5G Multi-Connectivity with Non-Ideal Backhaul: Distributed vs Cloud-Based Architecture

Diomidis S. Michalopoulos*, Andreas Maeder*, Niko Kolehmainen†
* Nokia Bell Labs, Munich, Germany, Email: {firstname.lastname}@nokia-bell-labs.com
†Magister Solutions Ltd., Jyvaskyla, Finland, Email: niko.kolehmainen@magister.fi

***Abstract*—We investigate the throughput performance of multi-connectivity over Heterogeneous Networks (HetNets). Specifically, we examine the effect of non-ideal backhaul on the throughput for two architecture scenarios, namely: i) The *distributed HetNet* scenario where the multi-connectivity anchor is co-located with the macro cell; ii) the *cloud-based HetNet* scenario, where the multi-connectivity anchor point is located at a centralized network point. An extensive set of simulations is conducted, followed by an investigation of the effect of the backhaul latency, packet file size and offered load on the throughput. It is shown that the cloud-based HetNet architecture leads to a superior throughput performance than that of the distributed architecture. Moreover, it is shown that the backhaul delay considerably affects the overall throughput for both architecture options.**

I. Introduction

Mobile broadband is increasingly gaining interest in the consumer market. From the network's perspective, this entails critical deployment challenges, since the Radio Access Network (RAN) is pushed to its limits in order to cope with such high traffic demand. In particular, it has become clear that the spectrum available for existing 4G technologies is not sufficient for meeting the 5G traffic, pushing thus towards higher carrier frequencies for 5G networks [1].

Nevertheless, besides higher carrier frequency, the telecommunications industry is also directed towards a more dense cell deployment. This facilitates the coverage of so-called "hot spots" since traditional cell deployments involving "rooftop" installations of base stations are unlikely to provide the required level of capacity. Consequently, densifying cell deployment gives rise to the notion of *Heterogeneous Networks (HetNets)*, where different layers of cells (e.g., macro and small cells) are utilized in the same area and time [2]. Typically, the coverage area of a set of small cells overlaps with that of a macro cell, which usually coincides with the area where network coverage focuses on.

A. The concept of Multi-Connectivity

An immediate technological followup of the concept of HetNets is that of multi-connectivity. Multi-connectivity refers to the case where the User Equipment (UE) establishes multiple physical link connections to the RAN, carried out at the same time. There exist, in general, several variations of multi-connectivity, depending on the number of simultaneous physical connections between the UE and the RAN, on whether the links involved operate at the same frequency, etc.

Multi-connectivity is a major candidate for 5G, due to its ability to tackle the diverse requirements involved. It comes as the continuation of "dual-connectivity" which is included in the standards of Long Term Evolution (LTE) [3]. Besides the number of access points involved, the major difference between multi- and dual-connectivity lies on the fact that dual-connectivity is designed to exclusively increase throughput; multi-connectivity is more versatile and able to meet other Key Performance Indicators (KPIs) as well, such as reliability and latency. Moreover, multi-connectivity is not restricted to single Radio Access Technology (RAT) connections, but rather spans multi-RAT scenarios, e.g., connectivity across LTE and 5G cells.

Multi-connectivity in literature: Although a relatively recent topic, multi-connectivity has appeared in a number of relevant publications so far. An extensive view of the underlying architecture supporting multi-connectivity across different RATs is provided in [4], [5], where it was concluded that such multi-RAT architecture should employ common high layer protocol as well as centralized radio resource management. Specific characteristics and requirements of multi-connectivity when applied to millimeter-Wavelength (mmW) networks are provided in [6], [7]. Specifically, [6] highlights the benefit of centralized controller decisions with regards to handover and scheduling, while [7] emphasizes the need for low-band support of 5G high frequency networks in the form of multi-connectivity. A performance assessment of multi-connectivity when implemented across intra- or inter-site deployments is given in [8], while mobility related improvements of dense deployments by means of multi-connectivity are analyzed in [9]. Finally, the dual potential of multi-connectivity to simultaneously enhance throughput and reliability is highlighted in [10], by means of a useful diversity-multiplexing tradeoff analysis.

