RT2016
Padova, Italy
June 5-10, 2016

FPGA-based Network Interface Cards Implementing Real-time Data Transport for HEP Experiments

R. Ammendola[1], A. Biagioni[2], O. Frezza[2], G. Lamanna[3], F. Lo Cicero[2], A. Lonardo[2], M. Martinelli[2], P. S. Paolucci[2], E. Pastorelli[2], L. Pontisso[4], D. Rossetti[5], F. Simula[2], M. Sozzi[4], P. Cretaro[2], P. Vicini[2]

[1]Sezione di Tor Vergata, Istituto Nazionale di Fisica Nucleare, Rome, Italy, [2]Sezione di Roma, Istituto Nazionale di Fisica Nucleare, Rome, Italy,
[3]Laboratori Nazionali di Frascati, Istituto Nazionale di Fisica Nucleare, Frascati (Rome), Italy,[4]Sezione di Pisa, Istituto Nazionale di Fisica Nucleare, Pisa, Italy, [5]nVIDIA Corp, Santa Clara, CA, USA

Abstract

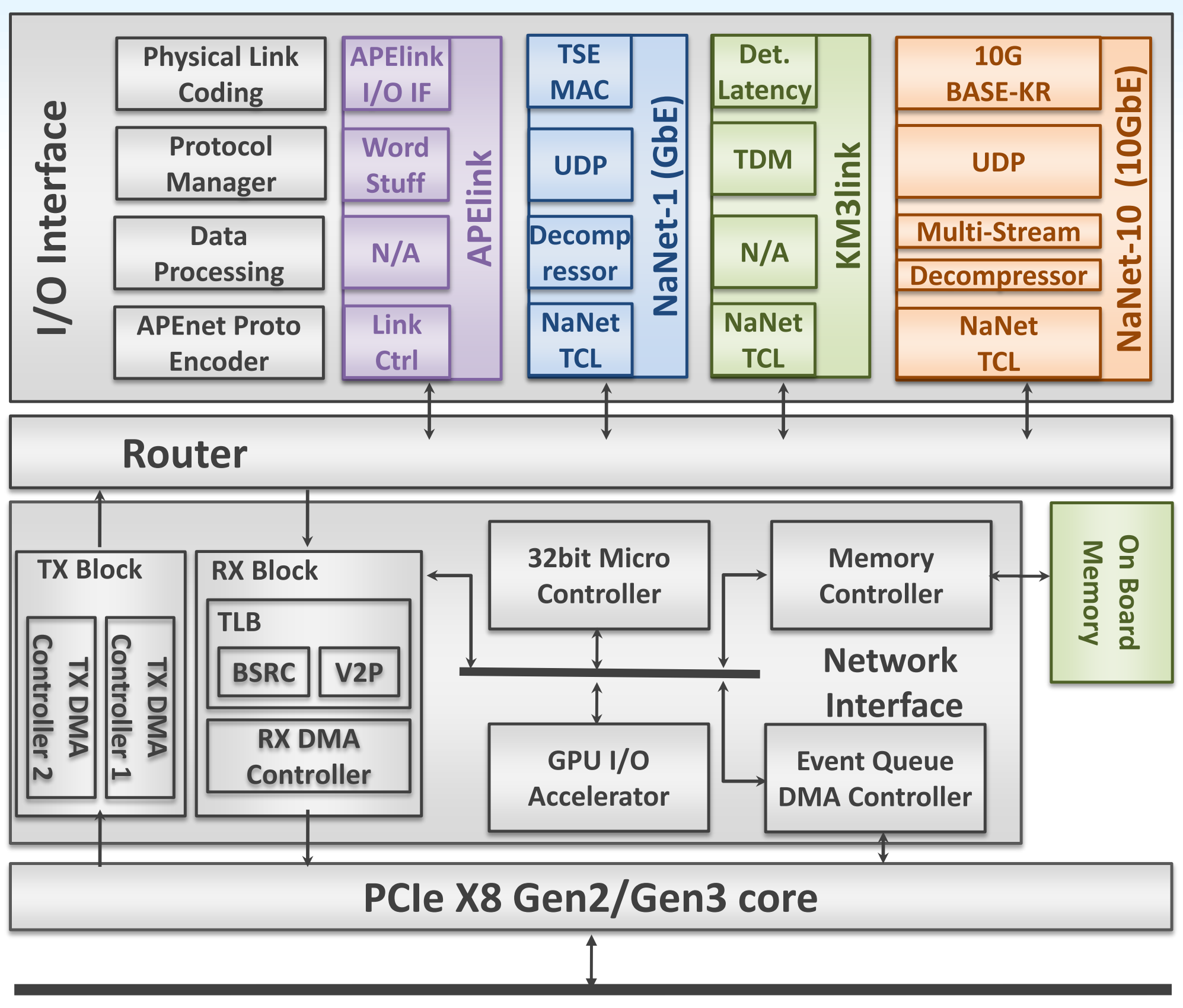
NaNet is a modular design of a family of FPGA-based PCIe Network Interface Cards implementing low-latency, real-time data transport between its network channels and the the host CPU and GPU accelerators memories.

The design feature a network stack protocol offloading module that operating in conjunction with a high performance PCIE Gen2/3 X8 core yields a low and predictable communication latency, making NaNet suitable for real-time applications.

A reconfigurable processing module is also available to implement application-specific processing on inbound/outbound data streams with highly reproducible latency.

As of now NaNet design has been specialized in the NaNet-1 (single 1GbE port) and NaNet-10 (four 10GbE ports) configurations employed in the GPU-based real-time trigger of the CERN NA62 experiment, and in the NaNet3 (four 2.5 Gbit optical channels) configuration adopted in the data acquisition system of the KM3NeT-Italia underwater neutrino telescope.

