Cross-Device Signaling Channel for Cellular Machine-Type Services

Chan Zhou, Egon Schulz
Huawei European Research Center
Riesstrasse 25 C-3.OG, 80992 Munich, Germany
Email: Chan.Zhou@huawei.com, Egon.Schulz@huawei.com

***Abstract*—The Machine-type services induce one critical challenge for the current cellular networks: The capacity of the conventional signaling channel in the network can hardly support the anticipated massive number of machine-type devices. According to the current network design, the signaling channel can easily be blocked up by a number of simultaneous signaling requests. In order to release the bottleneck of the signaling channel, we proposed a cross-device signaling mechanism to reduce the transmitted signaling data during the congestion-critical period. This mechanism removes the redundancy between the signaling messages of different devices. It has especially high performance when multiple highly synchronized devices generate simultaneous signaling requests. Thus, the mechanism can effectively mitigate the potential congestion in the signaling channel and further increase the design capacity of the network.**

I. Introduction

The market of machine-type communication and its accompanying services has been evolving quickly. It is expected that the number of connected machine-type devices will exceed 20 billion by 2020 [1], [2]. A large proportion of mobile operators worldwide are adapting their business models in order to offer machine-type services using the cellular infrastructure.

The integration of the machine-type communication into the cellular network can exploit the advantages of cellular services such as mobility, flexibility, reliability and good coverage. However, the machine-type communication has certain traffic characteristics which distinguish it from the human-centric communication: The machine-type communication generally has lower data rate and smaller packet size. The traffic is more uplink dominant and there are fewer interactions between the machine-type devices and the other side, which is usually an application server employed for the machine-type services.

Such traffic characteristics and the massive number of machine-type devices have strong impact on the design of the mobile and wireless systems, in particular the requirements on the signaling channels. It is shown in [3] that if the density of the terminal devices in the current LTE network increases to a certain degree, the signaling congestion will become a severe problem that might lead to the breakdown of the network.

Signaling congestions typically take place when a large amount of devices try to access the network in a highly synchronized manner [4], [5]. Some typical scenarios are:

- Recurring data transmissions are generated at precisely synchronous time intervals.
- A malfunction in the application or the server requests the devices to keep resending their data.
- An event triggers high numbers of machine-type devices to attach the network all at once, for example, massive metering devices becoming simultaneously active after a power cut.
- Many machine-type devices are simultaneously handed-over to another base station, which might be caused by highly synchronized mobility or a base station outage.

In the aforementioned scenarios, the network has to confront a high peak of signaling requests which severely challenges the design capacity of the signaling channel.

To date all existing solutions to reduce the signaling congestion follow one of the two basic principles: blocking or dropping the non-urgent signaling requests; batch signaling handling.

By the blocking/dropping methods [6], [7] priority classes will be defined for all kinds of applications. The low priority signaling requests will be blocked or dropped in order to guarantee the high priority requests. Some variations proposed hybrid schemes such as in [8], which combine convention-based and reservation-based multiple access. Obviously, the drawback of these methods is that not all requests will be responded in time. The low priority signaling requests have to endure the long delay or the neglect.

By batch signaling handling similar signaling messages will be grouped and handled together [9]. The drawback of these methods is also the additional delay introduced by the aggregation of the signaling message. Furthermore, messages from each device have to be aggregated at a central entity, e.g. the Machine-to-Machine (M2M) gateway first. This mechanism is only suitable for M2M area network where centralized signaling processing is possible.

However, by observing the signaling messages during the congestion period, it can be seen that there is a very high coincidence between the signaling information sent by the devices, particularly these belong to similar types, have similar service purpose and are located in a proximate area.

Part of this work has been performed in the framework of the FP7 project ICT-317669 METIS, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in METIS, although the views expressed are those of the authors and do not necessarily represent the project.

For example (Fig. 1) in the LTE system [10], the Device 1 and Device 2 sent their *RRCConnectionSetupComplete* message to base station. This message consists of sections such as PLMN-ID (Public Land Mobile Network-Identity), MME-ID (Mobility Management Entity-Identity), which are usually the same for the proximate and content-related devices. Thus, in this example, the message sections PLMN-ID and MME-ID sent by Device 1 and Device 2 are redundant.

