WP413 (v1.0) March 1, 2012

100G Dual Gearbox: Improving Port Density on Line Cards in Core Network Equipment

By: Harpinder S. Matharu

Insatiable demand for bandwidth, fueled by the rapid growth of multimedia content and the rise of cloud computing, calls for ever-faster network infrastructure powered by high-capacity network equipment. While 10G ports in the network are becoming mainstream, a shift to 100G ports in the core and aggregation nodes has become important to achieve scale, network simplicity, and overall cost reduction in running the network. FPGA technology has played a key role in increasing the speed of every generation of network equipment.

At this inflection point, a leading-edge, power-efficient, and higher density 28 nm FPGA technology from Xilinx is ready to steer the industry in clearing the initial hurdles encountered when attempting to migrate line cards and transmission equipment to perform seamlessly at 100G. This white paper focuses on a dualport 100G gearbox implementation on a Virtex®-7 FPGA that facilitates this new era of network performance.

Introduction

Perpetual growth in user demand for data and increasing reliance on the cloud computing for business operations is pushing the adoption of high-speed 100G ports and high port density in core, aggregation, data center routers, and transmission equipment. The IEEE Std 802.3ba 100G Ethernet Standard [Ref 1], ratified in June of 2010 and ITU-T defined OTN line rate of OTU4 for 100G [Ref 2], is guiding the industry to meet this growing market demand. However, 100G port deployments have ramped slowly due to the cost of high-speed optics modules. While high component prices continue to be a major adoption hurdle, network operational simplicity, scalability, and efficiencies are gradually building the ecosystem and supply chain.

These network operational benefits stem from routing efficiencies inherent in handling 100G flows as opposed to multiple 10G flows. A rapidly growing industry segment and supporting market trends are needed to create the necessary momentum to time the adoption. In this context, the cloud is having a serious influence in pushing 100G technology past the initial component price hurdles. The new wave of virtualized and distributed data center equipment is likely to act as the catalyst in building economies of scale and accelerating growth of supporting technology at the right price points.

To meet demand, the ability to scale up to support high-density and high-speed ports in the aggregation routers and transmission gear is critical. The size, form factor, and power consumption of optical transceivers are at the heart of this challenge. Currently, only up to four 100G form-factor optics modules can be mounted on the faceplate of the line cards. This limits line cards in core network equipment to up to four fiber connections on the line side, thereby capping maximum data capacity at 400 Gb/s bandwidth.

CFP MSA [Ref 3] [Ref 4] optics modules are soon expected to be replaced by power- and space-efficient CFP2 and CFP4 modules, improving port density and data bandwidth on the line cards. For the same space and almost equivalent power of four CFP connectors, a line card can support up to eight 100G CFP2 (up to 800 Gb/s data throughput) or up to sixteen 100G CFP4 connectors (up to 1,600 Gb/s data throughput).

The CFP2 module uses 10X 10/11G or 4X 25/28G serial links and CFP4 module uses 4X 25/28G serial links at the system side interface to gain density and power advantages. This necessitates a need for a 100G Gearbox device on line cards to connect existing 10X 10/11G Ethernet MAC or OTL4.10 supporting OTN ASICs and Framers to 4X 25/28G CFP2 and CFP4 optics modules.

A gearbox is a logic function or a device that maps data between an input and an output, where the input and output data path lane widths and line rates are not evenly divisible. It does this by performing multiplexing, de-multiplexing, and shift operations on the received data stream.

100G Optics Modules

Effectively upgrading networks to handle exponential growth in data usage is critical. This requires power and size/port density improvements in the high-speed optics modules while keeping the costs low. As mentioned in the Introduction, bandwidth is currently limited to 400 Gb/s (limited by the number of optical modules that can be placed on a front plate of a line card). In Figure 1, the CFP MSA optics module for 100G Ethernet LR4 (10 km) and ER4 (40 km) use four 25G lasers to drive four wavelength channels to achieve an overall throughput of 100G.

