FPGA based high speed DAQ systems for high-energy physics experiments: potential challenges

Dr. Shuaib Ahmad Khan

Scientific Officer

VECC (DAE), Kolkata

Outline

- Journey to High speed Data Acquisition System
- DAQ Overview in context of HEP experiments
  - Example of ALICE at CERN
  - Requirements and Features
- R&D on CRU
  - Choice of Location & FPGA
  - Readout Scheme
  - Latency measurement of the GBT link
  - BER analysis
  - Eye Diagram test
  - Phase alignment Logic
  - Multi-Gigabit Transceiver Optimization
- Hardware complexities and Assembly issues

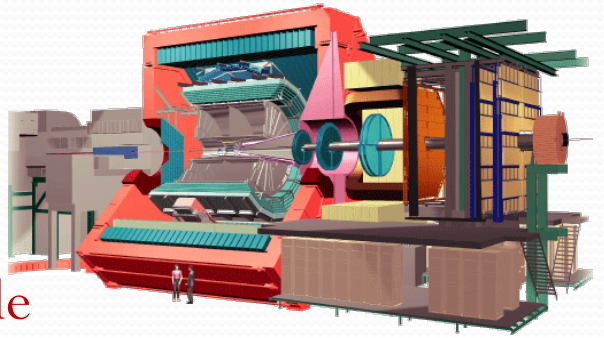
Journey to high speed DAQ

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

DAQ

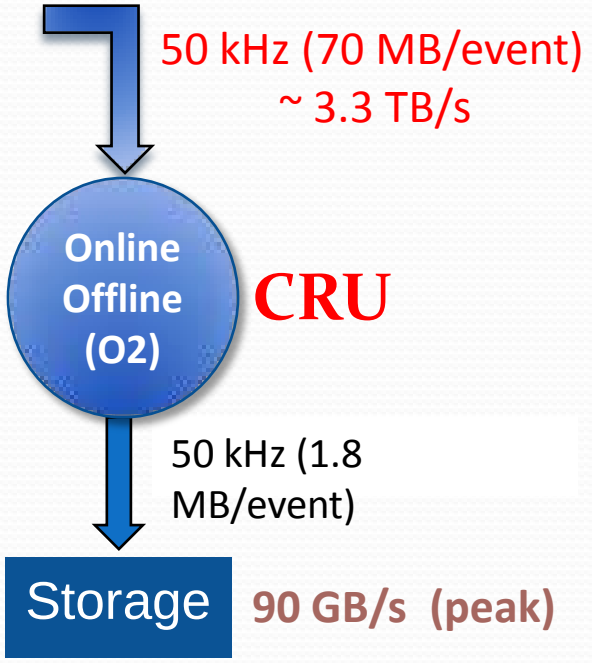
Context of ALICE Experiment

ALICE@Run3 and Role of FPGA based CRU



ALICE statistics before and after the scheduled upgrade

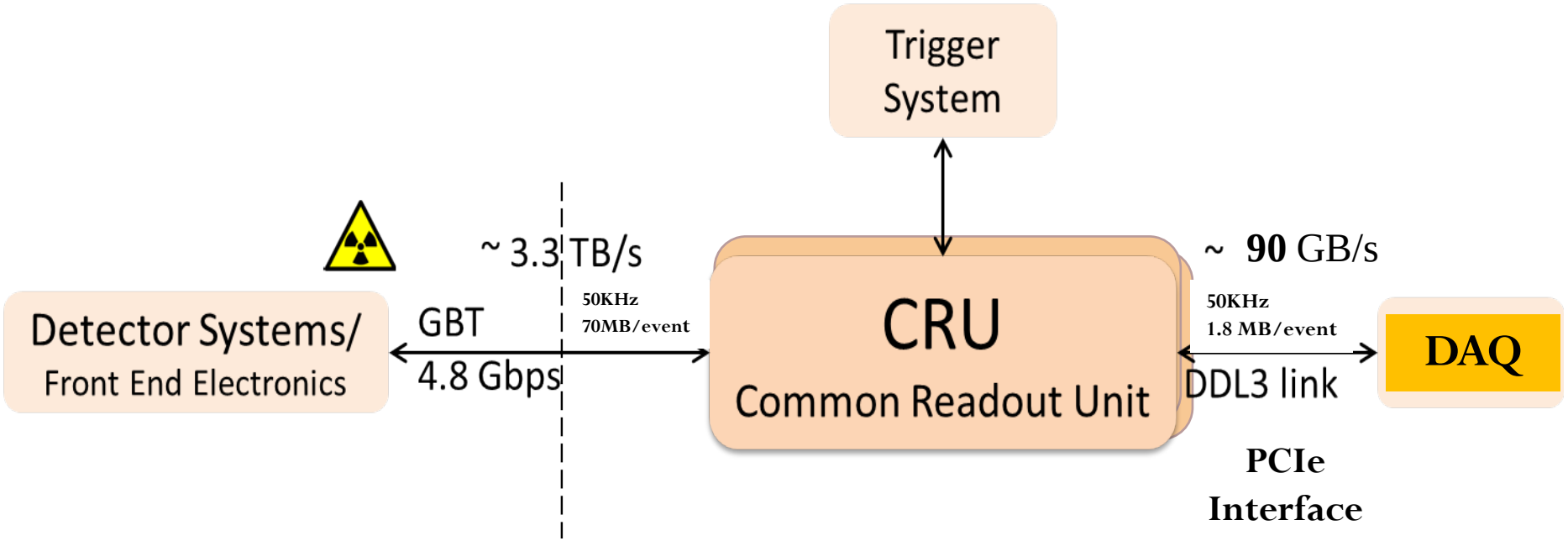
Parameter	RUN-2 (2014-2018)	RUN-3 (2021-2023)
Luminosity	10^{27} cm^{-2} s^{-1}	6 x 10^{27} cm^{-2} s^{-1}
Collision rate	8 kHz (PbPb)	50 kHz (PbPb)
Max Readout rate	500 Hz (PbPb)	PbPb: 50 kHz pp & pPb: 200 kHz



Requirement and Features for a Common Readout Unit

- To preserve smart features of the present DAQ system, like:
  1. Common interfaces between the various detectors and the common Online computing farm.
  2. Optimizing the system level costs by aggregating the digital data from very high number of sources to high bandwidth optical data links.
- DDL3 links, higher data throughput requirements of operation in Run3
- Reduce the number of different link technologies presently used
- Read-out the very large number high bandwidth, serial detector side links, and multiplex to common, even higher bandwidth, server-side links (uplinks, or DDL3).
- Minimize the number of physical links between the different nodes of the system

Common Readout Unit (CRU): FPGA based Readout board



CRU Task: To reduce recorded data volume by online reconstruction.

- Data concentration, reconstruction, multiplexing
- Gets the data from detector electronics and Trigger
- Moves the data to DAQ

Links: GBT Link (Radiation Tolerant High Speed Optical Link)

10 Gigabit Passive Optical Network (PON)

DDL3 link (PCIe Gen 3 x16)

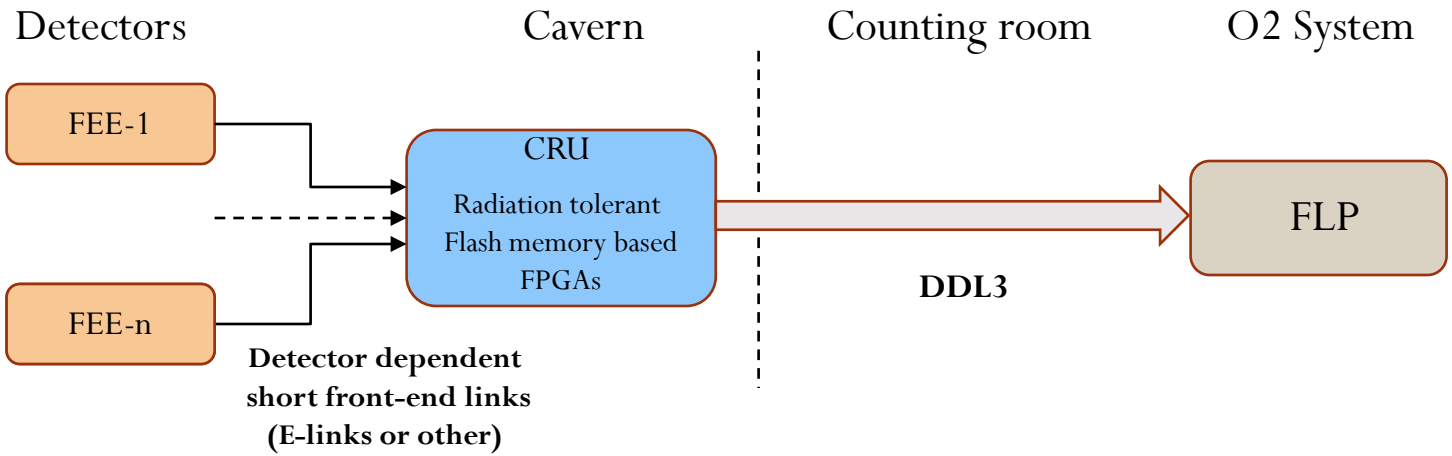
Choice of Location

- The application specific functionalities, require Common Read-out Units be implemented as electronics boards with custom designed, programmable functionality based on up-to-date FPGA technology.

