

Schede PCI Express riconfigurabili per il trasferimento di dati a bassa latenza negli esperimenti di Fisica delle Alte Energie

Andrea Biagioni
INFN – Sezione di Roma
per la collaborazione NaNet

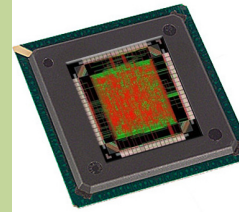
XV Incontri di Fisica delle Alte Energie
30 Marzo -1 Aprile 2016

Progettazione e implementazione di una famiglia di schede di rete (NIC) basate su FPGA e con interfaccia PCIe:

- ❑ Funzione di ponte fra l'elettronica di front-end e i nodi di calcolo su cui viene eseguito il software di trigger
- ❑ Supporto per sistemi di trasmissione dati a molteplici canali e con diversi protocolli di rete
- ❑ Latenza di comunicazione bassa e controllata (deterministica)
- ❑ Alta bandwidth
- ❑ Possibilità di aggiungere uno stadio di processamento in tempo reale sul flusso di dati (compressione/decompressione, riordinamento dei dati, unione di frammenti di eventi)
- ❑ Ottimizzazione del sistema di trasferimento dati verso le GPU

TRIGGER ONLINE

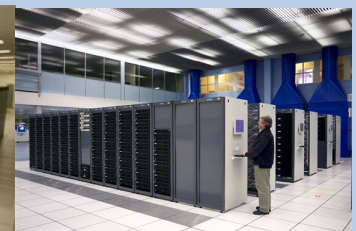
- ❑ Sistemi hardware: ASIC, FPGA
 - Elevata efficienza per primitive di trigger “semplici”
- ❑ Server (CPU)
- ❑ Acceleratori COTS
 - Possibilità di sviluppare algoritmi di trigger raffinati
- ❑ Potenza di calcolo, latenza, throughput



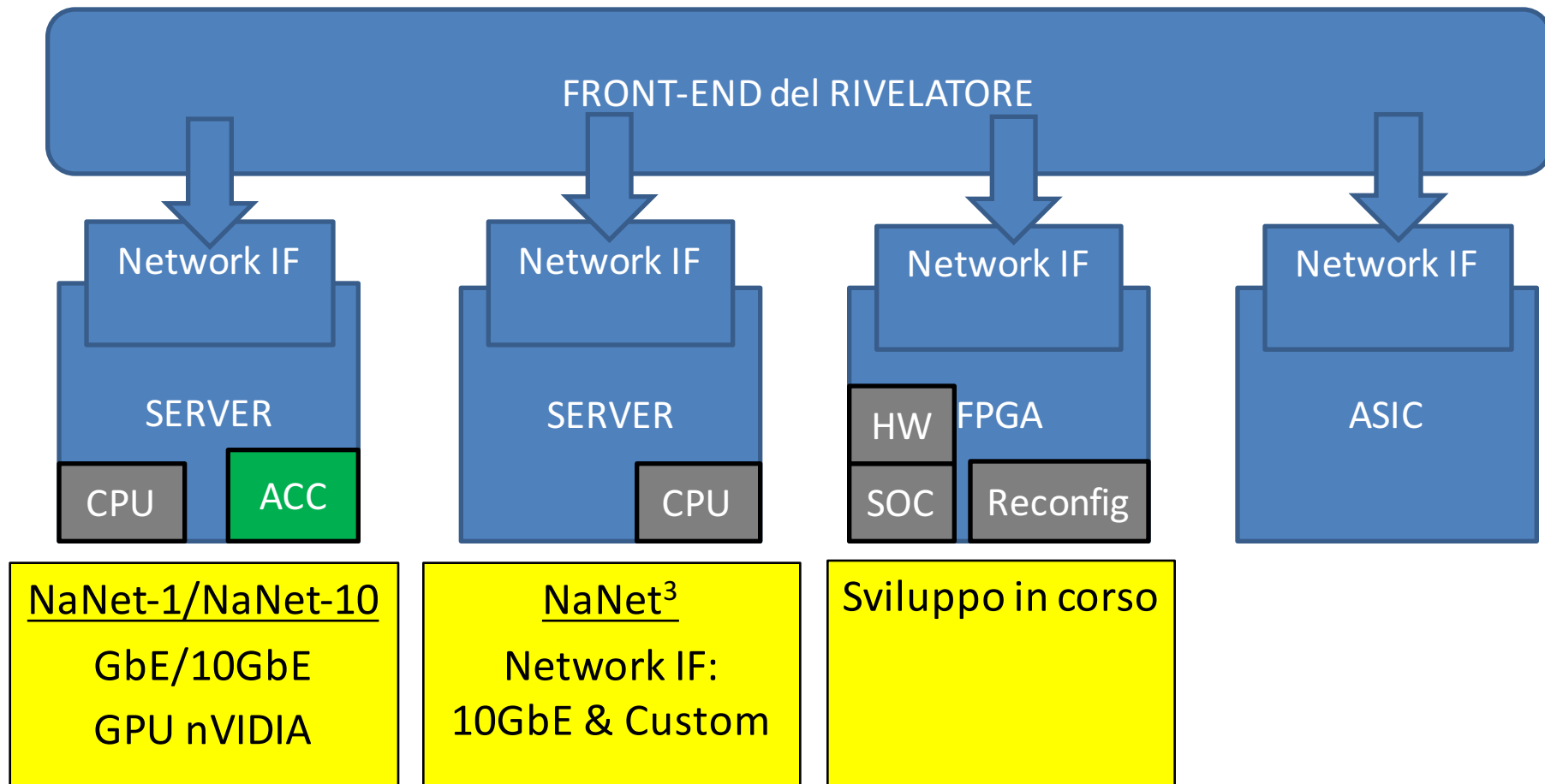
TRIGGER OFFLINE

Potenza di calcolo: PC farm, Supercomputer

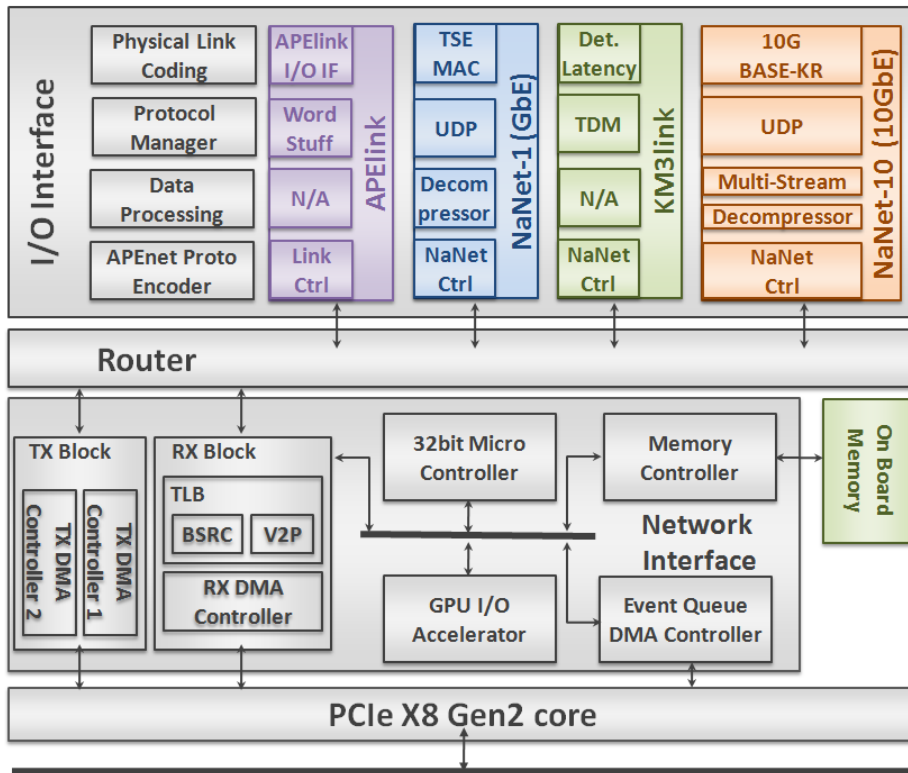
- ❑ Alta bandwidth
- ❑ Database
- ❑ Storage
- ❑ Endpoint Consolidation

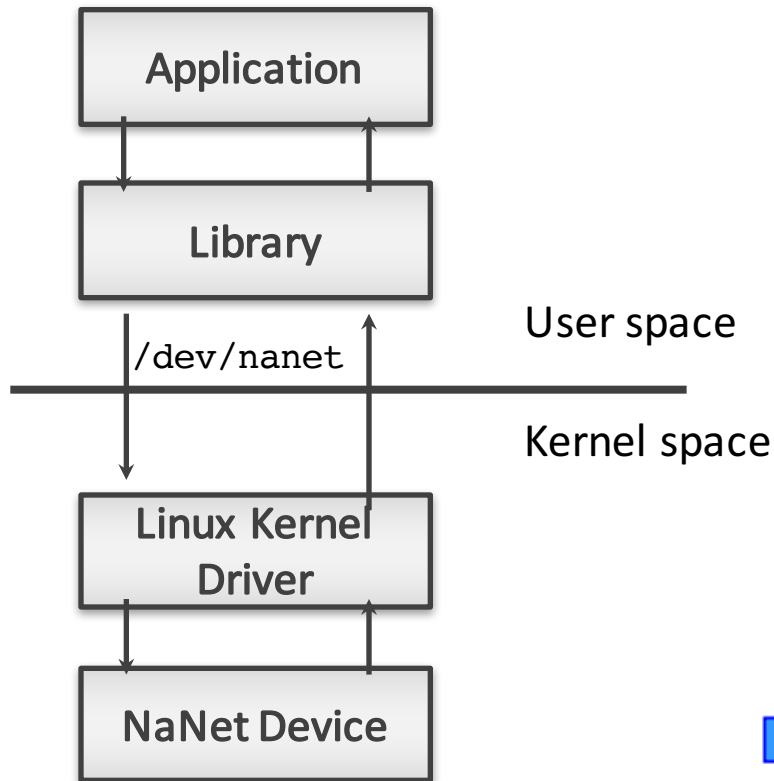


- ❑ Interfaccia di rete: protocollo standard, custom o combinazione di entrambi
- ❑ Acceleratori: GPU, Intel MIC
- ❑ Sistemi FPGA: firmware puro, ARM, logica riconfigurabile (openCL)

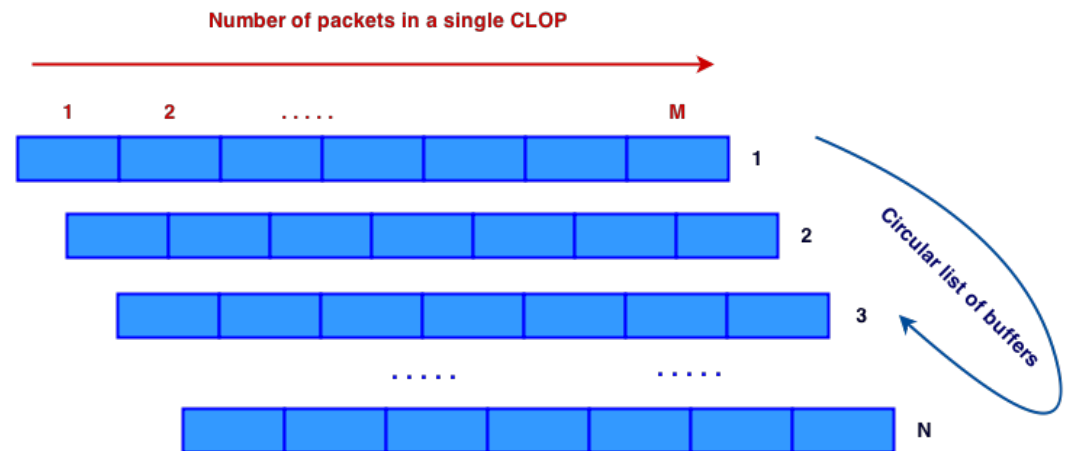


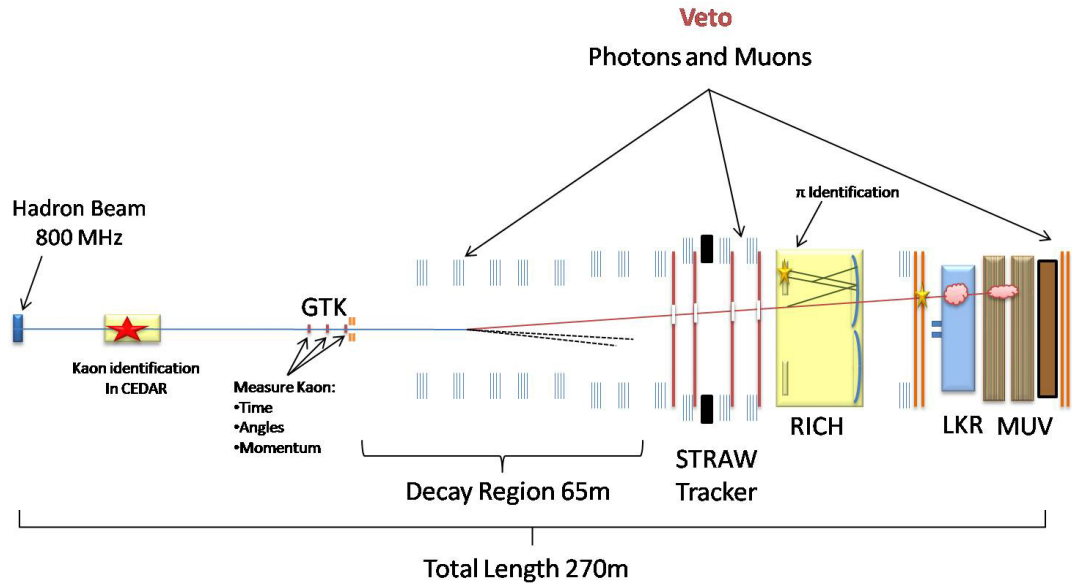
- ❑ I/O Interface
 - Supporto per molteplici canali
 - Protocolli di rete
 - Off-the-shelf: 1GbE, 10GbE
 - Custom: APElink (34gbps/QSFP), KM3link
- ❑ Router
 - Interconnette dinamicamente le porte di I/O con le porte della NI
- ❑ Network Interface
 - Gestisce i pacchetti TX/RX da e verso la memoria GPU/CPU
 - TLB & Nios II Microcontroller
 - Gestione della memoria virtuale
- ❑ PCIe X8 Gen2 (Gen3) Core





