DRINF

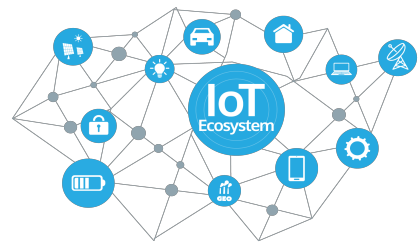
DOTTORATO di RICERCA in INGEGNERIA dell'INFORMAZIONE

Centro Stampa d'Ateneo

DOCTORAL SCHOOL OF MEDITERRANEA UNIVERSITY OF REGGIO CALABRIA

**Olga VIKHROVA**

# GROUP-BASED COMMUNICATIONS IN CELLULAR IOT




Advisor: Prof. Antonella MOLINARO
Co-advisor: Prof. Petar POPOVSKI
Coordinator: Prof. Tommaso ISERNIA
S.S.D. ING-INF/03
XXXIII Ciclo


**DOCTORAL SCHOOL**
UNIVERSITA' MEDITERRANEA DI REGGIO CALABRIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE, DELLE INFRASTRUTTURE E DELL'ENERGIA SOSTENIBILE (DIIES)

PHD IN
INFORMATION ENGINEERING

S.S.D. ING-INF/03
XXXIII CICLO

# Group-based Communications in Cellular IoT

CANDIDATE
Olga VIKHROVA
[redacted]

ADVISOR
Prof. Antonella MOLINARO
[redacted]

CO-ADVISOR
Prof. Petar POPOVSKI
[redacted]

COORDINATOR
Prof. Tommaso ISERNIA
[redacted]

REGGIO CALABRIA, APRIL 2021

Finito di stampare nel mese di **Gennaio 2021**

Edizione 


**Quaderno N. 52**
Collana *Quaderni del Dottorato di Ricerca in Ingegneria dell'Informazione*
Curatore *Prof. Tommaso Isernia*

**ISBN** 978-88-99352-46-2

Università degli Studi *Mediterranea* di Reggio Calabria
Salita Melissari, Feo di Vito, Reggio Calabria

OLGA VIKHROVA

# Group-based Communications in Cellular IoT

The Teaching Staff of the PhD course in
*INFORMATION ENGINEERING*
consists of:

Tommaso Isernia (coordinator)
Giuseppe Araniti
Giuseppe Ruggeri
Francesco Buccafurri
Claudio De Capua
Francesco Della Corte
Antonio Iera
Antonella Molinaro
Salvatore Coco (UNICT)
Giacomo Morabito (UNICT)
Lorenzo Crocco (IREA-CNR)
Giuseppe Coppola (IMM-CNR)
Rosario Morello
Giacomo Messina
Giuliana Faggio
Patrizia Frontera
Pier Luigi Antonucci
Aimè Lay Ekuakille (univ. del Salento)
Fabio Filianoti
Andrea Francesco Morabito
Domenico Rosaci
Sofia Giuffrè
Gianluca Lax
Domenico Ursino
Dominique Dallet (Institute Polytechnique de Bordeaux)
Giorgio Graditi (ENEA)
Voicu Groza (univ. Ottawa)

**Abstract**

Machine-Type Communications (MTC) or Internet of Things (IoT) communications refer to smart devices' interconnections, with reduced human intervention enabling them to participate more actively in everyday life. Enhancing IoT connectivity solutions with timely, energy-efficient, and reliable group-based communications will bring tangible benefits to society by rapidly evolving healthcare, industry, and intelligent monitoring systems. In the last two decades, there have been efforts in academia and industry to enable IoT connectivity over the legacy communications infrastructure. It is becoming more and more evident that many IoT service characteristics and requirements hardly fit the legacy solutions designed for human interactions and intensive broadband traffic. Therefore, IoT-specific communications systems and protocols have gained profound attention. The commercial potential of cellular IoT and its exponentially growing market trend for upcoming years makes Narrowband-IoT (NB-IoT) and Long Term Evolution for Machines (LTE-M) the dominant technologies for massive MTC.

In this context, the majority of IoT-based solutions have focused on access challenges and energy-efficiency of short packet transmissions from multiple sources to a common destination; on the contrary, network-originated transmissions to multiple destinations lacked sufficient attention. The increasing importance of *group-based* communications for disruptive IoT applications, such as remote control, monitoring, and sensing, calls for novel communication solutions addressing the challenges to provide Point-to-Multipoint (PTM) support for autonomous devices and coping with the inherent limitations of cellular IoT systems.

The present work is devoted to characterizing state-of-the-art technologies to enable PTM connectivity in cellular IoT networks, identify their limitations, and contribute to their performance assessment and enhancements. We mainly focus on the group-based communications and their applications in supporting massive and critical IoT communications. The main contributions presented in this work include: (i) a novel PTM transmission framework for supporting critical services in cellular IoT systems; (ii) radio resource management techniques for servicing unicast and multicast MTC-users with heterogeneous service requirements; (iii) a set of analytical models of the group-based communications and methods for their key performance metrics evaluation; (iv) advanced analysis of the timeliness of information for the status update in cellular IoT application scenarios. Our study indicates the need for advanced and adaptive techniques for efficient information distribution in highly autonomous IoT systems with diverse requirements and capabilities.

## Sommario

Le comunicazioni Machine-Type Communications (MTC) o dell'Internet of Things (IoT) si riferiscono alle interconnessioni dei dispositivi intelligenti con intervento umano limitato che consente loro di partecipare più attivamente nella vita di tutti i giorni. Il miglioramento delle soluzioni di connettività IoT con comunicazioni di gruppo tempestive, efficienti dal punto di vista energetico e affidabili apporterà vantaggi tangibili alla società, alla sanità, alle industrie e ai sistemi di monitoraggio intelligenti in rapida evoluzione. Negli ultimi due decenni, il mondo accademico e industriale hanno fatto molti sforzi per abilitare la connettività IoT sull'infrastruttura di comunicazione legacy. Sta diventando sempre più evidente che le caratteristiche e i requisiti dei servizi IoT sono molto diversi dalle soluzioni legacy progettate per l'interazione umana e per l'intensivo traffico a banda larga. Pertanto, i sistemi di comunicazione e i protocolli specifici dell'IoT hanno guadagnato una profonda attenzione. Il potenziale commerciale dell'IoT cellulare e il suo trend di mercato in crescita esponenziale per i prossimi anni rendono le tecnologie Narrowband-IoT (NB-IoT) e Long Term Evolution for Machines (LTE-M) dominanti per le MTC massive.

La maggior parte delle soluzioni basate sull'IoT si concentra sulle sfide di accesso e sull'efficienza energetica delle trasmissioni di piccoli pacchetti da più fonti a una destinazione comune, mentre le trasmissioni originate dalla rete verso più destinazioni mancano di attenzione. La crescente importanza delle comunicazioni basate sul gruppo per le applicazioni IoT dirompenti richiede nuove soluzioni che affrontino le sfide di fornire supporto Point-to-Multipoint (PTM) per dispositivi autonomi e inerenti alle limitazioni dei sistemi IoT cellulari.

Il presente lavoro è dedicato alla caratterizzazione delle tecnologie all'avanguardia per abilitare la connettività PTM nell'IoT cellulare, alle loro limitazioni e ai nostri contributi nella valutazione e al miglioramento delle prestazioni. Ci concentriamo principalmente su soluzioni basate sul gruppo e sulle sue applicazioni per supportare comunicazioni IoT massive e critiche. I principali contributi presentati in questo lavoro includono (a) un nuovo framework di trasmissione PTM per supportare servizi critici nei sistemi IoT cellulari; b) tecniche di gestione delle risorse radio per fornire agli utenti MTC unicast e multicast dei servizi con requisiti eterogenei; (c) una serie di modelli analitici delle comunicazioni di gruppo e metodi per la valutazione delle relative metriche chiave di prestazione; (d) analisi avanzata della tempestività delle informazioni di aggiornamento di stato per gli scenari di applicazione IoT cellulare. Il nostro studio indica la necessità di tecniche ancora più avanzate e adattive per una distribuzione efficiente delle informazioni in sistemi IoT altamente autonomi con requisiti e capacità diversi.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| 5G | Fifth Generation |
| 3GPP | Third Generation Partnership Project |
| 5G-S-TMSI | 5G Serving Temporary Mobile Subscriber Identity |
| ACB | Access Class Barring |
| ACK | Acknowledgement |
| AoI | Age of Information |
| AS | Application Server |
| AS | Access Stratum |
| BLER | Block Error Rate |
| BM-SC | Broadcast Multicast Service Center |
| BS | Base Station |
| BW | Backoff Window |
| CDF | Cumulative Distribution Function |
| cDRX | Connected DRX |
| CE | Coverage Enhancement |
| CIoT | Cellular Internet of Things |
| CN | Core Network |
| CoAP | Constrained Application Protocol |
| CP | Cyclic Prefix |
| CRC | Cycle Redundancy Check |
| CRS | Cell-specific Reference Signal |
| CRT | Contention Resolution Time |
| CRW | Contention Resolution Window |
| CSI | Channel State Information |
| CTMC | Continuous-Time Markov Chain |
| DCI | Downlink Control Information |
| DL | Downlink |
| DLT | Distributed Ledger Technology |
| DMRS | Demodulation Reference Signal |

| | |
|---|---|
| DoNAS | Data over Non Access Stratum |
| DRB | Data Radio Bearer |
| DRX | Discontinuous Reception |
| DTCH | Dedicated Traffic Channel |
| DTMC | Discrete Time Markov Chain |
| E-UTRA | Evolved Universal Terrestrial Radio Access |
| E-UTRAN | Evolved Universal Terrestrial Radio Access Network |
| EC-GSM-IoT | Extended Coverage GSM Internet of Things |
| EDGE | Enhanced Data rates for GSM |
| eDRX | extended Discontinuous Reception |
| EDT | Early Data Transmission |
| eGP | enhanced Group Paging |
| eMBB | enhanced Mobile Broadband |
| eMBMS | enhanced Mobile Broadcast and Multicast Services |
| eMTC | enhanced Machine Type Communications |
| eNB | eNodeB |
| EPC | Evolved Packet Core |
| EPS | Evolved Packet System |
| F-RAN | Fog Radio Access Network |
| FCFS | First Come First Served |
| FDD | Frequency-Division Duplex |
| G-RNTI | Group Radio Network Temporary Identifier |
| GEO | Geostationary-Earth Orbit |
| GERAN | General Radio Access Network |
| GGSN | Gateway GPRS Support Node |
| GID | Group ID |
| GP | Group Paging |
| GPRS | General Packet Radio Services |
| GSM | Global System for Mobile Communications |
| GSMA | Global System for Mobile Communications Association |
| GTP | GPRS Tunnel Protocol |
| H-SFN | Hyper System Frame Number |
| HARQ | Hybrid Automatic Repeat Request |
| HD-FDD | Half-Duplex Frequency-Division Duplex |
| HLR | Home Location Register |
| HSS | Home Subscriber Service |
| HTC | Human-Type Communication |
| HTTP | Hypertext Transfer Protocol |
| I-RNTI | Inactive Radio Network Temporal Identity |
| i.i.d. | Independent and Identicaly Distributed |

| | |
|---|---|
| ICT | Information and Communication Technology |
| ID | Identifier |
| IMSI | International Mobile Subscriber Identity |
| IoT | Internet of Things |
| IP | Internet Protocol |
| KPIs | Key Performance Metrics |
| LCFS | Last Come First Served |
| LEO | Low Earth Orbit |
| LLC | Logical Link Control |
| LOS | Line-of-Sight |
| LPWAN | Low-Power Wireless Networks |
| LST | Laplace-Stiltjes Transform |
| LTE | Long Term Evolution |
| LTE-M | Long Term Evolution for Machines |
| M2M | Machine-to-Machine |
| MAC | Medium Access Control |
| MBMS | Mobile Broadcast and Multicast Services |
| MBMS-GW | MBMS-Gateway |
| MBSFN | multicast and broadcast single frequency network |
| MCE | Multicast Coordination Entity |
| MCL | Maximum Coupling Loss |
| MCS | Mission Critical Services |
| MDP | Markov Decision Process |
| MEC | Mobile Edge Computing |
| MGF | Moment Generation Function |
| MIB | Master Information Block |
| ML | Machine Learning |
| MME | Mobility Management Entity |
| mMTC | Massive Machine-Type Communications |
| MPDCCH | MTC Physical Downlink Contol Channel |
| MSC | Modulation Coding Scheme |
| MSC | Mobile Switching Center |
| MTC | Machine-Type Communications |
| NAS | Non-Access Stratum |
| NB-IoT | Narrowband IoT |
| NCCE | Control channel element |
| NeGP | New enhanced Group Paging |
| NPBCH | Narrowband Physical Broadcast Channel |
| NPDCCH | Narrowband Physical Downlink Control Channel |
| NPDSCH | Narrowband Physical Downlink Shared Channel |

| | |
|---|---|
| NPRACH | Narrowband Physical Random Access Channel |
| NPSS | Narrowband Primary Synchronization Signal |
| NPUSCH | Narrowband Physical Uplink Control Channel |
| NPUSCH | Narrowband Physical Uplink Shared Channel |
| NR | New Radio |
| NRS | Narrowband Reference Signal |
| NSSS | Narrowband Secondary Synchronization Signal |
| NTN | Non-Terrestrial Networks |
| NVF | Network Function Virtualization |
| OFDM | Orthogonal Frequency Division Multiplexing |
| OFDMA | Orthogonal Frequency-Division Multiple Access |
| P-GW | Packet Data network Gateway |
| P-RNTI | Paging Radio Network Temporary Identifier |
| PAoI | Peak Age of information |
| PBCH | Physical Broadcast Channel |
| PDCCH | Physical Downlink Contol Channel |
| PDCP | Packet Data Convergence Protocol |
| PDF | Probability Density Function |
| PDN | Packet Data Network |
| PDSCH | Physical Downlink Shared Channel |
| PDU | Protocol Data Units |
| PF | Paging Frame |
| PGF | Probability Generation Function |
| PLMN | Public Land Mobile Network |
| PNB | Paging Narrowband |
| PO | Paging Opportunity |
| PRACH | Physical Random Access Channel |
| PRB | Physical Resource Block |
| PRS | Positioning Reference Signal |
| PSM | Power Saving Mode |
| PSS | Primary Syncronization Signal |
| PSTN | Public Switched Telephone Network |
| PTM | Point-to-Multipoint |
| PTP | Point-to-Point |
| PTW | Paging Transmission Window |
| PUCCH | Physical Uplink Control Channel |
| PUSCH | Physical Uplink Shared Channel |
| PWS | Public Warning System |
| QoS | Quality of Service |
| r.v. | Random Variable |

| | |
|---|---|
| RA | Random Access |
| RACH | Random Access Channel |
| RAN | Radio Access Network |
| RAO | Random Access Opportunity |
| RAR | Random Access Response |
| RAT | Radio Access Technology |
| RB | Resource Block |
| RC | Radio Control |
| RE | Resource Element |
| RLC | Radio Link Control |
| RNTI | Radio Network Temporary Identity |
| RoHC | Robust Header Compression |
| RRC | Radio Resource Control |
| RRM | Radio Resource Management |
| RSS | Resynchronization Signal |
| RU | Resource Unit |
| S-GW | Serving Gateway |
| S-TMSI | Serving Temporary Mobile Subscriber Identity |
| SC-DRX | Single-Cell Discontinuous Reception |
| SC-FDMA | Single Carrier Frequency Division Multiple Access |
| SC-MBR | Single-Cell Multimedia Radio Bearer |
| SC-MCCH | Single-Cell Multicast Control Channel |
| SC-MTCH | Single-Cell Multicast Traffic Channel |
| SC-PTM | Single-Cell Point-to-Multipoint |
| SC-RNTI | Single-Cell Radio Network Temporary Identifier |
| SCTP | Stream Control Transmission Protocol |
| SDN | Software Defined Network |
| SDU | Service Data Units |
| SFN | System Frame Number |
| SGSN | Serving GPRS Support Node |
| SI | System Information |
| SIB | System Information Block |
| SIB-1 | System Information Block Type 1 |
| SIB-20 | System Information Block Type 20 |
| SIB1-NB | System Information Block Type 1 Narrowband |
| SIB2 | System Information Block Type 2 |
| SIoT | Satellite Internet of Things |
| SLA | Service Level Agreement |
| SNDCP | Sub-Network Dependent Convergence Protocol |
| SNR | Signal-to-Noise Ratio |

| | |
|---|---|
| SP | Standard Paging |
| SRB | Signaling Radio Bearer |
| SRC | Sounding Reference Signal |
| SRS | Sounding Reference Signal |
| SSS | Secondary Syncronization Signal |
| SUCI | Subscription Concealed Identifier |
| SUPI | Subscription Permanent Identifier |
| TAU | Tracking Area Update |
| TB | Transport Block |
| TBS | Transport Block Size |
| TC-RNTI | Temporary Cell Radio Network Temporary Identifier |
| TDD | Time Division Duplex |
| TMGI | Temporary Mobile Group Identity |
| TTI | Time Transmission Interval |
| UCI | Uplink Control Information |
| UDP | User Datagram Protocol |
| UE | User Equipment |
| UL | Uplink |
| URLLC | Ultra-Reliable Low Latency Communications |
| VF | Virtual Frame |
| VoLTE | Voice over LTE |
| WSN | Wireless Sensor Networks |
| WUR | Wake Up Radio |
| WUS | Wake Up Signal |
| XML | Extensible Markup Language |

# 1

# Introduction

*In this chapter, we explain the motivation of this research and its aim. We define central research objectives and summarize our contributions. The chapter also covers our research methodology and the structure of the thesis.*

## 1.1 Motivation

Internet of Things (IoT) is part of a transformation that is affecting our society, including industries, consumers, and the public sector. It is a critical enabler in the unfolding digital transformation of the management of physical processes. It provides better insights and allows for more efficient operation by offering the capability to embed electronic devices into physical world objects and create smart objects that allow us to interact with the physical world through sensing or actuation. The IoT system, depicted in Fig. 1.1, is the enabler for services like intelligent transportation, predictive medicine, smart metering and emissions reduction, logistics, and industrial control that continuously improve the quality of our everyday lives.

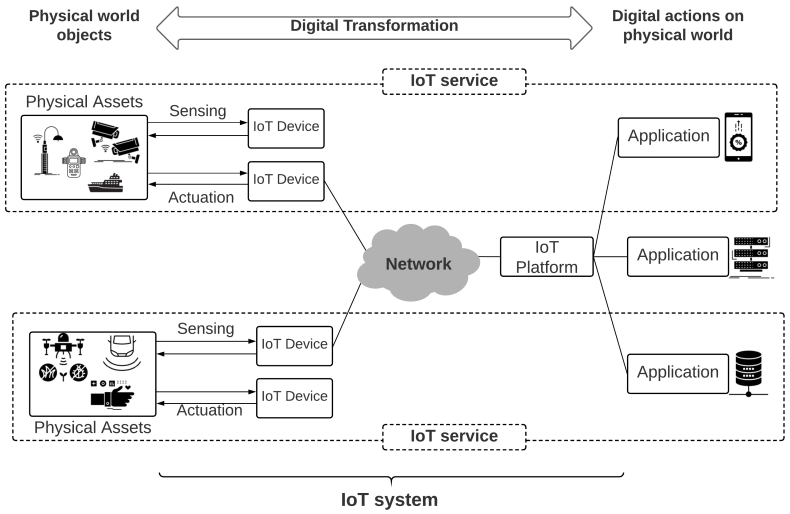



**Fig. 1.1.** IoT system for a digital representation of the physical world.

Starting from simple, proprietary, and even wired solutions for data collection from sensors, the IoT technology has evolved into a large-scale wireless system supporting different

types of data traffic and diverse service requirements. The Third Generation Partnership Project (3GPP) has evolved their cellular technologies to target new requirements for machine connectivity and a wide variety of IoT use cases. The new dedicated cellular IoT technologies are becoming globally available within a short time and boost IoT market and further evolution of the sector. The number of IoT connections is set to surge over the next years. The Global System for Mobile Communications Association (GSMA) Intelligence forecasts the total number of IoT connections will grow up to 25.2 billion by 2025 [1]. It estimates that 3.1 billion of these connections will employ cellular technologies.

The IoT landscape includes not only delay tolerant and low-payload Machine-Type Communications (MTC), but it is also expected to support the growing demand for critical and multimedia data traffic to be communicated in smart environments [2].

The 3GPP standardization of cellular networks is trying to address the requirements of novel IoT use cases to ensure that the technology standards evolution is addressing future market needs. The need for fast deployment and smooth integration of new functionalities within existing infrastructure is of utmost importance for large-scale IoT systems. To this end, 3GPP, for example, made it possible to enable the new features of the last releases through a software/firmware upgrade of the existing equipment.

Point-to-Multipoint (PTM) cellular communications perfectly fit the requirements of massive IoT update/upgrade applications (e.g., software bug fixes, device configuration commands) since they allow a file delivery to a theoretically infinite number of devices simultaneously. However, this is not the only case when PTM transmissions are beneficial and efficient from the service latency, network resources, and energy consumption perspectives. Remote control and synchronization of many smart devices in urban and rural areas, computing task allocation in Mobile Edge Computing (MEC) scenarios, distribution of schedules, commands, and alerts towards a subset of devices are examples of group-based communications between a source of information and multiple receivers.

The state-of-the-art PTM solution for cellular IoT is based on the Mobile Broadcast and Multicast Services (MBMS) framework in the form of Single-Cell Point-to-Multipoint (SC-PTM) transmission introduced in Release 14. Though MBMS was originally designed as an alternative to digital terrestrial television services, it was considered as a feasible enabler of PTM services for cellular IoT. However, it requires a rigorous evaluation and optimization, as the inherent requirements of battery-powered IoT devices impose a set of new challenges [3].

Not all IoT devices are expected to support full MBMS application protocol stack or MBMS-related procedures. Some IoT device categories, which comprise devices with limited processing, battery, or memory capabilities, unlikely support Hypertext Transfer Protocol (HTTP) and full Extensible Markup Language (XML) processing for the service subscription. IoT supports only a reduced MBMS profile, meaning that neither *Streaming* nor *Group Communication* in a form of existing services for the Public Warning System (PWS) and Mission Critical Services (MCS) are applicable for IoT. MBMS is a human-oriented service by design. The fundamental difference between machine-type and human-type communica-

tions is that the MTC devices are not involved in the decision-making process. A device-owner or an application has to decide which group of devices will receive content before the transmission. Moreover, the legacy service announcement procedures used for device notification about upcoming multicast transmissions over the MBMS radio bearers need a revision as it may violate the requirements of delay-sensitive applications. When neither sending nor receiving any data, IoT devices exploit power-saving techniques, such as Discontinuous Reception (DRX) /extended Discontinuous Reception (eDRX) or Power Saving Mode (PSM), to conserve their battery [4]. Being in *idle* mode, devices become unreachable for any *network-originated* services until the end of their sleep period.

Since devices might have different activity cycles, an efficient solution for service announcements and content transmission is needed to address challenging use cases of *initially unplanned* data delivery to a group of devices. With a proper paging solution, the latency and, consequently, device energy consumption of the group-based communications can be improved.

The problem of supporting delay-critical IoT applications is not sufficiently addressed in the literature. For instance, in [5], the performance of the delay-tolerant firmware updates over unicast and multicast links has been analyzed for Narrowband IoT (NB-IoT). The work [6] deals with the resource allocation problem for the multicast transmission in the presence of background unicast traffic. Both works lack analytical approaches and solutions for paging, which are the focus of this thesis.

Another critical aspect of MTC is the freshness of communicated information, e.g., a monitoring object's status, scheduling information, and synchronization command. Conventional delay- or latency-based analysis is incomplete in applications when the receiver, e.g., a system monitor or an IoT device, must have up-to-date information. A relatively new metric called Age of information (AoI) is used to quantify the freshness of the received information and complement the system performance and optimality knowledge [7, 8].

AoI is a useful metric in many applications such as vehicular, sensor, and broadcast networks. Most of the results are obtained for queuing systems that model either a single communication link between a sender and a receiver or just a packet scheduling mechanism at one of the system nodes [9]. However, the packet carrying timely information may go through more than one link to reach the destination or be delayed by a device before the transmission. Therefore, more advanced communication system abstractions should be considered to study the timeliness of transmitted content. The different practical methods for the average AoI derivation exist. However, this metric provides limited information about the system behavior in complex communication scenarios. The distribution of the AoI or its Moment Generation Function (MGF) is more informative but tricky to analyze [10]. The thesis also addresses this research gap.

## 1.2 Objectives and Methodology

Based on the challenges and research gaps identified in Section 1.1, four **research objectives** (ROx) have been formulated and subsequently addressed to answer the **central research question** of this thesis *how to provide group-based connectivity for cellular IoT devices efficiently.*

- **RO1. Identify the feasibility of legacy PTM solutions to serve a large number of autonomous IoT devices without human interventions.**
  Due to the human-oriented design, the state-of-the-art PTM technologies for cellular IoT do not fit battery-powered devices with low duty cycles and asymmetric uplink and downlink traffic. The legacy *service announcement* procedure is based on the idea of informing all subscribed users about the upcoming multicast transmission far in advance and let them decide when to join or leave the multicast session. However, in IoT systems, this decision is delegated to a device owner or a manufacturer. Therefore, the channel configuration and scheduling commands should be efficiently communicated to device groups.
- **RO2. Improve PTM communication latency and device energy consumption for delay-critical IoT application use cases.**
  When the arrival of multicast traffic cannot be planned, the latency of *delay-critical* applications and, as a consequence, device energy consumption could be unacceptably high. Therefore, this objective aims to improve the delivery of *unplanned critical* traffic after, e.g., bug fixes, system re-configuration or status changes. It calls for adaptive paging and scheduling solutions to distribute critical content as soon as possible.
- **RO3. Improve the performance of multicast and unicast communications in cellular IoT systems.**
  Due to the diverse requirements of multicast and unicast services, the group-based transmission may experience meaningful Quality of Service (QoS) degradation when overlapping in time with unicast transmissions or impact the performance of important unicast services instead. Managing both types of services is even more challenging in systems with reduced bandwidth, such as NB-IoT and Long Term Evolution for Machines (LTE-M).
- **RO4. Analyze the freshness of information in status update communication systems.**
  The problem of sampling updates in communication systems from the perspective of information freshness rather than packet delay is of utmost importance for conventional IoT applications with dominating uplink traffic and for group-based scenarios with a lack of feedback from devices. This objective is aimed at analyzing the timeliness of updates transmissions in two-hop communication systems, representing the most general scenario of IoT systems.

To attain the research objectives formulated above, a general research approach, illustrated in Figure 1.2, was adopted in this work. It consists of several stages, defined as follows.

- **Study the state-of-the-art technologies and research methods.** This stage aims at acquiring knowledge of the system of interest necessary to develop analytic and simulative frameworks and their subsequent cross-validation.
- **Use case analysis and problem statement.** Focusing on the challenging use cases and solutions available in the literature, different research gaps are identified at this stage and translated into a subset of problems.
- **Solution design**. Indicated challenges and problems are investigated to formulate a hypothesis or propose a solution for further testing and verification.
- **System model definition**. Each identified problem is addressed in a mathematical model design that provides relevant and illustrative Key Performance Metrics (KPIs) for assessment and validation of the proposed solutions. The solutions are evaluated analytically and validated via advanced simulators. Optimization techniques are employed to reach the target system performance.
- **Model validation**. Model assumptions and accuracy are verified at this stage by extensive simulations.
- **Analysis.** At this stage, the obtained results are analyzed and discussed to highlight the root cause of the addressed problem or revise the hypotheses and solutions.

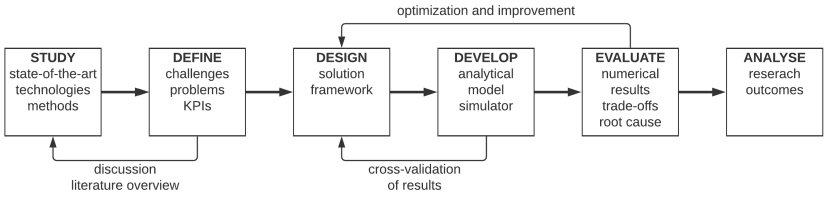



**Fig. 1.2.** Schematics of the research approach.

We apply probability theory and advanced queuing theory as powerful tools for modeling and analyzing communication systems of different topologies, properties, and complexity. Whenever closed-form solutions for the steady-state probabilities are not available, we propose low-complexity algorithms to compute probabilities and metrics of interest. To validate the developed analytical models for LTE-M and NB-IoT systems, we employ our calibrated MATLAB-based simulator. All analytical results have been validated by Monte Carlo simulations. Our simulation models have been run long enough to collect statistically representative data during the steady states.

The preliminary results of this work have been presented at leading workshops and symposia in the field for early-stage feedback. The central research results have been published in high-ranked conference proceedings and submitted to reputable journals for peer revision as a part of the result verification. Some of the important outcomes of this work are obtained in close collaboration with peers from Aalborg University.

## 1.3 Contributions

The main contributions of the thesis can be summarised as follows:

- a solution for improving PTM communications in cellular IoT networks that i) significantly reduces the latency of the service announcement stage and device energy consumption for the critical applications by optimizing paging and group-based transmission scheduling parameters;
- a rigorous analytical framework that accurately models all procedures entailing multicast connectivity, such as paging, system configuration for enabling PTM reception, and group-based data delivery;
- resource allocation and scheduling solutions for concurrent delivery of multicast and unicast traffic in cellular systems with reduced bandwidth;
- an analytical model of the group-based content delivery aimed at optimizing communication latency considering dynamic paging parameters;
- extensive analysis of the information freshness in a two-hop communications system, for which we obtain the distributions and average values of the AoI, Peak Age of information (PAoI), and system delay.

### 1.3.1 Research outputs

The detailed contributions and relevant publications are organized in four blocks, each corresponding to a research objective from the list presented in Section 1.2. Contributions C1 and C2 are included in chapter 3, chapter 4 contains our contribution C3, chapter 5 describes contribution C4.

**C1. Contributions in analyzing challenges to support PTM connections for massive IoT in cellular networks.**

Multicast communications perfectly fit the requirements of massive IoT since they provide a means for simultaneous content delivery to a theoretically infinite number of devices. However, the 3GPP solution for PTM is based on the *human-oriented* enhanced Mobile Broadcast and Multicast Services (eMBMS) specification while MTC requires very different solutions due to their i) limited capabilities and ii) specific use case scenarios. The first issue demands lighter protocols and procedures to let IoT devices join a multicast session without any human intervention and operate on battery for many years. Another challenge is the different latency requirements and traffic characteristics of group-based application use cases. For instance, a schedule for the software update can be *planned* and announced in advance so that all devices in an area of interest know when the multicast transmission starts. In contrast, the update after bug fixes should be distributed among devices on-the-fly once it is released.

In relation to these challenges, we give a concise overview of the legacy solution's inherent limitations of PTM communications in IoT networks. We propose a *machine-oriented*

multicast framework that significantly reduces the latency of the *unplanned* group-based transmissions when the beginning of a transmission cannot be communicated in advance. In our framework, instead of the long legacy service announcement procedure, we rely on a paging solution to wake up idle devices and employ Random Access (RA) procedure to provide devices with the configuration of a multicast transmission. Our framework considerably reduces the latency of the group-based services by replacing the legacy service announcement stage with adapted RA. In addition, we analyze the performance of different paging solutions and their impact on the service latency and device energy efficiency of group-based communications.

These contributions have been included in a book chapter and conference proceedings:

- O. Vikhrova, S. Pizzi, A. Molinaro, A. Iera, K. Samouylov, G. Araniti, "Energy Efficient Paging in Cellular IoT Networks", book chapter in LPWAN Technologies for IoT and M2M Applications, Elsevier, 2019.
- O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, K. Samuylov, G. Araniti, "Performance Analysis of Paging Strategies and Data Delivery Approaches for Supporting Group-Oriented IoT Traffic in 5G Networks", 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Jeju, 2019.

**C2. Contributions in improving the performance of delay-critical group-based communications in cellular IoT.**

Many emerging group-based IoT communication scenarios impose stringent delay requirements while expect to transmit a few bytes of an update, a command, or a task. Critical message broadcast services enabled in cellular networks target mobile devices with a rechargeable battery and short idle period, while IoT devices are inactive most of the time. It makes network-originated communications towards IoT devices with latency guarantees challenging.

We propose a new paging solution optimized for content delivery of delay-critical applications in response to this challenge. In particular, we adapt the paging scheme's parameters, such as the paging group size and the interval between the paging transmissions, and multicast transmission scheduling interval, to reduce the service latency and improve device energy consumption. We develop an accurate analytical model, validated through simulations, to evaluate our proposed scheme's performance and compare it against similar solutions from 3GPP specifications and literature. The numerical results indicate an almost double reduction of service latency and a significant energy-efficiency improvement.

The preliminary results have been presented in a conference, while the major contributions have been published in a journal:

- O. Vikhrova, S. Pizzi, A. Molinaro, A. Iera, K. Samouylov, G. Araniti, "Group-based Delivery of Critical Traffic in Cellular IoT Networks," Computer Networks, volume 181, p. 107563, 2020.

- O. Vikhrova, S. Pizzi, A. Molinaro, K. Samouylov, G. Araniti. "Group-Oriented Services for Critical Machine Type Communications in 5G Networks," 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC-2018.

**C3. Contributions in supporting unicast and multicast services with heterogeneous requirements in cellular IoT systems.**

Scheduling multicast traffic in bandwidth-reduced IoT systems requires special care when unicast transmissions run in parallel. We propose different resource allocation schemes, namely with resource reservation and dynamic resource sharing, to provide multicast services for NB-IoT devices more efficiently. We develop analytical frameworks to assess the performance of proposed allocation schemes and evaluate the performance of PTM and Point-to-Point (PTP) services under given resource constraints.

We investigate the impact of dynamic paging parameters on the group-based communication latency and find the optimal points that minimizes communication latency and time to wait for the multicast transmission.

Our research on resource allocation and latency optimization problems has been published in the following papers:

- O. Vikhrova, S. Pizzi, A. Molinaro, G. Araniti, "Paging Group Size Distribution for Multicast Services in 5G Networks", IEEE INFOCOM 2020, Toronto, ON, Canada, pp. 484-489, 2020.
- O. Vikhrova, S. Pizzi, I. Sinitsyn, A. Molinaro, A. Iera, K. Samuylov, and G. Araniti, "An Analytic Approach for Resource Allocation of IoT Multicast Traffic," ACM MobiHoc Workshop on Pervasive Systems in the IoT Era (PERSIST-IoT '19), ACM, New York, NY, USA, pp. 25-30, 2019.

**C4. Contributions in analysing timeliness of information in status update IoT application use cases.**

In IoT tracking applications that rely on real-time status updates, the classic packet delay metrics are not informative since it measures only the propagation time between the sender and the receiver, but it does not tell how old the received information is. The basic scenario assumes that devices sporadically transmit their updates to a monitor at which the AoI of received packets is calculated. However, in group-based communications, IoT devices that receive, e.g., scheduling information, also need to know whether it is up to date. The updates may arrive from different sources and over channels with different characteristics, therefore, the packets arriving at the receiver cannot be treated equally. To this end, we consider a general communication system composed of two queues in tandem and two independent flows of updates with different entry points, namely at the first queue of the tandem and at the second queue, respectively. Packets that cross the two queues are prioritized with respect to the packets arriving directly to the second queue to compensate for the delay in reaching the receiver.

Most of the literature results provide only first-order statistics of the AoI and its maximum value PAoI due to the complexity of the analysis. However, the distribution of both metrics and their higher-order statistics are more representative and informative for network engineering and optimization. We propose a powerful tool that allows us to obtain the Laplace-Stiltjes Transform (LST) of the distribution and statistics of a classic packet delay and novel age-related metrics for the system of interest with general service distribution time at the second node. Moreover, the closed-form expression for the two-node system with a single packet flow for the PAoI distribution has been presented. The analysis highlights the factors impacting the AoI and demonstrates that the systems' optimal parameters can be achieved.

These contributions have been published in the following papers.

- O. Vikhrova, F. Chariotti, B. Soret, A. Molinaro, G. Araniti, P. Popovski, "Age of Information in Multi-hop Networks with Priorities", IEEE Globecom'20 (in press).
- F. Chariotti, O. Vikhrova, B. Soret, P. Popovski, "Peak Age of Information Distribution for Edge Computing with Wireless Links," IEEE Transactions on Wireless Communications, 2020. (submitted).

### 1.3.2 Publications

The complete list of the author's publications produced during the Ph.D. period includes 8 papers related to the subject of the thesis and mentioned in Section 1.3.1, and 4 papers not included in the thesis.

1. F. Chariotti, O. Vikhrova, B. Soret, P. Popovski, "Information Freshness of Updates Sent over LEO Satellite Multi-Hop Networks," IEEE Journal on Selected Areas of Communitions, 2020, (submitted).
2. F. Chariotti, O. Vikhrova, B. Soret, P. Popovski, "Peak Age of Information Distribution for Edge Computing with Wireless Links," IEEE Transactions on Wireless Communications, 2020, (submitted).
3. O. Vikhrova, F. Chariotti, B. Soret, A. Molinaro, G. Araniti, and P. Popovski, "Age of Information in Multi-hop Networks with Priorities," in IEEE GLOBECOM 2020 - IEEE Global Communications Conference, 2020, (in press).
4. F. Rinaldi, S. Pizzi, O. Vikhrova, A. Molinaro, G. Aranit "Paging IoT Devices in 5G-Enabled Non-Terrestrial Networks," 71-st International Astronautical Congress (IAC) – The Cyber Space Edition, 12-14 October 2020.
5. O. Vikhrova, S. Pizzi, A. Molinaro, A. Iera, K. Samouylov, G. Araniti, "Group-based Delivery of Critical Traffic in Cellular IoT Networks," Computer Networks, volume 181, p. 107563, 2020.
6. O. Vikhrova, S. Pizzi, A. Molinaro, G. Araniti, "Paging Group Size Distribution for Multicast Services in 5G Networks", in IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2020, pp. 484–489.

7. O. Vikhrova, S. Pizzi, A. Molinaro, A. Iera, K. Samouylov, G. Araniti, "Energy Efficient Paging in Cellular IoT Networks", book chapter in LPWAN Technologies for IoT and M2M Applications, Elsevier, 2019.
8. O. Vikhrova, S. Pizzi, I. Sinitsyn, A. Molinaro, A. Iera, K. Samuylov, and G. Araniti, "An analytic approach for resource allocation of IoT multicast traffic," PERSIST-IoT'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 25 – 30.
9. O. Vikhrova, C. Suraci, A. Tropeano, S. Pizzi, K. Samouylov, and G. Araniti, "Enhanced radio access procedure in sliced 5G networks," in $11^{th}$ International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2019, pp. 1 – 6
10. F. Rinaldi, O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, and G. Araniti, "Joint Device-to-Device and MBSFN transmission for eMBB service delivery in 5G NR networks," in Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, 2019, pp. 599 – 609.
11. O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, K. Samuylov, and G. Araniti, "Performanceanalysis of paging strategies and data delivery approaches for supporting group-orientediot traffic in 5G networks," in 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2019, pp. 1 – 5.
12. O. Vikhrova, S. Pizzi, A. Molinaro, K. Samouylov, and G. Araniti, "Group-oriented services for critical machine type communications in 5G networks," in 2018 IEEE $29^{th}$ Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2018, pp. 824 – 828.

## 1.4 Thesis outline

The thesis encompasses different aspects of group-based communications focusing on massive IoT applications in 3GPP networks. It provides a solid background on the protocols and architecture of the LTE-M and NB-IoT systems, analysis of the system limitations, and frameworks to evaluate and optimize system parameters to improve the quality of IoT services. The thesis is organized into 6 chapters, their content is briefly described below.

- **Chapter 1. Introduction** contains the motivation, structure, and contributions of this work.
- **Chapter 2. Fundamentals of Cellular IoT** introduces central concepts of the IoT technologies in cellular networks, including network architecture, protocols, and relevant procedures that serve as a necessary background for the following chapters.
- **Chapter 3. Group-based Communications in Cellular IoT Networks** covers the state-of-the-art PTM IoT technologies, reviews the requirements of different group-based application scenarios in cellular IoT systems and the challenges to provide PTM services for such applications, and introduces our framework for supporting delay-critical group-based application use cases.

- **Chapter 4. Resource Allocation for PTM Communications in Cellular IoT Networks** is dedicated to the analysis of the resource allocation strategies for providing multicast and unicast services with different QoS requirements in cellular networks and to the optimization of parameters of the paging procedure.
- **Chapter 5. Age of Information for IoT Applications** presents methods for the advanced age-related metrics analysis in emerging IoT scenarios.
- **Chapter 6. Conclusions** includes the summary of research outcomes and discussion of future research avenues.

The final part of the document includes the bibliography.

# 2

# Fundamentals of Cellular IoT

*This chapter gives a concise overview of 3GPP standardization efforts to support IoT in cellular networks and covers design aspects of LTE-M and NB-IoT .*

## 2.1 Introduction

IoT is a driving force of modern and future wireless communications. It is one of the most ambitious projects in Information and Communication Technology (ICT) industry aiming to connect virtually anything with everything. According to the Ericsson forecast [11], about 29 billion devices will have direct wireless connections to the Internet by 2022. The most promising and life-improving IoT applications include wearables, smart homes, smart cities, healthcare, automotive, asset tracking, retail, and drones.

Introduction of the Machine-to-Machine (M2M) communication was the first attempt to connect devices with applications. Most M2M communication solutions are purpose-built, often with a dedicated communication stack and proprietary networking protocols, and designed to satisfy a very particular application and communication needs. Typical applications of M2M are controlled lighting, electric appliances, home monitoring. However, to unfold the full potential of the IoT concept, solutions for connecting devices and smart objects should be based on a common and interoperable Internet Protocol (IP)-based connectivity framework.

The 3GPP have evolved their cellular technologies to target a new IoT use cases. Cellular communication systems have introduced support for IoT since the second generation, however 3GPP have made additional effort to design *dedicated solutions* for Cellular Internet of Things (CIoT) connectivity.

Future market needs dictate the technology standards evolution. It has become clear that the variety of IoT use cases cannot be fully described with two broad categories of use cases addressing IoT requirements defined for Fifth Generation (5G) cellular systems [12], summarized in Fig. 2.1.

**Massive Machine-Type Communications (mMTC)** aims to address communication of a large number of light devices with small and intermittent traffic. The main requirements of this category are:

- massive deployment,
- scalability,
- extended coverage.

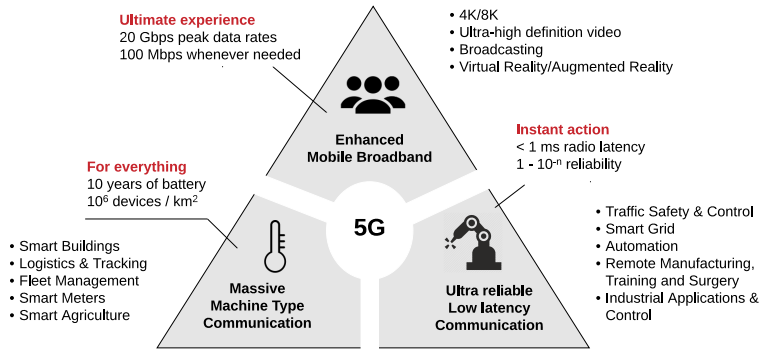



**Fig. 2.1.** Categories of use cases supported in 5G.

The target deployment density motivates for *ultra-low complex* device design to ensure reasonable deployment and operation cost. The most *common examples of mMTC* are utility metering and monitoring, fleet management, telematics, sensor sharing in the automotive and manufacturing industry, inventory management, asset tracking and logistics.

