electronics

MDPI

Article

Scalable AI + DSP Compute Frameworks Using AMD Xilinx RF-SoC ZCU/VCU Platforms for Wireless Testbeds for Scientific, Commercial, Space, and Defense Applications

Buddhipriya Gayanath *, Gayani Rathnasekara, Kasun Karunanayake and Arjuna Madanayake

Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174, USA; grath003@fiu.edu (G.R.); kkaru003@fiu.edu (K.K.); amadanay@fiu.edu (A.M.)
* Correspondence: bgaya003@fiu.edu

Abstract

This paper describes recent engineering designs that allow full-duplex SerDes connectivity between a number of cascaded Xilinx radio frequency system-on-chip (RF-SoC) and VCU FPGA systems. The design allows for unlimited scalability with all-to-all connectivity across FPGA systems and RF-SoCs that allow for bidirectional data transport in streaming mode at a capacity of 50 Gbps per ADC-DAC channel. A custom massively parallel systolic-array architecture supporting 8 parallel data streams from time-interleaved ADC/DACs allow real-time matrix–vector-multiplication (MVM). The MVM can be 8×8, 8×16, ..., 8×1024 in supported matrix size, and is demonstrated in real time sustained throughput of 1 Tera-MAC/second, for matrix size 8×512. The MVM is the building block supporting machine learning and filtering, with the computational graph split across FPGA systems using the SerDes connections. The RF data processed by the FPGA chain can be further utilized for higher-level AI workloads on an NVIDIA DGX Spark platform connected to the system. We demonstrate two platforms in which ZCU111 and ZCU1285 RF-SoC boards perform direct-RF data acquisition, while compute engines operating in real time on VCU128 and VCU129 FPGA boards showcase both digital beamforming and polyphase FIR filterbanking in a real-time bandwidth of 1.0 GHz.

Keywords: Aurora 64B/66B; DGX Spark; matrix–vector multiplier; RF-SoC; radio frequency machine learning; SerDes; wireless testbeds

check for updates

Academic Editor: José V. Frances-Villora

Received: 25 November 2025
Revised: 10 January 2026
Accepted: 19 January 2026
Published: 20 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Emerging artificial intelligence (AI) enabled next-generation wireless testbeds require custom compute architectures mapped to field programmable logic array (FPGA) fabrics. While radio frequency system-on-chip (RF-SoC) platforms, such as the Xilinx ZCU111 and ZCU1285, integrate high-speed data converters capable of multi-GS/s operation, their on-chip programmable logic resources are limited compared to large standalone FPGAs [1,2]. This motivates heterogeneous integration of multiple RF-SoCs with downstream FPGA accelerators that provide significantly greater logic capacity, thereby enabling complex digital signal processing (DSP) pipelines at the radio edge.

1.1. SerDes Technology and Multi-Gigabit Transceivers

Serializer/deserializer (SerDes) technology forms the foundation of high-speed interconnects in modern digital communication systems [3]. As the demand for ever-higher

throughput grows in both electrical and optical domains, advanced SerDes transceivers now deliver multi-gigabit-per-second (Gbps) data rates to support high-performance computing, cloud datacenters, and next-generation wireless communication networks [4–8]. A SerDes architecture converts wide parallel data streams into high-speed serial bit streams for transmission over differential signaling links, such as low-voltage differential signaling (LVDS) or other high-speed physical layers. At the receiver, the data is deserialized into parallel form for subsequent DSP. This conversion reduces interconnect complexity and mitigates high-frequency signal integrity challenges, making SerDes a crucial technology for scalable and reliable high-speed communication links. In FPGA-based systems, multi-gigabit transceivers (MGTs) serve as the SerDes blocks capable of operating at data rates in the multi-gigabit range enabling board-to-board communication, high-speed memory access, and network interface connectivity [3,9–12].

1.2. The Need for Scalable Compute Frameworks for Wireless Communication

Next-generation wireless systems including massive multiple-input multiple-output (MIMO) systems [13–17], terahertz band communication links [18–21], integrated sensing and communications (ISAC) [22–25], and AI-native physical-layer signal processing [26–29] demand unprecedented computational throughput at the radio edge. These systems must process multi-gigahertz bandwidth streams from massive antenna arrays, resulting in high-dimensional spatio-temporal data that must be handled in real time. While both traditional and modern RF-SoC platforms provide significantly enhanced sampling rates (often operating in the multi-GS/s range), they remain constrained by the limited on-chip processing density for advanced signal processing and AI inference workloads [30–32]. Therefore, scalable compute frameworks that partition workloads across heterogeneous devices are essential, with RF-SoCs responsible for high-speed direct-RF sampling, and downstream FPGA accelerators and graphics processing unit (GPU) clusters performing large-scale digital signal processing to satisfy real-time latency and throughput requirements.

1.3. Contribution of the Paper

This paper focuses on engineering design aspects for implementation of heterogeneous multi-FPGA systems, highlighting approaches to FPGA partitioning, high-speed streaming, and SerDes-based interconnects. We demonstrate two fully realized testbeds in which ZCU111 and ZCU1285 RF-SoC platforms (from AMD Xilinx, San Jose, CA, USA) are interconnected with Xilinx VCU128 and VCU129 FPGA accelerator boards (from AMD Xilinx, San Jose, CA, USA) over multiple 25 Gbps and 20 Gbps links, achieving maximum sustained real-time transfer rates of up to 50 Gbps per ADC–DAC channel. On the downstream VCU128/VCU129 platforms, FPGA-based MVM hardware cores are implemented using 8-phase polyphase decomposition, and the processed data are streamed back to the RF-SoCs for high-speed digital-to-analog converter (DAC) output. Additionally, we demonstrate system-level integration with an NVIDIA DGX Spark platform (from NVIDIA Corporation, Santa Clara, CA, USA), enabling higher-level AI/ML processing on the RF data as part of a scalable, AI-enabled wireless compute framework. The results showcase the potential of SerDes-enabled multi-FPGA frameworks as flexible, real-time DSP platforms for next-generation RF and AI-RF applications.

1.4. Organization of the Paper

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of related work on multi-FPGA platforms, including large-scale cloud and datacenter-scale deployments, wireless communication and RF testbeds, and distributed AI accelerators. Section 3 introduces the proposed scalable AI + DSP compute framework, discussing the theoretical background of the MVM core and the integration of the NVIDIA

DGX Spark platform into the testbed. In Section 4, we describe the digital design aspects of the realized prototype testbed in detail, covering the soft IPs utilized, the RF-SoC and FPGA platforms employed, the associated hardware and IP configurations, and the overall system-level dataflow architecture. Section 5 presents the experimental setups and results for both configurations, including transmission coefficient measurements that demonstrate the practical viability of the testbed. Section 6 provides a detailed discussion of the implementation challenges encountered during system realization, along with the practical workarounds adopted to address them. Finally, Section 7 discusses the implications of the proposed architecture and outlines potential directions for future research.

2. Review

Recent advances in multi-FPGA platforms, particularly those interconnected through high-speed multi-gigabit transceivers, have enabled scalable and high-performance processing architectures that are increasingly essential for next-generation RF signal processing, wireless communication, and AI-accelerated systems. Prior work in this area can be broadly categorized into (i) cloud- and datacenter-scale FPGA deployments, (ii) multi-FPGA wireless communication and RF testbeds, and (iii) distributed FPGA accelerators for AI and signal processing. This section reviews representative efforts in each category and highlights the limitations that motivate the proposed scalable AI + DSP compute framework.

2.1. Cloud and Datacenter-Scale FPGA Platforms

Several notable efforts have demonstrated the deployment of reconfigurable logic at cloud and datacenter-scale. Microsoft's Catapult project [33,34] integrates FPGAs directly into datacenter servers to accelerate large-scale workloads such as web search, machine learning inference, and network processing. The Catapult fabric interconnects Altera Stratix V D5 FPGAs in a 6×8 two-dimensional torus topology using custom lightweight serial links providing up to 20 Gbps bidirectional bandwidth over FMC and PCIe interfaces. Similarly, Amazon EC2 F2 instances [35] deploy up to eight AMD Virtex UltraScale+ HBM VU47P FPGAs per node, interconnected via PCIe Gen3, enabling cloud based development and deployment of reconfigurable hardware accelerators. Furthermore, earlier large-scale efforts such as the Reconfigurable Computing Cluster (RCC) project [36] explored petascale computing using network of Xilinx Virtex FPGAs interconnected through custom RocketIO MGT links operating at approximately 2.5 Gbps and employing Aurora 8B/10B protocol cores. While these platforms demonstrate impressive scalability, they primarily target cloud workloads and lack direct integration with high-speed RF data converters or real-time wireless signal processing pipelines.

