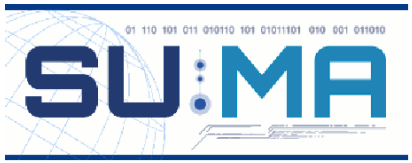
INFN Hardware developments on "GPU Direct" Technology in interconnect systems for computing and supercomputing solutions.

Roberto Ammendola

Istituto Nazionale di Fisica Nucleare, Sezione Roma Tor Vergata

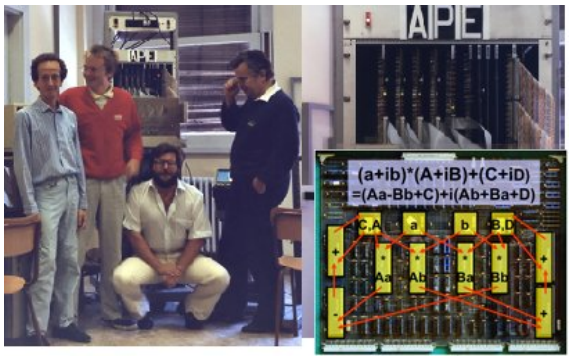
Workshop Commissione Calcolo e Reti 2013, Genova – 28 May 2013

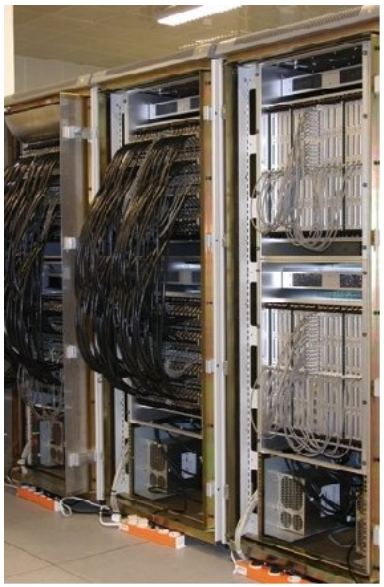


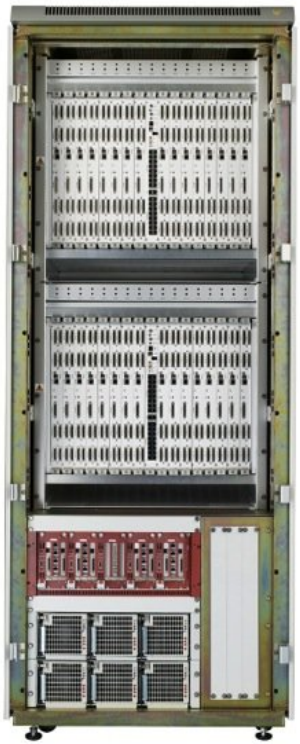
A little bit of history: from APE1 to apeNEXT

APE is a 25 years old project

- MPP (APE1, APE100, APEmille, apeNEXT) & PC Cluster interconnection network (apeNET)
- FP Engine optimized for application + Smart dedicated 3D Torus interconnection network
- APE1 (1988) 1GF, chipset Weitek
- APE100 (1992) 25GF, SP, REAL "Home made" VLSI processors
- APEmille (1999) 128GF, SP, Complex Italy+France+Germany collaboration
- apeNEXT (2004) 800GF, DP, Complex







