UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Elço João dos Santos Junior

Wireless RAN slicing of heterogeneous services for 5G and beyond systems: channel allocation and multiple access solutions

Florianópolis
2023

Elço João dos Santos Junior

Wireless RAN slicing of heterogeneous services for 5G and beyond systems: channel allocation and multiple access solutions

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina para obtenção do título de Doutor em Engenharia Elétrica.
Orientador Prof. Richard Demo Souza, Dr.
Coorientador Prof. João Luiz Rebelatto, Dr.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

dos Santos Junior, Elço João

Wireless RAN slicing of heterogeneous services for 5G and beyond systems : channel allocation and multiple access solutions / Elço João dos Santos Junior ; orientador, Richard Demo Souza, coorientador, João Luiz Rebelatto, 2023.

100 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis, 2023.

Inclui referências.

1. Engenharia Elétrica. 2. 5G and beyond. 3. Heterogeneous users. 4. Channel allocation. 5. Non orthogonal multiple access. I. Demo Souza, Richard. II. Rebelatto, João Luiz. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título.

Elço João dos Santos Junior

Wireless RAN slicing of heterogeneous services for 5G and beyond systems: channel allocation and multiple access solutions

O presente trabalho em nível de Doutorado foi avaliado e aprovado, em 23/06/2023, por banca examinadora composta pelos seguintes membros:

Profa. Victoria Dala Pegorara Souto, Dra.
Instituto Nacional de Telecomunicações

Prof. Glauber Brante, Dr.
Universidade Tecnológica Federal do Paraná

Prof. Marcelo Eduardo Pellenz, Dr.
Pontifícia Universidade Católica do Paraná

Certificamos que esta é a **versão original e final** do trabalho que foi julgado adequado para a defesa de doutorado.

ICP-Edu
Documento assinado digitalmente
Telles Brunelli Lazzarin
Data: 27/07/2023 11:08:58-0300
CPF: ***.400.719-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Telles Brunelli Lazzarin, Dr.
Coordenação do Programa de
Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Richard Demo Souza
Data: 24/07/2023 09:38:50-0300
CPF: ***.267.379-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Richard Demo Souza, Dr.
Orientador

Florianópolis, 2023.

Esse trabalho é dedicado aos meus pais, Elço e Maria,
e à minha esposa, Ana Paula.

AGRADECIMENTOS

Agradeço aos meus pais, Elço e Maria, e às minhas irmãs, Marilene e Arlene, pelo carinho e apoio;

à minha esposa, Ana Paula, pelo amor, compreensão e por me ajudar a entender que desafios nada mais são do que oportunidades de crescimento;

aos meus sobrinhos, Ana Júlia e João Pedro, pela constante reafirmação de que a vida pode ser muito mais simples do que enxergamos. Espero ser exemplo de que é possível realizar seus sonhos e se tornar o que quiser;

ao meu orientador, Doutor Richard Demo Souza, por ser, além do melhor professor e orientador que eu poderia ter, um amigo que sempre pude contar;

ao meu co-orientador, Doutor João Luiz Rebelatto, por todo suporte, discussões e contribuições durante o desenvolvimento da tese;

e a todos os pesquisadores e professores que pavimentaram o caminho usado como base para desenvolvimento desse trabalho.

RESUMO

Este trabalho considera o problema de compartilhamento de recursos de rádio entre usuários *enhanced Mobile Broadband* (eMBB) e *Ultra-Reliable and Low Latency Communications* (URLLC) conectados a uma mesma *Base Station* (BS) no escopo de 5G e além (B5G). Inicialmente, visando aumentar a confiabilidade do serviço eMBB, foi proposto o uso do método *Max-Matching Diversity* (MMD) para realizar a alocação de canais, considerando a divisão de recursos com base nos métodos de Multiple Acesso Ortogonal (OMA), Multiple Acesso Não Ortogonal (NOMA) e Multiple Acesso Híbrido (HMA). Os resultados indicam que a aplicação de MMD resulta em um aumento na taxa dados dos usuários do eMBB e na confiabilidade do serviço URLLC simultaneamente para todos os métodos de acesso múltiplo adotados. Além disso, mostra-se que o método HMA apresenta resultados superiores quando comparado ao OMA e NOMA. Posteriormente, foi utilizado o mesmo modelo de sistema, mas considerando múltiplos usuários URLLC que compartilham recursos entre si através do método *Rate-Splitting Multiple Access* (RSMA), onde um dispositivo URLLC divide sua transmissão em duas submensagens com potência parcial, que são potencialmente recuperadas na BS por *Successive Interference Cancellation* (SIC). Para estudar o desempenho desse método na presença de um usuário eMBB, foram consideradas as abordagens de fatiamento de rede OMA e NOMA. Como resultado, foi demonstrado que, em geral, o RSMA melhorou o desempenho em termos de taxa e confiabilidade, mesmo quando transmitindo simultaneamente com um usuário eMBB. Os resultados também mostraram que a taxa do URLLC pode ser aumentada ajustando adequadamente o fator de divisão de taxa com base na *Signal to Noise Ratio* (SNR) média, não exigindo *Channel State Information* (CSI). Para finalizar a tese, aplicamos o método MMD para atribuir usuários *Grant-Based* (GB) a canais, sendo que no mesmo recurso de rede que atende um usuário GB, um usuário *Grant-Free* (GF) é alocado por meio de um protocolo de contenção distribuído que considera o *Quality-of-Service* (QoS) do usuário GB, realizando o pareamento que cause o mínimo de interferência, para que a presença de um usuário GF seja transparente para o usuário GB. Ambos os usuários admitidos no mesmo recurso físico operam em NOMA ou RSMA, dependendo dos ganhos do canal. Expressões exatas para a probabilidade de *outage* deste sistema *Semi-Grant-Free* (SGF) são fornecidas e comparadas com resultados de simulação, mostrando que a estratégia de transmissão SGF com usuários GB auxiliados por MMD reduz a *outage* e aumenta a taxa alcançável de ambos os usuários.

Palavras-chave: 5G e além. Usuários heterogêneos. Alocação de canal. RSMA.

RESUMO EXPANDIDO

Introdução

Nos últimos anos, houve um crescimento acelerado no tráfego de dados móveis. Em 2021, esse valor chegou a 67 Exabytes por mês, com um aumento estimado para 282 Exabytes em 2027 (ERICSSON, 2023). Essa alta se dá devido a diversos fatores, tais como: aumento no número de dispositivos conectados; avanço/criação de tecnologias de comunicação; e surgimento de novas aplicações, como a *Internet of Things* (IoT). Com isso, o avanço das tecnologias de redes celulares se fez necessário, surgindo então a quinta geração, denominada 5G. Quando comparado às gerações anteriores, a 5G visa não somente o aumento da taxa de dados, mas também comunicação de baixa latência e alta confiabilidade, além de um grande número de nós conectados simultaneamente (POPOVSKI, Petar *et al.*, 2018). Estes requisitos, muitas vezes conflitantes, dão origem a três diferentes serviços, denominados *Ultra-Reliable and Low Latency Communication* (URLLC), *massive Machine-Type Communication* (mMTC) e *enhanced Mobile Broadband* (eMBB) (SHAFI *et al.*, 2017). Cada um dos serviços apresenta grandes desafios para implementação. No cenário URLLC, serviço não existente nas gerações anteriores da tecnologia, é requerida baixa latência de comunicação e alta confiabilidade (POPOVSKI, P., 2014). No mMTC, deseja-se permitir que vários dispositivos acessem o canal e transmitam mensagens de forma concorrente. Para o serviço eMBB, espera-se altas e moderadas taxas de dados. No caminho para os sistemas de comunicação sem fio 5G e além (B5G), é razoável supor que estes três serviços heterogêneos do 5G possam ser adaptados para suportar novos sub-serviços ou até mesmo que eles sejam combinados, dando origem a novas classes de serviços (FLAGSHIP, 2019; SAAD *et al.*, 2020; NGUYEN *et al.*, 2022; TARIQ *et al.*, 2020). Assim, um requisito desafiador dos sistemas B5G é oferecer suporte a novos casos de uso heterogêneos, como indústria 4.0 e robótica conectada, telepresença holográfica, etc. (FLAGSHIP, 2019; SAAD *et al.*, 2020; GIORDANI *et al.*, 2020; TARIQ *et al.*, 2020), além de fornecer os serviços atuais.

Ao analisar o fatiamento da *Radio Access Network* (RAN), geralmente o método de acesso múltiplo é ortogonal, ou seja, os recursos no domínio do tempo, frequência, potência, código, entre outros, são alocados ortogonalmente, diminuindo a interferência entre usuários/serviços (DAI *et al.*, 2015). A versão atual do 5G, *Release* 17, adotou o método *Orthogonal Frequency-Division Multiple Access* (OFDMA) para transmissões de *uplink* e *downlink*, seguindo o padrão 4G (SPECTRUM, 2021). No entanto, para atender aos requisitos dessas novas aplicações, a coexistência de serviços na mesma infraestrutura de rede requer métodos de acesso múltiplo mais robustos que combinem maior eficiência espectral com atraso limitado, alta confiabilidade e alto número de dispositivos conectados. No *Release* 17, as principais melhorias implementadas para o eMBB foram baseadas em múltiplas antenas, *multiple Transmission and Reception Point* (mTRP) e expansão da banda FR2 de 52,6 GHz até 71 GHz (ERICSSON, 2022; NEW, 2022). Porém, essas são soluções caras em termos de hardware/licença do espectro e recursos computacionais. Alguns trabalhos como (BAI *et al.*, 2010; SANTOS *et al.*, 2020) apresentaram métodos para melhorar a capacidade do eMBB através de software e podem ser explorados. Em aplicações como IoT, onde vários sensores e inúmeros dispositivos conectados são esperados, e também quando o foco está no *trade-off* entre baixa latência e confiabilidade satisfatória para o serviço URLLC, o método *Grant-Free* (GF) foi considerado uma boa solução (AZARI *et al.*, 2017; LIU, L. *et al.*, 2018; SAMAD *et al.*, 2019; KASSAB *et al.*, 2022). A abordagem GF permite transmissões sem longos protocolos de *handshake* para conceder acesso aos dispositivos (BAYESTEH *et al.*, 2014), em contraste com o acesso *Grant-Based* (GB), que depende de mensagens piloto para estimar o canal dos usuários e permite a troca de informações quando as condições são favoráveis para atingir a taxa alvo sob uma determinada probabilidade de erro aceitável. Em outras palavras, os usuários GF podem transmitir mensagens sempre que tiverem novos dados para enviar, sem solicitar permissão à *Base Station* (BS).

Em conjunto com a redução da latência, visando melhorar a eficiência espectral, alguns métodos não-ortogonais têm sido propostos nos últimos anos para substituir o Orthogonal Multiple Access (OMA), tais como *Non-Orthogonal Multiple Access* (NOMA), *Rate-Splitting Multiple Access* (RSMA) e *Hybrid Multiple Acces* (HMA). Desta forma, os usuários não precisam esperar para serem atendidos com blocos ortogonais disponíveis, resultando em uma redução na latência experimentada pelos usuários. Este é um recurso útil para suportar a comunicação URLLC (CHEN, H. *et al.*, 2018). Para o mMTC, o maior desafio é suportar a conectividade em massa considerando a escassez de recursos em frequência, para o qual métodos não-ortogonais também são uma solução interessante, pois permite que os usuários compartilhem seus recursos em vez de ocupá-los sozinhos (MA *et al.*, 2018). Embora a filosofia não-ortogonal permita que os dispositivos compartilhem os mesmos recursos de canal em um bloco de tempo com poucos erros de colisão (LIU, Yuanwei *et al.*, 2017), como não há controle centralizado para limitar o número de usuários, os protocolos GF tornam-se ineficazes quando há muitos usuários/dispositivos ativos, o que é um cenário provável para URLLC e mMTC. Assim, as frequentes situações de colisão levam à incapacidade de detectar múltiplos usuários, representando o desafio mais crucial para a estratégia de transmissão GF (LIU, L. *et al.*, 2018). Um método para lidar com o desafio mencionado é empregar um protocolo de contenção, já que vários usuários podem escolher o mesmo canal para transmitir ao mesmo tempo, o que é conhecido como esquemas de transmissão *Semi-Grant-Free* (SGF), que podem ser vistos como um meio-termo entre GF e esquemas GB convencionais (DING *et al.*, 2019; ZHANG, C. *et al.*, 2021). Em um cenário com usuários heterogêneos, por exemplo, todos os canais da rede estão disponíveis para transmissão GF, independentemente de terem sido reservados por usuários de GB. No entanto, para garantir que os requisitos de *Quality-of-Service* (QoS) dos usuários GB sejam atendidos, a contenção entre os usuários GF deve ser cuidadosamente gerenciada, tal que os usuários GF transmitam apenas quando não causem degradação significativa no desempenho dos usuários do GB.

Objetivos

O objetivo do trabalho é estudar técnicas que viabilizem a coexistência de serviços heterogêneos em uma mesma infraestrutura de rede, focando no ponto de vista da camada física. Isto é, comparar, via análises matemáticas e simulações computacionais, estratégias potenciais a serem aplicadas na *divisão* e *alocação* dos recursos de rádio de modo que os requisitos de confiabilidade, latência e taxa de dados sejam alcançados, focando nas mensagens de *uplink*.

Metodologia

Foi desenvolvido um método matemático para realizar a análise do desempenho em termos da capacidade do canal de comunicação sem fio, para um sistema composto por usuários eMBB/GB e URLLC/GF, dispostos em uma única célula de uma rede B5G, considerando somente as mensagens de *uplink*. A modelagem destas mensagens, por não existir coordenação entre os usuários da rede, se torna uma tarefa mais complexa e relevante do que a considerada no caso das mensagens de *downlink*. Através deste modelo, foram realizadas simulações computacionais e os resultados foram avaliados.
Primeiramente, consideramos a multiplexação de usuários eMBB e URLLC com uma estratégia híbrida de divisão de recursos, focando na melhoria da eficiência espectral. O método de alocação *Max-Matching Diversity* (MMD) (BAI *et al.*, 2010) é usado para alocar canais para usuários eMBB. Os resultados são comparados com os resultados de (SANTOS *et al.*, 2020).
Na sequência, visando aumentar a eficiência espectral do URLLC, permitimos que mais de um usuário URLLC realize compartilhamento não-ortogonal de recursos de frequência e tempo por meio de *rate-splitting*, o que chamamos de U-RSMA. No esquema proposto, combinamos os benefícios da decodificação RSMA, SIC e diversidade de frequência, tanto no fatiamento de rede OMA quanto no NOMA com usuários eMBB. O esquema U-RSMA proposto é então comparado com os chamados esquemas U-NOMA e U-OMA, onde o acesso múltiplo entre dispositivos URLLC é realizado por meio de NOMA e OMA, respectivamente.
Para finalizar a tese, estudamos um sistema de transmissão SGF. Usamos a aborda-

gem MMD para atribuir usuários GB a canais. Um usuário GF é atribuído ao mesmo recurso de rede que um usuário GB por meio de um protocolo de contenção distribuído que considera a QoS do usuário GB e os emparelha de maneira que cause a menor interferência possível, tornando a presença de um usuário GF quase transparente para o usuário GB. Dependendo dos ganhos do canal, ambos os usuários admitidos no mesmo recurso físico operam em NOMA ou RSMA. Esse método foi nomeado RSMA-MMD-SGF. Expressões exatas para a probabilidade de erro deste sistema SGF são fornecidas e comparadas com resultados de simulação.

Resultados e Discussão

Em um primeiro momento, ao se comparar o desempenho dos métodos ortogonal, não-ortogonal e híbrido, pôde-se concluir que o último é capaz de obter taxas de dados maiores quando $\bar{\gamma}_U > \bar{\gamma}_B$, onde $\bar{\gamma}_U$ e $\bar{\gamma}_B$ são as *Signal to Noise Ratios* (SNRs) médias do URLLC e do eMBB, respectivamente. Por outro lado, quando $\bar{\gamma}_B > \bar{\gamma}_U$, o método híbrido iguala os resultados do método ortogonal na maior parte dos valores de SNR simulados. Como a interferência gerada pelos dispositivos eMBB é maior neste caso, apenas a partição OMA no esquema HMA dará um resultado significativo na taxa, portanto, o resultado obtido é equivalente ao OMA. Isso mostra que o método HMA consegue extrair o melhor dos métodos OMA e NOMA, pois independente da relação das SNRs, obtêm resultados melhores ou pelo menos iguais aos métodos originais. Além disso, ao realizar as simulações para valores de confiabilidade URLLC mais restritos, o HMA mostrou que pode alcançar taxas superiores. Esse resultado foi possível devido à natureza do método híbrido, que consegue adaptar a divisão dos recursos conforme a necessidade de cada serviço em um determinado momento. O segundo cenário avaliado no trabalho, onde os recursos de rádio são compartilhados entre um usuário eMBB e múltiplos usuários URLLC, mostra que o método U-RSMA é capaz de atingir taxas mais altas quando o fator de divisão de potência é configurado corretamente, mesmo com requisitos de confiabilidade mais elevados. Além disso, é possível observar que a divisão de recursos não-ortogonal entre usuários heterogêneos é capaz de atingir o maior par de taxas para URLLC e eMBB simultaneamente. Isso nos leva a outro cenário interessante, em que a combinação de U-RSMA e NOMA é uma ferramenta poderosa para atender às demandas do B5G. Além disso, pôde-se concluir que U-OMA precisa de mais largura de banda para superar outros métodos, o que é um fator limitante. Nesse caso, U-RSMA é a melhor escolha para segmentos menores de espectro, resultando em maior eficiência. Para finalizar a tese, os resultados do método RSMA-MMD-SGF foram apresentados, nos levando à conclusão de que é possível aumentar o desempenho dos usuários de GB e manter a probabilidade de erro do GF simultaneamente, o que não ocorre nos resultados da literatura. Também

foi possível notar que a abordagem RSMA-MMD-SGF é capaz de atingir taxas-alvo mais altas para usuários de GF, mantendo sua confiabilidade controlada. Ademais, se fixarmos a taxa alvo do GB, a abordagem proposta é capaz de aumentar o desempenho do GF em termos de taxa, aproveitando os pontos alcançáveis de capacidade pelo RSMA e a melhor condição do canal experimentada pelos usuários GB auxiliados por MMD

Considerações Finais

Consideramos, inicialmente, a divisão de recursos de uma rede B5G composta por usuários eMBB e URLLC. Foi proposta a abordagem HMA, onde uma alocação híbrida de recursos pode ser realizada, com foco em extrair as vantagens dos métodos OMA e NOMA. Com esta técnica, foi possível obter taxas maiores para eMBB quando comparadas às estratégias ortogonal e não-ortogonal. Na segunda parte do trabalho, consideramos o problema de divisão de recursos de rádio entre eMBB e vários dispositivos URLLC. Avaliamos o desempenho da taxa de três métodos de acesso múltiplo para URLLC, sendo eles, U-OMA, U-NOMA e U-RSMA, quando operando sob divisão de rede OMA e NOMA com eMBB. Nossos resultados mostraram que U-RSMA é capaz de atingir taxas mais altas quando o fator de divisão de potência é configurado corretamente, mesmo com requisitos de confiabilidade mais rígidos. Além disso, mostramos que o fatiamento de rede não-ortogonal é capaz de atingir o maior par de taxas para URLLC e eMBB simultaneamente. Para finalizar a tese, consideramos o compartilhamento de recursos de rádio da rede entre os usuários GB e GF. O método proposto recorre à abordagem MMD para aumentar a diversidade de frequência dos usuários GB e a um protocolo de contenção distribuído para emparelhar um usuário GF com um usuário GB. Além disso, os usuários multiplexados podem ser sobrepostos com NOMA ou RSMA. Apresentamos expressões exatas para o sistema SGF proposto e mostramos que ele é capaz de aumentar a taxa alvo e reduzir a probabilidade de erro de ambos os serviços.

Palavras-chave: 5G e além. Usuários heterogêneos. Alocação de canal. RSMA.

ABSTRACT

This work considers the problem of sharing radio resources between enhanced Mobile Broadband (eMBB) and Ultra-Reliable and Low Latency Communications (URLLC) users connected to a common Base Station (BS) in the scope of 5G and beyond (B5G). Initially, aiming to increase the reliability of the eMBB service, the use of the Max-Matching Diversity (MMD) method to perform channel allocation was proposed, considering the division of resources based on Orthogonal (OMA), Non-Orthogonal (NOMA) and Hybrid (HMA) multiple access. The results indicate that MMD is able to increase the data rate of eMBB users and the reliability of the URLLC service simultaneously for all multiple access methods under consideration. Furthermore, it is shown that the HMA method presents superior results when compared to OMA and NOMA. Subsequently, the same system model was used, but considering multiple URLLC users who share resources with each other through the Rate-Splitting Multiple Access (RSMA) method, where a URLLC device splits its transmission into two sub-messages with partial power, which are potentially recovered in the BS through Successive Interference Cancellation (SIC). To study the performance of such method in the presence of a eMBB user, the OMA and NOMA slicing approaches were considered. As a result, it was shown that, in general, RSMA has improved performance in terms of rate and reliability, even when simultaneously transmitting with a eMBB user. The results also showed that the URLLC rate can be increased by properly adjusting the rate splitting factor based on the average signal-to-noise ratio (SNR), not requiring instantaneous Channel State Information (CSI). Finally, we apply the MMD method to assign Grant-Based (GB) users to channels. In the same network resource that serves a GB user, one Grant-Free (GF) user is allocated through a distributed contention protocol that considers the Quality-of-Service (QoS) of the GB user, performing the pairing that causes the minimum interference, so that the presence of a GF user is transparent to the GB user. Both users admitted to the same physical resource operate in either NOMA or RSMA, depending on the channel gains. Exact expressions for the outage probability of this Semi-Grant-Free (SGF) system are provided and compared to simulation results, showing that the SGF transmissions with MMD-aided GB users strategy reduces the outage and increases the achievable rate of both users.

Keywords: 5G and beyond. Heterogeneous users. Channel allocation. RSMA.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

2G	Second Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation
ARS	Adaptive Rate Splitting
AWGN	Additive White Gaussian Noise
B5G	Beyond 5G
BLER	Block Error Rate
BPCU	Bits per Channel Use
BS	Base Station
CDF	Cumulative Density Function
CSI	Channel State Information
eMBB	enhanced Mobile Broadband
FDMA	Frequency Division Multiple Access
FR	Frequency Range
FRS	Fixed Rate Splitting
GB	Grant-Based
GF	Grant-Free
HARQ-ACK	Hybrid Automatic Repeat Request-Acknowledgement
HMA	Hybrid Multiple Access
H-NOMA	Heterogeneous Non-Orthogonal Multiple Access
H-OMA	Heterogeneous Orthogonal Multiple Access
IoT	Internet of Things
LOTUS	Law of the Unconscious Statistician
LTE	Long Term Evolution
LTE-M	Long Term Evolution Machine Type Communication
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input Multiple-Output
MMD	Maximum Matching Diversity
MMSE	Minimum Mean Squared Error
mMTC	massive Machine Type Communications
mmWave	millimeter Wave
MRC	Maximum Ratio Combining
mTRP	multiple Transmission and Reception Point
NB-IoT	Narrowband Internet of Things
NOMA	Non-Orthogonal Multiple Access

OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OMA	Orthogonal Multiple Access
PDF	Probability Density Function
PHY	Physical Layer
QoS	Quality-of-Service
RaaS	Resource as a Service
RAN	Radio Access Network
RSMA	Rate-Splitting Multiple Access
RVRHK	Random Vertex Rotation based Hopcroft-Karp
SC	Selection Combining
SCMA	Sparse Code Multiple Access
SDMA	Space Division Multiple Access
SGF	Semi-Grant-Free
SIC	Successive Interference Cancellation
SIMO	Single-Input Multiple-Output
SNR	Signal-to-Noise Ratio
TDMA	Time Division Multiple Access
TS	Time Slot
U-NOMA	URLLC Non-Orthogonal Multiple Access
U-OMA	URLLC Orthogonal Multiple Access
URLLC	Ultra-Reliable and Low Latency Communication
U-RSMA	URLLC Rate Splitting Multiple Access

LIST OF SYMBOLS

$\bar{\gamma}_U$	Average SNR of URLLC user
$\bar{\gamma}_B$	Average SNR of eMBB user
$\mathcal{P}_s(\bar{\gamma}_B)$	eMBB outage probability of a single channel communication
F_B	Number of channels allocated to eMBB service
F	Number of independent channels in the bandwidth
w_f	f-th subcarrier
$U_{B,m,f}$	m-th GB user at subcarrier f
f	Channel index
$h_{i,f}$	Channel coefficient of user i in channel f
$\bar{\gamma}_i$	Average SNR of user i
$G_{i,f}$	Channel gain of user i in channel f
S	Number of mini-slots in a time frame
F_U	Number of channels allocated to URLLC service
a_U	Activation probability of an URLLC user
ϵ_B	Communication reliability of eMBB service
P_B	eMBB or GB transmission power
R_U	Rate of URLLC user
ϵ_U	Communication reliability of URLLC service
R_B^{MMD}	eMBB rate with MMD allocation
R_U^{OMA}	Rate of URLLC user in orthogonal slicing
$G_{B,f}^{\mathrm{tar}}$	Target SNR of eMBB user in channel f
$\bar{\gamma}_B^{'}$	Average SNR of eMBB user after definition
$G_{B,f}^{min}$	Threshold SNR of eMBB user in channel f
R_U^{NOMA}	Rate of URLLC user in non-orthogonal slicing
$G_{B,f}^{\mathrm{tar},1}$	Target SNR of eMBB user in channel f considering orthogonal slicing
$G_{B,f}^{\mathrm{tar},2}$	Target SNR of eMBB user in channel f considering non-orthogonal slicing
R_U^{HMA}	Rate of URLLC user in hybrid slicing
R_B	Rate of eMBB user
R_B^{HMA}	eMBB rate in hybrid slicing
R_B^{NOMA}	eMBB rate in non-orthogonal slicing
R_B^{OMA}	eMBB rate in orthogonal slicing
a_B	Activation probability of an eMBB user
n_U	Number of URLLC users sharing the same mini-slot
F'_U	Number of channels of an URLLC user in U-OMA
$\sigma_{n,f}$	SINR of URLLC user n in channel f
$G_{U,n,f}$	Channel gain of URLLC user n in channel f
$R_{U,n}$	Rate of URLLC user n

α	Power splitting factor of RSMA
$\epsilon_U^{\text{U-OMA}}$	URLLC reliability in U-OMA
$\epsilon_U^{\text{U-NOMA}}$	URLLC reliability in U-NOMA
$\epsilon_U^{\text{U-RSMA}}$	URLLC reliability in U-RSMA
K	Number of GF users
$U_{k,f}$	k-th GF user at subcarrier f
M	Number of eMBB or GB users
$h_{k,f}$	Channel of the k-th GF user at subcarrier f
$h_{B,m,f}$	Channel of the m-th GB user at subcarrier f
$\hat{\tau}_f\left(\lvert h_{B,m,f}\rvert^2\right)$	Inteference threshold to GB user
$\hat{R}_B$	GB user target rate
$\mathcal{M}$	MMD maximum matching set
$R_{F,k,f}$	Achievable rate of the k-th GF user at subcarrier f
P_F	GF transmission power
$R_{B,m,f}$	Achievable rate of the m-th GB user at subcarrier f
$R_{\text{I},F,k,f}$	Achievable rate of the k-th GF user at subcarrier f in Subgroup 1
$R^{'}_{\text{II},F,k,f}$	Achievable rate of the k-th first virtual GF user at subcarrier f in Subgroup 2
$R^{''}_{\text{II},F,k,f}$	Achievable rate of the k-th second virtual GF user at subcarrier f in Subgroup 2
$R_{\text{II},F,k,f}$	Achievable rate of the k-th GF user at subcarrier f in Subgroup 2
$\hat{R}_{\text{F}}$	GF user target rate
E_k	Event that there are k GF users in Subgroup I
E_0	Event with no user in Subgroup I
E_K	Event with no user in Subgroup II
$G_{B,f}$	Instantaneous channel gain of eMBB user in channel f
$G_{B,f}^{\text{MMD}}$	Instantaneous channel gain of eMBB user in channel f after MMD allocation
S_U	Number of URLLC users connected
$\bar{E}_U$	Event that considers that all URLLC messages are properly decoded
E_U	Event that considers that all URLLC messages are incorrectly decoded
E_B	Event of eMBB message not being correctly decoded

CONTENTS

1 INTRODUCTION

1.1 BACKGROUND AND THESIS OVERVIEW

In recent years, there has been an accelerated growth in the amount of mobile data traffic. In 2021, this value reached 67 Exabytes per month, with an estimated increase to 282 Exabytes in 2027 (ERICSSON, 2023). This increase is due to several factors, such as: number of connected devices; higher resolution videos; advancement/creation of new communication technologies; and the emergence of new applications, such as Internet of Things (IoT). The Fourth Generation (4G), which still dominates the commercialization of cellular network technology, is based on a one-size-fits-all concept (LI et al., 2017) using broadband communication, which is not able to support the amount of data and heterogeneity expected for applications emerging in the near future, such as smart cities, autonomous and connected cars, public safety, localization and sensing (FERRUS et al., 2018; NGUYEN et al., 2022).

Due to such factors, the advancement of cellular network technologies was necessary, resulting in the start of Fifth Generation (5G) standardization by 3rd Generation Partnership Project (3GPP) in early 2016, which still evolving, currently in Release 17 (CHEN, W. et al., 2023). When compared to previous generations, 5G aims not only at increasing the data rate, but also at reducing latency, increasing the communication reliability, as well as improving the number of nodes that can be connected simultaneously. These requirements are divided into three different services, called Ultra-Reliable and Low Latency Communication (URLLC), massive Machine Type Communications (mMTC) and enhanced Mobile Broadband (eMBB) (SAAD et al., 2020), which in themselves already present major challenges for implementation, as discussed below:

- **URLLC**: requires low latency and high communication reliability (POPOVSKI, P., 2014; FUENTES et al., 2020), with values around 1 millisecond and 10^{-5}, respectively (ITU-R, 2023). To reduce latency, one could choose to use a Grant-Free (GF) access method, where there is no prior allocation of resources to the devices (WANG, C. et al., 2017; LIU, Yan et al., 2021), and transmit small packets with none or the minimum of metadata. To increase communication reliability, diversity techniques and improved source/channel coding could be applied.

- **mMTC**: for the mMTC service, where multiple devices access the channel and transmit messages concurrently, it is necessary to apply techniques capable of decoding numerous overlapping signals, such as Successive Interference Cancellation (SIC) (BOCKELMANN et al., 2016). To reduce propagation errors, the decoding sequence can be ordered from the user with the highest Signal-to-Noise Ratio (SNR) to the user with the lowest SNR (ZAIDI, Ali et al., 2018). Furthermore, for mMTC, one must adopt a modulation that allows the devices to be simple, to reduce cost, and to have low energy consumption (BOCKELMANN et al., 2016).

