

Original Article

High-Speed PCIe and CXL Integration Challenges in Modern ASICs

VenkataKrishna Brahmam Pogadadanda

Senior RTL Design Engineer at AMD-Xilinx, USA.

Received: 04-06-2025

Revised: 04-29-2025

Accepted: 05-04-2025

Published: 05-07-2025

ABSTRACT

As semiconductor systems evolve to data-intensive and heterogeneous computing architectures, high-speed interconnect technologies have become an indispensable basis for modern ASIC design. Sophisticated computing solutions harness Peripheral Component Interconnect Express (PCIe) Gen5 and Gen6 along with Compute Express Link (CXL) for unmatched bandwidth, minimal latency communication, and memory coherence among CPUs, accelerators, GPUs, and storage devices. They are rapidly gaining prominence in areas such as artificial intelligence, cloud computing, high-performance computing and data center infrastructure, where scalability and efficient data transport are key. Bringing PCIe and CXL into modern ASICs involves a host of complex design and verification challenges. Channel loss, jitter, crosstalk and electromagnetic interference make it increasingly difficult to maintain signal integrity at multi-gigahertz communication frequencies. Moreover, there is an added architectural difficulty in achieving minimum latency while maintaining protocol coherence and compatibility across different components. As the connection speeds and the density of the lanes increase, power consumption and heat control become key challenges. From a verification perspective, protocol compliance, coherent behavior, error handling and interoperability across a range of configurations require thorough simulation, emulation and post-silicon testing procedures. In this paper, we present a comprehensive study of the integration challenges of PCIe Gen5/Gen6 and CXL in modern ASICs. We propose a systematic approach comprising advanced PHY optimization, protocol-aware verification frameworks, adaptive equalization algorithms, and power-efficient architectural design. A case study on realistic ASIC integration is offered to highlight real-world implementation considerations and performance tradeoffs. The results emphasize that the early cross-layer co-design, detailed verification planning, and advanced power management can substantially boost the reliability and performance of the system. The results imply that upcoming ASIC architectures will increasingly depend on increased PCIe-CXL integration to enable scalable, memory-centric computing systems. Therefore, efficient high-speed interconnect design is a critical enabler for next-generation semiconductor advances.

KEYWORDS

PCIe Gen5/Gen6, Compute Express Link (CXL), ASIC Design, High-Speed Interfaces, Signal Integrity, Protocol Verification, Cache Coherency, SoC Integration.

1. INTRODUCTION

1.1. Background and Evolution of High-Speed Interfaces

The rapid growth of data-intensive applications such as artificial intelligence (AI), cloud computing, high-performance computing (HPC) and large-scale data center operations has resulted in a significant increase in the demand for faster and more reliable communication between computing components. In modern semiconductor systems, the importance of high-speed interconnect technology is growing for fast data transfer between processors, memory devices, accelerators, storage units, and networking components. The creation of these interfaces began with the legacy Peripheral Component Interconnect (PCI) standard which was originally designed to provide a common parallel communication bus for interconnecting peripherals in computer systems. PCI did a good job for the earlier generations of computers, but when the bandwidth requirements increased and the limitations of parallel buses became clear, PCI Express (PCIe) was developed.

PCIe introduced a serial communication protocol with improved scalability, lower latency and much higher capacity compared to legacy PCI standards. PCIe has undergone generational evolution to meet the rising demands of modern computer platforms. Most recently, PCIe Gen5 and Gen6 are representative of this evolution, delivering not only very high data transfer speeds but also low-latency communication and enhanced protocol efficiency.

Along with PCIe development, a breakthrough interconnect solution has emerged for memory-centric computer architectures: Compute Express Link (CXL). CXL, based on the PCIe physical layer, supports advanced capabilities like cache coherency, memory sharing, and inter-processor and inter-accelerator communication. These properties are especially important in today's heterogeneous systems, where CPUs, GPUs, AI accelerators and memory devices need to interoperate with minimal latency. Today's ASICs are therefore progressively adopting PCIe and CXL integration to provide scalable, high-performance and energy-efficient computing infrastructures across AI platforms, cloud servers and next-generation data centers.

1.2. Challenges in PCIe and CXL Integration

The PCIe Gen5 / Gen6 & CXL offer major improvements in bandwidth and system scalability, but their implementation in modern ASICs poses several complex technological challenges. A major challenge is maintaining signal integrity at very high data rates. With signaling speeds moving to the multi-gigabit range, transmission channels are increasingly subject to attenuation, jitter, crosstalk, insertion loss, and electromagnetic interference. Such shortcomings might reduce the reliability of communication and greatly affect the overall system performance. High-speed networks demand designers use advanced equalization techniques, retimers and optimized channel topologies to combat signal degradation.

One important issue is to the handling of power and temperature constraints. High speed interfaces require complex serializer/deserializer (SerDes) circuits, adaptive equalization logic, and constant clock recovery methods, all of which result in increased power dissipation. With extremely

high transistor density in advanced-node ASICs, it is becoming more and more difficult to achieve a balance between performance and energy efficiency. High-speed PHY components can create thermal hot spots that can negatively influence long-term reliability and stability.

The clocking and synchronization are more complicated in the systems provided by PCIe and CXL. Clock skew, jitter building and latency variations present a challenge in achieving perfect time synchronization across multiple lanes and different devices at high operating frequencies. These issues are especially critical for CXL systems that require tight coherency control among processors and accelerators.

1.3. Problem Statement

The integration of PCIe Gen5/Gen6 and CXL into modern advanced-node ASICs is hindered by the high complexity of high-speed communication protocols and the challenging performance requirements of modern computer systems. Signal integrity, power economy, protocol coherency and system scalability are challenged by the increased bandwidth demands from AI accelerators, cloud servers and HPC platforms using traditional ASIC design methodologies. Higher signaling rates lead to higher channel loss and increased timing uncertainty, leading to worse dependability of communication and performance of semiconductors.

