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

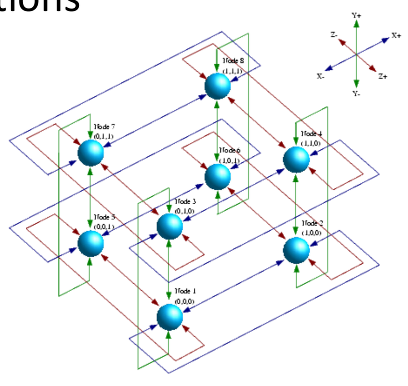
R. Ammendola[1], A. Biagioni[2], O. Frezza[2], F. Lo Cicero[2], A. Lonardo[2], P. S. Paolucci[2], D. Rossetti[2], F. Simula[2], L. Tosoratto[2] and P. Vicini[2]

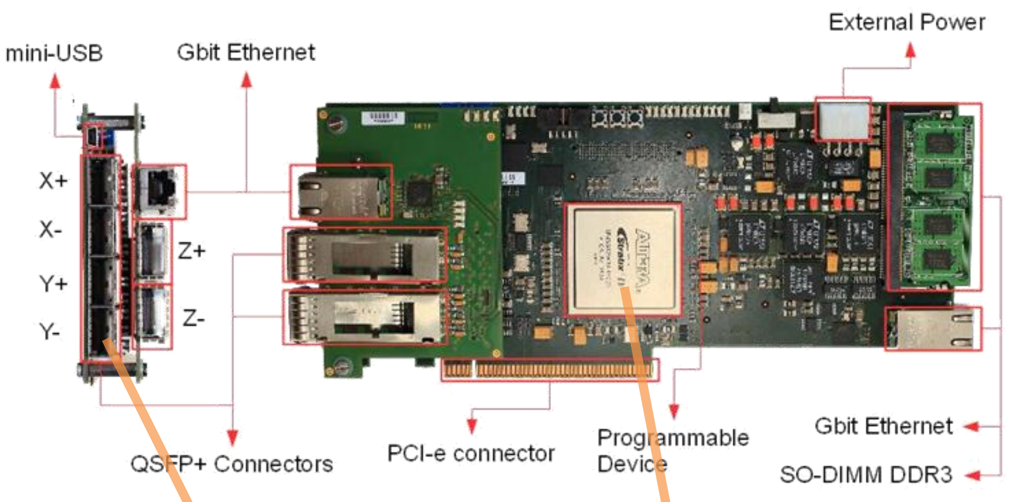
[1] *INFN Tor Vergata, Italy* [2] *INFN Roma, Italy*