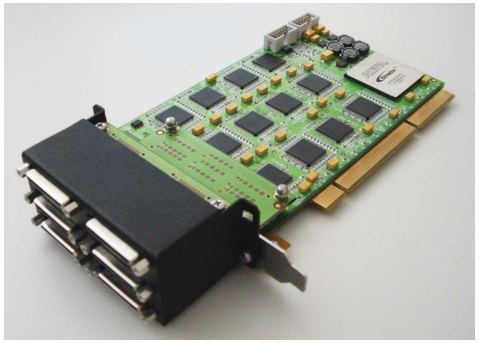
A little bit of history: APENet

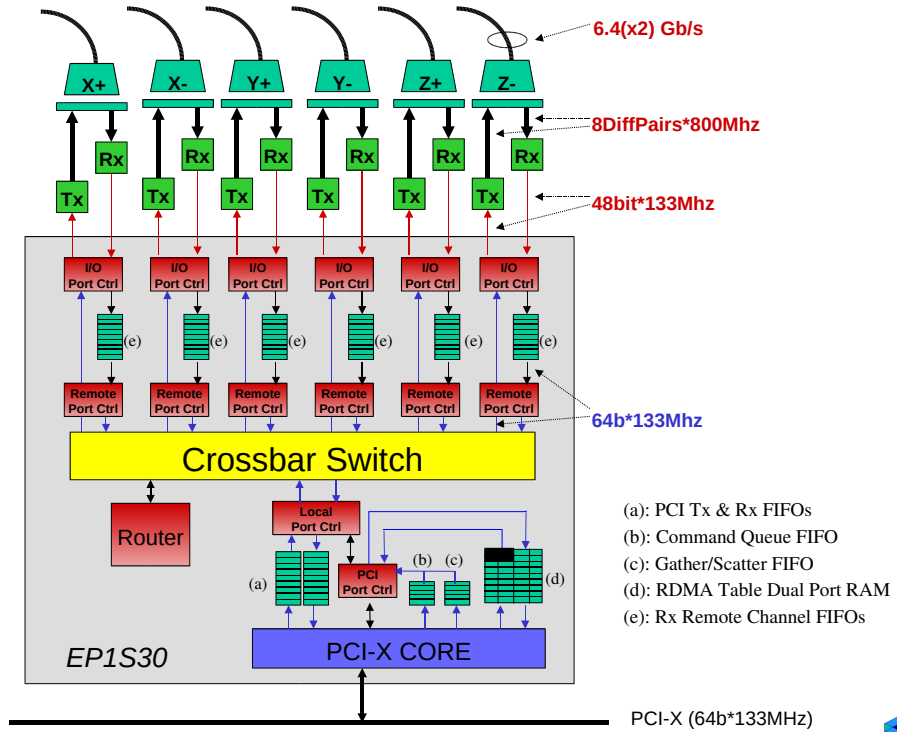
APEnet: PC Cluster 3-d torus network

- Integrated routing and switching capabilities
- High throughput, low latency, "light-weight" protocol
- PCI Interface, 6 Links full-bidir on torus side

History

- 2003-2004: APEnet V3 (PCI-X)
- 2005: APEnet V3+ same HW with RDMA API
- 2006-2009: APEnet goes embedded DNP, D(istributed) N(etwork) Processor EU SHAPES project co-development
- 2011: APEnet+ PCI Express, enhanced torus links





Is there still room for custom technology?

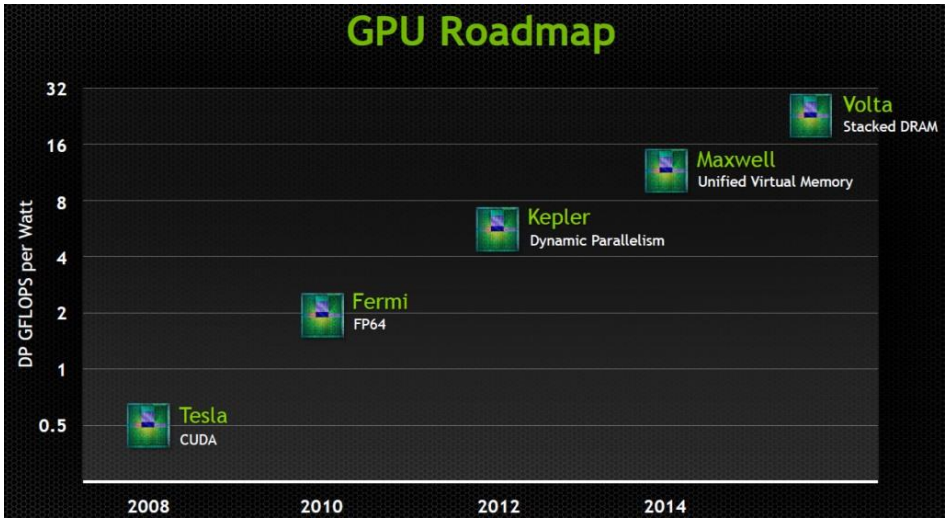
Time changes but some facts are still true.

- The hunger for floating point computing power remains unchanged
- Towards the ExaFlops, main keywords remain unchanged
  - Floating Point Engines allowing efficient execution of scientific applications $\longrightarrow$ high ratio flop/watt
  - Smart and efficent specialized interconnection system to scale up to systems made of huge number of computing nodes (100-10000-100000-...) $\longrightarrow$ 3D Torus network [facts: Bluegene network, Cray Gemini bought by Intel]
- NRE costs for custom developments of ASICs, systems, networks reached absurd high levels

Can we use commodities technologies to meet the demands of computing power of modern scientific applications?

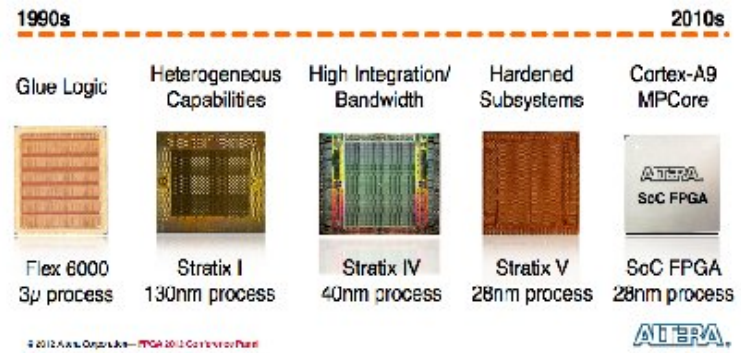
Peta(Exa)Flops scale enabling technologies: GPGPU

- General Purpose Graphic Processing Unit: impressive peak performance ($N \times$ *TFlops* per chip)
- Videogames market i.e. 10 G$/yr unified gaming and HPC chip architectures
- Architecture and characteristics fit with HPC scientific application (LQCD as an example) requirements
  - Many-Core (»100) SIMD-like architecture
  - High local memory bandwidth (140 GB/s $\rightarrow$ 500 GB/s)
  - "Green" and cost effective
- Aggressive but (really!) feasible roadmap: much room for performance scaling
  - Easy peak performance scaling allowed by "tiled" architectures
  - New features added generation by generation
  - Adoption of new technologies to improve performance