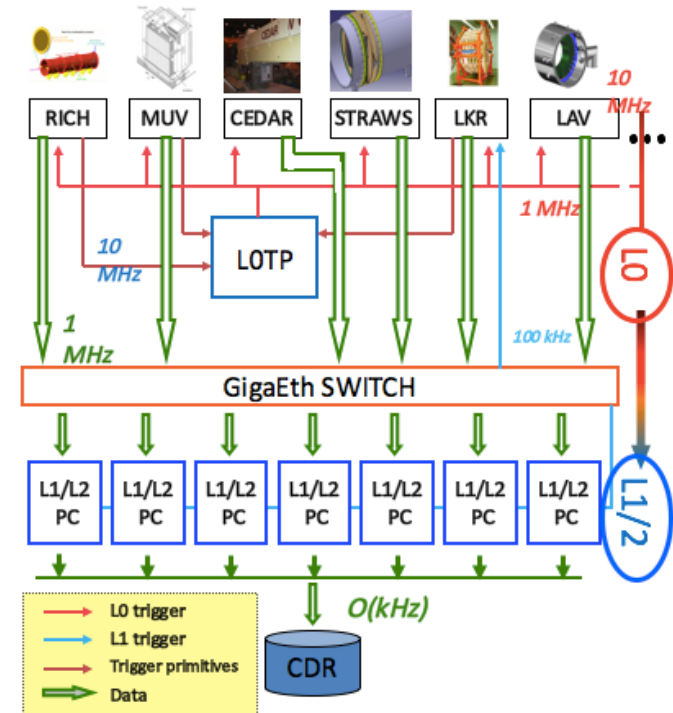
- ❑ Host
 - Linux Kernel Driver
 - User space Library (open/close, buf reg, wait recv evts, ...)
- ❑ Nios II Microcontroller
 - Programma a singolo processo per eseguire la configurazione ed inizializzazione del sistema

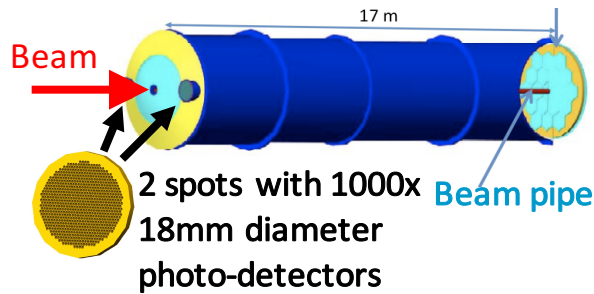




- ❑ Misura del decadimento ultra raro del Kaone $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ ($BR \sim 8 \times 10^{-11}$)
- ❑ Fascio di adroni ad alta intensità del (6% Kaoni)
- ❑ Il kaone decade lungo il percorso
- ❑ Trigger L0: riduzione del rate di eventi da 10 MHz a 1 MHz
 - Latenza: 1 ms

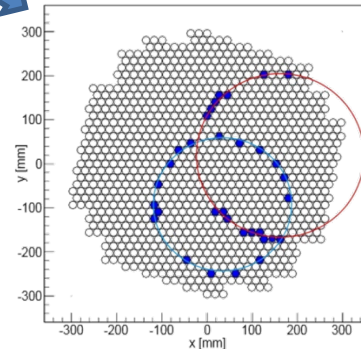
DAQ e TRIGGER



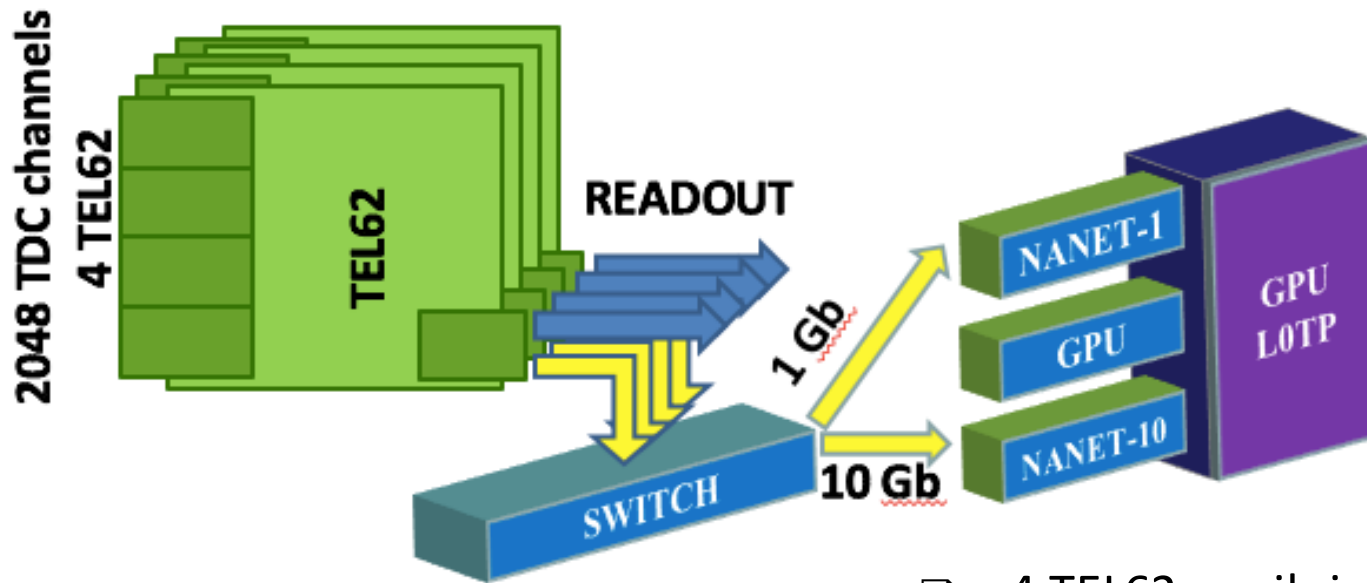


- ❑ Lunghezza: 17m, diametro: 3m
- ❑ Discriminazione pione-muone (inefficienza $< 1\%$ nel range 15-35 GeV)
- ❑ Risoluzione temporale : 70ps
- ❑ Rate di eventi 10 MHz: 20 hit (~ 40 byte per evento)
- ❑ 2 schede Read Out per ogni spot (ognuna 1000 fotomoltiplicatori)

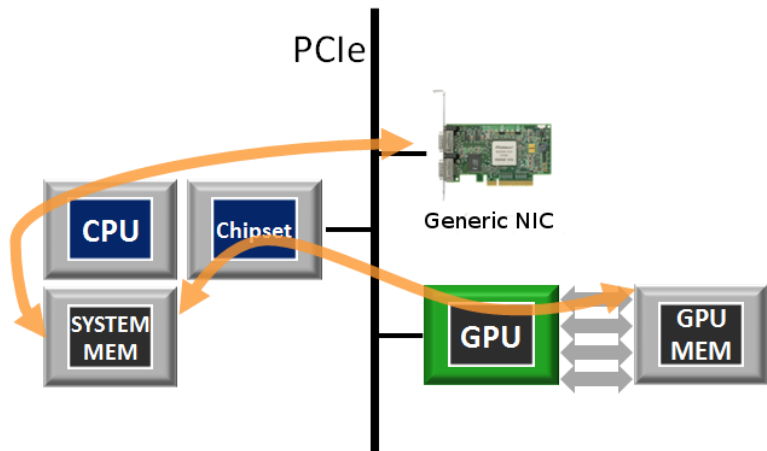
**Trigger di livello 0
basato su GPU per la
ricostruzione di cerchi**



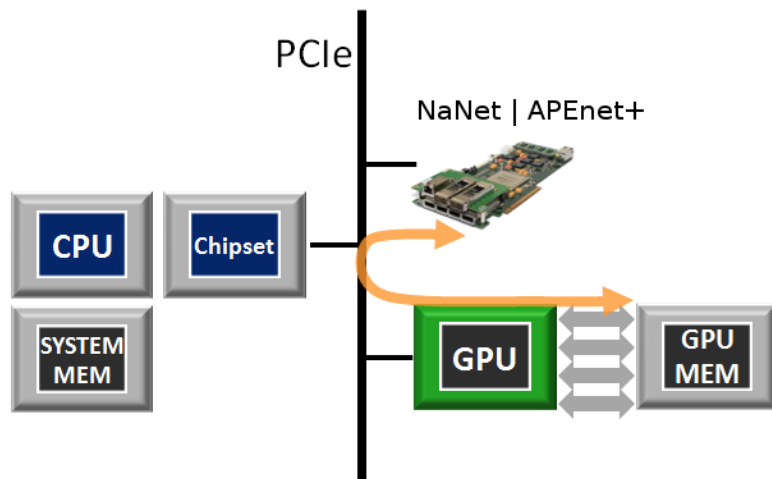
- ❑ Obiettivo del progetto: confrontare trigger implementato su FPGA con trigger basato su GPU
- ❑ Realizzare algoritmi di trigger più selettivi
 - Programmabili
 - Aggiornabili
- ❑ Buona efficienza della implementazione su GPU



- ❑ 4 TEL62 per il rivelatore RICH
- ❑ 8×1Gb/s links per data ReadOut
 - 4×1Gb/s trigger primitives
 - 4×1Gb/s GPU trigger
- ❑ Rate di eventi: 10 MHz
- ❑ L0 trigger rate: 1 MHz
- ❑ Max Latency: 1 ms

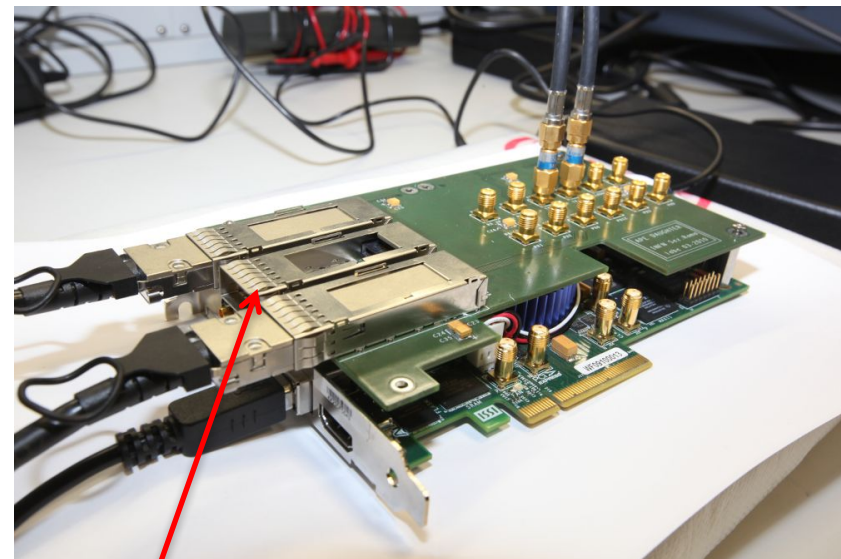


- ❑ Flusso dati con una scheda di rete standard (senza supporto GPUDirect)
- ❑ Buffer intermedio nella memoria della CPU per operazioni di I/O



