Response Time of the S1 and X2 Handover Procedures Between (H)eNBs in a Virtualized Environment

Ziyu Xiao[1,2] and Harry Perros[2]

[1]China Mobile Design Institute
Beijing, P.R. China
[2]Dept.of Computer Science,
NC State University
Raleigh, NC, USA

Abstract. We study the response time of the S1 and X2 handover procedures between (H)eNBs in a CRAN and a virtualized EPC, and compare it to that for the existing non-virtualized case.

Keywords: eNB, HeNB, EPC, CRAN, S1, X2, Virtualization, response time

1. Introduction

Home eNBs (HeNBs) will be introduced in LTE as part of the future G5 deployment in order to provide mainly indoor coverage in homes, offices, and shopping centers. Inter-operator roaming and network interoperability handover actions will be permitted as part of the anticipated massive deployment of HeNBs,. Therefore, it is necessary that network operators can execute horizontal handovers efficiently so as to provide service continuity between (H)eNBs. In this paper, we use simulation techniques to study the response time of the S1 and X2 handover procedures between (H)eNBs in a CRAN and a virtualized EPC, and compare it to that for the existing non-virtualized case. Other performance studies of handovers can be found in: [1], [2], and [3]. The paper is organized as follows. In section 2, we discuss how the X2 and S1 handover procedures can be implemented in a virtualized EPC and CRAN. In sections 3 and 4, we present simulation results of the response time of X2 and S1 for the existing non-virtualized and future virtualized case, respectively. The conclusions are given in section 5.

2. Future Virtualization of EPC and EUTRAN

The X2 procedure is used to handover a UE from a source (H)eNB to a target (H)eNB for which the MME is unchanged. Two procedures are defined depending on whether the serving GW (S-GW) is relocated, see [4], [5]. The S1 procedure is used by the source eNB to initiate a handover through the S1-MME interface, see [4]. This procedure may relocate the MME and/or the S-GW. The MME should not be relocated during an inter-eNB handover unless the UE leaves the MME Pool Area where the UE is served. S-GW relocation takes place as determined by the MME.

It is anticipated that the 5G EPC core network will migrate to the cloud, and the user plane of the EPC core network will use SDN-based switches and routers with extended GTP-U functions. In addition, the radio access network will be also virtualized in to a Cloud Radio Access Network (CRAN), which will centralize the execution of the control functions of all base stations. Figure 1 shows the EPC cloud and CRAN with the S1 interface between EPC and CRAN, and X2 interface between CRANs or intra-CRAN.

There have been several studies as to how the EPC can be virtualized, see for instance [6] and [7]. In general, as shown in Figure 2, the control plane network elements MME, HSS, PCRF, S-GW and P-GW will run in a centralized data center in VMs. The user plane functions of the S-GW and P-GW will run on merged SDN-enabled switches, with the SDN controller running in the control plane. OpenFlow (OF) will be used in the southbound interface to control the switches and routers. S1-C/SCTP/IP will be used between CRAN and MME, and S1-U/GTP tunnels will be established between SDN-enabled switches and routers. X2-C/SCTP /IP will be used between separate CRANs interfaces with X2-U/GTP tunnels in the user plane.