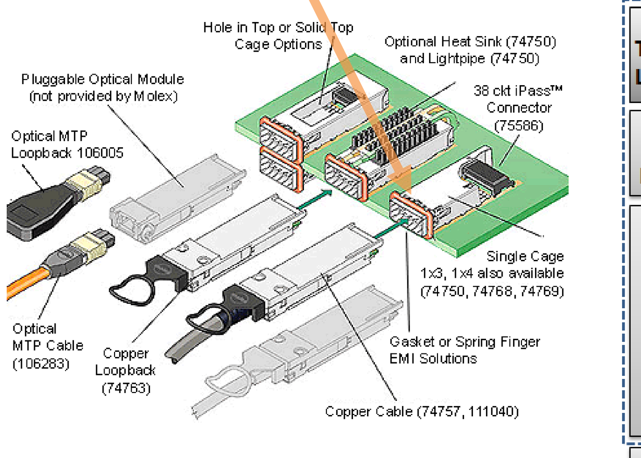
The APEnet+ Card

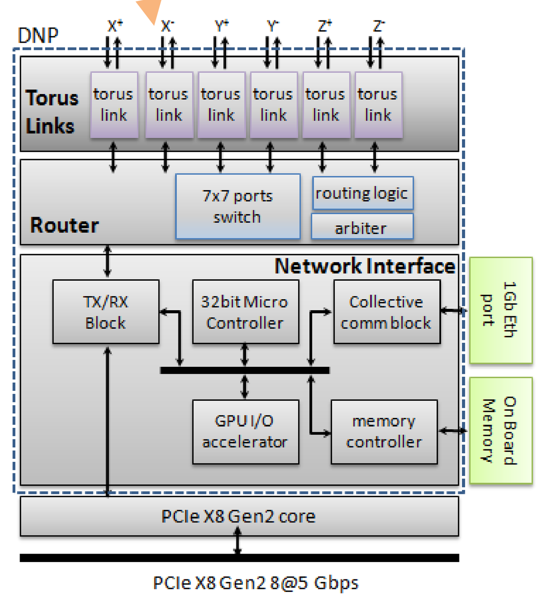
APEnet+ is the high performance, low latency interconnect card developed at INFN targeting hybrid CPU-GPU-based HPC platforms:

- ✓ 2D/3D toroidal mesh topology granting point-to-point dead-lock free communications
- ✓ PCIe board with signaling capabilities for up to X8 Gen2 (**4+4 GB/s** peak bi-directional bandwidth with the host PC)
- ✓ 6 full bi-dir links on 4 bonded lanes over **QSFP+** cables
- ✓ raw bandwidth up to **34Gb/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.

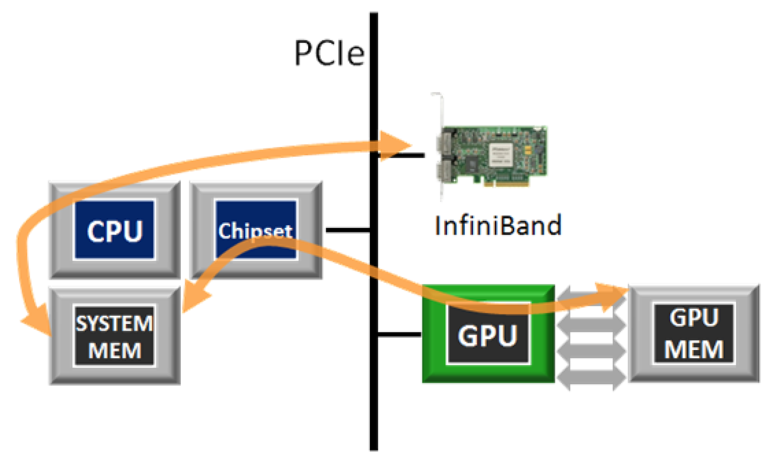






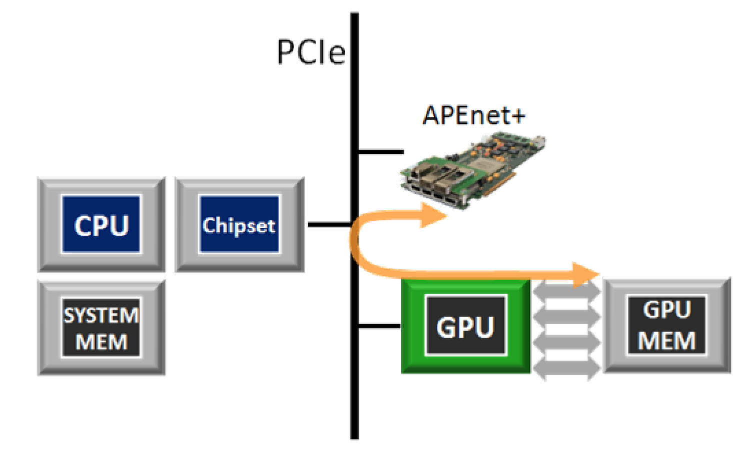


TRADITIONAL DATA FLOW



- ✓ Transmission of data residing on GPU memory, with a non-P2P adapter, *e.g.* Mellanox Infiniband, requires the CPU to:
  - Wait for current GPU Kernel to finish.
  - Copy data from GPU to an intermediate, CPU memory buffer.
  - Issue network transfer command on this memory buffer.
  - … and vice-versa on the receive side.

