OPTIMIZATION AND PERFORMANCE ANALYSIS OF THE COMMON READOUT UNIT FOR THE ALICE EXPERIMENT AT CERN

By

SHUAIB AHMAD KHAN

ENGG04201404004

Variable Energy Cyclotron Centre, Kolkata

A thesis Submitted to the
Board of Studies in Engineering Sciences

In partial fulfillment of requirements
for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

January, 2019

STATEMENT BY AUTHOR

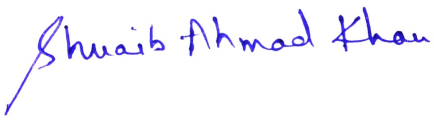
This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

(Shuaib Ahmad Khan)

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.



(Shuaib Ahmad Khan)

List of Publications

Refereed Journal Publications

1. "A potent approach for the development of FPGA based DAQ system for HEP experiments."

   Shuaib Ahmad Khan, Jubin Mitra, Erno David, Tivadar Kiss and Tapan Kumar Nayak.

   Journal of Instrumentation (2017) (an IOP and SISSA Journal)

   doi.org/10.1088/1748-0221/12/10/T10010 Vol. 12, 27 Oct. (2017).

2. "Optimization of multi-gigabit transceivers for high speed data communication links in HEP Experiments."

   Shuaib Ahmad Khan, Jubin Mitra, Tushar kanti Das, Tapan K. Nayak.

   Nuclear Instruments and Methods in Physics Research A: Accelerators, Spectrometers, Detectors and Associated Equipment.(2019)

   https://doi.org/10.1016/j.nima.2019.02.030 Volume 927, May 2019, Pages 14-23

3. "Trigger and timing distributions using the TTC-PON and GBT bridge connection in ALICE for the LHC Run 3 Upgrade."

   Jubin Mitra, Erno David, Eduardo Mendez, **Shuaib Ahmad Khan**, Tivadar Kiss, Sophie Baron, Alex Kluge, Tapan Nayak.

   Nuclear Instruments and Methods in Physics Research A: Accelerators, Spectrometers, Detectors and Associated Equipment.(2018)

   https://doi.org/10.1016/j.nima.2018.12.076 Volume 922, April 2019, Pages 119-133

Proceedings Published in Refereed Journal:

1. "Development of a high speed data acquisition system for the detectors at high luminosity LHC."

   Shuaib Ahmad Khan, Jubin Mitra and Tapan K Nayak.

   Published in Springer Proceedings in Physics, vol 203. Springer, Cham; 2018 page (223-226).

   https://doi.org/10.1007/978-3-319-73171-1_50

2. "GBT Link testing and performance measurement on PCIe40 and AMC40 custom design FPGA boards."

   Jubin Mitra, **Shuaib Ahmad Khan**, Manoel Barros Marin, Jean-Pierre Cachemiche, Erno David, Frederic Hachon, Frederic Rethore, Tivadar Kiss, Sophie Baron, Alex Kluge and Tapan K. Nayak.

   Published in Journal of Instrumentation, Volume 11, March 2016; IOP Publishing Ltd and Sissa Medialab srl.

   doi.org/10.1088/1748-0221/11/03/C03039

3. "Common Readout Unit (CRU) - A new readout architecture for the ALICE experiment."

   J. Mitra, **S.A. Khan**, S. Mukherjee and R. Paul.

   Published in Journal of Instrumentation, Volume 11, March 2016; IOP Publishing Ltd and Sissa Medialab srl.

   doi.org/10.1088/1748-0221/11/03/C03021

Conference Proceedings

1. "Development status of Common Readout Unit at India for the ALICE detector at CERN.", **Shuaib Ahmad Khan** and Subhasis Chattopadhyay, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 63 (2018) page (1166-1167).**

2. "Implementation of I2C bus master controller for CRU Slow Control in ALICE at LHC.", **Shuaib Ahmad Khan**, Filippo Costa, Erno David, Jubin Mitra, Tivadar Kiss, S. Mukherjee, Rourab Paul, Tushar K. Das, A. Chakrabarti, Tapan K. Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 61 (2016) page (1070-1071).**

3. "GBT link testing and performance measurement on Altera Stratix-V FPGA.", **Shuaib Ahmad Khan**, Jubin Mitra, and Tapan Kumar Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 60 (2015) page 1052-1053**

4. "Common Readout Unit (CRU) - A New Readout Architecture for the ALICE experiment at the CERN-LHC.", **Shuaib Ahmad Khan**, Jubin Mitra, Tapan Kumar Nayak, Tivadar Kiss, Erno David, and Alexander Kluge, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 59 (DEC 08-12, 2014) page 972-973**

5. "Design and Simulation of 10 Gigabit Ethernet on Altera FPGA.", **Shuaib Ahmad Khan**, Jubin Mitra, Erno David, Tivadar Kiss, and Tapan Kumar Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 59 (DEC 08-12, 2014) page 974-975**

6. "Testing and Performance Analysis of 10 Gigabit Ethernet.", **Shuaib Ahmad Khan**, Jubin Mitra, Erno David, Tivadar Kiss, and Tapan Kumar Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 59 (DEC 08-12, 2014) page 884-885**

7. "Common Readout System in ALICE.", J. Mitra, **S.A. Khan**, **7th International**

Conference on Physics and Astrophysics of Quark Gluon Plasma 1-5 February, 2015, Published in PoS ICPAQGP2015 (2017) pp.098 .

8. "Error Resilient Secure Multi-gigabit Optical Link Design for High Energy Physics Experiment.", Jubin Mitra, **Shuaib Ahmad Khan**, Rourab Paul, Sanjoy Mukherjee, Amlan Chakrabarti, Tapan Kumar Nayak, **29th International Conference on VLSI Design and 15th International Conference on Embedded Systems (VLSID) 2016; Published in IEEE xplore 2016, page (472-473).** DOI: 10.1109/VLSID.2016.108

9. "An efficient approach to manage DMA descriptors and evaluate PCIe based DMA performance for ALICE Common Readout Unit(CRU).", S. Mukherjee, F. Costa, R. Paul, A. Chakrabarti, **S. A. Khan**, J. Mitra, T. Nayak, **Advanced Detectors for Nuclear, High Energy and Astroparticle Physics 2017, Published in Springer Proceedings in Physics, vol 201. Springer, Singapore; 2018 page (107-118).** doi.org/10.1007/978-981-10-7665-7_10

10. "Implementation and evaluation of custom logic for various configuration schemes based on I2C and HDLC protocol for ALICE Common Readout Unit(CRU).", S. Mukherjee, F. Costa, R. Paul, A. Chakrabarti, **S. A. Khan**, J. Mitra, T. Nayak, **Advanced Detectors for Nuclear, High Energy and Astroparticle Physics 2017, Published in Springer Proceedings in Physics, vol 201. Springer, Singapore; 2018 page (217-222).** doi.org/10.1007/978-981-10-7665-7_23

11. "Channel Processor in 2D Cluster Finding Algorithm for High Energy Physics Application.", Rourab Paul, Amlan Chakrabarti, Jubin Mitra, **Shuaib A. Khan**, Sanjoy Mukherjee, and Tapan Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 61 (2016) page (1046-1047)**

12. "An efficient approach to evaluate PCIe DMA design and DMA performance for Com-

mon Readout Unit(CRU).", S. Mukherjee, Filippo Costa, Rourab Paul, A. Chakrabarti, **S. A. Khan**, J. Mitra, T. Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 61 (2016) page (1078-1079)**

13. "Electronic Data Aggregation Architecture for High Energy Physics Big data taking Experiments.", Jubin Mitra, Swagata Mandal, **Shuaib Ahmad Khan**, Jogender Saini, Subhashis Chattopadhyay, Tapan Kumar Nayak, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 61 (2016) page (1086-1087)**

14. "Optimization methodology of high speed transceivers for interfaces in HEP Experiments.", **Shuaib Ahmad Khan** and Subhasis Chattopadhyay, **Proceedings of the DAE-BRNS Symp. on Nucl. Phys. 64 (2019) (In Press).**

15. "Unfolding design strategies and the critical aspects for development of DAQ at the HEP experiments.", Fahad khan, Eram Taslima, Atiya Fatima Usmani, Jubin Mitra, **Shuaib Ahmad Khan**, **International Conference On Cutting-edge Technologies in Engineering 2019, IEEE Conference Record No. 47290, IEEE Digital Explore. (In Press)**

16. "Design of FPGA based phase reconfiguration technique.", Atiya Fatima Usmani, Eram Taslima, **Shuaib Ahmad Khan**, **3rd International Conference on Electronics, Materials Engineering & Nano-Technology 2019, IEEE Digital Explore. (In Press)**

Internal Notes

1. GBT BER measurement and Transceiver Optimization. (Internal Note)
   Jubin Mitra, **Shuaib Ahmad Khan**
   twiki.cern.ch/twiki/pub/ALICE/CruFwGbt/GBT_BER_REPORT.pdf

2. GBT-FPGA Firmware Debug Report for Altera Stratix V. (Internal Note)
   Jubin Mitra, **Shuaib Ahmad Khan**

twiki.cern.ch/twiki/pub/ALICE/CruFwGbt/GBT_Report_with_appendix_NEW.pdf

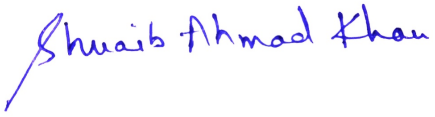
3. GBT Implementation in Stratix-V FPGA on AMC40 board. (Internal Note)

   Jubin Mitra, **Shuaib Ahmad Khan**

   twiki.cern.ch/twiki/pub/ALICE/CruFwGbt/gbt_work_summary_Nov2014.pdf

4. Receiving and HW Testing of the PCIe40_v1 CRU Cards. (ALICE Internal Note)

   T. Kiss, E. David, F. Costa, V. Baid, **S.A. Khan** (ALICE), and K. Arnaud, J.-P. Cachemiche, P.-Y. Duval, F. Hachon, M. Jevaud, R. Le Gac and F. Rethore (LHCb)



(Shuaib Ahmad Khan)

In the name of Allah

The Most Gracious and the Most Merciful

All Glory belongs to Him, He is The Most Exalted and The Most High

ACKNOWLEDGEMENTS

My deep gratitude goes to **(Prof.) Dr. Tapan Kumar Nayak**, for his valuable guidance, supervision and constant encouragement through my doctoral research. His broad vision of knowledge, unwavering enthusiasm for physics and rising basic questions in the subject helped me to think progressively in this field. Working with him is a pleasant experience and nutured me to become a better human being.

I would like to thank the deepest appreciation to my PhD committee members for their valuable suggestions and critical comments provided during the annual progress review meetings. I am particulary grateful to **Dr. Subhasis Chattopadhyay** for his continuous support to continue my research work.

My sincere appreciation extends to my technical advisor **Dr. Alexander Kluge** and CRU technical co-ordinator **Tivadar Kiss** for their mentoring, encouragement and early insights into the research work. I tender my grateful thanks to the VECC colleagues **Dr. Jubin Mitra**, **Dr. Swagata Mandal**, **Vinod Singh Negi**, **Vikas Singhal**, **Jogender Saini**, **Rama Narayana Singaraju**, **Partha Bhaskar**, **Dr. Zubayer Ahammed**, **Dr. Anand Dubey** and **Tushar K. Das**, for nurturing my research acumen. I am thankful to **Dr. Anuraag Misra** and **Dr. Saurabh Srivastava** from VECC, **Sanjoy Mukherjee**, **Dr. Supriya Das** and **Dr. Sidharth K. Prasad** from Bose Institute, **Dr. Rourab Paul** and **Dr. Amlan Chakraborty** from University of Kolkata for numerous engineering and technical suggestions. I thank to all the staff members from VECC for their time, effort and encouragement from them. I am extremely grateful to my entire CRU team members due to whom this journey is made possible. As part of my research, I have worked in multiple laboratories, without whom my research work would not have matured. From CERN electronics team I would like to thank **Dr. Sophie Baron**; from CPPM lab **Jean Pierre Cachemiche** and his team members; from Wigner Research Institute **Erno David**, **Josef Imrek** and **Dr. Filippo Costa** from CERN DAQ team.

A special thanks to my parents for all their sacrifices and prayers that made me sustained thus far. I would also like to thank all of my friends who supported me in writing, and giving impetus to strive towards my goal.

CONTENTS

Contents **Page No.**

CONTENTS **xiii**

SUMMARY **xvii**

LIST OF FIGURES **xix**

LIST OF TABLES **xxiv**

Chapter 1: Introduction **1**

1.1 Journey to high speed DAQ systems 3

1.2 The Large Hadron Collider (LHC) at CERN 5

1.3 The ALICE experiment at LHC 9

1.4 Research Motivation 10

1.5 Scope of the thesis 14

1.6 Structure of the thesis 14

Chapter 2: The ALICE experiment and its upgrade **17**

2.1 Introduction 17

2.2 The ALICE detector 18

2.2.1 The present ALICE detector: before Long Shutdown-2 18

2.2.2 ALICE Data Acquisition and computing for Run1 and Run2 29

2.3 ALICE upgrade overview: Upgrade of sub-detectors and their readout electronics 31

2.4 Upgrade of Data Acquisition methodology 34

2.4.1 Trigger system 35

2.4.2 Detector Data links . . . 37
2.4.3 Detector Data Readout Cards in ALICE . . . 37
2.5 Common Readout Unit (CRU) . . . 38
2.5.1 ALICE Readout scheme using CRU . . . 40
2.6 Summary . . . 44
Chapter 3: ALICE readout system and Common Readout Unit (CRU) 45
3.1 Introduction . . . 45
3.2 CRU design strategy . . . 47
3.2.1 Version A: CRU in the Cavern . . . 48
3.2.2 Version B: CRU in the Counting Room . . . 49
3.3 Features and functionalities of CRU . . . 50
3.4 Interfaces of CRU: Different Interfaces used, its selection and survey . . . 53
3.4.1 Detector side interface . . . 53
3.4.2 Server side Interface (aka DDL3) . . . 55
3.4.3 Trigger interface . . . 57
3.5 Selection of FPGA and the CRU hardware . . . 58
3.6 CRU peculiarites, complexities and the resolution . . . 64
3.6.1 Design requirements . . . 64
3.6.2 Hardware Complexity . . . 65
3.7 Summary . . . 68
Chapter 4: Link characterization and the signal integrity analysis 71
4.1 Introduction . . . 71
4.2 Readout architecture and the interfaces . . . 72
4.3 Front End Interface: GBT . . . 74
4.3.1 Design implementation on FPGA . . . 76
4.3.2 Calibration logic . . . 79
4.3.3 GBT Qsys Model . . . 88

4.3.4 FPGA Resource utilization . . . 88
4.3.5 Power Consumption . . . 89
4.3.6 Latency Measurement . . . 90
4.3.7 Eye diagram measurement . . . 92
4.3.8 BER analysis . . . 93
4.4 Back End Interface . . . 95
4.4.1 Communication over the PCI Express interface . . . 95
4.4.2 10-Gigabit Ethernet . . . 97
4.4.3 Performance Evaluation . . . 104
4.5 Avago MiniPOD performance tests . . . 108
4.6 Resource Estimation . . . 109
4.7 Summary . . . 109
Chapter 5: Optimization of multi-gigabit transceivers for high speed data communication links 113
5.1 Introduction . . . 113
5.2 Transceiver optimization . . . 114
5.2.1 Optimization Technique . . . 115
5.3 Test setup . . . 119
5.4 Methodology . . . 122
5.5 Results and discussion . . . 125
5.5.1 Eye Diagram analysis . . . 125
5.5.2 BER Results . . . 125
5.5.3 Improvement in Transmission . . . 129
5.6 Summary . . . 133
Chapter 6: CRU hardware development and tests 135
6.1 Introduction . . . 135
6.2 Development of CRU board . . . 136

6.3 Functional tests 140
6.3.1 Basic Electrical Tests 140
6.3.2 Assembly of the power mezzanine cards 145
6.3.3 Preparation for the Interface tests 146
6.3.4 Configuration of the Card for the first use 151
6.3.5 Programming of the Arria 10 FPGA 153
6.3.6 Hardware Tests 154
6.4 Summary 170
Chapter 7: Summary and Future Scope **171**
7.1 Summary 171
7.2 Future scope 176
REFERENCES **179**

Summary

The ALICE experiment at the CERN Large Hadron Collider is devoted to the research in heavy-ion physics, where the goals are to study the formation of Quark-Gluon Plasma (QGP), a de-confined matter consisting of quarks and gluons. To extend the physics reach and to understand the QGP matter in greater detail, ALICE is upgrading the detectors for data taking in the year of 2021, where the beam luminosity for Pb-Pb will increase six times to $6 \times 10^{27} cm^{-2} sec^{-1}$ at center-of-mass energy of 5.5 TeV. The increased interaction rates and the requirement of acquiring all the events information will result in an unprecedented dataflow of ~3 TB/sec from the detectors to the readout system. One of the major goals of the thesis is to design an efficient readout system to cope with the upsurge in data volume by acquiring data at a high rate and recovery from data error against the multi-bit upsets in radiation environments. A new FPGA based common readout unit (CRU) has been designed which acts as a coalesce between different interfaces and requires detector specific processing logic and firmware. The CRU receives data from the detector front-end electronics (FEE) boards located in the harsh radiation zone, performs online data processing and transfers to the back-end servers and storage located in the non-radiation areas.

As a part of the thesis work, optimization and performance analysis of the CRU in the context of the ALICE have been performed. The design aspects, principal tasks and complexities of the CRU have been discussed in detail. The prototype development of the CRU hardware has been illustrated and the detailed qualification tests are executed. The thesis presents performance analysis, evaluation of signal integrity and characterisation tests

on the high-speed interfaces. The measurements of resource utilisation, power consumption, critical path latencies, eye diagrams and bit error rate (BER) constitute the figures of merit for efficient system performance. Emphasis is given on the implementation and testing of the error-resilient 4.8 Gbps GBT links and its Qsys model for system integration is also proposed. Signal quality of the GBT core is characterised at the targeted BER of the order of 1 bit in 10^{12} bits. Total jitter is in the range of picoseconds only. The margin of receiver sensitivity is found to be 2.1 dBm for the two encoding schemes of GBT. An approach to handle the requisites for the testing, performance monitoring and parameter tuning of optical interconnects in FPGA-based systems is presented. A strategy is designed and developed for the latency-optimized implementation of the link to align the phase of the clocks. CRUs are associated with high rates of data transmission. Hence, optimization methodology for multi-gigabit transceivers is designed and tested to address the challenges of the high-frequency losses during the data transfer. It is implemented on the state-of-the-art 20nm Arria-10 FPGA manufactured by Intel Inc. The setup has been validated for three available high-speed data transmission protocols, namely, GBT, Timing-Trigger and Control over passive optical networks (TTC-PON) and 10-Gbps link. The improvement in the signal integrity is gauged by two metrics, the BER and the eye diagram. It is observed that the technique improves the signal integrity and reduces the BER.

The research and development summarized in the thesis is of high relevance for the firmware calibration and the hardware alignment purposes. The work could be further extended to design a load prediction model for efficient data distribution scheme and to architect a dynamic switching topology for the sudden rise of data volume.

LIST OF FIGURES

No. **Title** **Page No.**

Figure 1.1 Scheme for the succession of particle accelerators at the CERN accelerator complex. The LHC ring is indicated along with four major experiments ATLAS, ALICE, CMS and LHC-b. 6

Figure 1.2 Lattice layout of the LHC machine. 7

Figure 2.1 ALICE DAQ architecture for Run2 and the interface to the HLT system. 31

Figure 2.2 Scheme of the Major Sub-detectors of ALICE approved for upgradation during the long shutdown2 32

Figure 2.3 Data processing capability in the online systems 39

Figure 2.4 A system block diagram of ALICE readout scheme. Config I : CRU is used as trigger distribution system and read-out processor. Config II : CRU is used as read-out processor. The trigger distribution is done from the CTP/LTU directly to the on-detector electronics. Config III : CRU is not used for the detectors do not upgrade their front-end electronics and use detector-specific read-out cards. 41

Figure 2.5 Flow of data in the ALICE from the detector front-end electronics up to the O2 system. (FLP: First Level Processor, EPN: Event Processing Node) 42

Figure 3.1 Implementation version A: CRU as FPGA boards in the cavern. . . . 48

Figure 3.2 Implementation version B: CRU as FPGA cards in the counting room. 49

Figure 3.3 PCIe40 - candidate board for CRU development 63

Figure 4.1 Different communication forms between the CRU, Trigger & FEE Link architecture with the GBT ecosystem. (Courtesy: Erno David) 73
Figure 4.2 Link architecture with the GBT ecosystem. 74
Figure 4.3 Block diagram of a GBT link in FPGA 75
Figure 4.4 GBT link encoding scheme 76
Figure 4.5 Clock phase compensation between PCS and PMA 78
Figure 4.6 Data format for GBT protocol. 78
Figure 4.7 Clock Distribution Scheme for the Stratix-V FPGA 79
Figure 4.8 Multiplication or Division of Clock frequency 80
Figure 4.9 Word clock and Frame clock phase mismatch 81
Figure 4.10 Logic for Phase Calculation. 83
Figure 4.11 Phase Calibration logic. 84
Figure 4.12 Synchronization Register Chain for Gearbox 85
Figure 4.13 State Machine for Phase Calibration 85
Figure 4.14 Temperature sensor IP core for Stratix-V 86
Figure 4.15 Showing the data stability vs temperature variation of two FPGAs when GBT operating in (a)Transmit Latency Optimized mode (b) All other modes. 87
Figure 4.16 GBT Qsys Model. 88
Figure 4.17 Screen shot of the GUI for the Qsys model during run. 89
Figure 4.18 Test Setup for the GBT latency measurement. 90
Figure 4.19 Test setup for BER measurement. 93
Figure 4.20 BER measurement for GBT Frame coding and GBT wide-Bus mode. 94
Figure 4.21 Overview of Software Stackup for CRU. 95
Figure 4.22 Overview of Software Stackup for CRU. 96
Figure 4.23 Position of 10GbE in OSI model. 98

Figure 4.24 (a)10GbE Intel MAC IP core block diagram (b) 10GBASE-R PHY with Hard PCS and PMA in Intel devices. 99

Figure 4.25 Test system implementation using Qsys system integration tool. . . . 100

Figure 4.26 Interface of Avalon-MM and Avalon-ST with source and sink SGDMA data transfer. 101

Figure 4.27 User Logic. 102

Figure 4.28 Simpified digital communication optical link 102

Figure 4.29 Three level of frequency translation in 10GbE communication. 105

Figure 4.30 MAC to XGMII data payload conversion scheme 106

Figure 4.31 Variation of Eye width and Eye Height with PRBS type. 106

Figure 4.32 Plot for tuning of transceiver parameter optimized settings at PRBS31. 107

Figure 4.33 Eye Diagram for the 10Gb Ethernet on FPGA. 107

Figure 4.34 Bit error rate as a function of received optical power at 10Gbps. . . . 108

Figure 4.35 MiniPOD^{TM} performance with 10 GbE protocol (10.312 Gbps). . . . 108

Figure 5.1 Voltage output differential (VOD) and tunable pre-emphasis taps with flexible polarity in the embedded transceiver of FPGA. 116

Figure 5.2 The pre-emphasis signal generation technique at the 1st post-tap in embedded FPGA transceivers, $0 < x < 1$ is the tap weight. 117

Figure 5.3 Pre-emphasis 2nd post-tap (Inverted) compared with pre-emphasis 1st post-tap and their effect on the signal without pre-emphasis. 118

Figure 5.4 Pre-emphasis 1st pre-tap and the 2nd pre-tap (Inverted) and their effect on the signal without pre-emphasis. 118

Figure 5.5 Arria-10 FPGA card inserted in PCIe x16 slot of server. The optical signal from the externally pluggable SFP+ is looped back via the fibre equipped with the variable optical attenuator (VOA). 119

Figure 5.6 Typical BER test loopback logic on FPGA using Qsys tool. PRBS patterns are generated. The serialised data is transmitted, looped back and checked for the flipped bits at the receiver. 120
Figure 5.7 Stepwise flow diagram for the Transceiver Optimization. Data transmission is started with the Intel default parameters and a Solution matrix is derived to achieve the optimized signal integrity 124
Figure 5.8 Changes in the Eye height and Eye width with PRBS variation for optical links at three line rates. 126
Figure 5.9 Time to achieve BER of 10^{-12} for the Line rate of GBT, TTC-PON and 10 Gbps optical links having different CL. 128
Figure 5.10 BER versus received optical power(dBm) for transceiver at Intel FPGA default settings for different PRBS operating in three line rates. 129
Figure 5.11 Eye diagram at the Intel FPGA default and at the Optimized settings of transceiver. 130
Figure 5.12 Multivariate kiviat diagram showing the solution space and the Intel FPGA default values for three different link rates. 131
Figure 5.13 Comparison of BER versus the received optical power for default and optimized transceiver settings separately for three line rates. 132

Figure 6.1 Showing an open circuit between layer 1 and layer 2. 137
Figure 6.2 Bare board (Left) and its X Ray image (Right). 137
Figure 6.3 (Left) FPGA after solder balling (1932 pins) and (Right) 2D X-Ray image of the BGA package after mounting on PCB. 138
Figure 6.4 Thermal profile characteristics using the Oven VP800-64. 139
Figure 6.5 Shows the bare ribbon cable assembly drawing. 139
Figure 6.6 Shows the myopic view of the First prototype of the CRU card. . . . 140
Figure 6.7 Map of test points for measurements of power rail shorts. 141
Figure 6.8 Measurement points for 3.3V_SEQ and 3.3V_MMC. 142

Figure 6.9 MEZZ15A_6A power mezzanine module. Identical schematic for Module-1, 2, 3 and 4. 142
Figure 6.10 MEZZ60A power mezzanine module-5, top side (left) and bottom side (right). 143
Figure 6.11 (Left) Test adapter for Module-1 to Module-4, (Right) Schematic guide for the test adapter. 144
Figure 6.12 (Left) Test adapter for Module-5, (Right) Schematic guide for the test adapter. 144
Figure 6.13 Mounting of the Power Mezzanine Cards (top side of PCIe40). 146
Figure 6.14 Overview of the Power Sequencer. 147
Figure 6.15 (Left) CAD drawing of the Face-Plate developed using Wire Grid machine, (Right) Faceplate installed on PCIe40. 148
Figure 6.16 (Left) Heat Sink for Arria-10 installed on the board, (Right) Side View showing the arrangement. 148
Figure 6.17 MiniPODs and the on-board flexible ribbon cables 149
Figure 6.18 MiniPOD mapping. 150
Figure 6.19 Ready card with all accessories. 150
Figure 6.20 Jumpers, Switches and Connectors Locations. 151
Figure 6.21 PCIe40 JTAG Programming Scheme. 152
Figure 6.22 Options for Programming of the Arria-10 FPGA. 153
Figure 6.23 Configuration of PLX-8747. 163

LIST OF TABLES

No. **Title** **Page No.**

Table 1.1 Journey to high speed DAQ system 4

Table 1.2 Timelines for LHC operation 9

Table 1.3 An Overview of ALICE statistics before and after the scheduled upgrade 12

Table 2.1 Trigger level in ALICE detectors 30

Table 3.1 Advantages of the Proposed approach over the Conventional approach 50

Table 3.2 CRU requirements in the system 52

Table 3.3 Comparison table between PCIe interface and 10 Gigabit Ethernet interface 56

Table 3.4 Detailed comparison of the specifications of the high speed interface links used in high-energy physics experiments. 58

Table 3.5 FPGA selection parameters 59

Table 3.6 FPGA device and its family with maximum SerDes speed 61

Table 3.7 Important Specifications of Arria-10 FPGA 62

Table 3.8 CRU Hardware Complexities 66

Table 4.1 CRU requirements in the system 75

Table 4.2 FPGA resource utilization for the GBT-FPGA reference design. 89

Table 4.3 Power consumption with the GBT Encoding Scheme. 90

Table 4.4 GBT link latency measurement. 91

Table 4.5 The comparison of the total path delay for AMC40 and PCIe40. . . . 92
Table 4.6 Showing the Eye Diagram for GBT encoded data. 92
Table 4.7 GBT Jitter Measurement for the two FPGAs. 93
Table 4.8 FPGA resource utilization for the 10GbE design. 104
Table 4.9 Latency estimation for data transfer(1 clock cycle = 156.25MHz). . . . 105
Table 4.10 Shows the comparison of MiniPOD^{TM} performance for GBT protocol and 10GbE Protocol 109
Table 4.11 An integrated design for elementary firmware including low level interfaces. 109

Table 5.1 Components used in the test setup, their role and specifications. . . . 120
Table 5.2 Transceiver parameters, range of operations for the manual optimization. 123
Table 5.3 Comparison of Optical power(dBm) to attain BER of 10^{-12} for the three high speed interface links. 133
Table 5.4 Comparison of optical power for CDR for the three high speed interface links. 133

Table 6.1 Voltage levels from the power mezzanine modules. 144
Table 6.2 Voltage values measured 147
Table 6.3 JTAG switch settings (S = source, T = target) 153
Table 6.4 PCIe end point detection 163
Table 6.5 PCIe end point detection 164

Chapter 1

Introduction

Nuclear and particle physics experiments at high energies, often referred to as High Energy Physics (HEP) experiments, study the constituents of matter and their fundamental interactions. By colliding proton on proton or heavy-ions, such as, Au on Au or Pb on Pb at relativistic energies; one reproduces conditions that are prevalent within a microsecond after the birth of our universe. The evolution of our knowledge of the fundamental particles, their interactions as well as connections to the early Universe, has been proportional to the evolution of the available beam energies in the particle accelerators. The collisions produce zillion of highly energetic particles which are to be recorded by the experiments. The increase of collision energy and beam interaction rates demand for sophisticated and hightech detectors, electronics and data acquisition (DAQ) systems. In addition, the radiation levels in the proximity of the detectors have also been growing, which calls for radiation tolerant systems. The readout electronics in the harshly radiated area are highly prone to damage due to the total dose, single event upsets and non-ionizing energy loss [1] depending on the type of radiation. This poses various challenges for particle detectors, readout electronics, and DAQ systems. To cope with it, DAQ systems should have the ability to support high data rate, error resiliency against the multi-bit upsets in radiation environments, efficient data aggregation and processing schemes, easy reconfigurable with quick upgradation and

compactness of hardware due to space constraints. The research and development presented in the thesis is based on the optimization and performance analysis of a completely new readout system, which will be capable of high data rate communication in context of the ALICE experiment at the Large Hadron Collider (LHC) at CERN.

CERN, the European Organization for Nuclear Research is the world's largest particle physics laboratory, founded in 1954 with twelve founder member states from Europe. Scientists from all over the globe contribute here to advance our knowledge about the past, present and future of the universe. CERN was one of the first European joint ventures. It has become a shining example of "*Physics for the benefit of mankind*"; collaborating with the nations that had once been against one another during the second world war. India has a special status of observer since 1991 and has been invited by CERN management to become an associate member of the organization in the year of 2016. Today, there are twenty-two member states (including India) and four associate members. India has played a significant role at CERN in terms of contribution for the building of Large Hadron Collider (LHC) as well as participation in the sophisticated experiments.

The LHC project is the latest and the largest particle accelerator in the world, commissioned in the year of 2008 [2]. It allows the scientists to comprehend the new chapters of the universe; explores the principle ingredients of matter that the universe was composed of at the time. At LHC, the two counter-rotating particle beams at ultrarelativistic energies, collide at four places known as interaction points. Complex detector systems are installed at each interaction point. A Large Ion Collider Experiment (ALICE) is one of the four major detectors at the LHC. It has been recording unprecedented amount of data for proton on proton and heavy ions since the beginning of the LHC program. ALICE is specifically designed for the study of heavy-ion collisions and the goal is to analyze the formation of Quark-Gluon Plasma (QGP). During the last ten years of its operation comprising of Run 1 and Run 2, ALICE has confirmed the formation of QGP and has given us a wealth of information regarding the system formed in the collisions.

In order to extend the physics reach and to understand the QGP matter in greater detail, ALICE is upgrading the detectors for data taking in Run 3 foreseen in 2021. During Run 3, the beam luminosity for Pb-Pb will increase by six times viz. $6 \times 10^{27} cm^{-2} sec^{-1}$ at center-of-mass energy of 5.5 TeV. The corresponding rate of interactions will also be increased from 8 KHz at present to 50 KHz. The high interaction rate and the requirement of acquiring all the events information will result in an unprecedented dataflow of ~3 TB/sec from the detectors to the readout system. This huge data will be experienced for the first time in the ALICE experiment. To handle the high volume of generated data an upgrade is required for the experiment. For this purpose; the upgrade of particle detectors, their readout electronics, and DAQ systems including data storage is planned. The upgrade approach for the data acquisition is based on a new Common Readout Unit (CRU) being developed. CRU for ALICE is the major thrust of the thesis. The thesis aims on the research and development for the challenging new readout system to handle high data rate. It also highlights the design of engineering solutions and methodologies developed during the R&D phase of CRU.

1.1 Journey to high speed DAQ systems

Traditional DAQ systems [3, 4] of the last century, could handle low data rate and less data errors against multi-bit upset in radiation zone. The modern DAQ systems for HEP and nuclear physics experiments is a result of continuous evolution [5]. The overview of the journey and the different methodologies adopted in the field of high speed DAQ are summarized in the Table 1.1.

In the early two decades of 1960-1980 the DAQ issues had been acknowledged by custom designed readouts that were framed after the characteristics of the individual experiments. Detectors have several tens to few hundred of readout channels with a readout rate of the order of megabyte per second only. They use non-standard interconnects. The introduction of Nuclear Instrumentation Module (NIM) standard and a modular computer-controlled

Table 1.1: Journey to high speed DAQ system

Parameter	**1960-1980**	**1980-2000**	**2000 onwards**
No. of Readout Channels	~100s	~ $10^3 - 10^6$	~ $10^6 - 10^9$
Data Rate	~1 MB/sec	~1 GB/sec	~10 GB/sec to few TB/sec
Readout Standard	Front End Electronics Non standardized	Parallelism feature of distributed computing	Heterogeneous Computing
Technology Evolution (Year Wise)	**1964: NIM standard** (backplane bus not defined) **1969: CAMAC based** centralized backplane, but lacked parallelism, BW limited to 1 MB/sec, **1970-1980:** NIM based Front End read by minicomputer and CAMAC readout bus	**1986: FastBus** BW: 40-60 MB/sec Support parallelism **1982-1987: VME** development with microprocessors. BW: 40 MB/sec **1990:** NIM, CAMAC, Fastbus and VME coexisted **1997**: VME320 with BW: 320 MB/sec	Point to point High speed links **2003:** PC based computing farms with Ethernet and PCIe bus **Present:** upto 400Gbps Ethernet, PCIe 5.0 specifications released in June 2017. Boosting on-board and local processing with FPGAs
Example System	Experiments at TRIUMF, BNL	Experiments at SPS, LEP at CERN	Experiments at CERN LHC

bus named Computer Automated Measurement and Control (CAMAC) focused on the standardisation of front end and back end respectively. However this lacked parallelism with limited data rate and channel count. In the decades of 1980-2000, Fast-Bus standard supported parallelism but the advent of microprocessors leads to the Versa Module European (VME) standard and VME Inter-Crate bus specifications (VIC). The NIM, CAMAC, Fast-Bus and VME coexisted to address the challenges of HEP experiments. In the current century point to point high speed links are evolved, specifications for Ethernet and Peripheral Component Interconnect Express (PCIe) protocol are continuously expanded. With the advent of highly dense latest Field Programmable Gate Arrays (FPGAs), we are heading towards the data rates of Terabytes/sec with high channel count and on-board and local processing.

One of the examples of the demand for modern DAQ systems is at CERN's LHC, operate detectors up to billions of electronics channels. An efficient system is required at the present time to handle the key issues in the design of DAQ for HEP experiments like collecting the

data at high rate and channel count with error resilience against the multi-bit upsets in the harsh radiation environment. Before proceeding to the details of the new readout system, it is worthwhile to outline the LHC at CERN, the requirements at the LHC and the design criteria.

1.2 The Large Hadron Collider (LHC) at CERN

CERN has always remained at the cutting edge in the accelerator, detector, electronics and computing technology along with an objective to pursue the nuclear research of scientific and fundamental character. At present, the CERN accelerator complex is a systematic progression of particle accelerators to attain an ultra-relativistic velocity with very high energy [6]. Each accelerator boosts the speed of particles in the beam, before injecting into the next one in the process. The final destination is the LHC as shown in Figure 1.1. It accelerates the particles at a record energy of 7 TeV. To consolidate the journey of CERN's accelerators; a 600 MeV synchrocyclotron was commissioned in the year of 1957. In 1971, the Intersecting Storage Ring (ISR) was commenced which is the world's first proton-proton collider. In the year of 1974, the Super Proton Synchrotron (SPS) was developed, caged in a tunnel of 7 km of circumference. SPS has been the dynamo of CERN's particle physics program and continued to deliver beams for the LHC. The notable highlights of the LHC machine, experiments and the subsystems are briefly summarized in the context of the present dissertation.

LHC is one of the most ambitious mega-science projects ever undertaken; established on the Franco-Swiss border at the Geneva country side. It is installed in a tunnel of 26.7 km of circumference situated up to a depth of 170 m below the Alps and Jura mountain ranges. LHC is a superconducting hadron accelerator and collider with two parallel aligned circular beamlines (or beam pipes). The beamlines contain particle beams counter-rotating in circular trajectories around the ring. The particle beams consist of lead nuclei, proton,

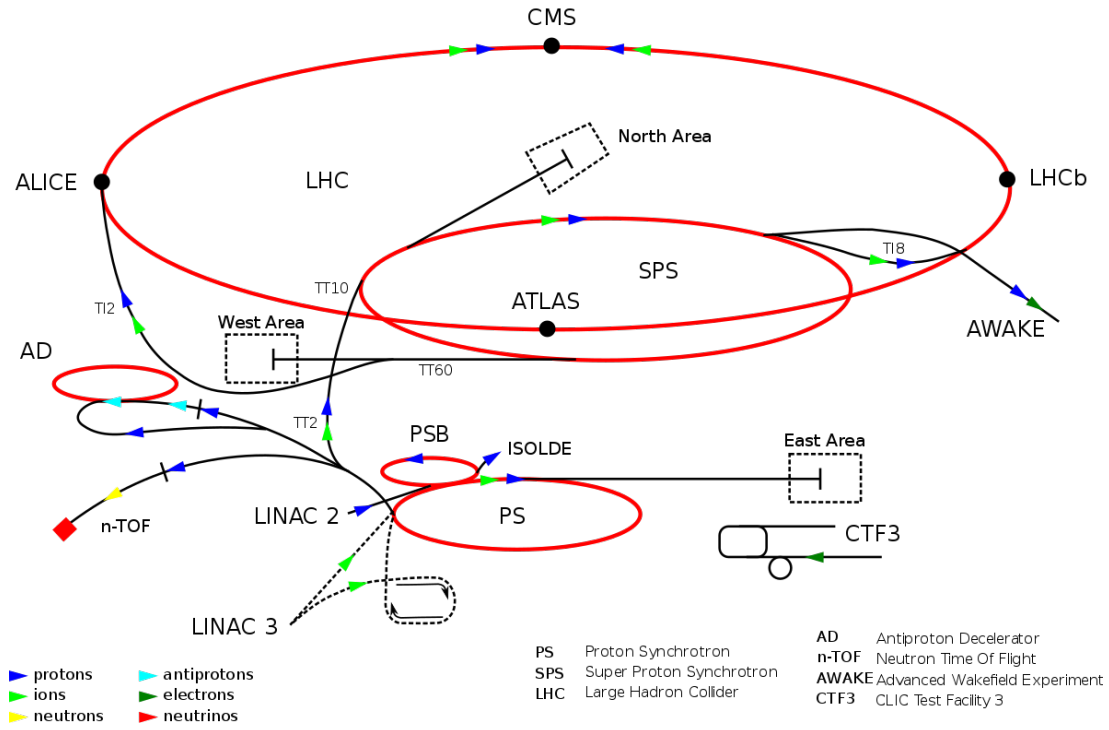


Figure 1.1: Scheme for the succession of particle accelerators at the CERN accelerator complex. The LHC ring is indicated along with four major experiments ATLAS, ALICE, CMS and LHC-b.

or proton and lead nuclei. Protons are obtained for the LHC by removing electrons from hydrogen atoms and Lead ions (Pb) from a source of vaporized lead.

A beam is formed of hundreds of particle bunches. A bunch is composed of about 10^{11} protons. The bunches in a beam cross every 25ns, this leads to around 600 million collisions each second. Most of the protons do not collide due to low cross section and continue to move. The beam size is squeezed down to 64 microns at the interaction points to enhance the probability of collision. The beams continue to circulate for hours; and the filling of beams in the LHC require several cycling time before it attains the full targeted energy. LHC uses magnetic fields to contain the particles in well-defined beams and electric fields to accelerate the charged particles up to almost the speed of light. The accelerated beams collide at four places known as interaction points where complex detector systems namely, A Toroidal LHC ApparatuS (ATLAS) [7], Compact Muon Solenoid (CMS) [8], A Large Ion Collider Experiment (ALICE) [9] and Large Hadron Collider Beauty (LHC-b) [10] are installed as

shown in Figure 1.2. Interactions at the crossing points are referred as events. Discovering

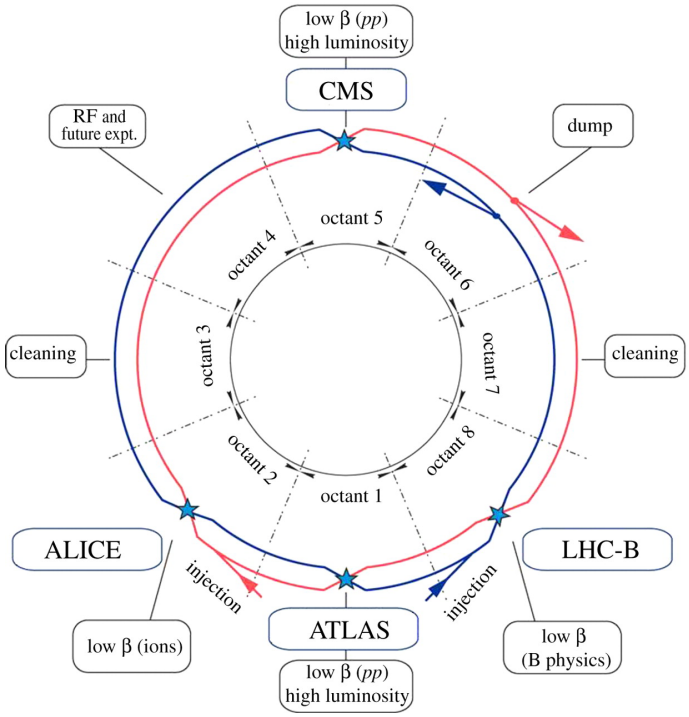


Figure 1.2: Lattice layout of the LHC machine.

the Higgs bosons has been the main goal behind building the LHC. In addition, search for the super-symmetric partners of known particles, verifying string theory predictions of extra-dimension, search and study of the primordial state of matter in the form of QGP, matter and anti-matter asymmetry in the universe and understanding dark matter and dark energy have been on the top of the LHC's agenda of physics topics. The ATLAS and CMS are two large experiments diametrically opposite in pits 1 and 5, respectively. These two proton-proton experiments are dedicated for new physics searches and the precision measurements. The ALICE in pit 2 is a dedicated heavy ion physics experiment to search and study the formation of QGP matter, a de-confined matter consisting of quarks and gluons. Whereas the LHC-b in pit 8 is dedicated for the search behind matter and anti-matter asymmetry and other rare phenomenon in the decay of *Beauty particles*.

In the high-energy accelerators like the LHC, particles reach very close to the speed of light. At the LHC; particles revolve at 0.999999991 times the speed of light at top energy of

7 TeV [11]. In these situations; the increase in the speed of the particle is minimal with the increase of energy, instead there is a gain of the relativistic mass. Therefore, particle physics is dealt in terms of the particle's energy, rather than its speed. In a collider experiment, the energy of the collision, known as centre-of-mass energy is the sum of the energies of two colliding particle beams. Interactions at the crossing points are referred as events. The number of events generated per second viz. interaction rate (R) in a LHC collision is proportional to the cross section σ_{event} for the event as shown in equation (1.1). The constant of proportionality is known as $Luminosity(L)$.

$$\mathcal{R} = L\sigma_{event} \tag{1.1}$$

The beam profile has a Gaussian distribution and the luminosity decays with the degradation in the circulating beams. Hence, the parameter, Integrated Luminosity is used to accumulate up the luminosity over a run. LHC is designed to collide proton beams with a centre of mass energy ($\sqrt{s}$) of 14 TeV (7 TeV each beam) and at an unprecedented luminosity of $10^{34} cm^{-2} s^{-1}$. It can also accelerate and collide heavy lead ions (Pb-ions) with an energy of 2.75 TeV per nucleon centre-of-mass energy ($\sqrt{s_{NN}}$) of 5.5 TeV at a peak luminosity of $10^{27} cm^{-2} s^{-1}$ [2].

The HEP experiments at CERN's LHC system are operative since the year 2009 to deduce the missing fragments in the current description of the fundamental structure of matter [12, 13]. The first phase (Run1) of data collection is concluded, and the next phase (Run2) will be completed by 2018. To sustain and coup the full discovery potential of the experiment [14]; a major progressive upgrade is planned for the third phase (Run3) of LHC running. Run3 will start from the year of 2021 followed by a High-Luminosity Large Hadron Collider (HL-LHC) in the year 2025. The LHC timelines are summarized in the Table 1.2. Under this road-map for the upgrade strategy; the LHC will progressively increase its luminosity and hence the collision rate. The integrated luminosity is aimed to

Table 1.2: Timelines for LHC operation

LHC Timelines			
Run1	2009 - 2012	LS1	2013 - 2014
Run2	2015 - 2018	LS2	2019 - 2020
Run3	2021 - 2023	LS3	2024 - 2025
Run4 (HL-LHC)	2025 - 2029	LS4	2030
and further			

increase by a factor of ten times more than the initial design value by the year 2030. These large event statistics will help for the analysis of the rare events.

The increase in the interaction rates due to the increase in the luminosity will result in a high volume of data flow. It is of the order of few TB per second from the detector systems to the Data Acquisition system. To handle the high data rate, there is a need for the upgrade of different detector subsystems, electronics and readout, and the data acquisition systems. There are different detectors at LHC viz. ATLAS, CMS, ALICE and LHC-b; each experiment has its own physics motivation and the behaviour of data flow [2]. Consequently, to tackle the upgrades, the experiments at LHC need efficient parallelized readout schemes with high-tech electronics and DAQ techniques to apprehend the unparalleled amount of produced data. The thesis is oriented towards the upgrade of ALICE experiment, focussing on the CRU. CRU helps to handle the high data rate and an integral part of the ALICE DAQ upgrade.

1.3 The ALICE experiment at LHC

A Large Ion Collider Experiment (ALICE) is one of the four major detectors at the LHC and devoted to the study of heavy-ion collisions. The goal is to address the physics of strongly interacting matter, in particular, the formation of QGP at extreme values of energy density and temperature in nucleus-nucleus collisions. ALICE will also address several topics of quantum chromodynamics (QCD); the theory of the strong interaction between quarks and

gluons, for which ALICE is complementary to other LHC detectors using the study of proton-proton collisions. In addition, the proton data will be used as a reference data for the heavy-ion-program. A detailed study of the hadrons, electrons, muons and photons produced in the collision of heavy nuclei are the main observables. The ALICE detector has been completed by a collaboration including over thousand physicists and engineers from several institutes. The experiment consists of various sub-detector systems arranged in cylindrical shells around the interaction point as well as in the forward directions. It is a challenge for the experiment to detect majority of the several thousands of particles produced in nuclear collisions with high precision. Different detector subsystems are installed in the experiment to identify different types of particles. The sub-detectors at ALICE are grouped into three sections viz. the central Barrel detectors, the muon spectrometer and the forward detectors. Each detector has its own design constraints and specific choice of technology motivated both by the physics aspects and the experimental conditions expected at LHC. Overall dimensions of ALICE are $16x16x26m^3$ with a total weight of approximately ten thousand tonnes.

During last ten years of operation of ALICE, two phase of data collection, Run-l and Run-2 are completed by 2018. By analysing data from a large number of collisions or events, ALICE has confirmed the formation of QGP and has given us a wealth of information regarding the system formed in the collisions. The third phase (Run3) of LHC running is foreseen to start from the year of 2021 with increased beam energy and luminosity after the shut down of two years scheduled from the year of 2019 [15].

1.4 Research Motivation

Before proceeding to the details of the new readout system, it is worthwhile to outline the requirements at the ALICE and the design criteria which motivated for the upgrade. During the LHC operations in Run-1 and Run-2 till the year 2018, it had provided proton-

proton collisions at maximum centre of mass energy of 13 TeV, Pb-Pb collisions at 5 TeV. In addition, LHC had proton on lead and Xenon on Xenon collisions. The Run 3 of LHC after Long Shutdown2 (LS2) will start in the year 2021 with increased beam energy (proton-proton at 14 TeV and Pb-Pb at 5.5 TeV) as well as increased beam luminosity. The luminosities for Pb-Pb collisions will be increased progressively from $1x10^{27}cm^{-2}s^{-1}$ to $6x10^{27}cm^{-2}s^{-1}$ at center-of-mass energy of 5.5 TeV. The corresponding rate of interactions will also be increased from 8 KHz at present to 50 KHz. To cope-up with the high data rates, ALICE experiment plans for a major upgrade during LS2. During this upgrade of the detectors, their readout electronics need to be upgraded, and a new data acquisition approach based on Common Readout Unit (CRU) is being developed [16]. Research and development of the challenging new readout system to handle high data rate is the major thrust of the thesis.

ALICE had collected $1nb^{-1}$ Pb-Pb collisions before LS2 at peak instantaneous luminosities of $1x10^{27}cm^{-2}s^{-1}$ analogous to the collision rate of 8 KHz. The maximum readout rate of the present ALICE detector is 500 Hz of Pb-Pb events. ALICE upgrade strategy is based on collecting full detector information i.e data from each collision. Data will be transferred to the DAQ system in a self-triggered mode or upon a minimum bias trigger. The specifications for the ALICE detector upgrade is set by the collision rate of 50 kHz for Pb-Pb and 200 kHz for pp and p-Pb collisions. ALICE will collect greater than $10nb^{-1}$ of Pb-Pb ion collisions at L upto $6x10^{27}cm^{-2}s^{-1}$ corresponding to the collision rates of 50 KHz. It will also collect $6pb^{-1}$ of p-p ion collisions at the equivalent Pb-Pb nucleon energy and $50nb^{-1}$ of p-Pb collisions, both at a levelled collision rate of 200 KHz. The running scenario of the ALICE detector before and after the LS2 are summarized in Table 1.3

Motivated by the past operation experiences and successful physics results, R&Ds for the ALICE upgrade have started. The physics of the upgrade is intended at precision measurement of the QGP, which will be accessible through measurements of heavy-flavour transport parameters, quarkonia down to zero transverse momentum and low mass di-leptons. These processes do not exhibit signatures that can be chosen by hardware triggers, they can only

Table 1.3: An Overview of ALICE statistics before and after the scheduled upgrade

Parameter		**ALICE (Run-2)**	**ALICE(Run-3)**
Luminosity (L)	(Pb-Pb)	$1x10^{27}cm^{-2}s^{-1}$	$6x10^{27}cm^{-2}s^{-1}$
	(p-p)	$10x10^{34}cm^{-2}s^{-1}$	$5x10^{34}cm^{-2}s^{-1}$
Collison Rate (R)		8 KHz (Pb-Pb)	50 KHz (Pb-Pb)
Integrated Luminosity		$1nb^{-1}$ (Pb-Pb)	$>10nb^{-1}$ (Pb-Pb) $>6pb^{-1}$ (p-p), $50pb^{-1}$ (p-pb)
Max Readout rate		50 kHz (Pb-Pb)	200 kHz (p-p and p-pb)

be collected by a zero bias (minimum bias) trigger. The upgrade strategy for data readout in ALICE is based on collecting all the data corresponding to all the collisions in a self triggered fashion by reading out full detector information in continuous mode [17]. The upgrade helps to extend the physics reach and fully exploit the scientific potential of the experiment. The increase in luminosities and hence the interaction rates helps to collect high statistics of data. It will allow the ALICE detector to inspect the measurements of rare probes and sustain the advance research in state-of-the-art HEP experiments.

