Um Mecanismo de Controle em Redes Programáveis Baseado em Telemetria para Aplicações Sensíveis à Latência

Vinícius S. Simão, Rodrigo de B. Lira, Lucas V. Monteiro, Paulo Ditarso Maciel Jr., Ruan D. Gomes, Leandro C. de Almeida

[1] Laboratório Smart4i – Polo de Inovação do IFPB
Instituto Federal da Paraíba (IFPB) – João Pessoa – PB – Brasil

{vinicius.simao,rodrigo.lira,lucas.monteiro}@polodeinovacao.ifpb.edu.br,
{paulo.maciel,ruan.gomes,leandro.almeida}@ifpb.edu.br

Abstract. *Latency-sensitive applications, such as interactive streaming and real-time industrial systems, require fast and stable responses, even under adverse network conditions. However, traditional traffic control methods struggle with variability and congestion, compromising quality of service (QoS). This paper proposes a telemetry-based control mechanism for programmable networks, evaluated through a proof of concept with video-specific scenarios. The results show that the solution ensures greater stability and maintenance of QoS, standing out for its rapid adaptation to traffic variations.*

Resumo. *Aplicações sensíveis à latência, como streaming interativo e sistemas industriais em tempo real, demandam respostas rápidas e estáveis, mesmo sob condições adversas de rede. No entanto, métodos tradicionais de controle de tráfego enfrentam dificuldades diante da variabilidade e do congestionamento, comprometendo a qualidade de serviço (QoS). Este artigo propõe um mecanismo de controle baseado em telemetria para redes programáveis, avaliado através de uma prova de conceito com cenários de transmissões de vídeo. Os resultados mostram que a solução garante maior estabilidade e manutenção da QoS, destacando-se pela rápida adaptação às variações de tráfego.*

1. Introdução

A demanda por aplicações sensíveis à latência tem crescido com a rápida adoção de tecnologias como veículos autônomos, realidade aumentada/virtual (AR/VR), sistemas de controle industrial e *streaming* de vídeo [Avan et al. 2023]. Essas aplicações exigem respostas em tempo real, em que atrasos na transmissão de dados podem resultar em uma degradação significativa de desempenho ou em falhas críticas. Neste sentido, os desafios no gerenciamento de aplicações sensíveis à latência incluem a variabilidade dinâmica das condições da rede, a complexidade na priorização de tráfego e a necessidade de garantir uma qualidade de serviço (QoS, do inglês *Quality of Service*) em ambientes com múltiplos fluxos de dados. Tradicionalmente, técnicas como a otimização de protocolos de rede, o uso de redes de entrega de conteúdo e a implementação de algoritmos de escalonamento têm sido empregadas para mitigar esses problemas [Yang et al. 2023]. No entanto, tais abordagens muitas vezes são insuficientes para lidar com a dinamicidade das redes atuais.

Impulsionadas por tecnologias como redes definidas por software (SDN, do inglês *Software-Defined Networking*) e virtualização de funções de rede (NFV, do inglês

Network Function Virtualization), as redes programáveis possibilitam o controle preciso e dinâmico do tráfego de rede [Ray and Kumar 2021]. Além disso, a programabilidade no plano de dados e a telemetria em tempo real permitem uma medição contínua do estado da rede com um alto nível de granularidade [Arslan and McKeown 2019]. Com estas tecnologias, é possível implementar mecanismos de monitoramento e atuação para tornar a rede mais responsiva às variações das condições de tráfego.

Nesse contexto, um mecanismo de controle que integra telemetria em tempo real a um plano de dados programável é proposto neste artigo. Tal mecanismo coleta métricas dos *switches* da rede de transporte e, com base em políticas predefinidas, faz ajustes dinamicamente, priorizando fluxos críticos, com o objetivo de garantir baixa latência para aplicações sensíveis ao atraso. Como forma de validar o mecanismo proposto, elaboramos uma prova de conceito (PoC, do inglês *Proof-of-Concept*) para ilustrar a aplicabilidade do controle integrado ao plano de dados em cenários de transmissão de vídeo na Indústria 4.0. Uma avaliação de desempenho, realizada com três cenários distintos, demonstrou que a abordagem proposta reduz a latência fim a fim em até 30%, ao passo que mantém as métricas de qualidade do vídeo recebido estáveis.