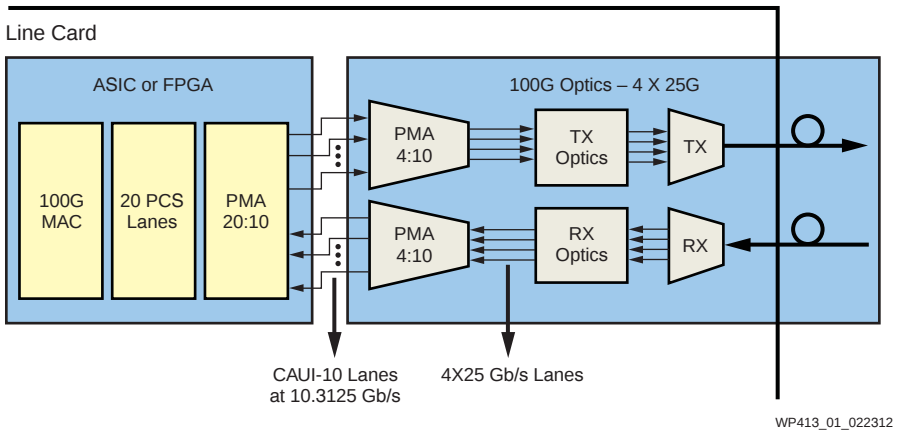

Figure 1: **100G Optical Module — Key Architectural Blocks**

The use of four lasers to drive four wavelengths in a CFP MSA optical module reduces the component count and complexity when compared to needing ten lasers to drive ten wavelengths in 10X 10 MSA optical modules (2 km and 10 km) [Ref 5]. The CFP MSA optical modules use 10X 10/11G serial links to connect to 100G Ethernet MAC or OTN ASIC and OTN Framers. Internally, CFP modules use a PMA (10:4) gearbox block to convert 10X 10/11g serial links to 4X 25/28G serial links that can drive 25/28G Baud lasers.

Implementing a Programmable 100G Gearbox in a Virtex-7 FPGA

As a first step to lower power and improve port density, the PMA (10:4) gearbox block within the optics module can be removed to support a direct four-lane 25/28 Gb/s interface. This helps lower pin count, space, power, and cost. The CFP2 optical module (under the purview of CFP MSA) is being defined to use a 10X 10/11G or 4X 25/28G system-side interface to the optical modules. Transition to smaller CFP2 optical modules with 4X 25/28G system side interface permits up to eight 100G optical modules on the line card, thereby increasing the total throughput per line card to 800 Gb/s.

External multiport 100G gearbox components are needed on the line card to connect a 10X 10/11G serial link based 100G Ethernet MAC ASIC or OTN Framer to a CFP2 module using a 4X 25/28G system side interface. In the case of an OTN line card, the gearbox connects OTL4.10 interfaces to OTL4.4 system-side interfaces. Programmable FPGA devices with 28 Gb/s-capable SerDes can be used to implement a dualport gearbox with enhanced test and debug functionality. Figure 2 shows how a Virtex-7 HT FPGA can be used to build a programmable and extendable dualport 100G gearbox device.

