AI-enabled 5G base-station with Hardware Acceleration for Non-Terrestrial Networks on a Space-Grade System-on-Chip

Michael Petry[1,2] and Raphael Mayr*[1]

[1] *Telecom & Navigation Department, Airbus Defence and Space GmbH, Munich, Germany*
[2] *Professorship for Big Geospatial Data Management, Technical University of Munich, Munich, Germany*

Although 5G Non-Terrestrial Networks envision regenerative satellites with on-board 5G modems to offer the highest degree of user experience, current base-station technology prohibits practical deployment in space. We bridge this gap by presenting a shared platform that unifies gNodeB and Artificial Intelligence (AI) functionality on a hardware-accelerated, space-grade System-on-Chip. By porting OpenAirInterface's gNodeB onto this architecture and offloading DSP-intensive processing tasks of the Physical layer to its Field Programmable Gate Array, benchmarks indicate a reduction in resource requirements by over one order of magnitude, enabling in-space processing and parallel execution of future AI workloads. Moreover, hardware-accelerated support for the Linux Industrial I/O Subsystem interface enables efficient connectivity to space-ready RF frontends. A discussion on further steps towards core network integration concludes this paper.

1 Introduction

The expansion of 5G connectivity through Non-Terrestrial Networks (NTNs) marks a significant advancement in the pursuit of global communication coverage. By leveraging satellite and space-based platforms, NTNs promise to extend high-speed cellular networks to remote and underserved regions, facilitating critical applications in disaster recovery, maritime and aeronautical communications, and the Internet of Things (IoT). Among other use-cases, such as business-to-business (B2B) and machine-to-machine (M2M), NTNs initiated a race towards space-dominance that culminated in the creation of multiple satellite mega constellations [1, 2].

One of the key goals of 5G NTN is to enable unmodified smartphones to directly connect with satellite constellations, referred to as Direct-To-Cell (D2C) access. Primarily led by private companies, different approaches with respect to constellation type, and more importantly, on-board radio frequency (RF) technology are being developed. First successful field trials, such as Lynk's Sub-1-GHz GSM/2G-based test in 2020 [3, 4], or most recently SpaceX's LTE/4G-based demonstration in 2024 to D2C-modified satellites in their Starlink fleet [5], confirmed principle feasibility. Expansion into geostationary earth orbit (GEO) is also being explored [6].

Despite its potential, the deployment of 5G NTNs entails formidable challenges. These can be broadly categorized into wireless propagation-related issues, the integration of NTNs in existing Terrestrial Networks (TN), and computational-related challenges. Additionally, efforts for augmenting the 5G stack with Artificial Intelligence (AI) across various layers, especially the physical, MAC, and application layer, are gaining momentum, as evidenced by the deployment of embedded AI accelerators in state-of-the-art smartphones. This trend is expected to expand to the base-station side [7, 8], necessitating on-board AI execution capabilities. While the first two categories are being addresses in the 5G NTN standard development, practical realization of the gNodeB on space-grade hardware remains a largely unaddressed issue. Unique environmental challenges in space, such as lower power capacities, radiation, extreme temperatures, and distinct processing technologies, necessitate innovative solutions to transition from testbed setups to commercially scalable implementations.

This paper presents a processing solution by realizing a 5G gNodeB on AMD-Xilinx's space-grade Versal AI Core adaptive System-on-Chip (aSoC) series. The main contributions are the following: By utilizing the OpenAirInterface's (OAI) gNodeB implementation [1], the 5G stack is ported to the ARM-based Versal aSoC, while DSP-intensive components are offloaded via hardware acceleration to the Field Programmable Gate Array (FPGA). Furthermore, we extend OAI's radio head with support for the Linux In-

