Adaptive Transmission and Multiple-access for Sparse-traffic Sources

Kaijie Zhou*, Tania Villa*, Navid Nikaein*, Raymond Knopp*, and Ruben Merz†
*EURECOM, France, Email: firstname.name@eurecom.fr
†Telekom Innovation Laboratories, Berlin, Germany, Email: firstname.name@telecom.de

***Abstract*—M2M/online gaming are considered as key applications in LTE and LTE-advanced networks. However, for most of these applications whose traffic is sporadic and some of them require very low latency, they are not well supported by the current LTE and LTE-advanced systems due to the large signaling overhead. This paper proposes two methods to address this problem. The first method provides a co-optimization method for AMC and HARQ when CQI is outdated or unavailable and there is a latency constraint. The second method presents a contention based access method to reduce uplink channel access latency. Simulation results show that with these two methods a significant improvement in spectral efficiency can be achieved while greatly reducing latency or maintaining a latency constraint.**

I. Introduction

High-performance online gaming, and machine-to-machine (M2M) are emerging massive applications for cellular networks. Both applications are expected to create an increasing number of connected devices over the following years, which in case of M2M this number will exceed the human-to-human communications (50 billions machines against seven billion people for 2011). It is predicted that these applications, in addition to conventional voice and Internet traffics, will be an integral part of the traffic transported by LTE [1].

Emerging application scenarios for online gaming class are first-person shooter (e.g. OpenArena), racing (e.g. kart rider), and sports, and for M2M are smart city, e-health, e-vehicle, remote monitoring/control. Analysis of such scenarios has reveled that in majority of cases, packets are sporadic, short and small in number, and uplink dominant. Low-latency is critical in both applications. In the online gaming case, it offers high-level of interactivity and fairness among active players and provides the best game experience as possible, while in the M2M case, it provides very fast reaction time to a realtime event from an overall system latency point of view to prevent potential accidents (e.g. when pressure drops through the gaz/oil pipelines). Low-latency communication is very challenging due to the co-habitation of M2M and online gaming traffic with conventional user traffic coupled with the potential of a rapid increase in the number of machines connected to cellular infrastructure. This is because such systems are primarily designed for a continuous flow of information, at least in terms of the timescales needed to send several IP packets (often large for userplane data), which in turn makes the signaling overhead manageable. Therefore for such a sporadic traffic, further optimizations and cost reduction are needed to lower the signaling overhead and optimize the AMC and HARQ techniques and MAC scheduling.

The work in [2] presents a method to provide QoS guarantees to facilitate M2M applications over LTE (applicable to online gaming). In [3], authors propose an architecture design method for M2M over LTE. Mobility management for M2M with LTE is proposed in [4]. Compared to the related work, our work differs significantly. We propose two different methods to lower the latency and cost of M2M/online gaming over LTE. The first method deals with cases where there is a latency constraint and it concentrates on the HARQ, adaptive modulation and coding (AMC) and physical resource allocation mechanisms. We address cases with no or only outdated channel-quality information (CQI). In such cases, the scheduler must operate blindly with respect to AMC and can only benefit from binary feedback after the first HARQ transmission round (in the form of ACK/NACK signaling). Furthermore, we do not perform any power control because it is impractical or simply not feasible for M2M scenarios. Rather, we develop an optimized rate adaptation policy that changes the number of physical resources (i.e. dimension) across rounds. Our optimized policies are applicable for both downlink and uplink data. With our solution we show that only one bit of feedback (ACK or NACK) of the HARQ protocol is sufficient for significant improvements in packet error rate. Even without any CQI, our results show dramatic error-rate reductions and improvements of the spectral efficiency. To address the cases that further need latency reduction in the uplink channel access, we propose a second method, which is referred as contention based access (CBA). With CBA, UEs are not allocated with specific resources, but rather with a pool of resources where they randomly select for the data transmission. The collision may happen if more than two UEs use the same resource. In this case, dedicated resources are allocated for data retransmissions provided that RNTI of the collided UEs can be correctly decoded based on the MU-MIMO detection technique. The latency gain is therefore achieved by bypassing the scheduling request (SR), and buffer state report (BSR) procedures used in regular scheduling methods.

The rest of the paper is organized as follows. In Section II, we present our joint optimization for HARQ and AMC. In Section III, we expose the method of contention based access. Finally, in Section IV we point out some conclusions.