B. Contribution and Structure

In the context of this work we focus on the broadband aspects of multi-connectivity. Specifically, we study *multi-connectivity over HetNets*, and particularly on the special case of non-overlapping small cells, where the UEs connect simultaneously to one macro cell and one small cell within its coverage zone. Two architecture options are considered, namely the distributed and the cloud-based architecture. Our main focal point is the *effect of the backhaul delay and the distributed vs cloud-based architecture to the throughput* performance of multi-connectivity. This effect is assessed via system simulations, which involve HetNets scenarios with different architecture topologies and different assumptions on the backhaul delay.

The structure of this work is as follows. For deriving our concluding remarks in Section VI, we conduct an extensive set of simulations whose setup is presented in Section III, while the corresponding results in Sections IV and V. The simulation setup corresponds to given assumptions in terms of the considered architecture, where two major cases are distinguished and discussed in the ensuing, Section II.

II. Multi-Connectivity Architecture

As illustrated in Fig. 1, two architecture and deployment options are considered for realizing multi-connectivity in a HetNet deployments. Such options are referred to henceforth as *Architecture (a)* and *Architecture (b)*, respectively, and correspond to a *distributed* and a *cloud-based* HetNet deployment, respectively.

The HetNet deployment consists of a macro cell coverage layer and a small cell capacity booster layer with cell radii of 50 m and below, depending on the carrier frequency and radio propagation scenario. These small cells can be densely deployed in urban areas with very high capacity demands such as in megacity hotspots. The resulting high traffic volume needs to be supported by an edge transport network of sufficiently capacity. Following [11], we assume a star edge network topology which is connected to an aggregation node of sufficient capacity of the mobile network operators transport network.

In the left part of Fig. 1, the distributed HetNet deployment is depicted. In this scenario, the aforementioned aggregation node is not co-located with any RAN functions. The Macro base station is logically connected via an Xn interface (equivalent to an X2 interface in LTE architecture) to a set of small cell base stations (both denoted as generalized Node B (gNB) in 3GPP nomenclature). A so called "split" data radio bearer (DRB) is established, with the user plane (UP) anchor point co-located with the macro cell. In such "split" DRB, one leg of the radio bearer is directly terminated at the macro gNB, while the other is routed via the small cell gNB over the Xn interface to the macro gNB, traversing the aggregation node. The packet data convergence protocol (PDCP) layer is responsible for aggregation and splitting of end-to-end traffic to the two radio legs. A flow control protocol as part of the Xn interface it ensures that buffer starvation in the small cell gNB is avoided [14].

In the right part of Fig. 1, the cloud-based architecture scenario is depicted. In this scenario, the aggregation node is enhanced with some computational capacity such that it can host a centralized unit (CU) which executes some of the higher L2 functions of the radio protocol stack, including PDCP and radio resource control (RRC). The lower layer part of the gNBs are located in distributed units (DUs) which are logically connected via the F1 interface to the CU [18]. The user-plane part of the F1 interface maintains a functionally identical flow control mechanism as the Xn interface, avoiding buffer starvation at the DUs. Since the PDCP layer is located in the CU, the split DRB is now anchored in the CU as well. The capacity requirements of the F1 interface are very close to

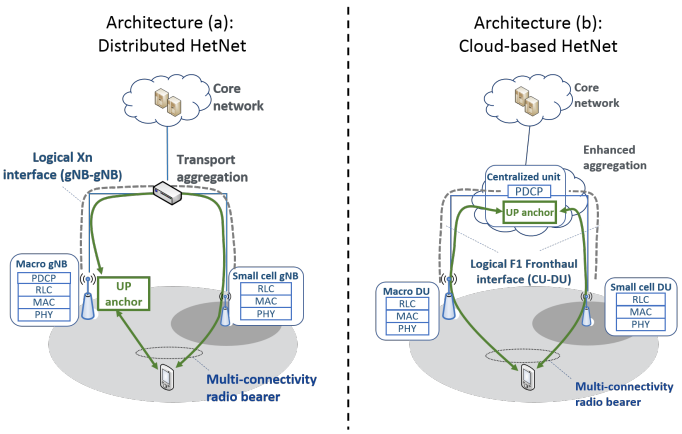

Fig. 1. Distributed HetNet vs. Cloud-Based HetNet architecture for multi-connectivity

that of the Xn interface on the user plane, such that the same transport network can serve both architecture options.

