Multi-dimensional Torus networks for current and next generation HPC systems: APENet+ status and perspectives

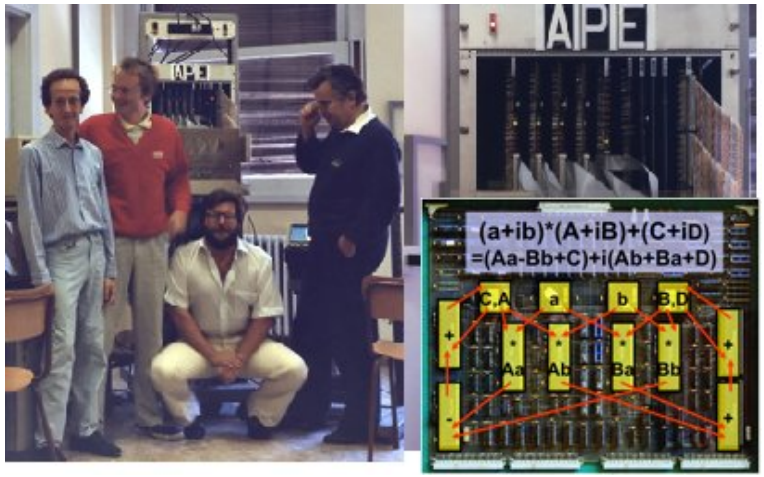
Roberto Ammendola
on behalf of the APE Group

Istituto Nazionale di Fisica Nucleare, Sezione Roma Tor Vergata

Supermassive Computations in Theoretical Physics,
Trento – February 12, 2015

The legacy of APE research

- Array Processor Experiment is a 25 years old project at INFN.
- Developing fully custom and hybrid parallel computing machines.
- Research advances in floating point engines, interconnection networks, system integration, compilers, software libraries, ...
- Since 2003 we started exploring a way to use commodity hardware (a.k.a. clusters) with a custom interconnect.
- Programmable chip are the key technology for this work.



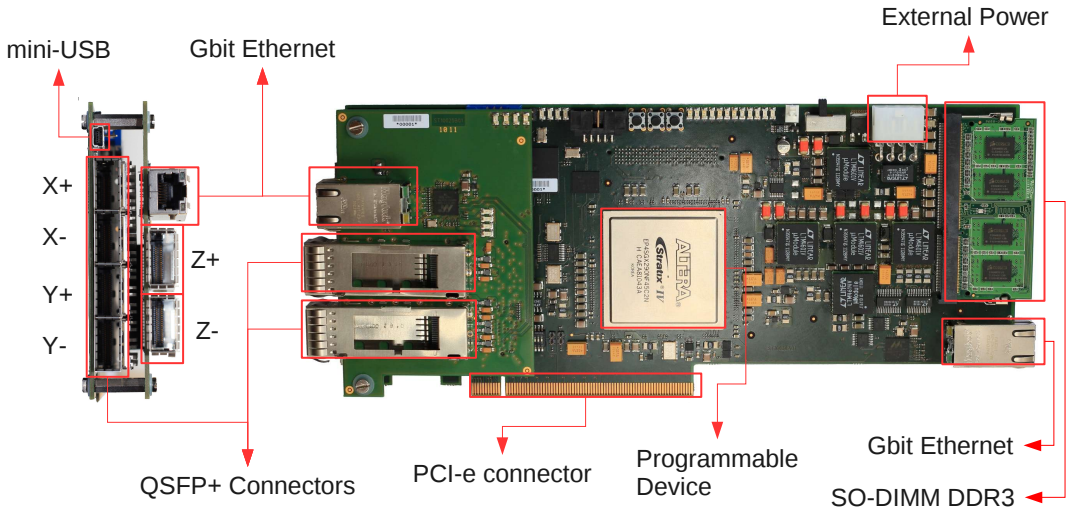
1988

2006

APENet+ Custom interconnect

Aimed for:

