Small Cells Solution For Enhanced Traffic Handlling In LTE-A Networks

Raid Sakat
Dept. of Electronic and Comp. Eng
Brunel University London
United Kingdom
raid.sakat@brunel.ac.uk

Raed Saadoon
Dept. of Electronic and Comp. Eng
Brunel University London
United Kingdom
raed.saadoon@brunel.ac.uk

Maysam Abbod
Dept. of Electronic and Comp. Eng
Brunel University London
United Kingdom
maysam.abbod@brunel.ac.uk

Abstract— **in contrast to the previous generations, the next generation mobile network will be largely driven by data. Data are increasing in exponential way due to the increase number of smart mobile devices and network connection capabilities. Mobile connectivity has become essential for many network users, and demand for mobile broadband to access the different applications and services on the internet is increasing in a rapid way. Therefore, data delivery will represent a challenging area for the mobile networks, taking in consideration the adaption of advanced multimedia application, video content and IoT applications. New methods, strategies, and techniques that provide Interworking and harmonization between computer and telecom network in a unified heterogenous ecosystem that can provide computing capabilities in the edge should be considered for the implementation of next generation of mobile networks.**

Keywords— next generation mobile network, MEC, latency, content delivery, IT.

INTRODUCTION

The current mobile network, 4G, has provided users with considerably faster data speed and lower latency rate, changing the way people accessing and using the internet on their mobile devices in a dramatic way.

Combining the 4G ability with smart devices capabilities lead to wide adaption of advanced multimedia applications that contributes to the increase in mobile traffic. The number of mobile subscribers has grown in a rapid way and according to Cisco Visual Network Index (VNI) [1], almost half a billion mobile devices were added to the network in 2016 and smart phones accounted for most of that growth (46%). By 2021 it is expected that global mobile devices will grow to 11.6 billion and nearly three-quarters of all devices will be smart [1]. The traffic generated by these smart devices will represent (86%) of the total mobile traffic by 2021, in which the overall global mobile data traffic is expected to increase 7 fold between 2016 and 2021 at a compound annual growth rate (CAGR) of (47%) to reach 49 exabytes/month. In addition, due mobile video content has much higher bit rates than other mobile content types, mobile video content will represent (78%) of the global mobile data traffic. In parallel to the growth of the smart devices, new use cases and applications related in one way or another to the IoT, also contribute to the global data generation and this growth of the data by such application is clear indication of the IoT growth. Example of such technologies is machine to machine (M2M), device to device (D2D), sensor networks, that are in continuous development in order to form the IoT networks, that will connect buildings, cars, people, and things. This will result in more traffic being generated, more bandwidth to be allocated, and more power will be consumed. This explosive growth of data traffic, the massive increase in the number of devices, with continues emergence of new applications and services will exhaust the current system resources and capabilities. The result will be direct impact in the performance and efficiency of the current mobile network KPIs, specially the backhaul capacity and long latency.

This is why next generation of the mobile network will require an ecosystem of changeable components in order to be able to fulfil the increasing demand for data required by the new applications and use cases presented in daily manner, and to be more relevant to both the user and the operator. Therefore, the next phase of the system must be adaptive to change, thus it should be heterogeneous ecosystem with more virtualisation so that new scenarios can be shaped in a timely manner. Convergence between computer and telecom networks could be a promising technique that gives the ability to bring the required data to the edge of the network using the content delivery mechanism and cashing, in the same way, many of the core network capabilities could be placed at the edge of the network benefiting from the virtualization functionality provided by the computer network mainly the mobile edge computing system. In the other hand, Heterogeneous networks (HetNets) that consists of macro, micro, small and femto cells are classified as one of the main steps toward meeting the future requirements for handling the infinite growth in mobile data [2].

Though interacting HetNet had offered great benefits, it also resulted and still leading to several design changes that need new solutions to make it a true success [3, 4]. There are multiple researches examining further integration of macro and small cell functionalities to fully utilize the concept of HetNet deployment. Many techniques are proposed and listed in [5, 6, &7]. With different carrier frequencies deployed at macro and small cell layers, dual connectivity (DC) which complements the LTE-A Carrier Aggregation (CA) functionality to allow the user equipment (UE) to receive data

from both a macro and a small cell eNodeBs simultaneously is a promising new technique. Dual Connectivity is one of the solutions standardized by 3GPP for Release 12 small cell enhancements. It aims to improve user throughput performance by utilizing radio resources offered by more than one eNodeB [8].

