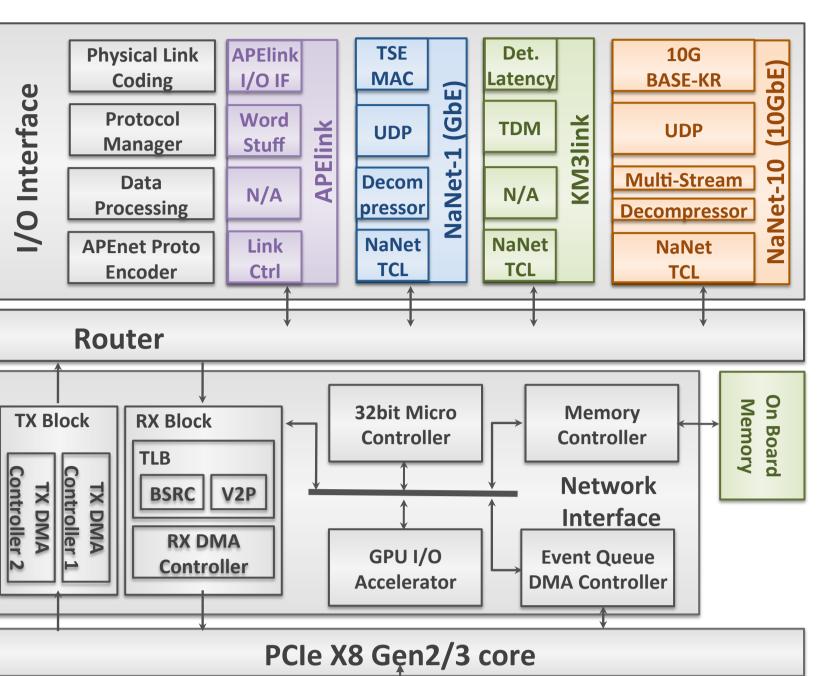
The NaNet Project: Heterogeneous Real-Time Stream Processing in the Low Level Trigger of the NA62 Experiment

R. Ammendola^(b), M. Barbanera^(d), A. Biagioni^(a), P. Cretaro^(a), G. Lamanna^(c,d), F. Lo Cicero^(a), A. Lonardo^(a), M. Martinelli^(a), P.S. Paolucci^(a), E. Pastorelli^(a), R. Piandani^(d), L. Pontisso^(a), <u>D. Rossetti</u>^(a,e), F. Simula^(a), M. Sozzi^(c,d), P. Vicini^(a)

(a) INFN Sezione di Roma (b) INFN Sezione di Roma Tor Vergata (c) Università di Pisa (d) INFN Sezione di Pisa (e) NVIDIA Corporation

Our work aims at improving the performances of the NA62 low-level trigger implementing a real-time stream processing architecture based on an orchestrated combination of **heterogeneous computing devices** (CPUs, FPGAs and GPUs). To enable it we devised NaNet, a FPGA-based PCI-Express Network Interface Card with processing and GPUDirect capabilities, which supports multiple link technologies (1/10/40GbE and custom ones). We have demonstrated the effectiveness of the method by retrofitting the RICH detector to generate refined physics-related primitives. Results obtained during the first months of 2017 run are presented and discussed, along with a description of the latest developments in the NaNet architecture.

NaNet Design

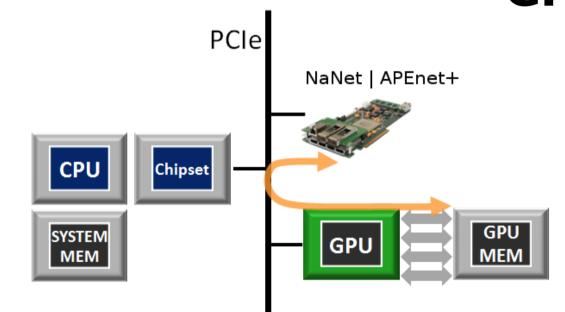


Design and implementation of a **family of** FPGA-based PCIe Network Interface Cards:

- Bridging the front-end electronics and the software trigger computing nodes.
- Supporting multiple link technologies and network protocols.
- Enabling a low and stable communication latency.
- Having a high bandwidth:
 PCI Gen2:
 - CPU: 2.8 GB/s Read, 2.5 GB/s Write GPU: 2.5 GB/s Read & Write
 - PCI Gen3: CPU: 4.8 GB/s Read, 4.2 GB/s Write
- Processing data streams from the network channels on the fly (data compression/ decompression, re-formatting ...).