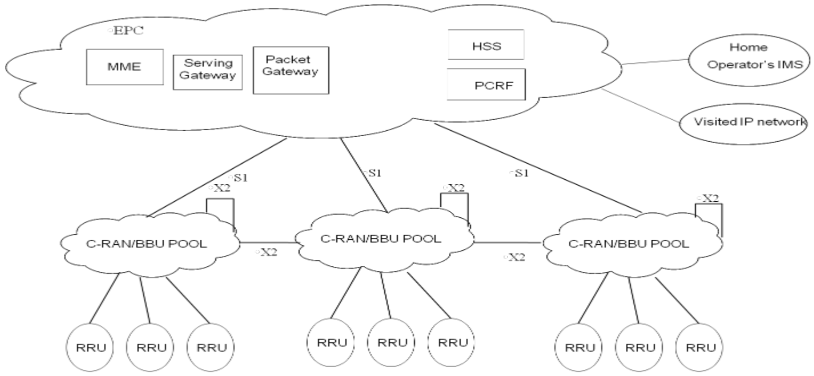

Fig.1: Virtualized EPC and EUTRAN

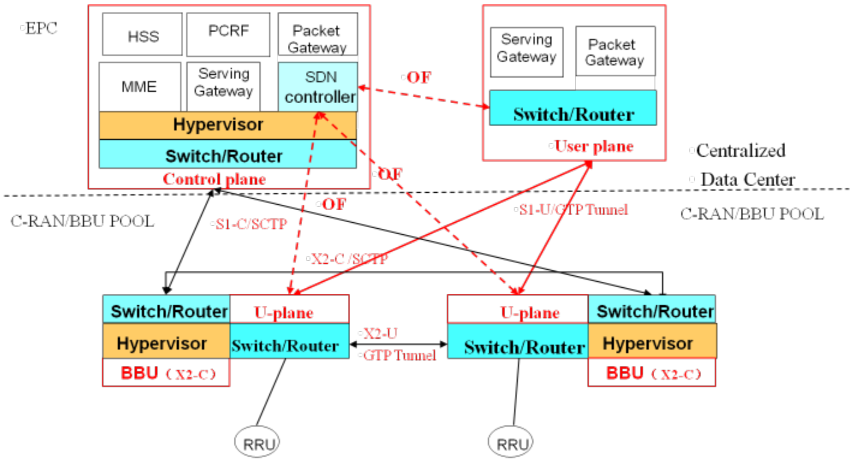

Fig. 2: The EPC cloud architecture

It is anticipated that a cloud-based EPC will serve an entire region/province, with the largest transmission distance between EPC core network and eNBs being over 1000 km. Also, several regions/provinces may share one or two data centers. Also, using CRAN technology, the control functions of eNB will be centralized in a CRAN pool, located no more than 20 km from antennas. The user plane of S-GW and P-GW run on OF-enabled switches/routers, as shown in Figure 2.

The X2-based handover signaling procedures will run almost unchanged with the exception of the messages CREATE SESSION REQUEST and RESPONSE, which should be sent from the MME to the control plane of S/P-GW running on VMs of the same cloud, see Figure 3.

In the virtualized environment, the S1 procedure will be modified as shown in Figure 4. The source HeNB sends a HANDOVER REQUIRED message to the source MME in the EPC cloud. The source MME selects the target MME with the target TAI which may or may not be in the same cloud. The target MME is virtualized in a single VM together with the control plane of S-GW, which we refer to as the MME/S-GW-C.