APEnet+ DATA FLOW



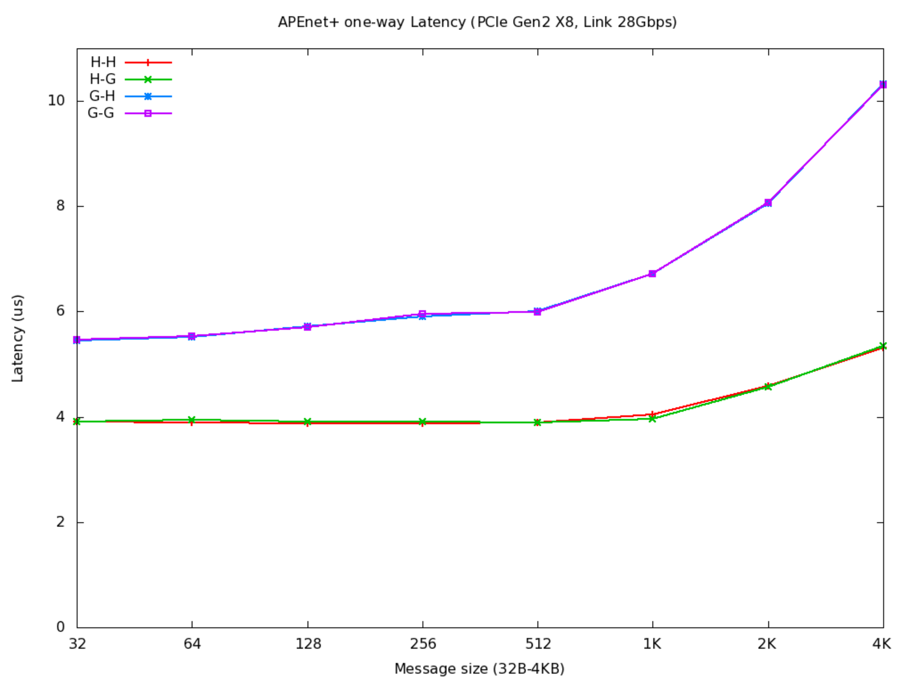
- ✓ P2P between Nvidia Fermi and APEnet+
  - Joint development with NVidia.
  - APEnet+ board acts as a peer.
- ✓ No bounce buffers on host. APEnet+ can target GPU memory with no CPU involvement.
- ✓ GPUDirect allows direct data exchange on the PCIe bus.
- ✓ Real zero copy, inter-node GPU-to-host, host-to-GPU and GPU-to-GPU.
- ✓ Latency reduction for small messages.

The DNP is the INFN custom-designed IP at the core of APEnet+. Its basic blocks are:

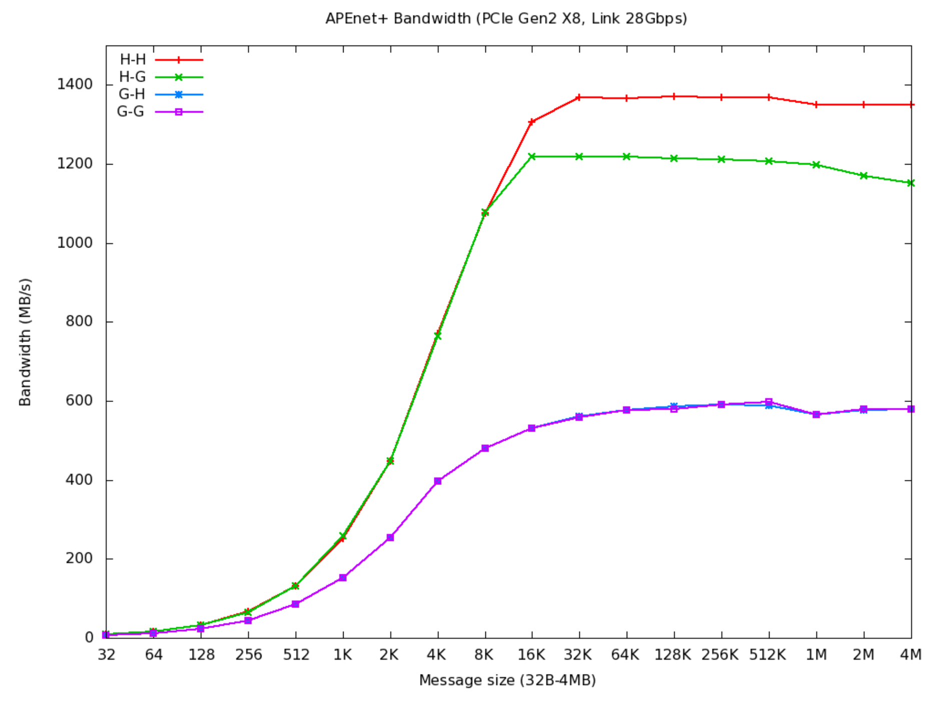
- ✓ **torus links** – bi-dir DC-balanced Ser/Des with word-stuffing CRC-protected low-level packet protocol;
- ✓ **router** – for packet arbitration and dimension-ordered routing, guaranteed deadlock-free by using virtual channels (60ns routing latency);
- ✓ **network interface** – for packet injection and processing logic comprising host interface, TX/RX logic and two auxiliary blocks:
  - **micro controller** – part of the FPGA, relieves the DNP core from some chores of RDMA implementation (for fast LUT management on its on-board memory)
  - **GPU/IO accelerator** – custom block for acceleration of GPU-initiated network operations