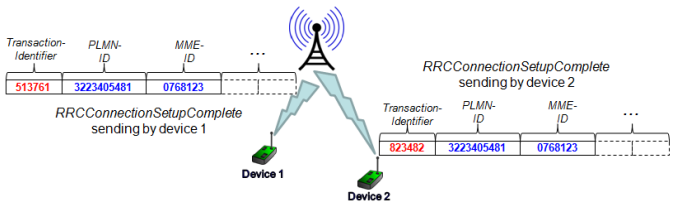

Fig. 1. Signaling messages sent by two devices containing high coincidence

In this paper, we introduce a solution to remove such redundancy in the signaling messages. Hence, it can effectively mitigate the signaling congestion in the wireless network and further increase the design capacity of the network. Firstly, a detailed description of the system and the considered problem is given in the next section. Then a cross-device signaling mechanism, including some of its variants, is illustrated in section III. The performance of the mechanism is analyzed in section IV. Finally, some conclusions are drawn in section V.

II. Signaling in MMC scenario

We consider a cellular network to support Massive Machine Communication (MMC) traffic. Signaling messages are used to coordinate the data connections between the base stations and a number of mobile devices (denoted as "device" for short in the follow text), e.g. UEs, sensors, actuators. The signaling messages are exchanged between the base stations and the devices via the wireless channels, which are also called signaling channels in the literature.

It is assumed that signaling messages are transmitted in both uplink and downlink, that is to say, from the devices to the base stations and vice versa. The signaling messages in the uplink are transmitted either via an unscheduled channel, e.g. the random access channel, or in a scheduled channel, e.g. the uplink shared channel. The signaling messages in the downlink are transmitted either in a broadcast channel, a shared channel, or a dedicated channel. The request for signaling can either be initialized by the base stations, or by the devices.

The information contained in each message is quantified by its entropy, i.e. $H(X_n)$ for the message X_n of the n-th device. According to the conventional network design, the signaling messages are sent and received by the device independently. Thus, the required date transmission volume in the signaling channel is

$$R_{sig,ind} = H(X_1) + H(X_2) + ... + H(X_N) \qquad (1)$$

if there are N devices in the network.

The entropy $H(X_n)$ gives the minimum amount of the required signaling information for device n. In practice it can be approached by a well-designed set of signaling instructions. The author in [11] described a mechanism to compress the signaling message based on the syntax of the historical messages which have been sent by the same device before. If the device wants to send a message repeating one of the old messages, then the device can send a short index of the old message instead of sending the whole message again. This mechanism is applied to a single communication link consisting of one transmitter and one receiver. The redundancy in the signaling messages in one single link is removed by the compression. However, the compression is always limited by the information-theoretic lower bound, i.e. the entropy $H(X_n)$.

If the signaling channel is observed from a system-level respective, the required signaling information is in fact given by the joint entropy

$$R_{sig,joint} = H\left(X_1, X_2, ..., X_N\right). \qquad (2)$$

As it is introduced in section I, the signaling in MMC scenario is correlated, in particular among a group of associated devices. The joint entropy $R_{sig,joint}$ is actually lower than the sum of the individual entropy. During the congestion-critical period, namely during the peak time of the signaling requests, the correlation is much higher, since the signaling traffic at this moment is mainly caused by the associated devices complying with similar behavior pattern.

The difference between the joint entropy $R_{sig,joint}$ and the sum of individual entropy $R_{sig,ind}$ is represented by the sum of the mutual information

$$\begin{aligned}\Delta R &= \sum_{n=1}^{N} H\left(X_n\right) - H\left(X_1, X_2, ..., X_N\right) \\ &= \sum_{n=1}^{N-1} I\left(X_{n+1}; X_1, ..., X_n\right), \end{aligned} \qquad (3)$$

which shows the potential to achieve a lower signaling overhead than the conventional signaling scheme. In the above expression (3) it can be seen that the potential gain even increases with larger number of associated devices.