- Two basic versions, depending on the physical location of the CRUs.
  - Version A: CRU in Cavern, close to the detectors.
  - Version B: CRU in the Control room.

Located at Cavern

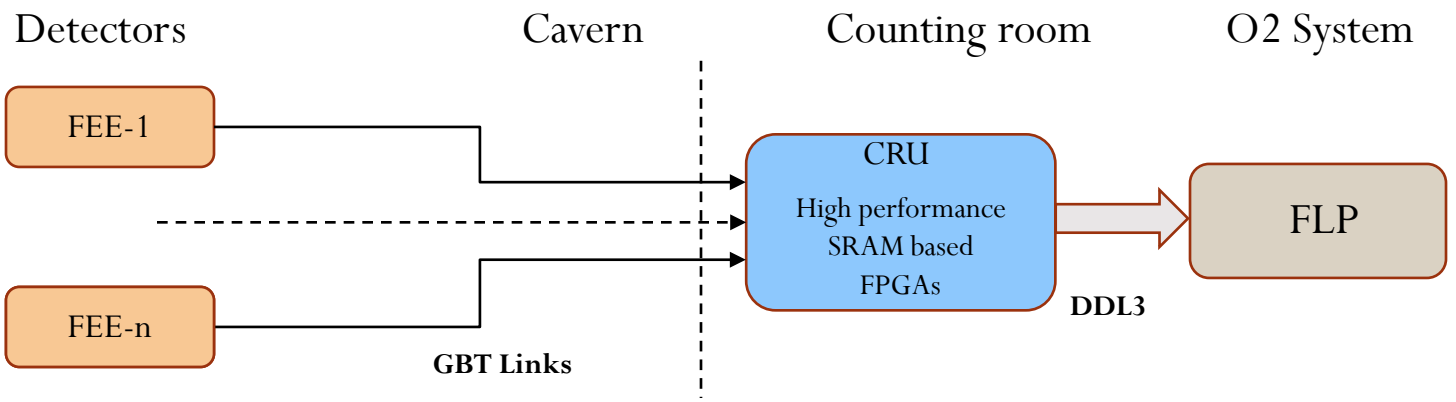
- CRU is placed in the ALICE cavern, mounted on/near the detector structures.
- Subject to radiation, must be custom designed radiation tolerant electronics.
- Detector systems are connected via short electrical/optical cables to CRU.
- Connection to the DAQ via long, high performance links (DDL3)



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

Located at Control room

- Located in the control room, CRU doesn't have to be radiation tolerant.
- Detector systems are connected to the CRU via GBT links and long optical fibers.
- Interfacing with the DAQ online computing is done with short, high performance industrial standard links (DDL3).
- Option to put the CRU as a PCI Express card in the computers of online system



Implementation version B: CRU as FPGA boards in the control room.

CRU Location contd..

- Location of CRU-control room would be in the ground level counting room, thus it is accessible during operation. No additional electronics, cabling, cooling needs to be installed or maintained on the detector.
- The CRU-control room electronics is not subject to radiation, thus state-of-the art high performance FPGAs be employed along with COTS components.
- Low impact on cavern infrastructure, ecosystem and legacy support.

Choice of FPGA

CRU in Cavern	CRU in Counting Room
Radiation Tolerant Flash memory based FPGA	Not radiation Tolerant SRAM based FPGAs
Micro-semi Smart Fusion 2	Xilinx Virtex 7 or Intel Stratix V GX/Arria-10

Advantages of the Proposed approach over the Conventional approach

Parameter	Conventional approach	Proposed approach
DPU location	Experiment area: radiation environment	Counting room: Controlled or no radiation
DPU Technology	Radiation hard electronics	Commercially 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 DPU (b)Between DPU 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