To support the mMTC category of use–case scenarios, 3GPP has presented evolutionary solutions namely:

- **Extended Coverage GSM Internet of Things (EC-GSM-IoT)** - backward compatible with Global System for Mobile Communications (GSM). It has been designed to support IoT connections under challenging radio coverage conditions (600 kHz frequency deployments). It can be installed onto existing GSM systems reducing the time and cost of deployment.
- **LTE-M** - backward compatible with Long Term Evolution (LTE), it reuses most of LTE design principles and is capable to support high-end IoT applications with stringent latency and throughput requirements using a flexible system bandwidth of 1.4 MHz.
- **NB-IoT** - clean-slate design, it is a brand new Radio Access Technology (RAT) but it can operate within an LTE carrier. The system occupies a narrow spectrum of 200 kHz providing extreme coverage and high uplink capacity.

**Ultra-Reliable Low Latency Communications (URLLC)** category addresses demanding use cases with very *high reliability* and extremely *low communication latency* including critical IoT scenarios. Remote, cooperative and autonomous driving, real-time sensor sharing, automation in a smart grid, remote control in manufacturing are all *examples of critical IoT* application use cases. The 3GPP support for critical IoT has been defined in a form of **LTE URLLC** and **New Radio (NR) URLLC**.

To better define the requirements that a certain use case puts on the devices and the supporting network a novel widely adopted classification has been proposed in [13]. As explained in Fig. 2.2 CIoT can be segmented into (i) **massive IoT**, (ii) **broadband IoT**, (iii) **critical IoT** and (iv) **industrial automation IoT**.

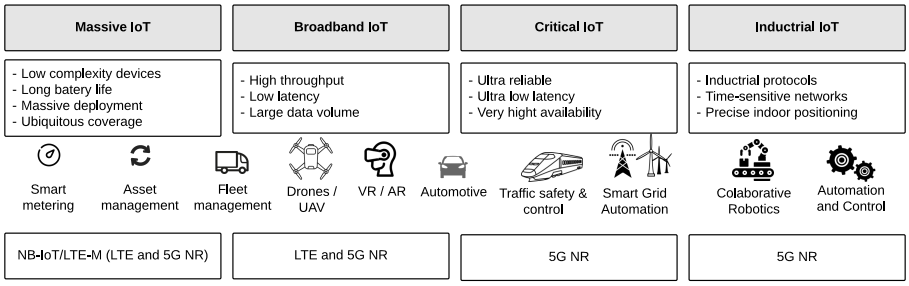



**Fig. 2.2.** Cellular IoT segments.

The development of LTE-M started in 2011 with the 3GPP study item [14]. The initial goals were:

- to extend the LTE device capabilities;
- to provide an alternative to General Packet Radio Services (GPRS) devices with complexity and cost on par with GPRS;
- to improve coverage by 20 dB beyond normal LTE coverage to facilitate deep indoor coverage.

The work on EC-GSM-IoT and NB-IoT was initiated in the Cellular IoT study item [15]. The central objective it this study was to develop solutions that would make CIoT devices competitive in the Low-Power Wireless Networks (LPWAN) market represented mainly by technologies for unlicensed operation with (i) **ultra-low complexity**, (ii) **extreme coverage range**, and (iii) **long device battery life** design in mind. Table 2.1 summarizes the performance objectives of CIoT study.

A reduction in the maximum supported transmission and reception bandwidths, and adopting a single radio frequency receive chain, have reduced the complexity and manufacturing costs comparable to that of a GPRS modem. Moreover, going to a single radio frequency chain helps to improve downlink coverage, while the lower transmit power in uplink causes a corresponding uplink coverage reduction. Another efficient solution for coverage extension is signal repetitions that allow achieving target Maximum Coupling Loss (MCL). An extension of the total system synchronization time has made it possible for IoT devices to switch to the extensive PSM with ultra-low energy consumption level.

## 2.2 Cellular IoT system architecture

### 2.2.1 Network Architecture

The cellular systems architecture comprises a Radio Access Network (RAN) part and a Core Network (CN) part. The RAN connects a device to the network via the radio interface, also known as the *access stratum*, while the CN connects the RAN to an external network. This

**Table 2.1.** Performance objectives for the cellular IoT.

| Objective | Description | Requirements |
|---|---|---|
| Coverage | The MCL of the system was required to surpass GPRS coverage by 20 dB. The coverage reference equal to 144 dB | 164 dB |
| Throughput | The candidate technologies should support a data rate of at least 160 bps at the coverage limit | 160 bps |
| Latency | A latency requirement of 10 s was set for high-priority reports | 10 s |
| Capacity | Capacity supporting 40 connected devices per household in downtown London with roughly 1500 households/km$^2$ was required | 60,000 d/km$^2$ |
| Power efficiency | The operation without the access to a mains power source and with a 5-watt-hour (Wh) battery was expected to last for at least 10 years in the most extreme coverage situations | 10 years |
| Complexity | An ultra-low device complexity was required to support mass production and deployment of devices of ultra-low cost | Ultra-low |

can be a public network such as the Internet or a private enterprise network. The overall purpose of the radio and core network segment is to provide an efficient data transfer between the external network and the devices served by the cellular system.

Fig. 2.3 illustrates two possible connection scenarios for CIoT devices namely:

- connection to GSM/GPRS network for EC-GSM-IoT,
- connection to LTE/Evolved Packet Core (EPC) network for LTE-M,
- connection to both networks for NB-IoT.

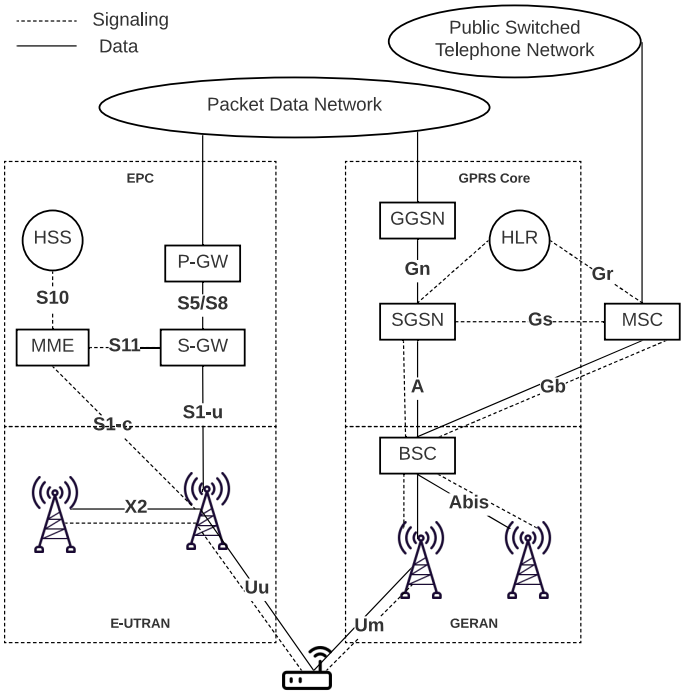



**Fig. 2.3.** Network architecture for cellular IoT.

The LTE radio network, also known as Evolved Universal Terrestrial Radio Access Network (E-UTRAN), together with EPC, define the Evolved Packet System (EPS). The Packet Data network Gateway (P-GW) provides the connection to an external packet data network. The Serving Gateway (S-GW) routes user data packets from the P-GW to an eNodeB (eNB) that transmits them over the LTE radio interface (Uu) to an end-user device. The connection between the P-GW and the device is established by means of a so-called EPS bearer, which is associated with certain QoS requirements. These correspond to the data rate and latency requirements expected from the provided service.

Data and control signaling is separated by means of the *user plane* and *control plane*. The Mobility Management Entity (MME), which is responsible for, e.g. idle mode tracking, is connected to the eNB via the control plane. The MME also handles subscriber authentication and is connected to the Home Subscriber Service (HSS) database for this purpose. It maps the EPS bearer to radio bearers that provides the needed QoS over the LTE radio interface.

In the GPRS core the Gateway GPRS Support Node (GGSN) acts as the link to the external packet data networks. The Serving GPRS Support Node (SGSN) fills a role similar to the MME and handles idle mode functions, as well as authentication toward the Home Location Register (HLR), which keeps track of the subscriber information. It also routes the user data to the radio network. In an LTE network the eNB is the single infrastructure node in the RAN. In case of General Radio Access Network (GERAN), the eNB functionality is distributed across a Base Station Controller (BSC) and the Base Transceiver Station (BTS). One of the most fundamental differences between GSM/Enhanced Data rates for GSM (EDGE) and the EPS architectures is that GSM/EDGE supports a circuit-switched domain for the handling of voice calls, in addition to the packet-switched domain. The EPS only operates in the packet-switched domain. The Mobile Switching Center (MSC) is the GSM CN node that connects the classic Public Switched Telephone Network (PSTN) to GERAN.

### 2.2.2 Radio protocol architecture

Fig. 2.4 depicts the LTE radio protocol stack including the **control** and **user plane** layers as seen by a device.

**User plane protocol stack**

- **IP layer** carries application data and terminates in the P-GW. IP is not a radio protocol, but mentioned to introduce the interface between the device and the P-GW. The IP packet is transported between the P-GW, the S-GW and the eNB using the GPRS Tunnel Protocol (GTP).
- The **Non-Access Stratum (NAS)** and Radio Resource Control (RRC) layers are unique to the control plane. A message-based IP transport protocol (Stream Control Transmission Protocol (SCTP)) provides a reliable message transfer between the eNB and MME.

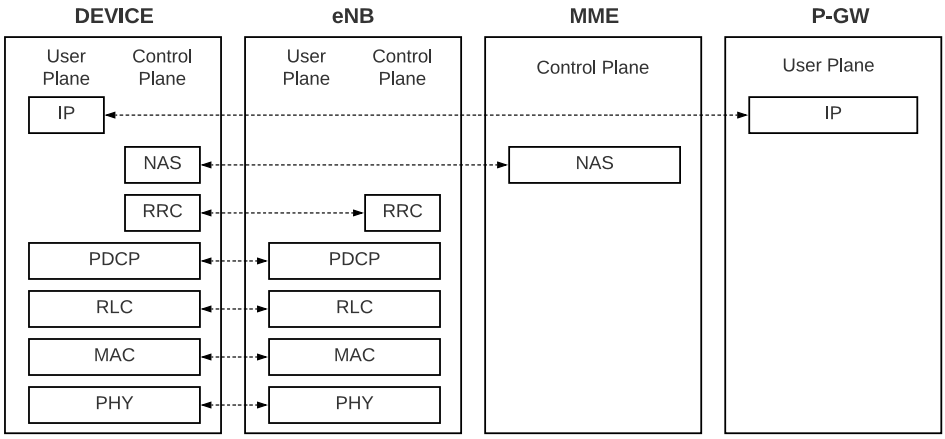



**Fig. 2.4.** Radio protocol stack for cellular IoT.

- The **RRC** handles the overall configuration of a cell including the Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC) and physical layers. It is responsible for the connection control, including connection setup, (re-)configuration, handover and release.
- The **PDCP**, the RLC, the MAC and the physical layers are common to the control and user planes. The PDCP performs Robust Header Compression (RoHC) on incoming IP packets and manages integrity protection and ciphering of the control plane and ciphering of the user plane data sent over the access stratum. It acts as a mobility anchor for devices in RRC connected mode. It buffers, and if needed, retransmits packets received during a handover between two cells. The PDCP packets are transferred to the RLC layer, which handles a first level or retransmission in an established connection and makes sure that received RLC packets are delivered in sequence to the PDCP layers.
- The **RLC** layer handles concatenation and segmentation of PDCP Protocol Data Units (PDU) into RLC Service Data Units (SDU). The RLC SDUs are mapped on RLC PDUs which are transferred to the MAC layer. Each RLC PDU is associated with a *radio bearer* and a logical channel. Two types of radio bearers are supported: Signaling Radio Bearer (SRB)s and Data Radio Bearer (DRB)s. The SRBs are sent over the control plane and bears the logical channels. The DRBs are sent over the user plane and are associated with the data transfer channel. The distinction provided by the bearers and the logical channels allows a network to provide a requested QoS for different types of signaling and data services.
- **MAC** manages multiplexing of bearers and their logical channels with MAC control elements according to specified and configured priorities. The MAC control elements are used to convey information related to an ongoing connection, e.g., the data buffer status report. MAC is responsible for the random-access procedure and Hybrid Automatic Repeat Request (HARQ) retransmissions. The MAC PDUs are forwarded to the

physical layer, which is responsible for the physical layer functions and different services (encoding, decoding, modulation, and demodulation).

The NAS protocol was dedicated to support only signaling. Since Release 13 the NAS may also carry user data. This exception to the general architecture was introduced as part of the **Control plane CIoT EPS optimization** [16] feature. The control plane NAS messages sent between the device and MME are transparent to the eNB.

Fig. 2.5 shows the data transfer through the protocol stack. At each layer, a header (H) is appended to the SDU to form the PDU, and at the physical layer also a Cycle Redundancy Check (CRC) is attached to the transport block.

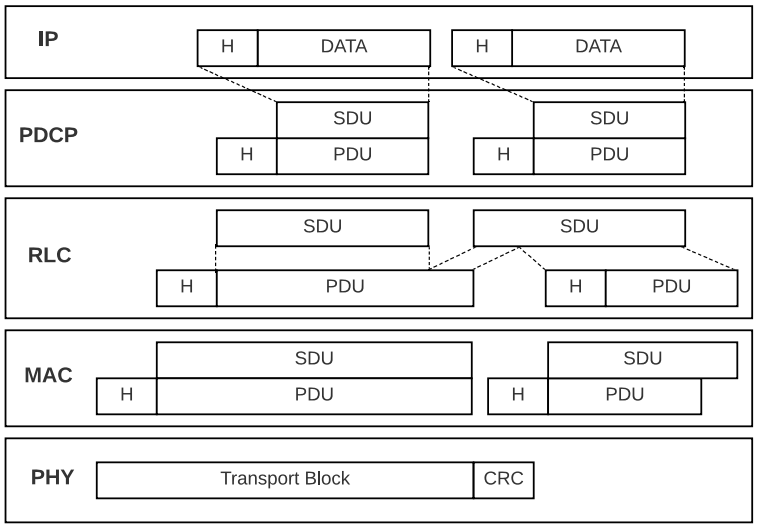



**Fig. 2.5.** Data flow in cellular IoT.

The GPRS protocol stack also includes the RRC, PDCP, RLC, MAC and physical layers. Although the same naming conventions are used in GPRS and LTE, the functionality belonging to the different layers has evolved. GPRS non-access stratum signaling between the device and SGSN is defined by means of the Logical Link Control (LLC) and Sub-Network Dependent Convergence Protocol (SNDCP) protocols. LLC handles encryption and integrity protection, while SNDCP manages RoHC. This functionality is similar to that provided by the LTE PDCP for providing compression and access stratum security. As a comparison remember that the PDCP terminates in the E-UTRAN, while LLC and SNDCP terminates in the GPRS CN.

### 2.2.3 Control Plane and User Plane Optimization

Fig. 2.6(a) illustrates a significant signaling overhead to transmit a small packet in the uplink. In 3GPP study [17], several critical protocols and procedure enhancements have

been introduced to handle small data transmissions efficiently and optimize the device energy consumption. In 3GPP Release 13, two solutions were adopted for streamlining the setup procedure in RANs to support small and infrequent data transmissions.

- **User plane CIoT EPS Optimization** or **RRC Resume procedure**. It allows a device to resume a connection previously suspended including the PDCP state, the Access Statum (AS) security and RRC configurations. This eliminates the need to negotiate AS security and configuring the radio interface. It supports the PDCP to use RoHC already from the first data transmission in a resumed connection. This functionality is based on a resume identity, which identifies a suspended connection. It is signaled in the RRC Connection Release message from the network to a device when a connection is suspended. The device signals the resume identity back to the network when it wants to resume a connection using Message 3 including the RRC Connection Resume Request message. The RRC resume procedure allows uplink data to be multiplexed with the RRC signaling already in Message 5. This multiplexing between RLC packet data units containing user data and control signaling is achieved in the MAC layer.
- **Control Plane CIoT EPS optimizations** or **Data over Non Access Statum (DoNAS)**. It uses legacy connection setup message flow, but in the RRC Connection Setup Complete message, a NAS container is used for transmitting uplink user data over the control plane. For the NAS interface terminated in the MME, security is negotiated when a device attaches to a network. This means that data sent in a NAS container is both integrity protected and ciphered. To support uplink and downlink transmissions after the connection establishment procedure Release 13 specifies two RRC messages (uplink (UL)/downlink (DL) information transfer) that only carry a NAS container where user data can be inserted. Since DoNAS transfers data over an SRB the quality of service frame work developed for LTE DRBs does not apply.

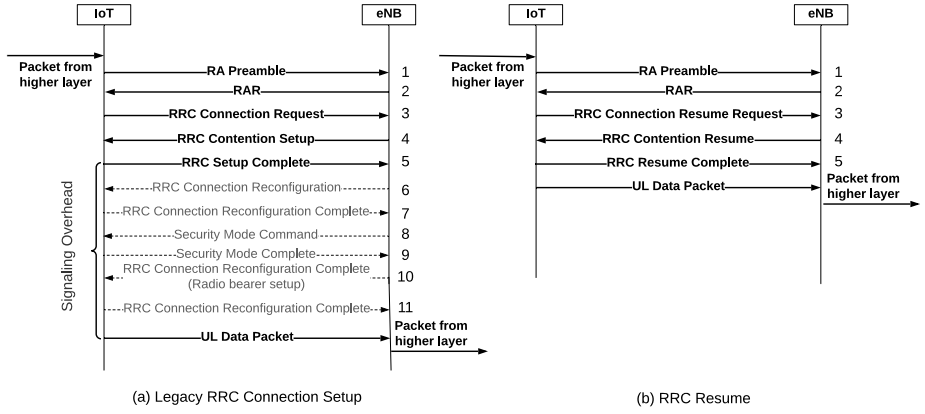



**Fig. 2.6.** Message flow diagram for the RRC Connection Setup and RRC Resume procedures.

### 2.2.4 Early Data Transmission

For a Release 13 or 14 device, uplink data and downlink data can at the earliest be delivered in RRC resume complete and the next one, respectively. 3GPP Release 15 introduces Early Data Transmission (EDT) that allows a device to transmit its uplink data in the RRC Connection Request and its downlink data in RRC Connection Setup, respectively. In this case, the device can complete its data transmission in idle mode without transitioning to connected mode.

EDT is limited to small data payloads. Its configuration is provided in System Information Block Type 2 (SIB2). A device can use the EDT procedure to transmit its uplink data only when the number of payload bits is less than the maximum Transport Block Size (TBS) permitted. A device indicates its intent to use EDT by initiating a random access procedure with an Narrowband Physical Random Access Channel (NPRACH) preamble randomly selected from a set of preambles configured for the EDT procedure. In this case, upon detecting the NPRACH preamble, the base station knows that the device attempts to transmit its uplink data through EDT, and, therefore, it can include an EDT uplink grant for Message 3 in Message 2. The base station can reject the device's EDT request. The EDT uplink grant includes information about the Modulation Coding Scheme (MSC) and number of repetitions associated with the maximum TBS.

EDT is enabled for mobile-originated access both for the User Plane CIoT EPS optimization procedure and Control Plane CIoT EPS optimization procedure. The user plane version builds on the RRC Resume procedure and makes use of the RRC Connection Resume Request instead of RRC Connection Request. RRC Resume Complete can be used for a potential downlink data piggybacking. Both of them include the needed NAS container carrying the data on an SRB over the NAS.

## 2.3 LTE-M

The LTE extensions for improved support for MTC and IoT have their origin in the 3GPP study item *Study on provision of low-cost MTC User Equipments based on LTE* [14], which is usually referred to as LTE-M. Since then a number of successive work items have been completed to form the basis for Release 12 – 15, which can be summarised as follows:

- **Release 12 or MTC** work item: *Low cost and enhanced coverage MTC User Equipment (UE) for LTE* [18], introduced LTE device category 0 (Cat-0);
- **Release 13 or eMTC** work item: *Further LTE Physical Layer Enhancements for MTC* [19], introduced the Coverage Enhancement (CE) modes A and B, as well as LTE device category M1 (Cat-M1);
- **Release 14 or feMTC** work item: *Further Enhanced MTC for LTE* [20], introduced various improvements for support of higher data rates, improved Voice over LTE (VoLTE)

support, improved positioning, *multicast support*, as well as the new LTE device category M2 (Cat-M2);

- **Release 15 or efeMTC** work item: *Even Further Enhanced MTC for LTE* [21], introduced further improvements for reduced latency and power consumption, improved spectral efficiency, and new use cases.

*All the Releases 14 and 15 features can be enabled through a software upgrade of the existing LTE network equipment. In many cases, it may also be possible to upgrade the software/firmware in existing devices to support the new features.*

### 2.3.1 Radio Access Design Principles

The physical layer changes in LTE-M are motivated by the requirements on low device cost, deep coverage, and long battery lifetime. The low device cost is mainly enabled by reduced transmit and receive bandwidths. The deep coverage is achieved through repetition techniques. The long battery lifetime is made possible by the introduction of long sleeping cycles and efforts to keep the overhead from both higher and lower layer control signaling as small as possible.

- **Coverage enhancement**. Release 13 introduced two CE modes: CE mode A, supporting up to 32 repetitions for the data channels, and CE mode B, supporting up to 2048 repetitions. Evaluations show that the initial coverage target of 20 dB can be reached using the repetitions available in CE mode B.
- **Long device battery lifetime**. Release 12 has introduced PSM, Release 13 - eDRX to support for a device battery lifetime of many years. A device can shut down its transceiver and only keep a basic oscillator running a clock. The reachability during PSM is set by the Tracking Area Update (TAU) procedure timers, the TAU forces devices to update their cell presence from time to time. eDRX significantly extends the maximum device sleep cycle defined by DRX. Mode details on the power saving solutions in CIoT are given in section 3.2.
- **Deployment flexibility** LTE-M can be deployed in a wide range of frequency bands. Both paired bands for Frequency-Division Duplex (FDD) operation and unpaired bands for Time Division Duplex (TDD) operation are supported, and new bands have been added in every release.
- **Coexistence with LTE**. The fundamental downlink and uplink transmission schemes are the same as in LTE, meaning that LTE-M transmissions and LTE transmissions can coexist in the same cell on the same carrier, and the resources can be shared dynamically between LTE-M and LTE users. If an operator has a large spectrum allocation for LTE, then there is also a large bandwidth available for LTE-M traffic.

### 2.3.2 Deployment and numerology

The LTE-M design builds on the solutions available in LTE. It uses Orthogonal Frequency Division Multiplexing (OFDM) in DL and Single Carrier Frequency Division Multiple Access (SC-FDMA) in UL. LTE-M reuses LTE numerologies for channel raster, subcarrier spacing, Cyclic Prefix (CP) lengths, resource grid, and frame structure. LTE-M supports frequency bands 1–8, 11–14, 18–21, 25–28, 31, 39–41, 66, 71–74, and 85. It operated at the same system bandwidths as LTE: 1.4, 3, 5, 10, 15, and 20 MHz. LTE-M reuses LTE's pilot and synchronization signals, which are located in the center of the LTE system bandwidth aligned with a channel raster of 100 kHz.

The overall time frame structure is illustrated in Fig. 2.7. One time slot consists of 7 OFDM symbols in case of normal CP length and 6 OFDM symbols in case of extended CP length giving in total 0.5 ms. The normal CP length is designed to support propagation conditions with a delay spread up to 4.7 ms. The extended CP is intended to support deployments where the delay spread is up to 16.7 ms. Two slots are grouped into a subframe of 1 ms, 10 consecutive subframes build a frame, 1024 frames compose a hyperframe. Each subframe can be uniquely identified by a Hyper System Frame Number (H-SFN), a System Frame Number (SFN), and subframe number.

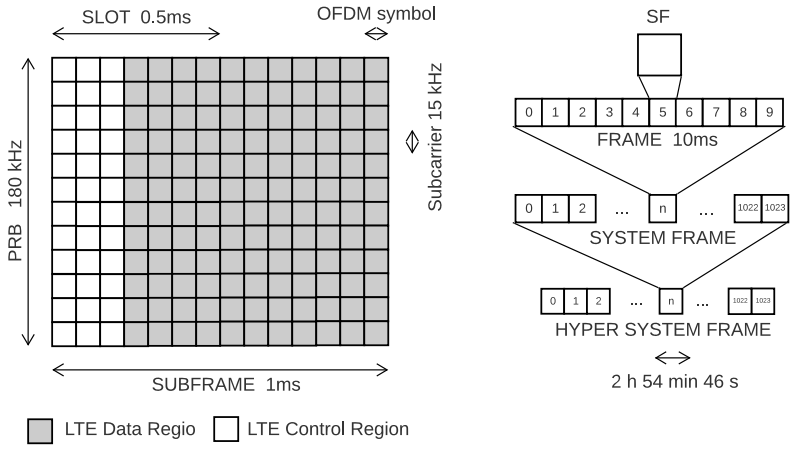



**Fig. 2.7.** Resource grid and Time Frame structure in LTE-M.

One Physical Resource Block (PRB) spans 12 subcarriers with the 15-kHz subcarrier spacing corresponing to 180 kHz. When full-PRB transmission is used, the smallest time-frequency resource that can be scheduled to a device is one PRB pair mapped over two slots, which for the normal CP length case (with 7 OFDM symbols per slot) corresponds to 12 subcarriers over 14 OFDM symbols as illustrated in Fig. 2.7.

LTE-M supports both TDD operation and FDD operation. It can be deployed both in paired FDD bands and unpaired TDD bands. In FDD operation, two different carrier

frequencies are used for downlink and uplink. If the device supports Half-Duplex Frequency-Division Duplex (HD-FDD) operation, it can perform reception and transmission at the same time, whereas if the device only supports HD-FDD operation, it has to switch back and forth between reception and transmission. LTE-M devices support HD-FDD operation type B due to the only one local oscillator for carrier frequency generation in downlink and uplink. A guard subframe is inserted at every switch from downlink to uplink and from uplink to downlink, giving the device time to retune its carrier frequency.

Low-cost LTE-M devices are only required to support a maximum channel bandwidth of 6 PRBs for transmission and reception. The transmissions are restricted to take place within one out of a number of non-overlapping narrowbands of size 6 PRBs as for the 15-MHz system bandwidth case. For all system bandwidths except for the smallest one, the system bandwidth cannot be evenly divided into narrowbands, which means that some PRBs that are not part of any narrowband. Therefore, the system with the bandwidth ranging from 1.4 till 20 MHz has 1, 2, 4, 8, 12, and 16 narrowbands, respectively.

### 2.3.3 Overview of channels and signals

To indicate which downlink subframes are valid for LTE-M transmission, the network broadcasts a cell-specific subframe bitmap corresponding to the subframes within 1 or 4 frames. It indicates subframes that are reserved for special signals, or are *invalid* for LTE-M due to the non LTE-M transmissions that occupy the entire bandwidth, e.g. multicast and broadcast single frequency network (MBSFN) transmissions.

The downlink subframe structure in LTE consists of a control region and a data region, as shown in 2.7. The control region consists of one or more OFDM symbols at the beginning of the subframe (white blocks). The data region consists of the remaining OFDM symbols in the subframe (gray blocks). In LTE-M, both the control channel and the data channel are mapped to the LTE data region to avoid collisions between the LTE control channels and LTE-M transmissions, meaning that the downlink subframe structure in LTE-M only uses a part of the downlink subframe Resource Elements (RE)s. The starting symbol for LTE-M transmissions is cell-specific. The possible LTE-M starting symbols are the second, third, and fourth symbol in the subframe, except for the system bandwidth 1.4 MHz, where the possible LTE-M starting symbols are the third, fourth, and fifth symbol.

LTE-M supports the set of **downlink** channels and signals as follows:

- **Synchronization signals:**
  - **Primary Syncronization Signal (PSS) and Secondary Syncronization Signal (SSS)**. These signals occupy subframes #0 and #5 in FDD, and subframes #0,#1,#5 and #6 in TDD, and spread over 62 subcarriers. PSS and SSS are used for the time and frequency synchronization, and cell identification as a part of the cell selection procedure. PSS and SSS are transmitted periodically. Therefore, a device can accumulate the received signal over multiple frames to achieve sufficient

acquisition performance, without additional repetitions but at the cost of increased acquisition delay.

- **Resynchronization Signal (RSS)** is introduced for enhancing energy efficiency when a device needs to re-acquire time and frequency synchronization toward a cell. It is optional for the base station to transmit the RSS, and optional for the device to use it.
- **Cell-specific Reference Signal (CRS)**. Downlink reference signals are transmitted by the Base Station (BS) to allow the device to estimate the downlink propagation channel to be able to demodulate the downlink physical channels, and perform downlink reference signal strength or quality measurements. CRS is used for demodulation of the broadcast and data channels.
- **Demodulation Reference Signal (DMRS)** can be used for demodulation of data and control channel as it is transmitted on the same logical antenna port as the associated channels.
- **Positioning Reference Signal (PRS)** is used for the observed time difference of arrival multilateration positioning method, where the position of a receiving device is determined based on differences in time of arrival between PRS signals from different time-synchronized BSs.

- **Physical Downlink Contol Channel (PDCCH)** occupies subframes #0 and #9 for FDD, #0 and #5 for TDD, and 72 subcarriers. The channel is used to deliver the Master Information Block (MIB) that provides essential information for the device to operate in the network.
- **MTC Physical Downlink Contol Channel (MPDCCH)**. It is used to carry Downlink Control Information (DCI). An LTE-M device needs to monitor MPDCCH for the uplink power control command, uplink grant information, downlink scheduling information, paging indication, notification of changes in multicast control channel and HARQ Acknowledgement (ACK) feedback.
- **Physical Downlink Shared Channel (PDSCH)**. It is primarily used to transmit unicast and multicast data, system information, paging message, and radio access related messages. The data packet from higher layers is segmented into one or more transport blocks, and PDSCH transmits one Transport Block (TB) at a time.

The transmission from a device needs to be contiguous in the frequency domain. Therefore, resources for random access and control channel are usually allocated near the system bandwidth's edges. The uplink reference signals are transmitted together with uplink data and control channels or separately for the radio channel's sounding. LTE-M supports the set of uplink channels and signals as follows:

- **Physical Random Access Channel (PRACH)**. The device use PRACH to initialize connection. It allows the serving BS to estimate the arrival time of uplink transmission. The PRACH configuration is cell-specific and supports different configurations in terms of mapping the signal on the subframe structure.

- **Uplink Reference Signals** are predefined signals transmitted by the device to allow the BS to estimate the uplink propagation channel to be able to demodulate uplink physical channels, perform uplink quality measurements, and issue timing advance commands.
  - **DMRS** is dedicated for data and control channel demodulation.
  - **Sounding Reference Signal (SRS)**. The network can reserve the last SC-FDMA symbol of some uplink subframes in a cell for SRS transmission for sounding of the radio channel.
- **Physical Uplink Shared Channel (PUSCH)** is primarily used to transmit unicast data. The data packet from higher layers is segmented into one or more TB, and PUSCH transmits one TB at a time. PUSCH is also used for the transmission of Uplink Control Information (UCI) when a periodic Channel State Information (CSI) transmission is triggered by setting the CSI request bit in DCI, or in case of collision between data and control channels.
- **Physical Uplink Control Channel (PUCCH)** is used to carry uplink scheduling request, downlink HARQ feedback and downlink CSI.

## 2.4 NB-IoT

The study group for [15] focused on the evolution of GSM/GPRS technologies, while some GSM operators considered refarming GSM spectrum to LPWAN dedicated for IoT services. This consideration triggered the study on non-GSM backward compatible technologies that provided a firm ground for the NB-IoT technology standardized in Release 13.

Since its first release in 2016, NB-IoT has gone through additional releases as follows:

- **3GPP Release 14** introduced category NB2 (Cat-NB2) device and support of 2 HARQ processes to improve device energy efficiency; multicast support, non-anchor carrier paging, and random access, mobility enhancement, the DL channel quality reporting during random access procedures for system performance improvement. A new 14 dBm device power class and positioning feature was introduced.
- **3GPP Release 15** introduced EDT, Wake Up Signal (WUS), quick RRC release procedure, and relaxed cell re-selection monitoring. It improved access barring, system information acquisition, measurement accuracy, device power headroom reporting, and device differentiation. Finally, the support of small cell deployment, extend cell radius up to 120 km, flexible uses of stand-alone carriers was introduced.

### 2.4.1 Radio Access Design Principles

NB-IoT is designed for ultra low cost mMTC. Low device complexity is one of the main design objectives. It is designed to offer substantial coverage improvements over GPRS and enable a long battery lifetime. NB-IoT has been designed to give maximal deployment flexibility.

- **Low Device Complexity and Cost**. NB-IoT is designed for allowing low-complexity receiver processing during initial cell selection and connection. For initial cell selection, a device needs to search for only one synchronization sequence to synchronize to the network with time and frequency. In connected mode, low device complexity is facilitated by restricting the DL TBS. Instead of using the turbo code, NB-IoT adopts a simple convolutional code, i.e., the tail-biting convolution code, in the DL channels. In addition, NB-IoT does not use higher-order modulations or multi-layer multiple-input multiple-output transmissions. Furthermore, a device needs to support only half-duplex operation and is not required to listen to the DL while transmitting in the UL, and vice versa. All the performance objectives of NB-IoT can be achieved with one transmit-and-receive antenna in the device. NB-IoT is designed for allowing relaxed oscillator accuracy in the device.
- **Coverage Enhancement** is mainly achieved by trading off data rate for coverage. NB-IoT has been designed to use a close to constant envelope waveform in the UL. This is an important factor for devices in extreme coverage- and power-limited situations because it minimizes the need to back off the output power from the maximum configurable level.
- **Long Device Battery Lifetime**. 3GPP Releases 12 and 13 introduced both eDRX and PSM to support this type of operation and optimize device power consumption. NB-IoT also adopts DRX as a major tool for achieving energy efficiency. In Release 13, the Connected DRX (cDRX) cycles were extended from 2.56 to 10.24 s for NB-IoT.
- **Support of Massive Number of Devices**. NB-IoT achieves high capacity in terms of the number of devices supported on a single narrowband carrier. NB-IoT UL waveforms include various bandwidth options. While a waveform of wide bandwidth (e.g., 180 kHz) is beneficial for devices in good coverage, waveforms of small bandwidths are more spectrally efficient from the system point of view for serving devices in bad coverage.
- **Deployment Flexibility**. To support maximum deployment flexibility and prepare for re-farming scenarios, NB-IoT supports three modes of operation: stand-alone, in-band, and guard-band.

### 2.4.2 Deployment and numerology

- **Stand-alone Mode**. NB-IoT can be deployed as a stand-alone carrier using any available spectrum with bandwidth larger than 180 kHz. An example is to deploy NB-IoT in the GSM band by re-farming part of its GSM spectrum.
- **In-Band and Guard-Band Modes**. NB-IoT is designed for deployment in LTE networks, either using one of the LTE PRBs, or using the LTE guard-band. NB-IoT can be deployed using one LTE PRB, or using the unused bandwidth in the guard-band. The guard-band deployment makes use of the fact that the occupied bandwidth of the LTE signal is roughly 90% of channel bandwidth when the LTE carrier bandwidth is 3, 5, 10, 15, or 20 MHz. There is roughly 5% of the LTE channel bandwidth on each side available as guard-band.

NB-IoT deployment inside an LTE carrier does not required any guard-band between NB-IoT and LTE PRBs. To minimize the impact on the existing LTE deployments and devices, NB-IoT physical layer waveforms must preserve orthogonality with the LTE signal in adjacent PRBs. NB-IoT should be able to share the same time-frequency resource grids as LTE the same way as different LTE physical channels share time-frequency resources. Because legacy LTE devices will not be aware of the NB-IoT operation, NB-IoT transmissions should not collide with essential LTE transmissions, e.g. PDCCH transmissions for scheduling information, paging indicators, Random Access Response (RAR), etc. The PDCCH spans over the entire range of LTE PRBs in frequency, and may span up to the first three OFDM symbols in every subframe in time. Therefore, the REs of the first three OFDM symbols in every subframe cannot be used by NB-IoT DL channels.

The MBSFN signal spans one entire subframe in the time dimension, and all PRBs in the frequency dimension. If one subframe is configured as an LTE MBSFN subframe, that subframe is not used by NB-IoT. Furthermore, resources used by LTE PSS, SSS, and Physical Broadcast Channel (PBCH) are protected by NB-IoT , avoiding to use any of the middle six PRBs in an LTE carrier in case of in-band deployment.

In case of stand-alone deployment, the NB-IoT anchor carrier can always be placed on a re-farmed GSM 200 kHz channel. However, for LTE in-band and guard-band deployments, it is not possible to place the NB-IoT anchor carrier with a center frequency exactly on the 100 kHz grid.

To facilitate an efficient initial cell selection for NB-IoT , in case of LTE in-band deployment, an anchor carrier is required to be configured on a PRB that has the smallest raster offset. A full list of PRB indexes for in-band deployment is given in Table 2.2.

**Table 2.2.** Suitable PRB indexes for NB-IoT anchor carrier in the in-band deployment.

| **Bandwidth** | **Allowed PRB Indexes for NB-IoT Anchor Carrier** | | **Raster Offset** |
|---|---|---|---|
| 1.4 MHz | Not supported | Not applicable | |
| 3 MHz | 2 | 12 | 7.5 kHz |
| 5 MHz | 2, 7 | 17, 22 | 7.5 kHz |
| 10 MHz | 4, 9, 14, 19 | 30, 35, 40, 45 | 2.5 kHz |
| 15 MHz | 2, 7, 12, 17, 22, 27, 32 | 42, 47, 52, 57, 62, 67, 72 | 7.5 kHz |
| 20 MHz | 4, 9, 14, 19, 24, 29, 34, 39, 44 | 55, 60, 65, 70, 75, 80, 85, 90, 95 | 2.5 kHz |

NB-IoT supports multicarrier operation. Because for NB-IoT it suffices to have one anchor carrier for facilitating device initial cell selection, the additional carriers may be located with an offset of up to 47.5 kHz outside the 100 kHz raster grid. These additional carriers are referred to as non-anchor or secondary carriers. A non-anchor carrier does not carry the physical channels that are required for device initial cell selection.

The overall time frame structure in NB-IoT is the same is in LTE-M. NB-IoT employs Orthogonal Frequency-Division Multiple Access (OFDMA) in the DL, and the concept of

PRBs is used for specifying the mapping of physical channels and signals onto REs. NB-IoT UL uses SC-FDMA with 15 kHz subcarrier spacing for multitone transmissions. In this case, the same numerologies as NB-IoT DL are used. Multitone transmissions may use 12, 6, or 3 subcarriers. In addition, single-tone transmissions are supported, and, in that case, the time-frequency resource grids can be based on 15 or 3.75 kHz subcarrier spacing.

For the UL, the Resource Unit (RU) is used to specify the mapping of the UL physical channels onto REs. The definition of the RU depends on the configured subcarrier spacing and the number of subcarriers allocated to the UL transmission. In the basic case, where 12 subcarriers using a spacing of 15 kHz are allocated, the RU corresponds to the PRB pair. In the case of sub-PRB scheduling assignments of 6, 3, or 1 subcarrier, then the RU is expanded in time to compensate for the diminishing frequency allocation. For the single subcarrier allocation, also known as single-tone allocation, the NB-IoT RU concept supports an additional subcarrier spacing of 3.75 kHz.

### 2.4.3 Overview of channels and signals

At a high level, the DL physical channels and signals are time-multiplexed, except for Narrowband Reference Signal (NRS), which are present in every subframe carrying Narrowband Physical Downlink Shared Channel (NPDSCH), or Narrowband Physical Downlink Control Channel (NPDCCH). Figure 2.8 shows time-multiplexing of different mandatory DL physical channels in a 20-ms period. The same pattern is repeated in subsequent periods. As shown, Narrowband Physical Broadcast Channel (NPBCH) and Narrowband Primary Synchronization Signal (NPSS) are transmitted in subframes 0 and 5 in every frame, respectively, and Narrowband Secondary Synchronization Signal (NSSS) is transmitted in subframe 9 in every other frame. The remaining subframes may be used to transmit NPDCCH or NPDSCH. The notion of invalid subframe can used when NB-IoT is deployed inside an LTE carrier having MBSFN subframes configured.

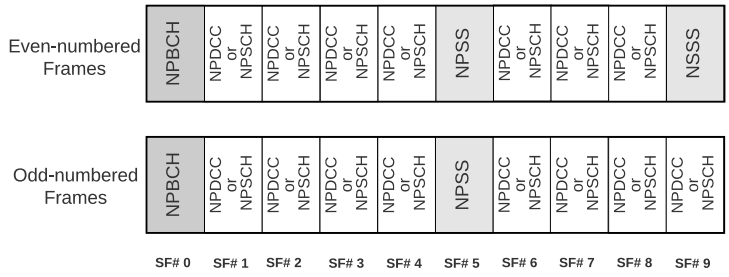



**Fig. 2.8.** Time-multiplexing of downlink physical channels on an NB-IoT anchor carrier.

NB-IoT supports the set of downlink channels and signals as follows:

- **Downlink synchronization signals:**

  - **NPSS and NSSS**. These signals allow a device to synchronize to a cell. They are transmitted in certain subframes based on an 80-ms repetition interval. By synchronizing to NPSS and NSSS, the device will be able to detect the cell identity number, and identify the framing information within an 80-ms NPSS and NSSS repetition interval. NPSS and NSSS are designed to allow a device to use an unified synchronization algorithm during initial acquisition without knowing the NB-IoT deployment mode. This is achieved by avoiding collision with REs used by LTE as much as possible. LTE may use up to the first three OFDM symbols in every subframe for PDCCH. To avoid potential collisions with LTE PDCCH, the first three OFDM symbols are not used in the subframes that carry either NPSS or NSSS. This leaves only 11 OFDM symbols per subframe available for NPSS and NSSS.
  - **NRS**. The NRS is used to allow the device to estimate the DL propagation channel coefficients, and to perform DL signal strength and quality measurements both in idle and connected mode procedures. It is mapped to certain subcarriers in the last two OFDM symbols in every slot within a subframe that carries NPBCH, NPDCCH, or NPDSCH. NRS may be transmitted also in subframes that do not have any NPDCCH or NPDSCH scheduled.
- **NPBCH** is used to deliver the NB-IoT MIB, which provides essential information for the device to operate in the NB-IoT network. NPBCH uses a 640 ms Time Transmission Interval (TTI), but within the TTI only subframe 0 is used in every radio frame.
- **NPDCCH** is used to carry Downlink Control Information (DCI). A device needs to monitor NPDCCH for three types of information: (i) UL grant information, (ii) DL scheduling information, (iii) Indicator of paging or System Information (SI) update. An NPDCCH subframe is divided into two narrowband Control channel element (NCCE)s. The number of REs available for one NCCE depends on NB-IoT deployment modes and the number of logical antenna ports, and for the in-band deployment it further depends on the configuration of the LTE cell.
- **NPDSCH**. The NPDSCH is used to transmit unicast and multicast data. The data packet from high layers is segmented into one or more glstbs, and NPDSCH transmits one TB at a time. NPDSCH is also used to transmit broadcast information such as SI messages. NPDSCH has a similar subframe-level resource mapping as NPDCCH. One NPDSCH subframe, however, can carry only one TB. The basic RU for NPDSCH is one PRB pair. The starting OFDM symbol in an NPDSCH subframe is always the fourth symbol in the subframe.

The supported NB-IoT uplink channels and signals are listed below.

- **NPRACH** is used by the device to initialize connection, and it allows the serving BS to estimate the time of arrival of the received NPRACH signal. This time reflects the round-trip propagation delay between the BS and device. Up to three NPRACH configurations can be used in a cell to support devices in different coverage classes. Different configurations are separated by using different time-frequency resources.

- **Narrowband Physical Uplink Shared Channel (NPUSCH)** is used to carry UL user data and control information from higher layers. Additionally, NPUSCH also carries HARQ acknowledgments for NPDSCH. The maximum device-scheduled bandwidth can be only one PRB.
- **DMRS** is always associated with NPUSCH, and is transmitted in every NPUSCH slot. The bandwidth of DMRS is identical to the associated NPUSCH. New DMRS sequences are introduced to support NB-IoT DMRS with a bandwidth smaller than 180 kHz.

## 2.5 LTE-M and NB-IoT Comparison

LTE-M and NB-IoT are both support a long list of frequency bands in the range 450 MHz to just below 3 GHz.