NaNet Design

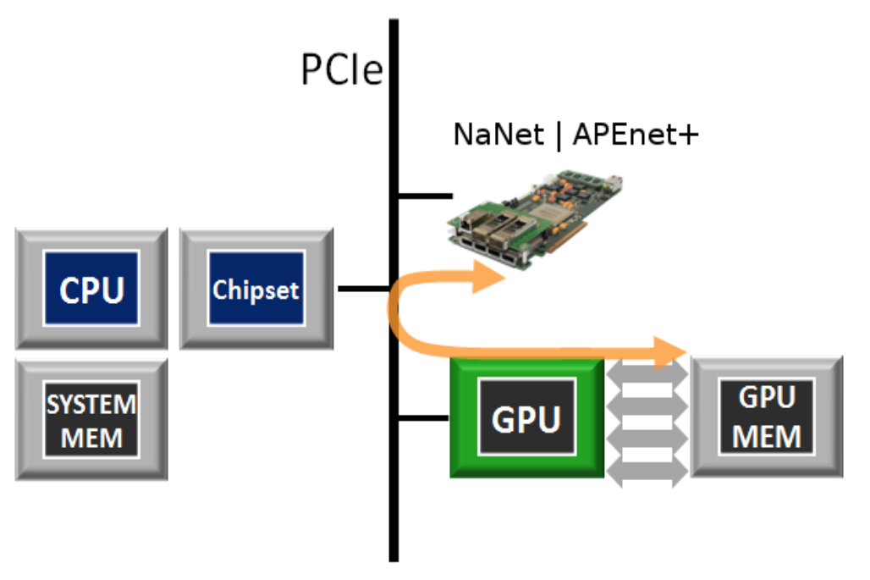


- ❑ I/O Interface
  - Multiple link.
  - Multiple network protocols.
    - **Off-the-shelf**: 1GbE, 10GbE
    - **Custom**: APElink (34 gbps/QSFP), KM3link
- ❑ Router
  - Dynamically interconnects I/O and NI ports.
- ❑ Network Interface
  - Manages packets TX/RX from and to CPU/GPU memory.
  - TLB & Nios II Microcontroller
    - Virtual memory management
- ❑ PCIe X8 Gen2 Core
  - CPU BW: 2.8 GB/s Read ÷ 2.5 GB/s Write
  - GPU BW: 2.5 GB/s Read & Write.
- ❑ Finalizing PCIe X8 Gen3 Core

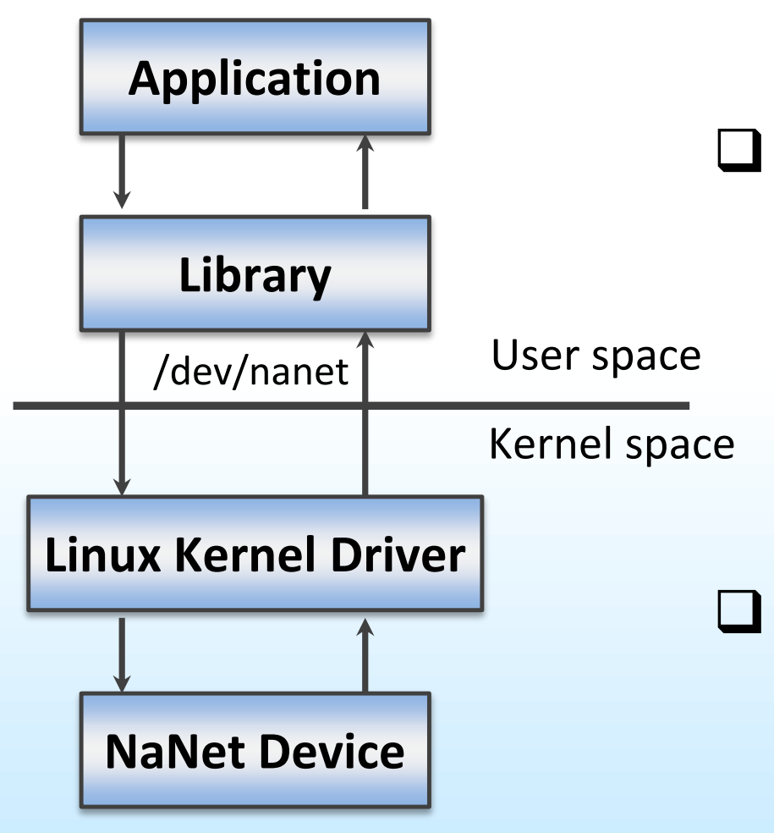
GPUDirect P2P/RDMA

GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement (zero copy)