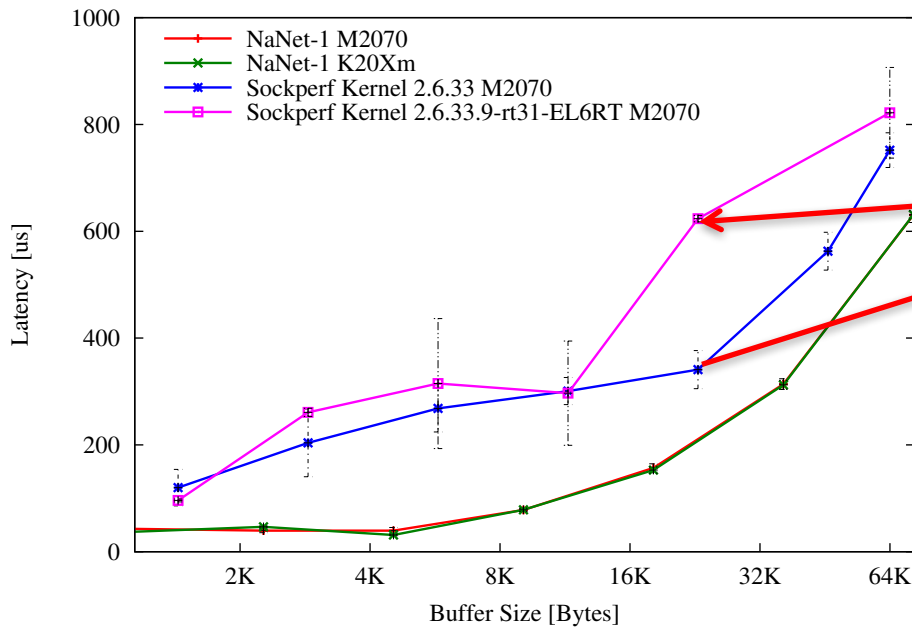
- ❑ GPUDirect P2P/RDMA consente lo scambio di dati diretto sul bus PCIe senza il coinvolgimento della CPU
- ❑ Nessuna necessità di usare buffer intermedi sulla memoria host
- ❑ Implementazione del I/O zero copy verso la GPU
- ❑ Riduzione della latenza per piccoli messaggi
- ❑ Compatibilità con nVIDIA Fermi; Kepler; Maxwell

NaNet-1: a 1 GbE NIC for NA62 GPU-based RICH L0 trigger prototyping



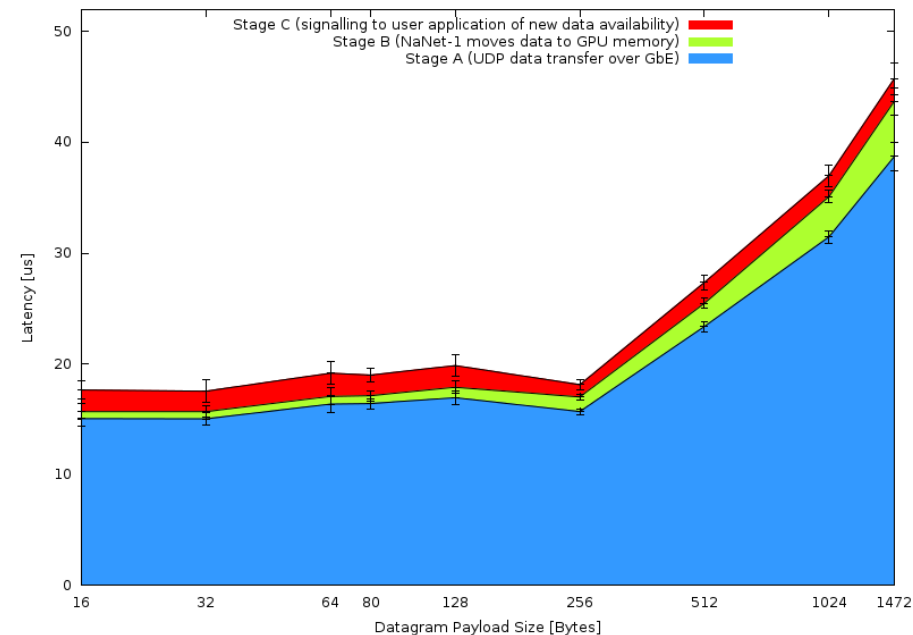
- ❑ Implementato su Altera Stratix IV
- ❑ GbE PHY Marvell 88E1111
- ❑ HSMC bus:
 - Supporto per 3 APElink (34 Gbps; HPC)
 - Supporto per TTC (INFN Ferrara)
- ❑ NaNet riceve il flusso TTC con informazioni di timing (40MHz, SOB, EOB) e i dati dall'esperimento

Communication Latency

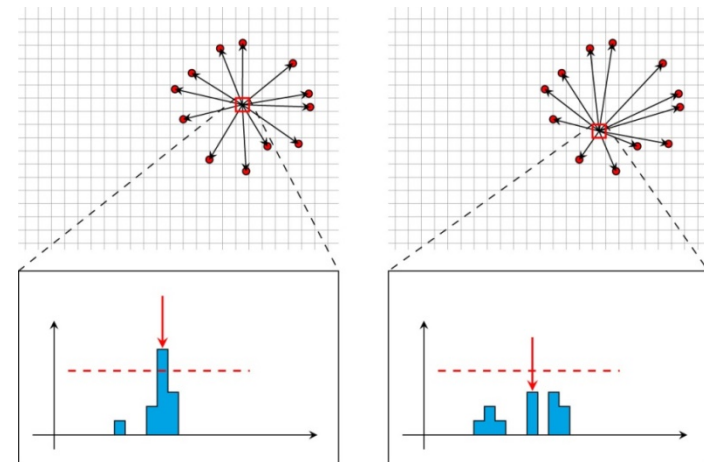
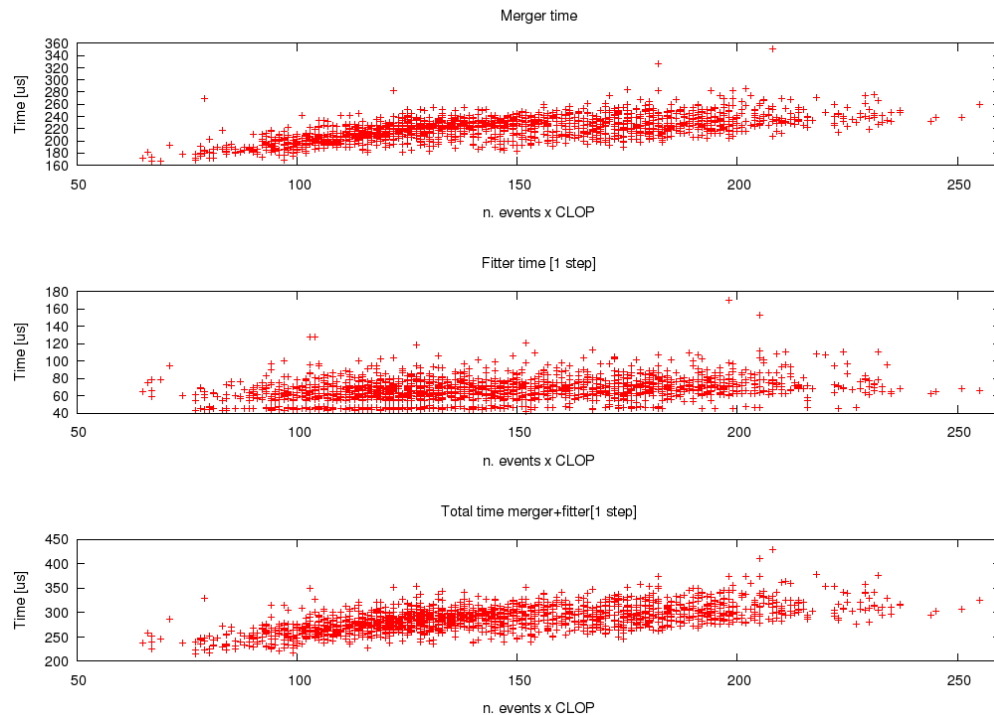


Real Time kernel: fluttuazioni ridotte rispetto ad un sistema standard, ma anche maggiore latenza....

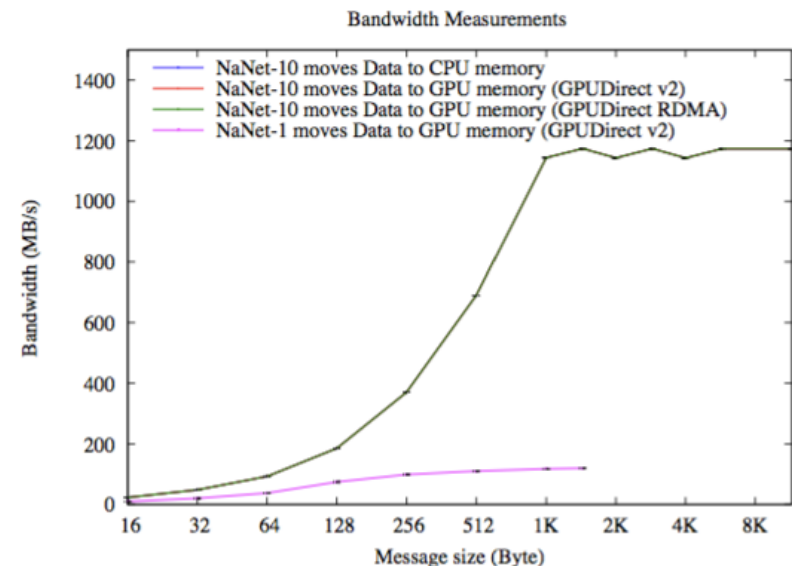
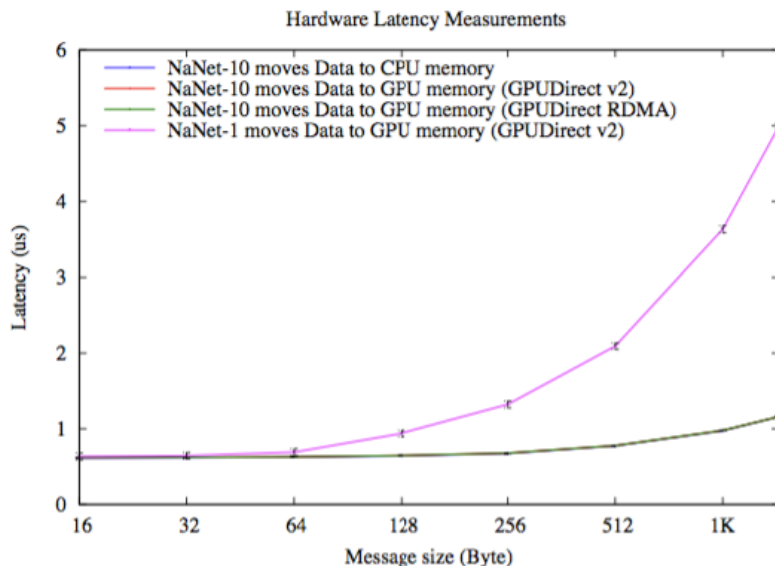
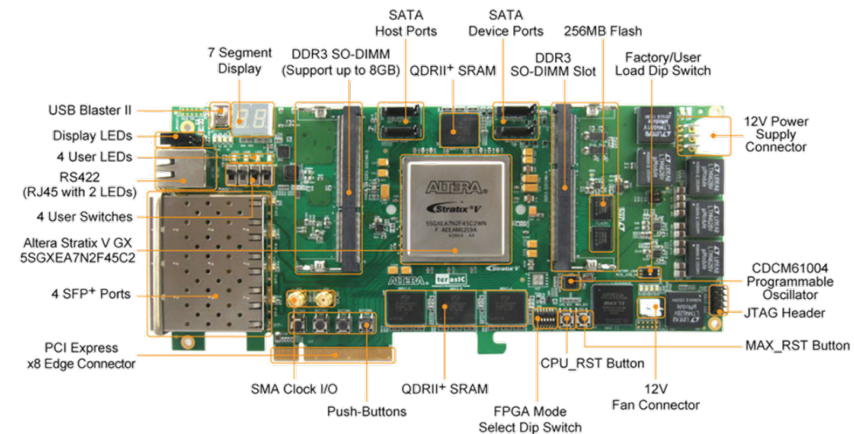
Latency measurements at different stages of data transfer from remote device to GPU memory