(Multi)PetaFlops scale enabling technologies: FPGA

- High-end FPGA-based systems are the ideal hardware to build custom network
- Most complex electronic devices leveraging on silicon process improving and state-of-the-art technologies
- Current devices (28nm) support Tflops, (multi)Terabits I/O bandwidth, hardIP uP cores
  - Dual ARM @800MHz
  - O(1) transceivers @28gbps, O(10) transceivers @10-14 gbps, O(100) transceivers @1-5 gbps
  - PCIe Gen1/2/3, 10G/40G/100G Ethernet, Serial RapidIO, CPRI (Fixed latency)
- Testbed for future interconnection technologies: Avago MicroPod up to 120gb/s full bidir + Altera
- Possibility to use OpenCL programming model on FPGAs



APEnet+ at a glance

3D Torus network

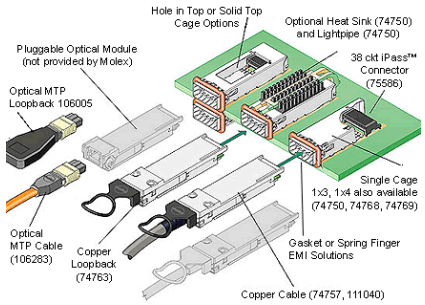
- ideal for large-scale scientific simulations (domain decomposition, stencil computation, ...)
- scalable (APENEt+ today up to 32K nodes)
- Cost effective: no external switches! 1 card+3 cables

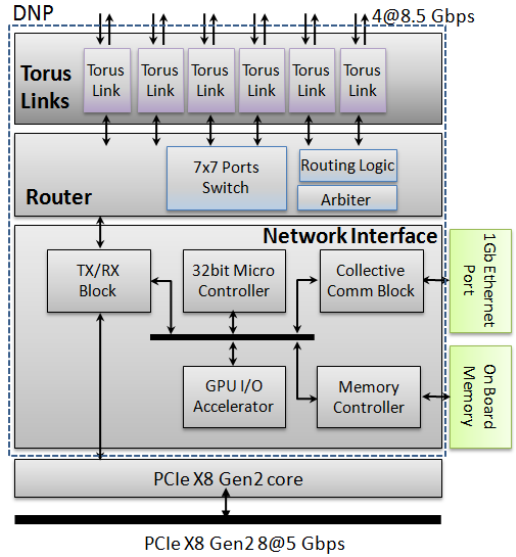
APEnet based on INFN DNP

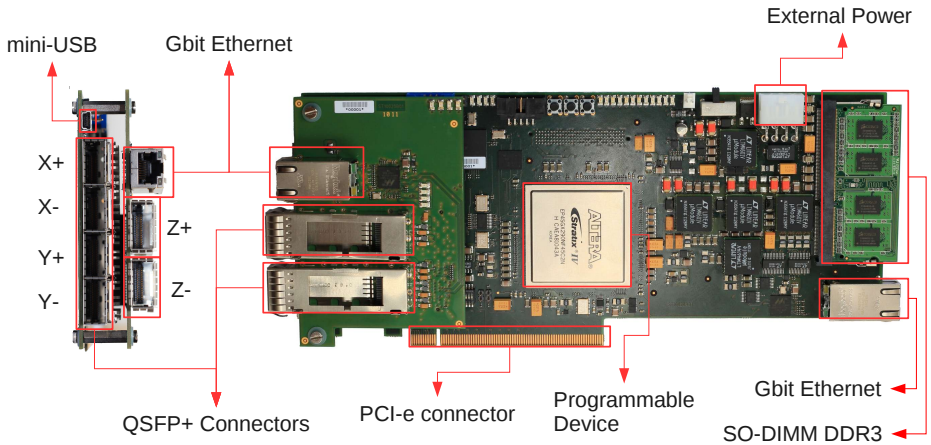
- RDMA: Zero-copy RX & TX !
- Small latency & high bandwidth

APEnet+ card:

- FPGA based (ALTERA Stratix IV)
- 6 full-bidirectional links up to 68 Gbps raw (400 Gbps)
- PCIe X8 Gen2 in X16 slot, peak BW 4+4 GB/s
- Network Processor, off-loading engine integrated in the FPGA
- Zero-copy RDMA host interface
- GPU Direct interface
- Industry standard QSFP+ cabling: Copper (passive/active), optical







