Efficient Mobility and Traffic Management for Delay Tolerant Cloud Data in 5G Networks

Athul Prasad, Petteri Lundén, Martti Moisio, Mikko A. Uusitalo, and Zexian Li
Radio Systems Research, Nokia Networks, P.O. Box 226, FI-00045 Nokia Group, Finland.
Email: {Athul.Prasad, Petteri.Lunden, Martti.Moisio, Mikko.A.Uusitalo, Zexian.Li}@nokia.com

Abstract—**The explosive growth of the demand for higher data rates in mobile networks have been mainly driven by the increasing use of cloud based applications by smartphones. This has led the industry to investigate new radio access technologies to be deployed as part of 5G networks, while providing mechanisms to manage user mobility and traffic in a more efficient manner. In this paper, we consider a mobility and traffic management mechanism that proposes a close interaction between the cloud data servers and the radio access network to enable efficient network operation. Such a management mechanism is enabled by utilizing the application-dependent delay tolerance properties of the cloud data, with the delay values conveyed to the radio access network and UE to manage the service requests for the cloud data. The mechanism was evaluated using LTE-Advanced heterogeneous network scenario and 5G dense-urban information society scenario from EU FP7 METIS project, and relative gains in terms of packet delays and throughput values are presented. The results indicate significant gains using the proposed management mechanism as compared to the reference case where no such enhancements are used.**

I. INTRODUCTION

Currently deployed radio access networks (RAN) are experiencing a significant increase in data rate and capacity demands, mainly due to mobile devices using data hungry applications, which are mostly cloud-based [1]. For such applications, even when the mobile devices are in close physical proximity, when they communicate for sharing information such as pictures or videos, the data essentially reaches the destination device after traversing through the radio access network, core network, to the cloud data server and back. Such communication paradigms cause significant overload on the radio access networks and mobile devices, which can be optimized. Such optimizations have been mainly studied as part of mobile cloud computing (MCC), which considers the integration of cloud computing with the mobile environment [2]. Some of the obstacles considered for MCC includes challenges related to performance, environment and security, a detailed survey of which is presented in [2]. The security aspects of such MCC systems, especially related to ensuring the privacy of user data in the cloud was studied in [3].

Mobile edge computing is another related topic that is currently receiving significant attention, with the main aim of reducing network latency and complexity by deploying mobile applications at the edge of the network [4], [5]. Some of the main characteristics of MEC include having the edge located on the premises where network is typically deployed, with close proximity to the source of information, enabling lower latency, with location awareness and enhanced use of network context information [5]. The work done in [6] analyzes the application of cloud computing model in LTE based radio access networks. Three major cloud computing platforms – OpenStack, Eucalyptus and OpenNebula, which has currently been deployed were considered for evaluation, and the analysis done the work indicated that none of the platforms could satisfy the requirements without further enhancements and extensions.

The application of cloud based computing paradigm is currently having an ever expanding scope, with use cases such as cloud-based vehicular networks [7] and radio access networks [8], [9] being investigated. In [8], the concept of RAN-as-a-Service is introduced, with a flexible architecture with centralized processing and increased interference handling capabilities in ultra-dense networks. It is proposed that such networks can reduce the network energy consumption, and improve the deployment cost-efficiency and network management capabilities. The work done in [9] describes how cloud technologies and flexible functionality assignment in RAN can enable the densification of networks, and centralized network operation using heterogeneous backhaul networks. The deployment of 5G cloud-based RAN using a realistic non-ideal backhaul was studied in [10], and an opportunistic hybrid automatic repeat request (HARQ) mechanism was proposed which was shown to perform close to the optimal HARQ scheme. The cloud-based RAN is an assumption used in this work as well, with gains possible by having close interaction between the RAN and cloud-data server studied.