- low latency
- high bandwidth
- CPU offloading

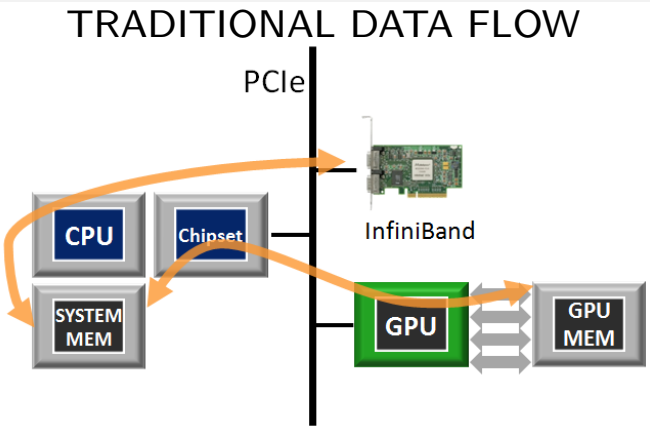


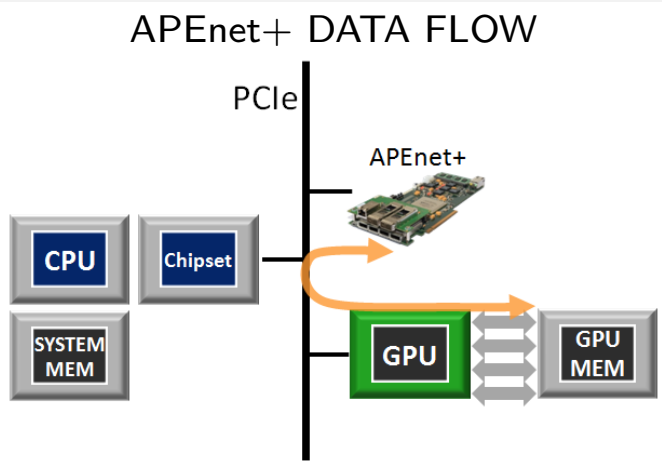
Altera Stratix IV based NIC:

- PCIe Gen2 x8
- 6 bidirectional Off-Board links
- 40 Gb QSFP+ interconnect fabric
- Nios microcontroller
- RDMA communication paradigm
- GPU Direct capable (NVidia peer-to-peer protocol)

Re-using IP cores also in High Energy Physics and Particle Physics data acquisition experiments.

GPU Direct feature





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 buffers
- Issue network transfer command on this memory buffers
- and vice-versa on the receive side

P2P between Nvidia Fermi and APEnet+

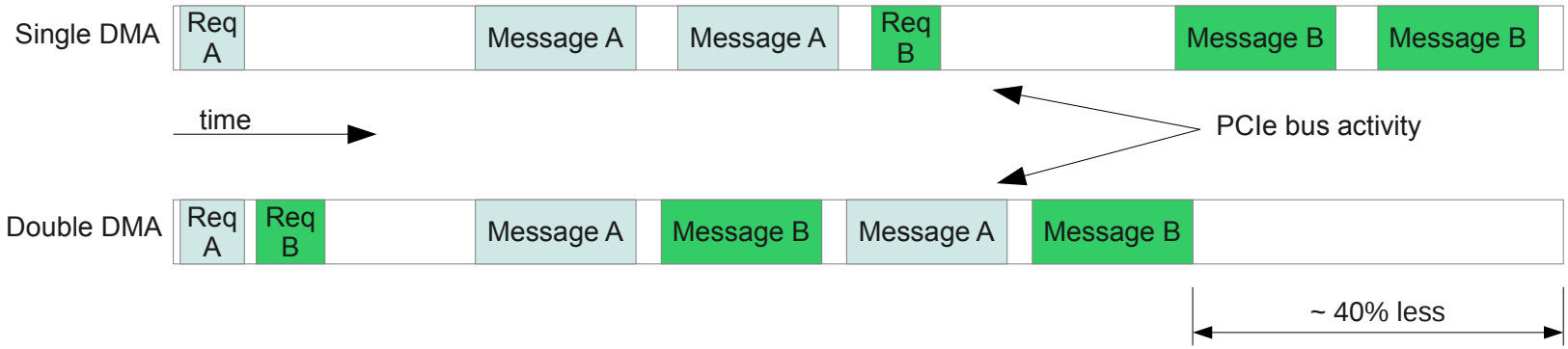
- No bounce buffers on host. APEnet+ can target GPU memory with no CPU involvement
- GPUDirect allows direct data exchange on the PCIe bus
- Real zero copy, inter-node GPU-to-host, host-to-GPU and GPU-to-GPU
- Latency reduction for small messages

Publications

- *APEnet+: a 3D Torus network optimized for GPU-based HPC Systems* Journal of Physics: Conference Series 396 (4), 042059
- *GPU peer-to-peer techniques applied to a cluster interconnect* Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International
- INFN Patent n. WO2013136355 A1 *Network interface card for a computing node of a parallel computer accelerated by general purpose graphics processing units, and related inter-node communication method*