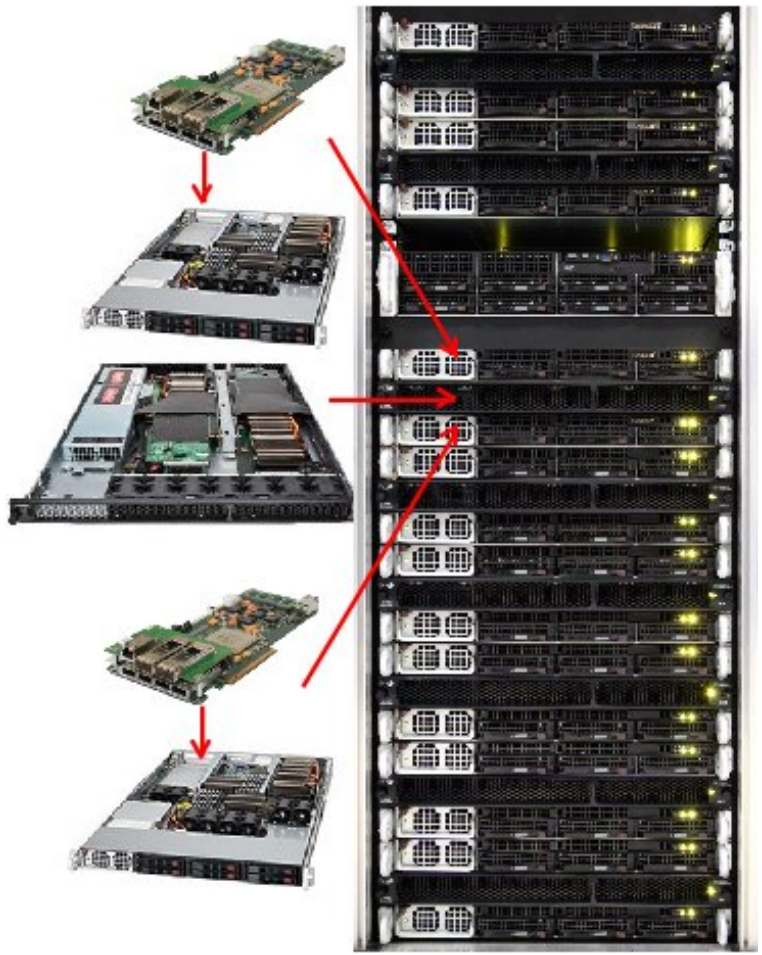
QUonG: GPU+3D Network FPGA-based

QUonG (QUantum chromodynamics ON Gpu) is a comprehensive initiative aiming to deploy an GPU-accelerated HPC hardware platform mainly devoted to theoretical physics computations.

- Heterogeneous cluster: PC mesh accelerated with high-end GPU (Nvidia) and interconnected via 3-D Torus network
- Added value:
  - tight integration between accelerators (GPU) and custom/reconfigurable network (DNP on FPGA) allows latency reduction and computing efficiency gain
  - Huge hardware resources in FPGA to integrate specific computing task accelerators (ASIP, OpenCL
- 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, Laser-plasma interactions, BioComputing, Complex systems, . . .)

QUonG assembly

- QUonG Hybrid Computing Node:
  - Intel Xeon E5620 double processor
  - 48 GB System Memory
  - 2 S2075 NVIDIA Fermi GPU
  - 1 APEnet+ board
  - 40 Gb/s InfiniBand Host Controller Adapter
- QUonG Elementary Mechanical Unit:
  - 3U Sandwich:
    - 2 Intel dual Xeon servers
    - 4 NVIDIA Tesla M2075 GPU
  - 2 Vertex on the APEnet+ 3d network
- Software Environment:
  - CentOS 6.3
  - NVIDIA CUDA 4.2 driver and dev kit
  - OpenMPI and MVAPICH2 MPI available
- Q2 2013: 16 nodes connected by the APEnet+ (4x2x2)
- Addition of few Tflops of Kepler GPUs during 2013



GPU Direct in APEnet+

APEnet+ is 1st non-NVidia device to implement Fermi P2P protocol

Peer-to-Peer means:

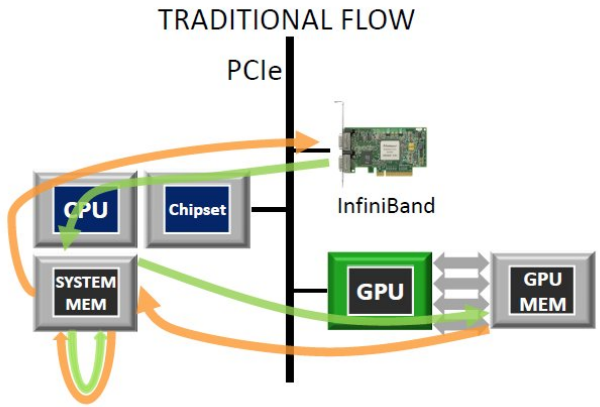
- Data exchange on the PCIe bus
- No bounce buffers on host

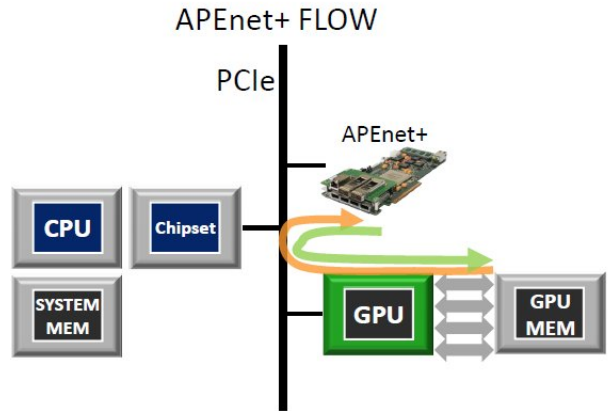
APEnet+ P2P support

- cutting-edge HW/SW technologies developed jointly with Nvidia
- APEnet+ board acts as a peer
- APEnet+ board can read/write directly GPU memory