Choice of FPGA

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	

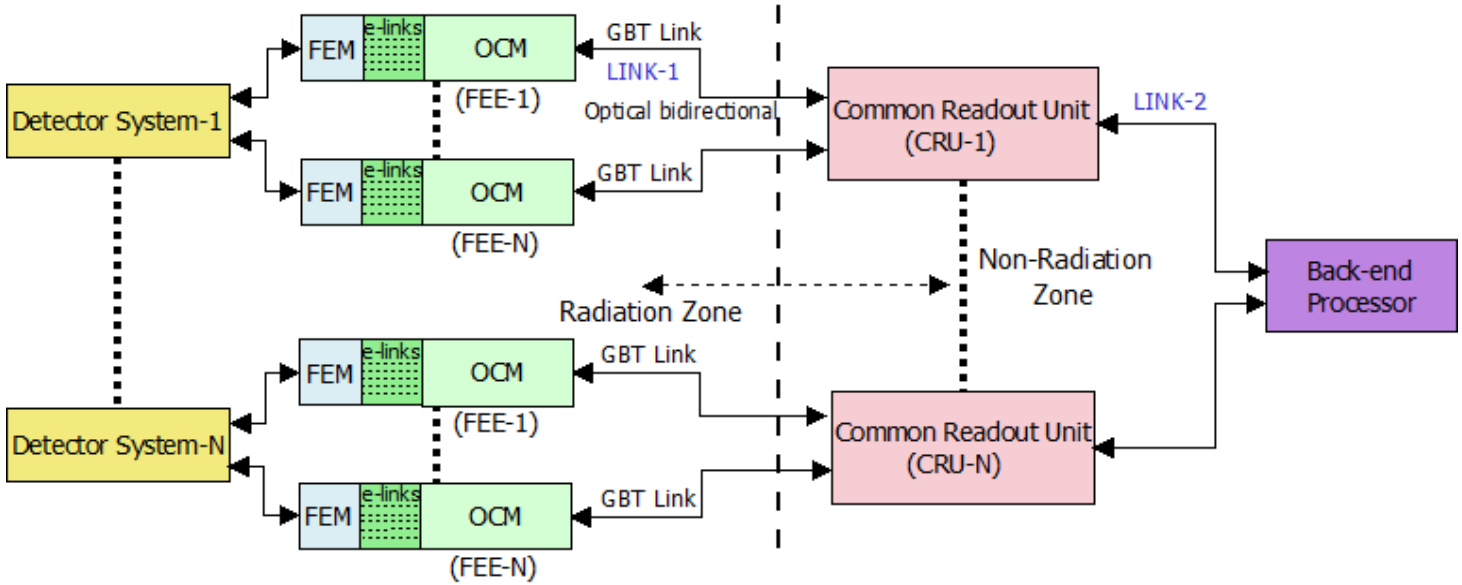
FPGA Selection Parameters

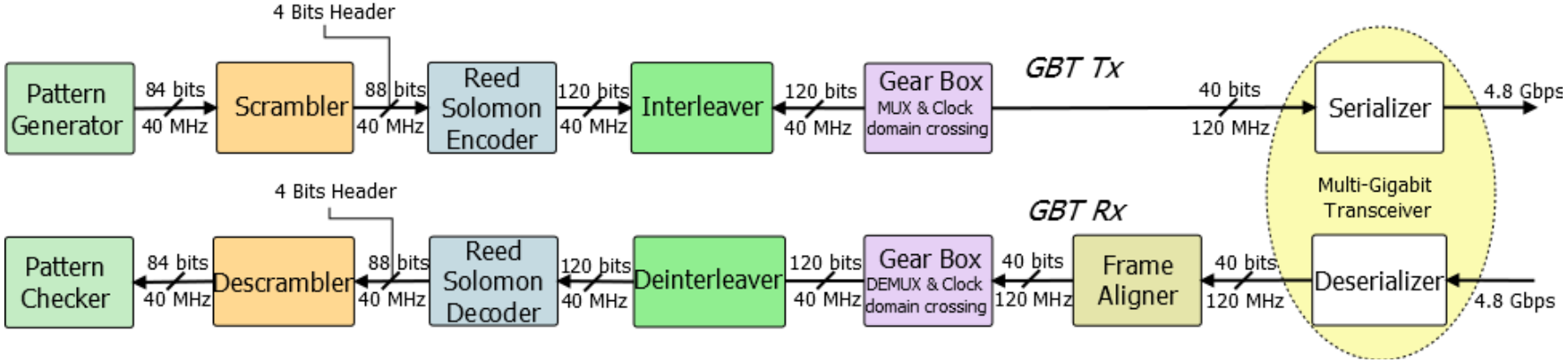
FPGA family with maximum SerDes Speed

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						

Readout Scheme

Readout Scheme

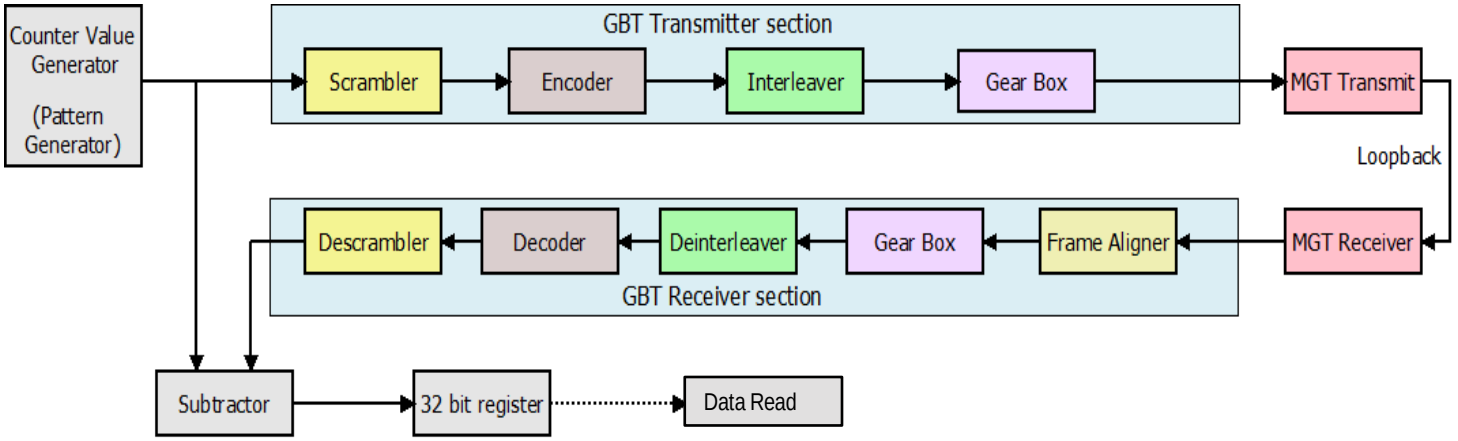




Gigabit Transceiver logic core firmware on FPGA

Latency Measurement for GBT

Latency measurement for GBT

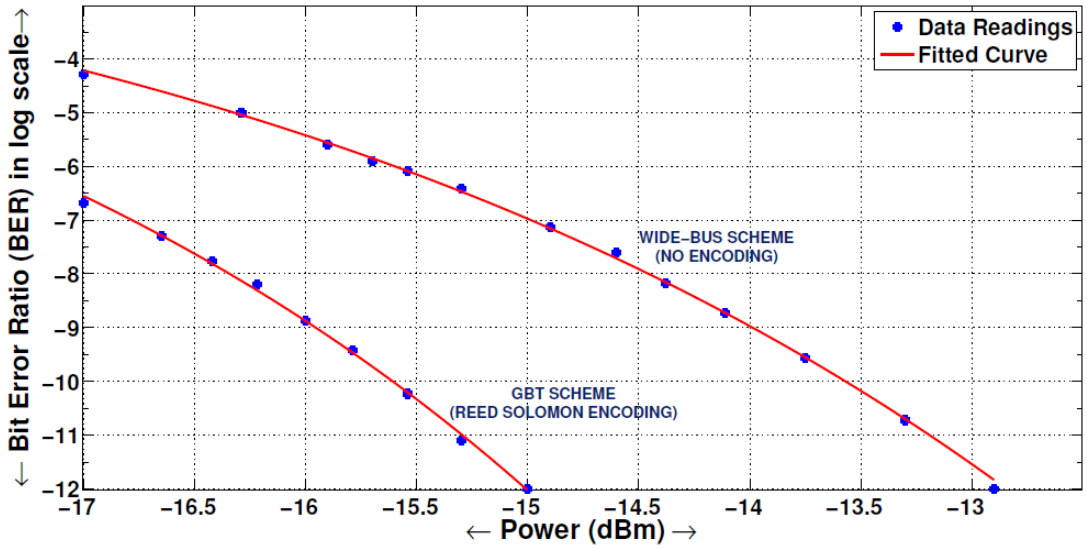


Latency for path	GBT Link mode of operation(ns)			
1 LHC Clock Cycle = 25 ns	Tx Latopt Rx Latopt	Tx Latopt Rx Std	Tx Std Rx Latopt	Tx Std Rx Std
$\mathcal{L}_1 = GBT\ Tx - MGT\ Tx - MGT\ Rx - GBT\ Rx$	150	350	350	600
$\mathcal{L}_2 = GBT\ Tx - GBT\ Rx$	75	275	250	425
$\mathcal{L}_3 = MGT\ Tx - MGT\ Rx$	75	75	100	175

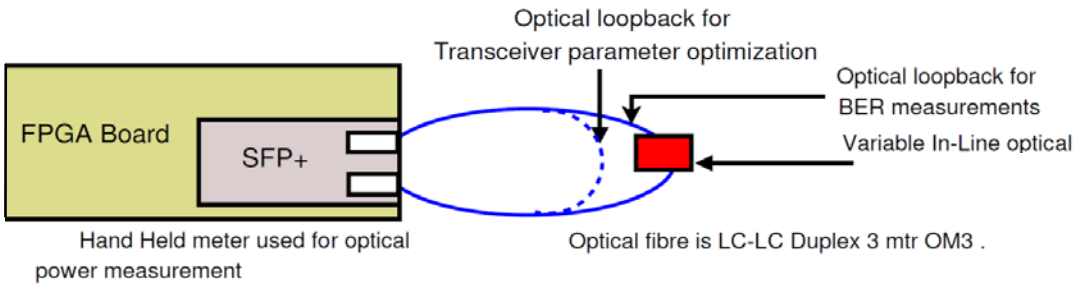
Publication: *Shuaib Ahmad Khan, et al. "A potent approach for the development of FPGA based DAQ system for HEP experiments." Journal of Instrumentation (2017) doi.org/10.1088/1748-0221/12/10/T10010 Vol. 12*

BER Analysis

BER measurement for GBT Frame coding and GBT wide-Bus mode



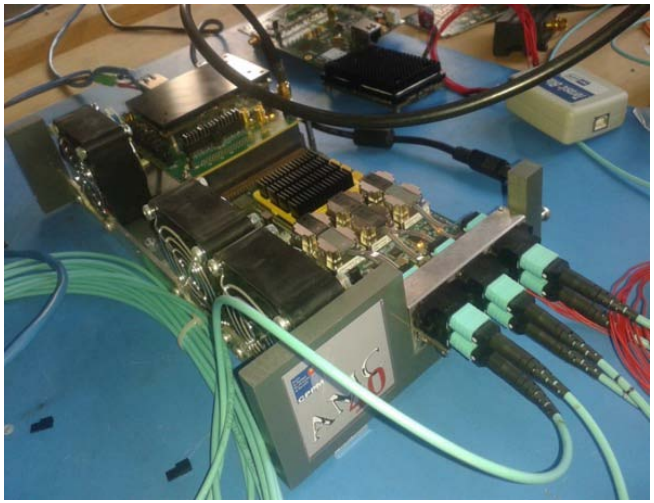
Margin of Receiver Sensitivity: 2.1 dBm

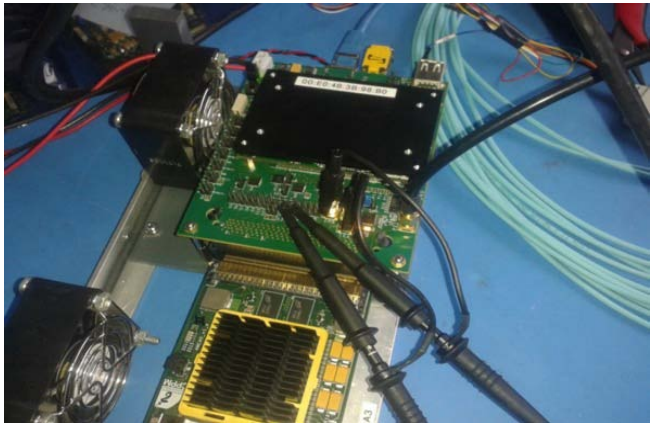


Publication: *Shuaib Ahmad Khan, et al. "A potent approach for the development of FPGA based DAQ system for HEP experiments." Journal of Instrumentation (2017) doi.org/10.1088/1748-0221/12/10/T10010 Vol. 12*