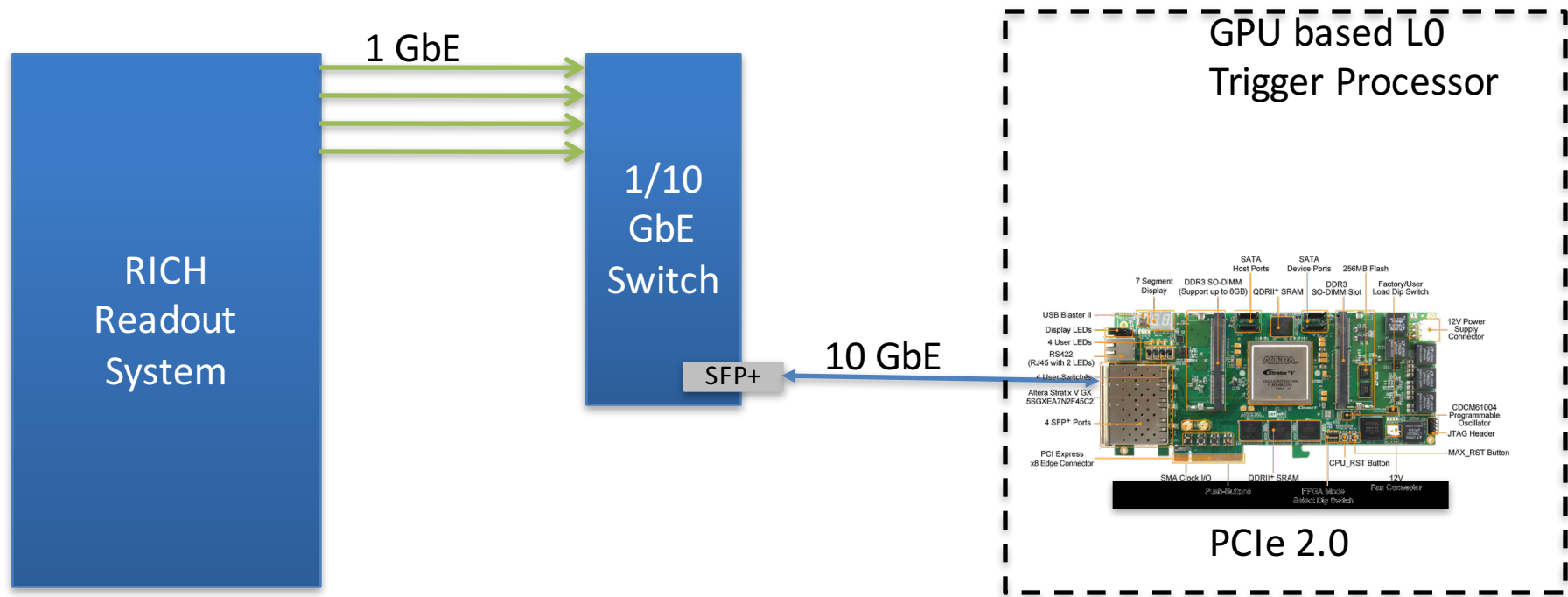


- ❑ Invio di dati reali dal RUN di NA62 del 2015
- ❑ Merging degli eventi nella GPU nVIDIA C2050
- ❑ Kernel Histogram
 - Piano XY diviso in una griglia
 - Viene creato un istogramma con la distanza fra i punti e gli hit dell'evento
 - I “cerchi” sono identificati analizzando i “bin” il cui contenuto eccede una soglia



- ❑ Implementato su Terasic DE5-NET basato su scheda di sviluppo Altera Stratix V
- ❑ PCIe x8 Gen3 (8 GB/s)
- ❑ 4 porte SFP+ (10GbE)
- ❑ 10GBASE-KR
- ❑ GPUDirect P2P/RDMA
- ❑ Gestione del network stack UDP in HW
- ❑ Stadio di merging degli eventi



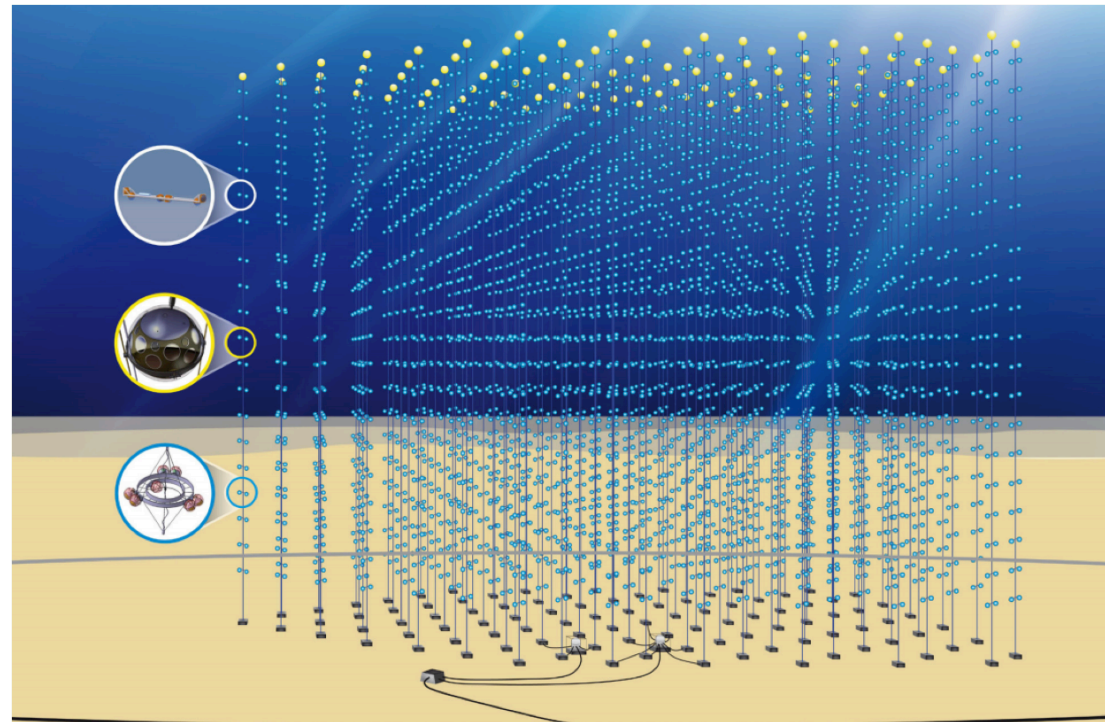


- ❑ NaNet-10 dovrà raggiungere le performance richieste per L0-TP
- ❑ Aggregazione del payload di 4 canali GbE in un singolo link 10 GbE attraverso lo Switch HP2920
- ❑ Funzionamento parallelo ed in modo “parassitico”
- ❑ Attualmente è in fase di test un sistema con singola porta 10GbE, PCIe Gen2 x8 da integrare nel setup di NA62 in Q2/2016

KM3NeT-IT è un apparato sperimentale sottomarino in costruzione per la rivelazione di neutrini di alta energia (TeV/PeV) basato sulla tecnica Čerenkov.

L'esperimento è localizzato a largo di Porto Palo di Capo Passero, Sicilia, e a circa 3500 m sotto il livello del mare. La stazione di terra è invece a Porto Palo di Capo Passero.

Il rivelatore sarà costituito da "stringhe" e "torri".



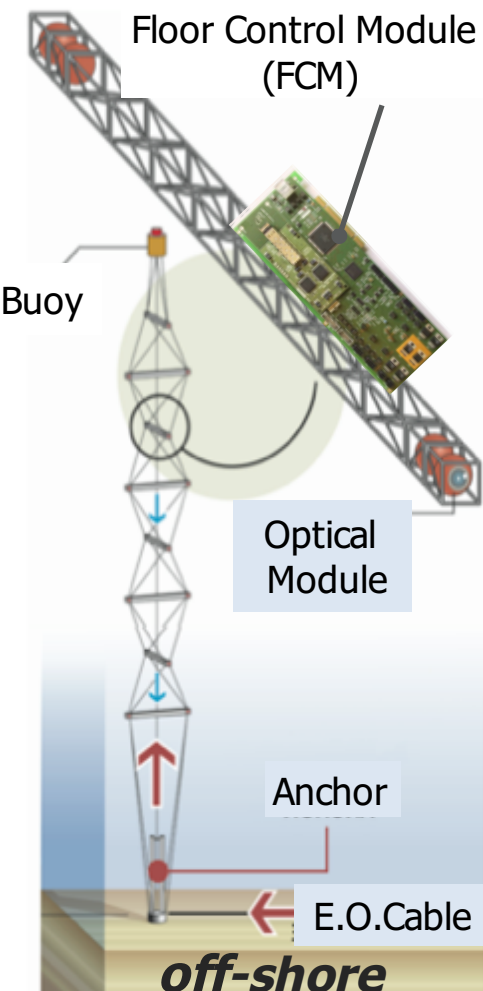
Una “torre” del rivelatore consiste in:

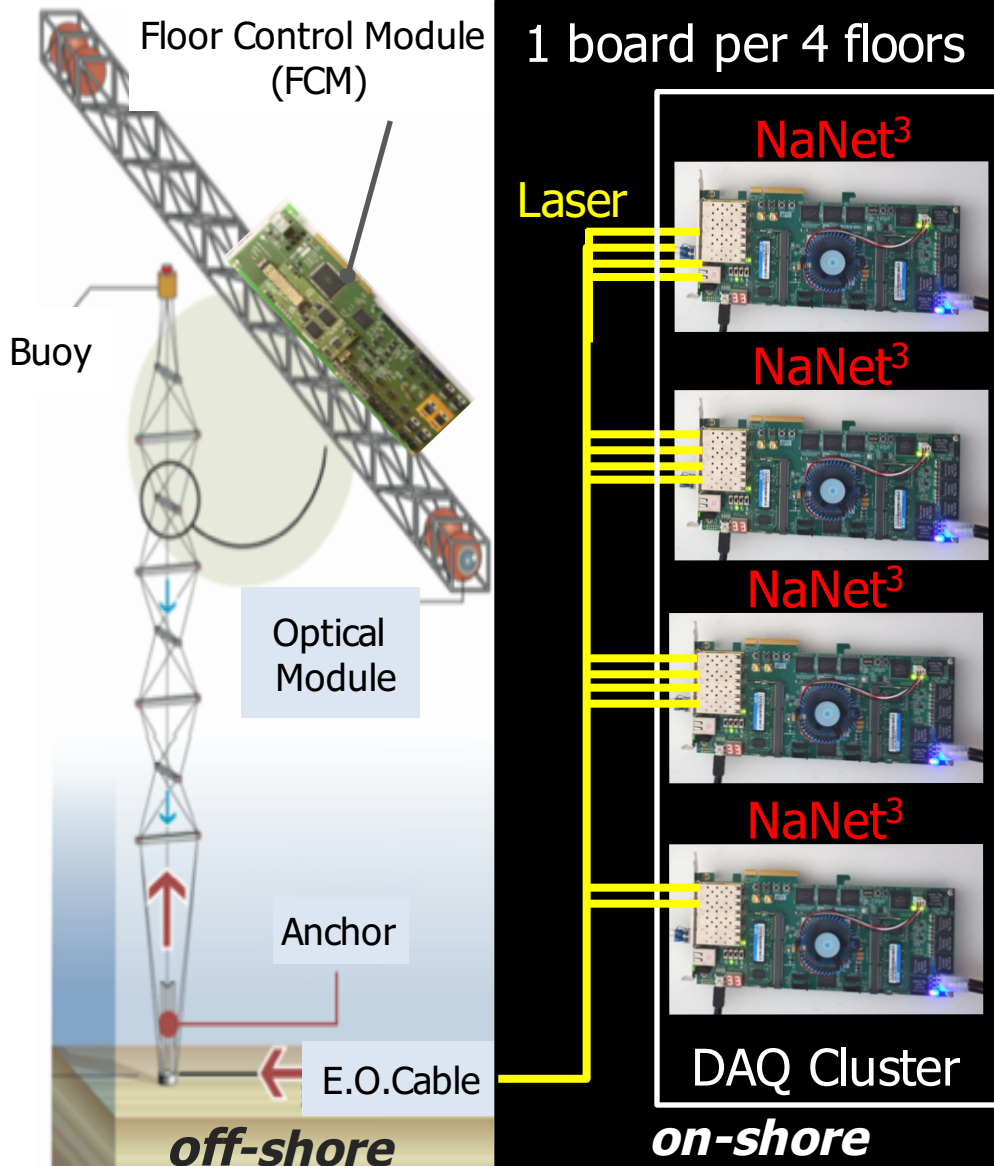
> 600 MB/s throughput

- ❑ **14 piani** distanziati fra loro di 20 m. Ogni piano ha:
 - Barre da 8m, con 6 Moduli Ottici (OM)
 - 2 idrofoni
 - Strumentazione oceanografica
- ❑ Una scheda FCM che gestisce la comunicazione tra il laboratorio a terra ed i dispositivi sottomarini (off shore), distribuendo le informazioni di timing (GPS clock) ed i segnali di slow-control ricevuti dai dispositivi di terra (on-shore)
- ❑ Un collegamento ottico con capacità di **2.5 Gb/s** per piano (800Mb/s payload) con protocollo TDM
- ❑ Progettazione iniziale: 2 schede FCM gemelle (on and off-shore) per piano