- **Cat-M1** and **Cat-M2** devices are capable of operating in Evolved Universal Terrestrial Radio Access (E-UTRA) FDD bands 1, 2, 3, 4, 5, 7, 8, 11, 12, 13, 14, 18, 19, 20, 21, 25, 26, 27, 28, 31, 66, 71, 72, 73, 74 and 85 in both half-duplex and full-duplex FDD mode, and in TDD bands 39, 40 and 41.
- **Cat-NB1** and **Cat-NB2** support E-UTRA bands 1, 2, 3, 4, 5, 8, 11, 12, 13, 14, 17, 18, 19, 20, 21, 25, 26, 28, 31, 41, 65, 66, 70, 71, 72, 73, 74 and 85 in HD-FDD, and band 41 for FDD.

Both technologies can operate in the NR bands corresponding to the same lists of E-UTRA bands. NB-IoT also supports operation in GSM spectrum thanks to its configurable mode of operation and small spectrum footprint.

LTE-M is natively supported on the LTE system bandwidths of 1.4, 3, 5, 10, 15 and 20 MHz. Cat-M1 devices operate on a channel bandwidth of 1.4 MHz (6 PRBs), while Cat-M2 case utilize 5 MHz (25 PRBs) for its transmissions. NB-IoT operates at a minimal radio frequency system bandwidth of 200 kHz, which is equivalent to 1 PRB of 180 kHz. The system capacity can be scaled by increasing the number of carriers. Carrier aggregation is not supported by NB-IoT, and so it is system capacity and not link capacity that scales with the number of deployed carriers. LTE-M suits well for deployments in narrow, fragmented frequency bands, and systems with larger bandwidth. NB-IoT is the most flexible system in terms of deployment.

LTE-M and NB-IoT demonstrate many similarities because of the parallel development in 3GPP. However, many NB-IoT features are distinct from LTE-M ones, as the former is aimed to be the low-cost massive IoT with advanced capabilities in terms of reachability, device power efficiency, and system capacity. NB-IoT Release 15 does not support voice, connected mode mobility, connected mode device measurements and reporting, and closed-loop power control. The mandatory Control Plane CIoT EPS optimization feature does not support RRC Reconfiguration of the access stratum for a device in RRC connected mode, nor does it support the MAC layer data radio bearer scheduling and prioritization to ensure QoS.

This is due to the data routing over SRB is available for Cat-NB devices. On contrary, LTE-M naturally support a richer set of features than NB-IoT. It supports advanced transmission modes, wideband transmissions, voice, connected mode mobility, and full duplex operation. The massive IoT application use cases go in many cases beyond small and infrequent data transmission, which make LTE-M more preferable for those use cases.

Both LTE-M and NB-IoT support the most extreme coverage and meet the 164 dB MCL requirement. For both technologies the PUCCH is the limiting channel required the longest transmission times to reach the 164 dB coverage target. For LTE-M the MPDCCH needs the maximum number of repetitions (256) to achieve the 1% BLER target set.

The range of data rates achievable for NB-IoT and LTE-M devices in downlink and uplink for the HD-FDD configuration is summarized in Table 2.3 and Table 2.4, respectively. The MAC-layer performance is compared for 164 dB MCL as an extreme scenario, and under error-free conditions.

**Table 2.3.** Half Duplex FDD PDSCH data rates.

| Technology | MAC-layer at 164 dB MCL | MAC-layer peak | PHY-layer peak |
|---|---|---|---|
| Cat-M1 | 279 bps | 300 kbps | 1 Mbps |
| Cat-M2 | $>$ 279 bps | 1.2 Mbps | 4 Mbps |
| Cat-NB1 | 299 bps | 26.2 kbps | 227 kbps |
| Cat-NB2 | 200 bps | 127.3 kbps | 258 kbps |

**Table 2.4.** Half Duplex FDD PUSCH data rates.

| Technology | MAC-layer at 164 dB MCL | MAC-layer peak | PHY-layer peak |
|---|---|---|---|
| Cat-M1 | 363 bps | 3375 kbps | 1 Mbps |
| Cat-M2 | 363 bps | 2.6 Mbps | 7 Mbps |
| Cat-NB1 | 293 bps | 62.6 kbps | 250 kbps |
| Cat-NB2 | 293 bps | 158.5 kbps | 258 kbps |

Note that the Cat-M1 uplink data rate can be further improved compared by means of the larger uplink TBS introduced in Release 14. The Cat-M1/M2 downlink data rates can be improved with HARQ bundling and 10 HARQ processes in the downlink. The cell-edge MAC-layer data rates for NB-IoT and LTE-M are similar and meet the 5G requirement of 160 bps. LTE-M can offer significantly higher data rates thanks to the larger device bandwidths and the lower processing times.

In LTE-M, the achievable latency for small data transmission at the 164 dB MCL ranges from 5 s to 7.7 s considering EDT and RRC Resume procedures, respectively. In the case of NB-IoT, the latency equal to 5.8 s and 9 s can be expected for the same scenario. LTE-M performs slightly better than NB-IoT because the MPDCCH can achieve 164 dB MCL within 256 ms, while NPDCCH requires 512 ms. In terms of device complexity, both technologies are not required to use more than one antenna to fulfill the performance requirements. Device categories with lower power classes have been introduced for both LTE-M and NB-IoT.

## 2.6 Conclusions

This chapter introduced the concept of CIoT defined by 3GPP for supporting massive and critical MTC categories of use cases. We summarized their service requirements and focused on the most important design aspects of LTE-M and NB-IoT technologies. We also indicated important 3GPP studies and releases to give an overview of the CIoT standard evolution and motivation. In the last section, we briefly compare two solutions for massive IoT in terms of their deployment, device complexity, performance, and applications use cases.

# 3

# Group-based Communications in Cellular IoT Networks

*This chapter presents an overview of mechanisms and procedures to deliver group-based services in cellular IoT networks. It concerns the basic power-saving solutions that let IoT devices operated on battery for years, and introduces challenges to deliver network-originated traffic to massive IoT. We introduce our proposal for supporting critical group-based services to improve service latency and device energy consumption. In conclusion, we discuss potential issues and research directions in relation to recent 3GPP solutions for secure paging and new WUS.*

## 3.1 Introduction

Emerging IoT use cases and scenarios move the focus from sporadic data transmissions in the uplink – such as smart metering devices that wake up once a day to send the consumption reports to the monitor – to continuous data transmission to multiple receivers in the downlink as in the example of software/firmware updates [22].

PTM communication is a natural way to interconnect the source of content with numerous receivers. The 3GPP specified MBMS initially in Release 6, and then introduced it for LTE in Release 9 as eMBMS. At the same release, a new MBSFN transmission mode was enabled to improve the quality of service for the edge users, and, overall network spectral efficiency by simultaneously transmitting identical waveforms from a group of well-synchronized BS over a set of conjunct cells [23]. Successively, the SC-PTM operation mode was introduced in Release 13 to support multicasting in a single cell. Finally, in Release 14, SC-PTM was made available for cellular IoT, but in idle mode only. Therefore, no acknowledgments on packet delivery or any device feedback are considered for multicast reception.

In early releases, IoT devices are considered to generate a sporadic uplink transmission and occasionally receive data in the downlink. Therefore, most of the time, devices are inactive. The DRX mechanism is a legacy 3GPP power saving solution that improves devices battery life. It instructs devices to power down or switch off their circuits to prevent battery drain while no data communication is expected. A new PSM or *deep sleep* mode targets use case scenarios when the extra long latency for network-originated services is not an issue. It allows devices to turn off most of their circuits to conserve the battery and reduce energy consumption level to a bare minimum. When in the deep sleep state, the device is unreachable for the network; in the idle state, it still consumes some energy while discontinuously listening

for *paging*. An essential difference between PSM and a full power-off state is that, in PSM, the device stays registered in the network, so it does not need to re-attach to the Packet Data Network (PDN) after wake-up. This feature reduces the signaling overhead and optimizes the device power consumption.

The longer devices remain in a deep sleep mode, the higher the energy saving. However, this solution introduces a long delay for network-originated services since data arriving at RAN cannot be immediately forwarded to devices in PSM. Therefore, it does not suir well delay-critical applications. In response to this issue, an eDRX mechanism was introduced to let devices spend more time in an idle state, but wake up at least once per cycle for paging.

The key driver for enabling SC-PTM in cellular IoT networks is a practical need to perform software/firmware updates in order to keep IoT infrastructure functional, secure, and ensure its coexistence with LTE and 5G. In 3GPP study [3], more general use cases for group-based MTC and potential solutions were presented.

- **Planned data delivery.** A file of interest is assumed to be available for download according to a predefined schedule, e.g., at a specific time and day of a month. The schedule can be embedded at devices by the manufacturer, or stored at devices once communicated by the network. In this case, devices wake up precisely at the given time.
- **Initially unplanned non-critical data delivery.** When a file is available for download, the network shall inform MTC devices about the new multicast schedule when they are reachable, i.e., at the very next paging opportunity, or right after the waking up for the periodic status update. The time interval between the subframe when the multicast schedule was announced and the subframe when the announced transmission starts must be larger than the longest PSM cycle in the multicast group to ensure that all group members are informed about the forthcoming data delivery session.
- **Initially unplanned critical data delivery.** A newly generated critical file/message must be delivered as soon as possible. The BS can inform devices about a new multicast transmission schedule only when they are available for paging. Data blocks can be repeated in several transmissions until all devices receive the content. The time between two successive transmissions is assumed to be fixed. The time between when a schedule for the critical file delivery is announced, and when the first group-based transmission starts, shall be less than the shortest PSM cycle of all group members. The BS can repeat transmission only to devices that have not received the schedule in the previous period.

The first two use case scenarios can be managed within the existing SC-PTM framework as the multicast transmission schedule is either known in advance or computed. The last scenario differs from the second one as the time between when the file gets available and when scheduled for the transmission is critical, and its computation is not trivial. We focus on the last and most challenging scenario to develop a solution for efficient critical file/message dissemination in cellular IoT systems.

In the following sections we provide a solid background on paging, MBMS and the SC-PTM feature for cellular IoT to introduce our multicast framework and paging solution for delay-critical PTM services. In section 3.2 we give more details on power saving solutions in cellular networks, in section 3.3 we explain how SC-PTM communication works, and in section 3.4 we present our solutions and research results. As conclusions we discuss essential paging extensions and new challenges.

## 3.2 Power saving solutions for Cellular IoT

### 3.2.1 DRX

The individual activity pattern of IoT devices is determined by their duty cycle, alternating short *connected* and long *idle* periods. This mechanism is called DRX. Devices are inactive most of the time to prevent their battery drain. The network specifies periodic occasions when devices have to leave the idle state and be available for *paging*.

A device can use DRX in both connected and idle state, as explained in Fig. 3.2. It discontinuously monitors the downlink channel for a Paging Radio Network Temporary Identifier (P-RNTI) indication in DCI. Once the P-RNTI is detected, the device wakes up to check the paging records list in the paging message. If it does not find its Identifier (ID) in the list, it stays in the idle state. Otherwise, it has to initiate an RRC connection procedure.

A device in PSM wakes up only when it wants to send data, or it has to update its cell registration via a TAU procedure. The PSM cycle is defined by the TAU timer T3412; the network can configure a cell-specific value for T3412 during the device attach procedure. The detailed description of each mode and the state transition triggers are summarized in Fig. 3.1.

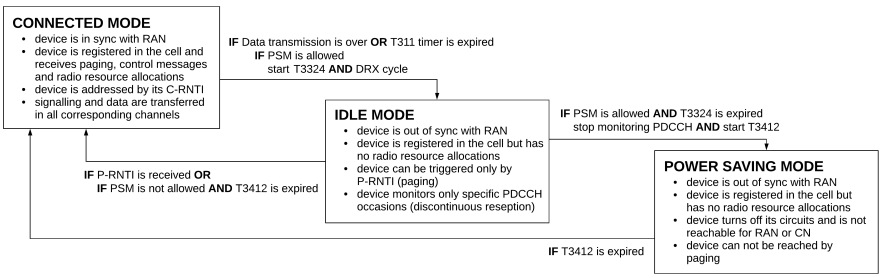



**Fig. 3.1.** Transition diagram between connected, idle, and deep sleep states.

The DRX mechanism tells how a device should monitor the downlink before switching off its receiving antenna. When a device detects the end of the data transmission, an *Inactivity Timer* starts, during which the device continuously listens to the PDCCH for a paging or scheduling indication. Note that LTE-M and NB-IoT use ad-hoc designed MPDCCH and

NPDCCH, respectively [15]. For the sake of brevity, in this chapter, we omit to specify the exact name of the different physical LTE-M and NB-IoT channels. If data arrives before the expiration of the Inactivity Timer, the timer resets. Otherwise, the *OnDuration* timer and DRX cycle starts. When the DRX cycle expires, the device continuously listens to the DCI during the OnDuration time.

Two parameters define the time instant when the device is available for paging, namely Paging Frame (PF) and Paging Opportunity (PO). For NB-IoT, the term Paging Narrowband (PNB) is used instead of PO to indicate the narrowband on which the device performs the paging message reception. The PF refers to a specific radio frame containing one or multiple POs. The PO defines the first subframe number within the PF from which a device has to listen to the paging indication. The number of POs per PF depends on the frame configuration, operating bandwidth and cell capacity. The position of POs within the PF is defined in [4]. For bandwidth reduced IoT devices, only subframe #1 and subframe #6 are available for POs. The network usually broadcasts a cell-specific default DRX that all devices can use. However, a device-specific configuration is also possible.

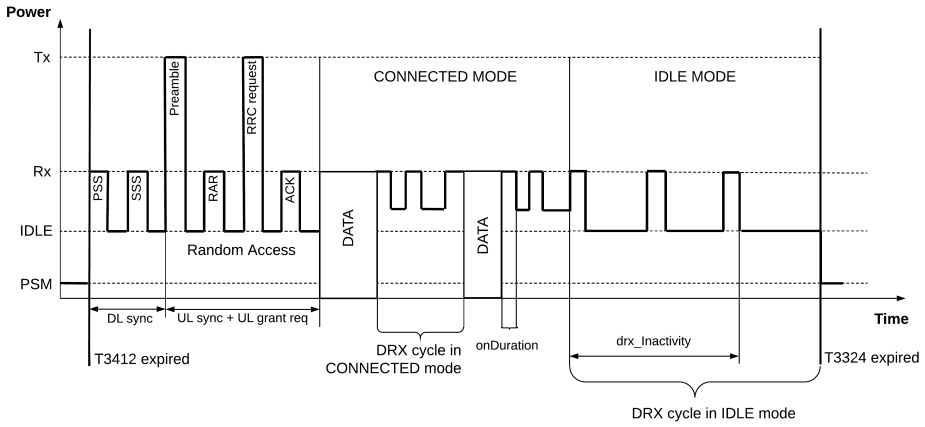



**Fig. 3.2.** Device energy consumption scheme in connected, idle, and PSM states.

When BS and the device agree on the DRX or *paging cycle* time T, they can independently calculate the PF from the equation: SFN mod $T = (T/N) \cdot$ (UE_ID mod N), where SFN takes values from 0 to 1023. The UE_ID can be derived from UE_ID = IMSI mod 1024, where International Mobile Subscriber Identity (IMSI) refers to a fixed unique device identifier, and N stands for the frequency of PF, and takes values 1, 1/2, 1/4, 1/8, 1/16, or 1/32.

Note, if P-RNTI is monitored on NPDCCH then UE_ID = IMSI mod 4096, and if device supports paging on a non-anchor carrier and P-RNTI is monitored on MPDCCH or NPDCCH then UE_ID = IMSI mod 16384. For NB-IoT devices, N is defined as the minimum between T and possible values of the nB parameter included in system information messages. Nb is equal to 4T, 2T, T, T/2, T/4, T/8, T/16, or T/32.

A short DRX cycle can be configured for the connected state in addition to the default long DRX cycle. An example of long and short DRX parameters is given in Fig. 3.3. The short DRX cycle is the first cycle a device follows after the successful reception of the paging indication in DCI. The number of consecutive short DRX cycles before using a long DRX cycle is defined by the *shortDRXcycle* Timer. The subframe number where the long DRX cycle should start is indicated by the *Start Offset*.

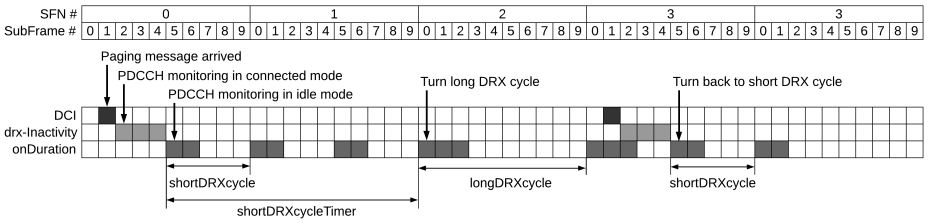



**Fig. 3.3.** Long and short DRX cycles in idle state.

### 3.2.2 Extended DRX

SFN and subframe number counters are used for time synchronization between the network and the devices, underpinning DRX mechanism. The longest time interval available for the synchronization is limited by the maximum number of SFN, which is equal to 1024. It means that devices can not remain in sleep mode for a time period longer than 10.24 seconds. The DRX cycle can be as long as 8, 16, 32, 64, 128, 256, and 512 ms. However, machine-type applications might require less frequent communication than every 10.24 seconds.

eDRX is a necessary step forward to address IoT use cases. To allow for a longer eDRX cycle, a new SFN counter, namely H-SFN, was introduced. One H-SFN contains 1024 SFNs, and builds an interval of 10,485,760 ms (almost 3 hours), while the system time is sufficiently extended to 1024 H-SFNs, or almost 413 days. With the new H-SFN feature, the maximum eDRX cycle value refers to 43.69 minutes. Since only one paging message is expected to be sent in a DRX cycle, any timing inaccuracy could cause the paging message loss. To avoid extra-long service latency and improve the paging reliability, every eDRX cycle can be configured with a Paging Transmission Window (PTW), during which a device can have more than one POs.

### 3.2.3 PSM

The PSM stands for a deep sleep state similar to the hibernation when almost all device circuits are switched off, except for the essential ones, e.g., clocks. PSM targets applications with relaxed latency requirements. Whenever a device wants to report its status change or send data to the network, it can leave the PSM. It is based on two timers: T3324 (Active

Time), and T3412 (extended TAU Time), which together frame the PSM cycle, as illustrated in Fig. 3.4. Any time a device leaves the PSM, timer T3412 resets.

Active Time refers to the time interval when a device is reachable for paging. The minimum recommended value for T3324 is 2 DRX cycles plus 10 seconds [24]. Since the DRX cycle value may vary, the relation between T3324 and T3412 should meet the condition (T3412 - T3324)/T3412 $>$ 0.9. T3412 is expected to be not less than 4 hours, but it can not exceed the maximum length of 413 days, as defined in [25].

When a device wants to enable the PSM feature, it has to request Active Time during the wake-up period in any RRC-related message. The network replies with an Active Time value indicating that the PSM is enabled. Upon T3324 expiration, a device enters PSM and becomes unreachable while T3412 is running, or until it requires to send any data to the network. An Active Time value can be either allocated by the RAN or CN, or be requested by a device. After receiving the request for the T3324 configuration, the network will choose an appropriate value taking into account whether the DRX or eDRX is currently enabled. If no device-originated transmission is expected within T3412, an IoT device remains in deep sleep mode for as long as the maximum value of the TAU timer.

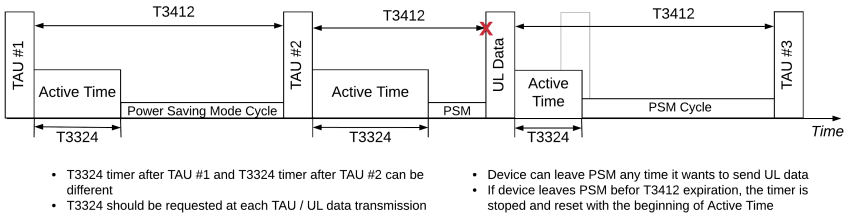



**Fig. 3.4.** Timers of the PSM.

## 3.3 MBMS Architecture and SC-PTM Mode

In cellular networks, a PTM transmission can be organized in two modes: over several adjacent cells, or on a per-cell basis. The former technology is called MBSFN, and it is based on the simultaneous transmission of the identical waveform by several highly synchronized BSs over the defined area. eMBMS transmission occupies the entire system bandwidth and up to 6 of 10 subframes of a radio frame. Multiplexing with unicast in the same subframe is not allowed, even though not all the radio resources in the frequency domain are utilized. SC-PTM was introduced in 3GPP release 13 to improve the resource allocation flexibility; it allows one cell to multicast the same content to a group of end-points multiplexing PTP and PTM data on the same PDSCH instead of using a dedicated physical channel for multicasting. Furthermore, it also benefits from a reduced end-to-end latency since the BSs synchronization stage is not needed.

SC-PTM is a supplementary radio bearer service that reuses the MBMS architecture and CN procedures, and partially reuses MBMS procedures in RAN. The SC-PTM support for LTE-M and NB-IoT was introduced in Release 14 to support narrowband operation for SC-PTM services [16]. The MBMS architecture, illustrated in Fig. 3.5, consists of three main components [23]:

- *Broadcast Multicast Service Center (BM-SC)*, usually located in the core network, is responsible for the service bearer activation/deactivation, subscription management, service announcement, and MBMS session transmission/re-transmission.
- *MBMS-Gateway (MBMS-GW)* forwards the MBMS-related data packets to all downstream nodes (e.g., BS) involved in the multicast/broadcast data delivery; it is in charge of session control management and RAN signalling.
- *Multicast Coordination Entity (MCE)* accomplishes admission control and radio resource allocation at the BS.

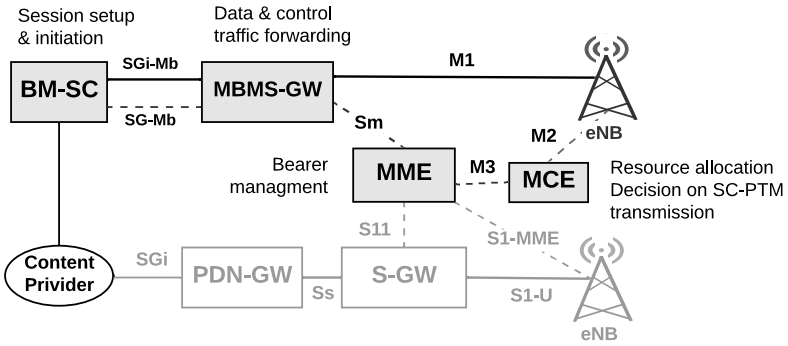



**Fig. 3.5.** MBMS architecture.

The MBMS is a subscription-based standard by design, which means that the users interested in an MBMS service first have to subscribe to it. The BM-SC informs subscribers about the available service and content through the *service announcement* procedure. Endpoints interested in the content reply to the MBMS-GW with a *joining request*. If at least one device *joins* the group, a multicast *session starts*. By forwarding a session start message, the MBMS bearer context is created in all involved downstream nodes, and specific control and data bearers are activated in the core and RAN segments, respectively. The context contains, among others, the description of the MBMS bearer service and the list of BSs that requested the service. When all corresponding bearers are setup and active, the *data transfer* begins. Once created, the bearer context remains in *active* mode until the *session stop* command changes its status to *standby*. A device may leave the group at any time by sending the *leave* request.

However, not all steps are necessary or even possible for MTC [6, 26]. The subscription step should be done by device owner or manufacturer as they are in charge of creating and

updating the multicast subscribers group. Also, legacy *join* and *leave* procedures are not applicable for IoT devices. Therefore, they have to be instructed by upper layers when to join or leave the group. Finally, carousel-like *service announcement* is not recommended for battery-constrained IoT devices and delivery of unplanned traffic [3].

In the SC-PTM framework, a new Single-Cell Multimedia Radio Bearer (SC-MBR) and two dedicated logical channels namely Single-Cell Multicast Control Channel (SC-MCCH) and Single-Cell Multicast Traffic Channel (SC-MTCH) have been introduced [16, 23]. The SC-PTM bearer service is identified by Group Radio Network Temporary Identifier (G-RNTI), while the content of both channels is carried by PDSCH and scheduled by PDCCH similar to unicast transmission as illustrated in Fig. 3.6.

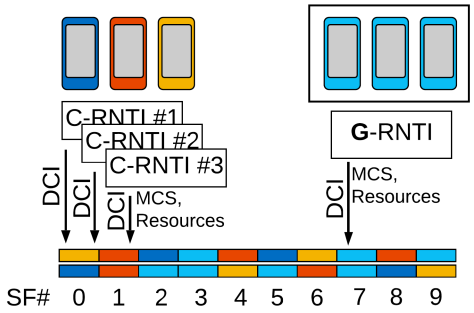



**Fig. 3.6.** Multiplexing of multicast and unicast transmissions in cellular IoT.

The Temporary Mobile Group Identity (TMGI) is allocated by a service provider to identify a multicast session in core and radio access network segments. This identifier, together with the multicast session scheduling information (i.e., scheduling period, scheduling window, and start offset), are provided in the *SC-PTM Configuration* message carried in the SC-MCCH. This information is broadcasted in the cell as a realization of the *service announcement*. The scheduling information for the SC-MCCH transmission is indicated in PDCCH by Single-Cell Radio Network Temporary Identifier (SC-RNTI), while the content of the SC-PTM configuration message is transmitted in the PDSCH. From the received message, users read the G-RNTI and multicast schedule. After that, they are ready for the group-based data reception over SC-MTCH.

To announce the delivery of *planned* group-based traffic a specific Single-Cell Discontinuous Reception (SC-DRX) cycle could be configured, which tells devices when SC-MCCH related transmission takes place. Similar to paging, a device can discontinuously listen to the SC-RNTI indication in PDCCH. Once it arrives, the device reads the scheduling information for SC-MTCH. A SC-MCCH opportunities defined by SC-DRX cannot overlap POs, therefore, SC-DRX is run on-top of regular paging forcing devices listen to both SC-MCCH and paging opportunities.

Such periodic monitoring of the SC-RNTI is an energy and resource-consuming approach for *unplanned* traffic delivery. In fact, devices always have to listen to the SC-MCCH channel even if there are no ongoing or scheduled group-based services.

Since SC-PTM services for LTE-M and NB-IoT are only supported in idle mode, the configuration of the SC-MCCH channel, including SC-RNTI, modification period, and offset, should be broadcasted to all devices in the cell. A new System Information Block Type 20 (SIB-20) is responsible for providing SC-MCCH configuration. Every time, when the SIB-20 changes, e.g., when a new multicast service is announced, the network should inform devices about changes in SIB-20. To do so, it uses System Information Block Type 1 (SIB-1) transmission with a specific flag indicating system information changes.

Information change notification applies the concept of *modification period*. It means that the system information content is not supposed to change within a modification period. During the first modification period, the BS informs devices that the SC-MCCH information is about to change, but the updated information itself is transmitted only in the next modification period. The minimum duration of one modification period is 640 ms [16]. The detailed example is illustrated in Fig. 3.7. The system change notification can be repeated several times to ensure that all devices get notified of the upcoming multicast transmission. Therefore, the latency of service announcements could be excessive when, e.g., the device owner intends to distribute unplanned critical updates or reconfiguration command among IoT devices. Such a delay motivates us to propose alternative solutions for the SC-PTM service announcement to reduce the overall latency of SC-PTM service delivery, as explained in section 3.4.

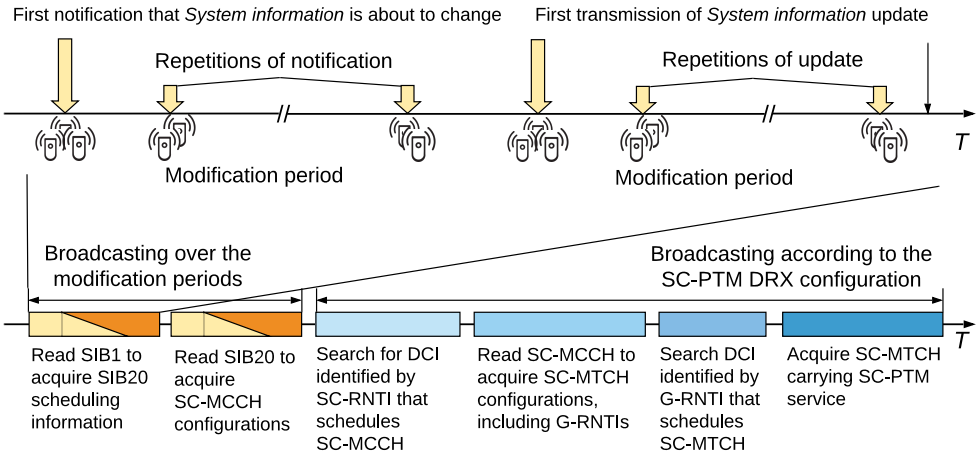



**Fig. 3.7.** Delay of the standard SC-PTM transmission.

## 3.4 Critical group-based communications

In conventional multicast scenarios, devices create a *multicast group* by subscribing to the content of interest and wait for the service announcement, which is usually broadcasted on a schedule. The service announcement stage usually runs for a long time to ensure that all group devices get ready for the content reception when multicast transmission starts.

The critical multicast transmissions can not be scheduled as described in the previous section because the content must be delivered to IoT devices with minimal delay. Since the arrival of critical content cannot be planned, and devices are not aware of its arrival, the network needs to send a paging message to notify them of an incoming transmission.

In relation to this challenge, we propose a multicast framework to improve critical PTM communications in cellular IoT to avoid long legacy service announcement procedure, and a new paging strategy that appropriately adjusts the paging interval and size of the paging group to improve PTM communication latency and device energy consumption.

### 3.4.1 A Framework for Critical Group-based Transmissions

We remind that each multicast session has a unique TMGI in core and radio access segments. Similar to paging, SC-PTM control and traffic transmissions are indicated by SC-RNTI and G-RNTI in DCI respectively. Once a device gets TMGI, G-RNTI and scheduling information for the SC-PTM transmission (i.e., scheduling period, scheduling window and start offset), it can receive the content.

When the content becomes available for download, IoT Application Server (AS) notifies BM-SC by allocating a TMGI. BM-SC triggers *service start* procedure to establish MBMS bearers in the core network providing session description information, as shown in Fig. 3.8. MBMS-GW forwards the message to the MCE, it reads operation mode field from the session start request to enable either SC-PTM or MBSFN modes.

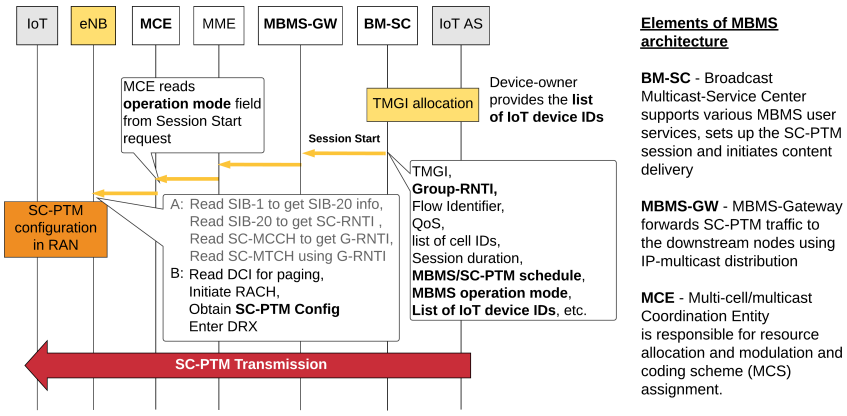
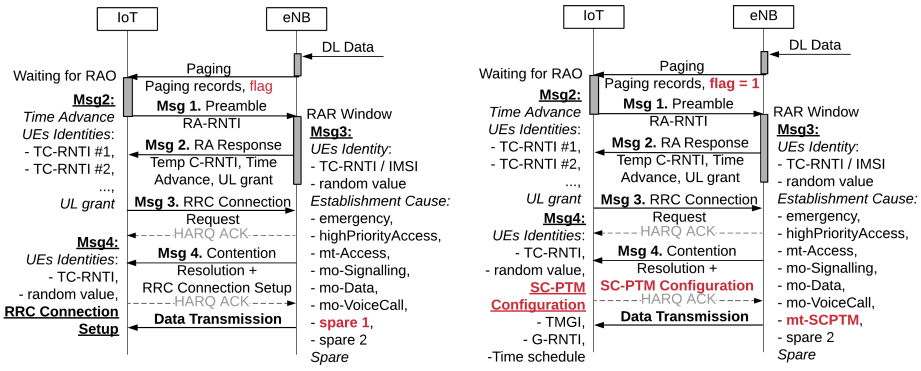



**Fig. 3.8.** Standard (Option A) and proposed (Option B) scheme to deliver SC-PTM traffic to IoT devices.

In the access network, BS provides the information about ongoing MBMS session, including TMGI, associated G-RNTI and scheduling information (i.e. scheduling period, scheduling window and start offset) in the *SC-PTM Configuration* message carried in SC-MCCH. Option A in Fig. 3.8 represent the sequence of steps to configure SC-PTM according to the 3GPP.

However, as discussed in section 3.3, the legacy way to announce a new transmission arrival does not fit delay-sensitive applications. SC-PTM configuration parameters could be provided to devices by means of *unicast transmissions*. Therefore, to avoid the significant configuration delay for SC-PTM reception due to the legacy service announcement, we propose to wake up devices with paging messages and piggyback SC-PTM configuration parameters into the Msg4 of the RA procedure replacing *RRC Connection Setup/Resume* message.

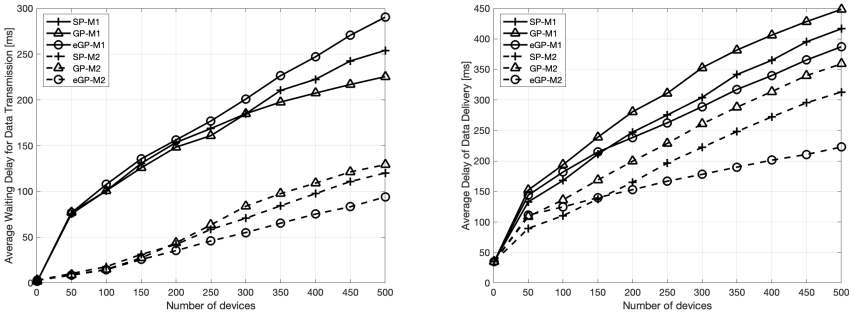
After receiving paging message a device initiates the RA procedure as illustrated in Fig. 3.9(a). It starts with sending a randomly chosen preamble (*Msg1*) over the PRACH scheduled at the specific Random Access Opportunity (RAO)s, defined by the PRACH configuration index. If the BS successfully decodes Msg1, then it replies with the RAR message (*Msg2*), containing the Temporary Cell Radio Network Temporary Identifier (TC-RNTI), the timing advance information, and an UL grant for the following glsrrc connection request (*Msg3*) transmission in the PUSCH. The device sends the Msg3 using the received grant and includes the received TC-RNTI and *Establishment Cause* into the *Msg3*. The Contention Resolution Time (CRT) window starts after Msg3 is sent. If two or more devices send the same preamble, they will receive the same UL grant and will collide during Msg3 transmission. If the BS decodes one of the colliding transmission, it replies with *Msg4* addressing the device identifiers from the decoded message. If both the returned and sent TC-RNTI match, the RA is successfully completed. Otherwise, the device will restart the procedure.




(a) 3GPP-based Random Access (b) Enhanced Random Access

**Fig. 3.9.** Enhanced and legacy RA procedure.

Fig. 3.9(b) illustrates the necessary modifications to the paging message and to the 3GPP compliant RA procedure to enable our solution. A *flag* in the paging message should be set




(a) Average waiting time (b) Average SC-PTM latency

**Fig. 3.10.** SC-PTM latency for M1 and M2 schemes.

to 1 to inform devices of the SC-PTM related paging. To emphasize that the SC-PTM configuration is requested, Msg3 is extended to let the device specify a new establishment cause in the corresponding spare field of Msg3 that we define as *mt-Multicast*.

Upon receiving SC-PTM configuration message, devices are ready for the multicast reception. However, due to the collisions, devices have to repeat the RA procedure until Msg4 is successfully received. Preamble re-transmissions can be triggered (i) due to the lack of DL resources for Msg2 transmission, or (ii) due to the collisions in Msg3. After a failed RA attempt, the device waits for the backoff and repeat the procedure. These re-transmissions are the main contributor to the RA delay and may cause device access failure.

Since devices experience different RA delays, they receive Msg4 at different times. Therefore, we need to define an appropriate scheduling scheme for the group-based transmissions. There are two general approaches to schedule multicast transmission: (M1) wait until all devices receive Message 4 to be fed by a single multicast transmission, or (M2) schedule identical transmissions over a fixed interval. In both cases, devices may need to wait for the beginning of the multicast transmission, but the waiting time is very different.

In [27], we have evaluated the performance of both approaches to understand which scheme better suits delay-sensitive scenarios. For paging strategies, we use the three reference solutions explained in section 3.4.2. Scheme M1 demonstrates a significant waiting time in Fig. 3.10(a), which in tern has a negatively impact on service latency, as depicted in Fig. 3.10(b). However, due to the non-optimized paging and interval between multicast transmissions, the collision rate of scheme M2 is relatively high as the overall delay, shown in Fig. 3.10(b), is much longer than the waiting time given in Fig. 3.10(a).

To optimize paging and multicast scheduling intervals, we propose to send paging messages to small subgroups of devices, and to schedule a multicast transmission in a short interval after paging, as illustrated in Fig. 3.11. Paging a large number of IoT devices may cause preamble re-transmissions, and may delay the RA completion. The fewer devices complete RA before the next schedule, the fewer devices join the multicast transmission. When the multicast subgroups are small, the radio spectrum is not efficiently utilized, and the total

SC-PTM service delay increases. Moreover, the next group of devices is paged only at the end of the RA stage of the previous group. The interval between two successive SC-PTM transmissions depends on the expected RA delay and SC-PTM transmission delay.

In section 3.4.3 we evaluate the performance of the proposed framework considering paging strategies available in the literature and our optimized paging solution named New enhanced Group Paging (NeGP). Section 3.4.2 presents related work on paging.

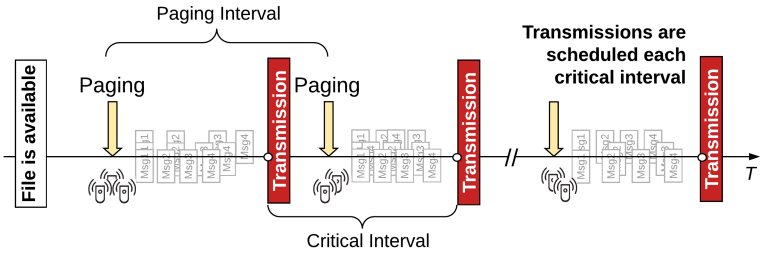



**Fig. 3.11.** Paging and Multiple-subgroups Multicast Transmissions.

### 3.4.2 Paging strategies

The trade-off between device availability and the energy consumption is well covered in the literature. For instance, the work in [28] discusses the impact of device active and sleep periods on the expected battery life cycle. In [29], device energy consumption for different active and sleep intervals and variable traffic rates have been analyzed, assuming unicast DL transmissions. The results demonstrate that both very short and very long intervals between paging indication and DL traffic arrival lead to an increase in device energy consumption. Similar results have been reported in [30] for more types of traffic and use cases. Device grouping is exploited in [31] to improve the energy consumption of IoT devices with similar UL traffic pattern and QoS. The grouping algorithm helps to avoid congestion in the UL when a huge number of devices try to access the network after receiving a paging indication. However, the mentioned works are mainly focused on the issue of *paging*, either to improve long-term device energy consumption with regular traffic. or to reduce device collision rate in the UL. In our framework, we address both paging and multicast traffic delivery aspects.

The solution for paging in [32] improves the device's battery life at the expense of a very long service delay that is unacceptable for critical applications. Authors in [33] obtained the optimal size for a paging group based on the limited capacity of the *RA* followed by paging. However, none of the mentioned works, except for [26], considers the impact of paging on multicast efficiency. For this reason, we propose a new paging solution that ldeparts from the general idea of the paging approaches discussed in section 3.2 (and proposed in [26] and [33]), but reinforces our SC-PTM transmission scheme for delay–critical IoT applications.

Paging is also used as a *pull-based* technique for RAN overload control to scatter massive MTC access attempts in time. Differently from *push-based* techniques that usually employ

access barring or re-transmissions with backoffs, the paging mechanism represents a proactive approach for handling access requests in a controlled manner. We describe three reference paging approaches used as benchmarks.

- **Standard Paging (SP).** In the legacy 3GPP paging, the number of unique device IDs, i.e., IMSI, that a paging message can carry depends on the size of each identity. The size of paging record usually varies between 25 and 61 bits, corresponding to maximum 16 for LTE-M and 8 for NB-IoT unique ID included into the paging records list [34]. The network will send several messages over the consequent POs until all devices are paged.
- **Group Paging (GP).** The capacity of legacy paging is very limited for massive IoT scenarios. The concept of Group ID (GID) was introduced to overcome this limitation. During cell registration, a group of devices can be assigned with an additional ID namely GID and be instructed to use a default IMSI for calculating their PO. Thus, all devices in the group can wake up at the same PO. Only one paging record is needed to address all members of the group. However, the collision probability at the access stage, if typically high as all devices reached by paging, will start their first RA attempt simultaneously.
- **enhanced Group Paging (eGP).** An adaptive approach was proposed in [26]. eGP is based on the idea of group-by-group paging where one message can reach more devices than in SP, but the interval between two consecutive paging transmissions is also increased to improve the collision probability. The interval should be long enough to ensure that the previous group of paged devices can complete RA.

### 3.4.3 System Model and Analysis

We consider a single-cell scenario with $N$ uniformly distributed devices. Let us define a *Virtual Frame (VF)* composed of $T_{VF}$ subframes as the time interval between two successive RAOs. The system time $T$ is slotted into $I = \lceil T/T_{VF} \rceil$ VFs, where $\mathcal{I} = \{1, \ldots, I\}$ denotes VF indexes. We assume that each VF has one PO and one RAO, as illustrated in Fig. 3.12.

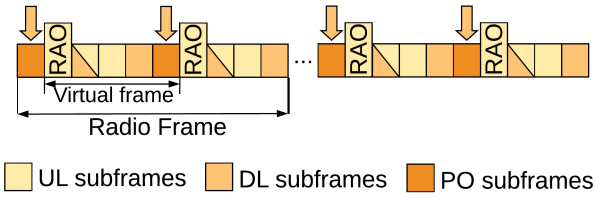



**Fig. 3.12.** System time model.

Let $Q$ denote the number of paging subgroups, $\mathcal{Q} = \{1, \ldots, Q\}$. If paging subgroup $q \in \mathcal{Q}$ has $n_q$ devices, then $n_1 + \ldots + n_Q = N$ and $n_q \leq N_j$, where $j \in \mathcal{J}$ denotes one of the paging schemes under consideration.

Let $\mathbf{P} = (\vec{P}_1, \ldots, \vec{P}_I)^T$ be the paging matrix composed of vectors $\vec{P}_i = (p_{i,q})_{i\in\mathcal{I},q\in\mathcal{Q}}$, whose element $p_{i,q}$ denotes the number of devices in the paging subgroup $q$ at the VF $i$. For a paging scheme $j \in \mathcal{J}$, we define $\mathcal{I}_j \subset \mathcal{I}$ as the subset of VF indexes in which paging messages should be sent. In particular, $\mathcal{J} = \{SP, GP, eGP, NeGP\}$. For the SP scheme, $\lceil N/N_{SP} \rceil$ VFs carry paging messages, where $N_{SP} = 16$ and paging interval is equal to one VF, therefore, $\mathcal{I}_{SP} = (1, 2, \ldots, \lceil N/N_{SP} \rceil)$. According to the GP scheme, all $N_{GP} = N$ devices can be reached by one paging message [35], so $\mathcal{I}_{GP}$ consists of only one element. The eGP scheme claims that a new paging group ($N_{eGP} = 36$) can be formed every $T_{eGP} = 30$ ms, i.e., every $i_{eGP} = \lceil T_{eGP}/T_{VF} \rceil$ VFs, thus $\mathcal{I}_{eGP} = \{1, 1 + i_{eGP}, \ldots, 1 + ([N/N_{eGP}] - 1)i_{eGP}\}$.