*Corresponding author. E-Mail: raphael.r.mayr@airbus.de

[1] https://openairinterface.org/

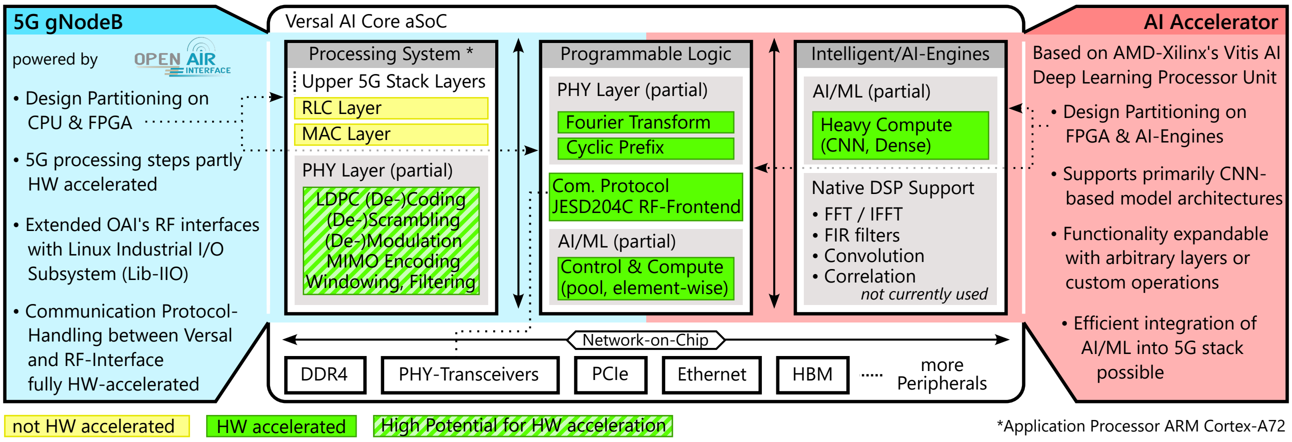

Figure 1: System Architecture for shared 5G gNodeB and AI/ML application on an AMD-Xilinx Versal aSoC.

dustrial Input/Output (IIO) Subsystem. Moreover we address future AI demands by implementing proof-of-concept (PoC) AI functionality on the platform.

The rest of this paper is structured as follows: In Sec. 2 we present the concept of the AI-augmented 5G base-station, outline design decisions w.r.t. processing offloading to the SoC's components, and introduce the PoC AI application. Sec. 3 analyzes the results of hardware acceleration in terms of speed and resource burden for three PHY configurations. Sec. 4 discusses the presented approach, discusses AI integration, elaborates on further optimization, and concludes this work.

2 AI-augmented 5G base-station

2.1 Concept and Target Hardware Platform

Artificial Intelligence is currently being explored on various fronts within cellular networks and is expected to benefit a variety of aspects, such as network and routing optimization, improved user experience and quality of service, energy efficiency, and more. While the integration of AI/ML computational capabilities into terrestrial base-stations is being investigated primarily with GPU-based approaches [9, 10], space-bound systems lack a hardware platform that can satisfy all demands.

With this work we present a shared system design between gNodeB and AI-Accelerator on the novel Versal AI Core System-on-Chip. The Versal facilitates heterogeneous processing with a state-of-the-art processor, hardware-accelerated DSP capabilities on programmable logic (FPGA), and a so-called AI-engine array, efficiently connected by a programmable Network-on-Chip (NoC). The central idea is that the gNodeB and AI Accelerator share this platform. The design split is visualized in Fig. 1, with the blue-shaded and red-shaded components denoting gNodeB and AI Accelerator, respectively. To exploit the heterogeneous architecture effectively, the gNodeB and AI Accelerator implement a design partitioning, distributing control and processing tasks to the most suitable hardware component. The detailed implementation of both components is discussed in the following two sub-sections.

