Copyright Notice

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

Multiplexing Gains Achieved in Pools of Baseband Computation Units in 4G Cellular Networks

Thomas Werthmann*, Heidrun Grob-Lipski† and Magnus Proebster*
*Institute of Communication Networks and Computer Engineering, University of Stuttgart, Germany
Email: thomas.werthmann@ikr.uni-stuttgart.de
†ALCATEL-LUCENT DEUTSCHLAND AG, BELL LABS GERMANY
Email: heidrun.grob-lipski@alcatel-lucent.com

***Abstract*—The tremendous increase of mobile user traffic load within the last few years forces us to efficiently use the wireless network and processing resources. Cloud computing and virtualization techniques offer an exciting opportunity to considerably reduce operation costs and provide flexible and dynamic systems. In this paper we present a simulation study for a cloud base station, which concentrates baseband processing functions of multiple radio sites. There we focus on the multiplexing-gains induced from user load and traffic heterogeneity. Our simulation results show that the data traffic influences the variance of the compute resource utilization, which in consequence leads to significant multiplexing gains if multiple sectors are aggregated into one single cloud base station. In addition, the spatial user distribution has a high impact on the compute resource load. These findings should be taken into account for the assessment of multiplexing gains in real networks.**

I. Introduction

During the last few years, the Internet as well as mobile terminals like smartphones and tablets have reached the mobile networks [1]. The numbers of mobile broadband subscribers using at least 3G bitrates have dramatically grown and between 2011 and 2012 the global wireless traffic has increased by 70 percent [2]. Applications like social networking and video have become popular and lead to new consumption paradigms and growing traffic demands within wireless networks [1], [2]. This fast increase in user traffic requires additional compute and transmission resources and network operators are confronted with expensive investments and high operation costs for radio access systems. Studies show that this trend will intensify due to the increasing number of 4G connections, which generate in average much more traffic than non-4G connections [2].

To satisfy the traffic demands and maintain or even improve the operators' economic perspective, intelligent systems and mechanisms are required, which support an effective use of the wireless network and compute resources. Emerging technologies like cloud computing and flexible sharing of resources through virtualization enable us to offer adaptive and dynamic systems. Simultaneously, they are able to reduce the operation costs significantly.

In [3], the authors propose the Centralized RAN (C-RAN) concept as a further development of the Radio Access Network (RAN). The approach moves the base station into the cloud and separates the radio units. By centralizing compute resources, the number of sites for baseband processing can be reduced considerably. This approach provides a concentration of base station functions, which also reduces backhaul connections and lowers maintenance costs. Unlike conventional radio access systems, where each antenna has dedicated compute resources, the mobile cloud system dynamically assigns compute resources to Remote Radio Heads (RRHs).

Conventionally, a base station is dimensioned to process maximum busy hour traffic for the respective radio cells. However, typical cellular deployments consist of cells of various sizes with heterogeneous traffic characteristics with different peak traffic loads. In addition, there are movements over time from areas at the periphery to the center of the cell cluster and vice versa [4]. These traffic variations in time and area hold a considerable potential for multiplexing gains by pooling compute resources.

The cloud base station concept introduced in [5] constitutes the base of the investigations in this paper. The future cloud-based RAN architecture has been derived to support flexible pooling on user respectively bearer level. This means that the dedicated user processing per user or per radio bearer in uplink and downlink are virtualized. If required, e.g. in order to reduce blocking, the computational effort for the dedicated user processing is offloaded to remote processors within the same Multi-Site/Multi-Standard Base Band Unit (MSS-BBU) or even to processors in remote pools.

Figure 1 depicts the future RAN architecture comprising multiple RRHs connected with high speed optical links to the associated MSS-BBU. Each MSS-BBU comprises several Base Band Units (BBUs). Multiple MSS-BBUs can be interconnected with each other via high speed optical links, e.g. via eX2, an enhancement of the X2-interface. In each MSS-BBU, a Distributed Cloud Controller (DCC) decides whether the processing of a bearer can be performed by a BBU within the MSS-BBU or by a BBU of a neighbor MSS-BBU.

