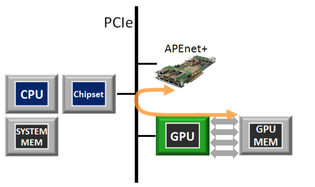
Development of a GPU aware NIC: from HPC to HEP experiments

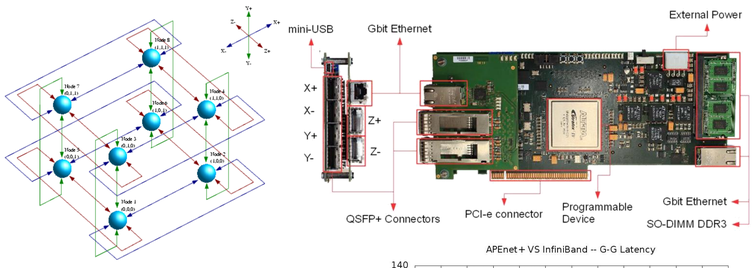
R. Ammendola[(a)], A. Biagioni*[(b)], O. Frezza[(b)], F. Lo Cicero[(b)], A. Lonardo[(b)], M. Martinelli[(b)]
P.S. Paolucci[(b)], E. Pastorelli[(b)] D. Rossetti[(b)(c)], F. Simula[(b)], L. Tosoratto[(b)], P. Vicini[(b)]
(a) INFN Sezione di Roma Tor Vergata (b) INFN Sezione di Roma (c) NVIDIA Corp.

The APEnet+ Card

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

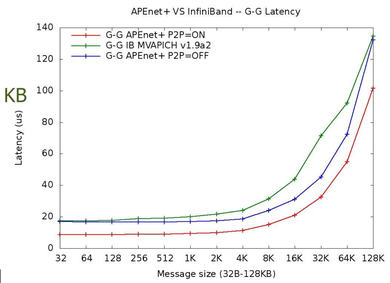
- 2D/3D toroidal mesh topology for point-to-point dead-lock free communications
- PCIe Gen2 X8 host interface (**4+4 GB/s**)
- 6 full bidirectional links over **QSFP+** cables
- raw bandwidth up to **34Gb/s** per direction
- **RDMA** transfer protocol – CPU offload
- **fast I/O interface with NVIDIA GPU:**
  - **GPUDirect v2 (Custom NVP2P)**
  - **GPUDirect RDMA**

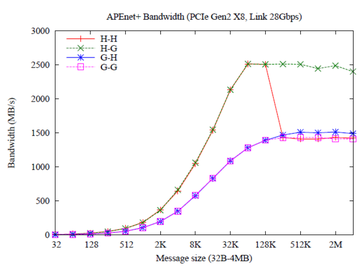




LATENCY (PCIe Gen2 x8)

- APEnet+ G-G latency lowest up to 128 KB
- APEnet+ P2P min latency ~8.2 µs
- APEnet+ staging min latency ~16.8 µs
- MVAPICH/IB min latency ~17.4 µs





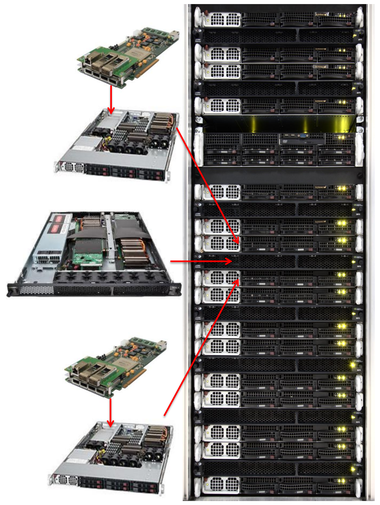
BANDWIDTH (PCIe Gen2 x8)

- CPU Memory Read ~2.8 GB/s
- GPU Memory Read ~1.5 GB/s
- APElink (Off-board) IF ~2.5 GB/s
- GPU Memory Write ≥ 2.5 GB/s
- CPU Memory Write ≥ 2.5 GB/s

QUonG HPC Platform

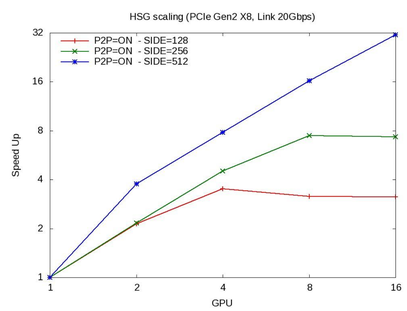
QUantum chromodynamics on Gpu (QUonG) is a comprehensive initiative aimed at providing a hybrid, GPU-accelerated x86_64 cluster with a 3D toroidal mesh topology, able to scale up to $10^4 \div 10^5$ nodes

- Current Status:
  - 16 nodes equipped with APEnet+ board (4x4x1 current topology)
- QUonG Hybrid Computing Node:
  - double Intel Xeon E5620 processor
  - 48 GB System Memory
  - 2x S2075 NVIDIA Fermi GPUs
  - 1 APEnet+ board
  - 40 Gb/s InfiniBand HCA
- Software Environment:
  - CentOS 6.4
  - NVIDIA CUDA 4.2 driver and dev kit
  - OpenMPI and MVAPICH2 MPI libraries available



QUonG with Kepler:

- **Q1/2014: 30 nodes equipped with APEnet+**
- **Few nodes with K20x and K40 for APEnet+ /Kepler integration validation**



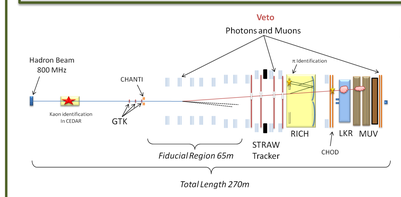
GRAPH500: Traversed Edges Per Second, Strong Scaling, number of graph vertices $|V| = 2^{20}$.

