System Design Considerations for Internet of Things (IoT) with category-M devices in LTE networks

Gurudutt Hosangadi
Nokia
600-700 Mountain Ave,Murray Hill
NJ 07974-0636, USA
Email:gurudutt.hosangadi@nokia.com

Dandan Wang
Nokia
600-700 Mountain Ave, Murray Hill
NJ 07974-0636, USA
Email:dandan.wang@nokia.com

Anil Rao
Nokia
12004 173rd Pl NE
Redmond, WA 98052, USA
Email:anil.rao@nokia.com

***Abstract*—Successful network deployment of the Internet of Things (IoT) requires many critical system design considerations. This paper highlights how an LTE system supporting Cat-M devices can be engineered to deal with the numerous constraints the 3GPP standard imposes for this new device type. Fundamental changes to the control channels, control and data timing relationships, the need to support half-duplexing, and variable repetition lengths for both data and control channels pose non-trivial challenges, particularly when attempting to satisfy the critical coverage KPI for Cat-M devices while at the same time preserving the capacity KPI for legacy LTE devices. In addition, the nature of IoT traffic is fundamentally different than legacy LTE data, requiring changes to existing system parameters and MAC algorithms. Finally, we will touch upon the very recent topic of supporting voice over IP traffic on Cat-M devices and the challenges therin.**

1. Introduction

The internet of things (IoT) buildup is well underway. The number of connected devices worldwide was around 5 billion in 2014 and is expected to reach 50 billion by 2020 [1]. These devices cover a wide range of applications: wearable devices, connected home appliances, remote sensing for utilities and smart cities to name a few. New applications are being devised daily which rely on the basic philosophy that any device with the ability to connect to the internet can be utilized in a smart way to either improve automation or provide connectivity in very small form factors. These massive number of devices communicating without human intervention constitute to what is commonly called as machine type communication (MTC) or referred to as IoT. Modern wireless cellular networks such LTE (Long Term Evolution) based on 3rd Generation Partnership Project (3GPP) are aptly placed to be an enabler of massive MTC. This is due to its all-inclusive-all-IP architecture, built in security, scalable traffic management capabilities and high spectral efficiencies. The traffic profile and requirement of IoT differs vastly from that of traditional mobile devices already supported in LTE cellular networks. Key differences include smaller traffic packet sizes and massive number of such devices. To support this change, 3GPP standards have been enhanced with new features. See [3.6] for a detailed coverage of the added features. In this tutorial, we focus on system design considerations of IoT devices. As we will show, the unique characteristics of IoT devices pose challenges to system design. Also, it is important for the design to minimize the impact to the key performance indicators (KPIs) of traditional service offerings given that traditional services and MTC traffic share the same wireless resources. Finally, we present some insights into performance based on our simulation results. For this, we specifically focus on the CaT-M feature within the LTE standards, which some major service providers are considering deploying to enhance their offered services for IoT.

2. Requirements

MTC devices pose a new set of requirements and challenges [2] to system design.Figure 1 shows the competing needs of MTC devices that affect system design. First and foremost is that the introduction of a large number MTC devices into the system should have a minimum impact on the operation of legacy devices. Next, MTC devices are expected to be low cost, low complexity, and have low power consumption. These attributes need to be considered in system design. Finally, MTC devices would be useful only if certain performance targets are met. We discuss each of these competing aspects next.

