additional Non-GBR traffic. The purpose of applying this traffic pattern is to evaluate how well NV and LB algorithms can track and response to the load variation.

TABLE III: SCENARIO SETTINGS

Parameter	Settings
Number of eNBs/cells	2 eNBs, 1 cell per eNB with 375m radius circular cell
Number of VOs per eNB	2 Virtual operators
GBR Margin Factor	40% (default)
Number of PRB	25 PRBs (5MHz) per VO per eNB
Mobility Model	Random Way Point (RWP), Vehicular A (120 Km/h)
Channel Model	Path loss: $128.1 + 37.6 \log_{10}(R)$, R in km [10] Slow fading: Correlated Log normal, zero mean, 8db std. and 50 m correlation distance Fast fading: Jake's like model
Number of users per VO per eNB	15 VoIP users, 5 Video users, 5 HTTP users, 5 FTP users
MAC scheduler	GBR with strict priority nGBR with proportional fair
MAC priority mapping	GBR: VoIP to QCI 1; Video to QCI 5 nGBR: HTTP to QCI 8, FTP to QCI 9
AMBR	3.6Mbps
VoIP traffic model	Encoder Scheme: G. 711 (64kbps) Talk period /Silence period: exponential (3s)
Video traffic model	24 Frames/sec, frame size: 1562 bytes (300kbps)
HTTP traffic model	Page size: constant 1MByte Inter-arrival time: exponential (10s)
FTP traffic model	File size: constant 3MByte Inter-arrival time: exponential (20s)
Traffic Activity	GBR (VoIP, real-time Video) Continuous call throughout the simulation: nGBR (HTTP, FTP) VO1eNB1, VO2eNB2: 100-300s and 500-700s VO2eNB1, VO1eNB2: 300-500s and 700-900s
Simulation Time	1000s (each run with 5 seeds)

1. eNB Virtualization/Sharing Case

In this case, the eNB virtualization function (which is based on the sharing algorithm proposed in section III) is enabled while

the LB function is switched off. This means, in each eNB the two VOs can dynamically share the available resources.

Figure 3 shows the spectrum resource demand and allocation in number of PRBs for the two VOs within the same eNB (with hypervisor resolution of 10s as an example). It can be seen, that when the VO has nGBR traffic, the demand is much higher than the other VO. Consequently, the hypervisor will allocate more PRBs to the VO with nGBR traffic. Similarly, when the VO only has GBR traffic, it needs less resource and thus the PRB allocations from the hypervisor decreases correspondingly. However, the PRB allocations are still larger than the demand (around 40%) according to the given GBR margin factor to guarantee the GBR performance. Compared to the case of no sharing, where each VO gets a constant 25 PRBs, the proposed eNB virtualization method can utilize the PRBs more efficiently in a more dynamic way.

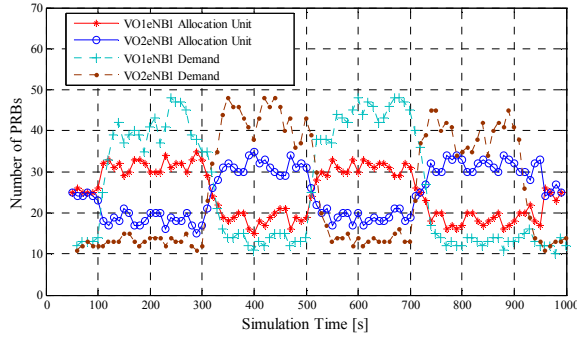


Figure 3: PRBs Demand and Allocation

Figure 4 to 6 compare the system behavior and application performance under different sharing strategies and hypervisor resolutions. The error bars show the 95% confidence interval of the simulation results. Figure 4 shows the PRB resource utilization ratio per VO per eNB during the whole simulation time for different sharing strategies. It can be seen, that more PRBs are used with sharing than without sharing. The figure reveals that with smaller hypervisor resolution, more PRB resources can be used, because the hypervisor can track and response the VO's demand more quickly. On the other hand, with larger hypervisor resolution, the efficiency decreases towards the no sharing case.

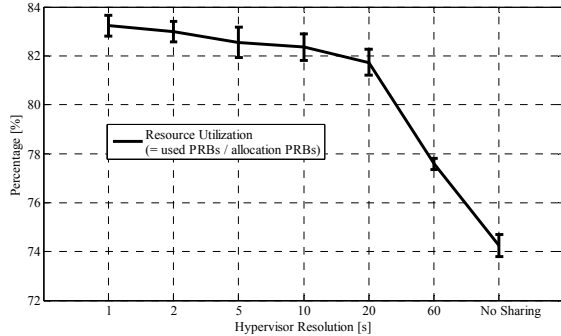


Figure 4: Resource Utilization

Figure 5 shows the average end-to-end delay (without coding/compressing delay) of VoIP and real-time video applications. For both VoIP and video, the no sharing scheme has the minimum delay. This is because with sharing, the allocated PRBs for one VO may be lower than 20PRBs (see figure 3) comparing to the constant 25 PRBs in case of no sharing. This means, the VO gets less available resources and