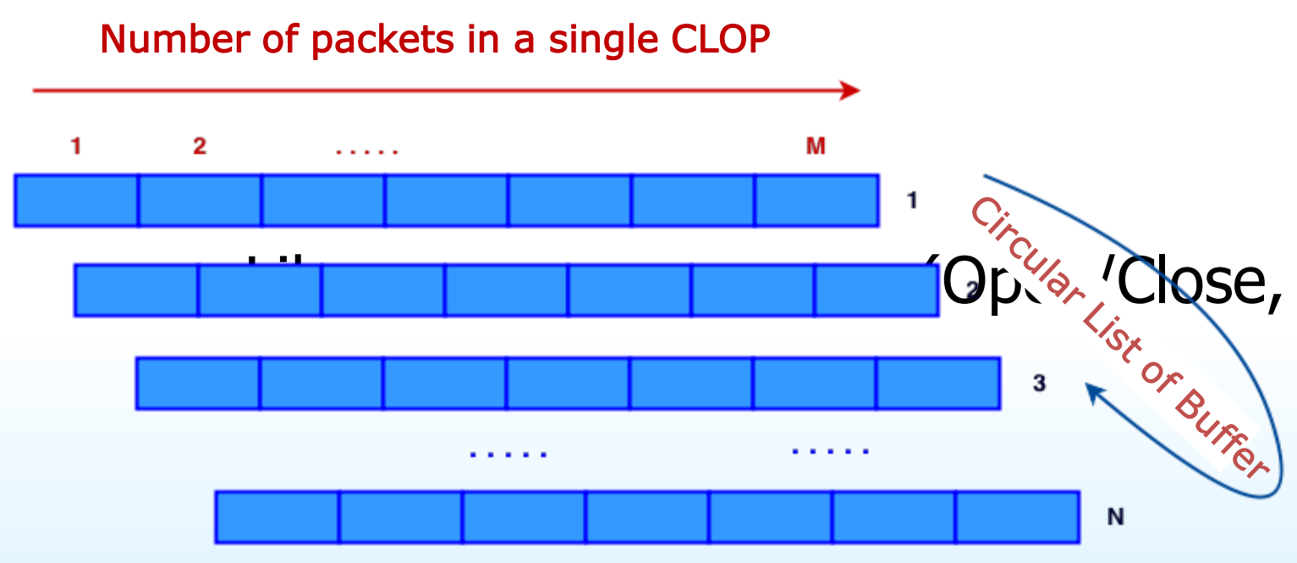
-> Latency reduction for small messages



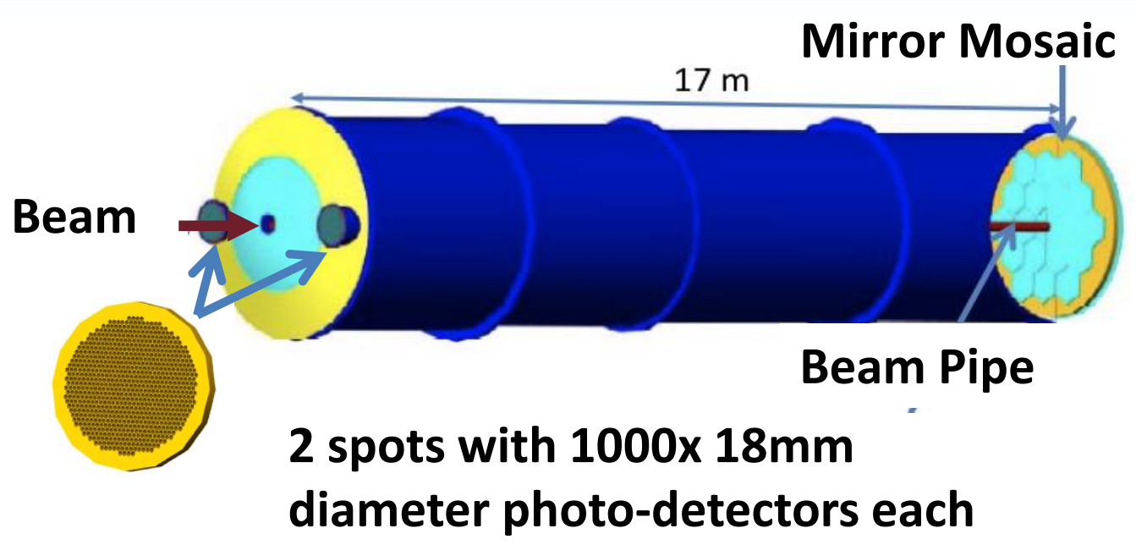
NaNet Software



- ❑ Host
  - User Space Application
  - User space Library (Open/Close, CLOP management,…)
  - Linux Kernel Device Driver
- ❑ NaNet Device
  - Nios II Microcontroller: single process software (bare metal) performing system configuration & initialization tasks



