Garantias de Latência sob Incerteza: Planejamento de Fatias URLLC em Redes B5G

Pedro Mendes da Silva Júnior[1] **e Flávio Geraldo Coelho Rocha**[1]

[1]Universidade Federal de Goiás (UFG) – Goiânia – GO – Brasil

{pedro_junior@discente.ufg.br, flaviogcr@ufg.br}

Resumo. *Este trabalho propõe um arcabouço estocástico para o planejamento de fatias URLLC em redes B5G, fundamentado no Cálculo de Rede Estocástico (SNC). A partir da modelagem probabilística dos processos de chegada e de serviço, é derivada uma expressão analítica em forma fechada que relaciona requisitos de latência, probabilidade de violação e, o mais importante, número máximo de usuários admissíveis por célula e fatia. O modelo permite dimensionar recursos de rádio com garantias probabilísticas de QoS, evitando o superdimensionamento de abordagens determinísticas. Resultados numéricos evidenciam ganhos significativos de capacidade quando comparados aos resultados obtidos utilizando Cálculo de Rede Determinístico (DNC). O arcabouço proposto fornece subsídios relevantes para o planejamento eficiente de fatias URLLC em ambientes multicelulares heterogêneos.*

Abstract. *This paper proposes a stochastic framework for planning URLLC slices in B5G networks based on Stochastic Network Calculus (SNC). By probabilistically modeling traffic arrivals and service processes, a closed-form analytical expression is derived that explicitly relates latency bounds, violation probabilities, and most importantly, the maximum number of admissible users per cell and slice. The proposed formulation enables principled radio resource dimensioning under probabilistic QoS guarantees, mitigating the overprovisioning inherent to deterministic approaches. Numerical results demonstrate substantial capacity gains compared to those obtained using Deterministic Network Calculus (DNC). The proposed framework provides valuable insights for efficient URLLC slice planning in heterogeneous multicellular scenarios.*

1. Introdução

As redes de comunicação da próxima geração enfrentarão desafios sem precedentes para atender às demandas de aplicações críticas que foram previstas para a era do B5G, mas que não foram materializadas ou popularizadas, como cirurgias remotas, serviços avançados com drones autônomos, veículos terrestres interconectados e ampla implementação das tecnologias para desenvolvimento da indústria 4.0. Todas essas aplicações são ancoradas no caso de uso do 3GPP denominado *Ultra-Reliable and Low-Latency Communications* (URLLC) com requisitos rigorosos por baixa latência [Adamuz-Hinojosa et al. 2023].

A tecnologia de fatiamento da rede (*network slicing*) nasceu com o B5G como uma solução promissora e principal viabilizadora dessas aplicações a serem executadas sobre uma infraestrutura física existente, visto que o fatiamento de rede fornece redes logicamente separadas e auto-contidas sobre uma infraestrutura física compartilhada [Choi et al. 2024].

Por meio de fatias URLLC, a rede pode ser planejada para suportar serviços que demandam transmissão de dados com latências inferiores a 1 ms e confiabilidade superior a 99%. No entanto, o planejamento e alocação de recursos para as fatias URLLC apresentam desafios significativos, particularmente na garantia de desempenho de longo prazo sob condições de tráfego e canal variáveis [Choi et al. 2024]. Os algoritmos e técnicas a serem utilizados para otimizar a alocação de recursos constituem uma questão em aberto [ETSI / 3GPP 2022, Banchs et al. 2020, del Prever et al. 2025].

Neste cenário, o Cálculo de Rede [Le Boudec and Thiran 2001] oferece ferramentas analíticas para estimativa da latência em redes de comunicações. O cálculo de rede pode ser dividido em *Deterministic Network Calculus* (DNC) e *Stochastic Network Calculus* (SNC) [Rocha et al. 2026]. Abordagens baseadas em DNC são tradicionalmente utilizadas no planejamento de redes e frequentemente levam à superestimação de requisitos de recursos devido à consideração do pior caso, ignorando os efeitos de multiplexação estatística [Adamuz-Hinojosa et al. 2023]. Em contraste, o SNC estende essas técnicas para um contexto probabilístico, permitindo limites de desempenho não assintóticos, em que o atendimento a níveis de serviço é garantido de maneira ampla, com probabilidade de violação a depender dos requisitos, e consequentemente, do *Service Level Agreement* (SLA) da aplicação [Adamuz-Hinojosa et al. 2023].

Sob uma perspectiva estocástica, a alocação de recursos para uma fatia URLLC passa a depender não apenas da quantidade absoluta de usuários, mas também de como esses usuários utilizam a rede, do perfil estatístico de suas transmissões, e da probabilidade de violação. Isso permite que o orquestrador da *Radio Access Network* (RAN) tome decisões dinâmicas, avaliando continuamente os riscos de violação de *Quality of Service* (QoS) das fatias e redistribuindo recursos de processamento, buffers ou largura de banda de acordo com o comportamento observado, garantindo que todas as fatias permaneçam dentro dos requisitos URLLC esperados.

Contribuições. Neste artigo, propõe-se um modelo baseado em SNC para planejamento de fatias URLLC em redes *Beyond 5G* (B5G), com foco específico no atendimento aos requisitos de latência em função do número de usuários conectados. A principal contribuição é uma equação baseada em SNC que relaciona o número de usuários admitidos com os parâmetros de QoS, considerando as distribuições estatísticas de chegada de pacotes, as condições de canal e a probabilidade de violação do requisito. Ao contrário de outros trabalhos da literatura sobre orquestração e alocação de recursos para diferentes fatias da rede que simplificam o cálculo da latência, este trabalho propõe um arcabouço matemático baseado em SNC para propor um orquestrador eficiente para redes com aplicações baseadas no caso de uso URLLC.

