AVALIAÇÃO DO USO DE SMARTNICS PARA *OFFLOADING* EM SISTEMAS DE REDES 5G PRIVADAS

RILBERT L. SILVA (PPGEE, IFPB, Campus João Pessoa), RUAN D. GOMES (PPGEE e PPGTI, IFPB, Campus João Pessoa), LEANDRO C. ALMEIDA (PPGTI, IFPB, Campus João Pessoa), MICHEL C. DIAS (UA3, IFPB, Campus João Pessoa)

E-mails: rilbert.lima@academico.ifpb.edu.br, {ruan.gomes, leandro.almeida, michel.dias}@ifpb.edu.br.
Área de conhecimento: 3.04.06.03-0 Sistemas de Telecomunicações.
Palavras-Chave: offloading; middleware; distribuição de video; indústria 4.0.

1 Introdução

As redes de Quinta Geração (5G) são impulsionadas por suportar casos de uso distintos como *Enhanced Mobile Broadband*, *Massive Machine Type Communications* e *Ultra-Reliable Low Latency Communication*, que proporcionam alta taxa de transmissão de dados, comunicação com milhares de dispositivos e latência reduzida, respectivamente. Essa nova arquitetura é caracterizada pela desagregação de funções baseada em serviços, juntamente com a divisão da Rede de Acesso de Rádio (RAN) em Unidade de Rádio (RU), Unidade Distribuída (DU) e Unidade Centralizada (CU) (DAHLMAN; PARKVALL; SKOLD, 2020), oferecendo flexibilidade e inúmeras possibilidades de conexão dos Equipamentos de Usuários (UEs). No entanto, ao aplicar em cenários da Indústria 4.0, essa arquitetura exige uma infraestrutura de processamento robusta, capaz de gerenciar pacotes de dados massivos em tempo real sem comprometer a eficiência de sistemas críticos, como a análise de fluxos de vídeo para controle de qualidade na produção (LIMA, 2022).

Entre as estratégias de otimização de desempenho nos servidores, tais como a memória cache, processamento assíncrono e serviços em nuvem, o *offloading* de processamento de funções para *hardware* especializados surge como uma solução candidata a redução da latência em sistemas de tempo real (XU et al., 2019). Neste sentido, o uso de *Smart Network Interface Cards* (SmartNICs), que são placas *PCI Express* equipadas com processadores próprios, como *System-on-Chip*, ou por dispositivos programáveis, como *Field Programmable Gate Array*, podem executar o *offloading* de processamento, como processos de criptografia, funções de rede virtualizadas e processamento de pacotes, liberando ciclos da CPU principal e mitigando gargalos do sistema (BOURENANE et al., 2024).

Na Indústria 4.0, focada no processamento distribuído de vídeo, mecanismos avançados em SmartNICs, implementados com a linguagem *Programming Protocol-independent Packet Processors* (P4) (P4 Language Consortium, 2022), permitem filtragem e priorização de pacotes de vídeo em tempo real, alocação dinâmica de Qualidade de Serviço (QoS) para atender à demanda dos fluxos de vídeo e mitigação de congestionamento com balanceamento de carga para melhorar a resiliência da rede. A capacidade de descarregar essas funções é crucial para a gestão de QoS (SALVA-GARCIA et al., 2024) e para a segurança da infraestrutura, sem comprometer o desempenho do sistema (BARSELLOTTI et al., 2022).

O P4 se destaca por possibilitar a programação flexível do plano de dados em dispositivos de rede programáveis, permitindo a definição personalizada de *parsers*, tabelas de encaminhamento e *pipelines* de processamento de pacotes. Essa flexibilidade permite obter um controle direto e flexível sobre o encaminhamento de pacotes, ideal para aplicações em computação de borda, tratando os desafios de latência e confiabilidade inerentes aos tráfego de vídeo de alta resolução em redes sem fio, incluindo 5G, e isolamento de fluxos críticos em ambientes industriais (ATUTXA et al., 2021; LIMA, 2022).

