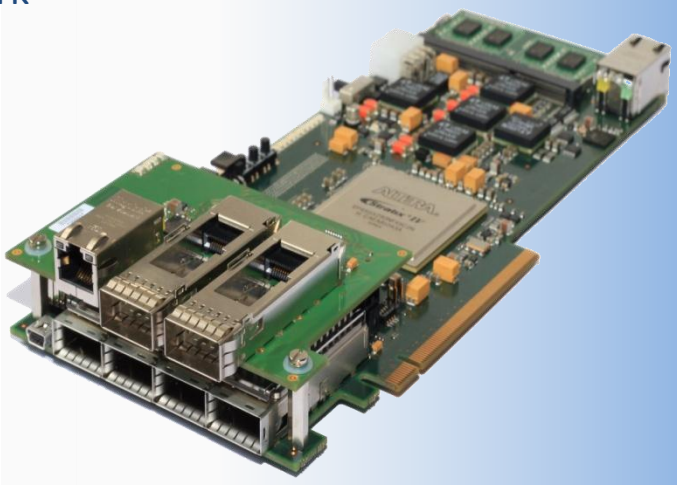
The APEnet+ interconnect

3D toroidal GPU-optimized network



The N-Dimensional torus network is a well established solution to interconnect parallel computing systems dedicated to a broad class of scientific applications (domain decomposition, stencil computation,…)

The APEnet is a 3D Torus network interconnect optimized for scientific computation on GPU-accelerated clusters.

APEnet+, the current release, is an FPGA-based full-length double-slot PCI-Express card.

APEnet+ is optimized for high bandwidth and low latency:

- Support for RDMA communication paradigm allowing for zero-copy transfers.
- Support for "NVIDIA GPUDirect" Peer-to-Peer communications: avoids buffer copies between GPU and host and enables excellent GPU-to-GPU latencies.

APEnet+ is a scalable and cost effective interconnection solution:

- Up to 32k computing nodes in the current implementation.
- No external switching hardware required, only card and cables.

APEnet+ advanced features:

- Hardware support for system fault tolerance.
- Hardware acceleration (on the FPGA) of specific software tasks.

APEnet development roadmap:

- PCI-Express Gen3 x16 introduction.
- Link speed enhancement (up to 56 Gb/s).

INFN Istituto Nazionale di Fisica Nucleare

Features Summary

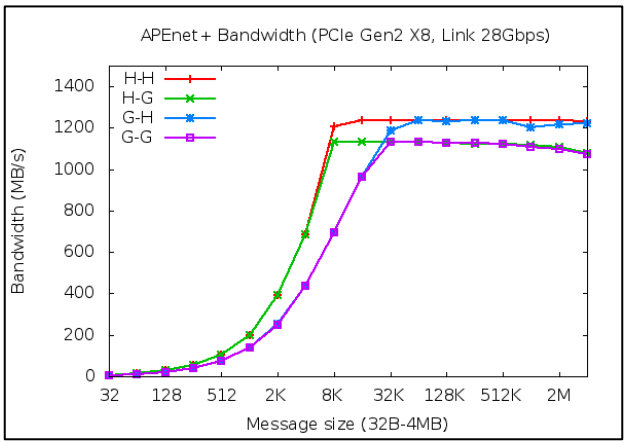
- **Host Bus Interface Specification**
  - PCI Express Gen2 x8
    - 2.5 or 5.0 GT/s link rate
    - Auto negotiate x8 x4 x2
- **Connectivity**
  - 6 Bi-Directional 34 Gbps Links
  - Passive Copper Cables
  - Optional Optical active Cables
  - QSFP+ connectors compliance
- **Physical Specification**
  - 11.0 x 26.5 cm
  - 2 PCI I/O slot wide in 6 links configuration
  - 1 PCI I/O slot wide in 4 links configuration
- **Software Tools**
  - Linux Host Driver
  - API Library
  - OpenMPI module
- **Advanced Features**
  - RDMA Communication Paradigm
  - NVIDIA GPUDirect P2P Communications
  - Up to 32^3 supported toroidal mesh
  - Auto-routing of packets
  - 2 Virtual Channels per Link
  - Tested on Linux x86_64
- **Under development**
  - Support for System fault tolerance
  - Collective offloading
  - Collective re-routing on on-board Ethernet