Preliminary benchmarks:

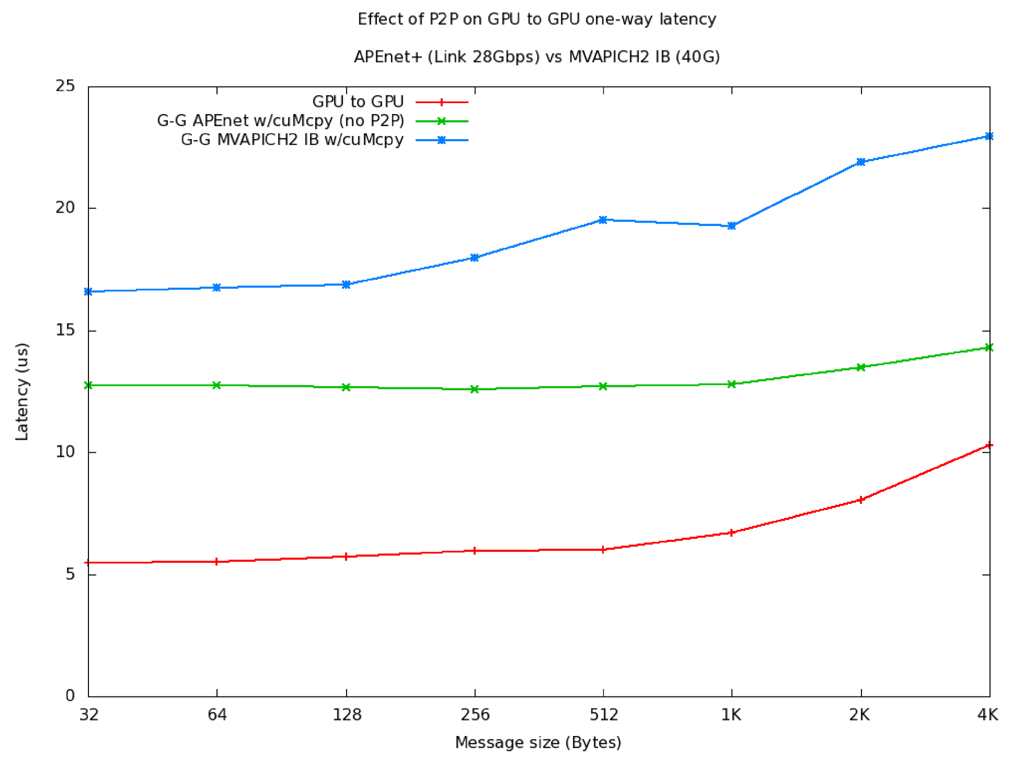
- ✓ Coded with APEnet+ RDMA API.
- ✓ CUDA 4.1.
- ✓ One-way point-to-point test involving two nodes.



