Approaches to Adaptively Reduce Processing Effort for LTE Cloud-RAN Systems

Thomas Werthmann
Institute of Communication Networks and Computer Engineering
University of Stuttgart, Germany
Email: thomas.werthmann@ikr.uni-stuttgart.de

Abstract—**Cloud-RAN is a novel architecture for LTE networks, where antennas with only limited processing capabilities are deployed in the field and all baseband and higher layer processing of the base stations is pooled in a central office. It has been shown that the centralization leads to multiplexing gains for signal processing hardware. When a system is able to efficiently cope with overload of the compute resources, significant savings are possible. This allows to install less resources (saving Capital Expenditure, CAPEX) and to switch off more resources during low-load periods (saving Operational Expenditure, OPEX). In this publication, the resource allocation of a system of LTE base stations is formulated as an optimization problem. The formulation includes the reduction of interference by not using radio resources for transmission and the flexibility of using different MIMO modes. The results of optimization runs are evaluated. They show that about 50 % of the compute resources can be saved without impacting the system performance. In overload situations, only 20 % of the resources are sufficient to deliver 87 % of the system performance. The most efficient approach to reduce processing effort is to adapt the MIMO mode and to reduce the number of virtual transmit antennas.**

I. Introduction

To handle the growing mobile communication demands, more and smaller cells have to be deployed. Cloud Radio Access Network (Cloud-RAN) is a novel architecture for 3GPP Long Term Evolution (LTE) and LTE-Advanced (LTE-A) networks, which has been proposed by IBM [1] and China Mobile [2] to make this growing infrastructure more efficient. The guiding idea of the Cloud-RAN architecture is to split the classical base station (BS, also called eNodeB) into a Remote Radio Head (RRH) and a BaseBand Unit (BBU). The RRH consists of AD/DA converters, power amplifiers and antennas. It is connected via high data rate interconnection with the BBU. The BBU performs all baseband processing (e.g., calculation of FFT, en- / decoding). In addition, it terminates higher layer protocols. The BBU is connected to the core network to forward user and control data.

By splitting antennas and baseband processing, the BBUs of multiple cells can be centralized. This centralization makes coordination between cells easier and simplifies maintenance. In addition, cloud concepts can be used to realize multiplexing gains: The BBUs of multiple cells can be executed by a shared pool of hardware units. Hardware resources not required for one cell (e.g., if the cell does currently not transmit data) can be used to serve other cells. We have shown that a significant amount of resources can be saved even at high load periods by exploiting the fluctuations in the data traffic demands and the load imbalance between the cells [3]. Additionally, during periods with reduced network load (e.g., at night), a part of the resources in the pool can be shut down to save energy [2]. Also, in the case of hardware failures, the remaining units of the pool can take over the load of the failed unit.

An orthogonal approach is to use General Purpose Processors (GPPs) instead of special hardware like custom ASICs, FPGAs, and DSPs for the baseband processing [4]–[7]. Although the implementation of complex signal processing algorithms on such processors is challenging, it has some advantages. Development is expected to be cheaper, because off-the-shelf hardware can be used and the main effort is software programming. In addition, generic hardware allows for flexible upgrades to support new technologies and standard versions.

A. Motivation

Compared to specialized hardware, the numeric compute power of GPPs is expensive. The compute power should therefore not be over-dimensioned. The compute effort depends, besides others, on the data traffic of the users, the channel conditions, and the transmission modes used by the system. In a sufficiently large pool, the probability that all cells are occupied with transmissions to users with ideal channel conditions is low. Typically, users have only average channel conditions, and some cells do not utilize all resources for transmission because there is no user traffic. In [3] we have shown that there is a significant difference between the typical load and the theoretical peak load. It is therefore desirable to dimension the processing power for a typical load situation, but not for the theoretical peak load.

Whenever the provided compute resources are not sufficient to handle the theoretical peak load, the system has to be able to cope with overload. This is the case if the resources have not been installed, but also if they have been switched off during low load periods and cannot be reactivated immediately. Efficient handling of overload is also useful in case of hardware failures or when the load grows over time to levels which have not been anticipated during initial rollout of the network.

Overload can dissipate after a short time, e.g., a couple of milliseconds, if it is caused by random fluctuations of the user traffic. It can also hold on for longer time, e.g., when a public event takes place in the area served by a BBU pool. While in a web cloud system some of the tasks would implicitly take more time to complete, this is not an option for a Cloud-RAN platform. Instead, overload has to be treated explicitly. The more efficient the system can cope with overload, the tighter the resources can be dimensioned without overly degenerating the network's performance.