2.2. Multi-FPGA Wireless Communication and RF Testbeds

In the context of wireless communication testbeds, several multi-FPGA platforms have been proposed to address scalability and real-time processing challenges. The FPGA-based digital wireless channel emulator [37] distributes signal processing workloads across multiple interconnected Xilinx Virtex FPGA nodes to emulate large-scale wireless networks with up to 1250 devices at 1.9 MHz per-channel bandwidth. Inter-FPGA communication is achieved using GTX transceivers with 8B/10B encoding, providing 4 Gbps per lane. In a related effort, the authors in [38] present a scalable clock-synchronization and data-communication framework for quantum-domain experiments, enabling deterministic clock distribution and fiber-based data transfer across multiple RF-SoC and FPGA boards. Their architecture relies on the Aurora 64B/66B protocol for high-speed optical communication, demonstrated using ZCU216 RF-SoC platforms. Similarly, Ref. [39] introduces a modular multi-FPGA wireless MIMO-OFDM testbed based on CompactPCI hardware and

interconnected Xilinx XC2V6000 FPGAs, enabling evaluation of spatial multiplexing and diversity techniques. While these platforms demonstrate distributed RF and baseband processing, they typically operate at relatively narrow bandwidths and do not explicitly address heterogeneous RF-SoC–FPGA partitioning for multi-gigahertz, AI-enabled wireless workloads.

2.3. Distributed FPGA Accelerators for AI and Signal Processing

Multi-FPGA acceleration has also been explored in the context of AI workloads. The authors in [40] propose multi-chip convolutional neural network (CNN) accelerators that partition computation across several FPGA devices interconnected via the Aurora protocol, demonstrated on Xilinx ZCU102 and VU19P platforms. These works highlight the potential of FPGA clusters for scalable AI inference and training. However, they are not integrated with direct-RF data acquisition and real-time wireless signal processing.

In contrast to prior work, the proposed compute framework tightly integrates direct-RF RF-SoC platforms with high-capacity downstream FPGA accelerators using multi-lane 20–25 Gbps SerDes links, enabling sustained real-time data rates of up to 50 Gbps per ADC-DAC channel. Unlike cloud-centric FPGA clusters or narrowband wireless testbeds, this work focuses on multi-gigahertz bandwidth RF data streaming, hardware-partitioned DSP pipelines, and closed-loop RF processing, while also supporting system-level integration with GPU-based AI platforms such as NVIDIA DGX Spark, for higher-level learning and inference. Therefore, the proposed testbed bridges a critical gap between scalable multi-FPGA compute architectures and AI-native, real-time wireless experimentation.

3. Proposed Testbed Architecture

In this section, we present an overview of the proposed multi-FPGA DSP framework, including its distributed computation architecture, the implementation of the MVM core used for real-time digital signal processing, and integration of NVIDIA DGX Spark platform, which will be leveraged in future work to execute higher-level AI algorithms on the processed RF data.

3.1. Overview of the Proposed Scalable AI + DSP Compute Framework

We envisage a scalable compute framework that leverages the heterogeneous integration of RF-SoCs and high-performance multi-FPGA platforms to enable AI-driven wireless testbeds. The proposed framework is illustrated in Figure 1. In this architecture, the RF-SoCs perform direct RF data acquisition and transmission through their integrated data converters, while the compute-intensive DSP and AI workloads are transferred to downstream FPGAs with significantly greater logic, memory, and DSP resources. The high-speed direct RF sampling rates require the support for both massive parallelism and multi-rate signal processing using polyphase digital systolic arrays [41]. The illustration depicts a subset of DSP and layers of a large multi-layer neural network implemented across multiple FPGA platforms in order to process the incoming polyphase ADC samples in training or inference. Due to the compute limitations in the RF-SoC, only a subset of the neural network layers is implemented, and the remaining layers are executed on the high-performance VCU128/VCU129 boards. Therefore, the output data of the RF-SoC implemented layers is transported through multiple 25 Gbps MGT links, which are realized across the selected FPGA platforms.

A single-layer implementation corresponds to multiple 8×8 MVM implementations. Each MVM core represents an 8 point convolution operation using 8 polyphase decomposition. Here, $f(x)$ is the activation function, which is ReLU/Sigmoid, etc. Each activation outputs are fully connected with the nodes in the subsequent layer. Furthermore,

during training stages, feedback information can be transferred to the preceding layers implemented on the upstream FPGA platforms through the same MGT links. Finally, the computed results can be routed back to the RF-SoC via the duplex MGT channels and output through the DAC, thereby enabling a closed-loop, real-time wireless testbed capability.

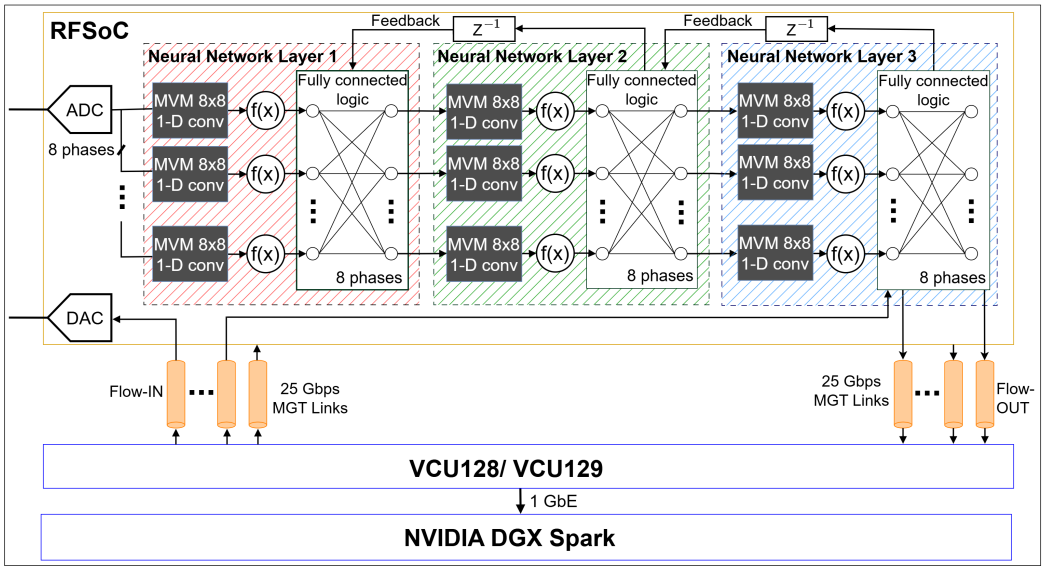


Figure 1. Proposed scalable AI + DSP compute framework.

As an extension to the testbed, the DGX Spark platform is connected to the FPGA system via a 1 GbE interface to retrieve processed RF data for executing more advanced AI algorithms.

3.2. Matrix–Vector Multiplier (MVM) Core

The hardware realized MVM core is a fully parallel multiply–accumulate (MAC) engine tailored for high-speed RF/AI applications. It is built upon an 8×8 block of input samples together with an 8×1 weight vector and performs one length-8 dot product at each FPGA fabric clock cycle. Notably, it performs a real-time convolution operation,

$$y(n) = h(n) * x(n) = \sum_{l=0}^{L-1} h(l)\, x(n-l) \tag{1}$$

where $h(n)$ denotes the convolution kernel used in AI/ML systems such as CNNs, $x(n)$ is a multi-GHz-rate wideband input signal, and L is the length of the sequence $h(n)$. This generic atomic unit can be realized on time-multiplexed or instantiated in parallel for higher-order matrix–vector operations, dense neural networks, and in polyphase finite impulse response (FIR) filters. In addition, the 8×8 MVM core can be used as a reconfigurable processing element (PE) in a tile-based AI + DSP compute fabric as proposed in this framework as illustrated in Figure 1.

In order to achieve higher bandwidths, the MVM core is designed as a M-phase polyphase architecture. Accordingly, M parallel samples are produced per FPGA fabric clock cycle. If the FPGA clock index is q, $n(k,q) = Mk + q$ for which $q = 0, 1, \ldots, M-1$ represents the effective high-rate sample index associated with the q-th interleaved ADC and DAC paths. Accordingly, the q-th output sample for each FPGA clock cycle k is

$$y_q(k) = \sum_{l=0}^{L-1} h(l)\, x(Mk - l - q). \tag{2}$$

In the prototype, the number of parallel phases is set to $M = 8$ and length of sequence $L = 512$. To support the required 512×8 polyphase convolution operation, the architecture instantiates 64 parallel copies of the 8×8 MVM cores. This parallelization enables the system to compute all $M = 8$ convolution outputs per FPGA clock cycle while sustaining the full $L = 512$ tap processing requirement. The MVM architecture was designed using the Xilinx System Generator (SysGen) tool box in MATLAB R2020a [42] to enable efficient development, verification, and seamless hardware realization. Once the design is simulated and verified, the complete MVM core was exported as a synthesizable SysGen IP. The generated SysGen IP was incorporated into the AMD Xilinx Vivado IP Integrator flow [43] using Vivado v2020.2 to construct the complete digital system and interface with the remaining IP blocks in the design. This IP-centric workflow enables tight coupling between high-level DSP design and the low-level hardware infrastructure, ensuring efficient utilization of FPGA resources. As a result, the prototyped MVM core performs 4096 MACs/clock cycle in a single data path, resulting in real-time operations across 8 parallel data paths processing at 250 MHz on digital fabric, leading to 1 TeraMAC/s throughput. For detailed implementation specifics, the reader is referred to our prior work [44].