The significantly improved ALICE detector will allow the experiment to collect hundred times more events during LHC Run-3 as compared to Run-1 and Run-2. This requires the development and implementation of a completely new readout and computing system. The new system is designed to combine all the computing functionalities needed in the experiment: detector readout, event building, detector calibration, data recording, data reconstruction, physics simulation and analysis. The total data volume forwarded by the front-end cards of the detectors will increase significantly, reaching a sustained data throughput of up to ~3 TB/s. This huge data will be handled first time by ALICE experiment. To minimize the requirements of the computing system for data processing and storage, the ALICE computing model is designed for a maximal reduction in the volume of data readout from the detectors at the earliest possible stage during the data processing. This is achieved by online processing of the data, including calibration of detector calibration and reconstruction of events in various steps synchronously with data taking. At its peak, the

estimated data throughput to mass storage is ~90 GB/s. So the challenge for the readout is to: (1) collect all the event data at the high rate from the FEE in a self-triggered fashion with high reliability, and (2) to perform online data reconstruction thereby reducing the data volume significantly.

For this purpose, the upgrades of most of the detectors, electronics and data acquisition system in ALICE have been planned since the year 2014. The readout methodology for the upgraded ALICE detectors is based on CRU. The CRU is an integral part of the DAQ upgrade. The upgraded DAQ system for Run-3 will be known as online and offline computing (O2) system [16]. CRU is common to all the constituent detector systems in ALICE. It is a FPGA based high performance hardware equipped with multi-gigabit transceivers having optical inputs and outputs. The development of CRU is quite a complex and challenging task concerning its design requirements and the peculiarities of the hardware development and the tests. Development of CRU requires an intensive survey and an extensive literature review for the selection of the different interfaces, the selection of the location of CRU in the experiment, choice of the FPGA used with the required transceiver speeds, the on-chip resources for the online data processing, the implementation scheme, its pros and cons and the features and functionalities of the present approach. CRU acts as an interface between the on-detector Front End Electronics (FEE), O2 system and the trigger system. It receives the data from FEE of different detectors located in the harsh radiation-hard zone using high-speed (4.8 Gbps) radiation tolerant Gigabit Transceiver (GBT) links. CRU does multiplexing and processes the data as per the specific requirements of sub-detectors and then forwards to the back-end computing nodes of O2 using high bandwidth links. Each sub-detector of ALICE requires a different number of CRUs depending on the number of FEE channels. The custom developed high-end complex board like CRU with 130 high-speed serial links and the readout methodology of implementing a common hardware for all the sub-detectors will be used for the first time in ALICE experiment. In the dissertation, the major focus is on the research and developement in context of CRU development, logic

cores, optimization, performance analysis, design using different interfaces and finally their implementation in hardware and the prototype tests.

1.5 Scope of the thesis

The thesis aims on the research and engineering development involved in the context of the CRU for high data rate upgrade of ALICE. Engineering solutions developed during the R&D phase of CRU are also highlighted and covered. An emphasis is attached to the different communication interfaces through the CRU module; particularly the design of the CRU interfaces on the FPGA, its signal integrity checks, optimization, performance analysis and the firmware tests. The development of CRU hardware and the integrated tests for the developed prototypes are discussed. The main research objectives of the thesis include:

1. Design of the CRU,
2. Optimization, characterisation and evaluation of error resilient high-speed gigabit transceiver link design on FPGA,
3. Phase alignment strategy for the latency optimised implementation of the link to calibrate the clocks,
4. Performance analysis and the signal integrity characterisation on the interfaces for efficient implementation of CRU,
5. Optimization of multi-gigabit transceivers for high-speed communication links,
6. Development of CRU hardware prototype and the detailed qualification tests.

1.6 Structure of the thesis

My contribution for the optimization and performance analysis of the CRU for the ALICE is organised in this thesis into seven chapters.

- In **Chapter 1**, I present the introduction to the thesis. Here I discuss the salient features of the LHC, the ALICE experiment, the need for upgrade, and the challenges for readout and DAQ systems.

- In **Chapter 2**, I give the details of the ALICE detectors and electronics. I discuss the role of CRU to handle the high data rates arising due to the LHC upgrade.

- In **Chapter 3**, I focus on the design aspects of the CRU, its features and the functionalities. A detailed discussion on the study of the location of the CRU, the choice of design scheme their pros and cons, different interfaces used, selection of FPGA, on-chip resources and the peculiarities are presented. The complexity in the design, selected hardware, features and the functionalities of the CRU are highlighted in this chapter.

- In **Chapter 4** of thesis, I present the interfaces of the CRU. I present the GBT-FPGA logic core firmware tests and characterisation of the interfacing links are presented. The latency measurements, detailed study of the critical path delays, FPGA logic resource utilization and the power estimation are given. Firmware stability with temperature variation and the newly developed auto-initialisation phase calibration logic for GBT is discussed. GBT is integrated using the Qsys platform designer tool. The signal integrity tests for the links are performed and presented.

- In **Chapter 5** of the thesis, optimization of multigigabit transceivers for high speed data communication links in HEP Experiments is presented. I address the challenges of the high frequency losses arising due to the increase of the data rates with the experimental upgrades. The technique improves the signal integrity and gauged by Eye diagram and BER. It is validated for the available data rates of three high speed transmission protocols: GBT, TTC-PON and 10 Gbps Ethernet, popular in the HEP experiments. The advantages of the devised methodology, test results on Arria-10 FPGA and the improvement in the metrics of signal integrity for different link speeds

are presented.

- In **Chapter 6** of the thesis, the development of CRU hardware and the tests for the developed prototypes are discussed. The tests are important for the reliability of the developed boards. The rigorous evaluations are performed both at the hardware development stage and for the functional qualification of the prototypes. The details of the step by step performed tests and their outcomes are explained in this chapter.

- The **chapter 7**, I present the summary of the thesis with a discussion on the future perspectives.

Chapter 2

The ALICE experiment and its upgrade

2.1 Introduction

The ALICE experiment plans for the major upgrade of the detector during LS2, which is at present foreseen to start in the year of 2019 [15]. The particle beam luminosities for RUN-3 will progressively be increased to six times from the present value of $1x10^{27}cm^{-2}s^{-1}$ to $6x10^{27}cm^{-2}s^{-1}$ for the Pb-Pb collisions. Hence the corresponding rates of collision will be raised from 8 KHz at present to 50 KHz. In order to match the specifications for the Run-3 of LHC and to fully exploit the scientific potential of the experiment; the sub-detectors and their readout electronics need to be upgraded. An upgrade in the approach of the ALICE data acquisition is also required to handle the high rate of data ~ 3.3 TB/s; generated during the collisions. The read-out strategy for the data acquisition upgrade is based on the special common data processing block called as the CRU. It is a new approach for data concentration and multiplexing [16].

In this chapter, the details for the upgrade of ALICE sub-detectors, their readout electronics and the updated approach for the acquisition of data are summarized. The role of CRU in the experiment to handle the high data rates arising due to the LHC upgrade is discussed. Section 2.2 briefs about the present ALICE detector, an overview of the upgrade

of its sub-detectors and their readout electronics is presented. Section 2.4 elaborates on the methodology of the upgrade of data acquisition along with trigger system and the detector data links (DDL). An outline of the various detector data readout cards used in Run1 and Run2 of ALICE is given to present a framework for the need of CRU in the experiment. Section 2.5 addresses the common readout unit; its objective in the experiment and the configuration of the readout scheme for ALICE using the CRUs are presented.

2.2 The ALICE detector

ALICE is a general-purpose, heavy-ion detector at the CERN LHC which focuses on QCD, the strong interaction-sector of the Standard Model. The experiment consists of the 18 different sub-detector systems. Each sub-detector has its own choice of specific technology and the constraints of design motivated by physics requirments and the experimental conditions. The various sub-systems are optimized to produce high-momentum resolutions along with particle recognition over a broad extent in momentum, up to the maximal multiplicities foreseen for LHC. This also allows extensive analysis of electrons, hadrons, muons and photons produced in the heavy nuclei collisons. The sub-detectors at ALICE are grouped into three sections viz. the Central Barrel detectors, the Muon spectrometer and the Forward detectors [9].

2.2.1 The present ALICE detector: before Long Shutdown-2

In the present configuration, ALICE is composed of various sub-detector systems which are arranged in cylindrical shells around the interaction point.

Central Barrel Detectors

The central barrel detectors cover polar angles from 45 deg to 135 deg and are embedded in a large solenoid magnet. The main purpose of the barrel detectors is to measure the

momentum and identity of particles produced in the $|\eta| \leq 0.9$ (η is Pseudorapidity) region over full azimuth. The present system implementation from the inside to out view of barrel is described as follows:

Inner Tracking System (ITS) It is the detector system closest to the interaction point. The basic functions of the inner tracker are i) secondary vertex reconstruction of heavy flavour and strange particle decays, ii) improvement of the impact parameter and momentum resolution and iii) particle identification and tracking of low-momentum particles. The ITS provides precise information of primary and secondary vertices. Also it could provide particle identification at small transverse momenta with resolution 10-12%. ITS is a system of six barrel layers of silicon detectors providing high-resolution spatial tracking and precise vertex information. It consists of three sub-detectors each of two planes of high-resolution detectors, starting from the centre and going outwards : the *silicon pixel detector (SPD), the silicon drift detector (SDD), and the silicon strip detector (SSD).* The innermost four layers are two-dimensional devices viz. SPD and SDD due to the high particle density and the outer layers have double-sided silion micro-strip detectors.

The *SPD* is the inner most detector in the ITS. It is a charged particle multiplicity detector in the region $|\eta| < 2.1$. The SPD active elements are small pixels on the face of a silicon sensor. It plays a vital role in the determination of primary and secondary vertex. It can operate at very high track densities about 50 $tracks/cm^3$ and relatively high radiation levels. The basic component of SPD are the hybrid silicon pixels in the form of a two-dimensional matrix of reverse-biased silicon detector diodes. Each diode is connected through a conductive solder bump to a contact on a read out chip that corresponds to the input of readout cell. The SPD contains 1200 readout pixel chips and a total of 10^7 cells. The SPD is also capable of generating a prompt trigger based on an internal Fast-OR. Each pixel chip provides a Fast-OR digital pulse whenever a pixel or a group of them detects a particle signal above the threshold. The SPD readout is binary and therefore no energy-loss information is available. The SPD does not contribute to the particle identification.

The *SDD* are the two intermediate layers (3rd and 4th layer) of the ITS. It can deliver highly precise position information (with a resolution of 35 μm) and the energy loss information of (dE/dx) which could be used for particle identification. On the 3rd layer of ITS there are 14 ladders with 6 modules each and on the 4th layer there are 22 ladders with 8 modules each. The modules and ladders are assembled in order to give a full angular coverage.

The *SSD* are the outermost layers of the ITS. The outer laters have double-sided silicon microstrip detectors. It is very crucial in matching the tracks from the TPC to the ITS. The SSD is composed of 1698 modules, each one consist of a 1536 strip double sided silicon sensor connected through a aluminium kapton micro-cables to the front end electronics. It can provide a two dimensional measurement of the track position. In addition it can provide the dE/dx information. The ITS tracking information is used to restrict the global tracking of particles in the central barrel detectors: tracks that do not seem to originate relatively close to the interaction point can be discarded as background tracks from cosmic rays, scattering in materials, or other such sources. The four outer layers employ analog readout for independent particle identification via dE/dx in the non-relativistic region; this provides the ITS with stand-alone capability as a low-P_T particle spectrometer. The SDD front end electronics is based on three types of ASICs; PASCAL, AMBRA and CARLOS and the readout of SSD is based on HAL25 chips [18]. The SDD and SSD, have slightly less granularity than the SPD. They provide further tracking points and charged particle multiplicity measurements.

Time Projection Chamber (TPC) The TPC is the heart of the ALICE detector and is the main tracking detector of the central barrel. It provides together with other central barrel detectors, charged particle momentum measurements, particle identification and vertex determination. The TPC is cylindrical in shape; the active volume has an outer radius of about 250 cm, an inner radius of about 85 cm, and an overall length along the beam direction of 500 cm. The detector covers a pseudo-rapidity of $|\eta| < 0.9$ for track

reconstruction with full radial length and of $|\eta| < 1.5$ for tracks reconstruction with 1/3 radial length (with reduced or no matching with the other detectors). The TPC covers the full azimuth (except the dead zones). It covers a large transverse momentum range (0.1 GeV/c $< P_T <$ 100 GeV/c) with good momentum resolution. It is a gas detector with a volume of 90 m^3 filled with Ne/CO_2/N_2 gas mixture. The primary electrons are transported over a distance of 2.5 m on either side of the central electrode to the readout plates. A high voltage of 100 kV is applied between the central electrode and the readout plates giving a high voltage gradient of about 400 V/cm, which results in a maximum drift time of about 90μsec. The TPC is the slowest detector in ALICE. The cylindrical TPC is based on a gated read-out mult-wire proportional chambers with cathode readout. There are about 560000 readout pads, to keep the occupancy low and to ensure the necessary specific energy-loss (dE/dx), position, and two track resolution. These are of three different sizes: 4 x 7.5mm^2 in the inner chambers, 6 x 10mm^2 and 6 x 15mm^2 in the outer chambers. Particle identification in the TPC is done by using the energy loss of particles in the gas. The TPC readout is based on the PASA chip for preamplification and shaping and ALTRO chip whiere the ADC and the digital circuits are contained.

Transition Radiation Detector (TRD) TRD is located outside the TPC barrel. It consists of a radiator and Multi Wire Proportional Chamber (MWPC) placed around the TPC at radial distance of 2.9 to 3.68 m ($|\eta| < 0.84$) over the full azimuthal angle. It was installed in ALICE to improve the electron detection for $P_T > 1GeV/c$. The working principle of TRD is based on the phenomena of transition radiation emitted by a relativistic charged particle passing through a medium of different dielectric constants. The high energy electrons that cross the threshold for transition radiation ($\gamma \sim 1000$) can produce X-ray photons within the energy range of 1 to 30 keV. Since the transition radiation photons are in the keV range they are detected by the gaseous detector. To optimize the absorption of X-rays a gas mixture of Xe/CO_2 (85/15) is used. The pion being heavier do not emit transition radiation below momentum $\sim 100GeV/c$. Since the TRD is very efficient in distinguish-

ing the electrons from pions with an efficiency of about 100%, this allows for the better measurement of J/ψ production through di-electron channel. The TRD is also designed to derive a fast trigger for charged particles with high momentum. The on-detector readout electronics are ASIC based; PASA and a tracklet processor (TRAP) chip are integrated on multi-chip module (MCM). After preprocessing of the analogue signals, the resulting tracks from the various detector layers have to be matched in 3-dimensions for transverse momentum reconstruction using the FPGA based global tracking unit.

Time Of Flight (TOF) The TOF detector in ALICE provides the particle identification in the intermediate momentum range, below about 2.5 GeV/c for pions and kaons, up to 4 GeV/c for protons, with Π/K and K/p separation better than 3σ, by measuring the span of time between the collision and the arrival of the particles in the detector. It provides a measurement of the time it takes for a particle to travel from the interaction point, through the magnetic field, to the outer rim of the barrel. The TOF is a gas detector based on Multi-gap Resistive Plate Chamber (MRPC). The basic unit of the TOF system is a 10-gap double-stack MRPC strip 122 cm long and 13 cm wide, with an active area of 120 x 7.4 cm^2 subdivided into two rows of 48 pads. The TOF consists of 90 modules. Every module of the TOF detector consists of a group of MRPC strips (15 in the central, 19 in the intermediate and external modules) closed inside a box that defines and seals the gas volume and supports the external front-end electronics and services. The detector covers a cylindrical surface and the modules are arranged in 18 sectors in ϕ and in 5 segments in z-direction. Five modules of three different lengths are needed to cover the full cylinder along the z-direction. The length of the central module is 117 cm, the intermediate ones 137 cm, and the external ones 177 cm. The overall TOF barrel length is 741 cm (active region). The TOF is located at radii from 2.70 to 3.99 m and covers a pseudo-rapidity of $|\eta| < 0.9$. The chambers have high and uniform electric field over the full sensitive gaseous volume of the detector. Any ionization produced by a traversing charged particle immediately starts a gas avalanche process which generates the observed signals on the pick-up strips. The setup

achieves a very good time resolution of about 40 ps. The TOF detector has about 160000 channels. The front-end electronics for the TOF is based on 'NINO' ASIC. It is complied with the basic characteristics of the MRPC detector; very fast differential signals from the anode and cathode readout pads and intrinsic time resolution better than 40 ps.

High Momentum Particle Identification Detector (HMPID) The Ring Imaging Cherenkov HMPID detector is placed at a distance of about 4.5m from the beam axis. Its purpose is to identify very high momentum particles. The HMPID exploits the fact that charged particles emit Cerenkov radiation when the velocity of the particle is larger than the speed of light in the medium traversed. The HMPID system enhances the particle identification capability of ALICE beyond the momentum range allowed by the energy loss measurements (ITS and TPC) and by the TOF. The HMPID detector has been designed to extend the useful range for the identification of Π/K and K/p, on a track-by-track basis, up to 3 GeV/c and 5 GeV/c respectively. The HMPID is designed as a single-arm array with an acceptance of 5% of the central barrel phase space. It is based on proximity focusing Ring Imaging Cherenkov (RICH) counters and consists of seven modules mounted in an independent support cradle, which will be fixed to the space frame, at the two o'clock position. The radiator is a 15mm thick layer of low chromaticity C_6F_{14} (perfluorohexane) liquid with an index of refraction of n = 1.2989 at λ = 175 nm corresponding to βmin = 0.77. Cherenkov photons, emitted when a fast charged particle traverses the 15 mm thick radiator, are detected by a photon counter, which exploits the novel technology of a thin layer of CsI deposited onto the pad cathode of a multi-wire proportional chamber (MWPC). The HMPID detector, with its surface of about 12 m^2, represents the largest scale application of this technique. The Cherenkov photons refracts out of the liquid radiator and reach the CsI-coated pad cathode, located at a suitable distance (the proximity gap) that allows the contribution of the geometrical aberration to the Cherenkov angle resolution to be reduced. The electrons released by ionizing particles in the proximity gap, filled with CH_4, are prevented from entering the MWPC sensitive volume by a positive polarization

of the collection electrode close to the radiator. The front-end electronics of the HMPID is based on the ASIC chips, GASSIPLEX and DILOGIC. There are 161280 readout channels.

Photon Spectrometer (PHOS) PHOS is a high resolution electro-magnetic calorimeter. It is located at a radial distance of 4.6 m with an azimuthal acceptance 220 degree to 320 degree ($|\eta|< 0.12$). It consists of lead tungsten crystals (PbWO4). It is a high-granularity calorimeter measuring photons and features an excellent energy resolution. It has 5 modules with 3584 crystals in each. There is a set of Multi-Wire Proportional Chambers infront of PHOS to reject the charged particles, called the Charged Particle Veto (CPV). The aim of the PHOS detector is to detect the direct photons, π^o and η mesons. The active volume is 14 mm thick gas mixture of Ar and CO2 in the ratio 80:20 at a pressure slightly above atmospheric pressure. PHOS readout electronics is based on the Hamamatsu S8664-55 Avalanche Photodiodes and the customised preamplifier and shaper. The data are further processed using the FPGA based trigger boards and the readout controller units.

Electromagnetic Calorimeter (EMCal) The ElecroMagnetic Calorimeter was installed to enhance the measurement of jets and high P_T photons and electron identification. The EMCAL is a Pb-scintillator sampling calorimeter. It is located at a radial distance of 4.5 m from the beam line opposite to the PHOS and covers an azimuthal angle of 80^o to 187^o ($|\eta| < 0.7$). It is, however, larger than PHOS with an acceptance of about 23% of phase space of the central region, but offers lower granularity and resolution. It can provide the measurement of transverse energy (E_T) in the region from 100 MeV to 100 GeV. It also provides a fast and efficient trigger (L0 and L1) for hard jets, photons and electrons. The detector is arranged in 12 supermodule units of two types: 'full size' which span $\Delta\eta = 0.7$ and $\Delta\phi = 20^o$, and 'one-third size' which span $\Delta\eta = 0.7$ and $\Delta\phi = 7^o$. The lower 2 supermodules are 'one-third size' type while the rest 10 are of 'full size' type. These supermodules are segmented into 12288 towers. The scintillation photons produced in each tower are captured by an array of Y-11 double-clad wavelength-shifting (WLS) fibres. Each fibre terminates in an aluminized mirror at the front face of the module and is integrated

into a polished, circular group of 36 at the photo-sensor end at the back of the module. The fibre bundle in a single tower terminates in a 6.8 mm diameter disk and connects to the Avalanche Photo Diode (APD) photo sensor through a short light guide.

A COsmic Ray Detector (ACORDE) The ACORDE provides a fast (Level-0) trigger signal, for the commissioning, calibration and alignment procedures of some of the ALICE tracking detectors, and it also detects in combination with the TPC, TRD, and TOF, single atmospheric muons and multi-muon events (so-called muon bundles) thus allowing us to study high-energy cosmic rays in the energy region of knee in the cosmic ray spectrum. The ACORDE is an array of plastic scintillator counters placed on the upper surface of the L3 magnet. It consists of two scintillator counters, each with 190 x 20 cm^2 effective area, placed on top of each other and read out in coincidence. The detector is arranged in 60 modules covering a pseudorapidity range $|\eta| < 1.3$.

Muon Spectrometer

The Muon spectrometer provides the measurement of the complete spectrum of quarkonia (J/ψ, ψ', Υ, Υ', Υ''), as well as $\Phi - meson$, via their decay in the $\mu^+\mu^-$ channel. The invariant mass resolution is of the order of 70 MeV in the J/ψ region and about 100 MeV close to the Υ. These values are good enough to separate out all five resonance states. The muon spectrometer consists of a passive front absorber to absorb hadrons and photons, a high-granularity tracking systems of 10 detection planes, a large dipole magnet, a passive muon filter wall, followed by four planes of trigger chambers, and an inner beam shield to protect the chamber from primary and secondary particles produced at large rapidities. The tracking system is made of 10 cathode pad/strip chambers arranged in 5 stations of 2 chambers each. To limit the occupancy within 5%, the full set of chambers has more than 1 million channels. The trigger system is designed to select heavy quark resonance decays. The selection is made on the P_T of the two individual muons. Four planes of Resistive Plate Chambers (RPCs) arranged in 2 stations and positioned behind a passive muon filter

provide the transverse momentum of each μ. The spatial resolution is better than 1 cm and the time resolution is 2 ns. The muon spectrometer covers a pseudorapidity range of $-4 < \eta < -2.5$ and has full azimuthal coverage for muon with p $>$ 4GeV/c. The front-end electronics of muon chambers consists of MANU (MAnas NUmerique) boards and the readout system is known as Cluster Readout Concentrator Unit System (CROCUS).

Forward Detectors

A number of detector systems placed at small angles from the beam line serve to provide global event characteristics, like triggering, primary vertex, and multiplicity.

Zero Degree Calorimeter (ZDC) Two identical sets of ZDC, one on each side relative to the interaction point (I.P.), provides the centrality trigger of the collision through the detection of spectator nucleons. In addition, the ZDC being a position-sensitive detector, can give an estimate of the reaction plane in nuclear collisions. It is composed of four calorimeters, two to detect neutrons (ZN) placed between the beam pipes at 0^o relative to the LHC axis and two to detect protons (ZP) placed externally to the outgoing beam pipe on the side where positive particles are deflected. They are located 116 meters away from the interaction point on both sides, exactly along the beam line. The measurement is complemented by an electromagnetic calorimeter (called ZEM, $4.8 < \eta < 5.7$) which measures the total forward energy at z = 7.25 m. The ZDCs are "spaghetti calorimeters", made by a stack of heavy metal plates grooved to allocate a matrix of quartz fibres. The metal plates are made of a special material namely a tungsten alloy for neutrons and brass for protons. The material of the metal plates is known as "passive material", while the quartz fibres are known as "active material". High energy protons and neutrons hitting the passive material create a cascade of particle, called "shower". When one of these shower particles crosses a fibre, if its speed is high enough, it can produce light (Cherenkov effect). This light propagates in the fibre by total reflection up to its end, where a photomultiplier converts the light into an electric signal. The amplitude of the electric signal is proportional

to the energy of the incoming protons and neutrons allowing to measure the energy carried away by the spectator nucleons and therefore, indirectly the size of the overlap region of the two colliding nuclei.

Forward Multiplicity Detectors(FMD) The FMD provides the measurement of charged particle multiplicity in the pseudorapidity range $-3.4 < \eta < -1.7$ and $1.7 < \eta < 5.0$, both in full azimuth. In addition, the information from FMD can be used to study the event-by-event multiplicity fluctuation, determination of reaction plane, and elliptic flow measurement within its pseudorapidity coverage. FMD can also be used to study the correlation between photons and charged-particles at forward rapidity. The FMD consists of 5 rings of Si semiconductor detectors with a total of 51200 individual strips. The rings are of two types: the inner type consist of 10 wafers subdivided into 20 sectors with 1024 strips each. The outer type is subdivided into 40 sectors each with 512 strips. The Si wafers are 300 micrometer thick and are manufactured out of 6 inch diameter Si disks. The FMD consists of 3 groups of detectors called FMD1, FMD2, and FMD3. FMD2 and FMD3 each consist of a ring of inner type Si sensors and a ring of outer type Si sensors. These are located on either side of the IP. FMD1 consists of a ring of inner type Si sensors and is placed opposite to the muon spectrometer to extend the charged particle multiplicity coverage.

Photon Multiplicity Detector (PMD) PMD has been designed to measure photon multiplicities in the forward region and to provide estimates of transverse electromagnetic energy. Using these measurements on an event-by-event basis, the PMD studies event shapes and fluctuations. It is located at a distance of 3.67 m from the interaction point in the A side of the ALICE and it covers a pseudo-rapidity $2.3 < \eta < 3.9$ with full azimuth. The PMD consists of two planes (charged particle veto and preshower) separated by a 3X0 thick converter (1.5 cm Pb and 0.5 cm Stainless Steel). It consists of a highly segmented detector placed behind a lead converter of suitable thickness known as preshower plane. A charged particle detector of similar granularity is placed in front of the converter to act as veto in order to improve the discrimination between charged particles and photons. Each plane

consist of 24 unit modules. PMD will not be further installed in the Run3 of ALICE because of the limitations of this detector to handle the high data rate of the experiment in Run3.

VZERO (V0) The V0 detector in ALICE provides the minimum bias trigger for the central detectors in pp and A + A collisions, triggers on centrality in Pb-Pb collisions via the multiplicity recorded in the event. It is used to reject the beam-gas events and to pre-trigger the TRD. The V0 consists of two arrays of scintillator counters, called as V0A and V0C, which are installed on two sides of the ALICE interaction point. The V0A is located 340 cm from the IP in front of PMD covering a pseudorapidity range $2.8 < \eta < 5.1$ while the V0C is located at 90 cm from the IP on the side of muon spectrometer covering a pseudorapidity range $-3.7 < \eta < -1.7$. Both V0A and V0C, are segmented into 32 individual counters which are distributed into four rings. The time resolution of individual counters are 1 ns. Each array provides two types of triggers. One is based on pre-adjusted time windows in coincidence with the time signals from the counters. Minimum bias, Beam-Gas, and Multiplicity Triggers are obtained by this method. The second type of triggers is based on the total charge collected by the arrays. The two centrality triggers are built starting from the quantities semi-central trigger (CT1) and central trigger (CT2)).

TZERO (T0) The TZERO detector consist of two arrays of Cherenkov counters, called T0A and T0C. The T0A is located at a distance of 3.75 m from the IP, opposite to the Muon spectrometer covering the range of $4.61 < \eta < 5.92$. The other component T0C is located on the opposite side of T0A at a distance of 7.27 m from the IP with coverage $-3.28 < \eta < -2.97$. The T0 detector can measure the collision time with a precision of 25 ps. This time is used as a reference time for the TOF to measure the flight time of particles. The T0 can also send a pre-trigger to the TRD. It can also generate minimum bias and multiplicity triggers.

2.2.2 ALICE Data Acquisition and computing for Run1 and Run2

The trigger mechanism decides the recording of the events for further analysis of data in back-end computing systems and rejects the non-interesting events. It is a complex system however it is required due to the limitation of computing power, data handling capability and storage capacity of the memory devices. The events are recorded on the basis of their properties. Parallelized trigger scheme is employed. This exploits the symmetry of the detector systems. Same nature of events on various parts of a detector system is recorded at the same time. Trigger system in HEP detectors are hierarchical in nature and divided into different levels. Each level gives the input to the next level which has more information and more time duration for a precise decision. As an example; ALICE experiment uses three different levels of trigger. The first level reduces event rate and is based on the electronics placed near the detector system. The second level is implemented using optimized software algorithm and it constitutes a segment of high level trigger (HLT). The third level trigger is also software based and filters the event. It is also as part of HLT. The first level trigger is always the fastest as compared to the higher level of triggers.

The ALICE trigger system consists of **Central Trigger Processor (CTP)** and a **High Level Trigger (HLT)** [18]. The HLT is a software trigger and CTP is a low level hardware trigger. The hardware trigger CTP in ALICE integrates the input from different detectors with fast trigger capability (for eg: ZDC, T0, V0, SPD, TOF, TRD, ACORDE, PHOS, EMCal, Muons). The trigger includes a flexible check against pile-up and an event prioritization scheme. This optimises both the acceptance of rare triggers and the overall throughput of accepted events. It is operated at several levels to satisfy the individual timing requirements of the different ALICE detectors. A pretrigger activates the TRD electronics shortly ($<$ 900 ns) after each interaction. The first level trigger is called the L0 which is delivered after 1.2 μsec, the second level is called L1 delivered after 6.5 μsec and the final trigger L2 is delivered after 100 μsec. L2 is issued after the end of the drift time in the TPC, the slowest detector in ALICE. After L2 trigger the event is stored. The CTP receives and

process trigger signal from trigger detectors as listed in Table 2.1. The CTP consist of

Detectors	TRD	TOF	EMCAL	PHOS	SPD	ACORDE	V0	T0	Muon Trigger	ZDC
Trigger level	L1	L0	L0/L1	L0	L0	L0	L0	L0	L0	L1

Table 2.1: Trigger level in ALICE detectors

24 Local Trigger Units (LTU) for each detector system. The output of the CTP goes to the LTUs of each detectors and then to the front end electronics of the detector through Low Voltage Differential Signaling (LVDS) cables and optical fibres. The CTP defines 50 independent trigger classes combining 24 no.s of L0 inputs, 24 no.s of L1 inputs and 12 no.s of L2 inputs.

The software based **HLT** is a framework of upto 1000 multiprocessor computers. HLT permits a firmware and software filtering mechanism to select interesting events or to compress the complete event information. The HLT receives a copy of the raw data via 454 Detector Data Links (DDL). Then it performs the basic calibration and extracts hits and clusters. Then the event is reconstructed for each detector individually. After that the selection of event is performed with the reconstructed physics observables. The HLT also perform compression on data allowing to reach a sustained rate to disk of more than 4 Gb/s. The generated data and decisions are transferred to dedicated local data collector machines. HLT-Readout receiver card (H-RORC) is an FPGA based card designed for receiving data from the ALICE detectors and transmitting the processed events out of the computing farms to the data acquisition system. H-RORC is an interface between the HLT system and the ALICE DDL.

The **ALICE DAQ** handles the data flow from the detector electronics to the permanent storage. The overall architecture of the present ALICE DAQ in Run2 and the interface to the HLT system is shown in Figure 2.1 [18]. A first layer of computers known as Local Data Concentrators (LDCs) read out the events from the optical DDLs. The DDLs are point to point links at 2 Gb/s and up to 12 of them can be connected to the same LDC. The LDCs assemble the data into the sub-events and then shipped to Global Data Collectors (GDCs).

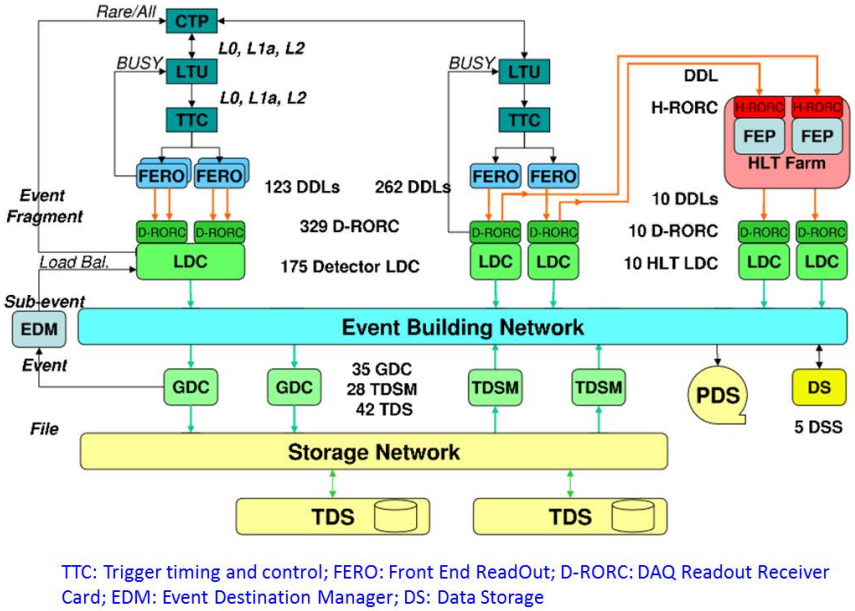


Figure 2.1: ALICE DAQ architecture for Run2 and the interface to the HLT system.

The GDCs recorded the events to a Transient Data Storage (TDS) and then it is migrated to the permanent data storage.

2.3 ALICE upgrade overview: Upgrade of sub-detectors and their readout electronics

For the planned upgrade to start from the year of 2019; the specification for the ALICE detector upgrade is set by the collision rate of 50 kHz for Pb-Pb and a collision rate of 200 kHz for pp and p-Pb. Also the concept of reading the full detector information either upon a Minimum Bias trigger or in a continuous fashion, requires the upgrade of the ALICE sub-detectors and systems, their readout electronics and the data acquisition methodology. The different constituents need to be upgraded is summarized in the subsequent sections. The ALICE sub-detectors that are going for the major upgrade as decided by experimentalists is shown in Figure 2.2.

To summarize the main detector modifications of different ALICE sub-detectors; it con-

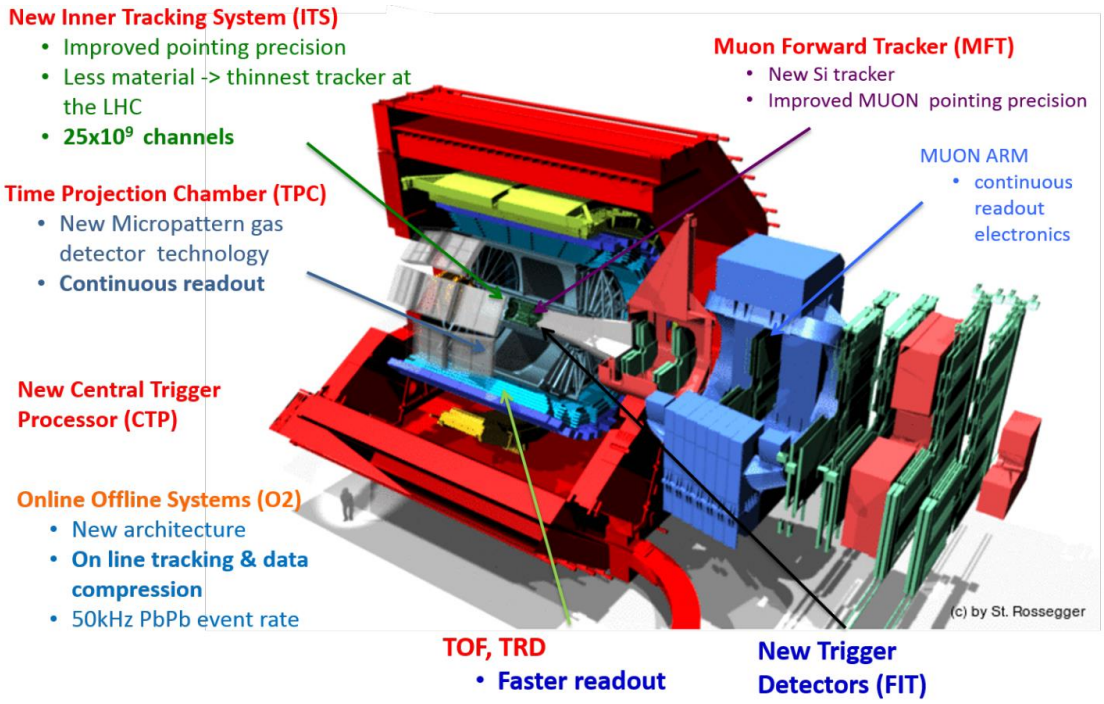


Figure 2.2: Scheme of the Major Sub-detectors of ALICE approved for upgradation during the long shutdown2

sists of replacing the present **silicon tracker ITS** by seven layers of monolithic silicon pixel detectors [19]. With this the ITS will be able to provide read-out at the rates of 100 kHz for Pb-Pb and 1 MHz for the p-p collisions.

The **Muon Forward Tracker (MFT)** is a new detector to be installed at the cavern during the upgrade which will work with muons. It is located in the forward region between the ITS and the front absorber. The detector technology of MFT is similar to the new ITS. MFT will consist of several discs of monolithic silicon pixel sensors. The sensors and read-out will exploit a maximum synergy with the ITS detector.

The present **TPC detector** based on gated readout with wire chambers limits the read-out rate to less than 3.5 kHz. During upgrade, the wire chambers will be replaced with gas electron multiplier (GEM) detectors [20]. It allows continuous operation to read out 50 kHz Pb-Pb collisions. The TPC electronics will send the time-stamped and digitised data to the acquisition system in a triggerless manner. The read-out of the TPC detector as well as muon chambers (MCH) detector whose readout rate is presently limited to 1 KHz; will

be performed by a dedicated new 32 channel ASIC called as SAMPA [21] that is presently being tested.

The upgrade trigger strategy does not foresee a muon trigger, all the events will be read upon the interaction trigger and the data are used offline for hadron rejection hence the **Muon Trigger detector (MTR)** will be designated as **Muon Identifer (MID)**. The resistive plate chambers (RPC) of the MTR will be operated in *genuine* avalanche mode with a significant reduction of the charge produced in the gas, hence limiting the ageing effects. The readout electronics upgrades of the MID system consist of replacement of all the read-out electronics, including local, regional and Dimuon-trigger ALICE Read-out Controller (DARC) cards. The FE chip, called ADULT [22] will be replaced by a new ASIC, named as FEERIC. Unlike ADULT, FEERIC will perform amplification of the analog signals from the RPCs.

The **Transition Radiation Detector (TRD)** is presently limited to a few kHz read-out rate. A trigger rate of 100 kHz for Pb-Pb and p-p can be achieved by reducing the data volume from the detector by using tracklets and increasing the data throughput of the off-detector electronics. On-detector electronics of TRD would be not be changed.

The **Time Of Flight (TOF)** detector is at present limited to the read-out rate of 40 kHz; by the throughput of the VME system located in the crates at the end of the detector modules [16]. An **upgrade of the VME** is planned. It will allow TOF to read out greater than 200 kHz Pb-Pb events, which easily satisfies the upgraded requirements.

To provide the Minimum Bias interaction trigger for the experiment, V0/T0/FMD (Forward Multiplicity Detector) detector system will be replaced by a single detector known as **Fast Interaction Trigger (FIT)** detector [16]. The FIT detector system will be located in the forward region of the ALICE detector at positions close to the present V0/T0 location. The trigger and read-out electronics of this detector system closely resembles to one of the present T0 system.

There are few detectors which will not be upgraded. The **Zero Degree Calorimeter**

(ZDC) is located at a distance of 115m from the interaction point. It will only change the read-out electronics to accept the increased trigger rate. The combinatorial logic for the discriminated signals will be implemented on FPGA rather than the NIM modules. The **Electro-Magnetic calorimeters (EMC)** and **Photon Spectrometer (PHO)** use the same readout electronics which is already in operation. They are capable of the 50 kHz operation and no upgrade is required. The **High Momentum Particle Identifier (HMP)** and **ALICE Cosmic Ray Detector (ACORDE)** will also not be modified.

2.4 Upgrade of Data Acquisition methodology

The physics objective of the detector upgrade is aimed at the precision measurements of the QGP. It will be accessible through the measurement of certain parameters and processes which exhibit traces that can only be claimed by a zero bias (also known as Minimum Bias) trigger. All the events need to be collected without being lost. To meet this requirement the methodology for the data acquisition needs to be modified from the previous Run1 and Run2 of ALICE. Although the back-end data processing complexity is reduced in the trigger based system like for Run1 and Run2, however it is not suitable when the motive of the experiment is to study the rare events. In order to capture the rare events, event rate should be very high. The method used for such a case is to recorded initially all the data from the front end and then there will be an online event selection mechanism that rejects the unnecessary background events prior to the storing of data for future analysis. This process is known as self-triggered mechanism. Apart from the decrease in probability of misdetection of rare events; its advantage is that the detector dead time is reduced, intricate event selection schemes can be easily be implemented in software and quick to adapt the new criteria when compared to hardware trigger.

2.4.1 Trigger system

The concept of upgrade strategy is to record all the interactions and apply a data reduction in the online-offline (O2) computing system. For this a combination of continuous trigger-less readout methodology and the minimum bias trigger based on the new forward FIT detector is used. However, a CTP will be employed in order to keep flexibility and to provide triggers for calibration and commissioning. It will also help to exclude the unwanted background noise. For each collision, one trigger signal is distributed to all detectors by the ALICE CTP to start the read-out of detectors. A trigger message is also sent to the detector read-out modules in addition to the trigger signal. The trigger message contains the orbit counter and bunch crossing information of the LHC beam.

In order to collect the event fragments in the O2 system by identifying the corresponding data, both counter values are tagged to the readout data packets. The CTP in ALICE will be upgraded to accommodate the higher interaction rate, providing trigger and timing distribution (TTS) to the upgraded detectors. Also the backwards compatibility to Run1 and Run2 detectors not upgrading their TTS interface is provided. These non-upgrading sub-detectors will therefore be read out whenever they are not *busy* and merged with the data from the other sub-detectors in the online computing system. (*Busy signal* is asserted when the trigger rate exceeds the detector read-out capabilities)

For the detectors operating in a trigger-less manner, the Heartbeat software trigger (a dedicated non-physics trigger) [17] is used to designate time frame boundaries for event building at the O2 computing system. The time frame duration of the heartbeat trigger made of configurable length is common to all the sub-detectors. It is long enough to minimise the number of events crossing the boundary of two successive time frames. A value of at least 100ms is foreseen which is comparable to the TPC drift time of 280us (the ion drift time of 180us from the sense wires to the gating grid with the electron drift time of 100us from the central electrode to the read-out chambers).

The time frames allow the segregation of the data stream into fragments for the event

reconstruction. The event building is based on the assembly of data recorded during this time frame. The heartbeat trigger will be directed by CTP with the highest concern and with a fixed period [23]. The time frame boundaries are broadcasted to the sub-detector read-out electronics via communication of non-physics heartbeat triggers. The detector read-out systems handles individual copies of the bunch crossing, orbit and trigger counters. The hardware compares these local copies of the counters with the global LHC counters only transmitted in full during a heartbeat event. In case of a difference, the detector electronics reshuffle the counters and the error is communicated to the online computing system. The information of the bunch crossing and orbit counter of the heartbeat trigger arrival is generated by each readout unit in the form of a heartbeat event containing no physics data. These events will be utilised for data segmentation, fault finding and recovery procedures by the online system. The detector electronics of the present detectors will be upgraded to support this combination of physics and heartbeat triggers.

For the trigger based readout of the non-upgraded detectors, receive one of the trigger levels; either LM or Level-0 or Level-1 based on the requirements of the trigger latency. The LM signal is produced by the FIT detectors, with a latency that is compatible with the timing requirements of the TRD wake-up signal. At nominal operation, this is the only trigger contributor and L0, L1 are the delayed versions of the LM signal [24]. All the trigger decisions are provided to the ALICE sub-detectors by CTP. Local Trigger Units (LTU) are used by the CTP for communicating triggers and timing information to each sub-detectors, using trigger and timing system (TTS). All trigger types provided by the CTP are sent via the LTUs and the TTS to the detector read-out systems.

Two communication paths are available between the CTP and the read-out electronics; TTS-Fast serial Trigger Links (FTL) and TTS-timing, trigger and control (TTC). For upgraded detectors the TTS pathway is based on asynchronous FTL. Detectors with non-upgraded electronics or legacy detectors will keep using CERN$'$s timing, trigger and control standard based equipments.

2.4.2 Detector Data links

DDLs are used to transfer the data between the on-detector electronics and the data acquisition system. There are different versions of DDLs developed to match the system requirement in the present Run1 and Run2 of ALICE data collection. They have different clocking speed and the form factors. The communication channel of the DDL consists of the Source Interface Unit (SIU) and the Destination Interface Unit (DIU).

The DDL1 is operated at 2.125 Gb/s in full duplex mode [25]; the SIU is implemented as a radiation tolerant mezzanine card plugged on the detector FEE and the DIU is a FPGA based Read-Out Receiver Card (RORC) [26] installed in the DAQ farm. The second version DDL2 [27] has been used during Run2 by the TPC and TRD detectors. The SIU of DDL2 is actuated as an Intellectual Property (IP) core and can be clocked at 4.25 or 5.3125 Gb/s according to the capabilities of the detector electronics using it. For the upgrades during Run3, CRU based higher performance read-out solution is developed This aims at higher bandwidths to assemble the data fragments into sufficient large blocks. It is discussed in detail in the subsequent chapters.

2.4.3 Detector Data Readout Cards in ALICE

During the system implementation at the Run1 and Run2 of the experiment; different read-out cards had been developed to serve the readout requirements of the specific sub-detectors. PCI-based Readout Receiver Card (PRORC) [28] and the DAQ Read-Out Receiver Card (D-RORC) [26] are the two versions of the RORC developed for the Run1. The PRORC is an ALTERA Programmable Logic Device (PLD) type card with 32bit/33MHz PCIe interface. The second version, D-RORC realized using an Intellectual Property (IP) implemented on the APEX-E PLD [29] with 64 bit/66MHz PCI interface form factor. After few iterations, D-RORC of type-1 and type-2 were used in the experiment. Both were based on Altera FPGA (Startix-II series) with 64bit/100MHz PCI-X bus interface and x4 PCI (Gen-1) bus interface respectively.

During Run1, High Level trigger RORC (H-RORC) cards were also used for the trigger distribution. Readout controller units (RCUs) were also used for the raw data readout from the detector. Both H-RORC and RCUs were based on FPGA co-processor architecture.

During Run2, the common version of the RORC cards (C-RORC) is developed [27]. These cards are based on Xilinx Virtex-6 FPGA with PCIe Gen2 (40 Gbps) bus plug-in and a support for 12 links of DDL1 (2.125 Gb/s) or DDL2 (4.25 Gb/s up to 5.3125 Gb/s) [25]. C-RORC has increased the processing capabilities in Run2 by unifying the task of the ALICE high level trigger systems and detector data acquisition on a single card.

In the present framework usage of Run2, ALICE gives a structure of common read-out and trigger interfaces. The DDL gives a standard SIU that associates the sub-detector read-out electronics optically to the RORC cards situated in the DAQ PCs. During Run3, the general approach is to readout all the events with increased beam luminosity. The large event size with high interaction rate will result in a data flow of ~3.3 TB/s from the detectors to the on-line system. Data processing capability in the online systems results in a peak data rate to storage of ~90 GB/s. The shift of paradigm in the readout strategy calls for the ALICE to develop new readout architecture.

ALICE will broaden the approach utilizing standard system interfaces for the Run3. The DDL will be upgraded to higher speed interfaces and complemented with a Common Read-out Unit (CRU) [30].

2.5 Common Readout Unit (CRU)

To handle the requirements of ALICE experiment in Run3, a new approach based on CRU is being developed. CRU is an integral part of O2 and detectors upgrade. The CRU acts as an interface between the on-detector electronics system, the O2 system and the trigger processor system as shown in Figure 2.3. It is a high speed, fault tolerant readout architecture with the ability to acquire the data from the harsh radiation environment, yet

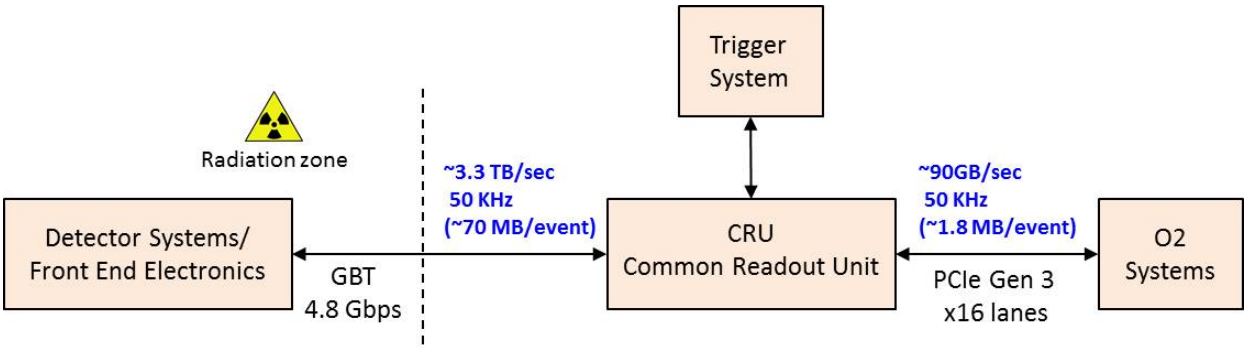


Figure 2.3: Data processing capability in the online systems

flexible enough to keep up with upgrades and instant reconfigurable. The uniqueness of the proposed scheme is that the readout is implemented as custom designed electronics boards with programmable functionality based on non-radiation hardened state-of-the-art Altera FPGAs. Non-radiation hard FPGAs have high processing power [31] in comparison to the radiation hard FPGAs. For this development, FPGAs are considered for rapid prototyping as they are field reconfigurable with large resources. It allows higher parallelism and pipelining thereby increasing the logic computation speed and minimizing the latency involved.

The on-detector interface of CRU is based on the Gigabit transceiver (GBT) and optical versatile link [32, 33] protocol and components. The data will be transferred with high reliability from the detectors in the harsh radiation-hard zone to the CRU situated in low or no radiation zone counting room through the developed 4.8 Gbps GBT links. These links can be operated in wide bus mode, with a payload data bandwidth of 4.48 Gb/s or in forward error correction mode, with a payload data bandwidth of 3.2 Gb/s. GBT interface standard for HEP is the suitable solution for error resilient data communication. The trigger interface to CRU is based on the Timing-Trigger and Control over passive optical networks (TTC-PON) [34]. TTC-PON standard is chosen for this interface. It has the high split ratio having fixed latency.