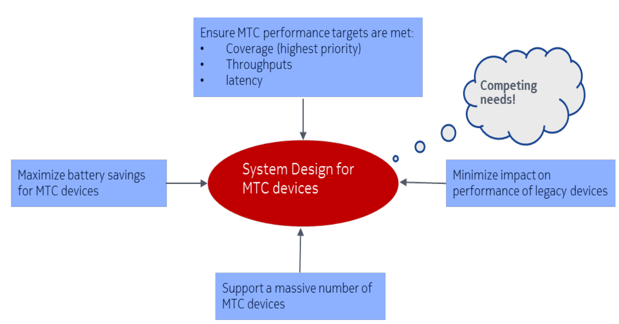

Figure 1. System Design for MTC devices

2.1. MTC Performance Targets

As illustrated in [3], enhanced coverage is one of the key requirements of MTC given the targeted application of these devices. The primary mechanisms that have been available in standards to improve coverage prior to Release 13 are hybrid automatic repeat request (HARQ) retransmissions. which primarily improve energy accumulation and can benefit from fading and interference diversity effects in the time domain. For dramatic improvements in coverage, a very large number of retransmissions will be required; for example, a 20dB coverage improvement would require approximately 100 retransmissions. This will consume resources for both uplink (feedback acknowledgement) and downlink (grants for HARQ retransmission) control channel resources. Further, due to the half duplexing nature of the Cat-M devices, the time duration to complete one HARQ round is much longer compared to legacy LTE devices operating in full duplex mode, resulting in large latencies. Finally, half duplexing makes the scheduling of packets on multiple HARQ processes in parallel more challenging (as explained in later sections) and prolongs the time that the UE will need to remain awake in order to transmit or receive a packet.

Due to the above-mentioned disadvantages of HARQ retransmissions, Rel 13 introduces a mechanism of repetition (upto 256) which is similar to transmission time interval (TTI) bundling in Rel 8 (upto 4 repeats) intended for voice over IP (VoIP) packets where consecutive TTIs are used to transmit the same packet. However, it should be noted that while all the transmission within the same repetitions have to be sent unconditionally, HARQ provides a potential resource saving since the retransmission of HARQ can be terminated once the packet can be successfully decoded. For example, an MTC device that does not require high coverage may be better served by using a small repetition size along with HARQ retransmissions compared to a cell edge UE which relies primarily on repetition.

Another system design tradeoff is between coverage and throughput/latency. While the peak rate is 1 Mbps (as per 3GPP), in reality, the achievable rates are impacted by resource availability, repetition used and the traffic profile of users in the system. In a later section we provide some examples on achievable rates observed in simulations for some sample configurations.

2.2. Low cost

This is mainly facilitated by standards which have reduced the operating bandwidth , maximum transmitted power, singlereceive RF chain (only 1 receive antenna is supported) and longer discontinuous reception (DRX) cycles.This is discussed in more detail later.

2.3. Minimal impact to legacy devices

A key selling point with the introduction of Cat-M devices in LTE is that it can work seamlessly together with legacy LTE devices on existing spectrum, with only a software upgrade needed for an operator to support Cat-M devices and capitalize on the MTC/IoT device explosion. To support this, 3GPP standards allows a MTC device to monitor and process a narrow bandwidth (1.4MHz for Cat.M1 and 200KHz for Cat.NB1)within the available bandwidth; this is referred to as a narrowband. An example of possible narrowbands is shown in Figure 2. One design choice may be to reserve this narrow band for Cat-M devices only while the legacy LTE devices use the rest of the available bandwidth. However, this simple design does not consider the instantaneous resource needs of Cat-M and legacy devices. For example, during a short time window, the reserved narrowband for Cat-M devices may go unused due to the sporadic nature of IoT traffic, while the network is at the same time heavily loaded with traffic from legacy LTE devices. A more comprehensive approach would be to have a unified mechanism to dynamically share the narrow band frequency resources between Cat-M and legacy devices taking into consideration quality of service (QoS) requirements.

3. System design

. In this section, we will discuss in more detail the different aspects of the system design to support Cat-M devices.

3.1. Location of Narrowbands

Of importance is the location of the narrow bandwidth used for MTC devices and the alignment of this bandwidth with resource block groups (RBGs) which may be used by the scheduler. The objective is to minimize the impact to RBG allocation. For example, Figure 2 illustrates the location of narrowbands for 10MHz. If we use any of NB0 through NB6, then a total of 9 PRBs cannot be used for RBG allocation. However, if we use NB7, fewer PRBs (8) cannot be used for RBG allocation.

