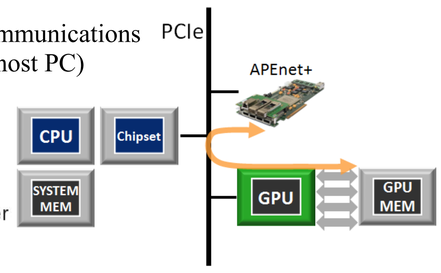
# Architectural improvements and 28nm FPGA implementation of the APEnet+ 3D Torus network for hybrid HPC systems

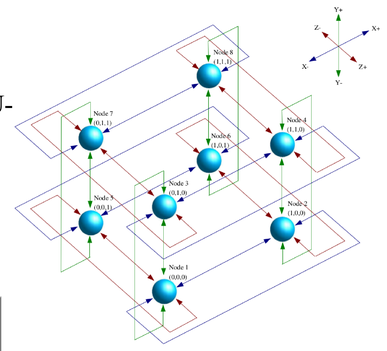
R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, A. Lonardo,
P. S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, P. Vicini
INFN – Istituto Nazionale di Fisica Nucleare
CHEP2013 - October 14 - 18, 2013 - Amsterdam, The Netherlands

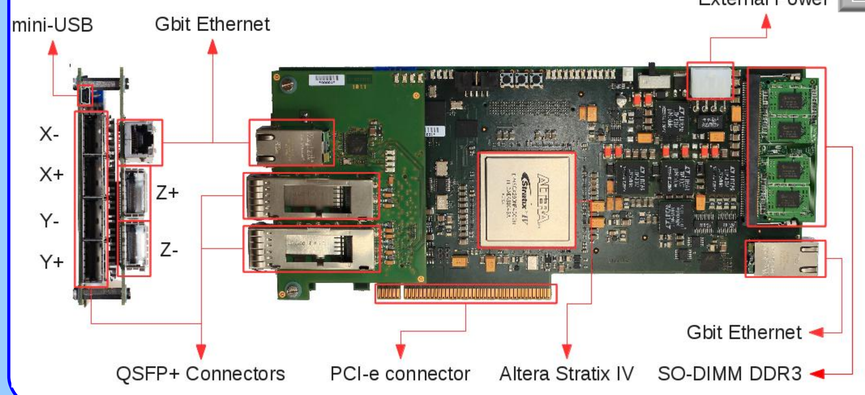
## APEnet+: a brief description

APEnet+ is the high performance, low latency interconnect system 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 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
- transfers are RDMA – CPU is not involved in data movement
- Hardware support for P2P GPUDirect RDMA (for Nvidia GPUs)






### NVIDIA GPUDirect

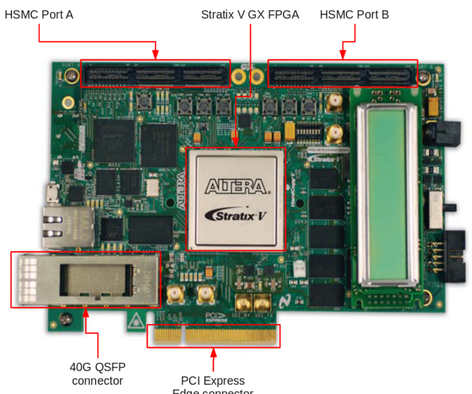
Peer-to-peer between Nvidia Fermi and APEnet+

- Joint development with NVidia.
- APEnet+ has been the first 3rd party device to implement it in hardware.
- 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.

## Designing next generation board

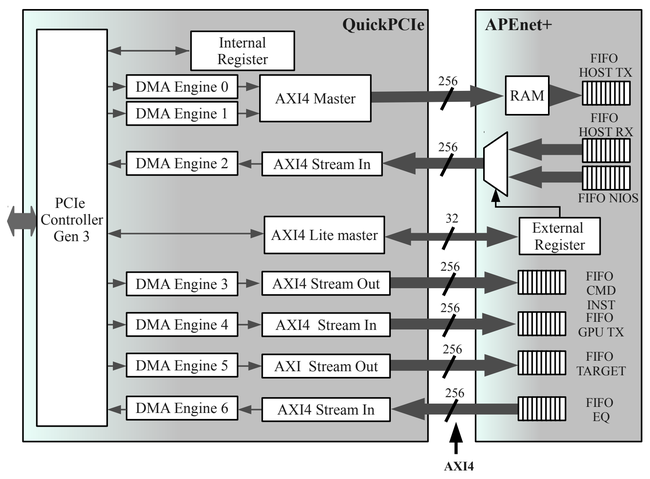
Newer FPGA families are now available on the market, driving re-design of two major hardware logic areas:

