

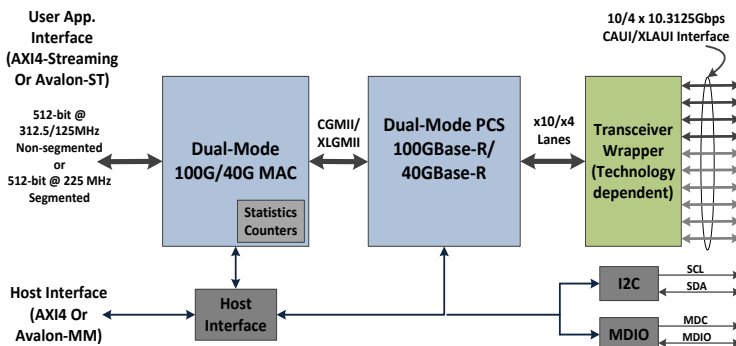
Dual-Mode 100G/40G Ethernet IP Solution

Product Brief (HTK-100G40G-ETH-320-FPGA)



The dual-mode 100Gbps/40Gbps Ethernet IP solution offers a fully integrated IEEE802.3ba compliant package for NIC (Network Interface Card) and Ethernet switching applications. As shown in the figure below, the 100G/40G Ethernet IP includes:

- 100Gbps/40Gbps dual-mode MAC core
- 100Gbps/40Gbps dual-mode PCS core
- Technology dependent transceiver wrapper
- Statistics counter block (for RMON and MIB)
- MDIO and I2C cores for optical module status and control



A complete reference design using a synthesizable L2 (MAC level) packet generator/checker is also included to facilitate quick integration of the Ethernet IP in a user design. A GUI application interacts with the reference design's hardware elements through a UART interface (PCIe option is also available). A basic Linux PCIe driver/API is also provided for memory mapped read/write access to the internal registers. See **Appendix A** for details.

MAC and PCS cores are designed with 320-bit data path operating at 312.5MHz.

As the transceiver wrapper is included with the Ethernet IP solution, the line side directly connects the 10.3125Gbps FPGA transceivers to the optical modules like QSFP+ (40Gbps only), CFP, CXP, 300Pin MSA, etc).

Ethernet IP solution implements two user (application) side interfaces. The register access port can either be a 32-bit AXI4 interface or a 32-bit Avalon-MM interface. IP solution provides a highly flexible 100Gbps traffic port interface options. Depending upon the application layer, user can select an AXI-4 streaming bus or an Avalon Streaming bus to interface with the MAC block. MAC interface bus is a 512-bit non-segmented bus that operates at 312.5MHz for 100Gbps mode and 125MHz for the 40Gbps mode. An interface wrapper is provided to support segmented operation at lower clock speeds.

100Gbps Ethernet IP supports advanced features like per-priority pause frames (compliant with 802.3bd specifications) to enable Converged Enhanced Ethernet (CEE) applications like data center bridging that employ IEEE 802.1Qbb Priority Flow Control (PFC) to pause traffic based on the priority levels.

Features Overview

- Enables the run-time selection of 100Gbps or 40Gbps Ethernet operation with a single FPGA image

MAC Core Features

- Implements the full 802.3 specification with preamble/SFD generation, frame padding generation, CRC generation and checking on transmit and receive respectively.
- Implements 802.3bd specification with ability to generate and recognize PFC pause frames
- Implements a 320-bit CGMII/XLGMII interface operating at 312.5 MHz for 100Gbps mode or 125 MHz for 40Gbps mode
- Implements Deficit Idle Count (DIC) mechanism to ensure maximum possible throughput at the transmit interface
- Implements logic for padding of frames on the transmit path if the size of frame is less than 64 bytes
- Implements fully automated XON and XOFF Pause Frame (802.3 Annex 31A) generation and termination providing flow control without user application intervention
- Pause frame generation additionally controllable by user application offering flexible traffic flow control
- Support for VLAN tagged frames according to IEEE 802.1Q
- Support any type of Ethernet Frames such as SNAP/LLC, Ethernet II/DIX or IP traffic
- Discards frames with mismatching destination address on receive (Except Broadcast and Multicast frames)
- Programmable Promiscuous mode to omit MAC destination address checking on receive EMAC
- Optional multicast address filtering with 64-bit HASH Filtering table providing imperfect filtering to reduce load on higher layers
- CRC-32 generation and checking at high speed using an efficient pipelined CRC calculation algorithm
- Implements logic for optional padding removal on RX path for NIC applications or forwarding of unmodified data to the user interface
- Discards runt frames (less than 64 Byte) at the core's reconciliation sublayer
- Implements logic for optional forwarding of the CRC field to user application interface
- Implements logic for optional forwarding of received pause frames to the user application interface
- Programmable frame maximum length providing support for any standard or proprietary frame length (e.g. 9K-Bytes Jumbo Frames)