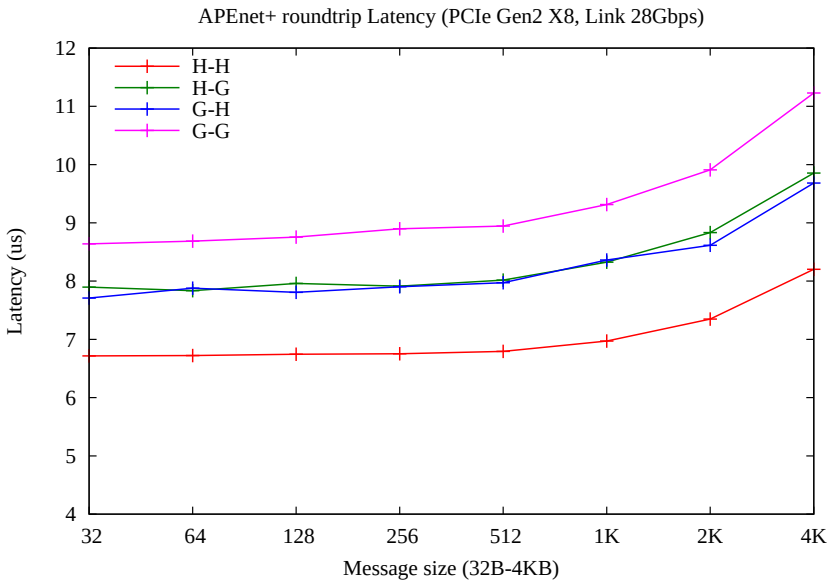
Direct GPU access

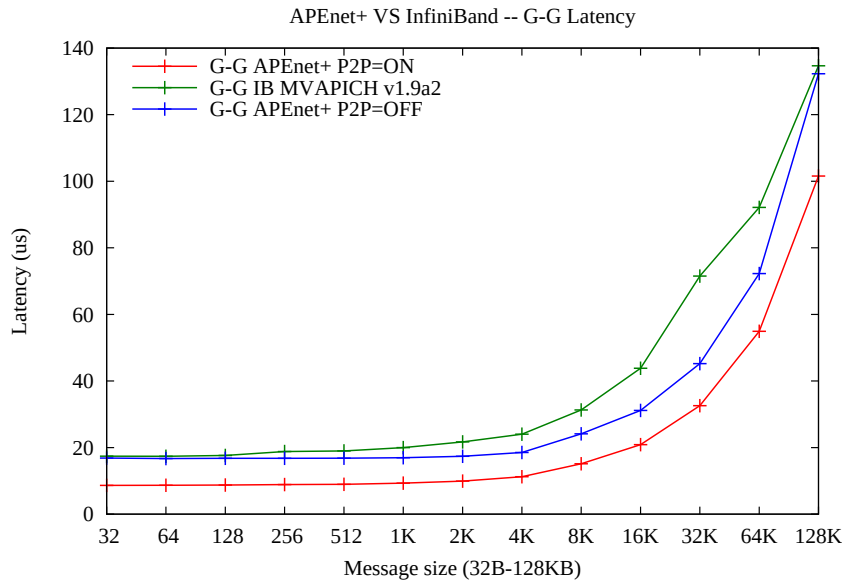
- Specialized APEnet+ HW block
- GPU initiated TX
- Latency saver for small size messages