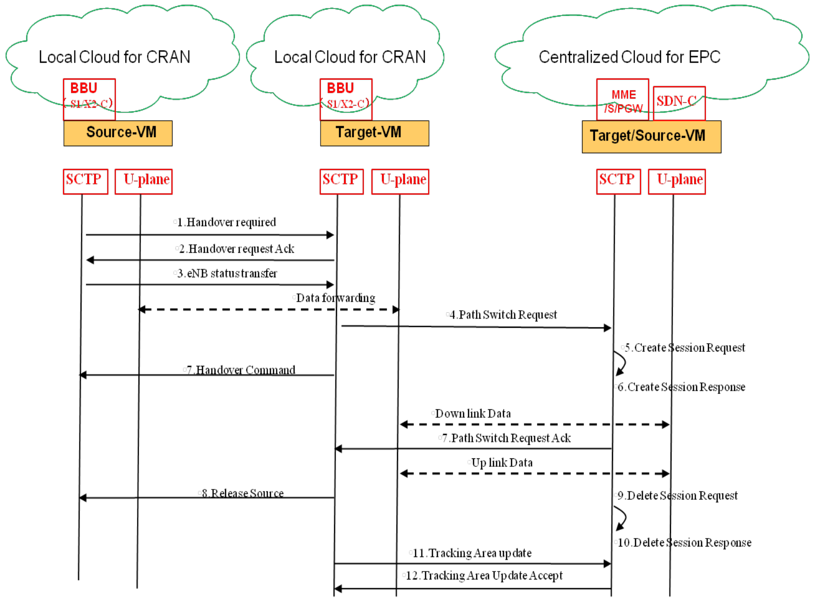

Fig. 3: X2-based handover procedures in virtualized EPC and CRAN

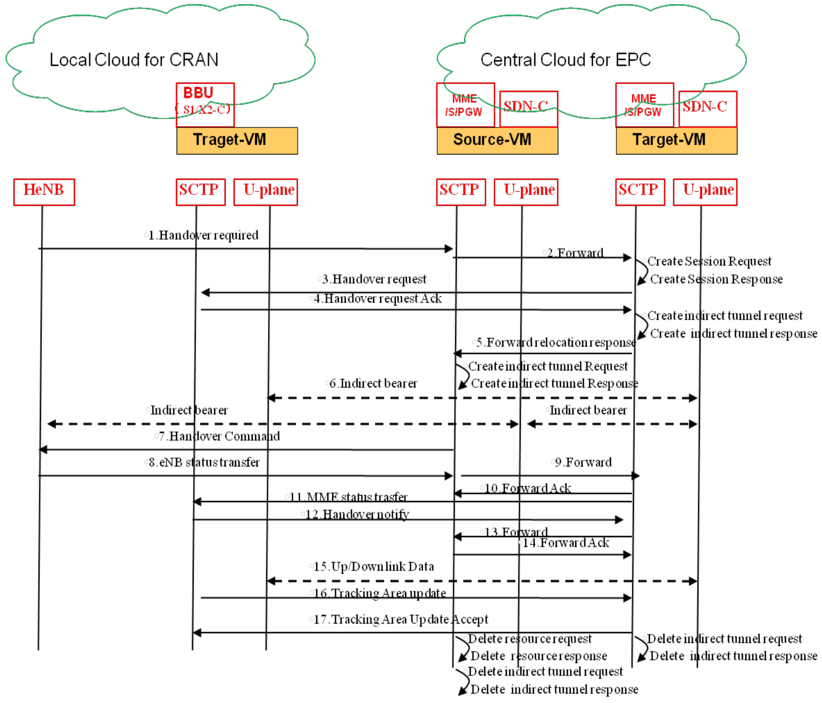

Fig. 4: The S1-based handover procedure for virtualized EPC and CRAN

In view of this, the MME does not need to send a CREATE SESSION REQUEST message to the S-GW. The target MME sends a HANDOVER REQUEST message to the target eNB in CRAN. The target eNB returns the HANDOVER REQUEST ACK message to the target MME. The target MME forwards the HANDOVER REQUEST ACK message to the source MME. The target MME/S-GW-C sets up forwarding parameters but it does not send a CREATE INDIRECT DATA FORWARDING TUNNEL REQUEST to the S-GW-C combined with MME. The source MME/S-GW-C also does not need to send a CREATE INDIRECT DATA FORWARDING TUNNEL REQUEST. The source MME sends a HANDOVER COMMAND message to the source HeNB. The source HeNB sends a STATUS TRANSFER message to the target eNB in CRAN via the MMEs. User plane connections are established through SDN. Then, the source HeNB starts forwarding downlink data to the target eNB via the source user plane and the target user plane. After the UE has successfully synchronized and accessed the target cell, downlink packets forwarded from the source HeNB can be sent to the UE. Also, uplink packets can be sent from the UE via the target eNB, and forwarded to the target user plane. The target eNB in CRAN sends a HANDOVER NOTIFY message to the target MME. The MME/S-GW-C sends a MODIFY BEARER REQUEST message to the P-GW's control plane. The SDN controller establishes user-plane connections. The UE initiates a TRACKING AREA UPDATE procedure since its tracking area has changed, and the source MME commands the source HeNB to release its resources related to the UE, and deletes the EPS bearer resources.