- **eMBB**: this service is considered the evolution of 4G, with the highest demand in terms of data rate (100 Mbit/s in the downlink and 50 Mbit/s in the uplink) and spectral efficiency (peaks of 30 bit/s/Hz in the downlink and 15 bit/s/Hz in the uplink) (FUENTES et al., 2020; ITU-R, 2023), focusing on the best possible user experience. Solutions used in the current 5G releases to meet these requirements are:
    - Increase the bandwidth available for the service. Compared to a typical Second Generation (2G) network, which has access to 20 MHz at carrier frequencies around 1 GHz, current 5G networks primarily cover a span of 100 MHz in the 3.5 GHz range (BJÖRNSON, E. et al., 2023). 5G also makes use of the millimeter Wave (mmWave) band (28-300 GHz) (SHAFI et al., 2017), mainly for applications that do not require long range, given the physical behavior of waves in this frequency range.
    - Another option is to densify the network, increasing the number of base stations covering the same area and installing hotspots for dedicated applications, such as in hospitals (SHAFI et al., 2017). However, there is also an increase in financial costs linked to this strategy, which is not always feasible in practice.
    - The third option is to increase the spectral efficiency (per cell and of the network as a whole), making it possible to transmit more information within the same amount of resources (BJÖRNSON, Emil et al., 2019). This can be achieved using techniques such as massive Multiple-Input Multiple-Output (MIMO) (TULLBERG et al., 2016), since it can significantly improve the system throughput without consuming extra bandwidth, and Orthogonal Frequency-Division Multiplexing (OFDM), where data is transmitted as a combination of orthogonal narrowband signals to reduce the effects of multipath fading.

It is important to highlight that 5G is already a reality, being deployed in different parts of the world (AMERICAS, 2023), mainly providing the eMBB service in the current release. Since Release 15, the efforts were to improve the eMBB performance as well as support primary versions of mMTC (mainly absorving Narrowband Internet of Things (NB-IoT) and Long Term Evolution Machine Type Communication (LTE-M) from 4G (GSMA, 2023)) and URLLC (SPECTRUM, 2021). In the current Release 17, some improvements were implemented to attend URLLC requirements, as Hybrid Automatic Repeat Request-Acknowledgement (HARQ-ACK), Channel State Information (CSI), intra-UE multiplexing and time-synchronization enhancements (ERICSSON, 2022). In summary, URLLC and mMTC services are still evolving, therefore, it is interesting to look at the problems anticipated for the next versions of this generation (Release 18 and 19, projected to 2023-2025) or even for Sixth Generation (6G) (CHENG-XIANG et al., 2023) networks, since the amount of data will continue to increase in the next ten years (UNION, 2015) and new applications will emerge.

1.2 PROBLEM STATEMENT

In the path to Beyond 5G (B5G) wireless communication systems, it is reasonable to assume that the three 5G heterogeneous services, namely eMBB, URLLC, and mMTC, could be tailored to support new sub-services or even combined, giving rise to new service classes (FLAGSHIP, 2019; SAAD et al., 2020; NGUYEN et al., 2022; TARIQ et al., 2020). Thus, a challenging requirement of B5G wireless systems is to support new heterogeneous use cases, such as pervasive connectivity, industry 4.0 and connected robotics, holographic telepresence, etc. (FLAGSHIP, 2019; SAAD et al., 2020; GIORDANI et al., 2020; TARIQ et al., 2020), while also supporting current 5G services.

To allow the coexistence of these heterogeneous services with diverse requirements within the same Radio Access Network (RAN) architecture, the concept of network slicing has been proposed (ZHANG, H. et al., 2017) and is one solution being used by companies (NOKIA, 2022) to support a heterogeneous network, in which the network functions are divided, through software, into sub-virtual networks that have resources allocated according to the services in use (FOUKAS et al., 2017). This can be performed thanks to network softwarization and virtualization, being considered the main enabler of Resource as a Service (RaaS) for B5G (TARIQ et al., 2020). The resources of the physical infrastructure of the network to be shared between the services include: processing and memory of the devices responsible for decoding, forwarding and storing the messages that travel in the network; frequency spectrum allocated to the technology; wireless communication channels accessed by devices (KALØR et al., 2018).

When analyzing RAN slicing from the point of view of the Physical Layer (PHY), usually multiple access is orthogonal, that is, the resources in the domain of time, frequency, power, code, among others, are allocated orthogonally, reducing interference between users/services (DAI et al., 2015). The current 5G version, Release 17, adopted the Orthogonal Frequency-Division Multiple Access (OFDMA) method for uplink and downlink transmissions, following the 4G standard (SPECTRUM, 2021). However, to attend the requirements of such novel applications, the coexistence of 5G services in the same network infrastructure requires robust multiple access methods that combine higher spectral efficiency with limited delay, high reliability and high number of connected devices.

In the current 5G Release 17, the main improvements implemented to eMBB were based in multiple-antenna, multiple Transmission and Reception Point (mTRP) and extending the Frequency Range (FR) 2 beyond 52.6GHz all the way up to 71GHz (ERICSSON, 2022; NEW, 2022). However, as discussed in Section 1.1, these are expensive solutions in terms of hardware/spectrum license cost and computational resources. Some works as (BAI et al., 2010; SANTOS et al., 2020) presented methods to improve the eMBB capacity through software. Software based methods to increase the capacity become very interesting in scenarios where cost is an important factor, such as private networks, or where the BS power consumption need to be reduced. Furthermore, this strategy has a low cost to increase the performance, different from options as increasing the bandwidth, thus being a policy that supports infrastructure sharing to remote areas.

Therefore, the problem to be tackled in this thesis is the design of software based solutions to improve the performance of 5G and B5G services without increasing hardware or spectrum costs.

1.3 POSSIBLE ENABLERS

1.3.1 Robustly multiplexing users

In applications as the IoT, where various sensors and a massive number of connected devices are expected, GF access has been envisioned to be a good solution to provide a satisfactory tradeoff between the URLLC low latency and reliability (AZARI et al., 2017; LIU, L. et al., 2018; SAMAD et al., 2019; KASSAB et al., 2022). The GF approach enables transmissions without lengthy handshaking protocols to grant devices access (BAYESTEH et al., 2014), in contrast to Grant-Based (GB) access, that relies on pilot messages to estimate the channel of the users and allows the exchange of information when the conditions are favorable to reach the target rate under a given acceptable outage probability. In other words, GF users are allowed to transmit messages whenever they have new data to send, without requesting permission from the Base Station (BS).

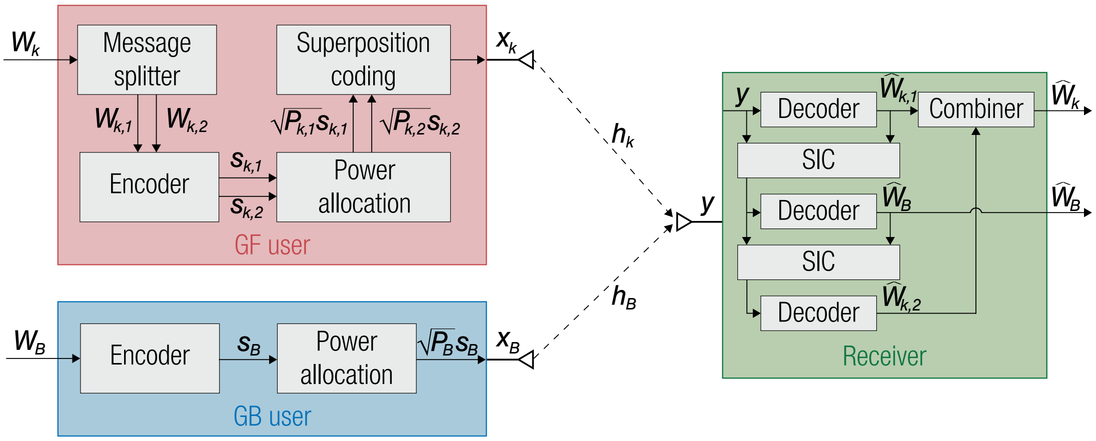
In conjunction to reducing latency, aiming to improve the spectral efficiency, Non-Orthogonal Multiple Access (NOMA) was identified as a promising technology that exploits, for instance, the power domain to allow multiple users to share the same resource block along the spectrum, time and/or code (LIU, Yuanwei et al., 2017). It is mainly indicated on scenarios where the use of multiple antennas is not feasible (CLERCKX et al., 2021). In this way, users do not have to wait to be served when orthogonal blocks are available, which results in a reduction in the latency experienced by users. This is a useful feature to support URLLC communication (CHEN, H. et al., 2018). Moreover, recall that a challenge to achieve mMTC is to support mass connectivity with scarce bandwidth resources, for which NOMA is a great solution as it encourages users to share their resources (MA et al., 2018). In order to recover the overlapped signals, the receiver of a NOMA-based communication system can apply the SIC algorithm, a method whose performance depends on the different power levels between the overlapped incoming signals (ISLAM et al., 2017). To this end, two approaches are commonly used to guarantee such power distinctiveness: (i) user pairing and (ii) power allocation. In the first case, users with distinct channel gains are separated in groups and paired (CHEN, X. et al., 2014; RAUNIYAR et al., 2020). In the second case, power allocation methods separate users (SEDAGHAT; MÜLLER, 2018; AZAM et al., 2019), even if random pairing is applied.

However, since in GF there is no centralized control to limit the number of users, such protocols become ineffective even with NOMA when there are too many active users/devices, due to the excessive number of collisions. This leads to the inability to detect multiple users, which represents the most crucial challenge for the GF transmission strategy (LIU, L. et al., 2018).

One solution to cope with the aforementioned challenge is to employ a contention protocol, as multiple users may choose the same channel to transmit at the same time, which is known as Semi-Grant-Free (SGF) transmission schemes, and that can be viewed as a compromise between GF and GB schemes (DING et al., 2019; ZHANG, C. et al., 2021). For instance, in a scenario with heterogeneous users, all channels in the network may be available for GF transmission, regardless of whether they have been reserved by GB users or not. However, to ensure that the Quality-of-Service (QoS) requirements of GB users are fulfilled, the contention among opportunistic GF users must be carefully managed, to ensure that GF users transmit only when they do not cause significant performance degradation to the GB users.

In NOMA, it is intuitive that the complexity of user pairing and power allocation increases with the number of users, turning its implementation unbearable in terms of latency in scenarios with a massive number of users. Also, in some cases, it is necessary to use CSI to adapt the transmission power, which entails extra latency and a potential loss in terms of reliability. In order to overcome issues of this nature, the Rate-Splitting Multiple Access (RSMA) method has gained significant attention recently, since it enables the achievement of the entire capacity region with successive decoding (RIMOLDI; URBANKE, 1996; TSE; VISWANATH, 2005a), providing superior performance over NOMA and over OMA methods like Space Division Multiple Access (SDMA) (CLERCKX et al., 2016; MAO et al., 2018). RSMA also shows robustness in cases with imperfect CSI (LEE et al., 2021) and is further used to mitigate problems of pilot contamination (MISHRA et al., 2022). Moreover, the basic principle of NOMA, which requires a single user in each group to decode the messages of other

Figure 1 – Illustration of the coding and decoding process in one transmission block of RSMA.



Source: The Author.

co-scheduled users, is an inefficient design in multi-antenna scenarios. In this case, RSMA also shows superior performance since the interference is partially decoded and partially treated as noise, which results in higher multiplexing gains (CLERCKX et al., 2021). In uplink RSMA, each user creates virtual users by splitting its transmission in two sub-messages. Although this procedure entails extra rounds in the SIC procedure, it automatically creates different arriving power levels among users, thus significantly reducing the implementation complexity when compared to NOMA. RSMA also increases the number of decoding orders, thus each part can be decoded flexibly at the receiver using SIC (YANG, Z. et al., 2020a; CLERCKX et al., 2023). Hence, if the GF user has a worse channel, one of its split streams can be decoded after the other streams, so it has lower interference and can reach higher rates. The case when it has a better channel is also valid. In this way, a higher decoding order flexibility enables higher achievable rates. This is also a big challenge in practical RSMA deployments and must be optimized, since the decoding order affects the achievable rate. Facing this issue, some works have proposed SIC receivers with dynamic decoding order to improve performance (ZHANG, Z.; HU, 2017; GAO et al., 2017; LIU, Ye et al., 2018).

Fig. 1 illustrates the RSMA coding and decoding process in a subcarrier with one GB and one GF user. We assume that only the GF admitted user message, W_k, is split into sub-messages, named $W_{k,1}$ and $W_{k,2}$. By independently encoding the two parts into $s_{k,1}$, $s_{k,2}$, respectively allocating transmit power $P_{k,1} = \alpha P_F$, $P_{k,2} = (1-\alpha)P_F$, where $0 \leq \alpha \leq 1$ is the RSMA power splitting factor, and superposing the two streams, the transmit signal of the GF user is given by

$$x_k = \sqrt{P_{k,1}}s_{k,1} + \sqrt{P_{k,2}}s_{k,2}. \tag{1}$$

At the GB user, the message W_B is directly encoded into s_B and a certain power P_B is allocated, resulting in the transmitted signal $x_B = \sqrt{P_B}s_B$. In each transmission block, the paired GB and GF users transmit their signals simultaneously, so the received signal at the BS can be written as

$$\begin{aligned} y &= h_B x_B + h_k x_k + n \\ &= \sqrt{P_B} h_B s_B + \sqrt{P_{k,1}} h_k s_{k,1} + \sqrt{P_{k,2}} h_k s_{k,2} + n, \end{aligned} \tag{2}$$

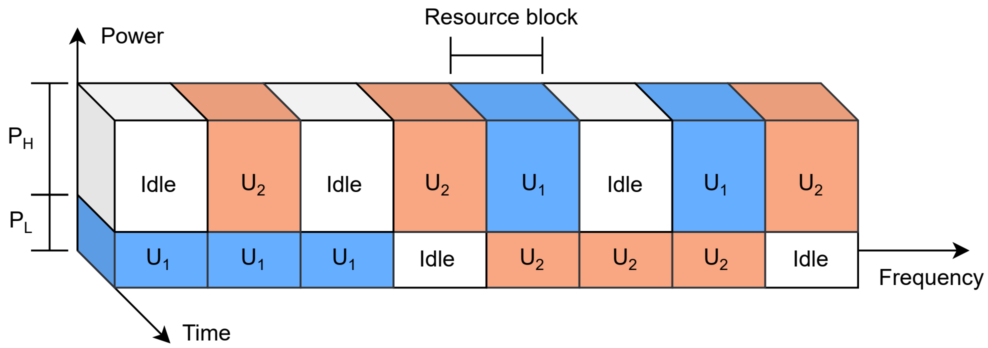
where $n \sim \mathcal{CN}(0, \sigma^2)$ is Additive White Gaussian Noise (AWGN) with zero mean and unit variance ($\sigma^2 = 1$).

Given the signal model of (2), in general two SIC layers are required to recover the messages of the two users at the receiver. However, for the cases of $\alpha = 0$ or $\alpha = 1$, the admitted GF user transmits the signal $x_k = \sqrt{P_F} s_k$ without splitting its message, thus NOMA is used, and only one SIC layer is needed.

Another multiple access technique that promises to be a good solution in B5G use cases and has been explored in the literature is the Hybrid Multiple Access (HMA) (MAHMOUDI et al., 2022; WANG, Q. et al., 2020; AL-ABBASI; SO, 2017; TANAKA et al., 2021; ELBAYOUMI et al., 2020; KIM; CHO, 2017), a method that combines both OMA and NOMA. As presented in (POPOVSKI, P. et al., 2018; ABREU et al., 2019; SANTOS et al., 2020; MAATOUK et al., 2019), OMA and NOMA can outperform each other in different setups. Therefore, a HMA method that dynamically adapt the slicing fraction, also called adaptive multiple access protocol, combining OMA and NOMA schemes according to the demands and requirements of each service (SAAD et al., 2020), may provide several benefits compared to using only OMA or NOMA.

Fig. 2 illustrates this hybrid scheme where, when operating under NOMA, users are separated in the power domain. The figure shows that the first resource block is reserved for user 1, the eighth resource block is reserved for user 2, and both users share the second resource block. This indicates that the HMA method has two advantages: it is less vulnerable to interference and requires fewer SIC processes than NOMA, and it is more spectrum-efficient than OMA as some users can share more resources compared to a completely orthogonal scheme.

Figure 2 – HMA scheme for two users in a bandwidth divided in eight channels.



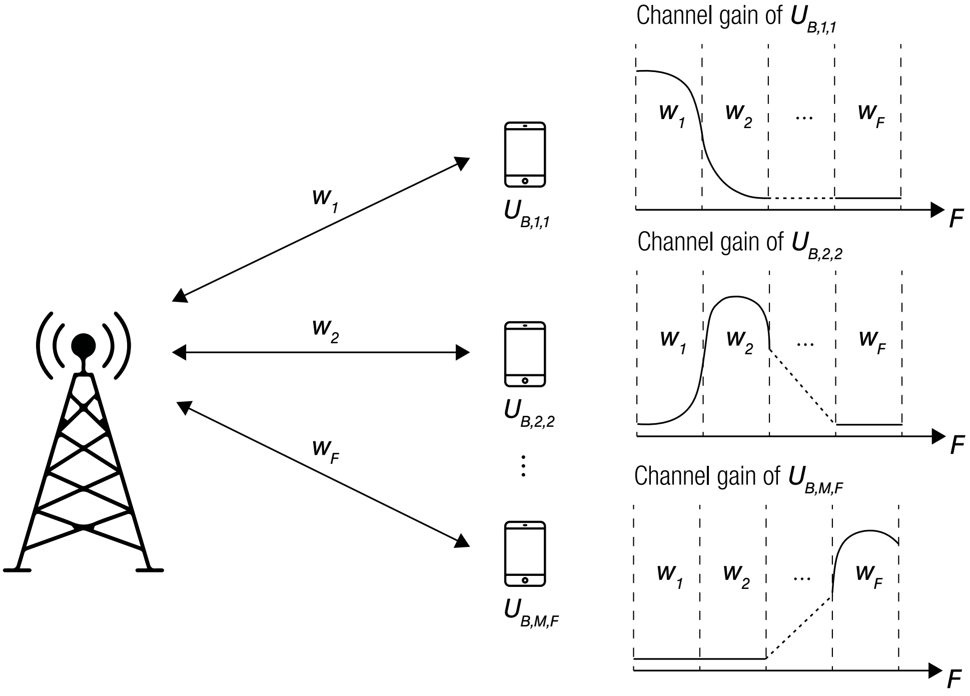
Source: The Author.

Furthermore, HMA can provide a better tradeoff between fairness and system efficiency compared to using only OMA or NOMA. OMA provides equal treatment to all users, while NOMA can provide higher throughput for some users at the expense of others. Finally, HMA can be more flexible, as it can adapt to different traffic conditions and user requirements. For example, when the number of users is low, the system can increase the OMA resource blocks to provide higher rates to all users, while when the number of users is high, preference can be given to NOMA to increase system spectral efficiency.

1.3.2 Allocating channels to GB through software

The outage probability in an OFDM system with F channels serving one user can be approximated for Rayleigh fading as $P_u \approx (\bar{\gamma}_B)^{-F}$, where $\bar{\gamma}_B$ is the average SNR. Thus, diversity order F is achieved. For an OFDMA system with $M \leq F$ users, the maximum frequency diversity gain for each user would be only F/M (BAI et al., 2010). Fig. 3 illustrates a set of different channel realizations among GB users and independent subcarriers. The bandwidth is divided in F channels, such that $\{w_f\}_{f=1}^{F}$ represents the f-th subcarrier. In this example, user $\{U_{B,m,f}\}_{m=1}^{M}$ has a good channel condition in subcarrier w_f, but it is in deep fading in other subcarriers.

Figure 3 – Example of channel realizations for different users and subcarriers.



Source: The Author.

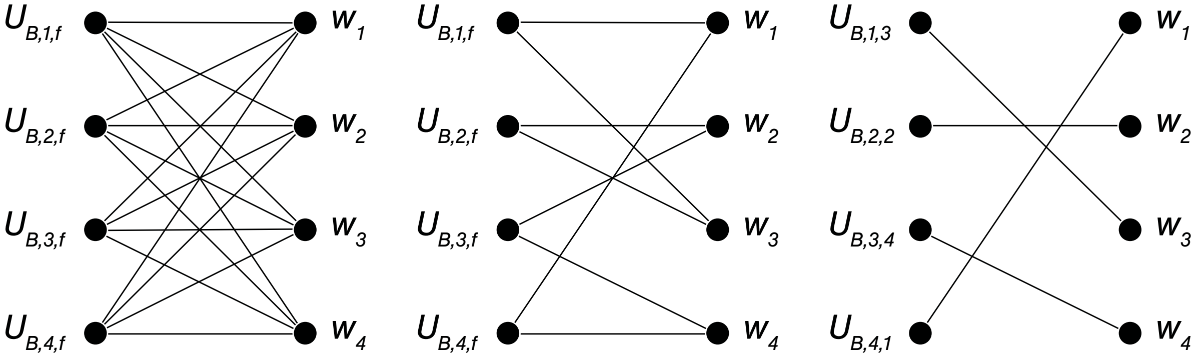
The authors from (BAI et al., 2010) presented an ingenious way to increase the frequency diversity gain in OFDMA, applying the Random Vertex Rotation based

Hopcroft-Karp (RVRHK) algorithm to assign users to channels. Taking as example Fig. 3, an obvious decision would be to allocate subcarrier w_f to user $U_{B,m,f}$. Surprisingly, the maximum frequency diversity gain achieved by the RVRHK algorithm is the same as that in point-to-point OFDM systems which serves only one user with F channels, while not decaying as the number of users increases. In this work, we refer to this allocation procedure as Maximum Matching Diversity (MMD), which has the following remarks (BAI et al., 2010): *i)* achieves optimal frequency diversity F through software, reducing the outage probability; *ii)* guarantees fairness between users; *iii)* inserts complexity of $\mathcal{O}(F^{2.5})$; and *iv)* adds 1 bit/channel/user of control data. In the following, more details about the allocation algorithm and the outage probability are presented.

1.3.2.1 RVRHK Algorithm

The RVRHK algorithm was designed to allow the largest amount of users to obtain non-outage channels, aiming at minimizing the outage probability of each user (BAI et al., 2010). To that end, the random bipartite graph model, which is effective to solve assignment problems, is used to formulate the channel allocation problem, where the users are one vertex set with the channels being the other set. All the non-outage channels are seen as edges between the user vertices and the channel vertices, that appear with some probability, and which could be intuitively described as: if $U_{B,m,f}$ is not in outage in w_f, join them with an edge.

Figure 4 – MMD allocation routine for M = 4 and F = 4.



(a) Complete graph. (b) Non-outage. (c) After allocation.

Source: The Author.

Fig. 4 illustrates each step of the allocation procedure for M = 4 users and F = 4 channels. This procedure occurs during the connection phase, being the BS responsible for executing the RVRHK algorithm. First, Fig. 4-a presents the complete graph, but after identifying the non-outage scenarios, only the possible edges, *i.e.*, the valid channel allocation options, remain connected. At this step, each user is connected to a set of non-outage channels, as for instance $\mathcal{N}(U_{B,1,f}) = \{w_1 U_{B,1,f}, w_3 U_{B,1,f}\}$ is the *matching subset* of $U_{B,1,f}$ (and as in Fig. 4-b). At the last step of the algorithm, the *maximum matching set* denoted by $\mathcal{M}$, which achieves the optimal frequency diversity, is determined. In this example, $\mathcal{M} = \{w_3 U_{B,1,3}, w_2 U_{B,2,2}, w_4 U_{B,3,4}, w_1 U_{B,4,1}\}$ is a possible solution. This max-matching $\mathcal{M}$ is found using the Hopcroft-Karp algorithm, but randomly rotating the users at the beginning to guarantee that every user gets an identical priority to be connected to a channel through an edge (this random user rotation is what differentiates the Hopcroft-Karp algorithm to the RVRHK algorithm). The m-th GB user is in outage either if $\mathcal{N}(U_{B,m,f}) = \emptyset$ or every w_f in $\mathcal{N}(U_{B,m,f})$ has been allocated to other users.

1.3.2.2 Outage of MMD-Aided OFDMA

Following (BAI et al., 2010), the outage probability of a user in a MMD-aided OFDMA system with F channels is

$$P_u = P_s(\bar{\gamma}_B)^F f(P_s), \tag{3}$$

where

$$f(P_s) = a_0 + a_1 P_s(\bar{\gamma}_B) + \cdots + a_{(M-1)F} P_s(\bar{\gamma}_B)^{(M-1)F}, \tag{4}$$

with $a_0 \neq 0$ and, for Rayleigh fading,

$$P_s(\bar{\gamma}_B) = 1 - e^{-\gamma_{min}/\bar{\gamma}_B}, \tag{5}$$

where $\gamma_{min} = 2^{\hat{R}_B} - 1$ is the threshold SNR and $\hat{R}_B$ is the target rate of the GB user.

Moreover, in the high SNR regime, the MMD outage probability from (3) can be approximated as (BAI et al., 2010)

$$\begin{aligned} \mathcal{P}_u &= a_0 \mathcal{P}_s(\bar{\gamma}_B)^F + O\left(\mathcal{P}_s(\bar{\gamma}_B)^F\right) \\ &\approx a_0 \mathcal{P}_s(\bar{\gamma}_B)^F, \end{aligned} \tag{6}$$

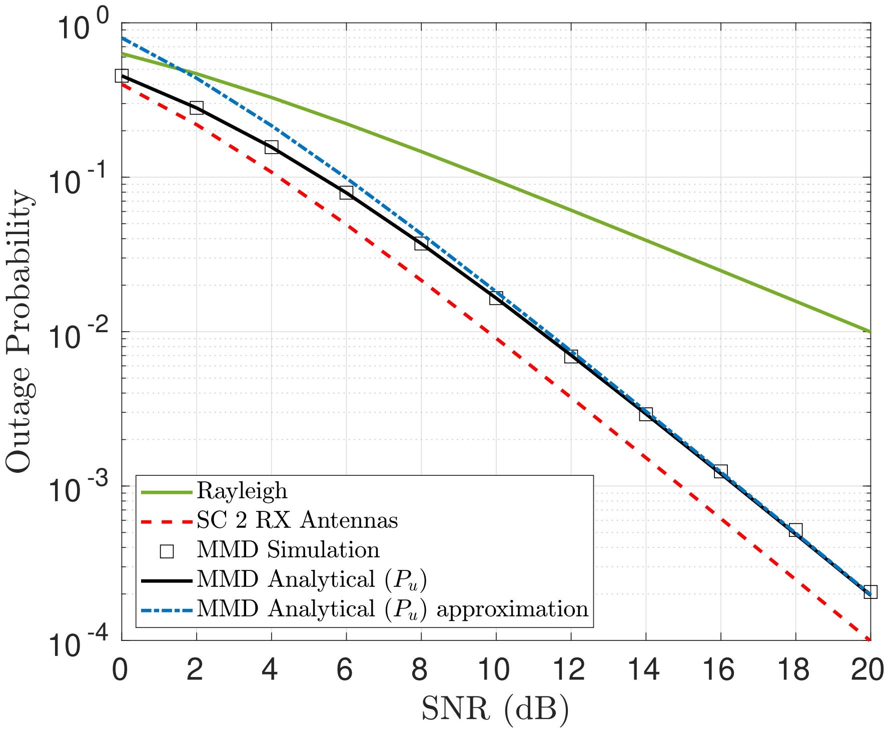
where $O(\cdot)$ is the higher order infinitesimal. Furthermore, a_0 corresponds to the multiplicity of the most relevant term of (6), which is $a_0 = 1$ for $M < F$ (user in outage only when this user is an isolated vertex of the bipartite graph) or $a_0 = 2$ for $M = F$ (user in outage either when this user or one subcarrier is an isolated vertex of the bipartite graph), as detailed in (BAI et al., 2010).

Thus, note from (6) that the MMD scheme maximizes the frequency diversity experienced by each user, which equals the number of channels F. Note also that such frequency diversity improvement comes exclusively due to the proper allocation provided by the RVRHK algorithm, i.e., one does not need to retransmit the same message in channels subjected to independent fading, which would require a larger bandwidth.

Example 1. *For $M = F = 2$, we have that (BAI et al., 2010)*

$$P_u = 2P_s(\bar{\gamma}_B)^2 - 2P_s(\bar{\gamma}_B)^3 + P_s(\bar{\gamma}_B)^4, \tag{7}$$

Figure 5 – MMD outage probability with $M = 2$ users and $F = 2$ channels for single antenna devices. We also illustrate the case without frequency diversity, but considering either a single or two receive antennas using SC.



Source: The Author.

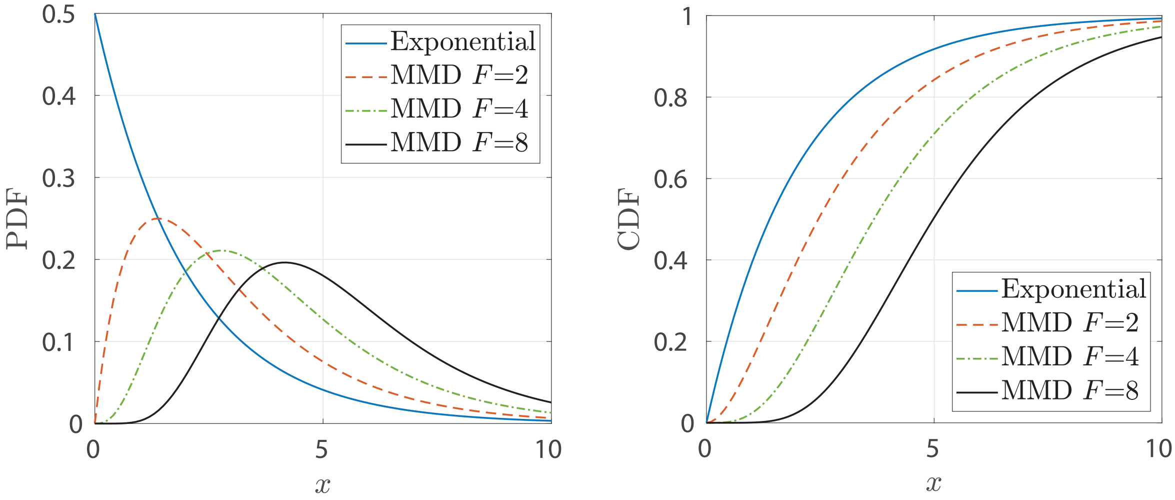
which is plotted in Fig. 5 along with simulation results. Moreover, we compare it with the outage probability in the case without frequency diversity, considering either a single or two receive antennas using Selection Combining (SC). The idea is to illustrate the diversity orders achieved with MMD exploiting frequency diversity from $F = 2$ channels (but without spatial diversity) with that of SC exploiting spatial diversity with two antennas (but without frequency diversity). It is important to mention that frequency diversity is obtained only through the proper allocation of channels, without retransmitting copies of messages on different channels. In other words, there is no redundancy in frequency. Clearly, the MMD method achieves diversity order $F = 2$ with single antenna devices, while the outage probability of MMD is two times higher because of the a_0 factor in (6)*.*

These results suggest that MMD is an optimal subcarrier allocation method in the sense of frequency diversity gain. This is achieved thanks to the impact created in the equivalent users' channels, whose SNR probability distribution can be expressed as a linear combination of F exponential Probability Density Function (PDF) as (SANTOS et al., 2020)

$$p_B(x) = \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} \frac{f e^{-fx/\bar{\gamma}'_B}}{\bar{\gamma}'_B}, \tag{8}$$

where $\bar{\gamma}'_B = \bar{\gamma}_B 2^{-1/F}$. The resulting distribution is less severe than the exponential distribution, as can be seen in Fig. 6, characterizing a more benign equivalent channel since the probability of being in deep fade reduces. Thus, applying MMD increases the frequency diversity gain experienced by the GB users, leading to a better equivalent channel when compared to the case without MMD.