The implications of these architecture options on multi-connectivity are as follows: In the distributed scenario, the downlink end-to-end traffic is routed over the aggregation node from the core network to the master node (MN - in this case the macro gNB), where it is split and one part sent over the macro cell radio leg to the UE, and the other part routed via the aggregation node to the secondary node (SN) and over the high capacity small cell radio to the UE. Assuming symmetric link latencies L between the aggregation node and the macro as well the small cell node, the latency on the Xn interface between the split/flow control function in the macro node and the small node is $2L$.

In the cloud-based scenario, the split and flow control functions are located in the CU at the aggregation node. Thus, assuming the same transport network as in the distributed scenario, a latency of L ms is introduced between CU and DU. Both radio legs experience the same latencies between UE and CU PDCP.

III. Simulation Setup

For assessing the throughput performance of multi-connectivity in scenarios with non-ideal backhaul with non-zero latency, we conducted an extensive set of simulations as explained below. We have used a fully dynamic system simulator with OFDM symbol level resolution in time and subcarrier resolution in frequency having both LTE and 5G NR simulation capabilities. However for speed-up in these simulations we have used resource block level Signal to Interference and Noise Ratio (SINR) calculation due to wide bandwidths in the scenario.

A. HetNets Layout

A HetNets simulation layout is used, which corresponds to the "Scenario 2a" considered for the 3GPP LTE Release 12 specifications [13, Appendix A.1.2], [14]. This scenario consists of a set of 21 macro cells and 84 small cells. The macro cells form a non-overlapping hexagonal grid of three sectors per

macro site, while the small cells are organized into clusters of four. That is, within the coverage area of each macro cell a cluster of four randomly located small cells is deployed. The macro cells are wide-area 5G macro cells, operating at a central frequency of 5.9GHz with 20MHz bandwidth and 46dBm maximum transmission power. The small cells operate in the millimeter-wavelength band with central frequency 28GHz, maximum transmission power of 35dBm and 100MHz bandwidth.

B. Channel and Traffic Model

The considered channel model is adopted from the International Telecommunications Union (ITU) guidelines for radio interface evaluation [15]. Specifically, the links between the UEs and the macro cells follow the urban macro model, while the links between UEs and small cells follow the urban micro model [15]. A 2x2 Multiple Input Multiple Output (MIMO) transmission scheme is assumed for all involved links, assisted by rank adaptation [16] as well as interference rejection combining [17].

The UEs are placed at random locations within the simulated area, and remain static until they complete a transmission session, referred to henceforth as "call". After a call has ended, the UEs are dropped to another random location in the area, selected in a uniform fashion. The calls consist of packet flows of configurable size, which varies from 0.5MB to 10MB. The flows are and generated from a Poisson process with a configurable average rate yielding an offered load per macro cell which varies between 20 Mbps and 140 Mbps. UDP transport protocol and downlink traffic are used in the simulations. The effect of the packet flow size and offered load is discussed later in Section V.

C. Multi-Connectivity Implementation

As regards the considered multi-connectivity process, our setup uses a *flow control* algorithm for coordinating the amount of data that flows via the macro and small cells. This flow control mechanism is assumed to be located i) at the macro cells in the case of Architecture (a); ii) at the centralized aggregation point (data center) in the case of Architecture (b) (see Fig. 1).

We adopt the flow control algorithm presented in [14, Eq. (5)], which is based on a request-and-forward mechanism that operates as follows. For architecture (a), the small cells periodically request data from the flow control entity. The amount of requested data depends on the buffer status of the small cells, their past throughput, as well as the volume of pending data requests which have been issued to the macro cells but not yet reached the small cells due to backhaul delay [14]. For architecture (b), flow control operates in a similar manner, yet data requests are issued from both small cells and macro cells towards the flow control entity, i.e., towards the central unit (see Fig. 1).

The flow control process is assumed to be repeated frequently enough to account for channel changes due to small-scale fading. Additionally, the small cell target buffering time (i.e., the parameter θ_S in [14]) is set large enough in order to minimize the probability that the buffer of the small cell becomes empty. In other words, the parameters of the flow control mechanism are set such that there is always enough data at the small cell to be delivered to the UE, thereby ensuring that the overall throughput depends only on the backhaul delay parameters and not in flow-control related configurations.