The evolution of edge computing and advancements in wireless networking to provide cloud computing and caching capabilities at the edge of cellular networks was largely covered in [9, &10], In both surveys, the authors make an exhaustive review on the state-of-the-art research efforts on mobile edge networks, including the ETSI MEC, Fog Computing, Cloudlet and edge caching. An overview of mobile edge networks including definition, architecture and advantages was given. The survey covers the issues on computing, caching and communication techniques, the applications and use cases. The key enablers of mobile edge networks such as cloud technology, SDN/NFV and smart devices are also included and discussed in the survey. Finally, open research challenges and future directions are presented as well.

In addition to the above, a Study on Context Aware Service Delivery in RAN for LTE was introduced by 3GPP TR 36.933 as part of the release 14, to provides descriptions and possible solutions of use cases for the Context Aware Service Delivery in RAN for LTE, some of the described problems was the backhaul long latency, throughput and video transmission. Local content caching and ETSI MEC work on mobile video delivery optimization was addressed as part of the solutions [11].

I. CACHING BENEFIT

In this section the cost structure of caching data at eNodeBs are calculated. The problem is considered from two standpoints: The Delay and the Bandwidth (hence the throughput). In the second scenario, it is shown that caching at eNodeB can lead to many benefits for both mobile operators and end users. Furthermore, it shows that the in-network caches can effectively improve the eNodeB's caching performance.

1) eNodeB Caching Delivery Cost

Selecting N to be the number of eNodeBs in the network, n_j be the number of data requests received by the jth eNodeB and P be the mean size (bytes) of a requested object. The cost components are denoted as follows:

i)   U is the cost per byte from UE to eNodeB.
ii)   C is the backhaul link cost per byte from eNodeB to PGW.
iii) W is the transit cost per byte between core network and the content provider.

If no caching is provided in the cellular network, each request will incur the costs U+C+W, and the total costs for complying all the requests in the network can be calculated as follows:

$$M_{nocache} = \sum_{j=1} (n_j) \times (U+C+W) \times P \qquad (1)$$

In the case of adding cache to the eNodeB, the following parameters for the equation are considered :

(i) n_{jc} be the number of request for objects that are cached at the j_{th} eNodeB.

(ii) K is the additional cost per byte of caching objects at the eNodeBs (server, storage cost, etc). With the existence of eNodeB caching, the cost for the mobile operator to serve the requests will be the sum of the following parameters:

$$M_1 = \sum_{j=1} (n_j - n_{jc}) \times (U+C+W) \times P \qquad (2)$$

where M1 is the cost when requested objects need to be fetched from the origin server.

$$M_2 = \sum_{j=1} n_{jc} \times U \times P \qquad (3)$$

where M2 is the cost incurred on the network path from the UEs to the caching eNodeB.

$$M_3 = \sum_{j=1} n_{jc} \times K \times P \qquad (4)$$

where M3 is the cost when adding cache to the eNodeB. This will result in:

$$M_{cache} = M_1 + M_{2+} M_3$$
$$M_{cache} = \sum_{j=1} n_{jc} \times U \times P + \sum_{j=1} (n_j - n_{jc}) * (U+C+W) \times P + \sum_{j=1} n_{jc} \times K \times P \qquad (5)$$

By subtracting (5) from (1) this will result in the benefit of adding cache to the network and as follows:

$$M_{Benefit} = \sum_{j=1} n_{jc} \times (C+W-K) \times P \qquad (6)$$

2) In-network Cache

It is observed that the new trend of in-network caching can achieve additional delay reduction for end users. Keeping in mind, the advantage of in-network cache might be a disadvantage if the in-network cache cannot be efficiently utilized by the main eNodeB. An example of this is the situation when the cached content is duplicated in both the eNodeB and the in-network caches, which most likely occur in the case of full load.