Dual-Mode 100G/40G Ethernet IP Solution

Product Brief (HTK-100G40G-ETH-320-FPGA)



- Status signals available with each Frame on the user interface providing information such as frame length, VLAN frame type indication and error information.
- Implements programmable internal XLGMII/CGMII Loop-back
- Implements statistics indicators for frame traffic as well as errors (alignment, CRC, length) and pause frames
- Implements statistics and event signals providing support for 802.3 basic and mandatory managed objects as well as IETF Management Information Database (MIB) package (RFC 2665) and Remote Network Monitoring (RMON) required in SNMP environments
- Implements a streaming user application interface. The application interface is designed as a 512-bit non-segmented (start of a new frame on next 512-bit word) bus operating at 312.5MHz for 100Gbps mode and at 125MHz for 40Gbps mode.
- An interface wrapper is provided for applications that implement a segmented (start of new frame within same 512-bit word) bus. In segmented mode, the 512-bit bus operates at @ 225MHz for 100Gbps.
- Implements memory-mapped host controller interface for accessing the core's register file

PCS Core Features

- Implements 40G/100GBase-R PCS core compliant with IEEE 802.3ba Specifications
- Implements a 320-bit CGMII/XLGMII interface operating at 125MHz/312.5MHz for 40G/100G Ethernet
- Implements 64b/66b encoding/decoding for transmit and receive PCS
- Implements 40G/100G scrambling/descrambling using 802.3ba specified polynomial $1 + x^{39} + x^{58}$
- Implements Multi-Lane Distribution (MLD) across 20 or 4 Virtual Lanes (VLs) for 100Gbps or 40Gbps operations respectively
- Implements periodic insertion of Alignment Marker (AM) on the transmit path and deletion on the receive path
- Implements 66-bit block synchronization and Alignment Marker Lock machines as specified in 802.3ba specifications
- Implements skew compensation logic in order to realign all the virtual lanes and reassemble an aggregate 40G/100G stream (with all 64b/66b blocks in the correct order)
- Implements lane reordering to support reception of

any virtual lane on any physical lane.

- Implements BIP-8 insertion/checking per Virtual Lane on transmit/receive respectively.
- Implements Inter Packet Gap (IPG) Insertion/Deletion for Alignment marker compensation while maintaining a minimum of 1 byte IPG.
- Implements gear-box logic to convert 66-bit blocks to 20/40-bit for 40/100G PCS. The 20/40-bit interface operate at the transceiver reference clock.
- Implements programmable internal CGMII/XLGMII loop-back which directs traffic received from core's receive path back to transmit PCS.
- Implements Bit Error Rate (BER) monitor for monitoring excessive error ratio. In addition, the core implements various status and statistics required by the IEEE 802.3ba such as block synchronization status, AM lock status, lane deskew and lane reordering status and BIP-8 error counters per virtual lane.

See **Appendix B** for functional details of MAC and PCS.

Licensing and Maintenance

- ***True sign once licensing with NO yearly maintenance fees***
- Basic core licensing for a single vendor (either Xilinx or Altera) compiled (synthesized) binary
- Additional vendor license provided at only 50% cost of the base license. This allows for cost effective multi-vendor designs with identical user and control interfaces.
- Other licensing options include:
 - Vendor and device family agnostic source code (Verilog) license
 - A ***low cost board locked*** license for low budget prototyping (upgradeable to full license)

Contact and Sales Information

Phone: +1-301-528-2244

Email: info@mantaro.com

Resource Utilization

The 40G/100G Ethernet solution is currently supported on Altera's FPGAs only. The core utilization summary for the 100G/40G Ethernet solution is given in following table.