In this publication, we present a detailed multi-layer model which describes data traffic, user distribution, mobile radio network, and compute resource requirements. Based on this model, we evaluate by simulation the short-term pooling effects induced from user load distribution and traffic heterogeneity within one cluster. The traffic and the corresponding compute effort vary over time and load peaks do not occur simultaneously in different cells. From the simulations we derive how this multiplexing gain scales with the number of aggregated cells. In addition, we investigate the influence of the spatial user distribution on the utilized compute resources.

Similar studies as ours have been performed in [6] and [7]. Both evaluate compute resource utilization with different

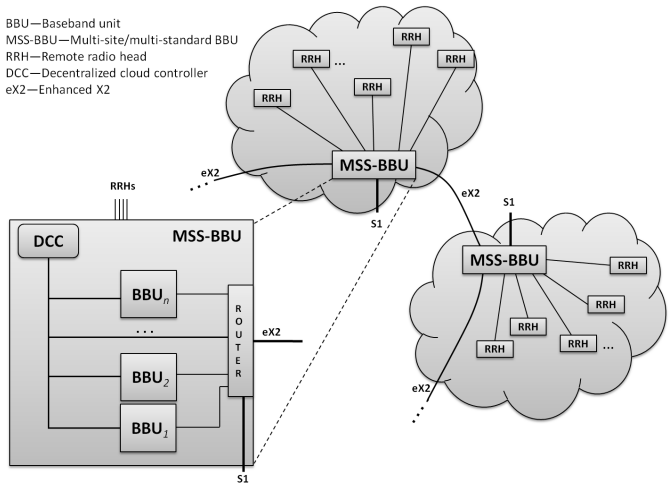

Fig. 1. Future cloud-based RAN architecture.

models for the compute effort. To assess the influence of the traffic, the authors of [6] use traffic measurements from a real Wideband Code Division Multiple Access (WCDMA) network. However, the data throughput could be measured only at coarse intervals of 15 minutes. This limits the level of detail available for the evaluation. Besides they do not provide any insight into the distribution of users over the area. The authors of [7] use the NGMN traffic model [8]. While focusing on the architecture, they did not investigate the influence of the number of aggregated sectors. In addition, they simply modeled a homogeneous user distribution.

In Section II, we give an overview of the user system model, comprising the user load distribution, the data traffic model, and the radio network model. Furthermore, we introduce our computation resource model derived from [9]. Then, Section III gives the simulation results for the load operation point in dependence of the offered traffic and user distribution. Finally, we conclude the paper in Section IV.

II. Evaluation Methodology

The aim of our studies is to analyze the possible multiplexing gain for RAN compute resources due to pooling of baseband resources. Therefore, we need to capture all influencing factors on the compute resource usage, namely the user and traffic model, the radio network model, and the compute resource usage model. With this model, we can then evaluate the multiplexing gain. Beyond that it may also be used to optimize the position of MSS-BBUs and the allocation of cells to MSS-BBUs. In our simulations, we assume unrestricted compute resources and study their utilization. Given that the operator accepts a certain probability that the capacity of the radio system is limited by compute resources, we are able to estimate possible savings of compute resources by looking at percentiles of the resource utilization.

We assume a 10 MHz LTE system. The base stations are placed in a hexagonal arrangement of 19 sites. Each base station supplies three sector cells, resulting in 57 sectors. We apply wrap-around to avoid border effects. In our simulations, we concentrate on downlink transmissions. Although uplink also causes high computational effort at the base station, it behaves similar as the downlink direction. For each evaluation, we compute a total of 10000 s (about 2.8 hours) of simulated time. To avoid transient effects, we prepend 500 s of transient time and do not utilize the output from this phase. Such long simulation times are required to eliminate distortions introduced by large downloads which require significant transmission time.