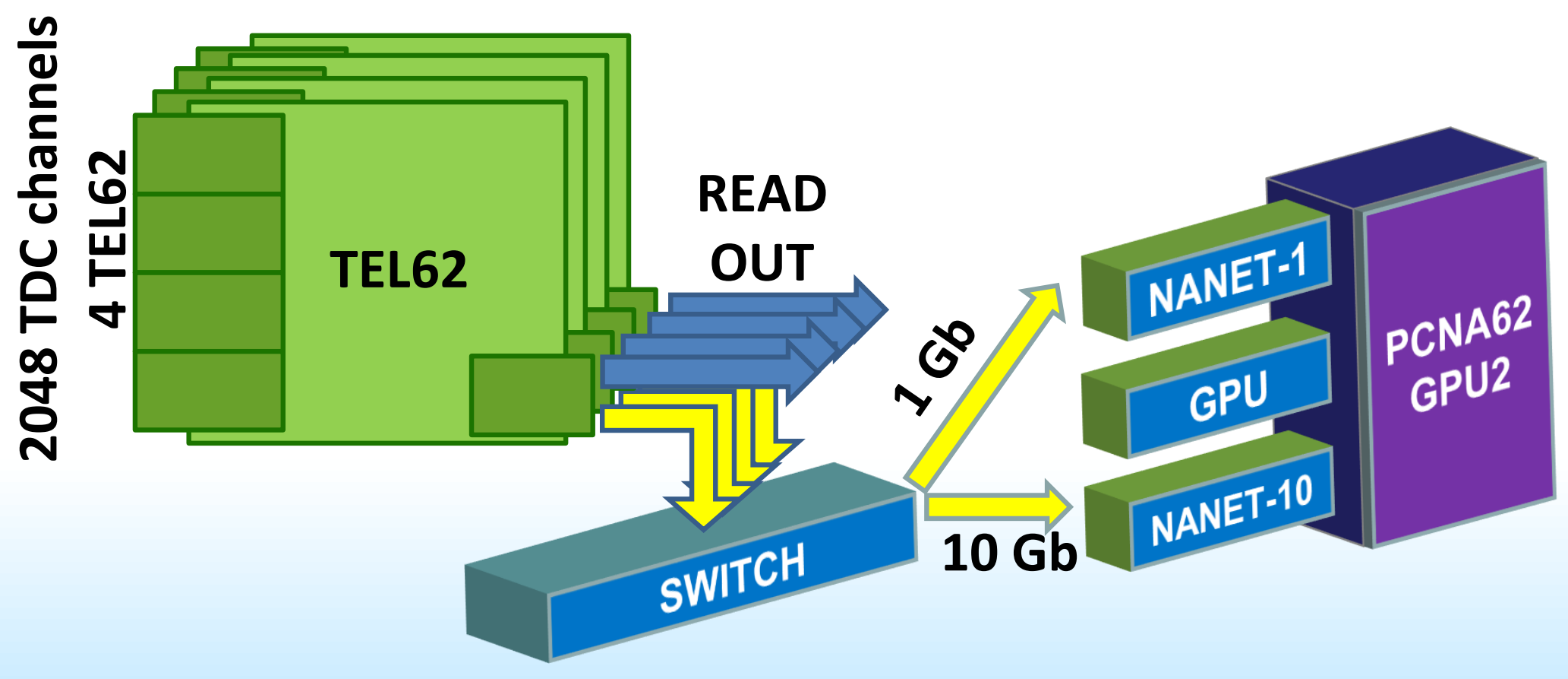
Case Study: NA62 RICH Detector



2 spots with 1000x 18mm diameter photo-detectors each

Ring-imaging Čerenkov detector

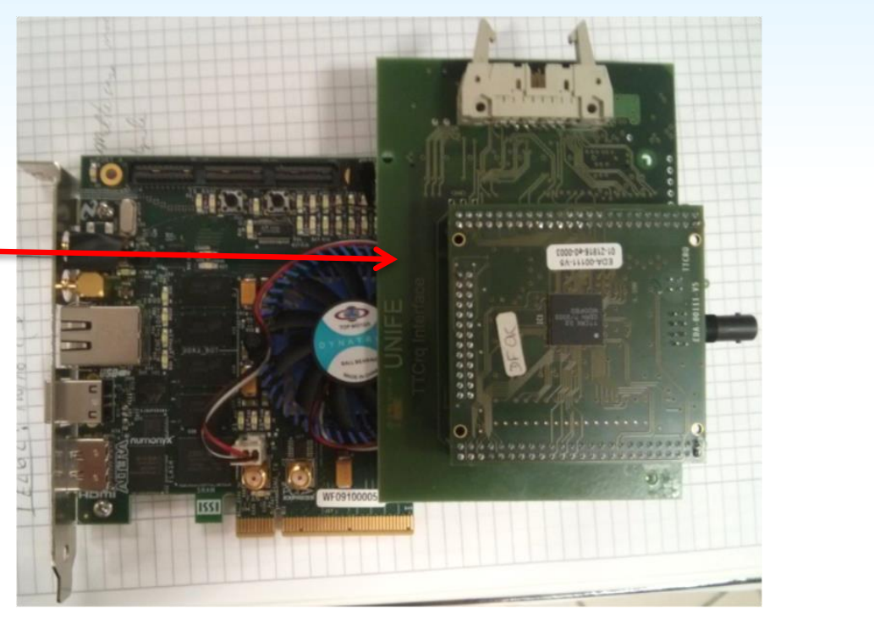
- ❑ Pion-Muon discrimination
- ❑ 70 ps time resolution
- ❑ 10 MHz event rate
- ❑ 20 photons detected on average per single ring event (hits on photo-detectors)
- ❑ 40 Byte per event



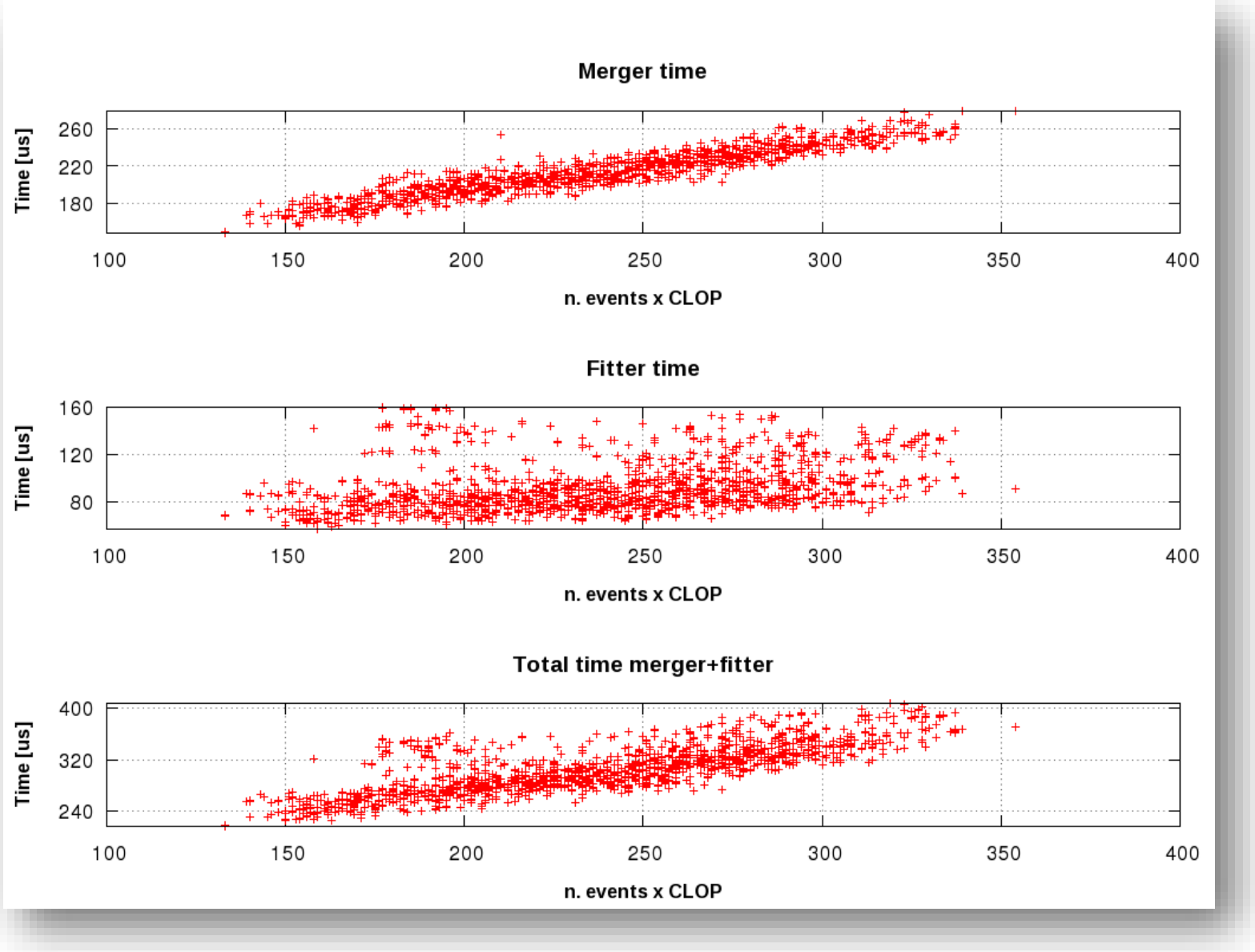
- ❑ 4 TEL62 for RICH detector
  - 8×1GbE links for data r/o
  - 4×1GbE trigger primitives
  - 4×1GbE GPU trigger
- ❑ Events rate: 10 MHz
- ❑ L0 trigger rate: 1 MHz
- ❑ Max Latency: 1 ms