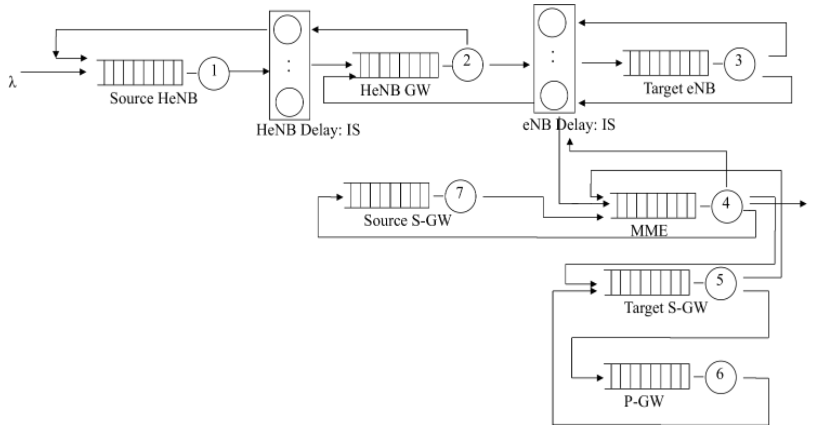

Fig. 5: The queueing model for X2-based handovers

Table 1. Service times (msec) for both X2 and S1 queueing models

Server's name	trgt/src MME	src/trgt S-GW	src/trgt (H)eNB	HeNB GW	P-GW	HSS
Source/target MME	50	50	50	50		100
Source/target S-GW	50				50	
Source/target (H)eNB	50		20	20/50*		
HeNB GW	50		20/50*			
P-GW		50				
HSS	100					

*For the X2 interface, source/target (H)eNB to target/source (H)eNB via HeNBGW is 20 msec. For the S1 interface, HeNB to MME via HeNBGW or MME to HeNB via HeNBGW is 50 ms.

3. The Response Time of X2 and S1: Non-virtualized Case

The flow of messages for the X2 procedure is modelled by the queueing network in Figure 5, which was analyzed by simulation using JMT [8]. (Due to the assumptions made, it does not admit a closed-form solution. The same applies to the remaining queueing models developed in this paper.) We consider the case where a UE is in connected status and it handovers from an HeNB to a macro target eNB via the X2 interface. We assume a large number of HeNBs, which are modelled indirectly through the arrival rate λ of handover requests. Specifically, it models the following seven servers: source HeNB, source HeNB GW, target eNB, source S-GW, MME, target S-GW, and P-GW. In addition, an infinite server (IS) queue is used to model the propagation delay between the source HeNB and the (source) HeNB GW, and between the source HeNB GW and the target eNB. The service times are exponentially distributed with an interface-dependent mean given in Table 1.

We assume that the maximum distance from an (H)eNB to the data center is no more than 1000 km, which give us an upper bound of 6.86 msec propagation delay from an HeNB to its HeNB GW. The eNBs are deployed one per square km, and they are connected to the operator's data center via the operator's IP backbone and edge network. The distance from a macro-eNB to the data center is no more than 1000 km. We assume an upper bound of 5.36 msec propagation delay from an eNB to MME.

The arrival of requests for handovers is assumed to be Poisson distributed with an arrival rate calculated as follows. We assume a scalable EPC network with a capacity varying from 1 to 5 million customers, of which 1% is in connected status. Of these connected status customers, 3% will request a handover between eNBs. This results to an arrival rate varying from 0.83 requests/sec to 4.17 requests/sec