3.3. Integration of NVIDIA DGX Spark Platform

The NVIDIA DGX Spark platform is introduced as a compact AI supercomputing solution aimed at individual developers, researchers, and students, addressing the traditional barrier of limited access to large-scale datacenter computation, which is implemented as a mini PC [45]. It is based on the Grace Blackwell GB10 superchip, which combines an Arm-based Grace CPU and a Blackwell-class GPU within a single package. This tightly coupled CPU-GPU architecture enables high AI computation throughput up to the petaFLOP scale within a constrained power envelope, making it suitable for laboratory-scale deployment alongside FPGA-based RF systems. The system provides a 128 GB unified memory space shared between the CPU and GPU, which facilitates efficient handling of large datasets and model parameters during training and fine-tuning, while the network connectivity is provided through an integrated high-speed Ethernet interface, enabling consistent data exchange with the FPGA testbed.

In addition to high-speed data converters, Xilinx Zynq UltraScale+ RF-SoC platforms such as the ZCU111 and ZCU216 integrate embedded ARM Cortex-A53 cores, tightly coupled with programmable logic (PL) on a single chip. Using the PYNQ framework [46,47], the hardware is exposed through Python and Jupyter notebook interfaces, enabling researchers and designers to capture and process the acquired RF data in a flexible, software-driven environment. Within this experimental framework, the NVIDIA DGX Spark platform serves as a complementary "AI brain in the lab", providing the heavy offline and near-real-time computational capability required for training neural receivers, channel models, synchronizer, and resource-allocation policies, as well as for conducting large-scale simulations, hyperparameter tuning, and architecture search on the acquired RF data using the PYNQ framework. The trained and optimized models can then be quantized and mapped back onto the multi-FPGA platform for inference, so that advanced AI-driven algorithms, developed using the computational power of DGX Spark, can be executed in real time over the air RF data on the proposed wireless testbed.

4. Digital Design Architecture

In this section, we discuss the digital design aspects of the realized multi-FPGA prototype testbed, detailing the soft IP cores employed in the system, as well as the hardware configuration of the RF-SoC and FPGA development boards. The overall platform is described in two architectural configurations: (i) an interconnected ZCU111-VCU128-VCU129

system linked through multiple 25 Gbps links, and (ii) an interconnected ZCU1285-VCU128 system operating over multiple 20 Gbps links. In addition, we outline the specialized cabling and hardware interfaces required to implement these high-speed connections in practice, which will help mitigate common challenges in digital-RF system integration for the broader research community.

4.1. A Starting Point to MGTs in Xilinx FPGAs

During our preliminary MGT configuration process, the Xilinx LogiCORE Integrated Bit Error Ratio Test (IBERT) IP for UltraScale/UltraScale+ GTY transceivers was employed to evaluate and monitor the high-speed serial links used in our system. This provides a reliable starting point for configuring MGT links. The IBERT core provides built-in pattern generators and checkers, enabling link verification through pseudo-random binary sequence (PRBS) testing and link status monitoring [48]. We utilized this IP primarily to configure the PLLs, determine the correct clocking configurations, including reference clock sources and reference clock rates for a desired line rate, and to identify the specific GTY transceiver quads on each board for the targeted interfaces, following the XmYn naming convention where m and n represent its (x,y) coordinates of the transceiver location out of the available transceivers. Furthermore, the IBERT design aids hardware-level debugging and validation of GTY transceiver configurations by enabling eye diagram visualization and link status monitoring for established serial connections. This ensured valid transceiver configurations, facilitating reliable point-to-point and loopback connectivity across the multi-FPGA platform.

4.2. Aurora 64B/66B IP Core

To facilitate the implementation of multi-gigabit transceivers for Xilinx FPGAs, Xilinx has developed the Aurora 64B/66B protocol IP, offering a scalable, lightweight, link-layer protocol for high-speed serial communication. As its name suggests, the protocol employs 64B/66B encoding, which achieves very low transmission overhead and improves line efficiency. It supports data lanes with throughput ranging from 0.5 Gbps to 400 Gbps and is capable of both full-duplex and simplex configurations based on user requirements [49]. The Aurora 64B/66B IP core supports the AXI4-Stream user interface, and its simple framing structure can be easily adapted to encapsulate user-specific data or even the data from other protocols, such as gigabit ethernet. Because of its low resource cost, scalable throughput, and flexible user interface, it is well-suited to chip-to-chip, board-to-board, and backplane links, as well as streaming applications [49]. It is designed to automatically initialize and maintain channel integrity by transmitting idle sequences in the absence of user payload, thereby ensuring continuous link activity while reducing design complexity for system integration.

To implement the proposed architecture, we employed the ZCU111, ZCU1285, VCU128, and VCU129 FPGA platforms, each equipped with GTY transceivers. The Aurora 64B/66B IP core is used to leverage these transceiver lanes for data transport at the designated line rates. The core's compatibility with AMD UltraScale+ and AMD Zynq device families enables seamless deployment across the selected platforms.

4.3. Multi-FPGA Prototype Testbed

We envision a scalable data processing pipeline that integrates multiple AMD Zynq UltraScale+ RF-SoCs with high-performance Xilinx FPGA platforms to serve as an emerging AI-native testbed.

The AMD Zynq UltraScale+ RF-SoC ZCU111 and ZCU1285 evaluation kits are employed as the primary RF data acquisition platforms in the proposed testbed. The ZCU111 provides eight 12-bit ADCs operating at up to 4.096 GSPS and eight 14-bit DACs at up to

6.554 GSPS, whereas the ZCU1285 extends this capability to sixteen ADC and DAC channels with comparable resolution but a reduced ADC sampling rate of 2.220 GSPS [50,51]. Despite the high RF sampling capability, the ZCU111 and ZCU1285 are constrained by relatively limited on-chip logic and DSP resources, with only 930 K system logic cells and 4272 DSP slices, which restrict their ability to implement complex AI-driven DSP workloads. To address this limitation, the acquired RF data is streamed to downstream FPGA platforms with substantially greater compute capacity. In particular, the AMD Virtex UltraScale+ VU37P FPGA on the VCU128 platform provides a significant increase in logic density, offering 2853 K system logic cells, 340.9 Mb of on-chip memory, and 9024 DSP slices, complemented by 8 GB of integrated high-bandwidth memory (HBM), thereby enabling efficient execution of compute-intensive DSP and matrix-based operations [52]. This capability is further extended by the AMD Virtex UltraScale+ VU29P FPGA on the VCU129 platform, which delivers even greater compute capacity with 3780 K system logic cells, 454.5 Mb of on-chip memory, and 12,288 DSP slices, supporting deeper processing pipelines and larger-scale parallel architectures [53]. Together, this heterogeneous architecture provides a flexible and scalable foundation for AI-native RF testbeds, where the RF-SoCs act as high-speed data acquisition front-ends and the VCU128/VCU129 platforms serve as powerful compute back-ends capable of sustaining real-time multi-gigabit dataflows.

4.4. Aurora 64B/66B IP Configuration

When working with the Aurora 64B/66B IP, proper IP configuration and the supporting logic are essential for the reliable establishment of MGT links across FPGA platforms. In our designs, the Aurora core manages GTY transceiver initialization, control character encoding and decoding, error detection, channel bonding, lane monitoring, AXI4-Stream data interfacing, and clock compensation, thereby ensuring reliable high-speed operation.

In each design, the Aurora IP was instantiated with GTY transceivers configured for the respective line rate and a lane count of four. To achieve the interested line rates of 25 Gbps and 20 Gbps, a GT reference clock of 156.25 MHz was used across all FPGA platforms, while the `INIT_clk` was configured at 250 MHz. Each channel was set in duplex mode with the framing interface enabled, and the supporting logic, including clocking and reset modules, was implemented following the guidelines of the provided example design of the Aurora 64B/66B IP.

In addition to these settings, the starting GT quad, GT lane parameter selection, and reference clock assignment should be carefully performed based on the target FPGA platform. A summary of such board-specific settings are shown in Tables 1 and 2 based on the setup.

Table 1. Board-specific Aurora 64B/66B IP configuration settings related to transceiver location for ZCU111-VCU128-VCU129 setup.

FPGA Platform	GT Connector Interface	Starting GT Quad	Starting GT Lane	GT Reference Clock
ZCU-111	SFP28	Quad X0Y1	X0Y4	MGTREFCLK1 128
VCU-128	QSFP28	Quad X0Y7	X0Y28	MGTREFCLK0 131
VCU-129	SFP28	Quad X1Y2	X1Y8	MGTREFCLK0 226

Table 2. Board-specific Aurora 64B/66B IP configuration settings related to transceiver location for ZCU1285-VCU128 setup.

FPGA Platform	GT Connector Interface	Starting GT Quad	Starting GT Lane	GT Reference Clock
ZCU-1285	Bullseye	Quad X0Y1	X0Y4	MGTREFCLK0 128
VCU-128	FMCP-HSCP	Quad X0Y0	X0Y0	MGTREFCLK0 124

4.5. Hardware Configuration

The prototype testbed is realized in two distinct configurations, which are differentiated by their operating line rates of 25 Gbps and 20 Gbps. The corresponding system-level dataflow architectures for the interconnected ZCU111–VCU128–VCU129–DGX Spark platform and ZCU1285–VCU128 platform are illustrated in Figure 2 and Figure 3, respectively.