II. GAIN ANALYSIS WITH DUAL CONNECTIVITY

In this section, gain increasing mechanism is demonstrated when introducing DC to the network and provisioning the layers to Macro and Small, first it is calculated theoretically according to Shannon-Hartely equation as follow:

$$C_i = B_i \log_2 (1+ SNR_i)$$

where C is the capacity (hence throughput), B is the bandwidth and SNR is the signal to noise ratio all related to cell (i) which is assumed to have the best (maximum)

throughput /user in both Macro and Small, i.e. im = arg max (Ci) for iϵ M and is = arg max (Ci) for iϵ S.

when there is no DC, the user is assumed to be served by a single cell, When introducing DC ,the users are assumed to receive data from the small cells and the controls from the macro cell. The candidate cells characterized by the best estimated throughput in both the macro and small cell layers are selected as the serving cells. The Shannon capacity for the user with DC is expressed as:

$$C_{DC} = C_{im} + C_{is}$$

The user throughput gain with DC will be

$$Gain = \frac{CDC - CnoDC}{CnoDC} \text{ X } 100\%$$

$$= \frac{Bq \log 2 \ (1 + SNRq)}{Bp \log 2 \ (1 + SNRp)} \text{ X } 100\%$$

Where q = arg min (Cq), and p= arg max (Cp)

If the same bandwidth is deployed in both two layers (i.e. Bq/Bq =1) it will explicitly show that DC is most beneficial for users exposed to similar channel conditions in both layers Plus 100 % DC gain when SNR difference is 0. Noting the DC gain cannot be larger than 100 %, due to the reason that for cases without DC the selected serving cell is assumed to have the highest estimated throughput from the candidate cells in the two layers.

The next section illustrates a framework presenting small cells solution where eNodeBs can use in-network cache information to improve performance. The model is simulated in order to find an optimal eNodeB caching solution to reduce end to end latency and to increase the throughput in the framework under several network load conditions.

III. SYSTEM ARCHITECTURE

Next generation mobile network will integrate different types of networks. computing and storage resources will be combined into one programmable and unified infrastructure, making use of the strong virtualisation ability of new system. This new model of unified system will allow for an optimized usage of all distributed resources, and the convergence of fixed, mobile and broadcast services [12]. The evolution towards next generation mobile networks is largely driven and characterized by an exponential growth of traffic caused by an increased number of user terminals. The majority of the terminals are smart powerful devices with IP based services supported by the all IP Lte-A network (4G).

IP-based services in the 4G network are a result of the convergence between Telecom and computer networks. 4G changes the way people access the internet as the use of IP-based services work across different devices in different networks in a reliable and cost effective way. It is important to meet the customers expectation and to increase the quality of experience (QoE) by employing the internet protocols (IP) to reduce the cost of transferring each bit of data. Current mobile networks are designed for typical internet traffic, and are not able to support this purpose efficiently; therefore, a general purpose IT platform, which allows virtualization and implementation of the centralized functionality of the telecom core network in a point nearest to the end user, could be seen as a promising way to allow for handling future data rate demands.

In such a scenario, bringing the computation and storage capacity from the core to the edge of the network, dense deployment of low-power, small-cell nodes in which the distance between the radio access points (RAP) and terminals is reduced, virtualization and centralized processing would improve the throughput while keeping the low latency and without adding additional overhead .

In this work, we consider the concept of adding computing and storing capacity to the main eNodeB, in which the content will be cached and stored in a server attached to the main eNodeB. Small cell nodes will be distributed with different frequency band under the coverage of the main eNodeB, and the small cells are connected to and controlled by the server attached to main eNodeB through fiber cables connection, as in the C-RAN architecture. In this way the resources of several cells can be pooled in a centralized entity. In Lte-A network, resource allocation takes place at the level of cells, and scheduling of the resource units called Resource Block (RB) took place every Transmission Time Intervals (TTIs). A UE is associated to a cell, and transmission of neighboring cells on the same RBs count as interference, interference-prone transmission imply lower Signal to Noise and Interference Ratio (SINR), leading to more RBs being used to transmit the same payload, this obviously reduces the capacity of the network, allowing fewer UEs to be served simultaneously, and effecting the quality of service being introduced to the end user, on the other hand, it will negatively affects the energy efficiency, which also depends on the number of bits per RB. [13]