Organização do artigo. O restante do artigo está organizado da seguinte forma: a Seção II apresenta os trabalhos relacionados, a Seção III introduz os fundamentos do Cálculo de Rede Estocástico, a Seção IV descreve o modelo do sistema, a Seção V deriva o modelo SNC para fatias URLLC, a Seção VI discute a alocação de recursos baseada em latência, a Seção VII apresenta os resultados, e a Seção VIII conclui o trabalho.

2. Trabalhos relacionados

Entre as ferramentas matemáticas para análise de desempenho, o Cálculo de Rede Estocástico (SNC) destaca-se por permitir o cálculo de limites estatísticos não as-

Tabela 1. Resumo de trabalhos relacionados sobre DNC, SNC, Slicing, URLLC, Alocação de Banda (AB) e Quantidade de Usuários (QU)

Ref.	DNC	SNC	Slicing	URLLC	AB	QU	Metodologia
[Adamuz-Hinojosa et al. 2023]	–	✓	✓	✓	✓	–	Otimização
[Banchs et al. 2020]	–	–	✓	✓	–	–	Otimização
[Bega et al. 2020]	–	–	✓	✓	✓	–	*Machine Learning*
[del Prever et al. 2025]	–	–	✓	✓	✓	–	Gerenciamento de Conflitos
[Fidler and Rizk 2015]	✓	✓	–	–	–	–	*Survey*
[García-Morales et al. 2019]	–	–	✓	✓	✓	✓	ILP
[Guo and Suárez 2019]	–	–	✓	–	✓	✓	Escalonamento
[Le Boudec and Thiran 2001]	✓	✓	–	–	–	–	Teórico/Livro
[Patra et al. 2017]	–	–	–	–	✓	–	Otimização
[Tang et al. 2019]	–	–	✓	✓	✓	–	Otimização
[Zanzi et al. 2021]	–	–	✓	✓	✓	–	Otimização
[Este Trabalho]	✓	✓	✓	✓	✓	✓	Analítico + Simulação

sintóticos para atrasos, expressos na forma *Probabilidade* [atraso $>$ orçamento] $\leq$ *probabilidade de violação*, mesmo considerando processos estocásticos complexos [García-Morales et al. 2019]. No entanto, essa generalidade tem um custo: em vez de soluções exatas, o SNC fornece estimativas menos rigorosas (aceitando pequenas violações com pequenas probabilidades) para tais limites. Um guia sobre esta ferramenta é apresentado em [Fidler and Rizk 2015]. De forma objetiva, [Le Boudec and Thiran 2001] estabelece a base teórica para sistemas determinísticos, cujos princípios são extensíveis para a análise estocástica. Aplicações específicas do SNC para a fatia URLLC em sistemas B5G são propostas por [Adamuz-Hinojosa et al. 2023], que desenvolvem uma abordagem para planejamento das fatias URLLC utilizando SNC, preenchendo uma lacuna identificada na literatura. A Tabela 1 apresenta os principais trabalhos relacionados.

No contexto de soluções específicas para proporcionar serviços de baixa latência, diversas abordagens foram exploradas na literatura. Trabalhos pioneiros como [Guo et al. 2019], [Patra et al. 2017], [Fidler and Rizk 2015] focam em requisitos de latência e confiabilidade, mas não abordam a tecnologia de *network slicing*, por ser mais recente. Utilizam, por outro lado, tecnologias como a Quantidade de Usuários (QU), Alocação de Banda (AB) e comparação entre DNC e SNC. Ferramentas que proporcionam embasamento nas condições que podem proporcionar as operações com SNC e superioridade dos resultados. Dessa forma, algoritmos dinâmicos de alocação de recursos para fatias da RAN, como os propostos em [Zanzi et al. 2021], [Bega et al. 2020], [Tang et al. 2019], [Guo and Suárez 2019] e [Rocha and Vieira 2019], concentram-se na operação em tempo real, mas sem o compromisso de atendimento a métricas rigorosas por fatia. Uma abordagem mais recente é apresentada por [Choi et al. 2024], em que os autores investigam a alocação de recursos baseada em baixa latência para comunicações via satélite *Low Earth Orbit* (LEO) auxiliadas por *Unmanned Aerial Vehicle*s (UAVs), um cenário de interesse emergente para URLLC.

Contudo, a interação entre o planejamento das fatias URLLC, que garante isolamento e desempenho, e a modelagem matemática do número de *User Equipment* (UEs) admissíveis com base em métricas de atraso que permitem uma probabilidade de violação, permanece como um desafio, particularmente quando se exige a consideração de cenários mais realistas com múltiplas fatias e células concorrendo pelos recursos do meio sem fio.

3. Background

Nesta seção serão apresentados os fundamentos do Cálculo de Rede (*Network Calculus*), com foco específico no SNC.

3.1. Princípio Fundamental do Cálculo de Rede

Para cada nó da rede, o cálculo de rede considera: (a) o processo de chegada de tarefas $A(\tau, t)$; e (b) o processo acumulativo de serviço (processamento) de tarefas $S(\tau, t)$. Estes processos estocásticos são considerados no intervalo de tempo $(\tau, t]$. Esses processos são do tipo envelope (por suas características acumulativas). Usando esses envelopes, parâmetros de desempenho como o limite de *backlog* B e o limite de atraso W de cada nó de rede podem ser derivados.

3.2. Função Afim para o Envelope de Chegada

Uma prática amplamente utilizada no Cálculo de Rede é assumir uma função afim para definir o envelope de chegada $\alpha(\tau, t)$:

$$\alpha(\tau, t) = \rho_A(t - \tau) + b_A, \tag{1}$$

onde $\rho_A > 0$ e $b_A \geq 0$ são os parâmetros de taxa de chegada e rajada, respectivamente. Para o SNC, e utilizando o modelo *Exponentially Bounded Burstiness* (EBB), obtém-se:

$$\mathbb{P}[A(\tau, t) > \alpha(\tau, t)] \leq \varepsilon_A, \tag{2}$$

onde $\varepsilon_A \geq 0$ é o limite de probabilidade de violação do envelope de chegada.