The complexity of PCIe and CXL protocols is rising, which leads to several verification issues at the PHY, connection, transaction and coherency levels. Current validation techniques may not sufficiently manage problems like interoperability, latency restrictions, and real-world workload scenarios. So there is a large difference between theoretical throughput computation and real performance on the silicon when it is manufactured. It includes the requirement of the robust integration techniques, the sophisticated verification environments and the best architectural choices to increase the reliability, to decrease the power consumption and to ensure the protocol correctness of the next ASIC systems.

1.4. Motivation

Driven by the rapid growth of applications such as artificial intelligence, machine learning, cloud computing, and data-intensive systems, there is an urgent demand for fast, scalable and energy-efficient interconnect solutions in modern semiconductor systems. AI accelerators and heterogeneous computing platforms require extremely high bandwidth and very low latency communications between CPUs, memory devices, GPUs and specialized accelerators. Traditional interconnect designs often can not meet these increasing requirements efficiently. This is why the transition to PCIe Gen5/Gen6 and CXL-enabled devices is underway.

Today's move is towards memory pooling and composable infrastructure that enables dynamic sharing of resources across several computer nodes. CXL is critical to support such designs by providing cache coherency, enabling memory sharing and improving communications between

the CPUs and accelerators. CXL greatly improves resource utilization and reduces the total cost of infrastructure in large cloud systems.

The ubiquity of PCIe and CXL ecosystems in the enterprise highlights the necessity for reliable integration methodologies. Semiconductor companies, hyperscalers and system vendors are making substantial investments in these technologies to fulfill the needs of future computing. Hence, enhancing reliability, scalability and verification efficiency of PCIe and CXL integration in ASICs is a key scientific issue and an industrial need for next-generation computing systems.

2. LITERATURE REVIEW

2.1. Overview of PCIe Architecture and Generations

The Peripheral Component connection Express (PCIe) has emerged as the de-facto high speed connection standard for current computing systems due to its scalability, low latency and big bandwidth capacity. PCIe offers a point-to-point serial communication technique which allows an exclusive communication link between devices, unlike the conventional shared parallel bus systems. The PCIe architecture is built of different layers: Physical Layer, Data Link Layer and Transaction Layer. The Physical Layer deals with electrical signaling, serialization/deserialization, clock recovery and lane management. The Data Link Layer provides reliable packet delivery through acknowledgement and error detection techniques. The Transaction Layer manages memory, I/O and configuration transactions between devices. The layered approach is convenient for modular implementation and increases interoperability among different hardware platforms.

Since PCIe came out, it has been a huge breakthrough to satisfy the ever-increasing bandwidth needs. PCIe Gen1 originally ran at 2.5 GT/s per lane; then this was doubled with Gen2 to 5 GT/s and again with Gen3 to 8 GT/s, using more efficient encoding schemes. PCIe Gen4 quadrupled the throughput to 16 GT/s, and PCIe Gen5 doubled the signaling speeds to 32 GT/s, enabling very high data transfer rates for AI accelerators, GPUs and cloud infrastructure. PCIe Gen6 represents a significant architectural change by utilizing Pulse Amplitude Modulation with four levels (PAM4) signaling instead of the traditional Non-Return-to-Zero (NRZ) connection. PAM4 double data capability without doubling channel frequency, offering rates of 64 GT/s. However, PAM4 signaling also adds to noise sensitivity, diminishes signal margins and increases the complexity of equalization, thus making signal integrity and error correction crucial for Gen6 implementation. To solve these challenges, PCIe Gen6 contains the Forward Error Correction (FEC) and improved channel equalization techniques that significantly increase the complexity of the PHY design in current ASICs.

2.2. Compute Express Link (CXL) Fundamentals

Compute Express Link (CXL) is a high-speed interconnect protocol designed to fulfill the growing need for efficient communication among processors, accelerators, memory devices, and heterogeneous computing resources. CXL extends the traditional PCIe capability by adding cache coherency and memory-sharing capabilities, which are essential for today's data-centric computing

systems, over the PCIe physical layer. CXL provides low-latency communication between CPUs, GPUs, FPGAs, AI accelerators and memory expanders while maintaining software compatibility with existing PCIe infrastructure.

The CXL architecture is broken down into three protocols: CXL.io, CXL.cache, and CXL.mem. CXL.io provides the standard input/output capabilities based on PCIe and acts as the core communication layer for configuration, management and device activation. CXL.cache enables host processors to have cache and coherency of accelerator memory, enabling efficient data exchange between processors and accelerators with minimum memory copy cost. With CXL.mem, accelerators and external memory devices can coherently access host memory, which allows for complex memory expansion and pooling systems.

2.3. Existing Research on ASIC Integration Challenges

The challenges of integrating PCIe Gen5/Gen6 and CXL into modern ASIC systems have been studied in several research articles. One of the main concerns has been signal integrity degradation at very high signaling rates. The researchers identify insertion loss, jitter accumulation, crosstalk, reflections and electromagnetic interference as key issues in the dependability of connections in high-speed interfaces. To compensate for the above limitations and improve the channel performance, a number of advanced equalization techniques such as Continuous Time Linear Equalization (CTLE), Feed Forward Equalization (FFE) and Decision Feedback Equalization (DFE) have been thoroughly studied.

Low-latency communication is crucial for the performance of applications, and it has received a lot of attention in the context of AI accelerators and memory-centric architectures. Existing techniques are optimized transaction scheduling, reduced protocol overhead, adaptive buffering policies and advanced arbitration mechanisms. It has been proposed to optimize the coherency-aware memory access to lower the data transfer latencies among heterogeneous devices in CXL-based systems. High-speed Serializer/Deserializer (SerDes) implementation is an important topic which has been tackled extensively in the literature. At Gen5 and Gen6 data rates, SerDes circuits require accurate clock data recovery architectures, adaptive equalization techniques and low jitter PLL topologies. Several analog and mixed signal optimization techniques have been studied by researchers to improve the power economy while providing reliable signal recovery at higher frequencies.