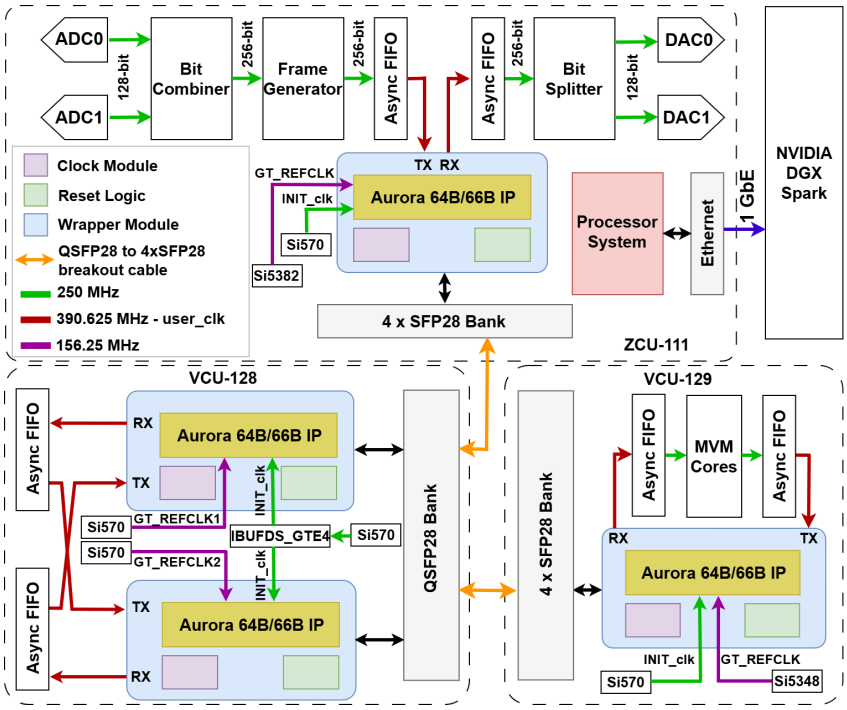


Figure 2. System-level dataflow architecture of the multi-FPGA DSP testbed incorporating the ZCU111, VCU128, VCU129, and DGX Spark platforms.

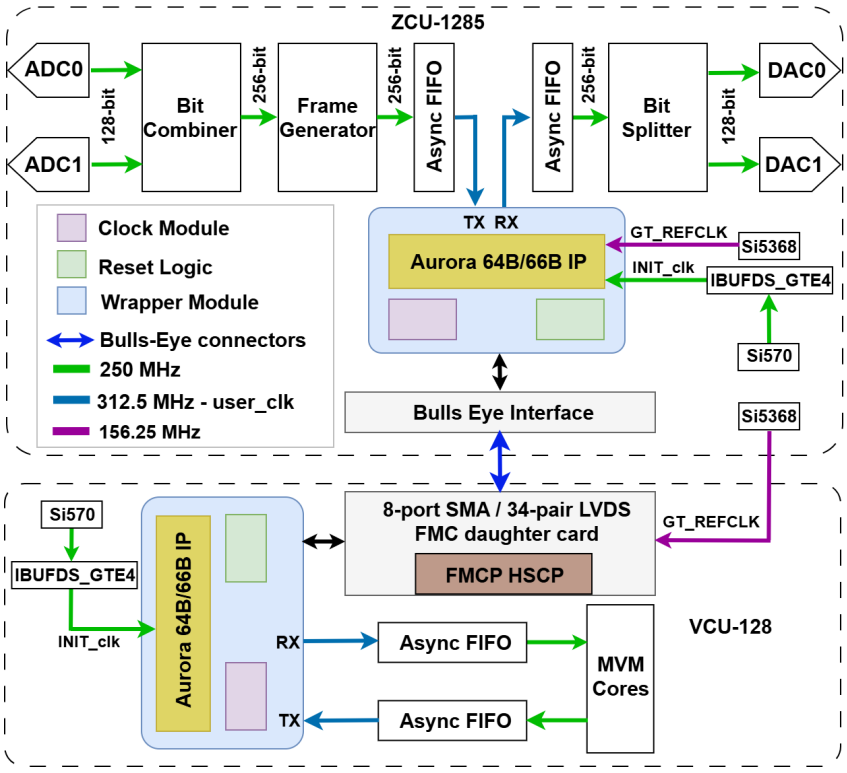


Figure 3. System-level dataflow architecture of the interconnected ZCU1285 and VCU128 platforms.

4.5.1. ZCU111–VCU128–VCU129 Platform: Interconnected at 25 Gbps

In this setup, two ADC/DAC pairs of the ZCU111 were configured with a sampling rate of 4.0 GSa/s. We employed a decimation/interpolation factor of 2, providing an effective instantaneous RF bandwidth of 1.0 GHz for each channel. The number of phases was set to 8, producing a 128-bit output word from each ADC at a 250 MHz FPGA fabric clock. These two 128-bit words, corresponding to the outputs of the two ADC channels, were repacked into a single 256-bit word. This aggregated data stream was then passed to a custom frame generator module, which generated the `_tlast` and `_tkeep` signals to define frame boundaries and formatted the output bus to comply with the AXI4-Stream protocol required by the Aurora IP. Based on the target line rate of 25 Gbps, the user logic interfacing with the Aurora core must operate at a clock frequency of 390.625 MHz. Therefore, an asynchronous first-in–first-out (FIFO) buffer configured in packet mode was employed to manage the clock domain crossing (CDC), with three CDC synchronization stages to ensure reliable data transfer. With the Aurora IP configured in a four-lane full-duplex architecture, where each lane operates at 25 Gbps, the interface achieves a combined throughput of 256 bits at 390.625 MHz, corresponding to an effective bit rate of 100 Gbps (i.e., 4 × 25 Gbps across the four lanes).

The sample data captured from the ZCU111 was then transmitted through its SFP28 interface using a 0.5 m Cisco QSFP-4SFP25G-CU50CM-compatible breakout cable (100G QSFP28 to 4 × 25G SFP28 passive direct-attach copper cables from FS.com Inc., New Castle, DE, USA), and it was received at the QSFP28 interface of the VCU128. The VCU128 design comprised two Aurora IP instances to facilitate connectivity with both ZCU111 and VCU129 platforms. In this configuration, the first Aurora instance received the incoming data, which was subsequently passed through to the second Aurora instance for transmission. Since the `user_clk` signals for the two Aurora instances were generated from independent clock sources, asynchronous FIFOs were inserted between them to ensure reliable duplex operation across the channel.

From the VCU128, the data was further transmitted to the VCU129 board via a similar QSFP28-to-SFP28 breakout cable connected to the SFP28 interface bank. On the VCU129, the AXI4-Stream data bus was routed to the DSP logic (referred to here as MVM cores) and its output is connected to the transmit interface of the Aurora IP instance, enabling the data to traverse the same boards in the reverse direction and return to the ZCU111 RF-SoC for DAC reconstruction.

In addition, the NVIDIA DGX Spark platform communicates with the ZCU111 over high-speed Ethernet using standard networking protocols (e.g., UDP/TCP sockets) through PYNQ framework to exchange RF data and control information, thereby enhancing the overall capability and flexibility of the testbed.

4.5.2. ZCU1285–VCU128 Platform: Interconnected at 20 Gbps

This setup exploits the GTY transceivers available in the ZCU-1285 RF-SoC interfacing through Samtec Bullseye connectors (Samtec RSP-200723-02-BEYE cable assembly from Samtec Inc., New Albany, IN, USA). On the VCU128 side, an 8-port SMA/34-pair LVDS FPGA Mezzanine Connector (FMC) daughter card (from HiTech Global, LLC, San Jose, CA, USA) is employed, connected through the FPGA Mezzanine card plus high serial pin count (FMCP HSCP) connector interface, which provides access to 8 serial transceivers, with the SMA cables from the RF-SoC connected to this module.

From the RF data converter side, two ADC/DAC pairs were configured with a sampling rate of 2.0 GSa/S with an effective instantaneous RF bandwidth of 1.0 GHz for each channel. The line rate was set to 20 Gbps, as the 25 Gbps configuration resulted in insufficient eye opening with the existing physical connections. As a result, the core operation

frequency is 312.5 MHz. With the Aurora IP configured in a four-lane full-duplex architecture, where each lane operates at 20 Gbps, the interface achieves a combined throughput of 256 bits at 312.5 MHz, corresponding to an effective bit rate of 80 Gbps. The digital design components used in this configuration were similar to those in the previous setup, including bit combiners/splitters, frame generators, and FIFO buffers. The captured RF data are transported to the VCU128 FPGA and fed to the MVM cores, after which the resulting output is returned to the ZCU1285 and transmitted through the DAC.