IV. Throughput Results: Effect of Backhaul Delay

This section contains the throughput results pertaining to the considered architecture options, for different assumptions on backhaul delay. The term "backhaul delay" is used here to refer to the delay caused by the link that connects the macro and small cell with the aggregation points. This corresponds to X2 and F1 interfaces for Architectures (a) and (b), respectively, as shown in Fig. 1. In the results presented in this section, the packet flow size is set to 2 MB, while the average rate of packet flow generation is set to 26.25 flows per second per the whole network, yielding an offered load of 20 Mbps per macro cell.

A. Throughput Distribution

It is apparent from the setup description of Section III that the delay in the backhaul, has a composite effect on the throughput. That is, the backhaul delay affects both the duration of packet delivery and the efficiency of the flow control mechanism. The latter holds since the backhaul delay impacts the time in which data requests are completed, as well as the volume of pending requests, as explained in Section III-C. Next, we quantify such effect by examining the throughput behavior for different assumptions on the backhaul delay, treating the cases of Architecture (a) and (b), as described in Section II, separately.

1) Architecture (a): Distributed HetNet: Figs. 2 and 3 depict the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the application layer throughput for Architecture (a), respectively. The main observation from both figures is that even a slight increase in the backhaul delay induces a substantial reduction in the throughput. Specifically, it is observed from Fig. 2 that increasing the backhaul delay from the ideal case of 0 ms to 1 ms leads to a decrease in the mode value (i.e., the value that is most frequently sampled) of approximately 100 Mbps. Moreover, it is observed from Fig. 3 that while for the ideal backhaul case of 0ms approximately 45% of the simulation time resulted in a throughput larger than 600 Mbps, for a small increase of 1ms in the backhaul delay this simulation time percentage drops to less than 30%.

2) Architecture (b): Cloud-based HetNet: Similarly as for Architecture (a), Figs. 4 and 5 illustrate the PDF and CDF of the application layer throughput for Architecture (b). Overall, we notice a similar effect as that of Architecture (a), in the sense that the distribution of the throughput is considerably affected by even a slight increase of the backhaul delay. In this regard, we notice a degradation of the throughput as the backhaul delay increases gradually from 0 ms to 10 ms; *nevertheless, the overall fallout is lower*. For instance, taking the same example as above, the percentage of simulation time where the throughput is above 600 Mbps is approximately 35% for

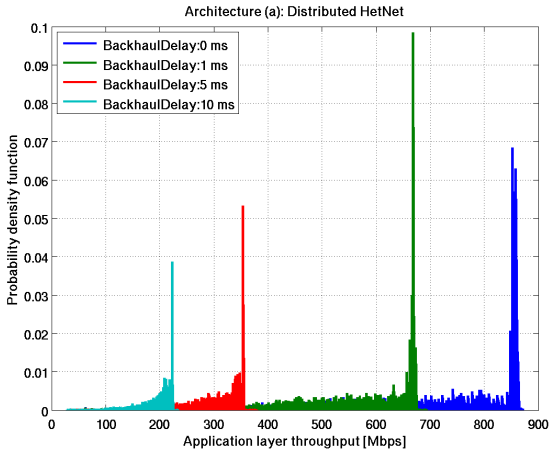

Fig. 2. Application-layer throughput probability density for Architecture (a): Distributed HetNet

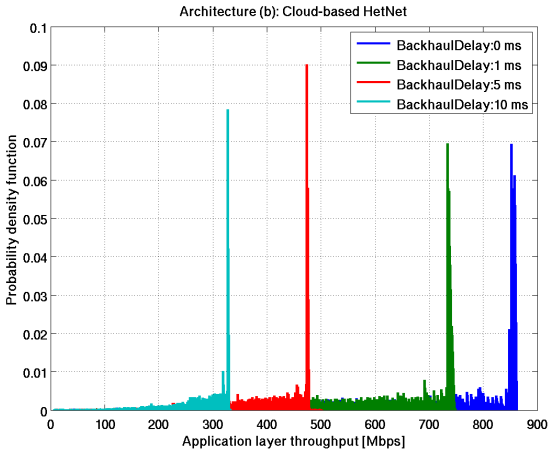

Fig. 4. Application-layer throughput probability density for Architecture (b): Cloud-based HetNet