- ✓ OSU-like one-way latency test for small msg sizes
- ✓ No small message optimizations
  - Copying of data in temporary buffers.
  - Reduced pipelining capability of the APEnet+ HW
- ✓ No large difference of perf with round-trip test
- ✓ **~ 5-6 μs on GPU-GPU test: record!**
- ✓ 1.5x due to GPU TX, working for improvements



- ✓ Very preliminary
- ✓ Host TX curves exhibit a plateau at msg size of 16KB
  - investigating about how to accelerate the receiving tasks performed by the μC
- ✓ GPU TX curves show a low asymptotic bw of 600MB/s.
  - P2P read protocol fully implemented, but
  - The overhead is still not overlapped among subsequent packet transmission, preventing the pipelining of the packet flow



- ✓ Comparison between APEnet+ GPU to GPU latency w or w/o P2P and MVAPICH2 over IB
- ✓ No P2P means use of `cudaMemcpyD2H/H2D()`
- ✓ `cuMemcpy()` costs ~3.5μs

The QUonG HPC platform

QUonG (QUantum chromo-dynamics 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 last 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.

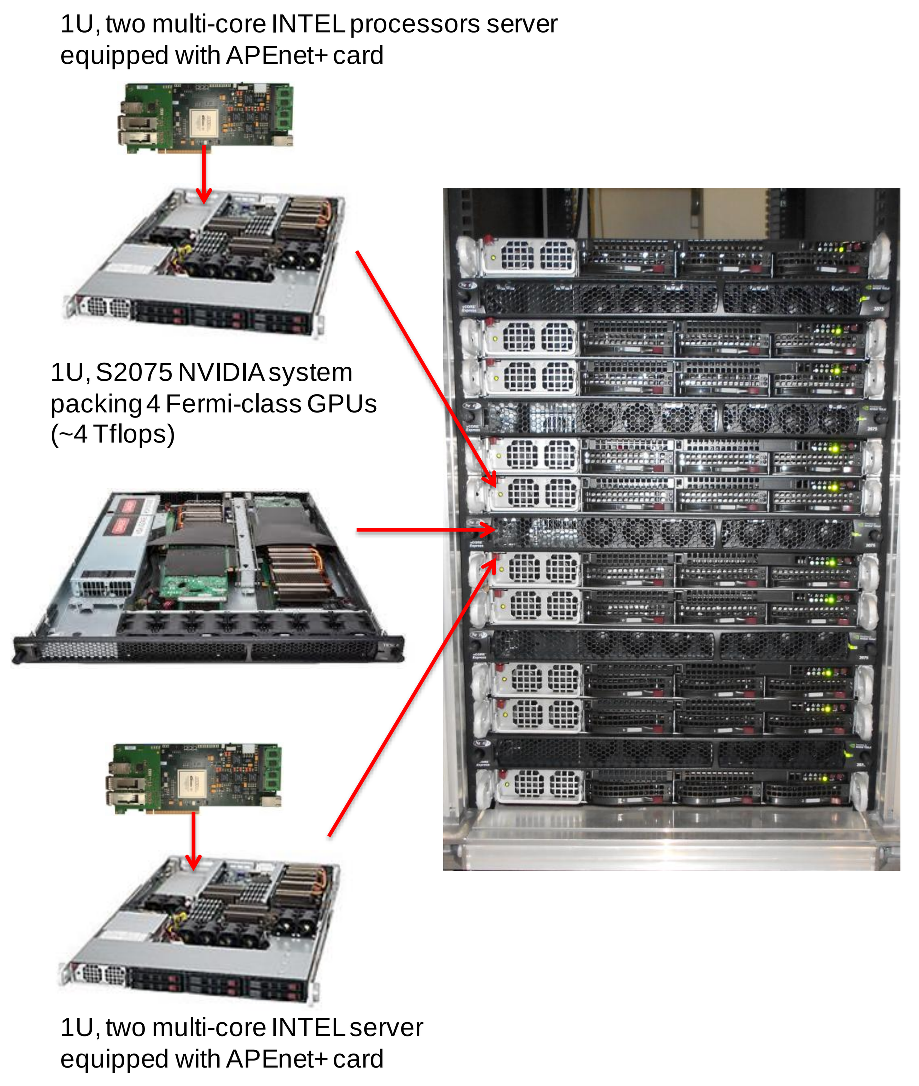
- ✓ Heterogeneous cluster: PC mesh accelerated with high-end GPU and interconnected via 3D torus network
- ✓ Tight integration between accelerators (GPU) and custom/reconfigurable network (DNP on FPGA) allowing latency reduction and computing efficiency gain
- ✓ Communicating with optimized custom interconnect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