Eye Diagram Test

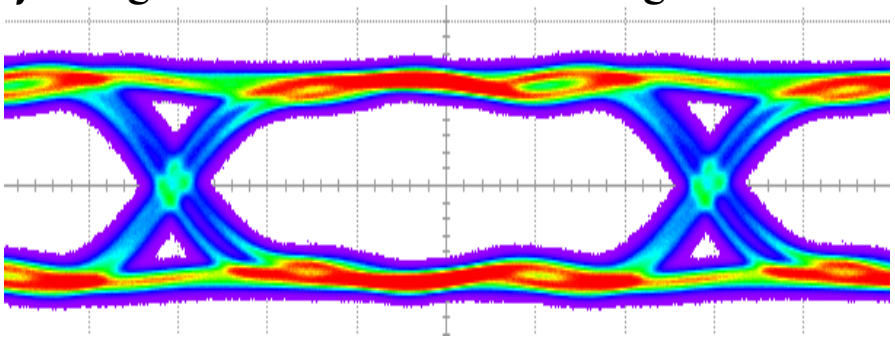
Eye Diagram Test





FPGA Test Setup

Eye Diagram for the GBT Link testing on Altera



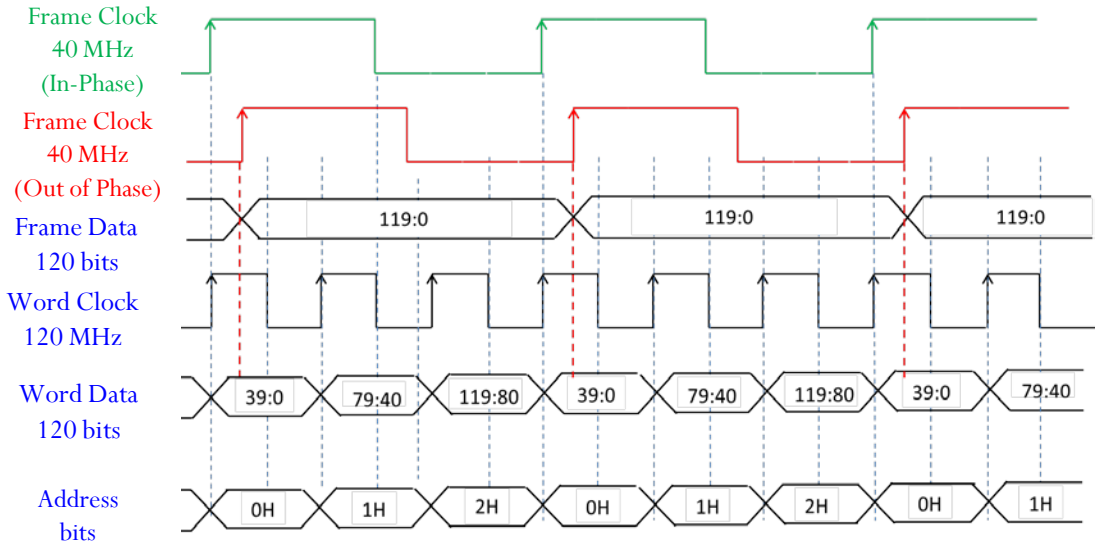
Eye Width is 176.8 ps, Eye Height is 373 mV
BER of 5.525×10^{-12}

Jitter Parameter	Unit Picosecond(ps)
Deterministic Jitter	5.503
Data Dependant Jitter (DDj)	11.228
Periodic Jitter (Pj)	6.75
Inter Symbol Interference (ISI)	11.095
Standard Deviation(s)	2.989
Duty Cycle Jitter (DCD)	2.000
Total Jitter (Tj)	51.148

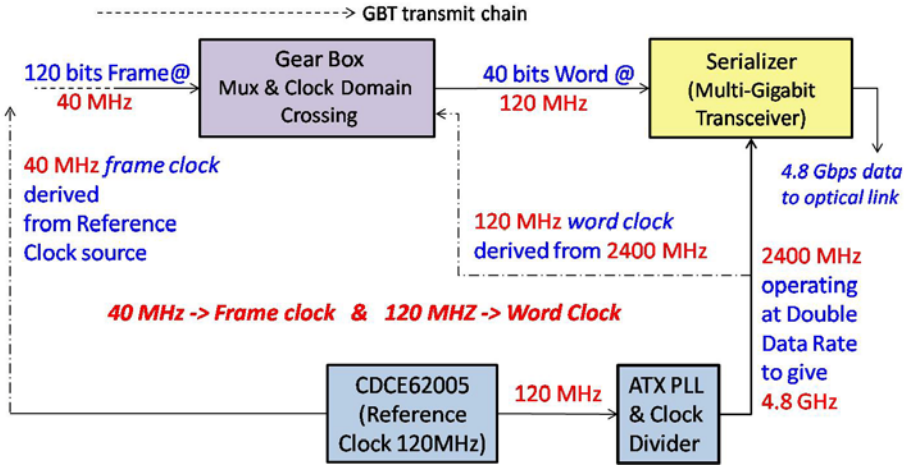
Publication: *Shuaib Ahmad Khan, et al. "A potent approach for the development of FPGA based DAQ system for HEP experiments." Journal of Instrumentation (2017) doi.org/10.1088/1748-0221/12/10/T10010 Vol. 12*

Phase alignment Logic

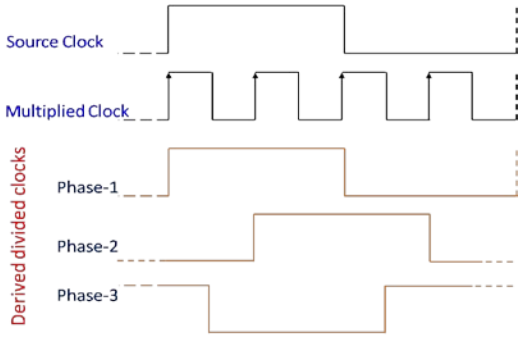
Auto-initialization alignment logic



Phase mismatch between the Word clock and Frame clock of GBT protocol

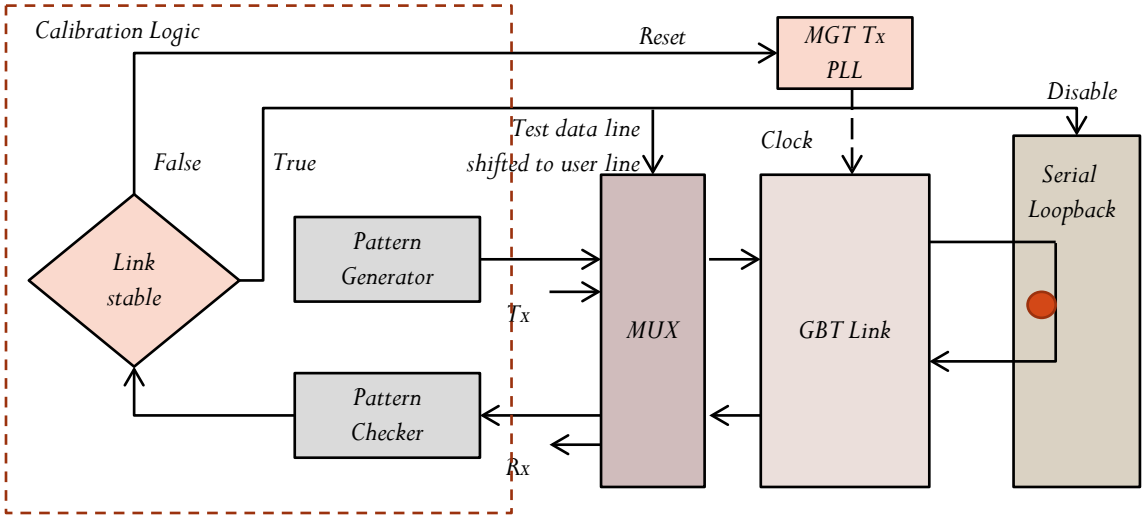


Distribution Scheme for clocking the Intel FPGA

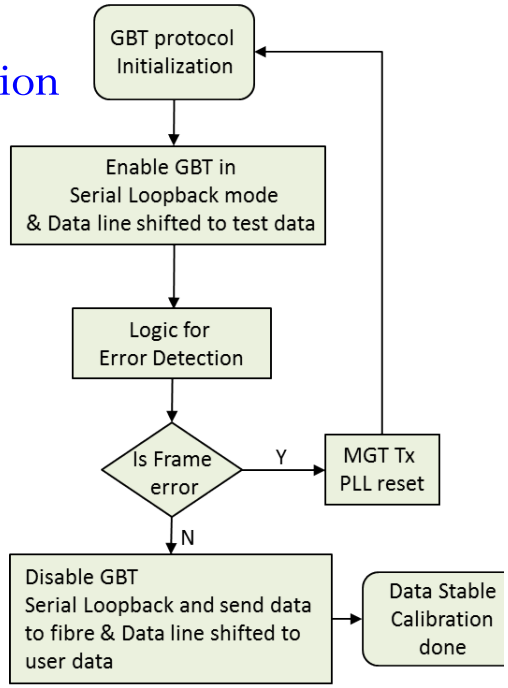


The rising edge of the derived divided clocks may be aligned to any of the triggering edges of the multiplied clock.

Phase alignment logic



State Machine for Phase Calibration



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

Logic for Error Detection

Publication: *Shuaib Ahmad Khan, et al. "Development of FPGA based phase alignment logic for the high speed protocol in HEP Experiments." Computer Physics Communication Volume 259, Feb 2021, doi.org/10.1016/j.cpc.2020.107649*

Multi-Gigabit Transceiver Optimization technique

What is the need!!!!!!