3.3. Função Afim para o Envelope de Serviço

Similarmente, b_S e ρ_S são rajada e taxa de serviço respectivamente, a função afim para o envelope de serviço é definida como:

$$\beta(\tau, t) = \rho_S \left[t - \tau - \frac{b_S}{\rho_S} \right]^+, \tag{3}$$

com a garantia probabilística:

$$\mathbb{P}[\exists \tau \in [0, t] : S(\tau, t) < \beta(\tau, t)] \leq \varepsilon_S. \tag{4}$$

A equação (4) expressa que a probabilidade de o processo de serviço estocástico $S(\tau, t)$ oferecer, em algum instante $\tau \in [0, t]$, uma quantidade de serviço inferior ao envelope determinístico $\beta(\tau, t)$ é limitada superiormente por ε_S. Esse parâmetro representa o limite máximo admissível para a probabilidade de violação do serviço, conforme especificado nos requisitos de QoS. Dessa forma, a curva $\beta(\tau, t)$ pode ser interpretada como um envelope determinístico conservador do serviço estocástico, válido com probabilidade ao menos $1 - \varepsilon_S$.

3.4. Limitantes para o *Backlog* e o Atraso

Uma vez definidas ou modeladas as curvas $\alpha(\tau, t)$ e $\beta(\tau, t)$, o cálculo de rede fornece o arcabouço necessário para se estimar limitantes para o *backlog* e o atraso:

$$B = \frac{\rho_A}{\rho_S} b_S + b_A \;\text{ e }\; W = \frac{b_A + b_S}{\rho_S}. \tag{5}$$

Estas equações formam a base para a análise de desempenho no SNC e serão estendidas no nosso modelo para fatias URLLC.

3.5. Modelos no Cálculo de Rede Estocástico

Os principais modelos utilizados no SNC são: *Moment Generating Function* (MGF), *Exponentially Bounded Burstiness* (EBB), e *Exponentially Bounded Fluctuation* (EBF) [Fidler and Rizk 2015]. Através da MGF, obtém-se uma representação completa dos processos estocásticos; com o EBB, obtém-se um limitante para a variabilidade do tráfego de entrada; e com o EBF, caracteriza-se a flutuação da capacidade de serviço. Essas abordagens permitem derivar garantias probabilísticas de desempenho do tipo $\mathbb{P}[\text{Atraso} > D] \leq \epsilon$, essenciais para aplicações críticas que exigem altos níveis de confiabilidade relacionadas a probabilidades específicas.

O modelo baseado na Função Geradora de Momentos (MGF) constitui uma ferramenta fundamental no Cálculo de Rede Estocástico para caracterizar processos estocásticos de forma probabilística [Fidler and Rizk 2015]. A MGF de uma variável aleatória X, definida como $M_X(\theta) = \mathbb{E}[e^{\theta X}]$ para um parâmetro real θ, captura toda a informação sobre sua distribuição de probabilidade. No contexto de tráfego de rede, ao aplicar a MGF aos processos de chegada $A(\tau, t)$ e serviço $S(\tau, t)$, pode-se derivar limites probabilísticos rigorosos para métricas de desempenho, como atraso e *backlog*. A principal vantagem desta abordagem reside na capacidade de utilizar a desigualdade de Chernoff, que permite transformar a MGF em limites superiores, no seguinte formato:

$$\mathbb{P}[X > x] \leq e^{-\theta x} M_X(\theta), \tag{6}$$

facilitando a análise de violações de QoS de forma analiticamente tratável [Adamuz-Hinojosa et al. 2023].

4. Modelo do Sistema

O modelo do sistema adotado neste artigo é composto por três componentes principais: o modelo de rede, o modelo de recursos de rádio e o modelo de canal. Cada um desses modelos é detalhado nas subseções a seguir.

4.1. Modelo de Rede

Considera-se um ambiente multi-celular B5G, no qual uma *Mobile Network Operator* (MNO) dispõe de uma infraestrutura de rede de acesso composta por um conjunto $\mathcal{I}$ de células. O conjunto de fatias da RAN é denotado por $\mathcal{M}$, sendo que cada fatia $m \in \mathcal{M}$ atende a um subconjunto $\mathcal{U}^m \subseteq \mathcal{U}$ de UEs.

A Figura 1 ilustra o ambiente multi-celular adotado, no qual múltiplos UEs são associados aleatoriamente a um conjunto de células com capacidades distintas de provisionamento de recursos para diferentes fatias de rede, identificadas na Figura 1 como SLICE 1 a SLICE N. Essas fatias diferenciam-se pelos tipos de serviços suportados, o que resulta em diferentes quantidades de recursos alocados aos UEs que utilizam tais serviços.

O objetivo é que, por meio do SNC, um orquestrador de recursos seja capaz de calcular e distribuir os recursos necessários entre as fatias de rede de forma a garantir, com probabilidades especificadas, o atendimento aos requisitos rigorosos de QoS de uma fatia URLLC, dada sua criticidade. Para tanto, define-se neste artigo que os recursos a serem distribuídos entre as fatias são os recursos de rádio no domínio tempo vs frequência, conforme modelado na subseção a seguir.