Some of the most fundamental challenges faced by 5G networks, including generic scenarios and test cases were studied in [11]. For the 5G related evaluations done in this work, we have considered the scenario settings defined in [11], with simulation guidelines following [12]. A similar context-aware cloud connectivity management mechanism was studied in [13], with the main focus on user speed, small cell proximity, traffic, UE battery conditions and user preferences. A lightweight mobile cloud offloading architecture (MOCA) is proposed in [14], which considers the use of an in-network cloud platform for efficient traffic offloading. In this work, we consider the traffic and mobility management based on signaling interactions between the cloud data server, unlike [13] which considers the use of context information obtained using local sensor nodes or other access points. We also consider the use of the cloud server and RAN determining the delay

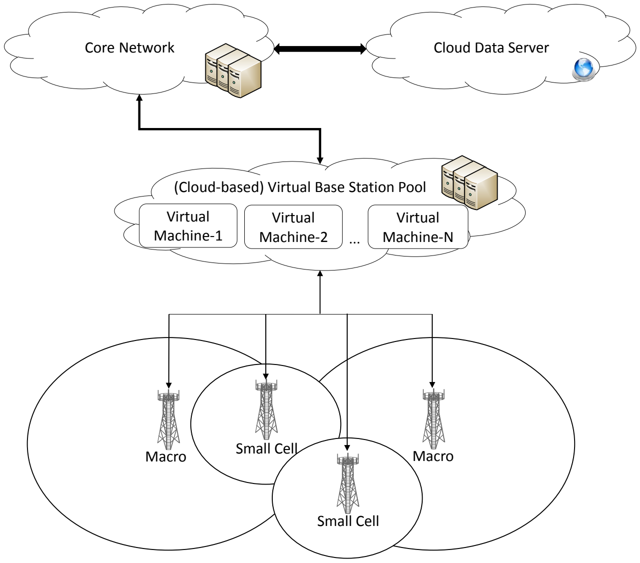


Fig. 1. Considered network architecture, based on [10].

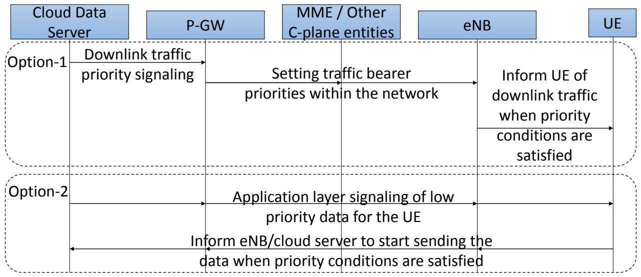
and offloading metrics of the traffic and providing assistance information to the UE, rather than based on UE autonomous decision-making. Also, compared to [14], the architecture considered in this work does not propose any modification within the RAN or core network within the 3GPP system.

In this work, we consider a MCC paradigm, where the network performance of the high-priority users is improved, making use of the heterogeneity of the environment, at the cost of delay tolerant, cloud server data which is assumed to have relatively lower priority. The work done in this paper considers a mechanism that could function within the framework of MEC, where the cloud server and radio access network interacts closely, to enable the provisioning of better Quality of Service (QoS) to the users. The mechanism proposed in this work considers the delaying of delay-tolerant data from being delivered in the radio access network, until the delay conditions expire, or when the UE connects to an appropriate cell which could be used for delivering such data. This enables the network to have better control over the radio resource provisioning in the network for higher priority data, without locally buffering the delay tolerant data. A brief outline of the work is also presented in [15], along with some of the results presented in this paper related to the METIS 5G scenario.

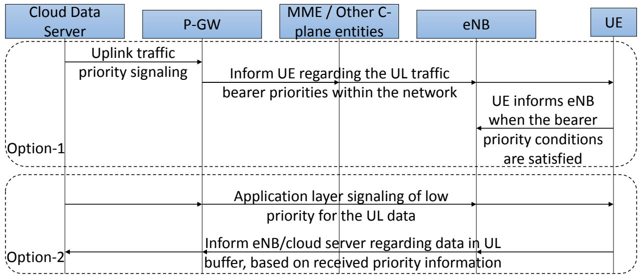
The rest of the paper is structured as follows: Section II gives an overview of the system model used. Section III discusses the evaluated mechanism. Section IV presents the simulation assumptions and system level parameters used for simulations, together with detailed performance results of the proposed scheme. Section V gives summary of the paper.

II. SYSTEM MODEL

In this work, we consider a heterogeneous network where UEs move around randomly, connecting to the best serving cell for its communication needs. The overall network architecture considered in this work is as shown in Fig. 1, based on the 5G network architecture considered in [10], with the macro base stations considered to be remote radio heads with higher transmit power and a wider coverage footprint, as compared to the small cells. The architecture within the radio access network is similar to the 5G cloud-based virtual base station pool system, with the cloud data server having a direct interface with the core network.The mechanisms considered in this paper is essentially independent of the network architecture and could work with a distributed 5G radio access network as well.