A. User and Traffic Model

The traffic model has a large influence on the resource usage. Opposed to a full buffer assumption, real Internet traffic is bursty and has a heavy-tailed object size distribution [10]. This leads to a fast-varying number of active users within a cell. As the traffic in the individual cells contributes to the behavior of the aggregated traffic a BBU has to handle, it is important to have a proper model for the per-user traffic demands.

We model traffic as pairs of request and response objects. This covers many of today's Internet applications. The objects are transmitted as quickly as possible, i.e. there is no rate limitation introduced by the sender. As we concentrate on the downlink in this publication, the uplink objects are not discussed further. Our model is based on the assumption that the network load is caused by a high number of independent users. The Inter-Arrival Time (IAT) of these request-response pairs follows a negative exponential distribution and is used to control the offered traffic in the system. We use an object size distribution measured on a campus link [10]. To avoid problems arising from very large objects, we clip the distribution at 10^8 bytes. Thereby, we cut off a part of the heavy tail of the distribution. However, objects above this size contribute only 0.7 % of the traffic volume.

Besides the traffic properties, in a cellular system, the location of the users is also important. To build an efficient network, an operator has to adapt the cell density to the traffic load per area. This results in an approximately equal load per cell. However, the planning is inaccurate, and the setup of new cells may be delayed to save cost or because of regulatory issues. In addition, the load typically changes over day. As a consequence, the cell density does not always match the user distribution. We model this with a configurable non-uniform user distribution by choosing the user location from a combination of a uniform and a non-uniform distribution. Uniform means that users are evenly spread over the whole area. The non-uniform part models a hot-spot situation, where the user density follows a normal distribution in the x- and y-dimension with the mean placed at the center of the scenario. We fixed the standard deviation of these normal distributions to 350 m, resulting in a broad hot-spot covering the area of the central site. By adjusting the ratio between the uniformly and the non-uniformly placed users, we can vary the unevenness in the user distribution. In the simulations, the user distribution is parameterized by the proportion of uniformly placed users. The remaining users are placed according to the normal distribution. To avoid spending a high number of resources for users with very low channel quality, we drop requests originating from mobiles which have an average SINR below -3.9 dB. This results in an outage of about 5 %.

In order to simulate changing user locations, each request originates from a new user with a new location. During

TABLE I. SYSTEM MODEL PARAMETERS

Property	Value
Cellular layout	19 sites, 3-sectors per site, wrap-around
Inter BS distance	500 m
BS TX power	46 dBm
UE TX power	23 dBm
BS/UE height	32 m / 1.5 m
Path-loss [dB]	$128.1 + 37.6 \cdot \log_{10} d\,[\mathrm{km}]$, from [12]
BS Antenna model	2D, 70° beamwidth
Shadowing	8 dB log-normal
UE velocity	0 km/h; for fast fading model: 3 km/h
Carrier frequency	2 GHz
System bandwidth	10 MHz
Frame duration	1 ms
Min. SINR	-3.9 dB

transmission, the users do not move. After the user has finished his transmission, he leaves the system. Note that, as users with low channel quality need more time to transmit their requests, the density of active users is higher at the cell edge. We apply a simple admission control, which drops arriving requests when there are more than 100 users active in the sector.

For our scenario, we are interested in the effects at the network layer and below. Therefore, we idealize transport layer effects and assume that both, the request and response objects arrive as a whole at the BBU respectively User Equipment (UE) buffers.

B. Radio Network Model

Besides the user and traffic model, also the radio network model is important to determine the required compute resources for a cell. For the radio propagation, we consider path-loss and shadowing. The parameterization of the radio propagation is summarized in Table I and complies with 3GPP specifications. From the transmit power and the signal degradation between all transmitters and the receiver as well as the noise level, we determine the mean Signal-to-Interference-and-Noise-Ratio (SINR) of a user.