In order to exploit this potential, we should go for a joint signaling scheme in which the signaling message X_n for the device n are coded based on the previous transmitted signaling messages $X_1, X_2, ..., X_{n-1}$ of other devices. Thus the minimum required data equals the conditional entropy $H(X_n \| X_1, X_2, ..., X_{n-1})$ instead of the independent entropy $H(X_n)$.

However, the joint signaling should fulfill the basic requirements on the control signals in the wireless communication system, namely

- *Lossless* - all the signaling messages should be received properly, both in the uplink and in the downlink.
- *Timeliness* - The signaling messages should be received with very low delay. The delay of signaling message will retard the whole communication process, sometime it will even impair the entire control mechanism, for instance in a fast time-varying radio environment.

These two requirements give us quite hard restrictions when we are trying to develop a joint signaling scheme. In the

following section, we introduce a cross-device signaling compression mechanism which effectively reduces the signaling redundancy without causing information loss and additional processing delay.

III. Cross-device signaling channel

We use a cross-device compression method to reduce the signaling overhead. The compression is achieved by utilizing the mutual information between the associated devices. Such devices are considered to have a certain relationship, e.g. being located in proximate area, serving same or related purpose. Thus they have a higher probability to have similar signaling requests within the same time period. The association between the devices can be determined in advance when additional context information is provided, such as location or service profile of the devices. The association can also be identified by a progressive learning process during the running time.

The whole mechanism consists of two phases: build up the dictionary for signaling messages and compress the new signaling message. By building up the dictionary, the previous messages transmitted in the signaling channel are collected and complied into a dictionary. This dictionary is available both at the side of the devices and at the side of the base stations. That is to say, each device and base station has a memory to save the dictionary. During the compression phase the original message is translated into a shorter message based on the dictionary. After receiving the shorter message from the transmitter, the receiver uses the same dictionary to regain the original message. The whole mechanism applied to uplink signaling channel and downlink signaling channel is different in certain points due to the particular limitations of these two channels. It is described in detail in the following sections.

A. Building up the dictionary for signaling messages

In the first step, a dictionary $\mathbb{D} := \{c_1, c_2, ..., c_K\}$ made up of code words c_i and indices will be compiled, which will be used later to compose the new signaling messages. The vocabulary size of the dictionary is denoted as K. The code words c_i are extracted from the previous signaling messages addressing the set of associated devices.

For the signaling messages in the uplink direction, i.e. from the device to the base station, there are two approaches for the devices to collect and store the signals sent from the other devices:

One approach is that the device listens to the uplink channel between the associated devices and the base station (see Fig. 2). For this approach the device should have the additional capability to receive the messages sent by other devices. This assumption gives a constraint of the distance between the associated devices and the receive capability of the listening devices. The acquired message is then stored at the device together with an index which unambiguously indicates the message. The base station saves the same message and the index in its memory. Various indexing methods can be used as long as the base station and the devices can obtain the same index for the same message according to a common agreement. For instant, the transmission order of the previous messages sent in the signaling channel can be used as the index.

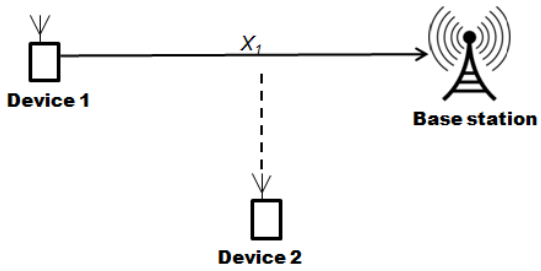

Fig. 2. Device 2 listens to the uplink signaling message from Device 1

Alternatively, the device can receive a set of instructions from the base station through the downlink channel (see Fig. 3). These instructions can consist of some representative signaling messages received by the base station before. Then the message is saved together with an index at the device and the base station. The index of each message can either be determined by the base station then be sent together with the corresponding message, or it can be composed according to a common agreement at the base stations and devices.

Then the dictionary is built up based on the saved messages. If the number of stored messages is large enough, entropy encoding algorithms, e.g. the Hoffmann coding [12] or arithmetic coding [13], which are based on the statistic model of the reference messages, can be used. The code words and the indices are calculated based on the stored reference messages.