(a) Downlink signaling flow



(b) Uplink signaling flow

Fig. 2. Uplink and Downlink signaling flows.

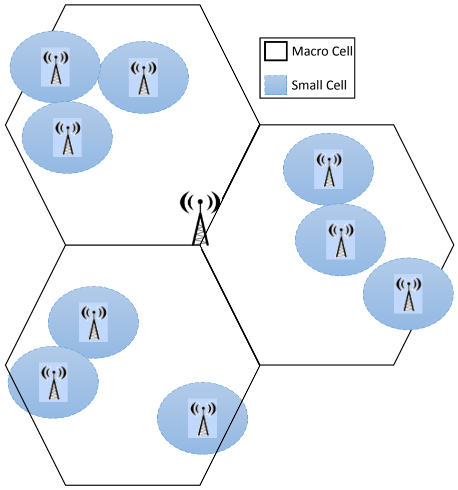
We assume the use of delay tolerant uplink and downlink data, where in uplink the cloud data server would be the target for data delivery, where as in downlink, the cloud data server would be the source of such data. We also assume that additional storage overhead required for storing cloud data server downlink data availability information could be handled in the core network without significant impacts on the network. While delay tolerance could vary from a few seconds to several hours, we consider low-priority traffic with significantly high values for this parameter in this paper, for applications such as non-essential data synchronization, which would not cause any impacts on the user experience. Here we assume that the core network entities have some capabilities, which would enable them to determine the QoS parameters of the delay tolerant data. This could include deep packet inspection capabilities, which would enable them to determine the specific application the data passing through the network is intended for, or using some advanced packet marking mechanisms, based on which the core network can determine the target application type.

III. MOBILITY AND TRAFFIC MANAGEMENT MECHANISM

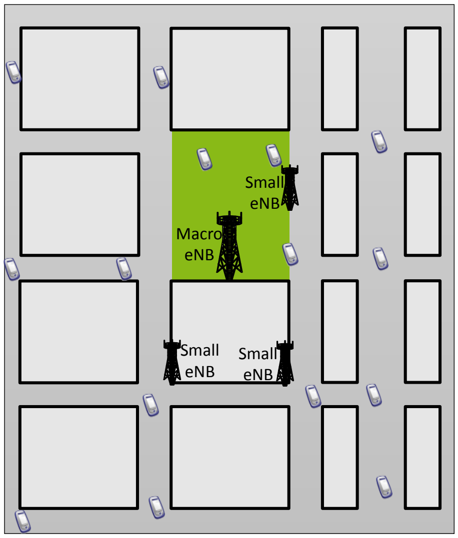
In this work, we propose an efficient mobility and traffic management mechanism for delay tolerant data, by selecting the appropriate cell for delivering data in downlink, and by configuring the UE with appropriate cell priorities for sending the data in uplink. In both uplink and downlink cases, we consider two approaches: i) assuming close inter-working between the 3GPP network and the cloud data server for delay tolerant data transfer, and ii) where the cloud data server informs the UE regarding the data delivery priorities transparent to the 3GPP network. Both approaches used in uplink and downlink is as shown in Fig. 2.

The signaling mechanisms involved in downlink is as shown in Fig. 2(a). Here Option-1 is based on the first approach, where the cloud data server informs the 3GPP radio access network, through the packet data network gateway (P-GW), mobility management entity (MME), regarding the presence of delay tolerant downlink data and the possible delay parameters, in terms of possible time scales and the appropriate serving cells suitable for data delivery. Here the key assumption is that the cloud server first indicates the availability of data to the 3GPP network, which then translates this message to derive the possible delay tolerance and other quality of service parameters. The use of appropriate serving cells could mean that the macro cell is used when it has appropriately low amount of load, or small cells which are deployed as capacity booster cells are used to minimize the load on macro cells, due to the delay tolerant, low priority data. In this case, it is also assumed that the information could be stored, for e.g. along with the context information of the UE at the MME, so that the UE can even enter idle state, until the criteria for the delivery of delay tolerant data is reached. The assumption here is also that in the first approach, the information regarding the delay tolerant data awaiting in the cloud server could also be passed on to the target eNB using the modified context information, so that when the data delivery conditions are satisfied, the data can be delivered using the appropriate cell. For Option-2, we assume that all the mechanisms described earlier is informed directly by the cloud data server to the UE, possibly using the related application.