→ Problema evidente nello scalare col numero di torri:

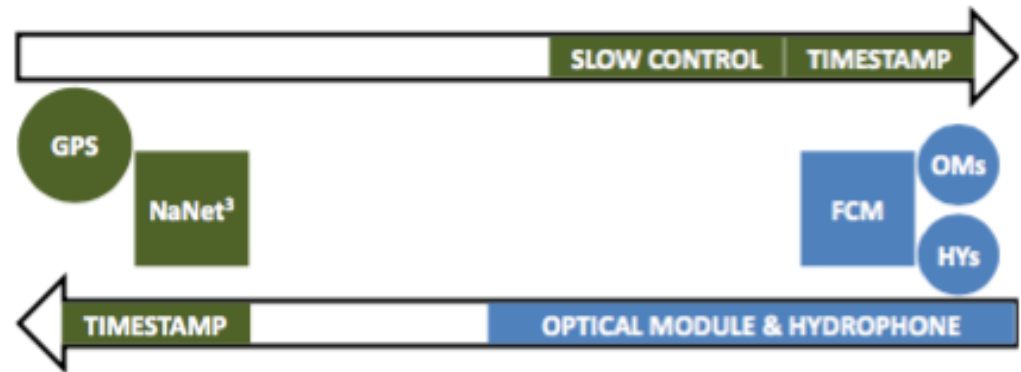
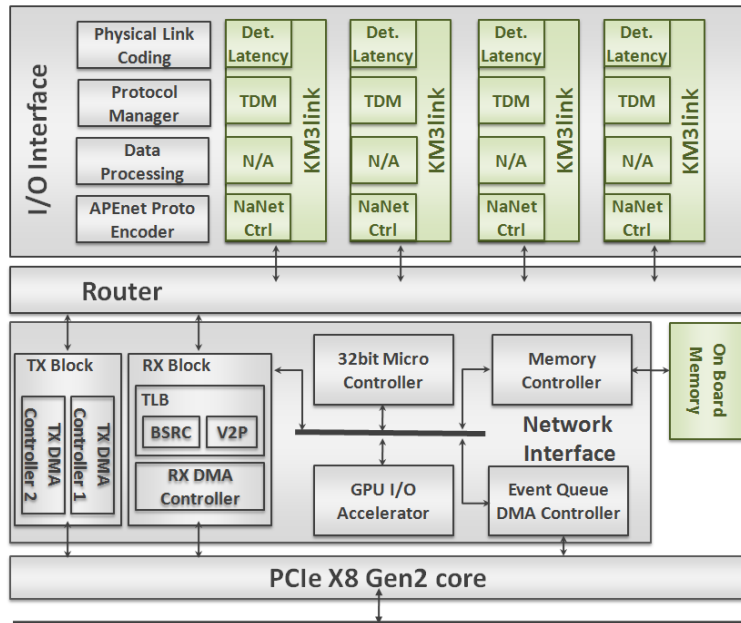
- Costi/Energia/Affidabilità del cluster DAQ che ospita le schede





- ❑ NaNet³ è attualmente utilizzata come controparte di 4 schede FCM a mare
 - Riduce di un fattore 3.5 il dimensionamento del cluster DAQ
- ❑ Requisiti: I canali devono sostenere la funzionalità di latenza deterministica per garantire un timing comune e un ritardo noto per il sistema di ReadOut spazialmente distribuito dell'esperimento
- ❑ Implementato su Terasic DE5-NET

- ❑ Specifiche di NaNet³ per la realizzazione di un sistema di RO ottimizzato:
 - Deterministic latency link: Distribuzione del clock con “latenza fissa” per il timestamp degli eventi sottomarini
 - Supporto al protocollo Time Division Multiplexing (TDM)
 - 4 canali per scheda per garantire riduzione di server, schede, ...



- ❑ Testbed: Server; NaNet; 4 FCM
- ❑ Test di slow-control
 - TX: da NaNet³ a FCM
- ❑ Test di acquisizione dati
 - Frame ID in slow_control_0
 - Data integrity; **168 ore di test**

- ❑ Famiglia di schede NaNet
 - 3 schede dal 2013: NaNet-1; NaNet³; NaNet-10
 - Sistema riconfigurabile e utilizzabile in contesti diversi
 - Tempi di sviluppo accettabili (elevata modularità e riuso di architetture consolidate)
 - Interfaccia ottimizzata nativa (GPUDirect P2P/RDMA) verso GPU nVIDIA
- ❑ Possibilità di sfruttare i vantaggi di sistemi basati su FPGA e di sistemi basati su architetture di calcolo più complesse (GPU, CPU)
- ❑ Sviluppi futuri:
 - Finalizzazione di NaNet-10
 - Migliorare le performance con tecnologie di punta (PCIe gen3, 40GbE, 100GbE, SOC, OpenCL)
 - Aumentare i campi di applicazione

❑ Esperimento di Gruppo V (Roma1, Roma2, LNF)
coordinatore A. Lonardo

❑ Collaborazione NaNet:

**F. Ameli^(a), R. Ammendola^(b), A. Biagioni^(a), A. Cotta Ramusino^(c),
M. Fiorini^(c), O. Frezza^(a), G. Lamanna^(d), F. Lo Cicero^(a),
A. Lonardo^(a), M. Martinelli^(a), I. Neri^(c), P. S. Paolucci^(a),
E. Pastorelli^(a), L. Pontisso^(f), D. Rossetti^(e), F. Simeone^(a),
F. Simula^(a), M. Sozzi^(f), P. Vicini^(a)**

(a) INFN Sezione di Roma

(b) INFN Sezione di Roma Tor Vergata

(c) Università di Ferrara e INFN Sezione di Ferrara

(d) INFN LNF and CERN

(e) nVIDIA Corporation, USA

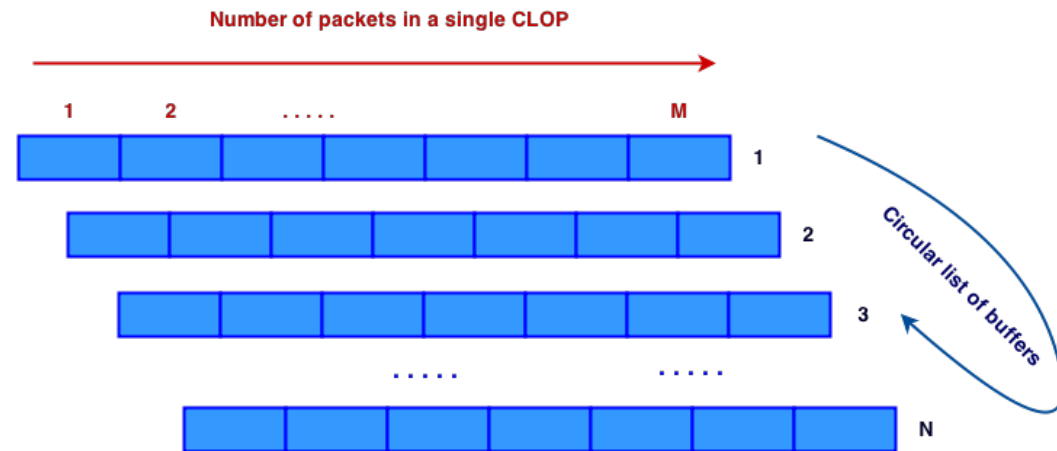
(f) INFN Sezione di Pisa and CERN

❑ CLOP: Circular List Of Persistent buffer

- A data frame (8kB) per buffer
- 2 pages of CPU memory

❑ Hardware

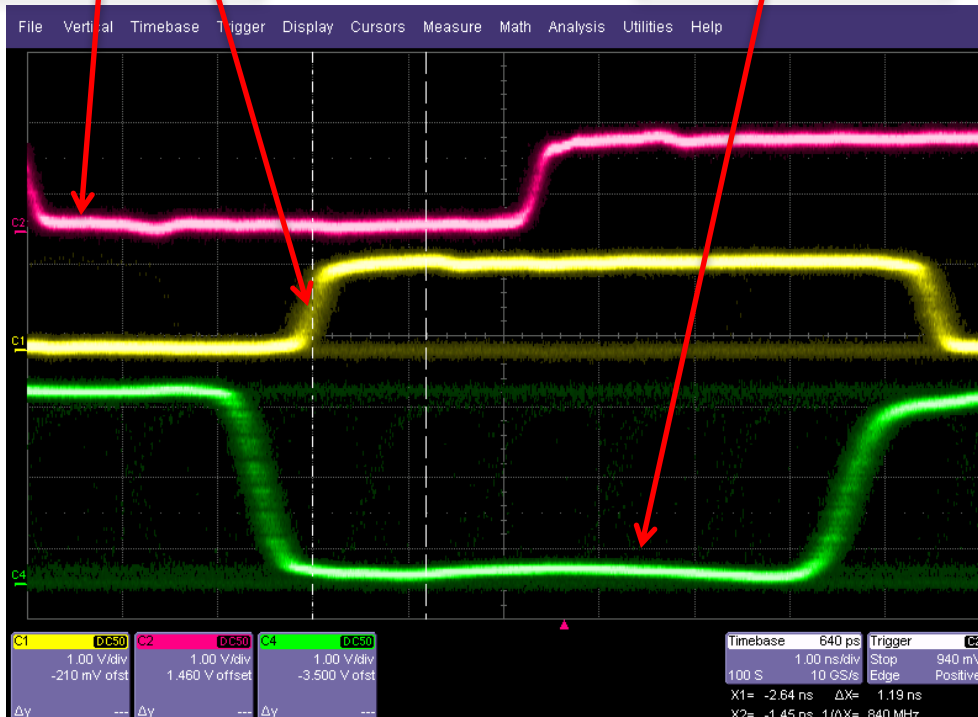
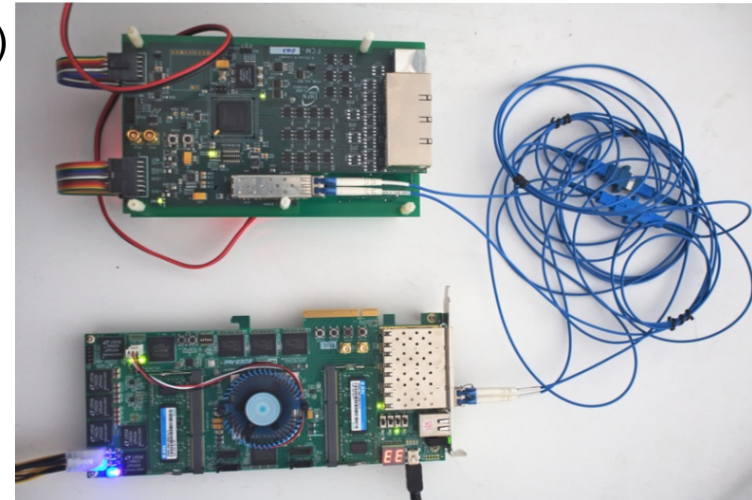
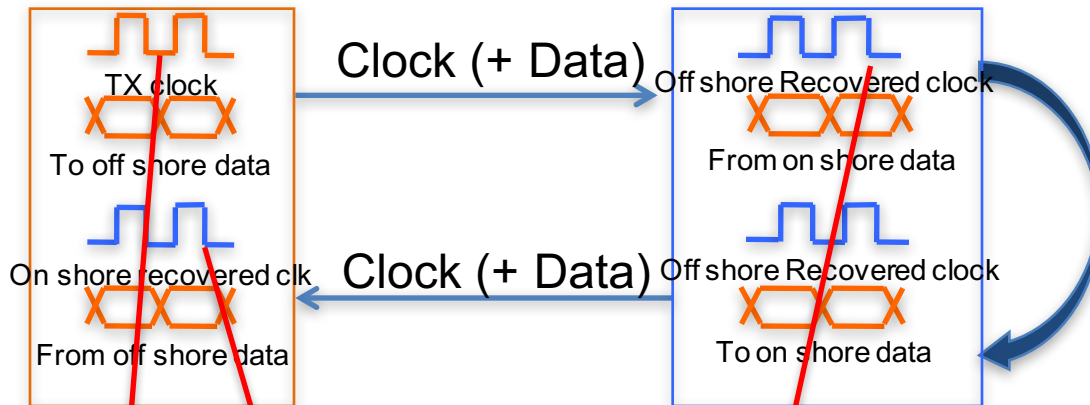
- Stable: no latency fluctuation
- Translation Lookaside Buffer:
 - 256 entries
 - 2 entries per frame
 - 32 buffer per channel
- Maximum execution time: 4 ms