Figure 6 – (a) PDF and (b) CDF of the exponential distribution and the MMD-based SNR distribution for $F \in \{2, 4, 8\}$.



(a) PDF.

(b) CDF.

Source: The Author.

1.4 SCOPE AND OBJECTIVES OF THE THESIS

In the path to B5G wireless communication systems, it is reasonable to assume that the three heterogeneous services could be divided into sub-services (MAHMOOD et al., 2020) or even combined, emerging new service classes (SAAD et al., 2020). Such services require robust multiple access methods that can combine higher spectral efficiency with strict delay and reliability requirements to attend applications like fully automated driving, where cooperation among cars for collision avoidance is vital (POPOVSKI, Petar et al., 2018; CHEN, H. et al., 2018).

Motivated by the aforementioned challenges, this thesis focuses on wireless access methods to perform the sharing of RAN resources between eMBB and URLLC users, as well as GB and GF users, in the same network infrastructure. To accomplish such task, mathematical analysis and computational simulations are used to investigate these potential strategies in the uplink, aiming to achieve the requirements of reliability, latency and data rate.

First, we consider the multiplexing of eMBB and URLLC users with a hybrid slicing strategy, focusing on improving the spectral efficiency. The MMD allocation method (BAI et al., 2010) is used to allocate channels to eMBB users. The results are compared to the MMD-aided versions of OMA and NOMA from (SANTOS et al., 2020). The achievable frequency diversity gain of MMD combined with HMA slicing is capable of improving the eMBB rate without decreasing the performance of URLLC devices, as well as of improving the URLLC reliability without reducing the eMBB achievable rate.

In the sequence, aiming to increase the URLLC spectral efficiency, we allow more than one URLLC user to perform non-orthogonal sharing of frequency and time resources through rate-splitting, which we refer to as URLLC Rate Splitting Multiple Access (U-RSMA). In the proposed scheme, we combine the benefits of RSMA, SIC decoding and frequency diversity, in both OMA and NOMA network slicing with eMBB. The proposed U-RSMA scheme is then compared to the so-called URLLC Non-Orthogonal Multiple Access (U-NOMA) and URLLC Orthogonal Multiple Access (U-OMA) schemes, where the multiple access between URLLC devices is performed by means of NOMA and OMA, respectively. To the best of our knowledge, this work is the first to apply RSMA to URLLC uplink transmission in a network slicing scenario, showing that RSMA can outperform both OMA and NOMA methods for URLLC service even in the presence of eMBB interference, specially for very strict reliability levels.

Finally, we study a SGF transmission system. We use the MMD approach in order to assign GB users to channels. One GF user is assigned to the same network resource as a GB user through a distributed contention protocol that takes into account the GB user's QoS and pairs them in a way that causes the least amount of interference, making the presence of a GF user almost transparent to the GB user. Depending on channel gains, both users admitted to the same physical resource operate in NOMA or RSMA. Exact expressions for the outage probability of this SGF system are provided and compared to simulation results.

1.5 CONTRIBUTIONS

The main contributions of this work are summarized as follows:

1. A new physical layer network slicing scheme that adapts to channel conditions and service demands, referred as HMA;
2. A rate-splitting design for URLLC uplink transmissions;
3. Detailed deriving of analytical expressions for eMBB users achievable rate, with and without MMD channel allocation;
4. Detailed deriving of URLLC outage probability expressions when operating solely, under U-RMSA, U-OMA and U-NOMA;
5. Exact outage expressions for GF users when sharing resources non-orthogonally with a MMD-aided GB user;

6. An improved SGF scheme, where both GB and GF users have superior performance when compared to a benchmark;

7. A comprehensive discussion of potential methods to address the heterogeneous services' coexistence problem in B5G networks.

1.5.1 Publications

The following scientific publications were authored during the development of this thesis.

1.5.1.1 Published

1. E. J. Santos Jr, R. D. Souza and J. L. Rebelatto, "*Hybrid multiple access for channel allocation-aided eMBB and URLLC slicing in 5G and beyond systems*", in **Internet Technology Letters**. 2021; 4:e294. doi:10.1002/itl2.294.

2. E. J. Santos Jr, R. D. Souza and J. L. Rebelatto, "*Rate-Splitting Multiple Access for URLLC Uplink in Physical Layer Network Slicing With eMBB*", in **IEEE Access**, vol. 9, pp. 163178-163187, 2021, doi: 10.1109/ACCESS.2021.3134207.

1.5.1.2 Submitted

1. E. J. Santos Jr, R. D. Souza and J. L. Rebelatto, "*Rate-Splitting Multiple Access for Semi-Grant-Free Transmissions with MMD-aided Grant-Based Users*", submitted. 2023.

1.6 THESIS OUTLINE

The rest of this thesis is organized as follows.

Chapter 2 starts with a literature review about HMA. Then, it presents the system model and outage formulation for MMD-aided eMBB and URLLC users when solely occupying the resources. In the sequence, this formulation is extended to the cases of OMA, NOMA and HMA network slicing. The succeeding section presents the simulation results, followed by a summary of the chapter and its major findings. Chapter 3 has a similar structure, but its main topic is the multiplexing of URLLC users with U-RSMA, U-OMA and U-NOMA and how the presence of a eMBB user impacts performance. Chapters 2 and 3 are complemented by appendices. Chapter 4 combines the benefits of MMD and RSMA methods to enhance the performance of a network composed by GB and GF users. After a literature review, the system model and outage formulation are presented, followed by the simulation results. Finally, Chapter 5 concludes the thesis and discusses some potential future works.

2 HYBRID MULTIPLE ACCESS FOR URLLC AND EMBB SLICING

This chapter discusses the system model, equations and results related to the application of the HMA method for slicing URLLC and eMBB (with MMD channel allocation) users in a single-cell network.

2.1 LITERATURE REVIEW

Previous works have paved the way for studying HMA. Authors from (SUGANUMA et al., 2019) proposed a HMA scheme using simultaneously NOMA and OMA in the same bandwidth of a 5G network, focusing on surpassing the negative points that NOMA technique presents in the power domain, such as failure in decoding inter-user interference when the difference among channels gains is small, cases where OMA presents superior performance. Results showed that, with a hybrid implementation, the overall system capacity can be increased.

In (MARCANO; CHRISTIANSEN, 2017), HMA was implemented to improve the capacity of broadband users using a power-domain NOMA strategy. A pairing algorithm based on Modulation and Coding Scheme (MCS) was proposed, that combined with extra transmission power setting, has the objective of increasing the SNR levels of users in the NOMA partition. To not exceed the regulated maximum limit of transmission power, a portion of the users is allocated in OMA slice, where the power levels can be lowered given the absence of interference. Results showed that the application of such method is capable of increasing the average user bit rate (up to 3.31-fold) and the system capacity (up to 1.78-fold), keeping the Block Error Rate (BLER) below 10%.

In (AL-ABBASI; SO, 2017), the authors presented a new resource allocation scheme for NOMA systems with a proportional rate constraint. This scheme is designed to allocate resources to users in a way that maximizes the sum rate of the system, while ensuring that each user achieves a desired proportional rate. On the top of that, a new HMA technique that combines the properties of NOMA and OFDMA was designed. To evaluate the performance of the proposed scheme, the authors performed computational simulations and compared their results with existing schemes in the literature, showing that the proposed scheme outperforms existing schemes both in terms of system sum rate, proportional rate and coverage probability.

The work in (ANWAR et al., 2020) proposed a novel method for downlink communication using HMA. An optimization problem was formulated and numerically solved to obtain the number of users that should be allocated in OMA which results in the optimal network throughput for HMA. Moreover, the formulation of outage probability was presented. Results showed that HMA always outperforms OMA, however, when compared to NOMA, the proposed scheme is capable of increasing the communication reliability in 40% with the disadvantage of a throughput loss of 3%.

In (MAHMOUDI et al., 2022), the authors proposed a new approach for clustering users, allocating resources, and selecting decoding order in a hybrid NOMA-OMA system to ensure equity among all users with regard to the probability of success. The users are grouped into multiple clusters, and each cluster utilizes channel resources using an orthogonal multiple access scheme. Within each cluster, power-domain NOMA is employed by the users. Simulation results show that the proposed scheme outperforms the existing schemes not only in terms of the fairness and minimum success probability of users, but also in terms of the total throughput. In (WANG, Q. et al., 2020), a similar model is used, but aiming to minimize the information freshness (also known as age of information) for the downlink transmission. However, none of the works have studied the coexistence of different services through hybrid access.

2.1.1 Novelty and Contribution

In this chapter, we study the multiplexing of eMBB and URLLC users proposing a HMA network slicing strategy to combine the current adopted method OMA with NOMA[1], focusing on improving the spectral efficiency. The MMD channel allocation method from (BAI et al., 2010) and adapted in (SANTOS et al., 2020) for eMBB users is used. The model for URLLC service, OMA and NOMA slicing are the same of (SANTOS et al., 2020), and will be reproduced here for comparison purposes. The main contributions of this work are summarized as follows:

- We derive the achievable sum-rate for eMBB (with MMD) and URLLC outage probability expressions in HMA slicing;
- We propose a network slicing method that can adapt the slicing depending on average channel conditions and services demand to improve spectral efficiency;
- We provide numerical results to compare the performance between HMA MMD-aided, NOMA MMD-aided and OMA MMD-aided schemes.

2.2 SYSTEM MODEL

The uplink of a single-cell network with eMBB and URLLC devices transmitting to a common BS is considered. The bandwidth is divided into F channels of index $f \in \{1, \ldots, F\}$, where each channel is subject to independent and identically distributed (i.i.d.). Rayleigh fading, since all devices are assumed to have a large enough spatial separation, so the fading realization observed by each device is uncorrelated from another. This fading is assumed to be constant during one transmission Time Slot (TS).

[1] In the literature, OMA and NOMA slicing are also referred to Heterogeneous Orthogonal Multiple Access (H-OMA) and Heterogeneous Non-Orthogonal Multiple Access (H-NOMA), respectively. Here, the *heterogeneous* term is omitted for brevity.

The channel coefficient of user $i \in \{B, U\}$ in channel f is thus $h_{i,f} \sim \mathcal{CN}(0, \bar{\gamma}_i)$, which represents that it is distributed as a circular-symmetric complex Gaussian channel (i.e., the channel coefficient is the same in all directions), where $\bar{\gamma}_i$ corresponds to the average SNR, being $G_{i,f} \triangleq |h_{i,f}|^2$ the channel gain, and where subscripts B and U refer to eMBB and URLLC devices, respectively. The number of channels allocated to service i is $F_i \leq F$, with $i \in \{B, U\}$. Moreover, each TS is divided into S mini-slots.

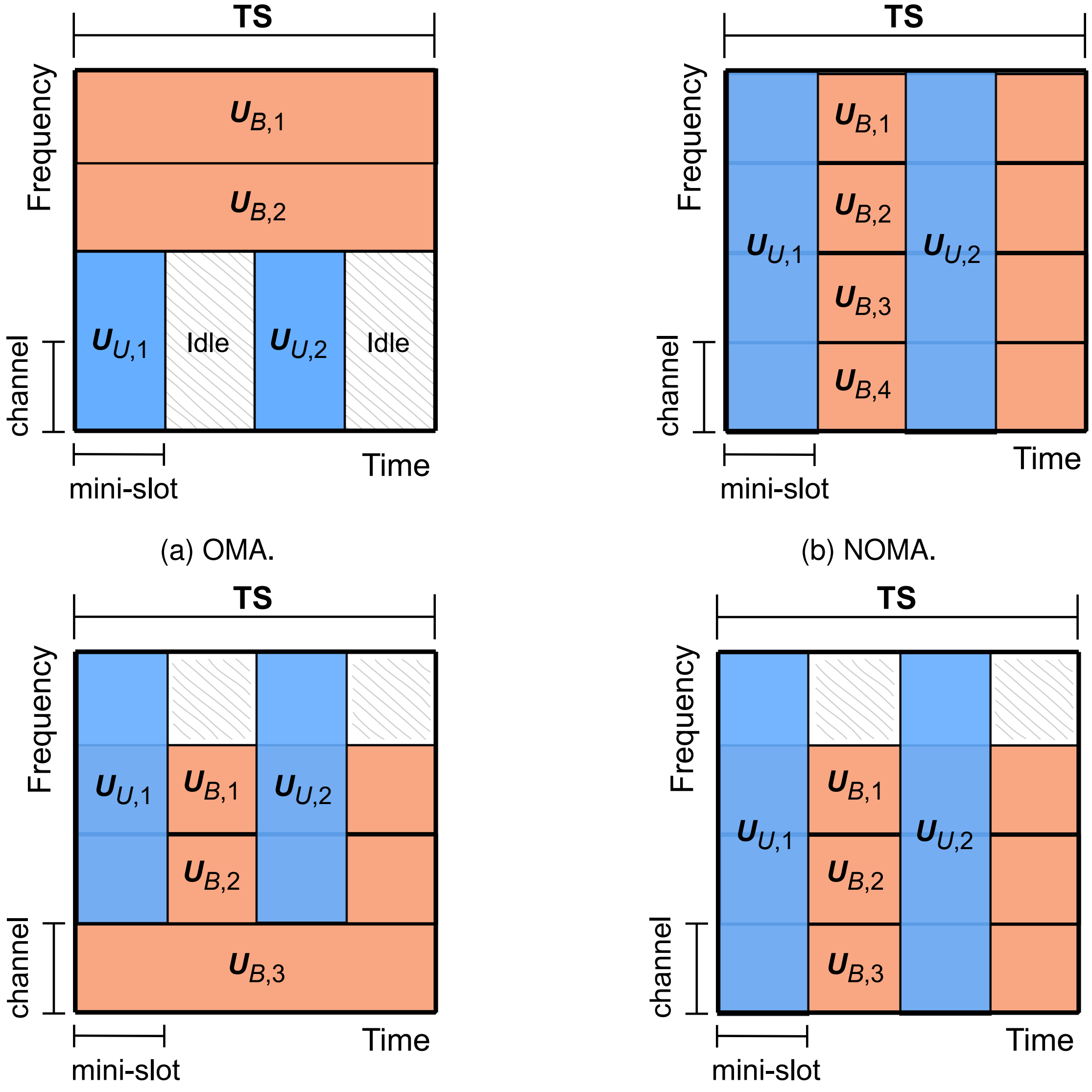
The following approaches regarding URLLC and eMBB transmissions are adopted:

- An URLLC device transmits in a pre-assigned mini-slot, where the resources are allocated in a GF fashion, without any scheduling request or resource allocation. Thereby, the communication latency can be reduced to meet the rigorous requirements and collisions are avoided. The latency considered is 1 mini-slot (the smallest possible delay in this work model). To increase the reliability, the device spreads the transmission over F_U channels. The activation probability of the device in a given pre-assigned mini-slot is a_U.

- An eMBB user transmits in a single channel f (dynamically allocated using MMD) among the F_B available channels, staying connected during the entire TS. We focus on the transmission phase, considering that radio access and competition among eMBB devices have been resolved before the considered time slot.

This time-frequency grid is illustrated in Fig. 7, considering OMA in Fig. 7a, NOMA in Fig. 7b and traditional HMA in Fig. 7c. Fig. 7d shows the HMA slicing optimized for eMBB-URLLC coexistence. As will be shown in Section 2.5.2, because of the high URLLC reliability and detection priority, the interference that it causes in eMBB traffic is minimal, so all bandwidth is allocated to URLLC to maximize its frequency diversity and the resource block allocated exclusively to eMBB is now shared. In this example, $S = 4$ is the quantity of mini-slots in the time domain, while $F = 4$ is the total number of channels available in the bandwidth.

Figure 7 – System model with F = 4 channels and S = 4 mini-slots.



(a) OMA.

(b) NOMA.

(c) Regular HMA.

(d) Optimized HMA (eMBB-URLLC).

Source: The Author.

2.3 EMBB WITH MMD

As presented in (SANTOS et al., 2020), a radio resource is allocated to one eMBB user through MMD method using the CSI information acquired before data transmission. The objective of such process is to maximize the eMBB data rate, given the requirements of reliability (ϵ_B) and average power constraint ($P_B = 1$). The MMD-aided eMBB rate is

$$R_B^{\text{MMD}} = \log_2\left(1 + G_{B,f}^{\text{tar}}\right), \qquad \text{(bits/s/Hz)} \tag{9}$$

where the target SNR is defined as

$$G_{B,f}^{\text{tar}} = \frac{\bar{\gamma}_B'}{\sum_{f=1}^{F_B}(-1)^{f-1}\binom{F_B}{f} f\Gamma\left(0, \frac{fG_{B,f}^{\min}}{\bar{\gamma}_B'}\right)}, \tag{10}$$

$G_{B,f}^{\min} = -\bar{\gamma}_B' \ln\left(1-\epsilon_B^{1/F_B}\right)$ and $\bar{\gamma}_B' \triangleq 2^{-(1/F_B)}\bar{\gamma}_B$. For the proof, please refer to Appendix B.

2.4 URLLC

For URLLC devices, it is assumed that no CSI is acquired, differently from eMBB users. This consideration is justified due to the fact of much higher latency restrictions, impeding the exchange of reference signals for channel state acquisition and/or reporting. As a result, no power and rate adaptation is used for URLLC transmissions. The URLLC device transmits data in all F_U i.i.d. channels of a mini-slot to increase its reliability through frequency diversity. The outage probability, in the absence of interference from eMBB, is (SANTOS et al., 2020)

$$P_{\text{out}}(G_{U,f}) = \Pr\left(\frac{1}{F_U}\sum_{f=1}^{F_U}\log_2(1+G_{U,f}) < R_U\right). \tag{11}$$

The target rate R_U is obtained by imposing the requirement $P_{\text{out}}(G_{U,f}) \leq \epsilon_U$ to (11), where ϵ_U is the URLLC reliability requirement.

2.5 SLICING FOR URLLC AND EMBB WITH MMD

2.5.1 Orthogonal network slicing

In OMA slicing, the F channels are divided in two slices, one containing F_B channels for eMBB users and other F_U channels for URLLC devices, such that $F_B + F_U = F$. In this scenario, the eMBB rate is obtained by considering the sum-rate of the active eMBB users

$$R_B^{\text{OMA}} = F_B\, R_B^{\text{MMD}}, \tag{12}$$

where R_B^{MMD} comes from (9) and R_U^{OMA} is computed from (11).

2.5.2 Non-orthogonal network slicing

In NOMA, all F channels are available simultaneously for both eMBB and URLLC to increase the spectral efficiency. However, this method introduces the inter-service interference that it is handled in our work considering that the BS performs SIC to successively demodulate and decode the URLLC messages with priority, to attend latency requirements, then re-encode and subtract its contribution to the received signal to decode eMBB traffic. If an error occurs during URLLC decoding, the eMBB packets are lost.

Following (SANTOS et al., 2020), the achievable sum-rate of a MMD-aided eMBB device in NOMA is

$$R_B^{\text{NOMA}} = F \log_2 \left(1 + G_{B,f}^{\text{tar}}\right), \tag{13}$$

where $G_{B,f}^{\text{tar}}$ is upper bounded by

$$G_{B,f}^{\text{tar}} = \frac{\bar{\gamma}_B^{'}}{\sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} f \Gamma \left(0, \frac{f G_{B,f}^{min}}{\bar{\gamma}_B^{'}}\right)} \tag{14}$$

and the threshold SNR is

$$G_{B,f}^{\text{min}} \leq -\bar{\gamma}_B^{'} \ln \left(\frac{1 - \epsilon_B^{1/F_B}}{1 - \epsilon_U (1 - (1 - a_U)^S)} \right). \tag{15}$$

For more details about (13)-(15) we refer the reader to Appendix C.

We can keep the target SNR approximately the same of the orthogonal case because the MMD channel allocation applied to eMBB already reduces the interference caused in URLLC transmissions, since eMBB can operate at lower transmission power levels, achieving the same (or even improved) performance. As the URLLC has more strict reliability requirements, the SIC process fails in decoding its messages only in few situations and the impact of URLLC transmissions in the eMBB decoding should be minimal. As $\epsilon_U << \epsilon_B$, the threshold from (15) may be approximately the same as the orthogonal case. In contrast, the eMBB interference in the URLLC traffic is supposed to be more critical, since it is treated as noise in URLLC decoding. As in (POPOVSKI, P. et al., 2018; SANTOS et al., 2020) the outage probability of URLLC under NOMA is

$$P_{\text{out}}^{\text{NOMA}}(G_{U,f}) = \Pr \left(\frac{1}{F_U} \sum_{f=1}^{F_U} \log_2 \left(1 + \frac{G_{U,f}}{1 + G_{B,f}^{\text{tar}}}\right) < R_U \right), \tag{16}$$

where eMBB interference is assumed to be always present in the URLLC decoding, due to their long period of activation. The URLLC achievable rate R_U^{NOMA} is numerically obtained by imposing the reliability constraint $P_{\text{out}}^{\text{NOMA}}(G_{U,f}) \leq \epsilon_U$.

2.5.3 Hybrid network slicing

In HMA network slicing, as portion of the channels operate under OMA and part in NOMA, the resulting rate for eMBB and URLLC is the summation of the OMA and NOMA partitions calculated before, with $F = F_N + F_B + F_U$, where F_N refers to the number of channels in NOMA partition, F_B is the quantity of eMBB users under OMA, and F_U is the number of channels in OMA slice for a given URLLC device.

The achievable sum-rate of a MMD-aided eMBB user in HMA is

$$\begin{aligned} R_B^{\text{HMA}} &= F_B \log_2\left(1 + G_{B,f}^{\text{tar},1}\right) + F_N \log_2\left(1 + G_{B,f}^{\text{tar},2}\right) \\ &\approx (F_B + F_N) \log_2\left(1 + G_{B,f}^{\text{tar},2}\right), \end{aligned} \tag{17}$$

where $G_{B,f}^{\text{tar},1}$ and $G_{B,f}^{\text{tar},2}$ are the target SNR defined in (14), but in $G_{B,f}^{\text{tar},2}$ the $G_{B,f}^{\text{min}}$ value is obtained from (15) for the NOMA slice. As previously discussed, the presence of URLLC devices in the same channel allocated to eMBB through MMD has minimal impact in eMBB detection $\left(G_{B,f}^{\text{tar},1} \approx G_{B,f}^{\text{tar},2}\right)$, making it possible to, instead of allocating an orthogonal slice to eMBB, share $F_B + F_N$ channels non-orthogonally with URLLC (the case of Fig. 7d). In this case, an URLLC device has always its transmission spread over all channels in the bandwidth ($F_U + F_N = F$), which can improve the value of R_U^{HMA}, since the frequency diversity factor will always be the maximum possible.

In the scenario of Fig. 7d, three eMBB users are connected to the network, occupying three non-orthogonal channels concurrently used by two URLLC devices that, besides using these three channels, also take advantage of one channel not used by eMBB to increase its frequency diversity, resulting in an orthogonal partition with one channel. Therefore, in the formulation of outage probability for URLLC under HMA, both orthogonal and non-orthogonal slices have to be considered, since the interference of eMBB users is relevant, and it is considered to be always present, resulting in

$$P_{\text{out}}^{\text{HMA}}(G_{U,f}) = \Pr\left(\frac{1}{F_U}\sum_{f=1}^{F_U} \log_2\left(1 + G_{U,f}\right) + \frac{1}{F_N}\sum_{f=1}^{F_N} \log_2\left(1 + \frac{G_{U,f}}{1 + G_{B,f}^{\text{tar},2}}\right) < R_U\right). \tag{18}$$

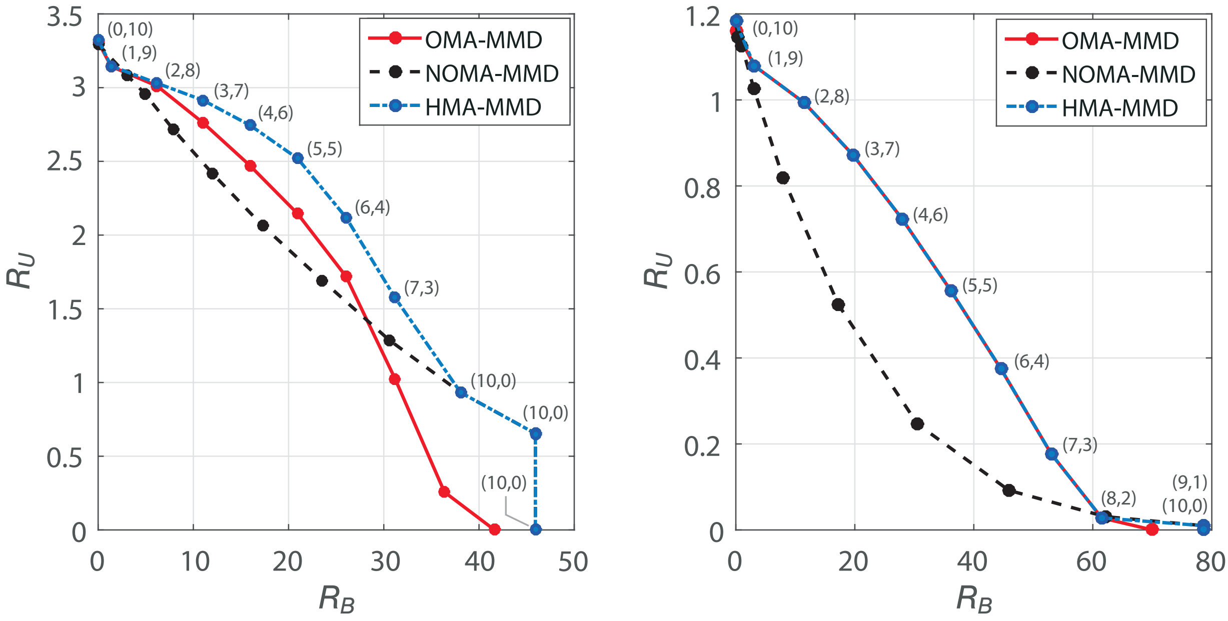
The URLLC achievable rate R_U^{HMA} is numerically obtained by imposing the reliability constraint $P_{\text{out}}^{\text{HMA}}(G_{U,f}) \leq \epsilon_U$.

2.6 NUMERICAL RESULTS

We resort to Monte Carlo numerical methods to evaluate the performance of the proposed HMA MMD-aided network slicing scheme. The results are obtained with the aid of MATLAB®, by averaging a number of 10^7 independent random runs for every particular scenario. Figs. 8a and 8b present the sum-rate pair for $S = 5$, $a_U = 0.1$, $F = 10$, $\epsilon_B = 10^{-3}$ and $\epsilon_U = 10^{-5}$, respectively for the cases $\bar{\gamma}_U > \bar{\gamma}_B$ and $\bar{\gamma}_B > \bar{\gamma}_U$. The curves of the OMA MMD-aided and NOMA MMD-aided from (SANTOS et al., 2020) are presented for comparison purposes. The labels represent the slicing combination (F_N, F_U) that optimizes the HMA sum-rate.

Considering the case when $\bar{\gamma}_U > \bar{\gamma}_B$, which may be a more realistic scenario given the URLLC reliability requirements, the HMA slicing with MMD allocation is capable of obtaining better results for R_B than OMA in the entire range, also presenting improved sum-rates over NOMA, mainly for high values of R_U, as shown in Fig. 8a. For $R_U \leq 1$, the values of R_B^{HMA} obtained are the same for R_B^{NOMA}, in that case, when focusing on higher rates for eMBB service, the slicing mechanism does not need to allocate specific channels for URLLC, thereby non-orthogonal division is prioritized. In this scenario, the resulting R_B calculated using (17) equals (13).

Figure 8 – Sum-rate region for eMBB and URLLC with $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$, $S = 5$, $a_U = 0.1$ and $F = 10$.



(a) $\bar{\gamma}_B = 10$ dB, $\bar{\gamma}_U = 20$ dB.

(b) $\bar{\gamma}_B = 20$ dB, $\bar{\gamma}_U = 10$ dB.

Source: The Author.

When $\bar{\gamma}_B > \bar{\gamma}_U$, the case of Fig. 8b, the HMA slicing with MMD applied to eMBB users equals the OMA MMD-aided performance for the majority range of values. Higher values of R_B are achieved only for really small values of R_U, matching the results found for NOMA MMD-aided. As the interference generated by eMBB devices is higher in this case, only the OMA partition in HMA scheme will give a significant result in the overall rate, thus the result obtained is equivalent of OMA MMD-aided.

Table 1 evaluates R_B for $\epsilon_U \in \{10^{-5}, 10^{-6}, 10^{-7}\}$, $R_U \in \{1, 2\}$ bits/s/Hz, $\bar{\gamma}_B = 10$ dB, and $\bar{\gamma}_U = 20$ dB. Considering the three columns of R_U fixed at 2 bits/s/Hz, it is possible to observe that R_B^{HMA} is always higher than R_B^{OMA} and R_B^{NOMA}. For $\epsilon_U = 10^{-5}$, it is obtained a percentage increase of 25.6 and 43.6 when compared to OMA and NOMA, respectively. For $\epsilon_U = 10^{-6}$, HMA can also achieve higher values like 20.5 bits/s/Hz when compared to NOMA with $\epsilon_U = 10^{-5}$, showing that the combination of

Table 1 – R_B *vs* ϵ_U, for $R_U \in \{1, 2\}$ bits/s/Hz.

ϵ_U	10^{-5}		10^{-6}		10^{-7}	
R_U	1	2	1	2	1	2
R_B^{OMA}	30.8	21.5	26.8	17.0	25.2	9.0
R_B^{NOMA}	36.9	18.8	34.1	15.0	30.0	10.0
R_B^{HMA}	36.9	27	34.1	**20.5**	30.0	**14**

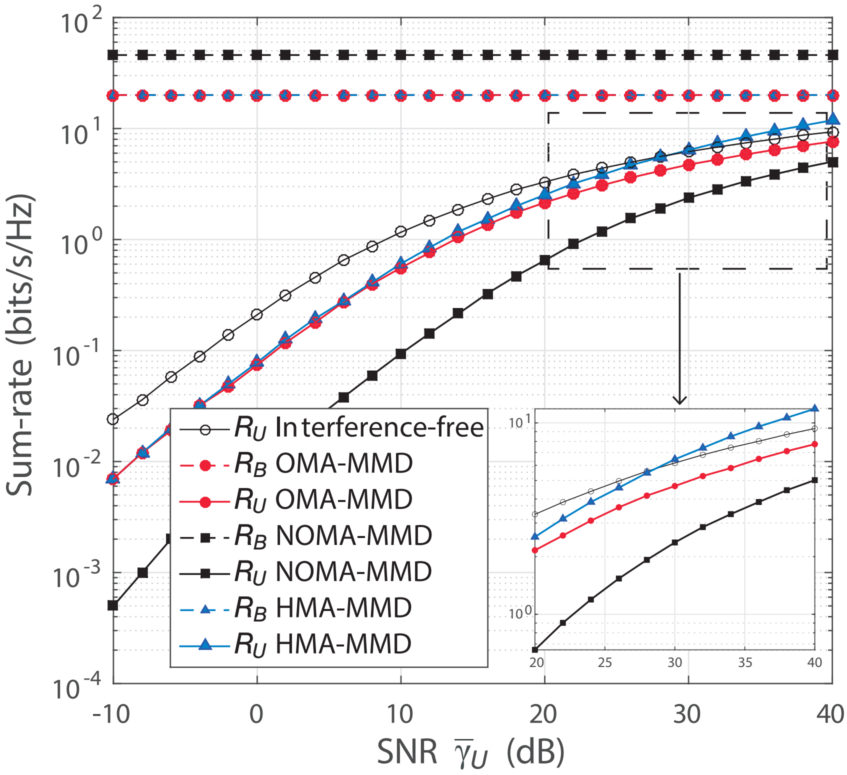
Source: The Author.