II. An optimized joint Hybrid-ARQ and AMC Policy

LTE-Advanced schedulers and resource allocation schemes are not optimized for M2M and online gaming traffic. Furthermore, because of the sparse traffic characteristic, of moderate to high mobility, of insufficient uplink CQI periodicity or of inter-cell interference, channel-quality information (CQI) may be outdated or unavailable. In the following, we describe a scheduling and resource allocation mechanism for latency-constrained networks.

A. Basic Idea

We derive analytical expressions, based on mutual information modeling, that capture the throughput performance of latency-constrained networks. We develop an optimized rate adaptation policy. This policy is based on the dynamic adaptation of the number of dimensions (resource blocks) used by each HARQ round which is a feature of the Rel-8/10 LTE coding and modulation subsystem. Surprisingly, this policy can operate with only one bit of feedback from the HARQ process. We also show that additional improvements are obtained when outdated channel-state information becomes available.

B. Modeling and Analysis

Without loss of generality, we consider OFDM signaling. The UL of an LTE system uses an SC-FDMA modulation. Our joint HARQ and AMC policy applies equally. Therefore, for a particular sub-carrier, let x denote the complex-valued transmitted symbol, z denote the additive white Gaussian noise (AWGN), and h denote the channel gain. Both z and h are modeled with a zero-mean and unit variance complex Gaussian random variable. Let l denote the discrete-time index i.e. $x[l]$ is the lth transmitted symbol. Using y to denote received symbols, the lth received symbol in a particular sub-carrier is

$$y[l] = h[l]x[l] + z[l], \qquad l = 1, 2, \ldots, N. \tag{1}$$

We consider a block-stationary Rayleigh fading channel model. Fading remains static for the duration of a HARQ round but varies between retransmissions. The HARQ feedback channel is assumed to be error-free. CQI can be received after each round. However, prior to the first round, CQI may or may not be available. We consider a one-shot transmission model where one transport-block of size N_{TB} arrives in subframe n and must be served at maximum spectral-efficiency under a latency constraint. We denote by N_R the maximum number of transmission rounds. To characterize code performance and the effect of the channel, we use the instantaneous mutual information in each transmission round. Let H_i denote the vector of channel realizations in the ith transmission round. Then $I\left(H_1, \ldots, H_{N_R}\right)$denotes the mutual information accumulated over N_R transmission rounds. In order to compute the mutual information, we assume Gaussian input signals (upper-bound on QAM modulation). For example, let us consider one sub-carrier of a SISO system without interference and let P denote the received power, $h_{0,i}$ is the channel response at round i and N_0 is the noise power, then

$$I(H_1, \ldots, H_{N_R}) = \sum_{i=1}^{N_R} \log_2\left(1 + \frac{P|h_{0,i}|^2}{N_0}\right). \tag{2}$$

Generalizing the notation from [5], the probability of decoding a transport-block in round n with N_j as the number of dimensions used in round j, and R_i, R_j the rate sequences at round i and j, is

$$\Pr\Big(I(H_1, \cdots, H_n) > R_n \sum_{j=1}^{n} N_j, \; I(H_1, \cdots, H_i) < R_i \sum_{j=1}^{i} N_j, \forall i < n\Big). \tag{3}$$

Let $\mu(n)$ denote the target transport-block error probability after n transmission rounds. The latency constraint is expressed by ensuring that the probability that the transport-block is not served after N_R transmission rounds is below $\mu(N_R)$. Under this framework, AMC is the optimization of the rate sequences R_i such that (1) the packet error probability remains below $\mu(N_R)$ after N_R transmission rounds and (2) the spectral-efficiency is maximized. The optimization is carried out as a function of the distribution of $I\left(H_1, \ldots, H_{N_R}\right)$.

We consider the case with two transmission rounds. Let B define the number of information bits to be transmitted. Let N_T denote the total number of dimensions available and let N_1 denote the number of dimensions used in the first round. Hence, the rate in the first round is $R_1 = \frac{B}{N_1}$, and the rate in the second round $R_2 = \frac{B}{N_T}$. We define $\lambda = \frac{N_1}{N_T}$. and we can relate R_1 to R_2 with $R_2 = \lambda R_1$. Let $\bar{R}$ denote the overall spectral efficiency. With $\mu(1)$ as the outage probability after the first round, we have

$$\begin{aligned} \bar{R} &= R_1\left(1 - \mu(1)\right) + \mu(1)R_2 \\ &= R_1\left(1 - \mu(1)\right) + \mu(1)\lambda R_1. \end{aligned} \tag{4}$$