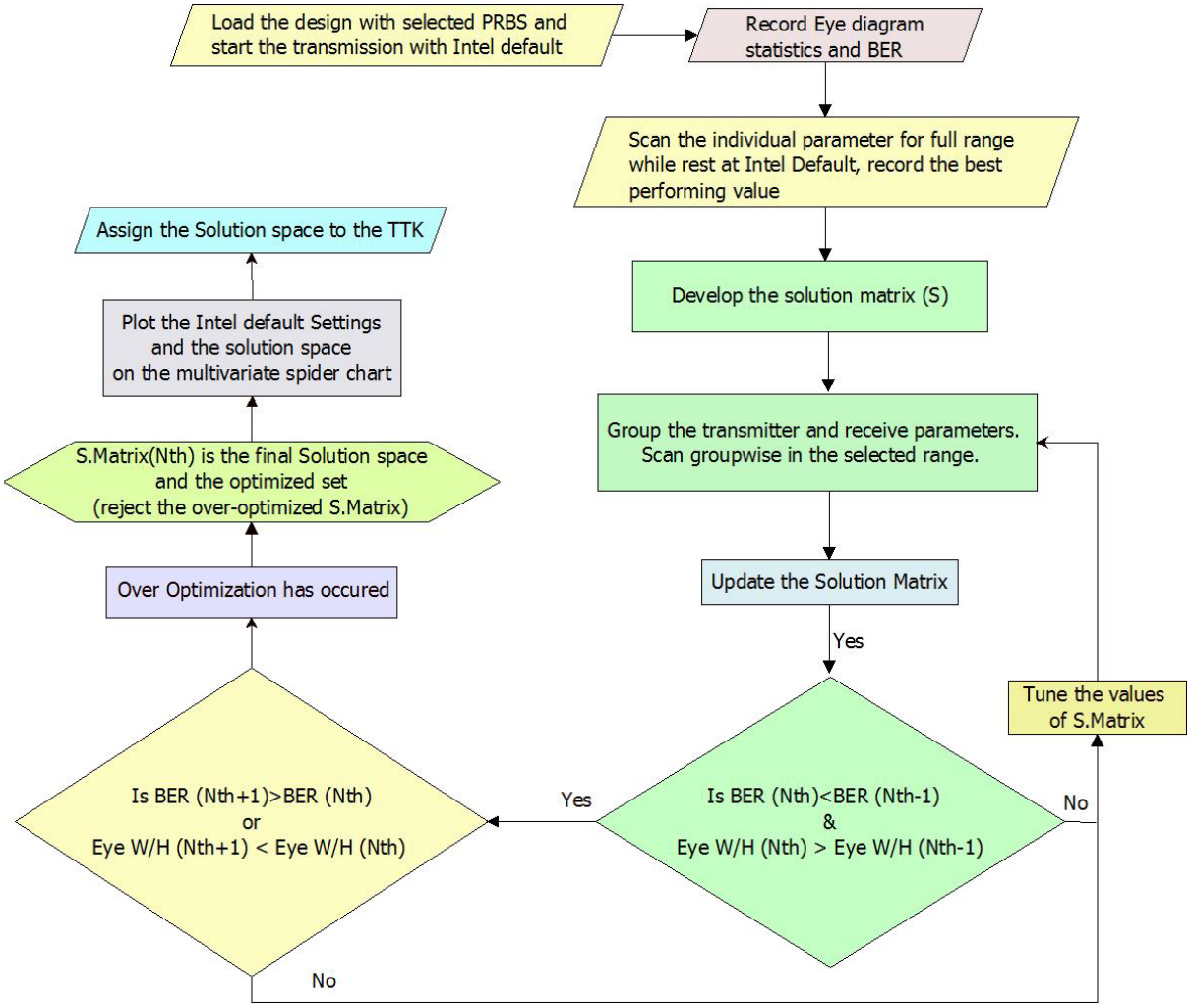
Transceiver parameter	Range of possible values	Number of iterations required
Transmitter Side		
VOD	0 to 31	32
Pre-emphasis 1st post-tap	-31 to 31	63
Pre-emphasis 1st pre-tap	-31 to 31	63
Pre-emphasis 2nd post-tap	-15 to 15	31
Pre-emphasis 1st post-tap	- 7 to 7	15
Receiver Side		
DC gain	0 to 4	5
Equalization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

32 x 63 x 63 x 31 x 15 x 5 x 16 x 8

$\sim 3.78 \times 10^8$ iterations

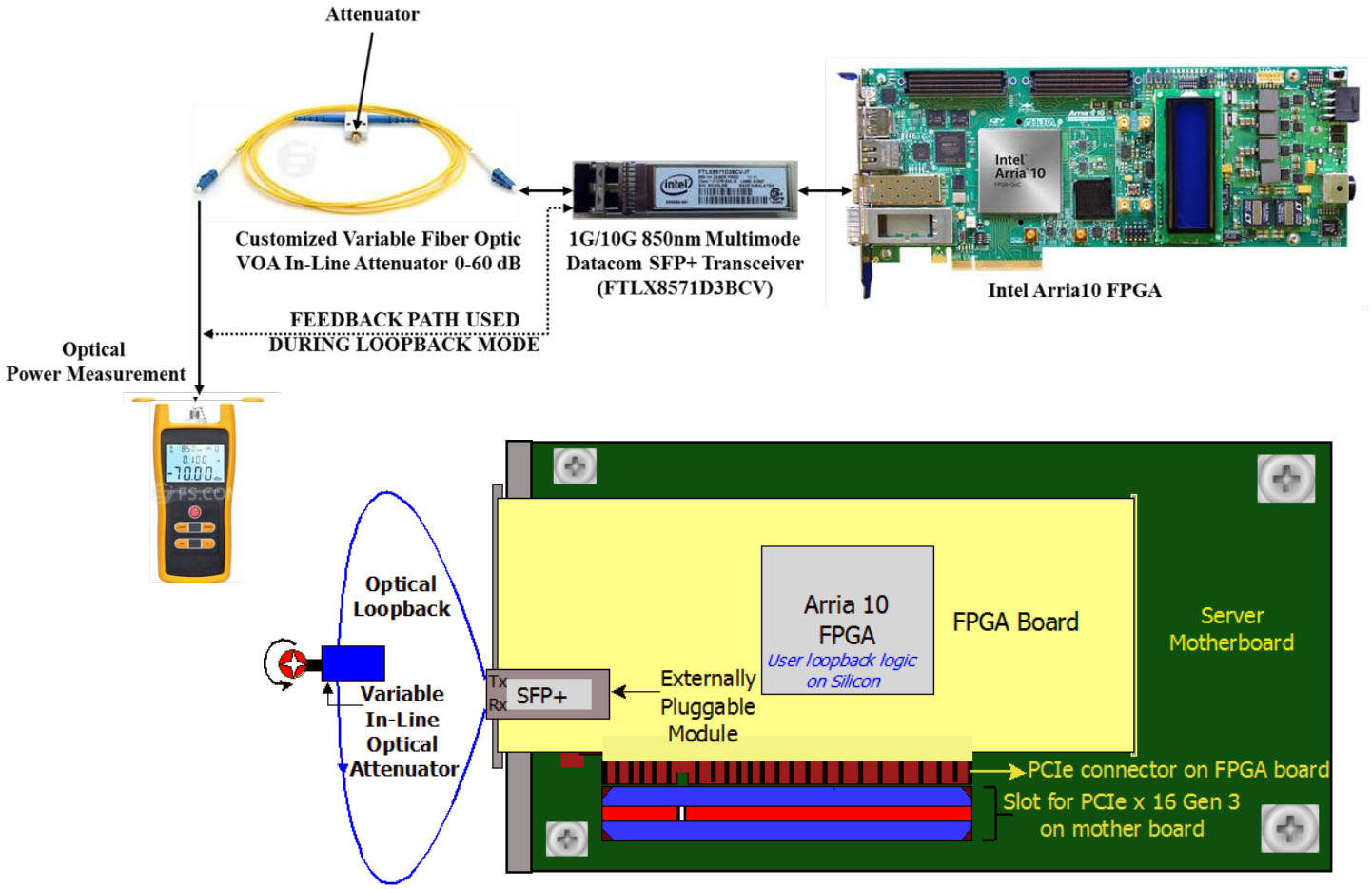
Multiple transceiver parameters with large span of operating range. Scan every combination is time consuming.