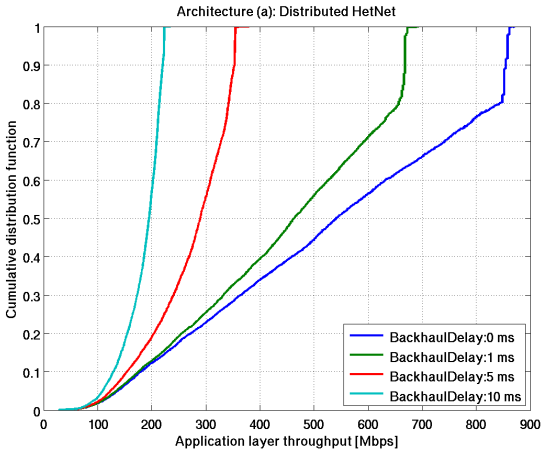

Fig. 3. Throughput cumulative distribution seen at the application layer for Architecture (a): Distributed HetNet

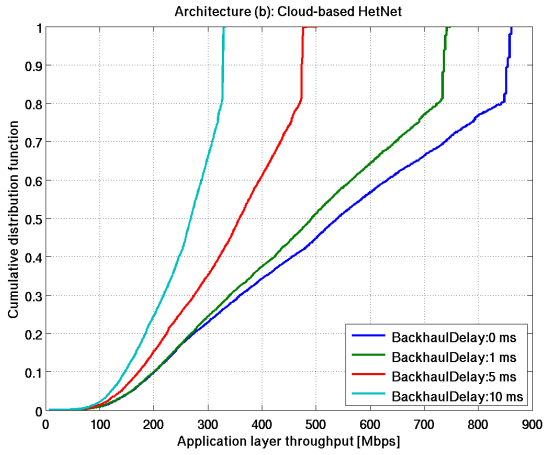

Fig. 5. Throughput cumulative distribution seen at the application layer for Architecture (b): Cloud-based HetNet

1 ms backhaul delay; recall that this value was below 30% for Architecture (a).

With reference to Fig. 1, this mitigated effect of the backhaul delay on the throughput of Architecture (b) stems from the fact that in the cloud-based HetNet architecture the multi-connectivity anchor point is centralized. As a result, the latency between the small cell node and the flow control operation point is lower. Considering that the small cell node is in principle associated with higher capacity than the macro cell node, the cloud-based architecture takes better advantage of the small cells and thus results in higher throughput in case of non-ideal backhaul. In this regard, it is noticed that, as expected, for the ideal case of 0 ms both architecture options yield the same throughput performance.

B. Mean and 5th-Percentile Throughput

Figs. 6 and 7 depict the mean throughput and 5-percentile throughput (that is, the lower 5% of the throughput performance) for the distributed and cloud-based architecture case, respectively. As seen from Fig. 6, the cloud-based architecture scenario (cloud-based HetNet) yields higher mean throughput. This observation is also in line with Figs. 3 and 5, where for any delay values larger than 0 ms higher mean throughput is obtained for Architecture (b) than for Architecture (a).

Similar observations are obtained from Fig. 7, where the comparison is made in terms of the 5-percentile throughput. Interestingly, we notice that as the value of backhaul delay increases, the difference on the 5-percentile performance of the two scheme decreases. This is in contrast to the mean throughput comparison in Fig. 6, where the difference in performance expands for large backhaul delay. This observation is partially explained by the fact that in HetNet deployments the 5-percentile performance mainly comes from the performance of the macro cell. Consequently, high values of backhaul delay have a higher impact on Architecture (b) than Architecture (a), since for Architecture (a) the UP anchor point is co-located with the macro cell, facilitating thus a faster packet delivery.