We want to maximize $\bar{R}$ such that the probability of outage after the second round is below the given constraint $\mu(2)$. For the first round, there is no feedback information. The outage probability $\mu(1)$ is unknown but it depends on H_1 and the signal-to-noise ratio (SNR). We can relate R_1 to $\mu(1)$ as:

$$\Pr\left(I(H_1) < R_1\right) = \Pr\left(\log_2\left(1 + SNR|h_1|^2\right) < R_1\right) = \mu(1). \tag{5}$$

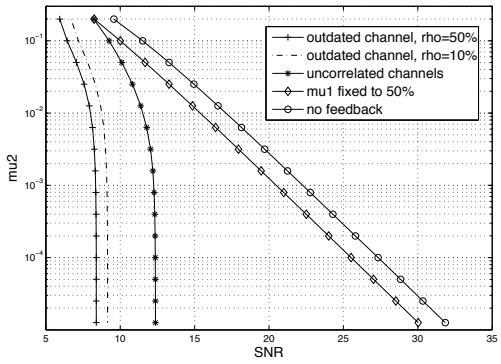
Consequently, we obtain

$$R_1 = \log_2\Big(1 - SNR\ln(1 - \mu(1))\Big). \tag{6}$$

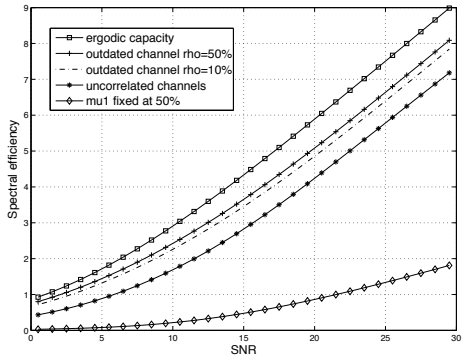
In the second round, feedback about the previous round is available. The outage probability is now given by

$$\Pr\Big(I(H_1, H_2) < R_2 | I(H_1) < R_1\Big) = \mu(2) \tag{7}$$

To find the optimal value of R_1 in the first round, we perform an extensive exploration on $\mu(1)$, given that we want to maximize equation (4) and subject to equation (7).



(a) Probability of outage.



(b) Spectral efficiency.

Fig. 1. We give the results in terms of the probability of outage and spectral efficiency. In (a), for different values of the probability of outage after the second round μ_2, we calculate the corresponding SNR for the different scenarios. And in (b), we give the spectral efficiency versus SNR for the different scenarios..

We also analyze the case when outdated CQI becomes available to the transmitter. We make the additional assumption that the channel remains constant over the two transmission rounds and let $h = h_1 = h_2$. Furthermore, we denote by h_0 the channel value that corresponds to the outdated CQI. In order to model a possible correlation between h_0 and h, we use the following model. Let ρ be the correlation parameter, then $h = \sqrt{\rho}h_0 + \sqrt{1-\rho}h^{'}$, where h_0 and $h^{'}$ are i.i.d. Gaussian-distributed random variables. Note that in this case, $\rho = \mathbb{E}[h_0 h^*]$. In addition, $|h|^2$ is a non-central Chi-square random variable with two degrees of freedom. We follow the same general procedure to obtain the throughput and probability of outage than in the previous cases. However, the spectral efficiency is a function of the outdated CQI and we have to average over the distribution of $|h_0|^2$. The overall spectral efficiency $\bar{R}$ over the two ARQ rounds can be written as

$$\bar{R} = \Pr(h > \gamma_1) R_1 + \Pr(h > \gamma_2, h < \gamma_1) R_2. \quad (8)$$

Where $\gamma_1 = \sqrt{\frac{2^{R_1}-1}{SNR}}$ is the outage threshold in the first round and $\gamma_2 = \sqrt{\frac{2^{R_2}-1}{SNR}}$ is the outage threshold in the second round. Then, $\Pr(h > \gamma_1)$ represents the probability of having a successful transmission in the first round, $\Pr(h < \gamma_1, h > \gamma_2)$ is the probability of being unsuccessful in the first round but successful in the second round, and $\Pr(h < \gamma_2)$ gives the probability of being in outage.

C. Performance Evaluation

In Figure 1, we present numerical results in terms of (1) the probability of outage and (2) the achieved spectral efficiency where we fix the spectral efficiency to 2 bits per channel use.