IP Improvements: Gaining efficiency on the PCIe bus

- we noticed that effective bandwidth on data transactions was quite low compared to the theoretical one (50%)
- this is due to the time elapsed between issuing a request on the PCIe bus and its completion
- the card must be able to manage more than one outstanding request on the PCIe bus
- we needed to implement two concurrent DMA engines fed by a prefetchable command queue
- we estimated an efficiency gain up of to 40% in time

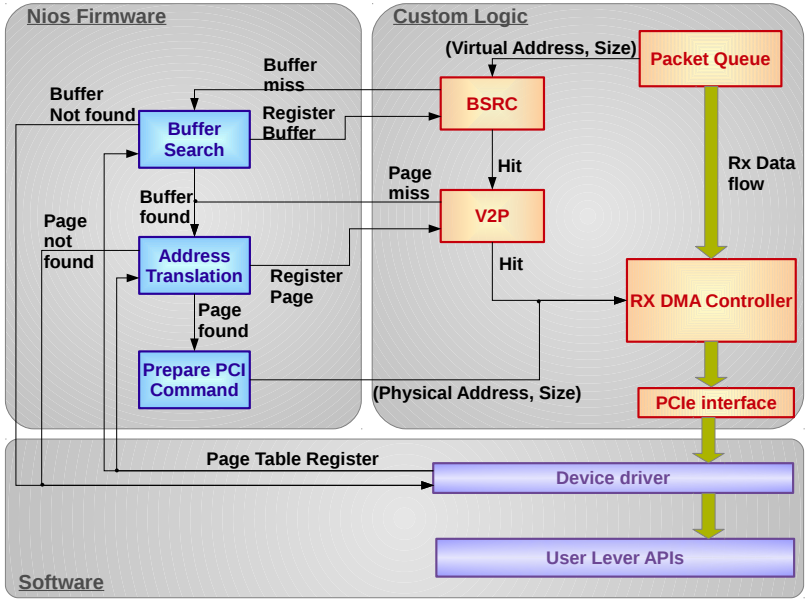


Publications

- *A 34 Gbps Data Transmission System with FPGAs Embedded Transceivers and QSFP+ Modules* Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE
- *APEnet+ 34 Gbps data transmission system and custom transmission logic* Journal of Instrumentation 8 (12), C12022
- *Architectural improvements and 28 nm FPGA implementation of the APEnet+ 3D Torus network for hybrid HPC systems* Journal of Physics: Conference Series 513 (5), 052002
- *Analysis of performance improvements for host and GPU interface of the APENet+ 3D Torus network* Journal of Physics Conference Series 06/2014; 523(1):012013
- *Architectural improvements and technological enhancements for the APEnet+ interconnect system* JINST 10 C02005

IP Improvements: Fast Virtual-to-Physical Address translation

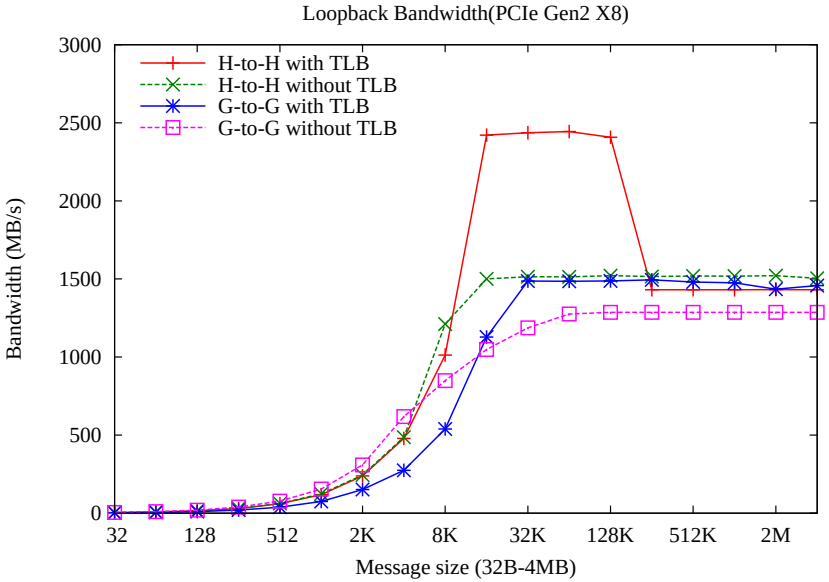
- Virtual to physical address translation is necessary in order to dispatch data payloads in the correct physical memory areas
- This task was initially executed by the Nios II embedded processor
- A novel implementation of a Translation Look-Aside Buffer (TLB) has been developed on the FPGA
- The TLB block can store a limited amount of page entries and, in case of page hit, the Nios II processor is completely bypassed
- A speedup of up to 60% in bandwidth on synthetic benchmarks has been measured

