

APENet+: a 34 Gbps data transmission system with FPGAs embedded transceivers and QSFP+ modules

R. Ammendola^(a), A. Biagioni^(b), O. Frezza^(b), A. Lonardo^(b), F. Lo Cicero^(b), P. S. Paolucci^(b), D. Rossetti^(b), A. Salamon^(a), F. Simula^(b), L. Tosoratto^(b), P. Vicini^(b)
 (a) INFN Sezione di Roma Tor Vergata (b) INFN Sezione di Roma



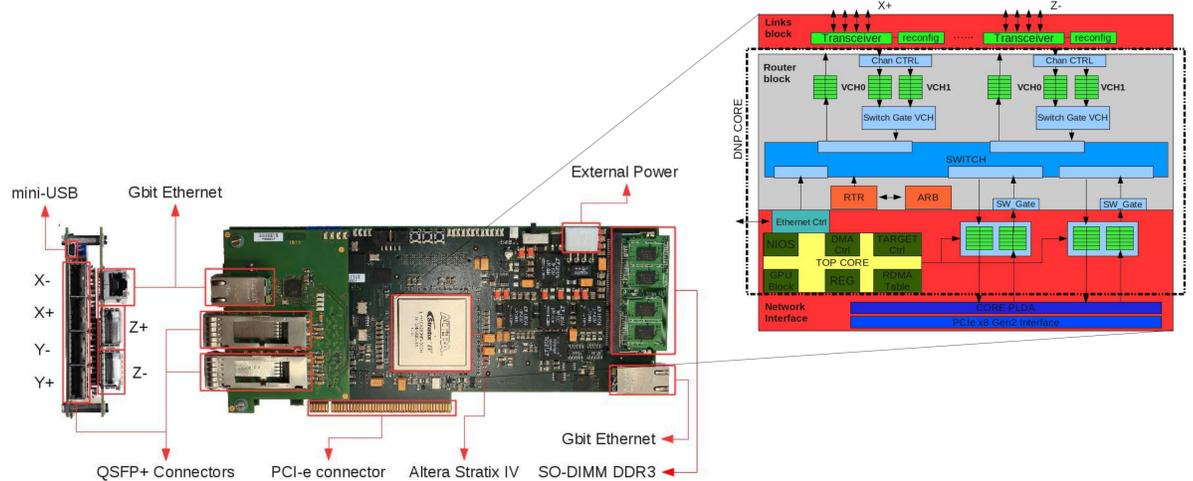
NSS-12 Nuclear Science Symposium - Anaheim, California, October 29 – November 3, 2012

APENet+ is our custom developed PCIe gen2 board based on an Altera Stratix IV FPGA. We demonstrate reliable usage of Altera's embedded transceivers coupled with QSFP+ (Quad Small Form Pluggable) technology. QSFP+ standard defines a hot-pluggable transceiver available in copper or optical cable assemblies for an aggregated bandwidth of up to 40 Gbps. We use embedded transceivers in a 4 lane configuration, each one capable of 8.5 Gbps, for an aggregate bandwidth of 34 Gbps per link. On Stratix IV 290 we can place up to 6 bidirectional links, together with a PCIe gen2 x8 hard IP. We describe design and implementation of this data transmission system.

APENet+ hardware outline

APENet+ is a custom PCIe board Gen2 x8:

- FPGA is Altera Stratix IV 290 F45 C2
- 6 QSFP+ module connectors (4 + 2 on daughter card)
- Daughter card uses Samtec SEAM-SEAF 18.5 mm stack height connector
- Embedded transceiver count is 32: 8 for PCIe, 24 for QSFP+ links
- Each link has 4 lane, connected to a single transceiver
- Each transceiver is capable of up to 8.5 Gbps data rate in full bidirectional mode
- Raw bandwidth up to 34Gb/s for any of the 12 directions
- Altera provides transceiver bonding up to 4 lanes
- Word alignment is up to application



Use cases

APENet+ is used in several experiments as interconnect system:

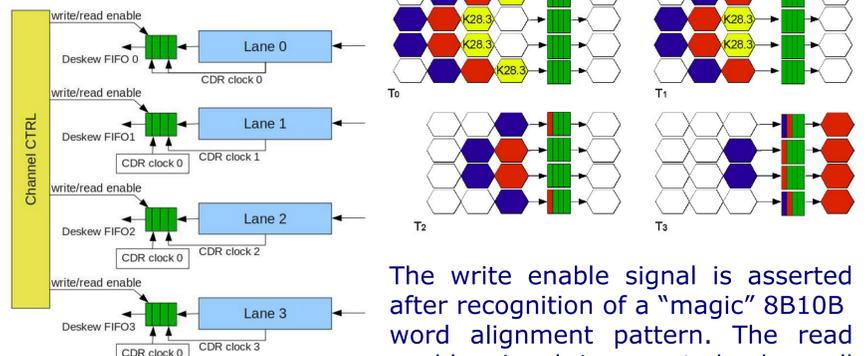
- **Quong**: 3D toroidal network with Nvidia peer-to-peer protocol, for high performance computing applications such as QCD.
- **LHCb**: real time event selection in HEP experiments using GPUs.
- **Euretile**: a EU FP7 funded project aimed to investigate and design a distributed brain-inspired massively parallel computing architecture.
- **Na62**: direct data injection from TDAQ level to GPUs with P2P techniques.
- **Biocomputing**: investigation on a fully connected parallel computing architecture.