5. Experiment Setup and Results

Figure 4 compares the measured eye diagrams obtained using the IBERT eye scan utility under different physical interconnects and line rates. As shown in Figure 4a, the 25 Gbps Bullsye-to-FMC connection shows a severely constrained eye opening, with an open-UI of 0%, indicating insufficient timing margin for reliable operation at this line rate. In contrast, reducing the line rate to 20 Gbps for the same Byllsye-to-FMC path, as shown in Figure 4b, significantly improves the signal integrity showing an open-UI of 44.44% and a comparatively larger open area, enabling stable link operation. Figure 4c demonstrates that for the SFP28 based interconnect, the 25 Gbps configuration achieves a considerably wider eye opening with an open-UI of 55.56%, confirming that the SFP28 based physical channel provides better signal integrity than Bullseye-to-FMC interconnect. These measurements further validate that the Bullseye-to-FMC path is the primary limiting factor at higher line rates, thereby motivating the use of reduced line rate of 20 Gbps for the ZCU1285–VCU128 platform for robust and reliable operation.

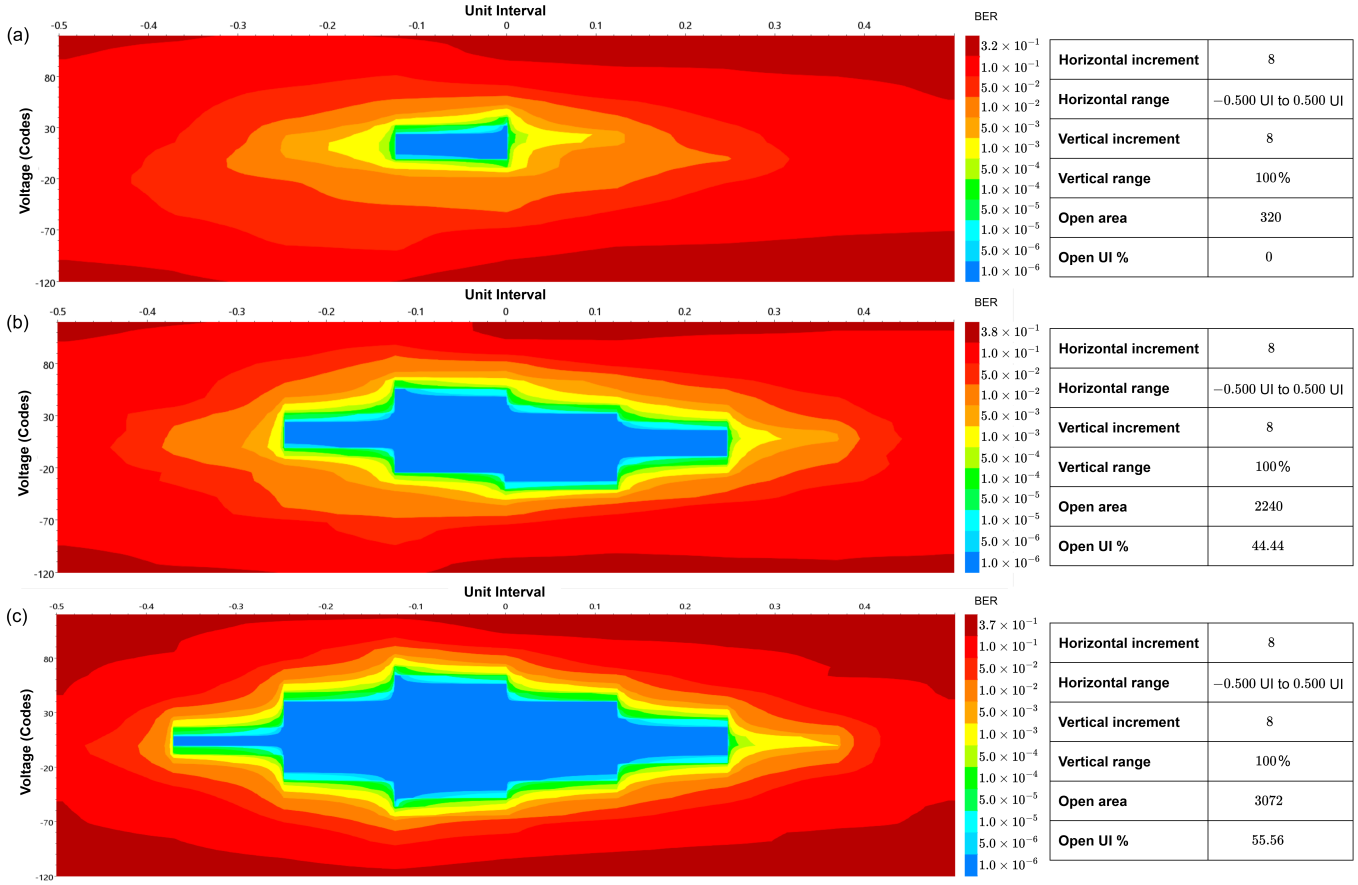


Figure 4. Integrated Bit Error Ratio Test (IBERT) eye diagram comparison for different physical interconnects and line rates: (**a**) 25 Gbps Bullseye-to-FMC connection, (**b**) 20 Gbps Bullseye-to-FMC connection, and (**c**) 25 Gbps SFP28-based connection.

Figure 5a presents the experimental setup of the ZCU111–VCU128–VCU129 platform interconnected via QSFP28-to-4×SFP28 breakout cables, with the DGX Spark connected through a 1-GbE link. For demonstration, the 512-tap MVM core, configured as a bandpass filter, is implemented on VCU-129 FPGA, with a single ADC sample data stream routed through the entire testbed. The demonstrated bandpass filter's impulse response $h(n)$ is designed using Equation (3).

$$h[n] = 0.974^n \cos\left(\frac{\pi(n-1)}{3}\right) \cdot \frac{\sin(0.15\pi(n-1)/4 + 0.01)}{0.5(\pi(n-1)/4 + 0.01)} \tag{3}$$

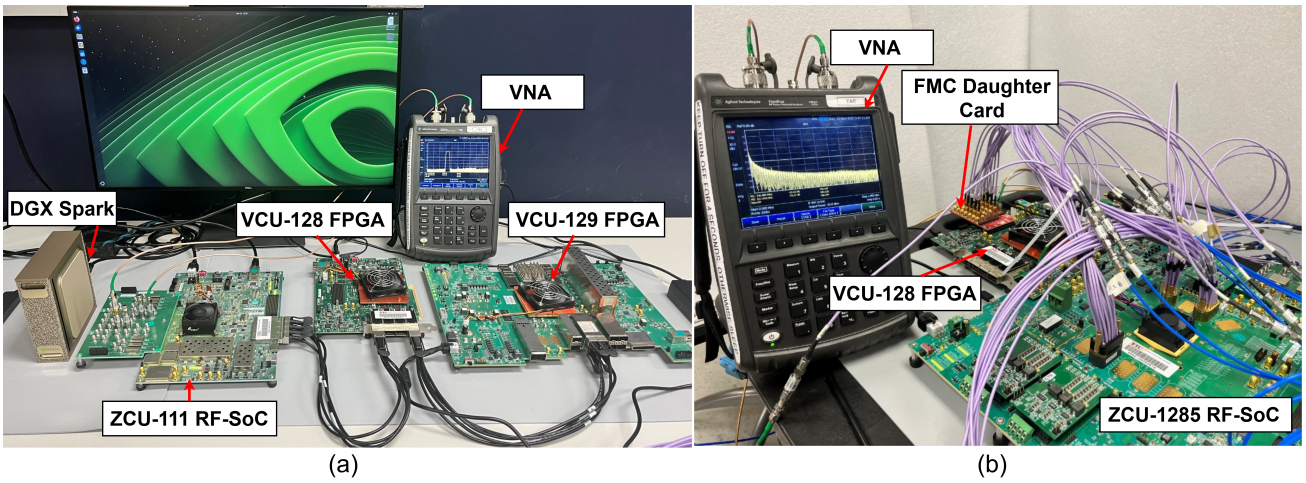


Figure 5. (**a**) Experimental setup of the ZCU111–VCU128–VCU129 platforms interconnected via QSFP28-to-4×SFP28 breakout cables, with DGX Spark connected over 1-GbE. (**b**) ZCU1285–VCU128 setup using Bullseye cables and an FMC daughter card for SMA interfacing.

The transmission coefficient ($S_{21}(dB)$) between the ADC and DAC ports was measured using an Agilent FieldFox N9923A (6 GHz) RF vector network analyzer (from Agilent Technologies, Santa Clara, CA, USA) calibrated up to 1 GHz with 10,001 points. Furthermore, the DGX Spark unit is connected to the ZCU111 with 1 GbE connection as illustrated. Figure 6a shows the normalized S_{21} magnitude response of the implemented bandpass filter. The result illustrates a sharply defined passband over the 320–360 MHz frequency range and strong attenuation across the stopband.