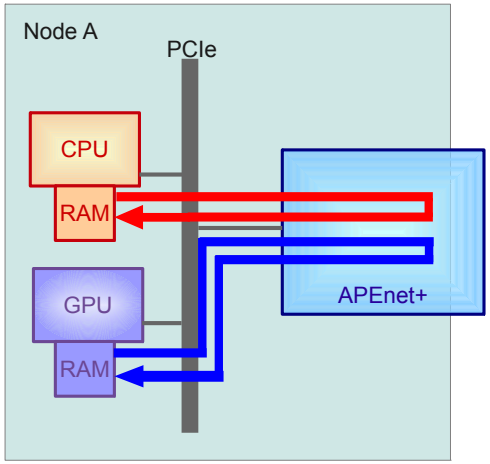


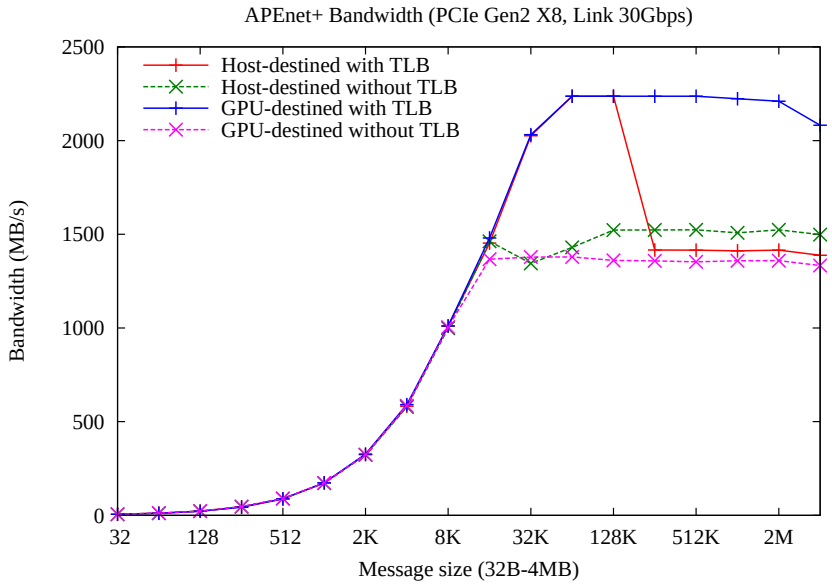
Publications

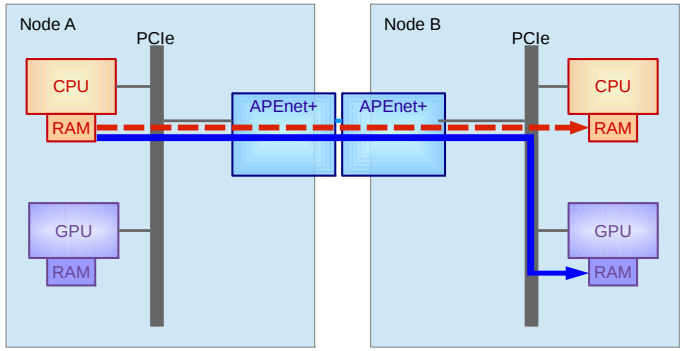
- *Virtual-to-Physical address translation for an FPGA-based interconnect with host and GPU remote DMA capabilities* Field-Programmable Technology (FPT), 2013 International Conference on, 58-65
- *ASIP acceleration for virtual-to-physical address translation on RDMA-enabled FPGA-based network interfaces* Future Generation Computer Systems, 2015