Lane alignment

Altera ensure transceiver channel bonding up to 4 lanes, but word alignment needs to be performed by custom logic.

As at PCS level we use 8B10B encoding to maintain the DC balance in the serial data transmitted, we can use special character for alignment purposes.

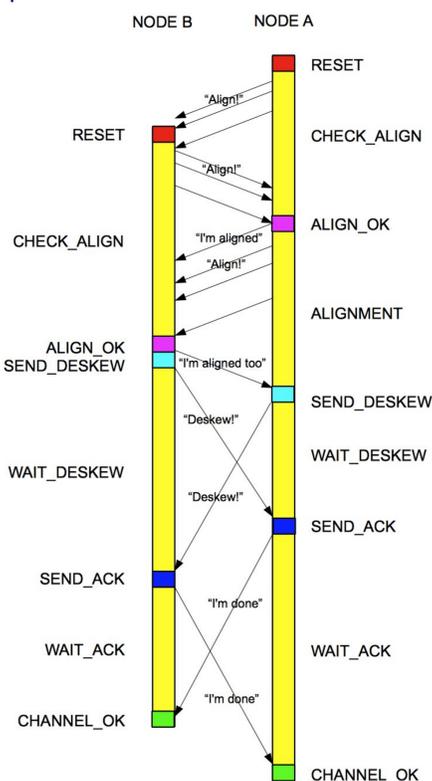
We implement four DESKEW FIFOs, each one directly connected to one lane of the channel, for bonding four independent lanes on a single channel. FIFOs use different clocks on the read and write sides: on the write side the clock used is the specific Clock provided by CDR, while on read side the clock used is the same for all FIFOs (Clock recovered by Lane 0).



The write enable signal is asserted after recognition of a "magic" 8B10B word alignment pattern. The read enable signal is asserted when all FIFOs are no longer empty.

Alignment process

When two independent boards are connected each other, a synchronization mechanism should take place on both sides. Special characters are used to align remote state machines at different phases.



Logic Utilization and Power Consumption

Logic utilization	42 %
Total Thermal Power Dissipation	13191 mW
Combinational ALUTs	70,673 / 232,960 (30 %)
Memory ALUTs	228 / 116,480 (< 1 %)
Dedicated logic registers	61,712 / 232,960 (26 %)
Total registers	61712
Total pins	242 / 1,112 (22 %)
Total block memory bits	7,533,432 / 13,934,592 (54 %)
DSP block 18-bit elements	4 / 832 (< 1 %)
Total GXB Receiver Channel PCS	32 / 32 (100 %)
Total GXB Receiver Channel PMA	32 / 48 (67 %)
Total GXB Transmitter Channel PCS	32 / 32 (100 %)
Total GXB Transmitter Channel PMA	32 / 48 (67 %)
Total PLLs	3 / 12 (25 %)

Logic Block	Combination al ALUTs	Memory ALUTs	ALMs	Dedicated logic registers	Block memory bits	Thermal Power Dissipation
Network Interface Block	32858	228	25893	29622	4609848	914 mW
Switch Block	12879	0	12810	12321	2875392	3464 mW
Single Link Block	4216	0	4413	3842	8032	1134 mW

Bit Error Rate Tests

Reconfig block allows dynamic reconfiguration of PMAs analog settings such as Equalization, Pre-emphasis, DC Gain and VOD (Voltage Output Differential). Manual tuning of the receiver channel's equalization stages involves finding the optimal settings through trial and error and then locking in those values at compile time.

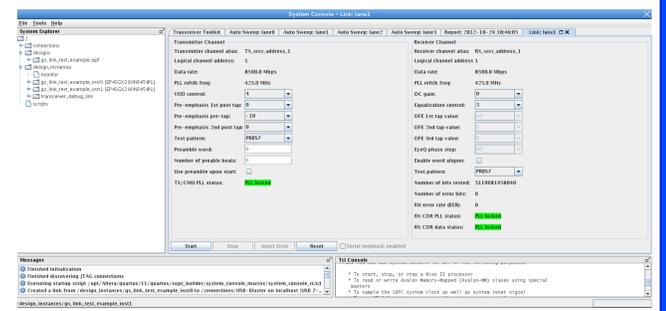
On the transmitter side, Pre-emphasis increases the amplitude of the high-frequency component of the output signal, and thus helps to compensate for the frequency dependent attenuation along the transmission line.

In order to choose the appropriate value of analog controls Altera endows the user with the Transceiver Toolkit feature.

On single lane tests we measured a BER < 10⁻¹⁵.

On 4 bonded lane tests we measured a BER < 10⁻¹³.

Test were performed on all channels and with 0.5, 1, 2, 3 and 5 meters long copper cables, and a 10 meters optical cables.



Future work

Next generation FPGAs (Stratix V) have been recently introduced in the market with transceivers working at up to 14.1 Gbps, which should permit to build a quad link of 56 Gbps, nominally.

A development kit will be soon available with a QSFP+ module, allowing tests at least at 40 Gbps.