In our proposed NeGP, we define $\mathcal{I}_{NeGP}$ by taking into account the RA and SC-PTM transmission delays. Specifically, $F$ VFs are needed to complete the 4-message handshake for the RA when $N_{NeGP}$ devices contend at the preamble transmission stage. Then, let $W$ denote the number of VFs required for the SC-PTM transmission. Thus, a new group of devices can be paged every paging interval $T_{NeGP} = (F + W) \cdot T_{VF} = i_{NeGP} \cdot T_{VF}$ ms, and $\mathcal{I}_{NeGP} = \{1, 1 + i_{NeGP}, \ldots, 1 + ([N/N_{NeGP}] - 1)i_{NeGP}\}$. The optimal number of devices in a paging group is equal to the maximum number of devices that can be acknowledged in *Msg2* during the RAR window, i.e. $N_{NeGP} = N_{RAR}$. By considering the RA control overhead of $\sigma = 30\%$, and the RAR message format [16], the maximum number of devices that can be acknowledged during the RAR window is computed as $N_{RAR} = [(1 - \sigma)D_0]\lceil T_{RAR}/T_{VF} \rceil$, where $D_0$ is the number of Resource Block (RB)s available for the DL transmission in a VF, and $T_{RAR}$ is the RAR window duration. For a given system configuration, $D_0 = 12$ and $T_{RAR} = T_{VF}$, which yields $N_{NeGP} = 8$.

An IoT device that receives a paging message in VF $i$ initiates the RA at the same VF. If the first RA attempt fails, the device may take up to $R$ attempts, $\mathcal{R} = \{1, \ldots, R + 1\}$. Let vector $\vec{\alpha}_{i,r}$ denote the number of devices having the RA attempt $r$ in VF $i$, where $i \in \mathcal{I}$, $r \in \mathcal{R}$.

When devices make the first RA attempt, i.e. $r = 1$,

$$\vec{\alpha}_{i,1} = \vec{P}_i, i \in \mathcal{I}. \tag{3.1}$$

The total number $\alpha_i$ of devices having Msg1 transmission in VF $i$ can be obtained as follows:

$$\alpha_i = \left( \sum_{r=1}^{R} (\vec{\alpha}_{i,r}) \right) \cdot \mathbf{1}, i \in \mathcal{I}, \tag{3.2}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T$, $|\mathbf{1}| = Q$.

The random access to $C$ preambles by $\alpha_i$ devices is an instance of the occupancy problem. The probability to pick a preamble by a device from $C$ available preambles is equal to $1/C$. If $\alpha_i$ devices contend at VF $i$, the probability $q_i(c)$ of using exactly $c$ out of $C$ preambles by at least one device can be given as in [36]:

$$q_i(c) = \binom{C}{C-c} \sum_{j=0}^{c} (-1)^j \binom{c}{j} \left(1 - \frac{C - c + j}{C}\right)^{\alpha_i}. \tag{3.3}$$

The expected number of used preambles $C_i$ in VF $i$, $i \in \mathcal{I}$, can be calculated as follows:

$$C_i = \left[ \sum_{c=1}^{C_i^*} c q_i(c) \Big/ \sum_{c=1}^{C_i^*} q_i(c) \right] \tag{3.4}$$

where $C_i^* = \min(C, \alpha_i)$. We normalize $\sum_{c=1}^{C_i^*} c q_i(c)$ because the sum of probabilities $q_i(c)$ for $c = \{1, \ldots, C_i^*\}$ does not hold 1 when the number of contending devices $\alpha_i$ is less than $C$. The probability $p_i$ of choosing a unique preamble in VF $i$ depends on the number of contending devices $\alpha_i$:

$$p_i = \left(1 - \frac{1}{C}\right)^{\alpha_i - 1}, i \in \mathcal{I}. \tag{3.5}$$

Collided devices, which have received the same UL grant in Msg2, collide again in Msg3 transmission, and they can repeat the RA attempt after the CRT window expiration. We denote $M = \lceil T_{CRT}/T_{VF} \rceil$ as the CRT window $T_{CRT}$ in number of VFs.

The expected number of contending devices in VF $i$ is the total number of devices that make the first RA attempt after paging, devices that failed to receive Msg2, and devices that collided at step 3 of the RA procedure. Let $\vec{\alpha}_{i,r}^*$ denote the number of devices that successfully received Msg2 in VF $i$ after $r$ RA attempts. Vectors $\vec{\beta}_{i,r}$ and $\vec{\beta}_{i,r}^*$ stand for the number of devices scheduled for the Msg3 transmission in VF $i$, and for the number of devices that successfully sent Msg3 in VF $i$ after $r$ RA attempts, respectively. Finally, let $\vec{\gamma}_{i,r,m}$ denote the number of devices that receive Msg4 in VF $i$ after $m$ VFs of the contention resolution time and $r$ RA attempts, while $\vec{\gamma}_{i,r,m}^*$ stands for number of devices that successfully received Msg4 in VF $i$.

Devices that failed the RA attempt retry after the Backoff Window (BW) $T_{BW}$ or $j$ VFs, $j = \overline{1, B}$, where $B = \lceil T_{BW}/T_{VF} \rceil$. Let $\varphi_j = 1/B$ be the probability of randomly choosing the backoff time. The expected number of devices contending in VF $i$ yields:

$$\begin{aligned} \vec{\alpha}_{i,r} &= H[i-1]\left(\vec{\gamma}_{i-1,r-1,M} - \vec{\gamma}_{i-1,r-1,M}^*\right) + H[i-k-M] p_{i-k-M} \cdot \vec{\beta}_{i-M,r-1} + \\ &+ \sum_{j=1}^{B} H[i-j-1]\left(\vec{\alpha}_{i-j-1,r-1} - \vec{\alpha}_{i-j-1,r-1}^*\right)\varphi_j, i \in \mathcal{I}, r \in \mathcal{R}, j = \overline{1, B} \end{aligned} \tag{3.6}$$

where $H[x]$ is a Heaviside function; it equals to 1 if $x > 0$, and takes 0 if $x \leq 0$.

The BS needs $T_{RA}$ ms to detect and decode transmitted preambles before sending Msg2. Thus, a device waits for $k = \lceil ((A-1)T_{VF} + T_{RA})/T_{VF}A \rceil$ VFs for the Msg2 reception. Let $N_{RAR}$ denote the system capacity for Msg2 transmissions in numbers of preambles that can be acknowledged by the BS. If devices contending in VF $(i-k)$ used less than $N_{RAR}$ preambles, then all devices receive Msg2. Otherwise, only a portion of them receives Msg2, that is given as follows:

$$\vec{\alpha}_{i,r}^* = \begin{cases} \vec{\alpha}_{i-k,r}, & C_{i-k} \leq N_{RAR} \\ \left[\vec{\alpha}_{i-k,r} N_{RAR}/C_{i-k}\right], & C_{i-k} > N_{RAR}. \end{cases} \tag{3.7}$$

The expected number of devices to be scheduled for the Msg3 transmission in VF $i$ can be given as follows:

$$\vec{\beta}_{i,r} = \vec{\alpha}^{*}_{i-1,r} + \left(\vec{\beta}_{i-1,r} - \vec{\beta}^{*}_{i-1,r}\right), \tag{3.8}$$

where $\left(\vec{\beta}_{i-1,r} - \vec{\beta}^{*}_{i-1,r}\right)$ counts for the devices that failed to send Msg3 in VF $i-1$ due to the lack of UL resources.

Let $U_0$ be the total number of UL resources available in VF $i$. Since the PRACH occupies a fixed number $U_P$ of RBs in the UL, the number of available UL resources in VF $i$ for Msg3 transmission equals to $U_i = U_0 - U_P$. The expected number of devices scheduled for the Msg3 transmission in VF $i$ can be given as follows:

$$\vec{\beta}^{*}_{i,r} = \begin{cases} \vec{\beta}_{i,r}, & \vec{\beta}_{i,r}\mathbf{u}^T \leq U_i \\ \left[\vec{\beta}_{i,r}U_i / \vec{\beta}_{i,r}\mathbf{u}^T\right], & otherwise, \end{cases} \tag{3.9}$$

where $\mathbf{u}^T$, $|\mathbf{u}| = Q$, denotes the average number of RBs required for the Msg3 transmission.

The expected number of devices to be scheduled for the Msg4 transmission in VF $i$ is either the number of devices that successfully sent Msg3 in the previous VF, or the number of devices that failed to receive Msg4 in the previous VF due to the lack of the DL resources:

$$\vec{\gamma}_{i,r,m} = \begin{cases} \vec{\beta}^{*}_{i-1,r}, & m = 1 \\ \vec{\gamma}_{i-1,r,m-1} - \vec{\gamma}^{*}_{i-1,r,m-1}, & otherwise. \end{cases} \tag{3.10}$$

where $i \in \mathcal{I}$, $r \in \mathcal{R} \setminus \{R+1\}$.

Let $D_0$ and $D_{RAR}$ be the total number of DL resources available in VF $i$ and the average number of resources required for the Msg2 transmission, respectively. The number of DL resources $D_i$ after the Msg2 transmission can be calculated as:

$$D_i = \begin{cases} D_0 - D_{RAR}, & \left(\sum_{r=1}^{R} \vec{\beta}_{i,r}\right)\mathbf{1}^T > 0 \\ D_0, & otherwise. \end{cases} \tag{3.11}$$

Therefore, the expected number of devices that successfully sent Msg4 in VF $i$ yields:

$$\vec{\gamma}^{*}_{i,r,m} = \begin{cases} \vec{\gamma}_{i,r,m}, & \vec{\gamma}_{i,r,m}\mathbf{d}^T \leq D_i \\ \left[\vec{\gamma}_{i,r,m}D_i / \vec{\gamma}_{i,r,m}\mathbf{d}^T\right], & otherwise, \end{cases} \tag{3.12}$$

where $\mathbf{d}^T$ denotes the average number of DL resources required for the Msg4 transmission, $|\mathbf{d}| = Q$.

After receiving Msg4 in VF $i$, devices can receive SC-PTM transmission scheduled in one of the next VFs. We assume that up to $S$ multicast transmissions can be scheduled within $I$ VFs, $\mathcal{S} = \{1, \ldots, S\}$. Let $i_s$ be the first VF of the SC-PTM transmission $s$. Then, the expected number $\vec{\delta_s}$ of devices ready for the SC-PTM transmission $s$ yields:

$$\vec{\delta}_s = \sum_{k=i_{s-1}}^{i_s-1} \sum_{r=1}^{R} \sum_{m=1}^{M} \vec{\gamma}^{*}_{k,r,m}, s \in \mathcal{S}. \tag{3.13}$$

Let $z$ define the critical interval between two successive SC-PTM transmissions. The first transmission should be scheduled with an offset to ensure that all devices of the first paging subgroup receive Msg4, while all next multicast transmissions are scheduled in $z$ VFs.

Let $\Theta$ be the multicast payload in terms of resources needed for the SC-PTM transmission. The residual number of resources $\theta_{l_s}$ required to complete transmission $s$ after the first $l_s - 1$ VFs is given as follows:

$$\theta_{l_s} = \begin{cases} \Theta, & l_s = 0 \\ \theta_{l_s-1} - D_{i_s^*+l_s}, & \theta_{l_s-1} > D_{i_s^*+l_s} \\ 0, & otherwise. \end{cases} \tag{3.14}$$

Let $l_s^*$ stand for the last VF of the SC-PTM transmission $s$ such that $\theta_{l_s^*} = 0$, i.e., denotes the duration of the SC-PTM transmission $s$. The expected number of devices $\vec{\delta}_s^*$ that successfully receive the multicast service after $l_s^*$ VFs equals to $\vec{\delta}_s$. We calculate the metrics of interest.

*Access success probability* $P_A$ is a ratio of the number of devices that completed the RA stage to the overall number of devices reached through paging

$$P_A = 1 - \left(\sum_{i=1}^{I} \vec{\alpha}_{i,R+1}\right) \mathbf{1}^T / \left(\sum_{i=1}^{I} \vec{\alpha}_{i,1}\right) \mathbf{1}^T. \tag{3.15}$$

*Average access delay* $D_A$ corresponds to the time to complete the RA:

$$D_A = \frac{1}{Q} \sum_{q=1}^{Q} \left(i_q^* - i_q\right) T_{VF}, \tag{3.16}$$

where $i_q$ stands for the VF at which group $q$ receives paging, and $i_q^*$ is given as follows

$$i_q^* = \left[ \left( \sum_{i=1}^{I} i \sum_{r=1}^{R} \sum_{m=1}^{M} \vec{\gamma}_{i,r,m}^* \right) \mathbf{e}_q^T \Big/ \left( \sum_{i=1}^{I} \vec{\alpha}_{i,1} \right) \mathbf{e}_q^T \right]. \tag{3.17}$$

*Average idle delay* $D_{Idle}$ is the time that elapses from the end of the RA stage until the beginning of the multicast transmission, therefore,

$$D_{Idle} = \frac{1}{Q} \sum_{q=1}^{Q} \left(i_q^{**} - i_q^*\right) T_{VF}. \tag{3.18}$$

where $i_q^{**}$ is given as follows

$$i_q^{**} = \left[ \sum_{s \in \mathcal{S}} i_s^* \left( \vec{\delta}_s \mathbf{e}_q^T \right) \Big/ \left( \sum_{s \in \mathcal{S}} \vec{\delta}_s \right) \mathbf{e}_q^T \right] - 1 \tag{3.19}$$

because not all devices of the same paging subgroup will be members of the same multicast subgroup for the SC-PTM reception.

*Average total delay* $D_{Total}$ includes the average access delay $D_A$, average idle delay $D_{Idle}$, and average SC-PTM transmission delay $D_{TX}$:

$$D_{Total} = D_A + D_{Idle} + D_{TX}, \tag{3.20}$$

where the average SC-PTM transmission delay can be computed as

$$D_{TX} = \frac{1}{S} \sum_{s=1}^{S} l_s^* \cdot T_{VF}. \tag{3.21}$$

*Total service delay* $D_{Service}$ is the total time to wake up all relevant devices and deliver the content of interest. Having $i_{S^*}$ and $l_{S^*}$ of the very last multicast transmission $S^*$, we compute the metric as follows

$$D_{Service} = (i_{S^*} + l_{S^*})T_{VF}. \tag{3.22}$$

*Average access energy consumption* $E_A$ can be given as an arithmetic mean of the average energy consumption per paging subgroup $E_{A_q}$:

$$E_A = \frac{1}{Q}\sum_{q=1}^{Q} E_{A_q}. \tag{3.23}$$

Let $t_1$, $t_2$, $t_3$ and $t_4$ be the average transmission delay of Msg1, Msg2, Msg3 and Msg4. The device energy consumption in transmission mode equals to $e_{TX}$ mW, in reception mode - $e_{RX}$ mW, devices in idle mode consume $e_{Idle}$ mW on average. In the access stage, devices of subgroup $q$ consume:

$$\begin{aligned} E_{A_q} =& (e_{TX}t_1 + e_{RX}t_2)r_q^2 + (e_{TX}t_1 + e_{RX}t_2 + e_{TX}t_3)(r_q^3 + 1) + \\ &+ e_{Idle}T_{BW}r_q^2 + e_{RX}t_4, \end{aligned} \tag{3.24}$$

where $r_q^2$ and $r_q^3$ denote the average number of re-transmission attempts due to failure after Msg2 and Msg3 transmission, respectively. The average number of RA attempts due to Msg2 or Msg3 failure is computed as the weighted mean:

$$r_q^2 = \frac{\left(\sum_{r=1}^{R} r \sum_{i=1}^{I} \left(\vec{\alpha}_{i,r} - \vec{\alpha}^*_{i,r}\right)\right)\mathbf{e}_q^T}{\left(\sum_{r=1}^{R}\sum_{i=1}^{I} \left(\vec{\alpha}_{i,r} - \vec{\alpha}^*_{i,r}\right)\right)\mathbf{e}_q^T}. \tag{3.25}$$

$$r_q^3 = \frac{\left(\sum_{r=1}^{R} r \sum_{i=1}^{I} \vec{\alpha}^s_{i,r}\left(1 - p_i\right)\right)\mathbf{e}_q^T}{\left(\sum_{r=1}^{R}\sum_{i=1}^{I} \vec{\alpha}^s_{i,r}\left(1 - p_i\right)\right)\mathbf{e}_q^T}. \tag{3.26}$$

*Average device energy consumption* is the total energy consumed during the access, idle and SC-PTM transmission stages by a device on average:

$$E_{Total} = \left(E_A + e_{Idle}D_{Idle} + e_{TX}D_{TX}\right). \tag{3.27}$$

*Resource utilization* $R_{UL}$ and $R_{DL}$ is the ratio between the number of occupied resources and the total number of available resources in $I$ VFs in the UL and DL, respectively:

$$R_{UL} = 1 - \frac{\sum_{i=1}^{I} U_i}{IU_0}, \tag{3.28}$$

$$R_{DL} = 1 - \frac{\sum_{i=1}^{I} D_i}{ID_0}. \tag{3.29}$$

### 3.4.4 Performance evaluation

We consider a symmetric radio frame configuration (with the same number of UL and DL subframes) with $A = 2$ RAOs, as shown in Fig. 3.12. The mentioned paging strategies have different number of devices per paging subgroup and different paging intervals. For the

**Table 3.1.** Notations of section 3.4.

| Notation | Definition | Value |
|---|---|---|
| $C$ | Number of available preambles | 54 |
| $R$ | Maximum number of preamble re-transmissions | 10 |
| $N_j$ | Paging group size, $j = \{SP, GP, eGP, NeGP\}$ | $\{16, N, 36, 8\}$ |
| $T_j$ | Paging interval, $j = \{SP, GP, eGP, NeGP\}$ | $\{5, 0, 30, 25\}$ ms |
| $A$ | Number of RA subframes in a radio frame | 2 |
| $d$ | Interval between two consecutive POs | 5 ms |
| $z$ | Critical interval | 25 ms |
| $T_{VF}$ | Virtual frame duration | 5 ms |
| $T_{RA}$ | Delay for the preamble detection and decoding | 5 ms |
| $T_{RAR}$ | RAR window | 5 ms |
| $T_{BW}$ | Backoff window | 20 ms |
| $T_{CRT}$ | Contention resolution time | 48 ms |
| $N_{RAR}$ | Number of devices that may receive RAR within $T_{RAR}$ | 8 |
| $U_0$ | Amount of resources available for the uplink transmission in each VF | 12 RBs |
| $U_P$ | Amount of resources occupied by PRACH in the UL | 12 RBs |
| $D_0$ | Amount of resources available for the downlink transmission in each VF | 12 RBs |
| $D_{RAR}$ | Amount of resources required for the RAR message transmission in DL VF $i$ | 6 RBs |
| $\mathbf{u}$ | Vector of the average number of resources for Msg3 transmission | (1,...,1) RBs |
| $\mathbf{d}$ | Vector of the average number of resources for Msg4 transmission | (1,...,1) RBs |
| $\Theta$ | Multicast traffic payload | $\{3, 12, 32\}$ RBs |
| $e_{Tx}$ | Average device power consumption in the transmit mode | 500 mW |
| $e_{Rx}$ | Average device power consumption in the receive mode | 80 mW |
| $e_{Idle}$ | Average device power consumption in the idle mode | 3 mW |

reader's convenience, we give definitions of the system model parameters and their corresponding values in Table 3.1.

The analytic results have been validated by simulations in MATLAB. Simulation parameters are set according to [15, 34], for radio interface, and to [37], for device energy consumption, as reported in Table 3.2. Data packets arriving in a burst of a given size are transmitted over a set of continuous subframes.

**Table 3.2.** System parameters of section 3.4

| Parameter | Value |
|---|---|
| Cell radius | 500 m |
| Carrier configuration | 1.4 MHz carrier bandwidth at 800 MHz |
| PHY numerology | TDD frame type 1, TTI 1 ms |
| RA capacity | 2 RAOs per radio frame |
| Resource allocation | PDSCH, PDCCH: 1 – 6 PRBs |
| | PUSCH, PUCCH: 1 – 6 PRB, |
| | PRACH: 6 PRBs |
| Device power class | 23 dBm |
| BS transmit power | 46 dBm |
| Power consumption | 500 mW (TX), 80 mW (RX), 3mW (Idle) |
| Traffic payload | $\{392, 1608, 4584\}$ bits |

In the following figures, analytical results are shown as solid lines with markers, and simulation results only as markers; an almost perfect match is observed. Results are plotted for a cluster of up to 500 devices camping on a single LTE-M narrowband. As explained in [38], the device arrival rate of 40.3 access attempts per second with a target outage probability below 1% corresponds to the LTE-M traffic capacity per narrowband equaled to $0,36 \cdot 10^6$ devices/km$^2$, the higher capacity of $10^6$ devices/km$^2$ can be achieved if three or more narrowbands are configured in a cell. Our NeGP paging solution allows 320 device arrivals per second with outage probability less than 1%, which ensures more than $10^6$ device/km$^2$ of supported connection density.

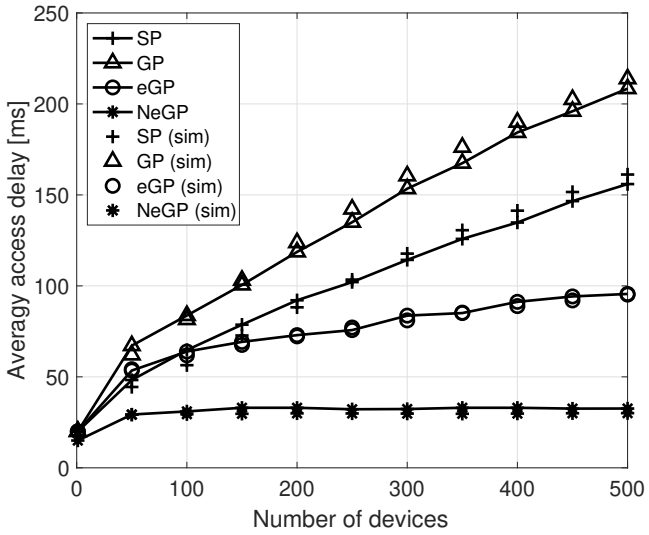



**Fig. 3.13.** Average access delay.

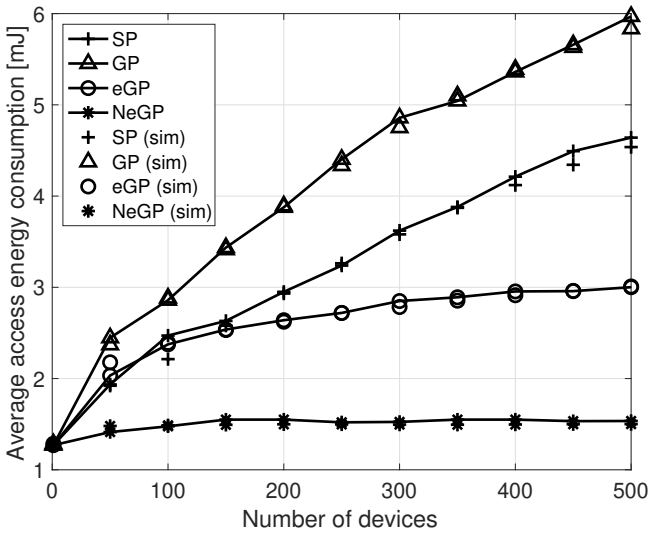



**Fig. 3.14.** Average access energy consumption.

Fig. 3.13 and Fig. 3.14 illustrates the average access delay and average device energy consumption for different paging strategies, respectively. The GP scheme introduces a significant delay and energy usage at the RA stage with respect to other schemes due to the high number of contending devices. For the SP and GP schemes both metrics grow almost linearly when the number of devices increases due to the preamble collisions and lack of radio resources. On the contrary, both metrics tend to saturate in the cases of the eGP and NeGP schemes. The eGP solution exploits the code-expanded preamble transmission technique that decreases collision rate and, consequently, the number of preamble re-transmission attempts [26]. However, our NeGP solution shows more than 50% reduction of both the average access delay and the average device energy consumption compared to the eGP scheme. The reason behind such performance gain is that the size of the paging groups and paging intervals in NeGP are optimized in such a way that devices complete the RA without any additional delay caused by preamble collisions or shortage of the radio resources.

Devices that complete the RA procedure remain in idle mode while waiting for the SC-PTM transmission, but keep listening to the DL from the last transmission until the end of the Inactivity timer defined by the DRX. If the timer expires before the SC-PTM transmission, devices switch off their receiving antenna and become unavailable until the next PO. Fig. 3.15 shows the average idle delay, i.e. the time to wait for the SC-PTM transmission after the reception of SC-PTM configuration parameters. The idle delay of the GP scheme grows fast under an increasing number of devices. In the case of SP and eGP, the metric increases mainly due to the short paging interval or high number of devices per group. To ensure that all paged devices receive the multicast transmission, the Inactivity timer should be higher than the idle delay. Fig. 3.16 illustrates the average device energy consumption under the assumption that the Inactivity Timer is set according to the experienced idle delay.

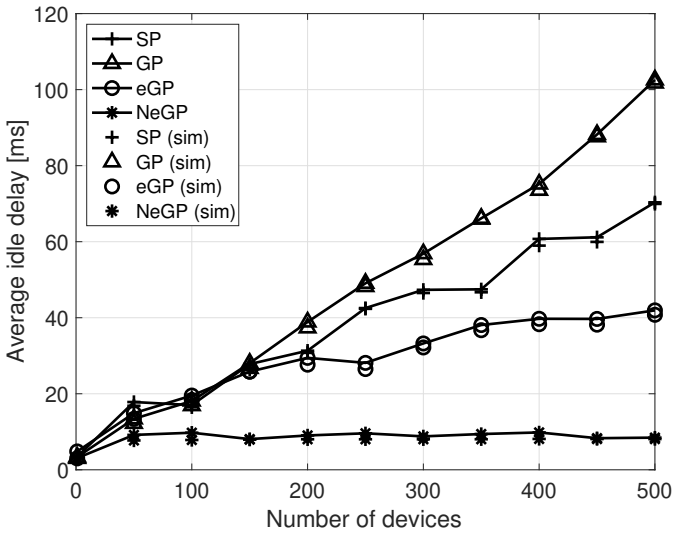



**Fig. 3.15.** Average idle delay.

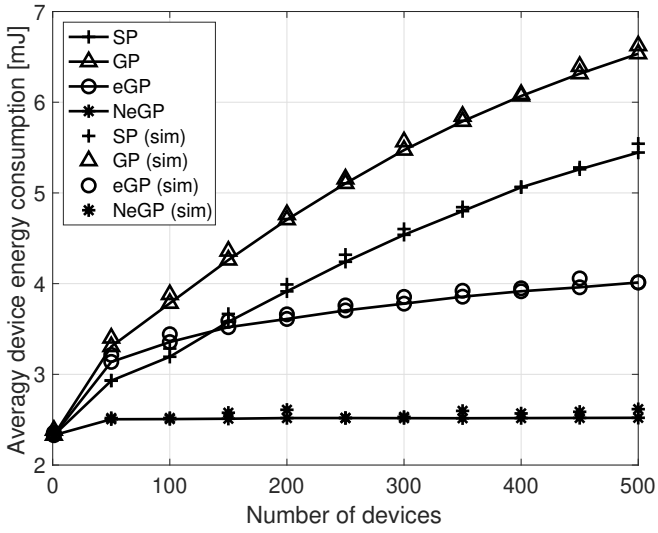



**Fig. 3.16.** Average device energy consumption.

The metric constantly grows under GP, SP and eGP strategies but it is almost constant for the NeGP scheme. This is an important result for battery-powered IoT devices.

Fig. 3.17–Fig. 3.19 show the average total delay for the variable SC-PTM payload. In particular, the size is set to 50, 200, and 500 bytes. For simplicity, we refer to these values as small (Fig. 3.17) medium (Fig. 3.18) and large (Fig. 3.19) payload, respectively. The total delay includes access delay, idle delay and the time to transmit an SC-PTM payload. The system performance is sensitive to the payload size because long multicast transmissions may overlap with the RA stage. Our NeGP paging and SC-PTM transmission design has been designed in order to avoid such an overlapping. As shown in Fig. 3.17–Fig.3.19, the increase

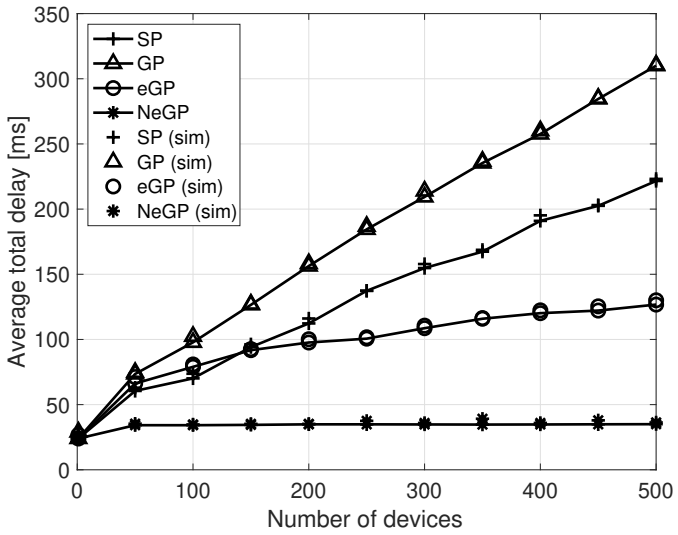



**Fig. 3.17.** Average total delay in case of small payload.

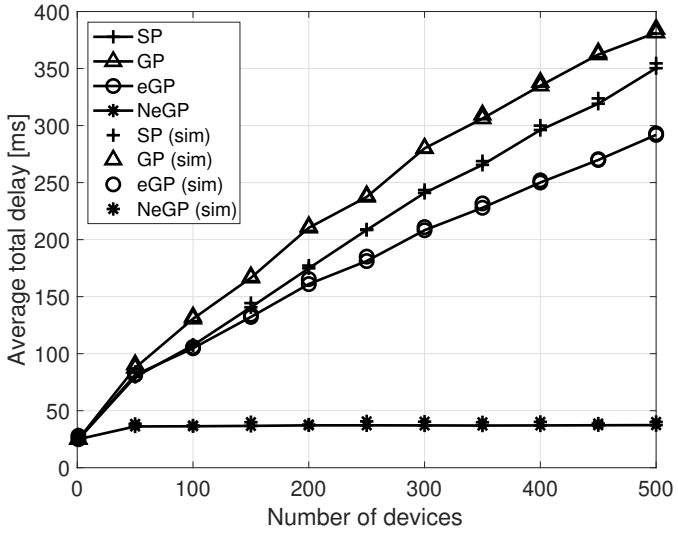



**Fig. 3.18.** Average total delay in case of medium payload.

of the SC-PTM payload does not lead to the significant performance degradation in the case of NeGP, and results only in an additional deterministic delay.

The access success probability is shown in Fig. 3.20. This metric also can be interpreted as the *service success probability* if necessary assumptions on the Inactivity Timer are made, as previously discussed. The failures are not only caused by preamble collisions, but also by re-transmissions after Msg2 and Msg3 failures. When the number of devices in the SP and GP schemes is increased, not all devices can successfully complete the RA. For a cluster of 500 devices, from 5% to 10% of devices fail the RA in the case of SP and GP strategies. Very

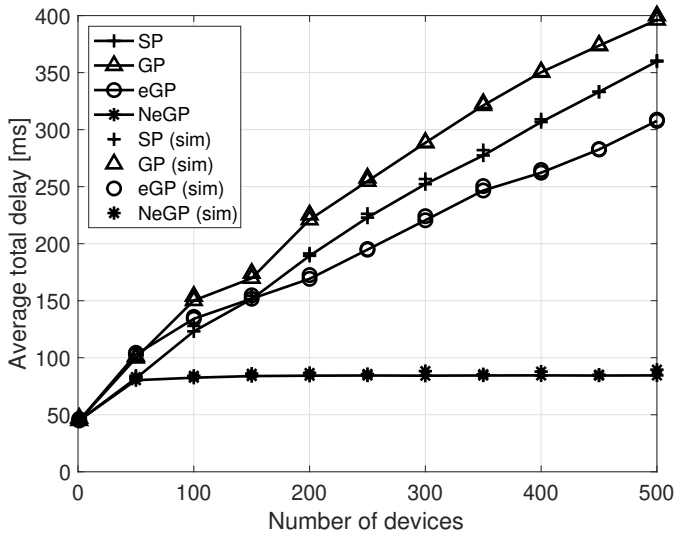



**Fig. 3.19.** Average total delay in case of large payload.

few devices lose the SC-PTM transmission if the eGP scheme is applied, while the NeGP guarantees the successful completion of the RA procedure by all devices.

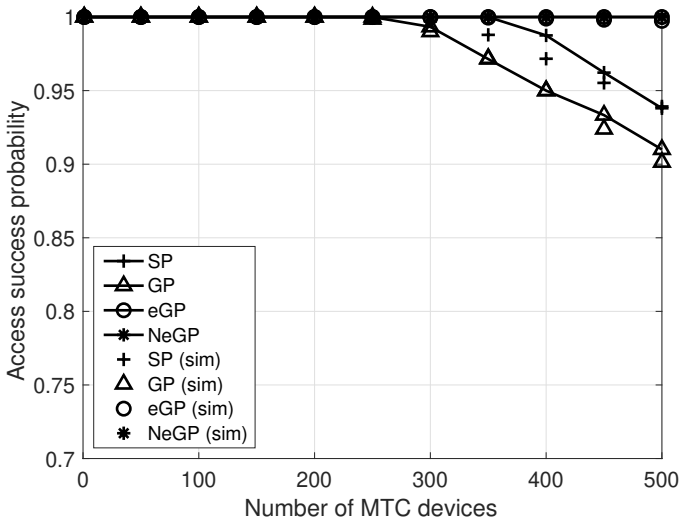



**Fig. 3.20.** Access success probability.

We compare the performance of our proposal with reference schemes in terms of radio resource consumption in the UL and DL for different payloads, as reported in Fig. 3.21 – Fig. 3.23. Regarding the UL utilization, the NeGP scheme requires less resources than SP, GP and eGP solutions, because it does not incur retransmissions of the RA messages. On the contrary, GP requires more UL resources than any other paging strategy due to the higher collision rate. Having more UL resources available is advantageous for the system

that can support other background traffic (e.g., from other IoT devices). The DL resource utilization depends on the number of multicast transmissions required to service all relevant devices. As expected, the NeGP solution requires more DL resources because it induces more SC-PTM transmissions. The difference in required DL resources becomes more evident when the payload size is larger, and more devices wait for the multicast service.

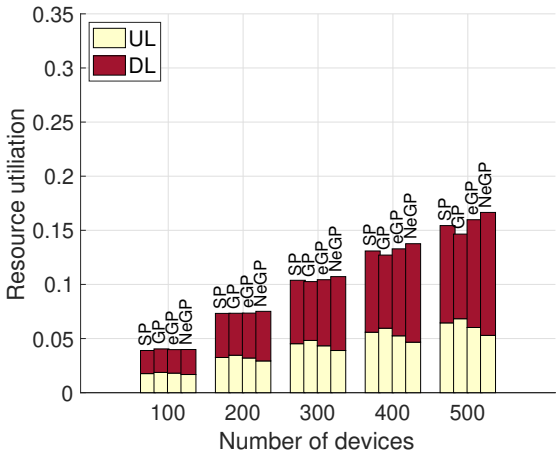



**Fig. 3.21.** UL and DL resources utilization in the case of small.

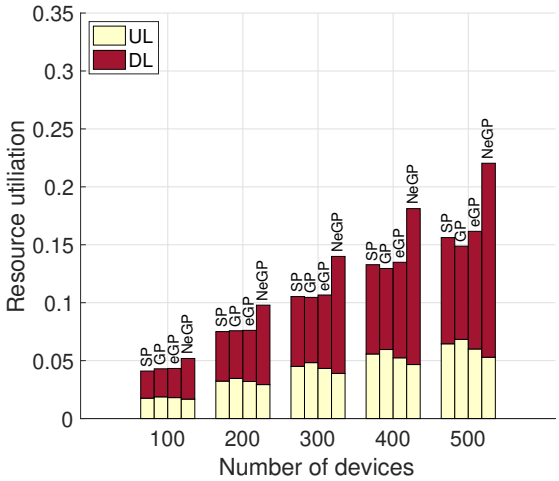



**Fig. 3.22.** UL and DL resources utilization in the case of medium payload.

## 3.5 Conclusions

This chapter covers essential aspects of group-based communications in cellular IoT systems, including paging and multicast transmission scheduling. We discussed different group-based

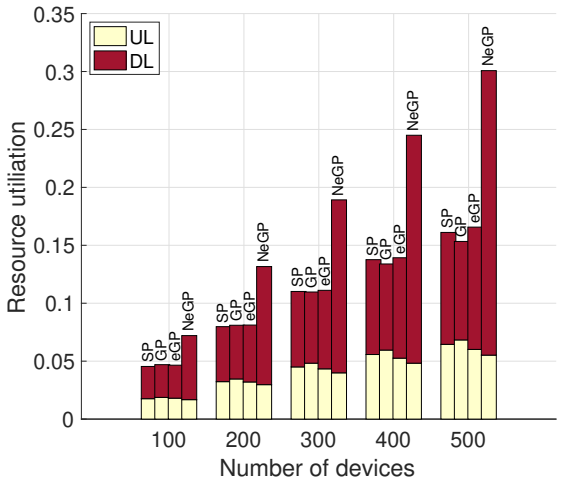



**Fig. 3.23.** UL and DL resources utilization in the case of large payload.

communication use cases, and highlighted the main challenges to support critical communications with tight delay requirements. A new framework based on the SC-PTM transmission mode and optimized group paging solution was presented. We proposed to schedule identical SC-PTM transmissions over an optimized interval to improve service latency, success probability, and reduce device energy consumption. We extensively evaluated the performance of our framework analytically and via simulations. We can conclude that paging significantly impacts the performance of critical SC-PTM communications when the arrival of multicast traffic can not be predicted. The optimal configuration of paging and SC-PTM scheduling guarantees that (i) all devices receive service, and (ii) the average service delay is insensitive to the total number of receivers.

In what follows, we present essential extensions concerning paging and, consequently, any network-originated transmissions.

### 3.5.1 Wake Up Radio

To improve the trade-off between network-originated service latency and device energy consumption, 3GPP has introduced a new WUS in 3GPP Release 15 [39]. Its design is based on one Zadoff-Chu sequence and scrambling code. WUS is a 1-bit indication sent to a device or a group of devices telling them whether to listen to the very next PO following the subframe where WUS is received.

With WUS support, devices can be configured with a shorter DRX cycle resulting in more opportunities for paging, but it can skip all “empty” PO. Indeed, since WUS is much shorter and requires less energy to detect and decode than the paging message needs, it improves device availability for paging with a slight energy consumption increase. Models and concepts of low-complexity wake-up receivers can be found in [40–42].

### 3.5.2 Secure paging

3GPP [43] has included privacy safeguards against IMSI catchers to ensure the secure paging for 5G systems. Privacy enhancement in the uplink is achieved by using a new concealed identifier Subscription Concealed Identifier (SUCI) instead of the IMSI analogue known as Subscription Permanent Identifier (SUPI). It is generated employing asymmetric cryptography every time when a device transmits data. As for downlink protection, the paging protocol has been enhanced by using new temporary identifiers 5G Serving Temporary Mobile Subscriber Identity (5G-S-TMSI) and Inactive Radio Network Temporal Identity (I-RNTI) instead of legacy long-term IMSI and temporary Serving Temporary Mobile Subscriber Identity (S-TMSI) identifiers [43].

Previously, paging timing was determined based on a permanent IMSI, but in 5G both PF and PO are calculated from temporal 5G-S-TMSI. This novelty makes it difficult for an over-the-air attacker to deduce information about a device's IMSI by monitoring the air interface and detecting which POs the device is monitoring. Furthermore, with native virtualisation and cloud-based RAN support, only a temporary identifier (5G-S-TMSI or I-RNTI) can be included into paging record list in 5G networks.

With the new security enhancements, if an attacker somehow obtains the device's long-term identifier in a 5G network, it still cannot attack the device because there is no long-term identifier based paging to start with. Moreover, to address a device, a new I-RNTI is used as a one-time identifier for paging. It must be refreshed after each paging. Unlike the S-TMSI that could be refreshed optionally, it is compulsory to refresh 5G-S-TMSI. By refreshing the 5G-S-TMSI, PO and PF are also changed, making it more difficult for an attacker to track a device during the paging [43].

Continuous changing a device's PO and PF improves protection from attackers; however, it makes the paging cycle dynamic and challenging to predict how many devices will be available at the same PO for a group paging.

# 4

# Resource Allocation for PTM Communications in Cellular IoT Networks

*This chapter addresses the problem of resource allocation and scheduling for the group-based communication. In particular, we propose two resource allocation schemes to balance multicast and unicast transmissions in resource-constrained NB-IoT systems. Additionally, we analyze the interplay between paging parameters and multicast transmission scheduling intervals.*

## 4.1 Introduction

Group-based communications in cellular IoT networks are gaining momentum as the number of connected devices keeps growing. Different sensors, video cameras, robots, and general-purpose IoT devices can be organized into groups for PTM transmissions to improve spectral efficiency and resource utilization in RAN. Thanks to the broadcast nature of the radio channel, a whole multicast group can be fed through a single transmission. However, group-based transmissions support is not trivial due to the heterogeneous requirements of different IoT applications in terms of a traffic pattern (e.g., periodic, unplanned, low data rate, heavy payload), latency (e.g., delay-tolerant, critical), and deployment density. It is getting even more challenging in bandwidth reduced system, like LTE-M and NB-IoT. Moreover, multicast transmissions are likely to overlap with unicast ones. Therefore, the background unicast traffic should be taken into account when allocating resources for group-based communications.

A comprehensive survey on multicast scheduling for OFDMA-based systems can be found in [44]. The most studied issue related to the PTM is a per-group transmission parameter optimization. The presence of cell-edge users forces the BS to use more robust MSCs to guarantee error-free reception to all receivers. Conventional Multicast Scheme [45] is the standard single-rate solution that conservatively selects the data rate for the multicast group based on the user with the worst channel quality [45]. Although this technique guarantees perfect fairness as resources are equally shared so that devices receive data with the same data rate, it suffers from low spectral efficiency. Alternative schemes have been proposed that opportunistically selects the best devices for multicast transmissions [46], or leverage users in good channel conditions to relay data [47] to other users with bad channel conditions. This solution

can not guarantee service to all multicast members. Another solution to enhance radio system utilization is clustering devices with similar channel quality for multicast reception [48]. With this approach, a better system capacity and session quality can be attained. Multi-rate solutions summarized in [49] exploit coding techniques to provide multiple versions of the content with various rates so that the subscribers could select the most appropriate rate according to their channel conditions.

In the area of multicast and unicast traffic management, different solutions have been presented. For instance, in [50], the resource allocation problem for multicast and unicast service is formulated as a maximization problem for the sum rate of unicast service under the fixed-rate guarantee for multicast service. A greedy-based algorithm is proposed to obtain an optimal subcarrier and power allocation solution. A scheduling scheme that guarantees both unicast and multicast users' minimum data rate is proposed in [51]. It dynamically assigns extra resources to either unicast or multicast service to improve the instantaneous rate. Differently, in [52], a resource allocation algorithm for mixed multicast and unicast traffic in MBSFN is proposed to maximize proportional fair utility. Multicast users with heterogeneous channel conditions are partitioned into groups so that the group members with poor signal strength do not impact group members' performance with good signal strength.