MMD and HMA benefits the eMBB rate and the reliability of URLLC service simultaneously, resulting in even larger values of R_B than the ones from (SANTOS et al., 2020). Moreover, it is possible to state that HMA network slicing, in the case of $\bar{\gamma}_U > \bar{\gamma}_B$, can increase the eMBB rate in even higher URLLC reliability levels $\left(\epsilon_U = 10^{-7}\right)$ when compared to the values obtained in (SANTOS et al., 2020) in that same level. For $R_U =$ 2 bits/s/Hz, $R_B^{\text{HMA}} = 14$ bits/s/Hz.

Fig. 9 presents the sum-rate, in different network slicing strategies, obtained for eMBB and URLLC when $\bar{\gamma}_U \in \{-10, \ldots, 40\}$ dB and $\bar{\gamma}_B$ is fixed at 10 dB. In the interference-free scenario, URLLC devices occupy all available channels ($F_U = F$) without interference of eMBB users ($F_B = 0$). For NOMA, $F_U = F_B = F = 10$, $F_B = F_U =$ 5 in OMA, $F_U = F_N = 5$ and $F_B = 0$ for HMA. Moreover, $S = 5$ and $a_U = 0.1$.

These curves give us the insight that, in high $\bar{\gamma}_U$ regimes, the HMA scheme can increase the value of R_U when compared to currently adopted OMA, without reducing the performance of eMBB, thanks to the contribution of the non-orthogonal slice. When compared to NOMA, it is also possible to increase URLLC rate, but only if we decrease R_B allowing fewer connections from this service. When compared to the interference-free, HMA can equals or even outperforms it for $\bar{\gamma}_U >= 30$ dB, while serving eMBB service in the same bandwidth. Besides, it is evident that the performance of eMBB does not change for different values of $\bar{\gamma}_U$, given the prior decodification of URLLC service. In the 20-30 dB range, the gains of the HMA method are 22.8-35.2% for R_U. For $\bar{\gamma}_U > 30$ dB, the URLLC rate increasing is up to 55.3%.

Figure 9 – URLLC and eMBB sum-rates for $\bar{\gamma}_U \in \{-10, \ldots, 40\}$ dB and $\bar{\gamma}_B = 10$ dB. The reliability levels are $\epsilon_B = 10^{-3}$ and $\epsilon_U = 10^{-5}$. Moreover, $S = 5$, $a_U = 0.1$, $F = 10$ in NOMA, $F_B = F_U = 5$ in OMA, $F_U = F_N = 5$ and $F_B = 0$ for HMA and $F_U = 10$ in Interference-free.



Source: The Author.

2.7 FINAL COMMENTS

We considered the network radio resource slicing between eMBB with channel allocation and URLLC users, comparing OMA and NOMA methods with the proposed HMA scheme. With this technique, the multiple access protocol could be dynamically adapted between OMA and NOMA to attend the services requisites. As the results showed, when focusing on the eMBB sum-rate, NOMA may be prioritized, however, for higher URLLC reliability levels or for moderate values of R_B and higher URLLC rates, a hybrid multiplexing is more suitable, mainly in scenarios where the URLLC devices have higher average SNRs than the eMBB users, which can be considered a realistic scenario given the type of applications that URLLC service is expected to address, such as mission-critical and Industry 4.0, that can be served by hot-spots or a nearby BS. This characteristic allows transmissions with large average channel gain to the target, then the high reliability requirement of URLLC traffic can be satisfied. On the other hand, the eMBB users are considered to be in arbitrary positions, maybe far from the BS or close to the cell boundaries, where high channel gain cannot be guaranteed. However, with channel allocation through MMD, this condition is compensated, and the service requirements are met. It was also shown that R_U can be enhanced without harming eMBB service when compared to the currently adopted OMA method. Furthermore, HMA MMD-aided can improve the eMBB rate under more restricted values of URLLC reliability, such as $\epsilon_U = 10^{-7}$, when compared to (SANTOS et al., 2020).

3 RATE-SPLITTING MULTIPLE ACCESS FOR URLLC IN SLICING WITH EMBB

This chapter discusses the system model, equations and results related to the application of the RSMA method to URLLC uplink in physical layer network slicing with eMBB users in a single-cell network.

3.1 LITERATURE REVIEW

Recently, several works studied different RSMA implementations in downlink wireless networks (CAO; YEH, 2007; JOUDEH; CLERCKX, 2016a, 2016b; MAO et al., 2019; CLERCKX et al., 2020), showing that RSMA can improve downlink rate and quality of service, achieving better performance than both NOMA and SDMA. For uplink RSMA systems, authors from (YANG, Z. et al., 2019, 2020b) study the problem of maximizing the sum-rate under proportional rate constraints for all users, by setting users transmission power and optimizing the decoding order at the BS through exhaustive search. As a result, they show that RSMA achieves better performance than NOMA and OMA techniques, such as Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA). However, the proposed strategy requires *a priori* CSI, not being in general applicable to URLLC users due to latency constraints. In (ZHU, Y. et al., 2017), the authors propose the use of RSMA to reduce the scheduling complexity of NOMA, since the transmission splitting by default diversifies the arriving power at the BS, avoiding the need of user pairing. In (LIU, H.; KWAK, 2019), the authors apply rate splitting to a pair of users under power-domain NOMA, considering that one of them is near the BS, while the other is far from the BS. Two techniques are studied, namely, Fixed Rate Splitting (FRS) and Adaptive Rate Splitting (ARS), where the power allocation factor that splits the messages of the near user can be fixed or dynamically designed based on CSI, respectively. This work is then extended in (LIU, H. et al., 2020), adopting cyclic prefixed single carrier transmissions. In both works, rate splitting has been shown to achieve superior outage performance when compared to NOMA.

In (ZENG et al., 2019), an exhaustive-search rate splitting algorithm was proposed to guarantee max-min fairness in Single-Input Multiple-Output (SIMO) NOMA networks, aiming at maximizing the minimum data rate and reduce the scheduling process. The receiver combines Minimum Mean Squared Error (MMSE) with SIC to identify the optimal detection order based on CSI. Results showed that rate splitting has higher minimum data rate and lower transmission latency than SIMO-OMA and SIMO-NOMA. The use of rate splitting in user cooperation networks is proposed in (ABBASI; YANIKOMEROGLU, 2021). Each user transmits its signal and receives the transmitted signal of the other user in the first mini-slot and, in the second mini-slot, relays the other user's message with amplify-and-forward protocol. The rate is split between mini-slots, generating space diversity at the uplink and consequently increasing reliability. At the receiver, Maximum Ratio Combining (MRC) is used to combine the received signals and SIC is applied to decode the superposed signal. Results prove that cooperative RSMA outperforms cooperative OMA and NOMA.

In scenarios with spectrum sharing among URLLC and eMBB services, several works compared OMA and NOMA network slicing (POPOVSKI, P. et al., 2018; ANAND et al., 2018; ABREU et al., 2019; ALSENWI et al., 2019; KORRAI et al., 2019; KASSAB et al., 2019). However, none of the aforementioned works considers multiple concurrent URLLC users in the same resource block. In (TOMINAGA et al., 2021b), URLLC users are assumed to share time and frequency resources through NOMA, in both OMA and NOMA slicing with eMBB service. It was shown that NOMA can leverage the URLLC sum-rate in some cases, considering that the SIC process is capable of attending the communication latency. Authors from (DIZDAR et al., 2021a) apply RSMA to URLLC in the downlink, showing its superior performance in terms of latency, allowing shorter block lengths. However, no interference from other services is considered.

3.1.1 Novelty and Contribution

Motivated by the above literature, in this work we focus on increasing the URLLC spectral efficiency, allowing non-orthogonal sharing of frequency and time resources through rate-splitting for URLLC users. In the proposed scheme, which we refer to as U-RSMA, the benefits of RSMA, SIC decoding and frequency diversity are combined, in both OMA and NOMA slicing with eMBB. The proposed U-RSMA scheme is then compared to the so-called U-NOMA and U-OMA schemes, where the multiple access between URLLC devices is performed by means of NOMA and OMA, respectively. In U-OMA, URLLC users only share the mini-slot, staying isolated from each other in the frequency domain. This could decrease its reliability, since the frequency diversity is reduced, however users may benefit from less interference.

To characterize the performance of eMBB and URLLC users, we evaluate each service sum-rate in different scenarios. To the best of our knowledge, this work is the first to apply RSMA to URLLC uplink transmission in a network slicing scenario, showing that RSMA can outperform OMA and NOMA methods for URLLC service even in the presence of eMBB interference, specially for very strict reliability levels.

3.2 SYSTEM MODEL

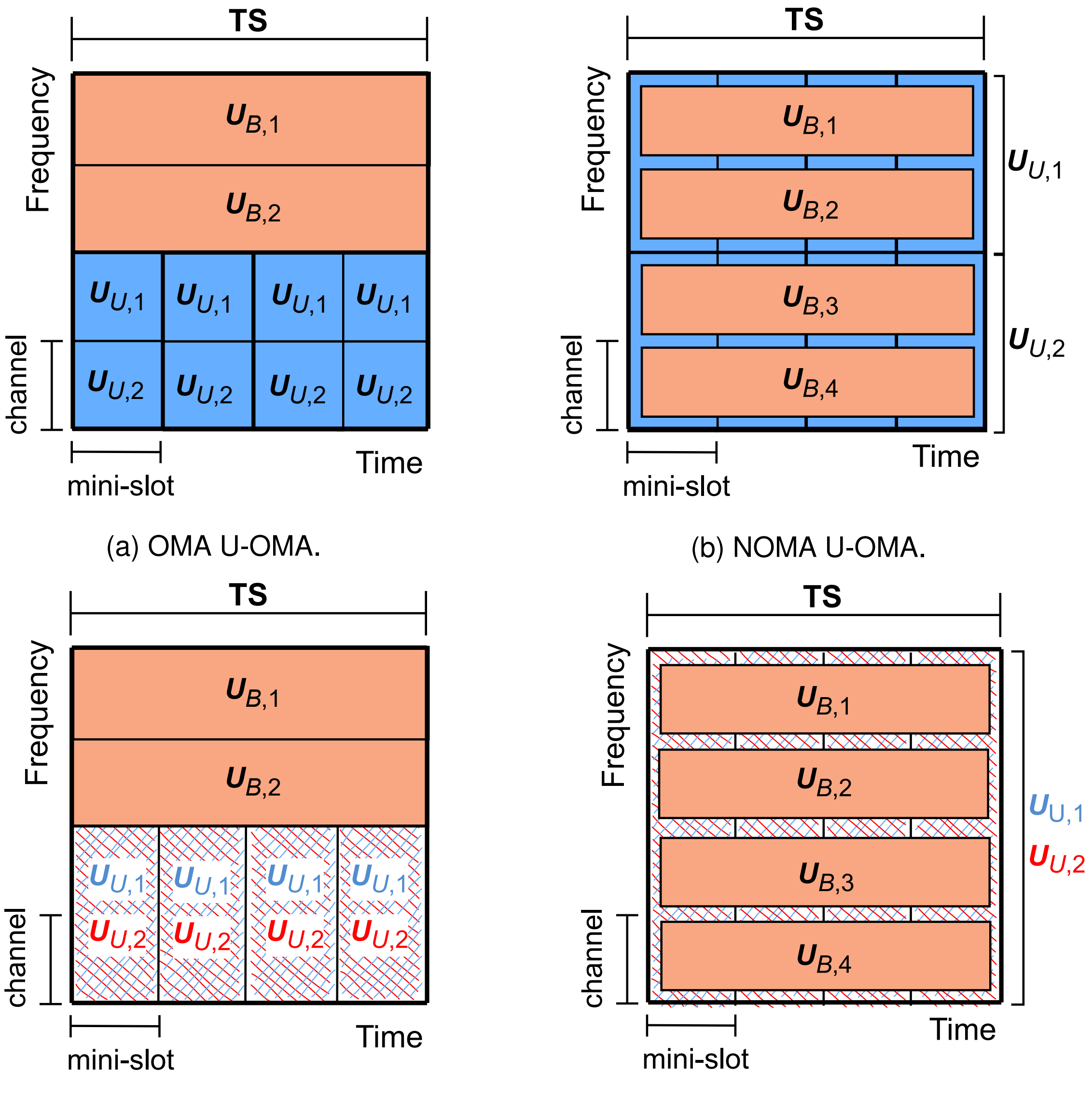
Similar to the system model presented in Chapter 2, here we evaluate the uplink of multiple eMBB and URLLC users when communicating to a common BS in a single-cell network with shared radio resources. The bandwidth is divided into F channels of index $f \in \{1, \ldots, F\}$ subject to i.i.d. Rayleigh fading. The fading realization observed by each device is uncorrelated from another due to the assumption that all devices have a large enough spatial separation. Furthermore, the fading is considered constant during one transmission TS, i.e., a block fading model where the TS is considered to be within the channel coherence time since its length is fairly small (TSE; VISWANATH, 2005b). As we assume that the average transmission power of all devices and the noise power at the BS are normalized to one, the received power equals the SNR for each device. Moreover, the channel fading realization for user $i \in \{B, U\}$ in channel f is $h_{i,f} \sim \mathcal{CN}(0, \bar{\gamma}_i)$, following a circular-symmetric complex Gaussian distribution, where $\bar{\gamma}_i$ corresponds to the average SNR, being $G_{i,f} \triangleq |h_{i,f}|^2$ the channel gain, and where subscripts B and U refer to eMBB and URLLC devices, respectively. The number of channels allocated to user i is $F_i \leq F$, with $i \in \{B, U\}$. Moreover, each TS is divided into S mini-slots, as considered in low latency scenarios (JI et al., 2018).

In accordance to (POPOVSKI, P. et al., 2018), we assume that an eMBB user is active with probability a_B and during the entire TS, occupying one random frequency channel f among F_B available channels. Furthermore, we model only the transmission phase, assuming that radio access and competition among eMBB devices have been resolved prior to the considered time slot, as usual in wireless cellular networks. Thus, the number of eMBB devices able to transmit in such TS is equal to the number of channels F_B. Moreover, we suppose that the eMBB devices and the BS have CSI as currently implemented in wireless standards such as LTE and 5G New Radio (TAKEDA et al., 2023; ZAIDI, A.; WANG, Z., 2023; 3GPP, 2023). Although channel estimation errors can occur in practice, for simplicity we consider a perfect CSI scenario in this work, as widely considered in the literature (ZAIDI, A. A. et al., 2023; POPOVSKI, P. et al., 2018). In contrast, an URLLC device spreads its transmission over $F_U \leq F$ channels to increase the reliability with the aid of frequency diversity, and sends, with some activation probability a_U, the entire information in only one mini-slot (the smallest time unit in our model) that was pre-assigned to meet latency requirements. We also consider that the protocol block length, which should be considered finite given the short transmissions, is long enough to justify an asymptotic information-theoretic formulation (YANG, W. et al., 2014). Moreover, in each mini-slot we have a maximum number of n_U users that share the resources following three distinct methods: U-OMA, U-NOMA and U-RSMA.

Different from eMBB users, we assume that the BS has no knowledge about the URLLC channel, given the high latency requirement which does not allow the exchange of reference signals for CSI acquisition. However, we do consider in U-RSMA that the BS sends (*e.g., in a synchronization mini-slot transmitted at the end of each TS*), the optimal power splitting factor based on $\bar{\gamma}_U$ from a look-up table, which results in power adaptation for the user that performs the splitting. Despite that, the overall transmission power is the same as in U-OMA and U-NOMA cases.

A time-frequency grid is illustrated in Fig. 10, considering that the heterogeneous URLLC and eMBB traffics are sliced in an OMA (Figs. 10a and 10c), and NOMA (Figs. 10b and 10d) fashion. In this example, $S=4$ is the quantity of mini-slots in the time domain, whereas $F=4$ is the total number of channels available in the bandwidth.

Figure 10 – System model with $F = 4$ channels and $S = 4$ mini-slots, composed by eMBB and URLLC users. Services are sliced in (a) (c) Orthogonal and (b) (d) Non-Orthogonal multiple access schemes.



(a) OMA U-OMA.

(b) NOMA U-OMA.

(c) OMA U-NOMA or U-RSMA.

(d) NOMA U-NOMA or U-RSMA.

Source: The Author.

Considering the OMA scenario, two channels are allocated to URLLC ($F_U=2$) and two for eMBB ($F_B=2$). There are $n_U=2$ URLLC active users, $U_{U,1}$ and $U_{U,2}$, in each mini-slot that spread their transmission over one channel, in the case of U-OMA, or over two channels when considering U-NOMA or U-RSMA, without interference from eMBB users. On the eMBB band, there are also two users, $U_{B,1}$ and $U_{B,2}$, connected to the BS. When considering NOMA, all four channels are available for both services ($F = F_U = F_B = 4$), which implies a multi-service interference, turning the detection at the BS more complex and prone to errors. The frequency diversity gain for URLLC users is higher in this case, and, as this device type does not necessarily transmit at every TS, the spectrum efficiency should increase because eMBB users can occupy a radio resource that might be unused for long periods, which is represented with the inclusion of new eMBB users $U_{B,3}$ and $U_{B,4}$.

3.3 OUTAGE FORMULATION AND SLICING SCHEMES

In this section, we discuss the achievable rates of the different services and slicing schemes.

3.3.1 eMBB

A given eMBB device transmits, with a certain instantaneous power and data rate, in the randomly allocated dedicated radio resource $f \in \{1, \ldots, F_B\}$, if the instantaneous channel gain is greater than a threshold SNR $G_{B,f}^{\min}$. This decision is made based on CSI. One can obtain the eMBB rate as

$$R_B^{\text{orth}} = \log_2\left(1 + G_{B,f}^{\text{tar}}\right), \qquad \text{(bits/s/Hz)} \tag{19}$$

where

$$G_{B,f}^{\text{tar}} = \frac{\bar{\gamma}_B}{\Gamma\left(0, \frac{G_{B,f}^{\min}}{\bar{\gamma}_B}\right)}, \tag{20}$$

and $G_{B,f}^{\min} = -\bar{\gamma}_B \ln(1 - \epsilon_B)$. For proof, please refer to Appendix A.

3.3.2 URLLC

3.3.2.1 U-OMA

The F_U channels available for URLLC are divided in n_U orthogonal slices with F_U' channels reserved to each $U_{U,n}$ user, with $n \in \{1, \ldots, n_U\}$. The outage probability of $U_{U,n}$, in the absence of interference from other services, is (POPOVSKI, P. et al., 2018)

$$P_{\text{out}}^{\text{U-OMA}}(E_U) = \Pr\left(\frac{1}{F_U'}\sum_{f=1}^{F_U'} \log_2(1 + \sigma_{n,f}) < R_{U,n}\right), \tag{21}$$

where $\sigma_{n,f}$, the Signal-to-Interference-plus-Noise Ratio (SINR) of the n-th active user in the frequency channel f, equals $G_{U,n,f}$, since for the moment there is no interference from other users. The target rate $R_{U,n}$ is numerically obtained by imposing the outage probability requirement $P_{\text{out}}^{\text{U-OMA}}(E_U) \leq \epsilon_U$ to (21). Thus, the sum-rate of the URLLC service is given by

$$R_U^{\text{U-OMA}} = \sum_{n=1}^{n_U} R_{U,n}. \tag{22}$$

3.3.2.2 U-NOMA

In U-NOMA, URLLC users share the F_U channels available in each mini-slot and the BS performs SIC to decode the multiple messages, which outperforms other techniques of multi-user detection, such as puncturing and erasure decoding (POPOVSKI, P. et al., 2018), and is a general receiver structure for non-orthogonal uplink (LIU, Yuanwei et al., 2017). As an user occupies more than one channel, we cannot simply define the decoding order in terms of the channel gain magnitude. Instead, the BS can order the users according to their mutual information (TOMINAGA et al., 2021b)

$$\mathbb{I}_n^{\text{sum}} = \sum_{f=1}^{F_U} \log_2(1 + \sigma_{n,f}), \tag{23}$$

where $\sigma_{n,f}$ is defined as

$$\sigma_{n,f} = \frac{G_{U,n,f}}{1 + \sum_{j>n}^{n_U} G_{U,j,f}}. \tag{24}$$

The decoding procedure starts with the strongest among all the active users in the current mini-slot. If correctly decoded, it is removed from the received signal and the operation continues, until a user cannot be decoded (an event that occurs with probability ϵ_U) or all users have been properly decoded. We consider that the BS is capable of decoding the n_U users within the mini-slot period, since each transmission carries a different message and the procedure must attend the latency requirement. The outage probability of the u-th user is

$$P_{\text{out}}^{\text{U-NOMA}}(E_U) = \Pr\left(\frac{1}{F_U}\sum_{f=1}^{F_U} \log_2(1 + \sigma_{n,f}) < R_{U,n}\right). \tag{25}$$

The target rate $R_{U,n}$ is numerically obtained by imposing the reliability requirement $P_{\text{out}}^{\text{U-NOMA}}(E_U) \leq \epsilon_U$ to (25). Thus, the sum-rate of the URLLC service is

$$R_U^{\text{U-NOMA}} = \sum_{n=1}^{n_U} R_{U,n}. \tag{26}$$

3.3.2.3 U-RSMA

Either under U-OMA or U-NOMA, URLLC users directly transmit their data to the BS once they are active. However, in U-RSMA, a user may first split its information into two sub-messages, creating the concept of “virtual users”. Each sub-message has transmission power defined by the so-called splitting factor $\alpha \in [0, 1]$.

As an example, let us consider the case with $n_U = 2$. In this two-user scenario, we assume that only one user, say $U_{U,1}$, splits its message[1], creating two virtual users referred to as $U_{U,1,1}$ and $U_{U,1,2}$. Without loss of generality, we consider that $U_{U,1,1}$ is always decoded before $U_{U,1,2}$. In this scenario, we have three possible decoding orders at the BS, namely: (i) $U_{U,1,1} \rightarrow U_{U,2} \rightarrow U_{U,1,2}$; (ii) $U_{U,1,1} \rightarrow U_{U,1,2} \rightarrow U_{U,2}$; and (iii) $U_{U,2} \rightarrow U_{U,1,1} \rightarrow U_{U,1,2}$, such that the proper decoding order is chosen based on the sum of mutual information from (23), similarly to U-NOMA.

While the decoding orders (ii) and (iii) achieve the same results of U-NOMA with $U_{U,1} \rightarrow U_{U,2}$ and $U_{U,2} \rightarrow U_{U,1}$, respectively (LIU, H. et al., 2020), it has been shown that (i) represents the optimal decoding order of RSMA (YANG, Z. et al., 2020b). Thus, in the SIC process, the receiver first attempts to decode a (virtual) user while regarding all the remaining messages as noise. Once the decoding is successful, its interference is removed out of the superimposed received signal, and the receiver then attempts to decode the next message following the pre-established decoding order. Upon adopting the decoding order from (i), the SINR of the virtual user $U_{U,1,1}$ is

$$\sigma_{1,1,f} = \frac{\alpha\, G_{U,1,1,f}}{1 + G_{U,2,f} + (1-\alpha) G_{U,1,2,f}}. \tag{27}$$

If $U_{U,1,1}$ is correctly decoded and canceled from the received signal, the SINR of $U_{U,2}$ becomes

$$\sigma_{2,f} = \frac{G_{U,2,f}}{1 + (1-\alpha) G_{U,1,2,f}}. \tag{28}$$

Finally, the SINR of the remaining virtual user $U_{U,1,2}$, subject to the correct decoding of the previous users, is

$$\sigma_{1,2,f} = (1-\alpha) G_{U,1,2,f}. \tag{29}$$

Then, the achievable rates of U-RSMA can be calculated from (25), by substituting $\sigma_{n,f}$ with the SINRs of U-RMSA presented in (27)-(29). The final rate of user $U_{U,1}$ is $R_{U,1} = R_{U,1,1} + R_{U,1,2}$. Thus, the sum-rate of the two-user U-RSMA URLLC service finally obtained as

$$R_U^{\text{U-RSMA}} = R_{U,1} + R_{U,2}. \tag{30}$$

It is worthy mentioning that, when compared to U-NOMA, U-RSMA requires an extra round in the SIC procedure, increasing the complexity of the decoding process.

3.3.3 Orthogonal network slicing

In Sections 3.3.1 and 3.3.2 we present, respectively, the achievable rates of eMBB and URLLC services when operating in standalone mode, without slicing the

[1] Following (RIMOLDI; URBANKE, 1996), only one out of the two users needs to split its message in order to achieve the capacity region.

network resources. When such slicing between the heterogeneous eMBB and URLLC services is designed in a orthogonal fashion, they are "isolated" from each other, thus for URLLC the only source of interference are the n_U users active with probability a_U in certain mini-slot occupying all $F_U \leq F$ channels, whereas eMBB experiences an interference-free scenario since users are allocated orthogonally within the remaining $F_B = F - F_U$ channels. The OMA performance is measured in terms of the sum-rate pair $\left(R_B^{\text{sum}}, R_U^{\text{sum}}\right)$, where R_B^{sum} can be defined as (POPOVSKI, P. et al., 2018)

$$R_B^{\text{sum}} = F_B \, R_B^{\text{orth}}, \tag{31}$$

where R_B^{orth} comes from (19) and R_U^{sum} is computed as presented in Section 3.3.2 for each particular multiple access method adopted by the URLLC service.

3.3.4 Non-orthogonal network slicing

In non-orthogonal slicing, eMBB and URLLC services simultaneously share all the F available channels, i.e., $F_B = F_U = F$. Due to latency and reliability constraints, it is assumed that the BS always attempts to decode the n_U active URLLC devices first, through SIC, while treating the eMBB traffic as interference. Therefore, the interference from URLLC transmissions into eMBB (and vice-versa) needs to be considered.

For eMBB users, the achievable rate in NOMA slicing is obtained following the same logic discussed in Appendix C, however, without considering the MMD allocation and with $n_U \times S$ URLLC users. This implies that

$$P_{\text{out}} \leq (1 - a_U)^S (1 - a_B) + \left(1 - (1 - a_U)^S\right) (\epsilon_U + (1 - \epsilon_U)(1 - a_B)), \tag{32}$$

can be rewritten as

$$P_{\text{out}} \leq (1 - a_U)^{n_U S} (1 - a_B) + \left(1 - (1 - a_U)^{n_U S}\right) \left(\epsilon_U + (1 - \epsilon_U)(1 - a_B)\right), \tag{33}$$

and the eMBB reliability constraint is imposed as $P_{\text{out}} \leq \epsilon_B$. Then, one can rewrite (33) as

$$a_B \geq \frac{1 - \epsilon_B}{1 - \epsilon_U \left(1 - (1 - a_U)^{n_U S}\right)}. \tag{34}$$

Having in mind that $a_B = \exp[-G_{B,f}^{\min}/\bar{\gamma}_B]$, it is possible to isolate the threshold SNR $G_{B,f}^{\min}$ from (34), resulting in

$$G_{B,f}^{\min} \leq -\bar{\gamma}_B \ln \left(\frac{1 - \epsilon_B}{1 - \epsilon_U \left(1 - (1 - a_U)^{n_U S}\right)} \right). \tag{35}$$

The target SNR $G_{B,f}^{\text{tar}}$ is obtained similarly to (20) as

$$G_{B,f}^{\text{tar}} \leq \frac{\bar{\gamma}_B}{\Gamma\left(0, \frac{G_{B,f}^{\text{min}}}{\bar{\gamma}_B}\right)}. \tag{36}$$

However, in the non-orthogonal case, $G_{B,f}^{\text{min}}$ is bounded by (35). Therefore, the maximum achievable rate of an eMBB device in NOMA is $R_B^{\text{n-orth}} = \log_2(1 + G_{B,f}^{\text{tar}})$.

The threshold from (35) indicates that the impact of URLLC transmissions in the eMBB decoding should be minimal, due to the fact that, by definition, $\epsilon_U << \epsilon_B$, which implies that a_B is close to $1 - \epsilon_B$. On the other hand, the eMBB interference in the URLLC traffic is supposed to be more critical, since URLLC is decoded prior to eMBB. As in (POPOVSKI, P. et al., 2018) the outage probability of URLLC under NOMA is

$$P_{\text{out}}^{\text{NOMA}}(E_U) = \Pr\left(\frac{1}{F_U}\sum_{f=1}^{F_U}\log_2\left(1 + \frac{\sigma_{n,f}}{1 + G_{B,f}^{\text{tar}}}\right) < R_{U,n}\right), \tag{37}$$

where it is assumed that the interference of eMBB is always present in the URLLC decoding.The value of $\sigma_{n,f}$ depends on the multiple access technique used by URLLC users, as discussed in Section 3.3.2. The URLLC achievable sum-rate R_U^{sum} is then numerically obtained by imposing the reliability constraint $P_{\text{out}}^{\text{NOMA}}(E_U) \leq \epsilon_U$, where the rates are separately calculated for all n_U transmitting URLLC users.

3.4 NUMERICAL RESULTS

In this section, we present some numerical results aiming at comparing the sum-rate performance of U-OMA, U-NOMA and U-RSMA under both OMA and NOMA network slicing strategies. These results were generated using Monte Carlo simulations in MATLAB®, where, for each particular scenario, we average a number of 10^7 independent random runs. Herein, we consider only the case of $n_U = 2$, as increasing the number of SIC iterations would consequently increase the URLLC latency. In U-RSMA, user $U_{U,1}$ splits its transmission according to α (which is optimized in each simulation step), creating two virtual users, namely $U_{U,1,1}$ and $U_{U,1,2}$. Furthermore, users that belong to the same service have the same average SNR, since we consider they are running identical applications. We consider that in each mini-slot there are always two URLLC users connected, i.e., $a_U = 1$ for each one of them, thus $F_U' = F_U/2$. Also, the number of eMBB users is F_B, equaling the number of channels available for the service. Moreover, one TS is composed of $S = 5$ mini-slots and the bandwidth is divided into $F = 8$ channels. The reliability requirement of eMBB service is $\epsilon_B = 10^{-3}$. For URLLC under U-OMA, the reliability is $\epsilon_U^{\text{U-OMA}} = 10^{-5}$, however, as for U-NOMA and U-RSMA the receiver employs SIC, we follow (DIZDAR et al., 2021a) and set the reliability target as $\epsilon_U^{\text{U-NOMA}} = \epsilon_U^{\text{U-RSMA}} = 5 \times 10^{-6}$ to ensure that the overall reliability does not exceed

Table 2 – Simulation parameters.