NP	APEnet+	OMPI/IB
1	6.7×10^7	6.2×10^7
2	9.8×10^7	7.8×10^7
4	1.3×10^8	8.2×10^7
8	1.7×10^8	2.0×10^8

NaNet: APEnet+ Customization for HEP Experiments

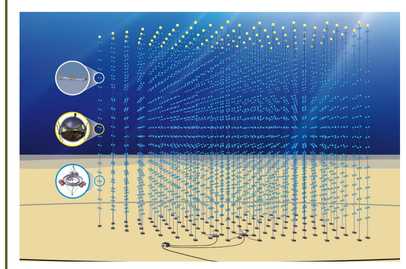
NaNet is an APEnet+ rehaul for real-time data acquisition able to directly relay data from the NIC into the CPU/GPU memory with no intermediate buffering, reusing the APEnet+ GPUDirect implementations. Moreover, it adds a network stack protocol management offloading engine to the logic to avoid OS jitter effects.

NaNet-1 developed in the framework of NA62 Experiment at CERN

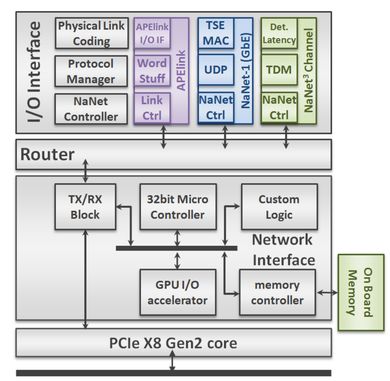


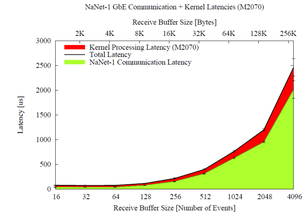
- **NaNet-1**: is a PCIe NIC featuring a standard GbE interface able to directly inject UDP data stream into the memory of a Fermi- or Kepler-class NVIDIA GPU. Moreover, it provides 3 bi-directional APElink channels, adopting a proprietary data transmission word-stuffing protocol.

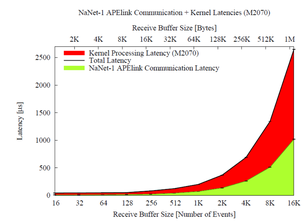
NaNet³ developed in the framework of KM3NeT Neutrino Telescope



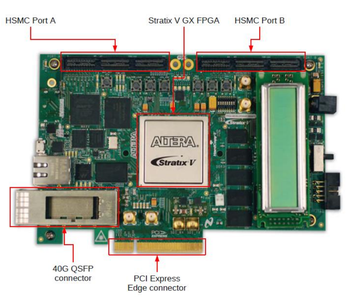
- **NaNet³**: the design employs Altera Deterministic Latency Transceivers with an 8B10B encoding scheme as Physical Link Coding and Time Division MultiPlexing (TDMP) data transmission protocol. Current implementation is being developed on the Altera Stratix V development board with a Terasic SFP-HSMC daughtercard plugged on top and sporting 4 transceiver-based SFP ports.







Future Work



APEnet V5 foresees the re-design of two major hardware logic areas:

- PCIe Gen3 featuring 8.0 Gbps/lane with a 128b130b encoding scheme
- Enhanced embedded transceiver for I/O interface to overcome 40 Gb/s limit.

NaNet-10 will support the more advanced 10-GbE industrial standard exploiting the 40G QSFP connector of Stratix V Dev Kit

Use Cases and References

- **HSG**: Heisenberg Spin-Glass simulation http://dx.doi.org/10.1016/j.jpdc.2012.09.006
- **GRAPH500** : Breadth-First-Search algorithms for graph traversal http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6650959
- **NaNet:** a NIC for real-time trigger systems based on GPUs arXiv:1311.4007
- **DPSNN**: Distributed Polychronous Spiking Neural Networks arXiv:1307.1270
- **KM3Net: Neutrino Telescope** www.km3net.org/home.php
- **Triggu**: Applications of GPUs to online track reconstruction in HEP experiments http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6551422

Contacts

- **The APE project:** http://apegate.roma1.infn.it/APE
- **The EURETILE project:** http://euretile.roma1.infn.it/mediawiki/index.php/Main_Page
- **QUonG initiative coordinator**: piero.vicini@roma1.infn.it
- **NaNet contact**: alessandro.lonardo@roma1.infn.it
- **Presenter contact:** andrea.biagioni@roma1.infn.it

This project was partially funded by the **EURETILE** European FP7 grant 247846