Transceiver Optimization Methodology



Publication: *Shuaib Ahmad Khan, et. al, "Optimization of multi-gigabit transceivers for high speed data communication links in HEP Experiments." NIMA: (2019) https://doi.org/10.1016/j.nima.2019.02.030 Volume 927, May 2019, Pages 14-23*

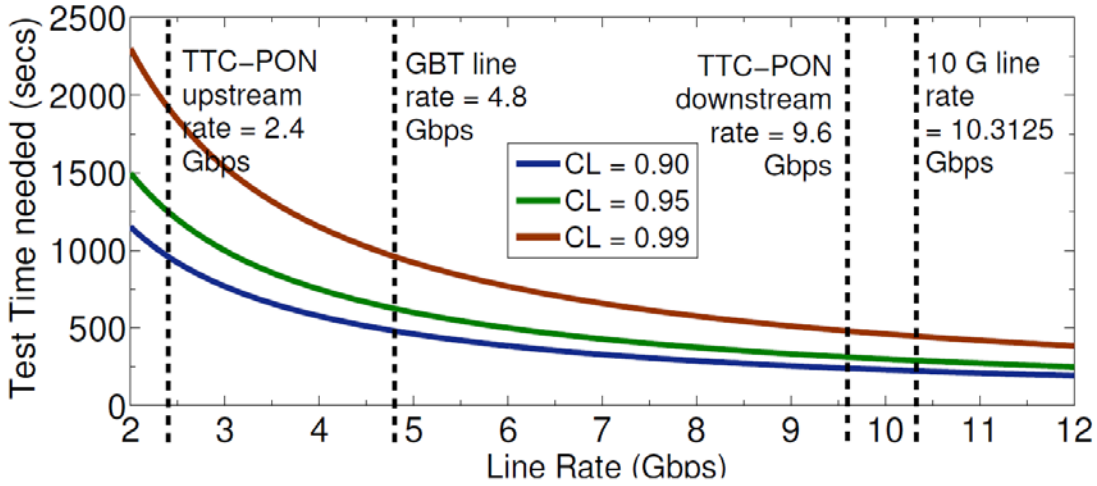
Transceiver tests and tuning Test Setup



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)

Time to achieve BER of 10^{-12} for the Line rate of GBT, TTC-PON and 10 Gbps optical links for different Confidence Level



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

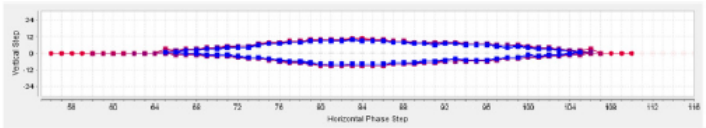
The minimum number of bits required to be tested for the BER measurement with a specific associated CL

Solution at N = 0

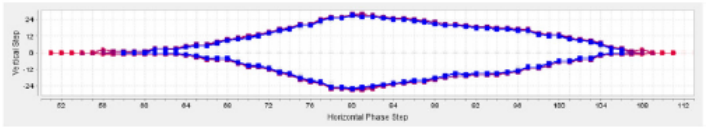
T is test time needed, R is the line rate

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

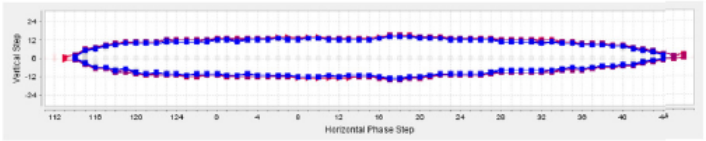
Eye diagram at the Intel-default and at the Optimized settings of transceiver deduced using the proposed technique for different line rates



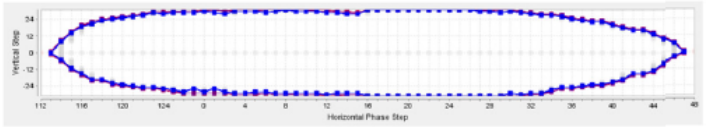
Vertical step(19)/Horizontal Phase step(41) for 10 Gbps at the Intel FPGA default settings



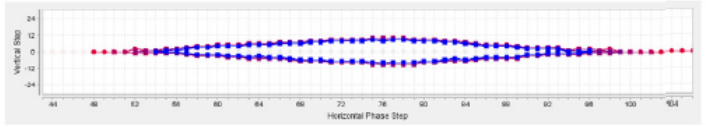
Vertical step(49)/Horizontal Phase step(54) for 10 Gbps at the Optimized FPGA settings



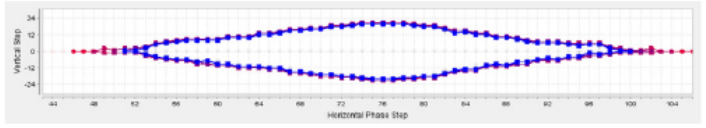
Vertical step(28)/Horizontal Phase step(59) for 4.8 Gbps at the Intel FPGA default settings



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

19/41

49/54

10 Gigabit

28/59

63/63

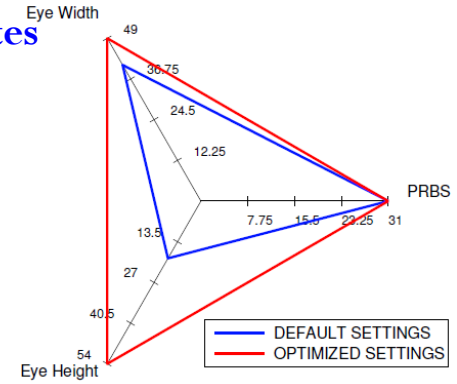
GBT

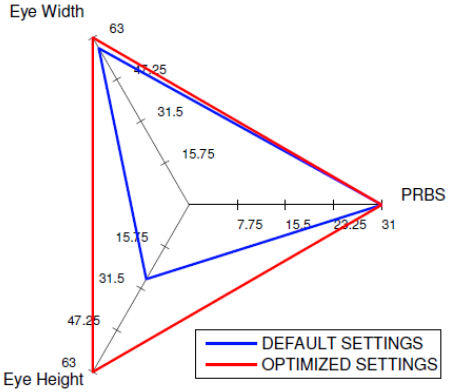
18/43

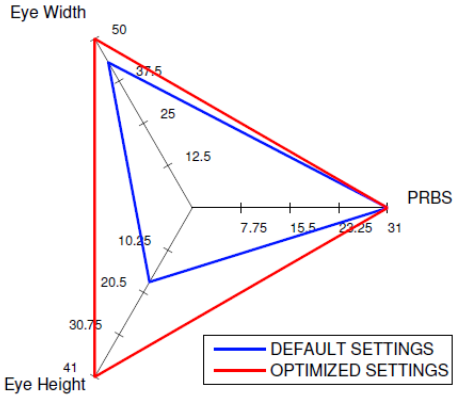
41/50

TTC-PON

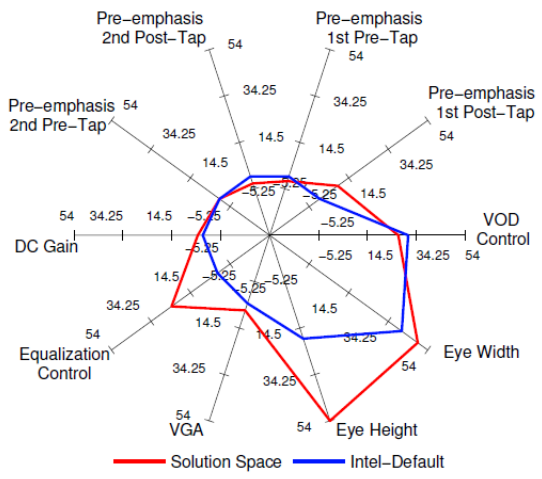
Spider plot for parameter comparison of Eye Diagram at PRBS31 for Default and the Optimized transceiver parameters settings



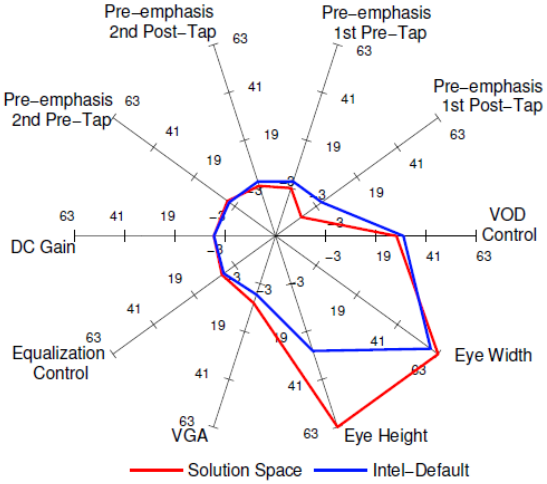




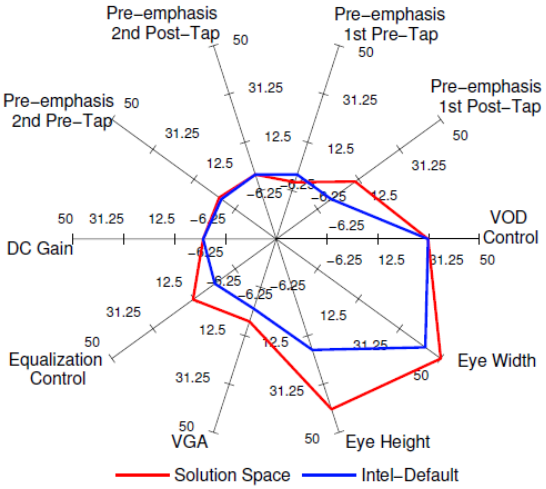
Multivariate spider plot for Solution space vs the Intel-default