Centralized processing of the resources would allow for efficient interference avoidance and cancellation algorithms across multiple cells as well as joint detection algorithms. In addition, the dense deployment of small cells under flexible centralization of the radio access network will allow for flexible functional split based on the virtualization functionality provided by the computing ability at the edge of the network, in this way, the main eNodeB could be used for the normal connection handling most of the system control signaling, while the small cells could be seen as hot spots used for downloading the required content. Fig (1) below shows the system diagram.
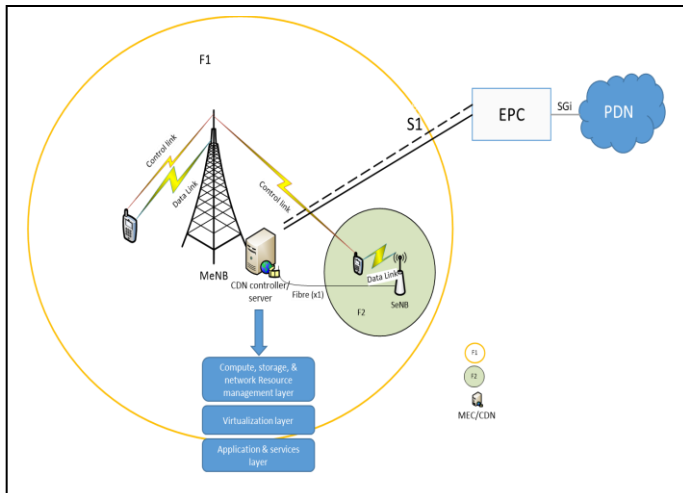
Figure 1: SYSTEM ARCHITECTURE

In this architecture, MeNodeB and SCs are operating in different frequency bands, (F1) for MeNodeB and (F2) for the SCs. All the SCs deployed under the coverage of a specific MeNodeB are operating in the same frequency (F2). The UE in the proposed form of dual connectivity maintain a normal connection with MeNodeB and will establish a U-Plane connection with a SC for the downlink of big data applications (i.e. videos) that could be saved in content delivery (CD) server located in or near the MeNodeB. (an ETSI MEC (Mobile Edge Computing) server could be used for this purpose), which can add computing capabilities for the radio access networks (RATs), or could be used as an aggregation point in the IP transport layer.

Dual Connectivity (DC) as in LTE release 12 specifications, LTE small cell enhancement by dual connectivity is a technology which extends carrier aggregation (CA) and coordinated multi-point (CoMP), in which the small cells are typically deployed as hotspots within macro cell coverage. where the UE can receive/transmit data from/to multiple eNodeBs simultaneously. The expected benefit from such enhancement such as:

- Increased UE throughput especially for cell edge UEs;

- Reducing signalling overhead towards the core due to frequent handover.

Information exchanged between the MenodeB server and UE may took place on different layers, such as MAC, PDCP and RRC layers. A UE in RRC – connected mode first obtain access to the MeNodeB and keep C-plane connection with this node, which is the only RAT element that is visible to the core Network (EPC), measurement and statistics information related to the UE gathered by the mobile network element based on the 3GPP signalling messages and Performance Measurements (PM) defined by 3GPP can be aggregated and processed by the controller of the MenodeB, a table of information will be generated that will also contain measurements considering the information coming from the