Parameter	Value
Number of URLLC users (n_U)	2
eMBB activation probability (a_B)	1
URLLC activation probability (a_U)	1
Number of channels (F)	8
Number of mini-slots (S)	5
eMBB reliability (ϵ_B)	10^{-3}
URLLC reliability in U-OMA ($\epsilon_U^{\text{U-OMA}}$)	10^{-5}
URLLC reliability in U-NOMA ($\epsilon_U^{\text{U-NOMA}}$)	5×10^{-6}
URLLC reliability in U-RSMA ($\epsilon_U^{\text{U-RSMA}}$)	5×10^{-6}
URLLC average SNR ($\bar{\gamma}_U$)	20 dB
eMBB average SNR ($\bar{\gamma}_B$)	10 dB

Source: The Author.

10^{-5}. Unless stated otherwise, we set $\bar{\gamma}_U = 20$ dB and $\bar{\gamma}_B = 10$ dB. Table 2 summarizes the simulation parameters.

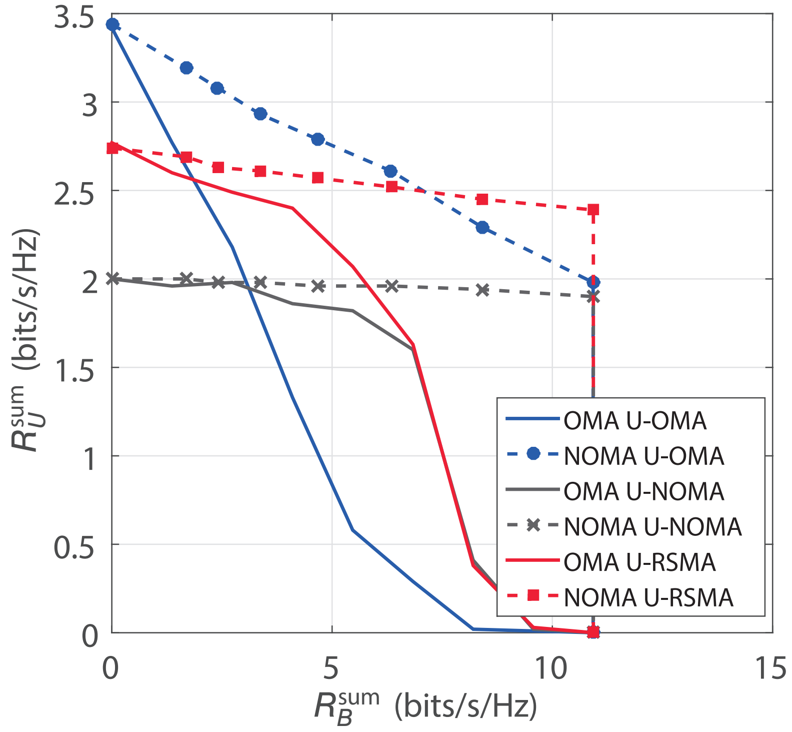
In Fig. 11 we plot the sum-rate pair $\left(R_B^{\text{sum}}, R_U^{\text{sum}}\right)$ for OMA and NOMA network slicing with URLLC operating under U-OMA, U-NOMA, and U-RSMA schemes. Comparing the NOMA slicing curves, U-OMA presents the highest rate pair values until $R_B^{\text{sum}} \approx 7$ bits/s/Hz, from where U-RSMA outperforms the other methods. Interestingly, R_U^{sum} remains almost constant as we increase R_B^{sum} in U-RSMA and U-NOMA, making these methods good options to achieve higher eMBB rates. In OMA slicing, U-OMA is the best method until $R_B^{\text{sum}} \approx 2.3$ bits/s/Hz, after that, U-RSMA achieves higher rates, presenting almost the same results as U-NOMA for high R_B^{sum} values.

Fig. 12 shows the URLLC sum-rate for different values of power splitting factor α. Note that, as expected, in U-OMA and U-NOMA we obtain constant values, since there is no message splitting. For U-RSMA, on the other hand, it is possible to observe that, as α increases, R_U^{sum} also increases, reaching the highest value when $\alpha = 0.8$ for NOMA and $\alpha \approx 0.75$ for OMA slicing.

The rates of users $U_{U,1}$ and $U_{U,2}$ when operating under U-RSMA are presented in Fig. 13, for both OMA and NOMA slicing. We see that $U_{U,1}$, the user that performs rate splitting, is capable of reaching higher rates when compared to $U_{U,2}$. Also, NOMA slicing is the best choice for this setup, achieving higher rates.

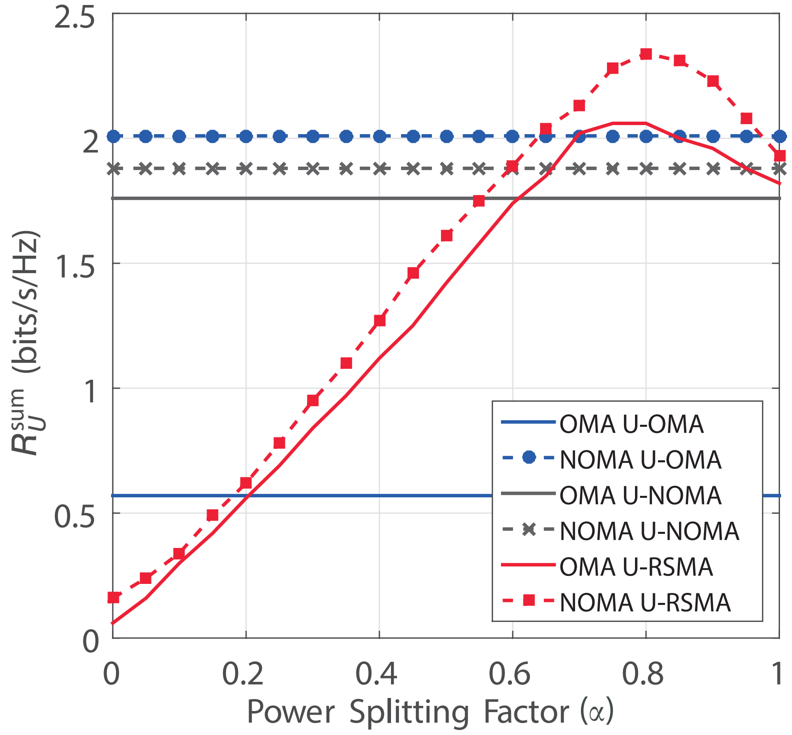
We consider that, during one TS, each eMBB user has the same target rate, since the channel gain is constant during this period over all channels. However, for URLLC, not imposing this requirement is beneficial, since different decoding orders provided by U-RSMA enable $U_{U,1}$ to reach higher rates, contributing to leverage the overall sum-rate, as shown in Figs. 14c and 14d, where we plot the URLLC per-user rate for $\bar{\gamma}_U \in \{0, \ldots, 20\}$ dB. Comparing U-RSMA and U-NOMA sum-rates in Figs. 14a and

Figure 11 – Sum-rate region in OMA and NOMA slicing with URLLC under U-OMA, U-NOMA, and U-RSMA schemes.



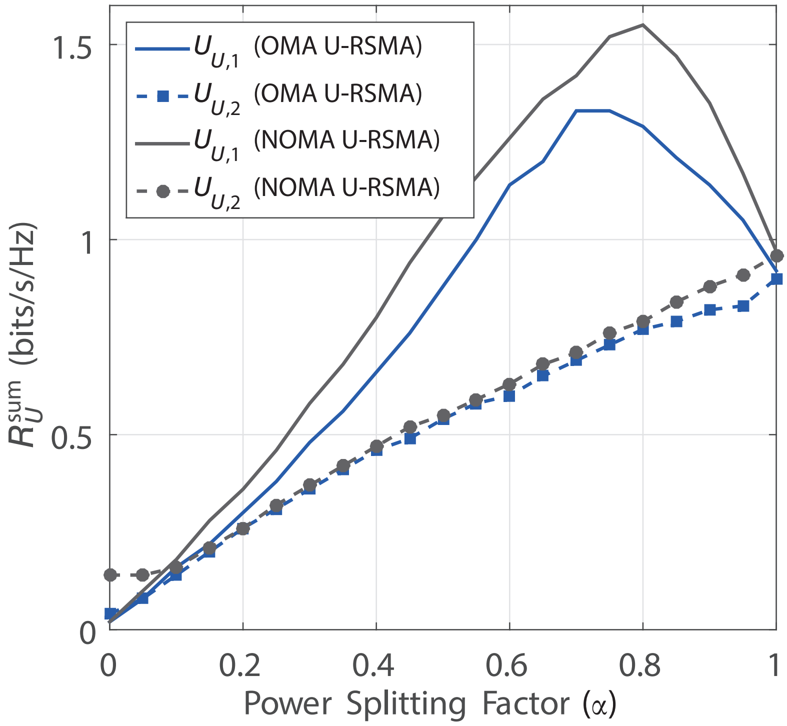
Source: The Author.

Figure 12 – URLLC sum-rate under U-OMA, U-NOMA, and U-RSMA schemes in OMA and NOMA slicing for $\alpha \in \{0, \ldots, 1\}$ and $F_U = F_B = 4$ in OMA.



Source: The Author.

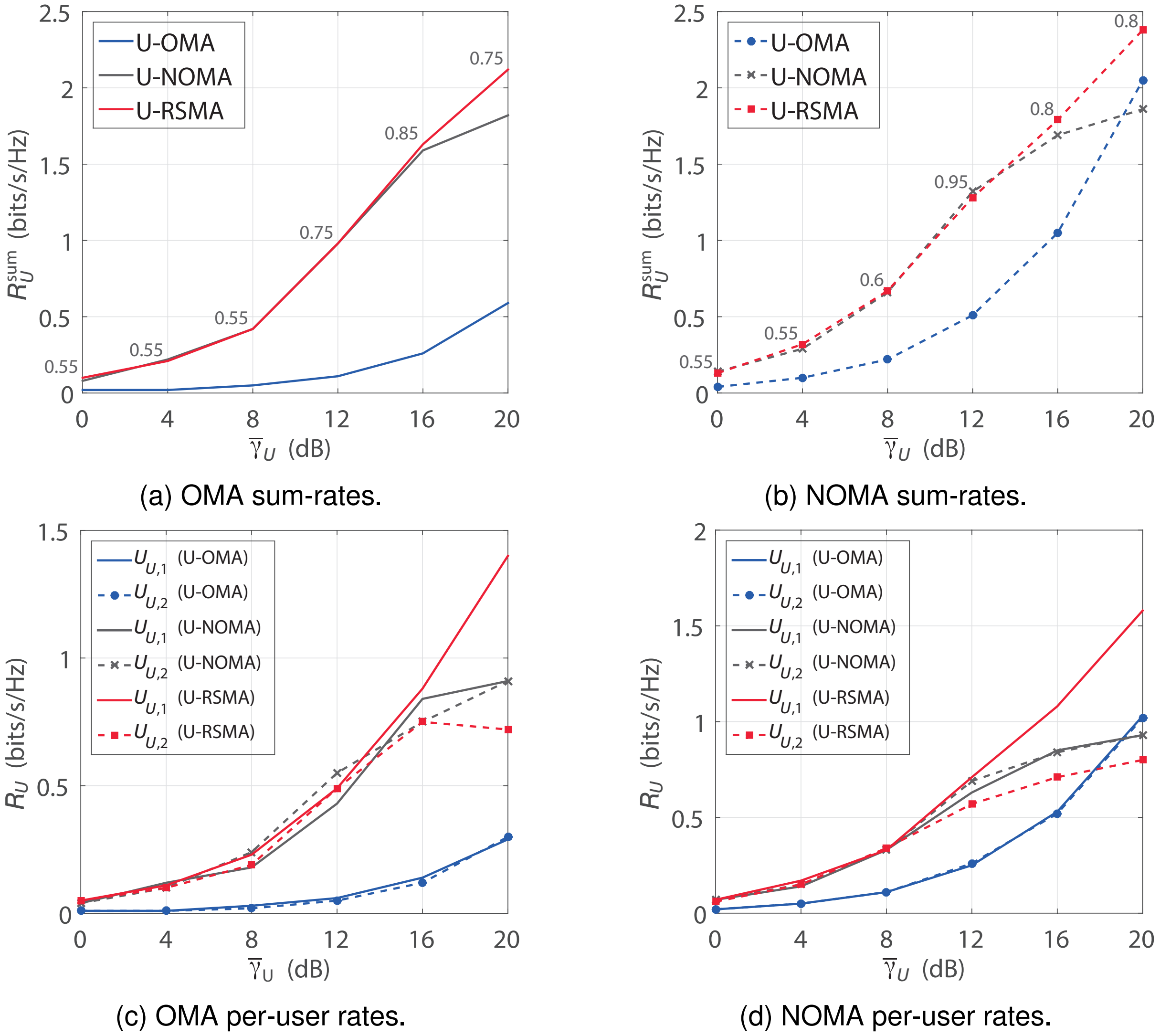
Figure 13 – URLLC per-user-rate under U-RSMA in OMA and NOMA slicing for $\alpha \in \{0, \ldots, 1\}$ and $F_U = F_B = 4$ in OMA.



Source: The Author.

14b, we see that the former is capable of operating with less performance degradation as the SNR increases, due to the fact that it is capable of handling the interference better, while the latter saturates as the SIC procedure fails to eliminate the interference.

Figure 14 – URLLC sum-rate and per-user rate under U-OMA, U-NOMA, and U-RSMA schemes in OMA and NOMA slicing for $\bar{\gamma}_U \in \{0, \ldots, 20\}$ dB and $F_U = F_B = 4$ in OMA.



(a) OMA sum-rates.

(b) NOMA sum-rates.

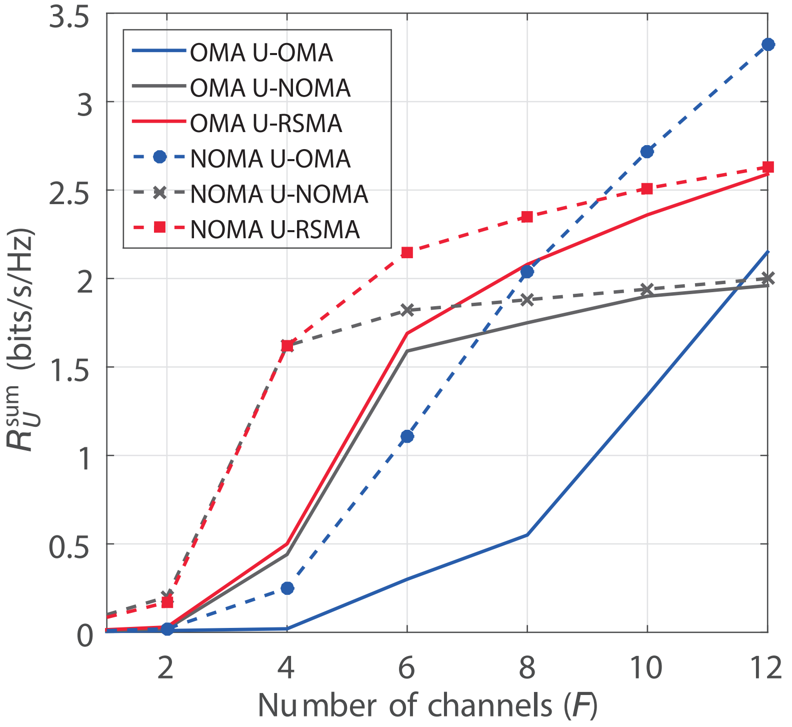
(c) OMA per-user rates.

(d) NOMA per-user rates.

Source: The Author.

From Fig. 15, considering the case of NOMA slicing, we conclude that U-OMA needs more bandwidth to outperform other methods, which is a limiting factor. Moreover, U-RSMA is the better choice for smaller chunks of spectrum, resulting in higher spectral efficiency since we can transmit more data with less bandwidth. In OMA, U-RSMA is better than other methods in all the evaluated range.

Figure 15 – URLLC sum-rate under U-OMA, U-NOMA, and U-RSMA schemes in OMA and NOMA slicing for $F \in \{1, \ldots, 12\}$.



Source: The Author.

3.5 FINAL COMMENTS

In this chapter, we considered the problem of radio resource slicing between eMBB and multiple URLLC devices. We evaluated the sum-rate performance of three multiple access methods for URLLC, namely U-OMA, U-NOMA, and U-RSMA, when operating under both OMA and NOMA network slicing strategies. Our results show that U-RSMA can achieve higher rates when the power splitting factor is properly configured, even with strict reliability requirements. Moreover, we show that non-orthogonal network slicing is capable of reaching the highest pair of rates for URLLC and eMBB simultaneously. This leads us to show another interesting scenario in which combining U-RSMA and NOMA is a powerful tool to attend B5G demands. The practical implementation of RSMA is still evolving, however, for our scenario, limiting the number of users in each mini-slot is a good strategy to reduce the decoding complexity and delay. Also, it is necessary to add a few bits of information in the message intended to the user that splits its data, sent in the synchronization slot to set the power allocation factor based on the average SNR. As some future research topics, the impact of imperfect CSI and applying rate-splitting among users of different services could be investigated.

4 RATE-SPLITTING MULTIPLE ACCESS FOR SEMI-GRANT-FREE TRANSMISSIONS WITH MMD-AIDED GRANT-BASED USERS

In this chapter, we apply the MMD method to assign GB users to channels. In the same network resource that serves a GB user, one GF user is allocated through a distributed contention protocol that considers the QoS of the GB user, performing the pairing that causes the minimum interference, so that the presence of a GF user is transparent to the GB user. Both users admitted to the same physical resource operate in either NOMA or RSMA, depending on the channel gains. Exact expressions for the outage probability of this SGF system are provided and compared to simulation results.

4.1 LITERATURE REVIEW

Usually, GF and GB users are allocated to orthogonal resources, operating without multiple-service interference. However, to increase the spectral efficiency, SGF transmissions could be applied to opportunistically admit GF users on the resource blocks occupied by the GB users, multiplexing them non-orthogonally (ZHANG, C. et al., 2021; DING et al., 2019, 2021; ZHANG, N.; ZHU, X., 2022). Thus, SGF can be viewed as a compromise between GF and GB schemes, where the BS still controls the multiple access, but with lower signaling overhead compared to conventional GB. In (DING et al., 2019), two contention protocols were proposed to prevent the system performance degradation of GB users, named *distributed contention protocol* and *open-loop protocol*. In the first case the number of GF users to be granted access is pre-defined, while in the latter this value is random, so the protocol still suffers from user collisions as in pure GF, and therefore the GB performance is degraded. In (DING et al., 2021), the authors proposed an extension of the distributed contention protocol from (DING et al., 2019), where the GB-related received power at the BS is used to calculate and broadcast a threshold to the GF users to facilitate distributed contention. In addition, they adopt a hybrid SIC decoding order to assure that the GB user can experience the same performance as in OMA.

Moreover, RSMA has gained significant attention recently, since it enables the achievement of the entire capacity region with successive decoding (RIMOLDI; URBANKE, 1996; TSE; VISWANATH, 2005a). In this scope, (LIU, Yuanwen et al., 2022) showed that RSMA can outperform NOMA in the network slicing between eMBB and URLLC devices. Furthermore, authors from (LIU, H. et al., 2021) applied RSMA to the system model of (DING et al., 2021), where instead of multiplexing the GB and GF users by means of power-domain NOMA, the GF user performs rate-splitting, expanding the SIC orders possibilities and consequently improving the transmission reliability for the admitted GF user when compared to previous works.

In heterogeneous systems, different services need synergy on their operating modes to maximize the network performance. For instance, as the access of the GF users depends on the channel quality of the GB users in (LIU, H. et al., 2021), a method that can reduce the contention threshold in a way to facilitate achieving the required QoS of the GB users, while potentially resulting in higher rates and reliability, would be welcome. In this sense, a possible strategy would be to employ diversity techniques to improve the channel quality of GB users, while maintaining the QoS fixed. Therefore, if the QoS target is fixed but the channel quality of the GB user improves, then the contention threshold can be reduced, making room for better GF performance.

In this sense, authors from (TOMINAGA et al., 2021a) employ space diversity in a heterogeneous scenario with eMBB and mMTC users, achieving higher rates as the number of antennas increases, at the cost of more complex and expensive receivers. In OFDMA, the multiple sub-bands allocated to the users may also provide frequency diversity, as long as each sub-band experiences independent fading (CLERCKX; OESTGES, 2013). However, in this case, diversity comes at the cost of more spectrum per user. Message repetition or retransmission schemes, *i.e.*, time diversity, are also used to alleviate the fading effects, but may potentially lead to higher packet collision, specially in networks with a massive number of connected users (OZAKU et al., 2020). Differently from the above, while exploiting power-domain NOMA between eMBB and URLLC users, in (SANTOS et al., 2020) the authors applied the MMD allocation method (BAI et al., 2010) to achieve frequency diversity for each eMBB user that equals the number of independent channels available in the network, not requiring multiple antennas, retransmissions or additional bandwidth. In this technique CSI is required prior to transmission, but, as the admission process of GB users already needs this information, there is no additional cost. The diversity is achieved by properly allocating users to channels, with the drawback of expending computational resources at the BS and a few more bits of control data.

4.1.1 Novelty and Contribution

Due to its known advantages over power-domain NOMA, in this work we consider the RSMA-based multiplexing of GB and GF users in the same network resource, as in (LIU, H. et al., 2021). However, we allocate channels to GB users through the RVRHK algorithm from (BAI et al., 2010), which is part of the MMD method. Moreover, differently from (SANTOS et al., 2020), we consider that the users are multiplexed through RSMA and the GF users are allocated using a distributed protocol. The proposed scheme, which we refer to as RSMA-MMD-SGF, combines the benefits of the frequency diversity provided by MMD to GB users and the distributed contention protocol used to allocate GF users, showing their synergy and capability of simultaneously improving the performance of both services. The outage formulation of this proposed scheme is derived and then compared to the QoS-SGF (DING et al., 2021) and RSMA-SGF (LIU, H. et al.,

2021) strategies by means of numerical and simulation results, showing that the proposed method can reduce the outage probability while simultaneously increasing the achievable rate of both GB and GF users.

4.2 SYSTEM MODEL

We evaluate the uplink of a single-cell network whose bandwidth is divided in F independent subcarriers (or channels), such that $\{w_f\}_{f=1}^{F}$ represents the f-th subcarrier. There are F groups[1] of K GF users $\{U_{k,f}\}_{k=1}^{K}$, and each group is dynamically paired with a GB user $\{U_{B,m,f}\}_{m=1}^{M}$, where $M = F$ is the number of GB users, thus, the number of groups of GF users equals the number GB users. The idea behind dividing the GF users in groups is that, as there is one GB user occupying each channel $\{w_f\}_{f=1}^{F}$, we want to have one GF user allocated is these same resources to maximize the spectrum efficiency, thus GF users are split to compete in a specific channel. This could also be done without groups, however, this division would be more effective because each GF user only needs to calculate its achievable rate in a specific channel for a given threshold, thus waiting less time to be allocated. Fig. 16 illustrates this model, where we show only one GF group and one GB user for simplicity. The admitted GF and GB users share the same radio resource to communicate to a common single-antenna BS. The GB user is assumed to be allocated to that specific resource f through the RVRHK algorithm, as it is detailed in Section 1.3.2. Furthermore, we assume that the channel gains of the GF users at the channel f are ordered as

$$|h_{1,f}|^2 \leq |h_{2,f}|^2 \leq \cdots \leq |h_{K,f}|^2 \tag{38}$$

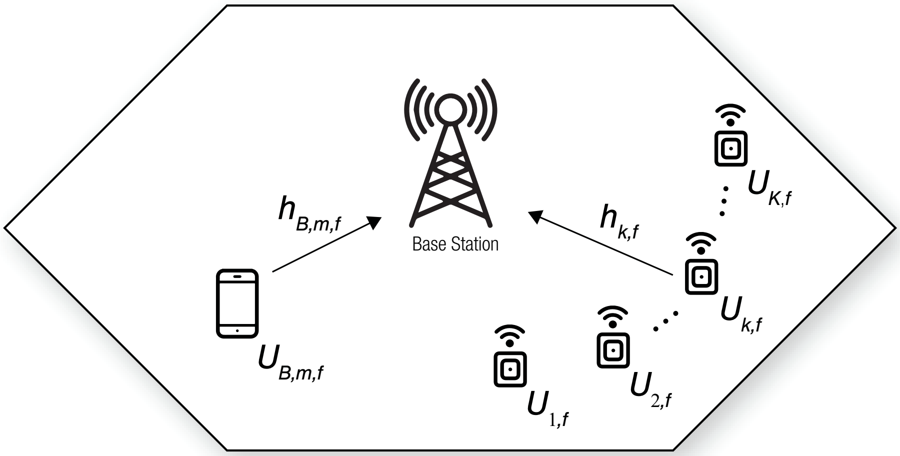
to facilitate the performance analysis. However, this information is not available to the BS and to the users. Here, $\{|h_{k,f}|\}_{k=1}^{K}$ represents the i.i.d. Rayleigh fading. In addition, we assume that prior to data transmission from the users, the BS broadcasts pilot signals, so that GB and GF users estimate their own channels, while the BS is informed of the GB user CSI, $h_{B,m,f}$, whose envelope is also modeled as i.i.d. Rayleigh fading. The BS is aware of the transmission power P_B of GB users. Due to latency and protocol complexity constraints, the same procedure is not carried out for GF users. Furthermore, the fading is considered constant during one TS, *i.e.*, a block fading model is assumed where TS is within the channel coherence time (TSE; VISWANATH, 2005b).

4.3 RSMA-MMD-SGF SCHEME

In this work, a GF user is allocated to the same radio resource as a GB user through the contention protocol proposed in (DING et al., 2019) and extended in (DING

[1] In our model, we consider that these groups are defined in a previous stage, while each group is assigned a certain resource f.

Figure 16 – System model with K GF users and one GB user.



Source: The Author.

et al., 2021). This protocol takes into account the QoS of the GB user to decide which out of the K GF users will be paired to perform the uplink simultaneously. Thus, the GB user experiences an OMA perception, without interference from another user. For this to be possible, the maximum interference power that the allocated GF user can impose to the GB user is defined by the following threshold (LIU, H. et al., 2021):

$$\hat{\tau}_f\left(|h_{B,m,f}|^2\right) = \frac{P_B|h_{B,m,f}|^2}{2^{\hat{R}_B}-1} - 1. \tag{39}$$

where $\hat{R}_B$ is the GB target rate.

As discussed in Chapter 1, applying MMD increases the frequency diversity gain experienced by the GB users, leading to a better equivalent channel when compared to the case without MMD. As a consequence, one can reduce the threshold $\hat{\tau}_f$ from (39), without compromising the QoS requirement. Then, it is expected that, by applying MMD before the GB user transmission, we would be able to increase the GB target rate without decreasing the GF performance and/or, alternatively, decrease P_B, reducing the power consumption of GB users.

During the channel estimation process, the BS broadcasts pilot signals to the GB and GF users. The GB users feedback their transmission power P_B and CSI $h_{B,m,f}$, which are inputs to the RVRHK algorithm run at the BS to allocate GB users to channels. After running the aforementioned algorithm, the BS informs the GB users about the allocation outcome, which requires 1 bit/channel/user of information. Then, as the maximum matching set $\mathcal{M}$ is determined at this point, the BS also calculates and broadcasts the interference threshold $\hat{\tau}_f$ to the groups of K GF users. Each user within a given group will compete to use subcarrier w_f assigned to the group. Based on their estimated channels and on $\hat{\tau}_f$, the GF users calculate their achievable rate $R_{F,k,f}$ and determine a backoff time inversely proportional to this value. The user with the lowest backoff time, *i.e.* the highest achievable rate, is the first to send an uplink identifying itself to the BS, being admitted to the same resource block as the GB user. The other GF users, by listening to the first GF user transmission, remain silent.

This simplified contention protocol is called SGF, being an intermediate solution between GB and GF that controls the number of users without adding too much overhead. After the GF user is admitted, it can share the radio resource with the GB user in different ways, such as NOMA (DING et al., 2021) and RSMA (LIU, H. et al., 2021) (in this case, the admitted GF user splits its transmission into two messages with power proportional to α and $1 - \alpha$, respectively, where $0 \leq \alpha \leq 1$ is the RSMA power splitting factor). Therefore, the SIC decoding modes are variable and determined by α, which can be set in two ways in our model depending on the relation of $P_F|h_{k,f}|^2$ with $\tau_f = \max\{0, \hat{\tau}_f\}$. Here, P_F is the GF user transmission power. This creates two possible user subgroups inside a group of K users, as described below:

1. Subgroup 1: $\alpha = 0$ when $P_F|h_{k,f}|^2 \leq \tau_f$. In this case, the BS decodes first the GB in the SIC procedure and the GF user does not perform the splitting (NOMA with order $U_{B,m,f} \longrightarrow U_{k,f}$). In this way, we have:

$$R_{B,m,f} = \log_2\left(1 + \frac{P_B|h_{B,m,f}|^2}{P_F|h_{k,f}|^2 + 1}\right) \tag{40}$$

   and

$$R_{\mathrm{I},F,k,f} = \log_2\left(1 + P_F|h_{k,f}|^2\right). \tag{41}$$

   The backoff time of each user in Subgroup I is set to be inversely proportional to $R_{\mathrm{I},F,k,f}$.

2. Subgroup II: $\alpha = 1 - \frac{\tau_f}{P_F|h_{k,f}|^2}$ when $P_F|h_{k,f}|^2 > \tau_f$. In this case RSMA is used and the decoding order will be $U^{'}_{k,f} \longrightarrow U_{B,m,f} \longrightarrow U^{''}_{k,f}$. The achievable rates for each user in this scenario can be expressed as:

$$R^{'}_{\mathrm{II},F,k,f} = \log_2\left(1 + \frac{P_F|h_{k,f}|^2 - \tau_f}{P_B|h_{B,m,f}|^2 + \tau_f + 1}\right), \tag{42}$$

$$R_{B,m,f} = \log_2\left(1 + \frac{P_B|h_{B,m,f}|^2}{\tau_f + 1}\right) \tag{43}$$

   and

$$R^{''}_{\mathrm{II},F,k,f} = \log_2\left(1 + \tau_f\right). \tag{44}$$

   The resulting GF user rate is $R_{\mathrm{II},F,k,f} = R^{'}_{\mathrm{II},F,k,f} + R^{''}_{\mathrm{II},F,k,f}$. Note that the backoff time of each user in Subgroup II is set to be inversely proportional to $R_{\mathrm{II},F,f,k}$.