- PCIe Gen2 → Gen3 migration to reach ~ 7.9 GB/s raw bandwidth on ×8 lanes towards host.
- Using faster transceivers for Off-board interface to overcome 40 Gb/s limit on the Off-Board links.




On a development board we implemented 3 bi-directional 4 lanes Altera custom PHY links:

- **X channel**
  - Achieved using 40G QSFP+ connector
  - Bandwidth =11.3 Gbps/lane (45.2 Gbps/channel)
  - Measured BER=0.0029 (without equalization, emphasis,.. and with cable of 40 Gbps)
- **Y and Z channels**
  - Implemented on the HSMC interfaces.
  - Bandwidth = 7.8 Gbps/lane (31.2Gbps/channel )




### Gen 3 features

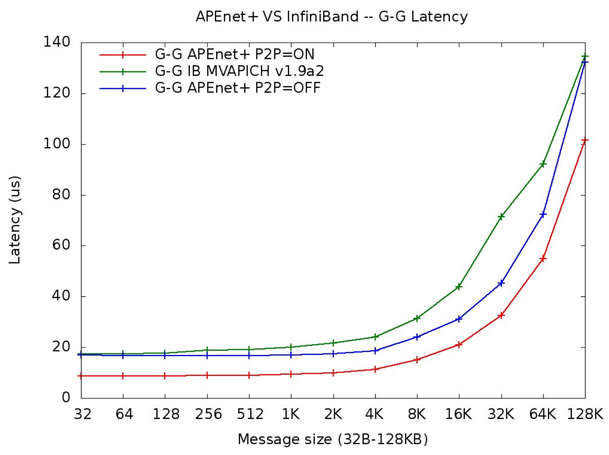
- 8.0 Gbps/lane
- 128/130 bit block encoding/decoding with an overhead of less than 1% (Gen1 and Gen2 overhead is 20%).
- Bus width on backend 256 bit
- Pcie_clk reference 250 MHz
- Bandwidth 7.877 GB/S
- PCIe core backend is AXI4 based: need to redesign APEnet internal SoC system

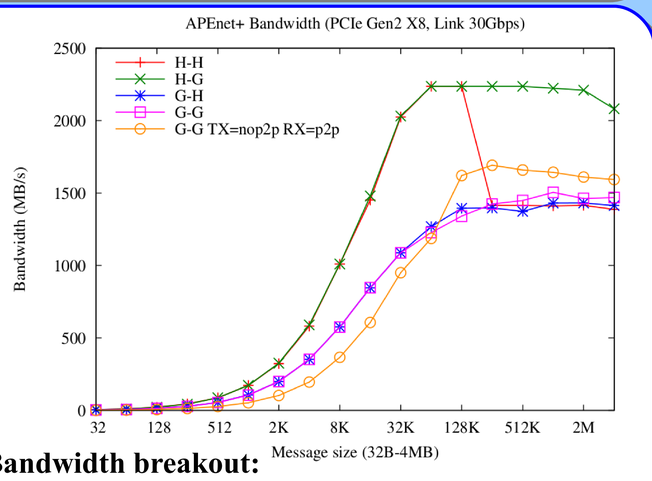
## Latency & Bandwidth synthetic tests

We updated latency and bandwidth measures thanks to the architectural improvements described. Significant performance gains are measured on bandwidth tests with respect to previously published results.