978-1-4673-6305-1/15/$31.00 ©2015 IEEE

B. Contributions

This publication deals with the compute requirements for downlink physical layer calculations in pooled BBUs. We evaluate different approaches to reduce these requirements without significantly degrading the mobile network performance. We present a system model and an optimization problem which can be solved in feasible time and allows to study the approaches under general conditions. From the solutions of the optimization problem we derive that adapting the Multiple-Input-Multiple-Output (MIMO) mode is the most promising approach to reduce compute requirements. By applying this approach, the system can maintain 84 % of the normal throughput when the available compute resources are reduced to 20 % of the resources required for theoretical peak load.

C. Related Work

There are many publications coping with the reduction of algorithmic complexity in communication. Two examples are [8], which treats efficient MIMO decoding, and [9], which discusses MIMO precoding. While designing efficient algorithms can reduce complexity, we here focus on dynamic savings of complexity. We assume that an overload situation is not the typical point of operation of a network. Therefore, we accept moderate performance degradation in this situation, while we want to keep high performance in the other cases.

The dynamic reduction of computational effort has also been investigated in literature. On the terminal side, [10] scales the accuracy of the Fast Fourier Transformation (FFT) calculations to the requirements which are defined by the modulation scheme. In [11], the authors investigate terminal power efficiency of Single-Input-Single-Output (SISO) and MIMO, including scalable turbo and MIMO decoders. Li et al. [12] propose a controller to adapt algorithmic accuracy to the users' requirements and thereby reduce processing effort. The power efficiency of SISO and MIMO modes at the BS is compared by [13]. There, the authors come to the conclusion that MIMO schemes with lower number of transmit antennas result in more efficient operation. However, none of the mentioned publications regards a pool of baseband units and the effects caused by interference between BSs in this pool. In this publication, we want to study the complexity of baseband computations in a general approach by solving an optimization problem. In contrast to studies which reduce the long-term average power consumption, we cope with a hard restriction of the compute effort.

D. Outline

This paper is structured as follows. Section II introduces the system model used in our studies. In section III we discuss three approaches to cope with overload. Thereafter, we define our optimization problem in section IV. In section V we then present the results of the optimization and insights gained from those results. Section VI concludes the publication.

II. System Model

A. Mobile Network Model

Our model of the mobile network mainly follows the 3GPP specifications [14]. The configuration of the model is specified in table I. To simplify the optimization problem, we do not regard frequency-selective channel effects. Instead, we assume that the channel characteristics of a single subcarrier apply to the whole bandwidth. On each Physical Resource Block (PRB), each BS either serves a single User Equipment (UE) or the PRB is left free (i.e. no multi user MIMO).

TABLE I. System model parameters

Property	Value
Cell layout	7 tri-sectorized sites, 500 m distance, wrap-around
BS / UE height	32 m / 1.5 m
Carrier frequency	2 GHz
System bandwidth	10 MHz (50 Physical Resource Blocks)
BS TX power	46 dBm
Path-loss [dB]	$128.1 + 37.6 \cdot \log_{10} d$, with d: distance in km [14]
Shadow fading	8 dB log-normal
MIMO channel model	3GPP Spatial Channel Model [15]
BS antenna model	3D, 15° tilt [14]
BS / UE antennas	8 / 4 cross-polarized with 0.5 λ distance

The transmitter uses precoding vectors as defined by 3GPP for closed loop precoding. At the receiver, we use zero-forcing to decode the MIMO signal. We assume ideal channel knowledge for the selection of the precoding vector and the MIMO processing at the receiver.

A transmission is interfered by other BSs transmitting simultaneously on the same PRBs. We do not want to look at interference alignment and similar approaches. Therefore, for interference calculation we assume that each interferer uses a random precoding vector.

We combine the Signal to Interference and Noise Ratio (SINR) values of multiple layers by using Mutual Information Effective SINR Metric (MIESM) to a single SINR value per codeword, roughly following the method described in [16]. Assuming ideal channel knowledge, we map this SINR value to a Modulation and Coding Scheme (MCS) with the help of Additive White Gaussian Noise (AWGN) block error tables taken from [17]. This MCS directly corresponds to a data capacity per PRB. To account for overhead, we assume that 3 Orthogonal Frequency Division Multiplex (OFDM) symbols per subframe are used for the Physical Downlink Control Channel (PDCCH). In addition, we assume that further 18 resource elements per PRB are used for cell specific reference symbols. The maximum throughput per cell is therefore 127 MBit/s, that of the whole system 2.66 GBit/s.

In LTE transmission mode 9, the BS is allowed to dynamically change the number of spatial streams and the applied precoding. We make use of this flexibility to reduce the number of independent transmit antennas. To reduce the computational complexity, we allow the BS to combine multiple physical antennas to a virtual antenna by just copying the transmitted signal. Thereby, the BS can use 8, 4, 2, or 1 virtual antenna to transmit data to it's UEs. Precoding vectors are selected from the tables defined for the respective number of antennas. As the UEs use demodulation reference symbols to decode the data, the BS is not required to explicitly notify the UEs of these decisions. The BS can acquire the channel knowledge required to select the precoding vectors by configuring multiple Channel State Information (CSI) processes, which implicates a small overhead. For our evaluations we assume ideal channel knowledge and do not regard the overhead for additional reference symbols and CSI signaling.