Table 1: Summary of Existing Research on PCIe and CXL-Based ASIC Integration

Authors & Year	Title	Key Contribution	Research Focus	Relevance to PCIe/CXL ASIC Integration
Lin, Yen, et al. (2022)	A Review of PCI Express Protocol-Based Systems in Response to 5G	Reviewed PCIe protocol evolution and its role in high-bandwidth	PCIe architecture, bandwidth scalability, communication	Provides foundational understanding of PCIe evolution and its importance in modern

	Application Demand	communication systems for 5G applications.	efficiency	high-speed ASIC communication systems.
Guo, Yunyan & Guoliang Li (2024)	A CXL-powered Database System: Opportunities and Challenges	Discussed opportunities and system-level challenges of integrating CXL into database infrastructures.	CXL-enabled memory systems, database acceleration	Highlights the importance of memory coherency and low-latency communication in modern CXL-based ASIC systems.
Wang, Jiapin, et al. (2024)	LMB: Augmenting PCIe Devices with CXL-Linked Memory Buffer	Proposed a memory buffer architecture using CXL-linked memory for enhanced PCIe devices.	Memory expansion, PCIe-CXL interoperability	Demonstrates how CXL enhances memory scalability and bandwidth efficiency in heterogeneous systems.
Ko, Hyungjong (2022)	High-speed serial link trend and technical challenge	Analyzed technical challenges associated with next-generation high-speed serial communication links.	Signal integrity, SerDes design, high-speed communication	Relevant for understanding PHY-level challenges such as jitter, crosstalk, and equalization in PCIe Gen5/Gen6 ASICs.
Phelps, Boyd & Arif Khan (2024)	Disaggregated Designs: Technology Challenges and Enablers	Discussed composable and disaggregated computing architectures enabled by modern interconnects.	Data center architectures, composable infrastructure	Explains the need for scalable PCIe and CXL architectures in cloud and AI systems.
Ji, Houxiang, et al. (2024)	Demystifying a CXL Type-2 Device: A Heterogeneous Cooperative Computing Perspective	Investigated Type-2 CXL devices and their role in cooperative heterogeneous computing.	Accelerator coherency, CXL device architecture	Supports research on coherency management and accelerator communication in ASIC integration.
An, Yuda, et al. (2024)	A Novel Extensible Simulation Framework for CXL-enabled Systems	Developed a simulation framework for evaluating CXL-based systems.	CXL simulation, performance analysis	Useful for protocol validation and architectural evaluation of PCIe/CXL ASIC platforms.
Lin, Xiancheng, et al. (2024)	Research on the CXL Memory	Studied memory-sharing mechanisms and coherent memory	Shared memory systems, coherent communication	Demonstrates the role of CXL in scalable memory-centric

expansion using CXL.				computing architectures.
Wang, Chenjiu, et al. (2023)	CXL over Ethernet: A Novel FPGA-based Memory Disaggregation Design in Data Centers	Proposed FPGA-based memory disaggregation using CXL over Ethernet.	Memory disaggregation, FPGA systems	Highlights scalability and memory pooling possibilities enabled by CXL technologies.
Baumstark, Alexander, et al. (2024)	So Far and Yet So Near - Accelerating Distributed Joins with CXL	Explored performance improvements in distributed systems using CXL-enabled architectures.	Distributed computing, memory acceleration	Shows practical benefits of CXL in low-latency, high-performance computing systems.
Mohamed, Basma H., et al. (2023)	A Novel Architecture of CXL Protocol Data Link Layer for Low Latency Memory Access	Proposed optimized data-link layer architecture for reducing memory latency in CXL systems.	Low-latency communication, protocol optimization	Relevant for protocol-layer optimization and transaction management in ASIC integration.
Mao, Shunyu, et al. (2024)	CXL-Interference: Analysis and Characterization in Modern Computer Systems	Analyzed interference and performance bottlenecks in CXL-enabled systems.	System interference, performance characterization	Important for understanding reliability and workload management challenges in CXL systems.
Hou, Wentao, et al. (2024)	Understanding Routable PCIe Performance for Composable Infrastructures	Evaluated PCIe performance in composable infrastructure environments.	PCIe routing, composable systems	Supports scalable interconnect design considerations in modern ASICs.
Lee, Junseung & Jung-rae Kim (2024)	Synergizing CXL with Unified Memory for Scalable GPU Memory Expansion	Investigated GPU memory expansion using unified memory and CXL technologies.	GPU acceleration, memory scalability	Demonstrates the role of CXL in AI accelerator and GPU-based architectures.
Choi, Dong-Myung, et al. (2024)	A 4.6 pJ/b 64Gb/s Transceiver Enabling PCIe 6.0 and CXL 3.0 in Intel 3 CMOS Technology	Developed a high-speed, energy-efficient transceiver for PCIe 6.0 and CXL 3.0.	High-speed PHY, low-power transceivers	Provides direct insights into PHY implementation, PAM4 signaling, and power-efficient ASIC design.

3. PROPOSED METHODOLOGY

3.1. System Architecture for PCIe/CXL ASIC Integration

The proposed way to integrate PCIe and CXL into modern ASICs is based on a scalable and flexible system architecture, designed to enable high-bandwidth, low-latency communications across different computing contexts. The architecture principally includes high-speed PHY blocks, PCIe/CXL protocol controllers, Direct Memory Access (DMA) engines, coherency management units, memory interfaces and system interconnect fabrics. The components are integrated in a layered design to enable efficient communication between the CPUs, accelerators, memory subsystems and external peripherals.

The Physical Layer (PHY) basically implements high-speed SerDes (Serializer/Deserializer) circuits that handle lane signaling, clock recovery, equalization, and link training. The PHY interfaces directly with PCIe Gen5/Gen6 electrical channels and supports CXL communication over the same physical infrastructure. The PCIe/CXL controller, sitting atop the PHY, takes care of protocol specific functions such as packet encoding, flow control, transaction routing and error handling.