P2P effects on latency

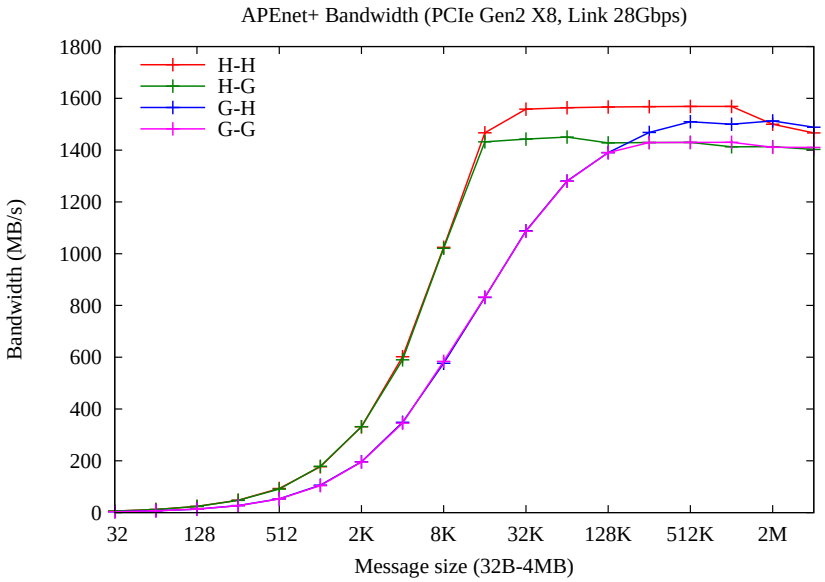


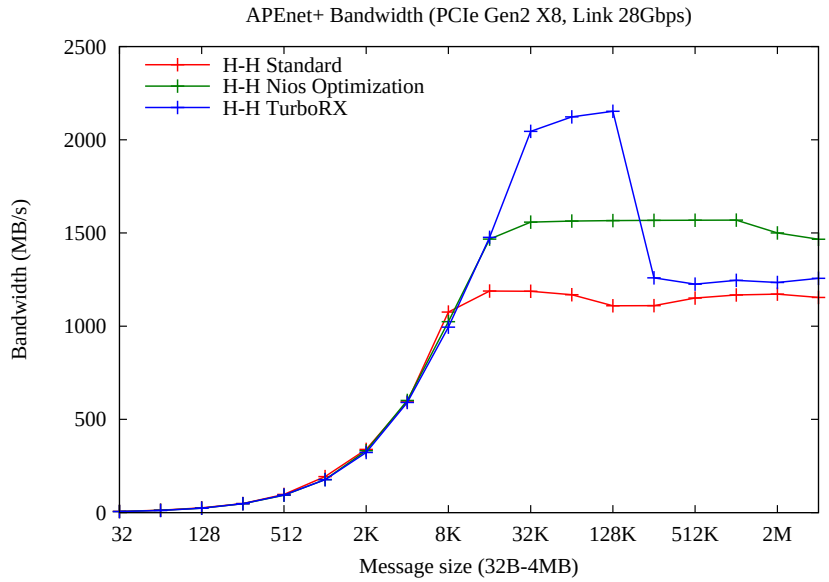


- APEnet+ G-G latency is lower up to 128KB
- APEnet+ P2P latency: $\sim 8.5\mu s$
- APEnet+ staging latency: $\sim 16.8\mu s$
- MVAPICH/IB latency: $\sim 17.4\mu s$

Rossetti D. et al: "GPU peer-to-peer techniques applied to a cluster interconnect" presented at CASS2013 Workshop (IPDPS 2013 conference)

APEnet+ bandwidth





- Host RX $\sim 1.6 GB/s$
- GPU RX $\sim 1.4 GB/s$
- Limited by RX LOGIC RDMA Virtual-to-Physical (V2P) Translation, most demanding task.

When handled by Nios II:

- Firmware not optimized: $\sim 1.2 GB/s$
- Nios Firmware Optimization $\sim 1.6 GB/s$

When handled by HW with a custom developed Translation Lookaside Buffer:

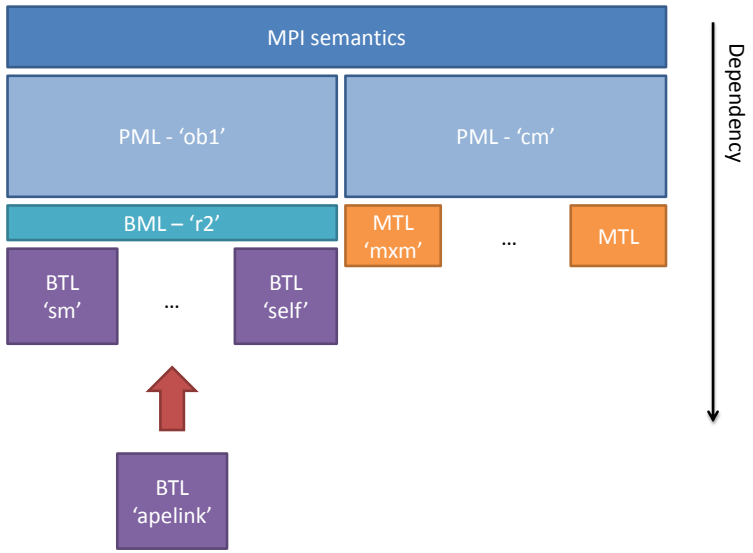
- First implementation. Host RX only! >2.2GB/s
- At hardware level, TLB hardware from 3000ns -> 124 ns (x30) for 128 KB buffer size. Work in progress

Ammendola et al: "Virtual to Physical Address Translation for an FPGA-based Interconnect with Host and GPU RDMA Capabilities" Submitted to FPL2013 conf.

SW stack and GPU optimizations

Current APEnet programming model

- native RDMA API:
  - RDMA buffer registration: pinning and posting combined
  - single message transmission async queue
  - async delivery of completion events (both TX and RX)
- MPI for APEnet+ OpenMPI based
  - The Byte Transfer Layer framework provides a set of components for raw data transfer for send receive and RDMA based inteconnects.
  - The apelink BTL uses the RDMA API to program the APEnet+ device
  - early prototype



From here, where?

- Can we do something more to increase application performances on hybrid CPU+GPU systems i.e. can our target application benefits from a more powerful torus network card?

- What about our commercial "competitors"?

LQCD Case Study

LQCD as an example

- Slightly modified performance model taken from Babich (STRONGnet 2010), from Gottlieb via Holmgren for a $64^3 \times 128$ total lattice
- Balance condition: perfect overlap between computing time and communication time

Unit local lattice	GPUs/unit	Req. BW (GB/s)	Total GPUs
$16^3 \times 32$	2	4.3	512
$16^3 \times 64$	2	4.0	256
$32^3 \times 64$	2	2.1	32
$16^3 \times 128$	4	7.4	256
$32^3 \times 128$	4	3.7	32

Facts:

1. Current PCIe implementation limits the performance and scaling
   - apeNET+ bandwidth (PCI Gen2 x8 / 34 Gbps per link) allows "strong scaling" up to few tens of GPU
2. Tight time budget i.e. specific HW GPU-APEnet+ optimizations are needed to reduce transfer latency and overheads

Network card competitors

Mellanox announced full support to RDMA GPU Direct ("Bar1 access") only for Kepler

- At GTC (GPU Tech. Conf, Mar 2013) presented a preliminary set of latency measures for Connect3-X InfiniBand adapter supporting GPU Direct protocol
- GPU Direct RDMA enabled board is available now to "selected customers"; availability in few months from now

APEnet game over?