SCs. As soon as a big size content is requested by a UE, the MenodeB will direct the UE (i.e. through the system information Block SIB) to connect to the best SC based on the parameters provided by the controller. Fig (2) below, shows the flow chart of this procedure.
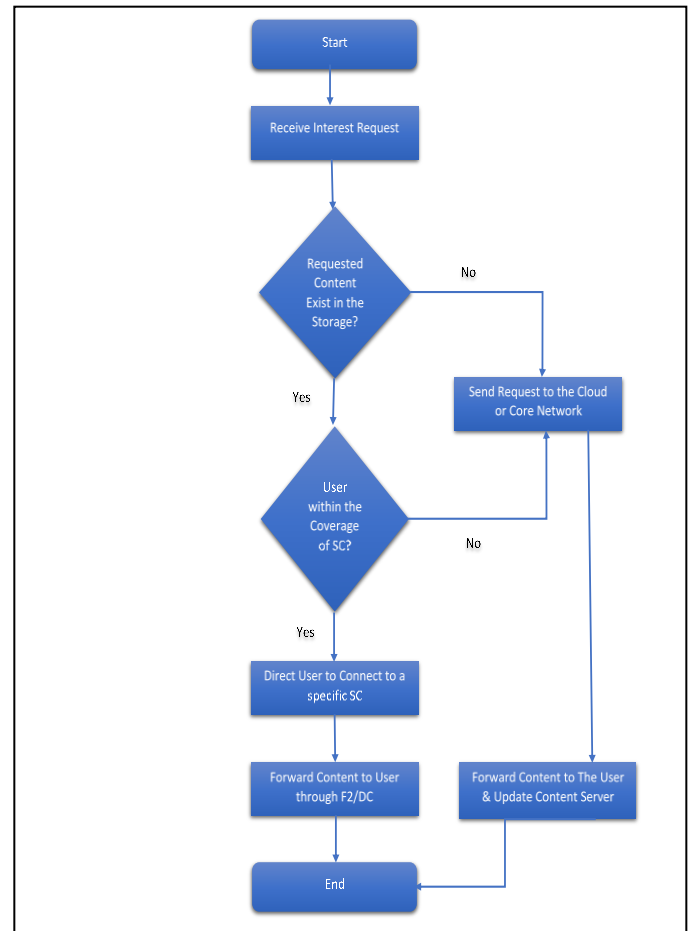

Figure 2: Content delivery procedure

The flowing steps, could explain such a procedure:

Step 1: the measurement report made by the UE and sent to the controller of the MeNodeB to be added to the measurement table.

Step 2: A content is requested by a UE

Step 3: MeNodeB decided to which node of the SCs the UE is to be connected considering the measurement parameters available in the measurement table.

Step 4: MeNodeB send the decision to the UE (through dedicated RRC signalling, i.e. RRC Connection Reconfiguration)

Step 5: UE connect to the node of small cell decided by the controller of the MeNodeB.

IV. PERFORMANCE EVALUATION

A. Simulation Setup

The proposed cached enabled dual connectivity architecture that integrate the SCs to the LTE MeNodeB at the PDCP layer, will be implemented using Riverbed 18.5 Modeller based on the 3GPP technical requirement for small cell enhancement [14]. The system model will be investigated on terms of throughput and delay.

The Riverbed (formerly known as OPNET) is a power simulation program, offers libraries that contains more than 400 'out of the box' protocols and vendor device models including TCP/UDP, IPv6, VoIP/Video/FTP/HTTP/Email, WiMAX, WLAN (a/b/g/n), and LTE to support accurate event driven simulation scenarios. Nevertheless, The LTE model features supported by this modeller are based on 3GPP release 8 & 9, that don't support dual connectivity. Therefore, a modification to the LTE node models is required. Furthermore, visualization functionality is not supported but there is huge amount of devices belongs to the computer networking that could be used instead to simulate the network using the program capabilities to assign tasks and choose the statistics of each node. A measurement entity is created for each UE, which records values of RSRP and RSRQ, thus the UE continuously measures RSRP and RSRQ for all nodes within its range.

The system performance is evaluated using multiple scenarios using riverbed simulator to investigate the optimal solution, with the same LTE simulation parameters, that are set according to 3GPP TR 36.842 [15] and TR 36.872 [16] summarized in Table I.

TABLE I.   SIMULATION PARAMETERS

Parameter	Value
Type of Service	HTTP with Video
Video Type	On demand
File Size	200 Mbytes
File inter-arrival distribution	Exponential
Average File inter-arrival Time	16s, 20s, 24s
Total MeNodeB Tx Power	46 dBm
SC Tx Power	30 dBm
Noise figure	9 dB in UE, 5 dB in eNodeBs
UE Tx Power	23 dBm
MeNodeB Carrier Frequency (F1)	2 GHz
SC Carrier Frequency (F2)	3.5 GHz
LTE Bandwidth/Duplexing	20 MHz/FDD
Sub-carrier spacing	15 kHz
Sub-frame length (TTI)	1 ms
Symbols per TTI	14
Data/control symbols per TTI	11/3

LTE system contains 1 MeNodeB, with variable number of hotspots (SCs) and UEs as shown in table II. The SCs and UEs are randomly distributed under the MeNodeB coverage. Adaptive modulation and coding was enabled in order to enable the UE to communicate with the eNodeB in variable channel conditions. The interference and multi-path are modelled. IP traffic is established between the UEs and HTTP server is connected to the LTE network through internet backbone.