The strong candidature for the link (also called as DDL3 link for Run3) between the CRU and the Data Acquisition system are presented by 10-Gigabit Ethernet links or the PCIe

interface. The link selected is Gen3x16 lanes PCIe interface and the hardware card is named as PCIe40 card [35]. The use of high speed interconnects like PCIe based backplane interface enables the DAQ system to handle data and trigger signal of the order of Gbps. Employment of PCIe based back-end interface and FPGA devices give flexibility to the designers for either local on-board data processing or PCIe based backplane data processing. CRU is a customized board based on Altera Arria-10 FPGA and qualified for the implementation of CRU. The details of CRU, its design, developement and iterfaces are presented in subsequent chapters. For detectors not upgrading their interfaces, backwards compatibility to Run1 and Run2 systems is provided. Depending on sub-detector specifications, the digitized data sent from the detector FEE to the CRU are multiplexed, processed and formatted before being forwarded to the back-end computing nodes.

2.5.1 ALICE Readout scheme using CRU

The block diagram of the proposed ALICE readout scheme for RUN3 using the CRUs is shown in Figure 2.4. The central trigger processor is situated in the experimental cavern connects to the TTS via the LTU, which depending on detector system, are based on either FTL or TTC [36] links. The on-detector electronics systems are connected to the ALICE-CRU via front-end GBT links for the upgraded sub-detectors. For the sub-detectors where the modification of the read-out electronics is not required, the on-detector electronics are connected to the detector-specific read-out systems.

The read-out systems are connected to the O2 system and the detector control system (DCS) via the ALICE different standard DDLs. There are three configurations for the ALICE read-out using the CRU

- In first configuration, the LTU makes use of FTL protocol to transmit the timing and trigger information directly to the CRUs located in the counting room via a trigger distribution module. The multiplexed data is transmitted to the detector FEE from the CRU via the GBT front-end links.

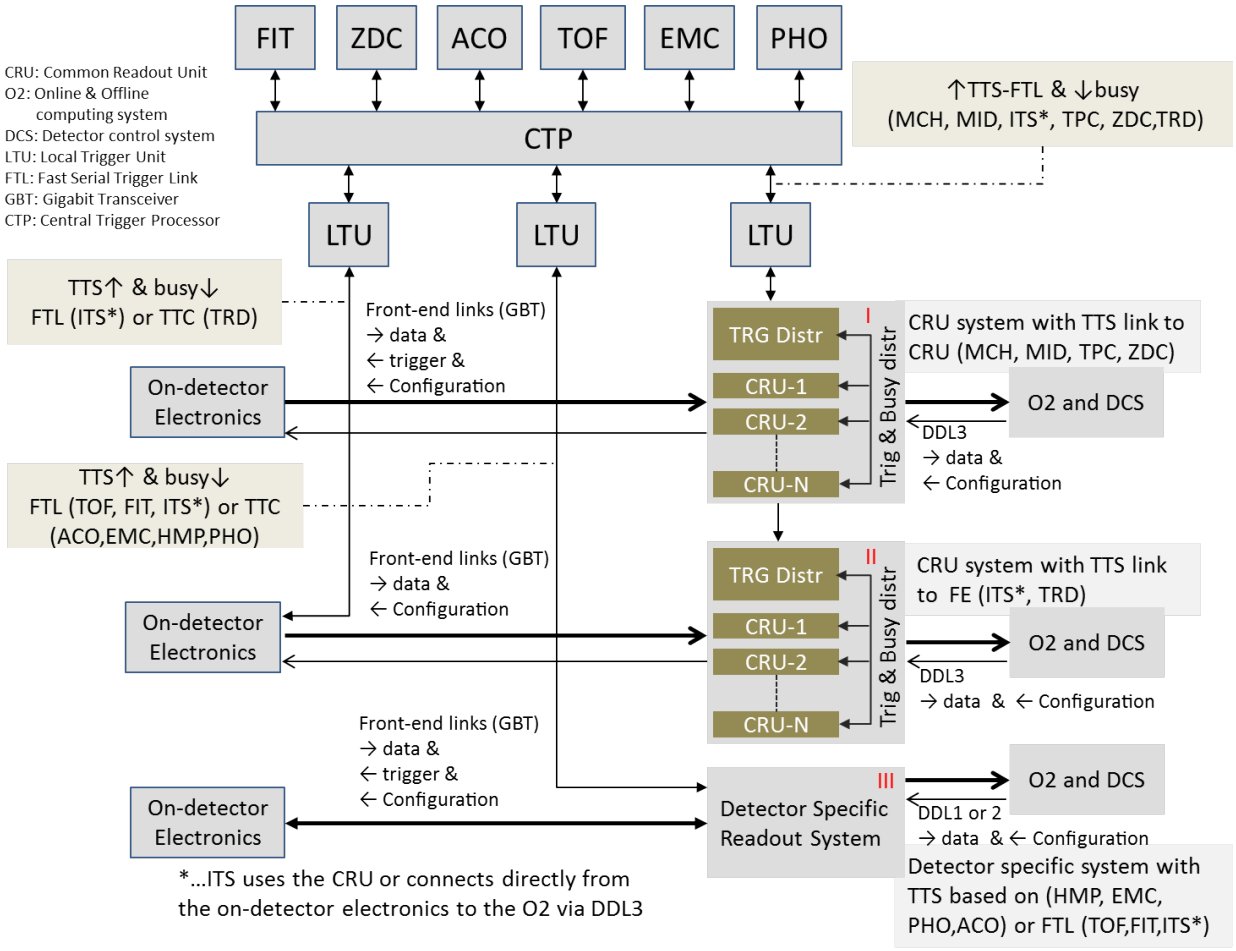


Figure 2.4: A system block diagram of ALICE readout scheme. Config I : CRU is used as trigger distribution system and read-out processor. Config II : CRU is used as read-out processor. The trigger distribution is done from the CTP/LTU directly to the on-detector electronics. Config III : CRU is not used for the detectors do not upgrade their front-end electronics and use detector-specific read-out cards.

- For the second configuration the LTUs connect the trigger and timing distribution links based on fast serial trigger protocol or TTC to the sub-detector front-ends, escaping the CRU in the trigger path. This configuration is used for the detectors which require a minimum latency trigger path along with the CRU for the read-out.

- The third configuration for the detector readout does not use the CRU. The trigger and timing distribution links based on fast serial trigger protocol or TTC from the LTU are used by the detector specific read-out systems and connect to the O2 system via DDLs.

The overall data-flow between the detector and the online-offline computing system is shown in Figure 2.5.

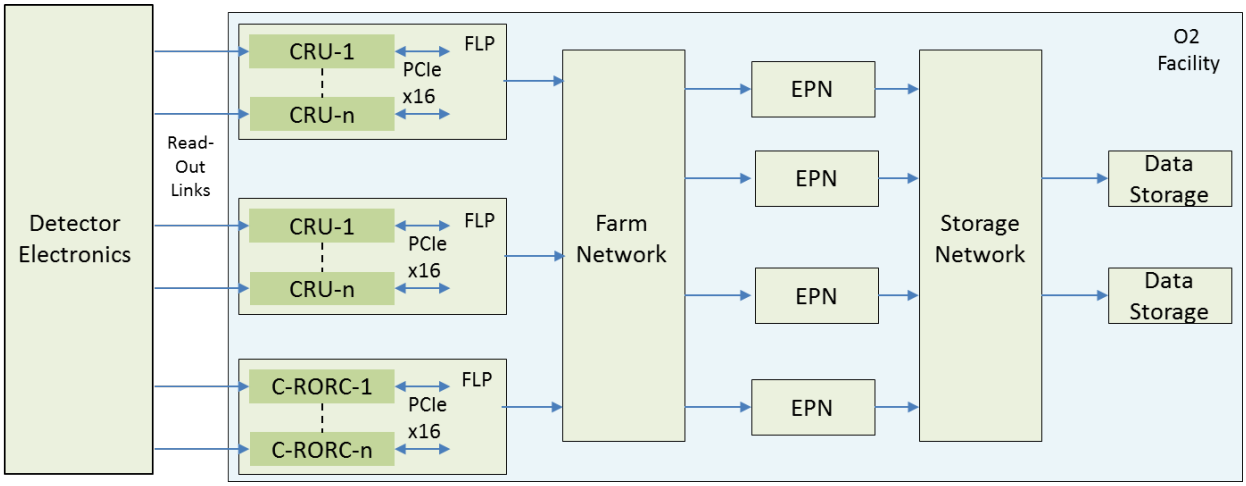


Figure 2.5: Flow of data in the ALICE from the detector front-end electronics up to the O2 system. (FLP: First Level Processor, EPN: Event Processing Node)

The generated data from the detectors are transported to the CRU in a trigger-less fashion or triggered mode using the GBT links or the system specific DDL based readout links. The triggered data will be labelled with the information of the LHC clock; like the case for Run1 and Run2.

The heartbeat triggers are distributed by the trigger system. It is distributed synchronously with the transfer of data over the trigger and timing links and the readout links.

The periodically occurring heartbeat triggers (HB); split the stream of continuous data into the heartbeat frames. Heartbeat identification is tagged with each frame. The data are aggregated, multiplexed and compressed in the CRUs and shipped to the memory of the first level processors (FLP). Each FLP combines several data streams and buffer them in memory. The CRU will send a acknowledge message to the CTP which contains the information about the successful delivery of heartbeat frame data or trigger data to the FLP and the occupancy status of the data buffer of CRU.

The heartbeat frames are collected into sub-time frames during a time period of around 22ms. The time frame duration of 22ms is selected to reduce the incomplete data at the boundaries of the time frame [17]. Sub-time frames are produces by all FLPs. It could be empty for the FLPs which receive data from the triggered detectors, inactive for the mentioned time duration. The sub-time frames are then sent to the Event Processing Nodes (EPNs) for aggregation. The sub-time frames from all the FLPs and related to the same time period are received by the same EPN and collected into a complete time frame (TF). The data will be processed in the EPNs synchronously with data taking. The data volume will be decreased by processing the data on the fly in the CRUs and EPNs and not by excluding the complete events. The original raw data will be replaced with the compressed data after the O2 system will perform a partial calibration and reconstruction. The produced data at this stage will be temporarily stored in the O2 system.

In order to achieve the required quality of data, a stage of second reconstruction will be performed asynchronously using the final calibration. All the FLPs, EPNs, networking and data storage together denotes the O2 facility. It is located at the experimental area at ALICE. O2 will also provide the interfaces with the Grid systems and the permanent data store at the Tier0 [37].

2.6 Summary

ALICE has been recording data for Pb-Pb, p-p and p-Pb collisions since the beginning of the LHC program (year of 2008). The LHC beam luminosities will be progressively increased, in order to extend the physics reach of installed experiments. With this the data rate will achieve an order of ~ 3.3 TB/s; which will be handled for the first time in ALICE. To cope with it; an upgrade is planned from the year of the 2019. The specifications of ALICE for Run3, scheduled in the year of 2021 are determined by the concept of reading the full detector information. This need an upgrade of sub-detectors their readout electronics, the trigger system, detector data links and a new approach for data acquisition formed on FPGA based CRU is adopted. In this chapter an overview of the present ALICE detector and the electronics till Run2 is discussed; along with the measures that will be taken to handle the high data rate of the Run3 are discussed. The chapter summarizes about the need of the CRU in the experiment, and also elaborated the readout scheme of ALICE system utilizing CRUs. In the next chapter; there is a focus on the design aspects of CRU, its features and the functionalities are discussed.

Chapter 3

ALICE readout system and Common Readout Unit (CRU)

3.1 Introduction

The major challenge for the data acquisition system in a HEP experiment is to transfer the data with high reliability between the different sub-detectors situated in the harsh radiation zone to the acquisition electronics and the data storage located in the non-radiation area. Also during the consecutive upgrades of the experiment it is essential to preserve the smart features of the present data acquisition system like having common interfaces between the various sub-detectors and the common online computing farm. In the present system during Run2 of ALICE, and also after the consolidations and upgrades during LS1, this had been ensured by data transmission through two generations of DDLs; DDL1 and DDL2. The present DDL links are fed by on-detector read-out concentrator units [27]. These on-detector concentrator units read out the ASICs of several front-end electronics cards typically on parallel electrical buses (copper flat cables). However this parallel electrical read-out technology has reached its limits, and the required new front-end cards/ASICs with higher performance are being developed with serial data transmission interfaces. As

a consequence, instead of sharing parallel buses, the new front-end electronics will have much higher number of serial data outputs. Thus for the higher data rate from large no. of detector channels and the continuous read-out requirements; especially for the re-designed detectors [38] during Run3, the updated approach of the third generation of links will be required with a significantly higher throughput in the form of CRU.

Thus, the main architectural role of the new CRUs is inherited from the earlier on-detector parallel read-out concentrators [26, 28], viz. they will have to read-out the very large number and higher bandwidth serial detector side front-end links, process and multiplex their data to common and even higher bandwidth server-side links. The hardware of CRUs is common to all the constituent sub-detector systems in ALICE however the sub-detector specific user logic/firmware varies as per the requirements. The versatile functionalities need CRU to be implemented as a custom designed electronics boards with re-programmable functionalities based on state-of-the-art FPGA technology.

In this chapter; the focus is on the design strategy and implementation aspects of CRU. Its different interfaces, qualification of the CRU harware and FPGA, the features and functionalities are detailed. The uniqueness of CRU, its complexities involved and the measures to handle are discussed. Section 3.2 discusses the design strategy, the location of CRU board in the experiment and elaborates on the implementation of scheme and its advantages. In section 3.3 features and the functionalities of CRU are explained. Section 3.4 presents the various communication standards used for the design of interfacing with different constituents of the system. Section 3.5 presents about the selection parameters of the FPGA and the hardware for the implementation of CRU. In Section 3.6 the complexity of the CRU hardware, peculiarities of design requirements and the resolutions are discussed.

3.2 CRU design strategy

In HEP experiments, because of the radiation hard requirements, there are three ways data can be readout and sent to the DAQ.

1. Using all radiation hard ASICs which will have dedicated and limited functionality without any versatile functions

2. Using all FPGAs, but radiation hard FPGAs are not yet available with as high performance as compared to the non rad-hard SRAM based FPGAs,

3. Using ASICs in stringent radiation environment and FPGAs for the rest of the part.

The application of CRU discussed concerns the third option which is presently being worked out for the experiments at the LHC. In this case, three major steps are involved. At first stage the data from the detectors (placed in radiation hard zones) are transferred to the Front End Electronics (FEE) through the detector backplane using dedicated ASICs. At second stage, data from the FEE are transferred by optical fibres to low or non-radiation zones and processed using FPGAs (CRU). At final stage, the processed data are further transferred to the DAQ (aka O2).

The most demanding features for the data processing in the HEP experiments involve handling the high data rate, error resilience, reusable modules with compact hardware for portability, quick upgradation and efficient data aggregation [39, 40]. Motivated by such requirements HEP experiments follow a hierarchical readout scheme [41] for HEP experiments. The detector readout system is broadly divided into two parts; FEE and the data processing unit (CRU in the case of ALICE experiment). FEE are located in the radiation zone with proximity to the detectors requiring custom built radiation-hard electronics. A FEE consists of front end module and the optical conversion module. At the first stage, FEE receives the data from the detectors. It amplifies, integrates and shapes the weak sensor signals over a given period, and provides robust signals to be transmitted off from the detector [42]. The optical conversion module consists of GBT chipset [43, 44] with radiation-hard

optical link ecosystem. Radiation tolerant GBT chipset is used for packaging of detector data and transmitting it to CRU in GBT standard at 4.8 Gbps using point to point optical link [45, 46]. Optical conversion is needed to communicate the high speed data over long distances with less channel noise and low power consumption. The digitized data received at CRU over optical links are multiplexed, processed and formatted as per the sub-detector specifications before being forwarded to the back-end computing nodes.

The physical location of the CRU in the readout chain is one of the major factors that affect the selection and design of the hardware. The design of the CRU as FPGA boards has two basic alternatives depending on the conditions determined by the physical location of the CRU in the readout chain. It can be located either near the individual sub-detectors, where there is certain level of radiation as the conventional approach or far from the experimental site in the controlled or no radiation zone as the proposed approach.

3.2.1 Version A: CRU in the Cavern

In the case of CRU located in the radiation zone, the FPGA based boards will be mounted directly on the detector structure in radiation, close to the detector front-end electronics. The power supplies require cabling inside the detectors like the present read-out concentrator units. The electronics design will have to fulfil radiation tolerance requirements, limiting the designers to use only certain type of FPGAs [47]. In this case the detector side links are relatively cheap, short board-to-board serial electrical links, while on the server side, the DDL3s have to be designed as long fibre optic links up to 200 m; as shown in the figure 3.1.

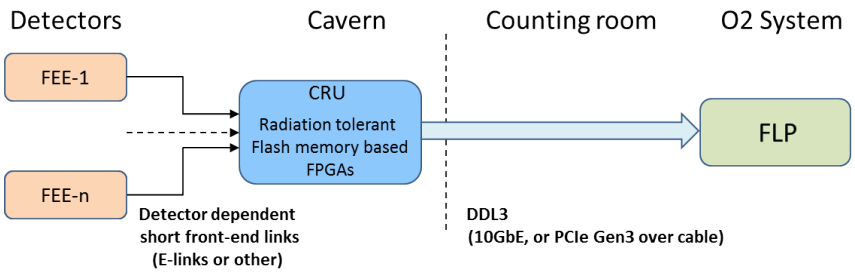


Figure 3.1: Implementation version A: CRU as FPGA boards in the cavern.

3.2.2 Version B: CRU in the Counting Room

The CRUs may be located in the counting rooms as the second option, close to the FLP server computers and in a non-radiating environment. In this case the CRU can be designed as as PCI Express plug-in cards placed directly onto the motherboard slots of the FLP server computers, or even as FPGA cards placed in industrial standard electronics racks or chassis, that will supply them with power [48]. This option also gives the possibility of using of one of the non-radiation tolerant, state-of-the-art high performance FPGAs. This solution requires large number of long, deterministic latency fibre optic links (GBT links as the custom developed solution for the HEP experiments) on the detector side. On the other side, the short-distance high-throughput server side links may be designed by a wider range of industrial standard technologies.The server side optical links, as an option, are omitted by a direct I/O bus connection, implementing the CRU as a Gen3 PCI Express plug-in card in the FLP computers. As a baseline location of CRU would be in

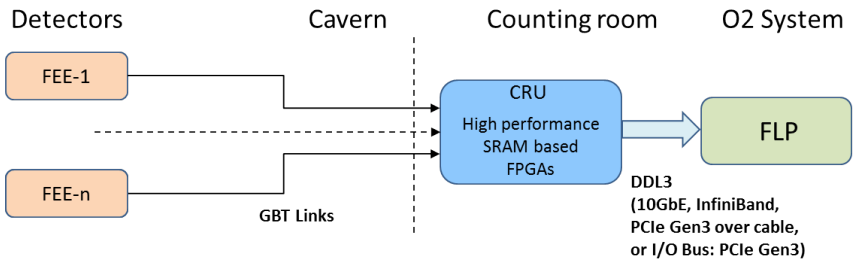


Figure 3.2: Implementation version B: CRU as FPGA cards in the counting room.

the ground level counting room, thus it will be accessible during operation. Also there will be no additional electronics, cabling, cooling needed to be installed and maintained on the detector. The GBT link based CRU located in the counting room presents a more robust system with higher processing power, flexible towards future requirements and lower impact on the cavern infrastructure. Therefore, CRU in counting room far from the radiation zone is chosen as the preferred location. The advantages of the proposed approach (CRU in counting room) over the conventional approach (CRU in cavern) against the critical design parameters and their implications for the CRU technology, available ecosystem and ease of

maintenance are listed in the Table 3.1.

Table 3.1: Advantages of the Proposed approach over the Conventional approach

Parameter	Conventional approach	Proposed approach
CRU location	Experiment area: radiation environment	Counting room: Controlled or no radiation
CRU Technology	Radiation hard electronics	Readily available components
FPGA	Radiation hard such as ACTEL or MICROSEMI FPGAs, Flash Memory based low performance,	Non-Radiation Hard Intel or XILINX Static RAM based, high performance FPGAs
Logic resources	Triple Modular Redundancy or voting logic, Low packing density of logic cells	logic redundancy not required, Densely packed logic cells
Radiation Campaign	Rigorous radiation tests to qualify the components	Not required
Cable lengths: (a)Between FEE and CRU (b)Between CRU and Back-end computing	(a) Short cables (b) Long Radiation tolerant link	(a) Long optical links, (b) short connection
Availability of ecosystem	Limited solutions, less choices for component selection	Ample solutions and more options for components selection
Impact on cavern infrastructure	High	Less or no impact
Accessibility, Maintenance, flexibility and adaptability	No or limited access, Difficult maintenance, Less flexible and low adaptability for upgrades	Easy accessibility Ease of maintenance. Highly flexible towards the future upgrades.
Cost	Relatively High	Advantageous over the conventional approach

3.3 Features and functionalities of CRU

CRU acts as the common interface between the on-detector systems/front-end electronics, the online and offline computing system and the central trigger processor. It shall be designed in the way, that the legacy and the new links shall provide compatible data formats and control interfaces in order to build a homogenous online system for the detectors of the experiment. The CRUs have to fulfil the various system level requirements; depending on that the features and functionalities of CRU are summarized below:

CRUs is in between the re-designed front-end electronics of the upgraded sub-detectors,

the O2 facility and the Detector Control System (DCS via the O2 facility), as well as the CTP via the Local Trigger Unit and the Trigger Timing and clock distribution System [49]. It aggregates data streams of the detector segments for the O2 system by multiplexing the data from a large number of high bandwidth detector side optical links to even high-performance common links on the server side. This increases the throughput and also optimizes the system level cost. As a common interface between the different systems in ALICE; the digitized data sent to the CRU are multiplexed, processed and formatted depending on the sub-detector specifications before being forwarded to the back-end computing nodes. CRU performs the required protocol conversions, framing, deframing, multiplexing and demultiplexing of data, control and trigger information between its different interfaces. The CRUs have a *deterministic latency* interface to the trigger system and to the detector FE electronics.

The other main system level role of CRUs is to reduce the number of physical links presently used for the data read-out, detector control, trigger and clock distribution, etc., by combining the data, control and trigger communication between the different nodes of the system i.e. between the CRU and the detector front end cards, between the CRU and FLP computers; wherever it is possible. Also, in the current run at ALICE experiment; data is transferred from the detector systems to a detector specific readout unit like CROCUS for PMD and Muon chambers [9], RCU card for the TPC detector etc. and from there it is sent to the DAQ at the rate of 1-Gbps by PMD and about 4-Gbps by TPC. However for Run3 all the intermediate detector specific units for the upgraded sub-detectors will be replaced by a congruent hardware viz. CRU. Hence CRU will also reduce the number of different link technologies presently used in the experiment. It will provide a homogeneous online system for the detectors and ease of maintenance and service.

The hardware design of CRU and board layout is identical for different sub-detectors under consideration. However, depending on the physics objectives there are sub-detector specific functionalities of CRU like number of optical links and data to be handled. Each sub-

detector has different requirements for the number of CRUs, the number of transceiver links on-board as shown in Table 3.2, and requires specific user logic/firmware. These application specific functionalities e.g. the protocol conversions of the different link types, multiplexing of data flows, embedding detector control and trigger information, etc., require CRU to be implemented as electronics boards with custom designed, programmable functionality based on up-to-date FPGA technology.

Table 3.2: CRU requirements in the system

User	Links needed (TX/RX)	No. of CRU cards	Available TX/RX per CRU	Used TX/RX no. of links connected	Links configurations for the CRUs
CRU team		04	24/24		
O2 team		10	24/24		
TPC	3276/6552	360	12/24	various	36 times (7/14, 8/16, 9/18, 9/18, 9/18, 9/18, 10/20, 10/20, 10/20, 10/20
MCH	624/624	32	24/24	various	8x 24/24, 4x 16/16, 4x 10/10, 8x 19/19, 8x 22/22
MID	32/32	2	24/24	16/16 for all	
TOF	72/72	4	24/24	various	2x 16/16, 2x 20/20
FIT	25/25	2	24/24	various	1x 19/19, 1x 6/6
ZDC	10/10	1	12/12	10/10 for all	
ITS	192/576	24	12/24	8/24 for all	
MFT	80/240	10	12/24	8/24 for all	
ACO	1	1	12/12	1/1 for all	
TRD	1044	36	0/36	0/29 for all	
CTP	21	1	24/24	21/21 for all	3/3 from CTP, 1/1 from each 18 LTUs

(Courtesy: Tivadar Kiss)

Data of the order of ~3.3 TB/s flows from the detectors to the online computing system. This huge data will be handled for the first time by ALICE experiment. To tackle this, the data are distributed over many CRUs (eg: 360 for TPC detector). A CRU could aggregate maximum 48 numbers of high speed input optical links however the actual number of links to be integrated depends on the detector design and requirements. Actual usage depends on the specific sub-detector eg: TPC will utilize 20 such readout links which corresponds to maximum aggregate throughput of 20*4.48 Gb/s = 89.6 Gb/s.

The system level requirements can be translated to the following requirements of the different interfaces of the CRUs viz. Detector side interface, Trigger interface and Server

side Interface.

3.4 Interfaces of CRU: Different Interfaces used, its selection and survey

Different high speed transmission protocols are implemented for the three interfacing links viz. Data link, DAQ Link, Trigger Link. Design of these links at ALICE requires custom development along with the other components because of the unique 40 MHz clock frequency of the LHC and not compatible with many readily available options. The system level requirements of the three interfaces and the selected protocol for the different interfaces of the CRUs are discussed.

3.4.1 Detector side interface

The data transmission links from the detector need to be upgraded to higher bandwidths for the Run3, in order to cope with the huge data being produced by the detectors. It is necessary to use lesser number of optical links at higher data rates rather than increasing the number of links. This helps to raise the bandwidth without paying a penalty on the material budget of the detector. The HEP detectors operate in the harsh radiation zone therefore various readily available data transmission protocols cannot be used as detector side interface with CRU. The robust error correcting code to protect the data upset due to the radiation effects is not present in commercial protocols for transmission. Also the data needs to be transmitted with a constant and deterministic latency.

The previously existed E-link protocol [50] is not suitable for this implementation. A radiation tolerant communication link is required between on-detector front-end electronics situated in harsh radiation environment to the CRU placed in a low radiation zone. To address this challenge, an efficient high speed error resilient Gigabit transceiver optical link is designed on FPGA. The detector front-end is read out by high-speed 4.8 Gbps optical

transceiver board with Gigabit Transceiver ASIC (GBTx) followed by CRU. However as per the treaty on the Non-Proliferation of Nuclear Weapons, whose objective is to prevent the spread of nuclear weapons and related technology; the GBTx ASIC cannot be made available at the Indian labs. Due to the non-availability of rad-hard GBTx ASIC; an emulator known as GBT FPGA logic core is developed. Radiation hardness is required near detectors, however it is not necessary for the CRU located away from the radiation zone; this feature is utilised to mimic the GBTx functionality on the non-radiation hardened SRAM based FPGAs. It is used for data, trigger timing and control distribution merged on a single data channel. This enable the CRUs for receiving the GBT datagram and also transmitting the control and timing signals from control room to the detectors. The GBT transmission is an asynchronous serial communication that is composed of a GBT transmitter, a multi-gigabit transceiver and a GBT receiver. The main functional custom developed blocks of the core are scrambler, error correcting block, interleaver, Gear Box for smooth frequency translation by modifying data bus width for clock domain crossing, and the transceiver. The GBT receiver performs descrambling, decoding and deinterleaving. In addition receiver has one more block of frame aligner. The frame aligner performs header detection and locking for frame synchronisation using an efficient pattern search algorithm to maintain synchronisation between the transmitted and the received data. Pattern generator and checkers are used for testing purpose, and they are replaced with buffered data for front-end acquisition. It is a radiation tolerant, error resilient and bidirectional optical link data communication standard having fixed latency support for high energy physics experiments. GBT link aims at providing simultaneous transmission of physics, trigger and experiment control data over the same link. The GBT ensures fixed, low and deterministic latency, and thus is ideal for use as the data link of HEP detectors.

The link rate of *4.8 Gigabit/sec* for GBT protocol is composed of *40MHz x 120 bits*. Frequency of 40 MHz is derived from the LHC bunch spacing time of 25ns [51] and 120 bits is technology parameter [52]. If FPGA transceiver reference clock is fixed at the 40 MHz

then the minimum input data width of the transceiver is 120 bits, so the GBT protocol with the link rate as 4.8 Gbps is chosen. Details of the GBT protocol are given in Table 3.4. Its implementation and testing on FPGA are discussed in subsequent chapter.

3.4.2 Server side Interface (aka DDL3)

The interface from the CRU to the back-end computing nodes requires a standard interface having large bandwidth with a multi-channel support and a proven transmission capability. For this purpose a custom designed or an available link could be used. Since the CRU will transfer data to a DAQ server, so a custom defined protocol for high speed link becomes difficult to maintain with poor future support. Hence, different high speed serial links options were explored that have large ecosystem with ample solutions [53, 54]. The links from CRU to the back-end computing nodes are not latency critical as the data packets have already been time-stamped. The two latest promising technology options available for this high speed interface are PCI express protocol and 10 Gigabit Ethernet (10GbE) protocol [55, 56, 57, 5]. However, for each experiment with its distinct set of requirements, it is not feasible to adopt a readily available solution. Hence it is best to perceive a standard high speed protocol and adapt it as per our requirements of HEP. The two most tangible options as mentioned are compared concerning the design requirements like the form factor, legacy support, ease of upgradation, flexibility and cost are listed in the tabular form in Table 3.3.

The design of the server side links highly depends on the hardware chosen for the implementation of the CRU as discussed in section 3.5. The ALICE CRU architecture was evolved based on the advance mezzanine card (AMC40) hardware developed in the framework of LHC-b read-out [58]. The system is based on the ATCA crate standard and 10GbE solution. However, with the evolution of the project requirements; the system shifted to the PCIe based custom developed FPGA card. The PCIe interface (Gen3 x16 lanes) is a viable solution chosen for the sever side interface of CRU. This architecture is compact

Table 3.3: Comparison table between PCIe interface and 10 Gigabit Ethernet interface

Comparison Parameter	10 Gigabit Ethernet Standard	PCIe Standard
CRU Form factor	ATCA or microTCA form factor	PCIe form factor
Legacy support	Good, Form factor and the network components allow backward compatibility.	Poor, Backward compatibility not maintained by PCI-SIG group
Ease of Upgradation	Retaining the key Ethernet architecture. Efforts optimized and reduced time of development	Less flexible for upgradation, incompatible with older systems, high cost and more development time
Flexibility	More number of cards installed in a single server PC using network switches It can be utilized as a fabric, however it has substantial hardware and software protocol processing requirements.	Only single CRU could be installed in one slot of the server PC. It is not a fabric and can only support the connectivity of small numbers of processors and/or peripherals. It serves as a bridge to fabric
Connectivity solutions	Ample peer to peer connectivity solutions	Natively PCI Express does not support peer-to-peer processor connectivity. Topology limitations.
Routability	Ethernet switches allow packets to be routed	Not a routable protocol; It defines a single large address space that devices are mapped into.
Line rate	10.312 Gbps	8 Gbps per lane (Gen 3 PCI Express:128Gbps)
Line Coding	64b/66b (3.125 percent overhead)	128b/130b (1.54 percent overhead)
Latency	Extra stage of communication to move data to or from processor memory Latency of Tens of microseconds	Tightly integrated with the memory subsystem in a system on chip device. Latency of Sub microseconds
Cost	Low cost of the network components and their backward compatibility reduces the cost of installation and maintenance.	Complex switching devices and proprietary solutions make it costly

with no optical links and cabling. PCIe being used as a chip-to-chip interconnect and is a growing up technology. It has evolved gradually and built a huge ecosystem in its existence. Both price and power for per Gbps transmission is lower without compromising with performance when compared to the 10GbE [59]. It is a low overhead protocol. Data loss and corruption in PCIe-based storage systems, are highly unlikely as no involvement of upper layers of the OSI model [60]. Although the designers had migrated to PCIe [55], still 10GbE has got future proof solutions. 10GbE is a system-to-system connection and already a mature technology with legacy support. 10GbE can be optimized for detector specific data and the Quality-of-Service is provided in the higher layers of Open Systems Interconnection (OSI) model [61]. The ability of the two technologies co-exist and complement one another. Hence, the 10GbE standard interface is always a good alternate for the high speed link to the computing system as summarized in Table 3.3. The CRU is based on PCIe Gen3 x16 lanes. This provides the freedom to the experimentalists to process the data and use the switching schemes on the PCIe based backplane nodes. Here processors in the backend should have multi-core CPU, and provide power and mechanical support to multiple PCIe

cards. The test results are shown in the subsequent chapters.

3.4.3 Trigger interface

The trigger distribution system for the ALICE upgrade using the CRU is a combination of multi-link technologies involving different protocol standards. The trigger interface to CRU is based on the Timing-Trigger and Control over passive optical networks (TTC-PON) [62] and the Giga-Bit Transceiver optical link (GBT); in conjunction to communicate the TTC information from the CTP to a detector through the CRU electronics system. TTC-PON standard is chosen for this interface as it has the high split ratio having fixed latency. TTC-PON protocol satisfies the strict timing mandate of 27 Km of LHC diameter unlike the limitations of the formerly presented solutions. Both GBT and TTC-PON are custom developed protocol standards to support the extreme experimental conditions at LHC.

In TTS-Fast serial Trigger Links (FTL), trigger is routed via CRU placed in counting room to feed the detectors in the cavern. The signal route involves two asynchronous serial link transitions (TTC-PON and GBT), each operating in fixed latency mode [63]. The downstream from CTP to CRU uses TTC-PON, which is time multiplexed trigger distribution topology, operates at the rate of 9.6 Gigabit per second and satisfies the timing specification of the LHC. The downlink path from CRU to LTU involves GBT protocol. The total routed path delay is to be 1.5 us for 150 m (= 75 m x 2) cable length. TTC-PON provides the LHC timing in proper phase for the on-detector electronics over the GBT downlink. Comparison for the features of different protocols, regarding the technology specification of the interfaces which are already used at the LHC experimental sites and the most tangible protocols as per the requirements for the three interfacing links viz Data link, Trigger Link and DAQ Link are summarized in Table 3.4.

Table 3.4: Detailed comparison of the specifications of the high speed interface links used in high-energy physics experiments.

Parameters	**GBT** [64]	**Mitra et al.** [65]	**TTC-PON** [62]	**10Gigabit Ethernet** [62]	**PCIe** [66]	**PCIe over fibre** [67]	**INFINIBAND** [54]	**(FC) Fibre Channel** [68]
Technology Specification	Custom	Custom	XGPON1 with modifications	802.3ae Specification Standard	PCI-Base specification	PCIe over IP	InfiniBand Trade Association	T11 Committee with rules from ANSI
Designer Group	CERN	NA	ITU-T with CERN modifications	IEEE	PCI-SIG	PCI-SIG	InfiniBand Trade Association	T11 Committee
Line Rate	4.8 Gbps	8.192 Gbps	Downstream: 9.6 Gbps Upstream: 2.4 Gbps	10.3125 Gbps per lane	2.5/5.0/ 8.0/16/ 32 GT/s	2.5/5.0/ 8.0 GT/s	14 Gbps	4X28.05 (max)
Payload Rate	3.2 Gbps	3.84 Gbps	Downstream: ~7.68 Gbps Upstream: 640 Mbps	10 Gbps per lane				
Payload Size	120 bits @40 MHz	128 bits	Downstream: 192 bits@40 MHz Upstream: 16 bits@40 MHz	64 bits@ 156.25 MHz	4 Kbytes	4 KBytes	4KBytes	2KBytes
Wavelength (nm)	850 nm-Tx (Multi-mode) 1310 nm-Rx (Single-mode)	850 nm (MM Tx) 1310 nm (SMTx)	Downstream: 1577 nm Upstream: 1270 nm	850 nm (10 Gb BASE-SR)		SFP+ modules (10 Gb Ethernet compliant)		SM (1270nm, 1290nm, 1310nm, 1330nm, 1490nm, 1550nm) MM (850-nm)
Network Topology	Point-to-Point	Point-to-Point	Point-to-Multipoint	Point-to-Point	Point-to-Point	Point-to-Point	Point-to-Point bidirectional serial links	1. Point-to-Point 2. Arbitrated Loop 3. Switched Fabric
Encoding	RS ECC with Block Interleaver	Golay ECC with Helical Interleaver and AES Encryption	8b/10b	64b/66b	128b/130b in Gen3	128b/130b in Gen3	8b/10b or 64b/66b	8b/10b or 64b/66b
Synchronous Trigger Support	Yes	Yes	Yes	No	Edge triggered	Edge triggered	Yes	Yes
Trigger Latency	~150 ns (Optical loop-back)	~703.125 ns (Optical loop-back)	~100 ns (Downstream) ~1.6 us (Upstream)	Variable (depends on Application layer protocol)	–	–	<100ns (shortest possible round trip latency)	

3.5 Selection of FPGA and the CRU hardware

Selection of FPGA

Why it is important? The selection of FPGA is important not only for high speed data transfer but modern acquisition systems use complex processing algorithm for purposes like clustering, feature extraction and data aggregation among others. Different detectors at ALICE like TPC also implement the data processing logic in CRU FPGA which require large number of resources. To design such algorithms sometimes arithmetic logic units present in general purpose processor may not be sufficient. With increase of the number of channels,

more datagrams enter into the DAQ in parallel fashion and a general purpose processor serializes them prior to processing. Use of FPGA devices give freedom to the designer for on-board data processing. On-board data processing scheme processes data in different level of hierarchical DAQ network and only required data will be forwarded to the computing nodes. This reduces the data processing load in the backend computing nodes and helps to distribute the processing loads throughout the DAQ network. The effciency of the system enhances manifolds if processing could be carried out parallely in DAQ. In the present era of embedded system, FPGAs are suitable to address the above mentioned problem due to presence of huge number of digital signal processing blocks which are equivalent to ALU for general purpose processors. These huge number of logic blocks support the implementation of various complex algorithms and process data in parallel. The on field programmability, modularity makes the FPGA more suitable for HEP system.

The selection of FPGA variant for the design of CRU hardware and DAQ firmware development is constrained by various factors like the availability of logic resources, High Speed Serial Interface (HSSI), Serializer-Deserializer (SerDes) on FPGA, market availability of FPGA etc. Different FPGA families are compared against various crucial parameters like the available logic resources, transceivers, Phase lock loops (PLL) and market availability for the choice of FPGA chip on CRU as listed in the Table 3.5.

Table 3.5: FPGA selection parameters

FPGA Family Name	Intel Stratix-V GX	Intel Stratix-10	Intel Arria-10 GX	Xilinx Virtex-6	Xilinx Vertex-7	Xilinx Virtex Ultrascale
Status	Available	End of 2017	Available	Available	Available	Available
FPGA part number	5SGXEA7	10SG280	10AX115	XC6VLX240T	XC7VX690T	XCVUI90
PLLs	28	48	32	12	20	60
>=10Gb/s Transeivers	48	144	96	24	80	60
Logic Elements/cells[M]	0.622	2.8	1.15	0.241	0.693	1.9
LUTs[M]	0.235	1.8	0.425	0.15	0.433	1.07
FFs[M]	0.939	7.4	1.7	0.3	0.866	2.14
18/20Kb RAM Blocks	2560	11721	2713	832	2940	7560
Total Block RAM(Mb)	50	229	53	15	53	133
PCIe x8,Gen3	4	6	4	2(Gen2)	3	6
Used for developing	AMC40 card		PCIe40 card	C-RORC board	MP7 card	

High speed transceivers integrated on the silicon in the high-end FPGAs form the base of the transmission system. The quantitative effects of Moores Law have driven qualitative

changes in silicon technology and FPGA architectures with optimized design productivity. The entire track of signal processing for digital communication and the high speed transceivers along with SerDes are packaged on FPGA chip. The result of the process node shrinkage are logic operations at high density and high data rate while maintaining lower BERs [69]. It gives an impetus to the shift of paradigm towards the FPGA based development of DAQ systems for HEP experiments [41]. The SerDes embedded in FPGAs are designed for non-deterministic transmission latency. However, an affirmed deterministic delay in data transmission is assured with the precise design of the protocol architecture and study of the clocking scheme configuration; as it is designed in the GBT protocol [30] and the development of Mitra et al. [65] as an example. The implementation method for the efficient FPGA firmware design in the development of DAQ depends on the FPGA device family used for the purpose. It is due to the differences in the transceiver architecture and the internal clocking schemes. Selection of the FPGA family with the required transceiver switching speed and architecture is a crucial parameter and time taking survey before designing the high speed protocols and the interfacing links. To ease the procedure for the designers, an extensive survey is presented regarding the silicon technology adopted by the various FPGA vendors, popularly available device families and the maximum SerDes speed is shown in Table 3.6.

Consequently the specifications of Arria-10 GX FPGA [70], the support from Intel-Altera and the timeline of its market availability are matched with the CRU project production and delivery deadlines. 20nm Arria10 is the mid range family of Intel-Altera FPGA, however Arria-10 has been chosen as it has enough available resources and meets the need of the TPC specific logic occupancy. The TPC sub-detector at ALICE is the major user (>75 percent) of the CRU boards.

Table 3.6: FPGA device and its family with maximum SerDes speed

FPGA Vendor	**Intel-ALTERA**		**XILINX**		**ACHRONIX**		**MICROSEMI**	
Silicon Technology	*Device Family*	*SerDes Max Speed*	*Device Family*	*SerDes Max Speed*	*Device Family*	*SerDes Max Speed*	*Device Family*	*SerDes Max Speed*
65 nm							IGLOO2	5 Gbps
							Smart Fusion2	5 Gbps
60 nm	Cyclone IV GX Variant	3.125 Gbps						
45 nm			Spartan 6 LXT	3.2 Gbps				
28 nm	Cyclone V		Artix 7	6.6 Gbps			PolarFire	250 Mbps to 12.7 Gbps. Optimized at 12.7 Gbps
	GX Variant	3.125 Gbps						
	GT Variant	6.144 Gbps						
	Arria V		Kintex 7	12.5 Gbps				
	GX Variant	6.5536 Gbps						
	GT Variant	10.3125 Gbps						
	GZ Variant	12.5 Gbps						
	Stratix V		Virtex 7					
	GS / GX Variant	14.1 Gbps	GTX Transceiver	12.5 Gbps				
	GT Variant	12.5 Gbps	GTH Transceiver	13.1 Gbps				
	GT Variant	28.05 Gbps	GTZ Transceiver	28.05 Gbps				
22 nm					Speedster 22i HD			
					(HD680, HD1000)	12.75 Gbps		
					(HD 1500)	28 Gbps		
					Speedster 22i HP			
					HP360, HP560	12.75 Gbps		
					HP560	28 Gbps		
20 nm	Arria 10		Kintex Ultrascale					
	GX / GT Variant	17.4 Gbps	GTX / GTY Transceiver	16.3 Gbps				
	GT Variant	25.78 Gbps						
			Virtex Ultrascale					
			GTH Transceiver	16.3 Gbps				
			GTY Transceiver	30.5 Gbps				
16 nm			Kintex Ultrascale +					
			GTH Transceiver	16.3 Gbps				
			GTY Transceiver	32.75 Gbps				
			Virtex Ultrascale + 32.75 Gbps					
14 nm	Stratix 10							
	GX Variant	17.4 Gbps						
	GXT Variant	30 Gbps						

Hardware for the design of CRU

The ALICE CRU hardware framework was advanced in view of the AMC40 data acquisition board developed for LHCb read-out as one of the suitable option. The AMC40 technique was based on the ATCA crate standard [48]. AMC40 had 36 bi-directional optical connections each with a bandwidth upto 10Gbps. However, with the evolution of the requirements; the system shifted to the PCIe based custom developed FPGA card. The next and the advanced execution option of the CRU framework depend on a solution where the CRU is directly plugged in to the PCIe slot of the server of the O2 system. In this the PCIe bus

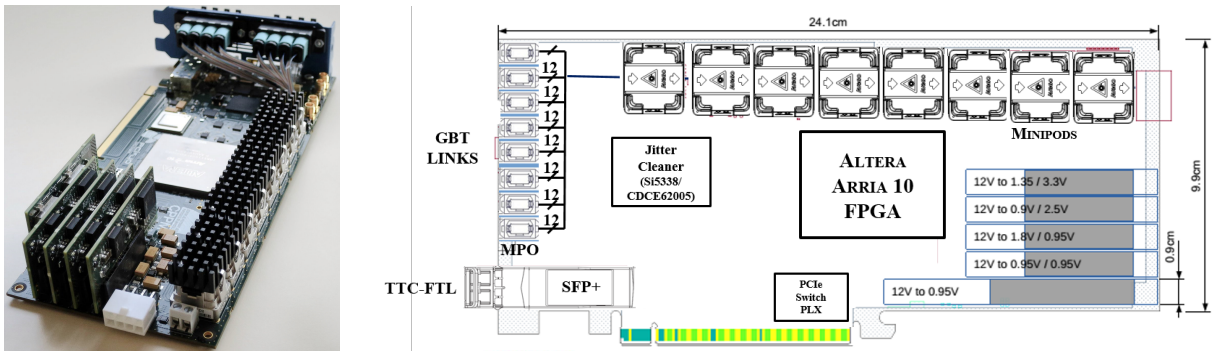
serves as the DDL3 and no optical associations are required to be introduced. Also vertical migration of the firmware is possible from the Arria-10 FPGA to the next higher series of Altera FPGA known as Stratix-10.

After thorough evaluation; the ALICE collaboration made a joint venture with LHCb experiment at CERN for the custom designed hardware known as PCIe40 DAQ engine [35] as shown in Figure 3.3. For ALICE experiment, the boards are developed and tested at India and the results are shown in chapter 6. It uses Arria-10 FPGA on-board. This is the largest one in the Arria-10 family with 1150 kLE. This FPGA is also pin compatible with the Stratix-10 FPGA of next higher series, with the same amount of cells which paves the way for a quick future upgrade. Compared with the Stratix-V FPGA implemented on the AMC40 pins we have 1.8 times more cells than before. This makes the implementation more comfortable. The main features are the following listed in Table 3.7 The nomenclature of

Table 3.7: Important Specifications of Arria-10 FPGA

Number of Logical Elements	115000
Internal memory	8.4 MB
Variable precision DSP blocks	1518
18 x 19 multipliers	3036
Fractional PLLs	32
I/O PLLs	16
High speed serial links at 12.5 Gbits/s	72
Maximum speed on serial links	12.5 Gb/s
Maximum internal speed	500 MHz

the PCIe40 board arises from its design as PCIe card and the LHC frequency of *40 MHz*. PCIe40 DAQ engine is a PCI Express Gen 3 x16 lanes (8 Gbits/s for each lane) form factor based readout board. It is equipped with 130 high-speed optical communication links and consumes a peak supply current of up to 60 A. Each card contains the mezzanine power modules mounted on a motherboard. Both the motherboard and the mezzanine modules house a wide range of active and miniaturized passive components (0201 minimum) of several



(a) PCIe40 (b) PCIe40 Schematic

Figure 3.3: PCIe40 - candidate board for CRU development

different package types.

There are 08 no. parallel optical receivers and transmitters in total on each board. These optical transceivers modules are from Broadcom Inc. known as Minipod. Each minipod on PCIe40 card handles 12 no. parallel optical ribbon fibre cables custom developed by Sylex Inc. Hence there are 48 no. optical receivers links and 48 no. optical transmitters links, leading to 96 no. of high speed optical links in total. Each optical link has the signalling rate upto 10Gbps. CRU connects to the server through PCIe 16-lanes edge-connector (each lane is dual simplex channels using two differential signaling pairs hence 32 nos of high speed links). It has 01 no. bidirectional link for the trigger interface uses SFP+ form-factor optical module for establishing bidirectional optical communication. The board uses latest 20nm Arria10 midrange FPGA from Intel-Altera. The Arria family is considered as mid-range. However the Arria10 is currently the most powerful device from Intel-Altera, as tabulated in Table 3.5 and Table 3.6. PCIe40 board has high-performance, low jitter clock generator CDCE62005 and the Si5338 chip embedded on it, useful for clock recovery from LHC.

3.6 CRU peculiarites, complexities and the resolution

3.6.1 Design requirements

The complexities of CRU and its uniqueness; taking into account the requirements of the system and design of CRU with the required resolutions are summarized as follows.

The huge data ~3.3 TB/sec will be handled for the first time in the ALICE system. CRUs have to collect all the event from the FEE with high reliability and also has to process the data on the fly before being sent to the O2 system via PCIe. It has to perform online data reconstruction thereby reducing the data volume significantly. The data volume will be reduced by processing the data and not by rejecting the complete events. The most demanding CRU implementation will be for the TPC due to the high data rate and signal processing like base-line correction, zero-suppression and cluster finding and then the reduced data is sent further. It is required as the estimated data throughput to be written to the mass storage at its peak is ~90 GB/sec.

Each CRU is connected to a detector specific local trigger unit over a bidirectional link. The CRU recovers the LHC clock from the incoming trigger signal from the LTU with deterministic phase and also extracts the trigger information with deterministic delay then propagates it toward the front-end cards over the GBT downlinks. This provides the LHC clock for the front end over the GBT downlink. The recovered LHC clock output from the GBTx ASIC must have a constant and deterministic phase relation to the original LHC clock between each GBT links. It has to be same for all the CRU units, also during power-on/off cycles and firmware updates [71]. CRU accesses the front-end GBTx ASIC registers and the GBT-SCA [72] features over the GBT links and provides software tools to adjust the reconstructed LHC clock phase on the front-end card. The maximum acceptable jitter for LHC clock is 300 ps (RMS) only and the clock skew should not exceed 1 ns (peak-to-peak); requirement specified by TPC detector. Hence accurate PLLs on-board and proper firmware techniques are implemented for the processing and the clock recovery. CRU also propagates

the detector busy information back to the LTU.

The CRU firmware and software shall provide a generic interface which is accessible on the FLP server on Linux application level which allows any detector control system (DCS) related application to access the control related front-end functionality in a transparent way. CRUs provides formatted data packages to the FLP computers with standardized common data headers. The CRU firmware supports the remote firmware upgrade and functionality control. The level of complexity in firmware could be divided in two parts; Low level interface (LLI) and the user logic firmware for FPGA. Using the LLI, different on-board components and interfaces could be programmed and monitored like PLLs, on-board temperature sensors, the PCIe Gen3 x16 lane interface, the transceiver settings, GBT interface, the other slow-control interfaces etc. [73]. The different interfaces for the DCS are integrated via the O2 system.

The other part is the user logic or the detector specific firmware to be designed on FPGA. The firmware of CRU is quite a complex hierarchical structure and its development is a challenging task as the CRU acts as interface between the different ALICE sub-systems and the upgraded ALICE trigger system which supports both the triggered and continuous mode for read out of detectors. The data sent to the CRU need to be multiplexed, processed and formatted for transmitting it to the O2 system.

3.6.2 Hardware Complexity

The CRU hardware is one of the major aspects of this development. The PCIe40 boards are custom designed for the high speed data transmission requirements of the HEP applications, here some of the complexities and the steps to ensure the proper functioning are listed:

It has large number of transceivers on it; there are 96 readout links in total (48 optical inputs and outputs each, serial signal up to 10 Gb/s) for GBT connections on-board, provided per CRU, which is the highest count of links integrated in a data acquisition board when compared to the earlier used boards in ALICE. CRU has a HDI (High Density In-

Table 3.8: CRU Hardware Complexities

CRU Parameter	Description
HDI PCB	More than 1750 components on the board
Layer, Drills, Vias	14, laser drilling, blind, buried and stacked
Optical Minipods	48 nos high speed (upto 10 Gbps) bidirectional optical links, 4.8 Gbps for GBT
PCI Express	16 Lane GEN 3
High power requirement	Supported by Mezzanine cards
Material	High Tg, Low Dk material for high speed, Low Loss ISOLA 408 HR, TUC 872 Lk
Thickness	1.57 mm +/- 10%
Track Width	3 mil track
FPGA with High user Inputs/outputs(ARRIA 10 Device)	1932 pin BGA package (ARRIA 10 Device overview) with 768 I/O and 76 Transceivers, production grade silicon, 20-nm technology.

terconnect) PCB with more than 1750 components on it. The complexity of the hardware is summarized in Table 3.8. It is a 14 layer board with blind, buried and the stacked vias. Also there are laser drills required for its fabrication due to the minimum via size of 0.2mm on the PCB. Power requirements are supported by five mezannine cards.