### Latency comparison:

- APEnet+ G-G latency is lower up to 128KB
- APEnet+ P2P latency = ~8.2 us
- APEnet+ staging latency = ~16.8 us
- MVAPICH/IB latency = ~17.4 us




### Bandwidth breakout:

- CPU Memory Read Bandwidth = ~2.4 GB/s
- GPU Memory Read Bandwidth = ~1.5 GB/s
- Off-Board Link Bandwidth = ~2.2 GB/s (@350 MHz)
- GPU Memory Write Bandwidth = ~2.2 GB/s
- CPU Memory Write Bandwidth = ~2.2 GB/s

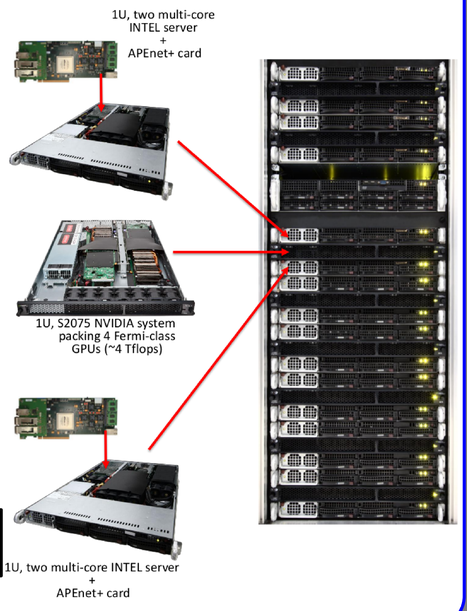
## Results on QUonG HPC platform

QuonG is our hybrid 16 nodes x86_64/dual GPU cluster with a 4 × 4 × 1 APEnet+ torus network, for testing, development and production run.
The following applications have been ported over the QUonG/APEnet+ HW with promising results:

- DPSNN: Distributed Polychronous Spiking Neural Network simulation using Izhikevich neuron model
- NaNet: A Custom NIC for Low-Latency, Real-Time GPU Stream Processing
- Triggu: Applications of GPUs to online track reconstruction in HEP experiments
- GRAPH500 : Breadth-First-Search algorithms for graph traversal
- HSG: Heisenberg Spin-Glass simulation

Talk by A. Lonardo at CHEP2013
"*NaNet: a low-latency NIC enabling GPU-based, real-time low level trigger systems*"
on 15 Oct from 13:30 to 13:50

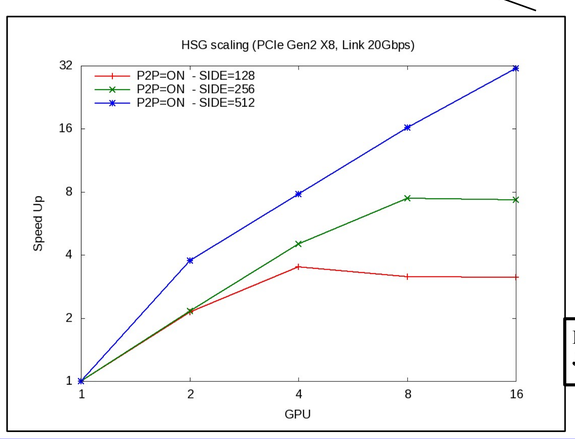
Talk by S. Amerio at CHEP2013
"*Many-core applications to online track reconstruction in HEP experiments*"
on 17 Oct from 14:10 to 14:30




| NProc | Run on APEnet+ | Run on IB |
|---|---|---|
| 1 | $6.7 \times 10^7$ | $6.2 \times 10^7$ |
| 2 | $9.8 \times 10^7$ | $7.8 \times 10^7$ |
| 4 | $1.3 \times 10^8$ | $8.2 \times 10^7$ |
| 8 | $1.7 \times 10^8$ | $2.0 \times 10^8$ |

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

Bernaschi et al. "*Breadth first search on APEnet+*"
**IAAA Workshop on Irregular Applications: Architectures&Algorithms**

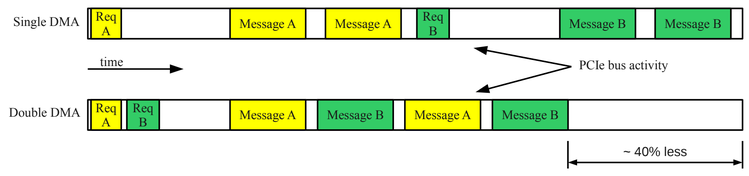



Bernaschi et al. "*Benchmarking of communication techniques for GPUs*"
**Journal of Parallel and Distributed Computing**

## Architectural improvements

APEnet+ is able to outperform IB for small-to-medium message size when using GPU peer-to-peer. For large message size host memory staging techniques are still winning, also due to better bandwidth of latest IB cards.
We worked in several parts of our architecture to improve overall performances.