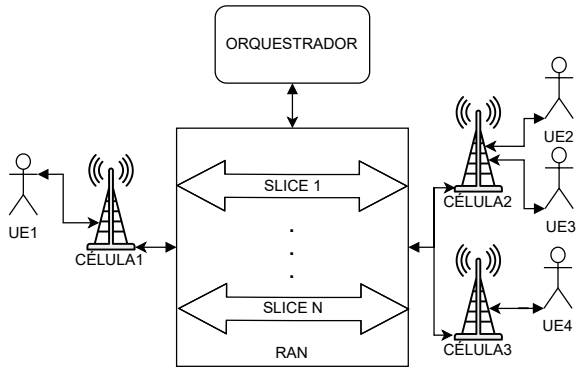


Figura 1. Modelo de sistema multi-celular com fatiamento da RAN

4.2. Modelo de Recursos de Rádio

Assume-se *Orthogonal Frequency Division Multiple Access* (OFDMA) como a tecnologia de acesso múltiplo dos usuários aos recursos de rádio das células. A célula i opera com uma largura de banda total BW_i. Essa largura de banda é dividida em N_i subportadoras OFDM, agrupadas em grupos de $N_{SC} = 12$ subportadoras. Cada grupo define um *Resource Block* (RB), que para a abordagem adotada, é a menor unidade de recursos que pode ser alocada para um UE.

O conjunto de RBs disponíveis na célula i durante cada *slot* de tempo t^{slot} é denotado por $\mathcal{R}_i$, e a quantidade destes RBs é dada por:

$$|\mathcal{R}_i| = \left\lfloor \frac{BW_i}{N_{SC}\Delta f}(1 - OH) \right\rfloor, \tag{7}$$

onde OH denota o fator de *overhead* devido a dados de plano de controle. Δf é a distância entre duas subportadoras.

No B5G NR com operação na faixa dos sub-6 GHz, o *Subcarrier Spacing* (SCS) é o espaçamento entre subportadoras definido pela numerologia $\mu \in \{0, 1, 2\}$ [3GPP 2023a], da seguinte forma:

$$\Delta f(\mu) = \Delta f_0\, 2^{\mu}, \qquad \Delta f_0 = 15 \text{ kHz}. \tag{8}$$

Para uma largura de banda B de 20 MHz e $\mu = 1$, tem-se 51 RBs disponíveis para alocação aos UEs por *slot* de tempo, o que corresponde a um *overhead* de 8% [3GPP 2023b].

4.3. Modelo de Canal

Para caracterizar um canal, a qualidade percebida pelo UE $u \in \mathcal{U}_i^m$, no RB r, é calculada considerando a *Signal-to-Interference-plus-Noise Ratio* (SINR) instantânea $\gamma_{u,r}^{(t)}$. Como a SINR sofre variações rápidas devido ao desvanecimento, mobilidade e interferência, são coletadas medições ao longo do tempo para todos os UEs da fatia m na célula i. A partir desse conjunto de amostras, obtém-se a *Cumulative Distribution Function* (CDF) da SINR, que fornece uma descrição estatística da qualidade do canal experimentada por essa fatia nessa célula.

A SINR é definida em (9), P_{TX} representa a potência de transmissão da gNB; h_r é o ganho de pequena escala associado ao receptor r; χ modela o sombreamento; $d_{u,i}$ é a distância entre o usuário u e a célula i. O α é o expoente de perda no caminho

(*pathloss*). No denominador estão contabilizadas as interferências das outras células; $\xi_{j,r}$ é uma variável de decisão utilizada para determinar se determinado recurso r da célula está ativo no instante t considerado, e P_N é a potência de ruído térmico no receptor.

$$\gamma_{u,r}^{(t)} = \frac{P_{\mathrm{TX}}\, h_r\, \chi\, d_{u,i}^{-\alpha}}{\sum\limits_{j \in \mathcal{I}\setminus\{i\}} \xi_{j,r}^{(t)}\, P_{\mathrm{TX}}\, h_r\, \chi\, d_{u,j}^{-\alpha} + P_N}. \tag{9}$$

A SINR é amostrada ao longo de uma janela temporal suficientemente longa para capturar as variações rápidas e lentas do canal. No caso específico deste trabalho, foi considerada uma janela de 0,5 ms, com o canal sob diferentes condições de interferência.

5. Estimação da Latência para uma Fatia URLLC utilizando SNC

Uma das contribuições deste trabalho consiste no emprego de uma expressão analítica em forma fechada para a caracterização da latência em um ambiente multicelular e multi-*slice*, fundamentada no *framework* de SNC, permitindo não apenas quantificar explicitamente o compromisso entre capacidade do sistema, requisitos de QoS e probabilidade de violação de atraso, mas também de determinar a quantidade de UEs no sistema formado por células heterogêneas, aspecto que não é diretamente capturado por abordagens clássicas baseadas em DNC.

5.1. Modelo de chegada de dados para fatias URLLC

O modelo de tráfego adotado é apresentado em (10) e baseia-se na MGF de (1) com parâmetro livre θ. Esse parâmetro é o que permite definir os momentos da MGF.

$$M_A(\theta) \leq e^{\theta[\rho_A(t-\tau)+\sigma_A]}. \tag{10}$$

Em [Adamuz-Hinojosa et al. 2023], os autores propuseram uma fatia URLLC com tráfego de entrada modelado por processo de Poisson para a quantidade de pacotes por unidade de tempo. Com base nisso, a MGF definida em (10) pode ser estendida para uma célula i e fatia m, conforme apresentado a seguir:

$$M_{A_{i,m}}(\theta) = \exp\left(\lambda_{i,m}\left(\sum_{u=1}^{|\mathcal{U}_{i,m}|}\left[\sum_{l\in\mathcal{L}_m} e^{\theta l}\, p_{m,l}\right]^{p_{i,m,u}-1}\right) t\right), \tag{11}$$

onde $\lambda_{i,m}$ é o parâmetro da distribuição de Poisson que representa a taxa média de chegada na fatia m e célula i; $p_{m,l}$ é a probabilidade de ter um pacote de tamanho l na fatia m e $p_{i,m,u}$ é a probabilidade de alocação de usuário u da célula i na fatia m.

