Programmable Packet-Optical Networks using Data Processing Units (DPUs) with Embedded GPU

Piero Castoldi[1,*], Rana Abu Bakar[1], Andrea Sgambelluri[1], Juan Jose Vegas Olmos[2], Francesco Paolucci[3], and Filippo Cugini[3]

[1] *Scuola Superiore Sant'Anna, Pisa, Italy*
[2] *NVIDIA, Denmark*
[3] *CNIT, Pisa, Italy*

[*] *piero.castoldi@santannapisa.it*

Abstract: Data Processing Units (DPUs) with embedded GPU have the potential to revolutionize optical networks functionalities at the edge. Use cases are presented for optical data monitoring with local AI processing and embedded security.

1. Introduction

Edge computing enables stakeholders to maintain IT infrastructures, Artificial Intelligence (AI) processing and service management closer to where data are produced and consumed. To achieve this target, edge computing resources need to support advanced networking capabilities with ultra-high bandwidth connectivity. Data Processing Units (DPUs), also called Smart Network Interface Card (SmartNIC) [1–3], originally designed for intra-data center operations, are emerging as attractive solutions to provide edge computing infrastructures with powerful and advanced networking capabilities. DPUs have recently been developed at rates of up to 400Gbps and equipped with embedded graphics processing unit (GPU). Furthermore, they have the potential to directly encompass coherent pluggable transceivers. This way, DPUs represent an innovative technology that cost-effectively combines packet, optical, and edge computing resources in a single networking element, opening the way to new use cases and solutions for programmable optical networking [4].

In this paper, we present two use cases that have the potential to benefit from the presence of DPUs equipped with GPU and coherent pluggable transceivers.

2. Enabling technology: DPU with embedded GPU and coherent pluggable modules

DPU is a specialized hardware component designed to deliver high-speed data processing. While DPUs are typically utilized in data centers, server setups, and supercomputers, there is growing interest in their potential application in edge networking scenarios. This is because DPUs possess the capability to manage data movement, storage, and processing for large datasets, enabling rapid computations and facilitating real-time analysis and acceleration of data-intensive applications. In contrast to traditional NICs, which mainly focus on low-level protocol acceleration, such as Ethernet, and rely on the server's Central Processing Units (CPUs) for other networking tasks, DPUs offer the advantage of programmability at higher layers. They can directly execute advanced in-network functions, thus freeing up processing resources for tenant and application services.

The current generation of DPUs is equipped with up to four interfaces operating at speeds of up to 400 Gb/s, advanced timing and synchronization capabilities, hardware encryption, and embedded security features. Additionally, they feature up to 16 ARM CPUs for handling embedded computing operations. Latest generation also includes embedded GPU resources. DPUs traditionally employ only flat-top connectors (e.g., OSFP, QSFP112/56/28 form-factors) and not yet the QSFP-DD form factor used in packet-optical switches supporting coherent pluggable modules. However, new generation of coherent transceivers fitting these form factors has been already announced for 100Gbps, while future DPU generations are expected to also support QSFP-DD for rates at 400Gbps and beyond. Once this gap is filled, two main advantages can be achieved in the context of optical metro/edge infrastructures.

The first advantage involves reducing the number of active standalone nodes and the need for intermediate Optical-Electrical-Optical (OEO) conversions. Fig. 1(top) illustrates the traditional optical network scenario with aggregation routers and standalone transponders. Fig. 1(center) shows an evolved scenario with white box equipped with coherent transceivers of type point-to-point (p2p) and point-to-multi-point (p2mp), which enables the removal of standalone transponders and aggregation routers respectively. Fig. 1(bottom) shows the innovative optical metro network scenario with edge computing nodes and DPUs equipped with coherent transceivers. This enables the consolidation of optical, packet, and computational resources on a single platform.

The second benefit revolves around latency reduction. This is achieved by collapsing computing and optical network resources on a single platform, thereby minimizing the use of OEO conversions, which, in turn, leads to a reduction in end-to-end latency. Additionally, embracing high-speed transceivers directly integrated into DPUs will

significantly enhance support for ultra-low latency services for access while capitalizing on hardware-accelerated network functions within DPUs (e.g., encryption).