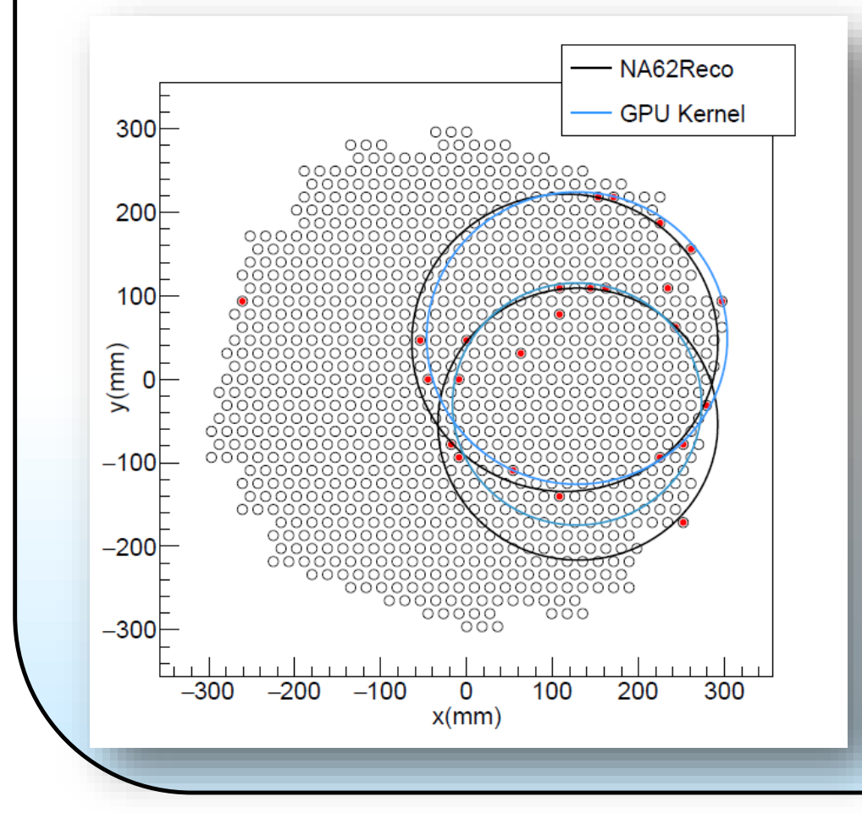
NaNet-1 in RICH low level trigger processor

- Implemented on Altera Stratix IV dev board
- TTC daughtercard with HSMC connector for timing (clock, SOB/EOB) and trigger signals



- ❑ Merger time depends on data size. Working on task speed up:
  - NOW performed on GPU
  - Finalizing FPGA implementation (tens of cycles latency)
- ❑ Computing time (K20c):
  - ~1 μs per event



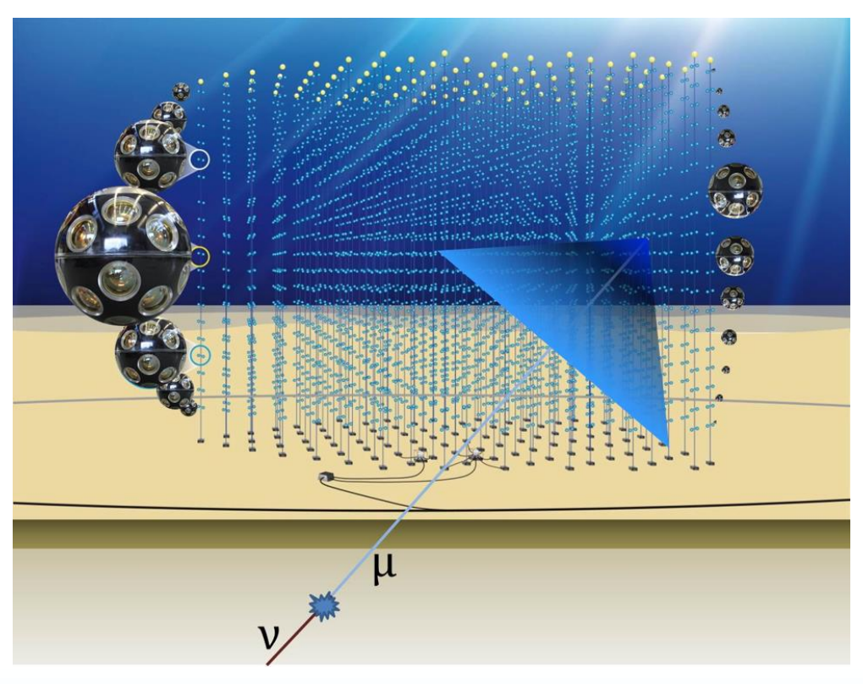


Rings pattern recognition and fit also performed on GPU:

- ❑ New algorithm (“Almagest”) developed for trackless, fast, and high resolution ring fitting
- ❑ Rough detection of particle speed (radius) and direction (centre).
- ❑ 0.5 μs per event (on nVIDIA K20x)

SEE ALSO THE TALK ON FRIDAY 8:30 A.M.