Este artigo investiga a otimização da distribuição de vídeo em redes 5G privadas em espectro licenciado e controle granular de QoS para aplicações industriais, por meio do *offloading* de processamento em dispositivos de rede. A hipótese central é que o *offloading* de partes um *middleware* de distribuição de vídeo para dispositivos programáveis, como *switches* BMv2 e SmartNIC, com P4 pode ser capaz de reduzir a latência entre quadros, melhorando assim a QoS da aplicação.

2 Conceitos Fundamentais

A infraestrutura das redes 5G, com sua arquitetura de *Core Network* (CN) virtualizada e a RAN desagregada, tem como possibilidade de suportar os rigorosos requisitos de diversas aplicações da Indústria 4.0, conforme demonstrada na Figura 1 (DAHLMAN; PARKVALL; SKOLD, 2020; PETERSON; SUNAY, 2020).

No CN 5G, destacam-se funções baseadas em serviço, como a Função de Gerenciamento de Acesso e Mobilidade (AMF), que registra dispositivos e gerencia sua mobilidade; a Função de Gerenciamento de Sessões (SMF), que estabelece as conexões; a Função de Plano de Dados (UPF), que distribui o tráfego de usuário de forma distribuída; e o Repositório

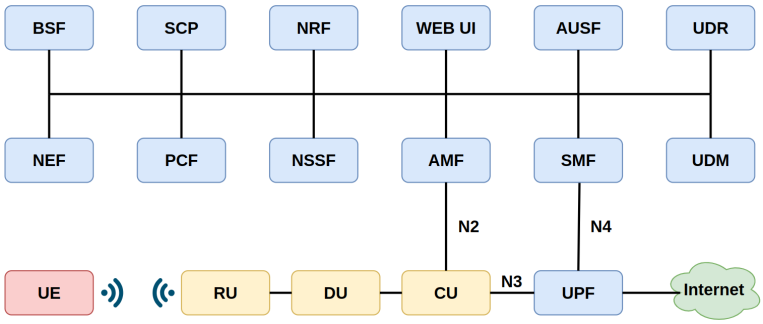

Figura 1: Implementação de uma Rede 5G Privada *open source* baseado em (PETERSON; SUNAY, 2020).

de Funções de Rede (NRF), que centraliza registros e permite a descoberta e coordenação das funções. Essa abordagem modular e escalável assegura a interoperabilidade, flexibilidade e eficiência da rede 5G (PETERSON; SUNAY, 2020).

A arquitetura da RAN 5G é projetada para atender exigências de alta capacidade de dados, baixa latência, alta densidade de dispositivos e uso de espectro privado licenciado. Nessa arquitetura, a CU centraliza funções de controle, como autenticação, alocação de recursos e gerenciamento de mobilidade, enquanto a DU processa sinais em tempo real e interage com a RU de transmissão, realizando funções como codificação e modulação (DAHLMAN; PARKVALL; SKOLD, 2020).

Os sistemas de distribuição de vídeo para visão computacional em tempo real se destacam por permitir o monitoramento preciso das linhas de produção, onde exigem baixa latência e alto desempenho, superando as limitações da computação em nuvem (CZIMMERMANN et al., 2020). A computação na borda aplicada com redes móveis 5G privadas oferece alta capacidade de transmissão, baixa latência e maior segurança, permitindo a análise de vídeo em tempo real, com confiabilidade. Para isso, são utilizados *middlewares*, que são componentes de *software* que realizam a gerência da conexão entre dois terminais e abstraem a complexidade da comunicação de rede. Contudo, mesmo em servidores de borda, o *overhead* de processamento gerado por *middlewares* e pelos múltiplos cabeçalhos de protocolos de rede, consomem ciclos adicionais de CPU, tornando-se um gargalo de desempenho e latência (DA CRUZ et al., 2018; ALIYU et al., 2018).