These benefits may lead to performance improvements to both low latency user applications as well as to Telco infrastructural elements. For example, they may facilitate the deployment of 5G functions closer to cell sites. Functions like UPF-DU-CU could be moved closer to the cell site and all co-deployed on powerful edge nodes, relying on (i) hardware-accelerated networking solutions provided by DPUs (e.g., deep packet inspection, cyber security, encryption) and (ii) direct p2p optical connections to cloud services and p2mp connections to multiple RUs. Furthermore, embedded GPU enable effective and fast AI-based predictive/proactive functions [5].

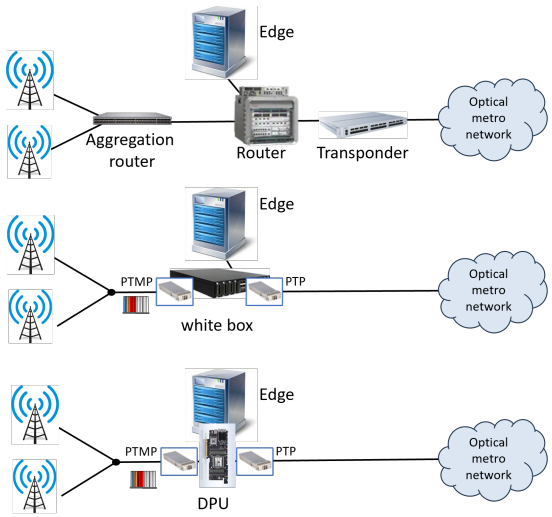

Fig. 1: Traditional optical network scenario with aggregation routers and standalone transponders (top); Evolved scenario with white box and coherent transceivers (center); innovative optical metro network scenario with edge computing nodes and DPUs with coherent transceivers (bottom).

3. Use cases exploiting DPUs with embedded GPU and coherent pluggable modules

3.1. Monitoring

The availability of DPUs equipped with coherent transceivers enables effective monitoring and correlation features directly at the network card. Received packet and optical data can be processed locally also leveraging powerful GPU resources. For example, in-band telemetry reporting per-packet information on experienced latency performance can be processed by embedded P4/DOCA libraries using AI-based algorithms, i.e. outperforming traditional threshold-based mechanisms which just forward data through telemetry towards remote management systems. Furthermore, the DPU has the capability to locally process optical parameters received by the local transceivers. As a proof of concept, we describe the experimental validation of an AI-based algorithm originally designed for soft-failure detection operated by a centralized SDN Controller [6], here deployed for local assessments within the DPU. Fig. 2 shows the considered network testbed. Two DELL PowerEdge Server are equipped with NVIDIA Bluefield2 DPU with embedded GPU. Since these DPUs can not yet support coherent transceivers, we rely on Edgecore switches equipped with 400ZR+ coherent transceivers. However, the control of the transceivers is performed through Rest interface by the DPU (i.e., not by an SDN Controller), as if the transceivers were installed within the DPU. The pair of transceivers is then interconnected by three optically amplified spans of 80-km each. Fig. 3(a) shows the inference time of processing four different sets of data each including N=30 optical parameters extracted from [6], i.e. emulating a single event of received signal. Results show that, depending on the presence of specific anomalies (e.g., soft failures), the inference time varies between 50ms and 70ms. To assess the scalability performance, we forced the system to simultaneously process N=5000 optical parameters. Fig. 3(b) shows that in this case inference time always remains below 360ms. Such fast processing open the ways to radically new decentralized control plane capabilities, where each node may receive telemetry data from multiple data plane nodes without being overwhelmed by scalability issues.

3.2. Cyber security

Traditionally, guaranteeing cyber-security to high-speed connections either overloads CPU performance or requires expensive dedicated hardware components (e.g., standalone firewall). The presence of hardware-accelerated security functions within DPUs provides native embedded security to edge infrastructures. A novel framework for

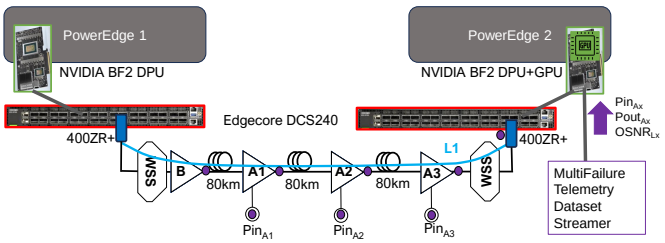

Fig. 2: Network testbed.