Adiante neste artigo, a Seção 2 apresenta conceitos importantes para o melhor entendimento da solução proposta. Na Seção 3, os trabalhos relacionados são descritos e comparados. A Seção 4 detalha o mecanismo de controle baseado em telemetria proposto e os cenários da PoC. Os resultados obtidos são apresentados na Seção 5. Por fim, a Seção 6 finaliza o trabalho com as considerações finais e perspectivas de trabalhos futuros.

2. Conceitos Fundamentais

Esta seção apresenta brevemente os conceitos fundamentais necessários para uma melhor compreensão deste artigo, i.e. redes programáveis e telemetria no plano de dados.

2.1. Redes Programáveis

Redes programáveis representam um novo paradigma no gerenciamento de redes, oferecendo maior flexibilidade e controle dinâmico. Diferente das redes tradicionais, cujo comportamento da rede é definido em hardware, as redes programáveis permitem controlar a infraestrutura via software e adaptar o seu funcionamento em tempo real, tornando-a mais responsiva. Essa mudança é possibilitada pela abordagem SDN, que separa o plano de controle (responsável por decisões de roteamento), do plano de dados (que executa o encaminhamento de pacotes). Com isso, é possível centralizar o controle da rede e programar comportamentos dinâmicos, como ajustes de rotas e priorização ágil de tráfego.

Adicionalmente, a programabilidade no plano de dados permite implementar lógicas de encaminhamento diretamente no hardware, sem a necessidade de recorrer ao plano de controle centralizado, sendo ideal para aplicações sensíveis à latência, em que a capacidade de responder rapidamente a mudanças nas condições da rede é essencial. A linguagem P4 (*Programming Protocol-Independent Packet Processors*) se destaca como artefato de programabilidade para o plano de dados [Bosshart et al. 2014].

2.2. Telemetria no Plano de Dados

Avanços em planos de dados programáveis permitiram que dispositivos de rede informem o estado da rede de forma autônoma, sem a necessidade da intervenção direta do plano de

controle [Arslan and McKeown 2019]. Os pacotes possuem instruções de telemetria em seus campos de cabeçalho para a coleta e o registro de dados da rede. Essas instruções de telemetria são definidas na especificação INT (*In-band Networking Telemetry*) [P4 2021].

A principal vantagem desta abordagem consiste na grande quantidade (granularidade fina) de dados de telemetria, uma vez que todos os pacotes que atravessam a rede carregam informações extras de todos os nós. Embora do ponto de vista do monitoramento isto seja positivo, pode haver penalidades de desempenho, que devem ser avaliadas. Isso ocorre devido à existência de uma carga adicional que representa o custo do *overhead* adicionado pelo INT. Além dos modos detalhados na especificação INT, existem abordagens alternativas para a coleta de metadados em redes programáveis. Uma dessas estratégias, denominada ONT (*Out-of-band Network Telemetry*), consiste na utilização de um "fluxo de telemetria exclusivo" para monitorar o estado da rede.

Na abordagem ONT, pacotes de sondagem são usados para medir o desempenho diretamente no plano de dados, sem alterar os pacotes da aplicação. Metadados são inseridos apenas nos pacotes do fluxo exclusivo de telemetria, à medida que atravessam os nós da rede, sendo coletados por um sistema de monitoramento. Com isso, o tráfego da aplicação permanece inalterado de ponta a ponta. A principal vantagem dessa abordagem é justamente permitir que o tráfego das aplicações atravesse a rede sem alterações, evitando problemas de fragmentação. No entanto, o uso de um fluxo de telemetria exclusivo (dedicado) para cada serviço pode gerar sobrecarga adicional na rede.