Figure 1(a) presents the minimum SNR necessary to achieve a given outage probability $\mu(2)$. For a given value of $\mu(2)$, we calculate the corresponding SNR for our rate adaptation policy. For the outdated CQI scenario, we consider a correlation coefficient with the actual channel of $\rho = 10\%$ or $\rho = 50\%$. For comparison purposes, we consider two more cases. First we evaluate a case where we force the probability of outage after the first round to 50%, fixing $\lambda = 0.5$ to make sure that 50% of the dimensions are used in each round. Typically, while conventional systems try to ensure a 10% outage probability per slot, we observe from our results that a higher value gives, in fact, a higher overall spectral efficiency for the case when the number of dimensions in each round is fixed. Second, we evaluate a case where no feedback at all is available i.e. when we can not even receive ACK/NACK from the HARQ process. This highlights the significant gain from adapting the rate across rounds with only one bit of feedback, even in the case without any CQI information. The gain is even higher when only outdated CQI information is available. Our rate adaptation policy gives a zero probability of outage without the need of having a high SNR. From the results in (a) we show that adjusting the dimensions used in each round results in almost causal feedback performance.

Figure 1(b) presents the overall spectral efficiency obtained for a given SNR. We set $\mu(2)$ to 1%. For the outdated CQI case, we consider $\rho = 50\%$ and $\rho = 10\%$. For reference purpose, we also plot the ergodic capacity (Rayleigh channel capacity), i.e. for perfect rate adaptation. Finally, we again consider a scenario when the rate in the first round is chosen so that the probability of outage after the first round is fixed to 50%. This value is chosen because it gives the highest spectral efficiency. Fixing the probability of outage after the first round to more or less than 50% gives a lower overall spectral efficiency. From our results we see a significant improvement in spectral efficiency even in the case without CQI. When we can benefit from outdated CQI, we achieve a performance close to the ergodic capacity.

Another important consideration regarding application on the UL is that the nature of the resource allocation policy is fundamentally related to power control, since we are assuming a constant transmit energy per channel dimension. This is also the adopted policy in LTE (assuming power adjustements are not made during retransmission rounds). Basically, low power is used during the first transmission and significantly more power is used in the second transmission if required.

III. CONTENTION BASED ACCESS

A. Basic idea

To address the problem of the inefficient signalling for uplink access, a new resource allocation signalling method, contention based access (CBA), is proposed. The main feature of contention based access is that the eNB does not allocate resources for a specific UE. Instead, the resources allocated by the eNB are applicable to all or a group of UEs, and any UE which has data to transmit randomly uses resource blocks among the available resources. The mechanism for CBA is as following. First, the eNB sends UEs the resource allocation information for CBA, which costs 0.5 ms provided that the CBA resource is available in each subframe. Then, with the resource allocation information which costs 3 ms for decoding, the UE selects resource randomly and sends packet on it. The latency for this whole procedure is 7.5 ms for the best case, which is much smaller than that of the regular scheduling case.

As the CBA resources are not dedicated but rather allocated for all or a group of UEs, collisions may happen when multiple UEs within a group select the same resource. To address the problem of collision, the following method is used. Each UE sends its identifier, C-radio network temporary identifier (C-RNTI), along with the data on the randomly selected resource. Since the C-RNTI is of very small size, therefore it can be transmitted with the most robust modulation and channel coding scheme (MCS) without introducing huge overhead. With the help of MU-MIMO detection, these highly protected C-RNTIs might be successfully decoded even if they are sent on the same time-frequency resource. Upon the successfully decoding for the collided C-RNTIs, the eNB triggers regular scheduling for the corresponding UEs. Therefore, the overall latency for this whole scheduling procedure is still not larger than that of the regular scheduling. For the collided UEs whose C-RNTIs are not decoded, retransmissions are performed if the regular scheduling information is not received.

B. Performance analysis

Let us denote the total number of resource elements allocated for one CBA transmission as N_{RACH}. This contains the amount of resource elements used for control information transmission, denoted as N_{ctrl} in addition to those reserved for data N_{data}. Therefore, the spectral efficiency of the control information is $R_c = 20/N_{ctrl}\text{bits/RE}$ under the assumption that the control information comprises 20 bits (16 bits for C-RNTI and 4 bits for MCS). Similarly, the spectral efficiency of the data is $R_d = M_{data}/N_{data}\text{bits/RE}$ where M_{data} is the bit of data payload.