Most of the discussed scheduling solutions rely on devices CSI feedback that is not provided by cellular IoT devices due to the limited battery lifetime and significant signaling overhead. Some multi-rate solutions are not feasible in IoT due to the need for the resource-consuming decoding at the device side. Moreover, all approaches focus on the best allocation and scheduling solution at a given moment in time, while we would like to capture a long-term gain of the proposed resource allocation schemes taking into account the dynamic of multicast and unicast transmissions.

In section 4.2 we propose two resource allocation strategies to improve group-based communications in NB-IoT networks. We evaluate multicast and unicast communications performance in different application scenarios, e.g., short message transmission to a group of devices or communication of a heavy file under varying unicast traffic load.

As mentioned in chapter 3, paging is a critical component in PTM communications. Its parameters impact the performance of group-based service delivery as the number of available resources is limited. Moreover, parameters of a paging strategy, such as the number of devices available at the same PO, and the interval between two consecutive paging transmissions, are not always fixed. In some cases, the number of devices per one PO is random due to, e.g., security paging enhancement, device clock drift, and loss of system time synchronization with the network. For in-band NB-IoT deployment, some of the subframes can be invalid due to the MBSFN transmission in LTE. If these subframes overlap with devices PO, the next DL subframe after the invalid one will be used for paging. In section 4.3, we study the impact of random paging parameters on group-based performance transmission.

## 4.2 Resource Allocation for Multicast and Unicast Services

### 4.2.1 Motivation

The literature on NB-IoT is mainly focused on analyzing the uplink direction, while only a little attention has been paid to the downlink. In [5], the performance of firmware/software updates over SC-PTM in NB-IoT has been studied, showing that the transmission takes a rather long time and may last even days. Work in [6] addresses the issue of resource allocation for multicast transmission in the presence of background unicast traffic.

As discussed in chapter 3, multicast traffic can be scheduled in such a way that all devices get synchronized at the beginning of a group-based transmission or over shorter intervals to address only a subgroup of devices. The formation of multicast groups impacts the latency of group-based transmissions and, consequently, resource utilization. Multicast and unicast transmission arrival rates are a critical component to take into account.

As explained in section 2.4, NB-IoT can be deployed in three different modes including the most constrained one in terms of available radio resources - inside an LTE carrier. The scarcity of radio resources and wireless links' heterogeneity are the main challenges of scheduling multicast and unicast traffic in cellular IoT systems. We remind the structure of the NB-IoT DL frame depicted in Fig. 4.1. It occupies 10 subframes, each lasting 1 ms. The NPSS is transmitted in subframe #5 in every frame, while NSSS occupies subframe #9 and is repeated every 20 ms. The NPBCH is transmitted in subframe #0 in every frame to carry the essential system and cell-specific information. The NPDCCH carries scheduling information for both DL and UL data channels. It further carries the HARQ acknowledgment information for the UL data channel, as well as a paging indication and RAR scheduling information. Data from the upper layers, paging messages, system information, network response messages are transmitted over the NPDSCH. However, in one narrowband radio frame, not more than 8 subframes can be allocated for NPDSCH transmissions [53]. SC-PTM transmission uses the same NPDSCH so that multicast and unicast transmissions are multiplexed in time.

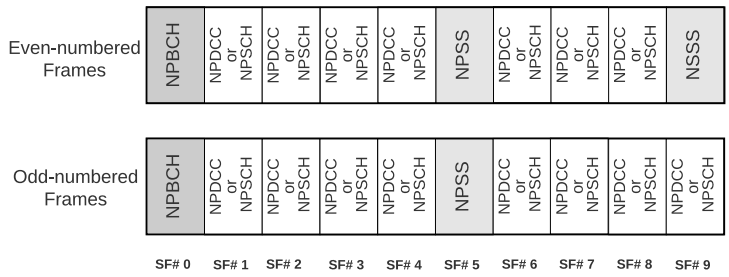



**Fig. 4.1.** NB-IoT DL frame structure.

Resource reservation scheme for SC-PTM transmissions in NB-IoT is a promising solution for delay-sensitive IoT applications. The latency of multicast services and device energy

consumption can be substantially improved if radio resources can be guaranteed to the PTM communication. However, in most IoT application use cases, unicast traffic has a priority over the group-based transmissions and usually spans over a low number of subframes since the network sends only a few bytes for the system or application acknowledgments. Moreover, unused resources due to the hard resource reservation for multicast transmission might be an issue.

We consider two strategies for scheduling group-based and unicast transmissions:

- **Multicast service with fixed guarantees**, where a portion of the resources is always guaranteed for group-based transmissions;
- **Non-Guaranteed multicast service**, where unicast and multicast transmissions share resources, and multicast transmission can be scheduled only if the required number of resources is available.

### 4.2.2 System Model

Let us consider a single-cell scenario with a BS in the center of the cell with radius $R = d_{IS}\sqrt{3}$ where $d_{IS}$ stands for the inter-site distance. We assume that devices are uniformly distributed in a cell.

Device density is calculated according to the typical urban scenario with 40 IoT devices per household, which results in 52,547 devices per cell. We assume that up to 70% of the users autonomously access the system to send reports with a 2-hour reporting interval [5]. Such a load corresponds to approximately $\lambda = 7$ arrivals per second. The remaining devices are involved in group-based communication, scheduled by the network each $\gamma^{-1}$ s on average. NB-IoT devices can join only one SC-PTM session at a time and, due to the limited system resources, we may assume no sessions in parallel. After waking up, devices wait for data transmission. We assume that a new multicast transmission starts after the previous one is terminated, and devices can join only a newly scheduled transmission.

Let $L_0 r^{\alpha}$ be the path loss, which a device experiences at the distance $r$ from the eNB, then the transmit power is $P_{TX} = P_0 L_0 r^{\alpha\beta}$, where $P_0$ is the open-loop transmit power, $L_0$ is the path loss at a reference distance of 1 m, $\alpha$ is the path-loss exponent, and $\beta \in [0, 1]$ is the power control factor. Without loss of generality, we assume that all devices belong to the normal coverage class. We refer to a MCL (dB) at the eNB to quantify the link budget needed for the target Signal-to-Noise Ratio (SNR). In particular, $MCL = P - S + G$, where $P$ (dBm) is the maximum power of the transmitter, $S$ (dBm) is the receiver sensitivity, and $G$ (dB) is the gain achieved by using coverage enhancement techniques, such as signal repetition or power spectral density boosting. For our case $P = 10 \cdot log_{10}(P_{max})$, and so the receiver sensitivity is calculated as $S = N_0 + P_{NF} + 10 \cdot log_{10}(W) + \sigma_k$ where $N_0$ (dBm/Hz) is the thermal noise density, $P_{NF}$ (dB) is the noise figure, $W$ (Hz) is the channel bandwidth, and $\sigma_k$ (dB) is the required SNR for NPDSCH at the MCL level $k$. The link performance for any physical channel is typically modelled using a set of SNR versus Block Error Rate

(BLER) curves for different MSC indexes [54]. These curves are then used to derive the highest MSC that can be used in a physical layer transmission, for a given SNR, such that the BLER is below the target value of 10 %.

For NB-IoT, $K = 14$ MSC levels are defined together with the different TBS at each level [55]. By knowing $\sigma_k$ for all $k \in \{0, 1, \ldots, K-1\}$, one can easily estimate the maximum distance $r_k$ to the eNB at which MSC $k$ is supported. We consider a fixed uniform deployment of IoT devices in a cell. The Probability Density Function (PDF) of the distance between a device and the eNB is given by $f(r) = 2r/R^2$. Thus, the probability of a device to use MSC $k$ is $\phi_k = (r_k^2 - r_{k-1}^2)/R^2$, which is the area between two circles of radii $r_k$ and $r_{k-1}$ divided by the total area of the cell.

We introduce parameter $\delta$ that partitions system resources between two types of DL traffic. In NB-IoT only 15 RBs are available for either unicast or multicast transmissions each 20 ms. Let $\delta$ be the portion of RBs allocated for the group-based transmission in two consecutive frames. If $B$ stands for the total number of RBs in two radio frames, the number of RBs available for the multicast transmission is $B_m = \delta B$. In the first scheme with guarantees, $B_m$ RBs are always reserved for multicast transmissions. While in the second scheme, $B_m$ RBs are not guaranteed, meaning that the multicast transmission can be blocked if unicast transmissions occupy more than $B - B_m$ RBs. The multicast transmission can be scheduled only when $B_m$ RBs are free.

### 4.2.3 Multicast Services with Fixed Guarantees

Devices interested in group-based content wake up according to a Poisson process with the rate $\epsilon$ and join a group for the subsequent group-based content reception. Once the group is formed, the transmission can start. A new transmission cannot begin if the previous one is not finished or if there are no devices in a group. We assume a deadline $W^*$ for a group formation that corresponds to an application's latency requirements.

The system of interest can be modeled as a single queue with batch service and finite waiting places equal to $M$. Device interarrival times $\xi_i = t_i - t_{i-1}$ follow an exponentially distributed Random Variable (r.v.) with parameter $\epsilon$, where $t_i, i > 0$ is the moment of a device activation. Without loss of generality, we assume that the system is empty at time $t_0 = 0$. Upon arrival, a device takes place in a buffer and waits until a batch is formed. Let $\tau_j$, $j > 0$, denotes the time instant when a batch can enter the server, i.e. the time instant when a multicast transmission is scheduled by the network. Intervals $\zeta_j = \tau_j - \tau_{j-1}$ between two consecutive transmissions are exponentially distributed with parameter $\gamma$. If at $\tau_j$ the server is busy or the buffer is empty, the batch will be formed after $\zeta_j + \zeta_{j+1}$. Let $\tilde{\tau_s}$ denotes the end of the batch service, the batch service times $\eta_s = \tilde{\tau_s} - \tau_j$ are also exponentially distributed r.v. with parameter $\nu$.

Let us calculate the multicast transmission rate that refers to the batch service time rate $\nu$ in our model. A device in a group supports the data rate $c_k$, $k = 1, \ldots, K$ with the probability

$\phi_k$. To ensure the reliability of multicast data transmissions the worst MSC should be chosen. Let $h(n,k)$ be the probability to choose MSC $k$ for a group of $n$ devices, therefore, the multicast data rate $C_m = \sum_{k=1}^{K} c_k \cdot \sum_{n=1}^{N} h(n,k)$. The probabilities $h(n,k)$ can be computed by a recursive formula $h(n,k) = \phi_k \sum_{i=1}^{k} h(n-1,i)$ with the initial condition $h(0,k) = \phi_k$. Since the multicast transmissions in our model carry the same payload $\theta_m$ (kbits), the batch service rate $\nu = C_m/\theta_m$.

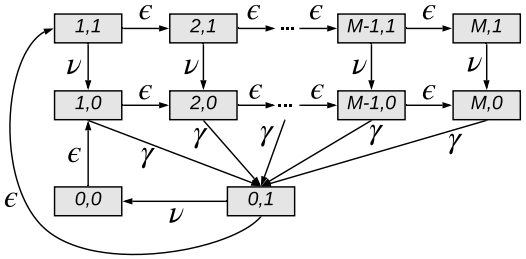



**Fig. 4.2.** The state transition diagram of the CTMC $\mathbf{X}(t)$.

Let $M(t)$ be the number of devices waiting in a buffer at time $t > 0$, while $S(t)$, $t > 0$, indicates if there is an ongoing multicast transmission. It takes value 1 if the server is busy, and equals 0 otherwise. We define a two-component CTMC $\mathbf{X}(t) = \{M(t), S(t), t > 0\}$ with the state space $\mathcal{X} = \{(m,s) : m = \overline{0,M}, s = \{0,1\}\}$. Let $p(m,s)_{(m,s)\in\mathcal{X}}$ denote the steady-state probabilities of $\mathbf{X}(t)$. The state transition diagram of the CTMC of interest is depicted in Fig. 4.2. The system of equilibrium equations based on the state transition diagram is given as follows:

$$
\begin{aligned}
\epsilon p(0,0) &= \nu p(0,1),\\
(\epsilon+\nu)p(0,1) &= \gamma \sum_{m=1}^{M} p(m,0),\\
(\epsilon+\nu)p(0,1) &= \gamma \left(p(1,0) + \cdots + p(M,0)\right),\\
(\epsilon+\nu)p(m,1) &= \epsilon p(m-1,1),\ m = \{1,\ldots,M-1\},\\
(\epsilon+\gamma)p(m,0) &= \epsilon p(m-1,0) + \nu p(m,1),\ m = \{1,\ldots,M-1\},\\
\nu p(M,1) &= \epsilon p(M-1,1),\\
\gamma p(M,0) &= \epsilon p(M-1,0) + \nu p(M,1).
\end{aligned}
\tag{4.1}
$$

The CTMC $\mathbf{X}(t)$ is not time-reversible; therefore, the expression for the steady-state probabilities can be given in a multiplicative form. However the probabilities $p(m,s)_{(m,s)\in\mathcal{X}}$ can be derived from the linear system equations (4.1) using linear transformations as follows:

$$
\begin{aligned}
p(m,1) &= p(0,0)\ \alpha \left(\frac{\alpha}{\alpha+1}\right)^{m},\ m=\overline{0,M-1},\\
p(M,1) &= p(0,0)\ \alpha^{2} \left(\frac{\alpha}{\alpha+1}\right)^{M-1},\\
p(m,0) &= p(0,0)\left[\left(\frac{\beta}{\beta+1}\right)^{m} + \sum_{i=1}^{m}\left(\frac{\alpha}{\alpha+1}\right)^{i}\left(\frac{\beta}{\beta+1}\right)^{m-i+1}\right],\\
& m=\overline{1,M-1},\\
p(M,0) &= p(0,0)\left[\beta\left(\frac{\beta}{\beta+1}\right)^{M-1} + \beta\sum_{i=1}^{M-1}\left(\frac{\alpha}{\alpha+1}\right)^{i}\left(\frac{\beta}{\beta+1}\right)^{M-i} + \right.\\
&\left. + \alpha\beta\left(\frac{\alpha}{\alpha+1}\right)^{M-1}\right],\\
p(0,0) &= \left[\sum_{m=1}^{M-1}\left[\left(\frac{\beta}{\beta+1}\right)^{M-1} + \sum_{i=1}^{M-1}\left(\frac{\alpha}{\alpha+1}\right)^{i}\left(\frac{\beta}{\beta+1}\right)^{m-i+1}\right] + \right.\\
&+ \beta\left(\frac{\beta}{\beta+1}\right)^{M-1} + \beta\sum_{i=1}^{M-1}\left(\frac{\alpha}{\alpha+1}\right)^{i}\left(\frac{\beta}{\beta+1}\right)^{M-i} + \\
&+ \alpha\beta\left(\frac{\alpha}{\alpha+1}\right)^{M-1} + \alpha\sum_{m=0}^{M-1}\left(\frac{\alpha}{\alpha+1}\right)^{m} + \\
&\left. + \alpha^{2}\left(\frac{\alpha}{\alpha+1}\right)^{M-1} + 1\right]^{-1},
\end{aligned}
\tag{4.2}
$$

where $\alpha = \epsilon/\nu$ and $\beta = \epsilon/\gamma$.

Having steady-state probabilities $p(m,s)_{(m,s)\in\mathcal{X}}$, we can compute performance metrics of interest. The average size of a multicast group is given as follows:

$$Q = \sum_{m=1}^{M} m\left(p(m,0) + p(m,1)\right). \tag{4.3}$$

The average waiting time for multicast transmission is:

$$W = \frac{1}{\epsilon(1-P_B)}\sum_{m=1}^{M} m\left(p(m,0) + p(m,1)\right). \tag{4.4}$$

Finally, the probability that the resources allocated for multicast services are idle gives:

$$P_{idle} = \sum_{m=1}^{M} p(m,0). \tag{4.5}$$

For this strategy, we can find such a multicast scheduling strategy $\gamma$ and resource allocation parameter $\delta$ that minimizes the probability $P_{idle}$, i.e. the portion of unused resources, for a given multicast payload $\theta_m$:

$$\min_{\gamma,\delta} P_{idle}(\gamma,\delta,\theta_m), \tag{4.6}$$

$$\text{subject to: } W(\gamma,\delta,\theta_m) \le W^{*}. \tag{4.7}$$

### 4.2.4 Non-Guaranteed Multicast Services

Recall that the unicast traffic arrives with rate $\lambda$. The multicast traffic arrival rate $\gamma$ and service rate $\nu$ are the same as in section 4.2.3. In the scheme with non-guaranteed multicast

service, all RBs are available for the unicast transmissions, while the number of resources occupied by the multicast transmission cannot exceed $\delta B$ RBs. The multicast transmission can be scheduled if there are enough resources. Otherwise, it will be blocked until the end of some unicast transmission when enough resources for multicast transmission will be released.

The system with non-guaranteed multicast service can be modeled as a tandem queueing system with a shared queue. The primary server serves unicast transmissions according to the processor sharing policy [56] with a finite number of parallel processing units $N$. Only one multicast session can be scheduled at once. The unicast transmission rate depends on the number of active transmissions and the presence of multicast traffic. All unicast devices equally share system resources. When a multicast transmission request arrives, the primary server hands ongoing transmissions over to the secondary server. Because of the multicast transmission running in parallel, less resources are available for unicast traffic. Therefore, the secondary server can service only $N^*$ devices at the same time.

Let $C$ (kbps) denote the system throughput. Only $C^* = (1-\delta)C$ (kbps) can be transmitted if the multicast transmission occupies part of the resources. The payload of a unicast transmission is equal to $\theta_u$ on average. The unicast traffic service rate becomes $\mu = C/\theta_u$ if no unicast and multicast transmissions are running on parallel, while $\mu^* = C^*/\theta_u$ when a multicast transmission is taking place.

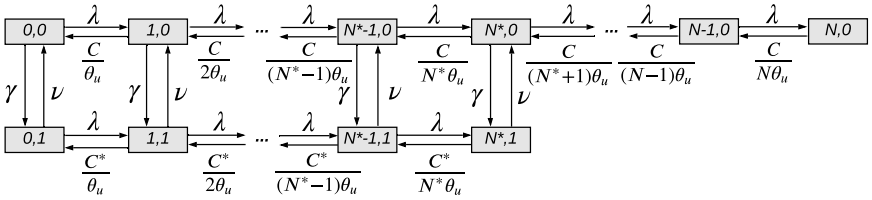



**Fig. 4.3.** State transition diagram of the CTMC $\mathbf{Y}(t)$.

Let $N(t)$ be the number of active unicast transmissions, and $L(t)$ corresponds to the number of ongoing multicast sessions at $t > 0$. We define a two-state CTMC $\mathbf{Y}(t) = \{N(t), L(t), t > 0\}$ over the state space $\mathcal{Y} = \{(n,l) : n = \overline{0,N} \text{ if } l = 0, n = \overline{0,N^*} \text{ if } l = 1\}$. Let then $q(n,l)_{(n,l)\in\mathcal{Y}}$ denote the steady-state probabilities of $\mathbf{Y}(t)$. The state transition diagram of the process of interest is given in Fig. 4.3. Therefore, we can define the system of equilibrium equations to obtain the steady-state probabilities as follows:

$$
\begin{aligned}
(\lambda+\gamma)q(0,0) &= \nu q(0,1)+\mu q(1,0),\\
(\lambda+\gamma+\frac{\mu}{n})q(n,0) &= \lambda q(n-1,0)+\frac{\mu}{n+1}q(n+1,0)+\\
&\quad +\nu q(n,1),\ n=\overline{1,N^*},\\
(\lambda+\frac{\mu}{n})q(n,0) &= \lambda q(n-1,0)+\frac{\mu}{n+1}q(n+1,0),\\
&\quad n=\overline{N^*+1,N-1},\\
\frac{\mu}{N}q(N,0) &= \lambda q(N-1,0),\\
(\lambda+\nu)q(0,1) &= \gamma q(0,0)+\mu^* q(1,1),\\
(\lambda+\nu+\frac{\mu^*}{n})q(n,1) &= \lambda q(n-1,1)+\frac{\mu^*}{n+1}q(n+1,1)+\\
&\quad +\gamma q(n,0),\ n=\overline{1,N^*-1},\\
(\nu+\frac{\mu^*}{N})q(N^*,1) &= \lambda q(N^*-1,1)+\gamma q(N^*,0).
\end{aligned}
\tag{4.8}
$$

To solve the system of equations (4.8) we first introduce unnormalized probabilities $q^*(n,j)_{(n,j)\in\mathcal{Y}}$, from which we can compute the steady-state probability as follows:

$$
q(n,l)=\frac{q^*(n,l)}{\sum_{(i,j)\in\mathcal{Y}}q^*(i,j)},
\tag{4.9}
$$

The unnormalized probabilities $q^*(n,j)$ can be computed using the recursive algorithms 1.

**Algorithm 1:** Recursive algorithm for computing steady-state probabilities

**Step 1:** compute coefficients $a_{k,j}$ and $b_{k,j}$, where $j=\{0,1\}$;
**input:** $a_{0,0}=1$, $b_{0,0}=0$, $a_{0,1}=0$, $b_{0,1}=1$; $a_{1,0}=\frac{\lambda+\gamma}{\mu}$, $b_{1,0}=-\frac{\nu}{\mu}$;
**for** *2 to* $N^*+1$ **do**
  $a_{k,0}=\frac{1}{\mu}\big[(\lambda+\mu+\gamma)a_{k-1,0}-\lambda a_{k-2,0}-\nu a_{k-1,1}\big]$;
  $b_{k,0}=\frac{1}{\mu}\big[(\lambda+\mu+\gamma)b_{k-1,0}-\lambda b_{k-2,0}-\nu b_{k-1,1}\big]$;
**for** *2 to* $N^*$ **do**
  $a_{k,1}=\frac{1}{\mu^*}\big[(\lambda+\mu^*+\nu)a_{k-1,1}-\lambda a_{k-2,1}-\gamma a_{k-1,0}\big]$;
  $b_{k,1}=\frac{1}{\mu^*}\big[(\lambda+\mu^*+\nu)b_{k-1,1}-\lambda b_{k-2,1}-\gamma b_{k-1,0}\big]$;
**for** $N^*+2$ *to* $N$ **do**
  $a_{k,0}=\frac{1}{\mu}\big[(\lambda+\mu)a_{k-1,0}-\lambda a_{k-2,0}\big]$; $b_{k,0}=\frac{1}{\mu}\big[(\lambda+\mu)b_{k-1,0}-\lambda b_{k-2,0}\big]$;
**Step 2:** Compute unnormalized probabilities $q^*(i,j)$, where $(i,j)\in\mathcal{Y}$;
**input:** $x=(\mu a_{N,0}-\lambda a_{N-1,0})/(\lambda b_{N-1,0}-\mu b_{N,0})$, $q^*(0,0)=1$, $q^*(0,1)=x$;
**for** $(i,j)\in\mathcal{Y}$ **do**
  $q^*(i,j)=a_{i,j}+b_{i,j}x$;

Having the steady-state probability distributions $q(n,l)_{(n,l)\in\mathcal{Y}}$, we can provide performance metrics of interest. The probability $P_u$ gives the portion of blocked unicast transmission due to the lack of the system resources and can be computed as follows:

$$P_u = q(N, 0) + q(N^*, 1) \tag{4.10}$$

while the probability $P_m$ corresponds to the portion of the blocked multicast transmission and is given as:

$$P_m = q(N^*, 1) + \sum_{i=N^*}^{N} q(i, 0) \tag{4.11}$$

The average time of a multicast transmission block yields:

$$T_m = \frac{1}{\lambda(1 - P_u)} \sum_{i=N^*}^{N} iq(i, 0) \tag{4.12}$$

### 4.2.5 Performance evaluation

We define four different thresholds for numerical analysis to indicate the portion of resources available for multicast transmissions. We start from the equal RBs partitioning ($\delta = 0.5$), i.e. unicast and multicast transmissions can occupy up to 50% of NPDSCH and NPDCCH subframes, towards lowering the number of subframes available for multicasting.

For the scheme with guaranteed multicast services, we consider three different scheduling policies given in terms of multicast transmission rate $\gamma = \{0.01, 0.1, 1\}$ (1/s). When $\gamma = 1$ multicast transmissions are scheduled by the network each 1 s on average. If $\gamma = 0.1$ and $\gamma = 0.01$ the transmissions are scheduled each 10 s and 100 s, respectively.

First, we analyze the average size of a multicast group under an increasing rate of $\epsilon$, i.e., the device wake-up rate. It is worth mentioning that the probability of choosing more conservative MSC grows as the number of devices in the multicast group increases. As shown in Fig. 4.4, the long interval between consecutive multicast transmissions increases the size of the multicast group significantly, even though it can be reduced with a higher threshold level to a certain extent. For instance, when only 12.5% of the resources are available for multicast transmissions, by shortening the scheduling interval (from $\gamma = 0.01$ to $\gamma = 0.1$ and $\gamma = 1$), the group size can be reduced by up to 30%, while with the higher thresholds 37.5% and 50% by up to 70% of group size reduction can be achieved. If we successively increase the threshold for a fixed $\gamma$, only 22% of improvement can be reported. It means that the scheduling policy impacts the multicast group size more than the resource allocation strategy.

Fig. 4.5 illustrates the average waiting time for a multicast transmission for different $\theta_m$. The long scheduling interval leads to a significant increase in waiting when the multicast payload is relatively small (100 kb). On the contrary, results with the heavier payload are different. The waiting times for different scheduling strategies are almost equal when the threshold is set to 12.5%, and the gap between blue and red bars in Fig. 4.5(b) slowly increases as more resources can be allocated to the multicast transmissions. We remind that devices cannot join ongoing multicast transmissions in our scenario, so they have to wait for the next one. A new transmission can start only after the end of the previous one. Therefore, the primary source of a high waiting delay in Fig. 4.5(a) is the long scheduling interval,

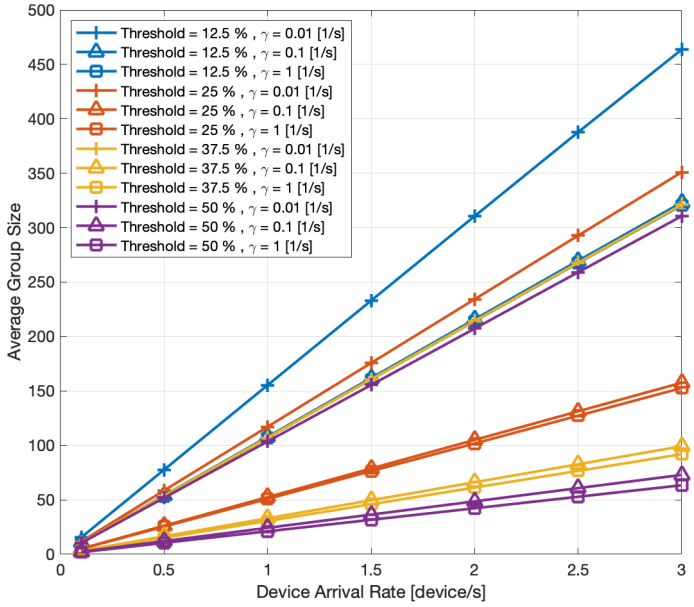
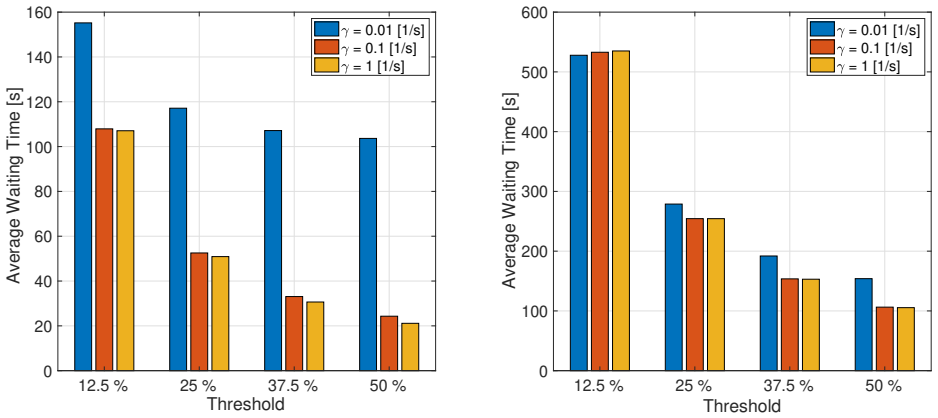



**Fig. 4.4.** Average group size.

meaning that most of the time allocated resources are left unused, which is not true for $\gamma = 0.1$ and $\gamma = 1$. In contrast, the lack of radio resources is the main reason for a long waiting time, as shown in Fig. 4.5(b). By increasing the threshold, the delay for all scheduling policies significantly improves.

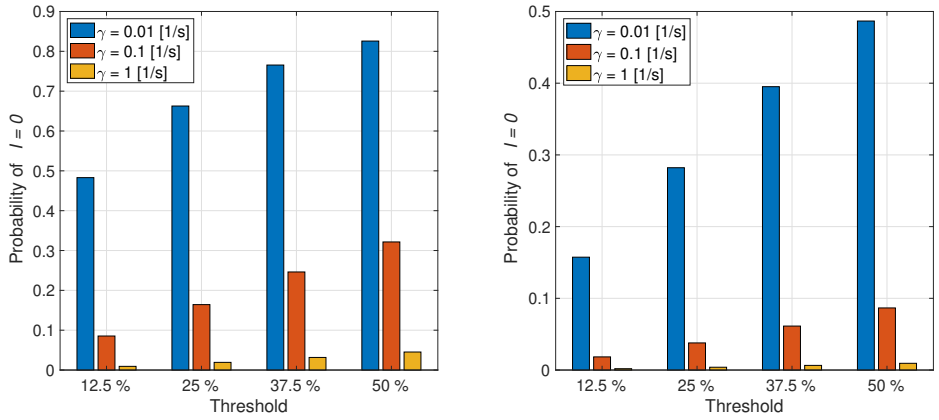



(a) $\theta_m = 100$ kb (b) $\theta_m = 800$ kb

**Fig. 4.5.** Average waiting time for the group-based transmission.

We have highlighted that not all resources are efficiently used. As shown in Fig. 4.6(a), up to 80% of the time, radio resources dedicated to multicast traffic remain unused. Note that probability $P$ decreases almost proportionally to the scheduling interval reduction. Fig. 4.6(b)

demonstrates better performance than the one shown in Fig. 4.6(a) mainly due to the higher traffic payload. Note that with a comparable waiting time for scheduling policies $\gamma = 0.1$ and $\gamma = 1$ depicted in Fig. 4.5(b), only few resources remain unused for $\gamma = 1$.




(a) $\theta_m = 100$ kb (b) $\theta_m = 800$ kb

**Fig. 4.6.** Probability of the radio resources being idle.

For the second scheme, we assume the maximum data rate for NPDSCH on the anchor carrier for the in-band operation close to 56.6 kbps. The effective data rate is lower since the unicast transmissions are scheduled in the NPDCCH, and are followed by the acknowledgment in NPUSCH. Furthermore, there are specific timing gaps between these channels. Therefore, the effective data rate $C$ of NPDSCH ranges from 0.36 kbps to 15.3 kbps [55]. The unicast payload $\theta_u = 520$ (kb), while multicast transmission carried $\theta_m = 800$ (kb) of useful information.

Fig. 4.7 illustrates the probability of a multicast transmission being blocked due to the resource shortage. As expected, the metric is mainly impacted by the threshold. It is low when only 12.5% of the total resources are requested for the multicast transmission. The blocking probability grows along with the unicast traffic rate and the number of resources allocated for group-based communication. A longer scheduling interval can provide only minor metric improvement.

To further analyze the impact of the multicast transmission blockage, we focus on the average time that the transmissions are blocked, as depicted in Fig. 4.8. Scheduling is a more critical parameter for blocking transmission delay than for blocking probability. For instance, the increase in scheduling interval reduces blocking delay by more than 2 times for different thresholds. However, with a growing intensity of unicast transmissions, the gain reduces fast.

Another interesting metric is the probability of the unicast transmission block, reported in Fig. 4.9. The probability grows naturally as the number of arrivals increases. Due to the short transmission time on average and sparse arrivals, there are enough opportunities to

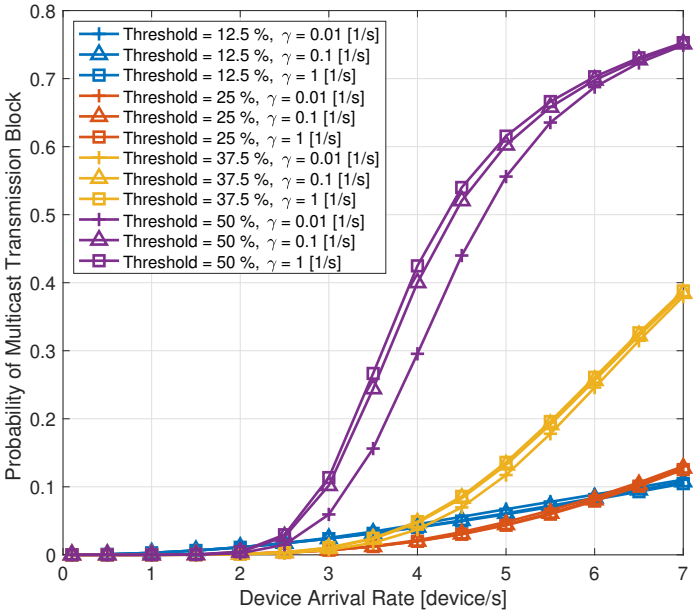



**Fig. 4.7.** Probability of multicast transmission block.

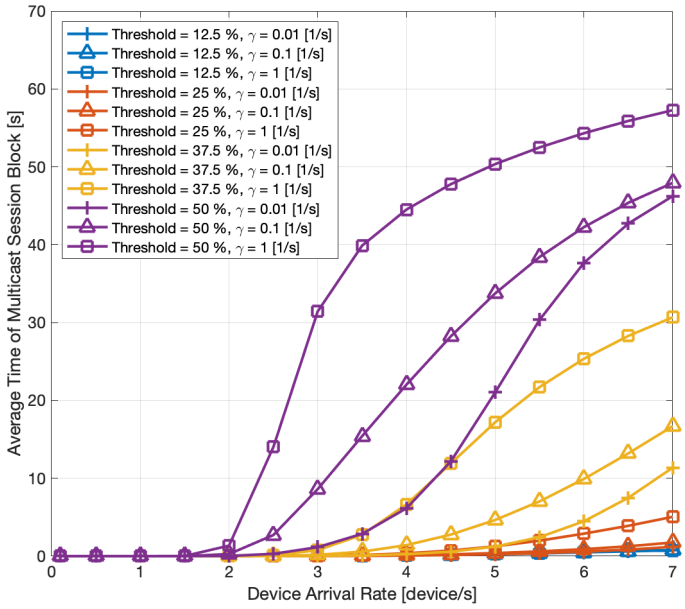



**Fig. 4.8.** Average time of multicast transmission block.

schedule multicast transmissions, which means less available resources for unicast transmissions. Comparing unicast and multicast blocking probabilities, we conclude that the scheme with dynamic resource allocation provides good fairness between unicast and multicast transmissions, even though the latter has only limited access to the system radio resources.

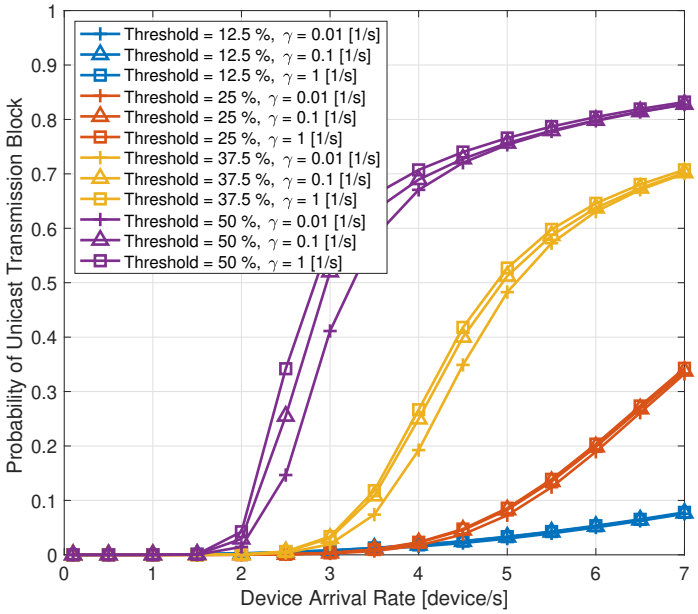



**Fig. 4.9.** Probability of unicast transmission block.

## 4.3 Paging optimization for Multicast Services

### 4.3.1 Motivations

As discussed in chapter 4, the 3GPP-based solution for PTM support in cellular networks may fail to provide reasonable latency in some delay-sensitive applications. SC-PTM is supported only in idle mode mainly to avoid excessive signaling overhead and Random Access Channel (RACH) overload having in mind massive IoT firmware/software update applications. Such transmissions may last for hours or even days [5].

In our multicast framework (section 3.4.1) we reuse the 4-handshake message exchange RA procedure to configure SC-MCCH and SC-MTCH for the group-based data reception. The time between when the multicast content is available in the RAN and when the multicast transmission is scheduled is critical for the performance of the group-based communication. This interval is depicted in Fig. 4.10 as *scheduling interval*. In a given illustrative example, the scheduling interval depends on the *paging interval* and the number of devices in each paging opportunity. A short paging interval with a high number of devices in a *paging subgroup* will increase the collisions and results in RA re-transmissions, which means that some devices will not joint the upcoming multicast transmission. A longer paging interval increases the overall service latency and waiting-for-transmission time for paged devices.

The number of devices in paging subgroups, i.e., devices that listen to the same PO and can be addressed in one paging message, is not necessarily the same. For instance, as mentioned in section 3.5.2 , after each paging transmission a device has to update its temporal identifier and, therefore, its PO and PF, as an new security measure in 5G. By

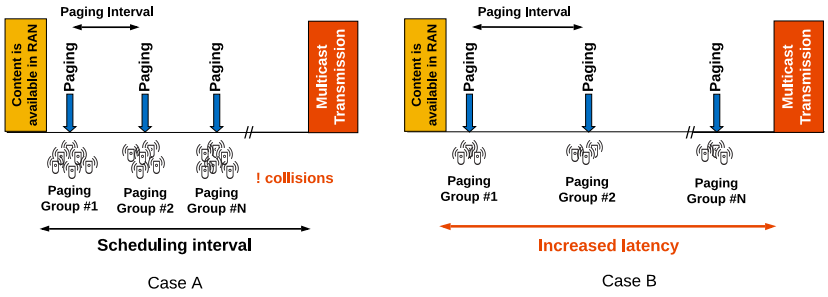



**Fig. 4.10.** Reference multicast transmission scenario.

exploiting GID, potentially any number of devices in a paging subgroup can wake up to initiate RA. Therefore, the size of the paging subgroup is another important parameter in scheduling interval optimization.

These considerations motivate us to study the impact of randomized paging on the performance of multicast services. In particular, we develop an analytical model to capture the main properties of paging and proposed group-based communication schemes to optimize the scheduling interval. We consider different paging group size distributions to investigate the impact of variance on the paging performance.

### 4.3.2 System Model

Let us consider a single-cell radio access system with $N$ devices. To inform them about the upcoming group-based transmission, the BS transmits paging messages at specific POs configured by DRX parameters. Let $T$ denote a default DRX cycle. Time offsets $\tau_i$ corresponds to the different POs within the cycle at which devices can be reached by paging messages. For the sake of mathematical tractability, we assume that intervals $\tau_k - \tau_{k-1}$ between two successive POs are Independent and Identicaly Distributed (i.i.d.) and exponentially distributed r.v. with parameter $\lambda$. Devices may have the same PO and, thus, could be paged in batches. Let $l_k$ denote the r.v. of the number of devices paged in a batch $k$, as illustrated in Fig. 4.11.

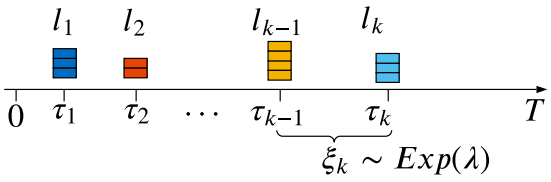



**Fig. 4.11.** Paging scheme.

After receiving the paging message, devices in a batch starts the RA procedure one by one if there is no ongoing RA transmission. Let the RA delays be i.i.d. r.v. with Cumulative

Distribution Function (CDF) $B(x)$ and mean $b$. All devices that successfully complete the RA join a multicast group and wait for the group-based transmission. Due to the collision and lack of resources, some devices will fail the RA attempt and cannot join the group. Let $C$ denote the capacity of the RA.

The reference scenario can be studied as an $M^{[X]}/G/1/C$ system with the Poisson arrival rate $\lambda l$, CDF $B(x)$ of the service time, and the finite queue length $C$, as illustrated in Fig. 4.12.

If a new batch of $i$ devices arrives and finds $j$ devices in the system, only $C-j$ devices can be served, while $i-C+j$ devices will be lost. Let $\pi$ denote the loss probability. Therefore, the service success probability that denotes the number of devices successfully received the group-based service to the total number of paged devices equals $1-\pi$.

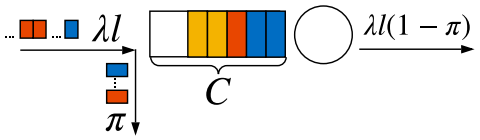



**Fig. 4.12.** System model.

Let $(l,\lambda)$ be a tuple such that $(l,\lambda) \in \{(l,\lambda)|l = \{0,1,...,l^*\}, 0 < \lambda < T\} = \mathcal{A}$, where $l^*$ stands for the maximum number of paging records in a paging message, i.e. the maximum size of a batch. Note that $l^* < \infty$ if the network addresses devices in the paging message by their unique identifiers, while $l^* = \infty$ if GID is used. The set $\mathcal{A}$ defines all possible settings of paging parameters in the system.

After the RA stage, devices join a multicast group and wait for the multicast transmission. The transmission starts when the last out of $N$ devices join the group. Let $t_n$ be the departure time of a device $n < N$. Provided that some devices may fail the RA, let $N^* = \lfloor N(1-\pi) \rfloor$ denote the number of devices that successfully join the multicast group, therefore, device $n$ waits for the content delivery $t_{N^*} - t_n$. Let $\Delta_n = t_n - t_{n-1}$ denote the time interval between two consecutive departures. The delay $D_n$ of waiting for the group-based transmission by device $n$ equals to the sum of intervals $\Delta_j$ for $j = n+1, \ldots, N^*$. Let $D(l,\lambda)$ denote the average *scheduling delay* for given paging parameters $l$ and $\lambda$, thus:

$$D(l,\lambda) = \frac{1}{N^*}\sum_{n=1}^{N^*} S_n = \frac{1}{N^*}\sum_{n=1}^{N^*}\sum_{j=n+1}^{N^*} \Delta_j = t_{N^*} - \frac{1}{N^*}\sum_{n=1}^{N^*} t_n. \tag{4.13}$$

To improve the overall multicast service latency and, consequently, device energy consumption, we define the following problem:

$$\arg A_{(l,\lambda)} \min D(l,\lambda), \text{subject to: } (l,\lambda) \in \mathcal{A},\ \pi(l,\lambda) \le \pi^*, \tag{4.14}$$

where $\pi^*$ is the target collision rate.

The problem in (4.14) can be reformulated as follows. Note that $\min D(l,\lambda) = \min t_{N^*}$ for $n = 1, \ldots, N^*$. Without loss of generality, let us assume that the departure time $t_{N^*}$ is

composed of three elements namely the arrival time of the batch with the last device $\tau_{[N/l]}$, average queuing time $\omega$, and average service time $b$. Therefore, $\min(t_{N^*}) = \min\big(\tau_{[N/l]}+\omega+b\big)$, where only $\tau_{[N/l]}$ and $\omega$ depends on $(l,\lambda)$, thus we define a reduced problem as follows:

$$\arg A_{(l,\lambda)} \min(\tau_{[N/l]}(l,\lambda)+\omega(l,\lambda)), \text{subject to: } (l,\lambda)\in\mathcal{A},\ \pi(l,\lambda)\le\pi^*. \tag{4.15}$$

To solve the problem in (4.15), we need to compute the loss probability $\pi$ and average queuing time $\omega$. In the following, we first consider an asymptotic case when $C=\infty$, and extend our results for the system with the finite queue size $C$. The notations used in this section are summarized in Table 4.1.