3. Trabalhos Relacionados

A crescente demanda por aplicações sensíveis à latência motivou diversos estudos recentes na literatura para otimização de redes, concentrados em três principais abordagens: **(i)** mecanismos de agendamento inteligente; **(ii)** técnicas de aprendizado de máquina para roteamento dinâmico; **(iii)** arquiteturas baseadas em redes programáveis. Esta seção analisa trabalhos representativos em cada vertente, destacando suas contribuições e limitações.

O artigo [Liu et al. 2023] aborda a complexidade do agendamento de fluxos em redes com requisitos rigorosos de latência. Os autores identificam que as propriedades dinâmicas das redes podem levar a altas latências e ineficiência na alocação de recursos, problemas críticos para aplicações como redes industriais e telemedicina. Como solução, é proposto um mecanismo de transmissão colaborativa que combina roteamento dinâmico e gestão ativa de filas, utilizando uma estrutura de pacotes inteligente para monitoramento em tempo real. Os resultados demonstram uma redução significativa na latência e um aumento da eficiência no uso dos recursos, comparado a métodos tradicionais.

No trabalho [Chen et al. 2023] é proposto um algoritmo de roteamento baseado em Aprendizado por Reforço Profundo (DRL) para aplicações sensíveis à latência. O método introduz um novo parâmetro chamado Tempo de Sobrevivência (ST), que complementa o TTL (*Time-To-Live*) das redes IP para melhor estimar atrasos e perdas de pacotes. O algoritmo, baseado em *Proximal Policy Optimization* (PPO), mostrou-se superior a protocolos tradicionais em termos de redução de atraso médio e perda de pacotes.

O estudo [Alkubeily et al. 2023] foca na melhoria de QoS em redes de casas inteligentes usando SDN e *OpenFlow*. Os autores propõem um roteamento *multipath* que balanceia dinamicamente o tráfego entre rotas, considerando tanto largura de banda quanto

atraso. Quando o atraso aumenta, o tráfego é redirecionado para caminhos com menor latência, otimizando o desempenho para aplicações como VoIP e *streaming* ao vivo. Os resultados mostram melhorias significativas no QoS em cenários de alta demanda.

A Tabela 1 apresenta uma análise comparativa das características-chave dos trabalhos relacionados com o mecanismo proposto. Mesmo com a comparação limitada em relação ao número de artigos analisados devido à limitação de espaço, destaca-se que a solução descrita no presente artigo avança o estado da arte ao integrar simultaneamente múltiplas filas, priorização de tráfego, telemetria em tempo real e redes programáveis.

Tabela 1. Análise Comparativa dos Trabalhos Relacionados.

Mecanismo de Controle	Múltiplas Filas	Priorização de Tráfego	Telemetria	Redes Programáveis
[Liu et al. 2023]	*x*	*x*	✓	*x*
[Chen et al. 2023]	*x*	✓	✓	*x*
[Alkubeily et al. 2023]	*x*	✓	*x*	✓
Este trabalho	✓	✓	✓	✓

4. Mecanismo de Controle Baseado em Telemetria

Esta seção descreve a implementação do mecanismo de controle baseado em telemetria ONT através de uma PoC e, em seguida, é detalhada a metodologia de experimentação.

4.1. Prova de Conceito

Esta seção apresenta uma prova de conceito voltada à validação do mecanismo de controle em redes programáveis, com foco na implantação de uma infraestrutura 5G privada em ambientes industriais. Na Indústria 4.0, a conectividade desempenha um papel central na integração de tecnologias avançadas e a utilização de uma rede 5G privada apresenta vantagens como alta velocidade e baixa latência na rede de acesso.