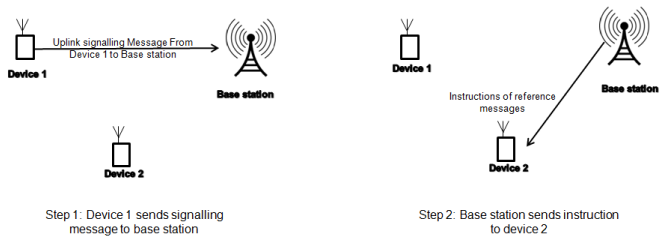

Fig. 3. Device 2 receives the instructions from base station. Step 1: Device 1 sends signaling message to base station. Step 2: Base station sends instruction to device 2

The same dictionary is available both at the side of the device and the base station. Each device may have a individual dictionary. In this case, the base station will keep multiple dictionaries which are used for each device individually.

For the second approach that the device receives the instructions from the base station, the instructions also can be processed signaling messages or even the whole dictionary itself. When applying the second approach, the associated devices are not necessary to be located in a close area. They can even belong to different cells. However, certain extra overhead in the downlink broadcast or multicast channel are required.

In the downlink direction, i.e. from the base station to the device, the signaling messages are sent in a broadcast channel or multicast channel such that all the associated devices are able to receive them. Then the messages are stored both at the base station and the devices. The dictionary is then built up based on the stored messages using the same method as the uplink message.

The stored signaling messages as well as the dictionary are constantly updated. Only the messages transferred within the last time period need to be stored. The messages sent beyond the time period are marked as outdated and are removed from the memory. The dictionary should be recompiled every time when the stored messages are updated. The length of the update time period is a predefined value which can be optimized regarding the efficiency of the dictionary and the available memory size. If the period length is large, more reference messages will be stored and the vocabulary of the dictionary becomes larger. Hence it is possible to find a longer code word to compose the new message and the compressed message might be more concise. However, the length of the indices also grows with the size of the vocabulary and it ultimately increases the length of compressed message. This trade-off should be considered by selecting the length of the time period.

B. Compressing the new signaling message

The new signaling message is compressed based on the dictionary. According to the available dictionary, diverse compression methods can be applied. The compression method should be lossless. Since the length of each signaling message is limited, the applied compression method should be efficient for short information sequences. One straightforward approach is to find the same code word in the new message and replace it with the index.

Fig. 4 shows an example of this approach.

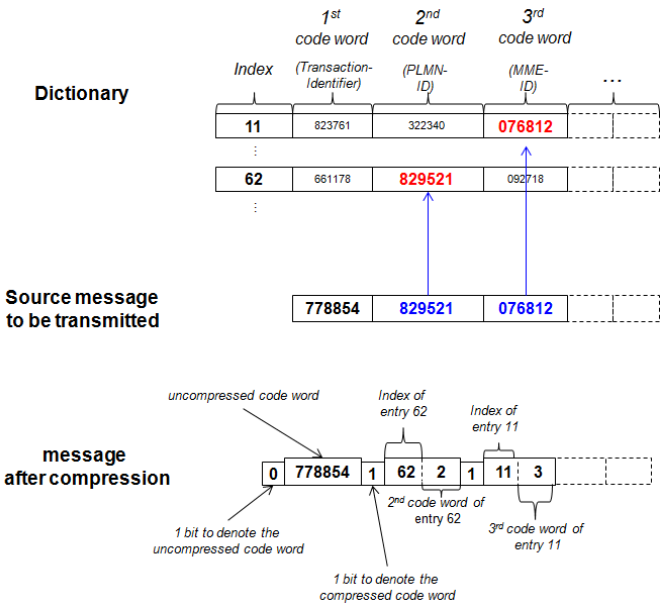

Fig. 4. Example for compression method

In this example, the dictionary is a collection of indexed messages. The source message to be transmitted is a new *RRCConnectionSetupComplete* message where the first code word is the transaction identifier "778854", which is usually unique and cannot be compressed. The second code word is the PLMN-ID "829521" and the third code word is the MME-ID "076812". Both of them can be found in the dictionary. The length of each code word is 6 digits, consists of 6 times 4 bits including the filling bits [10]. The code structure of *RRCConnectionSetupComplete* message is known at the side of the device and the base station thus it is not necessary to have additional bits to denote the start and the end of each code word.