Deterministic latency link inter-operability

NaNet3 On-shore (StratixV)

FCM Off-Shore (Virtex5)



- ❑ Testbed: FCM vs Terasic DE5-Net
 - Custom hw mode for FCM Transceivers (Xilinx)
 - Latency deterministic mode for Stratix V Transceiver
 - 2mt copper and 2 mt long fiber
- ❑ Test:
 - 12 hours of periodic (~s) Tx clock reset to verify pll locking and rx word alignment

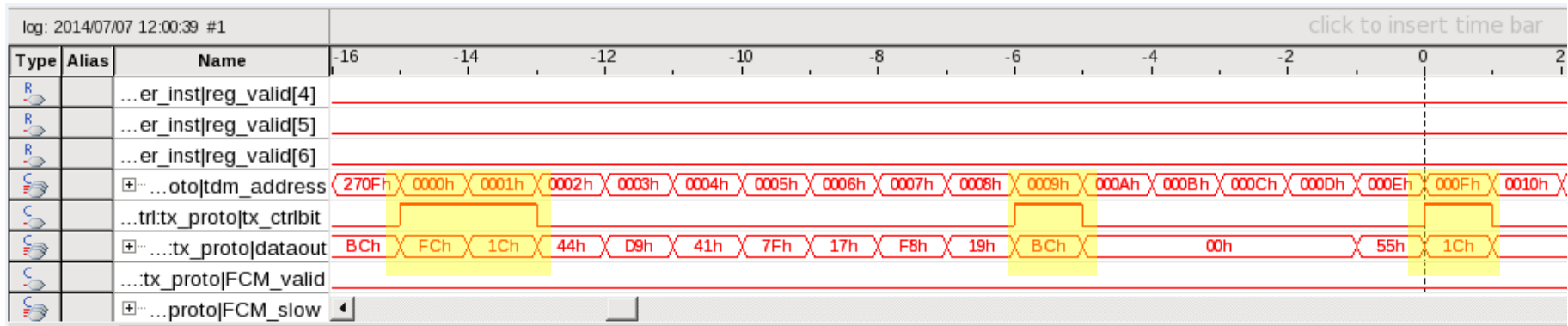
NaNet I/O interface Customization:

- TDM implementation: compliant with KM3NeT-IT specification, echo test OK!

	K Code	Kout	D
IDLE_CODE	K28.5	1	0xBC
FRAME_CODE	K28.7	1	0xFC
DATA_CODE	K28.0	1	0x1C

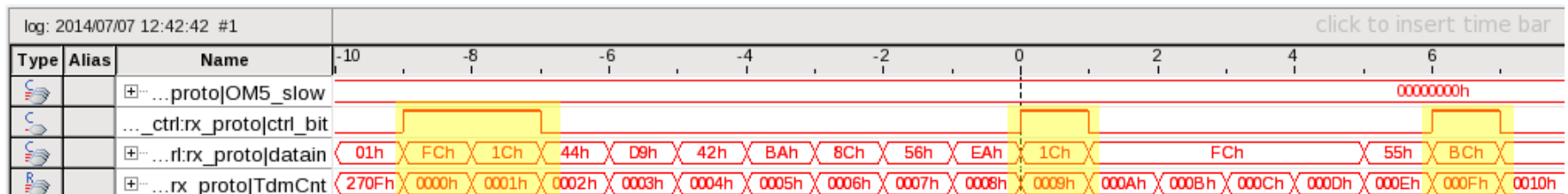
- NaNet³ TX

Field	Validity Code	Start Address	Stop Address	Length [byte]
Start Frame	FRAME_CODE	0x0000	0x0000	1
Frame Time	DATA_CODE	0x0001	0x0008	1+6+1
FCM Slow Control	DATA_CODE	0x0009	0x000E	1+4+1
OM0 Slow Control	DATA_CODE	0x000F	0x0014	1+4+1
OM1 Slow Control	DATA_CODE	0x0015	0x001A	1+4+1
OM2 Slow Control	DATA_CODE	0x001B	0x0020	1+4+1
OM3 Slow Control	DATA_CODE	0x0021	0x0026	1+4+1
OM4 Slow Control	DATA_CODE	0x0027	0x002C	1+4+1
OM5 Slow Control	DATA_CODE	0x002D	0x0032	1+4+1
Hydro data	No IDLE_CODE	0x0200	0x03FF	512
OM0 Phy Data	No IDLE_CODE	0x0400	0x07FF	1024
OM1 Phy Data	No IDLE_CODE	0x0800	0x0BFF	1024
OM2 Phy Data	No IDLE_CODE	0x0C00	0x0FFF	1024
OM3 Phy Data	No IDLE_CODE	0x1000	0x13FF	1024
OM4 Phy Data	No IDLE_CODE	0x1400	0x17FF	1024
OM5 Phy Data	No IDLE_CODE	0x1800	0x1BFF	1024

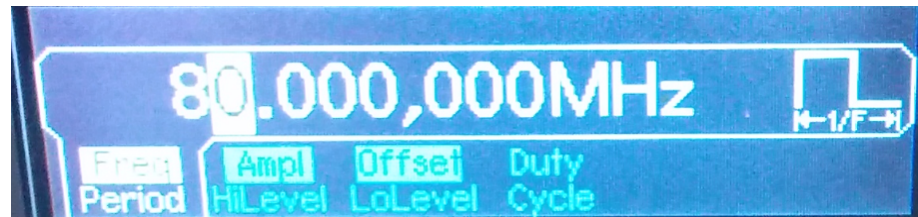
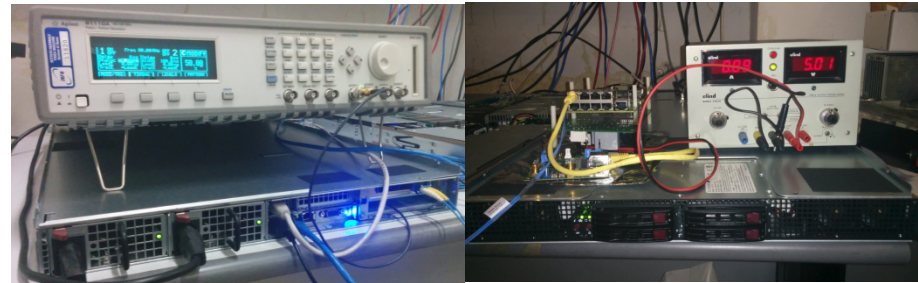


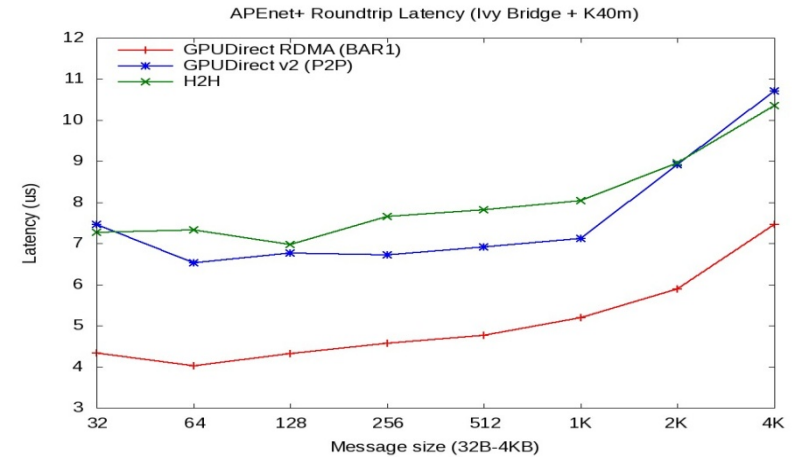
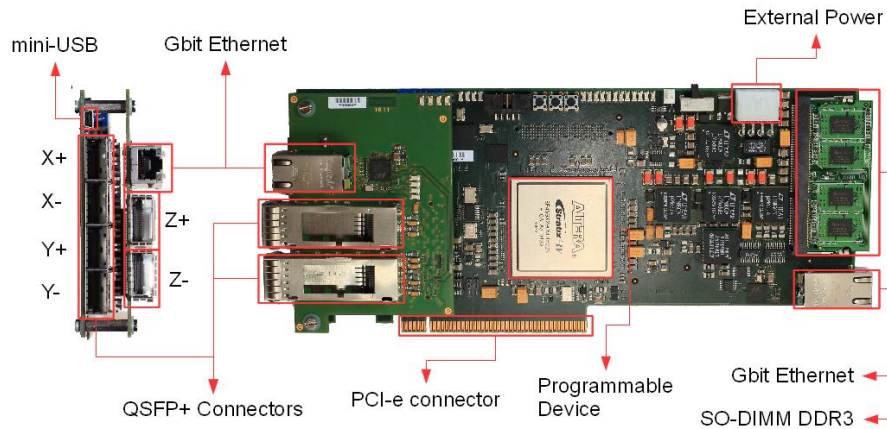
- NaNet³ RX:

Data flow: NaNet³ → FCM board → NaNet³

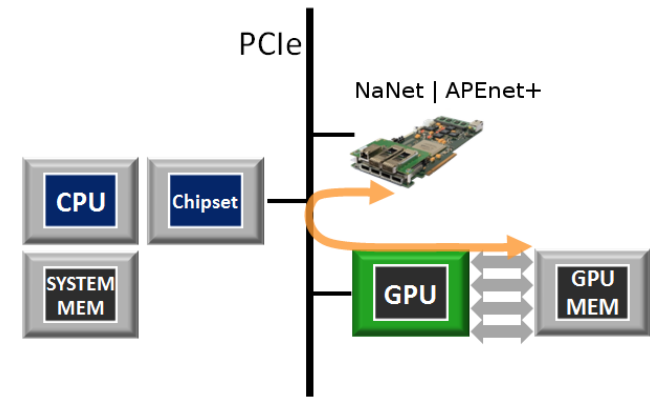


- ❑ Testbed Environment
 - Clock generator (80MHz)
 - FCMserver (see Matteo Manzali talk)
 - NaNet³
 - FCM + FEM
- ❑ Echo test (single channel)
 - NaNet/FCM roundtrip
- ❑ Test di slow-control
 - TX: from NaNet³ to FCM
- ❑ Data Acquisition test
 - Frame ID in slow_control_0
 - Data integrity
 - **48 hours test passed**
- ❑ Work-in-progress
 - Integration with FCMserver application (also to validate single channel implementation)
 - Multi-channel support



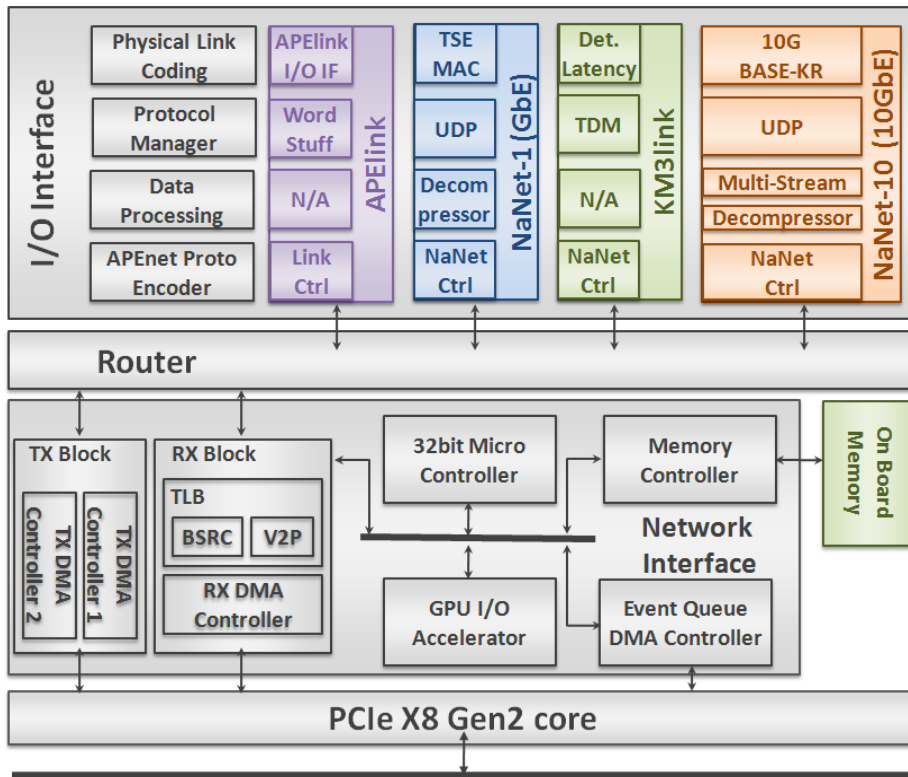


- ❑ High Performance Computing
- ❑ Point-to-point, low-latency, high-throughput implementing a 2D/3D torus topology
- ❑ PCI Express x8 gen2
- ❑ Up to 6 fully bidir 34 Gbps/channel over QSFP+.
- ❑ RDMA Hardware support for CPU offload
- ❑ NVIDIA GPUDirect RDMA/P2P HW support
 - GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement.
 - No bounce buffers on host memory .
 - Latency reduction for small messages!!!