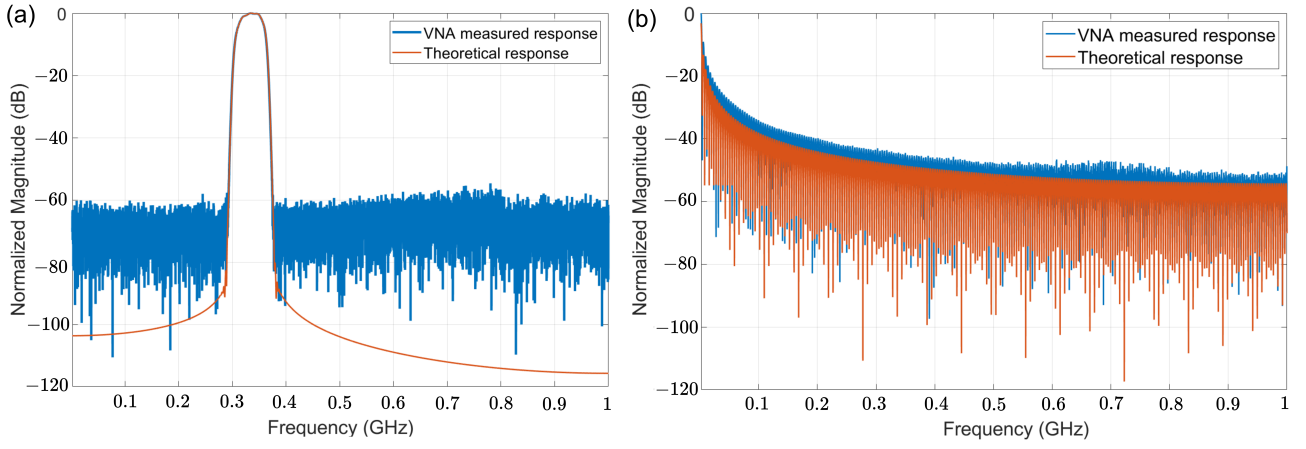


Figure 6. (**a**) Normalized measured S_{21} response of the ZCU111–VCU128–VCU129 setup, with the bandpass filter implemented on the VCU129, shown alongside the theoretical bandpass filter response. (**b**) Normalized theoretical versus measured response of the moving average filter implemented on the VCU128 in the ZCU1285–VCU128 setup.

Figure 5b depicts the corresponding setup for the ZCU1285–VCU128 platform where the MVM core configured as a moving average low-pass filter is implemented on VCU128. The low-pass filter is implemented with $h(n) = 0.9999$ using the 512-tap MVM core. Figure 6b shows the resulting normalized S_{21} magnitude response of the entire setup. Both measured responses exhibit close alignment with the MATLAB-simulated theoretical response across the full 1 GHz bandwidth. These results collectively confirm the correct operation and practical feasibility of the proposed scalable AI + DSP compute framework across the full RF bandwidth.

In addition, system-level integration of the DGX Spark platform with the proposed testbed was experimentally validated by transferring captured RF samples from the ZCU111 to the DGX Spark unit via the PYNQ framework. This integration confirms the feasibility of extending the testbed toward an AI-native architecture capable of supporting higher-level learning and inference on captured and processed RF data. While this work establishes the end-to-end dataflow and system integration, comprehensive evaluation of advanced learning and inference algorithms on the DGX Spark platform is reserved for future work.

6. Implementation Challenges and Practical Workarounds

The realization of this scalable, multi-FPGA AI + DSP computational framework using RF-SoC and high-performance FPGA platforms presents several nontrivial engineering challenges. This section discusses the key implementation issues encountered during system development, along with the solutions and workarounds adopted. These insights are intended to aid practitioners in reproducing and extending similar high-speed, SerDes-enabled RF systems.

6.1. Identification of Correct MGT Interfaces and GTY Quad Selection

A fundamental challenge during implementation was the correct identification of the MGT interfaces and their corresponding GTY quads for each FPGA platform. Each physical connector (i.e., SFP28, QSFP28, FMC, and Bullseye) maps to specific GTY transceiver locations following the Xilinx XmYn naming convention. Based on this convention, accurate selection of the starting GT quad and lane is essential for successful Aurora IP configuration. This required careful cross-referencing of FPGA evaluation board user guides [50–53] and its schematic documents provided on the Xilinx product pages. To validate the correctness of the selected interfaces, the IBERT example design [54] was extensively used. IBERT enabled verification of link establishment through lane and channel status indicators and facilitated confirmation of the correct XmYn mappings for the target interfaces. To assist readers in reproducing the design, Tables 1 and 2 under Section 4.4 list the connector interfaces, corresponding GTY quads, and starting lane selections for each RF-SoC and FPGA platform employed in this testbed.

6.2. Clocking Configuration

Selecting appropriate clocking configurations is one of the most critical aspects for stable operation of the Aurora 64B/66B IP at the target line rates. Two primary clock domains must be correctly defined: the GT reference clock and the system clock (`INIT_clk`).

For the evaluated line rates of 25 Gbps and 20 Gbps, a GT reference clock frequency of 156.25 MHz was selected across all platforms. This choice was motivated by three factors: (i) it is the default reference clock supported by the GTY transceivers on the evaluated boards, (ii) the required clock files are readily available through the system controller resources provided by the vendors, and (iii) the same reference clock configuration is recommended in IBERT reference designs for reliable high-speed operation. The specific GT reference clock sources used for each platform are listed in Tables 1 and 2 (Section 4.4).

The Aurora core additionally requires a free-running system clock (`INIT_clk`) for its initialization which can be sourced either from a dedicated FPGA fabric clock pin or from a reference clock input associated with one of the GTY transceivers. In this work, a 250 MHz `INIT_clk` was used, which consistently resulted in reliable lane initialization and channel bring-up. In cases where sufficient FPGA fabric clock pins were unavailable (specifically on the VCU128 platform), unused transceiver reference clocks were reused as the free-running system clocks. These clocks cannot be directly connected to fabric logic and must instead be routed through appropriate `IBUFDS_GTE*` primitives and constrained correctly in the Xilinx Design Constraints (XDC) file, as described in Ref. [55].

6.3. AXI4-Stream-Compliant RTL Modules

Another significant challenge involved ensuring correct AXI4-Stream protocol [56] compliance across custom digital modules used in the data path, including frame generators, and MVM core modules. Reliable inter-FPGA RF data transport requires strict adherence to AXI4-Stream protocol, particularly the correct alignment and timing of `_tvalid`, `_tdata`, `_tkeep`, and `_tlast` signals. Early design issues were addressed by updating the custom modules to ensure that the `_tvalid` signal was asserted only alongside valid data and by introducing packet-aware buffering using FIFOs configured in packet mode. These changes were essential for achieving stable, continuous data streaming across multiple FPGA devices at multi-gigabit rates.

6.4. Handling Non-Sequential GTY Lane Connectivity on QSFP28 Interfaces on VCU129

An additional practical consideration arose when configuring the Aurora 64B/66B IP on the VCU129 platform using QSFP28 interfaces. The QSFP28 connector maps to multiple GTY quads (Quads 124–127); however, within each quad, only a subset of the GTY lanes are physically connected. For instance, in Quad 124, only GTY0 and GTY2 are routed to the connector, while GTY1 and GTY3 are unconnected.

This non-sequential lane connectivity complicates GUI-based Aurora IP configuration, which typically assumes contiguous lane assignments. To address this limitation, explicit lane selection was performed during I/O planning, ensuring that only physically connected GTY lanes were enabled in the Aurora configuration. This approach is documented in the Xilinx I/O planning guidelines [57,58] and enabled successful 128-bit data transmission using multiple non-contiguous lanes within the same quad.

7. Discussion

Emerging AI-driven RF systems demand increasingly complex computational tasks, requiring scalable and efficient DSP frameworks. In this work, we present a scalable DSP compute framework tailored for AI-native RF testbeds by leveraging interconnected Xilinx multiple FPGA platforms through multi-gigabit transceiver links. We validated the implementation of the platform through two experimental setups. The ZCU111–VCU128–VCU129 platform achieves a combined throughput of 100 Gbps, corresponding to 50 Gbps per ADC stream, whereas the ZCU1285-VCU128 platform provides a slightly lower combined throughput of 80 Gbps, corresponding to 40 Gbps per ADC stream. Integrating both these configurations yields a fully interconnected, all-to-all wireless testbed that enables rapid prototyping of real-time wireless communication systems and beyond-5G/6G baseband pipelines, including higher-order modulation and demodulation schemes, polyphase filtering, real-time spectrum sensing, beamforming, and MIMO testbeds, as well as AI-assisted PHY and MAC-layer algorithms tightly coupled to high-speed RF front-ends. The experimental validation using real-time transfer function measurements across 1 GHz of instantaneous RF bandwidth, combined with the implementation of the MVM core on

high-performance FPGA platforms, further demonstrates the viability of the proposed AI + DSP computing testbed for broader scientific, commercial, space, and defense applications. Moreover, the integration with the NVIDIA DGX Spark platform provides a path toward executing higher-level AI algorithms on the RF data processed by the multi-FPGA system, enhancing its capability as a next-generation AI-enabled wireless research testbed.

Compared to prior multi-FPGA and wireless testbeds, the proposed system uniquely combines direct-RF RF-SoC front-ends with high-capacity downstream FPGA accelerators and GPU-based AI platforms using sustained multi-lane 20–25 Gbps SerDes links, enabling real-time processing of multi-gigahertz RF bandwidths that exceed the capabilities of existing implementations. Current limitations arise primarily from physical interconnect constraints at the highest line rates, specifically with the ZCU1285-VCU128 platform, while full-scale AI learning evaluations are reserved for future work.

