Abstract—Massive Machine Communication (MMC) is the next generation of mobile communication (5G) systems requires new Medium Access Control (MAC) and physical (PHY) layer concepts to handle massive access with low overhead. The recently developed concept coded random access is capable of resolving collisions in massive access at MAC layer. Furthermore, Compressive Sensing Multi-User Detection (CS-MUD) achieves joint activity and data detection from the PHY layer perspective by exploiting sparsity in sporadic multi-user detection. In [1], a joint design of these two concepts was semi-analytically evaluated combining CS-MUD with coded random access under certain assumptions, which showed a promising performance for supporting MMC. In this work, a full link-level analysis of the joint protocol will be presented to verify the previous results and provide deep insights on the link-level performance. Although the joint approach on the link-level shows a small throughput loss compared to the semi-analytical evaluation, high system flexibility can be achieved by tuning MAC and PHY resources dynamically based on the numerical analysis which can be referred to in the joint MAC-/PHY-layer design.

I. INTRODUCTION

The expected growth of Massive Machine Communication (MMC) raises new requirements for existing human-driven communication systems. Sporadic transmission is one property of MMC, which means that nodes are only occasionally active for data transmission with low data rates. Instead of high control overhead for access reservation to establish the connection in current cellular systems, direct random access with low signaling overhead and low data rate is more suitable for MMC. New techniques have been proposed to enable direct random access at both Medium Access Control (MAC) [2]–[4] and physical (PHY) layer [5]–[7]. Coded Slotted ALOHA (SA) is a promising coded random access approach for handling collisions in massive access, which is well demonstrated in [3], [4]. From the PHY perspective, Compressive Sensing Multi-User Detection (CS-MUD) as a novel PHY technique enables efficient direct random access and resolves collisions in multi-user detection by jointly detecting activity and data [5], [6].

Recently, it has been shown that these two concepts can be combined to form a new joint MAC and PHY protocol to handle the massive access efficiently [1]. From the MAC perspective, coded SA is able to resolve collisions in massive access by exploiting successive interference cancellation (IC) and control the user activity through broadcast messages in random access, which provides a sparse signal model for CS-MUD at PHY layer. This sparse multi-user detection problem can be solved nicely through CS-MUD exploiting the sparsity at PHY layer and achieving joint activity and data detection to facilitate efficient collision resolution and enable higher user density at MAC layer. In [1], a semi-analytical evaluation has been presented by combining a numerical performance description of CS-MUD with the and-or tree evaluation in coded SA, which has shown achievable optimum performance with ideal MAC and PHY configurations. However, there may be deviations from the semi-analytical evaluation in link-level performance without simplifying assumptions. This paper presents a full link-level analysis of the joint approach and shows how to efficiently distribute the MAC and PHY resources to reach high system flexibility by dynamically tuning the joint protocol while achieving significant throughput gains. For the sake of simplicity, we focus on the PHY performance in interaction with MAC protocol that is not specifically analyzed here.

The remainder of this paper is organized as follows: Section II presents the general system description. Section III introduces how the joint protocol works specifically. The effects of intra-/inter-slot IC are numerically analyzed in Section IV. In Section V, the performance evaluation is presented. The application of the link-level analysis is introduced in Section VI and the paper is concluded in Section VII.

II. SYSTEM DESCRIPTION

Fig. 1 shows that $N$ nodes are synchronized on a slot basis and contend for transmitting their data to the base station (BS) randomly within $M$ slots, called a contention period. The BS initiates the contention period with beacons which control the user activity. Each user is randomly and independently active in every slot, controlled by an activity parameter $\beta$ broadcast with beacons. The user activity can be modeled by different statistic distributions, e.g., Bernoulli distribution or Poisson distribution. Here, we define $\beta$ as the mean of Poisson distribution and all users will be active as Poisson distributed until the BS sends beacons to terminate the contention period.

A. MAC Layer

From the MAC layer perspective, coded SA is applied to handle the collision in a multi-user context which can be represented by the factor graph on the left-hand side in Fig. 1, where $N$ user nodes are on the left side and $M$ slot nodes are on the right. Edges from the same user represent that the user randomly selected slots in the contention period to transmit replicas of the same packet for enhanced collision resolution as indicated by different patterns on the right-hand side. Additionally, each replica contains pointers to all slots accessed by the same user. Therefore, the BS is aware of all the
access information of any node in the contention period once any replica of its packet is successfully decoded. From this perspective the activity parameter $\beta$ also controls the number of replicas. The main idea of coded SA is that once any user is recovered in any accessed slot by the PHY processing, its replicas will be subtracted from the current slot and the other accessed slots as well by successive IC. The coded SA iteratively subtracts replicas of recovered users and reduces the collisions encountered by other users. Therefore, the activity control mechanism of coded SA creates sporadic activity in each slot enabling CS-MUD at the PHY layer.