A arquitetura do *testbed* ilustrado na Figura 1 combina várias tecnologias: KVM como plataforma de virtualização; free5GC como 5G *Core* (5GC); UERANSIM para simular a rede de acesso via rádio (RAN); *switches* P4 com BMv2 para emular a rede de transporte. O ambiente foi provisionado de maneira automatizada com o Ansible [Lira et al. 2024]. Essa abordagem viabiliza a pesquisa experimental em programabilidade de redes, em um ambiente controlado e flexível. Os experimentos consistem em transmitir vídeo em tempo real, como cenário prático para validar o funcionamento e o desempenho do mecanismo de controle proposto baseado em telemetria. Na figura, percebe-se o tráfego de vídeo fim-a-fim diferenciado do tráfego de fundo em filas distintas. Os metadados capturados na rede de transporte habilitam o *loop* de controle no plano de dados dos *switches* programáveis e o gerenciamento ativo das filas dos dispositivos.

Para a aplicação em questão, foi utilizado um *middleware* de vídeo que implementa um *pipeline* completo de processamento, desde a captura até a entrega dos fluxos às aplicações, por meio de interfaces de vídeo virtuais (*/dev/video**) [Neto et al. 2024]. O *middleware* fornece métricas de desempenho em tempo real, incluindo taxa de quadros por segundo (FPS, do inglês *Frames Per Second*), taxa de bits e latência entre quadros. Optou-se por esta abordagem prática em detrimento de simulações, pois ela nos permite

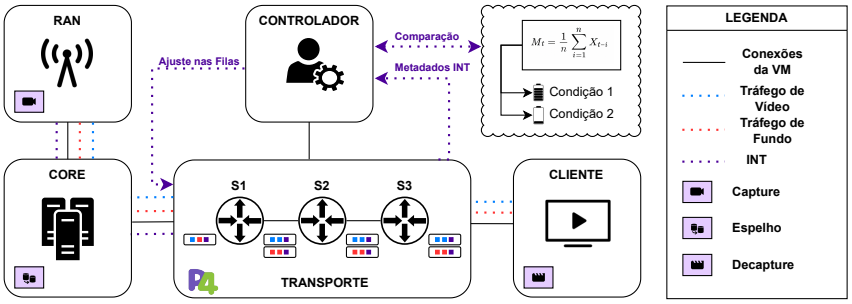


Figura 1. Ambiente de testes.

reproduzir condições reais de operação e obter dados mais representativos do desempenho da rede em cenários dinâmicos. A aplicação foi configurada para transmitir fluxos contínuos de vídeo enquanto registrava as métricas de desempenho.

4.2. Descrição dos Experimentos

Foram realizados experimentos em três cenários distintos: i) sem o mecanismo de controle; ii) com o mecanismo de controle atuando em todos os *switches* ao mesmo tempo; iii) com o mecanismo de controle atuando em cada switch individualmente.

O primeiro cenário, denominado Cenário 1, serviu como base para comparação, onde a rede foi configurada com uma única fila para todo o tráfego, sem qualquer mecanismo de priorização ou controle dinâmico. Ou seja, foi utilizada uma abordagem FIFO (do inglês, *First In, First Out*) de encaminhamento. Neste caso, o tráfego principal da aplicação de vídeo competiu diretamente com um tráfego de fundo gerado com o iPerf3.

O Cenário 2 introduziu o mecanismo de controle centralizado proposto, com a utilização de múltiplas filas com priorização explícita nos *switches* da rede de transporte. O tráfego principal (vídeo) foi alocado em uma fila prioritária, enquanto o tráfego de fundo (iPerf3) foi posicionado em uma fila secundária. O controlador, baseando-se nas medições coletadas pela telemetria ONT, ajustou dinamicamente as taxas de atendimento das filas. Quando o nível de ocupação das filas dos *switches* na rede atingiu um limiar predefinido (50% de ocupação), o controlador aplicou ajustes coordenados em todos os dispositivos, garantindo uma priorização consistente em toda a infraestrutura.