With our system level simulation, we want to look at effects on time scales of hundreds of seconds. Therefore, due to the computational complexity, it is difficult to model multipath-propagation. Instead, we use the model in [11] to consider fast-fading and frequency-selective scheduling with the commonly known proportional fair scheduler. This model uses the number of active users and their respective mean SINR to determine an effective SINR diversity gain. With the enhanced SINR, we derive the possible rate on the channel according to LTE Modulation and Coding Scheme (MCS). For this, we use Block Error Rate (BLER) tables generated from link layer simulations, including two Multiple-Input-Multiple-Output (MIMO) modes. Above an SINR of about 4 dB, we use 2x2 spatial multiplexing MIMO. At lower channel qualities, we apply Space-Frequency Block Coding (SFBC). We assume ideal channel knowledge at the base station and apply a target decode probability of 80 %. Failed transmissions are reinserted into the sending buffer after 8 ms.

C. Computation Resource Model

From the traffic and radio network models, we know which radio resources are actually in use and which transmission mode has been chosen (e.g. MIMO-diversity, MCS). With this, we are able to determine the required computational resources per user, per cell and for a whole BBU with the computation resource model described in the following.

In [9], the authors provide a detailed model of the power consumption in base stations of different sizes (macro, micro, pico, femto). They distinguish the components and functionalities of base stations in downlink and uplink. For the components, [9] gives the power budget. For the functionalities, the Giga Operations Per Second (GOPS) are defined per function block and how they scale with load and transmission mode.

We use this model as a baseline and extend the model according to our needs. Especially the numbers for computational complexity are of interest. First, we concentrate on the compute resources for physical layer calculations (Frequency-Domain processing (FD) and Forward Error Correction (FEC) in their publication). We extend the model by the separation of antennas and spatial MIMO layers: The computational effort for FD still scales with the number of antennas. However, we assume that the effort for FEC scales with the number of MIMO layers. This means that the effort for FEC is not significantly increased when a transmission uses SFBC instead of Single-Input-Single-Output (SISO).

The authors of [9] use a reference system and specify all input parameters relative to the reference values. In contrast, we define the compute effort as a function of the absolute values of the input parameters. To this end, we divide the absolute values by the respective values of the reference system. Then the following equation describes the compute resource effort in GOPS $P_{u,t}$ that is required to serve UE u at time t:

$$P_{u,t} = \left(30A_{u,t} + 10A_{u,t}^2 + 20\frac{M_{u,t}}{6}C_{u,t}L_{u,t}\right) \cdot \frac{R_{u,t}}{50} \quad (1)$$

where A is the number of used antennas, M the modulation bits, C the code rate, L the number of spatial MIMO-layers and R the number of Physical Resource Blocks (PRBs), each as allocated to UE u at time t. Note that the baseline model from [9] is based on an example implementation. The model needs to be adapted for other implementations.

To derive the compute resource load for a sector or for multiple sectors aggregated in an MSS-BBU, we sum up the load of the respective users $u \in U$:

$$P_{U,t} = \sum_{u \in U} P_{u,t} \quad (2)$$

III. SIMULATION RESULTS

In this section, we analyze the compute resource usage. We assume unrestricted compute resources and evaluate their utilization, determined by the user traffic, channel conditions, and bandwidth resources.

In the following, we first derive operation points for the later analysis, then we estimate the compute resource utilization for a system with uniform user distribution and finally we study the scenario with a non-uniform user distribution.

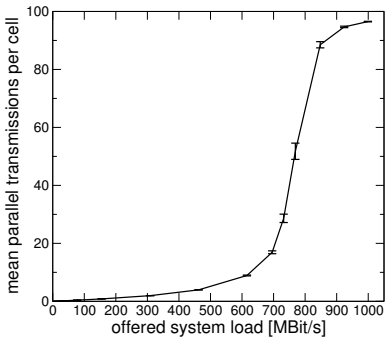

Fig. 2. Number of parallel transmissions per cell over total system load.