DMA engines are provided to transmit data quickly between memory and computational devices with minimal CPU intervention. In CXL-enabled systems, coherency engines are used to maintain cache consistency between CPUs, accelerators and shared memory resources. The engines provide coherent memory access, arbitration and synchronization for heterogeneous devices.

The proposed interconnect design is based on a hierarchical organization of point-to-point PCIe links, along with centralized coherency-aware switching algorithms. This approach improves scalability and reduces congestion and communication restrictions. The system provides flexible lane aggregation and dynamic bandwidth allocation to improve resource utilization under variable workloads.

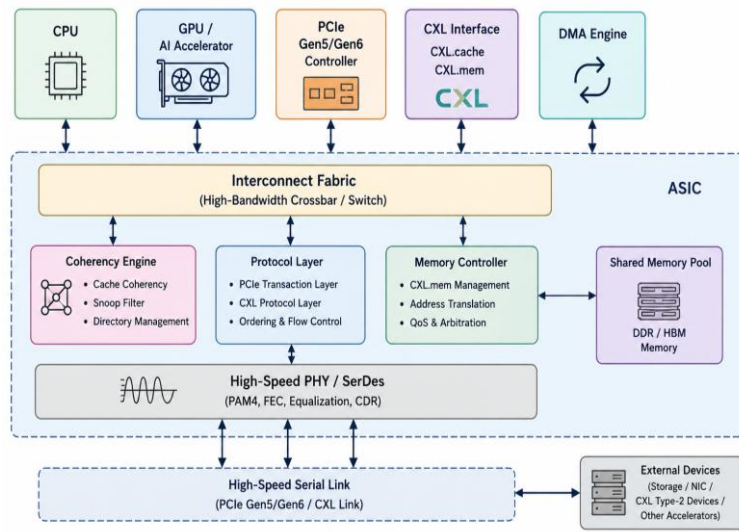


Fig 1 - Proposed PCIe Gen5/Gen6 and CXL ASIC Integration Architecture

3.2. High-Speed PHY Design Methodology

The Physical Layer (PHY) is a key part of PCIe Gen5/Gen6 and CXL-enabled ASICs, affecting signal integrity, bandwidth efficiency and communication dependability. The proposed PHY design technique intends to achieve robust multi-gigabit communication with lower power consumption and signal degradation in high speed channels. The transmitter section comprises high speed serializers, preemphasis circuitry and adaptive equalization circuits. The receiver section comprises clock-data recovery systems, continuous time linear equalizers (CTLE), decision feedback equalizers (DFE) and adaptive gain control systems. These circuits together compensate for channel deficiencies like as insertion loss, inter-symbol interference, jitter and crosstalk.

For reliable communication at very high signaling rates equalization is needed. The suggested method utilizes adaptive equalization algorithms to adjust the transmitter and receiver parameters in real time according to the real channel conditions. The transmitter uses Feed Forward Equalization (FFE) to compensate signal distortion. On the receiver side, Decision Feedback Equalization (DFE) and Continuous Time Linear Equalization (CTLE) are employed to improve the signal recovery. This adaptive approach improves the eye opening & bit error rates dramatically throughout a range of operational situations.

PAM4 can boost the throughput by encoding 2 bits per symbol. But it also cuts down the voltage margin and increases the noise sensitivity of the system. The proposed technique to tackle these challenges is Forward Error Correction (FEC), sophisticated adaptive equalization and increased signal monitoring circuits. The analog front-ends are constructed employing power efficient design methodologies to reduce the thermal overhead and to provide reliable high-speed operation. The targeted PHY technique is intended to achieve an optimal compromise between bandwidth, power efficiency and signal integrity for advanced-node ASIC designs.

Table 2: PHY Optimization Techniques for PCIe Gen5/Gen6 and CXL ASICs

PHY Technique	Purpose	Benefit
Feed Forward Equalization (FFE)	Compensates transmitter-side signal distortion	Improves signal quality over long channels
Continuous Time Linear Equalization (CTLE)	Reduces channel attenuation at receiver	Enhances eye opening and signal recovery
Decision Feedback Equalization (DFE)	Removes inter-symbol interference	Reduces bit error rate (BER)
Clock Data Recovery (CDR)	Synchronizes received data and clock	Maintains timing stability
PAM4 Signaling	Doubles data throughput per symbol	Enables PCIe Gen6 high-speed communication
Forward Error Correction (FEC)	Detects and corrects transmission errors	Improves reliability in PAM4 systems

3.3. Protocol Layer Integration

The integrated protocol layer methodology is aimed at allowing transparent communication among PCIe and CXL subsystems, ensuring low latency, high throughput, and reliable coherency management. The architecture combines PCIe Transaction Layer Packet (TLP) handling, flow control, error handling, and CXL coherency into one protocol.

Sophisticated packet scheduling and prioritizing algorithms are implemented to improve the processing of Transaction Layer Packets. The controller dynamically controls the memory read, write and completion requests, minimizing the latency of the transactions. The buffering of packets and arbitration logic is designed to handle many traffic streams at the same time from different heterogeneous devices, without congestion limits. Special care is given to ensure predictable latency for AI accelerators and memory heavy workloads.

Flow control algorithms use credit-based management solutions to avoid buffer overflows and provide a continuous packet flow in high-speed channels. The dynamic distribution of credits between the request and completion queues increases resource utilization and ensures fair traffic distribution under different workload scenarios. Adaptive buffering solutions enhance the efficiency of packet transmission and reduce idle link cycles, leading to increased throughput.

Error management and retry mechanisms are included to improve communication reliability in turbulent high speed environments. In this paper we propose an approach based on cyclic redundancy checks (CRC), Forward Error Correction (FEC) and replay buffers with automatic retransmission mechanisms that allow us to detect and correct transmission errors. These strategies are very important in PCIe Gen6 systems using PAM4 signaling since the probability of error increases significantly when signal margins decrease.