Remark 1. *In Subgroup II, if the resulting* α *equals 1, NOMA is applied and the BS first decodes the GF user* ($U_F \longrightarrow U_B$)*. Therefore, the rates will be:*

$$R_{II,F,k,f} = \log_2\left(1 + \frac{P_F|h_{k,f}|^2}{P_B|h_{B,m,f}|^2 + 1}\right) \tag{45}$$

and

$$R_{B,m,f} = \log_2\left(1 + P_B|h_{B,m,f}|^2\right). \tag{46}$$

Remark 2. *In Subgroups I and II, the GF user only transmits if* $R_{F,k,f} \geq \hat{R}_{\mathrm{F}}$*, where* $\hat{R}_{\mathrm{F}}$ *refers to the target transmission rate of all the GF users. Moreover, although the contention protocol is based on the achievable rate, the user transmits using the target rate.*

Remark 3. *Only two among the* K *GF users have the chance of being granted access, since the GF users' channel gains are ordered as in* (38)*, which entails that* R_F *is monotonically increasing with respect to* $|h_{k,f}|^2$*. The GF user having the largest* $|h_{k,f}|^2$ *in Subgroup I will be granted access if Subgroup II is empty. Furthermore, the GF user having the largest* $|h_{k,f}|^2$ *in Subgroup II will be granted access if Subgroup I is empty. When both subgroups are not empty, Subgroup II has the priority for admission since* $R_{\mathrm{II},F,k,f} > R_{\mathrm{I},F,k,f}$*, which always holds based on the fact* $R_{\mathrm{I},F,k,f} \leq \log_2(1 + \tau_f)$*.*

Algorithm 1 presents in detail the admission procedure, which is also summarized below:

- BS broadcasts pilot signals to assist GB and GF users to estimate their channels.
- The GB user feeds back its channel gain $h_{B,m,f}$ and P_B to the BS.
- The BS is then responsible for:
  - Executing the RVRHK algorithm to allocated GB users to channels.
  - Sending this information for each GB user (1 bit/channel/user).
  - Calculating the interference threshold τ and broadcasting it to all the K GF users in each group.
- Each GF user calculates its achievable rate and determines the associated backoff time.
- The user with the lowest backoff time, *i.e.* the highest achievable rate, is the first to send an uplink identifying itself to the BS, being admitted to the same resource block as the GB user.

Algorithm 1 Complete Admission Procedure

```
$\mathcal{M} \longleftarrow \emptyset$.
Acquire $h_{B,m,f}$ and $P_B$ of each GB user $\{U_{B,m,f}\}_{m=1}^{M}$ at each subcarrier $\{w_f\}_{f=1}^{F}$.
for $m \in \{1, 2, \ldots, M\}$ do
    Randomly choose a user $m$ among $M$ GB users.
    Allocate user $U_{B,m,f}$ according to its non-outage subcarriers $\mathcal{N}(U_{B,m,f})$.
    Inform user $m$ the channel it should transmit.
    Calculate the threshold $\hat{\tau}_f\left(|h_{B,m,f}|^2\right)$ of user $m$.
    Remove user $m$ and the allocated channel from the allocation queue.
end for
Broadcast $\hat{\tau}_f$ to the groups of $K$ GF users.
for each group of $K$ GF users do
    for $k = 1, 2, \ldots, K$ do
        if $P_F|h_{k,f}|^2 \leq \tau_f$ then
            User follows Subgroup I specification, sending a response to the BS after a backoff time $\propto 1/R_{\text{I},F,k,f}$.
        else
            User follows Subgroup II specification, sending a response to the BS after a backoff time $\propto 1/R_{\text{II},F,k,f}$.
        end if
    end for
end for
The BS admits the GF users that respond first.
```

4.3.1 Outage Formulation

In the following, we focus our analysis to one group of K GF users competing to be scheduled with one previously allocated GB user at a specific channel (as the example of Fig. 16). Thus, we omit the indexes m and f for convenience. Besides, as the outage formulation of MMD-aided GB users are the same as in OMA, which were presented in Subsection 1.3.2, we focus here on the outage formulation of the admitted GF user under RSMA-MMD-SGF, as presented in Theorem 1.

Theorem 1. *For $K \geq 2$, the outage probability of the GF user operating under the RSMA-MMD-SGF scheme is presented in* (47)*.*

$$
\begin{aligned}
P_{\text{out}} = & \frac{\varphi_0}{K(K-1)} \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} \mu_1 \nu(0, \mu_2) \\
& + \sum_{k=1}^{K-2} \varphi_k \sum_{n=0}^{K-k} \binom{K-k}{n} (-1)^{n} \sum_{\ell=0}^{k} \binom{k}{\ell} (-1)^{\ell} e^{\frac{\ell}{P_F}} \mu_3 \nu(\ell, \mu_4) \\
& + \frac{\varphi_0}{K-1} \sum_{\ell=0}^{K-1} \binom{K-1}{\ell} (-1)^{\ell} e^{\frac{\ell}{P_F}} \left(e^{\frac{1}{P_F}} \nu(\ell, \mu_5) - e^{-\frac{\epsilon_B+\epsilon_F+\epsilon_B\epsilon_F}{P_F}} \nu(\ell, \mu_6) \right) \\
& + \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} e^{\frac{\ell}{P_F}} \nu(\ell, 0) \\
& + \sum_{m=1}^{F} \binom{F}{f} (-1)^{f-1} \left(\left(1 - e^{-\eta_F}\right)^{K} e^{-\frac{f}{P_B}(1+\epsilon_F)\eta_B} \right. \\
& \left. + \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} e^{-\ell\eta_F} \frac{f}{P_B} \frac{1 - e^{-\left(\frac{f}{P_B}+\ell\eta_F P_B\right)\eta_B}}{\frac{f}{P_B} + \ell\eta_F P_B} \right),
\end{aligned}
\tag{47}
$$

where $\mu_1 = e^{\frac{K-\ell(1+\epsilon_B)(1+\epsilon_F)}{P_F}}$, $\mu_2 = \frac{K-\ell}{P_F\eta_B} - \frac{P_B\ell}{P_F}$, $\mu_3 = e^{\frac{K-k-n(1+\epsilon_B)(1+\epsilon_F)}{P_F}}$, $\mu_4 = \frac{K-k-n}{P_F\eta_B} - \frac{nP_B}{P_F}$, $\mu_5 = \frac{1}{P_F\eta_B}$, $\mu_6 = -\frac{P_B}{P_F}$, $\varphi_0 = \frac{K!}{(K-2)!}$, $\varphi_k = \frac{K!}{k!(K-k)!}$ for $1 \leq k \leq K-2$, $\epsilon_B = 2^{\hat{R}_B} - 1$, $\epsilon_F = 2^{\hat{R}_F} - 1$, $\eta_B = \frac{\epsilon_B}{P_B}$, $\eta_F = \frac{\epsilon_F}{P_F}$, $\nu(\ell, \mu) = \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} \frac{f}{P_B} \Theta(\ell, \mu, f)$, and

$$
\Theta(\ell, \mu, f) = \begin{cases} \epsilon_F \eta_B, & \text{if } \mu = -\frac{f}{P_B} - \frac{\ell}{P_F\eta_B}, \\ \dfrac{e^{-\left(\frac{\ell}{P_F\eta_B}+\mu+\frac{f}{P_B}\right)\eta_B} - e^{-\left(\frac{\ell}{P_F\eta_B}+\mu+\frac{f}{P_B}\right)\eta_B(1+\epsilon_F)}}{\frac{\ell}{P_F\eta_B} + \mu + \frac{f}{P_B}}, & \text{otherwise.} \end{cases}
$$

Proof. Following (DING et al., 2021; LIU, H. et al., 2021), the outage probability experienced by the admitted GF user can be expressed as

$$
\begin{aligned}
P_{\text{out}} = & \Pr\left(E_0, R_{\text{II},F} < \hat{R}_F\right) + \sum_{k=1}^{K-1} \Pr\left(E_k, R_{\text{II},F} < \hat{R}_F\right) \\
& + \Pr\left(E_K, R_{\text{I},F} < \hat{R}_F\right),
\end{aligned}
\tag{48}
$$

where $E_k = \left\{|h_k|^2 \leq \tau_f/P_F, |h_{k+1}|^2 > \tau_f/P_F\right\}$ is the event that there are k GF users in Subgroup I, $E_0 = \left\{|h_1|^2 > \tau_f/P_F\right\}$ is the event with no user in Subgroup I and $E_K = \left\{|h_K|^2 < \tau_f/P_F\right\}$ is the event with no user in Subgroup II. Knowing that $\tau_f = \max\left\{0, |h_B|^2/\eta_B - 1\right\} = 0$ when $|h_B|^2 < \eta_B$, the outage probability can be rewritten as

$$\begin{aligned} P_{\text{out}} = & \underbrace{\Pr\left(E_0, |h_B|^2 > \eta_B, R_{\text{II},F} < \hat{R}_F\right)}_{Q_0} \\ & \underbrace{+ \sum_{k=1}^{K-1} \Pr\left(E_k, |h_B|^2 > \eta_B, R_{\text{II},F} < \hat{R}_F\right)}_{Q_k} \\ & \underbrace{+ \Pr\left(E_K, |h_B|^2 > \eta_B, R_{\text{I},F} < \hat{R}_F\right)}_{Q_K} \\ & \underbrace{+ \Pr\left(|h_B|^2 < \eta_B, R_{\text{II},F} < \hat{R}_F\right)}_{Q_{K+1}}. \end{aligned} \tag{49}$$

The terms Q_0 and Q_k, which correspond to the first three terms in (47), are similar to the ones presented in (LIU, H. et al., 2021), except for the general expectation term $\nu(i, \mu)$, since the PDF of GB users is defined by (8) in our model.

4.3.1.1 Evaluation of $\nu(i, \mu)$

In (LIU, H. et al., 2021), a general expectation term is introduced to facilitate the calculation of the terms that compose the outage probability. Such a term is defined as

$$\nu(i, \mu) \triangleq \underset{\eta_B < |h_B|^2 < \eta_B(1+\epsilon_F)}{\mathcal{E}} \left\{ e^{-\left(\frac{i}{P_F \eta_B} + \mu\right)|h_B|^2} \right\}, \tag{50}$$

where $\mathcal{E}\{\cdot\}$ is the expectation operation. Let us define the variable $S_B = \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} \frac{f}{P_B}$. Since $|h_B|^2$ has the PDF defined in (8), in our case $\nu(i, \mu)$ can be calculated as

$$\nu(i, \mu) = S_B \underbrace{\int_{\eta_B}^{\eta_B(1+\epsilon_F)} e^{-\left(\frac{i}{P_F \eta_B} + \mu + \frac{f}{P_B}\right)x} dx}_{\Theta(i,\mu,f)}. \tag{51}$$

Solving $\Theta(i, \mu, f)$, we have

$$\Theta(i, \mu, f) = \begin{cases} \epsilon_F \eta_B, & \text{if } \mu = -\frac{f}{P_B} - \frac{i}{P_F \eta_B}, \\ \frac{1}{\sigma}\left[e^{-\sigma\eta_B} - e^{-\sigma\eta_B(1+\epsilon_F)}\right], & \text{otherwise,} \end{cases}$$

where $\sigma = \frac{i}{P_F \eta_B} + \mu + \frac{f}{P_B}$.

4.3.1.2 Evaluation of Q_K

The term Q_K in (49) is composed of two other terms, as presented in (LIU, H. et al., 2021, Eq. (A.25)). In our case, the first term $Q_K^{'} = \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} e^{\frac{\ell}{P_F}} \nu(\ell, 0)$

remains the same, but using $\nu(i,\mu)$ defined in (51). The second term can be calculated as follows:

$$\begin{aligned} Q_K'' &= \underset{|h_B|^2>\eta_B(1+\epsilon_F)}{\mathcal{E}} \left\{ \Pr\left(|h_K|^2<\eta_F\right)\right\} \\ &= \underset{|h_B|^2>\eta_B(1+\epsilon_F)}{\mathcal{E}} \left\{ (1-e^{-\eta_F})^K \right\} \\ &= \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} \frac{f}{P_B} \int_{\eta_B(1+\epsilon_F)}^{\infty} (1-e^{-\eta_F})^K e^{-\frac{f}{P_B}x} dx \\ &= (1-e^{-\eta_F})^K \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} e^{-\frac{f}{P_B}(1+\epsilon_F)\eta_B}. \end{aligned} \tag{52}$$

Thus, the term Q_K is the sum of Q_K' and Q_K''.

4.3.1.3 Evaluation of Q_{K+1}

This term can be calculated as

$$\begin{aligned} Q_{K+1} &= \Pr\left(|h_B|^2<\eta_B, \log_2\left(1+\frac{P_F|h_K|^2}{P_B|h_B|^2+1}\right)<\hat{R}_F\right) \\ &= \underset{|h_B|^2<\eta_B}{\mathcal{E}} \left\{ \Pr\left(|h_K|^2<\eta_F(1+P_B|h_B|^2)\right)\right\} \\ &= S_B \int_0^{\eta_B} \left(1-e^{-\eta_F(1+P_Bx)}\right)^K e^{-\frac{f}{P_B}x} dx \\ &= S_B \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} e^{-\ell\eta_F} \int_0^{\eta_B} e^{-\left(\frac{f}{P_B}+\ell\eta_F P_B\right)x} dx \\ &= S_B \sum_{\ell=0}^{K} \binom{K}{\ell} (-1)^{\ell} e^{-\ell\eta_F} \frac{1-e^{-\left(\frac{f}{P_B}+\ell\eta_F P_B\right)\eta_B}}{\frac{f}{P_B}+\ell\eta_F P_B}. \end{aligned} \tag{53}$$

Therefore, by combining Q_0, Q_k and Q_K' (considering $\nu(i,\mu)$ defined in (51)) with (52) and (53), the overall outage probability of the admitted GF user is obtained as shown in Theorem 1 and the proof is complete.

□

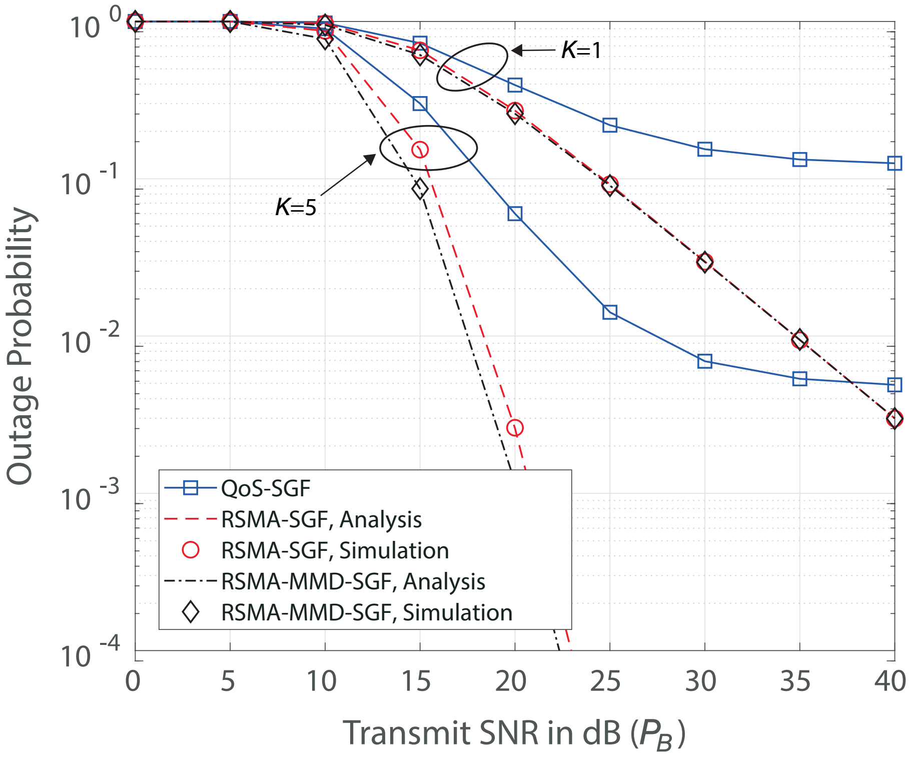
Finally, the outage probability experienced by an individual GF user, i.e. when $K=1$, can be easily obtained by following the same procedure as demonstrated in the proof of Theorem 1, resulting in

$$\begin{aligned} P_{\text{out}} = \sum_{f=1}^{F} \binom{F}{f} (-1)^{f-1} \times \Bigg(1 - e^{-\eta_F-\frac{f}{P_B}(1+\epsilon_F)\eta_B} \\ -\frac{f}{P_B} \frac{e^{-\eta_F}(1-e^{-(\frac{f}{P_B}+P_B\eta_F)\eta_B})}{\frac{f}{P_B}+P_B\eta_F}\Bigg) - e^{-\frac{\epsilon_B+\epsilon_F+\epsilon_B\epsilon_F}{P_F}} \nu(0,\mu_6). \end{aligned} \tag{54}$$

4.4 NUMERICAL RESULTS

We present numerical results comparing the GF user outage and data rate performance considering the proposed RSMA-MMD-SGF scheme with that achieved by the QoS-SGF (DING et al., 2021) and RSMA-SGF (LIU, H. et al., 2021) methods, by focusing on one group of K GF users competing to be scheduled with one previously allocated GB user at a specific channel. In general, the results presented in the following Fig. 17-Fig. 22 are analytically obtained from (47) and (54), which depend on the set of parameters $(\hat{R}_\mathrm{B}, \hat{R}_\mathrm{F}, P_\mathrm{B}, P_\mathrm{F}, K, F)$. Since each figure is generated for a particular setup, we present the values of the aforementioned parameters in the paragraph corresponding to each figure.Unless stated otherwise, we consider $F = 5$ independent channels. Figs. 17 and 18 present the outage probability versus P_B for different values of $\hat{R}_B$. The target rate pair is configured to $(\hat{R}_B, \hat{R}_F) = (1.5, 2)$ Bits per Channel Use (BPCU) in Fig. 17 and to $(\hat{R}_B, \hat{R}_F) = (4, 2)$ BPCU in Fig. 18. Besides, $P_F = P_B/10$ to reflect the scenarios in which the average channel conditions of the GF users are weaker than that of the GB user.

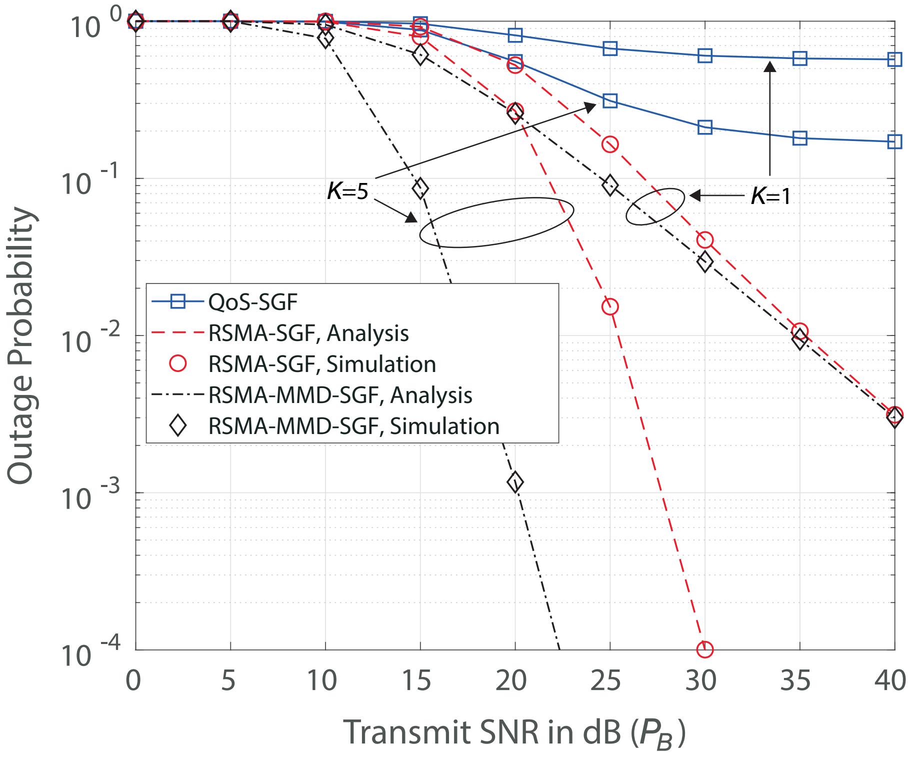
Figure 17 – SGF outage probability for $P_F = \frac{P_B}{10}$, $\hat{R}_B = 1.5$ BPCU and $\hat{R}_F = 2$ BPCU.



Source: The Author.

From Figs. 17 and 18 it is possible to observe the superior performance of the RSMA-MMD-SGF method for both $K = 1$ and $K = 5$, specially in Fig. 18 where we increase the GB target rate. These results lead us to the conclusion that it is possible to increase the performance of GB users and maintain the GF outage probability

Figure 18 – SGF outage probability for $P_F = \frac{P_B}{10}$, $\hat{R}_B = 4$ BPCU and $\hat{R}_F = 2$ BPCU.



Source: The Author.

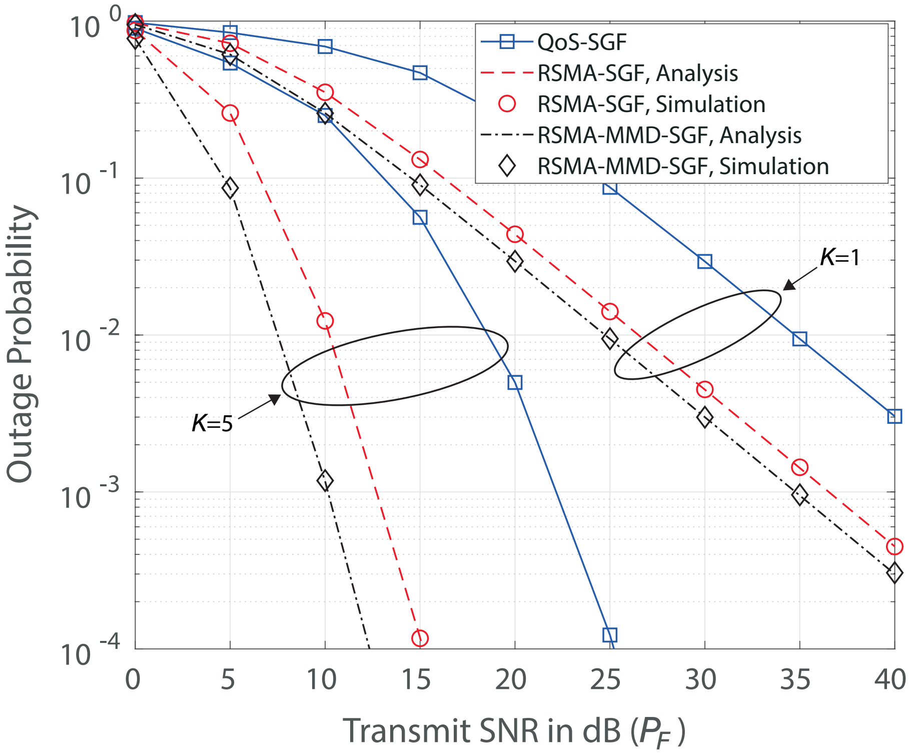
simultaneously, which does not occur with the QoS-SGF and RSMA-SGF methods. Interestingly, the RSMA-MMD-SGF method achieves better results for $K = 1$ when compared to QoS-SGF with $K = 5$ for all SNR values.

In Fig. 19, we investigate the GF user outage performance achieved by the considered SGF schemes for different values of P_F while P_B is fixed in 10 dB. The target rate pair is configured to $(\hat{R}_B, \hat{R}_F) = (1.5, 2)$ BPCU. We can see that, in this configuration, the proposed method reduces the GF outage probability in the whole SNR range. Note that the simulation results for RSMA-MMD-SGF scheme presented in Figs. 17, 18 and 19 match the analytical results perfectly, verifying the accuracy of the expressions in Theorem 1 and eq. (54).

Simulations on the impact of the target rate on the outage probability are discussed next. In Fig. 20, we set the transmission powers P_B and P_F to 15 dB and the GB target rate to 2 BPCU, while we plot the outage of the GF user for different values of $\hat{R}_F$. It is possible to note that the RSMA-MMD-SGF approach is capable of achieving higher target rates for GF users while keeping their outage controlled. For an outage of 10^{-2}, for instance, we can reach a value of $\hat{R}_F \approx 4$ BPCU, while for the RSMA-SGF scheme $\hat{R}_F \approx 3$ BPCU.

In Fig. 21 we keep the target rate of both services the same ($\hat{R}_B = \hat{R}_F$), while $P_B = P_F = 20$ dB. Here the increase in the target rate is even more relevant. For instance, for an outage of 10^{-2}, $\hat{R}_F$ and $\hat{R}_B$ go from ≈ 3.2 BPCU to ≈ 5.7 BPCU.

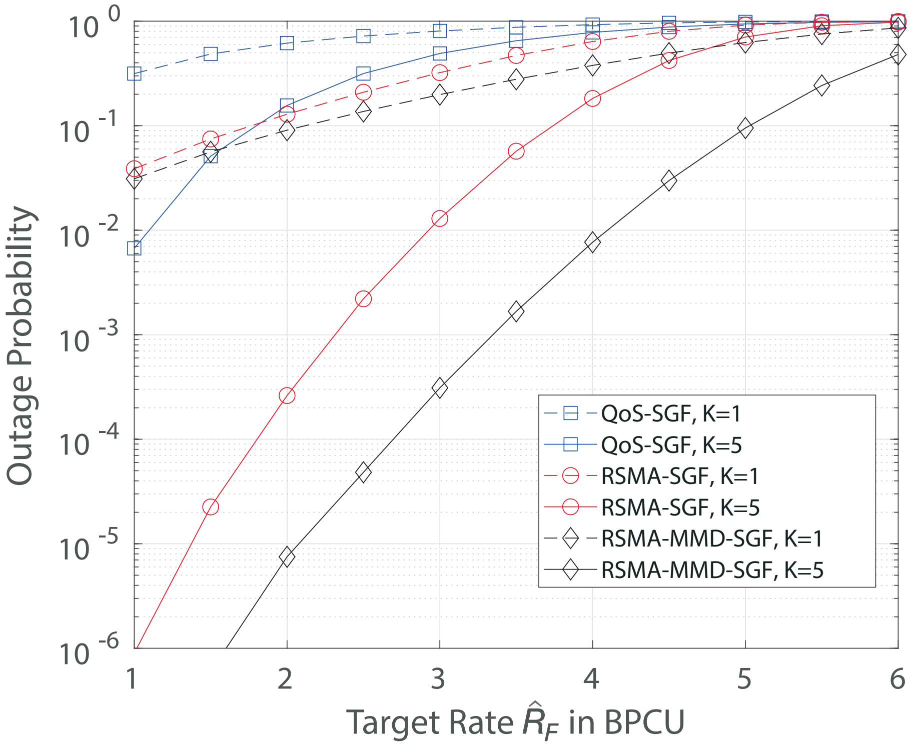
Figure 19 – SGF outage probability for fixed P_B = 10 dB, $\hat{R}_B$ = 1.5 BPCU and $\hat{R}_F$ = 2 BPCU.



Source: The Author.

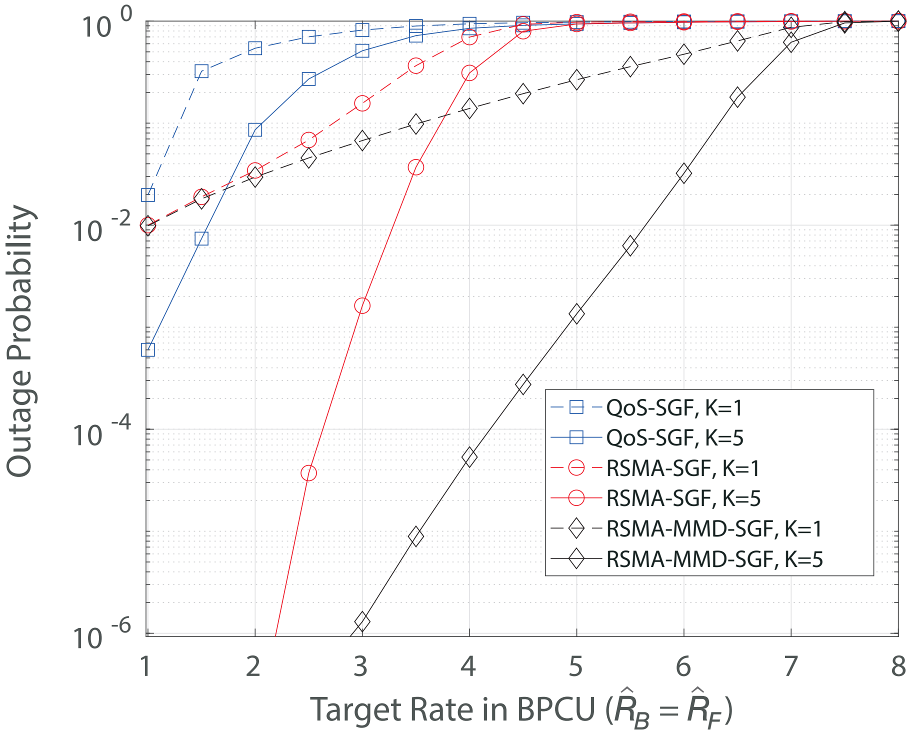
Surprisingly, for values of target rate $\geq$ 3.8 BPCU, the RSMA-MMD-SGF method with K = 1 is better than the other methods for both K = 1 or K = 5. As discussed in Subsection 1.3.2, it is clear that the MMD method increases the GB performance. However, from the previous results, we can affirm that if we fix $\hat{R}_B$, with the proposed approach it is possible to increase the GF performance as well, taking advantage of the capacity region points reachable by RSMA and the better channel condition experienced by the MMD-aided GB users, which allows more interference in the simultaneous uplink while keeping the QoS as in OMA.

Figure 20 – Impact of the target rate on the GF user outage probability. $\hat{R}_B = 2$ BPCU and $P_B = P_F = 15$ dB.



Source: The Author.

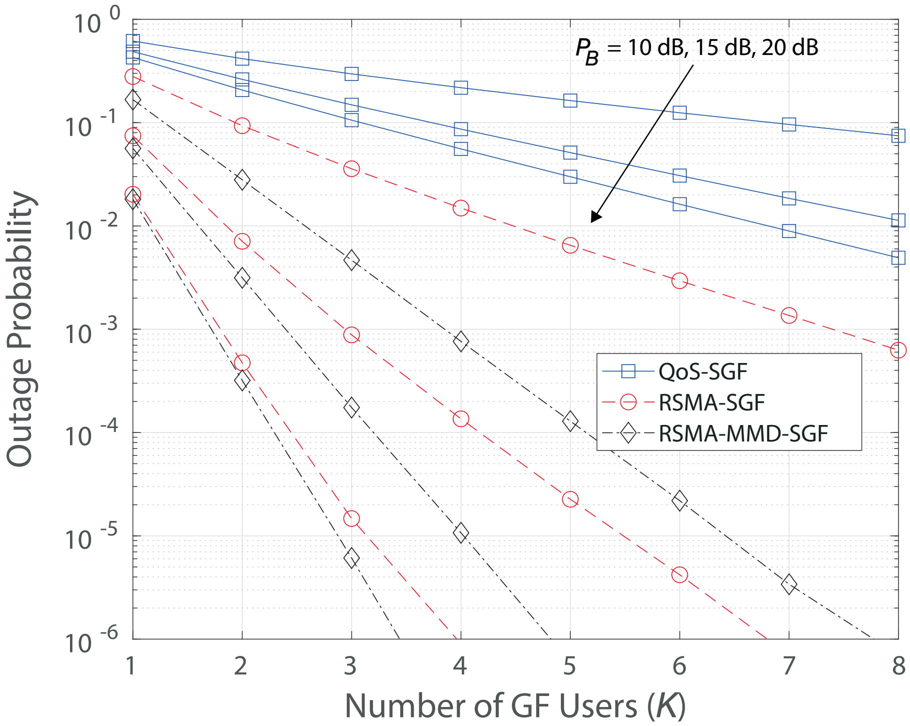
Figure 21 – Impact of the target rate on the GF user outage probability. $P_B = P_F = 20$ dB.