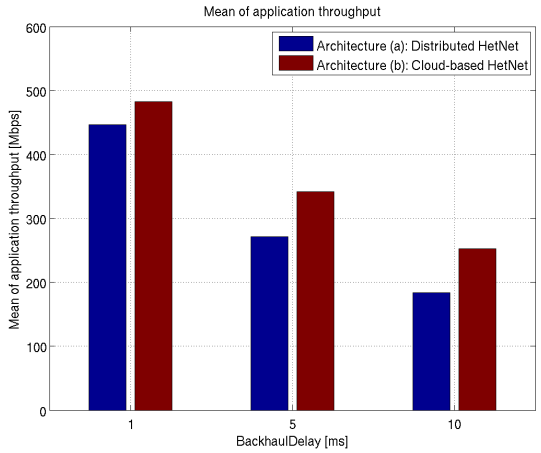

Fig. 6. Mean throughput seen at the application layer for Architecture (a) and Architecture (b), assuming a backhaul delay ranging from 1ms to 10ms

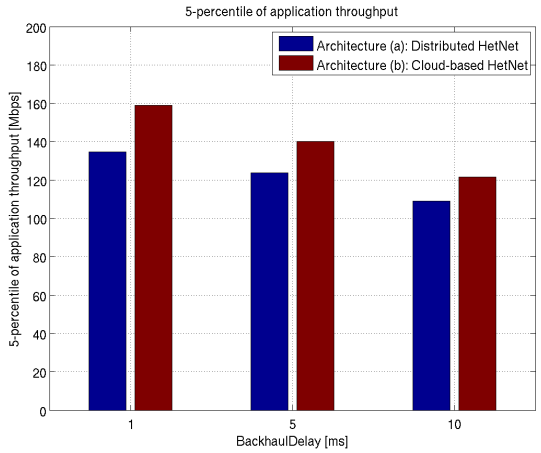

Fig. 7. 5-percentile throughput seen at the application layer for Architecture (a) and Architecture (b), assuming a backhaul delay ranging from 1ms to 10ms

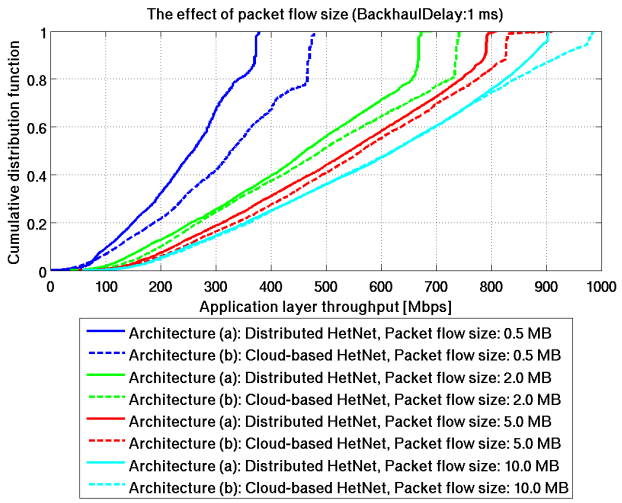

Fig. 8. The effect of packet flow size on the throughput cumulative distribution for backhaul delay 1ms

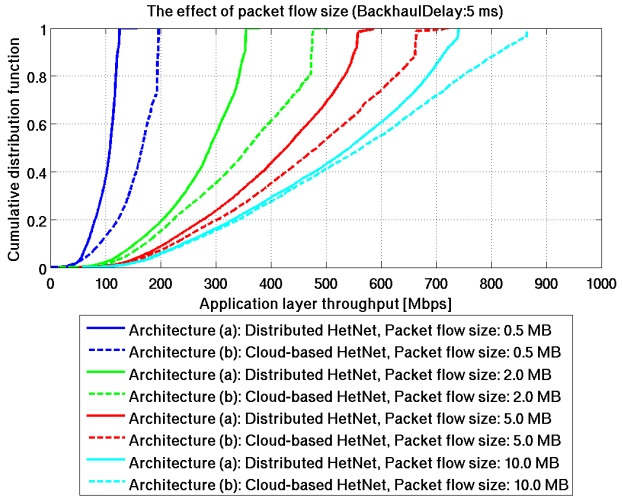

Fig. 9. The effect of packet flow size on the throughput cumulative distribution for backhaul delay 5ms

V. Impact of Packet Flow Size and Offered Load

Besides the backhaul delay, the throughput experienced by the UE is affected by the packet flow size, as well as the traffic load of the corresponding cell that serves such UE. In the remainder of this section, the effect of these two parameters on the throughput performance of Architectures (a) and (b) is investigated.