The authors plan to expand the platform by interconnecting a ZCU216 RF-SoC evaluation kit and a Stratix-10 AX FPGA development kit to support wider-band, higher-channel-count heterogeneous RF acquisition. This expanded architecture will enable the generation of advanced, high-fidelity RF datasets that will be highly valuable to the AI-enabled wireless research community.

Author Contributions: Conceptualization and system design by A.M. Multi-FPGA platform implementation and digital designs incorporating Aurora 64B/66B IP by B.G. Digital design and implementation of MVM core architecture by G.R. Integration of digital system with hardware–software co-design using the PYNQ platform and the DGX Spark by K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by NSF (under the grant DHARMA.AI 2509389).

Data Availability Statement: All original contributions from this study are documented within the article. For further details, please contact the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Analog-to-digital converter
AI	Artificial Intelligence
CDC	Clock domain crossing
CNN	Convolutional neural network
CPU	Central processing unit
DAC	Digital-to-analog converter
DSP	Digital signal processing
FIFO	First-in-first-out
FIR	Finite impulse response
FMC	FPGA Mezzanine Connector
FMCP HSCP	FPGA Mezzanine Card Plus High Serial Pin Count
FPGA	Field Programmable Logic Array
GbE	Gigabit Ethernet
Gbps	Gigabits per second
GPU	Graphics Processing Unit
GT	Gigabit transceiver
HBM	High-bandwidth memory
IBERT	Integrated Bit Error Ratio Tester
IP	Intellectual Property

ISAC	Integrated Sensing and Communications
LVDS	Low-voltage differential signaling
MAC	Multiply–accumulate
MGTs	Multi-gigabit transceivers
MIMO	Multiple-input multiple-output
MVM	Matrix–vector multiplication
OFDM	Orthogonal frequency division multiplexing
PLL	Phase-locked loop
PR	Processing element
PRBS	Pseudo-random binary sequence
QSFP	Quad Small Form-factor Pluggable
RAM	Random access memory
RCC	Reconfigurable computimg cluster
ReLU	Rectified linear unit
RF	Radio Frequency
SerDes	Serializer/deserializer
SFP	Small Form-factor Pluggable
SoC	System-on-chip
UI	Unit interval
XDC	Xilinx Design Constraints

References

1. De Silva, U.; Silva, H.; Madanayake, A. Augmented Envelope Neural Networks on RF-SoC for Digital Self-Interference Cancellation. *IEEE Access* **2024**, *12*, 44091–44103. [CrossRef]
2. Weerasooriya, H.; Rathnasekara, G.; Restuccia, F.; Madanayake, A. RF-SoC Platforms for RF-AI Spectrum Perception. In Proceedings of the 2024 International Applied Computational Electromagnetics Society Symposium (ACES), Orlando, FL, USA, 19–22 May 2024; pp. 1–2.
3. Athavale, A.; Christensen, C. *High-Speed Serial I/O Made Simple—A Designers' Guide, with FPGA Applications*; Xilinx Connectivity Solutions: San Jose, CA, USA , 2005.
4. Stauffer, D.R.; Mechler, J.T.; Sorna, M.A.; Dramstad, K.; Ogilvie, C.R.; Mohammad, A.; Rockrohr, J.D. *High Speed Serdes Devices and Applications*; Springer Science & Business Media: New York, NY, USA, 2008.
5. Lee, S.L.; Chen, J.; Liu, S.I.; Yang, C.F.; Tsao, H.W.; Hsu, S.H.; Lin, C.C.; Yang, C.L.; Zhang, Z.J.; Fu, K.L.; et al. Development of 400 Gb/s optical transceivers for SMF based datacenter optical interconnect. In Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC), Hualien, Taiwan 30 April–1 May 2018; pp. 1–4. [CrossRef]
6. Tang, M.; Zhu, H.; Zhu, Y.; Wang, S. Research of 800 Gbit/s OSFP DR8 Silicon Photonics Optical Transceiver Module. In Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC), Wuhan, China, 27–29 November 2024; pp. 187–191. [CrossRef]
7. Suzuki, N.; Miura, H.; Mochizuki, K.; Matsuda, K. Simplified digital coherent-based beyond-100G optical access systems for B5G/6G [Invited]. *J. Opt. Commun. Netw.* **2022**, *14*, A1–A10. [CrossRef]
8. Suzuki, N.; Miura, H.; Matsuda, K.; Matsumoto, R.; Motoshima, K. 100 Gb/s to 1 Tb/s Based Coherent Passive Optical Network Technology. *J. Light. Technol.* **2018**, *36*, 1485–1491. [CrossRef]
9. Cho, H.; Ha, U.; Roh, T.; Kim, D.; Lee, J.; Oh, Y.; Yoo, H.J. 1.2 Gb/s 3.9 pJ/b mono-phase pulse-modulation inductive-coupling transceiver for mm-range board-to-board communication. In Proceedings of the 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 17–21 February 2013; pp. 202–203. [CrossRef]
10. De Silva, U.P.; Lokumarambage, A.; Malavipathirana, H.; Mohottala, C.; Thayaparan, S. IEEE 802.3 100Gbps Ethernet PCS IP design challenges and solutions. In Proceedings of the 2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 30–31 May 2016; pp. 21–25. [CrossRef]
11. Mansour, W.; Janvier, N.; Fajardo, P. FPGA Implementation of RDMA-Based Data Acquisition System Over 100-Gb Ethernet. *IEEE Trans. Nucl. Sci.* **2019**, *66*, 1138–1143. [CrossRef]
12. Ruiz, M.; Sidler, D.; Sutter, G.; Alonso, G.; López-Buedo, S. Limago: An FPGA-Based Open-Source 100 GbE TCP/IP Stack. In Proceedings of the 2019 29th International Conference on Field Programmable Logic and Applications (FPL), Barcelona, Spain, 8–12 September 2019; pp. 286–292. [CrossRef]
13. Marzetta, T.L. Massive MIMO: An Introduction. *Bell Labs Tech. J.* **2015**, *20*, 11–22. [CrossRef]
14. Maruta, K.; Falcone, F. (Eds.) *Massive MIMO Systems*; MDPI: Basel, Switzerland, 2020.