Material requirements for the CRU PCB are also strict as CRU will be exposed to high thermal loads. Hence a high Glass Transition Temperature (Tg) PCB is required. Also Decomposition temperature (Td) is a measure of the degradation of the material. It is the point as which reliability is compromised and delamination may occur. Hence the PCB material chosen for CRU are ISOLA 408 HR/TUC 872 LK has the Tg ~ 200 degC and Td ~ 340 degC. CRU will aggregate the data from high speed GBT links at 4.8 Gbps each and will transfer to even higher speed PCIe Gen3x16 links (128Gb/s bandwidth) hence low dielectric constant (also called as relative permittivity) material is required to reduce the high frequency losses at high data rates. The dielectric constant (Dk) is closely related to the impedance of the circuits that will be fabricated on that material. Changes in a

PCB material's Dk, whether as a function of frequency, temperature, or other reasons, adversely affect the performance of broadband high-frequency high-speed digital as well as analog circuits because it will change the impedances of transmission lines. These unwanted changes in Dk and impedance result in distortion to the higher-order harmonics making up a high-speed digital signal, with loss of digital signal integrity. To ensure this the PCB material chosen is ISOLA 408 HR or TUC 872 LK with Dk $<$ 3.7 at 5 GHz, Df (loss tangent) $<$ 0.012 at 5 GHz.

For high speed applications track width is important. Track width requirement for the CRU PCB is 3 mil (1 mil = 0.001 inch). A 20 percent change in trace width can cause as much as a 10 percent change in impedance. As width increases, characteristic impedance decreases. To minimize crosstalk in high-speed interface implementations like CRU, the distance between two traces should be approximately 5 times the width of the trace. This spacing is referred to as the 5W rule. Control of trace width is both a function of process control by the board fabricator and to some degree the type of copper used on the base material. Hence an important point to be taken care during the development.

This PCB with all these complexities should only be 1.57 mm +/- 10 percent thick as it will be used in the PCIe slots of the backend servers. This board is much more complex with stringent requirements than the earlier DAQ boards used in the experiment. The Arria10 FPGA used on this board is one of the latest available FPGA with 20 nm technology from Altera. It is highly dense with large number of logic resources [70] which is required for detectors like TPC, so when all the links will be utilized all together the junction temperature could rise, this could change the trigger latency [71].

CRU as mentioned it is at the core of the data acquisition system through which the trigger, timing, control and data passes. Thus these boards has to be highly reliable as its failure during the run time of an experiment will lose some of the beam time before replaced, which is very much unwanted situation as the LHC beam is the result of many man-hours. The components on-board should be highly reliable. To ensure the reliability

of the boards several measures are applied. Although these are trivial and well known but worth mentioning for future references. The selected components should be RoHS compliant with lead-free soldering requirement. Reflow oven or vapour phase soldering technique should be used and soldering by hand should be avoided. To ensure the traceability of the components the date code, certificate of conformance and other specific references of each component used for the execution of the production are preserved. This allows to ascertain whether the whole production is affected or not, upon discovery of a faulty component during the tests. A strict discipline needs to be applied to protect all components in goods reception and storage areas. This is particularly important for surface mounted devices. The components shall be stored in conditions that minimize the growth of oxides on surfaces to be soldered and held in sealed bags or boxes containing dry desiccant or nitrogen purged desictors. In addition, before mounting, components shall be dried 24 hours in a stove. All items requiring component changes, re-soldering etc., shall be treated as untested and shall be passed through the entire inspection and test procedures again. Few other measures taken to ensure the reliability of this board are the firm producing and assembling the PCB must be ISO-9001 certified in the field of production and support of electronic modules. PCB production must be compliant to IPC-A600 and test certificate of compliance to this standard must be provided. Before the final delivery of the CRU, there are rigorous multilevel functional and electrical tests of the cards and discussed in chapter 6.

3.7 Summary

In this chapter the design strategy for the CRU in ALICE is well explained; taking into account the different implementation options. The location and its intricate issues are discussed. Location of CRU is an important aspect; as this effects the choice of the FPGA used, hardware design, component selection and the design of different communication in-

terfaces. Features and the functionalities of CRU are elabortaed. Different interfaces used for CRU; their selection and the survey is also presented. The choice of Arria-10 FPGA and the custom developed PCIe40 for CRU is detailed. The overall complexity of CRU in terms of design requirements, the peculiarities related to firmware and the hardware design; the measures to handle them are discussed. In the next chapter testing and performance analysis of different interfaces with main focus on the GBT link is presented.

Chapter 4

Link characterization and the signal integrity analysis

4.1 Introduction

The overall performance of the CRU depends on various factors and its constituents. The interfacing links of CRU plays an important role in this regard. This chapter presents the custom developed GBT-FPGA core firmware tests, its implementation on FPGA, back end interfaces and the signal integrity characterisation of the interfacing links. The performance of back-end interface links and the option of 10GbE as DAQ link have also been studied.

The chapter is organised as follows. Section 4.2 discusses about the flow of data through the CRU in the readout chain. Section 4.3 discusses about the features of custom developed GBT protocol and its implementation on FPGA. The strategy for clock calibration, explanation of the matter of contention and implementation of newly developed auto-initialization calibration logic for GBT is discussed in section 4.3.2. Different parameters are analysed as the indicators of efficient system performance. Firmware stability with temperature variation is measured. The various parameters of GBT are integrated using Intel-Altera system integration tool called as Qsys or the the Platform Designer tool. The Qsys model of GBT

is shown in section 4.3.3. Results of the FPGA logic resource utilization, estimation of power consumption, the critical path delays and the latency measurements are presented in section 4.3.4, section 4.3.5 and section 4.3.6 respectively. The signal integrity tests for the links are performed and presented in section 4.3.7 and 4.3.8. Back end interface of CRU is discussed in section 4.4. It presents the two most tangible options for the link viz. PCIe and the 10GbE solution for the interface. The calculation for the theoretical performance of the PCIe is given in section 4.4.1. Implementation of 10GbE on FPGA and the test setup is put section 4.4.2. Two models are presented in this implementation. Model-1 focus on the efficient method of high speed data transmission with minimum processor overload and the implementation of test system on Quartus II platform using Qsys. Model-2 focuss on the optical link testing. It presents an effective approach to address the challenges associated with the testing, performance monitoring and parameter tuning of optical interconnects in FPGA-based systems. The results of the performance evaluation of the two models are shown in section 4.4.3. Transceiver performance tests are dicussed in section 4.5. The resource utilization of FPGA for the integrated firmware is estimated in section 4.6 and the chapter is summarised in section 4.7.

4.2 Readout architecture and the interfaces

The CRU is a common functional block between the detector systems, O2 system and the trigger processor. The data flow scheme for CRU needs to be flexible enough to adapt to the demands of future detector and electronics upgrades.To address these challenges, we discuss an efficient DAQ scheme for error resilient, high speed data communication on state-of-the-art FPGA with optical links. The generic data flow of CRU setup with different communication links between on-detector FEE, CRU and the trigger system is shown in Figure 4.1, with the markings of major transactions of uplink and the downlink data during the transfer.

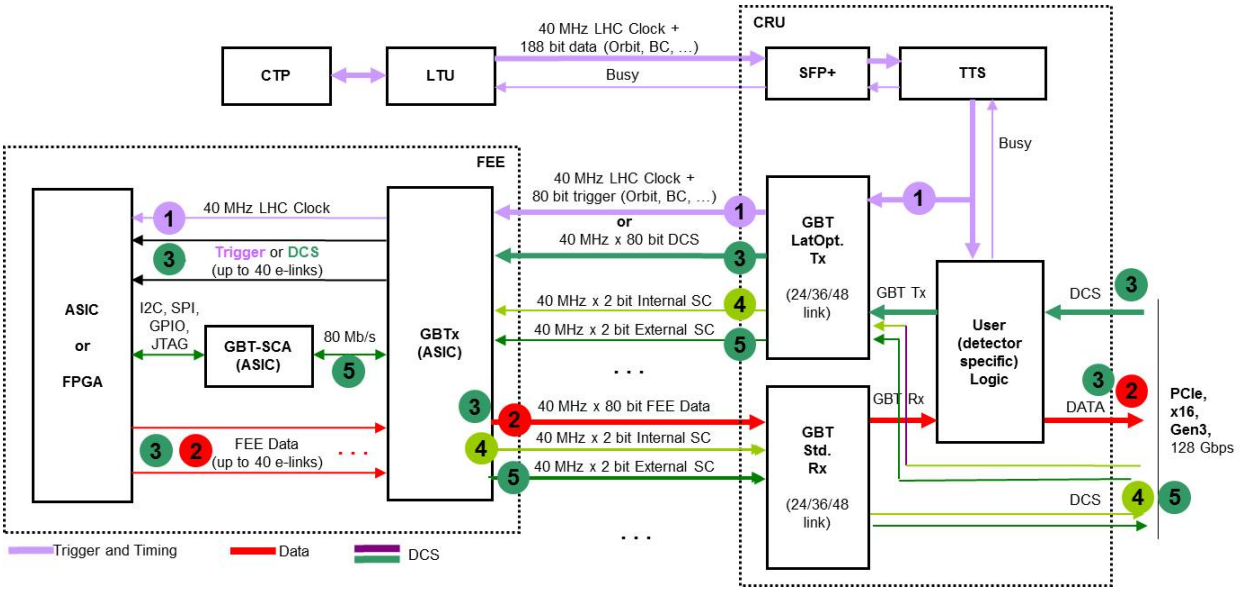


Figure 4.1: Different communication forms between the CRU, Trigger & FEE Link architecture with the GBT ecosystem.
(Courtesy: Erno David)

The markings are as as follows: 1. Delivering the TTS information or Read-out Control from CRU to FEE. 2. Receiving the detector data in serial and parallel form from the FEE. 3. Communicating custom packets between the CRU and the FEE FPGA in parallel form. 4. Sending and receiving packets to and from GBTx internal register block. 5. Sending and receiving packets to and from GBT-Slow Control ASIC. The various communication forms in this flow are organized into three groups based on the functionality viz. detector data transfer related communications, trigger related (clock, trigger, read-out) and detector slow control and command related communications [73]. The aim is to specify and test the front-end and the back-end interfacing link.The scheme utilizes GBT protocol to establish radiation tolerant communication link between on-detector front-end electronics situated in harsh radiation environment to the CRU placed in a low radiation zone. It is the front-end interface for digital data transfer from the on-detector electronics to the CRU. Data are multiplexed, processed and reformatted in the CRU depending on the detector specifications and sent to the server using the back end interface. A detailed measurement and analysis results are discussed in the subsequent sections.

4.3 Front End Interface: GBT

GBT link is the interface between Front-end/on-detector electronics and CRU. The framework allows high-speed time critical data communication with high error resilience to communicate reliably from cavern to readout electronics situated in low or non radiation zone. The GBT ecosystem, shown in the Figure 4.2 is composed of three segments; the GBT chipset which consists of radiation hard GBTx ASIC and the GBT-slow control ASIC (GBT-SCA), point to point optical link with the versatile link optoelectronic components operating in a single mode (1310 nm) or a multi-mode (850 nm) connecting the GBTx ASIC with the FPGA/COTS and FPGA programmed with the GBT logic core.

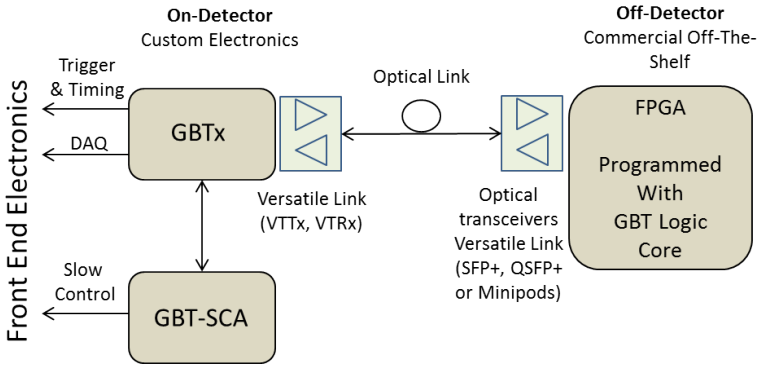


Figure 4.2: Link architecture with the GBT ecosystem.

Radiation hardness is required near detectors, however it is not necessary for the CRU located away from the radiation zone [74]; this feature is utilized for the realization of GBT functionality on the non-radiation hardened FPGAs. GBT-FPGA logic core firmware [32] is implemented on the FPGA based CRU. It mimics the GBT ASIC behaviour on the FPGA to enable the CRU for receiving the GBT datagram and also transmitting the control and timing signals from control room to the detectors. A short summary of the GBT protocol is tabulated in Table 4.1.

It is used for data, trigger and timing and control distribution merged on a single data channel. The GBT transmission is an asynchronous serial communication [75] that is composed of a GBT transmitter, a Multi-Gigabit Transceiver (MGT) [52] and a GBT receiver

Table 4.1: CRU requirements in the system

Parameters	GBT Data Transmission Protocol
Channel Data Throughput	4.8 Gbps
Raw data Throughput	3.2 Gbps
Bandwidth Utilization (Coding Efficiency)	73.33% (88 / 120)
Bandwidth Utilization (Data Efficiency)	66.67% (80 / 120)
Security	Not Applicable
Forward Error Correction	RS Encoding (15,11) with symbol size of 4 bits
Number of FEC block used	2
Error Detection	32 bits
Error Correction	16 bits
Burst Error Correction	16 bits
Latency optimized data transmission	Supported
Interleaver	Block
Supported Data Type	Idle/Control and Data

as shown in the Figure 4.3. Pattern generator and checkers are used for testing purpose only, and they are replaced with First In Fist Out (FIFO) buffers for FEE buffered data. In the transmitter, scrambler maintains the DC balance for accurate clock recovery without additional overhead by reducing the occurrence of a long sequence of continuous 1 or 0. Reed-Solomon (RS) encoder as shown in Figure 4.4 utilizes two double interleaved RS(15,

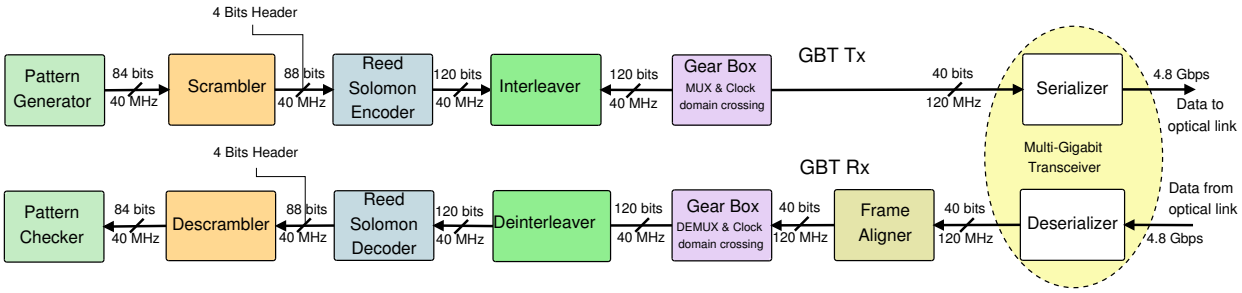


Figure 4.3: Block diagram of a GBT link in FPGA

11) encoded words each capable of correcting a double symbol error. Interleaving operation increases the error correction capability up to 4 symbols with each symbol of 4 bit. The whole process increases the code correction capabilities without any additional overhead. Gear Box, as shown in Figure 4.3 translates the frequency by modifying data bus width for Clock Domain Crossing. It consists of a dual port RAM, breaks down 120 bit frame to three word of 40 bits each. In the transmit chain, the input of Gear Box is 120bit@40MHz and output is at 40bit@120MHz keeping the data rate fixed at 4.8Gbps. Data frame is sent

from the GBT transmitter to a high speed Multi Gigabit Transceiver block. GBT Receiver

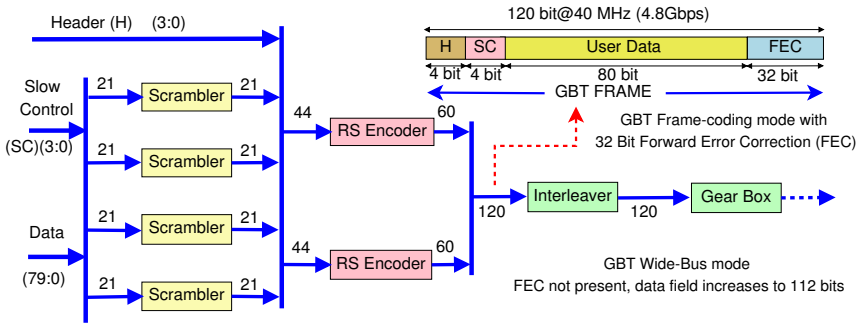


Figure 4.4: GBT link encoding scheme

performs descrambling, decoding and deinterleaving. The Frame aligner block [45] performs header detection and locking for frame synchronization using an efficient pattern search algorithm to maintain synchronization between the transmitted and the received data. A custom developed radiation hardened GBT chipset [44] is used for packaging of detector data and transmitting it in GBT standard. Data from the detector is framed into GBT standard using GBT chipset and transmitted to CRU via serial optical link. In this dissertation; the GBT link is implemented and functionally tested on Intel FPGA.

4.3.1 Design implementation on FPGA

In order to ease and speed-up the in-system implementation of the GBT-FPGA, the various components of the GBT-FPGA Core are combined in a unified module called as GBT Bank. The GBT Bank may include multiple GBT Links. Each GBT Link is composed by a GBT Tx, a GBT Rx (both together will be referred to as GBT Logic) and a Multi-Gigabit Transceiver (MGT). The clocking resources are external to the GBT Bank; so the designer can connect the different clocks as per the requirement. The idea behind the development of the GBT Core is to keep each GBT Bank as a self-sufficient entity regarding the logic and clocking resources. The GBT-FPGA Core facilitates the implementation of single-link and multi-link GBT-based systems. The systems, where the requirement for the number of GBT Links is more than that provided by a single GBT Bank, it is possible to add more

GBT Links by instantiating the GBT Banks in parallel. For example, if a sub-detector needs to have 20 links per CRU board then the DAQ designer could organise as three GBT banks of six links each and one GBT bank with two links. The GBT bank is the largest common cluster formed of the GBT links grouped together. The maximum number of GBT Links per Bank is limited by the architecture of the targeted FPGA.The maximum number of GBT Links per GBT Bank is limited by the architecture of the targeted FPGA.

The resource utilization of FPGA is an important aspect in the design of DAQ systems as a substantial amount of the FPGA resources will also be utilized for the sub-detector specific user logic. In this situation, the optimization of the GBT design to save the logic resources is an important task. Implementation of the GBT Links in the form of GBT Banks gives a new design approach to save the GBT specific periphery logic resources. The sharing of a single decoder block for several links in a GBT bank saves the FPGA resources. The implementation of CRU in the Arria-10 FPGA delivers one more level of optimization solution. Arria-10 FPGA has x6 Physical Medium Attachment (PMA) bonded mode; using this saves the clocking resources and also reduces the intra-link clock skew. The design uses PMA bonded mode configuration, to get equal latency through each bonded data-path. The PMA bonded mode in Arria-10 allows six GBT links to be packed closely in one Bank as illustrated in the Figure.

The Trigger Timing and Control systems in HEP experiments require a low, fixed and deterministic latency for the transmission of the data and clock to ensure correct event building. On the other hand, systems that are not time critical, such as DAQ, do not need to comply with this requirement. The GBT-FPGA core allows the implementation of one or several GBT links in two operational modes: *Standard* version (Std) for the non-time critical applications and the *Latency-Optimized* version (Latopt) for the latency critical systems either on Transmit, Receive or both [76]. In Std mode of GBT operation, an elastic buffer is used in between PCS and PMA blocks to compensate for the phase of the clocks that drive the two blocks as shown in Figure 4.5. Elastic buffer adds uncertainty in the

latency. To alleviate this effect, the elastic buffer is bypassed and an external phase aligner block is used to align the phase of the clocks between PCS and PMA. This helps to achieve a consistent delay for Latopt mode of GBT operation which is needed for the time critical data transfer. These links can be also configured to provide the different encoding mode offered

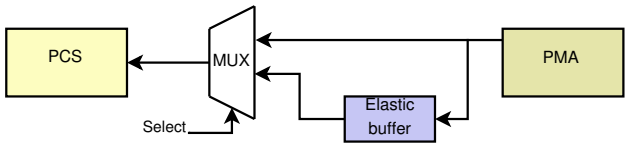
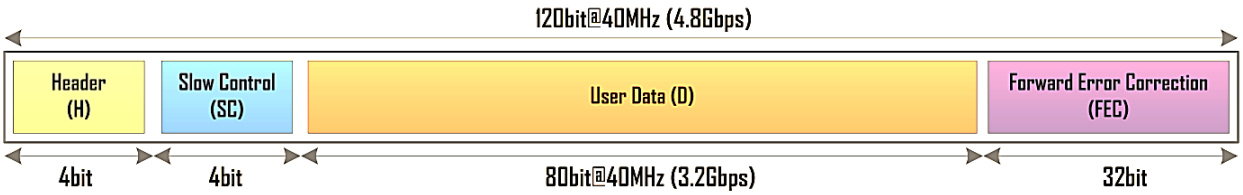
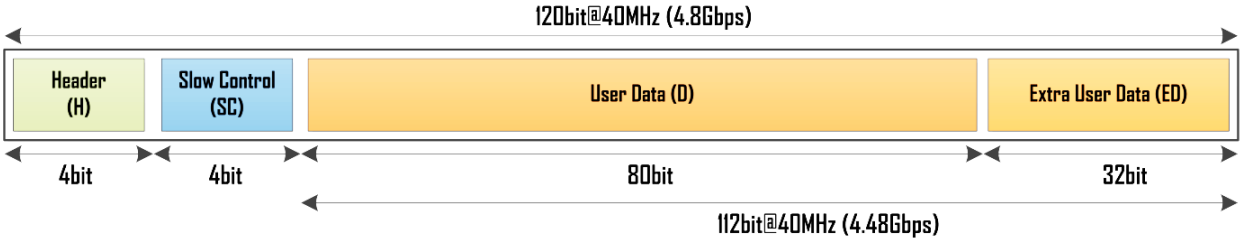


Figure 4.5: Clock phase compensation between PCS and PMA

by the GBT FPGA Core: the *GBT-Frame* mode (Reed-Solomon based) and the *Wide-Bus* mode (no encoding). The two frame formats are shown in Figure 4.6. GBT FPGA



(a) GBT-Frame encoding frame



(b) Wide-Bus encoding frame

Figure 4.6: Data format for GBT protocol.

Core design firmware is implemented on FPGA. Two custom DAQ boards PCIe40 [35, 77] and AMC40 [58, 78] based on latest Altera Arria 10 [70] and Stratix V [52, 79] FPGA respectively are used for the detailed comparative study, the design evaluation and tests for the prototype study.

4.3.2 Calibration logic

Why Required?

During the loopback tests of Latopt version, it was observed that the data at the input and the output of the link are mismatched. Clocking architecture for the embedded transceivers is studied in detail. The studies revealed that the *Tx word clock* (viz. clock at which the data is sent out of the link as shown in the Figure 4.3) is sampling the frame data (120 bit data received at the link) at shifted instances. Hence the formed data words for the transmission are shuffled and the incorrect data are received at the receiver. The clock distribution scheme for the implementation of transmitter section of the GBT link on Startix-V FPGA; where the issue is identified is shown in Figure 4.7.

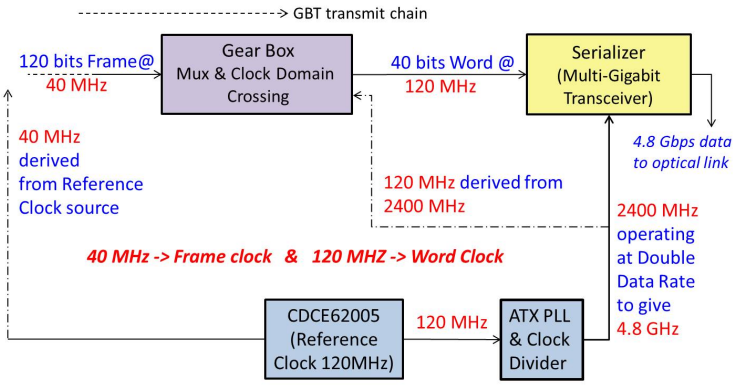


Figure 4.7: Clock Distribution Scheme for the Stratix-V FPGA

Explanation of the matter of contention

One of the potential points of uncertainty in logic design is the clock domain crossing (CDC) in the Gearbox of GBT link. The Gear Box as shown in Figure 4.3 is Mux/Demux where domain of the clocks are changed. In the transmit chain, it breaks down the 120 bit frame to three 40 bits words. In the transmit chain, it breaks down the 120 bit input frame at 40 MHz to three words of 40 bits each at 120 MHz. At the two ends of the Gear Box the data input is *120bit at 40MHz* and output is at *40bit at 120MHz*; keeping the data rate same i.e 4.8 Gbps at the two end points. 40 MHz is the *Tx frame clock* and 120 MHz is the *Tx word*

clock. The parallel fabric clock of 120 MHz (*Tx word clock*) that goes to one end of the Gear box is forwarded by the embedded transceiver of the FPGA. The *Tx word clock* drives the user logic data words each of 40 bits into the transceiver. Transceiver operates at the double data rate in *parallel in serial out mode*; the input frequency at transceiver Serdes is 2400 MHz and the operation at double data rate gives the output at 4.8 Gbps rate. The clock of 2400 MHz is obtained from the clock divider. The clock divider gets the input from external PLL whose input is from separate *reference source* of 120 MHz.

Tx Word clock of 120 MHz is derived from the serial clock of 2400 MHz; dividing it with a clock divider of serialization factor of 20 [80, 81]. The clock divider which gives an output of 2400 MHz; gets the input from external PLL whose input is from separate *reference source* of 120 MHz. The clock of 40 MHz (*Tx Frame clock*) is derived from the 120 MHz reference frequency. *Tx Frame clock* goes to the other end of the gear box and there is a phase difference between the two derived clocks viz. *Tx Word clock* and *Tx Frame clock*.

When the frequency of a source clock is multiplied first and then divided again to the original frequency; this leads to a non-deterministic phase difference between the source and divided clocks [75]. This occurs due to the concept that the rising edge of the divided clock may be locked onto any of the rising edges of the multiplied clock as shown in the Figure 4.8. At the two ends of the Gear Box; the two derived clocks viz. *Tx word clock*

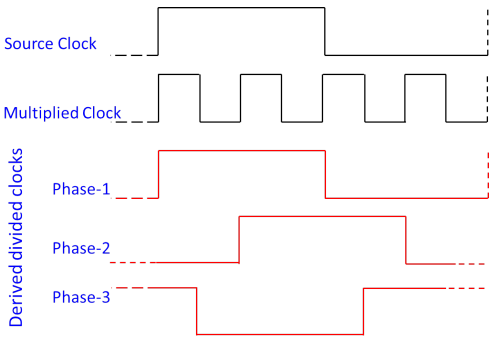
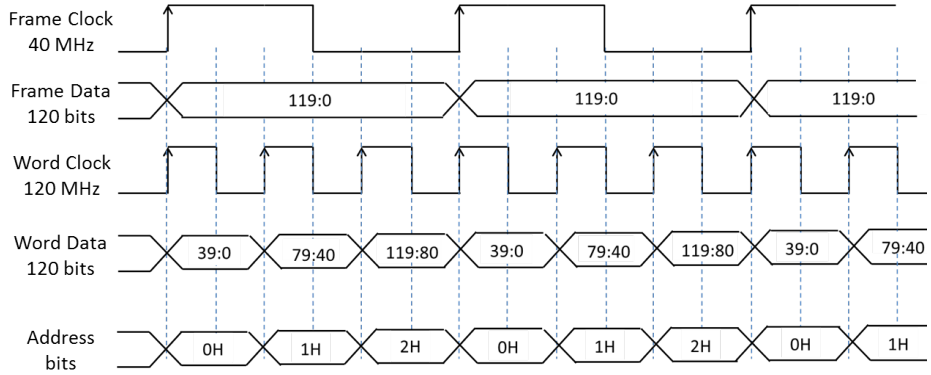


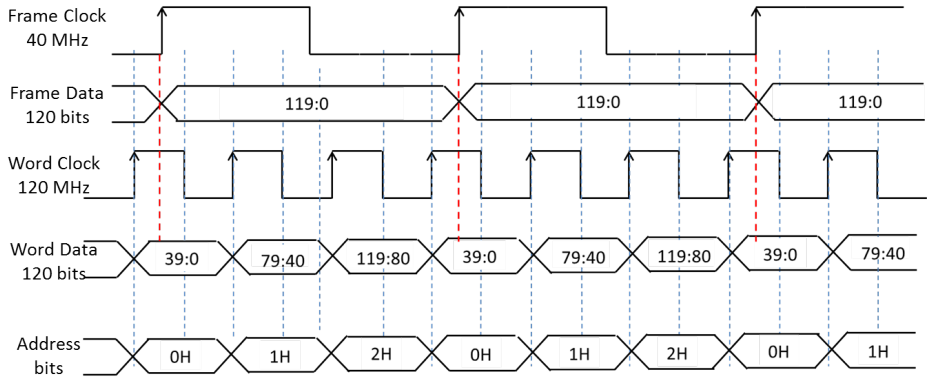
Figure 4.8: Multiplication or Division of Clock frequency

(clock at which the data is sent out of the transmit link) and the *Tx frame clock* (data received at the link) are not in phase. Thus there is a misalignment of the data words and

the data stream is rearranged incorrectly. The problem arising due to the phase mismatch between the two clocks is explained in the Figure 4.9. The ideal situation where the *Tx word clock* is locked at the correct phase and samples the frame data at the correct instances is shown in Figure 4.9a.



(a) Word clock locked at correct phase and sampling the frame data at correct point.



(b) Word clock not locked at proper phase and sampling the frame data at incorrect point.

Figure 4.9: Word clock and Frame clock phase mismatch

However in the actual scenario, due to the occurence of the phase shift between the two clocks, the data frame of 120 bits is sampled at the wrong instances as shown in Figure 4.9b. Hence the three words each of 40 bits are formed with shuffled data bits. In such case when there is a mismatch between the phase of the two clocks, it leads to the uncertainity and errors in the data also. The contention is due to the register based clock domain crossing like the Latopt mode of GBT; where data write and read occurs at the same time.

Futhermore, the sampling of the clocks should not lie within or close to the metastabilty zone. Sampling near this zone pushes the system back to the unstable regime even with a minor jitter of the clocks. The stability of the reference clocks in terms of jitter is important in this regard. Also in the systems with the multi-clock design and with multiple parameters like temperature variations, power on-off cycle, reset cycle, firmware upgrade, loss of lock in the transceiver [82], ageing of clock circuitry in phase-locked loop (PLL), fibres replacement etc.; may lead to the uncertainity in the clock phase. The phase shifts and metastability could not be avoided, however the detrimental effects could be neutralized. Thus, when dealing with systems featuring register based CDC; there is a need to develop a logic in the way that could calibrate and ensure the correct phase relationship between the clocks generated in different domains.

Implementation of the Calibration Logic

A unique auto-initialization calibration logic has been developed for the GBT firmware. This would allow to extract the relative phase information and to recalibrate the system when needed, to maintain the constant phase relationship. As shown in Figure 4.7 the clock divider used to obtain parallel *Tx word clock* gives a phase difference uncertainty of (Φ_p) with external clock from the *reference source*. The variable phase (Φ_p) is continual and could attain the integer values as shown in equation 4.1 and 4.2. For GBT at 4.8 Gbps; 1 UI is 208.33 picosec.

$$[\Phi_p] = 2mUI : m \;\; is \;\; integer, 0 \leq m \leq 19 \tag{4.1}$$

$$Maximum[\Phi_p] = 38\ UI = 7.916ns \approx 8ns \tag{4.2}$$

Logic for phase computation

It monitors the phase alignment between the *frame clock* and the *Tx word clock*; to ensure the sampling at the correct edge. The phase difference between the two clocks is measured

by XOR based phase detection technique[ref]. It is executed by sampling both the clocks (*frame clock* and *Tx word clock*) with a higher frequency of 600 MHz in this case. The logic for phase calculation is shown in Figure 4.10. The working principle is based on

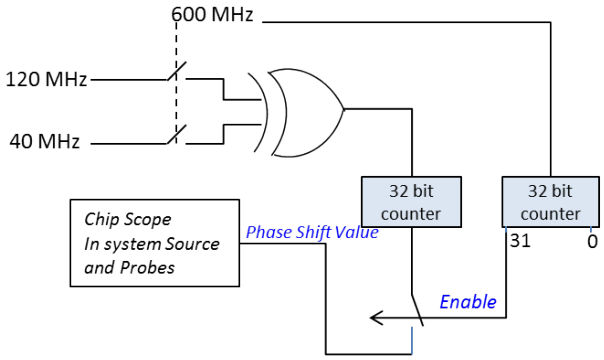


Figure 4.10: Logic for Phase Calculation.

the accumulation of the systematic samples over the phase detector signal. Counts of the XOR samples are added to estimate the creeped in phase difference. The phase shift value is stored in In-System-Source and Probes tool (ISSP) of Intel and displayed in the hexadecimal format. With each reset cycle of PLL, the phase transition is sequential and has a finite number of states. Hence it could be calculated.

Phase calibration logic

The characteristics of phase computation is used to design the proposed solution to calibrate and match the phase of the two clocks. The designed autocalibration logic has been shown in Figure 4.11.

The logic is implemented on FPGA without any additional circuitry and preserves a synchronous relationship between the clocks. In this calibration logic the test data pattern of 120 bits is at the input and the pattern checker at the receiver side. Pattern generator and checkers are used for calibration purpose only and they are replaced with FIFO buffers for the user data from the front-end electronics of the detectors. A static pattern is sent in the link and the receiver compares and checks for any data flips. If the data is not matched then there is a sequence of repetitive resets of the reference clock. During the calibration

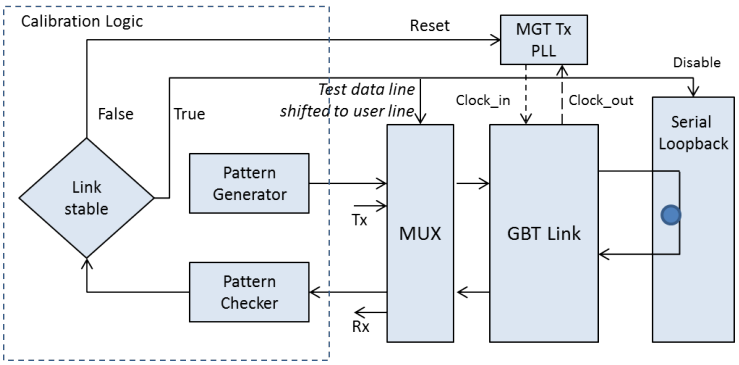


Figure 4.11: Phase Calibration logic.

process; data loopback from the transceivers is enabled in the serial loopback mode. Due to this the erroneous data is not sent from silicon to the optical fibre for the transmission and the data is looped back with in the silicon. The multiplexer between the GBT transmission link and the data input transfers the test data line to the user data once the calibration process is completed. By resetting the clock divider the phase shifts by an small amount every time and the phase values are repeated after 20 times reset cycles since 120 MHz clock can lock on to any one of the rising edge of 2400 MHz clock (2400 MHz/120 MHz). The data is stable when the phase is matched. The uncertainity of the phase in the clocks is mitigated with the phase calibration logic.

Metastability management

The metastability could creep in when signals are transferred between unrelated or asynchronous clock domains. Also it is caused when the drift of clocks occur due to the temperature or voltage variations. The register setup and hold time should be maintained to ensure the data read and write reliability. Hence a sequence of register chain is introduced in the destination clock domain as shown in Figure 4.12. Due to The phase transitions of forwarded *Tx word clock* may also lead the system to fall into this zone. Also *Tx word clock* needs initialization time after each reset cycle to align the data frame (120 bits) and the data word (40 bits each) in the Gearbox. Thus, the synchronizer reset to the Gearbox is delayed by few clock cycles from the global reset. This synchronizes the data to the new

clock domain [83] with sufficient settling time and any metastable conditions are resolved before the data is used for further processing.

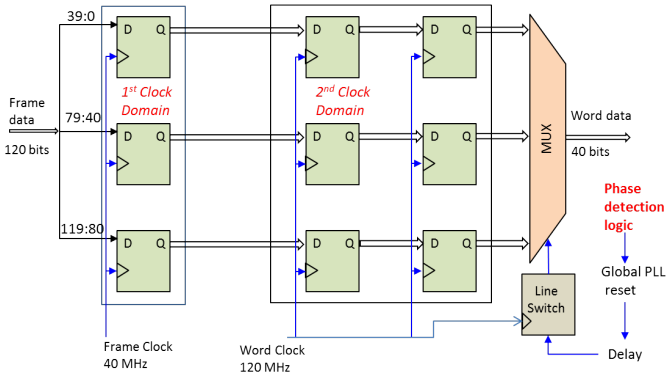
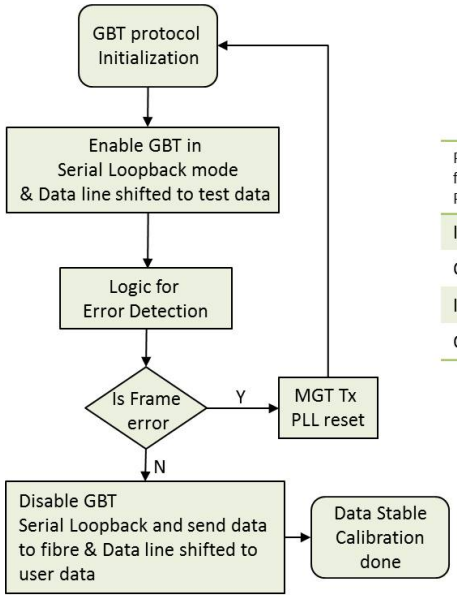


Figure 4.12: Synchronization Register Chain for Gearbox

State machine to implement calibration logic

The calibration trigger could be selected depending on metastability detection logic or phase error detection. Hence a state machine is developed as shown in Figure 4.13 which resets the divider network until the stable and synchronized data is recorded. After the phase is



Logic for Error Detection

Phase Shift Value from Logic for Phase calculation	Is Metastability Seen ?	Assert MGT Tx PLL Reset
In-Phase	Yes	Yes
Out of Phase	Yes	Yes
In-Phase	No	No
Out of Phase	No	Yes

Figure 4.13: State Machine for Phase Calibration

matched it is necessary to check the data frame error rate below a set threshold for various

data cycles. This eliminates the possibility of metastability also.

Temperature measurement and stability analysis

The temperature measurement and monitoring is inevitable for such systems. With temperature fluctuation the delays change due to the variation in the on/off time of transistors. The on-chip temperature sensing operation is carried with the help of an internal temperature sensor diode embedded on Stratix-V FPGA. It is read with a built-in 8-bit analog-to-digital converter (ADC) circuitry. The Figure 4.14 shows the top-level ports and the basic building blocks of the temperature Sensor intellectual proerty (IP) core for Stratix-V FPGA. The

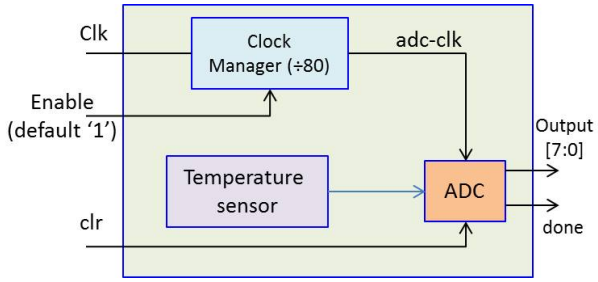
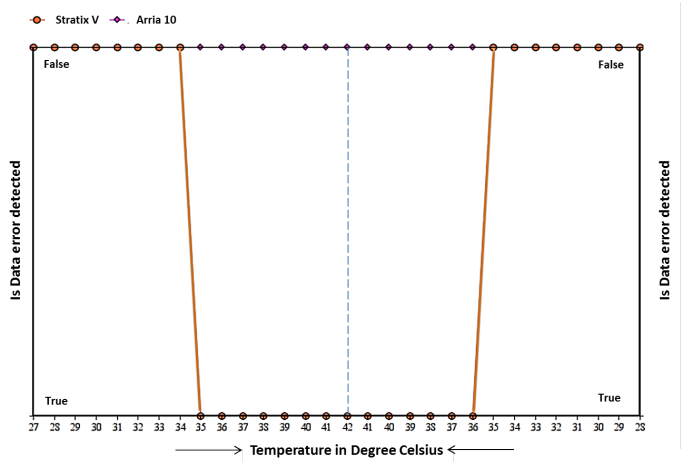


Figure 4.14: Temperature sensor IP core for Stratix-V

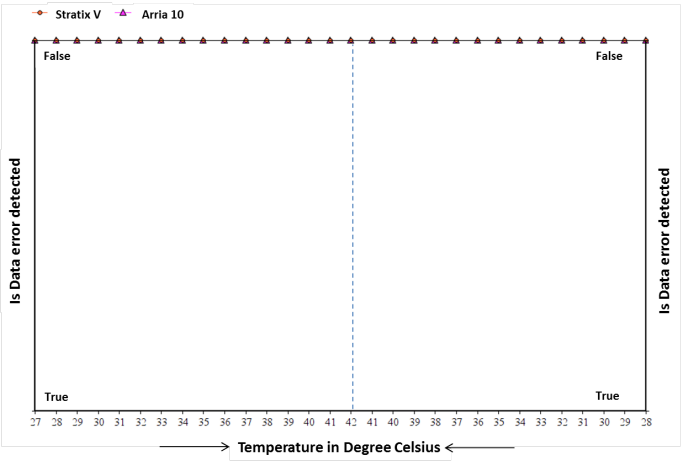
clock manager feeds the clk signal to ADC in a range of 1MHz or less. To read a valid ADC signal it is read after every 10 clock cycles of the ADC clock. A firmware module is developed to read the temperature values. The signals from the temperature sensor module are integrated with the calibration design at the register transfer level (RTL).

A study for the error in data with variation of temperature is carried out on the Stratix-V development board. The temperature is varied by controlling the on-board cooling system and noted for any data frame errors. The results for different operating modes of GBT are shown in Figure 4.15. This study is also conducted on the Arria-10 FPGA based development card. The results for the two FPGAs are plotted on the same canvas for the comparison. The results taken together depicts that the methodology of the clock distribution on stratix-V FPGA transceivers makes data unstable with the temperature fluctuation in the transmit Latopt mode. However, in the other modes the data are stable.

The Latopt mode on the transmitter side is necessary for the fixed and deterministic latency of the GBT operation and is the most utilized mode. At the receiver the data are time stamped and hence the constraints of timing are eased-off and allows the use of Std mode on the receiver side. It highlights the importance for the need of the calibration logic on the Stratix-V FPGA to operate in transmit Latopt mode. The entire logic is implemented



(a) Transmit Latency Optimized mode



(b) All other modes

Figure 4.15: Showing the data stability vs temperature variation of two FPGAs when GBT operating in (a)Transmit Latency Optimized mode (b) All other modes.

on the Gearbox in GBT Physical Coding Sublayer. To avoid any temperature fluctuations; for the stable operation of CRU, servers with special air cooling provisions are being used.

4.3.3 GBT Qsys Model

The Qsys is an Intel-Altera system integration tool also known as Platform Designer tool. A Qsys model of the GBT is designed as shown in Figure 4.16. It is designed in such a

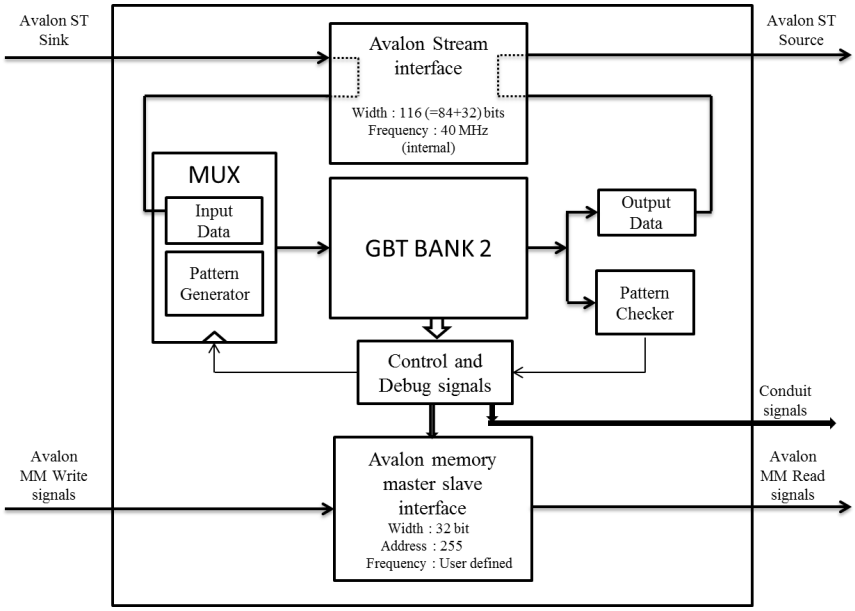


Figure 4.16: GBT Qsys Model.

way that all the debug signals are accessible to the users in a graphical user interface (GUI) which can not be directly probed at the VHDL/RTL level. A screenshot of the GUI is shown in Figure 4.17. Various signals like temperature read, mode of GBT operation, PLLs lock condition, Latency of the link are extracted out using the Avalon bus having provision for seven different kind of interfaces [84].

4.3.4 FPGA Resource utilization

The GBT-FPGA logic core firmware reference design is implemented on FPGA. It has two GBT banks. One Bank has a single link with Tx in Std mode and the Rx in the Latopt mode. The other Bank has four links, each with Tx in latopt and Rx in Std mode. Estimation for the utilization of FPGA logic resources is an important parameter for the choice of FPGA on the CRU where a large number of links need to be handled. Resource estimation is

```
Temp read = 31 °C

## Design type ## ************ Tx : Standard        CODING : GBT_FRAME   ************ Rx : Latency Optimized      CODING : GBT_FRAME

External PLLs are locked

SERIAL LOOPBACK IS ENABLED
 Do you want to give general reset ?? (0-No, 1-Yes)
0
GBT TX DATA (LINK 1) = 0000000000001D96A57DF            GBT RX DATA (LINK 1) = 0000000000001D96A57C0  Is Locked = 0
GBT TX DATA (LINK 2) = 0000000000001D96A5803            GBT RX DATA (LINK 2) = 0000000000001D96A57E4  Is Locked = 0
GBT TX DATA (LINK 3) = 0000000000001D96A5822            GBT RX DATA (LINK 3) = 0000000000001D96A5808  Is Locked = 0

GBT EXTRA WIDEBUS TX DATA (LINK 1) = D96A57E5           GBT EXTRA WIDEBUS RX DATA (LINK 1) = 00000000  Is Locked = 0
GBT EXTRA WIDEBUS TX DATA (LINK 2) = D96A5809           GBT EXTRA WIDEBUS RX DATA (LINK 2) = 00000000  Is Locked = 0
GBT EXTRA WIDEBUS TX DATA (LINK 3) = D96A5828           GBT EXTRA WIDEBUS RX DATA (LINK 3) = 00000000  Is Locked = 0
Is Error Seen in Link 1 = 0
Is Error Seen in Link 2 = 0
Is Error Seen in Link 3 = 0

Latency of Link 1 = 350 ns

Latency of Link 2 = 350 ns

Latency of Link 3 = 350 ns
```

Figure 4.17: Screen shot of the GUI for the Qsys model during run.

necessary to approximate the number of links that could be packed on an FPGA and to get an idea for the utilization of the hardware resources. The resource utilization is shown in Table 4.2

Table 4.2: FPGA resource utilization for the GBT-FPGA reference design.

Parameters	**Family: StratixV** (5SGXEA7N2F45C3)	**Family: Arria10** (10AX115S4F45I3SGE2)
PLLs	11/92 (12%)	10/176 (6%)
Registers	20060	18338
Logic utilization (in ALMs)	10,542/234,720 (4%)	9,055/427,200 (2%)
Block memory bits	202,752/52,428,800 (<1%)	126,464/55,562,240 (<2%)
RAM Blocks	56/2,560 (2%)	40/2,713 (1%)
HSSI PMA TX/RX SerDes	4/48 (8%)	4/72 (6%)

4.3.5 Power Consumption

It is important that the design modules consume least amount of power so that power consumption involving computation processes remains within the margin of power rating and prevents overheating. The power consumption for the Arria-10 FPGA using the Intel power estimator tool is summarized in Table 4.3 when a single GBT link under the different

encoding scheme consumes different power.

Table 4.3: Power consumption with the GBT Encoding Scheme.

Encoding Scheme	Power consumtion in FPGA (mw)	Power consumtion in Tranceiver bank (mw)
GBT Frame-coding mode	4492.8	1947.77
GBT Wide-Bus mode	4082.31	1462.93

4.3.6 Latency Measurement

Latency measurement for the GBT protocol is a crucial parameter. Data transmission from the detector to the CRU have to be time synchronized and a fixed latency is required for the application in the trigger and timing system. A detailed study for the estimation of the latency in terms of the clock cycles utilized in the GBT protocol for data processing and transmission is done for all the possible combinations of the mode of operation for transmit and the receive side [32]. The latency is measured by concurrently subtracting the transmitted and received value of the counter from the pattern generator in the loopback condition as shown in the Figure 4.18.

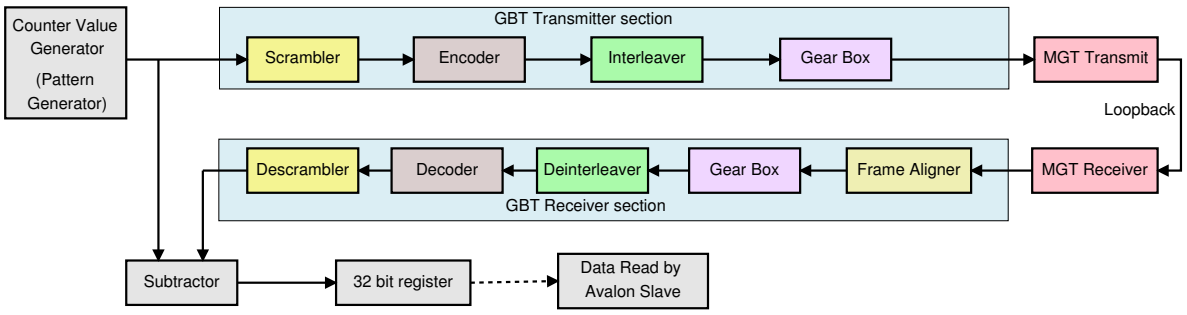


Figure 4.18: Test Setup for the GBT latency measurement.