Another consideration is that 3GPP standards allows frequency hopping (during repetition) of the narrowband to experience gains from frequency diversity. Since the hopping is on resources used by legacy devices, appropriate mechanisms to avoid collisions need to be designed.

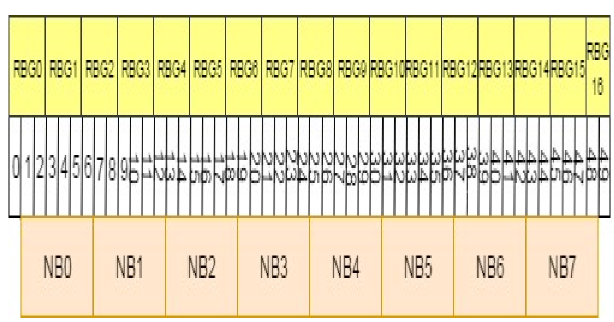

Figure 2. Narrowband reservation for CATM devices

3.2. Dormancy timer/Discontinous Reception (DRX) setting and Random Access Channel (RACH) impact

Since MTC devices are expected to carry very small amount of data for each device (unlike legacy devices), frequent RACH access requests will result in increased overhead especially if large repetitions are required to enhance RACH coverage. While there is no control on the number of initial RACH access requests, proper setting of dormancy time and DRX cycles can provide a good balance between subsequent RACH overhead and UE battery savings.

The use of a dormancy timer removes the RRC connection of the UE device from the evolved Node B (eNB) and thus, once it has new data to transmit or receive, it has to go through the RACH process again with the benefit of saving more UE battery. DRX, on the other hand, maintains the UEs Radio Resource Control (RRC) connection and avoids a RACH procedure. It saves battery by turning off some portion of the radio frequency chain during the DRX off cycle and waking up periodically to monitor the resource allocation. If there are only very small and infrequent packets sent/received by Cat-M devices, setting a smaller dormancy timer may be more desirable than configuring a DRX on/off pattern.

3.3. Channel State Information (CSI)/Scheduling Request (SR)

In legacy LTE system, one mechanism for eNB to acquire the downlink channel statistics of UEs is through CQI reporting. However, for Cat-M devices, due to the constraint of half duplexing, during the TTIs that a UE reports PCQI on PUCCH, this UE will not be eligible to receive any information on either the grant channel (MPDCCH) or dowlink data channel (PDSCH). This will impact the downlink throughput, especially when the uplink control channel (PUCCH) has to use a large number of repetitions in a coverage limited scenario. There is a similar impact when the UE reports SR on PUCCH. Thus it may be better to configure either a relatively large CQI period to reduce downlink impact or use aperiodic CQI mulitplexed with uplink data to reduce grant impact.

3.4. HARQ process management

Like legacy LTE devices, Cat-M devices can support multiple HARQ process at the same time, and, UL/DL HARQ processes are running in parallel independent of each other. However there are various aspects that need to be considered during system design such as:

- MPDCCH and PDSCH cannot be sent at the same TTI unless they are using the partial PRBs within the same narrowband.
- There is one TTI interval between receiving MPDCCH and receiving PDSCH.
- When MPDCCH, PDSCH or PUSCH is in the middle of transmission (repetition), this device is not eligible for another transmission.
- When a Cat-M UE sends feedback (ACK/NACK) on PUSCH/PUCCH, this UE is not eligible to be scheduled to receive on DL MPDCCH or PDSCH channel.

As an example, Figure 4 shows that due to the constraints, in UL, a maximum of 3 HARQ processes can be used at the same time.