How TLB impacted on performances





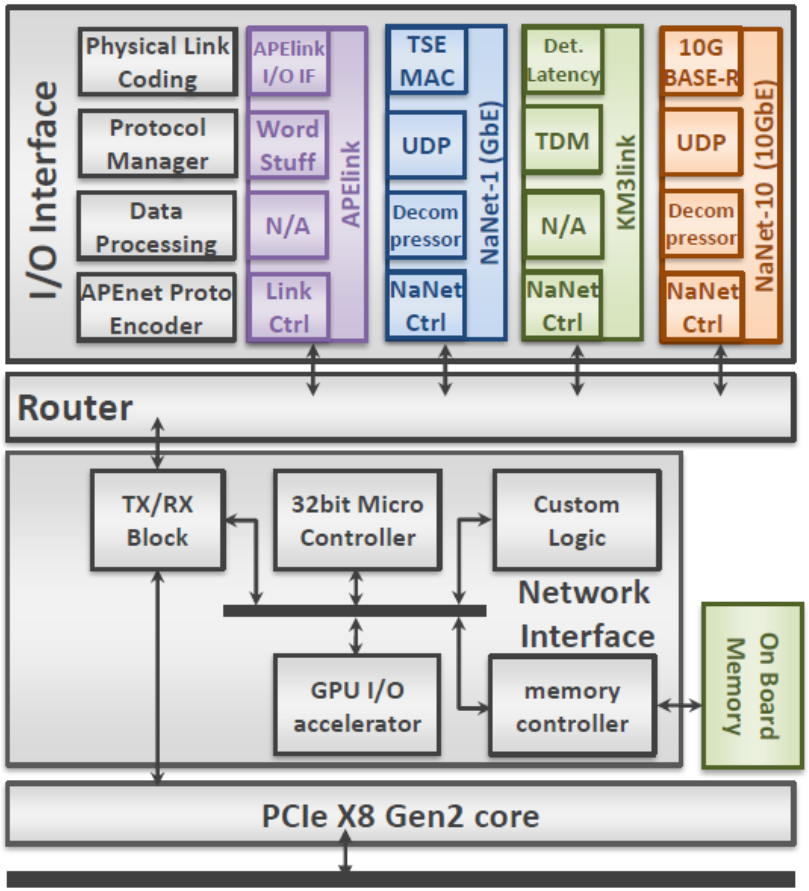




Making APENet+ a multi-purpose base platform

The APENet+ design integrates a number of parametric and reconfigurable IPs.
This modular design allowed a straightforward implementation of different APENet+ designs on more than one board model:

- APENet+ early prototype (3 links on StratixIV)
- APENet+ production board (6 links on Stratix IV)
- APENet+ with PCIe Gen3 prototype (1 link on Stratix V)
- NaNet-1 for low latency data acquisition through 1 Gb Ethernet for NA62 Experiment (1 link on Stratix IV)
- NaNet-10 upgrade to 10 Gb for NaNet-1 (4 links on Stratix V)
- NaNet^3 customization for KM3 HEP with deterministic latency feature



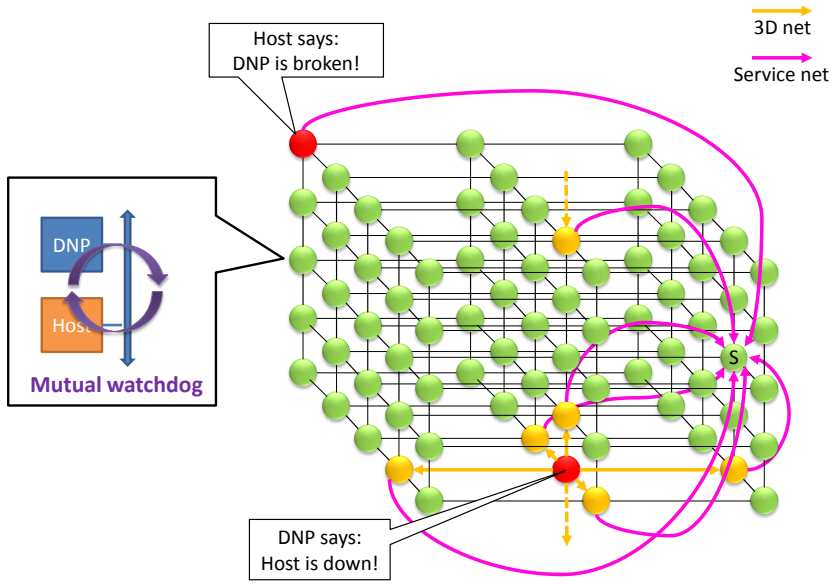
Publications

- *Design and implementation of a modular, low latency, fault-aware, FPGA-based network interface* Reconfigurable Computing and FPGAs (ReConFig), 2013 International Conference on; 12/2013

Studies on Fault Awareness

- Systemic fault awareness as first step for fault and critical events management on distributed systems

- Local Fault Monitor (LO|FA|MO) approach:
  - Fault detected locally to devices
  - Mutual watchdog mechanism between the Host (CPU) and the DNP (Network Interface): they monitor each other for liveness and faults
  - information about the nature of fault is propagated up to higher levels in the system hierarchy
  - double path for diagnostic messages via 3D network and service network ensures systemic awareness
  - Selected nodes (Supervisors) have the complete picture of the health status of the system