**B. PHY Layer**

As mentioned above, the average number of active users per slot is statistically controlled by the activity parameter $\beta$, which means that the number of active users per slot is unknown at the BS. At the MAC layer, coded SA is able to control the user activity to create a sparse signal model at the PHY layer, which can be described by a stacked multi-user vector $x_j$. Here, $x_j$ contains symbols from all $N$ user nodes in slot $s_j$ with active users transmitting $L$ symbols from a discrete modulation alphabet $A$ while inactive users are modeled as “transmitting” a frame of $L$ zeros. Hence, the elements of $x_j$ belong to the so-called augmented alphabet $A_0 = A \cup \{0\}$. For simplicity, we restrict our work to BPSK in this paper. Higher order modulations can be easily incorporated as well. The data transmission in any slot $s_j$ from the PHY layer point of view can be described as [5]

$$y_j = A_j x_j + n_j,$$  \hspace{1cm} (1)

where $A_j$ summarizes the influences of the channel and the application of spreading in a CDMA system as an example in this work, which generally can be other medium access scheme as well, e.g., OFDM. The additive white Gaussian noise in slot $s_j$ is represented by $n_j$ with zero mean and variance $\sigma_n^2$ and the vector $y_j$ denotes the receive signal. Due to the sporadic activity controlled by coded SA, $x_j$ turns out to be sparse. Therefore, the sparse multi-user detection problem (1) can be efficiently solved by CS-MUD. Further details of CS-MUD can be found in [5].

In combination, coded SA at the MAC layer determines how many users are accessing a slot via parameter $\beta$, which creates sporadic activity and allows CS-MUD to resolve collisions at the PHY layer. On the other hand, CS-MUD helps coded SA to support more users at the MAC layer by multi-user detection. Overall, both concepts are well matched and can be combined together. Additionally, to reduce the complexity, some assumptions are made as follows without loss of generality: (i) all users are independent and user channels are assumed to be i.i.d.; (ii) channel state information is assumed to be perfectly known at the BS; (iii) the receive power of all active users is the same on average.

**III. JOINT CS-MUD & CODED SA**

**A. Uplink Transmission**

As shown in Fig. 1, the uplink data transmission starts after the nodes receive broadcast messages from the BS indicating that a contention period begins. Every node having data to send will transmit a certain number of replicas of its packet, which is controlled by the activity parameter $\beta$ given by the beacons. The number of simultaneously active users in one slot, represented by $t_A$, follows a Poisson distribution with mean $\beta$. The contention period ends after the data transmission in the $M^{th}$ slot finishes, which is also controlled by BS beacons. There is no feedback in the downlink to the nodes in the contention period, which means that the active nodes have no knowledge of which replica is already recovered in previous slots and will not stop transmitting replicas randomly until the contention period ends.

**B. Base Station Process**

As mentioned before, from the MAC layer perspective, replicas of the same user data packet in coded SA enable successive IC which interacts with CS-MUD to resolve the collisions iteratively in the contention period. Fig. 2 shows the complete processing of MAC and PHY for the receive signal in slot $s_j$, where $y_j$ denotes the corresponding receive signal. The initial phase of the joint protocol is to perform (1) for the first slot, where no recovered user signal is available to perform (3). A cyclic redundancy check (CRC) after decoding determines whether the user is successfully recovered or not. One effect of the CRC block is the validation of packet integrity. Another is to exclude the false alarms for inactive users by CS-MUD. For the sake of simplicity, only perfect IC is considered. If no user signal is successfully decoded in the first slot, the process will continue to the next slot.

**Figure 2:** Process of joint CS-MUD and coded SA in slot $s_j$

The operation at the BS can be summarized as follows:

1) inter-slot IC (3) for receive signal $y_j$ (if any from previous slots);
2) decode remaining user signals by CS-MUD (1) and deliver recovered signals to MAC;
3) intra-slot IC (2) for the current slot $s_j$ (if any from (1));
4) inter-slot IC (4) for all slots $s_k, k < j$ (if any from (1));
5) goto (1) $\forall k \leq j$ until CRC fails.
As shown in Fig. 2, the joint access approach comprises three key components, i.e., intra-/inter-slot IC at the MAC layer and CS-MUD at the PHY layer. A simple toy example generated from the factor graph in Fig. 1 with $N = 4$ and $M = 4$ is shown in Fig. 3 to illustrate that these components, (1), (2), (3) and (4), interact with each other at BS, which helps to understand the numerical analysis of their roles.