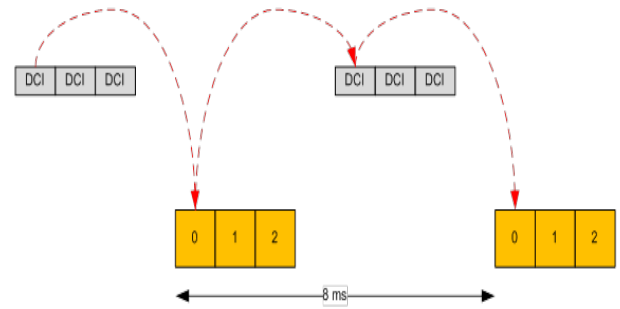

Figure 3. HARQ timing

3.5. MPDCCH Configuration

In legacy LTE, the maximum aggregation level (AGL) of PDCCH is 8 while it has been increased to 24 for Cat-M devices. As a result, the number of grants that can be supported per TTI is smaller. Thus, there is a tradeoff of coverage of MPDCCH and the number of grants can be sent per TTI. For example if a high AGL (e.g. 24) is used for a coverage limitied UE, only one user can be served in a given TTI, either in the uplink or downlink direction .

3.6. Use of closed loop algorithms

As illustrated in [6], the traffic pattern of Cat-M is very unique in the sense that Cat-M devices wake up very infrequently and then send/receive a very small amount of data. As such, Cat-M devices are expected to wake from RRC IDLE state when new data arrives and as a result it is very difficult for any closed loop algorithms to converge. We will revisit this later in the results section.

4. Results

In this section, we present some sample simulation results aimed at providing insight into how system design and configuration can affect performance of MTC devices.

4.1. Assumptions

We focus here on the use case where MTC devices are being served using the Cat.M1 feature support in LTE. Further it is assumed here that 1.4MHz bandwidth that is allocated is at a fixed location. Table 4.9 lists some of key assumptions that are applicable to the results discussed.

Parameter	Assumption
Frequency Bands	2GHz
Macro inter-site distance	500m
Shadow model	Shadow fading std. dev = 8 dB
Cat-M UE traffic	1K bits, mean/min reading time = 10s/2.5s
Dormancy timer	2 seconds
Fading Channel	ETU 3km/hr
Macro eNB antenna	17 dBi gain Vertical pattern: 10°@3dB beamwidth SLA = 20 dB, downtilt =15deg
Body and cable loss	1 dB (data terminal)
Mobile antenna	Omnidirection; -3 dBi gain
eNB Tx power	2x20W

TABLE 1. ASSUMPTIONS FOR SIMULATION RESULTS

4.2. Simulation Tool

The results presented in this section were generated using a C++ system level simulator, based on 3GPP LTE standard, which can simulate a multi-eNB layout including effects of the wireless fading channel, propagation environment and antennas. Algorithm focus is on the Medium Access Control (MAC) layer and specifically on scheduling and HARQ management. The physical layer is abstracted (to reduce simulation run time) and the simulator supports both legacy LTE devices and CATM devices. For the traffic profile, full buffer, VoIP and burst traffic profiles are supported. Key performance metrics such as throughputs, latencies and statistics on Signal to noise ratio (SINR), resource consumption, number of retransmission and errors rates are available for analysis from the simulator.

4.3. Impact of repetition length (RL)

We first look at the achievable rates for a sample configuration. The acheivable rate is highly impacted by configured repetition for the various channels. Figure 4 shows a sample configuration where the time between granting of resources to providing feedback is 20ms due to various constraints and configured repetition sizes resulting in a throughput of 43kbps.

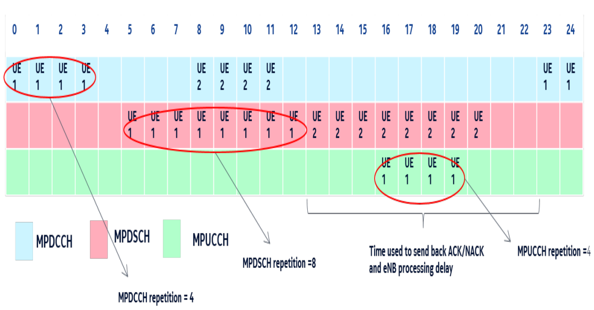

Figure 4. Illustration of timing relationship between channels

Next we consider the sensitivity of performance to varying the RL for a single user with full buffer traffic. We keep the RL of the downlink control channel (M-PDCCH) and uplink control channel (PUCCH) are fixed at 4 and 8 while the RL of downlink shared channel (PDSCH) is varied. It can be observed in Figure 5, which is for a burst traffic profile, that increasing the RL for cell edge users definitely improves the user experienced throughput while for near cell users which have a good SINR, using a smaller repetition number makes more sense.