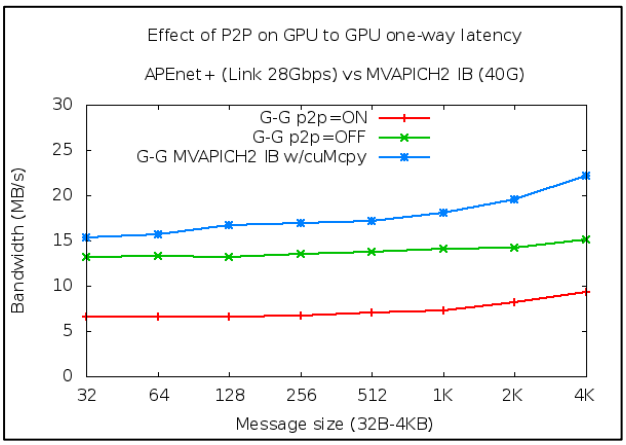
Benchmarking

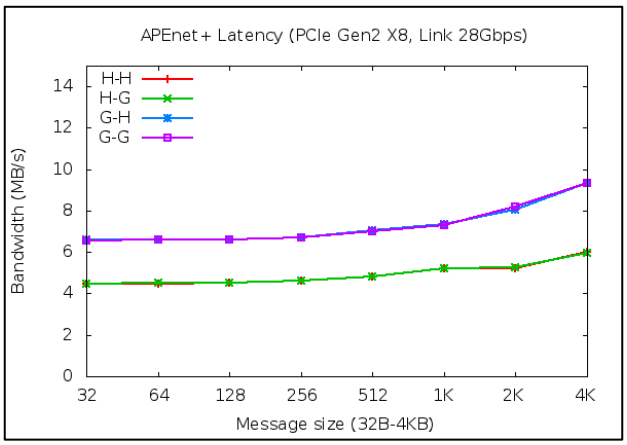
Minimum GPU to GPU latency for small messages of 6.6 µs

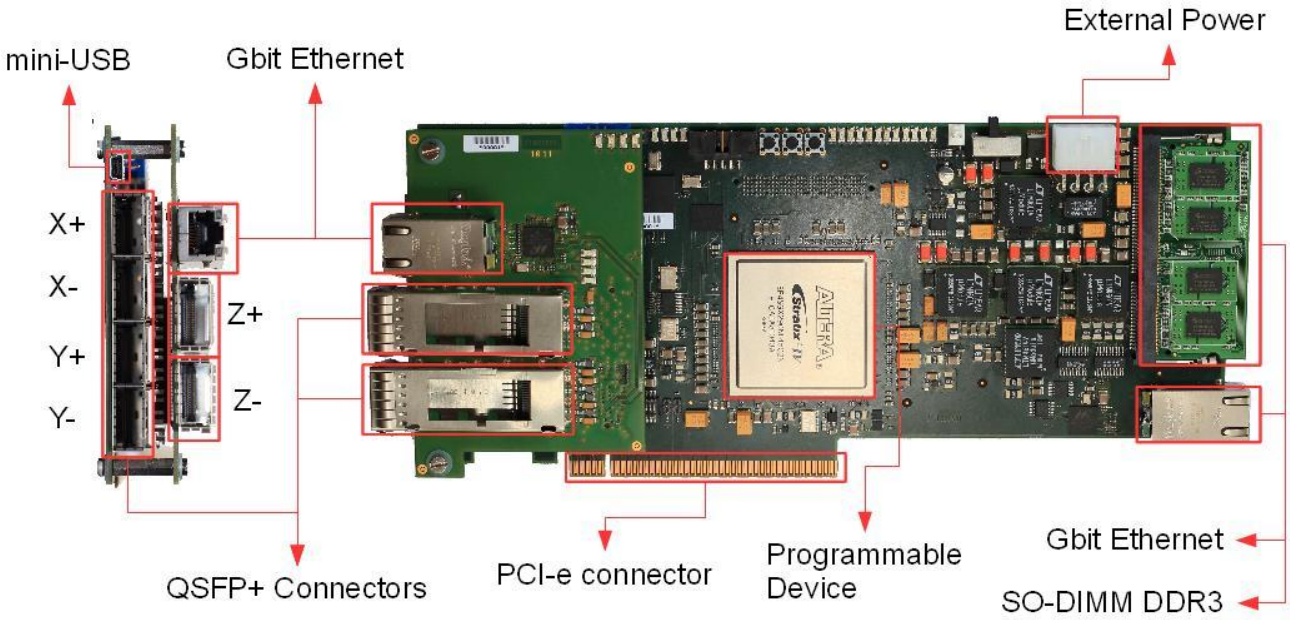
Testbed:

- 2 Xeon-based computing nodes each with
  - NVidia C2050
  - APEnet+
- Link speed @ 28 Gb/s
- PCIe @ Gen2 speed









APE group

The APE Group

The APE research group is active in the area of HPC and Embedded systems since 25 years.

Contact Info

http://apegate.roma1.infn.it

piero.vicini@roma1.infn.it - Tel/Fax: +39 0649914423

INFN Sezione di Roma
Piazzale Aldo Moro 2 - 00185 Roma, Italy