

# Architectural Design of a High-Performance ML-Based Processing Platform for 5G NTN

Maike Taddiken<sup>\*</sup>, Felix Prautzsch<sup>\*</sup>, Tim Düe<sup>†</sup>, Mohammad Amin Vakilifard<sup>†</sup>, Christopher Baumgärtner<sup>‡</sup>, Armin Dekorsy<sup>†</sup>, Jochen Rust<sup>§\*</sup>

<sup>\*</sup>*DSI Aerospace GmbH, Bremen, Germany. maike.taddiken@dsi-as.de*

<sup>†</sup>*ANT, University of Bremen, Bremen, Germany.*

<sup>‡</sup>*NXP Semiconductors Germany GmbH, Hamburg, Germany.*

<sup>§</sup>*Hamburg University of Applied Sciences, Hamburg, Germany.*

**Abstract**—The development of Non-Terrestrial Networks in modern communication standards like 5G includes a functional split of base station functionalities. This introduces the demand for high-performance processing on-board satellites. As a possible solution, this paper presents the architectural design for a high-performance processing platform as a demonstrator for Machine-Learning-based on-board baseband processing for 5G Non-Terrestrial Networks. A Versal AI Edge device is the centerpiece of the processing unit, providing specialized acceleration for ML algorithms. Additionally, a hardware platform to implement the functionality of the ground segment is presented.

## I. INTRODUCTION

The demand for high-performance data processing capabilities on-board satellites has been ever growing since the very beginning of space exploration. Novel algorithms for e.g. earth exploration and high-precision instruments have been a driver for the development in recent years. However, to meet the steadily increasing performance requirements, a comprehensive and costly re-design of subsystems is usually considered. With the introduction of space-terrestrial networks as a key aspect of next generation mobile communication systems, a new application with a need for high on-board computational performance has emerged. This demand is met with changes in the space industry, with a focus towards reducing costs and enhancing availability. The NewSpace market targets the application of commercial-of-the-shelf (COTS) components, suitable for lightweight and cheaper small satellites for short-lived missions in Low Earth Orbits (LEOs). However, the demand for robust and reliable high-performance processing also exists for critical missions and harsher environments, leading to the necessity of a trade-off regarding the selection of critical and non-critical components.

The 5G Open Radio Access Network (O-RAN) architecture comprises a general split of the base station functionality into three network elements: Radio Unit (RU), Distributed Unit (DU), and Central Unit (CU). The actual placement of the functionalities is determined in advance and also considers the possibility to collocate functionalities (see also Section II).

In this work, we present the architectural design for a High-Performance Data Processing Unit specialized for Artificial Intelligence Application (HPDPU-AI). This unit is capable of efficiently processing RU functionality on the satellite. To

achieve this goal, the HPDPU-AI provides efficient computation for the baseband signal processing and additional acceleration of Machine-Learning-(ML-)based algorithms. The DU/CU functionality is implemented on a ground segment with commercial, ML-capable hardware.

## II. 5G NTN: FUNCTIONAL SPLIT AND ML ALGORITHMS

Non-Terrestrial Networks (NTN) are wireless communication networks that expand the traditional terrestrial communication networks by exploiting airborne and space-borne platforms. The latter consist of satellites placed in LEO, Medium Earth Orbit (MEO), or Geostationary Earth Orbit (GEO). Airborne platforms are usually separated into High Altitude Platforms and Low Altitude Platforms, such as Unmanned Aerial Vehicles and drones [1, 2]. User Equipment as defined by the 3rd Generation Partnership Project (3GPP) can connect to the non-terrestrial component either directly or through a NTN Gateway, specifically for satellite communication in higher bands like K and Ka [3, 4].

The Functional Split (FS) concept has been introduced by the 3GPP since the 4G era. In 5G, the concept has been evolved to Radio Access Network (RAN) split and separation between User plane and Control Plane. The FS describes the separation of tasks among the NTN platforms. The 3GPP has so far introduced 8 split options for split-RAN. Technically speaking, these options are based on 3 logical nodes that refer to the RU, DU, and CU. An overview about the split options and according assignments to logical nodes is shown in Figure 1.