- LO|FA|MO is a lightweight approach: does not alter the system performance
- Towards automatic fault management: Supervisor could analyze the situation and initiate a reaction (e.g. migration of tasks)

Publications

- *A heterogeneous many-core platform for experiments on scalable custom interconnects and management of fault and critical events, applied to many-process applications* Vol. II, 2012 technical report
- *A hierarchical watchdog mechanism for systemic fault awareness on distributed systems* Future Generation Computer Systems, 2015
- *LO-FA-MO: Fault Detection and Systemic Awareness for the QUonG Computing System* 2014 IEEE 33rd International Symposium on Reliable Distributed Systems (SRDS)

The QUonG cluster

QuonG is our hybrid 16 nodes x86_64 dual GPU cluster with a 4 x 4 x 1 APEnet+ torus network, for testing, development and production run.

Current Status

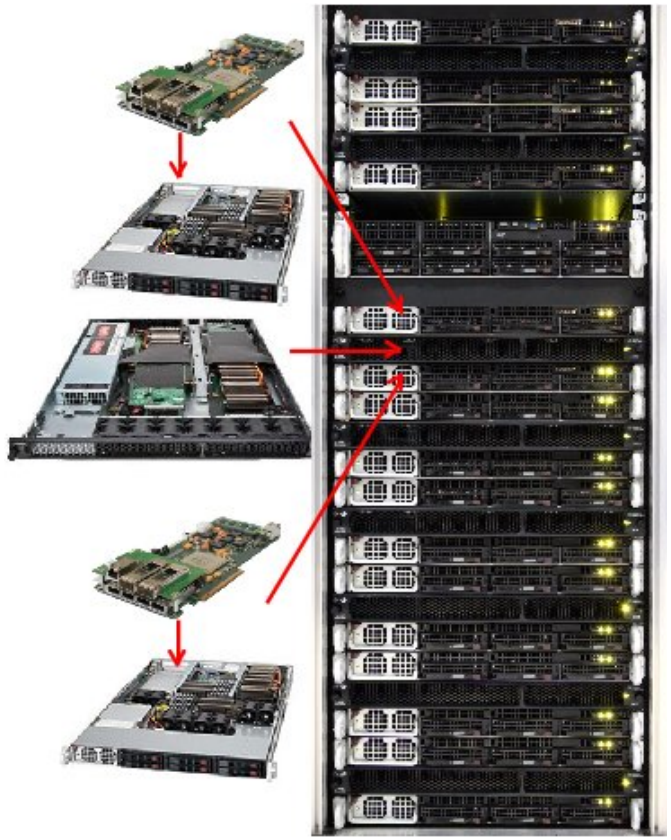
- 16 nodes equipped with APEnet+ board

QUonG Hybrid Computing Node

- double Intel Xeon E5620
- 48GB System Memory
- 2x S2075 NVIDIA Fermi GPUs
- 1 APEnet+ board
- 40 Gb/s InfiniBand HCA

Software Environment

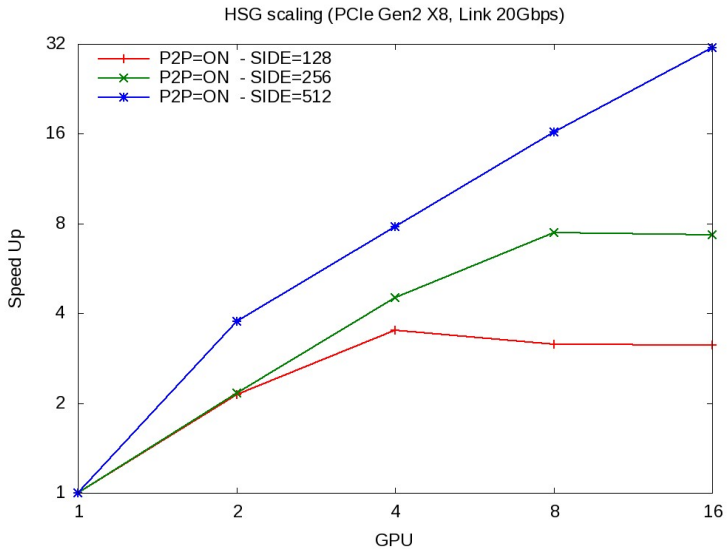
- CentOS 6.4
- NVIDIA CUDA 6.5 driver
- OpenMPI and MVAPICH2 MPI



Results on QUonG HPC platform

The following applications have been ported over the QUonG with APEnet+ HW with promising results:

