21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Contribution ID: **483** Type: **oral presentation**

Hardware and Software Design of FPGA-based PCIe Gen3 interface for APENet+ network interconnect system

Monday, April 13, 2015 5:45 PM (15 minutes)

The computing nodes of modern hybrid HPC systems are built using the CPU+GPU paradigm.
When this class of systems is scaled to large size, the efficiency of the network connecting GPUs mesh and supporting the internode traffic is a critical factor. The adoption of a low latency, high performance dedicated network architecture, exploiting peculiar characteristics of CPU and GPU hardware, allows to guarantee scalability and a good level of sustained performances.

In the attempt to develop a custom interconnection architecture optimized for scientific computing we designed APEnet+, a point-to-point, low-latency and high-performance 3D torus network controller which supports 6 fully bidirectional off-board links.
The first release of APEnet+ (named V4), was a board based on a high end 40nm Altera FPGA that integrates multiple (6) channels at 34Gbps of raw bandwidth per direction and a PCIe Gen2 x8 host interface. APEnet+ board was the first-of-its-kind to implement a Remote Direct Memory Access (RDMA) protocol to directly read/write data from/to Fermi and Kepler NVIDIA GPUs using the Nvidia "peer-to-peer"and "GPUDirect RDMA"protocols, obtaining real zero-copy, low-latency GPU-to-GPU transfers over the network and reducing the performance bottleneck due to the costly copies of data from user to kernel space(and vice-versa).

The last generation of APEnet+ systems (V5), currently under development, is based on state-of-the-art high end FPGA, 28nm Altera Stratix V, offering a number of multi-standard fast transceivers (up to 14.4 Gbps), huge amount of configurable internal resources and hardware IP cores to support main interconnection standard protocols.
APEnet+ V5 implements a PCIe Gen3 x8 interface, the current standard protocol for high end system peripherals, in order to gain performance on the critical CPU/GPU connection and mitigate the effect of the bottleneck represented by GPUs memory access.
Furthermore the FPGA technology advancement, allowed us to integrate in V5, new off-board torus channels characterized by a target speed of 56 Gbps.
Both Linux Device Driver and the low-level libraries, have been redesigned to support the PCIe Gen3 protocol, introducing optimizations and solutions based on hardware/software co-design.

In this paper we present the architecture of APEnet+ V5 and discuss the status of APEnet+ V5 PCIe Gen3 hardware and system software design. Measures of performance in terms of latency and bandwidth, both for the local APEnet+ to CPU-GPU connection (with Kepler class GPU) and host-to-host via torus links, will also be provided.

Primary authors: LONARDO, Alessandro (Universita e INFN, Roma I (IT)); BIAGIONI, Andrea (INFN); ROSSETTI, Davide (INFN Rome and NVidia Corp. (USA)); PASTORELLI, Elena (INFN Rome); LO CICERO, Francesca (INFN Rome); SIMULA, Francesco (INFN Rome); TOSORATTO, Laura (INFN); MARTINELLI, Michele (INFN Rome); FREZZA, Ottorino (INFN Rome); PAOLUCCI, Pier Stanislao (INFN Rome); VICINI, Piero (INFN Rome); AMMENDOLA, Roberto (INFN)

Presenter: MARTINELLI, Michele (INFN Rome)

Session Classification: Track 8 Session

Track Classification: Track8: Performance increase and optimization exploiting hardware features