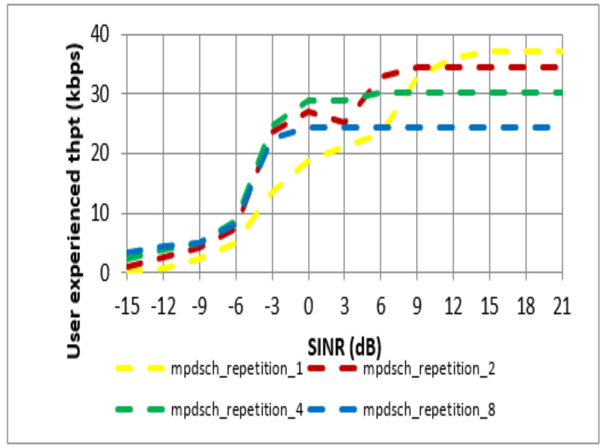

Figure 5. Impact of repetition on user traffic for bursty traffic

4.4. Link Adaptation and power control

Figures 6A shows that the performance is insensitive to the initial BLER setting (e.g. 10%) and step size chosen (e.g. slow or large step). Similarly Figure 6B shows that closed loop power control (CLPC) provides no benefit compared to the open loop (OLPC) case. In both these cases, the average time between packets is in seconds as a result of which the MTC device transmits only 1 measurement and then goes back to sleep and there is no time for any convergence.

Figure 6C highlights the importance of setting the correct OLPC setpoint. A lower set point allows multiple PRBs to be used which lowers the code rate and results in improved performance. It should be noted these results do not take into account channel estimation errors which increase as the SINR per PRB is lowered in which case the number of PRBs assigned should be limited to avoid the lower SINR per PRB range. Finally, Figure 6D shows the importance of setting the initial modulation and coding scheme (MCS). We touch upon this further next.

Therefore, the initial modulation and coding scheme (MCS) selection becomes very important. The initial MCS can be selected based on the reported CQI on RACH message 5 for DL or channel estimated based on received RACH message 3 for UL. Note that if the traffic pattern of Cat-M devices changes, the system should be designed so that link adaptation and power control will improve performance. For example, transmission of VoIP on Cat-M devices is currently being discussed in standards community, and in this case the Cat-M devices may benefit from both link adaptation and power control.

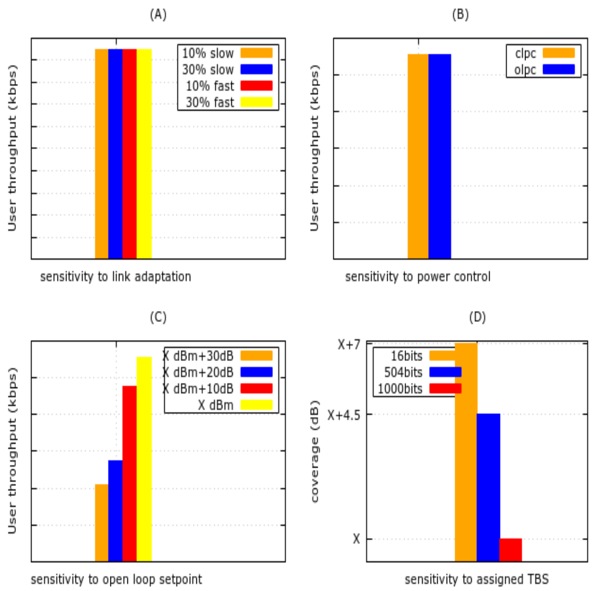

Figure 6. MTC performance sensitivity to design parameters (uplink, packet size = 1000bits)

4.5. Coverage

Coverage is expected to be one of the key metrics for MTC devices, maximizing this will drive system design. MTC devices in coverage limited locations are expected to be transmitting at their maximum power. With this resource exhausted, the transport block sizes allocated will impact coverage . Revisitng Figure 6D, we can see that the sensitivity of coverage to TBS assignment for a packet size of 1000 bits. Coverage here is defined as the maximum path loss at which residual BLER is below 2%. Overall a smaller packet size provides the best coverage performance.