The signaling mechanisms involved in uplink is as shown in Fig. 2(b). The both approaches considered are essentially similar to the downlink case, with the signaling occuring more proactively, rather than reactively, as was the case in downlink. For the first approach, shown as Option-1 in the figure, the core network entities such as P-GW and MME determines the uplink traffic signaling priorities, and this information is conveyed to the UE. The priority related parameters could include the cell information and the delay tolerance levels as well. When the UE satisfies all the priority conditions for the uplink traffic, the UE informs the eNB regarding the impending delay tolerant data for transmission using uplink buffer status reporting. Once the eNB receives this information, appropriate amount of resources is configured for the UE to initiate the uplink data transfer. In Option-2, the UE is informed of the delay tolerant data QoS parameters directly by the cloud data server, using application level signaling. In both approaches, the end result remains the same that uplink and downlink delay tolerant data is transmitted only when the priority conditions set are satisfied.



(a) Conventional HetNet Scenario



(b) METIS TC2 Scenario

Fig. 3. Considered scenarios.

IV. SIMULATION RESULTS

A. Simulation Assumptions

We consider both a conventional heterogeneous network (HetNet) scenario considered in 4G systems [16], [17], as well as the METIS dense urban information socity scenario (Test Case 2, TC2) [11] for evaluations in this work. Both the scenarios are as shown in Fig. 3. The main aim of considering both scenarios were to understand the essential differences in the deployments considered in 4G and 5G. The UEs are deployed randomly within the scenario, and they move in random directions until the end of the simulation world, and move towards a new random direction within the simulation world. For the 5G METIS TC2 scenario, it is further assumed that the UEs move within the outdoor regions only.

Dynamic system simulations were conducted using spatially correlated slow-fading, in order to emulate the system performance in real deployment scenarios. The detailed parameters used are as shown in Table I. The path loss models used for evaluation in the HetNet scenario follows 3GPP case 1 model defined in [18]. The path-loss is:

$$L_{\mathrm{M}} = 128.1 + 37.6 \log_{10}(R) \quad (1)$$

$$L_{\mathrm{P}} = 140.7 + 36.7 \log_{10}(R), \quad (2)$$

where L_{M} corresponds to distance dependent path-loss from macro BS to mobile terminal and L_{P} corresponds to the path loss from pico cell to the terminal. Here R [km] is the distance between the transmitter and receiver.

The parameters used for the HetNet scenario are similar to the ones defined in [16]–[18], while the METIS TC2 scenario parameters used follow [12]. The HetNet scenario was simulated with 12 picos randomly placed per macro cell, where as the METIS TC2 scenario using 3 picos in the scenario, randomly deployed along the pathways. The operating frequency used was 2 GHz for both the scenarios, with full buffer traffic model. The full buffer model was simulated by continuously generating a downlink traffic with files of 100 MB size, with a packet data unit (PDU) size 2 MB. For calculating throughput, the Shannon fitting formula considered in [19] was used with the bandwidth and SINR efficiency values as shown in Table I. The user throughput distribution is a combination of the distributions from macro and small cells. The PDU delay values indicate the amount of time taken by an user to send a PDU. The normalized values are calculated from the CDF distributions to present the relative gains of the CDTM mechanism, as compared to the baseline scheme.

The delay tolerant data is assumed to have indefinite delay capabilities, with the cell information for data delivery set as small cells. Thus the delay tolerant data is buffered at the UE or at the cloud data server, indefinitely, until the UE connects to a small cell. Various traffic ratios of 25, 50 and 75 % for the delay tolerant data and the remaining for the normal high-priority data was considered for evaluations, and the throughput and PDU delay metrics are presented. The results shown are applicable to both the approaches described in the previous section. The mechanism is abbreviated as Cloud Data Traffic Management (CDTM) with CDTM-Off indicating the baseline reference scheme with the traffic management mechanism disabled. The notation CDTM-n indicates various scenarios simulated with n % of the total traffic set to be the delay tolerant, low priority traffic.