IV. Numerical Analysis

In this section, the different roles of intra- and inter-slot IC will be numerically analyzed on the link-level. For the sake of consistency, the system setup is chosen the same as in [1] given in Tab. I.

<table>
<thead>
<tr>
<th>System Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium access</td>
</tr>
<tr>
<td># of nodes</td>
</tr>
<tr>
<td>Activity Parameter</td>
</tr>
<tr>
<td>Length of Contention Period</td>
</tr>
<tr>
<td>Spreading</td>
</tr>
<tr>
<td>Modulation</td>
</tr>
<tr>
<td># of Symbols Per Node</td>
</tr>
<tr>
<td>CS Algorithm</td>
</tr>
<tr>
<td>Channel Coding</td>
</tr>
<tr>
<td>Channel Type</td>
</tr>
<tr>
<td>Channel Impulse Resp.</td>
</tr>
</tbody>
</table>

Table I: Link-level setup of joint CS-MUD and coded SA protocol

Recalling the semi-analytical evaluation in [1], previous results obtained through and-or tree evaluation are based on the factor graph in Fig. 1. The general idea is to iteratively update the so-called capture probability over the corresponding graph edges, which is analogous to the iterative belief propagation erasure-decoding. The capture probability represents the probability of successfully recovering any single user out of a certain number of active users, which usually can be analytically provided by PHY layer processing techniques. However, it is still not feasible to analytically formulate the capture probability of CS-MUD so far. In [1], it has been proven that the capture probability can be evaluated in a combinatorial way using the recovery probability of CS-MUD, which is determined by PHY link-level simulations [1]. The recovery probability of CS-MUD represents the probability of recovering $k$ users out of $t_A$ active users given by

$$p(k|t_A) \approx \frac{\#(\Xi_{k}^{t_A})}{T_{sim}},$$

where $\Xi_{k}^{t_A}$ denotes the event of successfully recovering $k$ users out of $t_A$ active users and $T_{sim}$ defines the number of Monte Carlo trials. Based on Fig. 2, $p(k|t_A)$ characterizes the recover probability for a “clean” CS-MUD system marked as the inner dashed line without (2), (3) and (4). This recovery probability was combined with and-or tree evaluation to achieve the semi-analytical performance in [1], which models the PHY processing as ideal for the joint protocol. However, CS-MUD after (2), (3) or (4) will not averagely perform the same as a “clean” system due to the effects of IC, which motivates a more thorough link-level analysis of the joint protocol.

A. Scenario 1: Intra-slot IC + CS-MUD for Single Slot

As shown in Fig. 4b, we would like to find out how CS-MUD behaves after one iteration of intra-slot IC as shown in
Fig. 4a. With the assumption of \( k \) being the number of recovered users from CS-MUD at the beginning of the single-slot contention period, \( p_{IaC}(k'|t_A - k) \) represents the probability of successfully recovering \( k' \) users from the remaining \( t_A - k \). For the purpose of clear comparison, \( p(k'|t_A - k) \) is the probability of recovering \( k' \) users if we give the same number of \( t_A - k \) active users in the “clean” CS-MUD system.

Fig. 5a shows \( p(k'|t_A - k = 5, 10) \) and \( p_{IaC}(k'|t_A - k = 5, 10) \) using dashed and solid curves respectively, which are statistically evaluated based on Fig. 4a and 4b. With \( t_A - k = 10 \), a “clean” CS-MUD will choose the 10 best users with highest reliability and there will be averagely 8 users correctly chosen with a probability of around 0.3 while for the case of CS-MUD after one step intra-slot IC, the probability goes down to around 0.02. Comparing the results in Fig. 5a, the overall behavior changes completely towards recovering a much lower number of users after intra-slot IC.

To better understand the effect, the corresponding Cumulative Distribution Functions (CDF) w.r.t. \( p(k'|t_A - k) \) and \( p_{IaC}(k'|t_A - k) \) are plotted in Fig. 5b, where a clear shift towards lower number of recovered users after intra-slot IC is shown. The overall shift is caused by the condition of the recovery problem after intra-slot IC, where the active users with best reliability are already recovered. Even with perfect IC, the remaining users are still difficult to be separated. For example, considering GOMP in our link-level simulation, \( k \) users canceled by intra-slot IC are the best \( k \) ones based on the correlations with matrix \( \mathbf{A} \). The remaining \( t_A - k \) naturally show worse correlations compared to a randomly generated “clean” CS-MUD with \( t_A - k \) active users. Overall, the combination of CS-MUD and intra-slot IC already offers huge gains by resolving additional collisions compared to “clean” CS-MUD because a certain number of users haven’t been already recovered beforehand through IC without spending more slots. However, Fig. 5 shows that the assumption of CS-MUD performance in [1] was highly idealized leading to deviations in comparison to the semi-analytical results.