4.6. Latency

Most MTC applications are not expected to be time critical and latency will be a low priority. For example, a meter that is reporting power consumption at a specific node in a smart grid can tolerate delays on the order of seconds and the key requirement will be reliability.The latency experienced by the report will depend on three key factors:

- MTC traffic loading: higher loading will result in larger latencies
- Repetition length versus number of packet segments: using a higher RL instead of packet segmentation or retransmissions to combat coverage issues might be desirable as in the latter case contention with other MTC devices come into play and impacts latency
- Residual BLER i.e. if the packet does not go through successfully it will require a retransmission which will increase the latency

4.7. Interference

Given that coverage for low cost Cat-M devices is expected to be a key performance metric, interference management will play a key role in being able to meet this requirement. As discussed before, one option is to reserve a narrow bandwidth solely for MTC devices in which case the interference can be low. However, if this region is used to support VoIP traffic, the interference could begin to creep up. Another option is to simply share the resources with legacy LTE devices in which case the interference levels could be quite high and meeting coverage requirements of MTC devices could become more challenging. Therefore a system design consideration could be to consider doing some intercell frequency planning to ensure low interference on the reserved MTC narrowbands, which of course comes at the cost of capacity to the legacy LTE devices.

4.8. VoIP support using MTC devices

There is an ongoing discussion and studies in 3GPP to understand how VoIP could be supported via Cat-M devices, which may be desirable on wearables for example. While there is expected to be some relaxation on delay budget for the voice traffic packets, latency will still have to be controlled. In this section, we look at different aspects of system design affecting VoIP support using CatM devices.

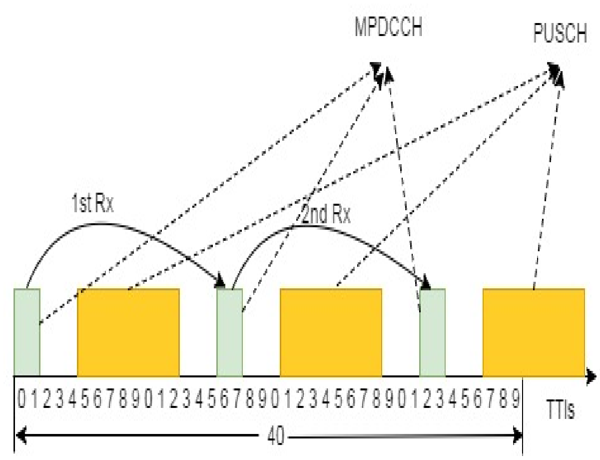

Figure 7. VoIP scheduling illustration

4.9. Impact of repetition length on VoIPcoverage

Larger RL will increase the per link coverage. However, as has been discussed in the context of non-VoIP transmissions with CatM devices, it also results in longer overall time duration for one HARQ transmission. Figure 7 illustrates an example of this in the context of VoIP transmissions. For the configured RLs, 2 HARQ transmissions can complete within 40ms. As VoIP packets arrives with a fixed pattern (every 20ms at talk spurt and 160ms at silent period) and the packets have to be transmitted within a delay budget,

PUSCH repetition	MCL
8	138
16	140
32	138

TABLE 2. IMPACT OF REPETITION ON COVERAGE

larger repetition number means we may have to aggregate more VoIP packets within one HARQ transmission which means a larger TBS. As the gain achieved by the extra repetition may be offset by the loss of the enlarged TBS, increasing the repetition may not necessarily result in larger coverage. For example, table 6 shows that when we increase the PUSCH repetition number from 8 to 16, with the delay budget of 200ms, we see the supported MCL (maximum coupling loss) increasing from 138dB to 140dB. However, when we further increase the repetition from 16 to 32, the MCL even becomes smaller Thus, there is a balance between increasing the repetition number and TBS increasing.