100G/40G Ethernet - Resource Usage for Altera Devices

<i>Device</i>	<i>User Interface Width</i>	<i>RMON and MDIO</i>	<i>COMB. ALUTs</i>	<i>Memory ALUTs</i>	<i>Registers</i>	<i>Memory M9K</i>
Stratix-IV (-2C Speed)	320-bit	Yes	47,426	2682	56,377	212
		No	46,422	2682	54,574	212
	512-bit (Seg.)	Yes	51,527	2682	60,014	244
		No	50,513	2682	58,211	244

Deliverables

- Compiled synthesizable binaries or encrypted RTL for the MAC and PCS cores
- Source code RTL (Verilog) for I2C, MDIO, RMON and Register-File blocks
- Self checking behavioral models and test benches for simulation
- Constraint files and synthesis scripts for design compilation
- A complete PCIe/UART host interface based reference design with:
 - Top level wrapper (source files, Verilog) for user specific customizations
 - Source files (Verilog) for the PCIe application layer
 - Binaries for a basic L2 packet generator and checker
 - PCIe driver/API (source files, C) for Linux
 - UART and command interpreter blocks with the optional UART host interface
 - GUI application (Linux only for PCIe, Linux and Windows for UART) for interfacing to the reference design
- Design guide(s) and user manuals
- USA based technical support by developers

A. Reference Design Details

A.1 Overview

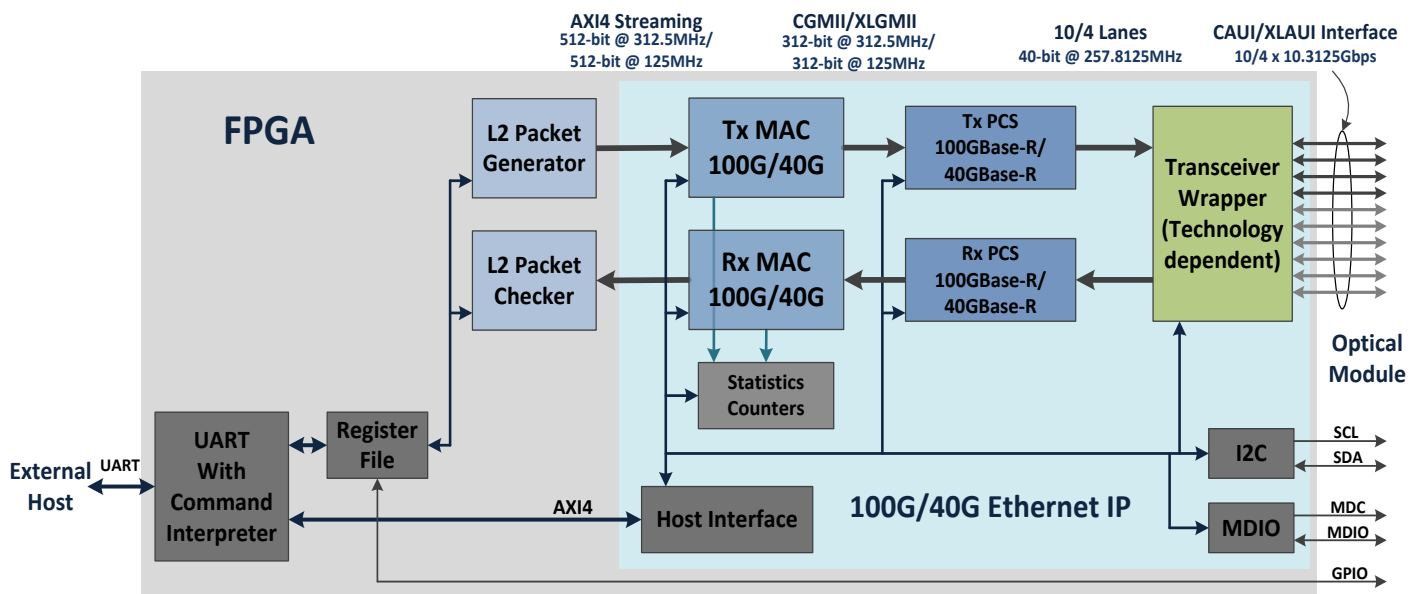
A 100G/40G reference design is included as part of the IP deliverable to facilitate quick L1 and L2 layer testing and verification of the 100Gbps and 40Gbps Ethernet operation on target platform. The capability to run the L1 PRBS pattern and configure each transceiver independently can be used for a fast module bring-up in the lab and can also be used for factory diagnostics.

The UART (normally through an onboard USB-to-UART converter chip) based 100G/40G Ethernet reference design can be seamlessly ported to various COTS FPGA networking and evaluation modules (see section for the list of verified modules). A GUI application controls the register read/writes to the FPGA through a UART core with integrated command interpreter. Both Linux and Windows platforms are supported for the UART based interface control.