**B. Scenario 2: Inter-slot IC + CS-MUD for Two Slots**

According to Section III, the main goal of inter-slot IC is to enable further CS-MUD and intra-slot IC recoveries in previous slots by adding more slots and to reduce the existing collisions for upcoming slots as shown in Fig. 2. For the former part as analyzed in Section IV-A, the intra-slot IC has already resolved most of the active users with highest reliabilities at every slot, which means that subtracting any of the remaining users may not help the CS-MUD process to recover further users. Consequently, inter-slot IC will not contribute as well as in the semi-analytical evaluation in previous slots. In order to determine whether inter-slot IC generally contributes, we specifically look into the analysis of inter-slot IC in the upcoming slot.

![Figure 5: a). recovery probability given the same number of active users with or w/o intra-IC; b). CDF w.r.t. a).](image)

![Figure 6: Recovery probability \( p_{IeC}(k_2|t_A - k_{sub}) \) of CS-MUD after one-iteration inter-slot IC and \( p(k_2|t_A - k_{sub}) \) for “clean” CS-MUD](image)
throughput w.r.t. different values of the activity parameter which is defined as throughput in the contention period and $T_p$ attention period in Fig. 4c have nearly no difference compared to analyzed given Tab. I with to a decreased performance of the joint protocol. Afterwards, which in this case deviates from the assumption of in Section IV -A.

Intra-slot IC and CS-MUD will behave the same as analyzed newly added slot can be achieved. However, the interaction of intra-slot IC with CS-MUD will not perform recovery probabilities. Therefore, the assumed performance of CS-MUDCS-MUD. Therefore, the assumed performance of CS-MUD in semi-analytical evaluation after applying inter-slot IC in the newly added slot can be achieved. However, the interaction of intra-slot IC and CS-MUD will behave the same as analyzed in Section IV-A.

In combination, the contribution of inter-slot IC highly depends on the performance of intra-slot IC and CS-MUD afterwards, which in this case deviates from the assumption of semi-analytical evaluation in [1] and this deviation may lead to a decreased performance of the joint protocol.

V. PERFORMANCE EVALUATION

In this section, the link-level throughput performance is analyzed given Tab. I with $T^*$ representing the optimum throughput in the contention period and $T$ as the individual throughput w.r.t. different values of the activity parameter $\beta$, which is defined as

$$T^* = \max_{10 \leq M \leq 10} \max_{10 \leq \beta \leq 40} \frac{\#(\text{recovered users})}{\#(\text{slot})}.$$  

A. Link-level vs. Semi-analytical

According to the analysis in Section IV, the interaction of intra- and inter-slot IC with CS-MUD will not perform as well as the one which has been assumed in [1], which leads to throughput loss as shown in Fig. 8. The black solid curve shows the simulated throughput without 2, 3 and 1 in Fig. 2 as the “lower bound”. In this case, additional slots in the contention period will lower the throughput due to the repetition of active users across slots. The red dashed line represents the semi-analytical throughput as the upper bound with both intra- and inter-slot IC applied, where CS-MUD is assumed to perform always the same as the “clean” CS-MUD [1]. Based on the semi-analytical evaluation, the joint protocol can significantly increase the system throughput.

The magenta curve with stars gives the results of optimum throughput. Obviously, a significant throughput gain compared to the lower bound is achieved with the help of IC. Due to the reduced performance of CS-MUD after intra-slot IC, there is still a throughput loss compared to the upper bound. It reaches the maximum with one slot because the repetitions of data packets are hurting the system without inter-slot IC across slots. As analyzed in Section IV, the remaining active users after intra-slot IC in every slot are already the most difficult ones to be separated. Even if some of the users remaining in previous slots are recovered in upcoming slots and subtracted, the newly initiated CS-MUD process can hardly deliver newly recovered users. Therefore, the throughput gain compared to a CS-MUD system without any IC mainly comes from intra-slot IC, which means that the PHY layer processing can improve the system performance more significantly.