B. User and Traffic Model

We place 105 to 420 UEs randomly and uniformly in the whole scenario. Although for simplicity we later specify the (average) number of UEs per cell, there is no cell-specific UE limit. Typically the compute load of a BBU pool would fluctuate due to user actions and movement. However, we do here look only at snapshots where overload is present. We assume that during the evaluated snapshot the UEs don't move and that they do not run out of data, i.e., we apply a full buffer model. Note that this is a worst case for the processing load, because in our model all cells have sufficient traffic demands to transmit on all PRBs.

C. Processing Model

We assume that our system has a single homogeneous pool of compute resources, and thereby disregard the problem of assigning tasks to discrete processing units. We concentrate on the effort required for the physical layer processing in downlink direction. The effort of calculating the inverse FFT (iFFT) is not regarded here, because it is constant and thereby not a promising candidate for realizing multiplexing gains. We use the processing effort model specified in (1), which is based on [18] and [3].

$$P = \frac{R}{10}\left(3A + AL + \frac{1}{3}MCL\right) \tag{1}$$

Equation (1) defines the compute resource effort P in Giga Operations Per Second (GOPS) that is required to transmit R PRBs. Here, A is the number of used antennas, L the number of spatial MIMO-layers, M the modulation bits, and C the code rate. In the following, A and L are referred to as MIMO mode, M and C as MCS. The compute effort for the whole system is equal to the sum over all PRBs of all BSs. Compared to the model in [18], we apply a small modification to differentiate between A and L.

For simplicity, and because the absolute numbers are not relevant for our evaluations, we normalize the compute effort by dividing it by the theoretical peak effort. In our model, the highest processing effort is required for the configuration $A = 8$, $L = 4$, $M = 6$, $C = 0.926$. When each of 21 BSs transmits with this configuration on $R = 50$ PRBs, the resulting peak effort is 6658 GOPS. In the remainder of this publication, we specify compute effort as percentage of this value.

III. Approaches to Reduce Processing Effort

This section gives an overview of the possible approaches to reduce processing effort. In principle, three components have an influence on the processing effort as modeled here:

- The MIMO mode determines the number of transmit antennas A and the number of spatial layers L.
- The MCS determines the modulation bits M and the code rate C.
- The whole effort is scaled with the number of used PRBs R.

To reduce processing effort, each of these components can be modified. However, each component also influences the data rate, so that a reduction of the processing effort goes along with a degradation of the network service.

Adapting the MIMO mode is promising, because transmit antennas and spatial streams have an influence on all terms of the processing effort formula. In addition, with realistic channels the throughput does typically not scale linearly with the number of antennas or streams, therefore the influence on the network performance could be limited.

When adapting the MCS, only the third term in the brackets is changed. However, the data rate scales linearly with modulation bits and code rate. Therefore, this approach will probably not serve to reduce compute effort while preserving a good service quality.

Leaving PRBs free linearly reduces the processing effort as well as the data rate. However, empty PRBs result in less interference for neighboring BSs. By applying Interference Coordination (IfCo), the system could use these resources to serve UEs in neighboring cells which suffer from high interference. This effect could partially compensate the reduced system throughput. It's impact depends on the fairness the network strives to achieve: If users with low channel quality shall be able to realize high data rates, reducing interference becomes fruitful. In contrast, a system which assigns all resources to UEs close to the BSs does not profit from reduced interference.

For the following evaluations, we allow the optimizer to adapt MIMO modes and PRB allocation. We look at different fairness configurations to see whether that influences the decisions. To limit the complexity, and because we regard that approach as not promising, we do not adapt the MCS but always select that one which delivers the best throughput for the given channel conditions.

IV. Optimization Model

The allocation of resources to UEs, it's effect on the interference, and the selection of MIMO modes can be formulated as an optimization problem. The available compute resources, denoted as P_{max}, are a constraint for the optimization. Fairness and throughput are the base for the objective and additional constraints.

The set of all UEs is denoted as U and the set of all BSs as B. $b_u \in B$ identifies the BS serving user u, and $U_b \subset U$ the set of UEs served by BS b. Each UE can be served with different MIMO modes. A MIMO mode is a feasible combination of a number of transmit antennas A and a number of spatial layers L (with $L \leq A$). The set of all MIMO modes is denoted as M. In principle, each user can be served with any MIMO mode, although some modes will yield low throughput for some users (i.e., when the mode has more layers than the respective channel supports).