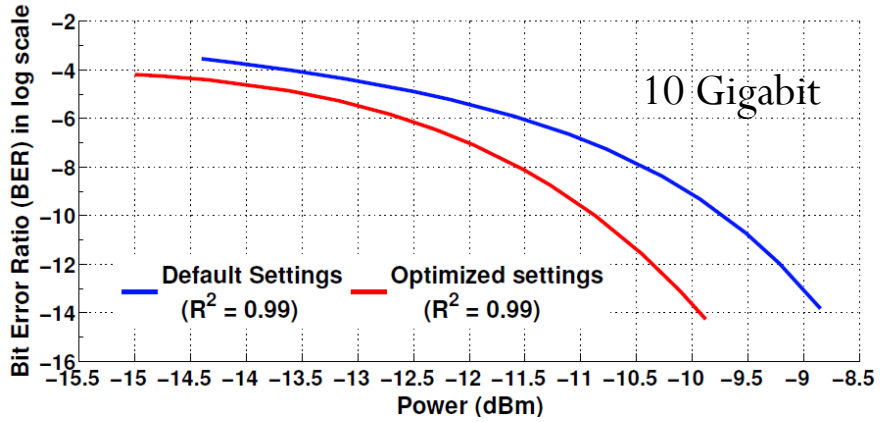
10Gbps



GBT Link

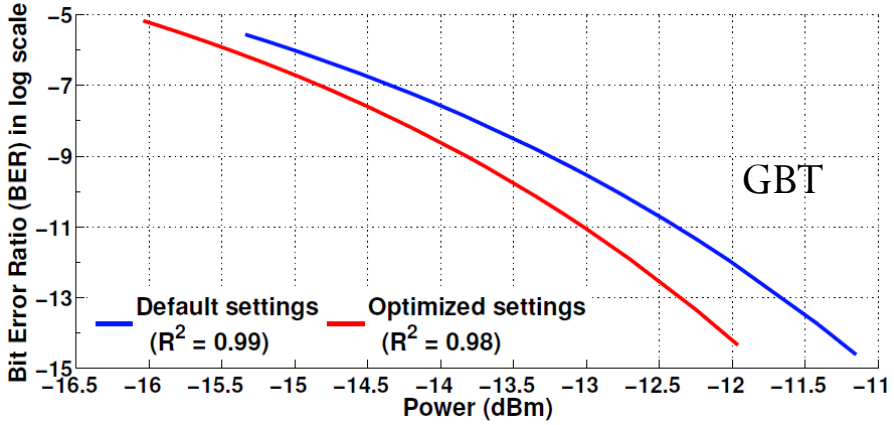


TTC-PON



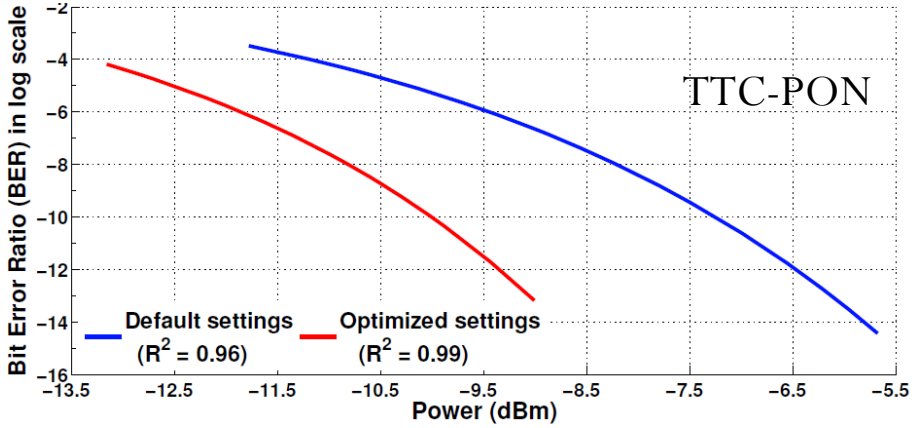
Link Speed (Protocol)	With default approach (dBm)	With optimization technique (dBm)
10 Gbps	-9.2	-10.35
4.8 Gbps	-11.9	-12.7
9.6 Gbps	-6.45	-9.3

Comparison of Optical power(dBm) to attain BER of 10^{-12}



Link Speed (Protocol)	With default parameters (dBm)	With optimization technique (dBm)
10 Gbps	-14.4	-15
4.8 Gbps	-15.34	-16.04
9.6 Gbps	-11.78	-13.2

Comparison of Optical power for CDR



Publication: *Shuaib Ahmad Khan, et. al, "Optimization of multi-gigabit transceivers for high speed data communication links in HEP Experiments." NIMA: (2019) https://doi.org/10.1016/j.nima.2019.02.030 Volume 927, May 2019, Pages 14-23*

Hardware complexities

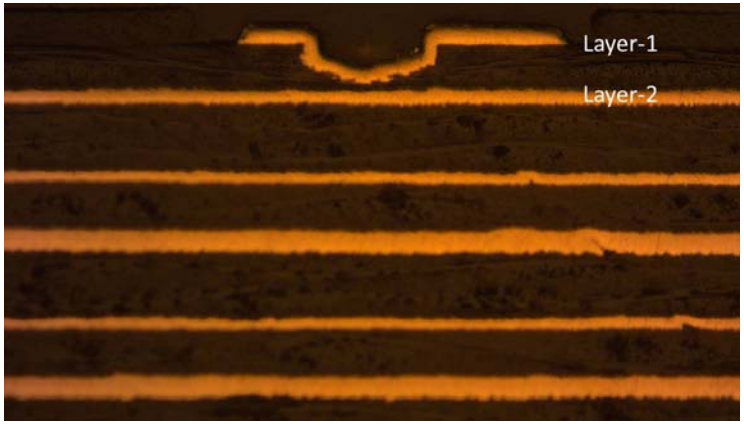
CRU Boards: Hardware Complexities

HDI PCB	More than 1750 components on the board
LAYER, DRILLS, VIAS	14 layer, laser drilling, blind, buried and stacked vias
OPTICAL MINIPODS	48 nos. high speed, 4.8 Gbps
PCI EXPRESS	16 Lane Gen x3
HIGH POWER REQUIREMENT	Supported by mezzanine cards
MATERIAL REQUIREMENTS	High TG, Low D_K material for high speed, low loss ISOLA 408 HR, TUC 872 LK
THICKNESS RESTRICTIONS	1.57 mm +/- 10%
FPGA CHIP with High user Inputs/outputs (ARRIA 10 Device specifications)	1932 pin BGA package (ARRIA 10 Device overview) with 768 I/O and 96 Transceivers, production grade silicon, 20-nm technology.

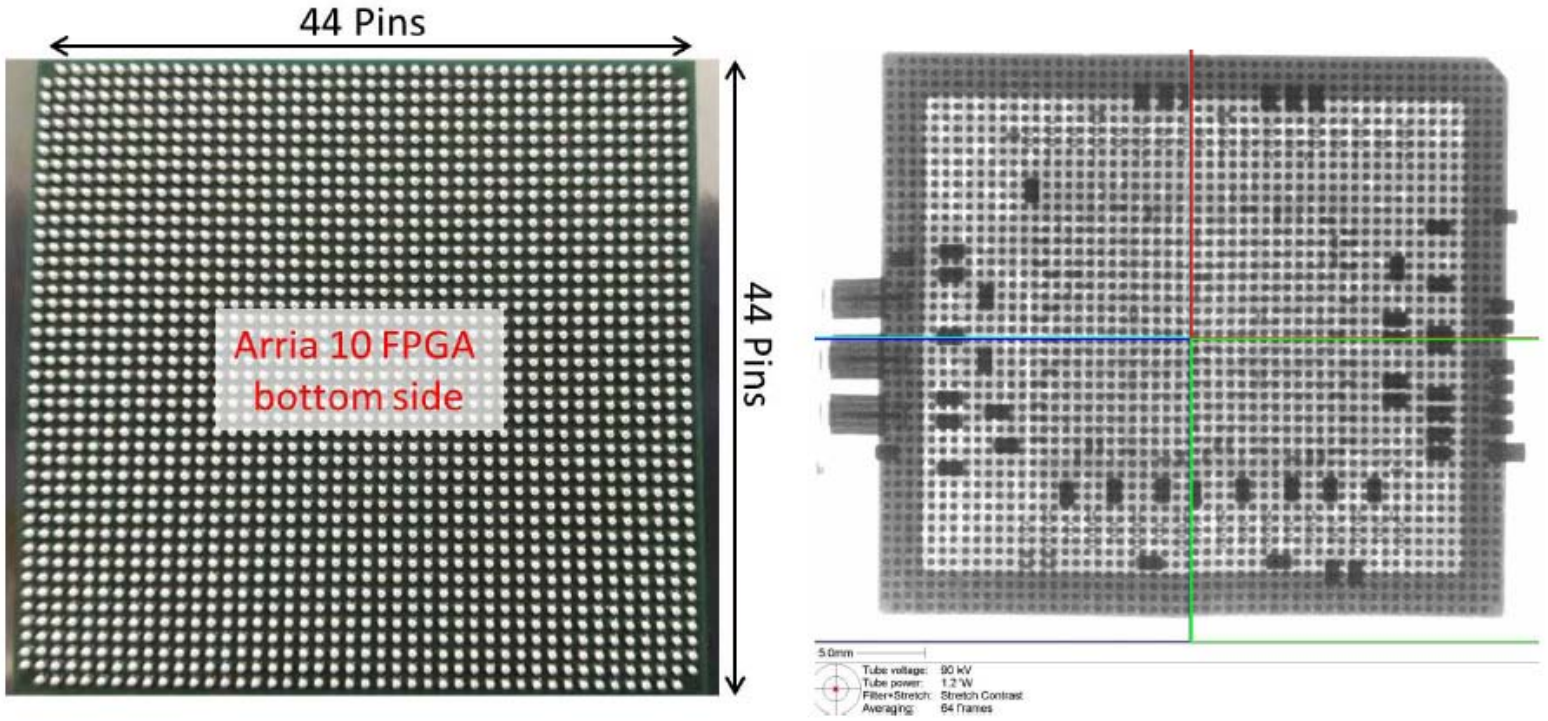
PCB: Thickness Measurement
1.64 mm (14 layer PCB)