B. Spreading vs. Slots

For the purpose of understanding the different roles of MAC and PHY layers, we analyze the performance w.r.t. the spreading resources and the number of slots. The throughput results shown in Fig. 9 except the upper and lower bound correspond to different $\beta$s. As shown in Tab. I, the system is 4 times overloaded with $N = 128$ users and length of spreading sequences $N_s = 32$. However, for small $\beta$, e.g., $\beta = 20, 25, 30$, the maximum throughput can be achieved with one slot. Additional slots will reduce the throughput because the spreading resources are still sufficient to support this low user activity and the contribution of intra-slot IC is more needed than inter-slot IC. Therefore, we can see that we don’t need additional slots to achieve the maximum throughput when there are sufficient spreading resources. As $\beta$ goes higher like $\beta = 35, 40$, the spreading resources become insufficient to support so many active users, which leads to a highly overloaded system and makes the collisions difficult to resolve on the PHY layer alone at a single slot. Now additional slots are required to achieve a higher throughput. If we add another slot, the active users which can not be recovered in the first slot due to the insufficient PHY resources might be active again in the new slot enabling a beneficial exploitation of repetition. When any of these users is recovered from the new

![Figure 7](image1.png)

**Figure 7:** Probability of recovering $k_1$ users in slot $s_1$ and $k_2$ users in slot $s_2$ with $t_A^1 = t_A^2 = h_{sub} = 5$ and 10

![Figure 8](image2.png)

**Figure 8:** Link-level throughput with $10 \leq \beta \leq 40$ compared to the lower and the upper bound
slot, inter-slot IC can reduce the number of remaining users in previous slot and make further PHY layer processing possible. As we can see from the green curve with \( \beta = 35 \), adding another slot brings in inter-slot IC to reduce the collisions in the first slot and help the CS-MUD and intra-slot IC to resolve the collisions in the first slot. Then the throughput reaches its maximum at the 2nd slot. Therefore, inter-slot IC can contribute to the system performance when the spreading resources are insufficient. However, the contribution is not high enough to overcome the effect from the increased number of slots, which leads to the consequence that the maximum achieved with more slots is still lower than before and the upper bound is still not reached.

Overall, the system throughput benefits from both the spreading resources and the number of slots in the CDMA based link-level simulation. The PHY processing mainly contributes with sufficient spreading resources to support relatively low user activity, which is more efficient than spending more time slots at MAC layer. When the user activity increases, one possible way to handle the highly overloaded situation is to spend more spreading resources promoting the PHY layer processing, which can be adapted in next contention period. However, the spreading resources can not be easily changed during a single contention period. Therefore, another option is to bring in more slots, which can be easily managed at MAC layer. With high active user density, more slots could mitigate the system overloaded situation and achieve a maximum throughput.

VI. APPLICATION IN 5G

To support MMC service in 5G, considering a system that has limited resources in terms of frequency and time, e.g., OFDMA in LTE system, the question is how to efficiently spend these resources to achieve the maximum gain. Assuming a multi-carrier scheme like OFDM or P-OFDM with a fixed amount of frequency and time resources for RACH, the PHY processing will efficiently do the work based on the link-level analysis of the joint protocol if the access density stays on the level that the resources are still sufficient. However, an instant increase of access density will lead to insufficient resources like the highly overloaded system in this paper. One option to handle this situation from the PHY layer perspective is to do more spreading over frequency and time resources. Clearly it cannot be easily done when we are in a single contention period as in our joint protocol. As mentioned in Section V, the spreading can be altered in the upcoming contention periods. Due to the limited amount of resources, another option is to introduce more slots on the MAC layer, which will highly reduce the complexity compared to adapting PHY layer processing. Therefore, the overall frequency and time resources can be dynamically allocated by actively controlling the system throughput to achieve the maximum gain. The link-level analysis of the joint MAC- and PHY-protocol points out the direction of system adaption to deal with limited resources in 5G services.

VII. SUMMARY

In this paper, a link-level system of joint MAC- and PHY-design is presented and analyzed to validate [1] and get insights on the “real world” performance of the joint protocol. The link-level analysis shows that sufficient design parameters, i.e., activity \( \beta \) and number of slots \( M \) can be chosen to achieve significant throughput gain and efficiently support MMC. Compared to the previous results [1], there is a throughput loss in link-level simulation due to the decreased performance of CS-MUD. However, in order to achieve maximum throughput, the use of spreading and slot resources are analyzed from the PHY and MAC layer respectively. Besides, the application of the link-level analysis in 5G is discussed. Considering the cross-layer design, the system flexibility can be achieved by dynamically tuning the number of slots in MAC layer and the spreading resources in PHY layer, which can be taken as system design criterion in 5G.

ACKNOWLEDGMENT

This work has been performed in the framework of the Horizon 2020 Project FANTASTIC-5G under Grant ICT-671660, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in FANTASTIC-5G.

REFERENCES