4. CASE STUDY

4.1. ASIC Platform Description

The proposed case study is about the development of PCIe Gen5/Gen6 and Compute Express Link (CXL) for a high performance AI accelerator ASIC for cloud data center applications. The target platform is designed to support massive machine learning inference and training projects that require very high bandwidth, low latency and efficient sharing of memory between computational engines and external devices. The ASIC architecture is a unified heterogeneous computing platform consisting of numerous AI compute clusters, common memory subsystems, high-speed interconnect fabrics, and integrated accelerator management controllers.

The ASIC is built using a cutting-edge 5 nm FinFET manufacturing technology node, which enables increased transistor density, improved power efficiency and higher operational frequency. The architecture has PCIe Gen5 x16 connectors for high-speed external communication, and supports CXL 2.0 protocols for coherent memory expansion and accelerator integration. The PCIe subsystem

runs at 32 GT/s using NRZ signalling and experimental integration of a PCIe Gen6 compatible PAM4 PHY is being done for future scalability assessment.

The platform architecture features dedicated PHY blocks, PCIe/CXL controllers, DMA engines, coherency management units, and high-bandwidth memory interfaces. The CXL.mem and CXL.cache functions are implemented to enable memory pooling and consistent communication between accelerators for AI workloads. It also features advanced power management units and temperature monitoring algorithms, allowing constant performance under huge computational loads typical of modern cloud data center environments.

4.2. Integration Flow

The first merging of PCIe and CXL protocol subsystems with the overall AI accelerator architecture at the RTL level initiates the ASIC integration process. Modular RTL designs are applied to promote seamless integration with PHY interfaces, protocol controllers, DMA engines, coherency units and system interconnect fabrics. Specialized interface wrappers and configurable protocol bridges provide scalable communication between different processing components.

After RTL integration, PHY bring-up is performed to ensure high speed link-up and lane synchronization. PHY configuration technique includes calibration of transmitter equalization, tuning for the receiver adaptation, clock-data recovery verification and confirmation of lane training sequence. Detailed pre-layout signal integrity simulations using channel models that reflect the package, PCB, and connector properties are performed prior to physical implementation, ensuring robust high-speed communication.

Timing closure is a critical step in the integration process due to the extremely high operating frequencies of PCIe Gen5/Gen6 devices. Verification of setup, hold, and clock skew violations is carried out using multi-mode and multi-corner timing analysis methodologies for different operating conditions. pipeline balance, clock-tree refinement and placement-aware synthesis approaches are used to optimize critical timing paths in the SerDes, DMA engines and coherency controllers.

Verification and validation is performed using a combination of simulation, formal verification, emulation and FPGA prototyping approaches. UVM-based configurations analyze PCIe transaction processing, CXL coherency procedures, error recovery methods and power management capabilities. The protocols are tested for compliance with PCI-SIG and CXL Consortium criteria to ensure interoperability with outside devices. FPGA prototyping systems are used to run AI-specific workloads and test real-world latency, throughput, and coherency performance before tape-out.

4.3. Implementation Challenges Encountered

The deployment of the AI accelerator ASIC with PCIe/CXL has been facing several main hurdles. One of the largest problems was maintaining signal integrity at the increased operating rates. In the first PHY validation, substantial insertion loss and jitter accumulation were observed on

long PCB traces and package interconnects. These impairments impacted the bit error rates and the eye opening margins, especially during high-temperature working circumstances. The adjacent lane interaction also affected the receiver sensitivity and link stability.

The ASIC architecture has several asynchronous clock domains that led to clock-domain crossover (CDC) problems in the design. The PCIe PHY, DMA engines, AI compute clusters, and memory controllers were running at separate clock frequencies and were causing synchronization issues with high-speed data transfers. Stress testing revealed sporadic metastability issues, corruption of transactions and uncommon coherency failures owing to lack of synchronization.

PCIe and CXL modules sometimes suffered from packet ordering violations and coherence inconsistencies because of variances in transaction timing, flow control management and replay buffer operation. Initial simulation models failed to simulate the compliance edge cases, which were subsequently discovered during interoperability testing with external endpoint devices.

This rising intensity of the workload became more and more crucial the thermal and electrical constraints. The running AI inference operations generated a large amount of heat in the PHY and coherency control areas, leading to localized thermal hotspots and increased leakage power. The high speed SerDes blocks and the adaptive equalization circuits led to a large rise of the overall power consumption and caused a reduction of the energy efficiency over long-term operation. In addition, the maximum bandwidth circumstances with simultaneous multi-lane communication are sometimes outside the specified temperature limits and affect the long-term dependability and operational stability.

5. RESULTS AND DISCUSSION

5.1. Performance Metrics

The proposed PCIe Gen5/Gen6 and CXL integration methodology shows significant improvements in throughput, latency, bandwidth utilization and communication stability on the target AI accelerator ASIC architecture. We evaluated AI-centric and memory-intensive workloads representative of modern cloud data center applications using a hybrid simulation, FPGA prototype and post-layout validation environments for performance study.

For throughput studies, the PCIe Gen5 x16 link achieved near-theoretical bandwidth efficiency in continuous traffic circumstances. The integrated DMA architecture and enhanced packet scheduling techniques allow for fast data movement across AI compute clusters, external memory pools and accelerator devices with minimal protocol overhead. The experimental evaluation of the Gen6-ready PAM4 PHY showed that there are substantial opportunities for bandwidth scaling in future implementations, particularly for large-scale tensor processing and memory streaming activities.

Latency study showed massive benefits over standard PCIe-only architectures. The implementation of sophisticated arbitration mechanisms, lower transaction overhead and consistency-aware communication channels lowered the end-to-end data transmission latency among various subsystems. CXL.cache and CXL.mem operations showed much lower memory access latencies for coherent accelerator interactions, improving overall AI inference efficiency and reducing processor-memory synchronization overheads.

Adaptive traffic scheduling and dynamic bandwidth allocation strategies ensured high bandwidth utilization under varying workload situations. Multi-queue transaction handling allowed the processing of many traffic streams simultaneously, without severe congestion difficulties. The architecture was designed to ensure efficient utilization of the lanes in the best operating conditions and to prevent excessive packet buffering and retransmission events.