For the first scenario set to start with 1 MeNodeB and 5 UEs randomly distributed within the MeNodeB coverage area, then for the succeeding scenarios, the number of SCs and UEs will be increased. To be as 0, 1, 2, 3 for the SCs, and 5, 10, 20 for the users. As shown in Table II. The simulation time has the duration of 30 minutes, there is a warm up time of 90 seconds approximately, before the start of the simulation and results collection.

TABLE II.   SIMULATION RESULTS

CSs	UEs	Throughput (bit/sec)
0	5	3,075,832.39
1	5	6,368,130.09
1	10	9,141,040.66
2	10	9,387,580.68
2	20	16,167,906.01
3	20	14,462,716.01

B. Results and Discussion

In this section we analyse the proposed scheme based on the scenarios and settings discussed in the preceding sections. The sent and received IP data packets all over the LTE network are examined.

The throughput and packet end-to-end (E2E) delay are chosen as the key performance factors. In the first scenario, the network is configured with low load traffic to decrease the probability of packet loss due to either the buffer overflow or repeated retransmissions due to the traffic congestion

Four main cases are considered;

1)   No content is cached.

2)   The UEs are connected to the M-eNodeB in the UL and to the SC-eNodeB in the DL, with Content cached in one SC-eNodeB and (5) UE's attached to it.

3)   The UE is connected to the M-eNodeB in the UL and to the SC-eNodeB in the DL, with Content cached in two SC-eNodeB and (5,10) UE's attached to them.

4)   The UE is connected to the M-eNodeB in the UL and to the SC-eNodeB in the DL, with Content cached in three SC-eNodeB and (10,20) UE's attached to them.

Fig. 3 below shows the response of the network in terms of E2E delay for the four scenarios. It can be observed that the delay and the response of the network is acceptable when the UE is connected to the M-eNodeB provided that no data is cached in the network, it jumps higher and when the number of UE's is tenfold. This is the expected response as the burden increase on the M-eNodeB assuming that same content is

routed in the network in every scenario. Whilst the incremental of SC-eNodeBs number in the entire network significantly drops the delay as the time elapsed to fetch data from the cloud is narrowed or sub-zeroed.

The explanation of drop in the delay is the fact that the distance to the M-eNodeB is quite larger than the distance to the SCeNodeB, provided that the DL frequency differs from the UL frequency which reduces the interference in the network in order to help decreasing the losses as proposed.
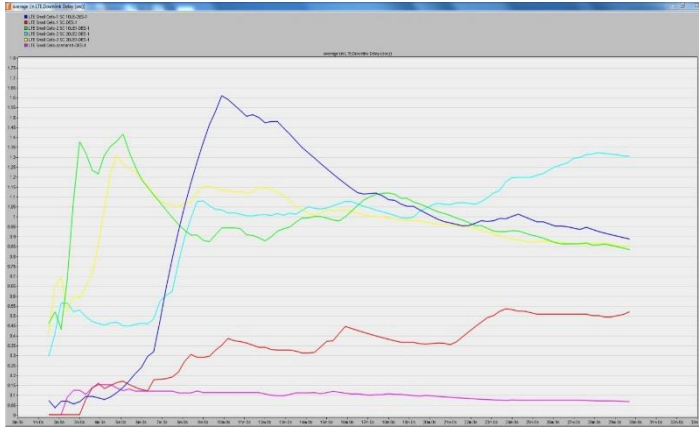

Figure 3: E2E delay

Considering the same network simulation we examine the results in terms of the throughput. Fig. 4 below illustrates the throughput delivered in bits/seconds. It can be observed that the throughput is increasing when the content server is getting closer to the UE, achieving its best when the UE downlink is connected to the SC-eNodeB and using the proposed scheme. Logically, the throughput (bits/sec) increases in 2 cases

- when the data traffic increases
- when the elapsed time decreases.

