APEnet+: a 3D Torus network optimized for GPU-based HPC Systems


INFN Tor Vergata, Italy
INFN Roma, Italy

The APEnet+ Card

APEnet+ is the high performance, low latency interconnect card developed at INFN targeting hybrid CPU-GPU-based HPC platforms:
- 20/40 toroidal mesh topology granting point-to-point deadlock-free communications
- PCIe board with signaling capabilities for up to x8 Gen 2
- (4x4 Gb/s peak bi-directional bandwidth with the host PC)
- 6 full-duplex links on 4 bonded lanes over QSPR cables
- raw bandwidth up to 840G/s for any of the 12 directions
- power envelope of 80W → power dissipation limited to 20W
- transfers are RDMA – CPU is not involved in data movement
- custom-designed network-to-GPU interface on top of PCIe P2P transactions available on Fermi-class NVidia GPUs → significant reduction in access latency for inter-node data transfers.

Preliminary benchmarks:
- Coded with APEnet+ RDMA API.
- CUDA 4.1.
- One-way point-to-point test involving two nodes.
- Comparison between APEnet+ GPU to GPU transfers and vice versa.
- Transmissions on NVidia Fermi and APEnet+ joint development with NVidia.
- APEnet+ board acts as a peer.
- APEnet+ allows direct data exchange on the PCIe bus.
- Zero copy, inter-node GPU-to-host, host-to-GPU and GPU-to-GPU.
- Very preliminary.
- Foot TX curves exhibit a plateau at 20W.
- No P2P means use of CUDA memcpy() (29/25 vs)
- Latency reduction for small messages.
- APEnet+: a 3D Torus network optimized for GPU latency (and w/o P2P and NVAPICH2 over IB)
- Reduced pipelining capability of the APEnet+ HW.
- Replacing the PCIe mesh with 3D Torus network.
- Cooperation of APEnet+ HW blocks and software components to monitor the system.
- Detection of APEnet+ faults (links malfunction, increasing temperature...).
- Collection of Host status.
- Propagation of the Host faulty status towards the node’s first neighbours via the 3D network.

The QuonG HPC platform

QuonG (GPU-centric dynamic-ones on Gpu) is an INFN initiative that aims to develop an HPC system dedicated to Lattice QCD computations; it is a massively parallel computing platform leveraging on commodity multi-core processors coupled with latest generation GPUs as computing nodes interconnected by the APEnet+ network 3D torus network. This network mesh is particularly suited to the transmission patterns of the set of algorithms LQCD belongs to.

Heterogeneous cluster: PC mesh accelerated with high-end GPU and interconnected via 3D torus network
- Tight integration between accelerators (GPU) and custom/reconfigurable network (DNP on FPGA) allowing latency reduction and computing efficiency gain
- Communicating with optimized custom interconnect (APEnet+), with a standard software stack (MPI, OpenMP...)
- Optionally an augmented programming model (cuIO)
- Community of researchers sharing codes and expertise (LQCD, GW, Bio-computing, Laser-plasma interaction)
- GPUs by NVidia
- Solid HD and high SW
- Collaboration with NVidia US development team to integrate “GPU” with our network

Fault-tolerance features

When scaling to petawollar-scale in HPC, usage of techniques that aim to maintain a low Time In Failure (TIF) ratio is mandatory.

Relying on the idea of splitting the fault-tolerance problem into fault awareness and fault Reactivity, APEnet+ provides a way to obtain the awareness, by monitoring itself and its host by means of watchdog techniques.
- Cooperation of APEnet+ HW blocks and software components to monitor the system.
- Detection of APEnet+ faults (links malfunction, increasing temperature...).

Next months R&D

APEnet+ update based on current and next generation (28nm) FPGA - i.e. more bandwidth, less latency:
- Architectural enhancements
- Larger buffers (larger packets handling).
- Optimized NV (low latency, direct access) interface to next-gen GPUs.
- Fault handling/tolerance capabilities to safely scale at multi-PFLOPS.
- Introduction of Dual PCI Gen3: 3x 4x
- Bipolar vs 508Gbps (Gen2), better encoding (128b/130b) vs 8b/10b – > 2x SW
- Increased nº of transceivers: allows for the integration of 2x PCIe Gen3 x8 → 2x Gen4
- Transceiver switching frequency increase
- 14 Gbps vs current 8 Gbps – > 2x bw on torus link

Contacts

INFN Roma, Italy, email: [luca.tosoratto@roma3.infn.it, giovanni.vinco@roma3.infn.it]
Web site: [http://aapegate.roma1.infn.it/APE](http://aapegate.roma1.infn.it/APE)

This project is partially funded by the EURELIUS EU Project.