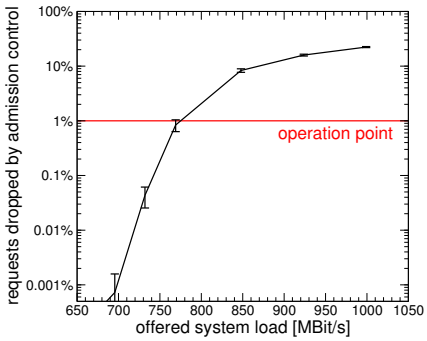

Fig. 3. Proportion of requests dropped by admission control over total system load.

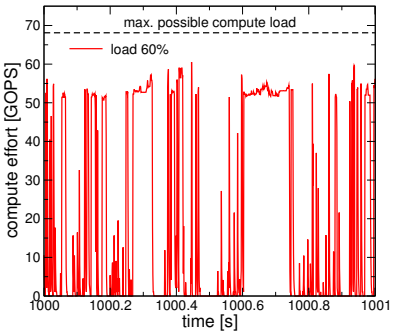

Fig. 4. Compute effort caused by a single sector, for 60% system load.

A. System Load Operation Point

Depending on the user distribution, the system can handle different maximum loads. If the users are distributed uniformly, load is equally spread over all sectors, so that the system capacity is high. When users are placed in the hotspot, the load concentrates in the central site. For a certain load, the central site becomes overloaded, and the admission control starts to drop a significant number of requests. We think that these conditions do not correspond to reality. Therefore, we use a different load operation point for each user distribution, which will be defined in the following paragraphs.

When the system load increases, the probability rises that multiple users are concurrently active in the same sector. This allows the system, for example, to utilize the channel diversity and thereby acquire a scheduling gain. However, as the users have to share resources with other users, the data rate achieved by each user decreases. This results in a longer duration for the users' file transfers, which in turn raises the likelihood that additional requests arrive before the previous ones are completed. Without admission control, this could lead to a system which becomes unresponsive under high load.

In the following figures, we depict the system behavior for a uniform user distribution. For other user distributions, the results are similar but the system capacity is lower. Figure 2 shows the number of parallel transmissions in dependence of the offered load. At about 700MBit/s the offered load reaches the system capacity which leads to a sudden increase in parallel transmissions, as individual transmissions last longer. For an offered load above 800MBit/s, the number saturates, because admission control drops new requests when there are already 100 ongoing transmissions. This can also be seen in Figure 3, which shows the ratio of dropped requests on a logarithmic scale. Our aim is to define the load operation point where the admission control has no noticeable effect on the simulation results. For each user distribution, we set our load operation point to 1% dropping ratio. This operation point is denoted as 100% system load in the following. For each user distribution, we run a preliminary simulation with a simple control loop steering the offered load to the desired admission control drop ratio.

Table II shows the resulting load levels for different configurations of the user distribution. For later simulations, we also use partial loads, which we define in relation to the respective load operation point. Note that 100% load does not directly correspond to the load where all PRBs just become occupied, because (a) the load is not constant and (b) even if all PRBs are used, more users could still be handled due to the gain from channel dependent scheduling.

TABLE II. LOAD OPERATION POINTS FOR THE USER DISTRIBUTIONS.

uniform users [%]	system load [MBit/s]
100	769
80	553
60	423
40	340
20	282
0	242

B. Compute Resource Usage for Uniform User Distribution

In a first evaluation, we investigate a uniform user distribution. In each sector, compute effort is spent to serve transmissions to the UEs. Due to the web-like data traffic behavior, a sector is typically either empty or all PRBs are spent to serve one or more UEs as quickly as possible. There are only some situations where few PRBs are sufficient to transmit a small object or the remainder of a larger one.