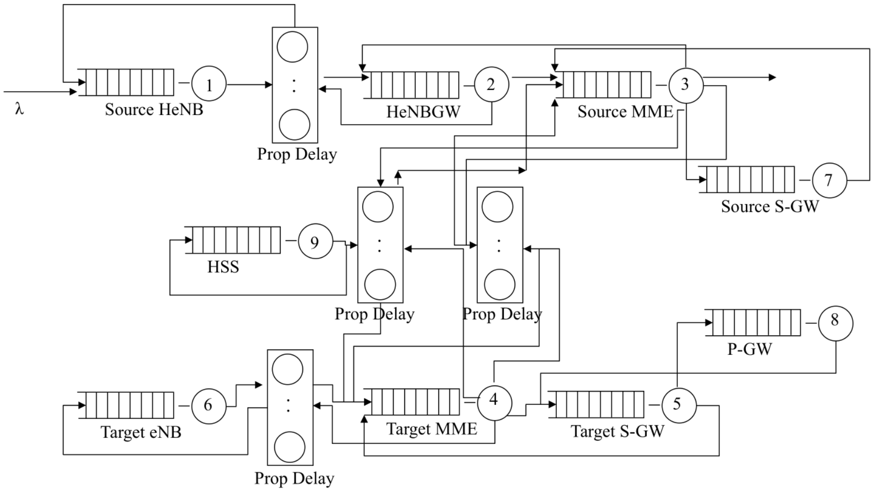

Fig. 6. The queueing model for S1-based handovers

The flow for an S1 procedure is depicted by the queueing network in Figure 6, which was analyzed by simulation using JMT for the same arrival rates. We define 9 servers, namely, source HeNB, (source)HeNB GW, source MME, target MME, target S-GW, target eNB, source S-GW, P-GW, and HSS. We simulate the scenario where an UE in connected status handovers from a HeNB to a macro target eNB via the S1 interface. We also introduce 4 delay nodes to depict the propagation delay between HeNB and HeNB GW, target eNB and EPC, source MME and target MME in separate operators, and MME and HSS. Since the source MME and the target MME belong to separate operators, the connections between them will go through different IP backbone and edge networks. As in the previous model, the service times are exponentially distributed with an interface-dependent mean given in Table 1.

The propagation delays are the same as in the X2-based model, that is, 6.86 msec from HeNB to the HeNB GW, and 5.36 msec from the target eNB to MME. We assume that the source MME is located no more than 2000 km from the target MME of the other operator's data center, via the operator's service gateway, an IP backbone and an edge network. Therefore, the upper bound for the propagation delay between the source MME and the target MME is 108.72 msec. Also, we assume that the HSS is located no more than 3000 km from the source and target MME of the different operator's data center via the operator's service gateway, an IP backbone and an edge network between operators. So, an upper bound on the propagation delay between the source MME and the target MME is 112.02 msec.

We first note that for the considered range of arrival rates, i.e., 0.83 to 4.17 requests/sec, the queueing networks become stable when each modelled entity consists of three servers is parallel. Therefore, all the single queues in the model are served by three servers, rather than one as depicted in the diagrams.

Figure 7 gives the mean and 95% percentile of the response time, which is the total time it takes for the system to complete a handover, as a function of the arrival rate λ of handover requests. For S1 handovers, the mean response time varies from 3.741 sec to 12.132 sec, and the 95th percentile from 4.646 sec to 28.580 sec. The last data point of 12.132 sec for the mean and 28.580 sec for the 95th percentile occurs when the arrival rate is 4.17 requests/sec, for which the utilization of the target MME is 97.5%, exceeding the typical threshold of 80% utilization for stable operations. For X2 handovers, the mean response time varies from 0.926 to 0.952 sec and the 95th percentile varies from 1.308 to 1.347 sec. An X2 handover is performed in less than 1 second on the average with a 95th percentile of less than 1.347 sec, as opposed to an S1 handover that takes at least 4 times more. This is due to the fact that the S1 procedure is a lot more complex than the X2 procedure.

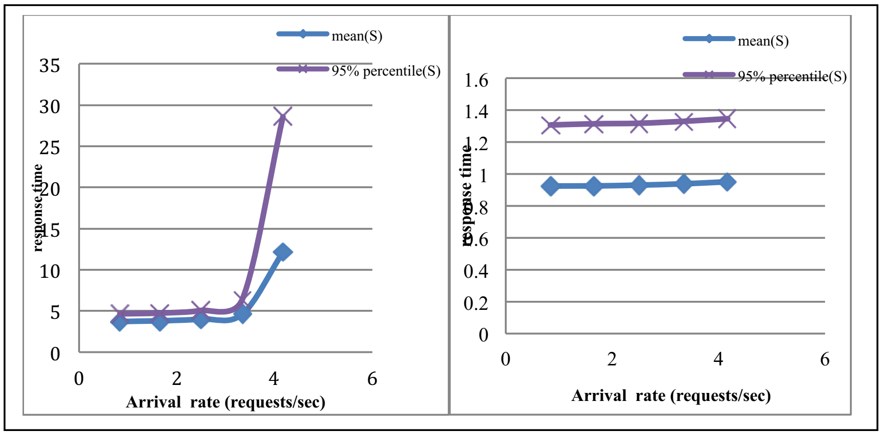

Fig. 7. Response time as a function of handover requests λ. (S1 left figure, X2 right figure)