### 4.3.3 Analysis

We start our analysis with the overview of the results for the $M^{[X]}|G|1|\infty$ system. Then we extend these results to investigate the probability characteristics of the system with finite queue ($C<\infty$).

Let $\alpha_k(t) = ((\lambda t)^k/i!)e^{-\lambda t}$ denote the probability of exact $k$ batch arrivals over time $t>0$, and $l_i$ and $l_i^k$ stands for the probability of having $i$ devices in a batch and the probability of having $i$ devices in $k$ batches, respectively. For a given $l_i$, let $l_i^k$ define a $k$-fold convolution $l_i^k = \sum_{j=0}^{i} l_{i-j}^{k-1} l_j$. Therefore, we can define the PGF of a batch size as follows:

$$L(z) = \sum_{j=0}^{\infty} l_j z^j, \tag{4.16}$$

where $l = L'(x)|_{z\to 1} = l_1 + 2l_2 + \dots$ is an average batch size. The PGF $L_k(z)$ of the number of devices in $k$ batches yields

$$L_k(z) = \sum_{i=0}^{\infty} l_i^k z^i = L^k(x). \tag{4.17}$$

Let $\beta_k = \displaystyle\int_0^{\infty} e^{-\lambda x}\frac{(\lambda x)^k}{k!}dB(x)$ be the probability of $k$ batches arriving over an arbitrary observation time with CDF $B(x)$.

The system load $\rho$ equals $\lambda^* b$, where $\lambda^*$ is the batch arrival rate. Having $l$ devices in a batch, the system utilization becomes $\rho = \lambda l b$.

We consider two stochastic processes $X_\xi(t)$ and $X_\eta(t)$ that give the total number of devices in the system and the queue length at time $t\ge 0$, respectively. Let $t_n$ denote departure times, and $X_\eta(t_n+0)=i$ stands for the number of devices at time $t_n$. Then $X_\eta(t_n+0)$ is a Discrete Time Markov Chain (DTMC) embedded at the departure points with the countable state space $\mathcal{X}=\{0,1,\dots\}$. Provided that the system is stable, i.e. $\rho<1$, we define the steady state probabilities $p_j$ and $q_j$ for $X_\xi(t_n)$ and $X_\eta(t)$ as follows:

$$p_j = \lim_{t\to\infty} P\{X_\xi(t)=j\}, j\in\mathcal{X}, \tag{4.18}$$

$$q_j = \lim_{n\to\infty} P\{X_\eta(t_n)=j\}, j\in\mathcal{X}. \tag{4.19}$$

**Table 4.1.** Notations of section 4.3.

| Notation | Description |
|---|---|
| $N$ | Number of devices interested in the content reception |
| $N^*$ | Number of devices in the multicast group |
| $T$ | Default DRX cycle |
| $\tau_i$ | PO and arrival time of batch $i$ |
| $\lambda^{-1}$ | Average interval between two consecutive POs |
| $t_n$ | Departure time of a device $n$ |
| $\Delta_n$ | Interval between device departures |
| $l, l^*$ | Mean and maximum batch size |
| $B(x), b$ | CDF and mean of access delay |
| $C$ | Number of waiting places in the first queue |
| $\pi$ | Portion of lost devices |
| $S(l, \lambda)$ | Multicast service delay |
| $\omega(l, \lambda)$ | mean time to wait for RA |
| $l_i$ | Probability of having $i$ devices in one batch |
| $l_i^k$ | Probability of having $i$ devices in $k$ batches |
| $\alpha_k(t)$ | Probability of having $k$ batches arrivals over time $t > 0$ |
| $L(z), L_k(z)$ | Probability Generation Function (PGF) of a number of devices in one and $k$ batches |
| $\beta_k$ | Probability of having $k$ batches arrived over time with CDF $B(x)$ |
| $X_\xi(t), X_\eta(t)$ | Stochastic process and DTMC that counts the number of devices in the $M^{[X]}/G/1/\infty$ system at time $t > 0$ |
| $\mathcal{X}$ | State space of $X_\xi(t)$ and $X_\eta(t)$ |
| $\rho$ | System utilization |
| $p_j, q_j$ | Steady state probabilities of the SP $X_\xi(t)$ and $X_\eta(t)$ |
| $P(z), Q(z)$ | PGF of the probabilities $p_j$ and $q_j$ |
| $R$ | Mean queue length |
| $\omega, \bar{\omega}$ | Mean waiting time in queue |
| $Y_\xi(t), Y_\eta(t)$ | Stochastic process and DTMC that counts the number of devices in the $M^{[X]}/G/1/C$ system at time $t > 0$ |
| $\mathcal{Y}$ | State space of $Y_\xi(t)$ and $Y_\eta(t)$ |
| $\bar{p}_j, \bar{q}_j$ | Steady state probabilities of the SP $Y_\xi(t)$ and $Y_\eta(t)$ |
| $P_Y(z), Q_Y(z)$ | PGF of the probabilities $\bar{p}_j$ and $\bar{q}_j$ |
| $\pi_j$ | Probability of having $j$ devices in the queue |

Let $Q(z) = \sum_{j=0}^{\infty} q_j z^j$ and $P(z) = \sum_{j=0}^{\infty} p_j z^j$ be the PGF of probabilities $q_j$ and $p_j$, respectively. We apply a $z$-transform approach to analyse the system of equilibrium equations:

$$q_j = q_0 \sum_{k=0}^{j} l_j^k \beta_k + \sum_{i=0}^{j} q_{i+1} \sum_{k=0}^{j-i} l_j^k \beta_k, j \geq 0. \tag{4.20}$$

We multiply both parts of (4.20) by $z^j$ and sum over all $j \in \mathcal{X}$, therefore,

$$Q(z) = q_0 \frac{(1-z)\beta(\lambda - \lambda L(z))}{\beta(\lambda - \lambda L(z)) - z} \tag{4.21}$$

Implying that $Q(1) = 1$, we obtain $q_0$ from (4.21) as follows:

$$q_0 = \lim_{z \to 1} \frac{z - \beta(\lambda - \lambda L(z))}{(z-1)\beta(\lambda - \lambda L(z))} = 1 - \rho, \tag{4.22}$$

and finally

$$Q(z) = (1-\rho) \frac{(1-z)\beta(\lambda - \lambda L(z))}{\beta(\lambda - \lambda L(z)) - z}. \tag{4.23}$$

Based on the method from [57, 58], we define probabilities $p_j$ as follows:

$$\begin{aligned} p_j = & \frac{1}{b} \Big( q_0 \int_0^{\infty} [1 - B(x)] e^{-\lambda x} \sum_{k=0}^{j} l_j^k \beta_k \\ & + \sum_{i=0}^{j} q_{i+1} \int_0^{\infty} [1 - B(x)] e^{-\lambda x} \sum_{k=0}^{j-i} l_j^k \beta_k \Big), j \geq 0. \end{aligned} \tag{4.24}$$

Applying again $z$-transform method to (4.24) we obtain

$$P(z) = \frac{1}{\lambda b} \left[ q_0 \sum_{i=0}^{\infty} \beta_i \sum_{k=0}^{i-1} L^k(z) + \frac{1}{z}(Q(z) - q_0) \sum_{i=0}^{\infty} \beta_i \sum_{k=0}^{i-1} L^k(z) \right], \tag{4.25}$$

while algebraic transformations yield

$$P(z) = \frac{1-\rho}{\lambda b} \frac{(1-z)(1 - \beta(\lambda - \lambda L(z)))}{(1 - L(z))(\beta(\lambda - \lambda L(z)) - z)}. \tag{4.26}$$

Once $P(z)$ is known, we can compute mean queue length $Q$:

$$Q = \lim_{z \to 1} P'(z) = \frac{\lambda l (b^{(2)}/b) + \lambda l b (l^{(2)}/l - 1)}{2(1-\rho)} \tag{4.27}$$

and mean waiting time $w$ as follows:

$$w = \frac{Q}{\lambda l} = \frac{b^{(2)}/b}{2(1-\rho)} + \frac{b((l^{(2)}/l) - 1)}{2(1-\rho)}. \tag{4.28}$$

The first component in (4.28) corresponds to the mean waiting time before the first device in a batch start its service time; the second component stands for he mean waiting time before a random device in the batch starts the service.

We now analyze processes $Y_\xi(t)$ and $Y_\eta(t)$, which keep the same definitions as processes $X_\xi(t)$ and $X_\eta(t)$ except for the state space $\mathcal{Y}$, which is now a finite set of non-negative integers $\mathcal{Y} = \{0, 1, \ldots, C\}$. Let $p_j^*$ and $q_j^*$ be the steady state probabilities of $Y_\xi(t)$ and $Y_\eta(t)$:

$$p_j^* = \lim_{t \to \infty} P\{Y_\xi(t) = j\}, j \in \mathcal{Y}, \tag{4.29}$$

$$q_j^* = \lim_{n \to \infty} P\{Y_\eta(t_n + 0) = j\}, j \in \mathcal{Y}, \tag{4.30}$$

*Lemma.* Let $\zeta_n$ and $\zeta_n^r$ be ergodic Markov Chain with the state space $\mathcal{Z} = \{0, 1, \dots\}$ and $\mathcal{Z}^r = \{0, 1, \dots, r\}$, and steady state probabilities $(s_j)_{j\in\mathcal{Z}}$ and $(s_j^r)_{j\in\mathcal{Z}^r}$, respectively. Let $S(z) = s_0\bar{S}(z)$ and $S^r(z) = s_0^r\bar{S}^r(z)$ denote their PGF. If transition probabilities $s_{ij}^r$ of $\zeta_n^r$ are such that

$$s_{ij}^r = \begin{cases} s_{ij}, j < r, i \in \mathcal{Z}^r, \\ \sum\limits_{i\geq r} s_{ij}, j = r, i \in \mathcal{Z}^r, \end{cases} \tag{4.31}$$

then $S^r(z) = s_0^r\bar{S}(z)|_r$, where

$$S(z)|_r = \sum_{j=0}^{r} s_j z^j. \tag{4.32}$$

Transition probabilities of $X_\eta(t)$ and $Y_\eta(t)$ meet the condition (4.31), thus

$$Q_Y(z) = \bar{q}_0 Q(z)|_C, \tag{4.33}$$

and $\bar{q}_0$ can be calculated from $Q_Y(1) = 1$. Based on (4.24) and (4.32) one can find the relation between $p_j^*$ and $q_j^*$, $j \in \mathcal{Y}$. The probability $\bar{p}_0$ that the server is idle equals $\bar{q}_0$ and $p_j^* = \bar{p}_0 p_j/(1-\rho)$, the probability $\bar{p}_C = 1 - \dfrac{\bar{p}_0}{1-\rho}\sum\limits_{j=0}^{C-1} p_j$.

Let $\pi_j$ be a probability of having $j$ devices in waiting places and seeing that $i - C + j$ devices of the batch will be lost:

$$\pi_j = \frac{1}{l}\sum_{i=C-j+1}^{\infty}(i - C + j)l_i. \tag{4.34}$$

Therefore, the loss probability $\pi$ can be calculated as follows:

$$\pi = \sum_{j=0}^{C} p_j^*\pi_j. \tag{4.35}$$

Finally, we define the PGF $P_Y(z)$ by using an already known PGF $P(z)$ and probability $\pi$:

$$P_Y(z) = \frac{1-\rho(1-\pi)}{1-\rho}\Big(P(z)|_{C-1} + \theta_c z^c\Big), \text{where} \tag{4.36}$$

$$\theta_c = \sum_{j=C}^{\infty} p_j - \rho\Big(\sum_{j=C}^{\infty} p_j + \sum_{j=0}^{C-1} p_j\pi_j\Big). \tag{4.37}$$

The mean waiting time $\omega^*$ in the system with $C < \infty$ waiting places is given as:

$$\omega^* = \frac{1}{\lambda l}\sum_{j=1}^{C} j p_j^* = \frac{P_Y'(1)}{\lambda l}. \tag{4.38}$$

### 4.3.4 Performance evaluation

We compare different distributions of the paging group size, which are referred to as batch size distributions in our system model. We consider three discrete random variables distributions, namely uniform (U), geometrical (G) and binomial (B). Their CDFs are illustrated in Fig. 4.13. The mean group size $l$ is given as the ratio $N/(\lambda T)$, implying that the paging time offsets for all $N$ devices limited within the default DRX cycle $T$. We also take into account a deterministic distribution (D) of the group/batch size to reflect a case when network can reach all $l^*$ devices in the paging record list from the paging message.

The paging rate equals to $\lambda$ POs per ms. The performance metrics were obtained under the assumption of exponentially distributed service times with mean $b$. The complete list of notations and corresponding values can be found in Table 4.2.

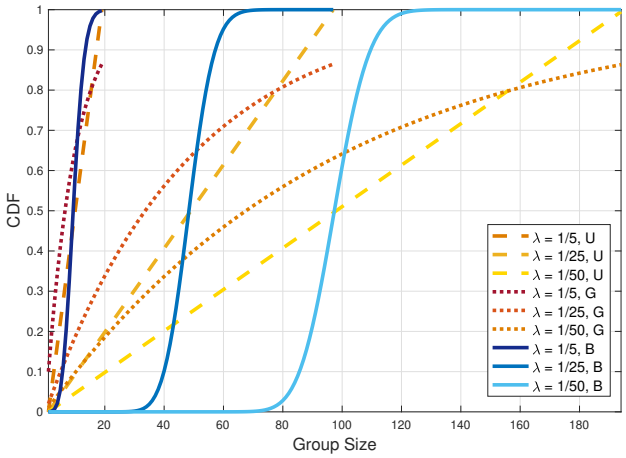



**Fig. 4.13.** Batch size distribution.

**Table 4.2.** System parameters of section 4.3.

| Parameter | Value |
|---|---|
| Number of devices, $N$ | 1000 |
| Default DRX cycle, $T$ | 512 ms |
| Queue length, $C$ | 30 |
| Mean service time, $b$ | $T/N$ |
| Mean paging group size, $l$ | {16,...,1000} |
| Maximum number of paging records, $l^*$ | $\{16, \infty\}$ |
| Mean paging interval, $\lambda^{-1}$ | {5,10,...,50} |

The first metric of interest is the loss probability $\pi$ that corresponds to the portion of devices that failed to join the multicast group and receive the content. By increasing the paging interval, fewer devices will simultaneously contend for the network resources. The fixed size of paging group (D) allows to achieve a better performance when the group size is relatively small, as shown in Fig. 4.14 for $l = 16$ and $l = 32$. The variable size of paging groups may result in local temporal congestion. When the arrived batch's size is greater than the number of available waiting places, the unserved devices will be lost. The geometrical distribution incurs the highest variance among the considered distributions. It gives rise to stronger performance degradation compared to the results demonstrated by other distributions.

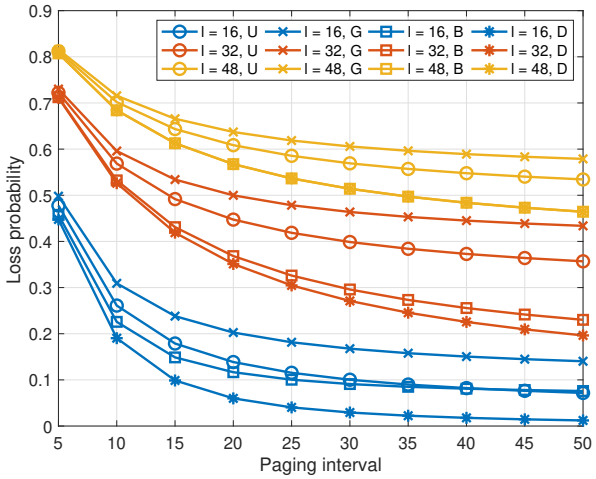



**Fig. 4.14.** Portion of devices lost due to congestion of the RA procedure.

The waiting time $\omega^*$ and system service time $b$ give in total the RA delay. Since the system service time does not depend on the paging procedure parameters, we discuss the results for a waiting time $\omega^*$, given in Fig. 4.15. We notice a local maximum when the mean paging interval equals 10 ms, paging groups contain 16 devices and follow binomial and deterministic distributions. However, if the paging group size is greater than $C$, the waiting time is a monotonic function that tends to its asymptotic value when the paging interval's length increases. The average scheduling delay, defined as $S(\lambda, l)$ in section 4.3.2, does not

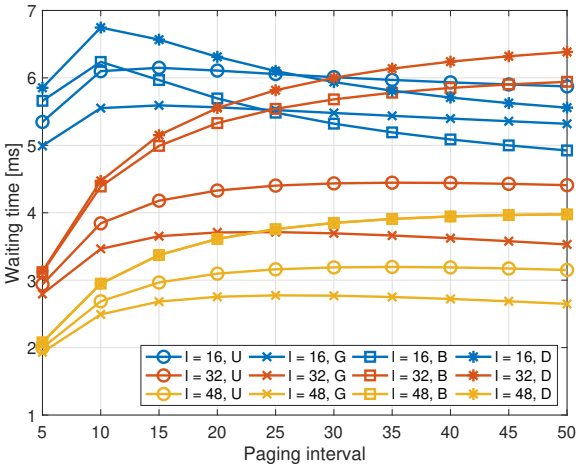



**Fig. 4.15.** Average waiting time $\omega^*$.

depend on the law of the paging group size distribution. The visualization of the set of all possible $S(\lambda, l)$ values when $\lambda^{-1} = \{5, \ldots, 50\}$ and $l = \{5, \ldots, 50\}$ is given in Fig. 4.17, while Fig. 4.16 shows only selected projections. The metric monotonously increases with increasing

paging interval. The optimal parameters $(l, \lambda)$ can be obtained from (4.14) for some given system parameters and $\pi^*$, as shown for $\pi^* = 0.5$ in Fig. 4.17. The metric is useful for the

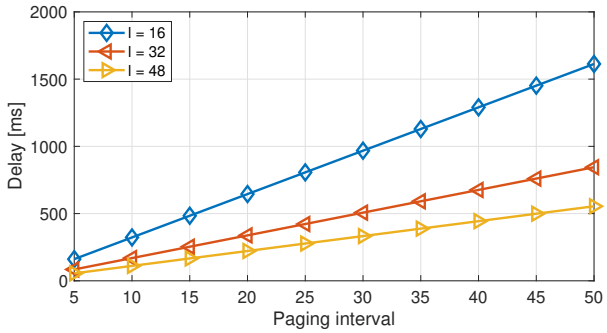



**Fig. 4.16.** Multicast scheduling delay for a set paging interval and multicast service delay

.

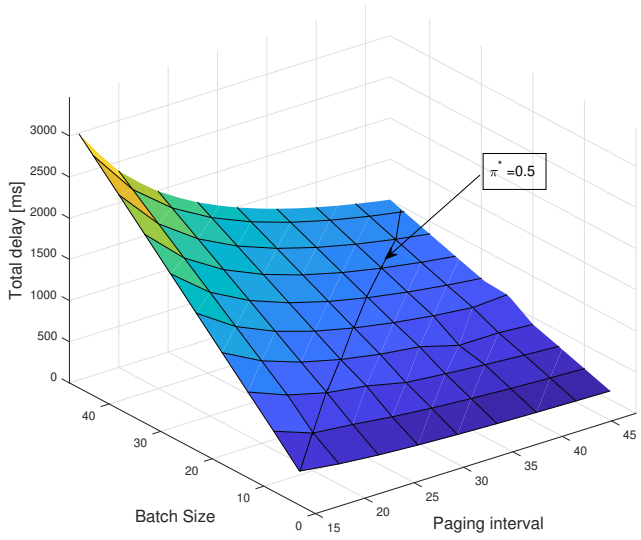



**Fig. 4.17.** Multicast scheduling delay and its optimal value.

multicast scheduling and device paging. For example, devices could be instructed to go idle at the end of the RA message exchange, and wake up for the data reception without any changes in their regular DRX cycles, or increase in energy consumption.

## 4.4 Conclusions

In section 4.2, we evaluated the performance of multicast and unicast transmissions in a NB-IoT system considering two resource allocation approaches and variable system parameters, e.g., multicast transmission scheduling rate and unicast user arrival rate. We discussed how metrics of interest could be improved and optimized in each applied resource allocation scheme.

The scheme with resource reservations for the group-based transmission works well for delay-sensitive applications with a small and intensive payload. Applications with relaxed latency requirements will benefit from a dynamic resource allocation strategy with a threshold for resources allocated to the multicast transmissions. Moreover, this strategy demonstrated better fairness among unicast and multicast users.

Section 4.3 presented a trade-off analysis for multicast transmission scheduling delay. We analytically study the interplay between the delay and paging parameters. Different paging group size distributions considerably impact communication and access delay, but the total scheduling delay is affected by group size variation. The waiting-for-transmission time is the main contributor to the total delay. As the number of devices willing to receive information in PTM mode grows, the delay becomes indifferent to the paging group size distribution.

# 5

# Age of Information for IoT Applications

*This chapter is dedicated to analyzing the AoI in IoT systems with a focus on the distribution or the LST of the age-related metrics. The results describe the aging of information packets traveling from the source to the destination through the relay node.*

## 5.1 Introduction

For decades, communication networks have considered an end-to-end delay, throughput, energy efficiency, and service reliability as the most critical performance metrics. Moreover, packet delay has been the only performance indicator to capture the latency requirements of a transmission. The concept of AoI was introduced in 2011 to characterize the information's freshness from the receiver's perspective [7]. It tells how old the last packet received by the destination is. Another age-related metric is the PAoI [59] that quantifies the maximum value of AoI. Both metrics have attracted intensive interest over the last years due to their novelty and advantages over conventional metrics. The AoI has been proved to be a useful metric in many real-time and context-aware IoT applications [60]. In these applications, a source generates information in the form of packets and sends them to a destination. The receiver is interested in up-to-date knowledge of the information available at the remote node, rather than in packet delay. Examples are sensor networks, vehicular networks, and tracking systems.

The *timeliness* of status updates was investigated for the first time in the *broadcast* scenario in the context of vehicular networks [61]. In particular, different sensor nodes that track velocity, acceleration, parking radar signals distribute time-critical content to other vehicles in proximity over wireless links to improve road safety and transportation intelligence.

Freshness of distributed information is a critical requirement in applications like Wireless Sensor Networks (WSN) for healthcare [62] and environmental monitoring [63]. It is of vital importance in content caching applications [64, 65]. In [64], a model for dynamic content requests based on its freshness and popularity has been proposed, while in [65] a new metric called Age of Synchronization has been introduced to measure the freshness of the local cache together with basic AoI.

Works in [66,67] have investigated information aging in wireless camera networks. Information from multiple views is processed to build a multi-view image that is then transferred to the monitor. In this work, communication and computing resources have been jointly

optimized to ensure an up-to-date view at the receiver. Moreover, a problem of *correlated observations* has been addressed. In [68] different policies have been proposed to minimize the average AoI of sensors with correlated observations.

AoI has been extensively studied in broadcast/multicast wireless networks for transmission scheduling over unreliable channels. Results from [69, 70] show that a greedy policy optimizes AoI, meaning that packets with the highest current age should be scheduled first. A similar scenario has been investigated in [71], where a scheduling algorithm with reduced complexity has been developed. An optimal multicast transmission scheduling algorithm aimed at improving device energy consumption has been proposed in [72]. The freshness of multicast information with hard deadlines in real-time IoT systems has been studied in [73], where the update transmission is terminated when the deadline reached, or a sufficient number of IoT devices receives the update.

There are many examples of age-sensitive IoT applications [74]. In [8], the authors consider a MEC system and investigate the impact of pre-processing the raw data collected from sensors on age performance. Another example is given in [75], which addresses the problem of the optimal status update generation in a wireless system where the source of updates runs applications with regular IoT traffic and AoI-sensitive traffic.

Finally, satellites' role in tracking applications for wide-area sensor and vehicular networks is growing due to their natural way to provide ubiquitous coverage for the massive IoT in areas where cellular communications are not available or are less cost-effective [76]. As explained in [77], Low Earth Orbit (LEO) satellites organized in a constellation may collect the status updates and forward them over the inter- or intra- satellite links to the ground station.

The timeliness of information indicated by its age can be used as a tool, performance metric, or a concept in vastly diverse systems [9]. The following sections present our contributions to this research area. In section 5.2, we give the definition of AoI, its peak value, and briefly explain some concepts used for the analysis. In section 5.3, we study the timeliness of status updates in a two-hop communication network with packet prioritization. Section 5.4 continues the analysis of two-hop communication systems with a focus on full PAoI distribution.

## 5.2 AoI definition and preliminaries

Let us consider a communication system between a source and a destination. The source observes a stochastic process $X(t)$ and generates update packets containing (i) a sample of the process $X(t_i)$ at time $t_i$ and (ii) the timestamp $t_i$. All packets $i = 1, 2, \ldots$ are sent over the communication system, updating the source's status at the destination upon arrival. The packets' arrival rate to the system is equivalent to the $X(t)$ sampling rate.

Let $t'_i$ be the time when the packet $i$ is received at the destination. The index of the most recently received update at time $t$ is $N(t) = \max\{i|t'_i \leq t\}$. Then the AoI of a given source

is a random process

$$\Delta_t = t - t_{N(t)}. \tag{5.1}$$

The average AoI of the process $\Delta_t$ can be defined as follows:

$$\mathbb{E}[A] = \lim_{T\to\infty} \frac{1}{T} \int_0^T \Delta_t dt. \tag{5.2}$$

Even though the average AoI is a useful metric, our aim is to obtain the distribution of the AoI or its LST since it is more informative in non-trivial communication systems with packet management or a complex topology. We focus on systems that can be modeled as two First Come First Served (FCFS) queues in tandem. The motivation behind this choice will be explained in sections 5.3.1 and 5.4.1.

A general formula for the stationary distribution of the AoI in ergodic information update systems has been defined [78]. Applying the sample-path approach, authors have proved that the stationary distribution of the AoI in *ergodic queueing systems* can be defined in terms of PAoI and system time. Below, we summarize the idea of the sample-path method and the main results for our analysis. The detailed proofs and assumptions are given in [78].

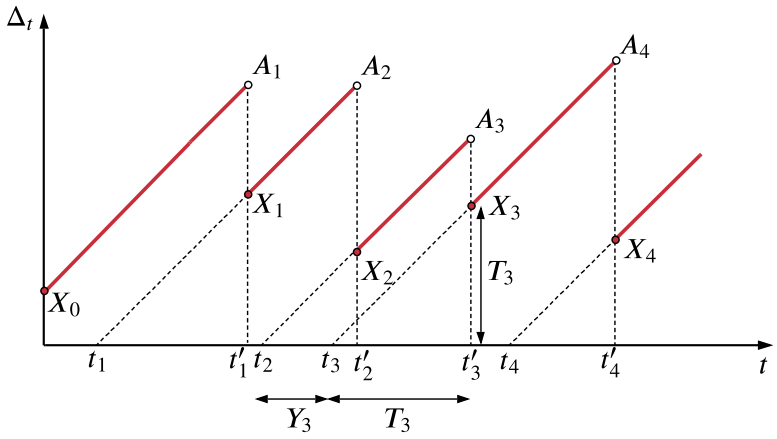



**Fig. 5.1.** Example of the AoI evolution.

The evolution of the AoI exhibits a sawtooth pattern, as illustrated in Fig. 5.1. Let $\Delta_0 = Z_0$ at time $t = 0$. The AoI $\Delta_t$ increases linearly with slope one until $t'_i$ when update $i$ arrives to the destination. At $t'_i$ it is reset to $Z_i = \Delta_{t'_i}$. Let $\{t'_i, i \geq 0\}$ be a deterministic point process such that $t' = 0$, $t'_{i-1} < t'_i$. Therefore, $Z_i$ is the AoI immediately after the $i$-th update. The AoI process $\Delta_t$ is non-negative, piece-wise non-decreasing, right-continuous process with jumps at $t'_i$ that is explicitly determined by a deterministic marked point process $\{Z_i, t'_i\}$, $i \geq 1$:

$$\Delta_t = Z_{i-1} + (t - t'_{i-1}),\ t \in [t'_{i-1}, t_i),\ i \geq 1. \tag{5.3}$$

The PAoI $A_i$ or the AoI immediately before the $i$-th update yields:

$$A_i = \lim_{t \to t_i'^-} = Z_{i-1} + (t_i - t'_{i-1}). \tag{5.4}$$

We then can define *asymptotic frequency distributions* of the AoI $\Delta^*(x)$, PAoI $A^*(x)$ and $Z^*(x)$ as a long-run fraction of time when $\Delta_t$, $A_i$ and $Z_i$ are not greater than an arbitrarily $x$:

$$\Delta^*(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{\Delta_t \leq x\}} dt \tag{5.5}$$

$$A^*(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{A_i \leq x\}} dt \tag{5.6}$$

$$Z^*(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{Z_i \leq x\}} dt \tag{5.7}$$

where $\mathbf{1}\{\cdot\}$ denotes an indicator function. Let $N(T) = \max\{n | t_n \leq T\}$ denote the number of arrivals by time $T$. For a finite number of arrivals over a finite interval where $\lim_{T \to \infty} N(T)/T = \lambda$, (5.8) can be defined as follows:

$$\Delta^*(x) = \lambda \int_0^x (Z^*(y) - A^*(y)) dy. \tag{5.8}$$

If a given communication system represents a general ergodic (i.e. stable) FCFS queuing system, where the timestamps of update packets are equal to their arrival times, then $Z_i$ equals to the system time $T_i = t'_i - t_i$. According to the definition given in [78], the general ergodic FCFS queuing system is a queuing system where informative packets arrive and depart in a FCFS manner. If a packet $j$ updates the information at the destination, i.e., resets the AoI at time $t'_j$ then it is assumed to be informative. If other packets in an ergodic system do not change packets' order of a given source, the system is assumed to be a general FCFS system.

Let $Y_i = t_i - t_{i-1}$ be the $i$-th interarrival time, therefore, the PAoI yields:

$$A_i = T_{i-1} + (t_i - t'_{i-1}) = Y_i + T_i. \tag{5.9}$$

Since the system is stationary, defined asymptotic distributions (5.5), (5.6), and (5.7) converge to the stationary distributions $F_\Delta(x)$, $F_A(x)$, and $F_T(x)$, respectively. Therefore, following (5.8), the PDF of the AoI can be given by

$$f_\Delta(x) = \lambda(F_T(x) - F_A(x)). \tag{5.10}$$

The LST of the AoI yields

$$\begin{aligned} \delta(s) = &\int_0^\infty e^{-sx} dF_\Delta(x) = \lambda \int_0^\infty e^{-sx} (F_T(x) - F_A(x)) dx \\ = &\frac{\lambda}{s}(\tau(s) - \alpha(s)), \end{aligned} \tag{5.11}$$

where $\tau(s) = \int_0^\infty e^{-sx} dF_T(x)$ and $\alpha(s) = \int_0^\infty e^{-sx} dF_A(x)$ stands for the LST of the system delay and PAoI by definition.

## 5.3 Age of Information in Tandem Queues with Priorities

### 5.3.1 Motivation

A close examination of the use cases and application presented in section 5.1 reveals the common features of the update systems and existing research gaps. A single queuing system can capture the timeliness of information only between two directly communicating instances, but it fails to give adequate results in multi-hop networks, i.e., when status updates are forwarded over one or several relay nodes. Another element is the existence of heterogeneous requirements and paths: different sources of updates should be treated according to their priority level, and update flows might use different entry points to the communication system.

This motivates us to consider a general multi-hop communication system with arrivals at the intermediate nodes and different status updates priorities. For our analysis, we take the illustrative case of two nodes, where status update packets sent via the relay (the first node) takes priority over the updates sent directly to the monitor (the second node), as shown in Fig. 5.2. Priority packets preempt all non-priority packets in the second node's queue but do not impact the ongoing service. This priority policy will improve the performance of the status updates that need the relay to reach the destination, reducing the difference in performance between the two paths.

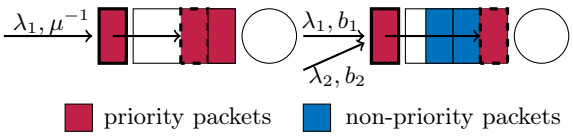



**Fig. 5.2.** System model as two FCFS queues in tandem with priorities.

In the considered system model, the second node can be a BS that schedules critical multicast traffic (red packets) and unicast traffic (blue), while the first node models a multicast group formation delay. In such a scenario, a device takes the role of a monitor. Another example is an aggregation of raw (blue) and processed (red) packets for further offloading to the monitor, assuming that the processed information is more valuable.

A system design similar to ours has been considered in [79]. Authors investigate the average AoI when the status update can be delivered either over the less reliable direct link or over the two-hop relay link with better reliability. However, all packets at the second node have been treated equally. In [8], only average PAoI is given for the two-hop tandem exponential queues with multiple sources. Authors in [80] study the average AoI of a two-hop system with packet arrivals only at the first node and a zero-waiting policy at the second node.

In [78] authors derive a general formula for the stationary distribution of the AoI in terms of the system delay and the PAoI for a broad class of $G/G/1$ systems with a single source under the general FCFS and Last Come First Serve (LCFS) packet management

policies with various preemption and packet discarding options. However, the LCFS policy can not be applied to the systems where packets carry incremental information and can not be discarded.

The idea of assigning different priorities to the update packets has been discussed for the first time in [81]. The average AoI is given for an exponential single-server system with a shared queue and LCFS discipline, where the arriving packet preempts another packet either in service or in waiting only if it has a higher priority. In [82] authors focus on a queuing system with $k$ classes of priorities, different buffer sizes, and queuing disciplines. In particular, the different combinations of infinite queues with FCFS and LCFS disciplines and queues with a single place to wait are considered. The exact expressions of the expected PAoI are given for the general service time distribution if the queues are infinite and for the exponential service time if the queue size is one, while the tight bounds have been calculated for the remaining scenarios. The works with the packet‘s prioritization mentioned above are limited to the single-node systems.

### 5.3.2 System model

We consider a two-hop network with intermediate traffic, as shown in Fig. 5.2. Status updates are generated according to a Poisson process with rate $\lambda$. With probability $p$ packets arrive at the first node and become priority packets. With probability $1-p$ all remaining packets arrive directly to the second node and become non-priority packets. The arrival rates are $\lambda_1 = p\lambda$ and $\lambda_2 = (1-p)\lambda$, respectively. Both queues apply the general FCFS discipline, but in the queue of the second node, all packets coming originally from the first node (priority packets) pre-empt in waiting packets coming directly to the second node (non-priority packets). Non-priority packets see the system as an $M/G/1$ queue with priorities, while priority packets pass through $M/M/1$ and $M/G/1$ queues connected in series.

Service times at the first node are limited to the exponential distribution for the sake of mathematical tractability, i.e. to ensure that the departure process from the first node is Poisson. Let $b_1$ and $b_2$ be the mean service times of priority and non-priority packets at the second node. The total system utilization equals to the second node utilization $\rho = \rho_1 + \rho_2$, where $\rho_j = \lambda_j b_j$, $j = \{1, 2\}$. Utilization of the first node $\rho_{11} = \lambda_1/\mu$, $\mu^{-1}$ is the mean service time at the first node.

Let $j, i$ denote packet $i$ of priority class $j$. Let $t_{j,i}$ and $t'_{j,i}$ be the time instances of packet $j, i$ arrival to the system (generation of a new status at source), and its departure from the system (updating the status at the monitor). Then $Y_{j,i} = t_{j,i} - t_{j,i-1}$ denotes the r.v. of packet $j, i$ *interarrival* time, and $T_{j,i} = t'_{j,i} - t_{j,i}$ corresponds to the r.v. of the packet's *system delay*. The AoI $\Delta_{j,i}$ at time $t > 0$ consists of the AoI $Z_{j,i-1}$ immediately after the departure of the packet $j, i-1$ and the time from $t'_{j,i-1}$ to $t$, i.e. $\Delta_{j,i} = Z_{j,i-1} + (t - t'_{j,i-1})$. As explained in section 5.2, in general FCFS systems, $Z_{j,i}$ equals to the system delay $T_{j,i}$ if all packets are time-stamped on their arrival. Therefore, the PAoI $A_{j,i} = t'_{j,i} - t_{j,i-1} = Y_{j,i} + T_{j,i}$.

In the ergodic system ($\rho < 1$), the probability PDF of the AoI can be defined as $f_{\Delta_j}(x) = \lambda_j(F_{T_j}(x) - F_{A_j}(x)),\ x \geq 0$, where $F_{T_j}(x)$ and $F_{A_j}(x)$ stand for the CDF of the system delay and PAoI, respectively [78]. The LST $\delta_j(s)$ of the AoI distribution, therefore, yields:

$$\delta_j(s) = \frac{\lambda_j}{s}(\tau_j(s) - \alpha_j(s)),\ s > 0, \tag{5.12}$$

where $\tau_j(s) = \int_0^\infty e^{-sx} dF_{T_j}(x)$ and $\alpha_j(s) = \int_0^\infty e^{-sx} dF_{A_j}(x)$.

Priority and non-priority packets arrive to the system independently, their interarrival times are exponentially distributed holding the LST $\lambda_j/(\lambda_j + s)$. System delay $T_{j,i}$ depends on the packets interarrival time $Y_{j,i}$ and the system delay $T_{j,i-1}$, it also depends on the arrival and departure processes of packets of another class. The r.v. $T_{1,i} = T_{11,i} + T_{12,i}$ while $T_{11,i}$ and $T_{12,i}$ are not independent. In the next section we define the PAoI for packet $j, i$ and then obtain the general distribution of $A_j$ for both classes of packets. A similar approach is applied for the calculation of the total system delay $T_1$ of priority packets.

Let us study the known distributions of the system delays at each node as preliminaries for further analysis. The system delay $T_{11}$ at the first node $(M/M/1)$ is exponentially distributed with parameter $\theta = \mu - \lambda_1$, the corresponding LST $\tau_{11}(s)$ equals to $\theta/(\theta + s)$. The LST of the system delay $T_{12}$ of priority packets, and the system delay $T_2$ of non-priority packets at the second node are given in [83]:

$$\tau_{12}(s) = \frac{s(1-\rho) + \lambda_2(1 - \beta_2(s))}{s - \lambda_1 + \lambda_1\beta_1(s)}\beta_1(s), \tag{5.13}$$

$$\tau_2(s) = \frac{(1-\rho)(s + \lambda_1 - \lambda_1\gamma(s))}{s - \lambda_2 + \lambda_2\beta_2(s + \lambda_1 - \lambda_1\gamma(s))}\beta_2(s), \tag{5.14}$$

where $\beta_1(s)$ and $\beta_2(s)$ are the LSTs of the service time distributions of priority and non-priority packets at the second node, $\gamma(s)$ stands for the LST of the distribution of the interval $G_1$, which elapses from the arrival of a priority packet in the empty queue of the second node until the end of continuous service of priority packets arriving afterwards. This interval is known as a busy period generated by a priority packet and its LST $\gamma(s) = \beta_1(s + \lambda_1 - \lambda_1\gamma(s))$. The busy period $G_2$ starts from the moment when a non-priority packet arrives to the empty node, therefore, its LST is $\beta_2(s + \lambda_1 - \lambda_1\gamma(s))$. For convenience we give the complete list of notations in Table 5.1.

### 5.3.3 Derivation of the AoI distribution and its statictics

#### 5.3.3.1 Priority packets

We first focus on priority packets. When packet $i$ arrives at the system, it can be queued in both nodes, queued only in one node, or go through two nodes without any queuing delay. The presence of non-priority packets at the second node hinders the derivation of the PAoI and system delay distributions. We assume that packet $i$ finds the second node free of non-priority packets with the probability $1 - \rho_2$.

**Table 5.1.** Notations of section 5.3.

| **Notation** | | **Definition** |
|---|---|---|
| $k$ | | Node index |
| $(j,i)$ | | Packet $i$ of priority class $j$ |
| $t_{j,i}$ | | Packet $(j,i)$ arrival time |
| $t'_{j,i}$ | | Packet $(j,i)$ departure time |
| $\lambda_j$ | | Arrival rate for class $j$ |
| $b_j$ | | Mean service time for class $j$ |
| $\rho_j$ | | Second node utilization by class $j$ |
| $b=\mu^{-1}$ | | Mean service time at the first node |
| $\theta$ | | Mean system delay at the first node |
| $\rho_{11}$ | | First node utilization |
| **RV** | **LST** | **Definition** |
| $Y_{j,i}$ | | Packets interarrival time |
| $S_{kj,i}$ | $\beta(s)$, $\beta_j(s)$ | Service time of packet $(j,i)$ at node $k$ |
| $W_{jk,i}$ | $\omega_{jk}(s)$ | Waiting time of packet $(j,i)$ at node $k$ |
| $T_{jk,i}$ | $\tau_{1k}(s)$, $\tau_j(s)$ | System delay of packet $(j,i)$ at node $k$ |
| $D_{j,i}$ | $\eta_j(s)$ | Supplementary to PAoI of packet $j,i$ interval as defined in Fig. 5.3 |
| $X_{1,i}$ | $\xi_1(s)$ | Supplementary to system delay of packet $j,i$ interval as defined in Fig. 5.3 |
| $G_{j,i}$ | $\gamma_j(s)$ | Busy period generated by a packet $j,i$ |
| $\tilde{Z}_{j,i}$ | $\tilde{\zeta}_{j,i}(s)$ | Residual time of interval $Z_{j,i}$ |
| $A_j$ | $\alpha_j(s)$ | PAoI of class $j$ |
| $\Delta_j$ | $\delta_j(s)$ | AoI of class $j$ |

There are six cases C1 – C6 that help to define system delay $T_{1,i}$ and PAoI $A_{1,i}$ of packet $i$ in the system of interest. Let us define intervals $D_{1,i}$ (bold red line) and $X_{1,i}$ (bold blue line) as illustrated in Fig. 5.3. Let also $\eta_1(s,C_m)$ and $\xi_1(s,C_m)$ denote the LST of the joint distribution of intervals contributing to $D_{1,i}$ and $X_{1,i}$ for a case $C_m$, $m=\{1,\ldots,6\}$, respectively.

We define the LST of the system delay $\tau(s,C_m)$ and the PAoI $\alpha(s,C_m)$ for each case. The resulting distributions will be given as a sum of LSTs of the six joint distributions, in particular

$$\tau_1(s)=\sum_m \tau_{1,i}(s,C_m) \tag{5.15}$$

$$\alpha_1(s)=\sum_m \alpha_{1,i}(s,C_m). \tag{5.16}$$

Below, we define each case in details.