Each node is capable of hosting different functions within 5G RAN [5, 4]. Higher-level splits e.g., Option 2 of 3GPP, assign most of the RAN processing units on the non-terrestrial platforms, enabling many advanced tasks like optimized routing, link adaptation, and signal processing to be performed on the non-terrestrial platforms, but also requiring a substantial amount of computation power. Lower-split options on the other hand, e.g. Option 7 of 3GPP, place only functionalities related to the physical layer processing on the non-terrestrial platforms, while the remaining tasks are executed on the ground side. This decreases the hardware requirements in the non-terrestrial platforms, but also disables the possibility

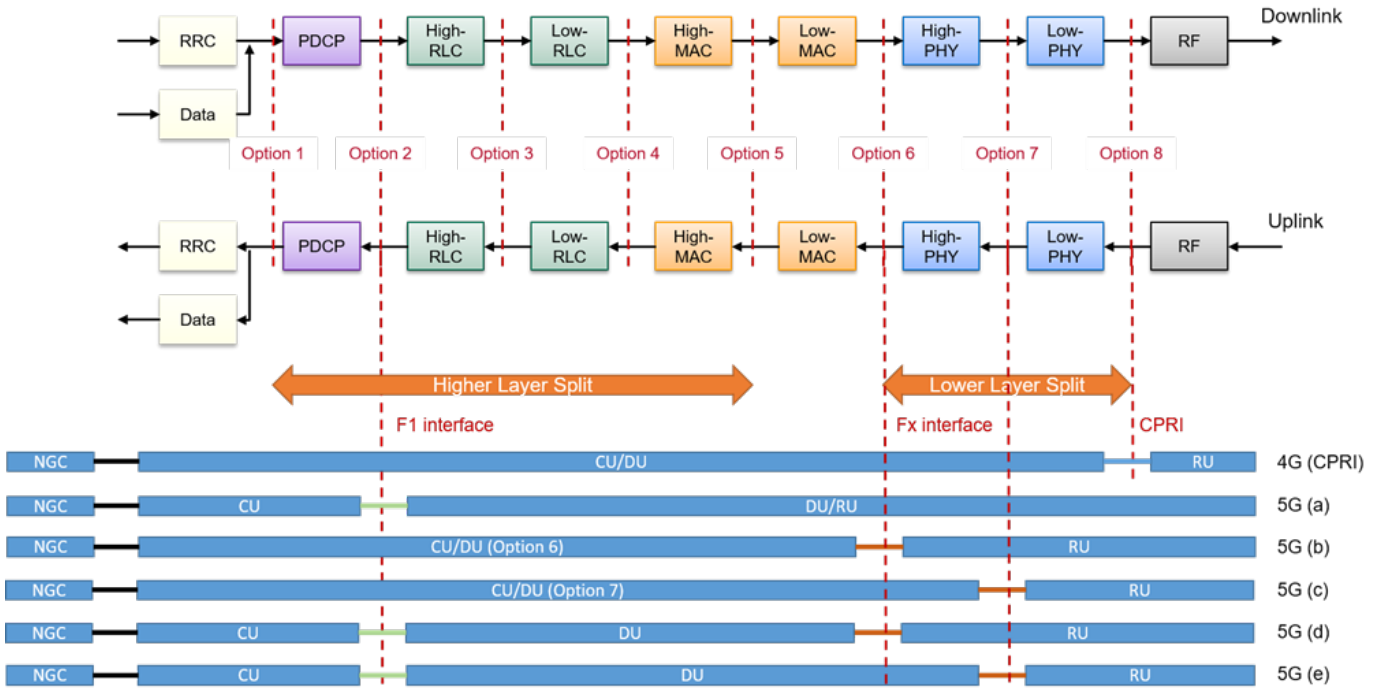


Fig. 1. Mapping of CU/DU/RU functions according to the optional split points.

of many advanced tasks being performed. The choice of FS depends on the desired system complexity, required data throughput, and necessary response time.

The continuously increasing number of tasks being performed through ML is also affecting communications with the inherent integration of ML being one of the fundamental goals of 6G [6]. Possible applications, that have already been implemented in Machine Learning (ML) or are currently researched, include e.g., beamforming, routing, coding, channel estimation, channel impact mitigation, and security [7]. One of the key factors in the success of ML lies within its ability to solve non-linear problems. In [8], the authors have shown that a ML-based channel equalizer can outperform a traditional linear equalizer, even if the linear equalizer is given perfect channel state information. The authors credit this to the linear estimation not being able to describe the channel that is non-linear by nature adequately enough. Results like this showcase the great potential of ML based approaches, especially due to the NTN many non-linear challenges [4].