For each contention based access transmission, we have the following events: (1) neither the control information nor the data are detected, which is denoted as E_1; (2) the control information is not detected but the data is detected, which is denoted as E_2; (3) the control information is detected but the data is not detected, which is denoted as E_3 and (4) both the control information and data are detected, which is denoted as E_4. In order to determine the probability of each event we take an approach based on instantaneous mutual information. This asymptotic measure yields a lower bound on the above probabilities for perfect channel state information at the receiver. To this end, the received signal on the m^{th} antenna at resource element k is

$$y_m[k] = \sum_{u=0}^{N_u-1} H_{m,u}[k]x_u[k] + Z_m[k], m = 0, \cdots, N_{\text{RX}} - 1 \quad (9)$$

where $H_{m,n}[k]$ is the channel gain for user u at antenna m, $x_u[k]$ is the transmitted signal, $Z_m[k]$ is the noise, and N_u is the random number of active users transmitted on this resource block. The normalized sum-rate for N_u contending users based on mutual information for both data and control portions is computed as

$$I_{\text{X}} = \frac{1}{N_u N_{\text{X}}} \sum_{k=0}^{N_{\text{X}}-1} \log_2 \det \left(\mathbf{I} + \sum_{u=0}^{N_u-1} \gamma_u \mathbf{H}_u[k]\mathbf{H}_u^*[k] \right) \quad (10)$$

where X is represents either control or data, $\gamma_n, n = 0, \cdots, N_u - 1$, is the received signal-to-noise ratio (SNR) and $\mathbf{H}_i[k] = \begin{pmatrix} H_{0,n}[k] & H_{1,n}[k] & \cdots & H_{N_{\text{RX}}-1,n}[k] \end{pmatrix}^T$. The use of this expression requires the two following assumptions. Firstly, all channels can be estimated at the receiver irrespective of the number of contending users. This has to make proper use of the cyclic shifts to guarantee contention-free access for channel estimation. In practice, for loaded cells with only CBA access, this will require association of UEs to orthogonal CBA resources (in time/frequency) and on a particular CBA resource a maximum of 12 contenting UEs can be accommodated. Secondly, the expression assumes Gaussian signals and that the eNB receiver uses an optimal multi-user receiver (i.e. it performs complete joint detection.) These expressions can be found in [6].

Under these assumptions, the probability for events are: $\Pr(E_1) = \Pr(I_{ctrl} < R_c, I_{data} < R_d)$, $\Pr(E_2) = \Pr(I_{ctrl} < R_c, I_{data} > R_d)$, $\Pr(E_3) = \Pr(I_{ctrl} > R_c, I_{data} < R_d)$, and $\Pr(E_4) = \Pr(I_{ctrl} > R_c, I_{data} > R_d)$. In general, the control information is more protected than the data, i.e., $R_c < R_d$, so $\Pr(E_2) \approx 0$. The probability for a packet being delivered only after n transmissions is

$$p_0(n) = \begin{cases} \Pr(E_4) & n = 1 \\ \Pr(E_1)^{(n-2)} \Pr(E_3) + \Pr(E_1)^{(n-1)} \Pr(E_4) & n \in (2, M) \end{cases} \quad (11)$$

where *M* is transmission limit. The average latency is then $T = \sum_{n=1}^{M} T_n p_0(n)$, where T_n is the time for a packet delivered with n transmissions.

C. Performance evaluation

To evaluate the performance of the proposed CBA method, the CBA is compared with two other methods: (i) regular scheduling with round-robin PUCCH allocations for scheduling requests, (ii) the method proposed in 3GPP TR 37.868 where the PRACH is configured with index 14 such that the PRACH with 64 preambles is available in each subframe (largest amount of resources). Regarding CBA method, the

DCI with format 0A is transmitted in each subframe, therefore CBA can be performed in every subframe. Furthermore, perfect power control is assumed yielding $\gamma_0 = \gamma_1 = ... = \gamma_{N_u} = \gamma$. The packet arrives uniformly over a period of 100 ms. Moreover, the packet size is assumed to be of small size, following exponential distribution with average packet size of 200 bits.

The results obtained from simulations using the regular scheduling method and the RACH methods are compared with the average latency T and shown in Fig. 2. The SNR γ is set to 5dB, and the number of receiving antennas is 2. For simplicity we have assumed a line-of-sight dominant channel model with randomized angle-of-arrival at the eNB receiver in order to model the $\mathbf{H}_i[k]$. It can be seen that the latency with CBA is much lower than that of regular scheduling and RACH method, which also implies the throughput of our methods is higher than the other two methods. Moreover, it is found that the performance of CBA depends on the rate of control information R_c. When the number of users is 750, R_c=0.2 achieves the lowest latency of 6.1ms. While when the number of users is 1250, R_c=0.15 achieves the best latency of 10.3ms. Therefore, R_c should be carefully configured, and is a topic for future research. A careful optimization of R_c is likely to be more important for larger packet sizes.