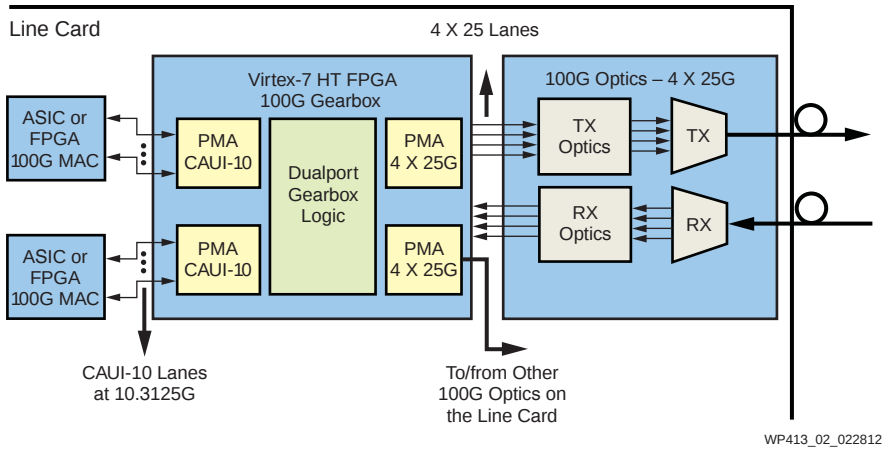

Figure 2: **100G Optical Module (CFP2) — CAUI-4 System Side Interface**

100G Gearbox Operation

To improve 100G port density on the line cards, multiport gearbox devices are needed to connect existing ASIC/FPGA/ASSP-based 100G Ethernet MAC or OTN functions. Based on the component functionality and layout of these components on a line card, dual or quad gearbox devices might provide a better solution in terms of optimal connectivity and lower cost in comparison to using a single port gearbox device. A combination of dual and/or quad gearbox devices can be used on a line card to connect to up to eight 4X 25/28G CFP2 optics modules or sixteen 4X 25/28G CFP4 optics modules. The MAC/NPU/Framer functions can natively support the appropriate interface for direct connectivity to the optical modules; however, this can take several years to come to fruition. Also, given the possibility of using several different interfaces with optics such as 4X 25G serial link interface, CPPI, OTL 4.4, or SFI-S, it might well be economical and expedient to keep the MAC/NPU/Framer functions independent of the optics interface.

For a MAC/NPU/Framer function supporting a 10X 10/11G serial link interface, the gearbox device translates a 10 lane 10/11G interface to a 4 lane 25/28G interface in order to connect to a 4X 25/28G CFP2 or CFP4 optics module. For example, in a 100G Ethernet MAC, the gearbox device does not need to implement the full PCS sub-layer (as defined by IEEE Std 802.3ba) because the translation from CAUI to the 4X 25G serial link interface can be done without performing scramble/unscramble and encode/decode functions on the twenty PCS lanes. To understand this better, a quick review of 100GBASE-R PCS is helpful.

100GBASE-R PCS Overview

The 100GBASE-R PCS sub-layer provides functions to map packets to 64B/66B blocks and then distributes them over twenty PCS serial streams or lanes going to the physical media attachment (PMA) sub-layer. The PCS layer encodes eight data octets to 66-bit blocks that it receives from the MAC and Reconciliation layer via a 100 Gb/s media-independent interface (CGMII). A two-bit synchronization header is added to the received 64-bit data to form a 66-bit block — a `"01b"` synchronization header for the data packets and a `"10b"` for the control packets. Subsequently, the 66- bit block is scrambled before distributing it over twenty PCS lanes. The synchronization header is not scrambled.

The PMA sub-layer receives data over these twenty PCS lanes and converts it to a CAUI-10 interface that has ten physical lanes, with each lane running at 10.3125 Gb/s, or an interface that has four physical lanes with each lane running at 25.78 Gb/s. However, serial streams encounter continuous changes in phase and skew that need to be compensated for at the terminating PCS receive layer before the packets are unscrambled and decoded.

Alignment Marker Insertion

For the receiver to deskew and perform PCS lane reordering, transmitters periodically insert alignment markers simultaneously on all PCS lanes after the transmission of every 16,383 blocks of data and control. The alignment markers are inserted after 64B/66B encoding and removed by receivers before 64B/66B decoding. The transmitter removes inter-packet gaps (IPG) to insert the alignment markers. On the terminating receive PCS, the state machine deletes the alignment markers and inserts idle control characters.

An alignment marker has a control synchronization header (`10b`) and a DC-balanced stream comprised of eight octets {M0, M1, M2, BIP3, M4, M5, M6, BIP7} such that the M4, M5, and M6 octets are bit-wise inverse of the M0, M1, and M2 octets. Each PCS lane (0, 1… 19) has a unique M0, M1, M2 coding to allow the receiver to decipher the PCS lane number. See Figure 3.

Note: The bit-interleaved parity check octet BIP7 is bitwise inverse of the bit-interleaved parity check octet BIP3. The bit-interleaved parity check octet is used to detect a bit error ratio (BER) over a PCS lane by computing even parity over 16,383 66-bit blocks (including the sync header) and the previous alignment marker. The current alignment marker is not used in this parity check.

The transmit PCS function distributes the stream onto twenty PCS lanes before sending the data to the PMA sub-layer or the FEC layer. A round-robin mechanism is used to distribute 66-bit blocks to the PCS lanes, starting with PCS lane 0.