Not yet, if

- I/O interfaces performance remain comparable (better?)
- coupling of GPU Direct RDMA, High speed low latency network and (huge) FPGA resources is fully exploited
  - Hardware system specialization driven by application requirements
  - Adding"processing on network", changing routing functions, changing physical network topology, implementing in HW exotic tasks, hw support to enhance system fault tolerance, ...

APEnet+ customization: NaNet

NaNet: APEnet + NA62 cern Experiment

GPU L0 TRIGGER for HEP Experiments

Implement a RO Board-L0 GPU link with:

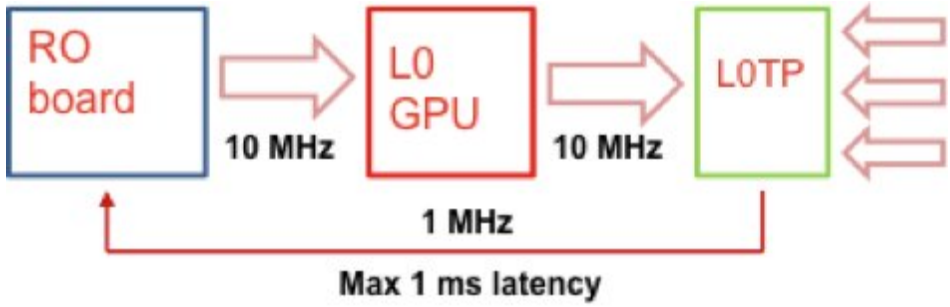
- Sustained Bandwidth $>$ 600 MB/s, (RO board output on GbE links)
- Small and stable latency

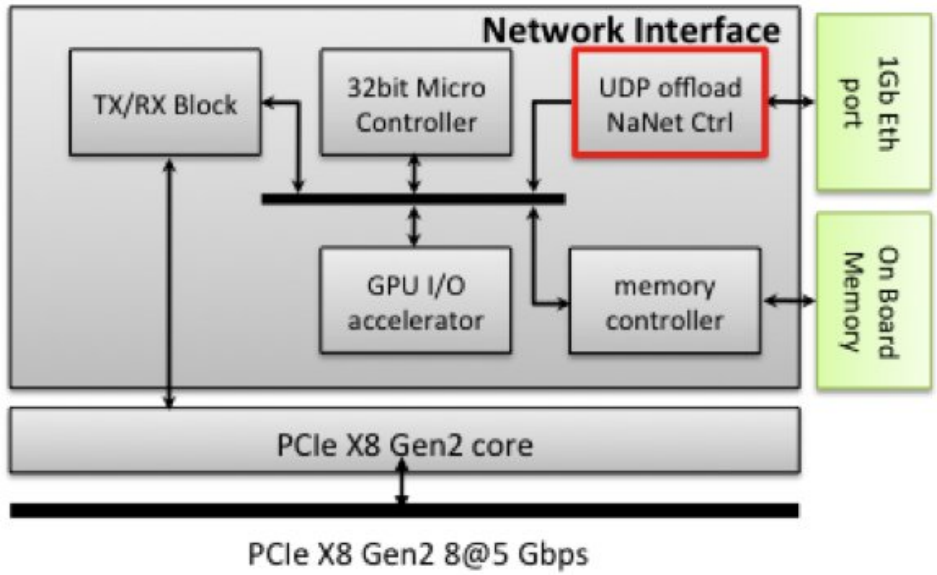
Problem: lower communication latency and its fluctuations. How?

- Offloading the CPU from network stack protocol management.
- Injecting directly data from the NIC into the GPU(s) memory.

NaNet solution:

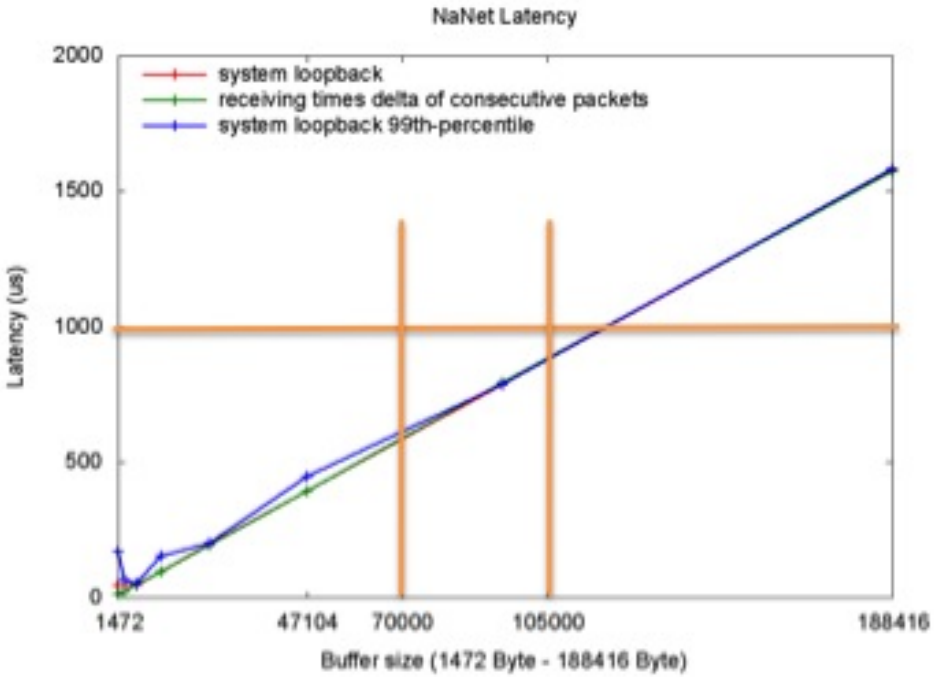
- APEnet+ FPGA-based NIC with an additional network stack protocol management offloading engine to the logic (UDP Offloading Engine).