Igualando o lado direito de (10) ao de (11), o que é válido para o caso da igualdade em (10), obtém-se:

$$\alpha_{i,m}(\tau, t) = (\rho_{A_{i,m}} + \delta)[t - \tau] + \sigma_{A_{i,m}}, \tag{12}$$

onde δ representa o argumento de caminho amostral [Fidler and Rizk 2015].

Considerando o sentido de *downlink* [Rocha and Vieira 2019], o processo de chegada $A_{i,m}(\tau, t)$ é a quantidade acumulada de bits que chega do núcleo da rede B5G para a fatia m da *Next Generation Node B* (gNB) i. Baseado na MGF, obtêm-se os parâmetros

do processo envelope de chegada, e assumindo que os dados não são armazenados para depois serem transmitidos ($\sigma_{A_{i,m}} = 0$) [Le Boudec and Thiran 2001], obtém-se:

$$\rho_{A_{i,m}} = \frac{\lambda_{i,m}}{\theta} \left[\sum_{u=1}^{|\mathcal{U}_i^m|} \left[\sum_{l \in \mathcal{L}^m} e^{\theta l} p_{m,l} \right] p_{i,m,u} - 1 \right]. \tag{13}$$

5.2. Modelo de serviço para fatias URLLC

De maneira similar ao cálculo da curva de chegada, a curva de serviço é estimada utilizando a abordagem da MGF proposta em [Adamuz-Hinojosa et al. 2023], onde o modelo de serviço para os dados de uma fatia URLLC m na célula i é dado por (14):

$$M_{S_{i,m}}(-\theta) = \exp\left(\frac{\ln\left(\sum_{q \in \mathcal{Q}} e^{-\theta c_q} p_q \right)}{t_{\text{slot}}} t \right). \tag{14}$$

A MGF para a curva de serviço em (14) é utilizada para determinar a taxa com que os dados serão transmitidos pelo meio sem fio no sentido *downlink* da gNB para os UEs de uma determinada fatia.

Dessa forma, a curva de serviço é definida como:

$$S_{i,m}(\tau, t) = \left[\rho_{S_{i,m}}(t - \tau) - \delta \right]^+ + \sigma_{S_{i,m}}, \tag{15}$$

onde $(\sigma_{A_{i,m}} = 0)$ representa a taxa de serviço e é calculada da seguinte forma:

$$\rho_{S_{i,m}} = \frac{-1}{\theta t^{slot}} \ln\left(\sum_{q \in \mathcal{Q}_i^m} e^{-\theta c_q} p_q \right). \tag{16}$$

5.3. Limitante de atraso para fatias URLLC

Para estimar a latência é necessário determinar a eficiência espectral, que por sua vez depende da condição do canal de comunicação, e a quantidade de RBs para aquela determinada distribuição de recursos. A taxa c_q é determinante para isso, pois é limitada pela equação de Shannon, em (17), que estabelece o *trade-off* entre largura de banda (expressa nesse trabalho por meio da quantidade de RBs, vide (19)) e a SINR (quantificada por meio da eficiência espectral, vide (18) que advém da equação no regime de bloco finito, obedecendo as condições AWGN).

$$C = B \log_2(1 + \gamma). \tag{17}$$

Em (17), C representa a taxa máxima de transmissão suportada pelo canal. O termo B corresponde à largura de banda do sinal transmitido, enquanto $\log_2(1 + \gamma)$ expressa a eficiência espectral sob a razão sinal-ruído γ, caracterizando a forma ergódica da capacidade de Shannon.

Em (18), essa formulação é estendida para considerar o regime de bloco finito, incorporando uma penalidade associada à confiabilidade da transmissão. Especificamente,

o segundo termo, subtraído da expressão clássica de Shannon, captura a degradação da taxa devido às limitações impostas pelo tamanho finito do bloco e pela probabilidade de erro de decodificação.

Dessa forma, diferentemente da de (17), que considera apenas as características ideais do canal, (18) fornece uma modelagem mais realista ao incluir os efeitos do meio e das restrições práticas de transmissão.

$$f^{-1}_{SE\to\gamma}(\gamma) = \log_2(1+\gamma) - \frac{1-\frac{1}{(1+\gamma)^2}}{\sqrt{n_{\text{block}}}} Q^{-1}(\varepsilon_{\text{dec}}) \log_2(e), \tag{18}$$

onde γ é advinda de (9), n_{block} representa o tamanho do bloco, Q^{-1} representa o inverso da função gaussiana Q e ε_{dec} é a probabilidade de erro para uma decodificação de pacote.

$$R^{pkt}_{r,i,m} = \frac{L_r}{t^{\text{slot}} \cdot N_{\text{SC}} \cdot \Delta f \cdot SE_r}, \tag{19}$$

onde L_r em (19) representa o tamanho do pacote para uma fatia m. Com (7), (19) e (18) é possível obter os valores possíveis de c_q e suas probabilidades. Substituindo (13) e (16) em (5), obtém-se (20), equação utilizada para calcular o limitante de atraso (*delay bound*) para uma fatia URLLC m na célula i.

$$W_{i,m} = \frac{2t^{slot}[\ln(\frac{\epsilon'_m}{2}) + \ln(1-e^{-\theta\delta})]}{\ln(\sum_{q\in\mathcal{Q}^m_i} e^{-\theta c_q} p_q) + \delta\theta t^{slot}}, \tag{20}$$

onde t^{slot} é intervalo para o qual é feita a decisão de alocação de recursos, correspondente ao tempo de cada *slot*, θ e δ são parâmetros livres provenientes dos modelos MGF e EBB, ϵ'_m representa a probabilidade de violação do requisito da fatia m, c_q é a taxa média de dados utilizada por usuário e p_q é a probabilidade da taxa c_q ser efetivamente alocada para o usuário, fatia e célula correspondente.

