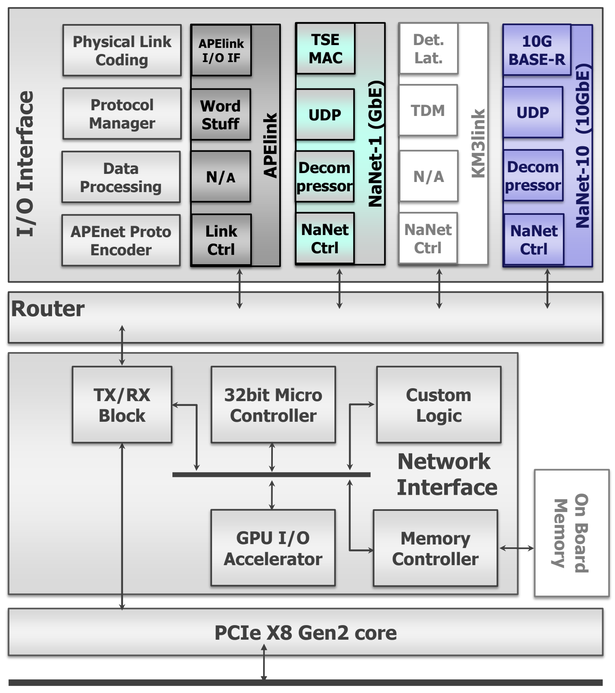
A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities

R. Ammendola[b], A. Biagioni[a], O. Frezza[a], G. Lamanna[c], F. Lo Cicero[a], A. Lonardo[a], M. Martinelli[a], P.S. Paolucci[a], E. Pastorelli[a], L. Pontisso[e], D.Rossetti[d], F. Simula[a], L. Tosoratto[a] and P. Vicini[a]

(a) INFN Sezione di Roma (b) INFN Sezione di Roma Tor Vergata (c) INFN LNF and CERN (d) NVIDIA Corp. USA (e) INFN Sezione di Pisa and CERN

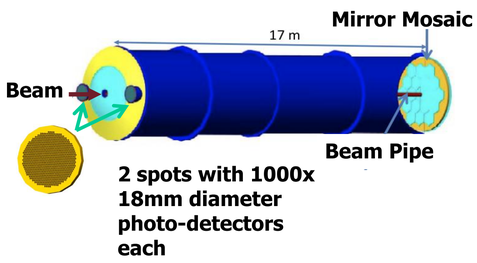
CHEP2015 Okinawa, Japan, 13-17 Apr 2015

Design of a FPGA-based Network Interface Card for Real-time GPU Computing



- Multiple link technologies
  - Standard: **1GbE** (1000Base-T), working on **10GbE** (10Base-R).
  - Custom: **APElink** (>20 Gb/s QSFP), deterministic latency **KM3link** (2.5 Gb/s optical).
- Network protocol offloading
  - **UDP**, **Time Division Multiplexing**.
- Processing on data stream
  - e.g. rearrange event data (size and alignment) in a GPU-friendly style.
- 6 ports full crossbar switch
  - 10 simultaneous 2.8 GB/s data streams
- Trigger and timing distribution system with HSMC daughtercard.
- **GPUDirect P2P/RDMA**
  - Direct data transfer between network and GPU memory (no bounce buffers on CPU memory).
- PCI Express x8 Gen2
  - Working on **PCIe X8 Gen3**.

NA62 RICH Detector

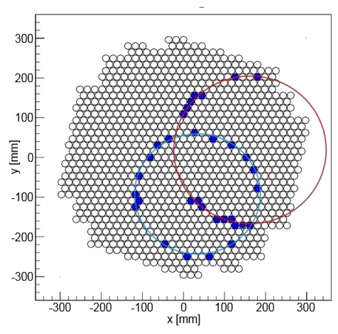


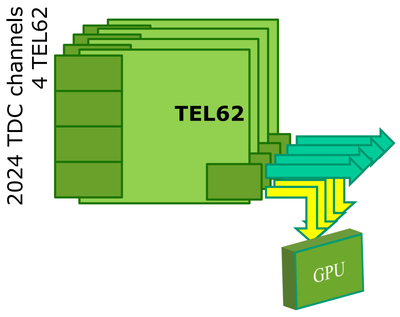
Ring-imaging Čerenkov detector

- Pion-Muon discrimination.
- 70 ps time resolution.
- **10 MHz** event rate
- **20** photons detected on average per single ring event (**hits** on photo-detectors)
- **40 Byte** per event

Rings pattern recognition and fit also **performed on GPU** (parasitic operation):

- New algorithm ("Almagest") developed for trackless, fast, and high resolution ring fitting.
- Rough detection of particle speed (radius) and direction (centre).
- **few µs per event** (on NVIDIA K20x).