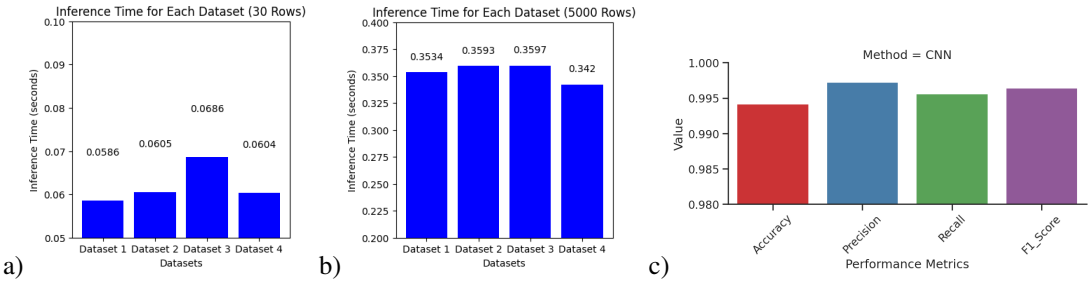

Fig. 3: (a)(b) Monitoring scenario Inference Time. (c) DDoS Attack Evaluation Results Using DPU.

live traffic analysis and intrusion detection on the DPU is here presented. It relies on a combination of DPDK libraries, DOCA library that provides Regular Expression (RegEx) pattern matching to DOCA applications, and the Suricata intrusion detection system. The framework leverages hardware acceleration to perform deep packet inspection (DPI) and deep learning-based intrusion detection. This enables the offloading of Distributed Denial of Service (DDoS) detection tasks to the DPU. Initially, the DPU receives incoming traffic and performs RegEx operations to detect malicious traffic. If the traffic is benign, it is forwarded to the edge CPU for further processing; otherwise, the DPU drops malicious traffic. The framework relies on a deep learning-based Convolutional Neural Network (CNN) model to detect DDoS attacks in real-time. The model was trained and tested using the CICD-DOS2019 dataset. Fig. 3(c) shows the main performance metrics of the AI model. Furthermore, results show that DDoS attacks are detected and blocked in real-time at a rate of up to 95 Gbps of attack traffic, with 99.45% accuracy and only 27.6% CPU utilization. This represents a significant improvement over a traditional non-hardware accelerated DDoS detection method that can process only up to 20 Gbps with 90% CPU utilization.

4. Conclusions

The latest generation of DPUs with embedded GPU, once equipped with coherent pluggable transceivers, has the potential to enable high- performance packet and optical networking within compact and power-efficient edge computing solutions. The expected key benefits include the tight integration of computing and networking resources, the reduction of active standalone nodes and intermediate electro-optical conversions, as well as the reduced latency. Three use cases are then presented to highlight the potential of the DPU technology in the context of metro/edge optical networks.

Acknowledgments. This work has been partially supported by the EU SNS SEASON Project (101096120), the KDT CLEVER Project (101097560), and by the EU under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”). Work is carried out within the Department of Excellence in AI and Robotics 2023-2027.

References

1. Y. Yan, A. F. Beldachi, R. Nejabati, and D. Simeonidou, “P4-enabled smart nic: Enabling sliceable and service-driven optical data centres,” J. Light. Technol. **38**, 2688–2694 (2020).
2. L. Barsellotti, F. Alhamed, J. J. V. Olmos, F. Paolucci, P. Castoldi, and F. Cugini, “Introducing data processing units (DPU) at the edge,” in *International Conference on Computer Communications and Networks (ICCCN),* (2022).
3. F. Cugini, M. Agus, M. Quagliotti, E. Riccardi, C. Castro, B. Spinnler, and A. Napoli, “Point-to-multi-point coherent optics on data processing units (DPUs) for beyond-5G low-latency applications,” in *ICTON,* (2023).
4. B. Niu, J. Kong, S. Tang, Y. Li, and Z. Zhu, “Visualize your IP-over-optical network in realtime: A P4-based flexible multilayer in-band network telemetry (ML-INT) system,” IEEE Access **7**, 82413–82423 (2019).
5. M. E. Morocho Cayamcela and W. Lim, “Artificial intelligence in 5G technology: A survey,” in *2018 International Conference on Information and Communication Technology Convergence (ICTC),* (2018), pp. 860–865.
6. M. F. Silva *et al.*, “Confidentiality-preserving machine learning algorithms for soft-failure detection in optical communication networks,” J. Opt. Commun. Netw. **15**, C212–C222 (2023).