thus slightly degrades the end-to-end packet delay performance. With a very high resolution value (e.g. 60s), the mean GBR demand estimation is more accurate. And with a very low resolution value (e.g. 1s), GBR demand can be tracked and responded very quickly. Nevertheless, the delay difference is not so significant (as the maximum delay degradation is less than 5ms). The difference for VoIP traffic is even smaller due to its highest priority set in the MAC scheduler. For all settings, the same Mean Opinion Score (MOS) of 3.78 for VoIP is achieved. The MOS is often used to evaluate the perceived quality (QoE) for the VoIP users. In our model it is estimated using the E-Model [11].

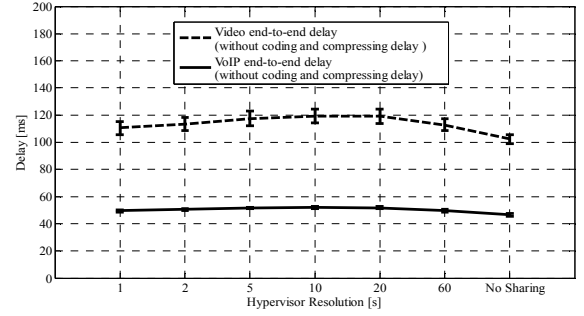


Figure 5: Video and VoIP delay (s)

The HTTP page download time and FTP file download time over different hypervisor resolutions are shown in figure 6. In contrast to the VoIP and video, the nGBR (i.e. HTTP and FTP) applications get significant benefit from the sharing. The gain for the HTTP traffic is not as obvious as the FTP because HTTP traffic has higher scheduling priority over FTP and thus there are relatively enough resources for HTTP in all scenarios. By comparing different hypervisor resolutions, we can observe, that the gain is obvious when the resolution is reduced from 60s to 10s. When the resolution is further reduced, a slightly better performance can be obtained, but of course more processing, computations and signaling is required at the hypervisor. As a result, 10s is a good trade-off between performance and computations/signaling.

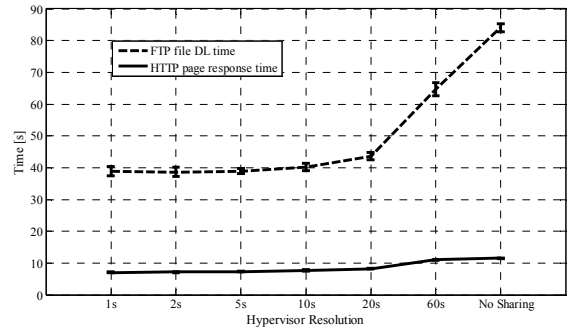


Figure 6: HTTP/FTP download time (s) over hypervisor resolution

Figure 7 and figure 8 compare the application delay performance under GBR margin factor (f) with a NV hypervisor resolution of 10s. A higher f implies a larger safety margin preserved for GBR traffic and thus less resource are left for sharing. It can clearly be seen, that with smaller f factor the end-to-end delay of the GBR traffic increases. However, the nGBR performance can be improved since there are more spectrum resources available for sharing.

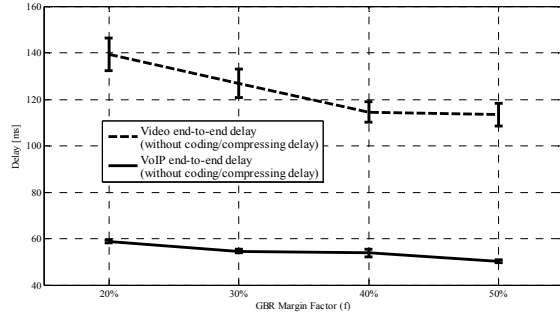


Figure 7. Video and VoIP delay (s) over GBR Margin factor

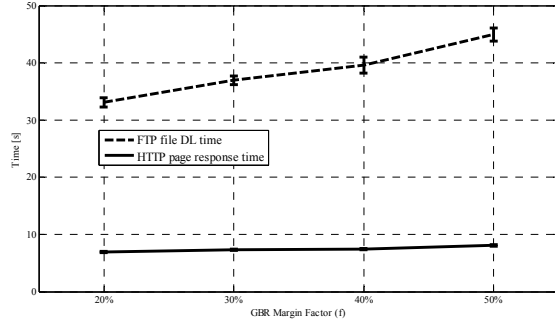


Figure 8: HTTP/FTP download time (s) over GBR Margin factor

2. Load Balancing Case

In order to compare the eNB virtualization with the load balancing, another set of simulations has been performed enabling the LB functionality, but disabling the virtualization (no sharing). From these simulations, it is found, that LB scheme has no impact on the GBR services as they have higher priority over nGBR and there is a constant amount of PRBs available at each eNB (25 per VO per eNB).