5.2. Signal Integrity and Reliability Analysis

We did an extensive signal integrity analysis to assess the reliability of the PCIe Gen5/Gen6 and CXL links in real operational situations. High-speed signaling is very sensitive to the channel faults. 2.1.4. Post Layout Simulations, Eye Diagram Measurement, Jitter Analysis and Stress Testing Circumstances A complete evaluation was conducted, including post-layout simulations, eye diagram measurement, jitter analysis and stress testing circumstances.

The Bit Error Rate (BER) analysis revealed a significant improvement with the application of adaptive equalization and channel optimization techniques. Due to insertion loss, inter-symbol interference and crosstalk effects over lengthy PCB traces and package interconnects better BER values were recorded during the initial PHY validation. However, the Bit Error Rate (BER) performance was significantly improved by tweaking the transmitter Feed Forward Equalization (FFE) and the receiver side Continuous Time Linear Equalization (CTLE) and Decision Feedback Equalization (DFE). The communication channels were still within the error thresholds allowed by the PCI-SIG and CXL compliance criteria.

Also, the signal conditioning procedures were proved using the eye diagram analysis. The preliminary eye patterns showed lower eye opening and temporal uncertainty under peak bandwidth conditions. The eye diagrams got better in the vertical and horizontal margins after equalization tuning and clock recovery optimization, showing better signal clarity and timing stability. Our investigation of PAM4 signaling for PCIe Gen6 demonstrated the viability of FEC-based techniques to overcome the reduction in voltage margins associated with multi-level signaling schemes.

The results of the channel-loss compensation showed that the adaptive equalization approaches may reduce the insertion loss and frequency-dependent attenuation in high-speed communication channels. Dynamic receiver adaptation allowed stable operation under diverse process, voltage and temperature settings and hence permitted reliable signal recovery.

Additional reliability tests in stress circumstances proved the stability of the suggested architecture. Persistent AI inference traffic, concurrent memory transactions and high heat conditions with longer execution of workload did not reveal any noticeable protocol failures or coherence violations. Thermal-aware traffic balancing and adaptive clock management methods are proposed to increase stability for lengthy durations of high-performance operation. Signal integrity and reliability were studied to validate the suggested PCIe/CXL integration methodology for advanced ASIC design.

5.3. Verification Outcomes

The proposed verification methodology provided significant functional and protocol coverage for PCIe Gen5/Gen6 and CXL subsystems. The UVM-based verification environment was used to perform extensive constrained-random verification scenarios to evaluate the transaction handling, coherency management, error recovery and protocol compliance in various operational contexts. Functional coverage was higher than existing project objectives, reaching over 95% of protocol features over PHY, Link, Transaction and coherency layers.

The bug classification study showed many types of implementation problems at initial verification stages like packet order violations, replay buffer sync problems, clock domain crossing failures and coherency mismatches between PCIe and CXL modules. Assertion based verification (ABV) methods have been very useful for finding timing violations and protocol errors that are difficult to discover using simulation alone. Most of the reported vulnerabilities were addressed by design improvements, synchronization improvements and better transaction arbitration.

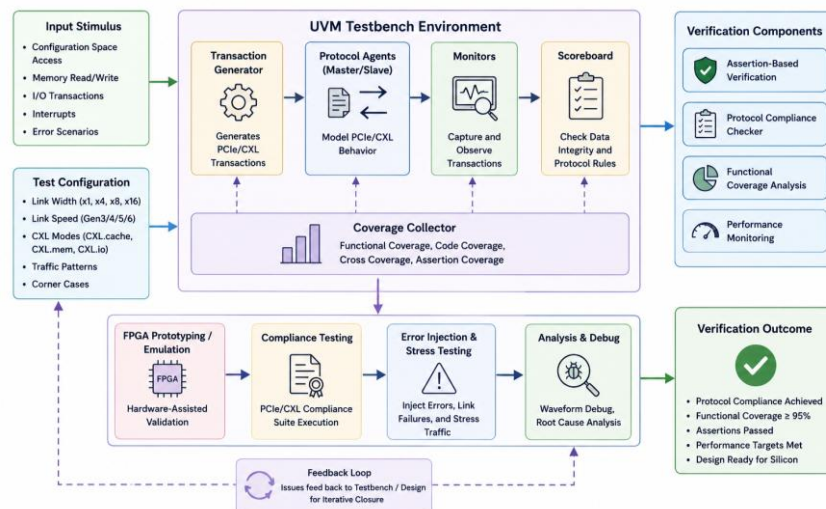


Fig 2 - UVM-Based PCIe/CXL Verification and Validation Flow

The compliance testing demonstrated robust interoperability performance with multiple endpoint devices and protocol configurations. PCI-SIG and CXL Consortium compliance suites validated proper implementation of flow control behavior, link training sequences, error handling mechanisms and coherency procedures. “FPGA prototyping allowed us to run real-world AI

workloads successfully before tape-out, significantly reducing post-silicon debugging and increasing confidence in the final silicon.” The end-to-end validation flow has proven the effectiveness of the proposed methodology in achieving a robust and scalable PCIe/CXL ASIC integration.

6. CONCLUSION AND FUTURE SCOPE

6.1. Conclusion

The rapid development of AI, cloud computing, high-performance computing and data-centric applications has made it increasingly important to integrate PCIe Gen5/Gen6 and Compute Express Link (CXL) technologies into modern ASIC systems. These technologies provide the required high bandwidth, low latency and scaling coherence mechanisms for modern systems of heterogeneous computing. Advanced node ASICs that use high speed interconnect protocols have many technology challenges in the areas of signal integrity, protocol complexity, power efficiency, thermal management and large-scale verification.

The paper provided a detailed way to overcome the huge difficulties of integrating PCIe and CXL into future-generation ASICs. The suggested framework covers scalable system architecture design, high-speed PHY optimization, improved protocol integration, robust verification methodology and performance-oriented optimization algorithms. Special emphasis was given to multi-Gigabit signal integrity challenges, adaptive equalization techniques, clock synchronization strategies, coherency management and protocol interoperability. The testing process also incorporated UVM based testing, assertion-based verification, compliance validation and FPGA prototyping to boost the functional coverage and to reduce the post-silicon risks.