2.2 Hardware-Accelerated 5G gNodeB

The 5G gNodeB's processing architecture leverages a Hardware/Software Codesign approach. Computationally intensive and repetitive tasks are offloaded to the FPGA, while tasks that require a more generic processor, such as control functionality and higher layers of the 5G stack are handled by the Versal's Application Processing Unit (APU). This functional split maximizes platform utilization and enhances performance and power efficiency, addressing a crucial issue in non-terrestrial environments.

2.2.1 Functional Split

Figure 1 illustrates the 5G stack's layers and their corresponding positions within the processing platform. As indicated by the green-shaded items, all major Physical-layer components, such as Coding, Modulation, Filtering, etc., and their inverse operations, are highly suitable for offloading. Currently, Discrete and Inverse Discrete Fourier Transforms (DFT/IDFT) including cyclic prefix handling are offloaded to the FPGA. Furthermore, OAI's data interface to the Radio Frequency (RF) frontend is extended with Lib-IIO capabilities, which utilize an FPGA implementation of the underlying JESD204C protocol driver, fully alleviating the CPU from any communication overhead.

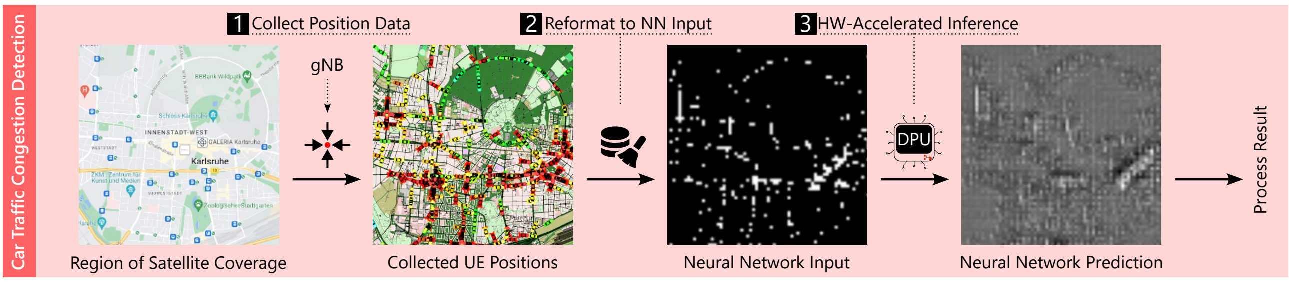
Car Traffic Congestion Detection
1 Collect Position Data
2 Reformat to NN Input
3 HW-Accelerated Inference
gNB
DPU
Region of Satellite Coverage
Collected UE Positions
Neural Network Input
Neural Network Prediction
Process Result

Figure 2: Visualization of the Proof-of-Concept AI car traffic congestion detection application dataflow.

These tasks are significantly more efficient and performant when executed on the PL section. Notably, the building blocks remain configurable via a bus interface from the ARM CPU, ensuring the flexibility of the gNodeB.

2.2.2 Hardware-Software Synchronization

Since the processing blocks on the FPGA have to be controlled differently compared to processing threads in software, a novel control and synchronization mechanism is introduced to the 5G Stack. This mechanism redirects the data flow from being written into Random Access Memory (RAM) to special buffers, which are directly accessible by the processing blocks on the FPGA via DMA. Additionally, a client thread manages the initiation and termination of processing on the PL and ensures synchronization between hardware processing and the calling thread through POSIX Inter-Process Communication (IPC) methods.

2.3 AI Application

As briefly mentioned in Sub-Sec. 2.1, the Versal AI Core series features an AI-Engine array which is a special hardware feature used to accelerate AI/ML inference. On 400 Very Long Instruction Word (VLIW)-Single Instruction Multiple Data (SIMD)-like vector processors, typical ML operations, such as matrix multiplication and convolution, can be performed in a massively data-parallel manner, reaching a performance of up to 100 Int-8 TOPS. The manufacturer's Intellectual Property (IP) core Deep Learning Processor Unit accelerates ML inference on generic CNN- and Dense-Layer-based model architectures by utilizing both AI-engine array and PL, as indicated by the red-shaded region in Fig. 1.