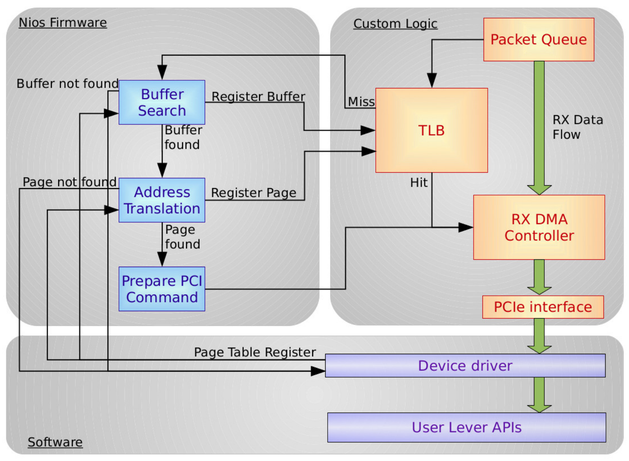
- Transmitter side speed-up: Double DMA Channel




Doubling the number of transaction request on PCIe bus allows an efficiency gain in multiple data transactions (40% less time **measured**).

will be presented at ReConFigurable Computing Conference 2013

- Receiver side speed-up: on-board memory management moved to HW functions.




A novel implementation of a Translation Look-Aside Buffer (TLB) has been developed, to accelerate virtual-to-physical address translation at hardware level.

will be presented at Field Programmable Technology Conference 2013

- Off-board Interface with higher efficiency.
  Data Link Layer protocol optimization depending on some HW structural parameters.

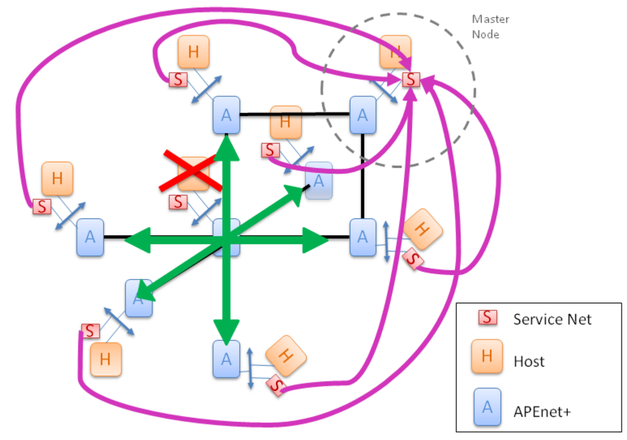
| FIFO Depth | Efficiency | BW@28Gbps | BW@34Gbps |
|---|---|---|---|
| 512 | 0.595 | 1666 MB/s | 2023 MB/s |
| 1024 | 0.784 | 2195 MB/s | 2665 MB/s |
| 2048 | 0.862 | 2414 MB/s | 2931 MB/s |
| 4096 | 0.898 | 2514 MB/s | 3060 MB/s |

presented at TWEPP Conference 2013

## Studies on Fault Awareness

Fault awareness is the first step when applying a Fault Tolerance technique in HPC (e.g. task migration, checkpoint/restart,...).
**On the QUonG platform, thanks to some APEnet+ hw features, each node is able to be aware of faults and critical events occurring to its components and to components of its neighbor nodes.**

- Even in case of multiple faults no area of the mesh can be isolated and no fault can remain undetected at global level.
- At the core of this approach, named **LO|FA|MO** (LOcal FAult MOnitor), there is a lightweight mutual watchdog protocol between the host node and APEnet+ and the 3D network topology.
- The time from the fault occurrence to the global fault awareness is dominated by the watchdog period: @WD 500 ms, Ta = 0.9 s.
- In the time range of interest for HPC (watchdog period 1-$10^3$ ms), the addition of LO|FA|MO features has no impact on data transfer latency.




## Contacts

**INFN Roma, Italy**, email: roberto.ammendola@roma2.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.