O Cenário 3 também utilizou o contexto de múltiplas filas, porém com uma abordagem diferente de controle. Nesta configuração, o controlador continuou recebendo as medições oriundas da telemetria ONT. No entanto, o mecanismo de atuação foi realizado separadamente em cada *switch* da rede de transporte. Neste caso, quando as medições indicam degradação de desempenho em um *switch* específico (limiar de congestionamento alcançado), o controlador aplica ajustes apenas no dispositivo afetado, sem interferir nos demais componentes da rede. Esta abordagem permitiu avaliar os benefícios de uma resposta localizada em comparação com a estratégia de atuação global do Cenário 2. Além disso, observamos que esta abordagem minimiza o *overhead* de gerenciamento na rede.

A metodologia de experimentação analisou o desempenho da rede sob determinado padrão de carga, usando um fluxo contínuo de vídeo por 10 minutos. Após o primeiro minuto, foram adicionados oito fluxos paralelos de tráfego de fundo com o iPerf3, alternando entre dois minutos ativos e um de pausa. O objetivo foi observar a resposta da

rede em períodos de alta e baixa demanda. Durante os testes, foram coletadas métricas como latência, *jitter*, taxa de bits, FPS e latência entre quadros consecutivos. Todos os códigos desenvolvidos estão disponíveis publicamente no GitHub[1].

5. Resultados

A Figura 2 apresenta a taxa de bits disponível ao longo do tempo para cada cenário avaliado. Trata-se de um gráfico em linhas sobrepostas, onde cada curva representa um dos cenários avaliados. Linhas verticais tracejadas em vermelho indicam os momentos exatos de início e fim dos períodos contendo tráfego de fundo, permitindo correlacionar claramente as variações de desempenho com os eventos de carga na rede.

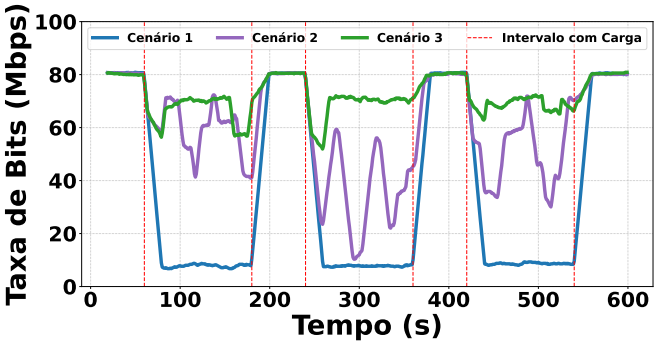


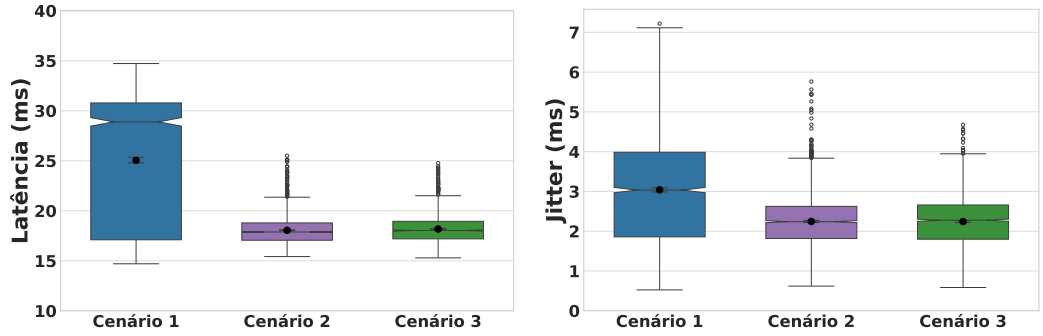
Figura 2. Taxa de bits para o fluxo de vídeo em cada cenário.