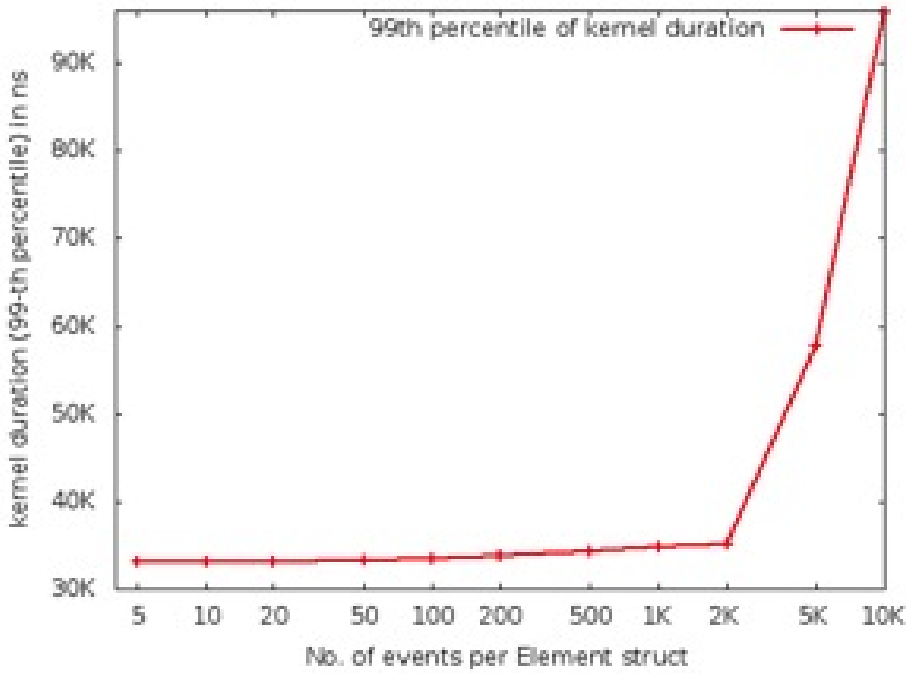




Lonardo A. "Building a Low-latency, Real-time, GPU-based Stream Processing System" GTC2013 Conference

NaNet results





- Several benchmarks have been performed on prototype board with APENet firmware.
- In the 1 GbE link L0 GPU-based Trigger Processor prototype the sweet spot between latency and throughput is in the region of 70-100 Kb of event data buffer size, corresponding to 1000-1500 events.

APEnet: 2013-2014 (main) activities

APEnet++: adoption of 28nm FPGA

- PCIe Gen2 -> Gen3 (x2 data rate)
- Torus links speed-up: from current 8.5 Gb/s to 14.5 Gb/s (x2)
- Explore the PCI Gen3 x16 (with PLX technology bridge)
- Explore the use of the embedded dual-core ARM processor to increase performance (Virt2Phys translation, P2P support, . . .)

Explore V5 porting on EUROTECH Tigon systems

Push on NVIDIA joint activities

- Further optimization of P2P GPU-APEnet+ and Kepler, Maxwell, . . . integration

APEnet+ customization: add specific I/O interface and accelerators in FPGA

- Low Level Trigger GPU-based for HEP collider (NA62, Atlas, . . .)
- Distributed read-out for KM3 Neutrino Telescope (under evaluation)
- Low latency coupling of read-out system and GPU computing for E-ELT (European Large Telescope) and for X-Ray microscopes imaging (LBNL)
- Brain simulation: dedicated network for high speed connectoma simulation (DPSNN model)

Conclusions

APEnet+ and QUonG

- APEnet+ V4 in "massive" construction phase.
  - Demonstrated advantages of HW GPU Direct RDMA mechanism introduction in hybrid systems (CPU+GPU)
  - Assembled a medium-size prototype (32 Tflops) almost "ready to use"
  - Software optimization (MPI, RDMA API) in progress

Roadmap 2013-2014

- APEnet+ Versione 5 (V5) based on 28nm FPGA: more room for performance improvements
  - PCIe Gen2->Gen3, Torus link speed enhancements, HardIP ARM coupling to APEnet exploration

Also, fast moving towards new scientific environments

- High-Low level trigger of future HEP experiments
- APEnet+ as low latency GPU interface to read-out system
- Network specialization for computing platform for not "APE traditional" fields: complex systems, Brain Simulation, molecular dynamics, ...

Thank you! Questions or comments?

Roberto Ammendola

Andrea Biagioni

Ottorino Frezza

Francesca Lo Cicero

Alessandro Lonardo

Pier Stanislao Paolucci

Davide Rossetti

Francesco Simula

Laura Tosoratto

Piero Vicini