Latency occurs in both transmitting and receiving directions, depending on the media and path involved. The total path $\mathcal{L}1$ in the loopback mode as shown in Figure 4.18 and is given by equation (4.3). It consists of GBT transmitter (GBT Tx), Multigigabit transmitter

(MGT Tx), Multigigabit receiver (MGT Rx) and GBT receiver (GBT Rx).

$$\mathcal{L}_1 = (GBT\ Tx - MGT\ Tx - MGT\ Rx - GBT\ Rx) \tag{4.3}$$

The number of clock cycles utilized in the GBT transmit and receive section and the MGT transceiver section are estimated separately. MGT transceiver is removed from the GBT protocol, and the GBT transmitter is coupled to the receiver section at the firmware stage. The two paths $\mathcal{L}2$ and $\mathcal{L}3$ is given by equation (4.4) and equation (4.5).

$$\mathcal{L}_2 = GBT\ Tx - GBT\ Rx \tag{4.4}$$

$$\mathcal{L}_3 = MGT\ Tx - MGT\ Rx \tag{4.5}$$

The clock cycles utilized are observed using the simulation models. The Latency in the MGT section is calculated by the difference of $\mathcal{L}1$ and $\mathcal{L}2$ path delays. Measurement within FPGA is always dependent on the data rate, hence the delay is measured in terms of clock cycles (1 clock cycle = 25 ns). The resolution is 25ns, which is the frequency of LHC bunch crossing. Transmission latency is measured for all the possible combinations of mode of operation of GBT protocol in AMC40 and tabulated as shown in Table 4.4. The comparison of the delay for the total path $\mathcal{L}1$ for AMC40 and PCIe40 is shown in Table 4.5 in terms of the clock cycles utilized. The information is useful for the designers to optimize the data acquisition firmware.

Table 4.4: GBT link latency measurement.

Latency for path	**GBT Link mode of operation(ns)**			
	Tx Latopt Rx Latopt	Tx Latopt Rx Std	Tx Std Rx Latopt	Tx Std Rx Std
$\mathcal{L}_1 = GBT\ Tx - MGT\ Tx - MGT\ Rx - GBT\ Rx$	150	350	350	600
$\mathcal{L}_2 = GBT\ Tx - GBT\ Rx$	75	275	250	425
$\mathcal{L}_3 = MGT\ Tx - MGT\ Rx$	75	75	100	175

Table 4.5: The comparison of the total path delay for AMC40 and PCIe40.

Latency for path ($\mathcal{L}_1 = GBT\ Tx - MGT\ Tx - MGT\ Rx - GBT\ Rx$)	**GBT Link mode of operation** 1 Clock cycle (LHC bunch crossing Rate) = 25 ns			
	Tx Latopt Rx Latopt	Tx Latopt Rx Std	Tx Std Rx Latopt	Tx Std Rx Std
AMC40	6	14	14	24
PCIe40	6	8	14	18

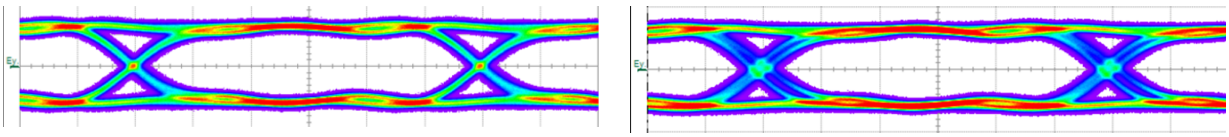
4.3.7 Eye diagram measurement

An Eye diagram [85, 86] analysis is used to indicate the quality of signals in high-speed digital transmissions. Data from the transceivers embedded in silicon (FPGA) are transmitted to the onboard MiniPOD [87] optical modules. They are coupled with flat ribbon cable using a Prizm connector and terminated on the other side with industry standard MTP connector. Eye diagram serves as an indicator of the link performance and is used as a target parameter for the link optimization. The channel performance of the transceiver link was studied by interpreting the eye diagram pattern in a Lecroy Serial Data Analyser (SDA) oscilloscope. BER analysis is done using the pattern generator and checker from the TTK.

The signal quality of the GBT protocol operating at 4.8 Gbps is measured for both Stratix-V FPGA (AMC40 board) and Arria-10 FPGA (PCIe40 board) using the Eye diagram analysis and the jitter measurements The signal to noise ratio of this high-speed data signal is directly indicated by the amount of eye closure or Eye Height. The Eye diagram and the details of jitter measurements are shown in Table 4.6 and Table 4.7 The signal to

Table 4.6: Showing the Eye Diagram for GBT encoded data.

(a) AMC40 (b) PCIe40



	AMC40	PCIe40
Eye Height	378 mV	373 mV
Eye Width	190.4 ps	176.8 ps

noise ratio of this high-speed data signal is directly indicated by the amount of Eye closure

Table 4.7: GBT Jitter Measurement for the two FPGAs.

Jitter Paramete(Unit Picosecond)	PCIe40	AMC40
Deterministic Jitter	13.125	5.503
Data Dependant Jitter (DDj)	21.647	11.228
Periodic Jitter (Pj)	7.24	6.75
Inter Symbol Interference (ISI)	21.657	11.095
Standard Deviation(s)	2.898	2.989
Duty Cycle Jitter (DCD)	1.912	2.000
Total Jitter (Tj)	58.185	51.148

or Eye Height. Eye width/height is 176.8 ps/373 mV for Arria-10 and 190.4 ps/378 mV for Stratix-V at the BER of the order of 10^{-12}. The measured total jitter is of the order of few picoseconds only. It is important that the clock jitter is within an acceptable hold time of the pipeline registers for error free data transmission. The data obtained is beneficial for further studies.

4.3.8 BER analysis

The test setup for the BER measurements of the high speed optical link is shown in Figure 4.19. It consists of an integrated FPGA system with embedded transceivers along with

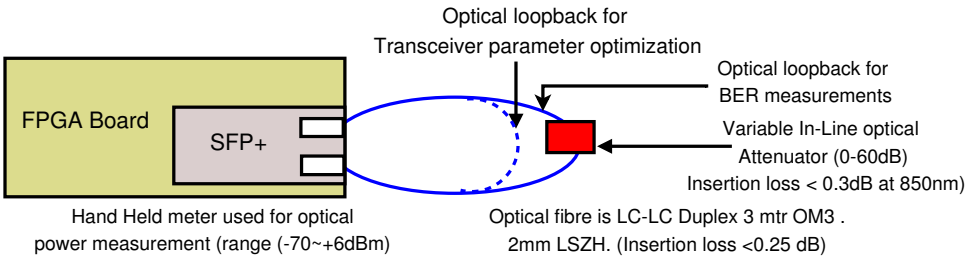


Figure 4.19: Test setup for BER measurement.

the Serial Form-factor Pluggable (SFP+) optical transceiver module and the Multi-Mode Fibre with connectors. A manually controlled In-line Variable optical Attenuator (VOA) is introduced in the fibre loopback to induce optical power degradation. The optical power output after the attenuation is measured using a handheld optical power meter with an insertion loss of $< 0.3dB$ at the 850nm range of operation. The output from the attenuator is

looped back as shown in Figure 4.19. A Pseudo-Random Binary Sequence (PRBS) is transmitted across the transceiver link to evaluate the BER function with the pattern checker. BER at different attenuation levels were measured. This test characterizes the sensitivity of the receiver and the minimum optical power required for achieving a specified BER in a system. We restrict the scope of present performance measurements to the physical layer parameters, keeping aside the issues of the Physical coding sublayer (PCS) sublayer.

BER measurements for GBT protocol with respect to the two encoding schemes viz. GBT Frame coding and the Wide-Bus as shown in Figure 4.4 is plotted in Figure 4.20. An exponential fit to the data is implemented. BER measurement cannot be pursued below -17 dBm of receiver sensitivity, due to the loss of recovered clock. However, the plot can be extrapolated based on standard complementary error function 'erfc' nature of the curve, assuming the Gaussian nature for noise.

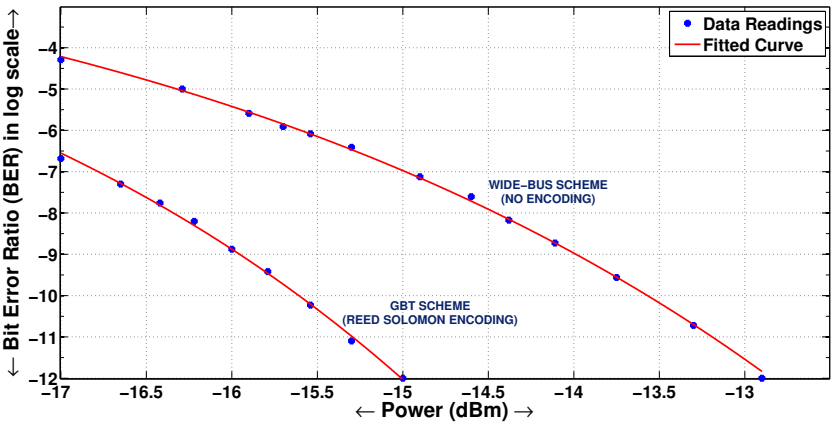


Figure 4.20: BER measurement for GBT Frame coding and GBT wide-Bus mode.

Margin of Receiver Sensitivity for targetted BER of 10^{-12} *between both the schemes is*

$$= (15 - 12.9)\ dBm = 2.1\ dBm \tag{4.6}$$

2.1 dBm as given in equation (4.6) is in close agreement to the measurement conducted for GBT protocol implemented on Xilinx FPGA [88] which is around 2.5 dBm.

4.4 Back End Interface

The two latest promising technology options available for the back end interface of CRU are PCIe express protocol [55] and 10GbE protocol [89]; which is also evident from the reference [5]. For CRU the most favourable options were AMC40 board which is 10GbE based and the PCIe40 board which is PCIe express based hardware.

4.4.1 Communication over the PCI Express interface

The different software layers and the underlying hardware interfaces needed to communicate through the CRUs are shown in Figure 4.21.

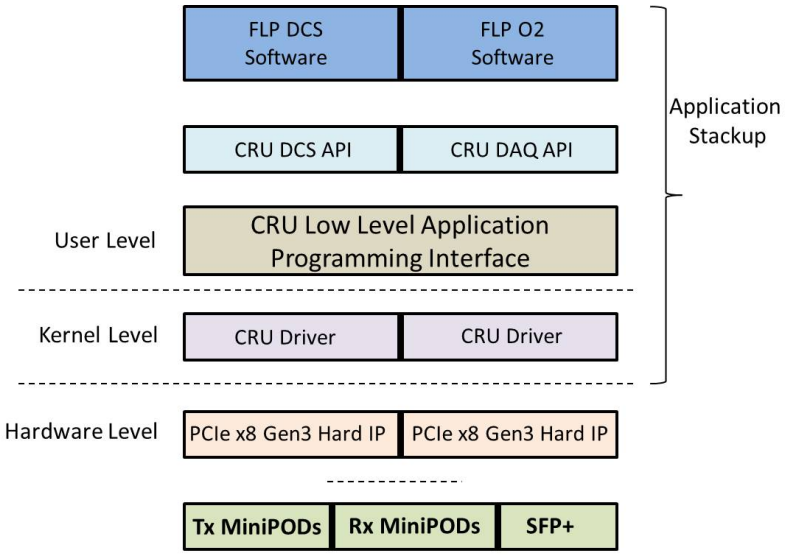


Figure 4.21: Overview of Software Stackup for CRU.

The portable driver architecture (PDA) driver is used by CRU. It belongs mostly to the user space and only has a small and easily maintainable adapter in the kernel. It provides the low-level enumeration, access to the base address registers (BAR) of PCIe, DMA memory handling, interrupt handling and synchronization support for multi-threading.

The CRU application programming interface (API) layer is a purely user space library. This layer provides controlled access to the features of the driver, low-level DMA data transfer management functions, C functions for accessing the register based interfaces provided

by the CRU firmware (for example functions for slow control and JTAG port), initialization, configuration and control of the device.

The CRU DAQ API and detector control system API rests on top of the CRU API layer. DAQ API provides a high-level C++ interface to access the data transfer functions of the CRU. The CRU DCS API is required for the safety of the experiment. The DCS controls and checks the status of the FEE even when the data taking is not active. For this reason, a dedicated communication channel between CRU and DCS is established. The DCS system configures and monitors the FEE using the CRU card connected to several GBT links. A software interface provides the communication channel for the DCS to control the CRU and execute the functions. The CRU DCS API gives the access to root level operations like WRITE/READ a register to/from the CRU. The CRU low-level API interface will be concealed to the DCS framework as shown in Figure 4.22. The minor details of the

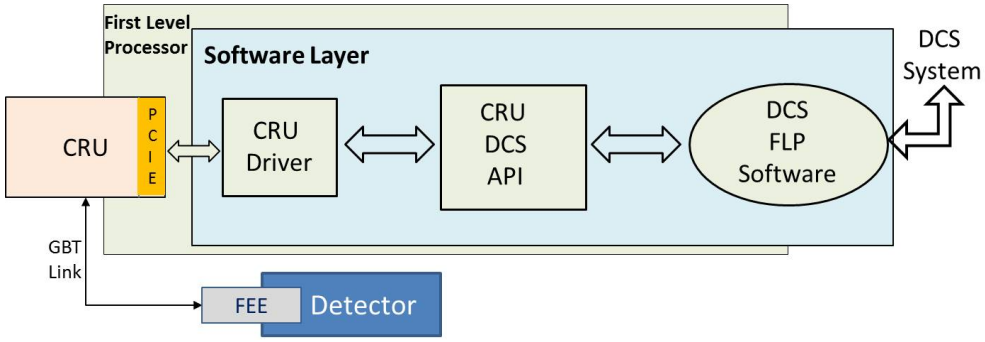


Figure 4.22: Overview of Software Stackup for CRU.

interface to commence the communication with the CRU and initialize the data transfer will be managed by the CRU DCS API. The interface will also provide information to communicate the status of the current command execution. In case of error during the configuration of the FEE the software will propagate it to the DCS system.

PCIe Performance

The CRU is engrossed with PCIe Gen3 x16 lanes; to provide output bandwidth around 128 Gbps. The DMA is used to move the data over PCIe; from and to CRU independent

of the processor. The Altera Avalon Memory-Mapped (Avalon-MM) DMA PCI Express for Arria 10 is used for bench marking PCIe Gen3 x16 performance in PCIe40. It is interfaced with application layer using QSYS interconnect with 256 bit interface. The Arria-10 ES2 and production chips run in GEN3. PCIe Gen3 links use 128b/130b encoding scheme to maintain DC balance; two extra synchronization bits are added. The encoding overhead is 1.56 percent. Hence, the effective theoretical link bandwidth for single lane is reduced to ~ 7.876 Gbps instead of 8 Gbps [90]. The Transaction layer Packet (TLP) has different kinds of headers. The overhead associated with a single memory write ranges from 5-7 double words (one double word = 32 Bits) i.e. from 160 bit to 224 bits maximum and the design supports maximum 256 byte as data payload. The TLP bridge interface between DMA engine and PCIe-Hard IP is 256 bits wide hence the 5/7-dword header requires a single bus cycle. The 256-byte payload requires 8 bus cycles. The maximum throughput for a write is estimated as:

Throughput = Clock Cycles for payload size ÷ Clock Cycles for (payload size + overhead) × effective link data rate = (8/(8+1)) × 7.877 Gbps = 6.93 Gbps Thus, maximum throughput for ×8 lanes equals to 6.93 Gbps × 8 = 6.93 GigaByte/sec. This test is repeated for both the PCIe banks and shown in chapter 6.

Although designers had migrated to PCIe based solutions, still 10GbE has got future proof solutions with legacy support and ease of upgradation. 10GbE can be optimized for detector specific data and the Quality-of-Service is provided in the higher layers of Open Systems Interconnection (OSI) model [61]. The option of l0GbE for the DAQ link, has also been studied in the initial stage of the CRU development and summarized in the subsequent section.

4.4.2 10-Gigabit Ethernet

10GbE protocol stack consists of Physical layer and the Data link layer according to OSI model. Data link layer is formed of Media Access Control (MAC) and the Logical Link

Control (LLC) as shown in Figure 4.23. Detector specific data payload is submitted to the

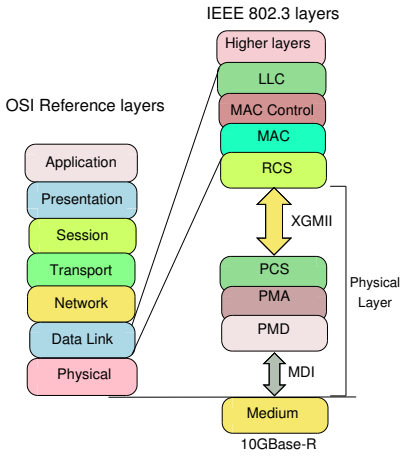


Figure 4.23: Position of 10GbE in OSI model.

MAC layer [56]. MAC layer initializes, controls and manages the peer to peer connection to prevent from transmission failure due to data collision. It acts as a bridge between the Physical layer and the data link layer. An interconnection between MAC layer and Physical Layer (PHY) is a 72 bit wide 10-Gigabit Media independent interface (XGMII) [56]. The parallel data path of the XGMII and the serial data stream of MAC is mapped by Reconciliation Sublayer (RCS). 10GbE MAC IP core and 10G-Base-R Physical layer (PHY) IP core [91] from Intel are used in this scheme. The internal block diagrams of the two IP cores are shown in the Figure 4.24. Both the Physical Coding Sublayer (PCS) and Physical Medium Attachment Sublayer (PMA) of the 10G-BASE-R PHY IP are implemented as hard IP blocks in FPGA to save resources. 10G-BASE-R PHY IP Core shown in Figure 4.24b delivers serialized data at a line rate of 10.3125 gigabits per second (Gbps) and supports optical communication. On the transmit path, the 10GbE MAC IP accepts the data frames and constructs the Ethernet frames. Data is parsed to the PCS module through RCS layer. Encoding, scrambling, rate matching is performed in the PCS sublayer, and the processed data is sent to the PMA sublayer module. The PMA serializes the encoded data into a bit stream suitable for serial bit-oriented physical devices and passes the stream to the Physical Media Dependent (PMD) [56] layer. Reverse action is performed on the receive path.

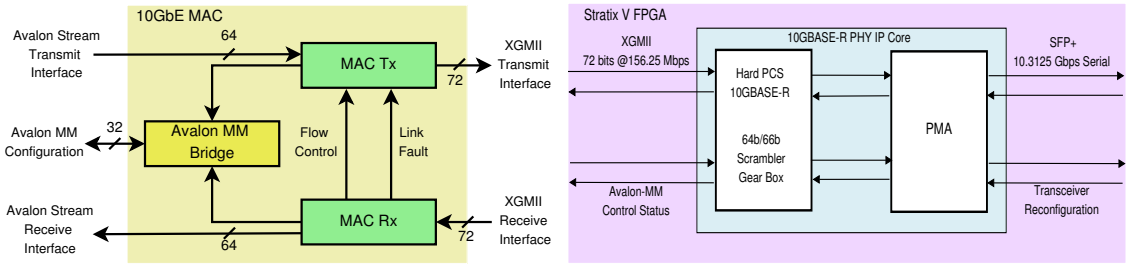


Figure 4.24: (a)10GbE Intel MAC IP core block diagram (b) 10GBASE-R PHY with Hard PCS and PMA in Intel devices.

Test setup

Development of the DAQ scheme involves the integration of the constituent components. Aim of the setup is to test the individual modules. It is important as the performance of the DAQ depends on the interactions between different constituents. Setup focusses on the testing of the interface links on the FPGA. It tests the compatibility of the components to transfer the valid data at the correct instance across the interfaces. Transceiver test forms an important part as it is the hardware interface to receive the data from the FEE to CRU and transmit it to the back-end processor. The 10GbE link interface are implemented and functionally tested on Intel FPGA. 10GbE link was implemented using the Intel's system integration tool Qsys [85] to adopt a modular approach. The entire test setup is segregated into multiple test models for easy step by step debugging and rapid fault finding of the constituent modules in case of faulty behavior and non-functioning of the scheme.

Implementation of 10GbE on FPGA

A test setup is developed for 10GbE implementation on FPGA utilizing Intel IP cores, along with the associated firmware and the embedded software. It is implemented on Quartus-II using the Qsys system integration tool for the quick generation of the interconnect logic and the functionality is verified using the Intel's ModelSim simulation software. The implementation includes **two** models. The *Model-1* focusses on the efficient method of high speed data transmission with minimum processor overload. It performs the loopback tests. The

Model-2 focusses on the optical link testing. It presents an effective approach to address the challenges associated with the testing, performance monitoring and parameter tuning of optical interconnects in FPGA-based systems.

Model-1: Test System Implementation on Quartus II platform using Qsys

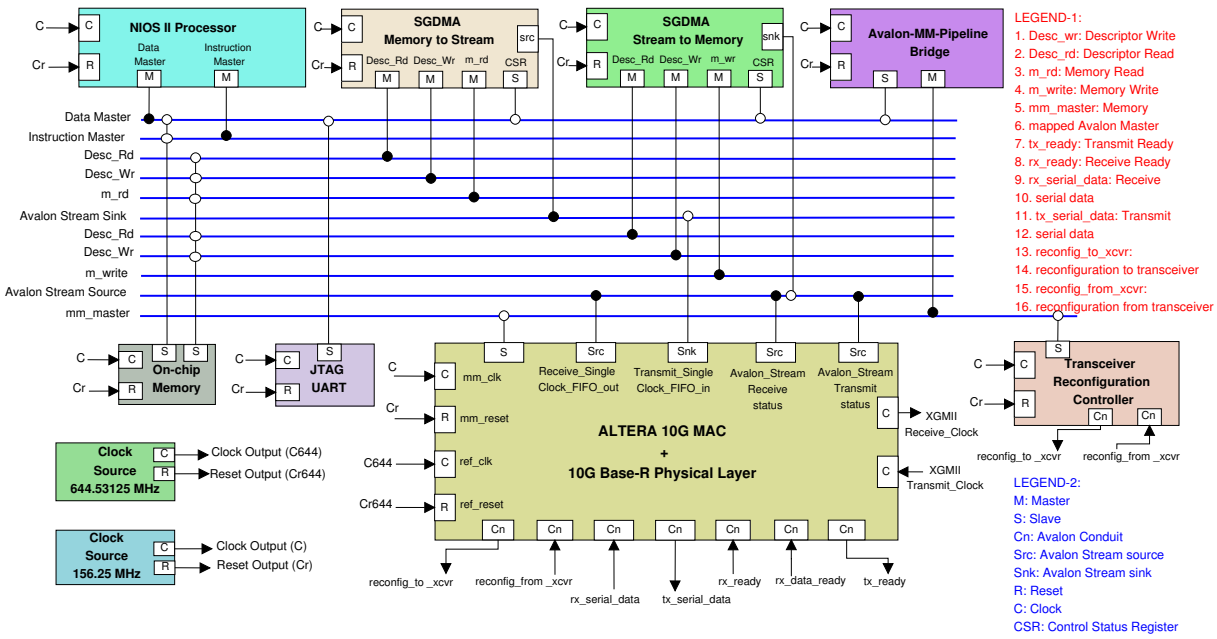


Figure 4.25: Test system implementation using Qsys system integration tool.

The architecture of the assembled system instantiated in FPGA and the interconnection between different sub-blocks is shown in Figure 4.25. The model consists of 10GbE Ethernet MAC IP core, NIOS-II processor IP, Scatter Gather Direct Memory Access (SGDMA) IP, JTAG UART [91], an On-chip memory, two On-chip dual clock FIFOs and a standard XGMII interface on the network side and configured to include 10G-Base-R PHY layer IP for optical communication. The implementation methodology is based on the resource intensive soft-core NIOS-II processor [92]. Its soft-core nature allows the designer to specify and generate a custom software over NIOS-II core. Nios II acts as a control unit in the loopback test, coordinates the design and provides overall system control. In this design, the SGDMA controller core is used for high-speed data transfer with minimal processor

hold-up [93]. It links the transfers to non-contiguous memory using a table of descriptors from memory. SGDMA improves the overall system performance as compared to the DMA cores. The On-chip memory stores the executable program, data, as well as descriptors for the SGDMA controllers. Dual clock FIFO buffers are used between the SGDMAs and the 10GbE MAC IP core for clock domain crossing.

Avalon Memory-Mapped (Avalon-MM), Avalon Stream (Avalon-ST) and Avalon conduit bus [84] are used as interface buses. A brief snapshot of the bus signalling is shown in Figure 4.26. Avalon-MM interfaces are used to implement the address-based read and write interfaces for the source and sink SGDMA. Avalon-ST interface on the client side is used to configure 10GbE MAC IP. Avalon-ST supports the unidirectional flow of data for the components that need low latency, high throughput point-to-point data transfer with data bursting and interleaving option. All the read/write signals and data transfer is synchronized with an associated clock interface. The control lines are implemented using Avalon-MM bus lines and data stream lines are implemented using Avalon-ST bus. Data buffers are

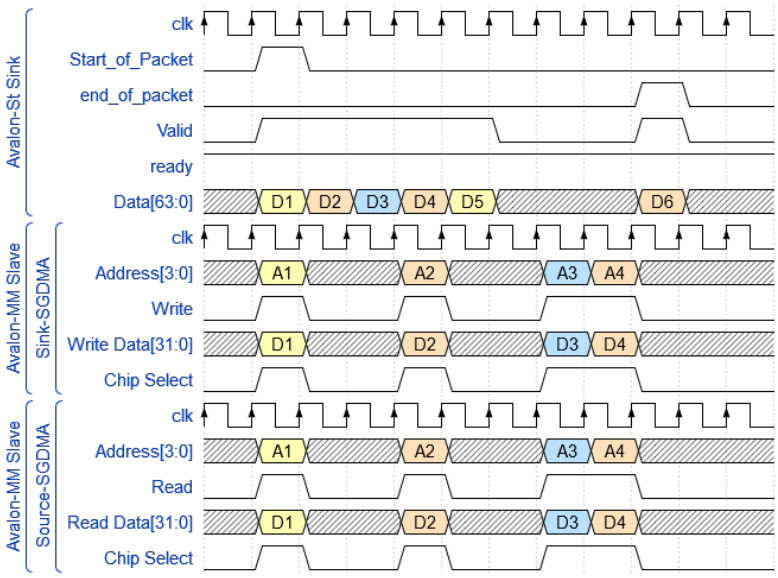


Figure 4.26: Interface of Avalon-MM and Avalon-ST with source and sink SGDMA data transfer.

transmitted through the system interconnect fabric maintaining Avalon standards as shown in Figure 5.6. The subsystem is programmed using the standard JTAG interface available on

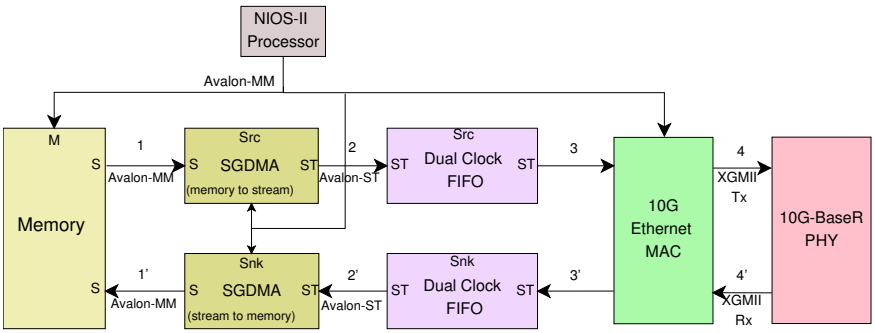


Figure 4.27: User Logic.

the Intel development board. The 10G-base-R PHY IP is operated in the internal loopback mode. Software based loopback test setup is developed using the Nios II Software Builder Tool (SBT) [94]. The NIOS-II processor runs the application program that handles the data transmission. It coordinates the design by allocating the memory to store the transmit and receive data buffers and the descriptor pairs. The test data is incremented in the transmit buffer; it populates the descriptor pair, writes the first descriptor pair to the SGDMAs, thereby starting the transfer, waits until both SGDMAs complete the transfer of all the data buffers. It also validates the received data with the transmitted data.

Model-2: Transceiver test

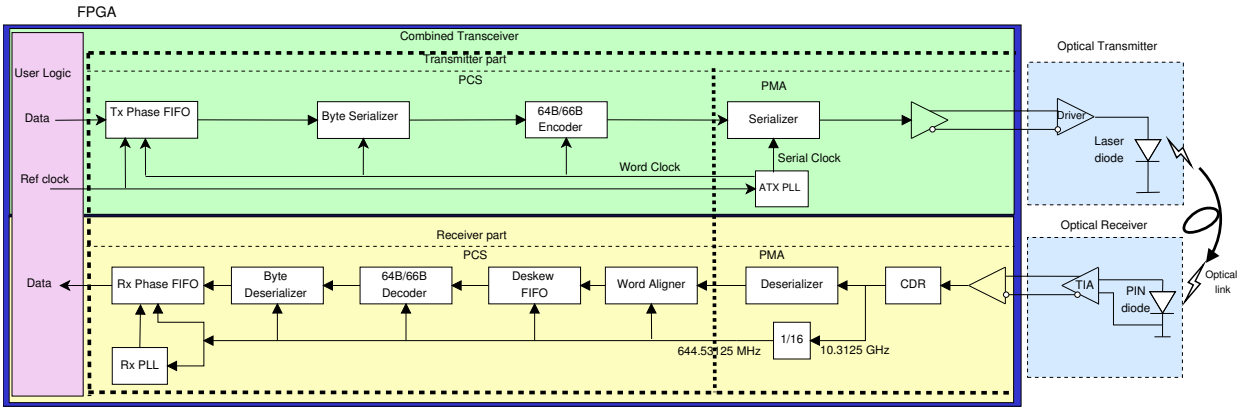


Figure 4.28: Simpified digital communication optical link

Optical link architecture for the digital communication in FPGA is illustrated in Figure 4.28. The data path consists of FPGA transceiver consisting of PCS and PMA, optical

transmitter (laser diode circuitry) and receiver (PIN diode circuitry) along with multimode optical fibre [95]. FPGA is connected to the transmission channel through the PMA block which generates the required clocks and perform serialization/deserialization. The digital processing between the PMA and the FPGA core is performed by PCS block. The PCS performs byte serialization/deserialization, byte ordering, rate matching, and 64B/66B encoding/decoding for the reliable digital data channel. However, we restrict the scope to the performance measurements of physical layer parameters, keeping aside the issues of the PCS sublayer. Transceiver Toolkit (TTK) [85] from Intel is used to validate the transceiver link signal integrity and to access and tune the transceiver settings in real time.

Tuning of the transceiver parameters is required for channel conditioning which affects the signal integrity and achieve the lowest possible bit errors. The major challenge lies in the fact that various components of the link have different parameter settings with a wide parameter optimization space and higher statistics are required to achieve low BER probability for a given confidence level [96].

The Auto-Sweep test is performed to identify the best PMA parameter settings [97]. Transceiver parameter settings like Voltage Output Differential (VOD), Pre-emphasis Pre-tap, 2nd Pre-Tap, 1st post Tap, 2nd Post tap, Equalization, DC gain and Variable Gain Amplifier(VGA) [98] are scanned and tuned for the optimal performance by the Auto-Sweep test in TTK. It also reports the signal quality of the received data in terms of eye diagram to understand the signal degradation mechanism. Eye diagram serves as an indicator of the link performance and is used as a target parameter for the link optimization. The test setup for BER measurements and to tune the transceiver parameters for the high speed optical link is shown in Figure 4.19. The best transceiver parameter values for targetted BER are called as solution space [97]. The output from the attenuator is looped back as shown in Figure 4.19. A Pseudo-Random Binary Sequence (PRBS) is transmitted across the transceiver link to evaluate the BER function with the pattern checker. BER at different attenuation levels were measured. This test characterizes the sensitivity of the receiver and

the minimum optical power required for achieving a specified BER in a system. Details are discussed in section 4.4.3.

4.4.3 Performance Evaluation

Test results and the performance analysis for the implementation of the interfacing links on FPGA are discussed in this section.

It includes two models. Model-1 presents the implementation results of 10GbE on FPGA and the analysis in terms of resource utilization, stages of the frequency translation, format of data transmission and the latency involved. Model-2 presents the transceiver tests, tuning to achieve solution space, spider chart, eye diagram and BER measurement as a function of optical power.

Model-1 results

The test setup shown in Figure 4.25 is implemented on FPGA and the logic resources utilized are summarized in Table 4.8. The data is transmitted from the fabric clock frequency of

Table 4.8: FPGA resource utilization for the 10GbE design.

Parameters	Family: StratixV 5SGXEA7N2F45C3
Logic utilization (in ALMs)	6,685/234,720 (3%)
Registers	11291
Block memory bits	586,560/52,428,800 (1%)
RAM Blocks	51/2,560 (2%)
HSSI 10G TX/RX PCSs	1/48 (2%)
HSSI PMA TX/RX Serializers/Deserializers	1/48 (2%)
PLLs	3/92 (3%)

156.25 MHz to 10.3125 Gbps at the transceiver. This frequency translation occurs at three stages as shown in Figure 4.29. Data output from the 10GbE MAC is transmitted to the 10G-base-R PHY over the XGMII parallel lines each at 156.25 MHz. These are multiplexed to 16 parallel lines, each operating at a frequency of 625 MHz keeping the bit rate same. The frequency of each bit is shifted to 644.53125 MHz after encoding in the PCS layer. At the Serializer/deserializer, data is serialized and each bit is transmitted from the silicon

to the physical media, with a data rate of 10.3125 Gbps. The reverse operation occurs at the receiver. Data Packet transmission in 10GbE MAC complies with the IEEE 802.3ae

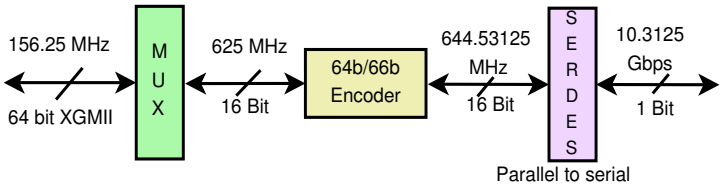


Figure 4.29: Three level of frequency translation in 10GbE communication.

Ethernet standard [56] when transmitting data frames on the XGMII interface. 10GbE MAC transmitter performs the endian conversion [99] and the frames received on the Avalon-ST interface from the user follows big endian format. The transmission on the XGMII interface follows little endian format by transmitting the frames from the least significant byte as shown in the Figure 4.30. In the receive data path, the 10GbE MAC Receiver decodes the data lanes coming through the XGMII. For all valid frames, the 10GbE MAC receiver removes the START, preamble, SFD, and EFD bytes and ensures the byte-wise frame alignment. The data transfer latency regarding the clock cycles is calculated for the user logic shown in Figure 5.6 and summarized in Table 4.9.

Table 4.9: Latency estimation for data transfer(1 clock cycle = 156.25MHz).

Path	L_{21}	L_{32}	L_{43}	$L_{2'1'}$	$L_{3'2'}$	$L_{4'3'}$
Latency(Clock Cycles)	0	3	30	0	9	12

Model-2 results

Transceiver testing is done as discussed in section 4.4.2 utilizing Intel TTK design operating in 10Gb mode. The transceiver is configured in far-end optical loop-back mode. PRBS31 is used for optimizing the parameters as it provides the most stressful boundary conditions to achieve a confidence level in the operating margins of design as shown in Figure 4.31. Autosweep test has been performed to scan the best performing case concerning Eye Width/Height at targetted BER of 10^{-12}. As indicated by the Auto-Sweep test,

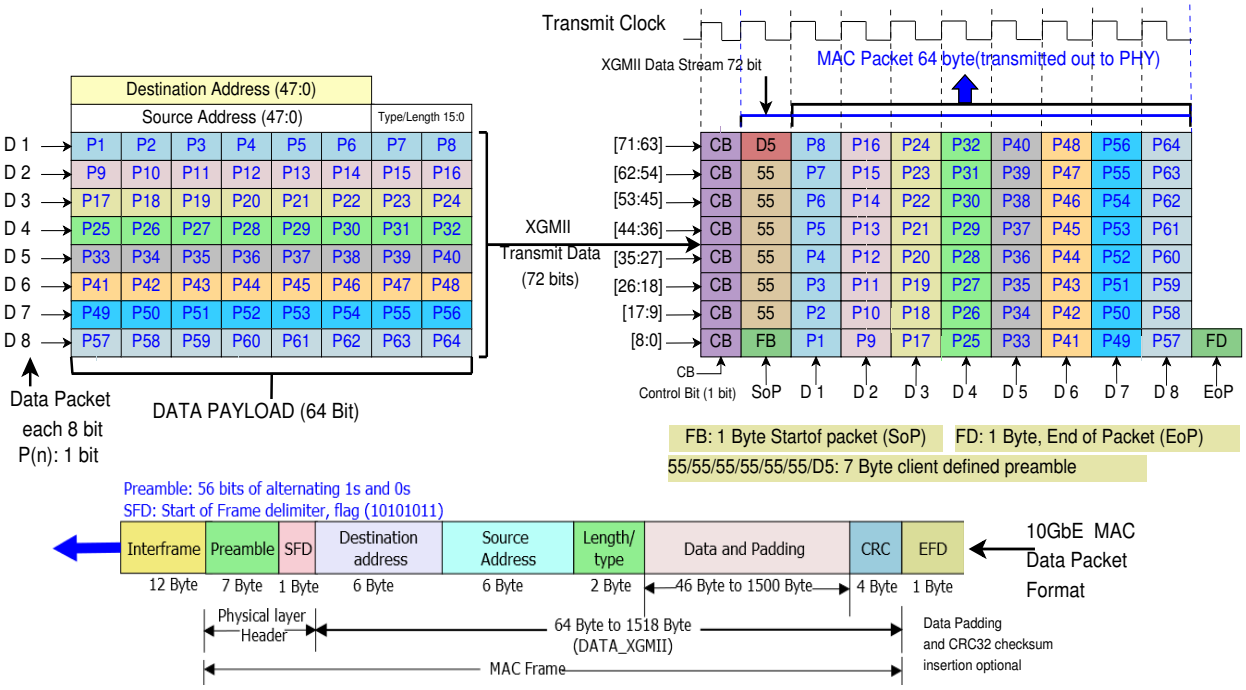


Figure 4.30: MAC to XGMII data payload conversion scheme

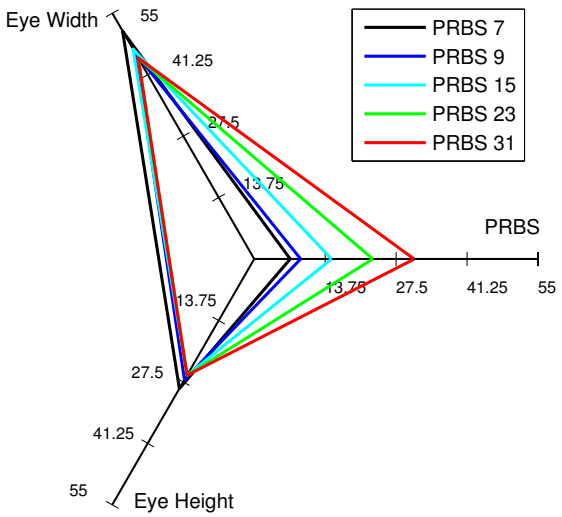


Figure 4.31: Variation of Eye width and Eye Height with PRBS type.

solution space is plotted in the form of spider chart as shown in Figure 4.32. The param-

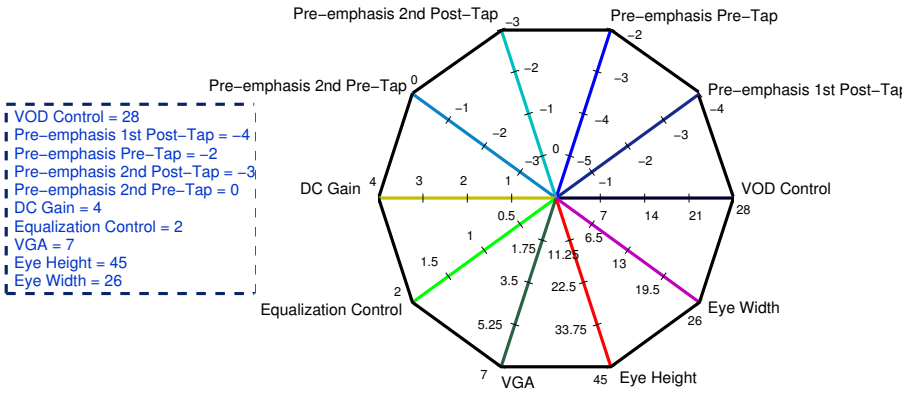


Figure 4.32: Plot for tuning of transceiver parameter optimized settings at PRBS31.

eters are fixed and the eye diagram is captured at the best PMA settings, and it is shown in the Figure 4.33 with Eye Width(Horizontal Phase Step)/Eye Height(Vertical Step) as 45/26. BER at different attenuation levels of optical transmitted power [100] are measured

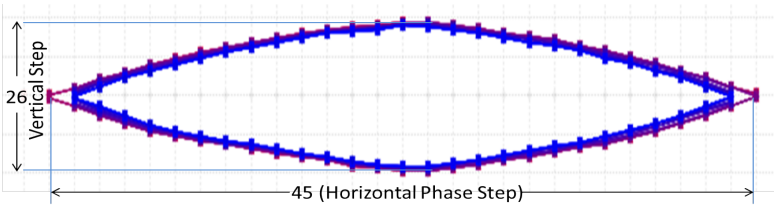


Figure 4.33: Eye Diagram for the 10Gb Ethernet on FPGA.

as shown in Figure 4.19. BER as a function of received optical power is shown in Figure 4.34. Optical transmitted power of around -11 dBm is required to achieve the BER of 10^{-12} for the transceiver under test. The exponential fit through the data points yields equation of the form $BER(dB) = a \times e^{b \times Power(dBm)}$, where coefficients a and b are -144.33 and 0.22887 respectively. The exponential fitting is done as BER is approximated by complementary error function 'erfc' and the system noise is Gaussian in nature; in logarithmic scale, it is approximated as exponential. The statistics for goodness-of-fit; the Sum of Squares due to Error (SSE), R-square, Adjusted R-square and Root mean squared error (RMSE) is 11.71, 0.9698, 0.9686 and 0.484 respectively.

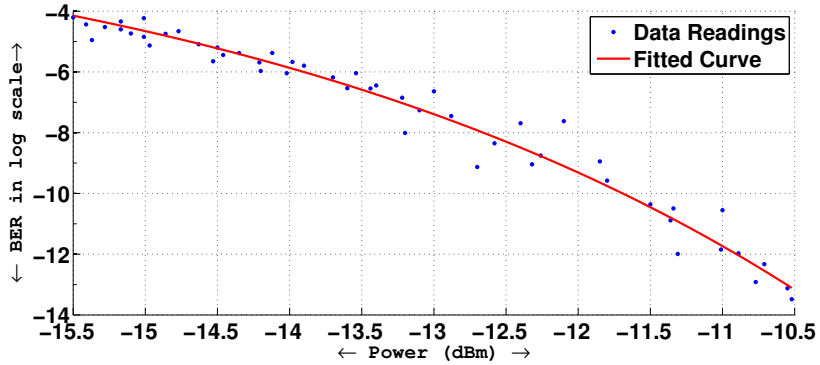
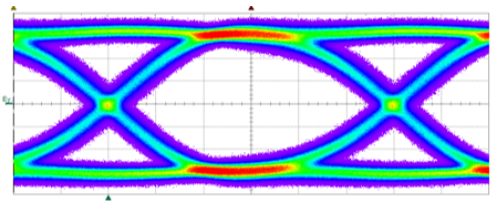


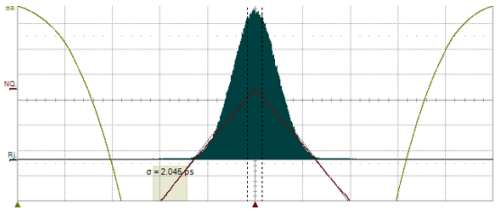
Figure 4.34: Bit error rate as a function of received optical power at 10Gbps.

4.5 Avago MiniPOD performance tests

Avago MiniPOD^{TM} (on-board transceiver) performance tests at transmission rate of 10Gb using Lecroy oscilloscope is shown in Figure 4.35. The increasing trace length in PCB



(a) 10 GbE Tx Eye Diagram Schematic



(b) 10 GbE Jitter Spectrum

Figure 4.35: MiniPOD^{TM} performance with 10 GbE protocol (10.312 Gbps).

material degrade the performance of the high-speed electrical signals. Also, there is a possibility of EMI containment with 12 optical channels interfaced in a single optical module, or MiniPOD [87]. Hence, a comparative signal integrity measurement for the PCIe40 is done using the measurements of 4.8 Gbps GBT protocol (as shown in Table 4.7) and the measurements of 10.3125 Gbps for 10 GbE protocol. The comparison is summarized in Table 4.10.

Table 4.10: Shows the comparison of MiniPOD^{TM} performance for GBT protocol and 10GbE Protocol

Parameters	**10 GbE Protocol**	**GBT Protocol**
Link Speed	10.312 Gbps	4.8 Gbps
Eye height	373 mV	373 mV
Eye Width	79.4 ps	176.8 ps
Deterministic Jitter (Dj)	5.172 ps	13.125 ps
Data Dependant Jitter (DDj)	9.158 ps	21.647 ps
Periodic Jitter (Pj)	0.675 ps	7.240 ps
Duty Cycle Jitter (DCD)	2.062 ps	1.912 ps
Inter Symbol Interference (ISI)	7.399 ps	21.657 ps
Random Jitter(Rj)	2.204 ps	3.204 ps
Total Jitter (Tj)	36.180 ps	58.185 ps
Standard Deviation(σ)	2.045 ps	2.898 ps

4.6 Resource Estimation

An elementary integrated firmware is designed which includes all the Low level interfaces with individual data pattern generators and the checkers. This is to estimate the amount of logic resources utilized; before the implementation of the full implementation of generic firmware. A preliminary resource estimation report is presented in Table 4.11, as part of integration test. The result in Table 4.11 is likely to evolve with firmware development. the FPGA used is Arria-10 (part number 10AX115S4F45I3SGE2). It also indicates that enough resources are present for detector specific firmware development. The intrinsic details related to the CRU firmware designs are available at the ALICE-CERN CRU TWiki page [101]

Table 4.11: An integrated design for elementary firmware including low level interfaces.

Aggregated Links without inter-connect glue logic	**Logic Utilization**	**HSSI SERDES Utilization**
48 GBT link + x16 PCIe Gen 3 + SFP+ (Transceiver Toolkit design)	34,614 / 427,200 (8.10%)	65 / 72 (90%)
36 GBT link + x16 PCIe Gen 3 + SFP+ (Transceiver Toolkit design)	32,247 / 427,200 (7.55%)	53 / 72 (74%)
24 GBT link + x16 PCIe Gen 3 + SFP+ (Transceiver Toolkit design)	29,771 / 427,200 (6.97%)	41 / 72 (57%)

4.7 Summary

The data flow of error resilient scheme to and from the CRU in connection with the other constituents of the system is presented in this chapter. For entire system to operate efficiently and correctly, the individual design needs to be optimized and be able to perform in

coherence to another blocks. The focus of the carried out study is on the implementation and detailed tests of the front-end and the back-end interfaces. The links are implemented on FPGAs. Two custom DAQ boards based on latest Stratix-V and Arria-10 FPGAs from Intel are used for the comparative study. The source code of the configured firmware are chronicled in CERN espace, and resources are accessible to clients on solicitations.

The phase calibration logic is implemented for the latency optimized mode of GBT operation. The calibration design is developed out of the necessity to measure small phase shift changes inside an FPGA. It is to ensure the synchronization of phase between the clocks from different domains. There are phase drifts due to temperature variation also and pushes the system in a unstable zone if the synchronized region lies close to the metastable region. Hence, the calibrated phase need to be chosen carefully and a study is done for the link stability with temperature variation also. Reliability study for the consistency of phase of the clocks is conducted by performing multiple iterations of reset assertion/de-assertion cycle (PFR cycle), power on/off cycle and the firmware upgrade. For easy system integration of the GBT firmware, an Intel-Altera QSYS framework is designed. Important features of the design like temperature, coding type, mode of operation, transmit and receive link data, state of phase locked loop and the latency value in nanoseconds are integrated in GBT QSYS model for ease of operation. From the detailed analysis it is concluded that the GBT protocol can veritably be implemented in both Stratix-V and Arria-10 Intel-Altera FPGAs.

The FPGA resource utilization and power consumption are measured. Latency calculation gives a measure of the clock cycles utilized for the data processing in the logic path and the distribution of the buffer (elastic or external phase aligner) in the transmission path. This information is a useful reference for the designers to optimize the data acquisition firmware. Latency in terms of clock cycles for Tx Latopt and Rx Std mode of GBT operation is found to be 14 clock cycles (350 ns) which is the most utilized mode for fixed latency operation. Tx Latopt mode is necessary to send the timing information in a deterministic way whereas on the receiver side the data comes padded with time stamp and hence the

timing constraint is relaxed on receiver side that allows the use of Rx Std mode. Signal quality of the GBT protocol is measured using eye diagram with BER of the order of 1 bit in 10^{12} and jitter range of picoseconds only. It is found that the measurement of BER for GBT protocol with respect to the optical power cannot be pursued below -17 dBm receiver sensitivity, due to the loss of recovered clock. However, the plot can be extrapolated based on standard complementary error function nature of the curve, assuming the Gaussian nature for noise. The margin of receiver sensitivity is found to be 2.1 dBm for the two encoding schemes of GBT at the targetted BER of ~ 10^{-12}.

For the back-end interface of CRU the PCIe Gen3 x16 is the opted link as per the form factor of the PCIe40 custom board. It will reduce the cabling and CRU will directly be inserted in server slots. The maximum achievable throughput for the Gen3 version of PCIe with 16 lanes is discussed and the exeperimental tests and results detailed in chapter 6. However, the other option of 10GbE as backend interface has also got future proof with ample ecosystem rather this could be usable with hardware of different form factor. As a study, 10GbE link is also implemented using the Qsys approach and the three level of frequency translation from the fabric frequency to the optical transmission is discussed. Endian conversion during the data packet transmission for the protocol is explained graphically. The number of clock cycles required to transmit the data buffer through the system interconnect fabric are calculated. Transceiver is tuned for the high-speed link using signal conditioning circuitry as it forms the important hardware interface for data transmission from GBT-chipset located at the FEE to CRU at 4.8 Gbps and from CRU to the DAQ server. Autosweep test is performed using Intel TTK and the multivariate data for the best case is displayed on a 2D spider chart also known as Kiviat diagram. The variation of BER at the speed of 10Gbps as a funtion of optical power is observed upto -15.5 dBm, below which the receiver sensitivity is lost. The deviation of the data set from the exponential fit is due to the various parameters for instance, opto-electronics conversion factor, gain, optical couplings, the insertion losses and the accuracy of the instruments used. The uti-

lization of the FPGA logics for the integrated firmware is estimated to optimize the usage of resources and to judge the suitability of the chosen FPGA Eye diagram analysis and jitter measurements at different data rates are performed; for the quantitative measurement of signal quality of the data from MinipPODs (transceiver). In the next chapter an efficient FPGA based technique is discussed for the optimization of multigigabit transceivers.

Chapter 5

Optimization of multi-gigabit transceivers for high speed data communication links

5.1 Introduction

Most of the DAQ systems are designed using the present available technology in such a way that it could be easily upgraded to match the requirements of the system. Since one of the major concerns is to efficiently acquire data for all the collisions, error resilient and efficient data transmission with minimal signal attenuation is required. Signal integrity is essential for the proper Clock and Data Recovery (CDR) [30, 75]. Thus it is a challenge to minimize the BER and improve signal integrity for increased data rates [102].

In this chapter we address the challenges of high-frequency losses arising due to the high data rates for the DAQ systems in HEP experiments. Using FPGA we present a heuristic optimization technique to tune the parameters of multi-gigabit transceivers for achieving the best performance at high-speeds for the transmission of data, trigger, timing and slow control information. The proposed technique helps to improve the system performance

in terms of signal integrity and is implemented on a state-of-the-art 20nm Intel Arria-10 FPGA [70]. It uses the Intel-Altera on-die Instrumentation tools [85] and does not require the probing of FPGA pins or transceiver attributes. The full setup is tested for the link rate of the high-speed communication protocols frequently used for data transmission in these experiments. The technique is useful for on-field system-level debugging, and the parameters can be reconfigured dynamically, allowing the user to configure the transceivers for optimum performance. The robustness of the optimization technique has been tested with Pseudo Random Binary Sequence31 (PRBS31) pattern, which represents the stressed and transitional data conditions. For the statistical reliability of the performed tests, a large number of data vectors are acquired. Different performance indicators, such as, BER and eye diagrams have been used to verify the improvement of the quality of data signal posterior to the execution of proposed optimization technique.