Hence in our model, when the content are cached in the M-eNodeB the network delivers and performs best at the begging of the simulation. However, it starts to perform even better in the third scenario when the DL is connected to the SC-eNodeB after 20% of the simulation time, this is due to SC-eNodeB initialization and time spent fetching the content from the cloud.
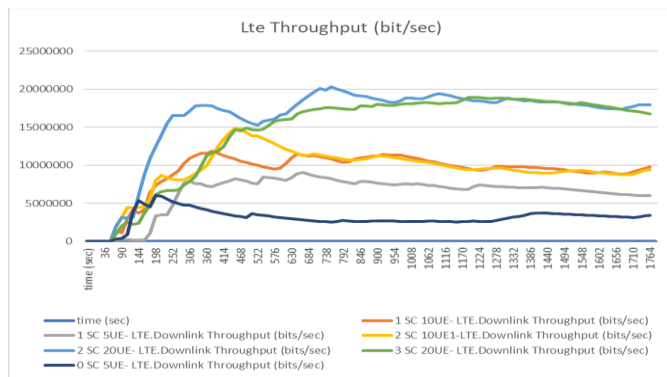

Figure 4: throughput delivered in bits/seconds

Finally we experienced the case when the same network runs and routes variable content full load scheme in its data plane. Fig. 5 below shows the response of the network in terms of E2E delay for the three scenarios. It can be observed that the delay is very high when the UE is connected to the M-eNodeB provided that no data is cached in the network while it is acceptable when the contents are cached in the M-eNodeB and has dropped remarkably when the UE is connected to the SC-eNodeB and using the proposed scheme.
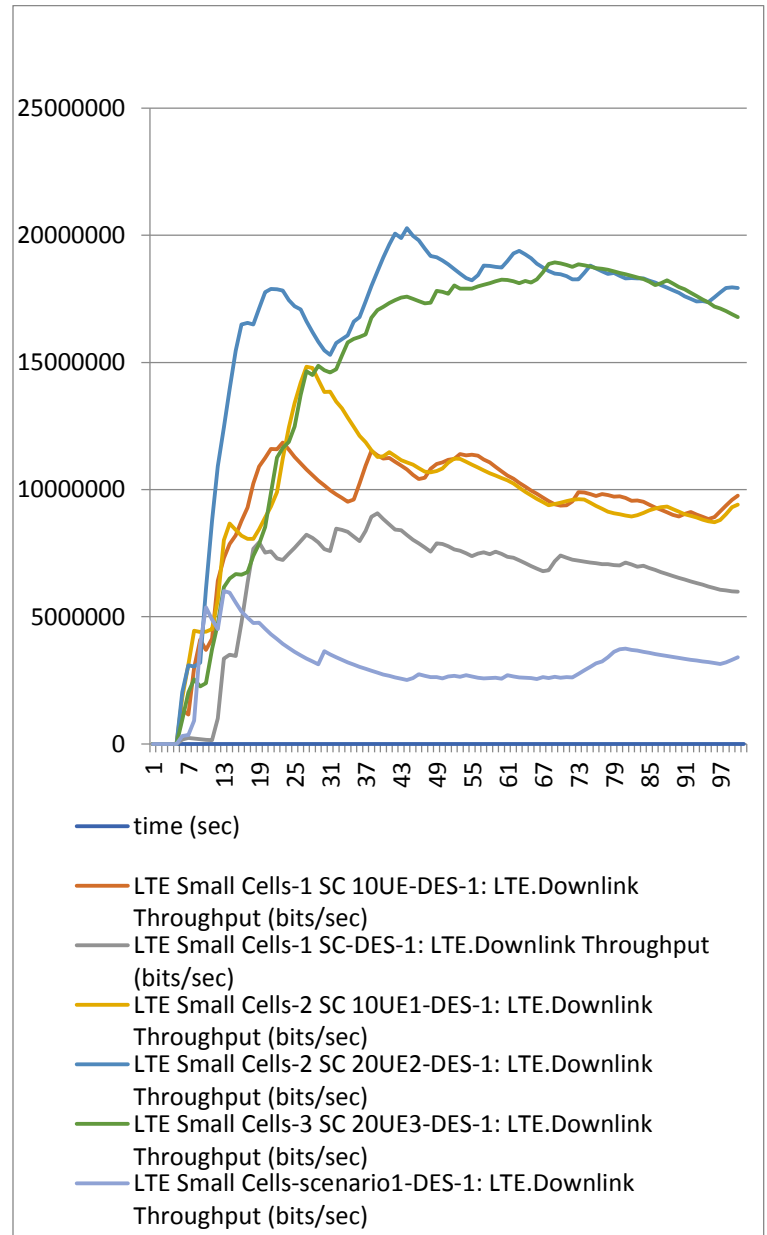