PCIe40 Bare board

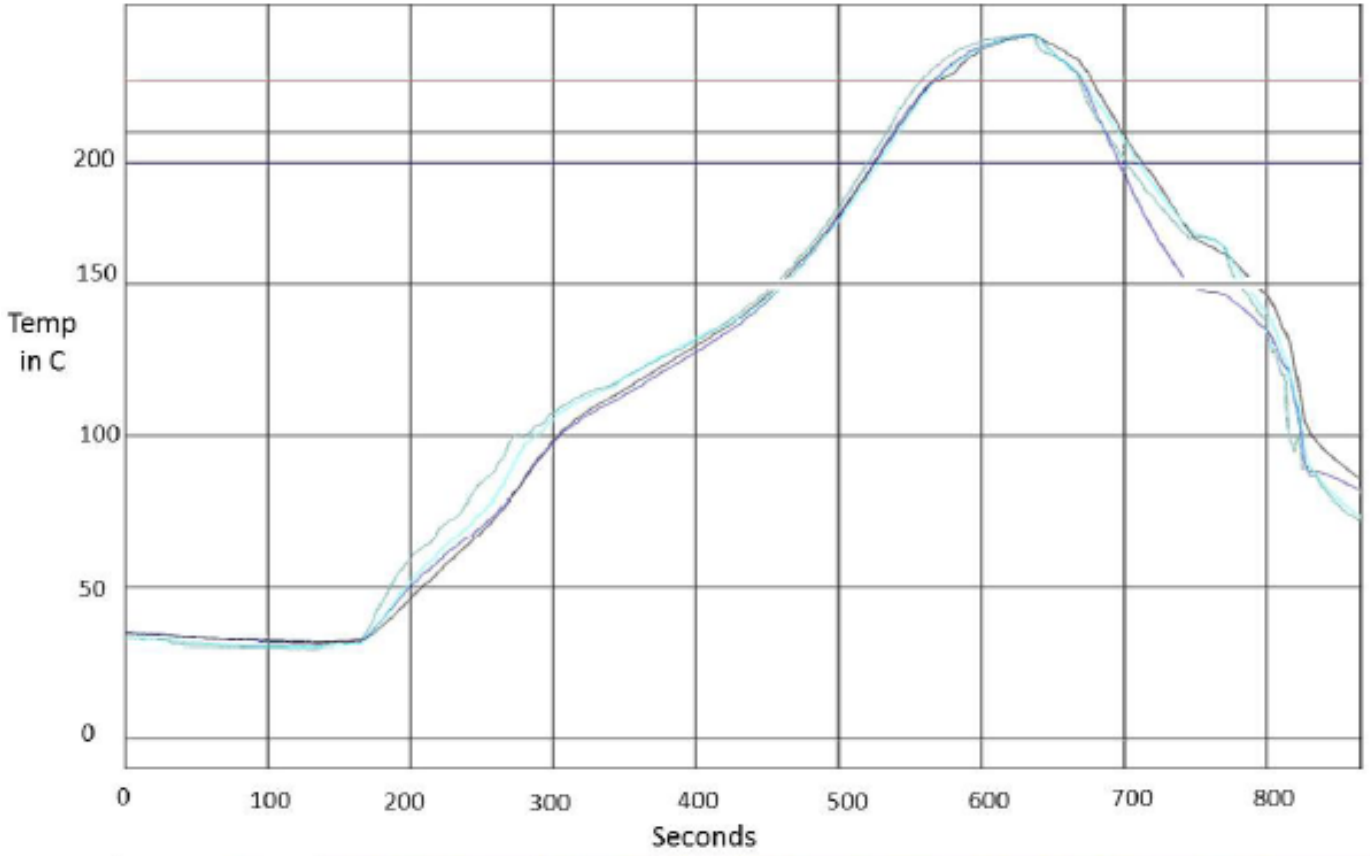
Assembly issues



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.



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



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

Temperature profiling

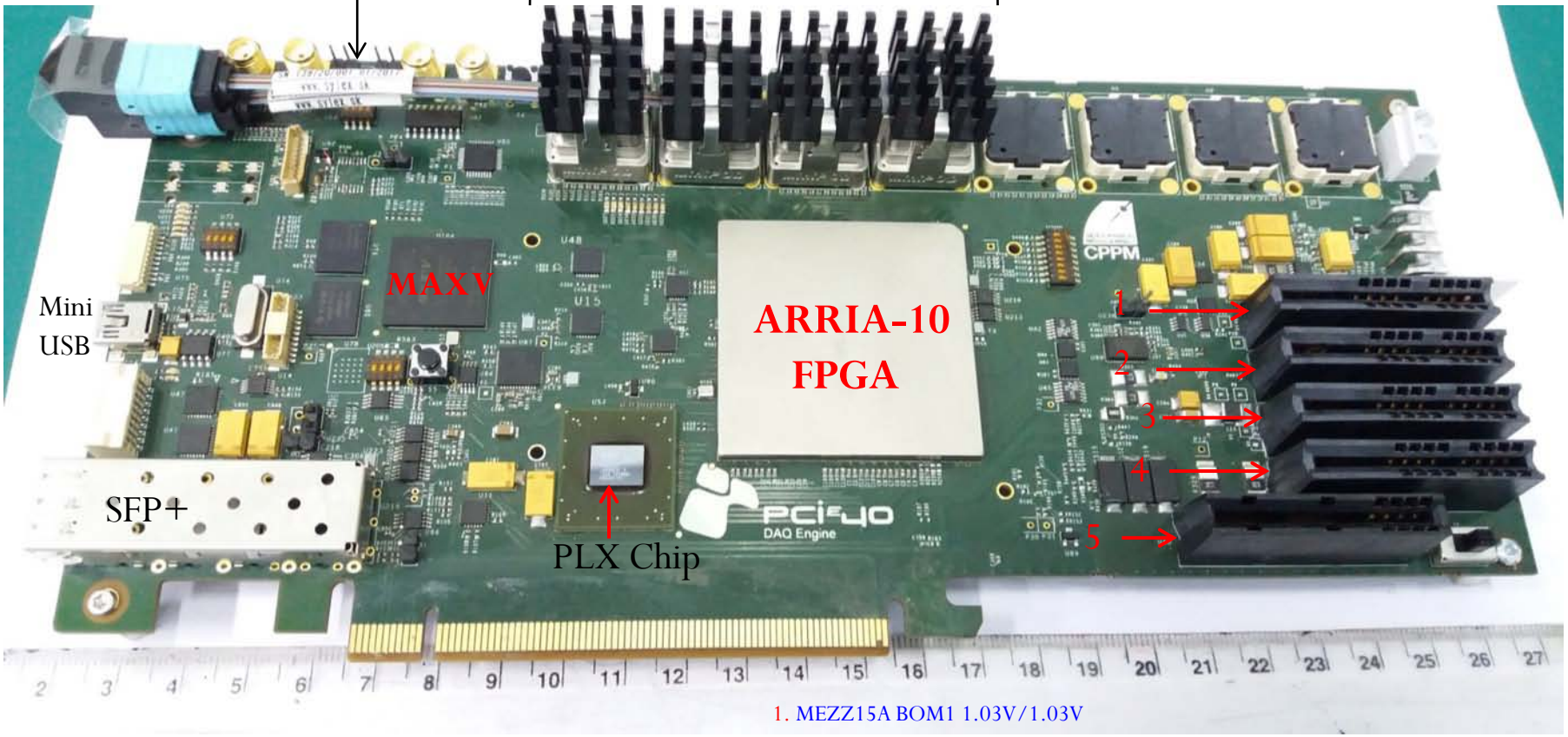
and

the Ramp-Spike method

First CRU card at VECC

2 Optical patch cords mounted –Syslex (custom)

2 Tx and 2 Rx minipods mounted



1. MEZZ15A BOM1 1.03V/1.03V
2. MEZZ15A BOM2 1.8V/2.5V
3. MEZZ15A BOM4 3.3V/3.24V
4. MEZZ15A BOM3 0.9V/0.9V
5. MEZZ60A 0.9V/0.95V


Thank you