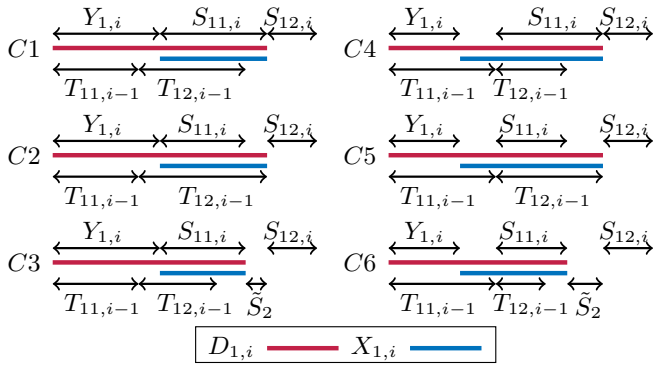



**Fig. 5.3.** Components of PAoI and system delay of priority packets.

- **Case C1.** Packet $i$ does not experience any queuing at nodes, therefore, the PAoI $A_{1,i} = D_{1,i} + S_{12,i}$ and system delay $T_{1,i} = X_{1,i} + S_{11,i}$. This happens if $T_{11,i-1} < Y_{1,i}$, $T_{12,i-1} + T_{12,i} < Y_{1,i} + S_{11,i}$ and if during the interval $Y_{1,i} + S_{11,i} - T_{11,i-1} - T_{12,i-1}$ all unserved non-priority packets complete their service and no new non-priority packets arrive. Since we assume that packet $i$ finds the second node free of non-priority packets with the probability $1 - \rho_2$, and service time $S_{12,i}$ is independent of other intervals, the LST of both metrics yields

$$\alpha_{1,1}(s, C_1) = (1 - \rho_2)\eta_1(s, C_1)\beta_1(s) \tag{5.17}$$

$$\tau_{1,1}(s, C_1) = (1 - \rho_2)\xi_1(s, C_1)\beta_1(s) \tag{5.18}$$

- **Case C2.** Packet $i$ finds the second node busy with packet $i - 1$, but its queuing delay at the first node $W_{11,i} = 0$, therefore, PAoI $A_{1,i} = D_{i,1} + S_{12,i}$ and system delay $T_{1,i} = X_{1,i} + S_{12,i}$ like in the case C1, but $D_{1,i} = T_{11,i-1} + T_{12,i-1}$, $X_{1,i} = T_{11,i-1} + T_{12,i-1} - Y_{1,i}$. This is true if $T_{11,i-1} < Y_{1,i}$ and $T_{11,i-1} + T_{12,i-1} > Y_{1,i} + S_{11,i}$. The PAoI and system delay distributions in this case give

$$\alpha_{1,2}(s, C_2) = \eta_1(s, C_2)\beta_1(s), \tag{5.19}$$

$$\tau_{1,2}(s, C_2) = \xi_1(s, C_2)\beta_1(s). \tag{5.20}$$

- **Case C3.** Packet $i$ finds the second node busy with a non-priority packet and its waiting time $W_{11,i} = 0$, thus the PAoI $A_{i,1} = D_{1,i} + \tilde{S}_2 + S_{12,i}$ and the system delay $T_{1,i} = X_{1,i} + \tilde{S}_2 + S_{12,i}$, where $\tilde{S}_2$ stands for the LST of the residual service time of a non-priority packet. This happens when $T_{11,i-1} < Y_{1,i}$ , $T_{1,i-1} < Y_{1,i} + S_{11,i}$ like in the case C1, but packet $i$ sees a non-priority packet in service with the probability $\rho_2$. The LST of the PAoI in the case C3 yields

$$\alpha_{1,3}(s, C_3) = \rho_2\eta_1(s, C_3)\tilde{\beta}_2(s)\beta_1(s). \tag{5.21}$$

LST of the $T_{1,i}$ gives

$$\tau_{1,3}(s, C_3) = \rho_2\xi_1(s, C_3)\tilde{\beta}_2(s)\beta_1(s), \tag{5.22}$$

where $\tilde{\beta}_2(s) = (1 - \beta_2(s))/s\mathbb{E}[S_2]$.

- **Case C4.** Packet $i$ is queued at the first node, but it finds the second node empty upon the arrival. The PAoI $A_{1,i}$ and system delay $T_{1,i}$ are defined as in the case C1, but in the case C4 $T_{11,i} > Y_{1,i}$ and $T_{12,i} < S_{S_{11,i}}$, in particular

$$\alpha_{1,4}(s, C_4) = (1 - \rho_2)\eta_1(s, C_4)\beta_1(s), \tag{5.23}$$

$$\tau_{1,4}(s, C_4) = (1 - \rho_2)\xi_1(s, C_4)\beta_1(s). \tag{5.24}$$

- **Case C5.** Packet $i$ is delayed by the packet $i-1$ in both nodes, if $T_{11,i} > Y_{1,i}$ and $T_{12,i} > S_i$. Given that $A_{1,i} = D_{1,i} + S_{12,i}$ and $T_{1,i} = X_{1,i} + S_{12,i}$ the distributions of PAoI and system delay yields:

$$\alpha_{1,5}(s, C_2) = \eta_1(s, C_2)\beta_1(s), \tag{5.25}$$

$$\tau_{1,5}(s, C_2) = \xi_1(s, C_2)\beta_1(s) \tag{5.26}$$

- **Case C6.** Packet $i$ is queued at the first node and finds the second node busy with a non-priority packet, then like in the case C3

$$\alpha_{1,6}(s, C_6) = \rho_2\eta_1(s, C_6)\tilde{\beta}_2(s)\beta_1(s), \tag{5.27}$$

$$\tau_{1,6}(s, C_6) = \rho_2\xi_1(s, C_6)\tilde{\beta}_2(s)\beta_1(s) \tag{5.28}$$

given that $T_{11,i} > Y_{1,i}$ and $T_{12,i} < S_i$.

We now need to calculate the LST of $D_{i,1}$ and $X_{1,i}$ for each case. These intervals are equally defined for the cases C1 and C3, and C4 and C6, therefore, we give their derivations with double indexes $\{13\}$ and $\{46\}$.

**Cases C1 and C3.** We denote the CDF of $D_{1,i}$ as $F_{D_1}(z, C_{13}) = \mathbb{P}\{D_{1,i} < z, C_{13}\}$. Given that $T_{11,i-1} < Y_{1,i}$ and $T_{1,i-1} < Y_{1,i} + S_{11,i}$ we calculate it as follows:

$$F_{D_1}(z, C_{13}) = \int_0^z dF_{Y_1}(y) \int_0^y dF_{T_{11}}(t) \int_0^{z-y} dF_S(x) \int_0^{x+y-t} dF_{T_{12}}(u). \tag{5.29}$$

The LST $\eta_1(s, C_{13}) = \int_0^\infty e^{-sz} dF_{D_1}(z, C_{13})$ yields:

$$\eta_1(s, C_{13}) = \frac{\lambda_1}{\lambda_1 + s}\beta_1(s)\tau_{12}(\lambda_1 + s) - \rho_{11}\beta^2(s)\tau_{12}(\mu + s). \tag{5.30}$$

Let $F_{X_1}(z, C_{13}) = \mathbb{P}\{X_{1,i} < z, C_{13}\}$ be the CDF of $X_{1,i}$, it can be calculated as

$$F_{X_1}(z, C_{13}) = \int_0^\infty dF_{Y_1}(y) \int_0^y dF_{T_{11}}(t) \int_0^z dF_S(x) \int_0^{x+y-t} dF_{T_{12}}(u), \tag{5.31}$$

and its LST $\xi_1(s, C_{13}) = \int_0^\infty e^{-sz} dF_{X_1}(z, C_{13})$ yields:

$$\xi_1(s, C_{13}) = \tau_{11}(s)\tau_{12}(\lambda_1) - \rho_{11}\tau_{11}(s)\beta_1(s)\tau_{12}(\mu + s). \tag{5.32}$$

**Case C2.** The CDF of interval $D_{1,i}$ and its LST $\eta_1(s, C_2)$ in the case C2 are given as follows:

$$F_{D_{1,i}}(z, C_2) = \int_0^z dF_{T_{11}}(t) \int_0^{z-t} dF_{T_{12}}(u) \int_t^{t+u} dF_{Y_1}(y) \int_0^{t+u-y} dF_S(z), \tag{5.33}$$

its LST, therefore, yields:

$$\eta_1(s, C_2) = (1-\rho_{11})\beta_1(s)\tau_{12}(s) - \beta_1(s)\tau_{12}(s+\lambda_1) + \rho_{11}\beta_1(s)\tau_{12}(\mu+s). \tag{5.34}$$

We define the CDF of interval $X_{1,i}$ as

$$F_{X_{1,i}}(z, C_2) = \int_0^z dF_{T_{11}}(t) \int_0^{z-t} dF_{T_{12}}(u) \int_t^{t+u} dF_{Y_1}(y) \int_0^{t+u-y} dF_S(z), \tag{5.35}$$

while its LST $\xi_1(s, C_2)$ gives

$$\xi_1(s, C_2) = \frac{\lambda_1}{s-\lambda_1}\tau_{11}(s)\tau(\lambda_1) - \rho_{11}\frac{\theta}{s-\lambda_1}\tau_{12}(s) + \rho_{11}\tau_{11}(s)\tau_{12}(\mu+s). \tag{5.36}$$

**Cases C4 and C6.** We define the CDF $F_{D_1}(z, C_{46})$ and $F_{X_1}(z, C_{46})$ in the cases C4 and C6 as

$$F_{D_1}(z, C_{46}) = \int_0^z dF_{T_{11}}(t) \int_0^t dF_{Y_1}(y) \int_0^{z-t} dF_S(x) \int_0^x dF_{T_{12}}(u). \tag{5.37}$$

$$F_{X_1}(z, C_{46}) = \int_0^\infty dF_{Y_1}(y) \int_y^{y+z} dF_{T_{11}}(t) \int_0^{z+y-t} dF_S(x) \int_0^x dF_{T_{12}}(u). \tag{5.38}$$

The LSTs of $D_{1,i}$ and $X_{1,i}$ give:

$$\eta_1(s, C_{46}) = \rho_{11}\tau_{11}(s)\beta^2(s)\tau_{12}(s+\mu). \tag{5.39}$$

$$\xi_1(s, C_{46}) = \rho_{11}\tau_{11}(s)\beta_1(s)\tau_{12}(s+\mu). \tag{5.40}$$

**Case C5.** The CDFs of $D_{1,i}$ and $X_{1,i}$ in the case C5 can be calculated as

$$F_{D_{1,i}}(z, C_5) = \int_0^z dF_{T_{11}}(t) \int_0^t dF_{Y_1}(y) \int_0^{z-t} dF_{T_{12}}(u) \int_0^u dF_S(x), \tag{5.41}$$

$$F_{X_{1,i}}(z, C_5) = \int_0^\infty dF_{Y_1}(y) \int_y^{y+z} dF_{T_{11}}(t) \int_0^{z+y-t} dF_{T_{12}}(u) \int_0^u dF_S(x). \tag{5.42}$$

The LSTs $\eta_1(s, A5)$ and $\xi_1(s, A5)$ in the case C5 yield:

$$\eta_1(s, C5) = \rho_{11}\tau_{11}(s)\beta_1(s)(\tau_{12}(s) - \tau_{12}(s+\mu)), \tag{5.43}$$

$$\xi_1(s, C5) = \rho_{11}\tau_{11}(s)(\tau_{12}(s) - \tau_{12}(s+\mu)). \tag{5.44}$$

Substituting (5.17),(5.19),(5.21),(5.23),(5.25),(5.27) to (5.16), we give the resulting LST of the PAoI distribution of priority packets as follows:

$$\alpha_1(s) = \Big[\frac{\lambda_1\nu}{\lambda_1+s}\beta_1(s)\tau_{12}(\lambda_1+s) - \frac{s}{s+\theta}\rho_{11}\beta_1(s)(\tau_{12}(s) - \tau_{12}(s+\mu)(1-\nu\beta_1(s)))\Big]\beta_1(s), \tag{5.45}$$

where $\nu = 1-\rho_2+\rho_2\tilde{\beta}_2(s)$.

The LST of system delay after substituting (5.18),(5.20),(5.22),(5.24),(5.26),(5.28) to (5.15) yields:

$$\tau_1(s) = \Big[\tau_{11}(s)\tau_{12}(\lambda_1)\Big(\nu - \frac{\lambda_1}{\lambda_1-s}\Big) + \tau_{12}(s)\Big((1-\rho_{11})\frac{\lambda_1}{\lambda_1-s} + \rho_{11}\tau_{11}(s)\Big)\Big]\beta_1(s). \tag{5.46}$$

Given (5.12) and (5.45)–(5.46) the LST of $\Delta_1$ yields:

$$\delta_1(s) = \Big[\tau_{11}(s)\tau_{12}(\lambda_1)\frac{\lambda_1}{s-\lambda_1}(\nu+\frac{\lambda_1}{s}(1-\nu)) + \frac{\lambda_1}{\lambda_1+s}\beta_1(s)\tau_{12}(s)(1+\frac{\lambda_1}{s}(1-\nu)) - \frac{\rho_{11}^3\beta(s)}{1-\rho_{11}}\tau_{11}(s)\tau_{12}(\mu+s)\tau_{11}(s)(1-\nu\beta(s)) + \nu\rho_{11}^2\tilde{\beta}(s)\beta(s)\Big]\beta_1(s). \tag{5.47}$$

Having the LSTs (5.45)–(5.47), we can calculate the average system delay, PAoI and AoI as $\mathbb{E}[T_1] = -\tau_1'(0)$, $\mathbb{E}[A_1] = -\alpha_1'(0)$, and $\mathbb{E}[\Delta_1] = -\delta_1'(0)$:

$$\mathbb{E}[T_1] = b + \frac{\lambda_1 b^{(2)}}{2(1-\rho_1)} + b_1 + \frac{\lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}}{2(1-\rho_1)}, \tag{5.48}$$

where $b_j^{(k)}$ denote the $k$-th moments of packet $j$ service time.

$$\mathbb{E}[A_1] = \Big(\frac{1}{\lambda_1} + b_1 + \rho_2\tilde{b}_2\Big)\tau_{12}(\lambda_1) - \rho_{11}(b+b\tau_{12}(\mu))(1-\rho_2+\rho_2\tilde{b}_2) + (1-\rho_{11})\Big(b_1+\mathbb{E}[T_{12}]\tau_{12}(\mu)\Big) + \rho_{11}(b_1+\mathbb{E}[T_{11}]+\mathbb{E}[T_{12}](1-\tau_{12}(\mu))) + \rho_{11}(1-\rho_2+\rho_2\tilde{b}_2)\Big(b_1+\mathbb{E}[T_{11}]+\mathbb{E}[T_{12}]\tau_{12}(\mu)\Big). \tag{5.49}$$

where $\tilde{b}_2 = b_2^{(2)}/2b_2$ is the average residual service time of non-priority packets.

Due to the cumbersome calculations, we give a very tight lower bound $\mathbb{E}[\underline{\Delta}_1]$ for the average AoI:

$$\mathbb{E}[\underline{\Delta}_1] = b_1 + \frac{1}{\lambda_1}\tau_{12}(\lambda_1) + \tau_{12}(\lambda_1)\mathbb{E}[T_1] + \rho_1^2\mathbb{E}[T_1] + \rho_{11}^2\Big(\frac{1}{\theta} - \frac{\rho_{11}}{\mu} + \frac{\rho_{11}}{\theta} + \frac{1}{\lambda_1} + \frac{\mu}{\lambda_1^2} - \frac{1}{\rho_{11}^2} - \frac{1}{\rho_{11}}\Big). \tag{5.50}$$

#### 5.3.3.2 Non-priority packets

Let us not move to non-priority packets. Packet $i$ can start service only if the second node is free of priority packets, i.e. at the end of the busy period $G_{2,i-1}$ or $G_1$, or if the node is empty. Let us introduce the interval $\Psi_{2,i-1} = W_{2,i-1} + G_{2,i-1}$, where $W_{2,i-1}$ stands for the waiting time of non-priority packet $i-1$. Intervals $W_{2,i-1}$ and $G_{2,i-1}$ are independent, therefore, the LST of $\Psi_{2,i-1}$ can be given as

$$\psi_2(s) = \omega_2(s)\beta_2(s+\lambda_1-\lambda_1\gamma(s)). \tag{5.51}$$

We consider three cases to define the PAoI $A_{2,i}$.

- **Case B1.** If $Y_{2,i} > \Psi_{2,i-1}$ and packet $i$ finds the second node empty it immediately goes to service, therefore, $A_{2,i} = Y_{2,i} + S_{2,i}$. At the end of interval $\Psi_{2,i-1}$ the node is empty, therefore, the probability that packet $i$ finds the node empty upon arrival equals to $1 - \rho_1$.
- **Case B2.** If $Y_{2,i} > \Psi_{2,i-1}$ and packet $i$ finds the node busy with a priority packet with probability $\rho_1$ it waits until the end of the ongoing busy period $G_1$, thus $A_{2,i} = Y_{2,i} + \tilde{G}_1 + S_{2,i}$, where $\tilde{G}_1$ denotes the residual time of interval $G_1$.
- **Case B3.** If $Y_{2,i} < \Psi_{2,i-1}$ packet $i$ finds the second node busy with non-priority packet $i-1$, therefore, $A_{2,i} = \Psi_{2,i-1} + S_{2,i}$.

The LST $\alpha_2(s)$ can be given as the sum of three LSTs defined above

$$\alpha_2(s) = \alpha_2(s, B_1) + \alpha_2(s, B_2) + \alpha_2(s, B_3). \tag{5.52}$$

**Case B1.** The LST of $Y_{2,i} + S_{2,i}$ if $Y_{2,i} > \Psi_{2,i-1}$ and the node is free of priority packets can be given as

$$\alpha_2(s, B_1) = (1 - \rho_1)\frac{\lambda_2}{\lambda_2 + s}\psi_2(s + \lambda_2)\beta_2(s). \tag{5.53}$$

**Case B2.** The LST of $Y_{2,i} + \tilde{G}_1 + S_{2,i}$ when $Y_{2,i} > \Psi_{2,i-1}$ and packet $i$ arrives during the busy period $G_1$ takes

$$\alpha_2(s, B_2) = \rho_1\frac{\lambda_2}{\lambda_2 + s}\psi_2(s + \lambda_2)\tilde{\gamma}(s)\beta_2(s), \tag{5.54}$$

where $\tilde{\gamma}(s) = (1-\gamma(s))/\mathbb{E}[G_1]s$ stands for the distribution of the residual time of the interval $G_1$.

**Case B3.** If $Y_{2,i} < \Psi_{2,i-1}$ the LST of the PAoI yields

$$\alpha_2(s, B_3) = (\psi_2(s) - \psi_2(s + \lambda_2))\beta_2(s). \tag{5.55}$$

The resulting LST of the PAoI distribution of non-priority packets gives

$$\begin{aligned}\alpha_2(s) = &\Big[(1 - \rho_1)\frac{\lambda_2}{\lambda_2 + s}\psi_2(s + \lambda_2) + \psi_2(s) - \psi_2(s + \lambda_2)\\ &+ \rho_1\frac{\lambda_2}{\lambda_2 + s}\psi_2(s + \lambda_2)\tilde{\gamma}(s)\Big]\beta_2(s).\end{aligned} \tag{5.56}$$

Having (5.12), (5.14) and (5.56) we give the LST of the AoI distribution of non-priority packets as follows:

$$\begin{aligned}\delta_2(s) = &\frac{\rho_2}{1 - \rho_1}\tau_2(s)\tilde{\beta}_2(s + \lambda_1 - \lambda_1\gamma(s))\\ &+ \psi(\lambda_2 + s)\beta_2(s)\Big(\frac{\lambda_2}{\lambda + s} + \rho_1\frac{\lambda_2}{\lambda + s}\frac{\lambda_2}{s}(1 - \tilde{\gamma}(s))\Big),\end{aligned} \tag{5.57}$$

where $\tilde{\beta}_2(s + \lambda_1 - \lambda_1\gamma(s))$ denotes the residual time of the busy period $G_2$ and equals to $(1 - \beta_2(s + \lambda_1 - \lambda_1\gamma(s)))/s\mathbb{E}[G_2]$.

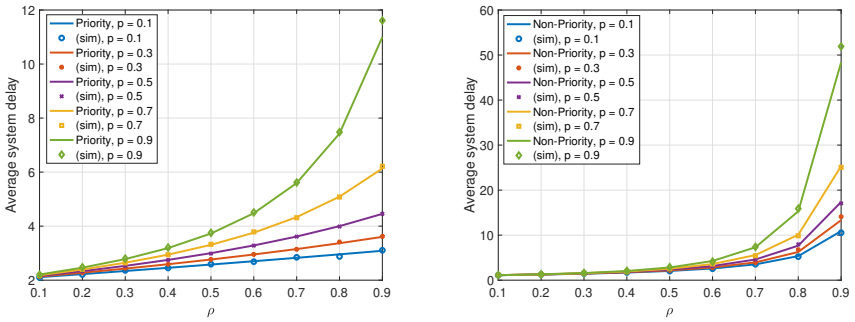
The straightforward calculation of $\alpha_2'(0)$ and $\delta_2'(0)$ gives the average PAoI $\mathbb{E}[A_2]$ and AoI $\mathbb{E}[\Delta_2]$:

$$\mathbb{E}[A_2] = b_2 + \frac{\lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}}{2(1-\rho)(1-\rho_1)} + \frac{b_1}{1-\rho_1} + \psi(\lambda_2)\frac{1}{\lambda_2} + \rho_1\psi(\lambda_2)\frac{b_2}{2(1-\rho_1)^2}. \tag{5.58}$$

$$\begin{aligned}\mathbb{E}[\Delta_2] =& \frac{\rho_2}{1-\rho_1}\Big(b_2 + \frac{\lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}}{2(1-\rho)(1-\rho_1)} + \frac{b_1}{1-\rho_1}\Big) + \\ &+ \psi(\lambda_2)\Big(\frac{1}{\lambda_2} + \frac{\rho_1\lambda_2}{2}\Big(\frac{b_1^{(3)}/b_1^{(2)}}{3(1-\rho_1)} + \frac{\lambda_1 b_1^{(2)}}{(1-\rho_1)^2}\Big)\Big) + \\ &+ \psi(\lambda_2)\Big(1 + \frac{\rho_1\rho_2}{2(1-\rho_1)^2}\Big)(b_2 + \psi'(\lambda_2)).\end{aligned} \tag{5.59}$$

### 5.3.4 Performance evaluation

The results of our analysis have been validated by Monte Carlo simulation. All data collected during the transient state has been discarded. We model arrivals, service, and departures of $10^5$ packets of the reference system. We calculate the average PAoI, AoI and system delay for different values of $p = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to capture the effect of the status updates generation rate on the AoI. The numerical results are given under the assumption of exponential service time with means $b = b_1 = b_2 = 1$. However, our results can be applied to general service time distributions, while the given closed-form expressions of the average AoI and PAoI requires the first and second order statistics of the service times. The results are displayed for variable utilization $\rho = \{0.1, \ldots, 0.9\}$ at the second node as it is the bottleneck in our system.

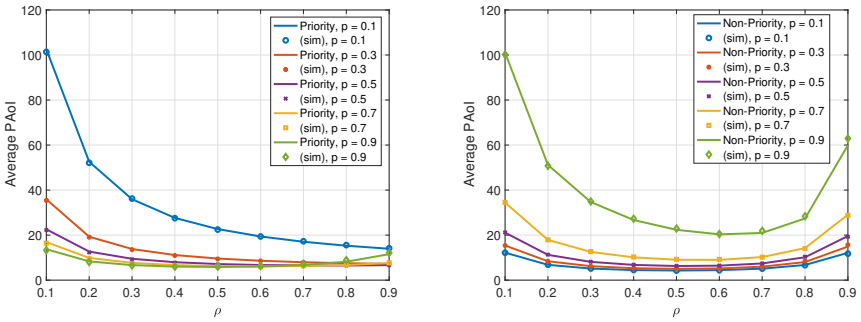



(a) priority packets (b) non-priority packets

**Fig. 5.4.** Average system delay.

First, we discuss timeliness of updates in terms of system delay illustrated in Fig. 5.4(a) and Fig. 5.4(b). The delay grows exponentially for priority and non-priority packets when the updates' arrival rate increases. It demonstrates an almost linear trend for priority packets when the arrival rate is below 50%. The high arrival rate of non-priority packets has a negligible impact on the delay of priority packets. While the priority packets instead entail a fast growth of the system delay for non-priority packets. Despite the fact that priority packets traverse through two queues while non-priority packets go through only one queue, the system delay of priority packets is much lower than that of non-priority packets. The

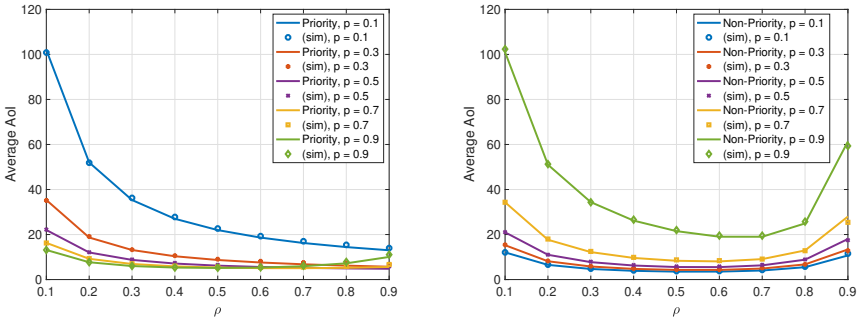
delay of non-priority packets at the second node mainly depends on the intensity of priority packets.




(a) priority packets (b) non-priority packets

**Fig. 5.5.** Average PAoI of (a) priority packets and (b) non-priority packets.

We proceed with the average PAoI analysis. Simulation results demonstrate a perfect fit with the analytical curves, as shown in Fig. 5.5(a) and Fig. 5.5(b). The average PAoI is very different from the system delay. It decreases as the arrival rate increases for priority packets meaning that the update generation rate is the leading factor of the priority packets aging. Only for the rate as high as 0.9 the average PAoI goes up. With sporadic updates (blue curve) of priority packets, the age is very high. The U shape of the average PAoI of non-priority packets tells that the system delay is a meaningful contributor to the aging, unlike priority packets. The age of non-priority packets is almost equal when the packet load is equally shared.




(a) priority packets (b) non-priority packets

**Fig. 5.6.** Average AoI of (a) priority packets and (b) non-priority packets.

The given lower bound for AoI is tight when the system utilization is low and becomes more visible when $\rho$ increases. In our system, the PAoI is a tight upper bound of the AoI due to the low correlation between interarrival and delay intervals of consecutive packets.

The results of the average AoI are presented in Fig. 5.6(a) and Fig. 5.6(b). The PAoI is a very tight upper bound of the PAoI in a given system. It is due to the interdependence of priority and non-priority flows of updates controlled by $p$ to prevent the overload of the second node that services both classes of traffic. The clear U shape of the age of non-priority packets means that the optimal performance can be reached. In Fig. 5.7 optimal values of update rates that jointly minimize the age of priority and non-priority packets are given for each value of $p$.

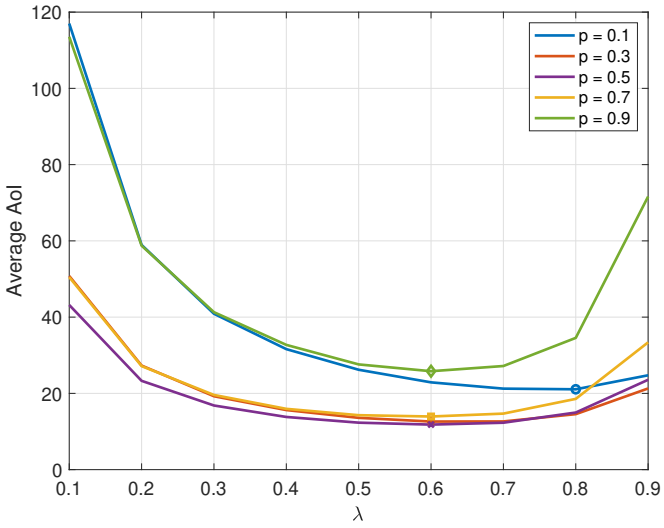



**Fig. 5.7.** Optimal arrival rate.

## 5.4 Peak Age of Information Distribution in Tandem Queues

### 5.4.1 Motivation

Being an essential metric for many real-time, safety, and critical control IoT applications, average AoI does not give an explicit knowledge of the system behavior and its bottleneck. In most age-sensitive applications, PAoI is more informative than the AoI as it describes the worst-case scenario. Knowledge of the full distribution of the PAoI is useful in communication system design and protocols optimization to guarantee reliability and timeliness of information distribution.

As mentioned in section 5.1, a relay network with at least one intermediate node with a buffer between transmitter and receiver represents better real communication scenarios and allows for advanced analysis. Even the minimal relay network that consists of only two queues in tandem, as illustrated in Fig. 5.8, can be applied in vehicular networks, satellite communications, sensor networks, blockchain technologies, and distributed learning. One of

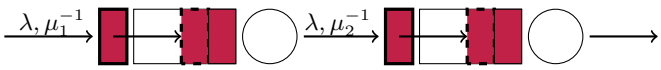



**Fig. 5.8.** System model as two queues FCFS in tandem.

the queues in tandem can model communication links, while another represents the data processing unit. In the system of interest, the age displayed at the destination is always higher than the age at the intermediate node since all updates traverse through the two nodes. However, the dynamics between them are not trivial.

A critical assumption for tandem queues analysis is the independence of service times at different nodes, and it fits well applications we have in mind. In modern vehicular, air-born, and non-terrestrial networks, a relay node that connects source and destination does not necessarily use the same radio technologies for forward and feeder links. Two queues in tandem with exponential service times is an adequate abstraction in the discussed systems [84].

Most theoretical results refer to single-node systems with different service disciplines. A few recent works focus their attention on the study of the AoI in tandem queues with Last Come First Served (LCFS) policy [85] or $M/M/1/1$ connected queues as these systems are more tractable for analysis. This line of research for a two-node system has been complemented with results in [86], and [87] for the different arrival processes. However, only first-order statistics for the AoI or PAoI have been derived. As an alternative to the average metrics, an upper bound of the quantile function of the AoI for two queues in tandem with deterministic arrivals has been obtained in a recent work [88] using the Chernoff bound technique. The complete PAoI distribution has only been derived for simple queuing systems [89].

This research gap motivates us to obtain and analyze the full distribution of the PAoI in a tandem queue consisting of two $M/M/1$ queues with independent service times and FCFS policy connected in line. The results of this analysis are useful in communication system design and defining PAoI thresholds in the network's specifications for reliability requirements.

### 5.4.2 System model

We consider two $M/M/1$ queuing systems connected in line. Packets arriving at the first system generated by a Poisson process with rate $\lambda$. Service time of the first system is exponentially distributed with rate $\mu_1$. When a packet exits the first system, it enters the second one, whose service time is an exponential r.v. with rate $\mu_2$. Both queues have infinite capacity and are unaware of the content of the packets. Therefore, packets with fresher information cannot preempt older packets, and all packets are services in an FCFS manner. As explained in the introduction, we assume that the service times in the two systems are independent; however, this is not true for the waiting times, as the queue at the second system depends on the output of the first one. In the following, we use the compact notation $p_{X|Y}(x|y)$ for the conditioned probability $p[X = x|Y = y]$.

The total system time $T_i$ of a packet $i$ in a tandem queue is the time elapsed from when a packet arrives at the first queuing system until its departure time in the second queue. As explained previously, when a packet is received, the AoI is equal to the system time of the packet $T_i$. If we denote the interarrival time $t_i - t_{i-1}$ as $Y_i$, the PAoI of packet $i$ is given by $A_i = T_i + Y_i$.

The system time $T_i$ depends on the interarrival time $Y_i$. Let us consider a simple example depicted in Fig. 5.9. On the right side, packet $i$ arrives after packet $i-1$ leaves the system. Therefore, its system time $T_i$ equals the service time $S_i$. On the contrary, on the left side, packet $i$ finds the previous packet in the system. Therefore, it has to wait until time $t'_{i-1}$, which corresponds to the waiting time $W_i$. Once packet $i-1$ leaves the system, packet $i$ can start the service.

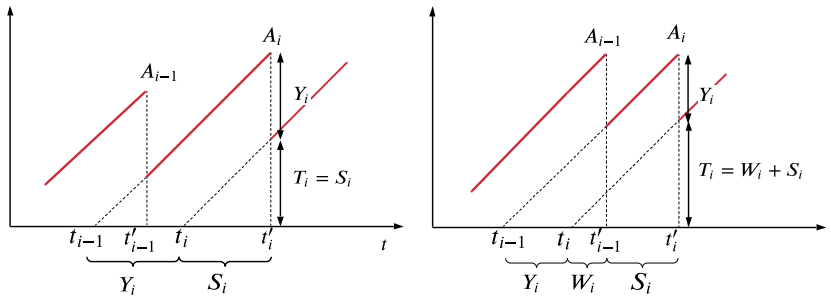



**Fig. 5.9.** Relation between packets interarrival time, system delay and PAoI.

The PDF of the PAoI $A_i$ can then be computed by using the conditional system time probability $p_{T_i|Y_i}(t_i|y_i)$:

$$p_{A_i}(\tau_i) = \int_0^{\tau_i} p_{Y_i}(y_i) p_{T_i|Y_i}(\tau_i - y_i|y_i) dy_i, \tag{5.60}$$

Therefore, we need to compute $p_{T_i|Y_i}(t_i|y_i)$. For each system $j$ in the tandem, $j = \{1, 2\}$, the system time $T_{i,j}$ is defined as the sum of the waiting time $W_{i,j}$ and the service time $S_{i,j}$. We also define $Y_{i,j}$, the interarrival time at system $j$:

$$Y_{i,j} = \begin{cases} Y_i & \text{if } j = 1; \\ Y_i + T_{i,k} - T_{i-1,k} & \text{if } j = 2. \end{cases} \tag{5.61}$$

The system times for the two queues are independent, as proven by Reich [90] using Burke's theorem [91] for exponential queues considering each system in steady state for packet $i-1$. If we consider system $j$ in steady state, then $T_{i-1,j}$ and $Y_{i-1,j}$ are independent. The system time $T_{i-1,j}$ of a $M/M/1$ system is exponentially distributed with rate $\alpha_j = \mu_j - \lambda$. However, the system is not in steady state for packet $i$ as $T_{i,j}$ and $Y_{i,j}$ are correlated.

We then define the *extended waiting time* $\Omega_{i,j}$ as the difference between the previous packet's system time and the interarrival time at the system, i.e., $\Omega_{i,j} = T_{i-1,j} - Y_{i,j}$. The reason we named $\Omega_{i,j}$ the extended waiting time is that $W_{i,j} = [\Omega_{i,j}]^+$, where $[x]^+$ is equal

to $x$ if it is positive and 0 if $x$ is negative. Knowing the PDF of the system time $T_{i-1,j}$, we can derive the PDF of the extended waiting time:

$$p_{\Omega_{i,j}|Y_{i,j}}(\omega_{i,j}|y_{i,j}) = \alpha_j e^{-\alpha_j(\omega_{i,j}+y_{i,j})} u(\omega_{i,j}+y_{i,j}), \tag{5.62}$$

where $u(\cdot)$ is the step function. The interarrival time at the first relay $Y_{i,1}$ is exponentially distributed with rate $\lambda$, while in the second system it is given by $Y_{i,2} = S_{i,1} + [-\Omega_{i,1}]^+$.

We can combine (5.62) with the definition of $Y_{i,2}$ to get:

$$p_{\Omega_{i,2}|S_{i,1},\Omega_{i,1}}(\omega_{i,2}|s_{i,1},\omega_{i,1}) = \alpha_1 e^{-\alpha_1(\omega_{i,2}+s_{i,1}+[-\omega_{i,1}]^+)} u(\omega_{i,2}+s_{i,1}+[-\omega_{i,1}]^+). \tag{5.63}$$

To compute the exact PDF of PAoI in the 2-system case ($j \in \{1,2\}$), we distinguish between free and busy systems at each node, i.e., we condition the PDF on the state of each system when packet $i$ arrives to it, and calculate it separately for the four possible combinations. An example of the relevant values in the four cases are shown in Fig. 5.10.

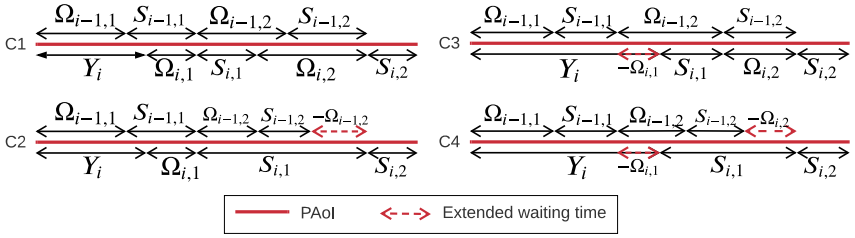



**Fig. 5.10.** Components of the PAoI.

- Case $C1 = \{\Omega_{i,1} > 0\} \cup \{\Omega_{i,2} > 0\}$. In case $C1$ packet $i$ is queued in both systems, as the previous packet is still in the system when $i$ arrives to each. The two extended queuing times (shown in red) are positive.
- Case $C2 = \{\Omega_{i,1} > 0\} \cup \{\Omega_{i,2} \leq 0\}$. In case $C2$, packet $i-1$ has already left the second system when packet $i$ leaves the first: the extended queuing time (shown in red with a dashed outline) is negative, and packet $i$ enters service in the second system as soon as it arrives.
- Case $C3 = \{\Omega_{i,1} \leq 0\} \cup \{\Omega_{i,2} > 0\}$. In case $C3$, it is the first system that is empty when packet $i$ arrives.
- Case $C4 = \{\Omega_{i,1} \leq 0\} \cup \{\Omega_{i,2} \leq 0\}$. In case $C4$ packet $i$ finds both systems empty.

The overall PDF of the PAoI is the sum of the four values in the four cases, multiplied by the probability of those cases happening, following the law of total probability:

$$p_A(\tau) = p_{A|C1}(\tau)\, p(C1) + p_{A|C2}(\tau)\, p(C2) + p_{A|C3}(\tau)\, p(C3) + p_{A|C4}(\tau)\, p(C4). \tag{5.64}$$

In case $C1$, packet $i$ is queued at both systems, and the packet will have the highest queuing delay and system time. The case with the lowest system time is case $C4$, in which the packet

experiences no queuing. However, these intuitive relations do not necessarily hold for the PAoI, as the interarrival time between update packets can play a major role. In the analysis of the four cases, we will omit the packet index $i$ wherever possible for the sake of readability. The overview of the notation used in the this section is given in Table 5.2.

### 5.4.3 Derivation of the PAoI distribution

To calculate (5.64), we first consider case $C1$, in which packet $i$ finds both systems busy, i.e., the $i$-th packet arrives before the departure of the $(i-1)$-th packet at each system. In this case, $\{\Omega_{i,1} > 0\} \cap \{\Omega_{i,2} > 0\}$. As the conditioned PDF of $\Omega_{i,j}$ was derived in the previous section, and we know that $Y_{i,1}$ is independent from $T_{i-1,1}$, as is $S_{i,1}$ from $T_{i-1,2}$, the probability of this case is denoted as $p(C1)$ and given by:

$$\begin{aligned} p(C1) &= p(\Omega_1 > 0)p(\Omega_2 > 0|\Omega_1 > 0) \\ &= p(Y_i < T_{i-1,1}, S_{i,1} < T_{i-1,1} - Y_i + W_{i-1,2} + S_{i,2}) \\ &= \int_0^\infty \int_0^{t_1} p_{y_1}(y_1) p_{T_1}(t_1) dy_1 dt_1 \int_0^\infty \int_0^{t_2} p_{S_1}(s_1) p_{T_2}(t_2) ds_1 dt_2 \\ &= \frac{\lambda}{\mu_1 + \alpha_2}. \end{aligned} \tag{5.65}$$

We start from the conditioned distribution of the system time on $\Omega_1$, $\Omega_2$, and $S_1$, so $S_2$ is the only remaining random variable:

$$p_{T|\Omega_1,\Omega_2,S_1,C1}(t|\omega_1,\omega_2,s_1) = \mu_2 e^{-\mu_2(t-\omega_1-s_1-\omega_2)} \cdot u(t-\omega_1-\omega_2-s_1). \tag{5.66}$$

We now uncondition on $\Omega_2$, and then on $S_1$, by using the law of total probability:

$$\begin{aligned} p_{T|\Omega_1,C1}(t|\omega_1) &= \int_0^{t-\omega_1} p_{S_1}(s_1) \int_0^{t-s_1-\omega_1} \frac{p_{\Omega_2|\Omega_1,S_1}(\omega_2|\omega_1,s_1)}{1 - P_{\Omega_2|\Omega_1,S_1}(0|\omega_1,s_1)} \cdot p_{T|\Omega_1,\Omega_2,S_1,C1} d\omega_2 ds_1 \\ &= \int_0^{t-\omega_1} \frac{\alpha_2}{\rho_2} \left( e^{-\alpha_2(t-\omega_1)} - e^{-\mu_2(t-\omega_1)+\lambda s_1} \right) \mu_1 e^{-\mu_1 s_1} ds_1 \\ &= \frac{\alpha_2\mu_2(\alpha_2+\mu_1)e^{-\alpha_2(t-\omega_1)}(\alpha_1 + \lambda e^{-\mu_1(t-\omega_1)} - \mu_1 e^{-\lambda(t-\omega_1)})}{\lambda\alpha_1\mu_1}, \end{aligned} \tag{5.67}$$

where the denominator in the first integral is required by the condition that we are in case $C1$. We then condition on $Y_1$ and uncondition on $\Omega_1$:

$$\begin{aligned} p_{T|Y_1,C1}(t|y_1) = \int_0^t p_{T|\Omega_1,C1}(t|\omega_1) \frac{p_{\Omega_1|Y_1}(\omega_1|y_1)}{1 - P_{\Omega_1|Y_1}(0|y_1)} d\omega_1 = \frac{\mu_2\alpha_2 e^{-\alpha_1 y_1}}{\lambda p(C1)} \cdot \\ \cdot \left( \frac{\alpha_1(e^{-\alpha_1 t} - e^{-\alpha_2 t})}{(\mu_2 - \mu_1)} + \frac{\lambda e^{-\alpha_1 t}(1 - e^{-\mu_2 t})}{\mu_2} - \frac{\mu_1(e^{-\alpha_1 t} - e^{-\mu_2 t})}{\mu_2 - \alpha_1} \right). \end{aligned} \tag{5.68}$$

We can now derive the PDF of the system time $T$:

$$\begin{aligned} p_{T|C1}(t) = \int_0^\infty p_{Y_1}(y_1) p_{T|Y_1,C1}(t|y_1) dy_1 = \frac{\mu_2\alpha_2}{\mu_1 p(C1)} \cdot \\ \cdot \left( \frac{\alpha_1(e^{-\alpha_1 t} - e^{-\alpha_2 t})}{(\mu_2 - \mu_1)} + \frac{\lambda e^{-\alpha_1 t}(1 - e^{-\mu_2 t})}{\mu_2} - \frac{\mu_1(e^{-\alpha_1 t} - e^{-\mu_2 t})}{\mu_2 - \alpha_1} \right). \end{aligned} \tag{5.69}$$

**Table 5.2.** Notations of section 5.4.

| Notation | Description |
|---|---|
| $\lambda$ | Packet generation rate |
| $\mu_j$ | Service rate of system $j$ |
| $\alpha_j = \mu_j - \lambda$ | Response rate of system $j$ |
| $S_{i,j}$ | Service time in system $j$ for packet $i$ |
| $Y_{i,j}$ | Interarrival time in system $j$ for packet $i$ |
| $\Omega_{i,j} = T_{i-1,j} - Y_{i,j}$ | Extended waiting time in $j$ for packet $i$ |
| $W_{i,j} = [\Omega_{i,j}]^+$ | Waiting time in system $j$ for packet $i$ |
| $T_{i,j} = S_{i,j} + W_{i,j}$ | Total time in system $j$ for packet $i$ |
| $A_i = Y_i + T_i$ | PAoI for packet $i$ |

Finally, we get the PDF of the PAoI, given by $T + Y_1$:

$$\begin{aligned} p_{A|C1}(\tau) =& \int_0^\tau p_{T|Y_1,C1}(t|\tau - t)\, p_{Y_1}(\tau - t)dt \\ =& \frac{\mu_1 + \alpha_2}{\lambda}\left(\frac{\alpha_2\mu_1\mu_2(e^{-\mu_1\tau} - e^{-\mu_2\tau})}{(\mu_2 - \mu_1)(\mu_2 - \alpha_1)} - \lambda e^{-\mu_1\tau}(1 - e^{-\alpha_2\tau})\right. \\ &\left. + \frac{\alpha_1\alpha_2\mu_2(e^{-\mu_1\tau} - e^{-\alpha_2\tau})}{(\mu_2 - \mu_1)(\mu_1 - \alpha_2)} + \frac{\alpha_1\mu_1\mu_2(e^{-\alpha_1\tau} - e^{-\mu_1\tau})}{(\mu_2 - \mu_1)(\mu_2 - \alpha_1)}\right). \end{aligned} \tag{5.70}$$