When a sector is nearly empty, only few compute resources are used. When all PRBs are occupied, the compute resource requirement is mainly determined by the used MCS and MIMO mode, which depends on the channel quality. Therefore, there are two major factors which influence the compute effort: data traffic and channel quality. Variations in the traffic load result in strong fluctuations of the compute effort per cell. Compared to that, the variations introduced by the channel quality are smaller. An exemplary trace of the resource utilization of a single sector is shown in Figure 4. The maximum possible compute effort of 68.1 GOPS is defined by the compute resource model (see subsection II-C) and the highest available MCS and MIMO mode. Typically, this is not reached even if all PRBs are used, because spatial multiplexing MIMO and higher order modulations schemes are rarely used.

Figure 5 shows a trace of the compute effort required by all 57 sectors in our simulation scenario for different load configurations. As expected, compared to a single sector trace, the sum of 57 sectors is more stable. When the system load increases, the compute effort approaches a level of about 3050 GOPS. For high load scenarios, high interference is caused for most transmissions, which limits the usage of higher order modulation schemes and spatial multiplexing

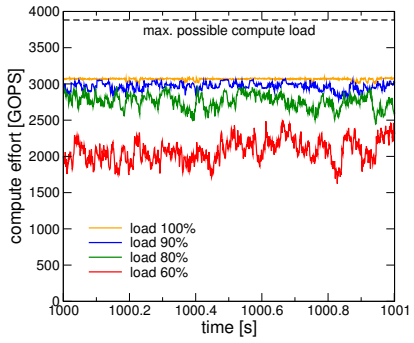

Fig. 5. Trace of the sum of the compute effort caused by 57 sectors.

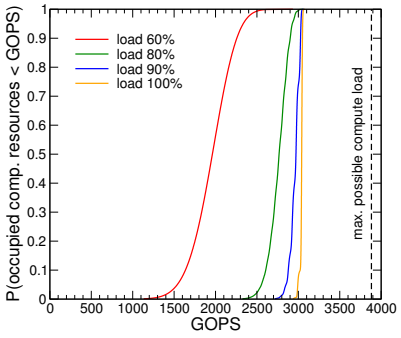

Fig. 6. CDF of the sum of the compute effort caused by 57 sectors.

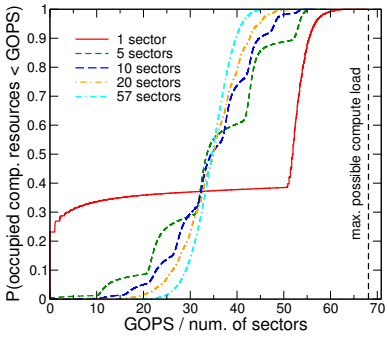

Fig. 7. CDF of the sum of the compute effort caused by different numbers of sectors, for a system load of 60%. (x axis normalized)

MIMO. This prohibits higher compute resource utilization. Figure 6 depicts the Cumulative Distribution Function (CDF) corresponding to the traces in Figure 5, which shows the same effect.

We now want to study how the aggregation of more sectors smoothes the resulting compute effort. By pooling multiple BBUs in one MSS-BBU, the overall load can be balanced. This reduces the variations of the compute load and thereby allows for a tighter dimensioning of the hardware resources. In this publication, we assume that the compute resources are homogeneous, i.e. there are no quantization effects and the load can be accumulated ideally.

Figure 7 depicts the CDF of the compute load for different numbers of aggregated sectors at a system load of 60 %. The x axis is normalized to the compute load of a single sector. As discussed above, a single sector is typically either nearly empty or uses all available PRBs. The latter results in a compute load depending on the MCS, typically between 50 and 60 GOPS. This is visible for the single sector case: In 23 % of the Transmission Time Intervals (TTIs), it does not use compute resources at all and in about 60 % of the TTIs, the sector uses 50 to 60 GOPS. Concentrating the compute load of more sectors reduces the variance, because it is unlikely that several sectors are simultaneously empty or that all sectors use all available PRBs. For small numbers of aggregated sectors, discrete states (full/empty) of the contained sectors are visible.