To model the influence of IfCo, the concept of system states is introduced. A system state s is defined by the set of BSs $B_{s,\text{active}} \subset B$ which transmit on a frequency resource, while all other BSs $B \setminus B_{s,\text{active}}$ do not transmit. On different frequency resources the system can be in a different state. The set of all states S is defined by all combinations of transmitting and not transmitting BSs (the power set of B: $S = \mathbf{P}(B)$). In total, our system has $|S| = 2^{|B|}$ states (here about 2 mio. states).

For each UE u, all BSs except the serving BS b_u can cause interference, i.e., they are the interferers I_u of u with $I_u = B \setminus \{b_u\}$. In principle, each UE has a different SINR for each system state, which determines data capacity and compute effort. For simplification, we only look at the $n_I = 3$ strongest interferers of each UE. We define these as the relevant interferers $I_{\text{rel},u}$, with $I_{\text{rel},u} \subset I_u$ and $|I_{\text{rel},u}| = n_I$. For the SINR calculations of u, all other BSs $I_u \setminus I_{\text{rel},u}$ are assumed to always cause interference, independent of the system state. We define the relevant interferers $I_{\text{rel},b}$ of a BS b to be the union of the relevant interferers $I_{\text{rel},u}$ of all UEs served by b, i.e., $I_{\text{rel},b} = \bigcup_{u \in U_b} I_{\text{rel},u}$.

For each BS b, we take the system states with $b \in B_{s,\text{active}}$. We divide those into state groups, such that for all system states in a state group t of BS b, the relevant interferers of b perform the same action, i.e.,

$$\forall s_i \in t, s_j \in t\colon \quad B_{s_i,\text{active}} \cup I_{\text{rel},b} = B_{s_j,\text{active}} \cup I_{\text{rel},b} \tag{2}$$

The set of state groups of BS b is denoted as T_b:

$$\bigcup_{t \in T_b} t = \{s \in S\colon b \in B_{s,\text{active}}\} \tag{3}$$

The states of a group cannot be differentiated by UEs served by BS b. The SINR, the data rate, and the compute effort are the same for all states in a state group. The grouping thereby serves to reduce the number of variables of the problem.

For each UE, MIMO mode, and state group, the SINR values of all layers are calculated. Then, the MCS (or MCSs if the mode contains two codewords) is selected which provides the highest throughput. The MCS defines the amount of data $D_{u,t,m}$, which can be delivered to a UE u per PRB if MIMO mode m is used and the system is in a state in state group t. Based on the same information, the compute effort $P_{u,t,m}$ required to deliver a single PRB to u is calculated by (1).

The optimization problem has two sets of variables: Assume that R_s is the number of PRBs for which the system is in state s and $R_{u,t,m}$ is the number of PRBs allocated to UE u, where the user is served with MIMO mode m and the system is in one of the states in t ($t \in T_{b_u}$). The optimization problem without fairness requirements is then:

maximize

$$O_{\text{sumrate}} = \sum_{u \in U} \sum_{t \in T_{b_u}} \sum_{m \in M} R_{u,t,m} D_{u,t,m} \tag{4}$$

subject to

$$\sum_{u \in U} \sum_{t \in T_{b_u}} \sum_{m \in M} R_{u,t,m} P_{u,t,m} \leq P_{\text{max}} \tag{5}$$

$$\forall b \in B, \forall t \in T_b\colon \quad \sum_{u \in U_b} \sum_{m \in M} R_{u,t,m} \leq \sum_{s \in t} R_s \tag{6}$$

$$\sum_{s \in S} R_s \leq R_{\text{max}} \tag{7}$$

Here, (4) denotes the sum of all data transmitted by the system. (5) ensures that the allocated compute effort does not exceed the available compute resources P_{max}. (6) makes sure that the variables R_s and $R_{u,t,m}$ are consistent, i.e., the number of PRBs assigned to the UEs in a cell for a state group t does not exceed the total number of PRBs for which the system is in one of the states in t. Finally, (7) guarantees that the system does not use more PRBs than available. R_{max} is configured to be the number of PRBs defined for the respective bandwidth, e.g., 50 PRBs for 10 MHz system bandwidth.

Depending on the desired fairness configuration, the optimization problem is modified by replacing the objective function and / or adding constraints. Note that we apply system-wide fairness definitions. Therefore the fairness also determines the optimal IfCo.