Figure 5: E2E delay for the three scenarios

It can be observed that throughput is increasing when the content server is getting closer to the UE, achieving its best when the UE downlink is connected to the SC-eNodeB and using the proposed scheme. Noting that when the content server is placed in the M-eNodeB the throughput drops

dramatically when the network is running full load, this is due to high traffic that is being processed and requests from the UEs to be fulfilled by one content server.

V. CONCLUSION

The convergence between the IT and Telecommunication Technologies bring new capabilities that will enable the deployment of new services at the edge of the network. In this paper we presented a new transmission/reception scheme based on the CoMP concepts when UL and DL links are physically separated into M-eNodeB and SCs in which the processing of the data and the control of the SC is within a controller at the edge of the network. When comparing the data traffic received in the DL direction under many network circumstances, it can be observed that the proposed method has successfully enhanced the performance of the network and suppressed any performance degrading that may occur due to a sudden high network load or conditions. The results show that the proposed scheme can reduce the packet loss and latency when routing full load in the network, and the model becomes more efficient and delivers better results when the number of users is increased.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021

[2] Damnjanonic, A., et al. (2011). A survey on 3GPP heterogeneous networks. IEEE Wireless Communications Magazine, 18(3), 10–21.

[3] Andrews, J. G. (2013). Seven ways that HetNets are a cellular paradigm shift. IEEE Communications Magazine, 51(3), 137–143.

[4] Ghosh, A., et al. (2012). Heterogeneous cellular networks: From theory to practice. IEEE Communications Magazine, 50(6), 54–64.

[5] Kishiyama, Y, et al. (2013). Future steps on LTE-A: Evolution toward integration of local area and wide area system. IEEE Wireless Communications Magazine, 20(1), 12–17.

[6] Nakamura, T., et al. (2013). Trends in small cell enhancements in LTE advanced. IEEE Communications Magazine, 51(2), 99–105.

[7] Soret, B., Wang, H., Pedersen, K. I., & Rosa, C. (2013). Multi cell cooperation for LTE-advanced heterogeneous network scenarios. IEEE Wireless Communications Magazine, 20(1), 27–34.

[8] 3GPP Technical Report 36.842. Small cell enhancements for E-UTRA and E-UTRAN—Higher Layer Aspects, V1.0.0 (2013-11). www.3gpp.org.

[9] Shuo Wang, et al. (2016). A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications.

[10] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016, pp. 1-8.

[11] 3GPP Technical Report 36.933. Study on Context Aware Service Delivery in RAN for LTE, V14.0.0 (2017-03)

[12] Raed Sahib Sadoon, "Explosion of Data (BIGDATA)", Chapter "3" in the "Internet of Things and Big Data Analysis: Recent Trends and Challenges", Nov, 2016, ISBN-10:0692809929.

[13] Nardini, G., Stea, G., Virdis, A. et al. Wireless Netw (2016) 22:11. Doi:10.1007/s11276-015-0948-6

[14] 3GPP TR 36.932. Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN; V14.0.0 (2017-03)

[15] 3GPP Technical Report 36.842. Small cell enhancements for E-UTRA and E-UTRAN—Higher Layer Aspects, V1.0.0 (2013-11).

[16] 3GPP TR 36.872. Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects; V12.1.0 (2013-12).