- DPSNN: Distributed Polychronous Spiking Neural Network simulation using Izhikevich neuron model
- GRAPH500: Breadth-First-Search algorithms for graph traversal
- HSG: Heisenberg Spin-Glass simulation



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

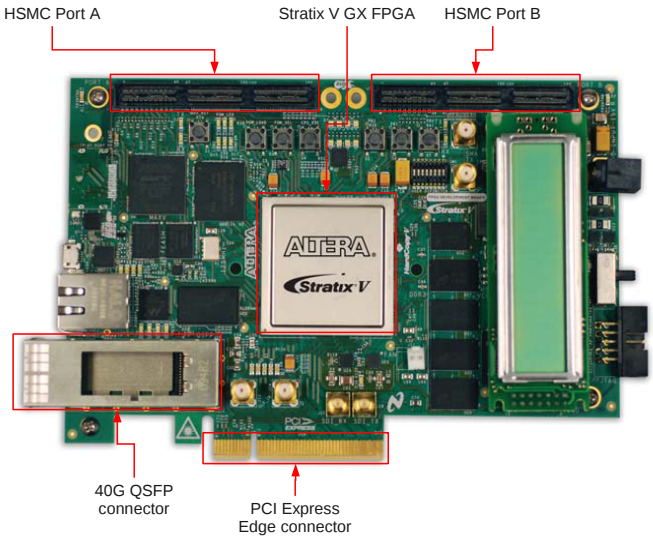
NP	APEnet+	OMPI/IB
1	6.7×10^7	6.2×10^7
2	9.8×10^7	7.8×10^7
4	1.3×10^8	8.2×10^7
8	1.7×10^8	2.0×10^8

References

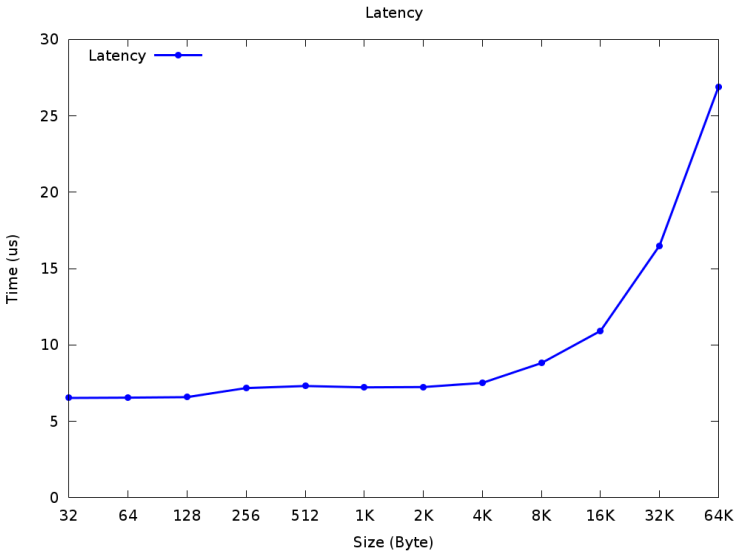
- Bernaschi et al. *Breadth first search on APEnet+* IAAA Workshop on Irregular Applications: Architectures&Algorithms
- Bernaschi et al. *Benchmarking of communication techniques for GPUs* Journal of Parallel and Distributed Computing

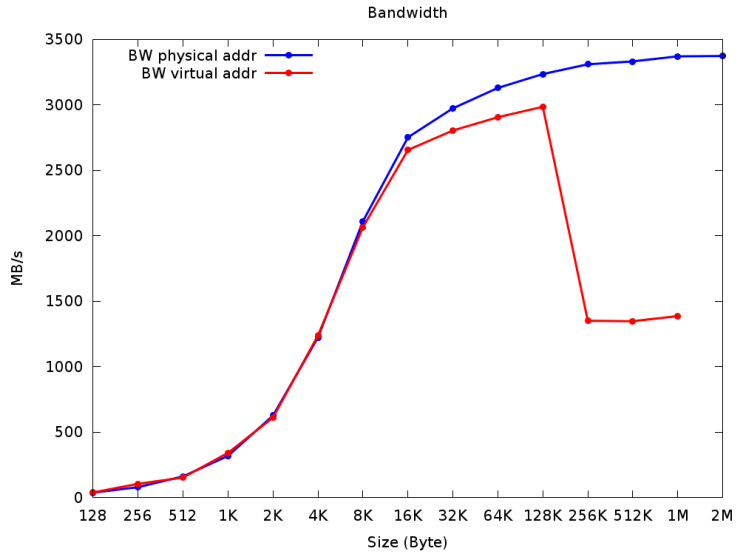
Enhancing APENet+: PCIe Gen3

- Upgrading FPGA technology to 28nm (Altera V generation)
- Take advantage of faster transceivers to support 40+ Gbps
- Increase host bandwidth with PCIe version 3.0 protocol