For proportional fairness as defined by Kelly [19], the objective function (4) is replaced by (8). Instead of maximizing the sum rate, which prefers UEs with good channel conditions, the logarithmic weighting ensures that UEs with worse channel conditions can also transmit at a non-zero rate. Here, D_{max} denotes the maximum bits which can be transmitted in a single PRB, assuming that the channel conditions are ideal.

$$O_{\text{propfair}} = \sum_{u \in U} \log \left(\sum_{t \in T_{b_u}} \sum_{m \in M} \frac{R_{u,t,m} D_{u,t,m}}{R_{\text{max}} D_{\text{max}}} \right) \tag{8}$$

In addition, two min. rate configurations are evaluated. The parameter $\rho_{\text{guarantee}}$ is the data rate (in bits per subframe) which is guaranteed for each UE. The original optimization problem with objective (4) is used together with the additional constraint (9):

$$\forall u \in U\colon \quad \sum_{t \in T_{b_u}} \sum_{m \in M} R_{u,t,m} D_{u,t,m} \geq \rho_{\text{guarantee}} \tag{9}$$

The most fair configuration is the max-min fairness. Here, the minimal data rate over all UEs is maximized. We define an additional variable ρ_{min} and add the constraint (10). At some point it is not possible to increase this rate further, but the system still has some degrees of freedom to assign resources to UEs which allow a higher rate. To get defined results in those situations and do not waste frequency resources, the objective of this configuration contains the average user rate multiplied with a small factor ϵ. We do here set $\epsilon = 0.001$, so that the influence of the second term is small but not ignored due to numeric inaccuracies. The objective of the optimization problem is then to maximize (11).

$$\forall u \in U\colon \quad \sum_{t \in T_{b_u}} \sum_{m \in M} R_{u,t,m} D_{u,t,m} \geq \rho_{\text{min}} \tag{10}$$

$$O_{\text{max-min}} = \rho_{\text{min}} + \epsilon \frac{O_{\text{sumrate}}}{|U|} \tag{11}$$

To get a linear optimization problem, we replace the $\log(\bullet)$ in (8) by a piecewise linear approximation. By choosing a sufficient resolution, the influence of this approximation is only marginal. To achieve feasible optimizer runtimes, we allow the optimizer to choose non-integer values for the variables R_s and $R_{u,t,m}$. Note that the given formulation of the problem allows a BS to serve one UE at the same time with different MIMO modes and MCSs. This could be prohibited by introducing additional integer variables, which we have avoided here.

We have decided to limit the evaluated IfCo schemes to those which either do not transmit or do transmit with full power on a resource. Schemes which use additional power levels, like soft reuse [20], could be supported by introducing additional system states. However, this leads to an extremely large solution space. As we are especially interested in reducing the computational effort, we only look at IfCo schemes which completely disable resources.

V. Evaluation

We use ILOG CPLEX to solve the optimization problems. For each parametrization, 20 independent problems (user drops) have been generated and solved. The plotted values correspond to the average over these 20 drops. The plotted error bars show the 95 % confidence intervals calculated via the Student-T test over the drop results.

A. Maximum Effort Depending on Fairness Strategy

In this subsection, we evaluate how the compute resource usage depends on fairness and number of UEs. For these evaluations the processing effort has not been limited ($P_{\text{max}} = \infty$). As in a typical system which is not designed to safe compute effort, the system does always use all transmit antennas, i.e., the set of MIMO modes M is restricted to those which use 8 transmit antennas.

Figure 1a shows the system throughput achieved with different fairness configurations and numbers of UEs. The value for 20 UEs and a minimum rate of 1 MBit/s is missing. For this configuration, some drops are infeasible, which means that the system cannot guarantee the minimum rate for all UEs.

As expected, the highest throughput is achieved with no fairness restriction. As more strict fairness requirements are enforced, the system throughput is lowered. More UEs result in more degrees of freedom in resource assignment. Therefore, more UEs result in increased throughput for most schemes. This effect is most prominent in the configuration with no fairness, because there in each cell the resources can be assigned to the UE with the best channel. When the number of UEs increases, the probability is high that a UE with a good channel is added. The fairer the system is configured, the less influence the number of UEs has on the throughput. For the max-min fair configuration, this effect is inverted: With more UEs, the probability increases that there is one UE in the system with bad channel conditions, which then determines the highest possible minimum rate.

Figure 1b shows the processing effort required to achieve the system throughput plotted in figure 1a. Note that the system is allowed to consume as many processing resources as required to achieve optimal performance. 100 % processing effort would be caused if all BSs transmit on all PRBs with the highest possible MIMO mode and MCS. This effort is reduced if (a) PRBs are not occupied, i.e., left free for interference reduction; (b) the highest MIMO mode is not used, i.e., the channel does not allow to transmit 4 parallel streams; or (c) the highest MCS is not used, i.e., a low SINR forces to reduce modulation or increase redundancy.

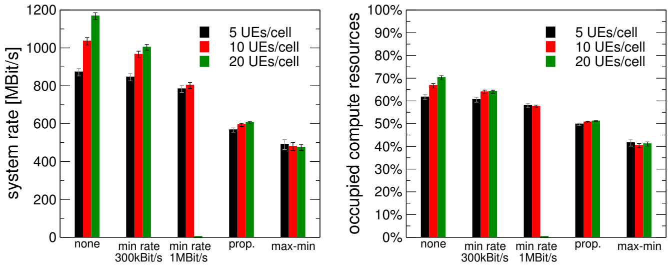

Fig. 1. System rate (a) and occupied compute resources (b) for unrestricted compute resources.