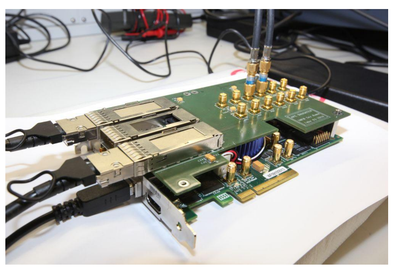




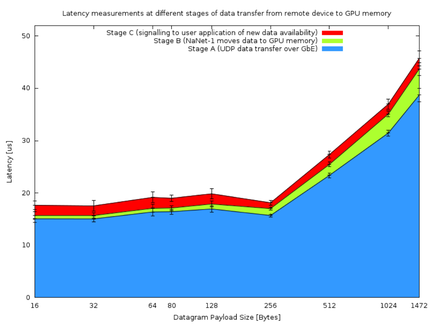
- 4 TEL62 for RICH detector
  - 8×1GbE links for data r/o
  - 4×1GbE trigger primitives
  - 4×1GbE GPU trigger
- Events rate: 10 MHz
- L0 trigger rate: 1 MHz
- Max Latency: 1 ms

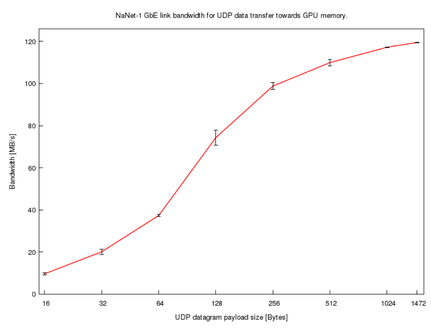
NaNet-1

- Implemented on Altera Stratix IV dev board (EP4SGX230KF40C2)
- 1GbE PHY Marvell 88E1111
- TTC daughtercard with HSMC connector for timing (clock, SOB/EOB) and trigger signals.
- Supports additional 3 APElink channels (20 Gb/s each) with HSMC daughtercard



NaNet-1 in RICH low level trigger processor





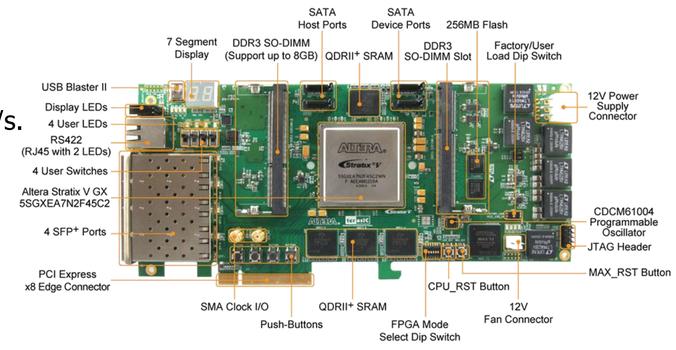
Contacts

alessandro.lonardo@roma1.infn.it
piero.vicini@roma1.infn.it
Presenter: andrea.biagioni@roma1.infn.it