This reference design can also be used on custom embedded design where the FPGA connects to the host processor via a PCIe interface. For the PCIe control interface, GUI application is hosted on a Linux platform (as PCIe driver/API is provided for Linux OS only).

A.2 Functional Description

Following figure shows the connectivity and the elements of the 100G/40G Ethernet IP reference design. Usually the UART interface from the FPGA connects to an external (can be on the same module as well) USB-UART converter. A Linux or Windows host (through a USB port) running the GUI application is used to configure and control the 100G/40G Ethernet. I2C, MDIO and GPIO interfaces included in the reference design can be used to control any optical module on the target platform including the 300Pin MSA (I2C), CXP (I2C) and CFP (MDIO) MSA compliant modules.

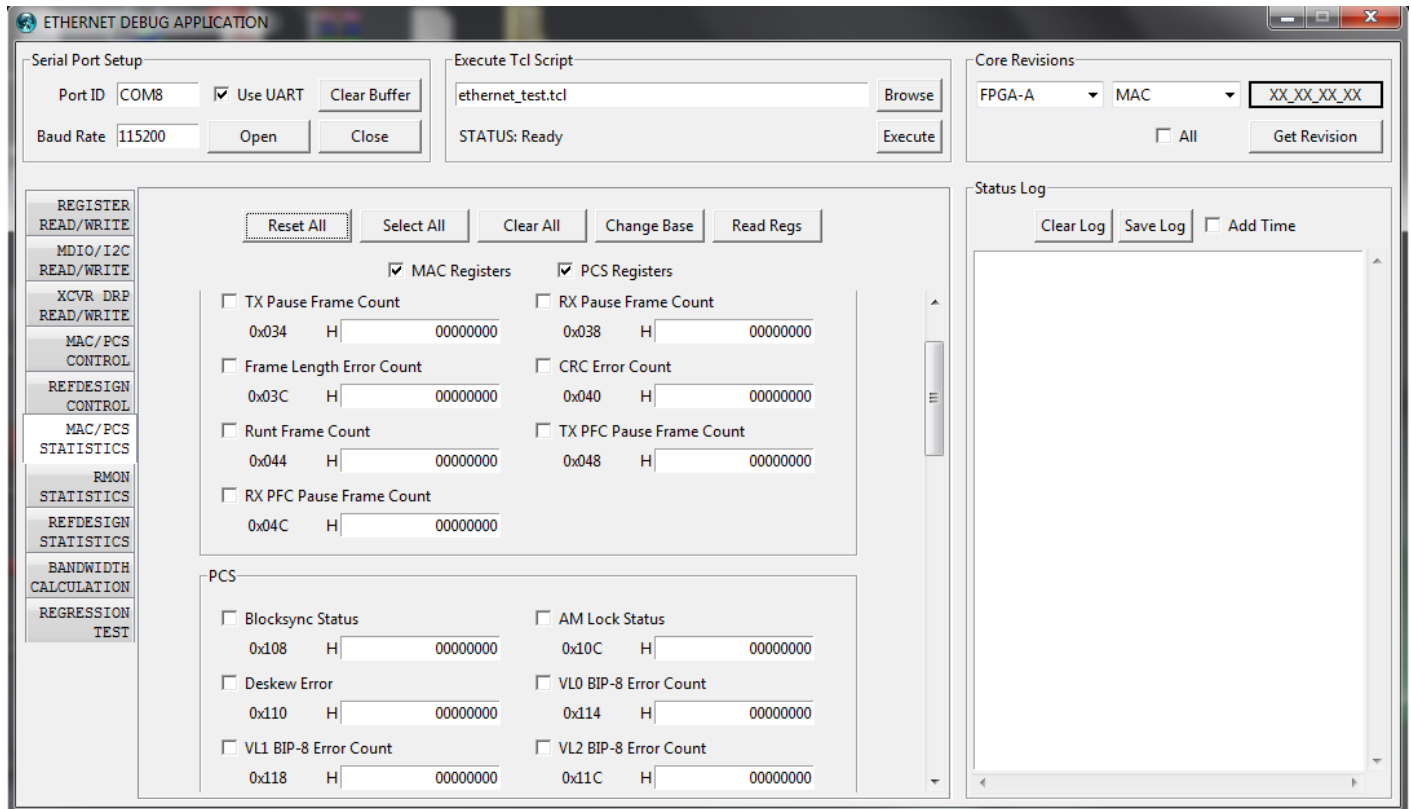


For L1 (physical layer verification and testing) GUI application provides an interface to independently control and configure 10/4 10.3125Gbps transceivers used for 100G/40G Ethernet transport. User can configure the transceivers to run various PRBS pattern and configure various transceivers parameters like transmit voltage, transmit pre-emphasis, receive equalization and receive gain.