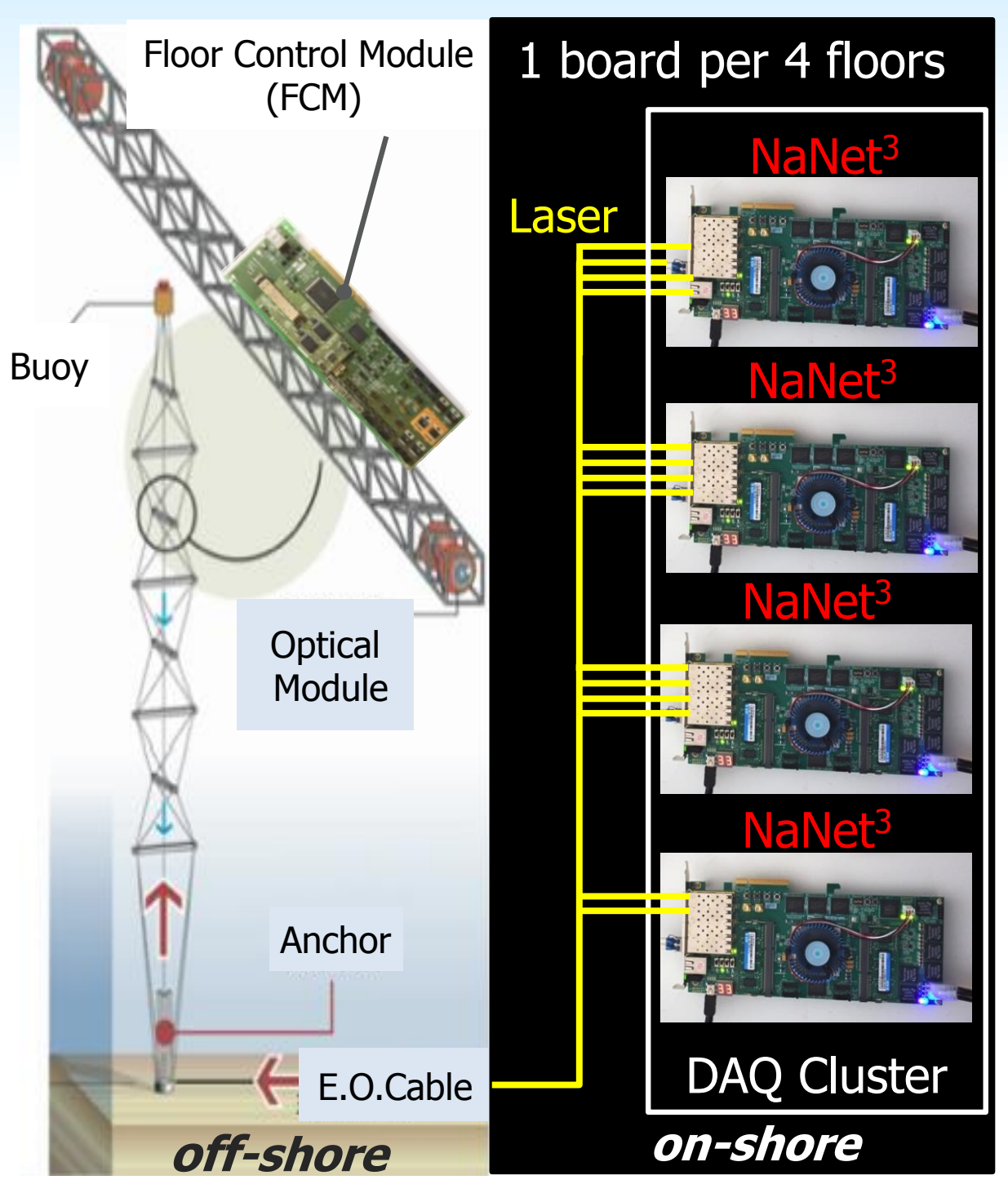
KM3NeT-Italia experiment



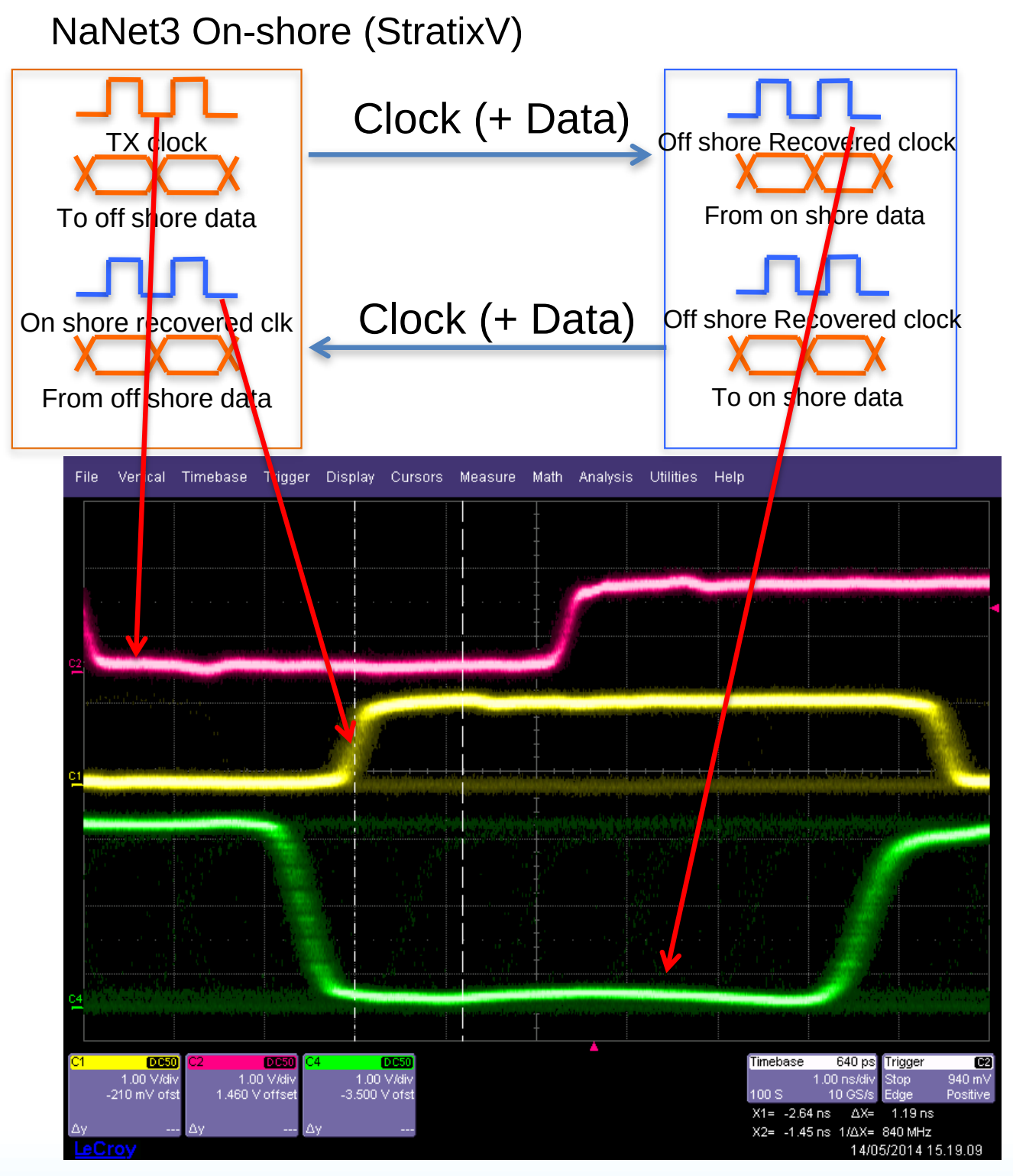
European deep-sea research infrastructure hosting a new generation of a neutrino telescope with a volume of several cubic kilometres located at the bottom of the Mediterranean Sea (~100km off-shore, ~3500m under the level of the sea).