The highest compute effort is caused by the configuration with no fairness constraints and 20 active UEs per cell. There, 99 % of the PRBs are occupied, but only 70 % of the available compute resources are used. This is caused by reasons (b) and (c), i.e., the channel quality does not allow to use the highest MIMO mode and MCS. By adding more UEs, the system would finally use all available compute resources. When the system is fairer, less compute resources are used, because more PRBs are left free for IfCo (only 70 % PRBs are used for max-min fairness and 20 UEs). In addition, more of the occupied PRBs are assigned to UEs with worse channel conditions and do therefore use a lower MCS and fewer MIMO layers.

The outcomes of our first evaluation are: Even if the system is fully loaded (full buffer model) and compute resources are unlimited, the system does not use all compute resources. A fair system uses less compute resources than an unfair system. At maximal fairness, only about 40 % of the compute resources are used, with the widely aspired proportional fairness about 50 %. This shows the large saving potentials of Cloud-RAN.

B. Coping with Limited Processing Resources

This subsection deals with the question how the system performance degrades when the compute resources are restricted. For simplicity, we restrict the evaluated configurations to those with 10 UEs per cell. To be better able to safe compute resources, the system is now allowed to reduce the number of transmit antennas as described in subsection II-A.

Figure 2 plots the throughput achieved with restricted processing resources. As expected, the resources can be limited to the amount required in the unrestricted case without impacting the system performance. Here, the values where degradation starts are slightly lower than those given in figure 1b, because some UEs prefer less transmit antennas even if sufficient compute resources are available. In fact, combining multiple antennas to virtual antennas introduces additional precoding vectors. Depending on the channel matrices, some UEs achieve higher throughputs with these additional precodings than with those defined for 8 Tx antennas.

When the available compute resources are restricted down to 20 %, the system performance worsens with a flat slope. The slope gradually grows steeper at further reduction of the resources, until the rate reaches zero when there are no compute resources available. With only 20 % of the resources, the proportional fair system still transmits with 87 % of the original data rate. This shows that the service degrades gracefully if the compute resource shortage is not extreme.

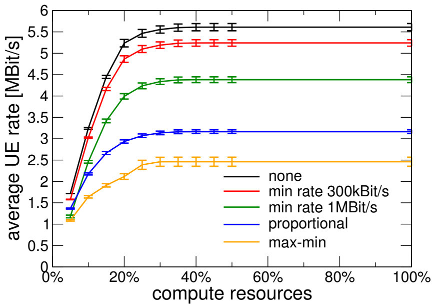

Fig. 2. Average UE rate with limited compute resources for 10 UEs per cell.

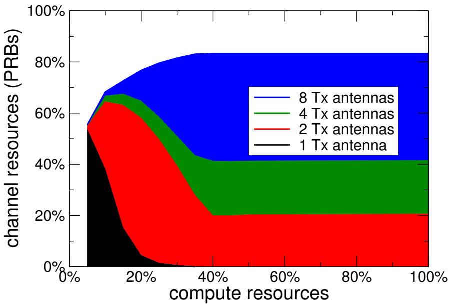

Fig. 3. Shares of PRBs transmitted with different MIMO modes.

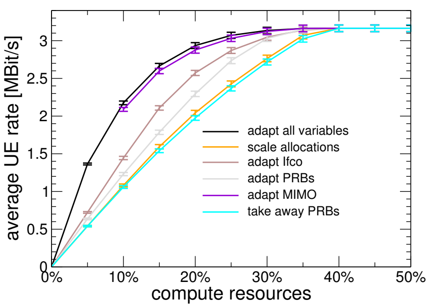

Fig. 4. Average UE rate with different restrictions of adapted variables.

To understand how the system maintains this high performance with few compute resources, we look at the PRB and MIMO mode usage of the proportional fair system in Figure 3. The colored areas depict how many PRBs are transmitted with the respective MIMO mode. For a clear representation, the MIMO modes which use the same number of transmit antennas are combined to one color. The upper outline of the colored areas shows how many PRBs are occupied, i.e., the white area above the plot corresponds to unoccupied PRBs.

For unrestricted compute resources ($>$ 45 %), most UEs are served with 8 Tx antennas. Only some UEs prefer 4 or 2 antennas. In the range between 45 % and 20 % of the compute resources, the number of occupied PRBs is reduced only slightly. In contrast, the shares of the MIMO modes change significantly. While at 45 % most UEs use 8 Tx antennas, at 20 % less than 5 % of the UEs use one of these modes. Instead, the usage of modes with one or two antennas increases from less than 2 % at 45 % resources to 32 % at 20 % resources. This means that it is more efficient to change the MIMO mode than to leave PRBs free and thereby reduce the interference. When the compute resources are restricted to values between 10 % and 20 %, the system performs both measures simultaneously. For some UEs, the losses of reducing the number of Tx antennas are passable. For other settings it is more efficient to reduce the scheduled PRBs, so that some UEs achieve a higher SINR. At 7 % of compute resources, most PRBs already use only one antenna, so further reduction is only possible with leaving PRBs free.