We note that in both procedures the utilization of the MME is significantly higher than the rest of the entities, and needs a CPU capacity of at least 30% more than the remaining entities in order to avoid being the bottleneck. (Please note that we do not provide supporting graphs due to lack of space.)

4. The Response Time of X2 and S1: Virtualized Case

The total propagation for the execution of an X2-based handover may or may not increase in a virtualized environment, depending upon the location of the eNBs and the CRAN. If we plan RRUs together with the same CRAN, most handovers will be intra-CRAN instead of inter-CRAN, and the propagation delay will not be increased. If we deploy an HeNBGW in CRAN, rather than in the centralized EPC cloud, then the propagation delay may increase a little, but no more than 16 msec. So, overall, we do not anticipate a change in the response time for the X2 procedure.

In order to study the response time of the S1 procedure we modified the simulation model developed for the non-virtualized case to reflect the changes described in section 2. The traffic arrival assumptions are the same as before. The service times are exponentially distributed with an interface-dependent mean given in Table 2. We assumed that the service time of the MME/S-GW-C is the same service time as in the non-virtualized MME because of anticipated future hardware and software upgrades.

Table 2. Service times (msec) for virtualized S1-based handover procedure

Server's name	MME/S-GW-C	(H)eNB/CRAN	HeNBGW	P-GW-C	HSS
MME/S-GW-C	50	50	50	50	100
HeNB/CRAN	50		50		
HeNBGW	50	50			
P-GW-C	50				
HSS	100				

We use upper bound values for the propagation delay as in the non-virtualized case. A maximum distance in a non-CRAN scenario is 1000 km since there is one eNB deployed per km. If a CRAN is located per 40 km, then the maximum distance is 1000-40 = 960 km. Since the difference in the propagation delay for 1000 km and 960 km is only 0.13 msec, the same propagation delay is used as before. The arrival of requests for handovers is assumed to be the same as in the non-virtualized case, ranging from 0.83 to 4.17 requests/sec.

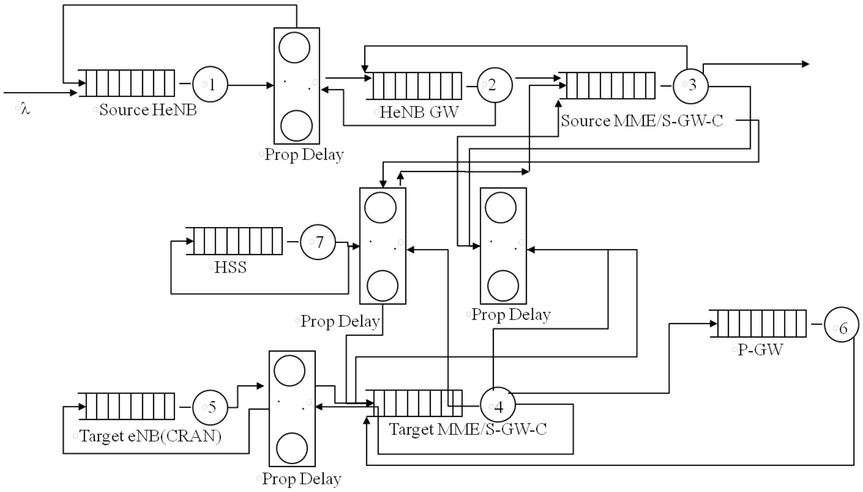

Fig. 8: The queueing model of the virtualized S1-based handover procedure

The flow of messages in the virtualized S1 procedure is depicted by the queueing network in Figure 8. As before, it was analyzed by simulation using JMT. We define 7 servers, namely, source HeNB, (source) HeNB GW, source MME/S-GW-C, target MME/S-GW-C, target eNB(CRAN), P-GW, and HSS. We simulate the scenario where a UE in connected status handovers from an HeNB to a target eNB located in CRAN via the S1 interface. As in the non-virtualized case, we also introduce 4 delay nodes to depict the propagation delay between source HeNB and (source) HeNB GW, target eNB (CRAN) and EPC, source and target MMEs located in different operators, and MME and HSS. As before, all the single queues in the queueing model are served by three servers, rather than one as depicted in the diagram.