### III. THE RADIO UNIT AS SPACE SEGMENT

The RU of NTN is considered as a space-borne platform. Due to the harsh environment, the common constraints are imposed on the selection of hardware components.

The HPDPU-AI is responsible for on-board computation of the physical layer (PHY) functionality of the 5G stack and provides the capability to compute ML-based processing of select 5G PHY functionalities. While the architecture of the presented demonstrator is based on COTS components, it has been designed to ensure that space-grade versions or suitable

replacements of the main components are available to facilitate a possible future deployment in orbit.

Driving factors of the HPDPU-AI's design are the very high data rates of up to  $100 \text{ Gbit s}^{-1}$  required by modern communication networks and the subsequent performance requirements to the signal processing capabilities. To meet these challenging demands, the signal processing of several functional blocks is implemented as hardware accelerators exploiting the programmable logic part of the Versal. Moreover, the ML-based processing will be performed by utilizing the Versal's dedicated AI engines [9].

#### A. HPDPU-AI Architectural Design

The Versal AI Edge device is part of AMD's Adaptive Compute Acceleration Platform (ACAP) line-up. These devices are intended to combine performance benefits of different processor categories like scalar or vector processors with adaptable processing capability provided by programmable logic [10]. The selected Versal device includes multiple CPU cores for application and real-time processing, digital signal processors, specialized accelerators for single instruction, multiple data (SIMD) workloads like ML inference called AI engines, and programmable logic for implementation of miscellaneous interface or processing functions. The HPDPU-AI's high-speed data input and output functions are realized as multi-gigabit ethernet interfaces implemented via the Versal's SerDes transceivers. The high-speed capability can be suitably supported by additional optical transceivers to limit the exposure of FPGA pins to electro-static discharge and mitigate potential problems of electromagnetic compatibility caused by the high signalling rates.