Source: The Author.

The impact of the number of the GF users on the outage probability is investigated in Fig. 22 using simulation results, for which we set $(\hat{R}_B, \hat{R}_F) = (2, 1.5)$ BPCU and $P_B = P_F$. As expected, as K increases the outage probability reduces thanks to the multi-user diversity of the distributed contention protocol applied to the GF users, which is achievable irrespective of the target rate value (LIU, H. et al., 2021). Comparing the three schemes, RSMA-MMD-SGF presents superior performance, which increases with P_B and P_F. For instance, for $K = 4$ and $P_B = P_F = 15$ dB, the GF outage is 10^{-5} for RSMA-MMD-SGF, while for RSMA-SGF this value is $\approx 10^{-4}$.

Figure 22 – Impact of the number of the GF users on the outage probability ($\hat{R}_B = 2$ BPCU, $\hat{R}_F = 1.5$ BPCU and $P_B = P_F$).



Source: The Author.

4.5 FINAL COMMENTS

We considered the sharing of physical network resources between GB and GF users. The proposed method resorts to the MMD approach to increase the frequency diversity of GB users and to a distributed contention protocol to pair one GF with a GB user. Moreover, the multiplexed users can be superimposed with NOMA or RSMA. We presented exact expressions for the proposed SGF system and showed that it is able to increase the target rate and reduce the outage probability of both services.

5 CONCLUSIONS

This work started by considering the RAN slicing of a B5G network composed by eMBB and URLLC users. Then, the HMA MMD-aided approach was proposed, where hybrid resource allocation is performed, focusing on extracting the pros of OMA MMD-aided and NOMA MMD-aided methods. With this technique, it was possible to obtain larger rates for eMBB when compared to OMA and NOMA slicing strategies. When prioritizing the eMBB rate, the network does not need to allocate specific channels to URLLC, then, the hybrid slicing is converted to non-orthogonal slicing. Furthermore, HMA MMD-aided can improve the eMBB rate under even more strict URLLC reliability, such as $\epsilon_U = 10^{-7}$, when compared to (SANTOS et al., 2020).

Then, we evaluated the sum-rate performance of three multiple access methods for URLLC, namely U-OMA, U-NOMA, and U-RSMA, when operating under both OMA and NOMA network slicing strategies. Our results show that U-RSMA is capable of achieving higher rates when the power splitting factor is properly configured, even with strict reliability requirements. Moreover, we show that NOMA network slicing is capable of reaching the highest pair of rates for URLLC and eMBB simultaneously. This leads us to show another interesting scenario in which combining U-RSMA and NOMA is a powerful tool for attending B5G demands. The practical implementation of RSMA is still evolving, however, for our scenario, limiting the number of users in each mini-slot is a good strategy to reduce the decoding complexity and delay. Also, it is necessary to add a few bits of information in the message intended to the user that splits its data, sent in the synchronization slot to set the power allocation factor based on the average SNR.

To finish the thesis, we considered the sharing of physical network resources between GB and GF users. The proposed method resorts to the MMD approach to increase the frequency diversity of GB users and to a distributed contention protocol to pair one GF with a GB user. Moreover, the multiplexed users can be superimposed with NOMA or RSMA. We presented exact expressions for the proposed SGF system and showed that it is able to increase the target rate and reduce the outage probability of both services. As a drawback, it takes longer to allocate GF users, since the RVRHK algorithm should be executed before defining the threshold τ.

5.1 FUTURE WORKS

For future works, regarding Chapter 2, it would be beneficial to investigate the network slicing between the eMBB and mMTC services. The study should entail the modeling of mMTC outage probability, evaluation of detection priority order using SIC at the receiver for non-orthogonal and hybrid scenarios, and assessing the interference between these services. Another potential alternative would be to improve the URLLC model by incorporating the specific duration of each mini-slot that a device utilizes.

This parameter could be adjusted to analyze the effect of URLLC transmission latency on eMBB users, or increased until an acceptable latency threshold is reached, while monitoring its impact on data transmission rate.

Regarding the RSMA method, a natural extension to the models presented in Chapters 3 and 4 is to allow more URLLC/GF users to share resources with eMBB/GB users in the same resource block. However, this entails in extra decoding complexity, as we have an increased number of decoding orders, and it is necessary to determine which users would split the messages.

Another possible extension of the thesis could involve proposing a new contention protocol to replace the one used in Chapter 4 to optimize, for example, the allocation delay caused by the exchange of grant messages. This could be performed developing more efficient algorithms or using machine learning techniques. Authors from (SHARMA; WANG, X., 2019) and (MEYER; TURAU, 2021) proposed a contention protocol based on reinforcement learning, more specifically, Q-learning algorithm, showing that it is possible to lower the network congestion maintaining the throughput and reliability in acceptable levels.

Other important topics for future research can be summarized as follows:

- Study of the impact of imperfect CSI on the design of SGF schemes;
- Consider a stochastic geometry model, to take into consideration the nodes position during the allocation procedure of GF users (ZHANG, C. et al., 2022);
- Investigate the performance of the proposed method in a simulation environment with different network configurations, such as user density, geographical distribution and varying traffic loads, to evaluate the scalability and efficiency of the system;
- Explore other multiple access techniques in addition to those used during the thesis, such as Sparse Code Multiple Access (SCMA), in the task of multiplexing heterogeneous users;
- Analyze the impact of user mobility on the network and how it affects the performance of different multiple access techniques (DIZDAR et al., 2021b);
- Investigation of techniques to reduce the complexity of decoding and delay in U-RSMA systems, such as the use of mini-slots with a limited number of users and the addition of information in the synchronization message to adjust the power allocation factor;
- Investigate the application of the proposed hybrid resource allocation approach in other network configurations, with different combinations of services and transmission rate requirements. This would allow for a more comprehensive performance

analysis of the approach and a comparison with other resource multiplexing techniques;

- Explore further the performance of the U-RSMA and U-NOMA methods, especially regarding their scalability to larger networks. Furthermore, it is possible to investigate how these methods can be combined with other resource allocation techniques, such as the proposed hybrid approach and MMD algorithm.

BIBLIOGRAPHY

3GPP. **5G, NR, physical layer procedures for data**. Available from: https://www.etsi.org/deliver/etsi%5C_ts/138200%5C_138299/138214/15.02.00%5C_60/ts%5C_138214v150200p.pdf. Visited on: 30 Apr. 2023.

ABBASI, Omid; YANIKOMEROGLU, Halim. Rate-Splitting and NOMA-Enabled Uplink User Cooperation. In: 2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). [S.l.: s.n.], 2021. P. 1–6. DOI: 10.1109/WCNCW49093.2021.9419994.

AL-ABBASI, Ziad Qais; SO, Daniel K. C. Resource Allocation in Non-Orthogonal and Hybrid Multiple Access System With Proportional Rate Constraint. **IEEE Transactions on Wireless Communications**, v. 16, n. 10, p. 6309–6320, 2017. DOI: 10.1109/TWC.2017.2721936.

ABREU, R.; JACOBSEN, T.; BERARDINELLI, G.; PEDERSEN, K.; MAHMOOD, N. H.; KOVACS, I. Z.; MOGENSEN, P. On the Multiplexing of Broadband Traffic and Grant-Free Ultra-Reliable Communication in Uplink. In: IEEE Vehicular Technology Conference (VTC). [S.l.: s.n.], 2019. P. 1–6. DOI: 10.1109/VTCSpring.2019.8746589.

ALSENWI, Madyan; TRAN, Nguyen H.; BENNIS, Mehdi; KUMAR BAIRAGI, Anupam; HONG, Choong Seon. eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach. **IEEE Communications Letters**, v. 23, n. 4, p. 740–743, 2019. DOI: 10.1109/LCOMM.2019.2900044.

AMERICAS, 5G. **5G and LTE Deployments**. Available from: https://www.5gamericas.org/resources/deployments/. Visited on: 18 Apr. 2023.

ANAND, Arjun; DE VECIANA, Gustavo; SHAKKOTTAI, Sanjay. Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks. In: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications. [S.l.: s.n.], 2018. P. 1970–1978. DOI: 10.1109/INFOCOM.2018.8486430.

ANWAR, Asim; SEET, Boon-Chong; LI, Xuejun. Hybrid multiple access for 5G downlink systems: Throughput and outage analysis. **AEU - International Journal of Electronics and Communications**, v. 117, p. 153100, 2020. ISSN 1434-8411. DOI: https://doi.org/10.1016/j.aeue.2020.153100. Available from: http://www.sciencedirect.com/science/article/pii/S1434841119316036.

AZAM, Irfan; SHAHAB, Muhammad Basit; SHIN, Soo Young. User Pairing and Power Allocation for Capacity Maximization in Uplink NOMA. In: 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). [S.l.: s.n.], 2019. P. 690–694. DOI: 10.1109/TSP.2019.8768824.

AZARI, Amin; POPOVSKI, Petar; MIAO, Guowang; STEFANOVIC, Cedomir. Grant-Free Radio Access for Short-Packet Communications over 5G Networks. In: GLOBECOM 2017 - 2017 IEEE Global Communications Conference. [S.l.: s.n.], 2017. P. 1–7. DOI: `10.1109/GLOCOM.2017.8255054`.

BAI, B.; CHEN, W.; CAO, Z.; LETAIEF, K. B. Max-matching diversity in OFDMA systems. v. 58, n. 4, p. 1161–1171, Apr. 2010. ISSN 0090-6778. DOI: `10.1109/TCOMM.2010.04.080478`.

BAYESTEH, Alireza; YI, Eric; NIKOPOUR, Hosein; BALIGH, Hadi. Blind detection of SCMA for uplink grant-free multiple-access. In: 2014 11th International Symposium on Wireless Communications Systems (ISWCS). [S.l.: s.n.], 2014. P. 853–857. DOI: `10.1109/ISWCS.2014.6933472`.

BJÖRNSON, E.; ELDAR, Y. C.; LARSSON, E. G.; LOZANO, A.; POOR, H. V. 25 Years of Signal Processing Advances for Multiantenna Communications. **IEEE Signal Processing Magazine**, 2023. DOI: `https://doi.org/10.48550/arXiv.2304.02677`.

BJÖRNSON, Emil; HOYDIS, Jakob; SANGUINETTI, Luca. Massive MIMO Networks: Spectral, Energy and Hardware Efficiency. In: [s.l.]: Foundations and Trends in Signal Processing, 2019.

BOCKELMANN, C.; PRATAS, N.; NIKOPOUR, H.; AU, K.; SVENSSON, T.; STEFANOVIC, C.; POPOVSKI, P.; DEKORSY, A. Massive machine-type communications in 5g: physical and MAC-layer solutions. **IEEE Communications Magazine**, v. 54, n. 9, p. 59–65, Sept. 2016. DOI: `10.1109/MCOM.2016.7565189`.

CAIRE, G.; TARICCO, G.; BIGLIERI, E. Optimum power control over fading channels. **IEEE Transactions on Information Theory**, v. 45, n. 5, p. 1468–1489, 1999. DOI: `10.1109/18.771147`.

CAO, Jian; YEH, Edmund M. Asymptotically Optimal Multiple-Access Communication Via Distributed Rate Splitting. **IEEE Transactions on Information Theory**, v. 53, n. 1, p. 304–319, 2007. DOI: `10.1109/TIT.2006.887497`.

CHEN, He; ABBAS, Rana; CHENG, Peng; SHIRVANIMOGHADDAM, Mahyar; HARDJAWANA, Wibowo; BAO, Wei; LI, Yonghui; VUCETIC, Branka. Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches. **IEEE Communications Magazine**, v. 56, n. 12, p. 119–125, 2018. DOI: `10.1109/MCOM.2018.1701178`.

CHEN, W.; LIN, X.; LEE, J.; TOSKALA, A.; SUN, S.; CHIASSERINI, C. F.; LIU, L. 5G-Advanced Towards 6G: Past, Present, and Future. **IEEE Journal on Selected Areas in Communications**, 2023. DOI: `https://doi.org/10.48550/arXiv.2303.07456`.

CHEN, Xiaohang; BENJEBBOUR, Anass; LI, Anxin; HARADA, Atsushi. Multi-User Proportional Fair Scheduling for Uplink Non-Orthogonal Multiple Access (NOMA). In: 2014 IEEE 79th Vehicular Technology Conference (VTC Spring). [S.l.: s.n.], 2014. P. 1–5. DOI: 10.1109/VTCSpring.2014.7022998.

CHENG-XIANG, W. et al. **On the Road to 6G: Visions, Requirements, Key Technologies and Testbeds**. [S.l.: s.n.], Feb. 2023.

CLERCKX, Bruno; JOUDEH, Hamdi; HAO, Chenxi; DAI, Mingbo; RASSOULI, Borzoo. Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution. **IEEE Communications Magazine**, v. 54, n. 5, p. 98–105, 2016. DOI: 10.1109/MCOM.2016.7470942.

CLERCKX, Bruno; MAO, Yijie; JORSWIECK, Eduard A.; YUAN, Jinhong; LOVE, David J.; ERKIP, Elza; NIYATO, Dusit. A Primer on Rate-Splitting Multiple Access: Tutorial, Myths, and Frequently Asked Questions. **IEEE Journal on Selected Areas in Communications**, p. 1–1, 2023. DOI: 10.1109/JSAC.2023.3242718.

CLERCKX, Bruno; MAO, Yijie; SCHOBER, Robert; POOR, H. Vincent. Rate-Splitting Unifying SDMA, OMA, NOMA, and Multicasting in MISO Broadcast Channel: A Simple Two-User Rate Analysis. **IEEE Wireless Communications Letters**, v. 9, n. 3, p. 349–353, 2020. DOI: 10.1109/LWC.2019.2954518.

CLERCKX, Bruno; OESTGES, Claude. Chapter 11 - Space-Time Coding for Frequency Selective Channels. In: CLERCKX, Bruno; OESTGES, Claude (Eds.). **Mimo Wireless Networks (Second Edition)**. Second Edition. Oxford: Academic Press, 2013. P. 385–418. ISBN 978-0-12-385055-3. DOI: https://doi.org/10.1016/B978-0-12-385055-3.00011-0. Available from: https://www.sciencedirect.com/science/article/pii/B9780123850553000110.

CLERCKX, Bruno et al. Is NOMA Efficient in Multi-Antenna Networks? A Critical Look at Next Generation Multiple Access Techniques. **IEEE Open Journal of the Communications Society**, v. 2, p. 1310–1343, 2021. DOI: 10.1109/OJCOMS.2021.3084799.

DAI, L.; WANG, B.; YUAN, Y.; HAN, S.; I, C.; WANG, Z. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. **IEEE Communications Magazine**, v. 53, n. 9, p. 74–81, Sept. 2015. ISSN 0163-6804. DOI: 10.1109/MCOM.2015.7263349.

DING, Zhiguo; SCHOBER, Robert; FAN, Pingzhi; POOR, H. Vincent. Simple Semi-Grant-Free Transmission Strategies Assisted by Non-Orthogonal Multiple Access. **IEEE Transactions on Communications**, v. 67, n. 6, p. 4464–4478, 2019. DOI: 10.1109/TCOMM.2019.2903443.

DING, Zhiguo; SCHOBER, Robert; POOR, H. Vincent. A New QoS-Guarantee Strategy for NOMA Assisted Semi-Grant-Free Transmission. **IEEE Transactions on Communications**, v. 69, n. 11, p. 7489–7503, 2021. DOI: `10.1109/TCOMM.2021.3100598`.

DIZDAR, Onur; MAO, Yijie; XU, Yunnuo; ZHU, Peiying; CLERCKX, Bruno. Rate-Splitting Multiple Access for Enhanced URLLC and eMBB in 6G: Invited Paper. In: 2021 17th International Symposium on Wireless Communication Systems (ISWCS). [S.l.: s.n.], 2021a. P. 1–6. DOI: `10.1109/ISWCS49558.2021.9562192`.

DIZDAR, Onur; MAO, Yijie; XU, Yunnuo; ZHU, Peiying; CLERCKX, Bruno. Rate-Splitting Multiple Access for Enhanced URLLC and eMBB in 6G: Invited Paper. In: 2021 17th International Symposium on Wireless Communication Systems (ISWCS). [S.l.: s.n.], 2021b. P. 1–6. DOI: `10.1109/ISWCS49558.2021.9562192`.

ELBAYOUMI, Mohammed; HAMOUDA, Walaa; YOUSSEF, Amr. A Hybrid NOMA/OMA Scheme for MTC in Ultra-Dense Networks. In: GLOBECOM 2020 - 2020 IEEE Global Communications Conference. [S.l.: s.n.], 2020. P. 1–6. DOI: `10.1109/GLOBECOM42002.2020.9322207`.

ERICSSON. **5G evolution toward 5G advanced: An overview of 3GPP releases 17 and 18**. Available from: `https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-evolution-toward-5g-advanced`. Visited on: 13 July 2022.

ERICSSON. **Ericsson Mobility Report**. Available from: `https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf`. Visited on: 30 Apr. 2023.

FERRUS, R.; SALLENT, O.; PEREZ-ROMERO, J.; AGUSTI, R. On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration. **IEEE Communications Magazine**, v. 56, n. 5, p. 184–192, 2018.

FLAGSHIP, 6G. Key drivers and research challenges for 6G ubiquitous wireless intelligence. **University of Oulu, Tech. Rep. 6G Research Visions 1**, v. 1, 2019.

FOUKAS, X.; PATOUNAS, G.; ELMOKASHFI, A.; MARINA, M. K. Network Slicing in 5G: Survey and Challenges. v. 55, n. 5, p. 94–100, May 2017. ISSN 0163-6804. DOI: `10.1109/MCOM.2017.1600951`.

FUENTES, M. et al. 5G New Radio Evaluation Against IMT-2020 Key Performance Indicators. **IEEE Access**, v. 8, p. 110880–110896, 2020. DOI: `10.1109/ACCESS.2020.3001641`.

GAO, Yichen; XIA, Bin; XIAO, Kexin; CHEN, Zhiyong; LI, Xiaofan; ZHANG, Sha. Theoretical Analysis of the Dynamic Decode Ordering SIC Receiver for Uplink NOMA Systems. **IEEE Communications Letters**, v. 21, n. 10, p. 2246–2249, 2017. DOI: `10.1109/LCOMM.2017.2720582`.

GIORDANI, M.; POLESE, M.; MEZZAVILLA, M.; RANGAN, S.; ZORZI, M. Toward 6G Networks: Use Cases and Technologies. **IEEE Communications Magazine**, v. 58, n. 3, p. 55–61, 2020. DOI: `10.1109/MCOM.001.1900411`.

GSMA. **Mobile IOT in the 5G future**. Available from: `https://www.ericsson.com/4ac64d/assets/local/reports-papers/5g/doc/gsma-5g-mobile-iot.pdf`. Visited on: 30 Apr. 2023.

ISLAM, S. M. Riazul; AVAZOV, Nurilla; DOBRE, Octavia A.; KWAK, Kyung-sup. Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. **IEEE Communications Surveys Tutorials**, v. 19, n. 2, p. 721–742, 2017. DOI: `10.1109/COMST.2016.2621116`.

ITU-R. **Minimum requirements related to technical performance for IMT-2020 radio interface(s)**. Available from: `https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf`. Visited on: 30 Apr. 2023.

JI, Hyoungju; PARK, Sunho; YEO, Jeongho; KIM, Younsun; LEE, Juho; SHIM, Byonghyo. Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects. **IEEE Wireless Communications**, v. 25, n. 3, p. 124–130, 2018. DOI: `10.1109/MWC.2018.1700294`.

JOUDEH, Hamdi; CLERCKX, Bruno. Robust Transmission in Downlink Multiuser MISO Systems: A Rate-Splitting Approach. **IEEE Transactions on Signal Processing**, v. 64, n. 23, p. 6227–6242, 2016a. DOI: `10.1109/TSP.2016.2591501`.

JOUDEH, Hamdi; CLERCKX, Bruno. Sum-Rate Maximization for Linearly Precoded Downlink Multiuser MISO Systems With Partial CSIT: A Rate-Splitting Approach. **IEEE Transactions on Communications**, v. 64, n. 11, p. 4847–4861, 2016b. DOI: `10.1109/TCOMM.2016.2603991`.

KALØR, A. E.; GUILLAUME, R.; NIELSEN, J. J.; MUELLER, A.; POPOVSKI, P. Network Slicing in Industry 4.0 Applications: Abstraction Methods and End-to-End Analysis. **IEEE Transactions on Industrial Informatics**, v. 14, n. 12, p. 5419–5427, Dec. 2018. DOI: `10.1109/TII.2018.2839721`.

KASSAB, Rahif; MUNARI, Andrea; CLAZZER, Federico; SIMEONE, Osvaldo. Grant-Free Coexistence of Critical and Noncritical IoT Services in Two-Hop Satellite and Terrestrial Networks. **IEEE Internet of Things Journal**, v. 9, n. 16, p. 14829–14843, 2022. DOI: `10.1109/JIOT.2021.3115483`.

KASSAB, Rahif; SIMEONE, Osvaldo; POPOVSKI, Petar; ISLAM, Toufiqul. Non-Orthogonal Multiplexing of Ultra-Reliable and Broadband Services in Fog-Radio Architectures. **IEEE Access**, v. 7, p. 13035–13049, 2019. DOI: `10.1109/ACCESS.2019.2893128`.

KIM, Nam I.; CHO, Dong-Ho. Hybrid Multiple Access System Based on Non Orthogonality and Sparse Code. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC). [S.l.: s.n.], 2017. P. 1–6. DOI: `10.1109/WCNC.2017.7925930`.

KORRAI, Praveen Kumar; LAGUNAS, Eva; SHARMA, Shree Krishna; CHATZINOTAS, Symeon; OTTERSTEN, Björn. Slicing Based Resource Allocation for Multiplexing of eMBB and URLLC Services in 5G Wireless Networks. In: 2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). [S.l.: s.n.], 2019. P. 1–5. DOI: `10.1109/CAMAD.2019.8858433`.

LEE, Byungju; SHIN, Wonjae; POOR, H. Vincent. Weighted Sum-Rate Maximization for Rate-splitting Multiple Access with Imperfect Channel Knowledge. In: 2021 International Conference on Information and Communication Technology Convergence (ICTC). [S.l.: s.n.], 2021. P. 218–220. DOI: `10.1109/ICTC52510.2021.9620766`.

LI, X.; SAMAKA, M.; CHAN, H. A.; BHAMARE, D.; GUPTA, L.; GUO, C.; JAIN, R. Network Slicing for 5G: Challenges and Opportunities. v. 21, n. 5, p. 20–27, 2017. ISSN 1089-7801. DOI: `10.1109/MIC.2017.3481355`.

LIU, Hongwu; KWAK, Kyung Sup. Adaptive Rate Splitting for Uplink Non-Orthogonal Multiple Access Systems. In: 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN). [S.l.: s.n.], 2019. P. 158–163. DOI: `10.1109/ICUFN.2019.8806098`.

LIU, Hongwu; TSIFTSIS, Theodoros A.; CLERCKX, Bruno; KIM, Kyeong Jin; KWAK, Kyung Sup; POOR, H. Vincent. **Rate Splitting Multiple Access for Semi-Grant-Free Transmissions**. [S.l.]: arXiv, 2021. DOI: `10.48550/ARXIV.2110.02127`. Available from: `https://arxiv.org/abs/2110.02127`.

LIU, Hongwu; TSIFTSIS, Theodoros A.; KIM, Kyeong Jin; KWAK, Kyung Sup; POOR, H. Vincent. Rate Splitting for Uplink NOMA With Enhanced Fairness and Outage Performance. **IEEE Transactions on Wireless Communications**, v. 19, n. 7, p. 4657–4670, 2020. DOI: `10.1109/TWC.2020.2985970`.

LIU, Liang; LARSSON, Erik G.; YU, Wei; POPOVSKI, Petar; STEFANOVIC, Cedomir; CARVALHO, Elisabeth de. Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things. **IEEE Signal Processing Magazine**, v. 35, n. 5, p. 88–99, 2018. DOI: `10.1109/MSP.2018.2844952`.

LIU, Yan; DENG, Yansha; ELKASHLAN, Maged; NALLANATHAN, Arumugam; KARAGIANNIDIS, George K. Analyzing Grant-Free Access for URLLC Service. **IEEE Journal on Selected Areas in Communications**, v. 39, n. 3, p. 741–755, 2021. DOI: `10.1109/JSAC.2020.3018822`.

LIU, Ye; DERAKHSHANI, Mahsa; LAMBOTHARAN, Sangarapillai. Outage Analysis and Power Allocation in Uplink Non-Orthogonal Multiple Access Systems. **IEEE Communications Letters**, v. 22, n. 2, p. 336–339, 2018. DOI: `10.1109/LCOMM.2017.2769088`.

LIU, Yuanwei; QIN, Zhijin; ELKASHLAN, Maged; DING, Zhiguo; NALLANATHAN, Arumugam; HANZO, Lajos. Nonorthogonal Multiple Access for 5G and Beyond. **Proceedings of the IEEE**, v. 105, n. 12, p. 2347–2381, 2017. DOI: `10.1109/JPROC.2017.2768666`.

LIU, Yuanwen; CLERCKX, Bruno; POPOVSKI, Petar. **Network Slicing for eMBB, URLLC, and mMTC: An Uplink Rate-Splitting Multiple Access Approach**. [S.l.]: arXiv, 2022. DOI: `10.48550/ARXIV.2208.10841`. Available from: `https://arxiv.org/abs/2208.10841`.

MA, Guoyu; AI, Bo; WANG, Fanggang; CHEN, Xia; ZHONG, Zhangdui; ZHAO, Zhuyan; GUAN, Hao. Coded Tandem Spreading Multiple Access for Massive Machine-Type Communications. **IEEE Wireless Communications**, v. 25, n. 2, p. 75–81, 2018. DOI: `10.1109/MWC.2018.1700107`.

MAATOUK, Ali; ASSAAD, Mohamad; EPHREMIDES, Anthony. Minimizing The Age of Information: NOMA or OMA? In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). [S.l.: s.n.], 2019. P. 102–108. DOI: `10.1109/INFCOMW.2019.8845254`.

MAHMOOD, Nurul Huda et al. White Paper on Critical and Massive Machine Type Communication Towards 6G. **arXiv:2004.14146 [cs, eess]**, May 2020. arXiv: 2004.14146. Available from: `http://arxiv.org/abs/2004.14146`. Visited on: 7 Dec. 2021.

MAHMOUDI, Ali; ABOLHASSANI, Bahman; RAZAVIZADEH, S. Mohammad; NGUYEN, Ha H. User Clustering and Resource Allocation in Hybrid NOMA-OMA Systems Under Nakagami-m Fading. **IEEE Access**, v. 10, p. 38709–38728, 2022. DOI: `10.1109/ACCESS.2022.3165756`.

MAO, Yijie; CLERCKX, Bruno; LI, Victor O.K. Energy Efficiency of Rate-Splitting Multiple Access, and Performance Benefits over SDMA and NOMA. In: 2018 15th International Symposium on Wireless Communication Systems (ISWCS). [S.l.: s.n.], 2018. P. 1–5. DOI: `10.1109/ISWCS.2018.8491100`.

MAO, Yijie; CLERCKX, Bruno; LI, Victor O.K. Rate-Splitting for Multi-User Multi-Antenna Wireless Information and Power Transfer. In: 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). [S.l.: s.n.], 2019. P. 1–5. DOI: `10.1109/SPAWC.2019.8815494`.

MARCANO, A. S.; CHRISTIANSEN, H. L. A Novel Method for Improving the Capacity in 5G Mobile Networks Combining NOMA and OMA. In: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring). [S.l.: s.n.], 2017. P. 1–5. DOI: `10.1109/VTCSpring.2017.8108677`.

MEYER, Florian; TURAU, Volker. QMA: A Resource-efficient, Q-learning-based Multiple Access Scheme for the IIoT. In: 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). [S.l.: s.n.], 2021. P. 864–874. DOI: `10.1109/ICDCS51616.2021.00087`.

MISHRA, Anup; MAO, Yijie; SANGUINETTI, Luca; CLERCKX, Bruno. Rate-Splitting Assisted Massive Machine-Type Communications in Cell-Free Massive MIMO. **IEEE Communications Letters**, v. 26, n. 6, p. 1358–1362, 2022. DOI: `10.1109/LCOMM.2022.3160511`.

NEW, RCR Wireless. **What is mTRP for eMBB?** Available from: `https://www.rcrwireless.com/20220422/5g/what-is-mtrp-for-embb`. Visited on: 13 July 2022.

NGUYEN, Dinh C.; DING, Ming; PATHIRANA, Pubudu N.; SENEVIRATNE, Aruna; LI, Jun; NIYATO, Dusit; DOBRE, Octavia; POOR, H. Vincent. 6G Internet of Things: A Comprehensive Survey. **IEEE Internet of Things Journal**, v. 9, n. 1, p. 359–383, 2022. DOI: `10.1109/JIOT.2021.3103320`.

NOKIA. **Automated network slicing**. Available from: `https://www.nokia.com/networks/network-slicing/`. Visited on: 23 Aug. 2022.

OZAKU, Shinichi; SHIMBO, Yukiko; SUGANUMA, Hirofumi; MAEHARA, Fumiaki. Adaptive Repetition Control Using Terminal Mobility for Uplink Grant-Free URLLC. In: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring). [S.l.: s.n.], 2020. P. 1–5. DOI: `10.1109/VTC2020-Spring48590.2020.9128403`.

POPOVSKI, P. Ultra-reliable communication in 5G wireless systems. In: 1ST International Conference on 5G for Ubiquitous Connectivity. [S.l.: s.n.], Nov. 2014. P. 146–151. DOI: `10.4108/icst.5gu.2014.258154`.

POPOVSKI, P.; TRILLINGSGAARD, K. F.; SIMEONE, O.; DURISI, G. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. **IEEE Access**, v. 6, p. 55765–55779, 2018. ISSN 2169-3536. DOI: `10.1109/ACCESS.2018.2872781`.

POPOVSKI, Petar et al. Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks. **IEEE Network**, v. 32, n. 2, p. 16–23, 2018. DOI: `10.1109/MNET.2018.1700258`.

RAUNIYAR, Ashish; ENGELSTAD, Paal; ØSTERBØ, Olav N. An Adaptive User Pairing Strategy for Uplink Non-Orthogonal Multiple Access. In: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. [S.l.: s.n.], 2020. P. 1–7. DOI: `10.1109/PIMRC48278.2020.9217383`.

RIMOLDI, B.; URBANKE, R. A rate-splitting approach to the Gaussian multiple-access channel. **IEEE Transactions on Information Theory**, v. 42, n. 2, p. 364–375, 1996. DOI: `10.1109/18.485709`.

ROSAS, F.; SOUZA, R. D.; VERHELST, M.; POLLIN, S. Energy-efficient MIMO multihop communications using the antenna selection scheme. In: IEEE Intern. Symp. on Wireless Commun. Systems (ISWCS). [S.l.: s.n.], 2015. P. 686–690.

SAAD, W.; BENNIS, M.; CHEN, M. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. **IEEE Network**, v. 34, n. 3, p. 134–142, 2020. DOI: `10.1109/MNET.001.1900287`.