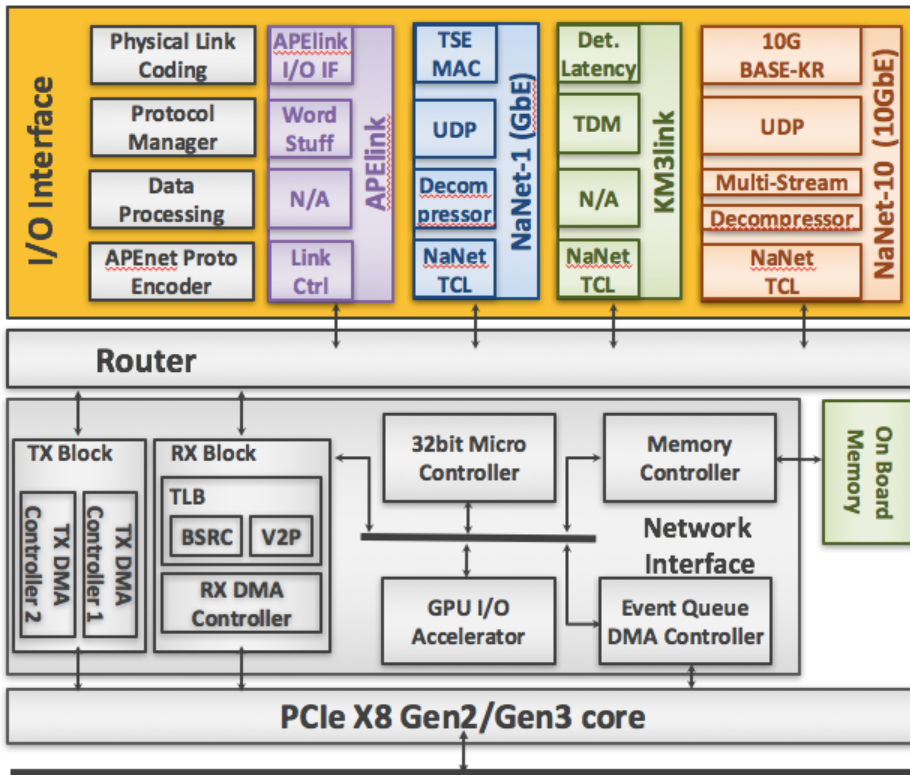


- ❑ Future HEP experiments on rare decays
 - Detection signature similar to that of the huge background
 - No simple selection algorithms in hardware
 - Not possible to work on a subset of detectors data
 - Trigger-less approach (see also LHCb upgrade...)
- ❑ Read out all detectors data in event builder computing nodes memories
 - Flexible I/O (different kind and number of I/O channels)
 - High Bandwidth (saturate computing node expansion bus)
- ❑ On the fly processing for data reduction
 - low and stable data transport latency (real-time operations)
- ❑ Integration of many-core accelerators to limit computing farm numerosity

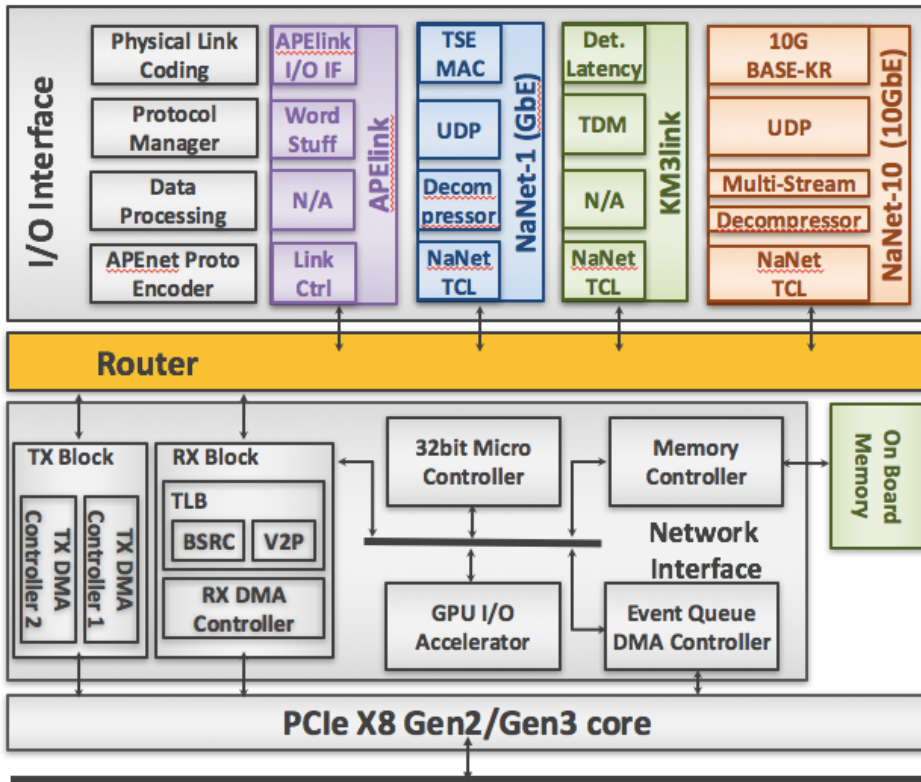
- ❑ **Interfaccia di I/O**
 - Supporto per molteplici canali.
 - Offloading della CPU dalla gestione dei
 - Stadio di processamento del flusso dati per ottimizzazione dell'applicazione
- ❑ **Router**
 - Interconnette dinamicamente le porte di I/O con le porte della NI
- ❑ **Network Interface**
 - Gestisce i pacchetti TX/RX da e verso la memoria GPU/CPU
 - Zero-copy RDMA
 - Interfaccia ottimizzata verso le GPU
 - TLB & Nios II Microcontroller per la gestione della memoria virtuale
- ❑ **PCIe X8 Gen2/Gen3 Core**



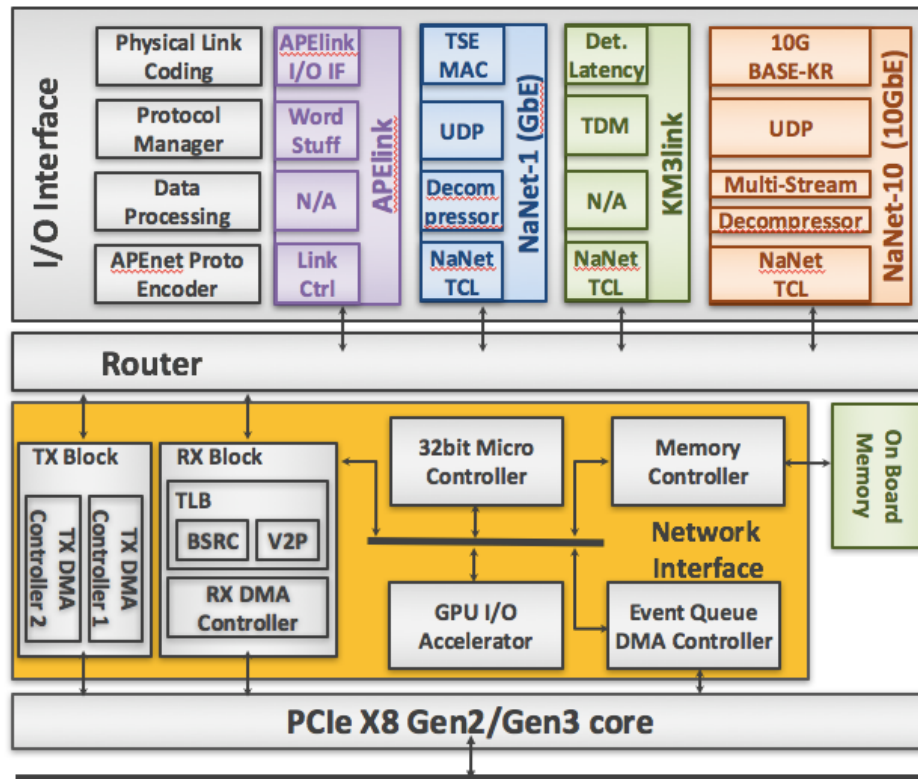
- ❑ Physical Link Coding
 - Standard: **1GbE** (1000Base-T), **10GbE** (10Base-KR)
 - Custom: **APElink** (34Gbps QSFP), **KM3link Deterministic Latency** (2.5 Gbps optical)
- ❑ Protocol Manager
 - Data/Network/Transport layers offload (UDP, TDM)
 - Minimizzazione delle fluttuazioni della latenza
- ❑ Data Processing
 - Processamento del flusso dati
 - **Decompressore & Merger** per l'esperimento **NA62**: riformattazione dei dati fusione di frammenti di evento prima dell'invio dei pacchetti verso la NI
- ❑ APEnet protocol encoder
 - Adattamento del protocollo tra on-board



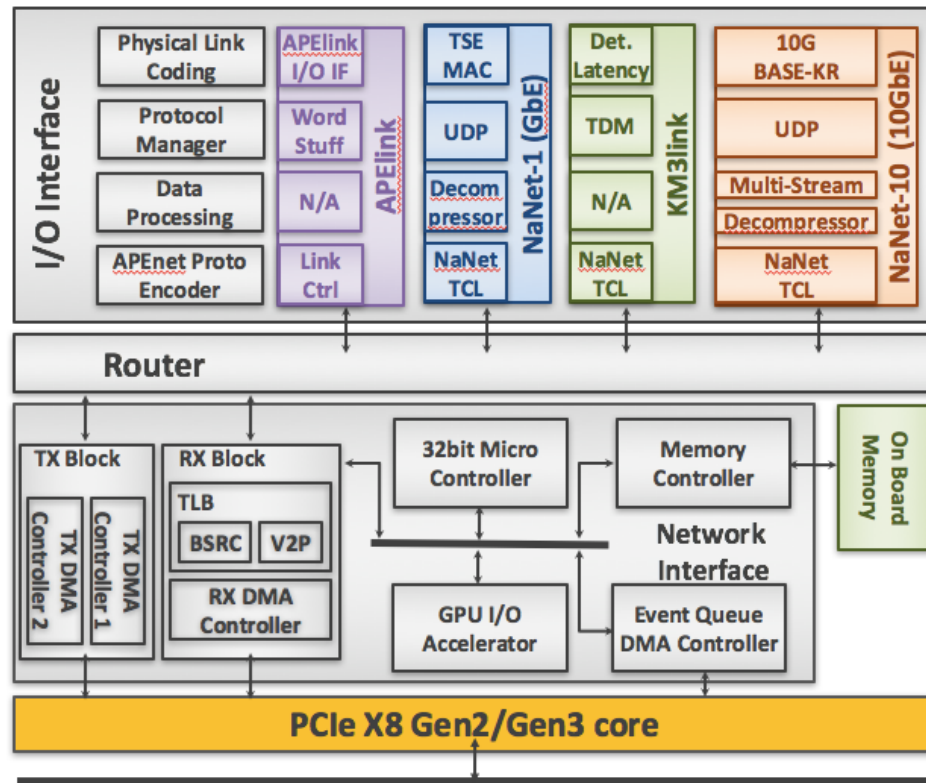
- ❑ Crossbar switch con 6 porte bidirezionali
- ❑ Router: interconnette dinamicamente le porte di I/O e NI
- ❑ Arbitraggio: risolve la contention delle richieste delle porte (static, round-robin)
- ❑ Supporta fino a 12 flussi contemporanei di 2.8 GB/s (la versione futura basata su interfaccia PCIe Gen 3 doppierebbe le attuali performance)
- ❑ Numero di porta riconfigurabile