For L2 testing, GUI application uses the 100Gbps capable packet generator/checker inside the FPGA to generate and check MAC frames up to full line rate. The packet generator supports a basic rate control mechanism to control the packet/data rate on the interface. The generator can be configured for fixed size as well as pseudo random packet size packet transmission. An incrementing counter is used as payload for the MAC frames. The checker on the receive side verifies the payload of receive MAC frames and reports error in the payload.

A comprehensive set of transmit and receive counters in the MAC core provide a detailed view of the packet statistics including various error types.

Following is a snapshot for the GUI application for the L2 packet test results screen.



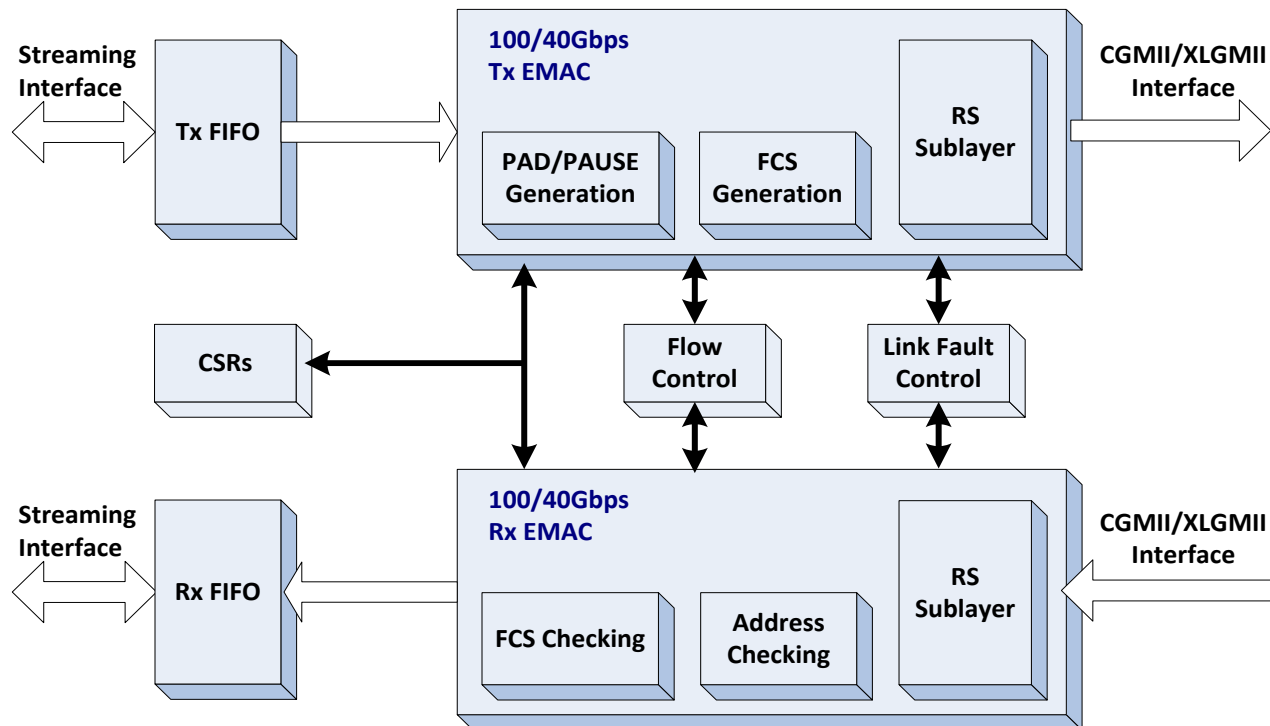
A.3 Validated/Ported Module List

1. HTG-S4Gx-100GIG module; Altera Stratix IV-GT FPGA, Evaluation Module with 2 AirMax Interlaken interfaces and FMC interfaces (http://www.mantaro.com/products/development_platforms/Altera_S4GXGT_100G.htm)

B. MAC and PCS Functional Details

B.1 MAC Functional Overview

The following figure shows the architecture for the 100/40 Gigabit Ethernet MAC. The MAC is basically the 320-bit data path 100Gbps MAC (operating at 312.5MHz) that runs at slower clock speed of 125MHz for the 40Gbps Ethernet mode.