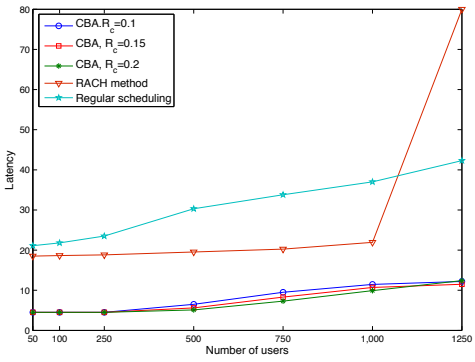


Fig. 2. Latency vs. network load

We also investigate the effect of the number of receiving antennas on the performance of CBA (The SNR γ is 5dB, and R_c=0.15). As shown in Fig. 3, it demonstrates the latencies under different number of antennas are almost the same when number of users is 50 and 250. However, when the number of users increases, using more antennas attains lower latency. This is feasible because when the number of users is large, the interference is very severe which causes lots of retransmissions and hence increases the latency. While with more antennas, the channel capacity is increased and retransmission is reduced.

IV. Conclusion and Future Work

This paper proposes two methods to enable efficient M2M communications and online gaming over LTE. The first method is a scheduling and resource allocation mechanism for sparse latency-constrained traffic. We show that adapting the number of dimensions used in each HARQ round results in almost casual feedback performance. We obtain a significant

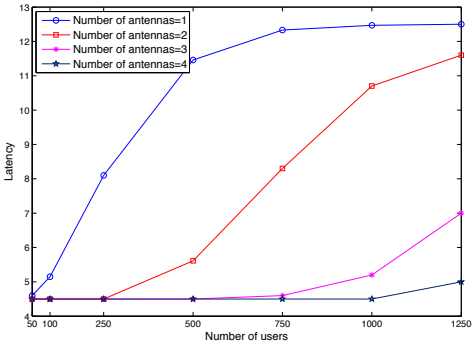


Fig. 3. Effect of number of antennas on the performance of CBA

improvement in spectral efficiency while maintaining a latency constraint, even in the case without CQI. Additional performance improvements are obtained when outdated CQI becomes available. With only one bit of feedback (ACK/NACK), we achieve a performance close to the ergodic capacity. The second method presents a contention based uplink channel access for MTC over LTE. With CBA, UEs select resource randomly without specific indications from eNB. To address the problem of collision, eNB empolies MU-MIMO detection to identify the collided UEs so that dedicated resources are allocated for them. The performance of the proposed method is compared with the regular scheduling method and RACH method. It shows that the CBA method can greatly reduce the latency. Moreover, it is also found that the performance of CBA depends on the coding rate of control information and the number of receiving antennas.

LTE is a flexible wireless system, which can accommodate various applications ranging from high rate HD video streaming to low rate M2M/online gaming applications. Therefore, there is some room for optimization to adapt to a specific application over LTE, which will be our future work.

V. Acknowledgments

The research leading to these results has received funding from the European Commission under the Seventh Framework Programme (FP7/2007-2013)/ ERC grant agreement no. 248993 (LOLA).

References

[1] N.Nakein and S.Krco, "Latency for Real-Time Machine-to-Machine Communication in LTE-Based System Architecture," *Proc. 11th European Wireless Conference-Sustainable Wireless Technologies*, Vienna, Austria, Apr. 2011, pp.263-268.

[2] S. Lien, K. Chen, and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications", *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66-74, Apr. 2011.

[3] Y. Chen, W. Wang, "Machine-to-Machine communication in LTE-A", *Proc. IEEE VTC fall*, Ottawa, Canada, Sept. 2010, pp. 1-4.

[4] L. Beom and K. Seong, "Mobility Control for Machine-to-Machine LTE Systems", *Proc. 11th European Wireless Conference-Sustainable Wireless Technologies*, Vienna, Austria, Apr. 2011, pp. 319-323.

[5] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971-1988, july 2001.

[6] Suard, B., Xu, G., Liu, H., Kailath, T.,"Uplink Channel Capacity of Space-Division-Multiple-Access Schemes," *IEEE Transactions on Information Theory,* vol. 44, no. 4, pp. 1468-1476, 1998.