O gráfico mostra que o Cenário 3 apresenta um desempenho superior, mantendo uma taxa de bits mais estável e consistente durante os períodos de alta carga, devido ao controle individualizado dos switches, que reduz o *overhead* de gerenciamento. Por outro lado, o Cenário 2, que utiliza controle unificado, apresenta flutuações mais evidentes nesses períodos, refletindo as limitações das ações globais de controle. Ainda assim, ambos os cenários com mecanismos de controle (2 e 3) demonstram desempenho superior ao Cenário 1, que, por não utilizar priorização, sofre quedas abruptas na taxa de bits durante o tráfego de fundo, prejudicando a transmissão de vídeo.

A Figura 3 apresenta os resultados de latência e *jitter*. Os gráficos mostram que o Cenário 1 (sem controle) apresenta alta variação de latência e *jitter*, com medianas acima de 30ms e de até 7ms, respectivamente, indicando desempenho instável para aplicações sensíveis ao atraso. Em contraste, os Cenários 2 e 3 (com controle) exibem uma menor variabilidade nos dados coletados, com latência abaixo de 20ms e *jitter* inferior a 3ms, comprovando a eficácia do controle. Apesar do Cenário 3 (descentralizado) apresentar valores ligeiramente menores que o Cenário 2 (centralizado), a diferença não é estatisticamente significativa. A estabilidade dos cenários controlados destaca a capacidade do mecanismo proposto de manter baixa latência e *jitter*, essencial para aplicações sensíveis ao atraso, como sistemas industriais baseados em transmissão de vídeo em tempo real.

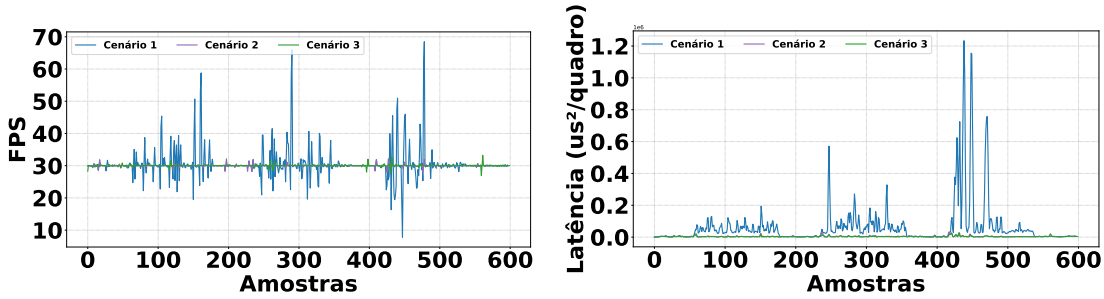
Além das métricas relacionadas à rede, é interessante observar o comportamento da aplicação de vídeo nos cenários avaliados. A Figura 4 mostra os resultados de FPS e latência entre quadros do vídeo, onde observa-se que o Cenário 1 apresenta significativa instabilidade durante períodos de congestionamento. O FPS varia entre 20 e 40,

[1]`https://github.com/viniciussimao/Mecanismo-de-Controle.git`



(a) Latência (b) *Jitter*

Figura 3. Resultados para as métricas de latência e *jitter* em cada cenário.



(a) FPS (b) Latência entre quadros

Figura 4. Resultados de FPS e latência entre quadros em cada cenário.

além de picos extremos abaixo de 10 e acima de 70 FPS. Essa flutuação provoca uma experiência de visualização fragmentada e insatisfatória, indicando baixo QoS. Em contraste, os Cenários 2 e 3 mantêm desempenho estável, com FPS consistentemente próximo de 30 FPS e latência controlada. A análise de latência corrobora esses resultados, mostrando como o Cenário 1 sofre sob condições adversas da rede, com pico acima de 1,2 μs^2/quadro, enquanto os demais cenários preservam excelente estabilidade.

A análise comparativa dos três cenários experimentais mostrou que a solução proposta reduz a latência, melhora a estabilidade do tráfego e garante QoS para aplicações de vídeo, enquanto soluções não otimizadas falham em situações de contenção da rede.