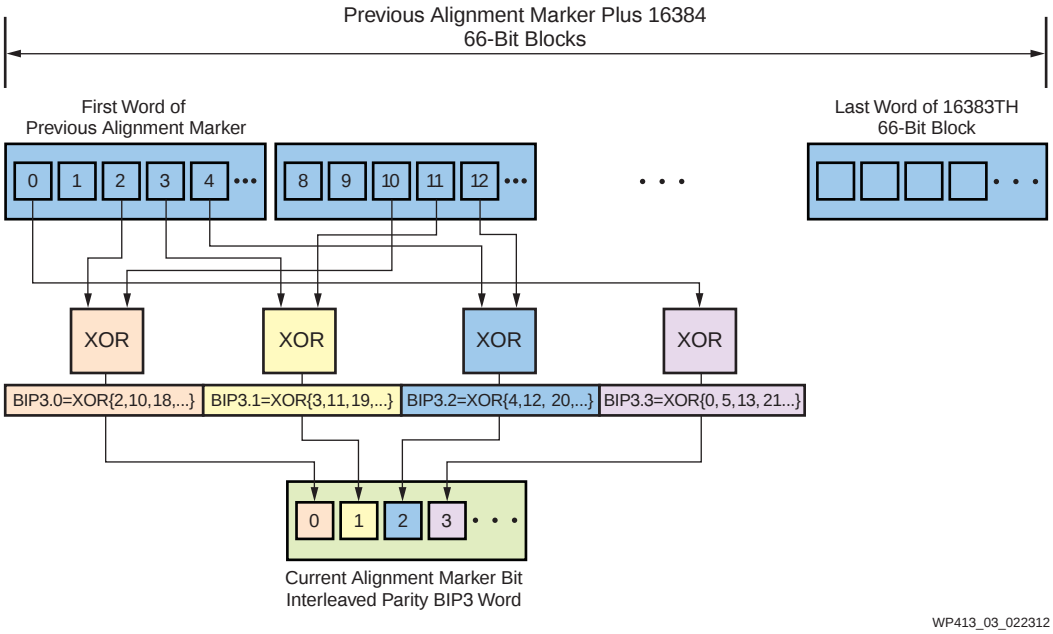

Figure 3: **Alignment Marker Insertion for PCS Lane Deskew and Lane Reordering**

Alignment Marker Detection and Lane Deskew/Reordering

The receiver uses a block header to attain block lock. After attaining block lock, the receiver looks for alignment markers, which provide the PCS lane numbers. After detecting the PCS lane numbers, the receive PCS sub-layer detects and corrects the PCS lane-to-lane skew, with the maximum allowed being 180 ns (~928 bits) with a maximum variation of 4 ns (~21 bits). Note that transmitted-stream PCS lanes can appear on different receive PCS lanes due to the path skew and multiplexing of streams at the intervening PMA sub-layers.

Basic Gearbox Functions

The basic function of the gearbox device is to take incoming 10X 10/11G serial streams and distribute them over a 4X 25/28G serial link interface using a PMA (20:4) mapping function. A standard gearbox function simply recovers the clock and performs bit shuffling. This gearbox functionality can be enhanced to help in system debug. As an example of an enhanced gearbox, following are some of the functions that can be performed by a gearbox translating a CAUI 10X 10G interface to a 4 25G serial links for a 100G MAC function:

- Attain block lock per PCS lane
- Detect alignment markers and PCS lane numbers
- Attain alignment marker lock
- Detect and report BER
- MDIO interface for gearbox registers (PMA/PCS-like and user specified)

Benefits of Implementing the100G Gearbox on an FPGA

FPGAs offer a high-performance, programmable, scalable, extendable platform with large amounts of logic, memory, high-speed serial I/O, and clocking resources. FPGAs are widely used to implement 100G MAC, NPU, Traffic Management/QoS, and Framer functions on a line card. Availability of 28G-capable serial links, such as in the Xilinx 28 nm Virtex-7 HT family, makes FPGAs the ideal platform to build a multiport, flexible 100G gearbox device at optimal cost and power.

As shown in Figure 4, the gearbox devices are required to support multiple interfaces, such as:

- 10X 10G CAUI interface to 4X 25G serial link interface
- Physical interface translation from OTL 4.10 to OTL 4.4
- 11.2 Gb/s 10-lane SFI-S interface to 28G 4-lane SFI-S interface (plus deskew lane).