Figure 9 gives the mean and 95% percentile of the system response time (total time it takes to complete a handover) in the virtualized case-as a function of the arrival rate of handover requests. The mean response time varies from 3.057 sec to 3.835 sec, vs. 3.741 sec to 12.132 sec in the non-virtualized case. Likewise, the 95th percentile varies from 3.906 sec to 5.447 sec vs. 4.646 sec to 28.580 in the non-virtualized case. In conclusion, the response time after virtualization is less than that in the non-virtualized case.

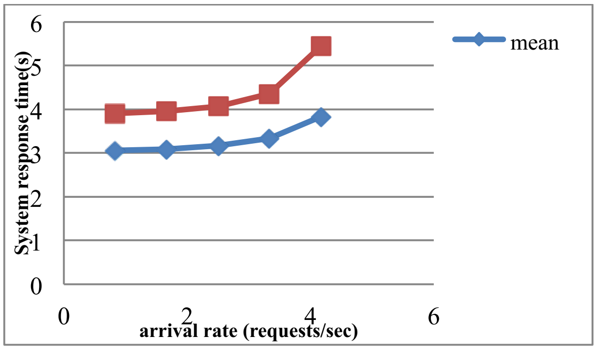

Fig. 9: Response time of S1 procedure as function of the handover requests λ

We note that the utilization of the target MME/S-GW-C is the highest varying from 16.1% to 79.5%, significantly lower than the target MME in the non-virtualized case which ranged from 19.3% to 97.5%, followed by the source MME/S-GW-C, whose utilization varied from 11.2% to 55.3% vs. the non-virtualized source MME that ranged from 15.1% to 76.1%. (No graphs are provided due to lack of space.) This is partially due to the fact that the MME/S-GW-C has a service time of 50 ms compared to the total service time of 100 ms for the MME and S-GW in the non-virtualized case. In addition, by combining the MME with the S-GW-C, we eliminate all inter-communication delays.

5. Conclusions

It appears that virtualization can decrease the response time of the complex S1 procedure and reduce server utilization, but there is no significant benefit for the X2 procedure since it does not need centralized EPC control. The virtualization of the EPC has not as yet been solidified and more performance evaluation studies of the response time of the two handover procedures are required.

References

1. A. Bajzik, P.Horvath, L. Korossy, C. Vulkan, Impact of Intra-LTE Handover with Forwarding on the User Connections; 16th IST, (2007).
2. K. Dimou, M. Wang, Y. Yang, M. Kazmi, Handover within 3GPP LTE: Design Principles and Performance; Vehicular Technology Conference Fall (2009).
3. A. Racz, A. Temesvary, N. Reider, Handover Performance in 3GPP Long Term Evolution (LTE) Systems; Mobile and Wireless Communications Summit, 16th IST, pp 1-5 (2007).
4. 3GPP TS 23.401 General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access V13.1.0 (2014-12)
5. 3GPP TS 36.300 V12.3.0 Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2 (Release 12).
6. Kempf, J., Johansson, B., Pettersson, S., Luning, H., Moving the Mobile Evolved Packet Core to the Cloud; Wireless and Mobile Computing, WiMob, 2012 IEEE.
7. Hoffmann, M., Staufer, M,Network Virtualization for Future Mobile Networks: General Architecture and Applications; Communications Workshops (ICC), 2011 IEEE.
8. M. Beetoli, G. Casale, and G. Serazzi, JMT: Performance Engineering Tools for System Modeling; ACM SIGMETRICS Performance Evaluation Review, 36 (4), pp.10-15 (2009).