Evaluations for the other fairness schemes show similar behavior (not plotted here). The fairer the system is, the more resources are left free for IfCo even in the case of unlimited resources. However, when the compute resources are reduced, all systems do first switch to less transmit antennas before leaving additional PRBs free.

These results lead to the following conclusions: When the resource shortage is moderate, the system performance degrades gracefully. To achieve this, the optimizer chooses to reduce the transmit antennas. To handle a more extreme resource shortage, it additionally decides to leave PRBs free to reduce interference.

C. The Approaches Applied Separately

In this subsection, we look at different resource saving approaches alone, to decide whether an approach with limited modification of an existing system is sufficient to cope with resource shortages. To do so, we run the optimizer twice for each drop. In the first run, we do not restrict the compute resources and allow the optimizer to adapt all variables. This models a reasonable system configuration which is not planned with a compute resource shortage in mind. In the second run, we restrict the compute resources, but fix selected variables to the values which resulted from the first run. The optimizer then adapts only the remaining variables to cope with the restricted resources. This models a system where only selected components are implemented to react on compute resource shortages. For simplicity, we limited this evaluation to a proportional fair system with 10 UEs per BS.

For this evaluation, we have split the problem in different ways: Fixing IfCo means that the variables R_s are fixed, but the variables $R_{u,t,m}$ can be adapted. Even if free resources are not coordinated (i.e., resources are allocated to a system state which allows a BS to transmit), the inequality in (6) allows to leave PRBs free by not allocating them to any UE. However, in this case UEs served by neighboring BSs do not benefit from the empty resources. To fix IfCo and local scheduling, the variables R_s as well as the terms $\sum_{m\in M} R_{u,t,m}$ are fixed, so that for each UE only the shares of the MIMO modes can be adapted. To fix MIMO, the shares of the MIMO modes for each UE, specified by the terms $\frac{R_{u,t,m}}{\sum_{m\in M} R_{u,t,m}}$, are fixed. When IfCo and MIMO are fixed, the variables R_s and $R_{u,t,m}$ are fixed, but the whole resource allocation of each UE is scaled by a variable factor.

Figure 4 shows the resulting average UE rates. The black curve shows the same values as plotted in figure 2 and serves as reference. The first question is whether it is important to adapt IfCo to a possible compute resource restriction. As dynamic IfCo requires a close collaboration of multiple cell schedulers, it is already a complex task when not considering the compute resources. The red curve shows that when not adapting the IfCo, the performance is only slightly lower than in the case where all variables are adapted.

Also fixing the local scheduling means that the optimizer is not allowed to leave additional PRBs free and also that it is not allowed to reallocate PRBs to different UEs. The result is that the optimizer has to consider fairness for the MIMO mode selection, and cannot compensate an unfair MIMO mode selection by shifting the number of allocated PRBs. The green curve shows that this leads to a small performance degradation below 20 % compute resources. In addition, with this configuration the optimizer cannot find a feasible solution for less than 10 % compute resources at all.

Fixing the MIMO modes results in a significant degradation of the system performance. The yellow curve shows the configuration where only the local scheduling is adapted to the available compute resources. The UE rates decline linearly when the compute resources are reduced to values below 45 %. In contrast to the previous configurations, adapting IfCo does make a difference here. The blue curve shows a better performance in all points, however it cannot achieve the performance achieved with MIMO mode adaptation.

The outcome of this last evaluation is that adapting the MIMO mode selection to the available compute resources is the most efficient way to cope with resource shortages. In addition, an adaptation of the local scheduling is desired for slightly better performance and to be able to handle extreme resource shortages. The complexity of adapting IfCo can be spared. The expected performance gain of leaving PRBs free to reduce compute requirements and interference at once appears only if MIMO is not adapted.

This outcome mainly results from the compute effort model defined in (1), which is dominated by the MIMO mode (terms A and L). However, the used model is rather abstract and not backed by measurements. For a real implementation, performance measurements could be performed to adapt the model. The efficiency of adapting IfCo also depends on the cell layout and on the user distribution, because those influence the impact of interference.

VI. Conclusion

In this publication, we have presented an optimization model which allows to study how a Cloud-RAN LTE system can cope with compute resource shortages. Our evaluations have shown that the required processing effort is typically much lower than the theoretical peak effort. Therefore, it is not efficient to dimension the compute resources for peak usage. The required compute resources depend on the configured fairness. Fair systems do leave more resources free for IfCo and do assign more resources to UEs with a lower channel quality. Thereby they consume less compute resources. The desired fairness should therefore be considered when deciding how many compute resources to install in a Cloud-RAN pool.