The AI accelerator case study showed the application of the proposed approach to the state-of-the-art 5 nm ASIC design in the real-world scenario. The experimental results showed significant improvement in the throughput efficiency, latency reduction, bandwidth efficiency and communication reliability for AI-centric applications. The impact of channel impairments on high-speed communications was effectively alleviated through the adoption of adaptive equalization and advanced clock recovery algorithms, as evidenced by the research of signal integrity. The verification results illustrate the robustness of the integrated PCIe/CXL architecture with full protocol coverage and good compatibility with a wide range of device combinations.

The suggested technique provides a scalable and efficient framework to realize PCIe Gen5/Gen6 and CXL enabled ASICs for future heterogeneous computing systems. The work highlights the importance of cross-layer optimization, workload-aware validation, and coherency-centric architectural design for designing reliable high-performance semiconductor systems for future data center and AI applications.

6.2. Future Scope

As computing systems evolve towards more data-demanding and memory-centric architectures, future work in PCIe and CXL integration is expected to focus on a number of emerging

technologies and design approaches. One big area of growth will be the development of PCIe Gen7, which aims to improve on data transfer speeds even more than current Gen6 specs. Higher signaling rates will pose new challenges in signal integrity, power supply, equalization complexity and protocol reliability, and more sophisticated PHY and channel optimization techniques will be required.

Future CXL ecosystems are expected to become large-scale memory fabrics that enable dynamic memory pooling, composable architecture, and shared coherent memory across geographically distributed computing environments. These advanced memory architectures are going to be vital for increasing scalability and resource efficiency in AI data centers and cloud platforms. Hence, the study of efficient coherency management and low-latency memory access approaches will remain important.

Promising is the use of AI to perform verification. In complex verification scenarios, machine learning techniques can be used instead of manual coverage optimization, bug prediction, anomaly detection, and production of stress scenarios. Sophisticated verification approaches can greatly shorten the development cycle of future high-speed ASICs and increase the efficiency of bug identification.

Chiplet-based architectures are seen as a feasible replacement for traditional monolithic ASIC designs, more and more. Future PCIe and CXL systems will likely rely more and more on chiplet integration to improve scalability, reduce manufacturing cost, and enable adaptable heterogeneous computing platforms. Also, 3D IC integration and advanced packaging technologies are expected to be increasingly important to improve interconnect density and bandwidth efficiency.

Finally, optical connectivity technologies can represent a sustainable alternative to overcome the electrical limitations of traditional copper-based signaling at very high data rates. Incorporating optical communication interfaces in future ASIC platforms could considerably boost bandwidth scalability, reduce power consumption and lead to next-generation exascale computing systems.

REFERENCES

- [1] Lin, Yen, et al. "A Review of PCI Express Protocol-Based Systems in Response to 5G Application Demand." *Electronics* 11.5 (2022): 678.
- [2] Parakala, A. (2024). Self-Learning Bots & Cloud-Native Platforms. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(4), 132-141. <https://doi.org/10.63282/3050-9246.IJETCSIT-V5I4P114>
- [3] Guo, Yunyan, and Guoliang Li. "A CXL-powered database system: Opportunities and challenges." *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024.
- [4] Suryadevara, S. S. K., & Nakirikanti, S. (2024). Blockchain-Backed Content Authenticity Verification Framework. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(1), 242-252. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I1P125>
- [5] Takkalapally, D., & Takkellapally, M. R. (2024). AI-SynPerf: Synthetic Data Intelligence Framework for 5G Mobile Performance Simulation. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 182-194. <https://doi.org/10.63282/3050-9246.IJETCSIT-V5I1P118>