- Data produced by OMs, hydrophones, and instruments, are collected by an electronic board contained in a vessel at the centre of the floor (FCM board)
- NaNet³ manages communication between the on-shore lab and the underwater devices, also distributing the timing information (GPS clock) and signals received from the on-shore equipment
- Deterministic latency links are required to obtain a common timing and known delay for the spatially distributed readout

NaNet³



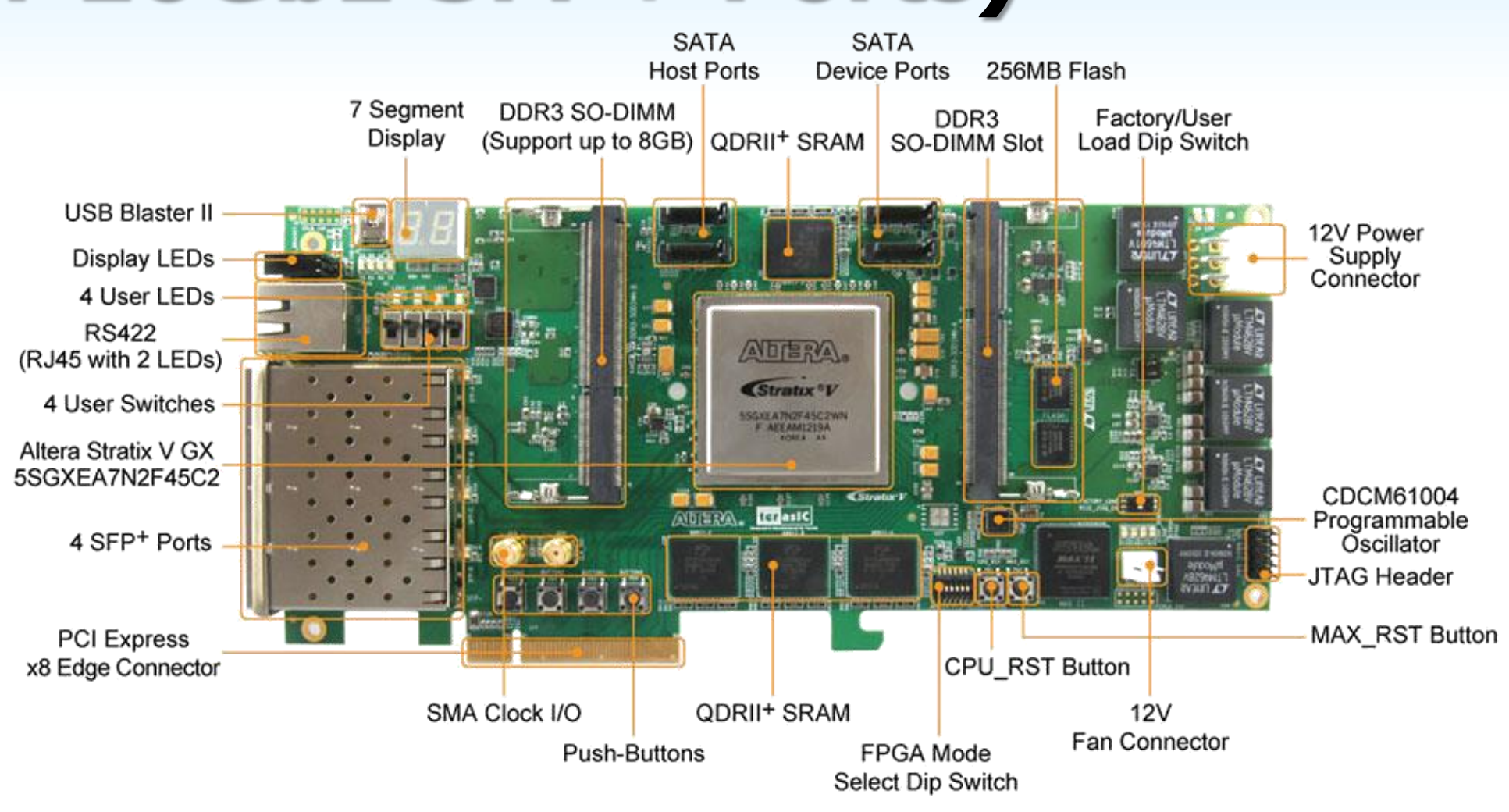
Is the on-shore endpoint for 4 offshore readout cards.