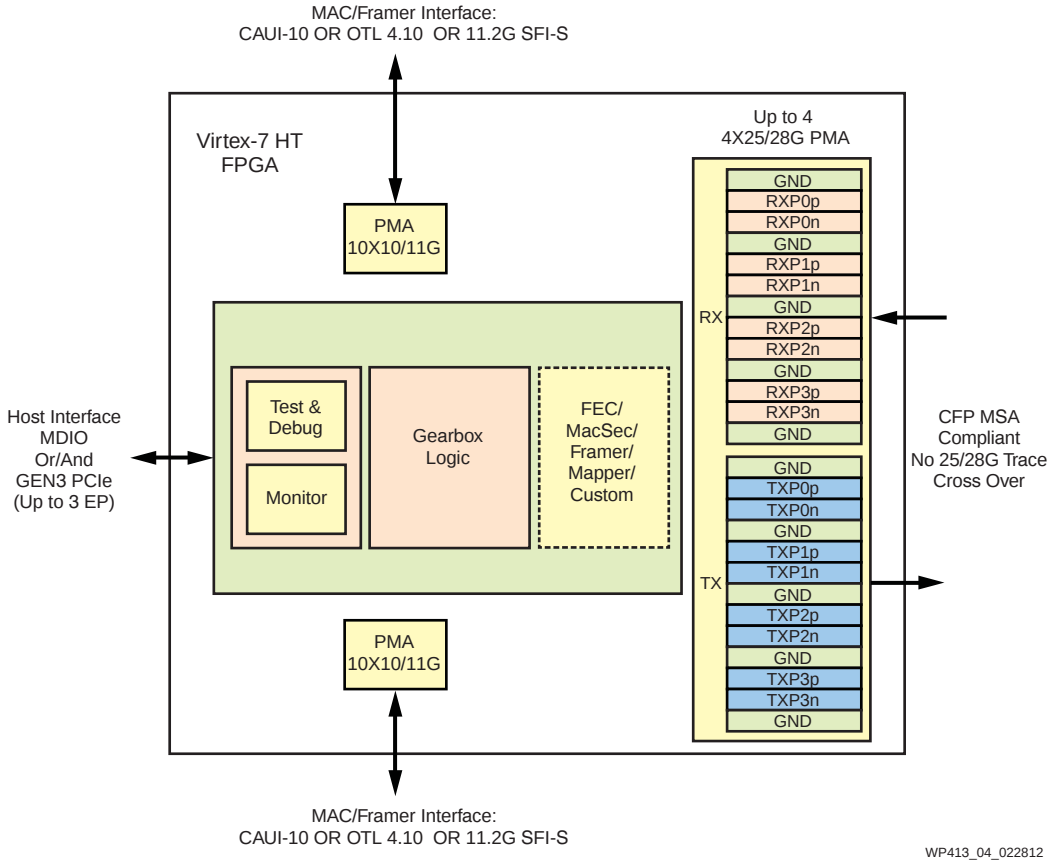

Figure 4: **Virtex-7 HT FPGA in Some Programmable, Extensible, and Flexible 100G Gearbox Applications**

FPGA devices are best suited to enable on-the-fly changes, creating much-needed economies of scale.

In addition to the industry-standard gearbox functionality, the Virtex-7 HT based dual gearbox solution also integrates advanced debug capabilities to replicate basic test equipment. A user can perform the following operations:

- Per transceiver 2D eye diagrams for both 13G and 28G SerDes using Xilinx's industry-leading transceiver debug tool, IBERT
- Per lane PRBS generation/checking
- PCS debug with per-PCS Lane BIP error counters, BER monitoring, block lock and alignment marker lock, and PCSL mapping status available via the MDIO interface
- PCS-level snooping for improved debug capabilities
- Access to the datapath to the 20 PCS lanes

Power and Board Space Efficiency on a Dualport 100G Gearbox

As shown in Figure 4, the Virtex-7 HT family offers up to seventy-two 10G-KR compliant 13.1 Gb/s SerDes and up to sixteen 28 Gb/s SerDes. The 28 Gb/s SerDes are compliant to the CFA MSA proposal to exclude any 28G trace crossover, thereby simplifying PCB layout and the connectors. The Virtex-7 HT devices have a PMA and associated clocking resources to support CAUI - 10X 10G, OTL 4.10, CPPI, 4X 25G serial interface, and OTL 4.4. Also, Virtex-7 HT devices can support:

- SFI-S protocol with 11 lanes at 11.2G (one lane for deskew) on any of the up to seventy-two 13.1 Gb/s SerDes
- SFI-S protocol with 5 lanes at 28G (one lane for deskew) on any of the up to sixteen 28 Gb/s SerDes

Hosting multiple ports on a single FPGA reduces device count on the line card and lowers total power consumption.