6. Alocação de recursos para fatias URLLC

No cenário analisado neste artigo, fatias com diferentes características competem por recursos de rádio quantificados em RBs, cujas taxas de transmissão variam em função das condições de canal. O objetivo é assegurar que as fatias associadas ao caso de uso URLLC disponham de garantias probabilísticas de atendimento aos requisitos de latência, mesmo sob incerteza. Em [Adamuz-Hinojosa et al. 2023], os autores utilizam o SNC para estimar a latência das fatias URLLC. O presente trabalho avança o estado da arte ao propor uma abordagem que estima a quantidade de UEs admissíveis na fatia URLLC com uma probabilidade específica definida no SLA, considerando um ambiente multi-celular com interferência intercelular, no qual cada célula atende tráfego URLLC e outros tipos de fatias. Essas demais fatias são agrupadas como demandas de *background*. A taxa $\psi_{i,m}$ é interpretada como a taxa de serviço efetiva da fatia m associada à célula i.

$$\psi_{i,m} = \frac{c_q}{t^{slot}}, q \in \mathcal{Q}^m_i. \tag{21}$$

Além disso, a taxa média experimentada por UE na fatia m da célula i é:

$$\zeta_{i,m} = \frac{\psi_{i,m}}{N_{i,m}}. \tag{22}$$

A partir de (20), isola-se o termo com a variável de interesse c_q obtendo:

$$\sum_{q \in \mathcal{Q}_i^m} e^{-\theta c_q} p_q = e^{\frac{2t^{slot}[\ln(\frac{\epsilon'_m}{2}) + \ln(1-e^{-\theta\delta})]}{W_{i,m}} - \delta\theta t^{slot}}. \tag{23}$$

Em (23), observa-se do lado direto c_q assumindo os valores da variável aleatória dos dados de chegada com p_q. Essa abordagem é utilizada para definir a quantidade de usuários para a fatia URLLC.

6.1. Estimativa de Quantidade Máxima de Usuários URLLC em uma Célula

Para estimar a quantidade máxima de usuários na fatia m, utiliza-se Chernoff *Bound* e MGF para os processos de chegada e serviço, conforme detalhado na Seção II, para o conjunto $|\mathcal{I}|$ de células, de onde obtém-se:

$$c_q = \frac{-\frac{2t^{slot}[\ln(\frac{\epsilon'_m}{2}) + \ln(1-e^{-\theta\delta})]}{W_{i,m}} + \delta\theta t^{slot} + \ln(p_q) + \ln(|\mathcal{I}|)}{\theta}. \tag{24}$$

Sabendo que:

$$c_q = N_{i,m} \cdot \zeta_{i,m} \cdot t^{slot}, q \in \mathcal{Q}_i^m. \tag{25}$$

Iguala-se (24) à (25) para obter a quantidade de UEs admissíveis, com garantia probabilística de atendimento ao QoS de latência, por fatia URLLC m da célula i:

$$N_{i,m} = \frac{-\frac{2t^{slot}[\ln(\frac{\epsilon'_m}{2}) + \ln(1-e^{-\theta\delta})]}{W_{i,m}} + \delta\theta t^{slot} + \ln(p_q) + \ln(|\mathcal{I}|)}{\theta\zeta_{i,m}t^{slot}}, q \in \mathcal{Q}_m^i. \tag{26}$$

6.2. Algoritmo de Alocação de Recursos

Uma vez proposto o arcabouço matemático de SNC para estimar latência e a quantidade de UEs por fatia m e célula i, defini-se o Algoritmo 1, disponível no repositório[1], para orquestração de recursos para admissibilidade de UEs, ciente do atendimento às métricas de QoS em ambiente de incerteza, decorrente da heterogeneidade das células, das fatias e das condições do canal.

Define-se a vazão total disponível na rede a partir de componentes associados a uma métrica k de QoS (e.g., requisitos de latência e probabilidade de violação da latência) da fatia especificada, da seguinte forma:

$$V = v_0 + v_1 + \ldots + v_K. \tag{27}$$

A distribuição de vazão entre células e fatias é então modelada por uma regra proporcional baseada em pesos $\omega_{i,m} \geq 0$, que capturam a criticidade da fatia m na célula

[1]https://github.com/pedrojuniormendes/URLLC-RAN-slice-B5G.git

i segundo a métrica de QoS adotada:

$$V_{i,m} = V \cdot \frac{\omega_{i,m}}{\sum_{j\in\mathcal{I}} \sum_{n\in\mathcal{M}} \omega_{j,n}}, \qquad \forall i \in \mathcal{I},\ \forall m \in \mathcal{M}. \tag{28}$$

Observe que (28) garante automaticamente a conservação de vazão, isto é, $\sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}} V_{i,m} = V$.

Para incorporar explicitamente o atendimento probabilístico de latência via SNC, pode-se definir $\omega_{i,m}$ como uma função crescente do "grau de urgência" da fatia:

$$\omega_{i,m} = \alpha_m \, g(\varepsilon_{i,m},\, d_{i,m})\,, \tag{29}$$

em que $d_{i,m}$ é o requisito de latência da fatia m na célula i, $\varepsilon_{i,m}$ é a probabilidade máxima de violação definida no SLA, α_m é um fator de prioridade da fatia, e $g(\cdot)$ é uma função de mapeamento a ser definida conforme a política do algoritmo. Para esse trabalho, assume-se uma distribuição normal padrão para as fatias com diferentes requisitos de QoS. No Algoritmo 1, as variáveis aleatórias, tal como a associada a função de mapeamento é amostrada a cada intervalo ΔT.