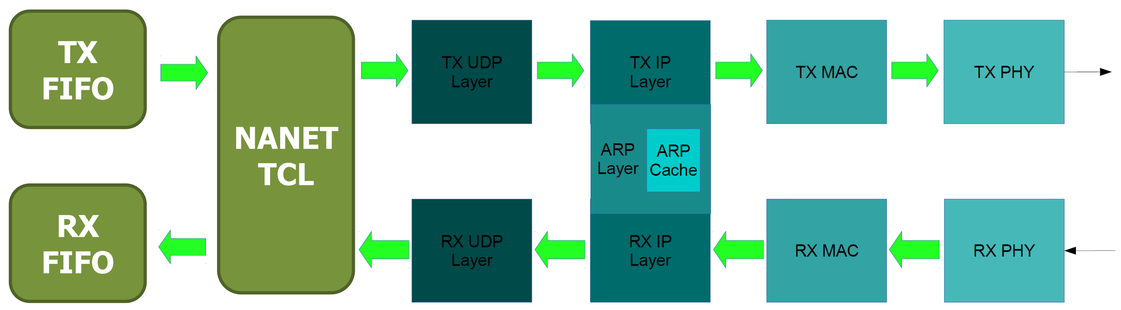
http://apegate.roma1.infn.it
http://euretile.roma1.infn.it
http://na62.web.cern.ch/na62/

NaNet-10 (four 10GbE SFP+ Ports)

- ALTERA Stratix V dev board.
- **PCIe x8 Gen3** (8 GB/s).
- 4 SFP+ ports (Link speed up to 10Gb/s.
- Implemented on Terasic DE5-NET board
- GPUDirect P2P/RDMA capability
- UDP offloas supports
- Available **2Q2015**.
- Planned **40GbE** development.