1) The Effect of Packet Flow Size: The throughput performance of the considered architecture schemes depends on the packet flow size as depicted in Figs. 8 and 9. Two effects need to be considered, which have different impact depending on the employed architecture: for the distributed case, transmission of the packet flow will start immediately after the data arrives at the user-plane anchor at the Macro gNB. However, only the macro layer is used for the first part of the transmission, since the packet flow needs to be split and sent via the backhaul link to the small cell, which adds an additional latency of two times the transport latency before the 100MHz small cell carrier can be used for transmission as well. In contrast, for the cloud-based architecture, transmission starts at both carriers at virtually the same time. This leads to a notable performance gain over the distributed case, notwithstanding the fact that for the cloud-based case, transport latency is counted to the overall transmission time of the packet flow.

We also notice that irrespective of the utilized architecture, larger flow sizes result in higher throughputs until a saturation point is reached. The reason for this behavior is that for larger packet flows the impact of the first phase of the transmission which is dominated either by the macro carrier (for the distributed case) or the backhaul delay (for the cloud case) is becoming smaller in relation to the overall transmission time compared with cases with smaller flow sizes. The effect diminishes for large flows since then the transmission over both carriers is dominating the overall transmssion time of the packet flow.

2) The Effect of the Offered Load: We now discuss the impact of the offered load (defined as the traffic per unit

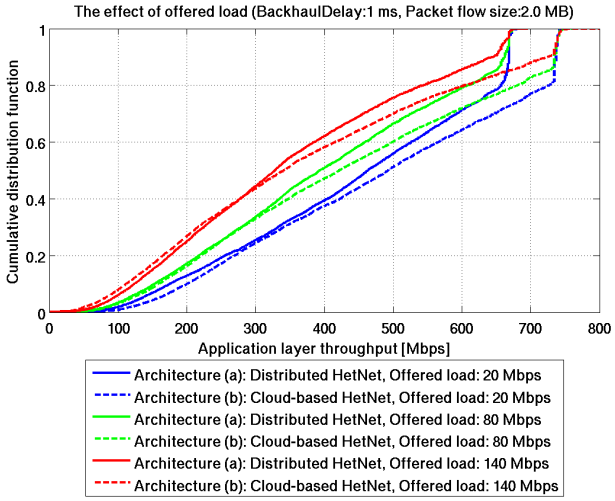

Fig. 10. The effect of offered load per macro cell on the overall throughput for backhaul delay 1ms and packet flow size 2MB

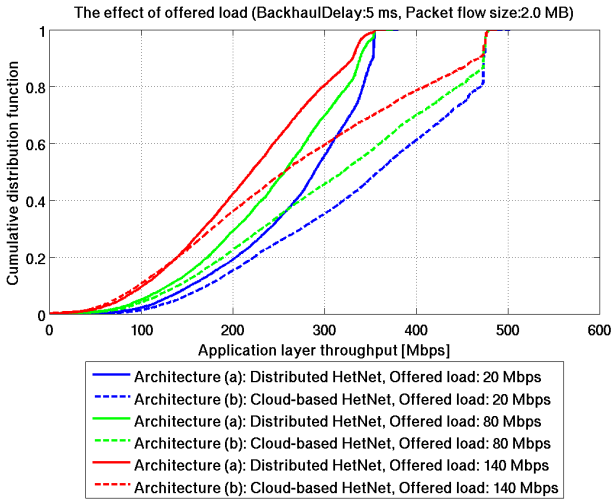

Fig. 11. The effect of offered load per macro cell on the overall throughput for backhaul delay 5ms and packet flow size 2MB

time and unit area) on the achieved per-UE throughput. Two cases are presented with 1 ms (Fig. 10) and 5 ms (Fig. 11) backhaul transport latency, respectively, considering the two architecture options with an offered loads of 20 Mbps, 80 Mbps, and 140 Mbps with a packet flow size of 2 MB.

In all cases there is a notable gap on the maximum achieved throughput between the distributed and the cloud-based architecture. This is again the impact of the initial transmisssion phase on the overall transmission time, which is dominated in the distributed case by the 20 MHz macro carrier. Consequently, for a higher backhaul latency, this gap is significantly larger than for the lower latency case.

The impact of offered load is visible in the throughput gap between low, medium and high loads. In case of a higher offered load, more packet flow transmissions are ongoing in parallel per cell, sharing radio resources. Consequently, the per-UE throughput becomes smaller. This effect is independent of the employed architecture.