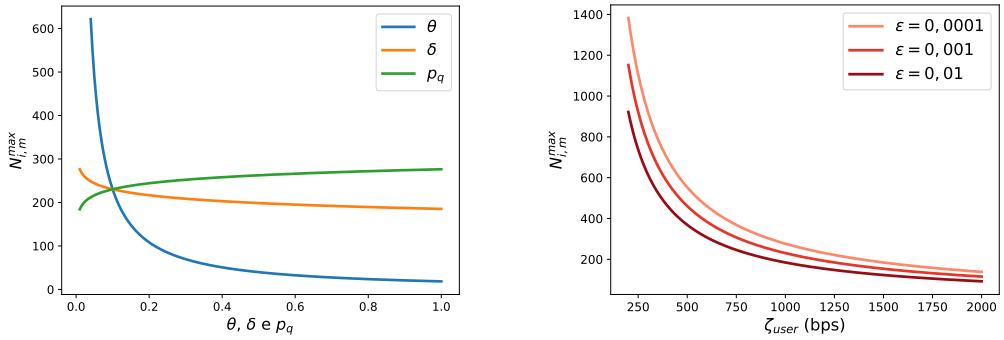
7. Resultados

A Tabela 2 apresenta a parametrização utilizada para a expressão proposta em (26).

Tabela 2. Valores utilizados em (26)

Variável	t^{slot}	ϵ'_m	θ	δ	$W_{i,m}$	p_q	$\lvert\mathcal{I}\rvert$	ζ_{urllc}
Valor utilizado	0,5 (ms)	0,001	0,1	0,1	1 (ms)	0,1	5	1000 (Bytes/s)

Além disso, a Figura 2 apresenta resultados de variação de diferentes parâmetros de (26), utilizando o *framework* de SNC. Na Figura 2(a), é mostrada a variação para os parâmetros δ, θ e p_q. Na Figura 2(b) varia-se ζ_{urllc} e ϵ, onde ζ_{urllc} é o valor médio de $\zeta_{i,m}$ para a fatia $urllc$ de todas as células da rede. A Figura 2(a) mostra que δ e p_q têm pouca influência nos resultados, enquanto a equação é altamente sensível a θ, sendo inversamente proporcional à quantidade de UEs.



(a) Variação de parâmetros de (26)

(b) Variação da probabilidade de violação

Figura 2. Análise de sensibilidade da quantidade de UEs

O parâmetro livre θ do SNC apresenta comportamento não-linear, com região ótima em torno de $\theta = 0.1$ para a configuração analisada. Valores muito baixos ou muito

altos de θ reduzem significativamente o número de usuários admitidos. O parâmetro p_q, representando a pmf de c_q, mostra uma relação logarítmica com $N_{i,m}$. Para $p_q = 0.1$, obteve-se $N_{i,m} \approx 230$, demonstrando a sensibilidade do modelo às distribuições de capacidade. Na Figura 2(a) em verde. A Figura 2(b) mostra o decaimento exponencial do número máximo de usuários admissíveis em relação a taxa de dados por usuário, assim como a relação com a probabilidade de violação do requisito de latência. Os parâmetros mostram a funcionalidade da expressão analítica encontrada e evidencia seu comportamento juntamente com as figuras expostas.

7.1. Impacto dos Requisitos de QoS na Latência

As análises desta seção assumem um cenário de rede com cinco células, com as seguintes distribuições de recursos BW_i: 2×10 MHz, 2×15 MHz e 1×20 MHz na banda n78 do B5G SA, em *Time Division Duplexing* (TDD), com numerologia $\mu = 1$. O tráfego de *background* neste trabalho simula todas as outras fatias da rede que não são URLLC, e a demanda do tráfego de *background* é aleatória entre 50 e 75% da capacidade de cada célula, simulando assim um ambiente de escassez de recursos para o URLLC. O canal segue o modelo *pathloss* de perda de propagação com parâmetros definidos na documentação 3GPP, e cujos detalhes de modelagem podem ser encontrados no repositório público do trabalho, assim como todos os códigos utilizados[2].

Tabela 3. Número de usuários admissíveis para diferentes requisitos de QoS

$W_{i,m}$ (ms)	ϵ'_m	$N_{i,m}$	Ganho Relativo
2,0	0,01	68	-
1,0	0,01	45	33,8%
1,0	0,001	38	15,6%
0,5	0,001	22	42,1%

A Tabela 3 demonstra o *trade-off* entre os requisitos de QoS e o número de usuários admissíveis. Requisitos mais rigorosos (menor latência e menor probabilidade de violação) reduzem significativamente a quantidade de usuários admissíveis.

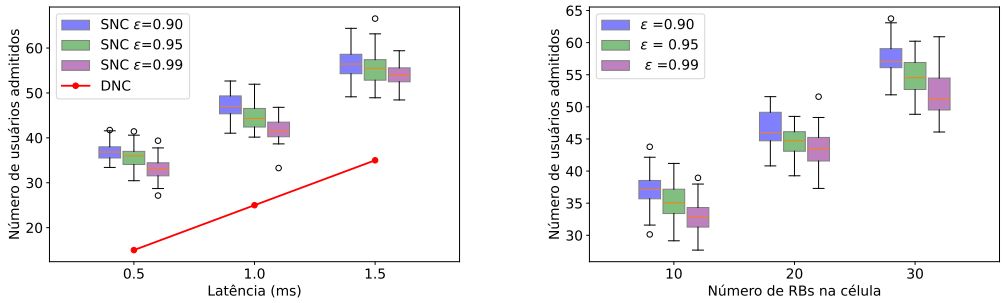
7.2. Comparação com Abordagem Determinística

A Figura 3(a) mostra que a abordagem baseada em SNC suporta significativamente mais usuários ($\approx$ 45 vs 25, para uma latência de 1 ms) que a abordagem baseada em DNC, enquanto mantém a latência dentro dos limites requeridos, demonstrando a superioridade da abordagem estocástica. Na Figura 3(b) é possível observar o número de usuários admitidos em relação ao número de RBs. Essa comparação evidencia o cenário estudado, que é entregar o SNC com garantias próximas ao do DNC.

