Contribution ID: **505** Type: **Poster**

Latest generation interconnect technologies in APEnet+ networking infrastructure

Tuesday, 11 October 2016 16:30 (15 minutes)

Hybrid systems are emerging as an efficient solution in the HPC arena, with an abundance of approaches for integration of accelerators into the system (i.e. GPU, FPGA). In this context, one of the most important features is the chance of being able to address the accelerators, whether they be local or off-node, on an equal footing. Correct balancing and high performance in how the network supports this kind of transfers in internode traffic become critical factors in global system efficiency.

The goal of the APEnet project is the design and development of a point-to-point, low-latency and high-throughput interconnect adapter, to be employed in High Performance Computing clusters with a 3D toroidal network mesh.

In this paper we present the status of the 3rd generation design of the board (V5) built around the 28nm Altera StratixV FPGA; it features a PCIe Gen3 x8 interface and enhanced embedded transceivers with a maximum capability of 50.0Gbps. The network architecture is designed according to the Remote DMA paradigm. APEnet implements the NVIDIA GPUDirect RDMA and V2 ("peer-to-peer") protocols to directly access GPU memory, overcoming the bottleneck represented by transfers between GPU/host memory.

The APEnet+ V5 prototype is built upon the StratixV Dev Kit with the addition of a proprietary, third party IP core implementing multi-DMA engines. Support for zero-copy communication is assured by the possibility of DMA-accessing either host and GPU memory, offloading the CPU from the chore of data copying. Current implementation shows an upper limit for the memory read bandwidth of 4.8GB/s. Here we describe the memory write process hardware optimization relying on the use of two independent DMA engines and an improved TLB and characterization of software enhancements aimed at exploiting the hardware capabilities to the most, e.g. using CUDA 7.5 features and the driver migration to user-space, this latter allowing us to either better pinpoint software-induced overhead - compared to a kernel-space only driver implementation - and to lower the perceived latency to the application.

The APEnet+ V5 prototype offers three APElink high performance data transmission channels. The X channel was implemented by bonding 4 lanes of the QSFP connector available on the board; the Y and Z channels were implemented onto the HSMC interface. In this paper we describe the Transmission Control Logic that manages the data flow by encapsulating packets into a light, low-level, "word-stuffing"proprietary protocol able to detect transmission errors via CRC. The current implementation of the APElink TCL is able to sustain the link bandwidth of about 5GB/s at an operating frequency of ˜312MHz. Measures of performance (latency and bandwidth) on host-to-host and GPU-to-GPU between two servers will be provided.

Finally, as regards future developments, we discuss work undertaken towards compliance with next generation FPGAs with hard IP processors on board.

Tertiary Keyword (Optional)

Secondary Keyword (Optional)

Network systems and solutions

Primary Keyword (Mandatory)

High performance computing

Primary authors: LONARDO, Alessandro (Universita e INFN, Roma I (IT)); BIAGIONI, Andrea (INFN); ROSSETTI, Davide (Universita e INFN, Roma I (IT)); PASTORELLI, Elena (INFN); LO CICERO, Francesca (INFN); SIM-

ULA, Francesco (INFN); TOSORATTO, Laura (INFN); MARTINELLI, Michele (INFN); FREZZA, Ottorino (INFN); PAOLUCCI, Pier Stanislao (INFN); VICINI, Piero (INFN Rome Section); AMMENDOLA, Roberto (Universita e INFN Roma Tor Vergata (IT))

Presenter: MARTINELLI, Michele (INFN)

Session Classification: Posters A / Break

Track Classification: Track 6: Infrastructures