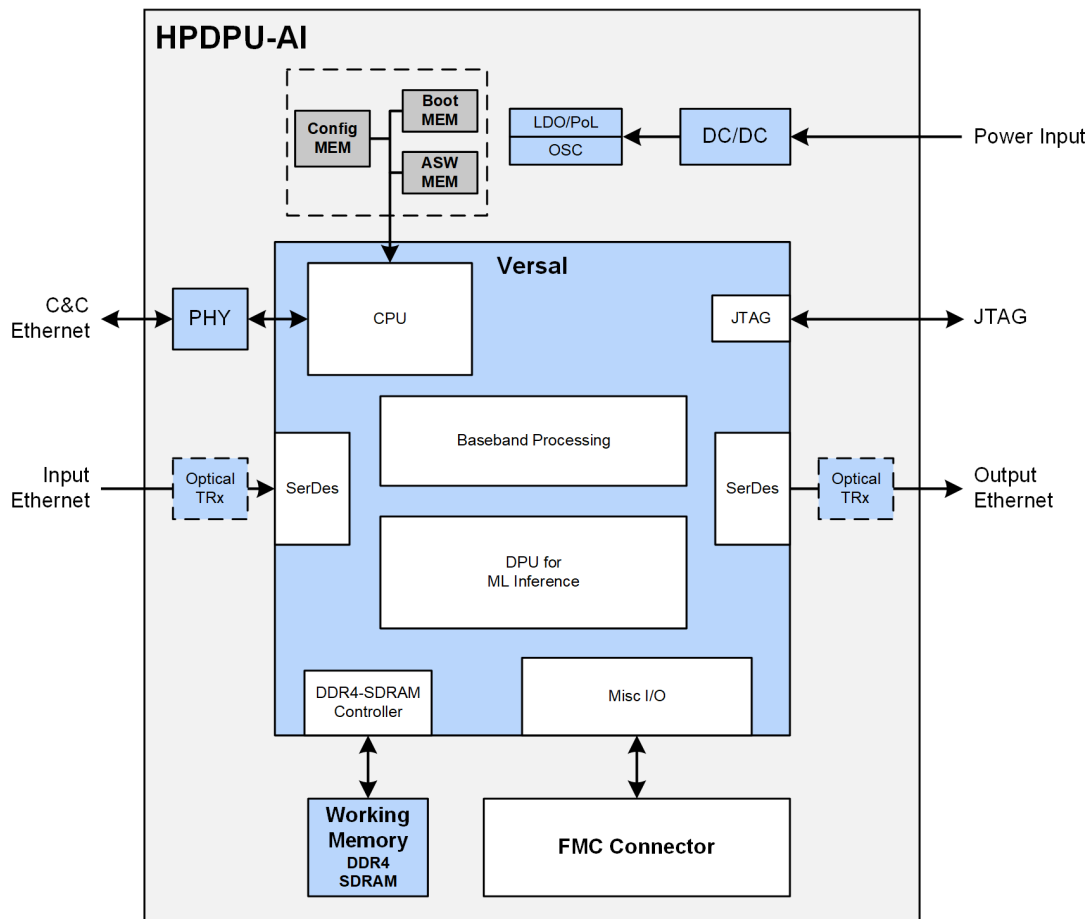


Fig. 2. HPDPU-AI architecture overview.

The HPDPU-AI further provides multiple memory types and devices to support the application. The configuration of the Versal device, i.e. the configuration of the programmable logic, boot software, and application software are stored on non-volatile PROM and Flash devices. The application is also supported by 4 GByte of DDR4-SDRAM working memory.

A dedicated interface for command and control of the unit is also provided. The chosen type on the demonstrator is a 1GBASE-T Ethernet interface, though well-established communication standards, such as CAN or SpaceWire can also be supported. Additional interfaces, or other functionality like ADC/DAC capabilities, can be added to the demonstrator via a provided mezzanine connector for testing purposes. An overview about the HPDPU-AI architecture is shown in Figure 2.

### B. Support of AI Inference Applications

Integration of ML and substituting established conventional algorithms with AI-based solutions is an active point of research in the field of communications technology [11]. As a logical consequence to this trend, the ML-capable Versal device family is being actively investigated for communication applications in space [12]. The HPDPU-AI by means of the

Versal AI Edge device correspondingly offers various means of supporting computation of ML algorithms.

Radiation-tolerant hardware usually significantly impairs the overall performance compared to COTS devices. This shortcoming has been solved traditionally by accelerating time-critical functions in programmable logic wherever possible. This also extends to the computation of ML algorithms by means of specifically designed accelerators [13].

In contrast, the Versal AI Edge device already offers multiple accelerators specifically designed for ML inference. Roughly speaking, AI engines are interconnected vector processors supported by high-speed memory and data interfaces [14]. The accelerators can be programmed manually to obtain a fine-tuned custom solution to efficiently compute a given algorithm; alternatively, for a subset of common ML computation types, the Deep Learning Processing Unit (DPU) provided by AMD can be exploited. This DPU combines multiple accelerators and combines them with additional scheduling and data handling logic implemented in programmable logic [15]. The adaptation of neural networks developed with common frameworks like TensorFlow or PyTorch for computation by the DPU is performed by the Vitis AI tool [16], which performs optimization steps such as quantization and pruning

before compiling the obtained model into a DPU-executable format.

Primarily driven by the high degree of flexibility and fast implementation on the hardware, enabling the evaluation of different options for ML based algorithms as part of the 5G physical layer implementation, the use of the DPU has been selected over the manual implementation.

#### IV. THE DISTRIBUTED UNIT AS GROUND SEGMENT

In addition to the HPDPU specialized for AI applications, the Layerscape architecture by NXP is selected for implementing the DU functionalities as part of the 5G O-RAN architecture. As part of the ground segment, the components of the DU are not subject to the hardware constraints imposed on the RU.

For the DU functionality, the considered choice is the Layerscape LX2160 high-performance Processor. This processor features 16 Arm Cortex-A72 cores optimized for packet processing, along with datapath acceleration for efficient handling of Layer 2/3 packet processing tasks. The LX2160 is well-suited for implementing the DU functionality, offering robust traffic management, security offload, and quality of service features.

In the overall architecture, considering the available hardware options and capabilities for implementation of DU and RU functions, the split option 6 or 7 defined by 3GPP is suggested as the most viable option. In this configuration, the DU functionality would be realized using the Layerscape LX2160 series, while the RU functionality would be implemented on HPDPU-AI. This approach leverages the strengths of each hardware component to achieve an efficient and scalable implementation of the O-RAN architecture.

Utilizing the Layerscape architecture LX2160 presents numerous advantages in implementing the DU functionalities within a 5G O-RAN architecture. The processor offers high-performance computing capabilities with multiple cores and optimized architecture, facilitating efficient packet processing and baseband processing tasks in 5G networks.