Other lossless algorithms can be used, e.g. the compression method based on Lempel-Ziv algorithm [14]. The compressed message consists of the index of the reference message, the position of the same message piece in the reference message and the length of the same message piece.

Regarding signaling messages containing numeric values, such as channel state information report or radio resource assignment, there is a high probability that the messages send by associated devices have approximate values. However, it is unlikely that two messages are completely identical. In this case, a low-range value is transmitted in addition to the reference message, in order to describe the deviation between the reference message and the source message. Alternatively, algorithms such as Run-Length-Code can be applied to code the deviation.

In uplink direction, the compressed signaling messages are sent only if a low compression rate can be achieved. Otherwise, if the achievable compression rate is not sufficient or there is no valid dictionary available, the original uncompressed message can be sent. The latter case can happen when the devices are just awoken from the sleep mode.

If in downlink direction the base station wants to simultaneously transmit diverse signaling messages to multiple devices, a special approach in the downlink direction can be used: The base station broadcasts a mixture of uncompressed and compressed signaling messages together at the same time. In Fig. 5 an example is given, where the base station transmits the signaling messages to Device 1 and Device 2 at the same time. M1 is an uncompressed message for Device 1. It is sent in the broadcast or multicast channel and received by Device 1 and Device 2 as well. M2 is the signaling message for Device 2. It is compressed based on M1 using one of the compression methods given before. Device 2 receives both M1 and M2 and decompresses the message in M2 using M1 as reference.

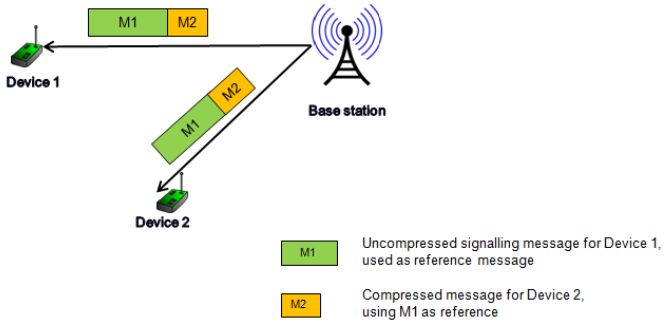

Fig. 5. Mixture of uncompressed message and compressed message

IV. Performance Analysis and Numerical Evaluation

The proposed cross-device signaling mechanism is evaluated in a LTE multi-cell network simulation environment. The

performance of the mechanism relies on the distribution of the signaling events and correlation between the signaling events of different devices. It is assumed that the signaling events follow a Beta distribution as in 3GPP studies [3]

$$p_i(t) = \frac{t^{\alpha-1}\left(T-t\right)^{\beta-1}}{T^{\alpha+\beta-1}Beta\left(\alpha,\beta\right)}, \tag{4}$$

in which $Beta\left(\alpha,\beta\right)$ is the Beta function with $\alpha = \beta = 50$. T is the event arrive circle with $T = 1000$ second, which fulfills

$$\int_0^T p_i(t)dt = 1. \tag{5}$$

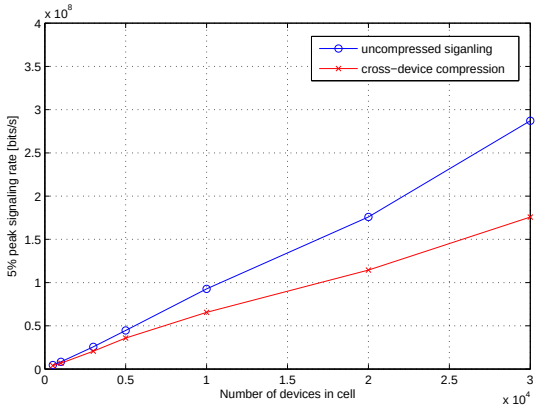

Fig. 6. Reduction of signaling peak rate

Fig. 6 shows the reduction of the signaling peak rate by applying the proposed mechanism. Here we observe the 5% signaling peak rate in the network when the network have in average $\bar{N} = 500, 1000, 3000, 5000, 10000, 20000, 30000$ devices in each cell. Further we assume that each associated device group has the size $N = 50$.