6. Conclusões

Este trabalho demonstrou que mecanismos de controle baseados em telemetria em redes programáveis melhoram o desempenho de aplicações sensíveis à latência. A análise comparativa mostrou que o cenário onde nenhum controle foi aplicado apresentou instabilidade, com grandes variações no FPS e picos de latência durante períodos de contenção da rede, comprometendo a qualidade da aplicação em tempo real. Em contraste, cenários onde o controle adaptativo foi aplicado mantiveram parâmetros estáveis de FPS e baixa latência, mesmo sob contenção na rede. O controle descentralizado dos *switches* apresentou ligeira vantagem sobre o controle global, especialmente na manutenção de uma taxa de bits estável durante tráfego de fundo. As métricas de rede confirmaram esses re-

sultados, mostrando que os cenários controlados reduziram significativamente a latência (<20ms) e o *jitter* (<3ms), comprovando a eficácia da abordagem proposta. Como trabalhos futuros, acredita-se que um mecanismo baseado em Inteligência Artificial pode ser utilizado como ferramenta de auxílio na tomada de decisões no plano de controle.

Agradecimentos

Este trabalho foi financiado pela EMBRAPII (BFA 2301.0001), Cisco, Prysmian, e MPT Cable. Os autores também agradecem ao CNPq (305536/2021-4), CPQD, Inatel, Taggen, Data Machina e o Polo de Inovação do IFPB.

Referências

[Alkubeily et al. 2023] Alkubeily, M. et al. (2023). Reducing Delay for Delay-Sensitive Applications in Smart Home Networks Using Openflow Protocol. In *the 5th Int. Youth Conf. on Radio Electronics, Electrical and Power Engineering*, volume 5, pages 1–5.

[Arslan and McKeown 2019] Arslan, S. and McKeown, N. (2019). Switches Know the Exact Amount of Congestion. In *Proceedings of the 2019 Workshop on Buffer Sizing*, BS '19, New York, NY, USA. Association for Computing Machinery.

[Avan et al. 2023] Avan, A., Azim, A., and Mahmoud, Q. H. (2023). A State-of-the-Art Review of Task Scheduling for Edge Computing: A Delay-Sensitive Application Perspective. *Electronics*, 12(12).

[Bosshart et al. 2014] Bosshart, P. et al. (2014). P4: Programming Protocol-Independent Packet Processors. *SIGCOMM Comput. Commun. Rev.*, 44(3):87–95.

[Chen et al. 2023] Chen, J. et al. (2023). Deep Reinforcement Learning Based Dynamic Routing Optimization for Delay-Sensitive Applications. In *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, pages 5208–5213.

[Lira et al. 2024] Lira, R., Monteiro, L., Simão, V. S., Almeida, L., Gomes, R., Neto, O. A. R., and Maciel Jr., P. D. (2024). Enabling Private 5G Experimentation with Network Programmability and Infrastructure as Code. Demo paper at the IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN).

[Liu et al. 2023] Liu, K. et al. (2023). Deadline-Constrained Multi-Agent Collaborative Transmission for Delay-Sensitive Applications. *IEEE Transactions on Cognitive Communications and Networking*, 9(5):1370–1384.

[Neto et al. 2024] Neto, O. R., Chaves, R., Nascimento, A., and Gomes, R. (2024). Middleware para Aplicações Distribuídas de Vídeo com Suporte à Computação na Borda na Indústria 4.0. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*, pages 215–222, Porto Alegre, RS, Brasil. SBC.

[P4 2021] P4 (2021). In-band Network Telemetry (INT) Dataplane Specification. Technical report, P4 Consortium.

[Ray and Kumar 2021] Ray, P. P. and Kumar, N. (2021). SDN/NFV architectures for edge-cloud oriented IoT: A systematic review. *Computer Communications*, 169:129–153.

[Yang et al. 2023] Yang, H. et al. (2023). A review on software defined content delivery network: a novel combination of CDN and SDN. *IEEE Access*, 11:43822–43843.