B.1.1 Reconciliation Sublayer (RS) Operation

The RS layer is responsible to map the data to/from the MAC sublayer to the XLGMII/CGMII interface. The RS layer provides a 320-bit interface at rising edge of XLGMII/CGMII clock. The data is organized into forty 8-bit lanes with a control bit available for each lane.

B.1.2 MAC Sublayer Operation

The MAC sublayer is responsible to perform transmit and receive media access control (MAC) operations. The transmit MAC block transmits frames from a user application interface to the reconciliation sublayer, which then transmits these frames to the XLGMII/CGMII physical interface. The receive MAC block receives Ethernet frames from the reconciliation sublayer, validates the Ethernet frame and transfers this frame to the user application interface. The following description defines the various functions performed by transmit and receive Ethernet MAC engines.

Transmit Ethernet MAC

The transmit Ethernet MAC performs the following main functions:

- Accepts data including Destination Address, Source Address and length field from the MAC client.

Dual-Mode 100G/40G Ethernet IP Solution

Product Brief (HTK-100G40G-ETH-320-FPGA)



- Appends preamble and SFD to the Ethernet frames.
- Pads the incoming frames from the user application interface to minimum frame size (64 Byte) whenever the frame size is less than 64 Byte.
- Calculates and Appends proper FCS (CRC-32) value to outgoing frames and verifies full octet boundary alignment.
- Delays transmission of frame data for specified inter-frame gap period.
- Controls Inter-frame gap timing by maintaining a Deficit Idle Count value between 0 - 7
- Inserts start and terminate control characters before frame transmission.
- Inserts Idle control characters between frames (Inter-frame gap) or when there are no frames available for transmission.
- Manages local device flow control by generating PAUSE control frames.
- Manages Remote device congestion by transiting to HALT state for a specified time quanta.

Receive Ethernet MAC

The receive Ethernet MAC performs the following main functions:

- Receives a frame from the RS sub layer via a 320-bit data bus.
- Presents to the MAC client sublayer frames that are either frames with group address or directly addressed to the local station (Address recognition).
- Filters Multi-cast frames using hash filtering algorithm.
- Discards all frames not addressed to the receiving station when promiscuous mode is disabled.
- Accepts all frames destined to the EMAC if promiscuous mode is enabled.
- Checks incoming frames for transmission errors by way of FCS and verifies octet boundary alignment.
- Discards received transmissions that are less than a minimum length (64 bytes) at the core's reconciliation sublayer
- Truncates frames with length greater than maximum frame length when Jumbo frames are not allowed to pass through.
- Optionally forwards pause frames to user application.

B.1.3 MAC Flow Control Operation

The MAC flow control block is responsible to maintain a proper flow of Ethernet frames through transmit and receive engines. It performs the following main functions:

- Prevents the receive EMAC FIFO congestion by sending pause control frames.
- Prevents the remote device congestion by responding to pause frames and going into idle state for specified number of slot times.

Automatic flow control is only available during the non-PFC (legacy single priority flow control) mode of operation. For PFC mode, user layer is responsible for managing the flow control operation.

B.1.4 Management Interface

The 100G EMAC core provides a set of signals which can be used to implement the statistics required in IEEE 802.3 basic, mandatory and recommended Management information packages. In addition the MAC core provides signals to generate the applicable objects of the Management Information Database (MIB, MIB II) according to IETF RFC2665.