TABLE I
SYSTEM LEVEL SIMULATION PARAMETERS

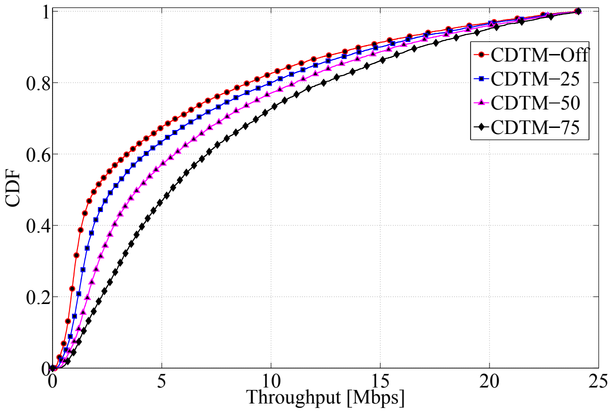
Basic Radio Configuration Parameters [17]		
Macro cell ISD	500 m	
Shadowing Standard Deviation	Macro (Pico) cell	8 (10) dB
Spectrum Allocation	10 MHz Channels Macro and Pico in separate carriers	
Macro (Pico) Max Tx Power [dBm]	46 (30)	
Antenna Gain [dB]	Macro (Pico)	15 (5)
UE Tx Power [dBm]	21	
Other Simulation Parameters		
Spectral Efficiency, S_{eff}	4.0	
No. of RBs, N_{RB}	50	
PRB size, RB_{s}	180 kHz	
Bandwidth Efficiency, B_{eff}	0.65	
SINR Efficiency, $SINR_{\mathrm{eff}}$	0.95	
User Placement	Random, 40 UEs per macro cell	

B. Simulation Results and Analysis

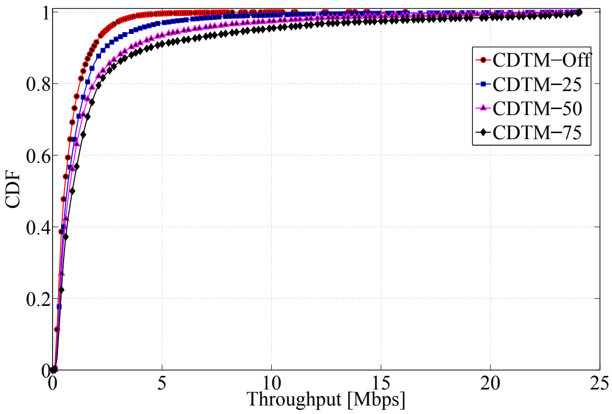
The user throughput distribution values are as shown in Fig. 4, for both the considered scenarios. The normalized mean, 5th and 95th percentile user throughput values are also shown in 5, with the values of CDTM-n schemes normalized to that of the reference mechanism with the traffic management scheme disabled. The normalized values would indicate the performance of the CDTM mechanism relative to the reference scheme. From figures, we can observe that the cloud traffic management mechanism gives significant throughput gains, mainly due to the delaying of low priority traffic in macro cells, giving higher throughput to the UEs having high priority downlink data. The losses for the high priority data while connected to small cells, through which the delay tolerant cloud data is scheduled is essentially not visible here, due to the full buffer traffic model assumed. It is also interesting to note that the mean user throughput values for the METIS scenario is significantly higher than the HetNet scenario, since the probability of connecting to a small cell is much higher due to the presence of high-rise buildings, which significantly degrades the macro cell link quality, prompting UEs to select small cells more often.

The PDU transmission delay results, based on the values taken from the delay distribution, are as shown in Fig. 6. The figure shows the normalized values, relative to the CDTM-Off mechanism. The cost of having a delayed traffic delivery mechanism can be observed here, with impacts less obvious in the HetNet scenario, due to UEs spending less time in small cells, even with the higher density. For the HetNet scenario, the delay of packet delivery is almost negligible. But for the METIS scenario, mean connection times decreases initially with the more UEs spending higher amount of time in small cells, and the remaining macro cell UEs receiving the full buffer data at a higher rate, thereby reducing the PDU delay. The PDU delay in small cells remain almost the same in both scenarios due to the full buffer traffic, explaining the difference in delays that are observed.