8. Conclusão e trabalhos futuros

Recentemente, a literatura avançou significativamente na tarefa de estimar a latência para usuários de aplicações críticas. Entretanto, poucos trabalhos apresentam equações analíticas fechadas para o cálculo do número de usuários admissíveis em casos de uso de baixa latência em cenários de múltiplas fatias e múltiplas células, nos quais a heterogeneidade é a regra, e não a exceção. Este trabalho demonstrou que a abordagem proposta baseada

[2]https://github.com/pedrojuniormendes/URLLC-RAN-slice-B5G.git



(a) Usuários admitidos vs Latência do SNC (ϵ = 0,90, 0,95, 0,99) comparado ao DNC

(b) Usuários admitidos vs quantidade de RBs para URLLC (ϵ = 0,90, 0,95, 0,99)

Figura 3. Análise da distribuição da quantidade de usuários

em SNC é superior à DNC, uma vez que, mesmo sob probabilidades extremamente baixas de violação das métricas de QoS, ainda é possível admitir um número significativamente maior de usuários à abordagem DNC.

A presença de tráfego de *background* do tipo *best effort* pode comprometer de forma relevante os recursos disponíveis, e a utilização do algoritmo proposto para oferecer garantias de QoS em cenários de incerteza mostra-se particularmente promissora para o planejamento de serviços URLLC em gerações pós-5G.

Uma limitação deste trabalho reside na não caracterização explícita de outras fatias por meio de equações analíticas específicas, tais como *enhanced Mobile BroadBand* (eMBB) e *massive Machine Type Communications* (mMTC). Como trabalhos futuros, pretende-se aprofundar a análise desses demais casos de uso, investigando quais abordagens e parametrizações se mostram mais adequadas a cada cenário. Ademais, almeja-se validar a proposta por meio de simuladores, como o *NS-3* e o *OMNeT++*.

Reconhecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e pela Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), através do projeto Centro de Excelência em Redes Sem Fio Inteligentes e Serviços Avançados (CERISE).

Referências

3GPP (2023a). NR; Base Station (BS) radio transmission and reception. Technical Specification TS 38.104, 3rd Generation Partnership Project (3GPP). Section 5.3.2: Channel bandwidth.

3GPP (2023b). NR; Physical channels and modulation. Technical Specification TS 38.211, 3rd Generation Partnership Project (3GPP).

Adamuz-Hinojosa, O. et al. (2023). A stochastic network calculus (snc)-based model for planning b5g urllc ran slices. *IEEE Transactions on Wireless Communications*, 22(2):1298–1312.

Banchs, A., De Veciana, G., Sciancalepore, V., and Costa-Perez, X. (2020). Resource allocation for network slicing in mobile networks. *IEEE Access*, 8:214696–214706.

Bega, D., Gramaglia, M., Banchs, A., Sciancalepore, V., and Costa-Pérez, X. (2020). A machine learning approach to 5g infrastructure market optimization. *IEEE Transactions on Mobile Computing*, 19(3):498–512.

Choi, J., Krishnan, S., and Park, J. (2024). Latency-optimal resource allocation for uav-aided leo communication. *IEEE Transactions on Vehicular Technology*, 73(8):12096–12108.

del Prever, P. B., D'Oro, S., Bonati, L., Polese, M., Tsampazi, M., Lehmann, H., and Melodia, T. (2025). Pacifista: Conflict evaluation and management in open ran. *IEEE Transactions on Mobile Computing*.

ETSI / 3GPP (2022). Etsi ts 129 536 v17.2.0 (2022-10): 5g; 5g system; network slice admission control services; stage 3 (3gpp ts 29.536 version 17.2.0 release 17). Technical Specification TS 129 536 V17.2.0, European Telecommunications Standards Institute. Accessed: 2025-01-14.

Fidler, M. and Rizk, A. (2015). A guide to the stochastic network calculus. *IEEE Communications Surveys & Tutorials*, 17(1):92–105.

García-Morales, J., Lucas-Estañ, M. C., and Gozalvez, J. (2019). Latency-sensitive 5g ran slicing for industry 4.0. *IEEE Access*, 7:143139–143159.

Guo, C., Liang, L., and Li, G. Y. (2019). Resource allocation for high-reliability low-latency vehicular communications with packet retransmission. *IEEE Transactions on Vehicular Technology*, 68(7):6219–6230.

Guo, T. and Suárez, A. (2019). Enabling 5g ran slicing with edf slice scheduling. *IEEE Transactions on Vehicular Technology*, 68(3):2865–2877.

Le Boudec, J.-Y. and Thiran, P. (2001). *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Springer.

Patra, M., Thakur, R., and Murthy, C. S. R. (2017). Improving delay and energy efficiency of vehicular networks using mobile femto access points. *IEEE Transactions on Vehicular Technology*, 66(2):1496–1505.

Rocha, F. G., Almeida, G. M., Cardoso, K. V., Both, C. B., and De Rezende, J. F. (2026). Optimal Resource Allocation with Delay Guarantees for Network Slicing in Disaggregated RAN. *IEEE/ACM Transactions on Networking*, 34:4249–4268.

Rocha, F. G. C. and Vieira, F. H. T. (2019). A channel and queue-aware scheduling for the lte downlink based on service curve and buffer overflow probability. *IEEE Wireless Communications Letters*, 8(3):729–732.

Tang, J., Shim, B., and Quek, T. Q. (2019). Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB. *IEEE Journal on Selected Areas in Communications*, 37(4):881–895.

Zanzi, L., Sciancalepore, V., Garcia-Saavedra, A., Schotten, H. D., and Costa-Pérez, X. (2021). Laco: A latency-driven network slicing orchestration in beyond-5g networks. *IEEE Transactions on Wireless Communications*, 20(1):667–682.