The chapter is organized as follows. Details of the transceiver optimization technique with its intricate features are presented in section 5.2. Section 5.3 describes the FPGA based test setup, and section 5.4 discusses the methodology to implement the proposed technique and its advantages. The test results are presented and discussed in section 5.5. The results are summarised in section 6.4.

5.2 Transceiver optimization

High-speed data communication suffers from the transmission losses and signal integrity issues; not seen at normal digital signalling levels [102]. The high-frequency content of the signal gets degraded due to dielectric losses, skin effect, discontinuities in connectors, reflections caused by the vias, inadequately placed traces, etc. We have developed a technique to optimize the transceiver parameters accurately and offer the best combination for a given high-speed link. This optimization of the transceiver parameters could take care of the transmission losses [103].

For the high-speed transmission channels with multi-gigabit rates, the unit interval (UI) for the data bit decreases. At high transmission rates, the PCB materials suffer from frequency dependent losses, hence become dispersive. This prevents the signal from reaching its full strength at the shrunk UI window, leading to jitter and intersymbol interference (ISI). It also disturbs the deciphering of the signal and the extraction of the embedded clock becomes difficult at the receiver end.

An increase of the signal strength is an obvious solution to overcome the attenuation. However, the issue of high-frequency roll-off remains, and the pattern dependent jitter gets aggravated. Consequently, the signal does not reach its optimal strength within the interval and may diffuse further into the next UI leading to ISI. Also for the increase of signal strength overall power consumption of the transceiver increases. Noise levels in the system also increase proportionally. All these lead to deteriorated metrics of signal integrity and reduced drive length. The effects are even more evident with the use of high-speed interfaces with the systems which were originally designed for low bandwidth applications.

To overcome these losses, we have developed the transceiver optimization technique and a proficient methodology for 20nm Arria-10 FPGA. This new FPGA with considerably large on-chip resources [70] are ideal for HEP experiments.

5.2.1 Optimization Technique

For the optimization, the high-frequency components in the data stream are boosted up on every switching, using the digital pre-emphasis taps of the on-chip transceiver. In addition, the low frequency components are reduced. This technique helps to achieve the same amount of emphasis with less power dissipation. The exaggerations are overridden by the attenuation during transmission and allow for the signal to be recovered accurately.

The optimization technique has been implemented on Intel Arria-10 FPGA development board with integrated reconfigurable transceiver architecture [70]. It incorporates additional circuitry in buffers for equalisation and pre-emphasis techniques. The transmitter of the

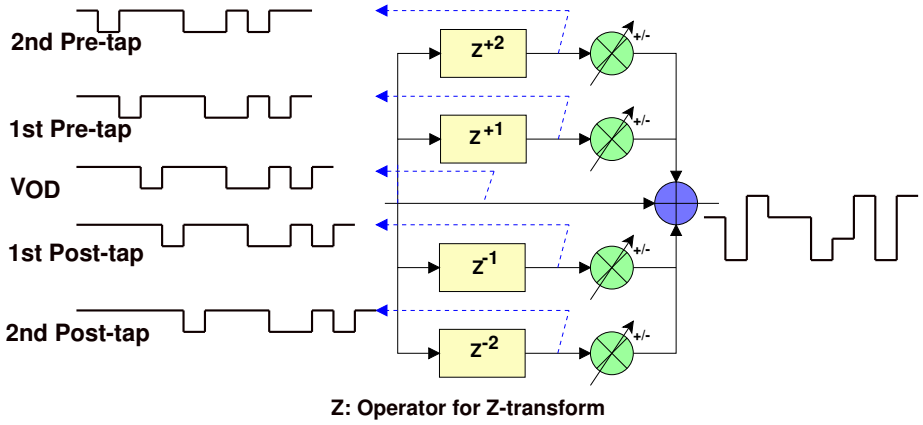


Figure 5.1: Voltage output differential (VOD) and tunable pre-emphasis taps with flexible polarity in the embedded transceiver of FPGA.

embedded transceiver has five programmable drivers as shown in Figure 5.1. Voltage output differential (VOD) controls the base amplitude. The four pre-emphasis taps are 1st pre-tap, 2nd pre-tap, 1st post-tap and 2nd post-tap. These taps also include polarity settings. The post taps are the causal taps and the pre-taps are the anti-causal taps. These multiple taps and choice of polarity could handle channel attenuating characteristics. Equalisation with DC gain and Variable Gain Amplifier (VGA) is on the receiver side of the transceiver. There are multiple transceiver parameters with a large span of operating range and so to scan the system performance for every combination of the parameters is a time-consuming process. Our goal had been to develop an efficient technique for optimization of transceiver parameters such that the signals impacted by the high-frequency losses are recovered.

It performs like a *Finite Impulse Response (FIR)* filter with different delays referred to as the taps as shown in the Figure 5.1. An FIR filter is based on a feed-forward difference equation. The pre-emphasis technique applies a delay to the signal and adds it back to the real signal with weight and inversion as and when required. Although depending on the transmission channel peculiarity, a simple delay, weight and inversion may not be able to provide the required compensation. For this reason, a combination of different delays, weights and the polarity are combined. In this configuration, the pre-emphasis 1st post-tap is the most useful parameter. It emphasises the immediate bit period after the transition.

The generation of the differential emphasised signal, applying the unit delay by the first post-tap is shown in Figure 5.2, assuming VOD = 1 and tap weight as $0 < x < 1$. The original positive signal Vp(T) is compared with Vp(T-1) which is the unit-delayed signal. The emphasised signal is the difference between the weighted x*Vp(T-1) signal and the Vp(T) signal. The negative signal is similarly generated. The pre-emphasised differential signal is differentiated from the positive and negative signals. The effect of 2nd post-tap

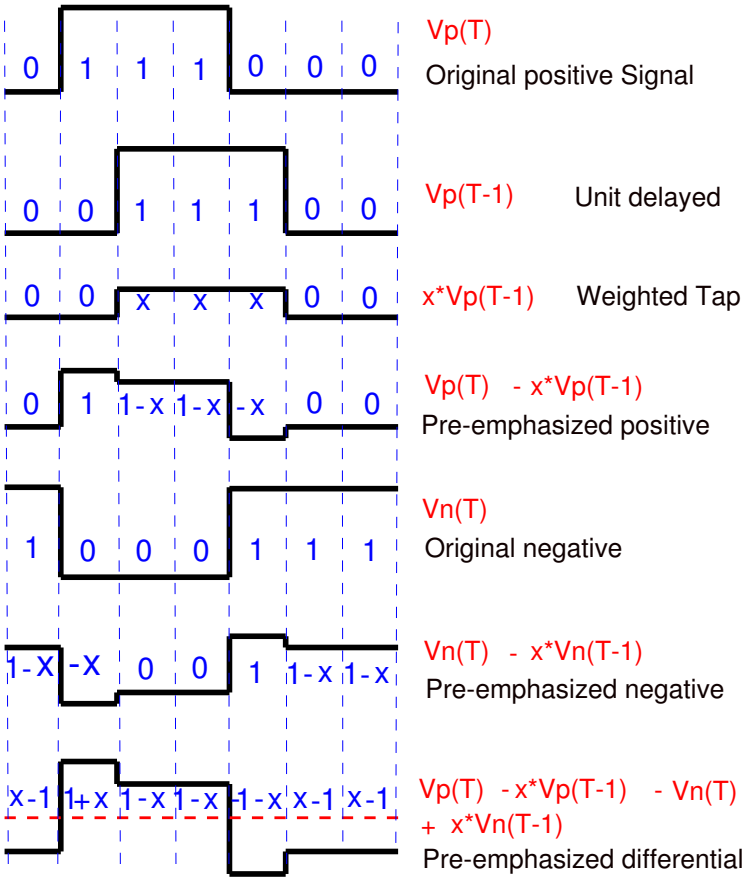


Figure 5.2: The pre-emphasis signal generation technique at the 1st post-tap in embedded FPGA transceivers, $0 < x < 1$ is the tap weight.

after the transition, depending on the chosen polarity setting is shown in Figure 5.3.

The pre-tap reduces the effect of pre-cursor ISI. Figure 5.4 shows the impact of 1st pre-tap and the 2nd pre-tap on the single and double bit period respectively, before the occurrence of high-frequency transition depending on the polarity. Both pre-cursor ISI and post-cursor ISI are handled by anti-causal and causal taps respectively. However, pre-emphasis alone

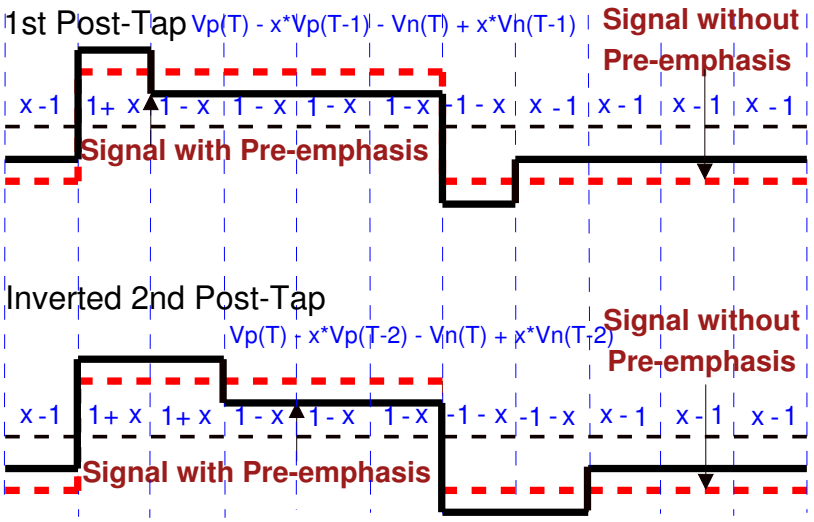


Figure 5.3: Pre-emphasis 2nd post-tap (Inverted) compared with pre-emphasis 1st post-tap and their effect on the signal without pre-emphasis.

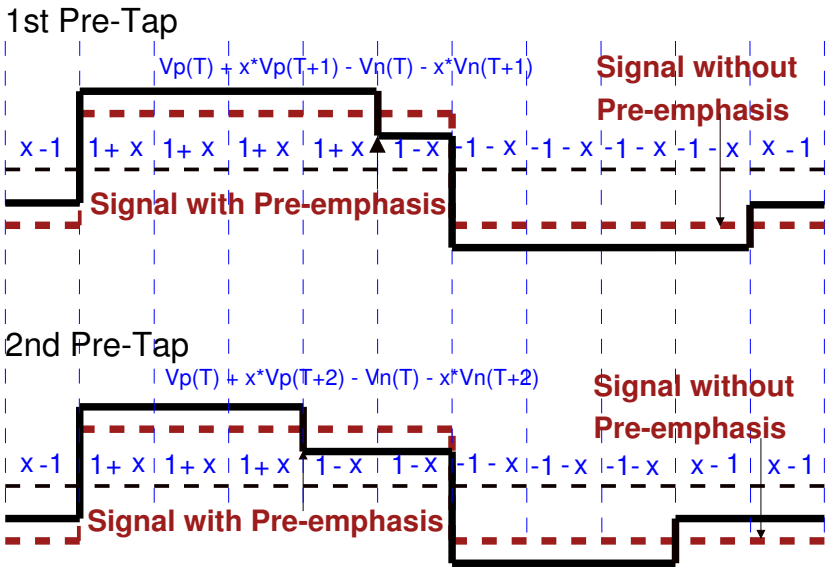


Figure 5.4: Pre-emphasis 1st pre-tap and the 2nd pre-tap (Inverted) and their effect on the signal without pre-emphasis.

cannot guarantee the performance of the system as it is implemented at the transmitter by pre-conditioning the signal before it is fed to the channel. There are high-frequency losses in the transmission channel itself. Hence an equalisation is required at the receiver end. It compensates for the low pass characteristics of the physical medium and amplifies the attenuated high-frequency components of the incoming signal. An equalizer on the receiver side lifts the contents inside a band of frequencies and attenuates the rest. The DC gain

circuitry gives uniform amplification to the received spectrum. It enables the transceivers to operate over longer distances. The VGA on the receiver optimizes the signal amplitude before the CDR sampling.

To achieve an optimal signal integrity performance, both transmitter and receiver parameters of the transceiver on FPGA chip augments each other and work combined to compensate for the high-frequency losses. However, the overcompensation degrades the signal quality and adds more jitter leading to the closed eye diagram rendering it futile for the receiver to identify the signal and hence should be avoided.

5.3 Test setup

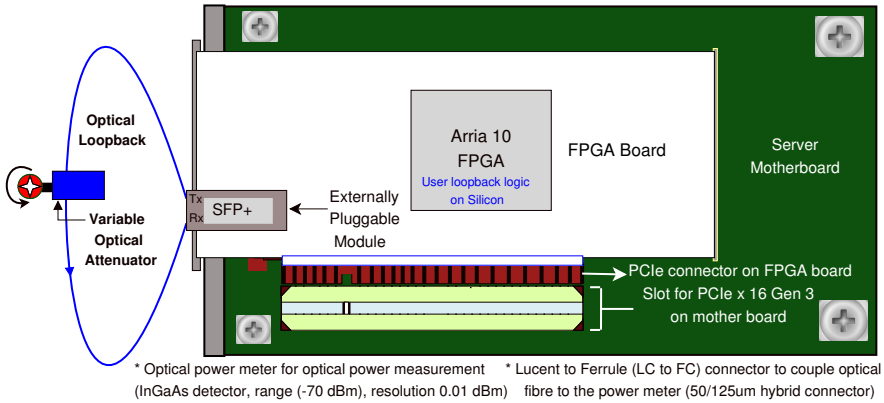


Figure 5.5: Arria-10 FPGA card inserted in PCIe x16 slot of server. The optical signal from the externally pluggable SFP+ is looped back via the fibre equipped with the variable optical attenuator (VOA).

An FPGA based setup has been developed to test the potency of the proposed optimization technique. The transceiver is tested for the high-speed links under the stressed conditions. The setup has been utilised to emulate the stressed high-speed link conditions and to investigate the high frequency losses in the transmission. It determines the capability of the transceiver system to recover the data from the degraded signals. Tests are performed at the system level to operate the setup at a prescribed BER equal to or better than 10^{-12} as per the IEEE standard.

Table 5.1: Components used in the test setup, their role and specifications.

Component	Role in test setup	Specification
FPGA Test Board	Integrated FPGA based design environment with embedded transceivers on silicon. PCIe connection. Slot for hot pluggable transceiver optical modules. Other accessories	Intel Arria10 FPGA, (20nm mid-range). Transceivers upto 17.4 Gbps [70].
Variable Optical Attenuator (VOA) with optical Fiber	Optical power attenuation in the fibre loopback path.	Range(dB)-0~60, Accuracy +/- 0.8dB. Fibre(850nm): Multimode 50/125um with Lucent connector (LC), Dia-2mm, Insertion loss <2.5dB, Length-2 m
Serial Form factor Pluggable (SFP+) module.	External transceiver modules to be coupled to the fibre. Laser at transmitter and PIN diodes at the receiver ends	Hot-pluggable footprint, upto 10Gbps, 850nm VCSEL laser, duplex LC connector. Link length of 300m [104].
Workstation with FPGA design platforms	FPGA board powered through PCIe Gen3x16 slot. Compile and generate the FPGA design with firmware development softwares	PCIe Gen3 x16 slots available. Quartus-II platform installed for firmware design and generation. FPGA programmed through USB blaster download cable.

The test setup, shown in Fig. 5.5, engrosses the Arria-10 FPGA development board (device 10AX115S2F45I1SG) for the implementation and testing of the optimization technique. The FPGA development card is installed on the PCIe 16 lane slot of the server, where the power is obtained from the server motherboard. The functions and specifications of each of the components of the setup are given in Table 5.1. Intel Quartus-II platform is

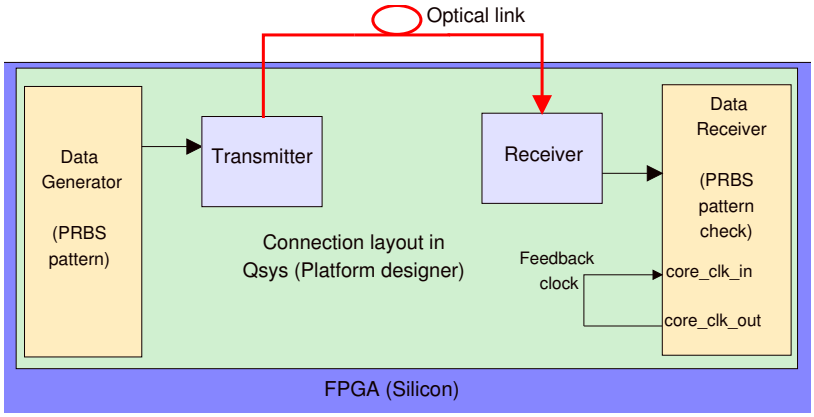


Figure 5.6: Typical BER test loopback logic on FPGA using Qsys tool. PRBS patterns are generated. The serialised data is transmitted, looped back and checked for the flipped bits at the receiver.

the firmware application package, implemented on the FPGA logic design. The transmission links at the specified data rates are implemented using Quartus-II Qsys tool. Qsys is Intel's system integration tool for the quick generation of the interconnect logic. The signal integrity of the transceiver links is validated using Transceiver Toolkit (TTK) feature of

Quartus-II with a GUI. The TTK is used to quickly access, tune and test the transceiver parameter settings in runtime through a combination of metrics. The TTK enables us to measure BER and the eye diagrams and also verify the signal integrity in external loopback mode. Details of the firmware-tools, such as, Quartus II, Qsys, TTK, PRBS patterns and auto-sweep features may be found in reference [85].

For the data loopback tests [105], multimode optical fibre equipped with Variable Optical Attenuator (VOA) and external pluggable SFP+ modules are used. The far end of the transceiver is coiled back to the receiving end. The received data is then verified by the data checker logic on FPGA for any erroneous bits as shown in Figure 5.6. To test the signal integrity a variety of data patterns can be used. However, in each case, a checker must be available for verification. PRBS patterns are injected into the test system as it generates the stressed and lengthy patterns with fewer memory consumption [106]. Another advantage of using PRBS patterns for the tests is that the boundary synchronisation is not necessary at the physical layer as the patterns are time correlated. The Intel soft logic cores are used for PRBS data pattern generator and checker [85].

The BER measurement approach was chosen with respect to the controlled attenuated optical power at the receiver with the help of VOA. It allowed us to rapidly characterise the transceiver sensitivity below which the embedded clock cannot be recovered from the data stream, and loss of lock occurs [103]. It also determines the minimum required optical power to achieve the targeted BER for a system operating at a specified data rate. Auto sweep feature of TTK is used to obtain the optimum settings of the best performing parameters of the transceiver for a specified BER. This optimized set of transceiver parameters delivers the best metrics of signal integrity and the eye diagrams by its height and width. In the next section, we elaborate the methodology for the optimization of high data rate on-chip transceivers to reduce the effect of high-frequency losses.

5.4 Methodology

The methodology to extract the optimized settings of the transceiver parameters has been explained in the flowchart in Figure 5.7. To start with, the optimization process scans the full range of each transceiver parameter using the TTK auto-sweep feature while the rest of the parameters are set at their Intel-default values. Then it records the best performing tap setting values for each transceiver parameter as indicated by eye parameters. At this instance, a Solution Matrix (**S**) at Nth iteration, set $N = 1$ is developed. Then, we separately group the receiver parameters viz. VOD, Pre-emphasis (1st pre-tap, 2nd pre-tap, 1st post-tap, 2nd post-tap) and the transmitter parameters (DC gain, Equalisation control, VGA). Then we scan again the transmission and receive parameters separately in the range of -3 ≤ **S** ≤ 3, while receive and transmit parameters respectively are set at the values enlisted in the **S**. Record again the best performing cases and update the **S** with newer values, increment N by 1. Assign the latest matrix values to the TTK and run the loopback test. If this does not result in the improved metrics of signal integrity (Eye diagram and the BER) than the one obtained at the Intel default set values; repeat the optimization loop with the adjusted **S** values in the range defined until the improvement in both eye diagram and BER is achieved.

The parameters cannot be declared as optimized until a stage of degradation in the signal integrity metrics from their peak values is observed. The degradation of metrics denotes the over-compensation and it marks the transition from the maxima of the transceiver parameters. Assign and update the **S** with the best performing case metric values rejecting the over-compensated value set. The final **S** values with the best performing metrics is known as *Solution Space* [103]. The deduced final values are fed to the transceiver for further analysis. The results are presented and discussed in the next section. The proposed technique has definite advantages over traditional method where the transceiver optimization may be carried out in an extremely time-consuming way by evaluating the signal integrity through a large number of permutations and combinations of the parameters. The parameters and

their possible ranges are listed in the Table 5.2.

Table 5.2: Transceiver parameters, range of operations for the manual optimization.

Transceiver parameter	**Range of possible values**	**Number of iterations required**
Transmitter Side		
VOD	0 to 31	32
Pre-emphasis 1st post-tap	-31 to 31	63
Pre-emphasis 1st pre-tap	-31 to 31	63
Pre-emphasis 2nd post-tap	-15 to 15	31
Pre-emphasis 2nd pre-tap	- 7 to 7	15
Receiver Side		
DC gain	0 to 4	5
Equalisation	0 to 15	16
VGA	0 to 7	8

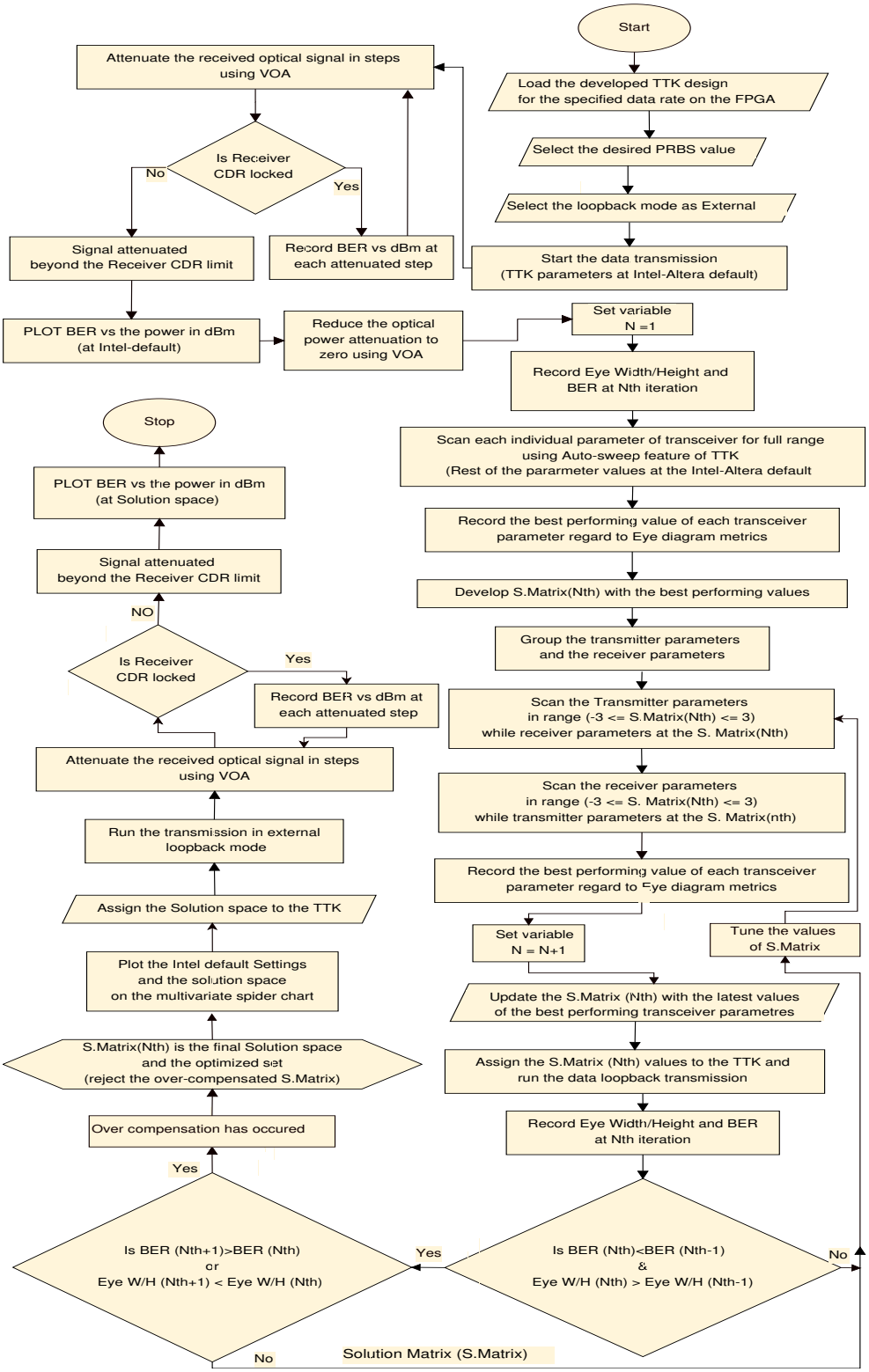


Figure 5.7: Stepwise flow diagram for the Transceiver Optimization. Data transmission is started with the Intel default parameters and a Solution matrix is derived to achieve the optimized signal integrity

5.5 Results and discussion

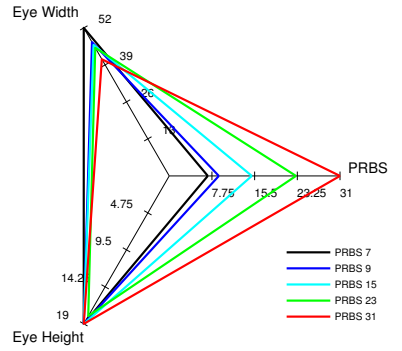
Results are demonstrated and validated for the three different high speed optical links: 10 Gbps links, 4.8 Gbps GBT protocol and 9.6 Gbps TTC-PON. The test system confronts the lock and hold capability of the CDR circuit, perturbs all the conceivable instances of ISI and analyses the receiver sensitivity for any probable drifts. Drifts at the receiver are caused due to long imbalanced runs of the data transition pattern. The PRBS31, $2^{31} - 1$ patterns integrate every alteration of 31 bits. It gives a random sequence of bits with high and low transitional values as defined by the logic levels of FPGA. The different combinations induce non-similar ISI configurations. It is required to stress the transceivers, test any innate ISI in a transmitter, and to assess the quality of transmission. PRBS patterns depict a white spectrum in the frequency domain and are injected to tests the robustness of the high-speed links. For the entire analysis, PRBS31 is used to stress the system. However, the variation of eye diagram and BER characteristics are also studied for PRBS7, PRBS9, PRBS15, PRBS23 in addition to PRBS31.

5.5.1 Eye Diagram analysis

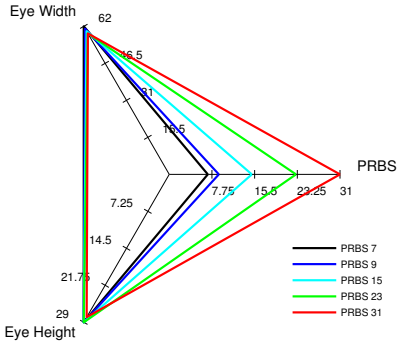
At the system startup, the transceiver parameters in TTK are set at the default values. Changes in eye diagram are compared for different PRBS stressed patterns as the first set of analysis. Eye Height and Width is plotted on a three axes plot with PRBS pattern on the third axes as shown in Figure 5.8. It is found that PRBS31 has the most stressed eye metrics and as anticipated a more closed eye is examined for all the three links speed.

5.5.2 BER Results

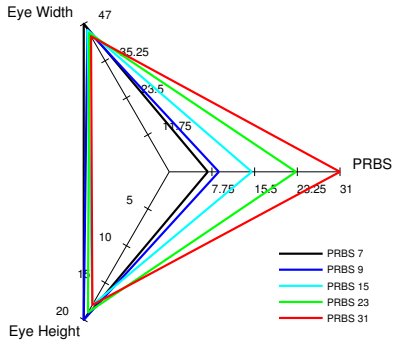
Another important metric of signal integrity is BER. Its measurement is a statistical phenomenon and the estimate is ideal only if the number of tested bits tends to infinity, which is not possible in a real lab test setup. Hence, a method was proposed in reference [107]



10 Gbps



GBT 4.8 Gbps



TTC-PON 9.6 Gbps

Figure 5.8: Changes in the Eye height and Eye width with PRBS variation for optical links at three line rates.

to limit the stressing time of a system to a feasible length and to measure the BER with high CL too. CL is used to quantify the quality of the estimate in percentage. It is the system's actual probability of error less than the specified limit. The minimum number of bits required to be tested for the BER measurement with a specific associated CL is given

in equation 5.1:

$$\left.\begin{aligned} n &= -\frac{\ln(1-CL)}{BER} + \frac{\ln\left(\sum_{k=0}^{N}\frac{(n*BER)^k}{k!}\right)}{BER} \\ T &= n/R \end{aligned}\right\} \tag{5.1}$$

T is test time needed, R is the line rate and when $N = 0$ the solution is trivial given in equation 5.2.

$$n = -\frac{\ln(1-CL)}{BER} \tag{5.2}$$

where n are the total number of bits transmitted and N are the number of errors that occurred during the transmission. There is a compromise between testing time and the required accuracy of the measurement as shown in equation 5.1.

For the 95 percent confidence level (CL), equation 5.2 reduces to n $\simeq 3/(BER)$. Hence to achieve the BER of 10^{-12} at 95 percent CL, total 3x10^{12} bits need to be tested, as a thumb rule.

BER analysis for various link speeds

The concept is further extended to find the minimum inspection time required to measure BER of 10^{-12} at a CL of 95 percent with no errors for GBT, TTC-PON and 10 Gbps links for different CL shown in Figure 5.9. In this paper, all the BER measurements are done for 3x10^{12} bits to achieve 95 percent CL. Variation of BER at Intel-default transceiver set is recorded with respect to the attenuation of the received optical power; following the methodology flowchart shown in Figure 5.7. This test is executed with the help of VOA attached to the loopback fibre. BER variation is recorded for different PRBS patterns and plotted for the links operating at 10 Gbps, 4.8 Gbps and 9.6 Gbps rates as shown in Figure 5.10.

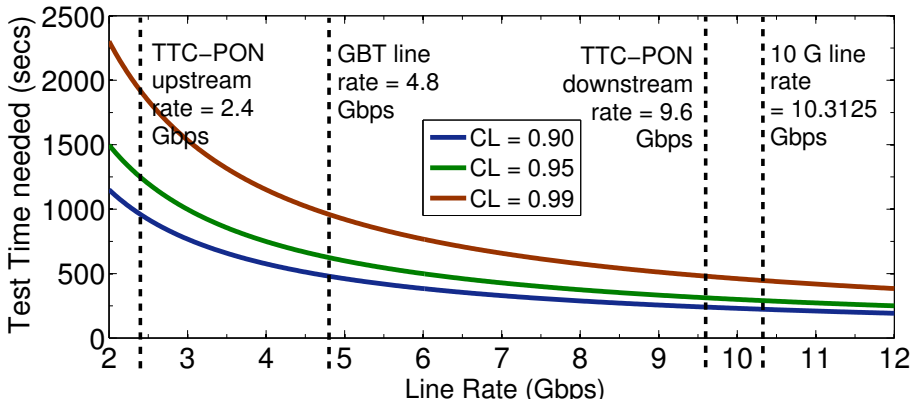


Figure 5.9: Time to achieve BER of 10^{-12} for the Line rate of GBT, TTC-PON and 10 Gbps optical links having different CL.

The exponential curve fitting is the best-suited approximation for the BER in logarithmic domain [108]. Double exponent fit function with constants is used to fit the BER data as it provides close fits in a variety of BER plot situations. It fits the BER data using unconstrained nonlinear optimization [109]. The statistics for goodness-of-fit in terms of R-Square (R^2) for different PRBS is marked in the Figure 5.10.

The test shown in Figure 5.10 highlights that at a specified CL higher number of errors are received in the transmission system for a given received optical power; when PRBS31 is injected as the test data pattern as compared to the other PRBS patterns. The outcome of the tests shown in Figure 5.8 and Figure 5.10 revealed the degradation of the metrics of signal integrity with the increase in the size of a unique word of data in the PRBS sequence. The results from these tests are as anticipated and well substantiated. It has further strengthened the usefulness of the PRBS31 as a strenuous test pattern to demonstrate the validation of the proposed methodology. However, there is a crossover point for 4.8 Gbps at BER~10^{-10}. It is kept beyond the discussion as our region of interest is better by two orders of magnitude which is BER~10^{-12}.

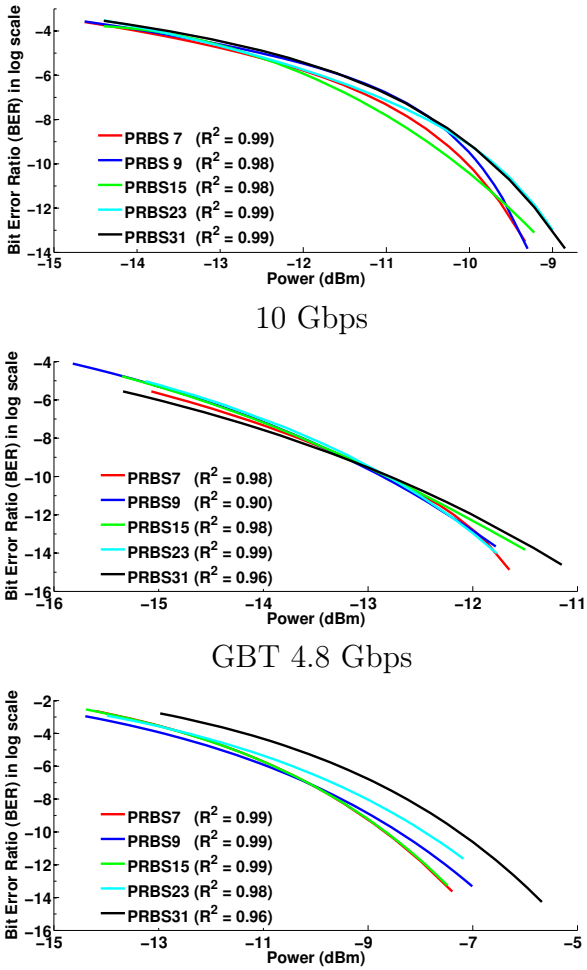
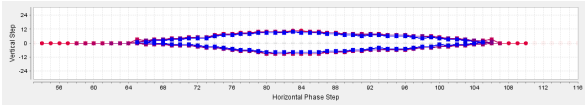


Figure 5.10: BER versus received optical power(dBm) for transceiver at Intel FPGA default settings for different PRBS operating in three line rates.

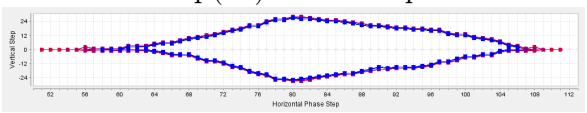
5.5.3 Improvement in Transmission

The improvement in the system performance is marked by two metrics of signal integrity viz. BER and Eye Diagram. The eye contour for the Intel-default settings and at the deduced optimized settings of the transceiver is captured using the EyeQ (a GUI feature of TTK). It helps to estimate and visualize the vertical and horizontal eye opening at the receiver as shown in Figure 5.11. After the application of the deduced transceiver parameter's settings using the proposed technique, there is a notable enhancement in width (Horizontal Phase Step) and height (Vertical Step) of the eye diagram. Hence the quality of signal transmission

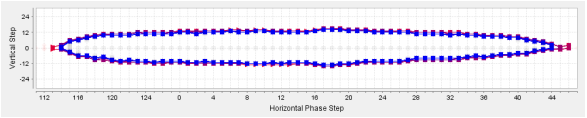
is improved.



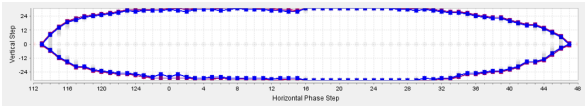
Vertical step(19)/Horizontal Phase step(41) for 10 Gbps at the Intel FPGA default settings



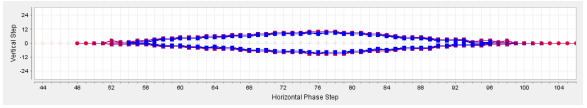
Vertical step(49)/Horizontal Phase step(54) for 10 Gbps at the Optimized FPGA settings



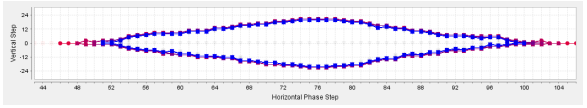
Vertical step(28)/Horizontal Phase step(59) for 4.8 Gbps at the Intel FPGA default settings



Vertical step(63)/Horizontal Phase step(63) for 4.8 Gbps at the Optimized FPGA settings



Vertical step(18)/Horizontal Phase step(43) for 9.6 Gbps at the Intel FPGA default settings

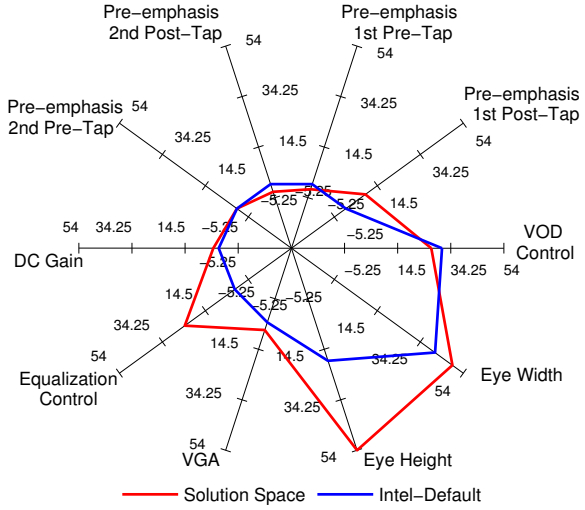


Vertical step(41)/Horizontal Phase step(50) for 9.6 Gbps at the Optimized FPGA settings

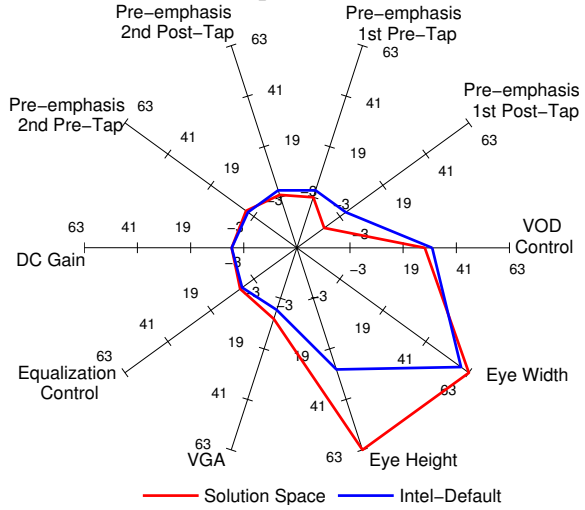
Figure 5.11: Eye diagram at the Intel FPGA default and at the Optimized settings of transceiver.

The optimized values of the transceiver parameters known as solution space, found from the proposed methodology for the targeted BER of 10^{-12} are plotted against the Intel-default set in the form of a multivariate kiviat diagram for all the three link speeds as given in Figure 5.12. It allows us to demonstrate a clear comparison of the individual parameters on each axis.

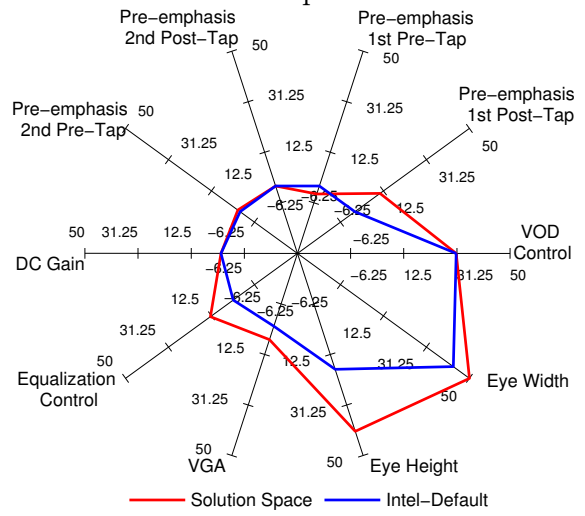
Variation in BER is plotted for the deduced solution space values of a transceiver and for the Intel default set; concerning the different attenuation levels of input optical power at the receiver. It is shown for PRBS31 for all the three links under observation in Figure 5.13.



10 Gbps line rate

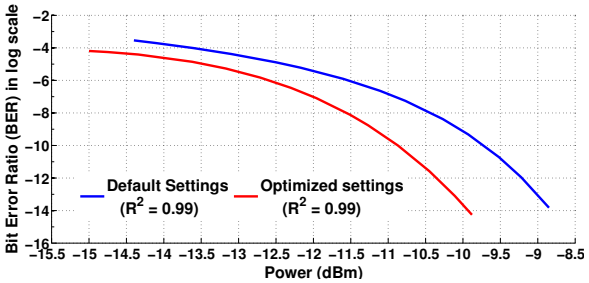


GBT 4.8 Gbps line rate

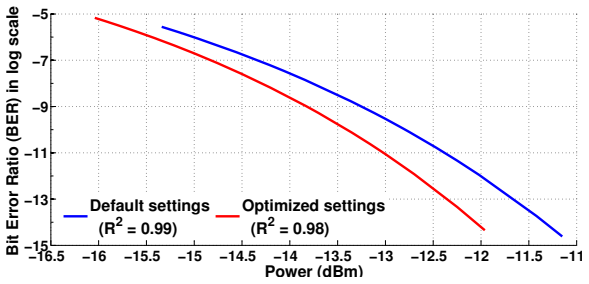


TTC-PON 9.6 Gbps line rate

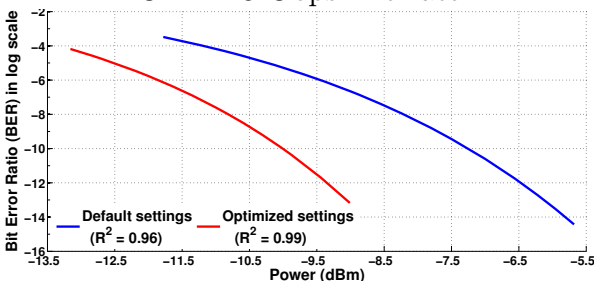
Figure 5.12: Multivariate kiviat diagram showing the solution space and the Intel FPGA default values for three different link rates.



10 Gbps line rate



GBT 4.8 Gbps line rate



TTC-PON 9.6 Gbps line rate

Figure 5.13: Comparison of BER versus the received optical power for default and optimized transceiver settings separately for three line rates.

Further analysing the results from Figure 5.13, the least optical power required at the receiver to attain a preferred BER or better could be determined from the curve. Also it shows, that a specific marked BER is achieved at a lower optical power when transceiver is operated at the deduced parameter values listed in solution space in comparison to the Intel default set. Here to mention the particular case as an example, the targeted BER of 10^{-12} for the optical link test as per IEEE standards is achieved at lower values of the optical power and the improvement at the mentioned BER is quantitatively listed in Table 5.3 for the three link speeds.

Table 5.3: Comparison of Optical power(dBm) to attain BER of 10^{-12} for the three high speed interface links.

Protocol	With default approach (dBm)	With optimization technique (dBm)	Difference (dBm)	Improvement (Percentage)
10Gb Ethernet	-9.2	-10.35	-1.15	12.5
GBT	-11.9	-12.7	-0.8	6.7
TTC-PON	-6.45	-9.3	-2.85	44.1

Another clear observation emerged from the data comparison of Figure 5.13 is that the receiver sensitivity below which the loss of lock occurs, is enhanced due to the reduction in the high-frequency losses with the application of the proposed optimization technique. This results in reducing the limit of the optical power required for the proper CDR and the signal is traceable for comparatively lower values of the received optical power. The quantitative comparisons are given in Table 5.4.

Table 5.4: Comparison of optical power for CDR for the three high speed interface links.

Protocol	With default parameters (dBm)	With optimization technique (dBm)	Difference (dBm)	Improvement (Percentage)
10Gb Ethernet	-14.4	-15	-0.6	4.17
GBT	-15.34	-16.04	-0.7	4.56
TTC-PON	-11.78	-13.2	-1.42	12.05

The test results shown in Figure 5.12 and 5.13 confirms that the effect of high-frequency losses on the link performance is controlled. It is achieved after the application of the deduced solution space values to the TTK and a significant improvement on the BER is noted at a particular received optical power. The tests and results validate the usefulness of the proposed technique to enhance the transceiver performance and the signal integrity by compensating for the high-frequency losses.

5.6 Summary

We have presented a novel transceiver optimization technique to reduce the high-frequency losses which occur due to the increased rates of data transmission in case of HEP exper-

iments. The technique has been implemented on the latest 20nm Intel-Altera Arria-10 FPGA. The scheme has been tested and validated for the link rates of three high-speed communication protocols, GBT, TTC-PON and 10 Gbps Ethernet, which are most commonly used for interfacing the detector front-end electronics, trigger and DAQ systems. The proposed scheme is an optimized approach which reduces number of iterations required.

The tests are performed with PRBS31 pattern at a confidence level of 95 percent. There is considerable gain in the system performance with the application of the proposed technique as specified by the two parameters of signal integrity, the BER and the Eye Diagram. The Intel FPGA set parameters and the solution space values are marked on the kiviat diagram for the fast comparison between the parameters. The results point that to attain the marked BER of 10^{-12}; the required optical power is reduced by 12.5%, 6.7% and 44.1% for 10Gbps, GBT and TTC-PON respectively. The BER is also improved over the received range of optical power. The CDR capability of the system is also enhanced as the least optical power required to recover the data traffic is reduced by 4.17%, 4.56% and 12.05% for 10Gbps, GBT and TTC-PON respectively. The technique improves the signal integrity and reduces the BER. This technique is a heuristic solution and has potential for practical applications as it provides rapid convergence of the solution space to achieve optimized transceiver settings. It makes the implementation of the new technique time efficient. This transceiver optimization technique and its implementation approach would lend itself well for other FPGAs users that allows on-chip assessment of signal quality like Eye diagram.

Chapter 6

CRU hardware development and tests

6.1 Introduction

The CRU is an integral part of O2 and the detectors upgrade. It is at the core of the ALICE data acquisition system through which the trigger, timing, data and control signals are channelled. It is a complex electronics board engrossed with the latest FPGA technology along with optical inputs and outputs and with mezzanine cards for powering. It has more than 1750 components on it and other accessories. The CRU board as a hardware needs to be highly authentic. Its failure during the run time of the experiment is unacceptable as the LHC beam time is a result of several man x hours. To ensure this, extensive tests are carried out for the verification of the developed prototype of CRU hardware. The CRU hardware; PCIe40 DAQ engine is a custom developed PCIe supported board, based on latest Intel-Altera Arria-10 FPGA The PCIe40 readout board is based on PCIe x16 lanes Gen3 standard for interface with the server. It has total 130 number of high-speed signal lanes. The detailed tests of the newly developed and assembled CRU prototypes are required for the predictable and faithful operation of the boards in the experiment. The tests at the fabrication stage as well as the basic electrical performance are performed. The boards are also tested for the functional verification of its communication interfaces and for the

qualification of the prototypes.

In this chapter, the rigorous evaluations performed at both the hardware development stage and the functional qualification tests of the prototype are presented. The details of the step by step performed examinations, verification and their outcomes are explained. The chapter is organised as follows. An overview of the CRU hardware assembly and development is discussed in section 6.2. Functional tests are elaborated in section 6.3. This section includes the basic electrical tests of the board along with the power mezzanine modules, step by step preparation for the interface test, its configuration for the first use and the programmability check is detailed in section 6.3.1, section 6.3.2, section 6.3.3, section 6.3.4, section 6.3.5 respectively. Hardware tests are grouped and detailed in section 6.3.6. The summary of the tests is presented in section 6.4

6.2 Development of CRU board

The PCIe40 card consists of a 14 layer PCB with a large FPGA and its associated logic. The multi-layered PCB of the CRU consists of blind, buried, stacked and the staggered vias. The most challenging process in the fabrication of PCBs is the development of the vertical interconnects or vias from the inner layers to external surfaces or between internal layers. The minimum size of the micro-via required for the PCB of CRU is 0.2 mm. The vias were initially attempted with mechanical drill technology as it could bring down the cost of the PCB fabrication. However, it was not successful and could not bring the desired results. The need for small drill hole diameter limits the use of the mechanical drilling. The via size of 0.2 mm is exactly the limit where the capability of the mechanical drill is restrained. The mechanical drill leads to an open circuit between the two layers; the microsection analysis of the board is shown in Figure 6.2. Laser drill offers greater resolution over the mechanical technique with its ability to produce via sizes well below 50um. Hence, the technique of laser drill [110] is used to fabricate micro-vias on the PCB used for PCIe40 DAQ engine.

The PCB manufactured as per the design files is shown in Figure 6.2. This 14 layer PCB

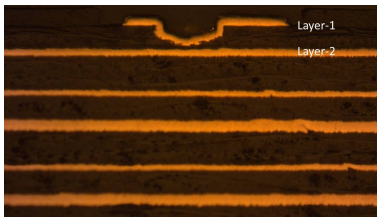


Figure 6.1: Showing an open circuit between layer 1 and layer 2.

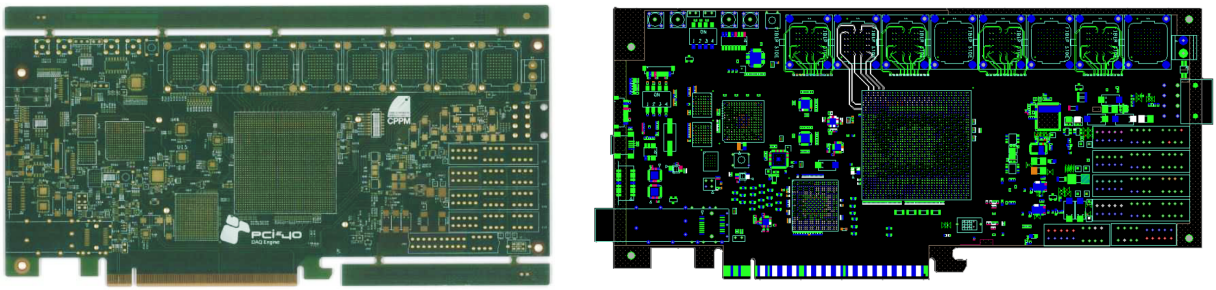


Figure 6.2: Bare board (Left) and its X Ray image (Right).

has the thickness restrictions of 1.57mm (with 10 percent tolerance). The developed PCB is 1.64 mm thick and meets the requirement. Its thickness is an important specification to be maintained as the board needs to be installed in the PCIe slot of the server. The on-board components are assembled using the vapour phase soldering technique also known as condensation soldering [111]. The heat for component soldering technique is provided using the latent heat of liquid vaporization in this technique. It is more precise with higher yield, but is costlier than reflow soldering. FPGA mounting is an important task. Arria-10 FPGA is 45mm x 45mm size having 1932 pins with 0.8mm ball-pitch ultra-fine ball grid array (BGA) package. The FPGA mounting needs special care; the air should not be trapped in the BGA solder balls as when heated it expands and leads to shorts with the nearby pins. Figure 6.3 shows the FPGA image after balling and a 2D X-Ray image of the FPGA after mounting it on the PCB. X-Ray scan is done at the fab house after the mounting of FPGA. It is helpful in finding any possible shorts in the solder ball points beneath the BGA package.