Whenever the compute resources are not dimensioned for the worst case, the system has to be able to cope with overload. Adapting the MIMO mode is the most promising approach to realize this. We have shown that, by doing so, the system performance degrades gracefully: A proportional fair system can sustain 84 % of the average UE rate with only 20 % of the resources required for the theoretical peak load.

In future studies, we plan to design a simple heuristic which realizes similar performance. This will also be evaluated in a dynamic scenario, so that effects from traffic dynamics and user mobility can be investigated. Studying the upling direction is also interesting. There, the compute effort is higher. Additional degrees of freedom exist, because in overload situations the decoding can be delayed by a small time.

References

[1] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, Jan. 2010.

[2] China Mobile Research Institute, "C-RAN The Road Towards Green RAN," http://labs.chinamobile.com/report/view_59826, Tech. Rep. v2.5, 2011.

[3] T. Werthmann, H. Grob-Lipski, and M. Proebster, "Multiplexing Gains Achieved in Pools of Baseband Computation Units in 4G Cellular Networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2013.

[4] X. Tao, Y. Hou, H. He, K. Wang, and Y. Xu, "GPP-based soft base station designing and optimization (invited paper)," in *Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on*, Aug. 2012, pp. 49–53.

[5] N. Kai, S. Jianxing, H. Zhiqiang, and K. K. Chai, "LTE eNodeB prototype based on GPP platform," in *Globecom Workshops, 2012 IEEE*, Dec. 2012, pp. 279–284.

[6] H. Zhiqiang, S. Jianxing, D. Ran, and Y. Chen, "Analysis for singal processing development with general purpose processor," in *Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on*, Aug. 2012, pp. 792–796.

[7] J. Weipeng, H. Zhiqiang, D. Ran, and W. Xinglin, "Major optimization methods for TD-LTE signal processing based on general purpose processor," in *Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on*, Aug. 2012, pp. 797–801.

[8] M. Chouayakh, A. Knopp, and B. Lankl, "Low-effort near maximum likelihood MIMO detection with optimum hardware resource exploitation," *Electronics Letters*, vol. 43, no. 20, pp. 1104–1106, Sep. 2007.

[9] M. Aghababaeetafreshi, L. Lehtonen, M. Soleimani, M. Valkama, and J. Takala, "IEEE 802.11ac MIMO transmitter baseband processing on customized VLIW processor," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 7500–7504.

[10] M. Li, B. Bougard, E. Lopez-Estraviz, A. Bourdoux, L. Van der Perre, and F. Catthoor, "The quality-energy scalable OFDMA modulation for low power transmitter and VLIW processor based implementation," in *Global Telecommunications Conference (GLOBECOM), 2007. IEEE*, Nov. 2007, pp. 2894–2898.

[11] C. Desset and R. Torrea Duran, "Reducing the power of wireless terminals by adaptive baseband processing," *Annals of Telecommunications - Annales des Télécommunications*, vol. 67, no. 3-4, pp. 161–170, Apr. 2012.

[12] M. Li, D. Novo, B. Bougard, C. Desset, A. Dejonghe, L. Van Der Perre, and F. Catthoor, "Energy aware signal processing for software defined radio baseband implementation," *Journal of Signal Processing Systems*, vol. 63, no. 1, pp. 13–25, 2011.

[13] F. Cardoso, S. Petersson, M. Boldi, S. Mizuta, G. Dietl, R. Torrea-Duran, C. Desset, J. Leinonen, and L. Correia, "Energy efficient transmission techniques for LTE," *Communications Magazine, IEEE*, vol. 51, no. 10, pp. 182–190, Oct. 2013.

[14] "Further advancements for E-UTRA physical layer aspects," 3GPP, Tech. Rep. TR36.814, Mar. 2010, v9.0.0.

[15] "Spatial channel model for multiple input multiple output (MIMO) simulations," 3GPP, Tech. Rep. TR25.996, Sep. 2003, v6.1.0.

[16] J. Colom Ikuno, "System level modeling and optimization of the LTE downlink," Ph.D. dissertation, Vienna University of Technology, 2013.

[17] C. Mehlführer, J. Colom Ikuno, M. Simko, S. Schwarz, M. Wrulich, and M. Rupp, "The vienna LTE simulators-enabling reproducibility in wireless communications research." *EURASIP J. Adv. Sig. Proc.*, vol. 2011, p. 29, 2011.

[18] 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*, Apr. 2012, pp. 2858–2862.

[19] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[20] N. Himayat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards," *Communications Magazine, IEEE*, vol. 48, no. 8, pp. 86–92, 2010.