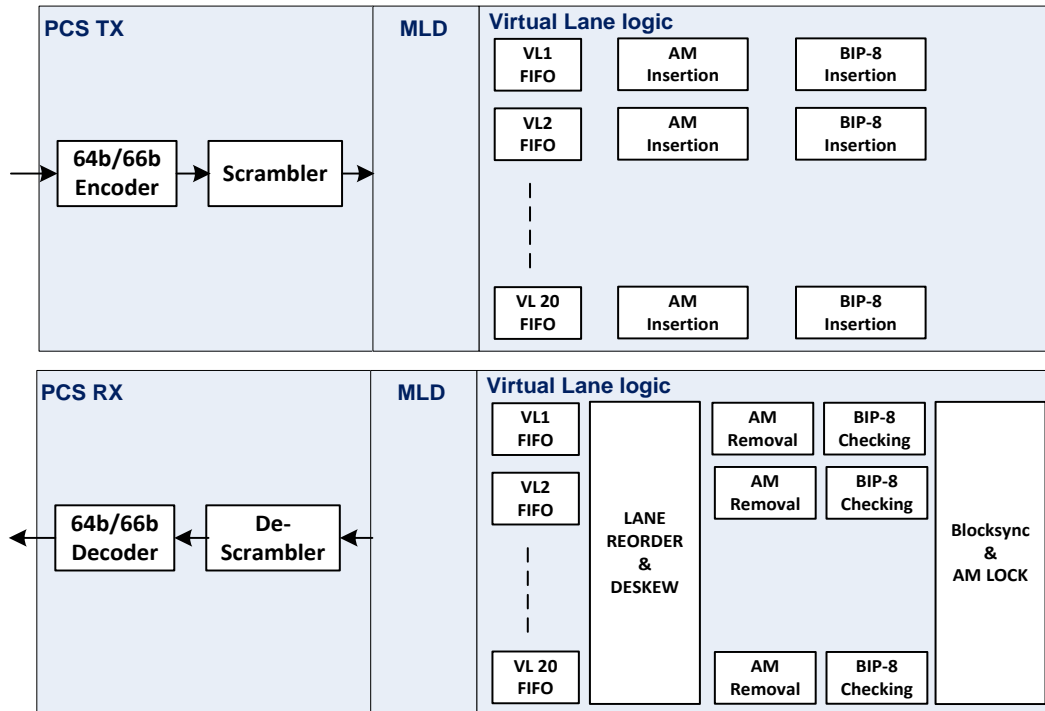
Dual-Mode 100G/40G Ethernet IP Solution

Product Brief (HTK-100G40G-ETH-320-FPGA)



B.2 100Gbase-R PCS Functional Overview

The following figure shows the architecture for the 100/40 Gigabit Ethernet PCS (100GBase-R/40GBase-R).



B.2.1 Transmit Physical Coding Sublayer

The transmit PCS receives 320-bit CGMII/XLGMII data from MAC Reconciliation Sublayer (RS) at 312.5MHz or 125MHz or for 100Gbps or 40Gbps data rates respectively. The received data from MAC is encoded into a continuous stream of 64B/66B blocks and then scrambled using a self-synchronizing scrambler. The synchronization headers of the 66-bit blocks allows establishment of block boundaries. The transmission code and scrambler ensures adequate transition density on any of the electrical lanes which in turn makes clock recovery possible at the receiver. The scrambled data is then transferred to the Multi-Lane Distribution Block (MLD). The MLD distributes the 66-bit blocks to 20 or 4 Virtual Lanes (VLs) on a round-robin basis for 100GBase-R PCS or 40GBase-R PCS. An alignment block is periodically added on each Virtual Lane in order to deskew and align all the Virtual Lanes at the receiving end. BIP-8 (Bit Interleaved Parity) is added per PCS lane to detect or isolate infrequent error events. Finally, 66:20 or 66:40 gear-boxes are used to transfer data streams to the integrated SERDES macros which transmit 40G/100Gbps data streams using 10/20 or 4 or high speed serial lanes for 100G or 40G PCS Sublayer Operation.

B.2.2 Receive Physical Coding Sublayer

The receive PCS is responsible for receiving data streams from 10/20 or 4 physical lanes and transferring 40G/100G aggregate stream to CGMII/XLGMII interface. The receiver de-multiplexes data received from 10/20 or 4 electrical lanes to 20 or 4 PCS Virtual Lanes. 66-bit block synchronization is performed on each Virtual Lane on the data received from 20:66 or 40:66 gear-boxes for 100Gbps or 40G mode operation respectively. Once 66-bit block synchronization is achieved on each of the virtual lanes, inter-lane alignment is established. This is accomplished by first obtaining Alignment Marker (AM) lock on each virtual lane and then removing any lane to lane skew to rebuild the 40G/100G aggregate stream with all 66-bit block in the correct position. Lane-reordering is then performed to support reception of any physical lane to any virtual lane. The data is then descrambled and 64B/66B decoded and 320-bit CGMII/XLGMII data is sent to the MAC RS layer. In addition, periodic idle deletion is performed for Alignment marker compensation.