Para endereçar este desafio, este artigo investiga a variação na latência por quadro decorrente da aplicação ou não de técnicas de *offloading* de processamento de pacotes de um fluxo de vídeo na resolução *Full High Definition* com camada de transporte em UDP e codificação H.264 em um dispositivo programável, como *switches* BMv2, a partir do uso da linguagem P4, na qual se utiliza uma infraestrutura de RAN 5G privada comercial da empresa *SunWave* e utilizando CN *open source* Open5GS para conexão da UE e transmissão dos fluxos de vídeo de uma webcam para o processamento distribuído através de um *middleware* próprio, onde são obtidas métricas de latência por quadro pela aplicação consumidora do vídeo.

3 Resultados e Discussão

Nesta seção iremos apresentar o cenário de avaliação, conforme demonstrado na Figura 2, que consiste em um *pipeline* P4 executando em um *switch* BMv2 simulando as funções de uma SmartNIC, capaz de analisar, processar e encaminhar os pacotes de vídeo, tratando de forma eficiente as camadas de encapsulamento. Ao descarregar a lógica de manipulação dos fluxos de vídeo para o *hardware* da rede, utilizando as abstrações do P4 como tabelas de correspondência-ação e analisadores customizados (P4 Language Consortium, 2022), buscou-se contornar a pilha de rede do sistema operacional, sendo crucial para redução da latência através da redução na quantidade de estágios de processamento dos pacotes.

Com isso, a contribuição reside na aceleração do processamento de um *middleware* de distribuição de vídeo, resultando em uma redução de 31,62% a 53,52% da latência por quadro obtida na aplicação de consumo de vídeo em testes iniciais utilizando *switches* BMv2, além de reduzir a carga de CPU por remover a obrigatoriedade do encaminhamento de pacotes do *middleware*, viabilizando um sistema de visão computacional mais eficiente e determinístico para o ambiente industrial.

4 Considerações Finais

Com base na fundamentação teórica apresentada, conclui-se que a utilização de SmartNICs para execução de *offloading* de processamento por meio da linguagem P4 representa uma abordagem promissora em cenários da Indústria 4.0 para a

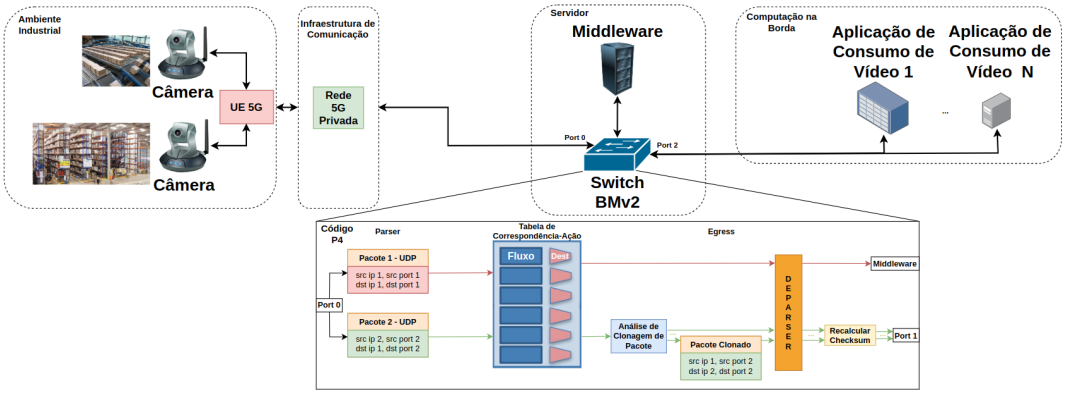

Figura 2: Implementação de *Offloading* de Processamento do *Middleware* de Distribuição de Vídeo com *SmartNIC*.

otimização de desempenho em sistemas de distribuição de vídeo em redes 5G privadas.