Preliminary synthetic benchmarks: latency and bandwidth

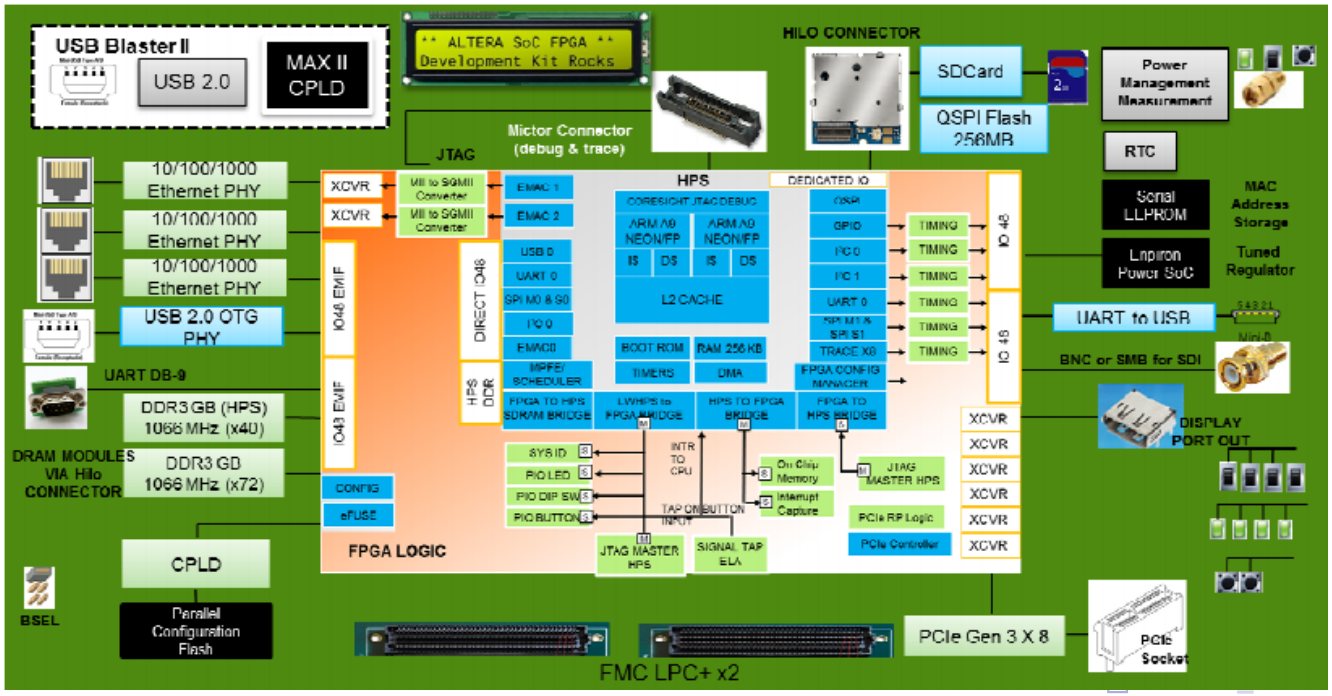




Work in progress and perspectives

Many concurrent developments are on-going:

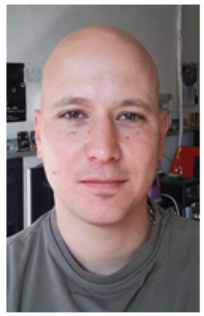
- PCIe Gen3 implementation with device driver redesign
- multi-channel 10 Gb with UPD/IP protocol for KM3 and NA62 collaborations
- MPI-level library optimizations
- studies on futher enhancements:
  - hardware multicast and all-to-all communications
  - hardware support for many-tasks access
  - enable byte-wide data transfers
  - enhanced TLB with embedded memory controller
- ARM integration (with Arria10 SoC FPGA soon arriving)



Thank you! Questions or comments?


Roberto Ammendola


Andrea Biagioni


Ottorino Frezza


Francesca Lo Cicero


Alessandro Lonardo


Michele Martinelli


Pier Stanislao Paolucci


Elena Pastorelli


Davide Rossetti


Francesco Simula


Laura Tosoratto


Piero Vicini