OTN OTL 4.10 Translation to OTL 4.4

The Optical Transport Network G.709 hierarchy [Ref 2] defines mapping 100G Ethernet to an optical channel data unit client (ODU4) using generic mapping procedure (GMP). In turn, the ODU4 client is mapped to an optical channel transport unit (OTU4). The OTU4 uses optical channel transport lanes (OTL4.10 or OTL 4.4) as an interface to the optics module. The ODU4 client bit rate is 104.79 Gb/s; the OTU4 bit rate is 111.809 Gb/s. The OTL4.10 interface carries an OTU4 payload over ten SerDes lanes, each running at (255/227) x 9,953,280 Kb/s = 11.18 Gb/s. Alternately, an OTL4.4 interface can be used to carry an OTU4 payload over four SerDes lanes running at (255/227) x 24,883,200 = 27.952 Gb/s. See Figure 5.

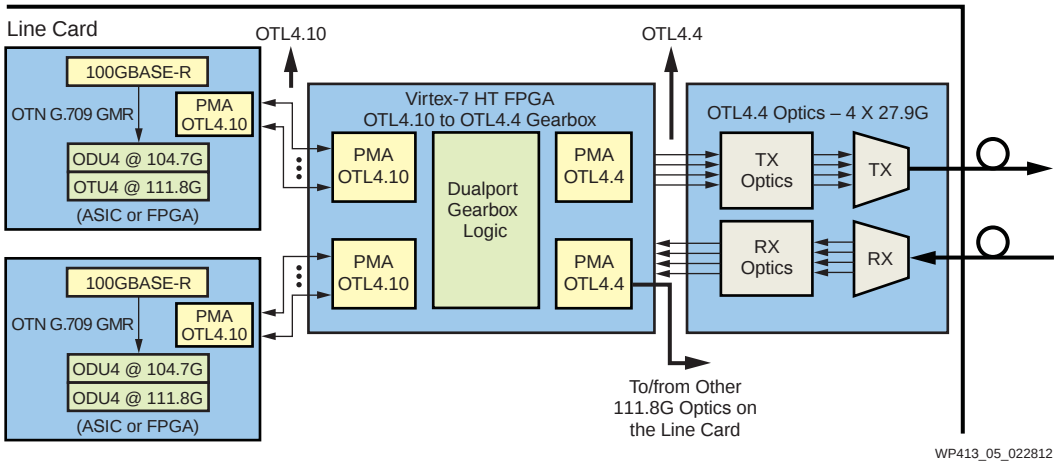

Figure 5: **Virtex-7 HT FPGA in OTL4.10 to OTL4.4 Gearbox Applications**

OIF-SFI-S-01.0 (Scalable SerDes Framer Interface) Standard

The OIF SFI-S 1.0 standard [Ref 6] defines a 4-to-20 lane scalable interface to support 80–160 Gb/s links between optical modules, an FEC processor, and a framer. An FPGA-based gearbox device can support interconnection between these devices using different SFI-S line rates and widths.

The SFI-S protocol is independent of the format of the data and could carry any protocol format on the transmit and receive data paths. To simplify the deskew algorithm and reduce SerDes complexity, the SFI-S protocol uses a separate deskew channel. This has a side effect, in the sense that an SFI-S protocol requires one extra SerDes lane. The SFI-S gearbox maps an 11.2 Gb/s SFI-S-based 100G framer to optics modules using a 28G SFI-S interface. A Virtex-7 HT FPGA-based gearbox is also capable of hosting an FEC processor, as shown in Figure 6.