A combinação da arquitetura desagregada da rede 5G com a capacidade de programação flexível do plano de dados viabiliza a execução de funções críticas diretamente na SmartNIC, que pode possibilitar a redução significativa da latência e o uso da CPU dos servidores de borda e contribuir para a confiabilidade e a segurança da infraestrutura.

Agradecimentos

Este trabalho é parcialmente apoiado pela EMBRAPII e pelas empresas Cisco, Prysmian e MPT Cable. Os autores também agradecem ao CNPq (305536/2021-4), ao IFPB e ao Polo de Inovação do IFPB.

Referências

ALIYU, A. et al. Towards video streaming in iot environments: Vehicular communication perspective. *Computer Communications*, v. 118, p. 93–119, 2018. ISSN 0140-3664. Disponível em: <https://doi.org/10.1016/j.comcom.2017.10.003>. Acesso em: 10 jun. 2025.

ATUTXA, A. et al. Achieving low latency communications in smart industrial networks with programmable data planes. v. 21, n. 15, 2021. ISSN 1424-8220. Disponível em: <https://doi.org/10.3390/s21155199>. Acesso em: 10 jun. 2025.

BARSELLOTTI, L. et al. Introducing data processing units (dpu) at the edge [invited]. In: *2022 International Conference on Computer Communications and Networks (ICCCN)*. [s.n.], 2022. v. 19 jun. 2024, p. 1–6. Disponível em: <https://doi.org/10.1109/ICCCN54977.2022.9868927>. Acesso em: 05 jun. 2025.

BOURENANE, A. et al. A programmable 5g du-ru smartnic based on mpsoc fpga. In: *2024 IEEE 25th International Conference on High Performance Switching and Routing (HPSR)*. [s.n.], 2024. p. 209–214. Disponível em: <https://doi.org/10.1109/HPSR62440.2024.10635977>. Acesso em: 05 jun. 2025.

CZIMMERMANN, T. et al. Visual-based defect detection and classification approaches for industrial applications—a survey. v. 20, n. 5, 2020. ISSN 1424-8220. Disponível em: <https://doi.org/10.3390/s20051459>. Acesso em: 03 jun. 2025.

DA CRUZ, M. A. et al. Performance evaluation of iot middleware. *Journal of Network and Computer Applications*, v. 109, p. 53–65, 2018. ISSN 1084-8045. Disponível em: <https://doi.org/10.1016/j.jnca.2018.02.013>. Acesso em: 07 jun. 2025.

DAHLMAN, E.; PARKVALL, S.; SKOLD, J. *5G NR The Next Generation Wireless Access Technology*. 2. ed. [S.l.]: Academic Press, 2020.

LIMA, V. N. *A TECNOLOGIA 5G E A INDÚSTRIA 4.0 NO BRASIL - OS DESAFIOS DA INDÚSTRIA NACIONAL*. [S.l.], 2022. Disponível em: <http://bit.ly/40fOYeT>. Acesso em: 20 mai. 2025.

P4 Language Consortium. *P4 16 language specification, version 1.2.3*. 2022. Disponível em: <https://p4.org/p4-spec/docs/P4-16-v-1.2.3.html>. Acesso em: 10 jun. 2025.

PETERSON, L.; SUNAY, O. *5G Mobile Networks: A Systems Approach*. 1. ed. [S.l.]: Morgan Claypool Press, 2020.

SALVA-GARCIA, P. et al. An ebpf-xdp hardware-based network slicing architecture for future 6g front- to back-haul networks. *IEEE Trans. on Netw. and Serv. Manag.*, IEEE Press, v. 21, n. 2, p. 2224–2239, abr. 2024. ISSN 1932-4537. Disponível em: <https://doi.org/10.1109/TNSM.2023.3329942>. Acesso em: 05 jun. 2025.

XU, X. et al. A heuristic offloading method for deep learning edge services in 5g networks. *IEEE Access*, v. 7, p. 67734–67744, 2019. Disponível em: <https://doi.org/10.1109/ACCESS.2019.2918585>. Acesso em: 23 mai. 2025.