The unified hardware and software model of Layerscape further provides scalability, enabling the creation of end-to-end network systems with varying complexities and performance requirements. Extensive connectivity options, including numerous SerDes interfaces, PCIe lanes, and support for high-speed DDR memory, ensure seamless communication between network elements. Additionally, the programmable and configurable nature of Layerscape processors allows for customization to meet specific application requirements, supporting different functional split options and proprietary applications. In particular, the LX2160 processor is optimized for packet processing tasks, making it well-suited for implementing the DU functionality with robust traffic management, security offload, and quality of service features.

Regarding the functional split options, dividing the functionality optimizes resource utilization and scalability. Several functions, demanding more computational resources, can be offloaded to dedicated hardware like the LX2160 processor,

while selected PHY functions can be efficiently handled by the RU hardware. This approach minimizes latency by distributing processing tasks closer to the point of data generation, enhances overall system efficiency and performance, and reduces the complexity of the system.

Considering the hardware options presented, the split option realized on the described hardware options makes use of the capabilities of the Layerscape LX2160 processor for DU functionality and the specialized HPDPU-AI hardware for RU functionality. This configuration maximizes performance, scalability, and efficiency while minimizing latency and overall system complexity.

#### V. CONCLUSION

This paper presents the functional split considered in the 5G O-RAN architecture and the subsequent necessity for space-borne high-performance processing capability to implement NTN. To this end the architectural design of a high-performance data processing unit based on the Versal AI Edge device for on-board baseband processing is shown. This unit includes advanced capabilities to efficiently compute ML algorithms as part of the physical layer functions. Furthermore, a ground segment implementation of the DU based on the NXP LX2160 processor is presented.

#### VI. ACKNOWLEDGEMENT

This work has received funding from the European Space Agency (ESA) research and innovation program in the project "AI for Satellite 5G Communications - AICoS" under grant number 4000139559/22/UK/AL .

#### REFERENCES

- [1] M. M. Azari *et al.*, "Evolution of Non-Terrestrial Networks From 5G to 6G: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 24, pp. 2633–2672, 2022.
- [2] M. Giordani and M. Zorzi, "Non-Terrestrial Networks in the 6G Era: Challenges and Opportunities," *IEEE Network*, vol. 35, pp. 244–251, 2021.
- [3] 3GPP, *TR 38.821: 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Solutions for NR to support non-terrestrial networks (NTN)(Release 16)*, 2023.
- [4] 3GPP, *TR 38.811: Technical Specification Group Radio Access Network; Study on New Radio (NR) to support non-terrestrial networks; (Release 15)*, 2020.
- [5] 3GPP, *TS 38.401: NG-RAN; Architecture description (Release 17)*, 2024.
- [6] S. Periannasamy *et al.*, "Analysis of Artificial Intelligence Enabled Intelligent Sixth Generation (6G) Wireless Communication Networks," in *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, 2022, pp. 1–8.
- [7] B. Ozpoyraz *et al.*, "Deep Learning-Aided 6G Wireless Networks: A Comprehensive Survey of Revolutionary PHY Architectures," *IEEE Open Journal of the Communications Society* 3, 2022.

- [8] M. Honkala, D. Korpi, and J. M. J. Huttunen, "DeepRx: Fully Convolutional Deep Learning Receiver," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3925–3940, 2021.
- [9] *Versal Architecture and Product Data Sheet: Overview*, v1.20. AMD, 2023.
- [10] *Versal: The First Adaptive Compute Acceleration Platform (ACAP)*, v1.1.1. Xilinx, 2020.
- [11] B. Ozpoyraz *et al.*, "Deep Learning-Aided 6G Wireless Networks: A Comprehensive Survey of Revolutionary PHY Architectures," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1749–1809, 2022.
- [12] M. Petry *et al.*, "Machine Learning on Telecommunication Satellite," in *Data Systems in Aerospace (DASIA)*, Eurospace, 2023.
- [13] E. Rapuano *et al.*, "An FPGA-Based Hardware Accelerator for CNNs Inference on Board Satellites: Benchmarking with Myriad 2-Based Solution for the Cloud-Scout Case Study," *Remote Sensing*, vol. 13, no. 8, 2021.
- [14] *AI Engines and Their Applications*, v1.2. AMD Xilinx, 2022.
- [15] *DPUCV2DX8G for Versal Adaptive SoCs*, v1.0. AMD, 2023.
- [16] *Vitis AI User Guide*, v3.5. AMD, 2023.