4.10. Impact of iBLER selection

In legacy LTE systems, a 10% iBLER target is generally used. As discussed previously, a longer HARQ duration for VoIP traffic will will result in collision of HARQ retransmission with the newly arrived voice packets. If we lower the iBLER target (for example to 5%), the HARQ retransmission probability is reduced but it will require more repetitions to support the same TBS at the same SINR. Thus, it is a trade-off between more HARQ vs. repetition . [7] shows that using HARQ retransmission can achieve higher coverage than without HARQ retransmission and very large repetition. The best combination of iBLER target and RL needs more study.

4.11. Impact of SID packets

In a typical voice conversation between 2 users, during a talk spurt of one user, there will be silence insertion descriptor (SID) packets sent by the other user. Thus, this user has a voice packet to transmit at the same time has to receive SID packets which due to the constraints of half duplexing is not allowed. This makes the VoIP scheduling more challenging. Another challenge is that we may not be able to use the same fixed TBS for VoIP packets anymore as SID happen less frequently compared with voice packets and whenever SID packets arrive, more aggregation of voice packets will happen.

4.12. Impact of Segmentation

Segmentation is generally used in legacy VoLTE to extend the coverage. However, in the case of Cat-M, due to the timing constraints caused by half duplexing, it is challenging to transmit multiple HARQ process at the same time especially when we use larger repetition numbers. Assume we only support one HARQ process, and one HARQ duration is 15ms (1+3+8+3 = 15 ms) with the assumption of MPDCCH repetition level of 1 and PUSCH repetition level of 8. Thus, every 20ms, there is only enough time to transmit one HARQ. If we segment a VoIP packets into multiple small segments, the coverage for each segment becomes better but the overall delay could be large.

5. Conclusions

We have outlined the numerous system design aspects which must be considered to successfully deploy an LTE network supporting Cat-M MTC devices. The numerous constraints as well as additional coverage/power-saving features the 3GPP standard has included for such devices poses significant challenges integrating support for such devices in an LTE network while minimizing the KPI impact to existing smartphone and other high performance data-centric devices. It has been shown that careful selection of system parameters such as the Cat-M dormancy timer, the number of HARQ transmissions and repetition factor used for Cat-M data and control channels, and the configuration needed to support VoIP on Cat-M devices involves many different tradeoffs, particularly between coverage and latency and also the capacity impact to the legacy LTE network. We have demonstrated that link adaptation features such as closed loop rate conrol and closed loop power control need to be revisited based on the nature of MTC traffic, and certain system settings such as the open loop power control setpoint and default initial MCS assignment become much more critical. It is important to highlight such considerations so that an operator can tailor the parameters and scheduler design aspects to achieve the desired trade-offs inherent in introducing MTC devices into an existing LTE network.

References

[1] www.machinaresearch.com

[2] Adyson M. Maia, Dario Vieira, Miguel F. de Castro, Yacine Ghamri-Doudane, *A Mechanism for Uplink Packet Scheduler in LTE Network in the Context of Machine-to-Machine Communication*, Globecom 2014.

[3] Alberto Rico-Alvario, Madhavan Vajapeyam, Hao Xu, Xiaofeng Wang, Yufei Blankenship, Johan Bergman, Tuomas Tirronen, and Emre Yavuz , *An overview of 3GPP enhancements for machine to machine communications*, IEEE Communications magazine, 2016.

[4] Jill Jermyn, Roger Piqueras Jovery, Ilona Murynetsy, Mikhail Istominy and Salvatore Stolfo, *Scalability of Machine to Machine systems and theInternet of Things on LTE mobile networks*, IEEE, 2015

[5] Erfan Soltanmohammadi, Kamran Ghavami, Mort Naraghi-Pour, *A Survey of Traffic Issues in Machine-to-Machine Communications over LTE*, IEEE Internet of Things Journal, 2015

[6] 36.888, *Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE*, version 12.0.0.