Figure 9 shows a comparison between NV and LB of different LB resolutions. It can be seen, that with LB, a better performance can be achieved for nGBR services as compared to the scenario with no LB. This is because when one eNB is congested (overloaded), some of its users will be forced to handover to the neighboring eNB having spare resources. In this way, the load can be reduced in the overloaded cell and the average user performance be improved. Similar to NV, with smaller LB resolution, a higher gain can be achieved. Nevertheless, the LB resolution is not recommended to be set to low values in the order of hundred milliseconds, since the path switch of handover will cause temporary service interruption (~30ms [12]) for the user. The Ping-Pong effect needs to be avoided. Figure 9 also shows the results for the case of NV but without LB, it can be seen that the proposed NV scheme performs much better than LB in the given example. The main reason is that with NV the available PRBs can go beyond 25 per eNB per VO introducing a larger multiplexing gain for nGBR users than in the LB case. In addition, NV will not interrupt the transmission. But with LB a handover might cause packet losses at the RLC and so forcing TCP into the fast retransmit or the slow start phase. Additionally, the PDCP buffers need to be forwarded to the target eNB.

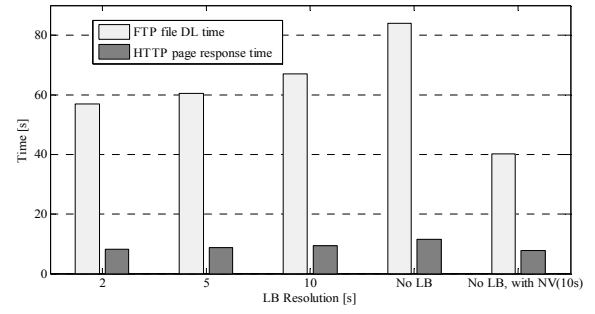


Figure 9. HTTP/FTP download time (s) over LB resolution

V. CONCLUSION AND FUTURE WORK

The paper proposes a spectrum sharing scheme for LTE eNB virtualization with multi-eNBs and multi-VOs system. In this work extensive simulations are performed to evaluate the improved efficiency of the spectrum resource utilization and the end user performance. With eNB virtualization, the resources can be utilized in a more efficient manner by multiple VO, and thus improving the system performance, especially for the nGBR services.

Furthermore, we investigate the optimal parameters settings for the NV, such as sharing intervals and safety margins. From the investigations, an optimum setting can be found as a trade-off between the user performance and the processing effort as well as signaling overhead. In addition, we also present a dynamic load balancing scheme that can lead to a significant gain of the user performance by offloading the excessive traffic from high loaded eNBs to low loaded ones. Due to the individual benefits of both NV and LB, we can consider applying both schemes in the system as a hybrid approach to further improve the overall efficiency. This coordinated NV and LB mechanism is part of our future work.

REFERENCES

- [1] J. Sachs and S. Baucke. Virtual Radio-A Framework for Configurable Radio Networks. In WICON'08, Hawaii, USA, Nov. 2008.
- [2] S. Paul and S. Seshan. GENI Technical Document on Wireless Virtualization. <http://groups.geni.net>, September 2006.
- [3] AKARI architecture conceptual design for new generation network. <http://akari-project.nict.go.jp>, 2008.
- [4] S. Baucke and C. Görg. Virtualization as a Co-existence Tool in a Future Internet. In ICT Mobile Summit - 4WARD Workshop, Stockholm, Sweden, June 2008.
- [5] L. Zhao, M. Li, Y. N. Zaki, A. Timm-Giel and C. Görg; "LTE virtualization: From theoretical gain to practical solution," International Teletraffic Congress (ITC), 2011 23rd International, vol., no., pp.71-78, 6-9 Sept. 2011
- [6] Y. N. Zaki, L. Zhao, C. Görg and A. Timm-Giel, "LTE wireless virtualization and spectrum management," Wireless and Mobile Networking Conference (WMNC), pp.1-6, 13-15 Oct. 2010
- [7] Y. N. Zaki, L. Zhao, C. Görg and A. Timm-Giel, "LTE Mobile Network Virtualization - Exploiting Multiplexing and Multi-User Diversity Gain," Mobile Networks and Applications, pp. 1-9, 2011
- [8] 3GPP TS 23.401 V9.2.0 (2009-09) GPRS enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access
- [9] OPNET official website, <http://www.opnet.com>
- [10] Anas, M.; Calabrese, F.D.; Mogensen, P.E.; Rosa, C.; Pedersen, K.I.; "Performance Evaluation of Received Signal Strength Based Hard Handover for UTRAN LTE," Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th, vol., no., pp.1046-1050, 22-25 April 2007
- [11] ITU-T Rec. P.800. "Methods for subjective determination of transmission quality", Approved in August 1996
- [12] Racz, A.; Temesvary, A.; Reider, N.; "Handover Performance in 3GPP Long Term Evolution (LTE) Systems," Mobile and Wireless Communications Summit, 2007. 16th IST, pp.1-5, 1-5 July 2007