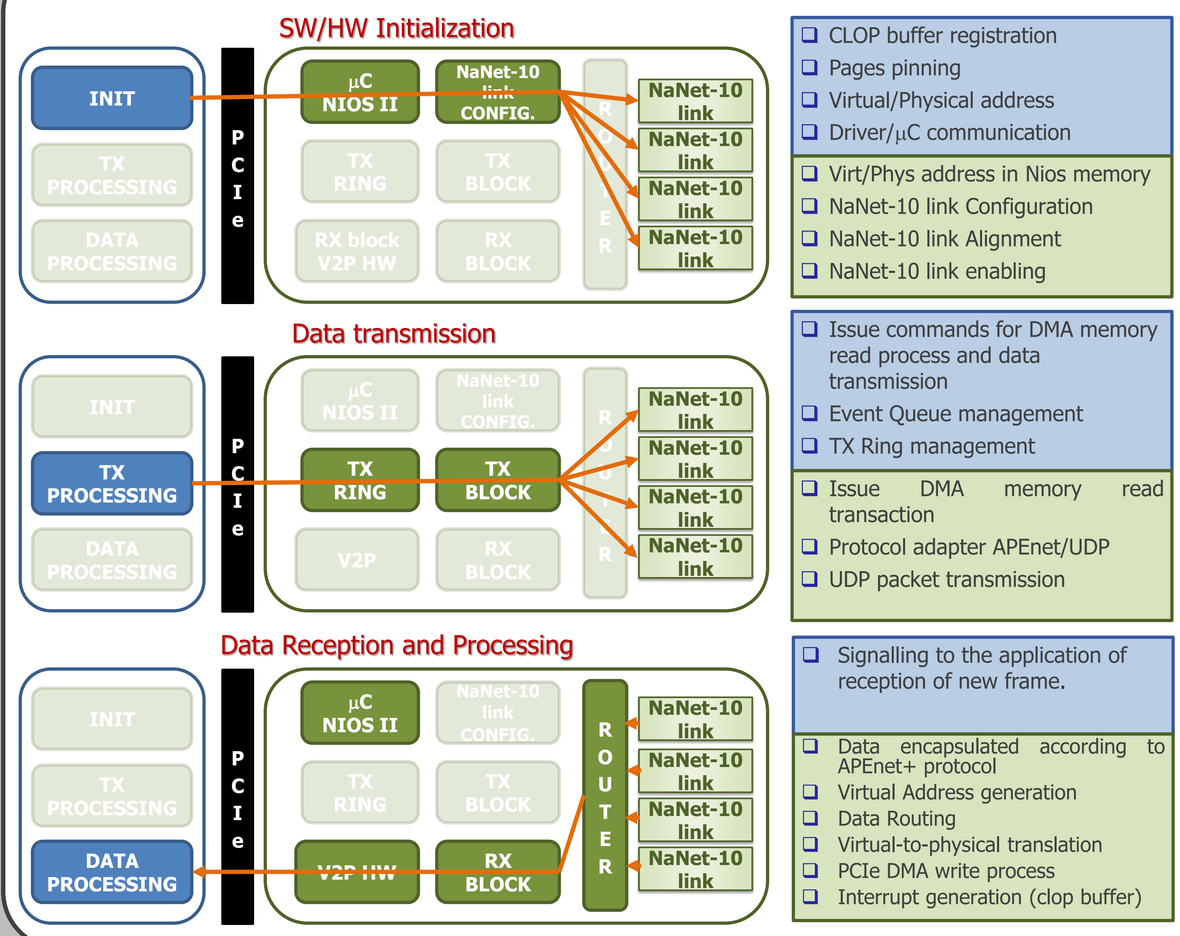


NaNet-10 link



- A fully UDP/IP 10GbE Link (IEEE 802.3-2005 compliant) has been implemented in hardware.
  - Altera 10GBASE-R PHY block
  - Altera 10 GbE MAC layer
  - 10 Gbit UDP/IP Core adapted from 1 Gbit UDP/IP core at www.opencores.org
  - AXI-lite data interface working @ 156.25MHz
  - Fully customizable IP and MAC address
  - ARP level functionalities. 256 entry cache for IP-to-MAC address translation
  - PHY block tested with an optical cable 3m long
- NaNet Transmission Control Logic
  - TX path: APEnet/AXI/UDP protocol translation
  - RX path: UDP/AXI/APEnet protocol translation. Virtual Address Generation. Data flow Manager.

Software/Hardware Overview



- CLOP buffer registration
- Pages pinning
- Virtual/Physical address
- Driver/µC communication

- Virt/Phys address in Nios memory
- NaNet-10 link Configuration
- NaNet-10 link Alignment
- NaNet-10 link enabling

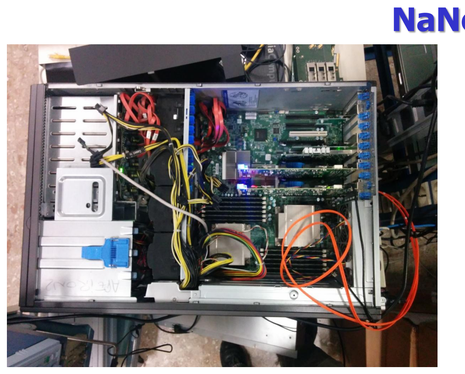
- Issue commands for DMA memory read process and data transmission
- Event Queue management
- TX Ring management

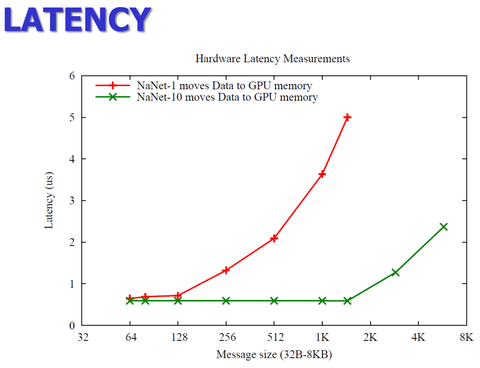
- Issue DMA memory read transaction
- Protocol adapter APEnet/UDP
- UDP packet transmission

- Signalling to the application of reception of new frame.

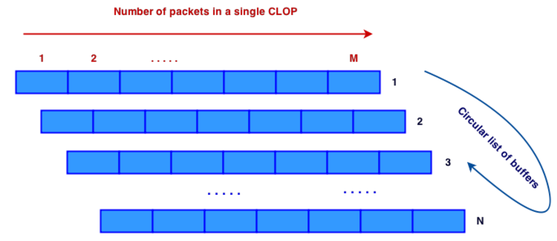
- Data encapsulated according to APEnet+ protocol
- Virtual Address generation
- Data Routing
- Virtual-to-physical translation
- PCIe DMA write process
- Interrupt generation (clop buffer)

NaNet-10 LATENCY





NaNet-10 Software



- Application Library
  - open/close NaNet-10 device.
  - Management Support for the circular list of persistent receiving buffers (CLOPs) in CPU/GPU memory
- Linux Kernel Driver
  - Status/Configuration registers.
  - TX registers interface.
  - Custom Event Queue management
- NIOS II Firmware
  - New BSP for NaNet10 board.
  - Initialization of NaNet10 channels.
  - Management of 4 concurrent data streams