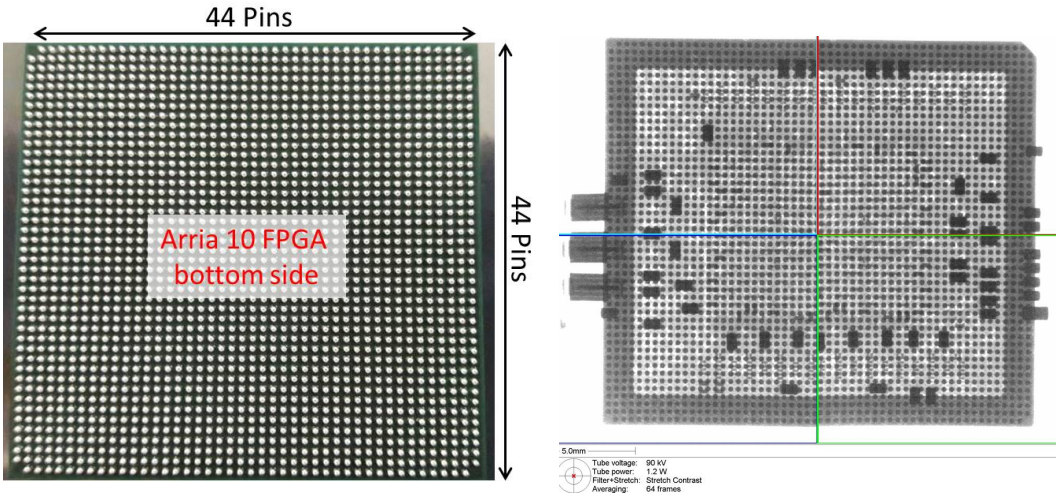
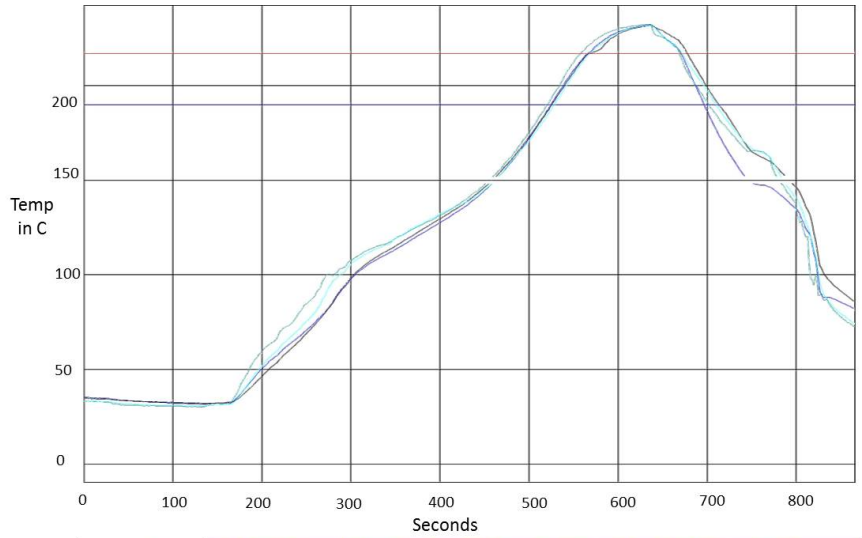


Figure 6.3: (Left) FPGA after solder balling (1932 pins) and (Right) 2D X-Ray image of the BGA package after mounting on PCB.

Temperature profiling and the *Bed of nails connectivity tests (In-circuit test)* are also performed at the assembly house. Temperature profiling [112] is performed using the *Ramp-Spike* method [113] as shown in Figure 6.4. The test is to quickly find the faults and identify the problems before they affect the production quality. Bed of nails connectivity test is performed to check for shorts, opens, resistance, capacitance, and other basic quantities which are used to validate the fabrication.

There are various components on the CRU board apart from the passives like resistors, capacitors and switches. There are clock generators ICs for GBT clocks and the other clocks required for the data transmission. There are other elements like CPLDs to store the program and to configure the transmission chain during the power ON, flash memories etc. The optical transmitter and the receiver modules are important constituent of the board. These transceivers are from *Broadcom inc.* and known as MiniPOD [87, 114]; consists of twelve no.s of 10G pluggable Transceiver Module with 300m range. It is the industry's one of the smallest physical footprint solutions for high density 850nm parallel optical fiber with multi-mode connections. Each optical interface of miniPOD uses flexible ribbon cable as shown in Figure 6.5; each with twelve fibre counts (1No.x12) and PRIZM Light-turn optical miniature detachable connectors at one end and multi-fibre push (MTP) connectors at the



PWI= 91%	Max Rising Slope		Max Falling Slope		Soak Time 150-190°C		Reflow Time /220°C		Peak Temp	
U85	0.72	-64%	-1.93	54%	62.84	-91%	104.38	48%	232.24	-70%
U1	0.70	-65%	-1.80	60%	66.89	-77%	110.28	68%	231.73	-77%
U80	0.75	-63%	-2.17	41%	65.56	-81%	111.42	71%	232.52	-66%
U104	0.71	-65%	-1.97	52%	66.31	-79%	103.63	45%	231.90	-75%
Delta	0.05		0.37		4.05		7.79		0.79	

Statistic Name	Low Limit	High Limit	Units
Max Rising Slope (Target=2.0) (Calculate Slope over 50 Seconds)	0.0	3.0	Degrees/Second
Max Falling Slope (Calculate Slope over 20 Seconds)	-5.0	-1.0	Degrees/Second
Soak Time 150-190°C	60	120	Seconds
Time Above Reflow - 217°C	60	120	Seconds
Peak Temperature	230	245	Degree Celsius

Figure 6.4: Thermal profile characteristics using the Oven VP800-64.

other end for fibre connection.

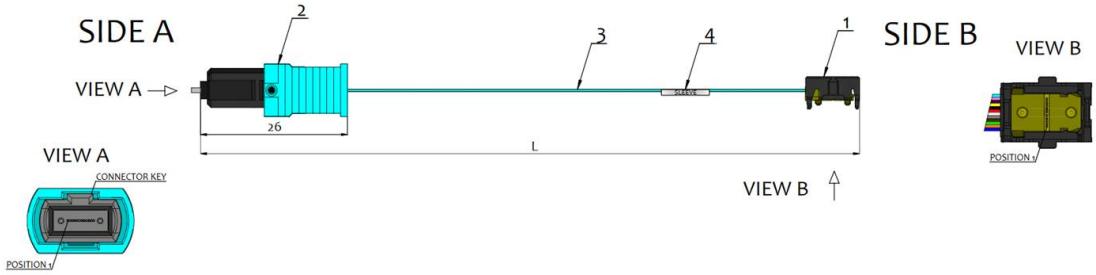


Figure 6.5: Shows the bare ribbon cable assembly drawing.

Figure 6.6 shows the myopic view of the first CRU prototype card with different components mounted on it. In this Figure all the miniPODs and flexible ribbon cables are not installed.

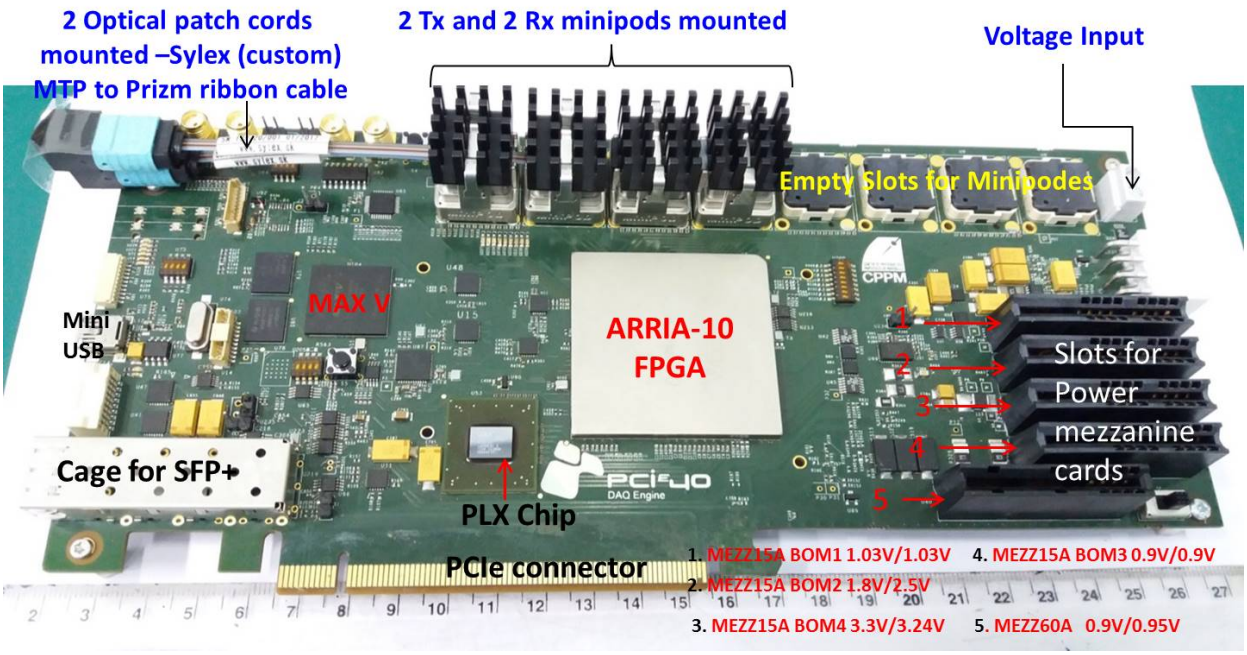


Figure 6.6: Shows the myopic view of the First prototype of the CRU card.

6.3 Functional tests

Once the cards are fully fabricated and assembled; the step by step procedure for functional verification of PCIe40 cards is carried out before the final installation in the experiment.

6.3.1 Basic Electrical Tests

After the reception of the developed prototype and assembled PCB boards; a careful visual inspection of the main card, daughter-cards (aka Power mezzanine cards) and the components is carried out. Reference design schematics are used for this purpose.

Checking for Shorts on the Power Rail

The resistance measurements on the power rails are carried out before powering on the PCIe40 card as shown in Figure 6.7. The aim of the measurements is to exclude the possible assembly failures resulting short(s) on the power rails of the PCIe40 card.

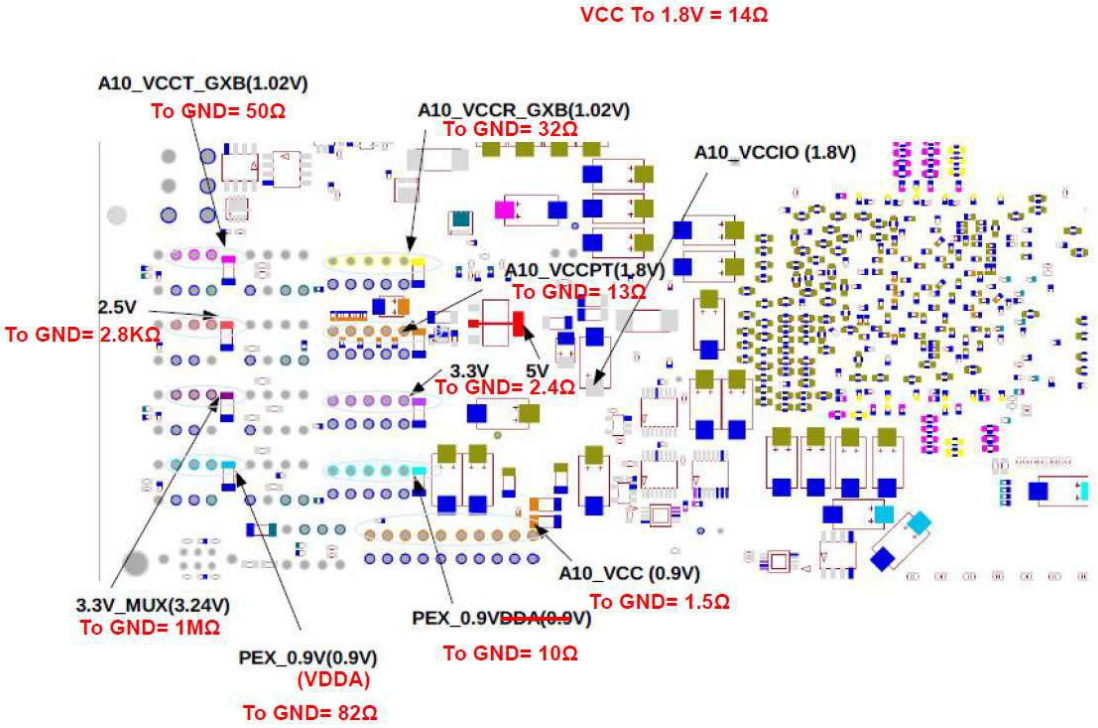


Figure 6.7: Map of test points for measurements of power rail shorts.

Switching the PCIe40 Card ON without the Power Mezzanines

The supply voltage of +12V is applied to the card without the power mezzanine cards mounted on it. The initial output current limit is set to 100 mA (to avoid any huge currents drawn resulting in damage). Without the mezzanine cards only few mA current is drawn. It is to take note of the supply current, and to pay attention to any unusual current drawn values if any. Also the presence of the voltages; +3.3V supply of the power-up sequencer (3.3V_SEQ) and +3.3V supply of the modular multilevel converter (MMC) (3.3V_MMC) on the board is measured as shown in Figure 6.8.

The next step is to mount the 5 power mezzanine daughter cards that provide supply voltages for the functional components of the PCB. Before mounting the daughter cards, the individual test of each mezzanine card is required, so that wrong voltages are not fed to the main board.

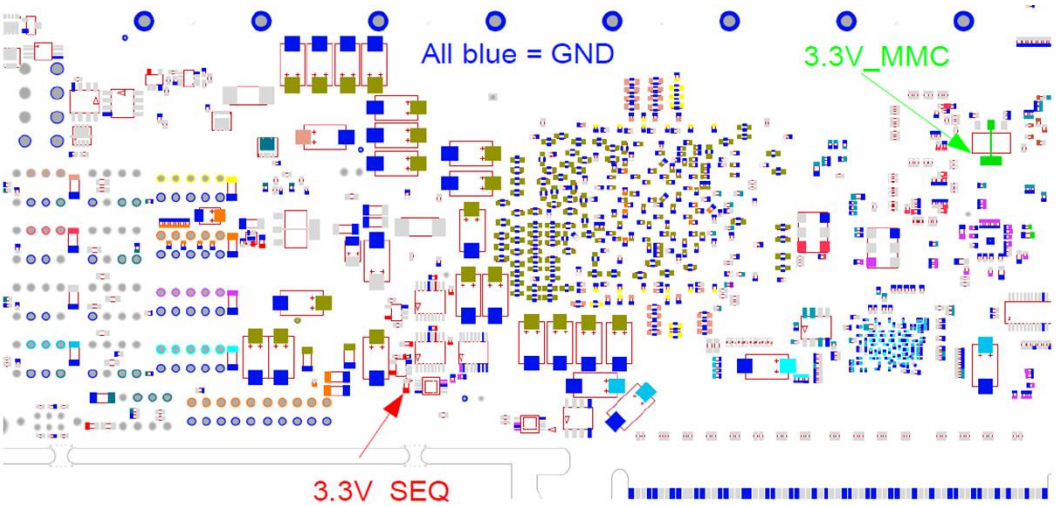


Figure 6.8: Measurement points for 3.3V_SEQ and 3.3V_MMC.

Testing of the Power Mezzanine modules

There are total five mezzanine modules to be mounted on the main PCIe40 board to provide the necessary power supply voltages. The five power mezzanine modules are of two different types; MEZZ15A_6A and MEZZ60A. The power mezzanine module shown in Figure 6.9 is of type MEZZ15A_6A. Module-1, module-2, module-3 and module-4 are same and uses

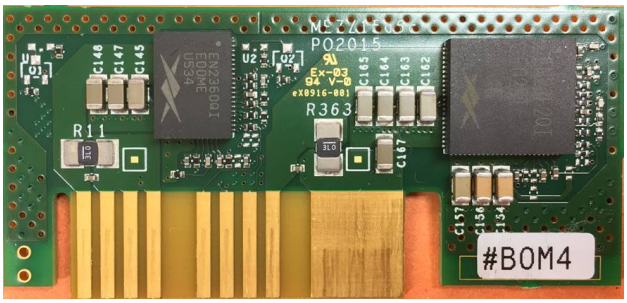


Figure 6.9: MEZZ15A_6A power mezzanine module. Identical schematic for Module-1, 2, 3 and 4.

the same schematic. However, each of them has different voltage settings. The different voltages are tuned by the change of the resistor values on the different modules. Each of the MEZZ15A_6A module is a double power supply unit with a 15A regulator (*Empirion EN23F2QI of ALTERA*) and a 6A regulator (*Empirion EN2360QI of ALTERA*). The Empirion regulators used on these double power modules are sensitive to the Electro-

static Discharge (ESD) The power mezzanine modules, especially the MEZZ15A_6A power modules, are sensitive to electrostatic discharge (ESD). Above the general requirements of handling the electronics items, a special care was taken when touching, inspecting, or installing the MEZZ15A_6A power modules (all module-1 to module-4 variants). A maximum level of ESD precautions was applied during the handling and operation of these modules. Hot-plugging of the power modules is strictly forbidden, on the main PCIe40 card and also on the test cards.

The power mezzanine module-5 shown in Figure 6.10 is of MEZZ60A type and has a different schematic. MEZZ60A is a 60A power supply unit and one no. of the MEZZ60A

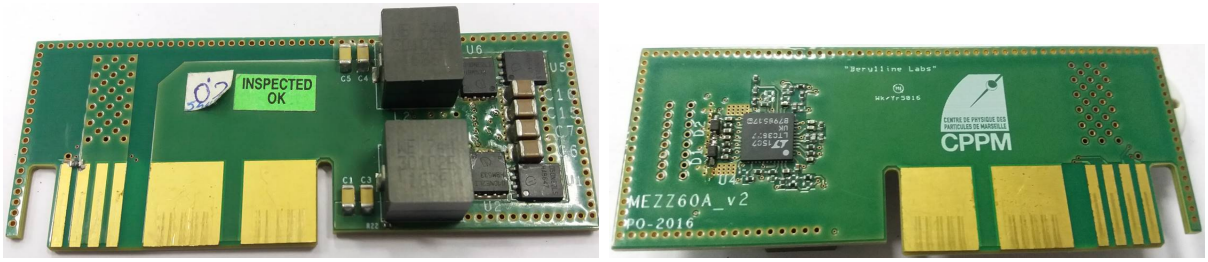


Figure 6.10: MEZZ60A power mezzanine module-5, top side (left) and bottom side (right).

is required for each PCIe40. It is designed with a LTC3877 controller circuit and power FETs. MEZZ15A_6A and MEZZ60A cards are produced using improved FR-4 raw material (with high glass transition temperature and low coefficient of thermal expansion). Power mezzanine cards are 6 layers board with 0.4 mm core (35/35 Cu), thin 1080 prepregs, standard (1.6 mm) thickness with 35 um initial and 70 um final copper thickness on the outer layers. It has edge connectors with electrical gold plated and edge milled for a specified for a given cross section.

Each mezzanine module gives two output voltages as per the voltage requirement of the main board. The voltage values for each module originally depend on the supply voltages required for the FPGA, its transceivers and different on-board components. Table 6.1 summarizes the different voltage levels provided by mezzanine cards.

All the power mezzanine cards are tested before their installation on the PCIe40 card

Table 6.1: Voltage levels from the power mezzanine modules.

Power Mezzanine Module	Voltage levels (Volts)
Module-1	1.03/1.03
Module-2	1.8/2.5
Module-3	0.9/0.9
Module-4	3.3/3.24
Module-5	0.9/0.95

using the dedicated power mezzanine test adapters. Test adapter for the MEZZ15A_6A is shown in Figure 6.11 along with its schematic guide. Also the test adapter for MEZZ60A cards along with its schematic guide is shown in the Figure 6.12.

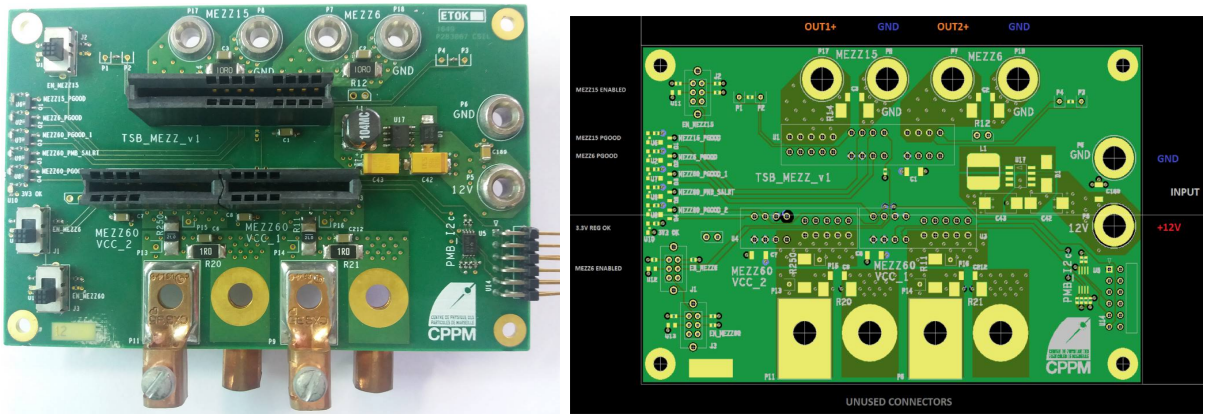


Figure 6.11: (Left) Test adapter for Module-1 to Module-4, (Right) Schematic guide for the test adapter.

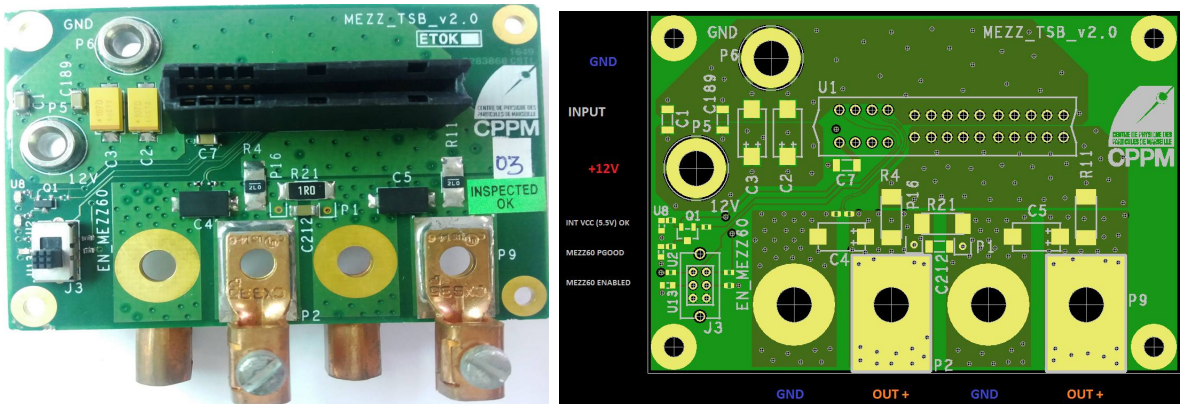


Figure 6.12: (Left) Test adapter for Module-5, (Right) Schematic guide for the test adapter.

The tests are done to check the output voltages of the power mezzanines modules at small

load before installation. It is important to test the modules with a nominal load because the modules should start up even without a load. However, the output voltage on an unloaded module does not necessarily mean that its regulator is functional. The power-on transients and small stand-by currents can load the output capacitors to the nominal voltage, even if the controller of the regulator does not start up, but as soon as a load is applied, the voltage disappears and the module may turn out being not functioning. This is why a real load is needed to be applied, even if a very moderate one. The load values are calculated for the different output voltages, so that at least 1A load (pin through-hole type) with appropriate power rating is applied to the outputs of the different power mezzanines. The simple test procedure is to power ON the given Mezzanine board while in the specified test adapter and checks the output voltage at no load. It should be exactly the nominal voltage. Apply the load and again check the output voltage and ripple (in the oscilloscope) after running a few minutes. After the successful tests the functional power mezzanine cards are mounted on the PCIe40 main card.

6.3.2 Assembly of the power mezzanine cards

The assembly of the mezzanine cards must follow the placement scheme carefully as shown in Figure 6.13 to avoid any damages due to wrong voltages.

Power-On, Sequencing and verifying the Power Supply Voltages on PCIe40 board

There is an on-board power-on voltage sequencer subsystem on the PCIe40 card. The following block diagram in Figure 6.14 illustrates the various parts of the logic state machine of the voltage sequencer circuitry. The supply current increases in four consecutive steps noticed in the lab power supply. All the voltages will be turned OFF in case a fault is detected by the sequencer.

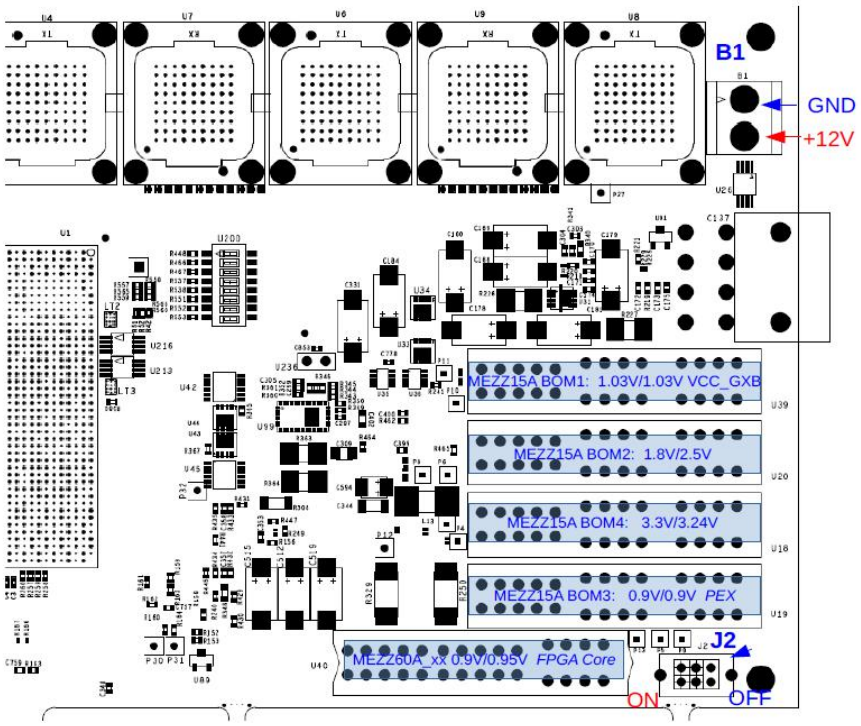


Figure 6.13: Mounting of the Power Mezzanine Cards (top side of PCIe40).

Verifying power supply voltages

The following internal supply voltages as listed in Table 6.2 are checked at the dedicated test points shown in Figure 6.7.

6.3.3 Preparation for the Interface tests

Before running the functional tests for the different interfaces of the PCIe40 card, the following operations are required to prepare the card for the functional tests:

Mounting of the Face Plate

The brackets (a.k.a. face plate, front panel) are mounted to the main board. The faceplate provides the mechanical support to the MTP/MTP connectors, JTAG, USB and the SFP+ connectors. These faceplate as shown in Figure 6.15 are machined using the special wire cut machines to meet the dimensions of the slots to be cut in.

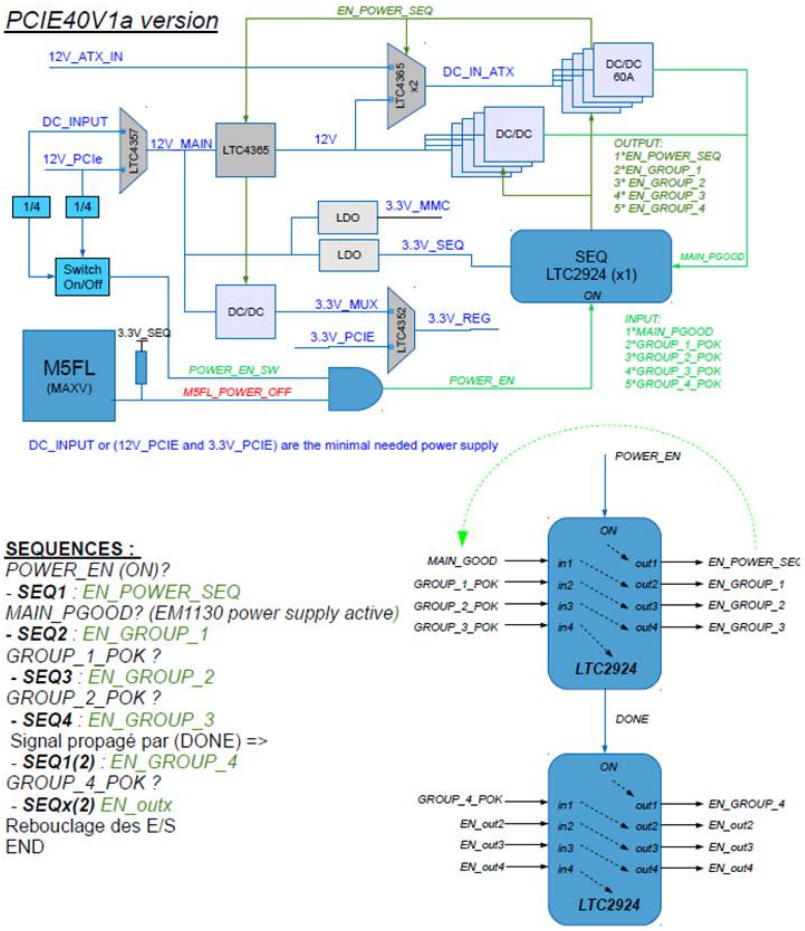


Figure 6.14: Overview of the Power Sequencer.

Table 6.2: Voltage values measured

Net name	Voltage levels (Volts)	Measured Levels (Volts)
A10_VCC	0.9	0.907
A10_VCCR_GXB	1.02	1.024
A10_VCCT_GXB	1.02	1.023
A10_VCCPT	1.8	1.855
A10_VCCIO	1.8	1.8
PEX_0.9V	0.9	0.922
PEX_0.9VDDA	0.9	0.92
2.5V	2.5	2.59
3.3V	3.3	3.35
3.3V_MUX	3.24	3.24
5V	5	5
3.3V_MMC	3.3	3.31
3.3V_SEQ	3.3	3.31

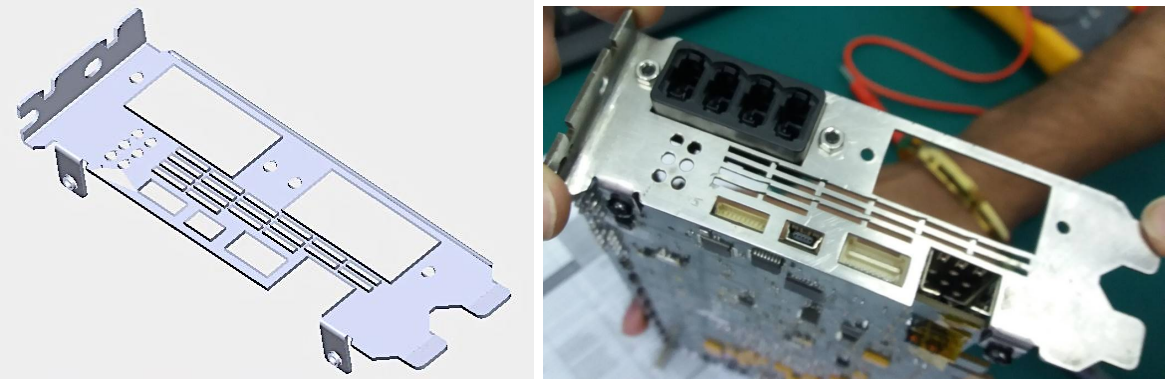


Figure 6.15: (Left) CAD drawing of the Face-Plate developed using Wire Grid machine, (Right) Faceplate installed on PCIe40.

Mounting of the heat sink over the Arria-10 FPGA

The aluminium heat sink is mounted over the Arria-10 FPGA using a set of screws. The heatsink was designed in such a way that it covers both the FPGA and the PLX-8747 chip (*PCIe Gen 3 switch device*). The extra columns (in red) as shown in the Figure 6.16 (on right) were added to avoid the mechanical constraints observed on the PCB, during the mounting of the heatsink.

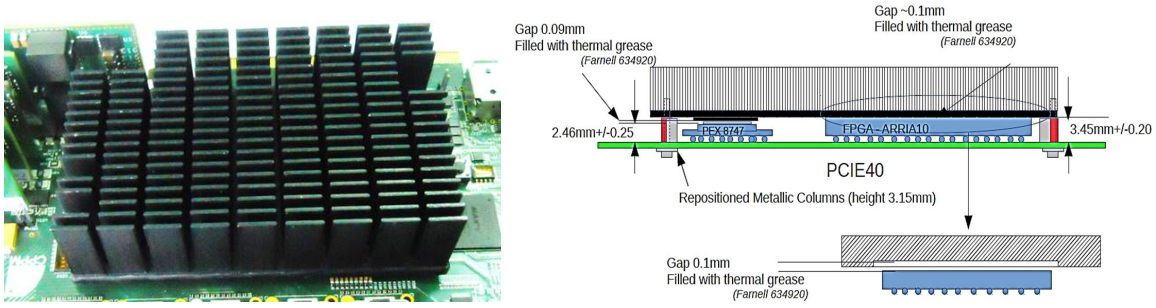


Figure 6.16: (Left) Heat Sink for Arria-10 installed on the board, (Right) Side View showing the arrangement.

Inserting an SFP+ optical transceiver

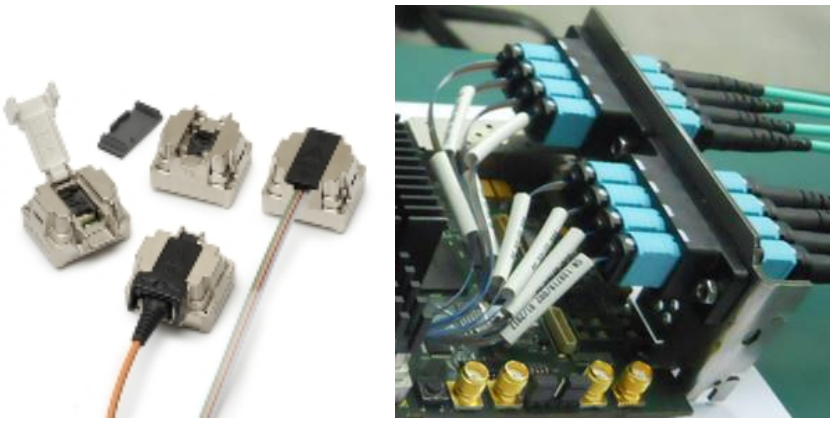
The board also embeds one SFP+ bidirectional device devoted to trigger and timing interface. It is also possible to insert a PON device to broadcast the clock to several Front-Ends.

Mounting of the Parallel Optical Transceivers (Minipods)

The board has 48 bidirectional optical links at 10.3125 Gbps each for a total bandwidth of 0.48 Tbits in each direction. This interface is implemented with 4 optical transmitters and 4 receivers also called as MiniPODs from Avago inc. as shown in Figure 6.17(a). Each of them handle 12 unidirectional optical links. Each transmitter or receiver is linked to a MTP/MTP connector on the faceplate.

Mounting of the Optical Ribbon Cables

The optical ribbon cables are assembled between the optical transceivers and the faceplate. Total 08 such flexible ribbon cables as shown in Figure 6.5 each with 12 fibres are used as the optical interface of miniPODs in one card. The ribbon cables used are of different lengths depending on the location of the miniPOD. The installed ribbon cables on the board are shown in Figure 6.17(b).



(a) MiniPODS (b) Installed Ribbon Cables

Figure 6.17: MiniPODs and the on-board flexible ribbon cables

The layout of the miniPODs on the card, the routing of the flexible ribbon cables and the mapping of miniPODs to the transceivers on the FPGA fabric is shown in Figure 6.18.

The card is ready for the functional tests and the fully assembled card with all its accessories is shown in Figure 6.19

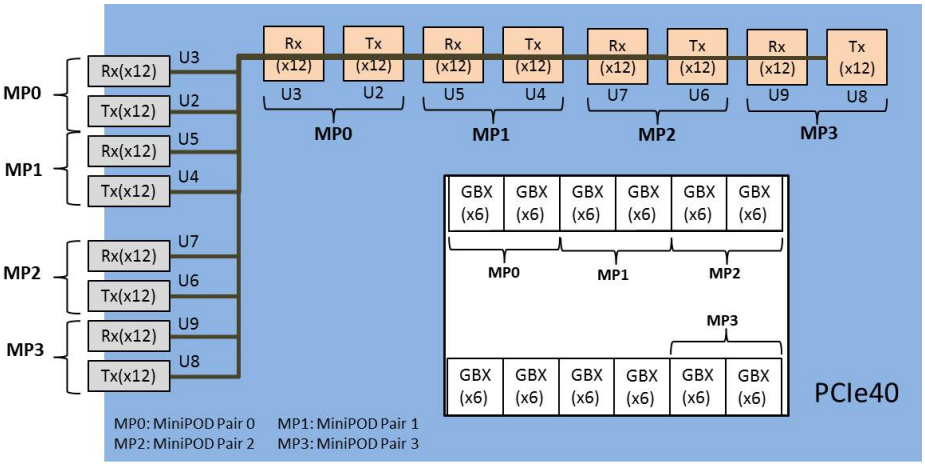


Figure 6.18: MiniPOD mapping.

Figure 6.19: Ready card with all accessories.

6.3.4 Configuration of the Card for the first use

PCIe40 cards require a configuration setting for the desired operation and the proper functional tests.

Setting of Jumpers and Switches

Jumpers and switches are positioned depending on the modes of operation. The location of different on-board jumpers, switches and connectors are referred in Figure 6.20.

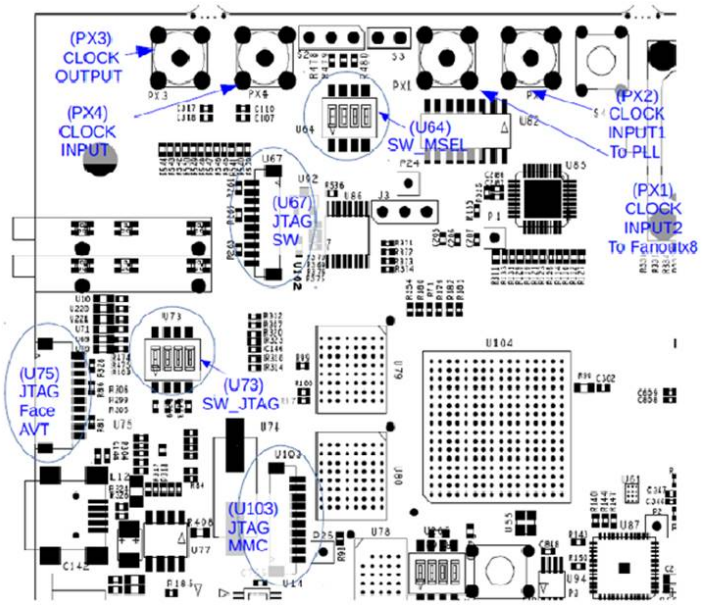


Figure 6.20: Jumpers, Switches and Connectors Locations.

Configuring Programmable Components

The PCIe40 card has two CPLDs (ALTERA MAX V devices) that are referred to as MAX-V JTAG Switch (MAXVSW) and MAX-V Flash (MAXVFL). These programmable logic devices are considered as parts of the hardware. They are permanently programmed during the first testing process. These devices complete the hardware circuit by interconnecting on-board components to each other in predefined ways. These connections are configured by on-board DIP switches.

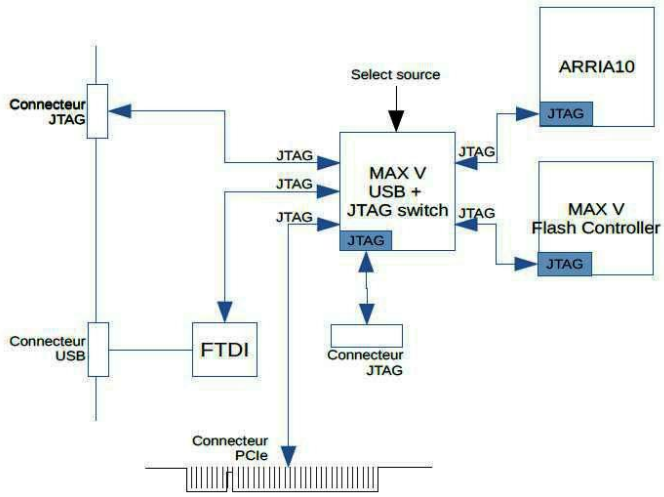


Figure 6.21: PCIe40 JTAG Programming Scheme.

Programming of MAX V JTAG Switch

There a dedicated JTAG interface to program the MAXVSW CPLD only as shown in Figure 6.21. The other devices of the PCIe40 could not be accessed through this port. To program the CPLD, JTAG cable is connected to an external USB Blaster to connector U67 (as labelled in the Figure 6.20), and load the specified configuration file (in the .pof format). After successful programming of the CPLD MAXVSW, the main JTAG chain of the PCIe40 board became operational, and is available through the faceplate JTAG connector.

Programming of the USB interface for the on-board Altera USB Blaster

The PCIe40 card has an on-board USB Blaster which needs programming and installation before it could be used for the first time.

The internal EEPROM of the on-board transceiver chip (*FTD245 USB 2.0*) is programmed with Altera USB Blaster device identification information. After this step, the host computer (running the Quartus programmer) recognizes the on-board USB peripheral as an Altera USB Blaster, and the appropriate Intel-Altera driver is installed.

JTAG Source and Destination Options

With the switch U73 as shown in Figure 6.20, the JTAG source and destinations of the main JTAG chain of PCIe40 are selected from the Table 6.3.

Table 6.3: JTAG switch settings (S = source, T = target)

Source	S1	S2	Target	T1	T2
USB Blaster front panel	OFF	OFF	Arria10 + MAX5_Flash	OFF	OFF
PCIe JTAG	ON	OFF	MAX5_Flash	ON	OFF
JTAG connector front panel	OFF	ON	Arria10	OFF	ON
Not connected	ON	ON	Not connected	ON	ON

Programming of MAX-V Flash

The CPLD Flash is programmed selecting the programming path source and destination with the U73 switch settings from the Table 6.3.

6.3.5 Programming of the Arria 10 FPGA

The most important component of the card is the Arria-10 FPGA. The FPGA device can be programmed in three different ways viz. JTAG, Flash memory or through the PCIe using Configuration via protocol (CvP) as shown in the Figure 6.22

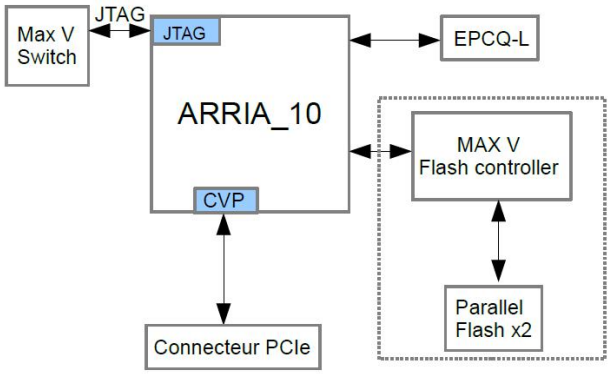


Figure 6.22: Options for Programming of the Arria-10 FPGA.

A successful programming of the FPGA through the JTAG configuration interface with

the test firmware results in blinking of the specified LEDs of the board. This step also verifies the functioning of the basic clock circuits. The advanced option of configuring the Arria-10 FPGA using the on-board Flash Memory enables the permanent storing of the Arria-10 firmware image on board.The CvP allows the FPGA fabric to be updated through the PCIe link without a host reboot or FPGA full chip reinitialization. In this case the complete firmware image is not stored on-board. This programming mode enables the automated programming part of the FPGA firmware from the driver/Software stack through the PCIe interface during the booting phase.

6.3.6 Hardware Tests

The functional tests for the board are categorized into three parts; the CRU hardware tests A, B, and C test the following peripherals of the PCIe40 card: *Hardware Test A* tests the front panel LEDs, on board DIP switches, front panel TFC connector, clock inputs, SMA (SubMiniature version A) clock input/output and the optical links of the SFP+ and the Minipod MP3. *Hardware Test B* performs the additional tests for the optical links of Minipod MP0, MP1, and MP2. *Hardware Test C* checks the DMA transactions, PCI express Register Access, checks for the proper access of the I2C chains and the on-board voltage and temperature sensors. For the *Hardware test A and B* the PCIe40 board can be powered on the desktop or be installed in a server computer. In case of testing of the card on table, a strong cooling airflow has to be maintained. It is strongly recommended, that the PCIe40 card be always operated in a server computer with proper airflow.

The card is designed to have a passive cooling only (there is no on-board fan to maintain the required airflow autonomously), so the cooling of the card relies on the airflow provided by server category computer chassis in which the card is installed. The cards need to be inserted in PCI express slot of the server computer. The produced heatneeds to be dissipated as excessive heating of the card reduces the performance and with the risk of damaging it. To handle the issue of the heat dissipation the ASUS ESC4000-G3 type of server computers are

comprehensively tested and are used for ALICE and LHCb experiments for the purpose of hardware testing and operation of the 'PCIe40' CRU cards. The computer has appropriate power supply and cooling system, and it has two easily removable bays for the PCIe add-in cards. This later feature enables the frequent and easy removal and installation of the PCIe cards which is an important feature when the server is mainly used for add-in cards testing purpose. It is important in the phase of prototypes, small series and mass production testing. The server chassis has an independent air intake for the add-in cards and for the CPU, so the airflow of the add-in cards is not heated by the CPU. The server has a strong airflow 2 m/s. One server can host two CRUs with double wide face-plates by reworking the metal back plates of the removable bays for the add-in cards. The server is declared suitable after it had passed the different tests performed with 'PCIe40' CRU cards like mechanical compatibility, external ATX power supply cables compatibility, temperature measurements of the cooled air for the PCIe add-in cards at 100 percent CPU load, CRU card FPGA junction temperature measurements loaded by a scalable test firmware, data throughput and integrity tests of the PCIe interface, testing of installation and removal of PCIe add-in cards and the possibility of reworking the back-panel of the chassis.

Basic electrical tests and the card Configuration for the first use could be performed by powering the PCIe40 card on the lab bench. These test points do not require cooling airflow. The functional tests require proper cooling of the card. The functional tests are performed while the cards are already installed in the server with cooling air flow. In exceptional cases *hardware test A and hardware test B* can be performed on the lab bench, provided that a strong airflow is ensured and the FPGA junction temperature is frequently monitored. For the *Hardware Test C*; the PCIe40 board must be installed in a server with PCIe x16 Gen3 slot.

The test firmwares are designed to verify the basic functionalities. The hardware test scripts are developed using the "Tool Command Language" (TCL) and python scripting. The test firmware is downloaded on the FPGA via the JTAG connection; and there is

a specified list of commands at the system console TCL shell prompt that invokes CRU hardware test script and is executed to test the board. The tests, procedure and the results are described in the subsequent section.

Hardware Test A

The SRAM object file (.sof); of the test firmware is uploaded to the FPGA. The commands at the TCL shell prompt invoke the required section of the firmware and executed to test the prototype hardware. These are as follows:

JTAG connection test: *enum* command is executed at the TCL shell prompt to enumerate the system console connections

outcome:

% enum

Enumerating System Console connections:

Devices:

1. /devices/10AT115S1@1#USB-0

Masters:

1. /devices/10AT115S1@1#USB-0/(link)/JTAG/(110:132 v1#0)/phy_0/master

Check:

The enum command lists the "phy_0/master" JTAG-to-Avalon module under the Masters section.

To print the generic status report about the test firmware:

Outcome:

% s

2017-03-26 20:59:43 info: Reading CRU Status

2017-03-26 20:59:43 info: Arria10Id: 00540002-18f2010a

2017-03-26 20:59:43 info: FirmwareBuildDate: 2017.03.21 04:45:30

2017-03-26 20:59:44 info: ModFullName(0x0E000000): cru_hw_test_v1_00_b (A 1.0)

2017-03-26 20:59:44 info: gbtAvalonClockFrequency: 100.00 MHz

2017-03-26 20:59:44 info: gbtRefClockFrequency: 240.00 MHz

2017-03-26 20:59:44 info: gbtAtxPllLock: 1

Check:

ModFullName(0x0E000000) should be cru_hw_test_v1_00_b (A 1.0) which means test firmware A v1.0, the Avalon clock frequency should be 100 MHz, the GBT reference clock should be 240 MHz and the gbtAtxPllLock should be 1.

To read the DIP switch positions:

Outcome:

% sw

2017-03-21 10:49:36 info: Reading DIP switch positions

2017-03-21 10:49:37 info: Arria10Id: 00540002-18f2010a

2017-03-21 10:49:37 info: DipSwitchStatus: 0xFF(binary: 1-1-1-1-1-1-1-1)

Check:

DIP Switch Status display the actual DIP switch positions. The DIP switch is changed manually and the command **sw** executed again to check different positions.

To get the frequency counter values

Outcome:

% f

2017-03-21 10:53:21 info: Reading frequencies

2017-03-21 10:53:21 info: A10_CAL_CLK_100MHZ: 100.00 MHz

2017-03-21 10:53:21 info: CLK_A10_100MHZ_P: 100.00 MHz

2017-03-21 10:53:21 info: A10_REFCLK_TFC_P: 240.00 MHz

2017-03-21 10:53:21 info: A10_REFCLK_2_TFC_P: 234.26 MHz

2017-03-21 10:53:21 info: A10_CLK_PCIE_P_0: 99.77 MHz

2017-03-21 10:53:21 info: A10_SI5338_CLK_40_P: 40.08 MHz

2017-03-21 10:53:21 info: A10_SMA_CLK_IN_P: 0.32 MHz

Check:

At least A10_CAL_CLK_100MHZ and CLK_A10_100MHZ_P should be 100 MHz, and A10_REFCLK_TFC_P should be 240 MHz. A10_SMA_CLK_IN_P should display the clock frequency connected to SMA input. The other values may depend on the PLL EEPROM settings

To run the internal loopback tests on the left side transceivers (SFP, MP3)

Outcome:

% lbi

2017-04-04 10:18:36 info: Checking CRU GBT-FPGA block in serial loopback mode

2017-04-04 10:18:36 info: Date: 2017-04-04 10:18:36

2017-04-04 10:18:36 info: AvalonMaster: /devices/10AT115S1@1#USB-0/(link)/JTAG/(110:132 v1#0)/phy_0/master

2017-04-04 10:18:36 info: Arria10Id: 00540002-18f2010a

2017-04-04 10:18:36 info: Enabling internal loopbacks and reseting the error counters ...

2017-04-04 10:18:36 info: loopback: 1, txcontrol: 0xcbbc0000, rxcontrol: 0xcbbc0000

2017-04-04 10:18:36 info: index: 0, bank: 0, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 1, bank: 0, link: 1, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 2, bank: 0, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 3, bank: 0, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 8, bank: 1, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 9, bank: 1, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 10, bank: 1, link: 4, atx_pll_lock: 1, locked_to_data , error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 11, bank: 1, link: 5, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 info: index: 12, bank: 2, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:36 passed: Internal loopback failing links (count: 0): (PASSED)

Check:

The test script reports on each link either OK, or error status and at the end an aggregated summary is reported *passed* when all links are without errors. (link index 0..11 means Mini-

Pod 3, and link index 12 means the PON SFP transceiver).