SAMAD, A.; NANDANA, R.; WALID, S. Fast Uplink Grant for Machine Type Communications: Challenges and Opportunities. **IEEE Communications Magazine**, v. 57, n. 3, p. 97–103, 2019. DOI: `10.1109/MCOM.2019.1800475`.

SANTOS, Elço João dos; SOUZA, Richard Demo; REBELATTO, João Luiz; ALVES, Hirley. Network Slicing for URLLC and eMBB With Max-Matching Diversity Channel Allocation. **IEEE Communications Letters**, v. 24, n. 3, p. 658–661, 2020. DOI: `10.1109/LCOMM.2019.2959335`.

SEDAGHAT, Mohammad Ali; MÜLLER, Ralf R. On User Pairing in Uplink NOMA. **IEEE Transactions on Wireless Communications**, v. 17, n. 5, p. 3474–3486, 2018. DOI: `10.1109/TWC.2018.2815005`.

SHAFI, M.; MOLISCH, A. F.; SMITH, P. J.; HAUSTEIN, T.; ZHU, P.; SILVA, P. De; TUFVESSON, F.; BENJEBBOUR, A.; WUNDER, G. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice. v. 35, n. 6, p. 1201–1221, June 2017. ISSN 0733-8716. DOI: `10.1109/JSAC.2017.2692307`.

SHARMA, Shree Krishna; WANG, Xianbin. Collaborative Distributed Q-Learning for RACH Congestion Minimization in Cellular IoT Networks. **IEEE Communications Letters**, v. 23, n. 4, p. 600–603, 2019. DOI: `10.1109/LCOMM.2019.2896929`.

SPECTRUM, IEEE. **3GPP Release 15 Overview**. Available from: `https://spectrum.ieee.org/telecom/wireless/3gpp-release-15-overview`. Visited on: 18 Mar. 2021.

SUGANUMA, H.; SUENAGA, H.; MAEHARA, F. Hybrid Multiple Access Using Simultaneously NOMA and OMA. In: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). [S.l.: s.n.], 2019. P. 1–2. DOI: 10.1109/ISPACS48206.2019.8986345.

TAKEDA, K.; HARADA, H.; OSAWA, R. **NR physical layer specifications in 5G**. Available from: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical%5C_journal/bn/vol20%5C_3/vol20%5C_3%5C_007en.pdf. Visited on: 30 Apr. 2023.

TANAKA, Fuga; SUGANUMA, Hirofumi; MAEHARA, Fumiaki. Hybrid Multiple Access Scheme Using NOMA and OMA Simultaneously Considering User Request. In: 2021 24th International Symposium on Wireless Personal Multimedia Communications (WPMC). [S.l.: s.n.], 2021. P. 1–5. DOI: 10.1109/WPMC52694.2021.9700453.

TARIQ, F.; KHANDAKER, M. R. A.; WONG, K. -K.; IMRAN, M. A.; BENNIS, M.; DEBBAH, M. A Speculative Study on 6G. **IEEE Wireless Communications**, v. 27, n. 4, p. 118–125, 2020. DOI: 10.1109/MWC.001.1900488.

TOMINAGA, Eduardo Noboro; ALVES, Hirley; LÓPEZ, Onel L. Alcaraz; SOUZA, Richard Demo; REBELATTO, João Luiz; LATVA-AHO, Matti. Network Slicing for eMBB and mMTC with NOMA and Space Diversity Reception. In: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). [S.l.: s.n.], 2021a. P. 1–6. DOI: 10.1109/VTC2021-Spring51267.2021.9448974.

TOMINAGA, Eduardo Noboro; ALVES, Hirley; SOUZA, Richard Demo; LUIZ REBELATTO, João; LATVA-AHO, Matti. Non-Orthogonal Multiple Access and Network Slicing: Scalable Coexistence of eMBB and URLLC. In: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). [S.l.: s.n.], 2021b. P. 1–6. DOI: 10.1109/VTC2021-Spring51267.2021.9448942.

TSE, David; VISWANATH, Pramod. **Fundamentals of Wireless Communication**. [S.l.]: Cambridge University Press, 2005a. DOI: 10.1017/CBO9780511807213.

TSE, David; VISWANATH, Pramod. **Fundamentals of Wireless Communication**. [S.l.]: Cambridge University Press, 2005b. DOI: 10.1017/CBO9780511807213.

TULLBERG, H.; POPOVSKI, P.; LI, Z.; UUSITALO, M. A.; HOGLUND, A.; BULAKCI, O.; FALLGREN, M.; MONSERRAT, J. F. The METIS 5G System Concept: Meeting the 5G Requirements. v. 54, n. 12, p. 132–139, Dec. 2016. DOI: 10.1109/MCOM.2016.1500799CM.

UNION, International Telecommunication. IMT traffic estimates for the years 2020 to 2030. **Report ITU-R M.2370-0**, v. 1, 2015.

WANG, Chao; CHEN, Yan; WU, Yiqun; ZHANG, Liqing. Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services. In: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring). [S.l.: s.n.], 2017. P. 1–6. DOI: 10.1109/VTCSpring.2017.8108593.

WANG, Qian; CHEN, He; LI, Yonghui; VUCETIC, Branka. Minimizing Age of Information via Hybrid NOMA/OMA. In: 2020 IEEE International Symposium on Information Theory (ISIT). [S.l.: s.n.], 2020. P. 1753–1758. DOI: 10.1109/ISIT44484.2020.9174163.

YANG, W.; DURISI, G.; KOCH, T.; POLYANSKIY, Y. Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength. v. 60, n. 7, p. 4232–4265, July 2014. DOI: 10.1109/TIT.2014.2318726.

YANG, Zhaohui; CHEN, Mingzhe; SAAD, Walid; XU, Wei; SHIKH-BAHAEI, Mohammad. Sum-Rate Maximization of Uplink Rate Splitting Multiple Access (RSMA) Communication. In: 2019 IEEE Global Communications Conference (GLOBECOM). [S.l.: s.n.], 2019. P. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013344.

YANG, Zhaohui; CHEN, Mingzhe; SAAD, Walid; XU, Wei; SHIKH-BAHAEI, Mohammad. Sum-Rate Maximization of Uplink Rate Splitting Multiple Access (RSMA) Communication. **IEEE Transactions on Mobile Computing**, p. 1–1, 2020a. DOI: 10.1109/TMC.2020.3037374.

YANG, Zhaohui; CHEN, Mingzhe; SAAD, Walid; XU, Wei; SHIKH-BAHAEI, Mohammad. Sum-Rate Maximization of Uplink Rate Splitting Multiple Access (RSMA) Communication. **IEEE Transactions on Mobile Computing**, p. 1–1, 2020b. DOI: 10.1109/TMC.2020.3037374.

ZAIDI, A.; WANG, Z. **Designing for the future the 5G NR physical layer**. Available from: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/designing-for-the-future-the-5g-nr-physical-layer. Visited on: 30 Apr. 2023.

ZAIDI, Ali; ATHLEY, Fredrik; MEDBO, Jonas; GUSTAVSSON, Ulf; DURISI, Giuseppe; CHEN, Xiaoming. 5G Physical Layer: Principles, Models and Technology Components. In: London: Academic Press, 2018.

ZAIDI, Ali A.; BALDEMAIR, Robert; ANDERSSON, Mattias; FAXÉR, Sebastian; MOLÉS-CASES, Vicent; WANG, Zhao. **Designing for the future: the 5G NR physical layer**. Available from: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/designing-for-the-future-the-5g-nr-physical-layer. Visited on: 30 Apr. 2023.

ZENG, Jie; LV, Tiejun; NI, Wei; LIU, Ren Ping; BEAULIEU, Norman C.; GUO, Y. Jay. Ensuring Max–Min Fairness of UL SIMO-NOMA: A Rate Splitting Approach. **IEEE Transactions on Vehicular Technology**, v. 68, n. 11, p. 11080–11093, 2019. DOI: `10.1109/TVT.2019.2943511`.

ZHANG, Chao; LIU, Yuanwei; DING, Zhiguo. Semi-Grant-Free NOMA: A Stochastic Geometry Model. **IEEE Transactions on Wireless Communications**, v. 21, n. 2, p. 1197–1213, 2022. DOI: `10.1109/TWC.2021.3103036`.

ZHANG, Chao; LIU, Yuanwei; YI, Wenqiang; QIN, Zhijin; DING, Zhiguo. Semi-Grant-Free NOMA: Ergodic Rates Analysis With Random Deployed Users. **IEEE Wireless Communications Letters**, v. 10, n. 4, p. 692–695, 2021. DOI: `10.1109/LWC.2020.3034725`.

ZHANG, Haijun; LIU, Na; CHU, Xiaoli; LONG, Keping; AGHVAMI, Abdol-Hamid; LEUNG, Victor C. M. Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges. **IEEE Communications Magazine**, v. 55, n. 8, p. 138–145, 2017. DOI: `10.1109/MCOM.2017.1600940`.

ZHANG, Ningbo; ZHU, Xuzhen. A Hybrid Grant NOMA Random Access for Massive MTC Service. **IEEE Internet of Things Journal**, p. 1–1, 2022. DOI: `10.1109/JIOT.2022.3222622`.

ZHANG, Zekun; HU, Rose Qingyang. Uplink Non-Orthogonal Multiple Access with Fractional Power Control. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC). [S.l.: s.n.], 2017. P. 1–6. DOI: `10.1109/WCNC.2017.7925935`.

ZHU, Ye; WANG, Xianbin; ZHANG, Zhaoyang; CHEN, Xiaoming; CHEN, Yan. A rate-splitting non-orthogonal multiple access scheme for uplink transmission. In: 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP). [S.l.: s.n.], 2017. P. 1–6. DOI: `10.1109/WCSP.2017.8171078`.

Appendix

APPENDIX A – EMBB EQUATIONS

This appendix discusses the mathematical model proposed in this work to analyze the performance of the network in terms of channel capacity and outage probability, demonstrating the deduction of equations related to eMBB.

A.1 EMBB

Let us consider a scenario where a radio resource $f \in \{1, \dots, F\}$ is allocated exclusively to an eMBB user. The main objective of eMBB is to maximize its data rate, subject to the reliability requirement ϵ_B and the average power constraint $P_B = 1$.

Assuming that the BS is capable of acquiring the CSI of eMBB devices, as considered in (POPOVSKI, P. et al., 2018) and utilized in current wireless standards such as Long Term Evolution (LTE) and 5G New Radio (TAKEDA et al., 2023; ZAIDI, A.; WANG, Z., 2023; 3GPP, 2023), a transmission with a determined instantaneous power and data rate only occurs when $G_{B,f}$, the instantaneous channel gain, is greater than a threshold SNR $G_{B,f}^{min}$. The outage probability of a point-to-point (single channel) communication is then

$$\begin{aligned} \mathcal{P}_s(\bar{\gamma}_B) &= \Pr[G_{B,f} < G_{B,f}^{min}] \\ &= \int_0^{G_{B,f}^{min}} p_{G_{B,f}}(x)dx, \end{aligned} \tag{55}$$

where $p_{G_{B,f}}(x)$ is the PDF of $G_{B,f}$, which is the function that represents the probability of $G_{B,f}$ be smaller than $G_{B,f}^{min}$. It is considered that the channel fading is modeled as Rayleigh, which implicates that the PDF can be expressed as

$$p_{G_{B,f}}(x) = \begin{cases} \frac{e^{-x/\bar{\gamma}_B}}{\bar{\gamma}_B}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{56}$$

The eMBB outage probability is then

$$\begin{aligned} \mathcal{P}_s(\bar{\gamma}_B) &= \int_0^{G_{B,f}^{min}} \frac{e^{-x/\bar{\gamma}_B}}{\bar{\gamma}_B} dx \\ &= \frac{1}{\bar{\gamma}_B} \times -\bar{\gamma}_B \times e^{-x/\bar{\gamma}_B} \Big|_0^{G_{B,f}^{min}} \\ &= -\left(e^{-G_{B,f}^{min}/\bar{\gamma}_B} - e^{-0/\bar{\gamma}_B}\right) \\ &= 1 - e^{-G_{B,f}^{min}/\bar{\gamma}_B}. \end{aligned} \tag{57}$$

After imposing the reliability constraint $P_{\text{out}} = \epsilon_B$, we obtain the threshold SNR from (57) as

$$\begin{aligned}
\mathcal{P}_s(\bar{\gamma}_B) &= \epsilon_B \\
1 - e^{-G_{B,f}^{min}/\bar{\gamma}_B} &= \epsilon_B \\
e^{-G_{B,f}^{min}/\bar{\gamma}_B} &= 1 - \epsilon_B \\
-\frac{G_{B,f}^{min}}{\bar{\gamma}_B} &= \ln(1 - \epsilon_B) \\
G_{B,f}^{min} &= -\bar{\gamma}_B \ln(1 - \epsilon_B)\,.
\end{aligned} \tag{58}$$

As already mentioned, it t is practical to consider that the BS is capable of obtaining the CSI $G_{B,f}$ of eMBB users. This information is used to select its transmission power based on power inversion scheme (CAIRE et al., 1999), where a value is chosen based on $G_{B,f}$ as

$$P_B(G_{B,f}) = \begin{cases} \frac{G_{B,f}^{\text{tar}}}{G_{B,f}}, & \text{if } G_{B,f} \geq G_{B,f}^{min} \\ 0, & \text{otherwise.} \end{cases} \tag{59}$$

This means that the eMBB device does not necessarily transmit in every slot allocated to it because of outage situations, then it is possible to increase the instantaneous power when the transmission occurs, so that the long-term average power $\left(P_B(G_{B,f}) = 1\right)$ is achieved. The target SNR $G_{B,f}^{\text{tar}}$ is then obtained by imposing the average power constraint on the expected value of the function P_B of the random variable $G_{B,f}$. This is calculated using the Law of the Unconscious Statistician (LOTUS) theorem as

$$\mathbb{E}\left[P_B(G_{B,f})\right] = \int_{G_{B,f}^{min}}^{\infty} p_{G_{B,f}}(x) P_B(x) dx = 1, \tag{60}$$

where $p_{G_{B,f}}(x)$ is the PDF of $G_{B,f}$, obtained from the Cumulative Density Function (CDF) $\mathcal{P}_s(\bar{\gamma}_B)$ as

$$\begin{aligned}
p_{G_{B,f}}(x) &= \frac{d}{dx}\left[\mathcal{P}_s(\bar{\gamma}_B)\right] \\
&= \frac{d}{dx}\left[(1 - e^{-x/\bar{\gamma}_B})\right] \\
&= \frac{1}{\bar{\gamma}_B} e^{-x/\bar{\gamma}_B}.
\end{aligned} \tag{61}$$

Replacing (61) in (60):

$$
\begin{aligned}
\mathbb{E}\left[P_B(G_{B,f})\right] &= \int_{G_{B,f}^{min}}^{\infty} \underbrace{\frac{1}{\bar{\gamma}_B} e^{-x/\bar{\gamma}_B}}_{p_{G_{B,f}}(x)} \underbrace{\frac{G_{B,f}^{\mathrm{tar}}}{x}}_{P_B(x)} dx = 1 \\
&= \frac{G_{B,f}^{\mathrm{tar}}}{\bar{\gamma}_B} \underbrace{\int_{G_{B,f}^{min}}^{\infty} \frac{e^{-x/\bar{\gamma}_B}}{x} dx}_{-\mathrm{Ei}\left(-\frac{G_{B,f}^{min}}{\bar{\gamma}_B}\right)} = 1,
\end{aligned}
\tag{62}
$$

where $-\mathrm{Ei}\left(-\frac{G_{B,f}^{min}}{\bar{\gamma}_B}\right)$ is obtained from the integral and can be classified as the upper incomplete gamma function $\Gamma(\cdot,\cdot)$ for $G_{B,f}^{min} > 0$. Then, (62) can be rewritten as

$$
\mathbb{E}\left[P_B(G_{B,f})\right] = \frac{G_{B,f}^{\mathrm{tar}}}{\bar{\gamma}_B} \Gamma\left(0, \frac{G_{B,f}^{min}}{\bar{\gamma}_B}\right) = 1. \tag{63}
$$

By isolating $G_{B,f}^{\mathrm{tar}}$ from (63), the target SNR of eMBB user becomes

$$
G_{B,f}^{\mathrm{tar}} = \frac{\bar{\gamma}_B}{\Gamma\left(0, \frac{G_{B,f}^{min}}{\bar{\gamma}_B}\right)}. \tag{64}
$$

Therefore, the eMBB rate is finally

$$
R_B = \log_2\left(1 + G_{B,f}^{\mathrm{tar}}\right). \qquad \text{(bits/symbol)} \tag{65}
$$

APPENDIX B – EMBB WITH MMD EQUATIONS

The main objective of MMD is to maximize the channel capacity of eMBB users, *i.e.*, enable the transmission, under a given reliability, of the highest possible number of bits per symbol, respecting the constraint of communication reliability in different coexistence schemes with URLLC devices.

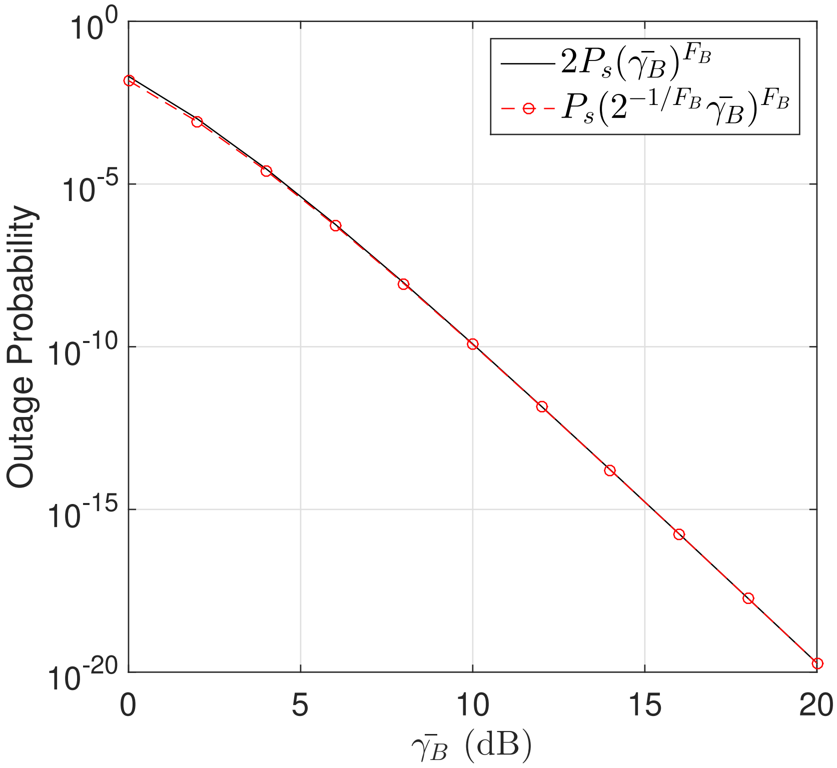
B.1 EMBB WITH MMD

When considering MMD channel allocation from (BAI et al., 2010) with F_B independent channels, the outage probability of a eMBB user, in a scenario where the number of users is equal to F_B (the case of our work), can be approximated following (6) as

$$P_{\text{out}} \approx 2\mathcal{P}_s(\bar{\gamma}_B)^{F_B} \approx \mathcal{P}_s(\bar{\gamma}_B^{'})^{F_B}, \tag{66}$$

where $\bar{\gamma}_B^{'} \triangleq 2^{-(1/F_B)}\bar{\gamma}_B$. This approximation is valid, as shown in Figure 23, where both functions were evaluated for $F_B = 10$ and $\bar{\gamma}_B \in \{0, \ldots, 20\}$ dB.

Figure 23 – Verification of $\mathcal{P}_s(\bar{\gamma}_B)$ approximation for $F_B = 10$ and $\bar{\gamma}_B \in \{0, \ldots, 20\}$ dB.



Source: The Author.

After imposing the reliability constraint $P_{\text{out}} \approx \mathcal{P}_s(\bar{\gamma}_B^{'})^{F_B} = \epsilon_B$, we obtain the threshold SNR from (57) and (66) as

$$
\begin{aligned}
\mathcal{P}_s(\bar{\gamma}_B^{'})^{F_B} &= \epsilon_B \\
\left(1 - e^{-G_{B,f}^{min}/\bar{\gamma}_B^{'}}\right)^{F_B} &= \epsilon_B \\
1 - e^{-G_{B,f}^{min}/\bar{\gamma}_B^{'}} &= \epsilon_B{}^{1/F_B} \\
e^{-G_{B,f}^{min}/\bar{\gamma}_B^{'}} &= 1 - \epsilon_B{}^{1/F_B} \\
-\frac{G_{B,f}^{min}}{\bar{\gamma}_B^{'}} &= \ln\left(1 - \epsilon_B{}^{1/F_B}\right) \\
G_{B,f}^{min} &= -\bar{\gamma}_B^{'} \ln\left(1 - \epsilon_B{}^{1/F_B}\right).
\end{aligned}
\tag{67}
$$

As already mentioned, it is practical to consider that the BS is capable of obtaining the CSI $G_{B,f}^{\text{MMD}}$ (the instantaneous channel gain after MMD) of eMBB users. This information is used to select its transmission power based on power inversion scheme (CAIRE et al., 1999), where a value is chosen based on $G_{B,f}^{\text{MMD}}$ as

$$
P_B(G_{B,f}^{\text{MMD}}) = \begin{cases} \frac{G_{B,f}^{\text{tar}}}{G_{B,f}^{\text{MMD}}}, & \text{if } G_{B,f}^{\text{MMD}} \geq G_{B,f}^{min} \\ 0, & \text{otherwise.} \end{cases} \tag{68}
$$

Similarly to the reason discussed in Appendix A, one can increase the instantaneous power so that the long-term average power $\left(P_B\left(G_{B,f}^{\text{MMD}}\right) = 1\right)$ is achieved. The target SNR $G_{B,f}^{\text{tar}}$ is then obtained by imposing the average power constraint on the expected value of the function P_B of the random variable $G_{B,f}^{\text{MMD}}$, calculated as

$$
\mathbb{E}\left[P_B(G_{B,f}^{\text{MMD}})\right] = \int_{G_{B,f}^{min}}^{\infty} p_{G_{B,f}^{\text{MMD}}}(x) P_B(x) dx = 1, \tag{69}
$$

where $p_{G_{B,f}^{\text{MMD}}}(x)$ is the PDF of $G_{B,f}^{\text{MMD}}$, obtained from the CDF $\mathcal{P}_s(\bar{\gamma}_B^{'})^{F_B}$ as

$$
\begin{aligned}
p_{G_{B,f}^{\mathrm{MMD}}}(x) &= \frac{d}{dx}\left[\mathcal{P}s(\bar{\gamma}_B')^{F_B}\right] \\
&= \frac{d}{dx}\left[(1-e^{-x/\bar{\gamma}_B'})^{F_B}\right] \\
&= \frac{F_B}{\bar{\gamma}_B'} \underbrace{\left[1-e^{-x/\bar{\gamma}_B'}\right]^{F_B-1}}_{\text{Apply binomial theorem}} e^{-x/\bar{\gamma}_B'} \\
&= \frac{F_B}{\bar{\gamma}_B'} e^{-x/\bar{\gamma}_B'} \sum_{f=0}^{F_B-1} (-1)^f \binom{F_B-1}{f} e^{-fx/\bar{\gamma}_B'} \\
&= \sum_{f=0}^{F_B-1} \frac{F_B!(f+1)}{(f+1)!(F_B-1-f)!} \frac{(-1)^f}{\bar{\gamma}_B'} e^{-(f+1)x/\bar{\gamma}_B'} \\
&= \sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} \frac{f e^{-fx/\bar{\gamma}_B'}}{\bar{\gamma}_B'}.
\end{aligned}
\tag{70}
$$

The summation in (70) is obtained applying the binomial theorem (ROSAS et al., 2015), so that the PDF of the SNR of F_B channels can be expressed as a linear combination of F_B exponential PDFs. Replacing (70) in (69):

$$
\begin{aligned}
\mathbb{E}\left[P_B(G_{B,f}^{\mathrm{MMD}})\right] &= \int_{G_{B,f}^{min}}^{\infty} \underbrace{\sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} \frac{f e^{-fx/\bar{\gamma}_B'}}{\bar{\gamma}_B'}}_{p_{G_{B,f}^{\mathrm{MMD}}}(x)} \underbrace{\frac{G_{B,f}^{\mathrm{tar}}}{x}}_{P_B(x)} dx = 1 \\
&= G_{B,f}^{\mathrm{tar}} \sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} \frac{f}{\bar{\gamma}_B'} \underbrace{\int_{G_{B,f}^{min}}^{\infty} \frac{e^{-fx/\bar{\gamma}_B'}}{x} dx}_{-\mathrm{Ei}\left(-\frac{f G_{B,f}^{min}}{\bar{\gamma}_B'}\right)} = 1,
\end{aligned}
\tag{71}
$$

where $-\mathrm{Ei}\left(-\frac{f G_{B,f}^{min}}{\bar{\gamma}_B'}\right)$ is obtained from the integral and can be classified as the upper incomplete gamma function $\Gamma(\cdot,\cdot)$ for $G_{B,f}^{min} > 0$. Then, equation (71) can be rewritten as

$$
\mathbb{E}\left[P_B(G_{B,f}^{\mathrm{MMD}})\right] = G_{B,f}^{\mathrm{tar}} \sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} \frac{f}{\bar{\gamma}_B'} \Gamma\left(0, \frac{f G_{B,f}^{min}}{\bar{\gamma}_B'}\right) = 1. \tag{72}
$$

Then, isolating $G_{B,f}^{\mathrm{tar}}$ from (72), the target SNR of eMBB user is obtained, resulting in

$$G_{B,f}^{\mathrm{tar}} = \frac{\bar{\gamma}_B^{'}}{\sum_{f=1}^{F_B} (-1)^{f-1} \binom{F_B}{f} f \Gamma \left(0, \frac{f G_{B,f}^{min}}{\bar{\gamma}_B^{'}} \right)}. \tag{73}$$

Therefore, the MMD-aided eMBB rate is

$$R_B^{\mathrm{MMD}} = \log_2 \left(1 + G_{B,f}^{\mathrm{tar}} \right). \qquad \text{(bits/symbol)} \tag{74}$$

APPENDIX C – THRESHOLD SNR IN NOMA SLICING

An eMBB message would not be affected by URLLC interference in two cases: (i) there is no URLLC device connected ($S_U = 0$); or (ii) there are URLLC transmissions ($S_U > 0$), but they were decoded and removed from the signal by the SIC decoder. In case (ii), either all URLLC messages are properly decoded (event $\bar{E}_U$) or they are all incorrectly decoded (event E_U), since interference from eMBB users are constant over all mini-slots. Thus, to formulate the eMBB outage probability in NOMA scenario, it is necessary to distinguish the case when eMBB is under interference of URLLC service, and the case in which is not. For this purpose, the total law of probability is applied as follows

$$\begin{aligned} P_{\text{out}} =& \Pr(S_U = 0)\Pr(E_B|S_U = 0) \\ &+ \Pr(S_U > 0)(\Pr(E_U|S_U > 0)\Pr(E_B|E_U, S_U > 0) \\ &+ \Pr(\bar{E}_U|S_U > 0)\Pr(E_B|\bar{E}_U, S_U > 0)), \end{aligned} \tag{75}$$

where E_B is the event of eMBB message not being correctly decoded. The only source of outage for eMBB when there is no URLLC signal interfering is when the SNR value is below the threshold SNR ($G_{B,f}^{min}$), which implies that the term $\Pr(E_B|S_U = 0)$ from (75) equals the outage probability for the orthogonal case, which is $1 - e^{-G_{B,f}^{min}/\bar{\gamma}_B'}$, that will be rewritten as $1 - a_B$ for simplification purposes. Moreover, it is considered that when the URLLC message is incorrectly decoded, the eMBB user is in outage, i.e., $\Pr(E_B|E_U, S_U > 0) \leq 1$. Besides that, the correct decodification and cancellation of URLLC signal has the same performance effect of the case when URLLC is not transmitting, thus, $\Pr(E_B|\bar{E}_U, S_U > 0) = \Pr(E_B|S_U = 0) = 1 - a_B$. With these considerations, (75) can be bounded as

$$P_{\text{out}} \leq (1 - a_U)^S(1 - a_B) + \left(1 - (1 - a_U)^S\right)(\epsilon_U + (1 - \epsilon_U)(1 - a_B)). \tag{76}$$

Imposing the eMBB reliability condition when the MMD allocation algorithm is used

$$P_{\text{out}} \approx \mathcal{P}_S(\bar{\gamma}_B')^{F_B} \leq \epsilon_B \Longrightarrow P_{\text{out}} \leq {\epsilon_B}^{1/F_B},$$

(76) can be rewritten as

$$
\begin{aligned}
&(1-a_U)^S(1-a_B)+\left(1-(1-a_U)^S\right)(\epsilon_U+(1-\epsilon_U)(1-a_B)) \leq {\epsilon_B}^{1/F_B}\\
&(1-a_B)\left((1-a_U)^S+1-\epsilon_U-(1-\epsilon_U)(1-a_U)^S\right)+\epsilon_U\left(1-(1-a_U)^S\right) \leq {\epsilon_B}^{1/F_B}\\
&(1-a_B)\left(\epsilon_U(1-a_U)^S+1-\epsilon_U\right)+\epsilon_U\left(1-(1-a_U)^S\right) \leq {\epsilon_B}^{1/F_B}\\
&\epsilon_U\left((1-a_U)^S-1\right)+1-\epsilon_U a_B\left((1-a_U)^S-1\right)-a_B+\epsilon_U\left(1-(1-a_U)^S\right) \leq {\epsilon_B}^{1/F_B}\\
&\epsilon_U a_B-a_B-\epsilon_U a_B(1-a_U)^S \leq {\epsilon_B}^{1/F_B}-1\\
&a_B \geq \frac{1-{\epsilon_B}^{1/F_B}}{1-\epsilon_U\left(1-(1-a_U)^S\right)}.
\end{aligned}
\tag{77}
$$

Knowing that $a_B = e^{-G_{B,f}^{min}/\bar{\gamma}_B'}$, it is possible to isolate the threshold SNR from (77), resulting in

$$G_{B,f}^{min} \leq -\bar{\gamma}_B' \ln\left(\frac{1-{\epsilon_B}^{1/F_B}}{1-\epsilon_U\left(1-(1-a_U)^S\right)}\right). \tag{78}$$

The target SNR $G_{B,f}^{\text{tar}}$ is obtained in the same way of (64), however, in the non-orthogonal case, $G_{B,f}^{min}$ is bounded by (78):

$$G_{B,f}^{\text{tar}} = \frac{\bar{\gamma}_B'}{\sum\limits_{f=1}^{F_B}(-1)^{f-1}\binom{F_B}{f} f\Gamma\left(0, \frac{fG_{B,f}^{min}}{\bar{\gamma}_B'}\right)}. \tag{79}$$

Therefore, the achievable rate of a MMD-aided eMBB device in NOMA is

$$R_B^{\text{NOMA}} = \log_2\left(1+G_{B,f}^{\text{tar}}\right). \qquad \text{(bits/symbol)} \tag{80}$$