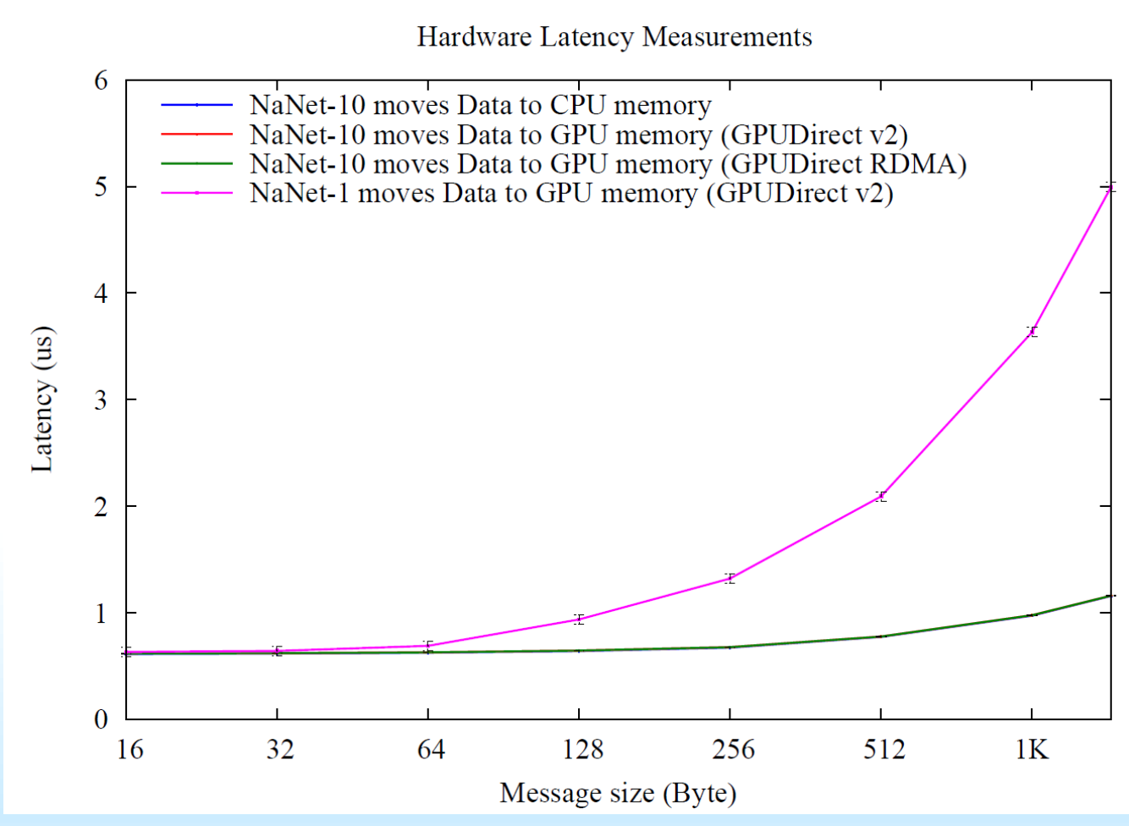


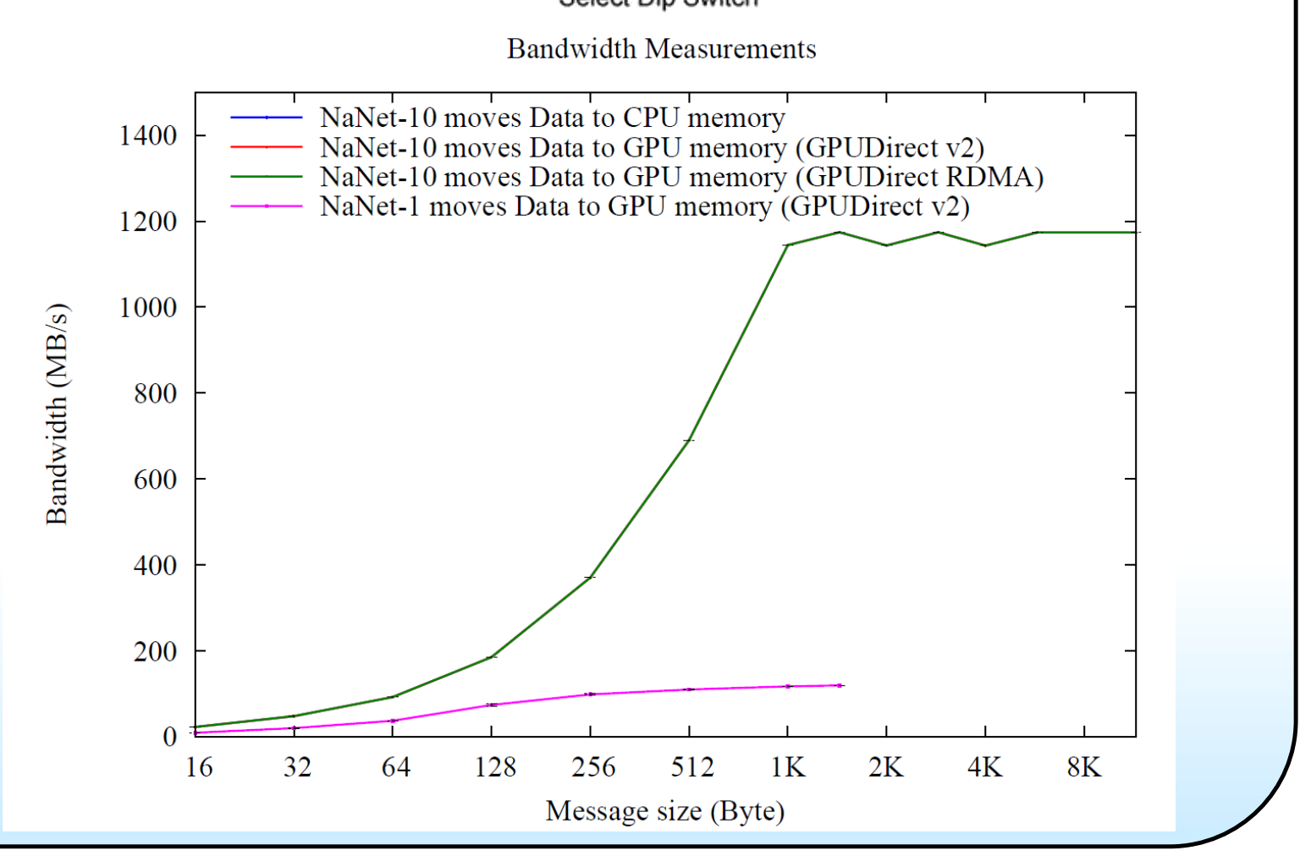
- Implemented on the Terasic DE5-NET Stratix V FPGA dev board
- 4 custom 2.5 Gbps deterministic latency optical channels
- Link speed up to 10 Gb/s
- GPUDirect P2P/RDMA capability
- Deterministic latency link: employs Altera Deterministic Latency Transceivers with an 8B/10B encoding scheme as Physical Link Coding and Time Division MultiPlexing (TDMP) data transmission protocol

NaNet-10 (four 10GbE SFP+ Ports)

- ALTERA Stratix V Terasic DE5-NET dev board
- 4 SFP+ ports (Link speed up to 10 Gb/s)
- Implemented on Terasic DE5-NET board
- GPUDirect P2P/RDMA capability
- UDP offload supports
- Realtime decompressing and event merging capability
- Planned **40GbE** development







Contacts:
NaNet project: http://apegate.roma1.infn.it/nanet
APE project: http://apegate.roma1.infn.it/APE
Presenter Contact: michele.martinelli@roma1.infn.it
NaNet coordinator: alessandro.lonardo@roma1.infn.it