To run the external optical loopback tests on the left side transceivers (SFP, MP3)

Outcome:

% lbo

2017-04-04 10:18:43 info: Checking CRU GBT-FPGA block in optical loopback mode

2017-04-04 10:18:43 info: Date: 2017-04-04 10:18:43

2017-04-04 10:18:43 info: AvalonMaster: /devices/10AT115S1@1#USB-0/(link)/JTAG/(110:132 v1#0)/phy_0/master

2017-04-04 10:18:43 info: Arria10Id: 00540002-18f2010a

2017-04-04 10:18:43 info: Reseting the error counters ...

2017-04-04 10:18:43 info: loopback: 0, txcontrol: 0xcbbc0000, rxcontrol: 0xcbbc0000

2017-04-04 10:18:43 info: index: 0, bank: 0, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 1, bank: 0, link: 1, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 2, bank: 0, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 3, bank: 0, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 8, bank: 1, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 9, bank: 1, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 10, bank: 1, link: 4, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 11, bank: 1, link: 5, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:18:43 info: index: 12, bank: 2, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 2768 <- error counter

2017-04-04 10:18:43 failed: Optical loopback failing links (count: 1): [12] (FAILED)

Check:

The test script reports on each link either ok or error status and at the end an aggregated summary is reported **passed** if all links are without errors (link index 0..11 means MiniPOD 3, and link index 12 means the PON SFP transceiver).

Hardware Test B

This is an additional loopback test for the optical links through MiniPOD MP0, MP1, and MP2. The firmware for this test is downloaded to the FPGA, the hardware test script checks the JTAG connection and the generic report for the different clock frequencies and firmware version is checked; as given in Hardware Test A also. It is an important step to verify the accessibility of the clocks and the correctness for the version of the firmware downloaded. The links are tested for both the internal loopback at the level of silicon and for the external loopback using the optical fibre. The tests and their outcome is detailed as follows

The internal loopback tests on the right side transceivers (MP0, MP1, MP2)

Outcome:

% lbi

2017-04-04 10:27:00 info: Checking CRU GBT-FPGA block in serial loopback mode

2017-04-04 10:27:00 info: Date: 2017-04-04 10:27:00

2017-04-04 10:27:00 info: Arria10Id: 00540002-18f2010a

2017-04-04 10:27:00 info: Enabling internal loopbacks and reseting the error counters ...

2017-04-04 10:27:00 info: loopback: 1, txcontrol: 0xcbbc0000, rxcontrol: 0xcbbc0000

2017-04-04 10:27:00 info: index: 0, bank: 0, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:00 info: index: 1, bank: 0, link: 1, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:00 info: index: 2, bank: 0, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:00 info: index: 3, bank: 0, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:00 info: index: 4, bank: 0, link: 4, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

. . .

2017-04-04 10:27:01 info: index: 32, bank: 5, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:01 info: index: 33, bank: 5, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:01 info: index: 34, bank: 5, link: 4, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:01 info: index: 35, bank: 5, link: 5, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:01 passed: Internal loopback failing links (count: 0): (PASSED)

Check: The test script will report on each link either OK or error status and at the end will report an aggregated summary passed if all links without errors (link index 0..11 means MiniPOD 0, link index 12..23 means MiniPOD 1, link index 24..35 means MiniPOD 2).

The external optical loopback tests on the right side transceivers (MP0, MP1, MP2)

Outcome:

% lbo

2017-04-04 10:27:03 info: Checking CRU GBT-FPGA block in optical loopback mode

2017-04-04 10:27:04 info: Arria10Id: 00540002-18f2010a

2017-04-04 10:27:04 info: Reseting the error counters ...

2017-04-04 10:27:04 info: loopback: 0, txcontrol: 0xcbbc0000, rxcontrol: 0xcbbc0000

2017-04-04 10:27:04 info: index: 0, bank: 0, link: 0, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:04 info: index: 1, bank: 0, link: 1, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:04 info: index: 2, bank: 0, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:04 info: index: 3, bank: 0, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

. . .

2017-04-04 10:27:05 info: index: 32, bank: 5, link: 2, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:05 info: index: 33, bank: 5, link: 3, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:05 info: index: 34, bank: 5, link: 4, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:05 info: index: 35, bank: 5, link: 5, atx_pll_lock: 1, locked_to_data: 1, error_counter: 0 <- ok

2017-04-04 10:27:05 Passed: Optical loopback failing links (count: 0): (PASSED)

Check: The test script will report on each link OK or error status and at the end will report an aggregated summary **passed** if all links are without errors (link index 0..11 means MiniPOD 0, link index 12..23 means MiniPOD 1, link index 24..35 means MiniPOD 2).

Hardware Test C

Test Steps

In the experimental setup, the CRU hardware will be installed in the PCIe x16 Gen3 slot of server. The motive of the test is to check the detection of the developed prototype cards in the PCIe slot of the server, transfer test data, and evaluate the data rate at which it is transferred. The different tests are run from the Linux command terminal. The HW Test C firmware (cru_hw_test_C_v2.sof) is loaded into the FPGA.

Results of Hardware Test C

The CRU is equipped with an Arria-10 FPGA; it provides PCIe Gen 3 x16 lane connection for data transfer. Each lane provides 8 Gbps to provide the output bandwidth of 128 Gbps. Gen 3 x16 is composed of two individual PCIe endpoints, each with Gen 3 x8 lanes. PLX8747 chip is used on CRU for converting two x8 lanes of PCIe into one x16 PCIe. The generic term is a PCI Express Switch.

The Arria-10 FPGA on the PCIe40 card can only manage PCIe Gen3 interfaces with 8 lanes whereas the CPU needs a PCIe Gen3 with 16 lanes interface. The adaptation is made with a PCIe bridge PLX8747 from PLX. The PX8747 has 48 PCIe Lanes, implemented as 16 lanes per station across three stations. The stations are connected to one another by the internal non blocking fabric. The adaptation 8 lanes/16 lanes is done by connecting two 8 lanes ports of the bridge to the FPGA and a 16 lanes port to the PCIe connector of the card as shown in Figure 6.23.

The detection of the two end points of PCIe; each with Gen3 x8 lanes is an important task for the newly fabricated cards. It verifies the integrity of the hardware path from the PCIe end points of FPGA to the slots on the mother board of the server via the mounted PLX chip on the card.

Executing roc-list-cards

Outcome:

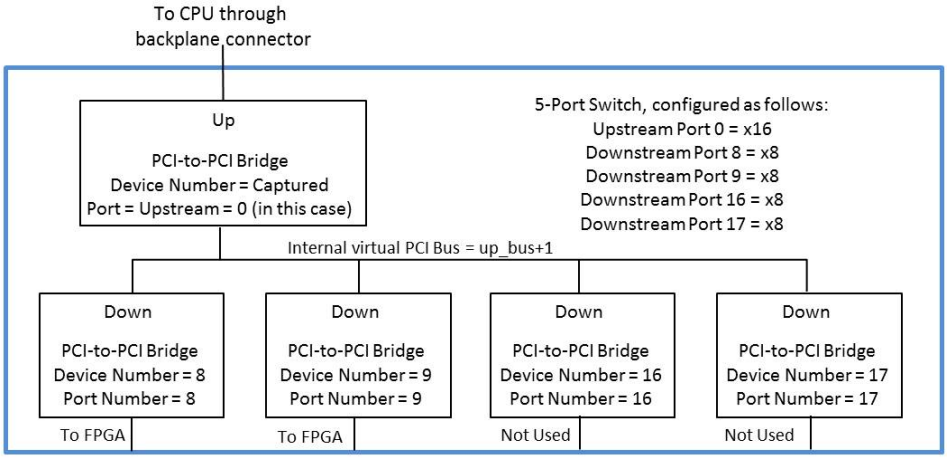


Figure 6.23: Configuration of PLX-8747.

% roc-list-cards

2017-10-24 22:22:23.756714 [pci=05:00.0 serial=0 channel=0] Acquiring DMA channel lock

2017-10-24 22:22:23.757742 [pci=05:00.0 serial=0 channel=0] Acquired DMA channel lock

2017-10-24 22:22:23.793336 [pci=05:00.0 serial=0 channel=0] Initializing memory-mapped DMA buffer

2017-10-24 22:22:23.795828 [pci=05:00.0 serial=0 channel=0] Scatter-gather list size: 1

2017-10-24 22:22:23.798812 [pci=05:00.0 serial=0 channel=0] Buffer is hugepage-backed

2017-10-24 22:22:23.798895 Enabling link(s): 0

2017-10-24 22:22:23.800028 [pci=05:00.0 serial=0 channel=0] Releasing DMA channel lock

2017-10-24 22:22:23.867106 [pci=06:00.0 serial=0 channel=0] Acquiring DMA channel lock

2017-10-24 22:22:23.867271 [pci=06:00.0 serial=0 channel=0] Acquired DMA channel lock

2017-10-24 22:22:23.893809 [pci=06:00.0 serial=0 channel=0] Initializing memory-mapped DMA buffer

2017-10-24 22:22:23.894304 [pci=06:00.0 serial=0 channel=0] Scatter-gather list size: 1

2017-10-24 22:22:23.896241 [pci=06:00.0 serial=0 channel=0] Buffer is hugepage-backed

2017-10-24 22:22:23.896276 Enabling link(s): 0

2017-10-24 22:22:23.897141 [pci=06:00.0 serial=0 channel=0] Releasing DMA channel lock

Table 6.4: PCIe end point detection

#	Type	PCI Addr	Vendor ID	Device ID	Serial	FW version
0	CRU	05:00.0	0x1172	0xe001	0	20171016-180322-6cce25a6
1	CRU	06:00.0	0x1172	0xe001	0	20171016-180322-6cce25a6

The two end points are detected using the script; the addresses are denoted in red as shown in Table 6.5.

PCIe DMA testing

The main task of CRU is to receive the data over GBT links at 4.8 Gbps each, process it depending on the individual detector requirements and transporting it to the memory of the readout server. There are total 48 input links on the CRU hence a huge amount of data needs to be directed to the server memory. The CRU is engrossed with PCIe Gen3 x16 lanes; to provide output bandwidth around 128 Gbps. The DMA is used to move the data over PCIe; from and to CRU independent of the processor [115]. The first motivation behind DMA evaluation is to know about maximum bandwidth that can be utilized out of 128 Gbps.

The script is used to test the CRU card to be recognized as a Gen3 x8 (for each end points) and to check the DMA performance

DMA CRU testing program

source dma_test.sh

Outcome:

- Setting up environment

hugeadm: WARNING: Directory /var/lib/hugetlbfs/global/pagesize-2MB is already mounted.

hugeadm: WARNING: Directory /var/lib/hugetlbfs/global/pagesize-1GB is already mounted.

/root/CRU_DMA_I2C/BASH

Insert ID of the PCIe endpoint as displyed above [0/1] : 1

Table 6.5: PCIe end point detection

#	Type	PCI Addr	Vendor ID	Device ID	Serial	FW version
0	CRU	05:00.0	0x1172	0xe001	0	20171016-180322-6cce25a6
1	CRU	06:00.0	0x1172	0xe001	0	20171016-180322-6cce25a6

Insert how much data you want to collect GB [10] : 10

Check the DATA [Y/n] : Y

Store DATA on the disk [Y/n] : n

CARD PCIE CAPABILITY AND STATUS

PCIe [06:00.0] DMA Running ... GB/s 6.75165

PCIe [05:00.0] DMA Running ... GB/s 6.88151

The program generates a *log file logdma_ FPGAID_ PCIEID e.g. logdma_ 54000218ee0509_ 6:0.0*. Each FPGA has a unique identification code.

Inter Integrated Circuit (I2C) scan testing

There are various chips on CRU hardware like clock generators, temperature sensors, optical transceivers modules etc., that are associated with the proper functioning and monitoring of the CRU. Different on-board parameters need to be configured before the installation of CRU in the experiment; also monitoring is required for certain parameters like on-board temperature. The I2C interface on the chips is used to monitor and configure the parameters from the FPGA. The I2C master is attached to the PCIe endpoints. A shell script checks all the chips detected on different I2C chains. It is an important verification test for the proper mounting of the on-board detected components and for the response to the commands.

Outcome:

source i2c_scan.sh PCIE_ID

Hello I2C SCAN Python script PCIe 6:0.0

BAR 0x500

ADD 0x70 CHIP FOUND [0x80000000]

ADD 0x71 CHIP FOUND [0x80000000]

ADD 0xf0 CHIP FOUND [0x80000000]

ADD 0xf1 CHIP FOUND [0x80000000]

On I2C chain [0x500] found 4 chips

Hello I2C SCAN Python script PCIe 6:0.0

BAR 0x600

ADD 0x18 CHIP FOUND [0x80000000]

ADD 0x29 CHIP FOUND [0x80000000]

ADD 0x4c CHIP FOUND [0x80000000]

ADD 0x4d CHIP FOUND [0x80000000]

ADD 0x4e CHIP FOUND [0x80000000]

ADD 0x4f CHIP FOUND [0x80000000]

ADD 0x77 CHIP FOUND [0x80000000]

ADD 0x98 CHIP FOUND [0x80000000]

ADD 0xa9 CHIP FOUND [0x80000000]

ADD 0xcc CHIP FOUND [0x80000000]

ADD 0xcd CHIP FOUND [0x80000000]

ADD 0xce CHIP FOUND [0x80000000]

ADD 0xcf CHIP FOUND [0x80000000]

ADD 0xf7 CHIP FOUND [0x80000000]

On I2C chain [0x600] found 14 chips

Hello I2C SCAN Python script PCIe 6:0.0

BAR 0x700

ADD 0x0 CHIP FOUND [0x80000000]

ADD 0x38 CHIP FOUND [0x80000000]

ADD 0x80 CHIP FOUND [0x80000000]

ADD 0xb8 CHIP FOUND [0x80000000]

On I2C chain [0x700] found 4 chips

Hello I2C SCAN Python script PCIe 6:0.0

BAR 0x800

On I2C chain [0x800] found 0 chips

Hello I2C SCAN Python script PCIe 6:0.0

BAR 0x900

ADD 0x28 CHIP FOUND [0x80000000]

ADD 0x29 CHIP FOUND [0x80000000]

ADD 0x30 CHIP FOUND [0x80000000]

ADD 0x31 CHIP FOUND [0x80000000]

ADD 0x58 CHIP FOUND [0x80000000]

ADD 0x59 CHIP FOUND [0x80000000]

ADD 0x68 CHIP FOUND [0x80000000]

ADD 0x69 CHIP FOUND [0x80000000]

ADD 0xa8 CHIP FOUND [0x80000000]

ADD 0xa9 CHIP FOUND [0x80000000]

ADD 0xb0 CHIP FOUND [0x80000000]

ADD 0xb1 CHIP FOUND [0x80000000]

ADD 0xd8 CHIP FOUND [0x80000000]

ADD 0xd9 CHIP FOUND [0x80000000]

ADD 0xe8 CHIP FOUND [0x80000000]

ADD 0xe9 CHIP FOUND [0x80000000]

On I2C chain [0x900] found 16 chips

Note: For all the cards the outputs must be the same, if they have equal hardware configuration.

I2C scan for miniPODs

Base address 0x900 scans the minipods, the list of chips varies with the number of installed minipods on the card. The minipods are configured to be transmit or receive. There are two variations in programs to read-out the different register configurations for the two type of minipods viz. transmit and the receive. It displays the operating parameters of transceivers.

To read the minipods

python txmp.py PCIE_ID

Outcome:

Hello I2C MINIPOD Python script PCIe 6:0.0

BAR 0x900

TX MINIPOD ADD= 0x28

BASIC MAKE [AVAGO]

PART [811]

CASETMP [70] degC

SIG RATE [10300-2500] Mb/s

WAVELENGTH [845] nm

TOLERANCE [15] nm

RANGE TEMP [95.0 to -5.0] degC

3.3V [3.565 to 3.035] V

2.5V [2.725 to 2.275] V

BIASi [10000 to 2000] uA

PWR [2000 to 125.9] uW

CURRENT TEMP [44.62890625] degC

HIVOL [3.3202]V

LOVOL [2.5295]V

BIASi [5976 | 6084 | 5892 | 5694 | 5988 | 5882]uA

BIASi [5890 | 5988 | 6070 | 5952 | 5930 | 6106]uA

PWRo [1005.2 | 994.0 | 1001.0 | 938.4 | 967.9 | 932.0]uW

PWRo [1045.6 | 822.3 | 872.6 | 914.7 | 885.2 | 880.1]uW

ETIME [4354]hrs

The miniPODs have several registers that are read-out. All their values are stored in log files: txmp_log.txt: for the transmit minipod and rxmp_log.txt : for the receive minipod. Similary tests are repeated for other transmit minipods and the receiver minipods also.

Temperature measurement

There are several sensors installed on the board to read to read out the temperature from

them also to read the temperature of the FPGA.

The script accesses the chip installed in the CRU to read the external FPGA temperature.

python max1619.py PCIE_ID

MAX1619 for FPGA temperature measurement

Outcome:

Hello I2C MAX1619 Python script PCIe 6:0.0

BAR 0x600

MFGID 0x4d

DEVID 0x4

TEMPERATURE

MAX1619 : 42 C

FPGA(remote) : 61 C

FPGA : 55 C

MONITOR TEMPERATURE OF THE CARD

python ltc2990.py PCIE_ID

Outcome:

Hello I2C LTC2990 Python script PCIe 6:0.0

BAR 0x600

CONFIGURE CHIP FOR TEMP ADD [0x4c]

CONFIGURE CHIP FOR TEMP ADD [0x4e]

CONFIGURE CHIP FOR TEMP ADD [0x4f]

CHIP TEMPERATURE [0x4c]

CHIP INT TEMP [42.125]

T1 : 34.9375

CHIP TEMPERATURE [0x4e]

CHIP INT TEMP [39.875]

T1 : 31.0625

T3 : 41.6875

CHIP TEMPERATURE [0x4f]

CHIP INT TEMP [40.0]

T1 : 36.1875

T3 : 51.875

6.4 Summary

In this chapter, the development of CRU hardware is discussed. The detailed step by step procedure for the testing of hardware is described in the pedagogical manner. The developed prototypes are verified for all the basic electrical tests along with the power mezzanine modules. The card is dressed with all the accessories and prepared for the interface testing. Once the prototype cards are fully ready for the first use; all the programmable components like the jumpers and switches are adjusted as per the proposed configuration. The programmability of on-board FPGA and CPLDs are verified. Various test are carried out for the functional validation and they are grouped as hardware tests A, B and C. The test scripts are developed on tcl and python platforms and executed for the prototype hardware validation. All the optical interfaces are functionally checked with internal and external loopbacks. The data are transferred to the server PC through the PCIe interface of the prototype boards. The speed check for the data transfer is an important test. The practical data transfer speed of the Gen3 x16 is ~6.8 GB/s which equates with the theoretical PCIe performance of ~6.9 GB/s as calculated in section 4.4.1 of chapter 4. The ICs mounted on-board, temperature sensors, miniPODs etc are scanned using the I2C protocol as per the developed tests scripts. The insight view of the prototype CRU hardware and the test procedure as a whole is summarized in this chapter. In the next chapter the thesis is concluded along with an overview for the future outlook.

Chapter 7

Summary and Future Scope

7.1 Summary

The major goals of HEP experiments are to probe the fundamental constituents of the matter and understand the nature of fundamental forces. Advanced research in HEP demands a progressive increase in collision energies and beam luminosities of the particle accelerators, which are essential for accessing rare probes with extremely low cross sections. The experiments are continuously upgraded with sophisticated detectors, electronics and DAQ systems. In this regard, the DAQ architectures have been evolving continuously to cope up with the demands of the experiments. The LHC at CERN will go through a major upgrade during the LS2 period, following which the beam luminosities will increase by about an order of magnitude from their present values. To handle this the experiments at the LHC are upgrading the detector and DAQ systems to allow for faster readout of the online data. In this thesis, the marked challenge for the acquisition of data during the high data rate Run-3 of the upgraded ALICE detector is emphasized. The presently used methodologies and the capabilities of different readout units used in the experiment will not be able to meet the upgraded specifications of the Run-3 for ALICE experiment. Hence to handle the high data rate and capture all the events; a major upgrade of the ALICE experiment

is scheduled to start from the year of 2019. It consists of the upgrade of sub-detectors, their readout electronics and the data acquisition system. The upgrade strategy of data acquisition is based on the CRU. The FPGA based CRU acts as an interface between the on-detector front end electronics, O2 system and the trigger system.

The introductory features of LHC, ALICE experiment and need for the upgrade considering high luminosity run of LHC, the statistical numbers for the crucial parameters before and after the scheduled upgrade are detailed in **Chapter 1**. The present ALICE detector and an overview for the upgrade of ALICE sub-detectors, their readout electronics and the updated approach for the data acquisition are presented in **Chapter 2**. The different detector data readout cards used previously in Run1 and Run2 of ALICE are studied to present a framework for the need of CRU in the experiment. In this chapter the role of CRU in the experiment to handle the high data rates arising due to the LHC upgrade is highlighted. The methodology of the upgrade of data acquisition along with trigger system and the detector data links is elaborated. The same CRU will be used for the trigger transmission, the DAQ group and the detector frond end electronics group with custom functionality also which is exceptional from previous runtime experience. The configuration of the readout scheme for ALICE using the CRUs is also presented in this chapter.

The implementation and integration of CRU for the ALICE data acquisition system is the main inspiration for this thesis where the novelty is claimed. The main postulates of the thesis claiming the research aspect are the implementation of CRU since inception to the transfer of data in the DAQ server including the firmware, software and hardware developmental challenges. The integration and tests of the radiation tolerant high speed interfaces, trigger interface and the DAQ interface are handled in this dissertation which is quite a complex and challenging task. The in-depth study on the peculiarities of the design aspects and implementation strategy of CRU is focussed in **Chapter 3**. In this chapter; the features of custom developed radiation tolerant high speed protocol GBT, the TTC-PON based trigger interface and PCIe based DAQ interface of CRU with their comparative

features are summarized. The strategy followed for the implementation of CRU taking into account the different implementation options, its physical location and the intricate issues in the experiment is described and the choice of this scheme where CRU is kept in less or no radiation zone and its advantages are elaborated. The uniqueness of the hardware of CRU, overall complexities involved and their resolution, selection parameters of Arria-10 FPGA and the features and functionalities of CRU are detailed.

For the efficient performance of the complete system its front-end and the back-end interfaces needs to be individually tested and optimized. In this regard **Chapter 4** is an important section of the thesis. The flow of data and control signals in the readout scheme from and towards the CRU are mentioned. The interfacing links are implemented on silicon. Results for FPGA resource utilization, protocol latency, signal integrity, transceiver tests, BER analysis for quantitative measurement of signal quality and functional tests are presented. The detailed measurement, testing and performance analysis of the links implementation are discussed. A phase alignment logic for the clocks from different clock domains is implemented on FPGA for the stable 120 MHz operation of GBT. The consistency of GBT operation is verified by numerous repititions of firmware upgrade, system resets and power off and on cycles. BER as a function of optical power and the effect of optical transceivers parameter settings on the signal strength is studied. The scheme is also portable to the different set of FPGAs; vertical migration of the firmware to the next higher series of Intel's FPGA i.e 14nm stratix-10 FPGA is also possible. All these factors befit the present approach to cope up with current and future needs of the experiments.

CRU and other high luminosity upgrades for the experiments at LHC are associated with high rates of data transmission. The data rates in the DAQ systems are expected to reach few TB/sec in the next years. However, the high data rates are correlated with the increase in high frequency losses; which leads to distortion in the detected signal and degradation of signal integrity. One of the major concerns is to efficiently acquire data for all the collisions; thus error resilient and efficient data transmission with minimal signal

attenuation is required. Signal integrity is also essential for the proper clock and data recovery. Thus it is a challenge to reduce the bit error ratio (BER) and improve signal integrity for increased data rates.

To address the challenges of the high frequency losses arising due to the increase of the data rates with the experimental upgrades; an optimization technique of the multigigabit transceivers for high speed data communication links in HEP experiments is developed and presented in **Chapter 5** of the thesis. It is implemented on the Arria-10 FPGA. The technique presents the heuristic solution to tune the transceiver parameters for achieving the best performance at high-speeds for the transmission of data, trigger, timing and slow control information. It enhances the signal integrity and ascertained by Eye diagram and BER analysis as performance indicators posterior to the execution of proposed optimization technique. It is validated for the link rate of the high-speed communication protocols frequently used for data transmission in the HEP experiments viz. GBT, TTC-PON and 10 Gbps Ethernet. None of the published articles describes the minute details for the technique and its elaborate implementation procedure on FPGA. Hence the details of the transceiver optimization technique with its intricate features, the FPGA based test setup, the methodology to implement the proposed technique, its advantages are presented in this chapter. Test results and the improvements in the metrics of signal integrity for different link speeds are discussed and presented using Kiviat diagrams. It uses the Intel on chip instrumentation tools and does not require the external probing of FPGA pins or transceiver attributes. The proposed scheme is an optimized approach which reduces the number of iterations required. It makes the implementation of the new technique time efficient. The technique is useful for on-field system-level debugging, and the parameters can be reconfigured dynamically, allowing the user to configure the transceivers for optimum performance. The robustness of the technique has been tested with PRBS31 pattern. A large number of data vectors are acquired to achieve statistical reliability of the performed tests.

The tests are performed at a confidence level of 95 percent. The Intel FPGA set parame-

ters and the solution space values are marked on the kiviat diagram for the quick comparison between the parameters. It is concluded from the results of measurements that to attain the marked BER of 10^{-12}; the required optical power is reduced by 12.5%, 6.7% and 44.1% for 10Gbps, GBT and TTC-PON respectively. The BER is also improved over the received range of optical power. The CDR capability of the system is also enhanced as the least optical power required to recover the data traffic is reduced by 4.17%, 4.56% and 12.05% for 10Gbps, GBT and TTC-PON respectively. This transceiver optimization technique and its implementation approach would lend itself well for the users of other FPGAs also.

CRU is a complex electronics boards with the state-of-the-art FPGA technology along with optical I/Os. It has mezzanine cards for voltage distribution. The board has more than 1750 components on it with other accessories. Development and tests of the CRU hardware prototypes are presented in **Chapter 6** of the thesis. It is of key importance for the ALICE experiment. It constitutes an integral part of O2 and the detectors upgrade. The LHC beam is a costly affair. The CRU hardware; PCIe40 DAQ engine is a custom developed PCIe supported board. The performance and the hardware integrity of CRU has a direct linkage with the accomplishment of the experiment. To ensure this, extensive tests are carried out during the development stage as well as the functional tests for the verification of the developed prototype of CRU hardware.

The CRU hardware; PCIe40 DAQ engine is a custom developed PCIe supported board, based on latest Intel-Altera Arria-10 FPGA The PCIe40 readout board is based on PCIe x16 lanes Gen3 standard for interface with the server. It has total 130 number of high-speed signal lanes. The detailed tests of the newly developed and assembled CRU prototypes are required for the predictable and faithful operation of the boards in the experiment. The tests at the fabrication stage as well as the basic electrical performance are performed. The boards are also tested for the functional verification of its communication interfaces and for the qualification of the prototypes. In this chapter, the details of the step by step rigorous evaluations performed at both the hardware development stage and the functional

qualification tests of the prototype are presented. The functional tests are grouped as per the attributes of the process required. All the electrical tests and the optical loopback functionality checks are performed and found successful. The detection of the various chips integrated on the CRU boards are scanned using the tcl and the python scripts. Furthermore, the data are successfully transferred to the server PC at a throughput of 6.8 GB/s per lane of PCIe Gen3 x16 which is about 98.5 percent of the theoretical calculated throughput. The thorough study of the prototype CRU hardware and the performance tests are concluded in this chapter. The research and development summarized in the thesis is of high relevance for the firmware calibration and the hardware alignment purposes. This acts as a golden reference for the DAQ designers and the major postulates of the thesis are principal association for the implementation and integration of CRU in the ALICE experiment.

7.2 Future scope

The details of the analysis and investigations elucidated in the thesis presents a refined solution for the updated requirements of RUN-3 of ALICE detector. However, some additional issues required to be further investigated for the performance improvement of the readout system.

With the increase of the data with the next upcoming RUNs of ALICE, if all the incoming channels need to be operated at near to full occupancy then the excessive traffic needs to be diverted to the nearby CRU for load balancing. For this a detailed data driven mathematical model for CRU needs to be worked out for the prediction of data traffic load with the use of Queuing analysis. An efficient load prediction model helps to design data distribution scheme and to architect a dynamic switching topology for a sudden upsurge of data. An elaborate queuing model of the system needs to be thoroughly formulated, to help designers to initiate data traffic diversion and to predict or prevent data congestion. Also to handle the high throttle (back pressure) in CRU and avoid overloading in the computing system; a

bandwidth throttling system needs to be formulated to smoothly reduce the effective rate of collected events. It also requires a real time analysis of data with the use of the trigger signal.

The signal integrity of the hardware could futher be investigated by changing the separations between the power planes and rearranging. Also in order to make use of CRU for high data rate digital transmission systems other than the HEP experiments; like space applications and for the other industrial application of high data rate; the data needs to be secured through the GBT protocol using advanced encryption schemes to prevent unauthorized acccess of classified information by encrypting the data while transmitting over the optical medium.

For the upcoming Run 3 and Run 4 of LHC, the data rate of all the experiments ALICE, CMS, ATLAS, and LHCb will increase ten-fold. However the increased capacities of compute farms and storage cannot compensate the increased data sizes. Thus, optimized online event re-construction are required to make better use of the available hardwares and storage systems. The use of hardware accelerators like graphics processing units (GPUs) needs to be evaluated while FPGAs are in use.

REFERENCES

[1] C. Inguimbert *et al.*, ""Effective NIEL" in Silicon: Calculation Using Molecular Dynamics Simulation Results," *IEEE Transactions on Nuclear Science*, vol. 57, no. 4, pp. 1915–1923, Aug 2010. 1

[2] A. Breskin and R. Voss, *The CERN large hadron collider: accelerator and experiments*. CERN Geneva, 2009. 2, 8, 9

[3] S. Castillo and K. Ozanyan, "Field-programmable data acquisition and processing channel for optical tomography systems," *Review of Scientific Instruments*, vol. 76, no. 9, 9 2005. 3

[4] C. C. W. Robson *et al.*, "An FPGA- Based General-Purpose Data Acquisition Controller," *IEEE Transactions on Nuclear Science*, vol. 53, no. 4, pp. 2092–2096, Aug 2006. 3

[5] J. Toledo, F. Mora, and H. Müller, "Past, present and future of data acquisition systems in high energy physics experiments," *Microprocessors and Microsystems*, vol. 27, pp. 353–358, 2003. 3, 55, 95

[6] L. Evans, "LHC Machine," *JINST*, vol. 3, p. S08001, 2008. 5

[7] ATLAS-Collaboration, "The ATLAS experiment at the CERN large hadron collider," *JINST*, p. S08003, 2008. 6

[8] CMS-Collaboration, "The CMS experiment at the CERN LHC," *JINST*, p. S08004, 2008. 6

[9] ALICE-Collaboration, "The ALICE experiment at the CERN LHC. JINST 3," *JINST*, p. S08002, 2008. 6, 18, 51

[10] LHCb-Collaboration, "The LHCb Detector at the LHC," *JINST*, p. S08005, 2008. 6

[11] ALICE-collaboration *et al.*, "CERN faq LHC the guide, CERNBrochure-2008-001-Eng, 2008," *CERN-Brochure-2017-002-Eng*, 2017. 8

[12] D. E. Morrissey, T. Plehn, and T. M. Tait, "Physics searches at the lhc," *Physics Reports*, vol. 515, no. 1-2, pp. 1–113, 2012. 8

[13] G. L. Kane and A. Pierce, *Perspectives on LHC physics*. World Scientific, 2008. 8

[14] L. Rossi, O. Brüning *et al.*, "High Luminosity Large Hadron Collider – A description for the European Strategy Preparatory Group," in *European Strategy Preparatory Group-Open Symposium, Krakow*, 2012. 8

[15] B. A. et al and the ALICE Collaboration, "Upgrade of the ALICE Experiment: Letter Of Intent," *Journal of Physics G: Nuclear and Particle Physics*, vol. 41, no. 8, p. 087001, 2014. 10, 17

[16] P. Antonioli, A. Kluge, W. Riegler, and for the ALICE Collaboration, "Upgrade of the ALICE Readout & Trigger System," *CERN Technical Design Report*, vol. CERN-LHCC-2013-019, ALICE-TDR-015, 2013. 11, 13, 17, 33

[17] A. Kluge and P. V. Vyvre, "The detector read-out in ALICE during Run 3 and 4," *Change*, vol. 1, p. 3, 2016. 12, 35, 43

[18] T. A. Collaboration, "The alice experiment at the cern lhc," *Journal of Instrumentation*, vol. 3, no. 08, p. S08002, 2008. [Online]. Available: http://stacks.iop.org/1748-0221/3/i=08/a=S08002 20, 29, 30

[19] B. A. et al and the ALICE Collaboration, "Technical Design Report for the Upgrade of the ALICE Inner Tracking System," Tech. Rep. CERN-LHCC-2013-024. ALICE-TDR-017, 2013. 32

[20] B. Ketzer *et al.*, "A time projection chamber for high-rate experiments: Towards an upgrade of the ALICE TPC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 732, pp. 237–240, 2013. 32

[21] S. Barboza *et al.*, "SAMPA chip: a new ASIC for the ALICE TPC and MCH upgrades," *Journal of Instrumentation*, vol. 11, no. 02, p. C02088, 2016. 33

[22] Arnaldi *et al.*, "Front-end electronics for the RPCs of the ALICE dimuon trigger," *IEEE transactions on nuclear science*, vol. 52, no. 4, pp. 1176–1181, 2005. 33

[23] R. Divia *et al.*, "Proposal of an Heartbeat trigger for ALICE Run 3," *O2 Project CWG4*, 2013. 36

[24] W. H. Smith, "Triggering at LHC experiments," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 478, no. 1, pp. 62 – 67, 2002, proceedings of the ninth Int.Conf. on Instrumentation. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016890020101720X 36

[25] F. Carena, W. Carena, V. C. Barroso, F. Costa, S. Chapeland, C. Delort, E. Dénes, R. Divià, U. Fuchs, A. Grigore *et al.*, "DDL, the ALICE data transmission protocol and its evolution from 2 to 6 Gb/s," *Journal of Instrumentation*, vol. 10, no. 04, p. C04008, 2015. 37, 38

[26] F. Carena, W. Carena, S. Chapeland, V. C. Barroso, F. Costa, E. Dénes, R. Divià, U. Fuchs, A. Grigore, T. Kiss *et al.*, "The ALICE data acquisition system," *Nuclear*

Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 741, pp. 130–162, 2014. 37, 46

[27] H. Engel and U. Kebschull, "Common read-out receiver card for ALICE Run2," *Journal of Instrumentation*, vol. 8, no. 12, p. C12016, 2013. 37, 38, 45

[28] W. Carena, J. Marin, J. Sulyán, C. Soós, P. Van de Vyvre, P. Csató, R. Divià, A. Vascotto, F. Carena, K. Schossmaier *et al.*, "PCI-based readout receiver card in the ALICE DAQ system," in *8th Workshop on Electronics for LHC Experiments, Colmar, France*, 2002, pp. 9–13. 37, 46

[29] ALICE collaboration, "Technical design report of the trigger, data acquisition, high level trigger and control system," *CERN, January*, 2004. 37

[30] J. Mitra, S. Khan, S. Mukherjee, R. Paul, and for the ALICE collaboration, "Common Readout Unit (CRU)-A new readout architecture for the ALICE experiment," *Journal of Instrumentation*, vol. 11, no. 03, p. C03021, 2016. 38, 60, 113

[31] M. C. Herbordt *et al.*, "Achieving High Performance with FPGA-Based Computing," *Computer*, vol. 40, no. 3, pp. 50–57, March 2007. 39

[32] J. Mitra, S. A. Khan, M. B. Marin, J. P. Cachemiche, E. David, F. Hachon, F. Rethore, T. Kiss, S. Baron, A. Kluge *et al.*, "GBT link testing an d performance measurement on PCIe40 and AMC40 custom design FPGA boards," *Journal of Instrumentation*, vol. 11, no. 03, p. C03039, 2016. 39, 74, 90

[33] Moreira *et al.*, "The GBT project," in *Proceedings of the Topical Workshop on Electronics for Particle Physics*, 2009, pp. 342–346. 39

[34] D.-M. Kolotouros, S. Baron, C. Soos, and F. Vasey, "A TTC upgrade proposal using bidirectional 10G-PON FTTH technology," *Journal of Instrumentation*, vol. 10, no. 04, p. C04001, 2015. 39

[35] M. Bellato *et al.*, "A PCIe Gen3 based readout for the LHCb upgrade," in *Journal of Physics: Conference Series*, vol. 513. IOP Publishing, 2014, p. 012023. 40, 62, 78

[36] B. Taylor, "TTC distribution for LHC detectors," *IEEE Transactions on Nuclear Science*, vol. 45, no. 3, pp. 821–828, 1998. 40

[37] M. Lamanna, "The LHC computing grid project at CERN," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 534, no. 1, pp. 1 – 6, 2004, proceedings of the IXth International Workshop on Advanced Computing and Analysis Techniques in Physics Research. 43

[38] P. Antonioli, A. Kluge, and W. Riegler, "Upgrade of the ALICE Readout and Trigger System," CERN, Tech. Rep. CERN-LHCC-2013-019. ALICE-TDR-015, Sep 2013. [Online]. Available: https://cds.cern.ch/record/1603472 46

[39] H. G. Essel, "FutureDAQ for CBM: on-line event selection," *IEEE Transactions on Nuclear Science*, vol. 53, no. 3, pp. 677–681, June 2006. 47

[40] A. D. Oancea *et al.*, "A resilient, flash-free soft error mitigation concept for the CBM-ToF read-out chain via GBT-SCA," in *2015 25th International Conference on Field Programmable Logic and Applications (FPL)*, Sept 2015, pp. 1–4. 47

[41] S. A. Khan, J. Mitra, E. David, T. Kiss, and T. K. Nayak, "A potent approach for the development of FPGA based DAQ system for HEP experiments," *Journal of Instrumentation*, vol. 12, p. T10010, 2017. 47, 60

[42] G. F. Knoll, *Radiation detection and measurement*. John Wiley & Sons, 2010. 47

[43] P. Moreira *et al.*, "The GBT-SerDes ASIC prototype," *Journal of Instrumentation*, vol. 5, no. 11, p. C11022, 2010. [Online]. Available: http://stacks.iop.org/1748-0221/5/i=11/a=C11022 47

[44] P. Vichoudis, S. Baron *et al.*, "The Gigabit Link Interface Board (GLIB), a flexible system for the evaluation and use of GBT-based optical links," *Journal of Instrumentation*, vol. 5, no. 11, p. C11007, 2010. [Online]. Available: http://stacks.iop.org/1748-0221/5/i=11/a=C11007 47, 76

[45] S. Baron *et al.*, "Implementing the GBT Data Transmission Protocol in FPGAs," in *TWEPP-09 Topical Workshop on Electronics for Particle Physics*, Paris, France, Sep. 2009, pp. 631–635. [Online]. Available: http://hal.in2p3.fr/in2p3-00468912 48, 76

[46] R. M. Lesma, F. Alessio, J. Barbosa, S. Baron, C. Caplan, P. Leitao, C. Pecoraro, D. Porret, and K. Wyllie, "The versatile link demo board (vldb)," *Journal of Instrumentation*, vol. 12, no. 02, p. C02020, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=02/a=C02020 48

[47] F. W. Sexton, "Destructive single-event effects in semiconductor devices and ICs," *IEEE Transactions on Nuclear Science*, vol. 50, no. 3, pp. 603–621, June 2003. 48

[48] M. Nomachi and S. Ajimura, "Serial data link on advanced TCA back plane," in *14th IEEE-NPSS Real Time Conference, 2005.*, June 2005. 49, 61

[49] S. A. Khan *et al.*, "Common Readout Unit (CRU)-A New Readout Architecture for the ALICE experiment at the CERN-LHC," in *DAE Symp. Nucl. Phys.*, vol. 59, 2014, pp. 972–973. 51

[50] G. Tambave and A. Velure, "High speed continuous DAQ system for readout of the ALICE SAMPA ASIC," in *2016 IEEE-NPSS Real Time Conference(RT)*, June 2016, pp. 1–4. 53

[51] A. Aloisio *et al.*, "High-Speed, Fixed-Latency Serial Links With FPGAs for Synchronous Transfers," *IEEE Transactions on Nuclear Science*, vol. 56, no. 5, pp. 2864–2873, Oct 2009. 54

[52] I. Altera, "Stratix V Device Handbook Volume 2: Transceiver," 2017. 54, 74, 78

[53] J. Reis, V. Shukla, D. Stauffer, and K. Gass, "Technology options for 400G implementation," in *Optical Internetworking Forum White Paper*, 2015. 55

[54] I. trade association. (2012) InfiniBand architecture specification, volume 2, release 1.3. 55, 58

[55] J. Cachemiche *et al.*, "The PCIe-based readout system for the LHCb experiment," *Journal of Instrumentation*, vol. 11, no. 02, p. P02013, 2016. 55, 56, 95

[56] I. C. Society, "IEEE Standard for Ethernet," *IEEE Std 802.3^{TM}-2015*, vol. 4, pp. 38–687, 2015. 55, 98, 105

[57] F. Costa *et al.*, "The New Frontier of the DATA Acquisition Using 1 and 10 Gb/s Ethernet links," *Physics Procedia*, vol. 37, pp. 1956–1964, 2012. 55

[58] L. Collaboration, "LHCb Tracker Upgrade Technical Design Report," Tech. Rep. CERN-LHCC-2014-001. LHCB-TDR-015, Feb 2014. 55, 78

[59] A. Technologies, "PCI Express vs Ethernet Selecting the Superior Technology for Real Time Embedded Systems," 2015. 56

[60] B. Batmaz and A. Dogan, "UDP/IP Protocol Stack with PCIe Interface on FPGA," in *Proceedings of the International Conference on Embedded Systems and Applications (ESA)*, 2015, p. 49. 56

[61] Y. Jiang, C.-K. Tham, and C.-C. Ko, "Challenges and approaches in providing QoS monitoring," *International Journal of Network Management*, vol. 10, 2000. 56, 97

[62] E. Mendes, S. Baron, D. Kolotouros, C. Soos, and F. Vasey, "The 10G TTC-PON: challenges, solutions and performance," *Journal of Instrumentation*, vol. 12, p. C02041, 2017. 57, 58

[63] J. Mitra, E. David, E. Mendez, S. A. Khan, T. Kiss, S. Baron, A. Kluge, and T. Nayak, "Trigger and timing distributions using the ttc-pon and gbt bridge connection in alice for the lhc run 3 upgrade," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 922, pp. 119 – 133, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900218318904 57

[64] M. B. Marin, S. Baron *et al.*, "The GBT-FPGA core: features and challenges," *Journal of Instrumentation*, vol. 10, p. C03021, 2015. 58

[65] J. Mitra, S. A. Khan *et al.*, "Error Resilient Secure Multi-gigabit Optical Link Design for High Energy Physics Experiment," in *VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID), 2016 29th International Conference on*. IEEE, 2016, pp. 427–432. 58, 60

[66] P. PCI-SIG, "Express Base Specification," *Revision*, vol. 1, p. 422, 2002. 58

[67] S. D. et al., "Method of transporting a PCI express packet over an IP packet network ," Patent, 2006. 58

[68] A. computer, "Technology Brief, Fibre Channel Basics," 2006. 58

[69] R. R. Tummala, "SOP: what is it and why? A new microsystem-integration technology paradigm-Moore's law for system integration of miniaturized convergent systems of the next decade," *IEEE Transactions on Advanced Packaging*, vol. 27, no. 2, pp. 241–249, 2004. 60

[70] I. Altera, "Intel Arria 10 Device Overview," 2018. 60, 67, 78, 114, 115, 120

[71] J. Mitra, S. A. Khan *et al.*, "Trigger and Timing Distributions using the TTC-PON and GBT Bridge Connection in ALICE for the LHC Run 3 Upgrade," *arXiv preprint arXiv:1806.01350*, 2018. 64, 67

[72] A. Caratelli, S. Bonacini, K. Kloukinas, A. Marchioro, P. Moreira, R. De Oliveira, and C. Paillard, "The GBT-SCA, a radiation tolerant ASIC for detector control and monitoring applications in HEP experiments," *Journal of Instrumentation*, vol. 10, no. 03, p. C03034, 2015. 64

[73] S. A. Khan, R. Paul, T. Nayak, F. Costa, A. Chakrabarti, E. David, S. Mukherjee, T. K. Das, T. Kiss, and J. Mitra, "Implementation of I2C bus master controller for CRU Slow Control in ALICE at LHC," in *DAE Symp. Nucl. Phys.*, vol. 61, 2016, pp. 1070–1071. 65, 73

[74] R. e. Schwemmer, "Evaluation of 400 m, 4.8 Gbit/s Versatile Link lengths over OM3 and OM4 fibres for the LHCb upgrade," *Journal of Instrumentation*, vol. 9, no. 03, p. C03030, 2014. 74

[75] B. Razavi, "Challenges in the design high-speed clock and data recovery circuits," *IEEE Communications magazine*, vol. 40, pp. 94–101, 2002. 74, 80, 113

[76] FPGA Working group – J. Mendez, "GBT-FPGA Tutorial," *TWEPP '16*, 2016. [Online]. Available: https://indico.cern.ch/event/489996/contributions/2291863/attachments/1345764/2028939/GBTTutorial_-_TWEPP2016.pdf 77

[77] F. Alessio, S. Baron, M. B. Marin, J. Cachemiche, F. Hachon, R. Jacobsson, and K. Wyllie, "Clock and timing distribution in the LHCb upgraded detector and readout system," *Journal of Instrumentation*, vol. 10, no. 02, p. C02033, 2015. 78

[78] J. P. Cachemiche *et al.*, "Recent developments for the upgrade of the LHCb readout system," *Journal of Instrumentation*, vol. 8, no. 02, p. C02014, 2013. 78

[79] Altera, "Stratix V GX FPGA Development Board Reference Manual," 2015. 78

[80] Altera Corporation, "Transceiver Clocking in Stratix-V Devices," *Stratix V Device Handbook*, vol. 3, 2012. 80

[81] Intel Corporation, "Intel® Arria® 10 Transceiver PHY User Guide," *Arria 10 Device Handbook*, p. 615, 2015. 80

[82] X. Liu, Q.-X. Deng, and Z.-K. Wang, "Design and fpga implementation of high-speed, fixed-latency serial transceivers," *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 561–567, 2014. 82

[83] Intel, "Understanding Metastability in FPGAs," *White Paper Intel-Altera*, 2009. 85

[84] Intel Inc., "Avalon Interface Specifications," *Interface specifications*, 2018. 88, 101

[85] I. Altera, "Quartus prime standard edition handbook volume 1: Design and synthesis," 2017. 92, 99, 103, 114, 121

[86] Intel. Corporation, "JNEye User Guide," *User guide*, 2017. [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/ug/ug_jneye.pdf 92

[87] R. H. Derek Vaughan and M. Fields, "Applications for Embedded Optic Modules in Data Communications," *Avago Technologies (White Paper)*, 2011. 92, 108, 138

[88] C. SOOS, "GBT protocol implementation on GBT protocol implementation on Xilinx FPGAs," *LHCB meeting*, 2008. 94

[89] F. Costa *et al.*, "The new frontier of the data acquisition using 1 and 10 gb/s ethernet links," *Physics Procedia*, vol. 37, pp. 1956 – 1964, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187538921201913X 95

[90] J. Lawley, "Understanding performance of pci express systems," *WP350 (v1. 2). Xilinx*, 2014. 97

[91] I. Altera, "Introduction to Intel FPGA IP Cores," 2017. 98, 100

[92] I. Nios, "Processor Reference Handbook," 2009. 100

[93] I. Altera, "Scatter-Gather DMA Controller Core Overview," 2009. 101

[94] I. Nios, "Getting Started with Nios II Software in Eclipse," 2014. 102

[95] A. Kuzmin and D. Fey, "Optical link testing and parameters tuning with a test system fully integrated into FPGA," in *The Fourth International Conference on Advances in System Testing and Validation Lifecycle*, 2012. 103

[96] Xiang *et al.*, "High-speed serial optical link test bench using FPGA with embedded transceivers," 2009. 103

[97] Altera Corporation, "High-Speed Link Tuning Using Signal Conditioning Circuitry in Stratix V Transceivers," *White Paper*, 2015. [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/an/an678.pdf 103

[98] Altera. Corporation, "Understanding the Pre-Emphasis and Linear Equalization Features in Stratix IV GX Devices," *Application Note*, 2010. [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/an/an602.pdf 103

[99] Altera, "10-Gbps Ethernet MAC MegaCore Function User Guide," 2014. 105

[100] S. Detraz *et al.*, "FPGA-based bit-error-rate tester for SEU-hardened optical links," *JINST*, 2009. [Online]. Available: http://cds.cern.ch/record/1236362/files/p636.pdf?version=1 107

[101] CRU Team Members. (2018) ALICE CRU Hardware, Firmware, and Software Development–ALICE Web TWiki. 109

[102] S. H. Hall and H. L. Heck, *Advanced signal integrity for high-speed digital designs*. John Wiley & Sons, 2011. 113, 114

[103] I. Altera, "High-SpeedLink Tuning Using Signal Conditioning Circuitry in Stratix V Transceivers," 2015. 114, 121, 122

[104] S. Committee *et al.*, "Sff-8431 specifications for enhanced small form factor pluggable module sfp+, revision 4.1, july 6, 2009." 120

[105] S. I. Green, "Multichannel bit error rate tester for fiber optic transceiver testing," *Review of scientific instruments*, vol. 73, no. 8, pp. 3125–3127, 2002. 121

[106] H. Badaoui, Y. Frignac, P. Ramantanis, B. E. Benkelfat, and M. Feham, "Prqs sequences characteristics analysis by auto-correlation function and statistical properties," *IJCSI*, p. 39, 2010. 121

[107] D. Mitić, A. Lebl, and Ž. Markov, "Calculating the required number of bits in the function of confidence level and error probability estimation," *Serbian Journal of Electrical Engineering*, vol. 9, pp. 361–375, 2012. 125

[108] L. J. Ippolito, "Appendix b: Error functions and bit error rate," *Satellite Communications Systems Engineering: Atmospheric Effects, Satellite Link Design and System Performance*, pp. 363–366. 128

[109] S. Chapra and R. P. Canale, "Numerical methods for engineers : with personal computer applications / steven c. chapra, raymond p. canale," 05 2018. 128

[110] E. K. Gan, H. Zheng, and G. Lim, *Laser drilling of micro-vias in PCB substrates*, 2000. 136

[111] "Convection vs vapour phase reflow in LED and BGA assembly, author=Dziurdzia, Barbara and Sobolewski, Maciej and Mikolajek, Janusz, journal=Soldering & Surface Mount Technology, volume=30, number=2, pages=87–99, year=2018, publisher=Emerald Publishing Limited." 137

[112] F. Sarvar and P. P. Conway, *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*. 138

[113] D. Suraski, "Benefits of a ramp-to-spike reflow profile," *SMT SURF MOUNT TECHNOL MAG*, vol. 14, no. 4, p. 3, 2000. 138

[114] "MiniPOD Embedded Optical Modules," *MiniPODTM AFBR-81uVxyZ 12-channel transmitter, AFBR-82uVxyZ 12-channel receiver high density, pluggable, embedded parallel-fiber-optics modules product brief*. 138

[115] S. Mukherjee *et al.*, "An Efficient Approach to Manage DMA Descriptors and Evaluate PCIe-Based DMA Performance for ALICE Common Readout Unit (CRU)," in *Advanced Detectors for Nuclear, High Energy and Astroparticle Physics*. Singapore: Springer Singapore, 2018, pp. 107–118. 164