Based on the evaluations done in this section, we can observe that significant gains can be obtained in terms of throughput using the traffic and mobility management mechanism considered in this work. Here the mobility management element is involved mainly due to the cell preference conditions that could be set for the low-priority delay tolerant data. With a full buffer traffic, we can observe that there is up to 50 % increase in mean user throughput in the HetNet scenario, and up to 150 % increase in the METIS TC2 scenario. The difference is mainly because of the difference in the probability of small cell selection for both scenarios. Significant increase in cell edge, 5th percentile user throughput performance can be observed for the HetNet scenario, while higher gains for cell center 95th percentile users can be observed in the METIS

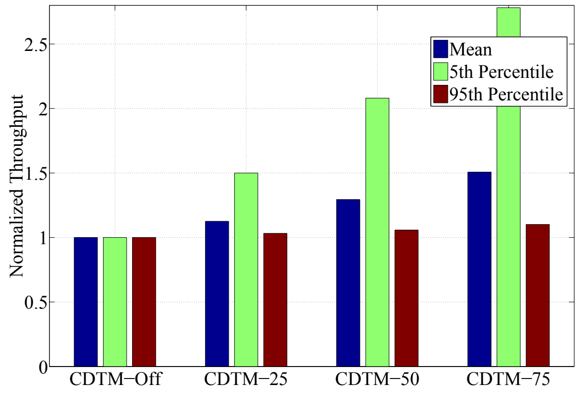


(a) HetNet Scenario

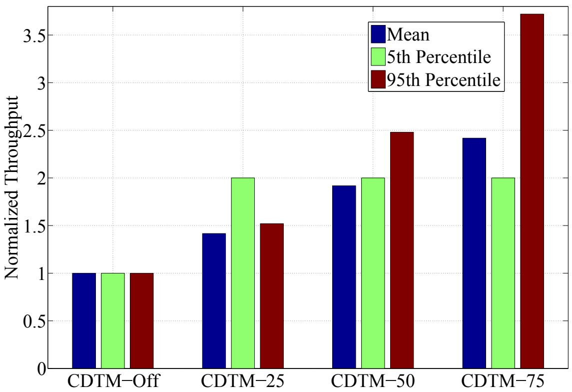


(b) METIS Scenario

Fig. 4. User throughput distribution for various scenarios.



(a) HetNet Scenario



(b) METIS Scenario

Fig. 5. Normalized throughput values for various scenarios.

TC2 scenario. Similar lowering of PDU delays can also be observed from the results.

V. CONCLUSION

In this work, we propose a mechanism for mobility and traffic management for delay tolerant cloud data, taking into account the close coordination between the cloud server and the 3GPP network. The mechanism was evaluated using LTE-Advanced (4G) HetNet scenario and 5G dense-urban information society scenario and related simulation settings. Based on the evaluations results, it can be observed that using the considered setting, significant mean throughput gains of up to 50 % for HetNet scenario, and 150 % for METIS TC2 scenario can be observed. For the cell-edge or 5th percentile users, gains from 50 to 150 % were observed for HetNet scenario and up to 100 % for METIS TC2 scenario. The gains are observed especially for the high-priority traffic, with acceptable packet delivery delays for the low-priority traffic. From the results, we can conclude that such mobility and traffic management schemes where the external cloud servers operate in cooperation with the mobile networks, could be an ideal enhancement for currently deployed 4G, as well as for future 5G systems.

One of the first enhancements for further studies could include evaluating the mechanism using more realistic traffic models. The use of more realistic mobility patterns, especially in the 5G METIS scenario, and the impact on the results would also be an interesting area for future work.

ACKNOWLEDGMENT

Part of this work has been performed in the framework of the FP7 project ICT-317669 METIS, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in METIS, although the views expressed are those of the authors and do not necessarily represent the project.

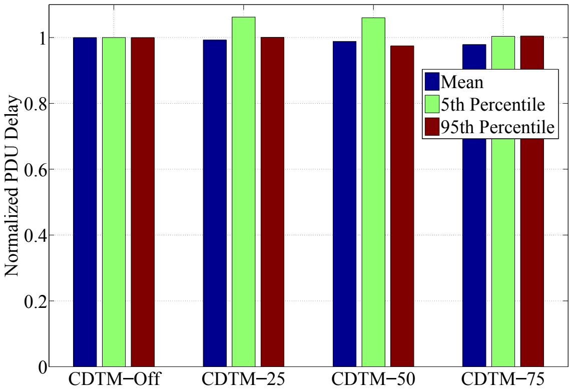
The authors would like to thank Dr. Kimmo Valkealahti for his assistance in the performance evaluations for the studies done in this work.