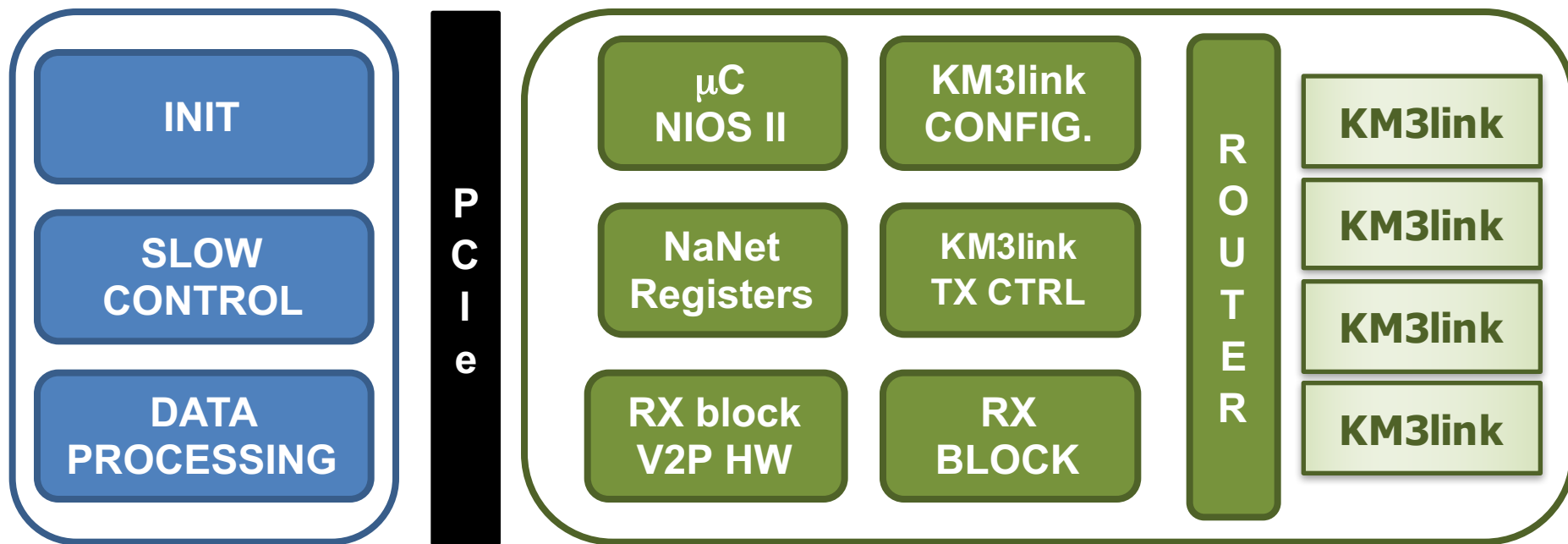


- ❑ TX: raccoglie i dati dall'interfaccia PCIe e li reindirizza verso la porta di destinazione
- ❑ RX: esegue l'operazione di ricezione dati **zero-copy RDMA** gestendo la traduzione di indirizzo CPU/GPU da virtuale a fisico
 - Translation Lookaside Buffer (memoria associativa)
 - Microcontrollore in caso di miss
- ❑ Accelerazione dell'interfaccia I/O verso la GPU (GPUDirect RDMA/P2P)
- ❑ TX/RX blocco
 - Istanza di transazioni PCIe simultanee tramite molteplici motori DMA
- ❑ Microcontrollore Altera NIOS II

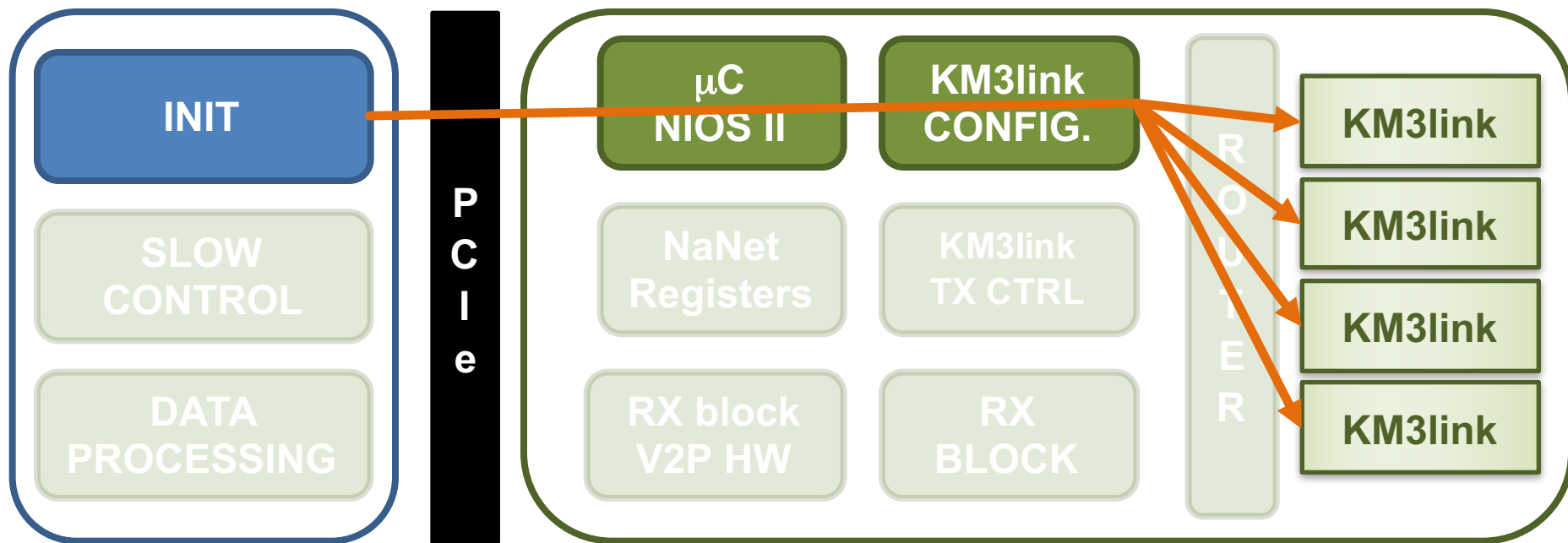


- ❑ PCIe x8 Gen2 Core (stabile)
 - Basato su PLDA EZDMA Gen2 hard IP core.
 - CPU/GPU BW: 2.8 GB/s write, 2.5 GB/s read
- ❑ PCIe x8 Gen3 Core (testing)
 - Basato su PLDA XpressRICH Gen3 soft IP core
 - CPU BW: 4.5 GB/s write, 4.0 GB/s read
 - GPU BW: 2.8 GB/s write, 3.0 GB/s read
- ❑ PCIe x8 Gen3 Core home-made (sviluppo)
 - Basato su Altera PCIe hard-IP
- ❑ Ivy Bridge, Kepler K20 setup



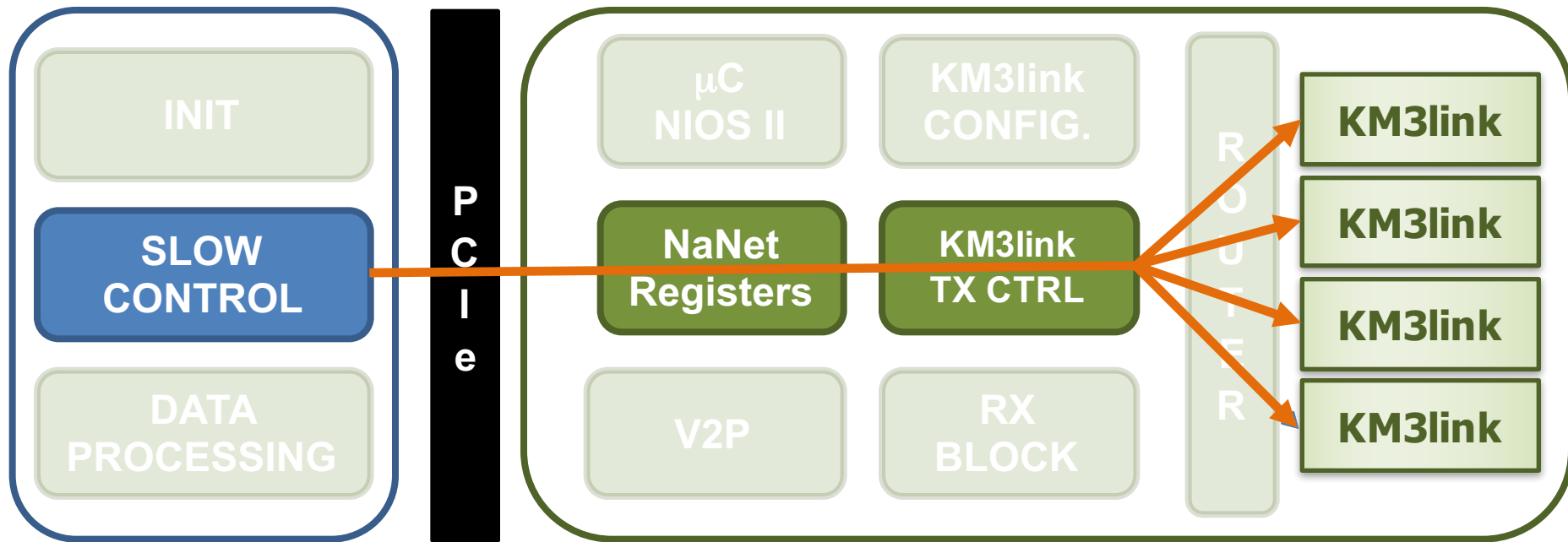


- ☐ SW/HW Initialization
- ☐ Slow Control Management
- ☐ Data Reception and Processing



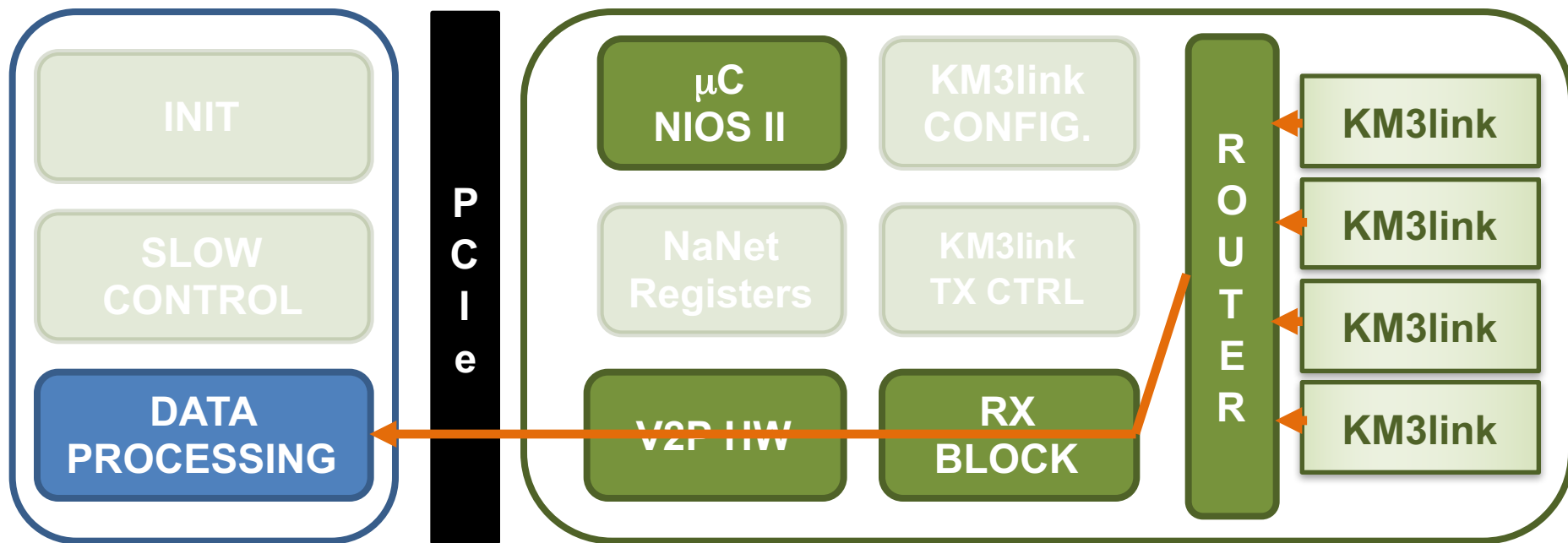
- ☐ CLOP buffer registration
- ☐ Pages pinning
- ☐ Virtual/Physical address
- ☐ Driver/μC communication

- ☐ Virtual/Physical address in Nios memory
- ☐ KM3link Configuration
- ☐ KM3link Alignment
- ☐ KM3link enabling



- ☐ TX registers interface
- ☐ Slow control buffer management

- ☐ Slow control messages distributed over the KM3links
- ☐ Slow control encapsulated in TDM streams
- ☐ Checksum for data corruption avoidance



- ❑ Signalling to the application (FCM server) 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 (clon buffer)