15. Abood, T.; Hburi, I.; Khazaal, H.F. Massive MIMO: An Overview, Recent Challenges, and Future Research Directions. In Proceedings of the 2021 International Conference on Advance of Sustainable Engineering and its Application (ICASEA), Wasit, Iraq, 27–28 October 2021; pp. 43–48. [CrossRef]
16. Nayebi, E.; Ashikhmin, A.; Marzetta, T.L.; Yang, H. Cell-Free Massive MIMO systems. In Proceedings of the 2015 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 8–11 November 2015; pp. 695–699. [CrossRef]
17. Björnson, E.; Sanguinetti, L. Scalable Cell-Free Massive MIMO Systems. *IEEE Trans. Commun.* **2020**, *68*, 4247–4261. [CrossRef]
18. Akyildiz, I.F.; Jornet, J.M.; Han, C. Terahertz band: Next frontier for wireless communications. *Phys. Commun.* **2014**, *12*, 16–32. [CrossRef]
19. Akyildiz, I.F.; Han, C.; Hu, Z.; Nie, S.; Jornet, J.M. Terahertz Band Communication: An Old Problem Revisited and Research Directions for the Next Decade. *IEEE Trans. Commun.* **2022**, *70*, 4250–4285. [CrossRef]
20. Petrov, V.; Pyattaev, A.; Moltchanov, D.; Koucheryavy, Y. Terahertz band communications: Applications, research challenges, and standardization activities. In Proceedings of the 2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Lisbon, Portugal, 18–20 October 2016; pp. 183–190. [CrossRef]
21. Song, H.J.; Lee, N. Terahertz Communications: Challenges in the Next Decade. *IEEE Trans. Terahertz Sci. Technol.* **2022**, *12*, 105–117. [CrossRef]
22. Lu, S.; Liu, F.; Li, Y.; Zhang, K.; Huang, H.; Zou, J.; Li, X.; Dong, Y.; Dong, F.; Zhu, J.; et al. Integrated Sensing and Communications: Recent Advances and Ten Open Challenges. *IEEE Internet Things J.* **2024**, *11*, 19094–19120. [CrossRef]
23. Sivasankar, S.; Kumarasiri, U.; Lawrance, K.; Karunanayake, K.; Sen, P.; Cintra, R.J.; Madanayake, A. Joint Communication and Sensing at D-Band SDR and at 5.8 GHz via Multibeam Arrays. *IEEE Trans. Microw. Theory Tech.* **2025**, *73*, 3236–3249. [CrossRef]
24. Elbir, A.M.; Mishra, K.V.; Chatzinotas, S.; Bennis, M. Terahertz-Band Integrated Sensing and Communications: Challenges and Opportunities. *IEEE Aerosp. Electron. Syst. Mag.* **2024**, *39*, 38–49. [CrossRef]
25. Pin Tan, D.K.; He, J.; Li, Y.; Bayesteh, A.; Chen, Y.; Zhu, P.; Tong, W. Integrated Sensing and Communication in 6G: Motivations, Use Cases, Requirements, Challenges and Future Directions. In Proceedings of the 2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S), Dresden, Germany, 23–24 February 2021; pp. 1–6. [CrossRef]
26. Wang, C.X.; Renzo, M.D.; Stańczak, S.; Wang, S.; Larsson, E.G. Artificial Intelligence Enabled Wireless Networking for 5G and Beyond: Recent Advances and Future Challenges. *arXiv* **2020**. [CrossRef]
27. Katsaros, K.; Mavromatis, I.; Antonakoglou, K.; Ghosh, S.; Kaleshi, D.; Mahmoodi, T.; Asgari, H.; Karousos, A.; Tavakkolnia, I.; Safi, H.; et al. AI-Native Multi-Access Future Networks—The REASON Architecture. *IEEE Access* **2024**, *12*, 178586–178622. [CrossRef]
28. Xu, S.; Kurisummoottil Thomas, C.; Hashash, O.; Muralidhar, N.; Saad, W.; Ramakrishnan, N. Large Multi-Modal Models (LMMs) as Universal Foundation Models for AI-Native Wireless Systems. *IEEE Netw.* **2024**, *38*, 10–20. [CrossRef]
29. Poddar, A.K.; Mayank, A.; Prabhu, A.S.; Krishan, G.D.; Kasyap P, A. A Survey on Distributed AI Systems: Architectures, Strategies, and Applications. In Proceedings of the 2025 International Conference on Intelligent and Secure Engineering Solutions (CISES), Greater Noida, Gautam Buddha Nagar, U.P., India, 11–13 August 2025; pp. 291–298. [CrossRef]
30. Glover, L.; Maguire, S.; Kerr, J. Software Defined Radio with Zynq UltraScale+ RFSoC. 2024. Available online: https://www.rfsocbook.com (accessed on 24 November 2025).
31. Chu, O. The Future of Radar Technology—Integrating RFSoC with Reconfigurable Computing for Reconfigurable Radar Systems. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 24–26 May 2024; pp. 609–615. [CrossRef]
32. Enbin, W.; Zhaojian, J.; Zhifeng, M.; Deshe, W.; Qiongzhi, W. AI Heterogeneous Radar Digital Back-end Platform Based on RFSoC. In Proceedings of the 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Beijing, China, 3–5 October 2022; pp. 1885–1889. [CrossRef]
33. Putnam, A.; Caulfield, A.M.; Chung, E.S.; Chiou, D.; Constantinides, K.; Demme, J.; Esmaeilzadeh, H.; Fowers, J.; Gopal, G.P.; Gray, J.; et al. A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services. *IEEE Micro* **2015**, *35*, 10–22. [CrossRef]
34. Ovtcharov, K.; Ruwase, O.; Kim, J.Y.; Fowers, J.; Strauss, K.; Chung, E. Toward Accelerating Deep Learning at Scale Using Specialized Hardware in the Datacenter. In Proceedings of the 27th IEEE HotChips Symposium on High-Performance Chips (HotChips 2015), Cupertino, CA, USA, 22–25 August 2015.
35. Amazon Web Services (AWS). Amazon EC2 F2 Instances. 2025. Available online: https://aws.amazon.com/ec2/instance-types/f2/ (accessed on 24 November 2025).
36. Sass, R.; Kritikos, W.V.; Schmidt, A.G.; Beeravolu, S.; Beeraka, P. Reconfigurable Computing Cluster (RCC) Project: Investigating the Feasibility of FPGA-Based Petascale Computing. In Proceedings of the 15th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM 2007), Napa, CA, USA, 23–25 April 2007; pp. 127–140. [CrossRef]
37. Buscemi, S.; Kritikos, W.; Sass, R. A Range and Scaling Study of an FPGA-Based Digital Wireless Channel Emulator. In Proceedings of the 2013 IEEE 21st Annual International Symposium on Field-Programmable Custom Computing Machines, Seattle, WA, USA, 28–30 April 2013; pp. 137–144. [CrossRef]

38. Xu, Y.; Rajagopala, A.D.; Fruitwala, N.; Huang, G. Multi-FPGA Synchronization and Data Communication for Quantum Control and Measurement. *arXiv* **2025**. [CrossRef]
39. Wouters, M.; Bourdoux, A.; Derore, S.; Janssens, S.; Derudder, V. An approach for real time prototyping of MIMO-OFDM systems. In Proceedings of the 2004 12th European Signal Processing Conference, Vienna, Austria, 6–10 September 2004; pp. 689–692.
40. Park, G.; Taing, T.; Kim, H. High-Speed FPGA-to-FPGA Interface for a Multi-Chip CNN Accelerator. In Proceedings of the 2023 20th International SoC Design Conference (ISOCC), Jeju, Republic of Korea, 25–28 October 2023; pp. 333–334. [CrossRef]
41. Gayanath, B.; Karunanayake, K.; Mohammad, H.; Venkatakrishnan, S.B.; Alwan, E.A.; Jornet, J.; Singh, A.; Madanayake, A. D-Band SDR with 64 GHz B/W on COTS Chiplets. In Proceedings of the 2025 International Applied Computational Electromagnetics Society Symposium (ACES), Orlando, FL, USA, 18–21 May 2025; pp. 1–2. [CrossRef]
42. Advanced Micro Devices, Inc. Introduction to System Generator. In *Vivado Design Suite Tutorial: Model-Based DSP Design Using System Generator (UG948)*; Advanced Micro Devices, Inc.: Santa Clara, CA, USA, 2020.
43. Advanced Micro Devices, Inc. Getting Started with Vivado IP Integrator. In *Vivado Design Suite User Guide: Designing IP Subsystems Using IP Integrator (UG994)*; Advanced Micro Devices, Inc.: Santa Clara, CA, USA, 2025.
44. Rathnasekara, G.; Madanayake, A. A Systolic Array Matrix Multiplier on RF-SoC for Wireless AI Applications. In Proceedings of the IEEE Radio and Wireless Week (RWW), Los Angeles, CA, USA, 18–21 January 2026.
45. NVIDIA Corporation. DGX Spark User Guide. 2025. Available online: https://docs.nvidia.com/dgx/dgx-spark/index.html (accessed on 24 November 2025).
46. AMD. PYNQ: Python Productivity for Zynq. 2025. Available online: https://www.pynq.io/ (accessed on 23 November 2025).
47. Karunanayake, K.; Singh, A.; Jornet, J.; Madanayake, A. PanTera: RF-SoC Testbed using PYNQ Platform for 147 GHz SDR with 64 Mbps at up to 2 km. In Proceedings of the MILCOM 2024—2024 IEEE Military Communications Conference (MILCOM), Washington, DC, USA, 28 October–1 November 2024; pp. 111–116. [CrossRef]
48. Xilinx. *IBERT for UltraScale GTY Transceivers v1.3—LogiCORE IP Product Guide*; 4 February 2021. Document No: PG196; Xilinx: San Jose, CA, USA, 2021.
49. AMD. *Aurora 64B/66B v13.0—LogiCORE IP Product Guide*; 11 December 2024. Document No: PG074 (v13.0); AMD: Santa Clara, CA, USA, 2024.
50. Xilinx, Inc. *ZCU111 Evaluation Board User Guide*; UG1271 (v1.2), 2 October 2018; Xilinx: San Jose, CA, USA, 2018.
51. Xilinx, Inc. *ZCU1285 Characterization Board User Guide*; UG1348 (v1.0), 16 July 2019; Xilinx: San Jose, CA, USA, 2019.
52. Xilinx, Inc. *VCU128 Evaluation Board User Guide*; UG1302 (v1.0), 21 December 2018; Xilinx: San Jose, CA, USA, 2018.
53. Xilinx, Inc. *VCU129 Evaluation Board User Guide*; UG1238 (v1.0), 21 June 2017; Xilinx: San Jose, CA, USA, 2017.
54. Advanced Micro Devices, Inc. *Generating and Implementing the IBERT Example Design*; Vivado Design Suite User Guide: Programming and Debugging (UG908); Advanced Micro Devices, Inc.: Santa Clara, CA, USA, 2025.
55. AMD—Adaptive SoC & FPGA Support. *67029—Using a Transceiver Reference Clock as System Clock for a Debug Core Such as IBERT*; 23 September 2021. Knowledge Article; AMD: Santa Clara, CA, USA, 2021.
56. AMD. How AXI4-Stream Works. In *Vitis High-Level Synthesis User Guide (UG1399), Section "How AXI4-Stream Works"*; AMD: Santa Clara, CA, USA, 2025.
57. Xilinx, Inc. *Vivado Design Suite User Guide—I/O and Clock Planning*; UG899 (v2022.1), 4 May 2022; Xilinx: San Jose, CA, USA, 2022.
58. AMD Adaptive Support Community. How to Configure Aurora 64B/66B IP on VCU129 Board with Nonsequential GTY Transceiver Connections for 128 bit Data Transmission. 2026. Available online: https://adaptivesupport.amd.com/s/question/0D54U00008fGDekSAG/how-to-configure-aurora-64b66b-ip-on-vcu129-board-with-nonsequential-gty-transceiver-connections-for-128bit-data-transmission?language=en_US (accessed on 8 January 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.