VI. Conclusions

We investigated the throughput performance of multi-connectivity in HetNet deployments. In particular, we considered two architecture options, corresponding to a distributed and cloud-based scenario, respectively. The main observations made were a) a noticeable degradation on throughput with the increase on the backhaul delay; b) the packet flow size and offered load substantially impact the throughput performance; c) overall, the cloud-based architecture scenario outperforms the distributed architecture scenario in terms of throughput.

Acknowledgment

This work has been performed in the framework of the H2020-ICT-2016-2 project 5G-MoNArch. The authors acknowledge the contributions of their colleagues. This information reflects the view of the consortium, but the consortium is not liable for any use that may be made of any of the information contained therein.

References

[1] Nokia Networks, *5G use cases and requirements*, White Paper, 2014.
[2] J. Acharya, L. Gao, and S. Gaur, *Heterogeneous Networks in LTE-Advanced*, Wiley Publishing, 2014.
[3] 3GPP TR 36.842, *"Study on Small Cell Enhancements for E-UTRA and E-UTRAN Higher layer aspects (Release 12)"*, Sep, 2014
[4] S. Chandrashekar, A. Maeder, C. Sartori, T. Hhne, B. Vejlgaard and D. Chandramouli, *"5G multi-RAT multi-connectivity architecture"*, IEEE International Conference on Communications Workshops (ICC), 2016
[5] A. Ravanshid *et al.*, *"Multi-Connectivity Functional Architectures in 5G"*, IEEE International Conference on Communications Workshops ,2016
[6] M. Giordani, M. Mezzavilla, S. Rangan and M. Zorzi, *"Multi-Connectivity in 5G mmWave cellular networks"*, Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), 2016
[7] D. Aziz, J. Gebert, A. Ambrosy, H. Bakker and H. Halbauer, *"Architecture Approaches for 5G Millimetre Wave Access Assisted by 5G Low-Band Using Multi-Connectivity"*, IEEE Globecom Workshops, 2016
[8] D. S. Michalopoulos, I. Viering, and L. Du, *"User-Plane Multi-Connectivity Aspects in 5G"*, International Conference on Telecommunications (ICT), 2016
[9] F. B. Tesema, A. Awada, I. Viering, M. Simsek and G. Fettweis, *"Evaluation of Context-Aware Mobility Robustness Optimization and Multi-Connectivity in Intra-Frequency 5G Ultra Dense Networks"*, IEEE Wireless Communications Letters, vol. 5, pp 608-611, Dec 2016
[10] A. Wolf, P. Schulz, D. Öhmann, M. Dörpinghaus, and G. Fettweis, *"Diversity-Multiplexing Tradeoff for Multi-Connectivity and the Gain of Joint Decoding"*, submitted to IEEE for possible publication. available: https://arxiv.org/abs/1703.09992, Mar 2017
[11] Next Generation Mobile Networks (NGMN) Alliance, *"Whitepaper on Small Cell Backhaul Reuirements"*, 2012
[12] W. Guo, S. Wang, X. Chu, J. Zhang, J. Chen and H. Song, *"Automated small-cell deployment for heterogeneous cellular networks"*, IEEE Communications magazine, vol. 51, pp 46-53, May 2013
[13] 3GPP TR 36.872 V12.1.0, *"Small Cell Enhancements for E-UTRA and E-UTRAN - Physical layer aspects (Release 12)"*, Dec, 2013
[14] H. Wang, C. Rosa and K. I. Pedersen, *"Inter-eNB Flow Control for Heterogeneous Networks with Dual Connectivity"*, IEEE Vehicular Technology Conference (VTC Spring), 2015
[15] ITU-R M.2135-1, *"Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced"*, Dec, 2013
[16] T. Jonna, S. K. Reddy and J. K. Milleth, *"Rank and MIMO mode adaptation in LTE"*, IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2013
[17] M. Lampinen et. al, *"System Level Modeling and Evaluation of Interference Suppression Receivers in LTE System"*, IEEE Vehicular Technology Conference (VTC Spring), 2012
[18] 3GPP, *"TS 38.401 V15.0.0; NG-RAN; Architecture description"*, Dec., 2017