- [6] Wang, Jiapin, et al. "LMB: Augmenting PCIe Devices with CXL-Linked Memory Buffer." *arXiv preprint arXiv:2406.02039* (2024).
- [7] Vppalapati, M. (2024). Cooling Domains as First-Class Failure Boundaries in Storage Architecture. *American International Journal of Computer Science and Technology*, 6(2), 96-106. <https://doi.org/10.63282/3117-5481/AIJCSST-V6I2P110>
- [8] Shiramalla, R. (2023). Optimizing Cross-Platform Enterprise Integrations Using Workato: A Case Study of Salesforce and Oracle SaaS Applications. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(1), 232-243. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P124>
- [9] Ko, Hyungjong. "High-speed serial link trend and technical challenge." *2022 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. IEEE, 2022.
- [10] Muppaneni, R. K. (2023). AI-Driven Forecasting in Dynamics 365 Sales: What Businesses Need to Know. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 168-176. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I1P117>
- [11] Srigadde, B. R., & Devaraju, J. M. (2024). Building a Reusable AI Connection Utility Class. *International Journal of Emerging Research in Engineering and Technology*, 5(2), 188-200. <https://doi.org/10.63282/3050-922X.IJERET-V5I2P119>
- [12] Phelps, Boyd, and Arif Khan. "Disaggregated designs: Technology challenges and enablers." *IEEE Micro* 45.1 (2024): 9-15.
- [13] Allenki, S. S. (2024). Building Scalable Data Replication Pipelines for Real-Time Analytics. *American International Journal of Computer Science and Technology*, 6(1), 71-81. <https://doi.org/10.63282/3117-5481/AIJCSST-V6I1P108>
- [14] Muppaneni, K., & Vejella, M. (2023). Security and Data Privacy in Redux Stores. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(4), 153-162. <https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P117>
- [15] Ji, Houxiang, et al. "Demystifying a cxl type-2 device: A heterogeneous cooperative computing perspective." *2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2024.
- [16] Gaddam, R. R. (2021). Vertex AI as a Unified Control Plane for MLOps. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(2), 92-102. <https://doi.org/10.63282/3050-9262.IJAIDSML-V2I2P110>
- [17] An, Yuda, et al. "A novel extensible simulation framework for cxl-enabled systems." *arXiv preprint arXiv:2411.08312* (2024).
- [18] Vppalapati, M. (2024). Power-Bound Storage Design: Architecting Systems for Electrical Scarcity. *International Journal of AI, BigData, Computational and Management Studies*, 5(1), 208-217. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I1P121>
- [19] Katangoori, S. (2024). Jupyter Notebooks as First-Class Citizens in Cloud-Native Data Workflows. *American International Journal of Computer Science and Technology*, 6(3), 127-138. <https://doi.org/10.63282/3117-5481/AIJCSST-V6I3P110>
- [20] Lin, Xiancheng, et al. "Research on the CXL Memory." *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. Vol. 6. IEEE, 2024.
- [21] Shiramalla, R. (2024). Secure Multi-Cloud API Orchestration between Salesforce, Oracle CPQ, and Azure. *American International Journal of Computer Science and Technology*, 6(3), 102-113. <https://doi.org/10.63282/3117-5481/AIJCSST-V6I3P108>
- [22] Kumar Doodala, A. N. (2024). Validating UX consistency Across Omnichannel Platform. *American International Journal of Computer Science and Technology*, 6(6), 87-97. <https://doi.org/10.63282/3117-5481/AIJCSST-V6I6P109>
- [23] Wang, Chenjiu, et al. "Cxl over ethernet: A novel fpga-based memory disaggregation design in data centers." *2023 IEEE 31st Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2023.
- [24] Gaddam, R. R. (2021). Hermetic ML Environments using Conda-Lock and Docker. *American International Journal of Computer Science and Technology*, 3(4), 22-34. <https://doi.org/10.63282/3117-5481/AIJCSST-V3I4P103>
- [25] Muppaneni, R. K. (2023). Low-Code Revolution: How Power Platform Extends Dynamics 365 Capabilities. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(3), 162-171. <https://doi.org/10.63282/3050-9262.IJAIDSML-V4I3P119>
- [26] Kommuru, Madhurima, and Appala Nooka Kumar Doodala. "Performance Bottlenecks Necks in Data Heavy Python Applications." *International Journal of Applied Data Science & Modern Computing* 6.2 (2023): 01-19.
- [27] Baumstark, Alexander, et al. "So far and yet so near-accelerating distributed joins with CXL." *Proceedings of the 20th International Workshop on Data Management on New Hardware*. 2024.

- [28] Suryadevara, S. S. K. (2024). Resilient Multi-CDN Delivery Model Using AI-Based Traffic Switching for Global AEM Deployments. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(3), 191-200. <https://doi.org/10.63282/3050-9246.IJETCSIT-V5I3P119>
- [29] Muppaneni, K. (2023). Virtual DOM vs Real DOM: Performance Benchmarks. *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 180-189. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P118>
- [30] Vanapalli, Satyendra Kumar. "Identifying Operational Inefficiencies through CRM Data Analysis." *International Journal of Data Engineering and Intelligent Computing* 7.2 (2024): 01-13.
- [31] Mohamed, Basma H., et al. "A novel architecture of CXL protocol data link layer for low latency memory access." *2023 International Conference on Microelectronics (ICM)*. IEEE, 2023.
- [32] Allenki, S. S. (2024). Automating Backups and Recovery: Reducing Manual Work by Over 50%. *International Journal of Emerging Research in Engineering and Technology*, 5(1), 166-176. <https://doi.org/10.63282/3050-922X.IJERET-V5I1P119>
- [33] Takkalapally, D. (2024). ShiftLeft-AI: Machine Learning Framework for Proactive Performance Assurance in CI/CD Pipelines. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(4), 285-296. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I4P126>
- [34] Mao, Shunyu, et al. "CXL-Interference: Analysis and Characterization in Modern Computer Systems." *arXiv preprint arXiv:2411.18308* (2024).
- [35] Katangoori, S. (2024). JupyterOps: Version-Controlled, Automated, and Scalable Notebooks for Enterprise ML Collaboration. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(3), 211-223. <https://doi.org/10.63282/3050-9246.IJETCSIT-V5I3P122>
- [36] Srigadde, B. R. (2024). Agents, LLMs, and Salesforce with Multi-Cloud Provider (MCP). *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(3), 277-288. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P127>
- [37] Hou, Wentao, et al. "Understanding Routable {PCIe} Performance for Composable Infrastructures." *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. 2024.
- [38] Vanapalli, Satyendra Kumar. "Business Process Reengineering Using CRM Workflow Automation." *International Journal of Intelligent Automation & Robotics Engineering* 7.2 (2024): 01-17.
- [39] Kumar Doodala, A. N. (2024). Service Virtualization for API-First development: A shift-Left Testing Strategy. *American International Journal of Computer Science and Technology*, 6(4), 50-58. <https://doi.org/10.63282/3117-5481/AIJCSIT-V6I4P105>
- [40] Parakala, A. (2023). Citizen-Facing Automation: Chatbots and Self-Service in Public Services. *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 108-118. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P112>
- [41] Lee, Junseung, and Jungrae Kim. "Synergizing cxl with unified memory for scalable gpu memory expansion." *2024 International Conference on Electronics, Information, and Communication (ICEIC)*. IEEE, 2024.
- [42] Kommuru, M. (2024). Assessing the Limitations of LangChain in Production Environments. *American International Journal of Computer Science and Technology*, 6(4), 71-83. <https://doi.org/10.63282/3117-5481/AIJCSIT-V6I4P107>
- [43] Choi, Dong-Myung, et al. "A 4.6 pJ/b 64Gb/s Transceiver Enabling PCIe 6.0 and CXL 3.0 in Intel 3 CMOS Technology." *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2024.
- [44] Taluri, R. (2024). A Cloud-Native Reference Architecture for Data Engineering, Generative AI, and Decision Intelligence Using AWS and Amazon Bedrock. *International Journal of AI, BigData, Computational and Management Studies*, 5(1), 218-227. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I1P122>