We now consider case $C2$, in which the first system is busy but the second one is free when packet $i$ reaches it, i.e., the packet is not queued at the second system. We have $\Omega_{i,1} > 0 \wedge \Omega_{i,2} \leq 0$, and this case happens with probability $p(C2)$:

$$\begin{aligned} p(C2) &= p(\Omega_1 > 0)p(\Omega_2 \leq 0|\Omega_1 > 0) \\ &= \int_0^\infty \int_0^{t_1} p_{y_1}(y_1)p_{T_1}(t_1)dy_1dt_1 \int_0^\infty \int_{t_2}^\infty p_{S_1}(s_1)p_{T_2}(t_2)ds_1dt_2 \\ &= \frac{\lambda\alpha_2}{\mu_1(\mu_1 + \alpha_2)}. \end{aligned} \tag{5.71}$$

In this case, the system time PDF is independent of $\Omega_2$, and we can just give the conditioned PDF as:

$$p_{T|\Omega_1,S_1,C2}(t|\omega_1, s_1) = \mu_2 e^{-\mu_2(t-\omega_1-s_1)}(1 - e^{-\alpha_2 s_1}) \cdot u(t - \omega_1 - s_1). \tag{5.72}$$

As for case $C1$, we condition on $Y_1$ and uncondition on $S_1$ and $\Omega_1$:

$$\begin{aligned} p_{T|Y_1,C2}(t|y_1) =& \frac{\mu_1(\mu_1 + \alpha_2)e^{-\alpha_1 y_1}}{\alpha_2}\left(e^{-\alpha_1 t}\left(1 - e^{-\mu_2 t}\right)\right. \\ &\left. - \frac{\alpha_2\mu_2\left(e^{-\alpha_1 t} - e^{-\mu_2 t}\right)}{(\mu_2 - \mu_1)(\mu_2 - \alpha_1)} + \frac{\alpha_1\mu_2(e^{-\alpha_1 t} - e^{-\mu_1 t})}{\lambda(\mu_2 - \mu_1)}\right). \end{aligned} \tag{5.73}$$

From this result, we derive the conditioned PDF of the system time $T$ for case $C2$:

$$p_{T|C2}(t) = \frac{\lambda e^{-\alpha_1 t}}{p(C2)}\left(1 - e^{-\mu_2 t} + \frac{\alpha_1\mu_2(1 - e^{-\lambda t})}{\lambda(\mu_2 - \mu_1)} - \frac{\alpha_2\mu_2\left(1 - e^{-(\mu_2-\alpha_1)t}\right)}{(\mu_2 - \mu_1)(\mu_2 - \alpha_1)}\right). \tag{5.74}$$

The unconditioned PDF of the PAoI for case $C2$ is given by:

$$\begin{aligned} p_{A|C2}(\tau) =& \frac{\mu_1}{p(C2)}(e^{-\alpha_1\tau} - e^{-\mu_1\tau}) - \frac{\lambda\mu_1 e^{-\mu_1\tau}(1-e^{-\alpha_2\tau})}{\alpha_2 p(C2)} \\ &+ \frac{\mu_1\mu_2\alpha_2\left((e^{-\mu_1\tau}-e^{-\alpha_1\tau})(\mu_2-\mu_1)+\lambda(e^{-\mu_1\tau}-e^{-\mu_2\tau})\right)}{(\mu_2-\mu_1)^2(\mu_2-\alpha_1)p(C2)} \\ &+ \frac{\alpha_1\mu_1\mu_2(e^{-\alpha_1\tau}-(1+\lambda\tau)e^{-\mu_1\tau})}{\lambda(\mu_2-\mu_1)p(C2)} \end{aligned} \tag{5.75}$$

We can then consider case $C3$, in which the $i$-th packet does not experience any queuing at the first system, i.e., $\Omega_1 \leq 0$, but there is queuing in the second system, i.e., $\Omega_2 > 0$. The probability of a packet experiencing case $C3$ is given by:

$$\begin{aligned} p(C3) &= p(\Omega_1 \leq 0, \Omega_2 > 0) \\ &= \int_0^\infty \int_0^{y_1} \int_0^\infty p_{y_1}(y_1) p_{T_1}(t_1) p_{S_1}(s_1) \int_{s_1-t_1+y_1}^\infty p_{T_2}(t_2) dt_2 ds_1 dt_1 dy_1 \\ &= \frac{\lambda}{\mu_2(\mu_1+\alpha_2)}. \end{aligned} \tag{5.76}$$

The conditioned PDF of the system time is:

$$p_{T|\Omega_1,\Omega_2,S_1,C3}(t|\omega_1,\omega_2,s_1) = \mu_2 e^{-\mu_2(t-s_1-\omega_2)} u(t-s_1-\omega_2). \tag{5.77}$$

As in case $C1$, we condition on $Y_1$ and uncondition on $\Omega_2$, $S_1$, and $\Omega_1$:

$$p_{T|Y_1,C3}(t|y_1) = \frac{\mu_2\alpha_2 e^{-\alpha_2 t}(e^{-\alpha_1 y_1} - e^{-\alpha_2 y_1})}{\lambda(\mu_2-\mu_1)p(C3)} \cdot \left(\alpha_1 - \mu_1 e^{-\lambda t} + \lambda e^{-\mu_1 t}\right). \tag{5.78}$$

We can now find the PDF of the system delay:

$$p_{T|C3}(t) = \frac{\alpha_2 e^{-\alpha_2 t}\left(\alpha_1 - \mu_1 e^{-\lambda t} + \lambda e^{-\mu_1 t}\right)}{\mu_1 p(C3)}. \tag{5.79}$$

The conditioned PDF of the PAoI is then:

$$\begin{aligned} p_{A|C3}(\tau) =& \frac{\mu_2}{(\mu_2-\mu_1)p(C3)}\left(\frac{\alpha_1\alpha_2(e^{-\mu_1\tau}-e^{-\alpha_2\tau})}{\alpha_2-\mu_1} + \lambda e^{-\mu_1\tau}\right. \\ &- \frac{\mu_1\alpha_2(e^{-\mu_1\tau}-e^{-\mu_2\tau})}{\mu_2-\mu_1} - \frac{\alpha_1\alpha_2(e^{-\alpha_2\tau}-e^{-\mu_2\tau})}{\lambda} \\ &\left. -\lambda e^{-(\mu_1+\alpha_2)\tau} + \alpha_2\mu_1\tau e^{-\mu_2\tau} - \frac{\lambda\alpha_2 e^{-\mu_2\tau}(1-e^{-\alpha_1\tau})}{\alpha_1}\right). \end{aligned} \tag{5.80}$$

Finally, we examine case $C4$, in which the packet experiences no queuing, i.e., $\Omega_{i,1} \leq 0 \wedge \Omega_{i,2} \leq 0$. This case happens with probability $p(C4)$:

$$p(C4) = p(\Omega_1 \leq 0, \Omega_2 \leq 0) = \frac{\alpha_1\mu_2(\mu_1+\alpha_2)) - \lambda\mu_1}{\mu_1\mu_2(\mu_1+\alpha_2)}. \tag{5.81}$$

Since the system time probability is independent of $\Omega_2$, we can just give the conditioned system time PDF as:

$$p_{T|\Omega_1,S_1,C4}(t|\omega_1,s_1) = \frac{\mu_1\mu_2 e^{-\mu_2(t-s_1)}(1-e^{-\alpha_2(s_1-\omega_1)})}{\alpha_1} \cdot u(t-s_1). \tag{5.82}$$

We then condition on $Y_1$ and uncondition on $S_1$ and $\Omega_1$:

$$\begin{aligned} p_{T|Y_1,C4}(t|y_1) =& \frac{\mu_1\mu_2}{(\mu_2-\mu_1)p(C4)}\left(e^{-\mu_1 t}(1-e^{-\alpha_1 y_1})\right. \\ &\left. -e^{-\mu_2 t}(1-e^{-\alpha_2 y_1}) + e^{-(\mu_2+\alpha_1)t}(e^{-\alpha_1 y_1} - e^{-\alpha_2 y_1})\right). \end{aligned} \tag{5.83}$$

The PDF of the system time is:

$$p_{T|C4}(t) = \frac{\mu_2(\mu_1-\lambda)e^{-\mu_1 t} - \mu_1(\mu_2-\lambda)e^{-\mu_2 t}}{\lambda(\mu_2-\mu_1)p(C4)} + \frac{\lambda e^{-(\mu_2+\alpha_1)t}}{p(C4)}. \tag{5.84}$$

We can now find the PDF of the PAoI in case $C4$:

$$\begin{aligned} p_{A|C4}(\tau) = & \frac{\mu_1\mu_2\lambda}{p(C4)}\Bigg(\frac{\tau(e^{-\mu_2\tau} - e^{-\mu_1\tau})}{\mu_2-\mu_1} \\ & + \frac{\alpha_1 e^{-\mu_2\tau} - \alpha_2 e^{-\mu_1\tau}}{\alpha_1\alpha_2(\mu_2-\mu_1)} + \frac{\left(e^{-\lambda\tau} + e^{-(\mu_1+\alpha_2)\tau}\right)}{\alpha_1\alpha_2}\Bigg). \end{aligned} \tag{5.85}$$

### Derivation of the PAoI distribution for $\mu_1 = \mu_2$

In this subsection, we consider a special case in which the general formula of the PAoI PDF is indeterminate, $\mu_1 = \mu_2$ (denoted simply as $\mu$ in the following), in which the overall system time is Erlang-distributed with shape $k = 2$ and rate $\mu - \lambda$. We follow the same steps as in the normal derivation, denoting the four cases with an apostrophe to distinguish them from the general form. The other cases in which the general formula is indeterminate, $\mu_1 = \alpha_2$ and $\mu_2 = \alpha_1$, are not derived in this paper.

In the special case $C1'$, i.e., for $\Omega_1 > 0 \wedge \Omega_2 > 0$, we have:

$$p(C1') = \frac{\lambda}{\mu+\alpha}. \tag{5.86}$$

Following the same steps as in the general case, we get:

$$\begin{aligned} p_{A|C1'}(\tau) = & \frac{e^{-\mu\tau}\left(\mu(e^{\lambda\tau}-1) + \lambda(e^{-\alpha\tau} - e^{\lambda\tau})\right)}{p(C1')} \\ & + \frac{\mu\alpha e^{-\mu\tau}\left(\alpha(1+(\tau\lambda-1)e^{\lambda\tau}) + \mu(1+\lambda\tau-e^{\lambda\tau}\right)}{\lambda^2 p(C1')}. \end{aligned} \tag{5.87}$$

The probability of being in case $C2'$, i.e., $\Omega_1 > 0 \wedge \Omega_2 \leq 0$, is:

$$p(C2') = \frac{\lambda\alpha}{\mu(\mu+\alpha)}. \tag{5.88}$$

The conditioned PDF of the PAoI is then:

$$\begin{aligned} p_{A|C2'}(\tau) = & \frac{\mu e^{-2\mu\tau}\left(\alpha e^{(\mu+\lambda)\tau} - \mu e^{\mu\tau} + \lambda e^{\lambda\tau}\right)}{\alpha p(C2')} \\ & + \frac{\mu^2 e^{-\mu\tau}}{\lambda p(C2')}\left(1+\lambda\tau-e^{\lambda\tau} - \alpha\frac{\lambda\tau(\lambda\tau+2) - 2e^{\lambda\tau}+2}{2\lambda}\right). \end{aligned} \tag{5.89}$$

Since the system is completely symmetrical and time-reversible, the probability of being in case $C3'$, i.e., $\Omega_1 \leq 0 \wedge \Omega_2 > 0$, is the same as for case $C2'$:

$$p(C3') = \frac{\alpha\lambda}{\mu(\mu+\alpha)}. \tag{5.90}$$

For the same reason, the conditioned PDF of $A$ in case $C3'$ is the same as in case $C2'$:

$$p_{A|C3'}(\tau) = p_{A|C2'}(\tau). \tag{5.91}$$

Finally, we look at case $C4'$, in which both systems are free:

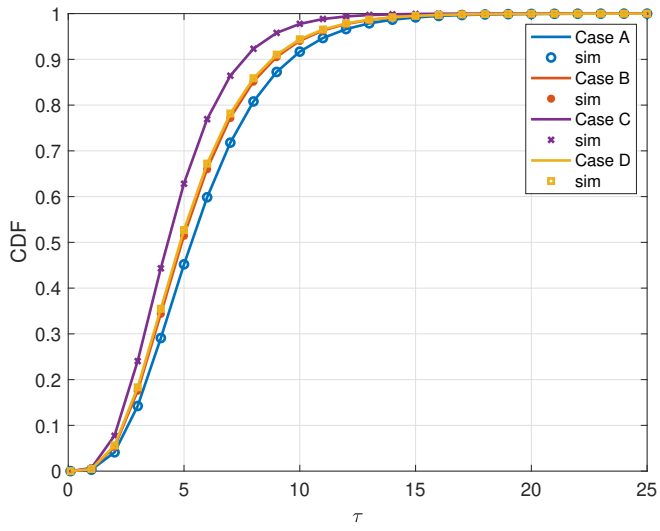



**Fig. 5.11.** CDF of the PAoI in the four cases for $\lambda = 0.5$, $\mu_1 = 1$ and $\mu_2 = 1.2$.

$$p(C4') = \frac{\alpha(\mu+\alpha)-\lambda}{\mu(\mu+\alpha)}. \tag{5.92}$$

We then have the conditioned PDF of the PAoI:

$$p_{A|C4'}(\tau) = \frac{\mu^2\lambda e^{-\mu\tau}(2\cosh(\alpha\tau)-\alpha^2\tau^2-2)}{\alpha^2 p(C4')}. \tag{5.93}$$

As in the general case, the overall PAoI is given by

$$p_A(\tau) = p(C1')p_{A|C1'}(\tau) + p(C2')p_{A|C2'}(\tau) + p(C3')p_{A|C3'}(\tau) + p(C4')p_{A|C4'}(\tau). \tag{5.94}$$

### 5.4.4 Performance evaluation

We verified the results of our analysis by Monte Carlo simulation, transmitting $10^6$ packets. The initial stages of each simulation were discarded, removing enough packets to ensure that the system had reached a steady state. We divided the packets in the four cases as illustrated in Fig. 5.11. As the derivation of the PAoI distribution does not involve any approximations, simulation results fit the theoretical curves perfectly.

Fig. 5.11 shows the PAoI CDF in the four subcases for $\lambda = 0.5$, $\mu_1 = 1$, and $\mu_2 = 1.2$. The PAoI is the lowest in case $C3$, and almost identical in cases $C2$ and $C4$. This difference is due to the effect of the interarrival times on the PAoI, as case $C4$ usually means that the instantaneous load of the system is low and packets are far apart, increasing the PAoI. In case $C3$, the faster system (system 2) is busy and the bottleneck is empty. Intuitively, this can reduce age, as the second system will probably be able to serve packets fast enough, but at the same time the instantaneous load will be high enough to avoid having a strong impact on the age.

Contrary to the system time, which increases with $\lambda$, the PAoI shows a different behavior, as Fig. 5.12 shows: the lowest PAoI is not attained at the lowest value of $\lambda = 0.25$, because

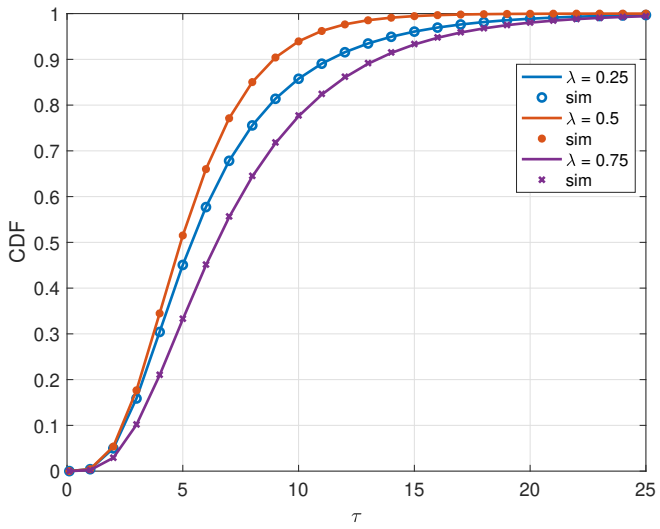



**Fig. 5.12.** CDF of the PAoI $A$ for different values of $\lambda$ with $\mu_1 = 1$ and $\mu_2 = 1.2$.

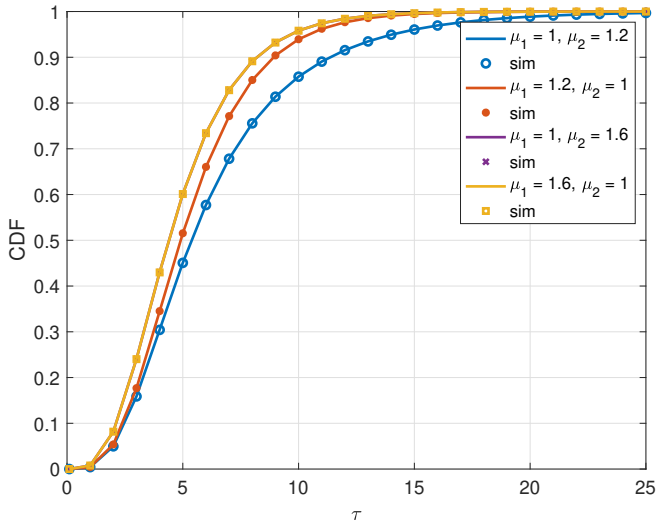



**Fig. 5.13.** CDF of the PAoI $A$ for different values of $\mu_1$ and $\mu_2$ with $\lambda = 0.5$.

for this small values of $\lambda$ the high interarrival times become the dominant factor. Instead, $\lambda = 0.5$ gives lower values, as it keeps a good balance between the frequency of the arriving packets and the length of the queues.

On the other hand, the values of $\mu_1$ and $\mu_2$ also have an important effect, as Fig. 5.13 shows: while the bottleneck always has a service rate 1, changing the service rate of the other link, and even switching the two, can have an impact on the PAoI. Naturally, increasing the rate of the other link from 1.2 to 1.6 slightly reduces the PAoI, but we note that, for both values, having the first system as the bottleneck reduces performance.

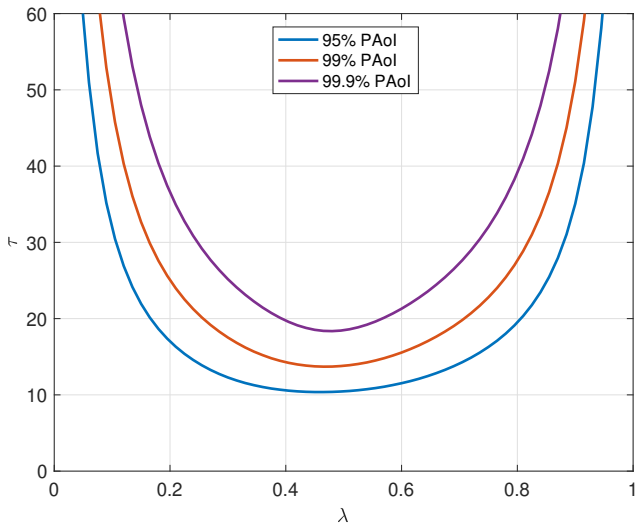



**Fig. 5.14.** Plot of different PAoI percentiles as a function of $\lambda$ with $\mu_1 = 1$ and $\mu_2 = 1.2$.

Finally, Fig. 5.14 shows how the worst-case PAoI, measured using the 95th, 99th and 99.9th percentiles, changes as a function of $\lambda$: if the traffic is very high, the queuing time is the dominant factor, causing the worst-case PAoI to diverge. The same happens if the traffic is too low, as the interarrival times can be very large: in this case, the system will almost always be empty, but updates will be very rare. The best performance in terms of PAoI is close to the middle. Depending on the desired reliability, system designers should choose the value of $\lambda$ that minimizes the given percentile of the PAoI in the specific conditions they consider.

## 5.5 Conclusions

This chapter analyzed the timeliness of information in communication systems in terms of the conventional system delay metric and more informative characteristics such as AoI and PAoI. The AoI captures the freshness of the information at the receiver side, and has been proven useful in network planning and protocol design. We focused on a two-hop communication scenario that includes different critical components of a specific tracking, status update applications, e.g., on-board computing, scheduling, and heterogeneous paths for information dissemination.

In section 5.3, we developed a versatile framework to analyze both the system delay and age-related metrics in IoT applications where the freshness of information is essential. We considered a tandem system with different paths of information updates and packet prioritization policy. Our model is indifferent to the service time distribution at the last node in line. The derived LST of the AoI, PAoI, and system delay distribution are the main contributions of this work. The high-order statistics can be directly derived from the given

transforms. We also presented closed-form expressions for the first moments of all three metrics and discussed the obtained results.

The extension to $N$ hops requires an exponential service time at the first $N-1$ hops, while the last hop that aggregates traffic from all previous nodes holds general service time distribution. Such an assumption is in line with many multi-hop systems from the reference literature. Other possible research directions are multi-source scenarios and the introduction of age-aware packet management policies.

In section 5.4, we derived the full PDF of the PAoI in a tandem system composed of two $M/M/1$ queues. This result can be used in the design of bounded AoI systems.

Potential future directions are the inclusion of error probabilities in the links and preemption-based policies, or the extension to a $M/D/1-D/M/1$ tandem, as these systems are often used to model real update applications. Finally, the optimization of $\lambda$ to minimize a given quantile of the CDF could be another interesting extension.

# 6

# Conclusions

*We conclude this thesis with a summary of the main research outcomes, we then discuss cellular IoT evolution, and present future research avenues.*

## 6.1 Summary

In this thesis, a set of mathematical frameworks and simulation tools has been developed to characterize the performance of group-based communications in cellular IoT networks. The introduced solutions were applied to evaluate the characteristics of multicast transmissions in different usage scenarios. Furthermore, specific technological enhancements for supporting delay-constrained PTM services have been presented and analyzed.

The study of this thesis has led to the following important conclusions:

- The 3GPP-based MBMS service announcement procedure does not fit the use-case scenarios of group-based communications with tight latency requirements. The examples are IoT critical updates, bug fixes, and commands.
- With optimized paging parameters and multicast transmission scheduling, the unicast-based service announcement outperforms the legacy solution, significantly reducing the service delay and device energy consumption.
- In NB-IoT systems with reduced bandwidth, the trade-off between multicast and unicast service latency can be improved with a dynamic resource allocation policy.
- The AoI in systems with different entry points for updates, e.g., schedules, can be improved by giving priority to updates with the longest path in the communication system. Moreover, the optimal operation for priority and non-priority flows can be achieved.

## 6.2 Future Research

The society as a whole, including different vertical sectors, is becoming increasingly digitalized. Providing ubiquitous wireless connectivity and integrating massive and critical use cases, cellular MTC are among the main enablers of such digitization at large.

LTE will remain the dominating RAT in mobile networks for many years to come, however, the 5G NR access technology is expected to reach almost 1.5 billion subscriptions by 2024 [11]. These networks encompass a broad range of cellular IoT use cases, giving support to massive IoT and critical IoT services. The segment of massive low-power wide-area IoT use cases is well addressed by LTE-M and NB-IoT. 3GPP standardization has ensured the tight integration of both RATs into an NR carrier.

The evolution of cellular IoT in supporting demanding sensors is highly anticipated [92]. These new type of devices transmit much higher data volumes than the conventional sensors. They will mainly address industrial IoT applications and advanced monitoring, e.g., acoustic sensing, machine vision. Another example of demanding sensors are alarms that require reliable very low latency, but still have to be cheap, simple, and powered on battery as they are expected to transmit only occasionally [93]. Such a new category of IoT devices could leverage the design principles of NR, including flexible numerology, TDD and the beam management.

### 6.2.1 On-demand sensing in Non-Terrestrial Networks

Space communications have been recently brought into the focus of 5G development to extend connectivity where the terrestrial coverage is limited, i.e., in moving platforms, remote and rural areas [94]. The need to interconnect a myriad of devices led to a new paradigm known as the Satellite Internet of Things (SIoT) that spreads the concept of the IoT to space communications [95]. For instance, LEO satellites provide (i) seamless IoT coverage during and after natural geological disasters and in remote areas (i.e., deserts, forests, and oceans), (ii) lower propagation delay than Geostationary-Earth Orbit (GEO) satellites, and (iii) robust access with different elevation angles for IoT devices thanks to the tolerance to terrestrial obstacles [96].

In many remote sensing and tracking applications, IoT devices communicate their status update to a monitor either when the device buffer is full or according to a schedule. This is the most typical push-based communication scenario for sensing applications, where devices usually have a very low duty cycle to prolong their battery life. However, in some tracking applications, the monitor may want a very recent status of a physical object. In this case, it will use a pull-based approach by paging relevant devices and forcing them to send up-to-date measurements.

In terrestrial cellular networks, the problem of device availability for paging in *on-demand status update* applications can be solved by employing WUS functionality. On the contrary, in Non-Terrestrial Networks (NTN), where satellites collect data from sensors on the ground, the communication distance between devices and the network is much longer and can negatively impact the WUS performance. Moreover, fast-moving LEO satellites have a very limited Line-of-Sight (LOS) connection with the sensing devices in a given area, therefore,

inter-satellite handover is another critical factor to take into account. The first step in this direction has been done in our work in [97].

### 6.2.2 Wireless monitoring based on Distributed Ledger Technology

Distributed wireless sensor systems are the critical enabler for many essential monitoring applications in industry, agriculture, healthcare, and urban environments. The way the sensed information is stored and collected in such systems raises concerns about data integrity, trust, security, transparency, and availability. A good example is related to the carbon dioxide emission measurements and reporting results to a regulatory body. The stakeholders clearly may have incentives to manipulate the data. The collected data can represent a mixture of private and public data, making it difficult to validate their origin and consistency. In many systems with centralized authorization, the data are vulnerable to tampering due to the man-in-the-middle attack.

Distributed Ledger Technology (DLT) is a key enabler for transparent and reliable information sharing among untrusted parties in distributed monitoring systems [98]. In DLTs, transactions containing sensing data are recorded to the ledger and then distributed among the nodes so that each node stores a copy of the ledger. An authentication process of DLT is based on consensus among all participants. Therefore, all records must be synchronized.

NB-IoT provides high downlink and uplink capacity, together with better coverage compared to other LPWAN [99]. However, the integration of DLT into the NB-IoT-based monitoring system brings an additional overhead to the resource-constrained system with a particular impact on the downlink. Maintaining a ledger in sync for all the nodes significantly increases the amount of downlink data and imposes strict requirements on the total distribution latency. This work is a natural extension of our efforts on improving group-based communications in NB-IoT systems.

### 6.2.3 AoI-based Data Collection in Fog Networks

With 5G, the cellular architecture is evolving from a BS-centric architecture to a fog-like set-up. The former is characterized by centralized processing at the edge node, while in the latter, network functionalities can be distributed more flexibly between centralized processing at the cloud and local processing at the edge. Network softwarization and fiber fronthaul links connecting edge nodes with the cloud are main enablers of a flexible Fog Radio Access Network (F-RAN) architecture [100].

Storing and distributing large amounts of sensing data in F-RAN is challenging due to the intrinsic bandwidth limitations and finite resources at the edge nodes. The achieved results in AoI can leverage in developing routing and packet management protocols to balance resource utilization and network load between edge nodes and cloud processor.

# References


1. GSMA Intelligence, "Internet of Thing by 2025," 2018, the Mobile Economy Report. [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2018/08/GSMA-IoT-infographic18-192.png
2. A. Nauman, Y. A. Qadri, M. Amjad, Y. B. Zikria, M. K. Afzal, and S. W. Kim, "Multimedia internet of things: A comprehensive survey," *IEEE Access*, vol. 8, pp. 8202–8250, 2020.
3. 3GPP, "Technical Specification Group Services and System Aspects; MBMS for IoT," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.850, 2018, version 16.0.0.
4. 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.304, 2018, version 15.0.0.
5. L. Feltrin, G. Tsoukaneri, M. Condoluci, C. Buratti, T. Mahmoodi, M. Dohler, and R. Verdone, "Narrowband IoT: A survey on downlink and uplink perspectives," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 78–86, 2019.
6. G. Tsoukaneri, M. Condoluci, T. Mahmoodi, M. Dohler, and M. K. Marina, "Group Communications in Narrowband-IoT: Architecture, Procedures, and Evaluation," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1539–1549, 2018.
7. S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2011, pp. 350–358.
8. C. Xu, H. H. Yang, X. Wang, and T. Q. S. Quek, "Optimizing Information Freshness in Computing-Enabled IoT Networks," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 971–985, 2020.
9. A. Kosta, N. Pappas, V. Angelakis *et al.*, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, vol. 12, no. 3, pp. 162–259, Nov. 2017.
10. R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Transactions on Information Theory*, 2020.

11. Ericsson, "Mobility Report," November 2019. [Online]. Available: https://wcm.ericsson.net/4acd7e/assets/local/mobility-report/documents/2019/emr-november-2019.pdf
12. ITU-R, "Minimum Requirements Related to Technical Performance for IMT 2020 Radio Interface(s)," International Telecommunication Union Radiocommunication Sector (ITU-R), Report M.2410-0, 2017.
13. Ericsson, "Cellular IoT evolution for industry digitalization," 2018. [Online]. Available: https://www.ericsson.com/en/white-papers/cellular-iot-evolution-for-industry-digitalization
14. 3GPP, "Study on Provision of Low-Cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.368, 2013, version 12.0.0.
15. 3GPP, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 45.820, 2014, version 13.1.0.
16. 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.331, 2018, version 15.0.0.
17. 3GPP, "Study on Machine-Type Communications (MTC) and Other Mobile Data Applications Communications Enhancements," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 23.887, 2016, version 12.0.0.
18. V. Group and et.al., "Revised Work Item on Low Cost and Enhanced Coverage MTC UE for LTE," 3rd Generation Partnership Project (3GPP), RAN meeting 63 RP-171427, 2014.
19. Ericsson and et.al., "Further LTE Physical Layer Enhancements for MTC," 3rd Generation Partnership Project (3GPP), RAN meeting 65 RP-171427, 2014.
20. Ericsson and et.al., "Revised WID Proposal on Further Enhanced MTC for LTE," 3rd Generation Partnership Project (3GPP), RAN meeting 75 RP-171427, 2017.
21. Ericsson and et.al., "Revised WID on Even further enhanced MTC for LTE," 3rd Generation Partnership Project (3GPP), RAN meeting 77 RP-171427, 2017.
22. G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Network*, vol. 31, pp. 80–89, 2017.
23. 3GPP, "Technical Specification Group Services and System Aspects; MBMS; Protocols and codecs," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.346, 2018, version 16.0.0.
24. GSMA, "LTE-M Deployment Guide to Basic Feature Set Requirements," Global System for Mobile Communications Association, White paper, April 2018. [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2019/08/201906-GSMA-LTE-M-Deployment-Guide-v3.pdf

25. 3GPP, "Mobile Radio Interface Layer 3 Specification; Core Network Protocols; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 24.008, 2018, version 15.0.0.
26. M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT Machine Age With 5G: Machine-Type Multicast Services for Innovative Real-Time Applications," *IEEE Access*, vol. 4, pp. 5555–5569, 2016.
27. O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, K. Samuylov, and G. Araniti, "Performance Analysis of Paging Strategies and Data Delivery Approaches for Supporting Group-Oriented IoT Traffic in 5G Networks," in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2019, pp. 1–5.
28. H. Ferng and T. Wang, "Exploring Flexibility of DRX in LTE/LTE-A: Design of Dynamic and Adjustable DRX," *IEEE Transactions on Mobile Computing*, vol. 17, no. 1, pp. 99–112, 2018.
29. S.-M. Oh, K.-R. Jung, M. Bae, and J. Shin, "Performance analysis for the battery consumption of the 3GPP NB-IoT device," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2017, pp. 981–983.
30. A. K. Sultania, P. Zand, C. Blondia, and J. Famaey, "Energy Modeling and Evaluation of NB-IoT with PSM and eDRX," in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–7.
31. S. Xu, Y. Liu, and W. Zhang, "Grouping-based discontinuous reception for massive narrowband Internet of Things systems," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1561–1571, 2018.
32. E. Kurniawan, P. H. Tan, K. Adachi, and S. Sun, "Hybrid Group Paging for Massive Machine-Type Communications in LTE Networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.
33. O. Arouk, A. Ksentini, and T. Taleb, "Group paging-based energy saving for massive MTC accesses in LTE and beyond networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1086–1102, 2016.
34. 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.213, 2019, version 15.7.0.
35. C. Wei, R. Cheng, and S. Tsao, "Performance Analysis of Group Paging for Machine-Type Communications in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3371–3382, 2013.
36. V. Savaux, A. Kountouris, Y. Louët, and C. Moy, "Modeling of Time and Frequency Random Access Network and Throughput Capacity Analysis," *EAI Endorsed Transactions on Cognitive Communications*, vol. 3, no. 11, p. e2, 2017.
37. R. Ratasuk, N. Mangalvedhe, D. Bhatoolaul, and A. Ghosh, "LTE-M Evolution Towards 5G Massive MTC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–6.

38. O. Liberg, M. Sundberg, Y.-P. E. Wang, J. Bergman, and J. Sach, Eds., *Cellular Internet of things: technologies, standards, and performance*. Academic Pres, 2018.

39. 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.211, 2018, version 15.0.0.

40. M. Lauridsen, "Studies on mobile terminal energy consumption for LTE and future 5G," PhD thesis, Aalborg University, January 2015.

41. N. Kouzayha, Z. Dawy, J. G. Andrews, and H. ElSawy, "Joint downlink/uplink RF wake-up solution for IoT over cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1574–1588, 2017.

42. S. Rostami, K. Heiska, O. Puchko, K. Leppanen, and M. Valkama, "Robust pre-grant signaling for energy-efficient 5g and beyond mobile devices," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

43. 3GPP, "System architecture for the 5G System (5GS)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 33.501, 2019, version 16.0.0.

44. R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 240–254, 2012.

45. J. Liu, W. Chen, Z. Cao, and K. B. Letaief, "Dynamic power and sub-carrier allocation for OFDMA-based wireless multicast systems," in *2008 IEEE International Conference on Communications*. IEEE, 2008, pp. 2607–2611.

46. T.-P. Low, M.-O. Pun, Y.-W. P. Hong, and C.-C. J. Kuo, "Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 2, pp. 791–801, 2010.

47. F. Hou, L. X. Cai, P.-H. Ho, X. Shen, and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1508–1519, 2009.

48. C. Tan, T. C. Chuah, and S. Tan, "Adaptive multicast scheme for OFDMA-based multicast wireless systems," *Electronics Letters*, vol. 47, no. 9, pp. 570–572, 2011.

49. D. Striccoli, G. Piro, and G. Boggia, "Multicast and broadcast services over mobile networks: A survey on standardized approaches and scientific outcomes," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1020–1063, 2018.

50. J. Shen, N. Yi, B. Wu, W. Jiang, and H. Xiang, "A greedy-based resource allocation algorithm for multicast and unicast services in OFDM system," in *2009 International Conference on Wireless Communications Signal Processing*, 2009, pp. 1–5.

51. H. Deng, X. Tao, and J. Lu, "QoS-Aware Resource Allocation for Mixed Multicast and Unicast Traffic in OFDMA Networks," in *2011 IEEE Vehicular Technology Conference (VTC Fall)*, 2011, pp. 1–5.

52. J. Chen, M. Chiang, J. Erman, G. Li, K. K. Ramakrishnan, and R. K. Sinha, "Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and

dynamics," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1266–1274.

53. Y. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A Primer on 3GPP Narrowband Internet of Things," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 117–123, 2017.
54. ZTE, "Further considerations on NB-PDSCH design for NB-IoT," 3GPP TSG RAN1 NB-IoT Ad-Hoc meeting #2, White paper R1 - 161860, March 2016, written contribution.
55. O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of things: technologies, standards, and performance*. Academic Press, 2019.
56. L. Kleinrock, "Time-shared systems: A theoretical treatment," *Journal of the ACM (JACM)*, vol. 14, no. 2, pp. 242–261, 1967.
57. R. Pyke and R. Schaufele, "Limit theorems for Markov renewal processes," *The Annals of Mathematical Statistics*, pp. 1746–1764, 1964.
58. R. Pyke and R. Schaufele, "The existence and uniqueness of stationary measures for Markov renewal processes," *The Annals of Mathematical Statistics*, pp. 1439–1462, 1966.
59. L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 1681–1685.
60. M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the Role of Age of Information in the Internet of Things," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 72–77, 2019.
61. S. Kaul, R. Yates, and M. Gruteser, "On piggybacking in vehicular networks," in *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011.* IEEE, 2011, pp. 1–5.
62. P. Dong, Z. Ning, M. S. Obaidat, X. Jiang, Y. Guo, X. Hu, B. Hu, and B. Sadoun, "Edge computing based healthcare systems: Enabling decentralized health monitoring in internet of medical things," *IEEE Network*, 2020.
63. Y. Gu, H. Chen, Y. Zhou, Y. Li, and B. Vucetic, "Timely status update in internet of things monitoring systems: An age-energy tradeoff," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5324–5335, 2019.
64. C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Information freshness and popularity in mobile caching," in *2017 IEEE International Symposium on Information Theory (ISIT).* IEEE, 2017, pp. 136–140.
65. J. Zhong, R. D. Yates, and E. Soljanin, "Two freshness metrics for local cache refresh," in *2018 IEEE International Symposium on Information Theory (ISIT).* IEEE, 2018, pp. 1924–1928.
66. Q. He, G. Dan, and V. Fodor, "Minimizing age of correlated information for wireless camera networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* IEEE, 2018, pp. 547–552.

67. Q. He, G. Dán, and V. Fodor, "Joint assignment and scheduling for minimizing age of correlated information," *IEEE/ACM Transactions on Networking*, vol. 27, no. 5, pp. 1887–1900, 2019.

68. A. E. Kalør and P. Popovski, "Minimizing the age of information from sensors with common observations," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1390–1393, 2019.

69. I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 844–851.

70. I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2637–2650, 2018.

71. Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Transactions on Mobile Computing*, 2019.

72. S. Nath, J. Wu, and J. Yang, "Optimum energy efficiency and age-of-information tradeoff in multicast scheduling," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

73. J. Li, Y. Zhou, and H. Chen, "Age of information for multicast transmission with fixed and random deadlines in IoT systems," *IEEE Internet of Things Journal*, Mar. 2020.

74. M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the role of age of information in the internet of things," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 72–77, 2019.

75. G. Stamatakis, N. Pappas, and A. Traganitis, "Optimal Policies for Status Update Generation in an IoT Device With Heterogeneous Traffic," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5315–5328, 2020.

76. M. Bacco, P. Cassarà, M. Colucci, and A. Gotta, "Modeling Reliable M2M/IoT Traffic Over Random Access Satellite Links in Non-Saturated Conditions," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 1042–1051, 2018.

77. B. Soret, S. Ravikanti, and P. Popovski, "Latency and timeliness in multi-hop satellite networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

78. Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "A General Formula for the Stationary Distribution of the Age of Information and Its Application to Single-Server Queues," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8305–8324, 2019.

79. B. Li, H. Chen, Y. Zhou, and Y. Li, "Age-Oriented Opportunistic Relaying in Cooperative Status Update Systems with Stochastic Arrivals," 2020.

80. Q. Kuang, J. Gong, X. Chen, and X. Ma, "Age of Information for Computation Intensive Messages in Mobile Edge Computing," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.

81. S. K. Kaul and R. D. Yates, "Age of Information: Updates with Priority," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2644–2648.

82. J. Xu and N. Gautam, "Peak Age of Information in Priority Queueing Systems," *arXiv: Information Theory*, 2020.

83. R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of scheduling*. Reading, Massachusetts: Addison-Wesley, 1967.

84. Y. Zhu, M. Sheng, J. Li, D. Zhou, and Z. Han, "Modeling and performance analysis for satellite data relay networks using two-dimensional markov-modulated process," *IEEE Transactions on Wireless Communications*, Mar. 2020.

85. A. M. Bedewy, Y. Sun, and N. B. Shroff, "Age-optimal information updates in multihop networks," in *International Symposium on Information Theory (ISIT)*. IEEE, Jun. 2017, pp. 576–580.

86. R. D. Yates, "Age of information in a network of preemptive servers," in *Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Apr. 2018, pp. 118–123.

87. C. Kam, J. P. Molnar, and S. Kompella, "Age of information for queues in tandem," in *Military Communications Conference (MILCOM)*. IEEE, Oct. 2018, pp. 1–6.

88. J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for AoI in single-hop and two-hop systems with periodic arrivals," *arXiv preprint arXiv:1910.09949*, Oct. 2019.

89. R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 721–734, Feb. 2019.

90. E. Reich, "Note on queues in tandem," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 338–341, Mar. 1963.

91. P. J. Burke, "The output of a queuing system," *Operations Research*, vol. 4, no. 6, pp. 699–704, Dec. 1956.

92. Ericsson, "New SID on NR MTC for industrial sensors," 3rd Generation Partnership Project (3GPP), RAN meeting 83 RP-190432, 2019.

93. Ericsson, "Motivation for new SID on NR MTC for industrial sensors," 3rd Generation Partnership Project (3GPP), RAN meeting 83 RP-190433, 2019.

94. 3GPP, "Technical Specification Group Radio Access Network, Study on New Radio (NR) to support non terrestrial networks," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.811, 2020, version 15.3.0.

95. I. F. Akyildiz and A. Kak, "The Internet of Space Things/CubeSats: A ubiquitous cyber-physical system for the connected world," *Computer Networks*, vol. 150, pp. 134–149, 2019.

96. Z. Zhang, Y. Li, C. Huang, Q. Guo, L. Liu, C. Yuen, and Y. L. Guan, "User activity detection and channel estimation for grant-free random access in leo satellite-enabled internet-of-things," *IEEE Internet of Things Journal*, 2020.

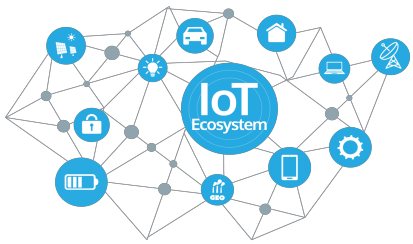
97. F. Rinaldi, S. Pizzi, O. Vikhrova, A. Molinaro, and G. Araniti, "Paging IoT Devices in 5G-Enabled Non-Terrestrial Networks," in *71$^{st}$ International Astronautical Congress (IAC) – The Cyber Space Edition, 12-14 October*, 2020.

98. P. Danzi, A. E. Kalor, R. B. Sorensen, A. K. Hagelskjær, L. D. Nguyen, C. Stefanovic, and P. Popovski, "Communication aspects of the integration of wireless iot devices with distributed ledger technology," *IEEE Network*, vol. 34, no. 1, pp. 47–53, 2020.

99. K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "A comparative study of lpwan technologies for large-scale iot deployment," *ICT express*, vol. 5, no. 1, pp. 1–7, 2019.

100. M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *Ieee Network*, vol. 30, no. 4, pp. 46–53, 2016.

DOTTORATO di RICERCA in INGEGNERIA dell'INFORMAZIONE
SCUOLA di DOTTORATO
Università degli Studi *Mediterranea* di Reggio Calabria




The present work is devoted to characterizing state-of-the-art technologies to enable PTM connectivity in cellular IoT networks, identify their limitations, and contribute to their performance assessment and enhancements. We mainly focus on the group-based communications and their applications in supporting massive and critical IoT communications. The main contributions presented in this work include: (i) a novel PTM transmission framework for supporting critical services in cellular IoT systems; (ii) radio resource management techniques for servicing unicast and multicast MTC-users with heterogeneous service requirements; (iii) a set of analytical models of the group-based communications and methods for their key performance metrics evaluation; (iv) advanced analysis of the timeliness of information for the status update in cellular IoT application scenarios. Our study indicates the need for advanced and adaptive techniques for efficient information distribution in highly autonomous IoT systems with diverse requirements and capabilities.

ISBN 978-88-99352-46-2


COLLANA DELLA SCUOLA DI DOTTORATO DELL'UNIVERSITA' DEGLI STUDI *MEDITERRANEA* DI REGGIO CALABRIA