To demonstrate the simultaneous operation of gNodeB and AI Accelerator on the same platform we deploy a Proof-of-Concept AI Application in the Application Layer with the assumption of using metadata from the gNodeB as its input to solve a given task. Here, we decided for a car traffic congestion detection algorithm by using the position of connected (assumed vehicular) users. Fig. 2 visualizes the data-flow starting with position information collected from the gNodeB, reformatting the data, and finally processing it with a neural network (NN).

3 Results

3.1 Partial Physical Layer Offloading

In this work, the reduction in CPU workload is used as primary metric to measure the effectiveness of hardware acceleration. Three Physical layer configurations[2] with varying computational loads as summarized in Tab. 1 were implemented. Table 2 summarizes the utilized hardware resources for CPU, FPGA, and RAM for each configuration in HW-accelerated and non-accelerated mode.

A significant reduction in CPU load is evident for Config-1 and Config-2. On average, the CPU load is reduced by about 25 %. For the largest configuration (Config-3), no significant CPU reduction of is observed, which can be attributed to the saturation of the CPU *both* in the non- and HW-accelerated case. This indicates that further offloading of other processing steps is required to free up CPU resources. The memory consumption remains unchanged independent of offloading, since only the processing domain changes. Overall, the required PL hardware resources are below 5 %.

3.2 AI Accelerator Performance and Resource Utilization

In this sub-section we evaluate the achieved NN inference performance and resource utilization. The U-Net-based NN [11] with a relatively small input resolution

[2] The Physical layer test configurations can also be found at `https://gitlab.eurecom.fr/oai/openairinterface5g` in the leo-5g-ntn branch under gnb.sa.band66.fr1.25PRB.usrpx300.conf, gnb.sa.band78.fr1.24PRB.usrpb210.conf, and gnb.band78.tm1.106PRB.usrpx300.conf, respectively.

PHY Config	Resource Blocks [#]	Sub-Carriers [# Total]	Sub-Carrier Spacing [kHz]	(I)FFT Size [#]	Sample Rate [MSamp/s]	Band [Nr.]	RF Head
Config-1	25	300	15	512	7.68	66	RF-SIM
Config-2	24	288	30	512	15.36	78	RF-SIM
Config-3	106	1.272	30	2048	61.44	78	RF-SIM

Table 1: Key properties of the implemented Physical layer configurations.

PHY Config	HW	FPGA					CPU	RAM
	Acc.	LUT	FF	BRAM	URAM	DSP	[% total]	[%]
Config-1	no	0	0	0	0	0	73 %	9.3 %
	yes	1.7 %	1.6 %	4.7 %	0	3.2 %	53 %	9.3 %
Config-2	no	0	0	0	0	0	75 %	9.6 %
	yes	1.7 %	1.6 %	4.7 %	0	3.2 %	55 %	9.6 %
Config-3	no	0	0	0	0	0	93 %	16.8 %
	yes	1.7 %	1.6 %	4.7 %	0	3.2 %	91 %	16.8 %
Resources Available	N.A.	899.000	1.799.000	34 MBit	130 MBit	1968	100 %	8 GB

Table 2: Summary of utilized hardware resources (CPU, FPGA, RAM) for each PHY configuration in HW-accelerated and non HW-accelerated mode.

DPU Arch.	#Batch	#AI-E	FPGA	Time
C32B1	1	32	9 %	1.36 ms
C32B3	3	96	23 %	0.73 ms
C64B5	5	320	40 %	0.61 ms

Table 3: Resource Utilization and Performance of the AI Accelerator.

of 128x128x1 pixels is evaluated on three DPU configurations, i.e., the smallest (C32B1), medium-sized (C32B3), and largest (C64B5) configuration, realizing a trade-off between performance and required resources, which can be observed in Tab. 3. In summary, an inference time of < 1 ms per batch is achieved, demonstrating the platform's AI capabilities. For a more detailed performance analysis w.r.t. different network architectures and sizes, please refer to [12].