REFERENCES

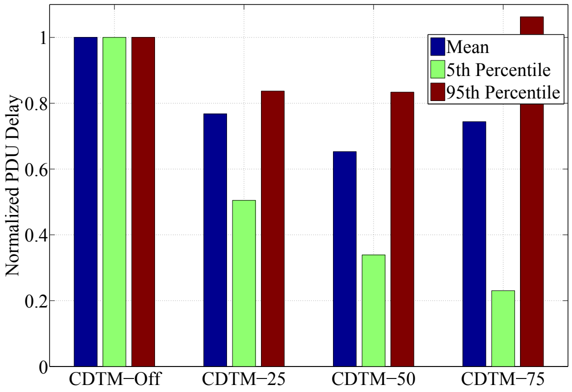
[1] T. Taleb and A. Ksentini, “Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks,” *IEEE Network*, vol. 27, no. 5, pp. 12–19, 2013.

[2] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, “A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches,” *Wireless Communications and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[3] D. Huang, Z. Zhou, L. Xu, T. Xing, and Y. Zhong, “Secure Data Processing Framework for Mobile Cloud Computing,” in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011, pp. 614–618.



(a) HetNet Scenario



(b) METIS Scenario

Fig. 6. Normalized PDU delay values for various scenarios.

[4] M. T. Beck and M. Maier, “Mobile Edge Computing: Challenges for Future Virtual Network Embedding Algorithms,” in *The Eighth International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP)*. IARIA, 2014, pp. 65–70.

[5] M. Patel et al., “Mobile-Edge Computing - Introductory Technical White Paper,” Sept. 2014, Issue 1.

[6] A. Staring and G. Karagiannis, “Cloud Computing Models and Their Application in LTE Based Cellular Systems,” in *IEEE International Conference on Communications Workshops (ICC)*, 2013, pp. 750–755.

[7] R. Yu, Y. Zhang, S. Gjessing, W. Xia, and K. Yang, “Toward Cloud-based Vehicular Networks with Efficient Resource Management,” *IEEE Network*, vol. 27, no. 5, pp. 48–55, 2013.

[8] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. Di Girolamo, and G. Giuliani, “RAN as a Service: Challenges of Designing a Flexible RAN Architecture in a Cloud-based Heterogeneous Mobile Network,” in *Future Network and Mobile Summit (FutureNetworkSummit)*, 2013, pp. 1–8.

[9] P. Rost and C. Bernardos et al., “Cloud Technologies for Flexible 5G Radio Access Networks,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[10] P. Rost and A. Prasad, “Opportunistic Hybrid ARQ - Enabler of Centralized-RAN Over Non-ideal Backhaul,” *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 481–484, Oct. 2014.

[11] METIS Deliverable D1.1, “Scenarios, Requirements and KPIs for 5G Mobile and Wireless System,” April 2013.

[12] METIS Deliverable D6.1, “Simulation Guidelines,” Oct. 2013.

[13] O. Yilmaz, C. Wijting, P. Lundén, and J. Hamalainen, “Optimized Mobile Connectivity for Bandwidth-Hungry, Delay-Tolerant Cloud Services Toward 5G,” in *IEEE ISWCS*, 2014, pp. 6–10.

[14] A. Banerjee, X. Chen, J. Erman, V. Gopalakrishnan, S. Lee, and J. Van Der Merwe, “MOCA: A Lightweight Mobile Cloud Offloading Architecture,” in *Eighth ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch)*, 2013, pp. 11–16.

[15] METIS Deliverable D4.3, “Final Report on Network-Level Solutions,” Feb. 2015.

[16] A. Prasad et al., “Energy Efficient Inter-Frequency Small Cell Discovery Techniques for LTE-Advanced Heterogeneous Network Deployments,” *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 72–81, May 2013.

[17] 3GPP TR 36.839, “E-UTRA; Mobility Enhancements in Heterogeneous Networks,” March 2012, ver. 0.5.0.

[18] 3GPP TR 36.814, “Further Advancements of E-UTRA: Physical Layer Aspects,” March 2010, ver. 0.4.0.

[19] P. Mogensen et al., “LTE Capacity Compared to the Shannon Bound,” in *IEEE VTC*, Dublin, April 2007, pp. 1234–1238.