QUonG full rack deployment:

- ✓ 42U standard rack system:
  - 60/30 TFlops/rack in single/double precision
  - 25KW/rack (0.4KW/TFlops)
  - 300K€/rack (<5K€/TFlops)

Roadmap to full QUonG rack:

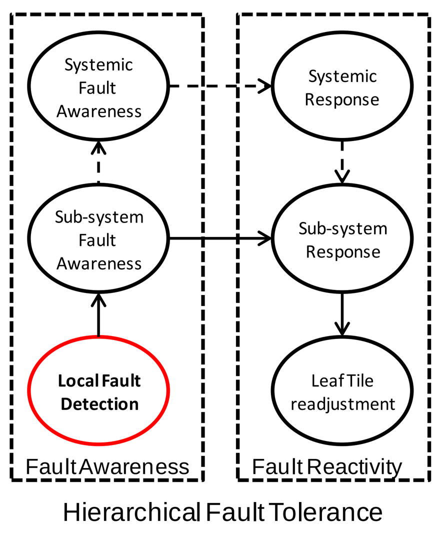
- ✓ 25 TFlops ready at 1Q/12
- ✓ Full rack ready at 4Q/12
- ✓ …waiting for Kepler GPUs



Fault-tolerance features

When scaling to peta/exa-scale in HPC, usage of techniques that aim to maintain a low Failure In Time (FIT) 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.



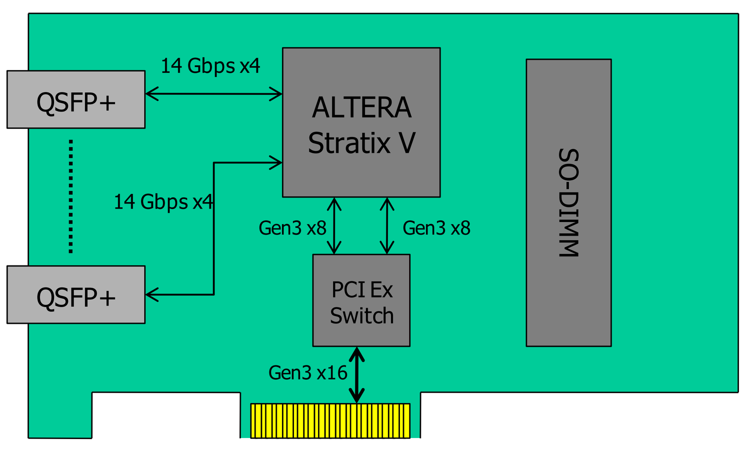
Hierarchical Fault Tolerance

- ✓ 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.

Next months R&D

APEnet+ update based on current and next generation (28nm) FPGA - *i.e.* more bandwidth, less latency:

- ✓ Architectural enhancements
  - Larger buffers (bigger packets handling).
  - Optimized HW (low latency, direct access) interface to next-gen GPUs.
  - Fault handling/tolerance capabilities to safely scale at multi-PFLOPS.
- ✓ Introduction of Dual PCI Gen3 -> **4x bw**
  - 8Gbps vs 5Gbps (Gen2), better encoding (128b/130b) vs 8b/10b -> 2x bw
  - increased # of transceivers allows for the integration of a 2nd PCIe Gen3 x8 -> 2x bw



- ✓ Transceiver switching frequency increase
  - 14 Gbps vs current 8 Gbps -> **~2x bw on torus link**

Contacts

INFN Roma, Italy, email: laura.tosoratto@roma1.infn.it , piero.vicini@roma1.infn.it

Web site: http://apegate.roma1.infn.it/APE

This project is partially funded by the EURETILE EU Project.