The dimensioning of the compute resources can be derived from a percentile of the compute load. E.g. an operator could decide to accept service degradation due to hardware limitations in 1 % of the TTIs. The hardware would then be dimensioned according to the 99 %-ile of the compute load, which can be derived from Figure 7. Compared to the same number of separate sectors, the aggregation of five sectors would gain 9 %, 20 sectors would gain 20 % and 57 sectors would gain 27 %. Note that, as we use a wrap-around scenario, interference causes a correlation of load even across the border of the scenario. In reality, such a correlation would not be present. This could further increase the multiplexing gain.

From these simulations, we have seen that the variance of the compute resource usage is mainly influenced by the data traffic behavior. Beyond this, the aggregation of multiple sectors in a single MSS-BBU can lead to significant multiplexing gains for the compute resource utilization.

C. Compute Resource Usage for Non-Uniform User Distribution

Figure 8 shows the CDF of the sum of the compute effort caused by 57 sectors for different user distributions. For each configuration, the system is operated at 60 % system load (see Table II). It is clearly visible that the compute load is much higher for the uniform distribution. However, this is mostly caused by the different load of the network. For a uniform user distribution, the system can handle over three times more load than for the configuration with 0 % uniformly placed users. To separate the effects, we performed two distinct evaluations, which are discussed in the following paragraphs.

First, we scaled the x axis of the curves in Figure 8 by the offered load used for the respective simulation. The outcome is shown in Figure 9. Even after scaling, 100 % uniformly placed users result in the highest compute load per transmitted bit. For more hotspot users, the average load is lower but its variance is higher. This can be explained by the different interference conditions: For the uniform user distribution, the average SINR per received PRB is 5 dB. Caused by reduced interference from lower loaded neighbor cells, the average SINR per received PRB is 9 dB for 0 % uniformly placed users. A lower SINR causes longer transmissions, resulting in an increased computational effort per transmitted bit. In contrast, a higher SINR allows to use spatial multiplexing MIMO and higher order modulation schemes, which contributes to the variance of the compute load.

In a second evaluation, we simulated all configurations of the user distribution with the same offered load of 242 MBit/s, which is feasible for all user distributions. For a uniform user distribution, this results in a medium load per cell. For the non-uniform configurations, the cells at the hotspot location are highly loaded while the remaining cells are even lower loaded than in the uniform case. Figure 10 shows the resulting CDF of the compute effort. Here, the highest compute effort is caused by the system with 0 % uniformly placed users. For more uniform distributions, the average load reduces but variance increases. This again can be explained by the SINR, which is 10.4 dB in average for the uniform distribution and 6.4 dB in average for the 0 % uniform case. These simulations have shown that the spatial distribution of the load also influences the compute resource usage. If multiplexing gains have to be evaluated in real networks, the user distribution should be taken into account.

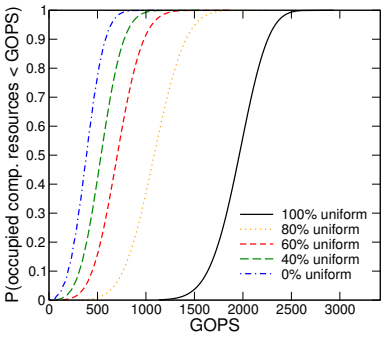

Fig. 8. CDF of the sum of the compute effort caused by 57 sectors, at 60 % of the respective load operation point.

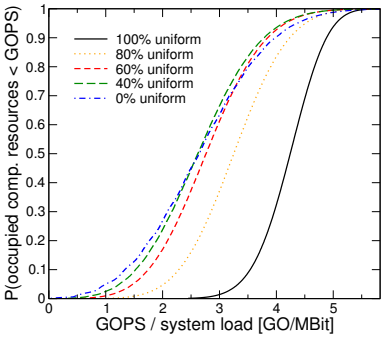

Fig. 9. CDF of the sum of the compute effort as in figure 8, scaled by the offered system load.