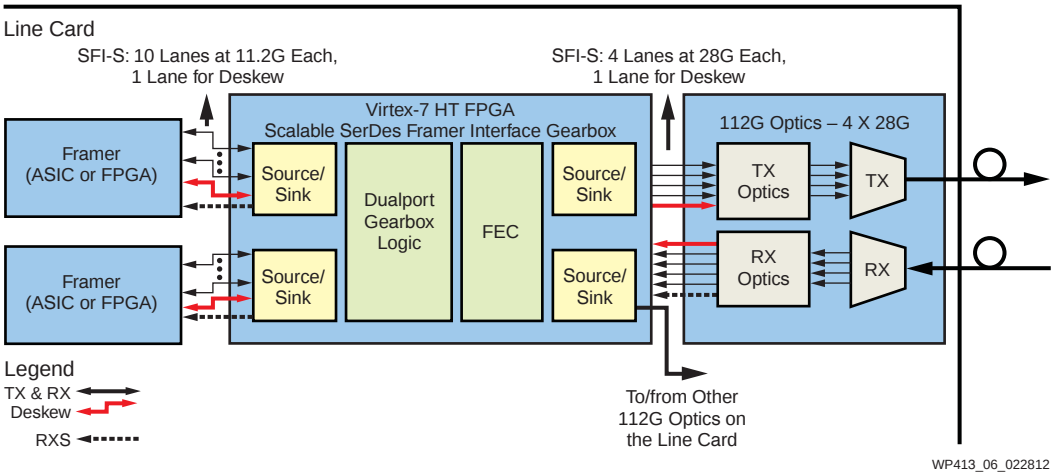

Figure 6: **Virtex-7 HT FPGA as 100G SFI-S Gearbox**

Host Logic to Switch Any PCSL to Any Port for 100G MAC function

The Virtex-7 HT devices have adequate logic and memory resources to build a switching fabric, thereby allowing flexibility and redundancy for failover. The Virtex-7 HT FPGA based dual gearbox solution can map any PCSL lane on the ingress 10 x 10G port to any of the 4 x 25G egress ports going to CFP2 or CFP4 optics modules.

Host In- or Out-of-Band FEC, MacSec, and Other Similar Functions

The Virtex-7 HT FPGA family has more than adequate logic resources to implement highly efficient and low latency industry-standard or proprietary FEC algorithms. In addition, the FPGA logic can be used to implement MacSec functions as well as standards-compliant or custom framer/mapper blocks.

Host Extensible Test, Debug, and Monitoring Functions Easily

The key benefit of using an FPGA in a gearbox is that it allows building extensive amounts of test, debug, and monitoring functions. The 100G testing methods are not well defined and are evolving. This makes an FPGA the ideal platform to continue extending the hardware test and debug functions as 100G testing procedures mature, without having to re-spin silicon every time. It is fairly easy, for example, to support a PRBS pattern generator and checker per PCS lane in a variety of loopback modes. The Virtex-7 HT devices have a built-in PRBS pattern generator that supports PRBS pattern generator and checker per physical lane for both 13.1 Gb/s and 28 Gb/s SerDes.

The Virtex-7 HT device has a significant amount of block RAM to support several milliseconds of known packet streaming to check received data with varied bit patterns and packet lengths.

The Xilinx dual gearbox solution allows the addition of PCS lane skew to the PCSL user interface of a gearbox. In combination with other gearbox functions, this allows the implementation of tests that simulate real-world skew and lane variations.

Virtex-7 HT FPGAs also support up to three PCIe® Gen 3 Endpoints (hard blocks) as an interface to a host processor. The PCIe Gen 3 Endpoints can be leveraged to control and monitor the gearbox functionality.

Summary: Xilinx Virtex-7 HT FPGA Family Overview

Virtex-7 HT FPGAs with integrated 28 Gb/s transceivers deliver the industry's highest bandwidth platform. This family offers the largest single-FPGA solution for 100G–400G line cards for the next generation of communication systems by delivering a total of 2.8 Tb/s full-duplex serial bandwidth. Virtex-7 HT devices support up to 72 10G-KR backplane-compliant 13.1 Gb/s transceivers for line and client side interfaces and up to sixteen high performance 28 Gb/s transceivers to interface with next-generation CFP2 optical modules. The 28G SerDes meets stringent OIF CEI-28G specifications by using a low-phase-noise LC Tank PLL. For more information, visit the Xilinx 28 Gbps Serial Transceiver Technology site.

References

1. IEEE 802.3ba Standard Specification-2010
2. ITU T-REC-G.T-REC-G.709-200912 http://www.itu.int/itu-t/recommendations/index.aspx?ser=G
3. CFP MSA Hardware Specification Revision 1.4, June 7, 2010
4. CFP MSA 100G Roadmap and Applications
5. 10x10 MSA Technical Specifications Rev 2.4 http://www.10x10msa.org/documents/MSA%20Technical%20Rev2-4.pdf
6. OIF-SFI-S-01.0 Standard Specification, November 2008

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
03/01/12	1.0	Initial Xilinx release.

Notice of Disclaimer