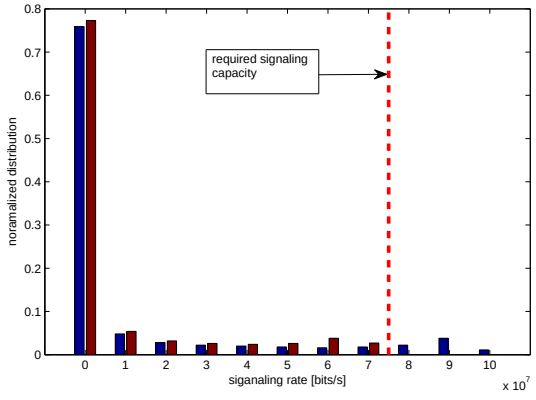

Fig. 7. Histogram of the signaling rate by $\bar{N} = 10000$

It can be seen that the requirement of the peak signaling rate increases linearly with the number of the devices, which challenges the network capacity. The proposed mechanism effectively reduced the peak signaling rate and relief the load in the signaling channel.

Fig. 7 gives the histogram of the signaling rate when there are in average $\bar{N} = 10000$ devices in a cell. It is shown that the peak rate over 80 Mbits/s is completely removed by applying the compressing mechanism. Thus, the probability of the congestion by $R > 80$ Mbits/s is eliminated.

V. Conclusion

In this paper we showed that a certain amount of mutual information exists among the signaling messages of the associated M2M devices. This mutual information provides us the potential to reduce the signaling overhead, particularly in the MMC scenario. In order to exploit the potential, the signaling messages should be coded cross the different devices.

According to our proposed mechanism, the devices and the base stations maintain a dictionary of reference messages, which consists of messages recently transferred in the signaling channel between the base stations and all associated devices. Then, in case that a new signaling message is to be transmitted, the message will be coded at the transmitter side based on the dictionary and reconstructed at the receiver side. The compression rate is especially high during the congestion-critical period, since at this moment, multiple correlated devices are sending similar signaling messages. Hence there is a higher probability that the code words in the dictionary can be reused.

Analytical and simulation results showed that the proposed mechanism can effectively reduce the congestion possibility and further increase the design capacity of the network.

References

[1] GSMA Intelligence, "From concept to delivery: the M2M market today," -, White Paper, Feb. 2014.

[2] Machina Research, "The need for low cost, high reach, wide area connectivity for the Internet of Things," -, White Paper, 2014.

[3] *TR 37.868, Study on RAN Improvement for Machine-Type Communication*, 3GPP Std., Rev. V 11.0.0, Sep. 2011.

[4] *TS 22.368, Service Requirements for Machine-Type Communications*, 3GPP Std., Rev. V 12.2.0, Mar. 2013.

[5] D. Boswarthick, O. Elloumi, and O. Hersent, *M2M Communications: A Systems Approach*. John Wiley & Sons, 2012.

[6] C.-Y. Liao, "Handling signaling congestion and related communication device," European Patent 11 006 466.4, 2011.

[7] S. Y. Lien, T. H. Liau, C. Y. Kao, and K. C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27 – 32, Jan. 2012.

[8] Y. Liu, C. Yuen, X. Cao, N. U. Hassan, and J. Chen, "Design of a scalable hybrid mac protocol for heterogeneous m2m networks," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 99 – 111, Feb. 2014.

[9] T. Taleb and A. Kunz, "Machine type communications in 3gpp networks: potential, challenges, and solutions," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 178 – 184, 2012.

[10] *TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*, 3GPP Std., Rev. V 10.1.0, Mar. 2011.

[11] F. Wartenberg, "Arrangement and method relating to messageing," European Patent 05 821 570.8, 2005.

[12] D. Huffman, "A method for the construction of minimum-redundancy codes," in *Proc. of I.R.E*, 1952.

[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *The Art of Scientific Computing*. Cambridge University Press ISBN 978-0-521-88068-8, 2007, ch. Section 22.6 Arithmetic Coding.

[14] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530 – 536, 1978.