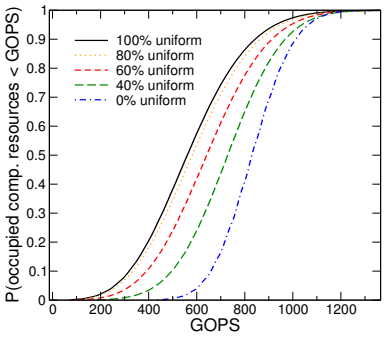

Fig. 10. CDF of the sum of the compute effort caused by 57 sectors, at a load of 242 MBit/s.

IV. Conclusion

In this publication, we have described a simulation model which is capable of capturing the effects from data traffic, user distribution, and radio transmissions. With our simulations we have shown that in typical scenarios, the compute resource utilization is limited to about 80 % of the theoretical maximum, which is mainly caused by the channel conditions. We also evaluated how the multiplexing gain increases when more sectors are aggregated in a single MSS-BBU. While the combination of five sectors already saves 9 %, the aggregation of 57 sectors saves more than a quarter of the compute resources. Finally, we have shown that the user distribution has a strong influence on the utilization of the compute resources. In future work, we plan to increase the size of the scenario and integrate smaller cells, so that the model better resembles typical urban network layouts. In addition, we will investigate how the load balancing between multiple adjacent MSS-BBUs can be achieved and what gains can be expected from this.

Acknowledgement

The authors would like to sincerely thank Dipl.-Ing. Bernd Haberland and his Bell Labs Mobile Cloud team for the valuable discussions and contributions.

References

[1] F. Pujol, "Mobile traffic forecasts 2010-2020 & offloading solutions," May 2011. [Online]. Available: http://www.ict-befemto.eu/fileadmin/documents/publications/workshop_2011/F._PUJOL_IDATE_15_05_2011.pdf

[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017," 2013. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[3] China Mobile Research Institute, "C-ran: The road towards green ran," Tech. Rep., 2011, version 2.5, Oct. 2011. [Online]. Available: http://labs.chinamobile.com/report/view_59826

[4] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, "Mobile landscapes: using location data from cell phones for urban analysis," *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, vol. 33, no. 5, p. 727, 2006.

[5] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, "Radio base stations in the cloud," *Bell Labs Technical Journal, General Papers Issue*, vol. 18, no. 1, June 2013.

[6] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "Cloudiq: a framework for processing base stations in a data center," in *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012, pp. 125–136.

[7] L. Guangjie, Z. Senjie, Y. Xuebin, L. Fanglan, N. Tin-fook, Z. Sunny, and K. Chen, "Architecture of gpp based, scalable, large-scale c-ran bbu pool," in *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE, 2012, pp. 267–272.

[8] R. Irmer (ed.), "Radio access performance evaluation methodology," NGMN White Paper V1.3, Jan. 2008.

[9] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. Gonzalez, H. Klessig, I. Godor, M. Olsson, M. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of lte base stations," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, April, pp. 2858–2862.

[10] F. Hernández-Campos, J. S. Marron, G. Samorodnitsky, and F. D. Smith, "Variable heavy tails in internet traffic," *Perform. Eval.*, vol. 58, no. 2+3, pp. 261–261, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2004.07.008

[11] J. Ellenbeck, J. Schmidt, U. Korger, and C. Hartmann, "A concept for efficient system-level simulations of ofdma systems with proportional fair fast scheduling," in *GLOBECOM Workshops, 2009 IEEE*, 30 2009-dec. 4 2009, pp. 1 –6.

[12] 3GPP, "Physical layer aspect for evolved universal terrestrial radio access (utra)," 3rd Generation Partnership Project (3GPP), Tech. Rep. 25.814, Oct. 2006. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25814.htm