4 Discussion

In this work we successfully implemented OAI's gNodeB and AMD-Xilinx's AI accelerator on a shared, embedded platform. This is a significant achievement, since OAI's official system requirements (OS Ubuntu 22.04., 8 cores x86_64 @ 3.5 GHz, 32 GB RAM) prohibit deployment on a practical space-grade system. As shown in Tab. 2, we reduced the required resources by over an order of magnitude. By using only a fraction of the PL and no AI-Engines for the gNodeB, sufficient resources are available for the AI-accelerator. However, as shown in Sub-Sec. 3.1, our implementation has reached the limit for computationally heavy configurations, necessitating further optimizations:

Further work The primary step is to offload more PHY components, such as Coding, Scrambling, Modulation, Filtering, and more, to alleviate the CPU from most DSP-based processing burden. As a consequence of realizing multiple components on the PL, an efficient CPU-free inter-communication and data buffering strategy is to be implemented. Afterwards, the focus might be shifted on higher layers.

Once the gNodeB itself is sufficiently HW-accelerated, its connections to other base-stations and the 5G core network must be taken care of. While this will definitely require a hardware-accelerated communication protocol implementation, the current Ethernet-based NG- (gNodeB to core network) and XN- (inter-gNodeB) interfaces must be replaced with a space-to-ground (ground station) or inter-satellite (relaying to ground station) link, respectively. This inherently requires a dynamic position-aware routing algorithm to be designed.

References

1. Kulu, E. Satellite Constellations - 2021 Industry Survey and Trends. *35th Annual Small Satellite Conference* (2021).
2. Knopp, A. *et al. HEUMEGA – Independent trend analysis on megaconstellations* tech. rep. (German Aerospace Center (DLR), Cologne, Germany, June 2021). https://www.dlr.de/de/ar/medien/publikationen/broschueren/broschuere-heumega/HeumegaStudie.pdf/@@download/file.
3. Jackson, Donny. *Lynk claims successful test of satellite-to-cell-phone communications, cites potential public-safety value* ,in Urgent Communications, https://urgentcomm.com/2020/03/19/lynk-claims-successful-test-of-satellite-to-cell-phone-communications-cites-potential-public-safety-value/. 2020.
4. Laursen, L. No More "No Service": Cellphones will increasingly text via satellite. *IEEE Spectrum* **60,** 52–55 (2023).
5. SpaceX. *SpaceX sends first text messages via its newly launched Direct-To-Cell satellites* , available at https://api.starlink.com/public-files/DIRECT_TO_CELL_FIRST_TEXT_UPDATE.pdf. 2024.
6. Kumar, S. *et al. 5G-NTN GEO-based over-the-air demonstrator using OpenAirInterface* in *39th International Communications Satellite Systems Conference (ICSSC 2022)* **2022** (2022), 110–114.
7. 3GPP. *"3rd Generation Partnership Project; Technical Specification Group Radio Access Network; New SID on AI/ML for NR Air Interface"* tech. rep. RP-212708 (3GPP, 2021).
8. 3GPP. *"3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Integrated Sensing and Communication (Release 19)"* tech. rep. TR 22.837 (3GPP, 2024).
9. Villa, D. *et al.* An Open, Programmable, Multi-vendor 5G O-RAN Testbed with NVIDIA ARC and OpenAirInterface. *Proc. of the 2nd IEEE Workshop on Next-generation Open and Programmable Radio Access Networks (NG-OPERA)* (2024).
10. DeepSig & Intel. White Paper: Amplifying 5G vRAN Performance with Artificial Intelligence & Deep Learning (2022).
11. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597 [cs.CV].
12. Petry, M. *et al. Accelerated Deep-Learning Inference on the Versal adaptive SoC in the Space Domain* in *2023 European Data Handling Data Processing Conference (EDHPC)* (2023), 1–8.