GPU: 3 GB/s Read & Write

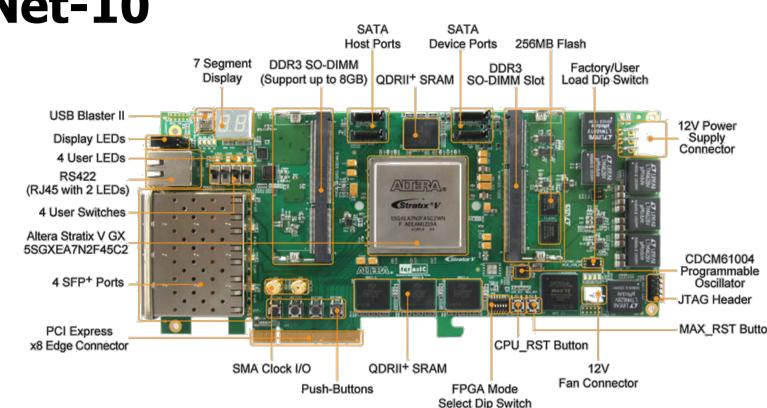
GPUDirect P2P/RDMA



- ☐ GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement.
- □ No bounce buffers on host memory.
- □ Zero copy I/O.
- □ Latency reduction for small messages.
- □ NVIDIA Fermi/Kepler/Maxwell/Pascal

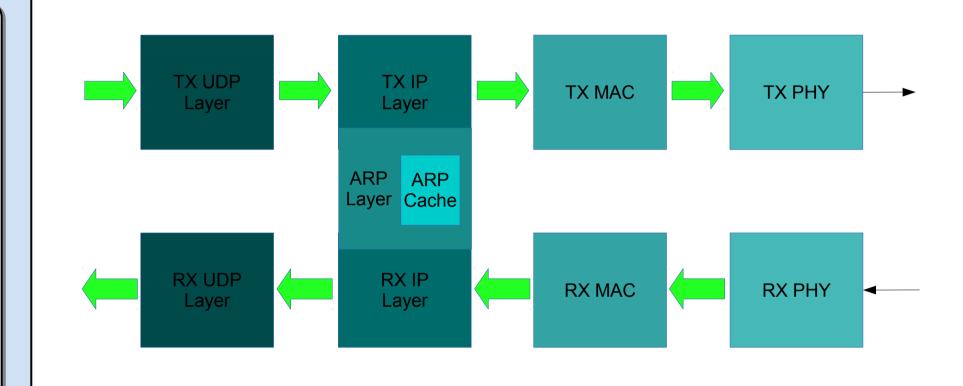
NaNet-10

- □ Altera Stratix V Terasic DE5-NET dev board
- □ 4 SFP+ ports (Link speed up to 10 Gb/s)
- □ PCIe X8 Gen2/3
- □ GPUDirect P2P/RDMA capability
- □ UDP offload support
- □ Processing stage for data decompression and events merging



NaNet-40

- □ A novel device is on advanced state of development in order to enable 40 Gb Ethernet with UDP/IP protocol hardware offload.
- □ Bittware S5-PCIe-HQ board with Altera Stratix V
- □ 2 QSFP+ ports
- Copper and Optical cable tested





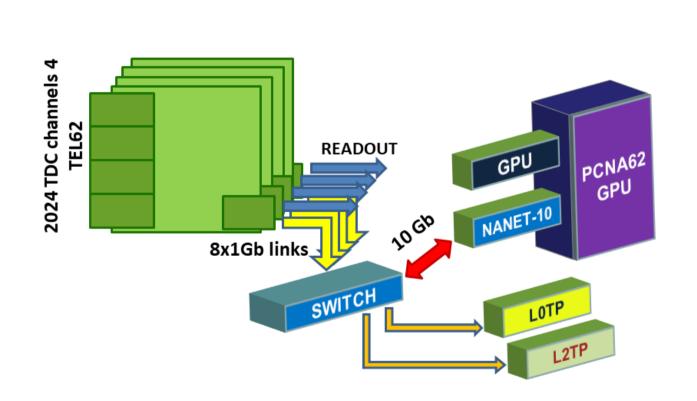
- Physical and MAC layer from Altera library
- □ IP and UDP protocol layer with hardware offloading
- □ Fully compliant ARP layer management with 256-entries ARP cache
- □ PCIe X8 Gen3

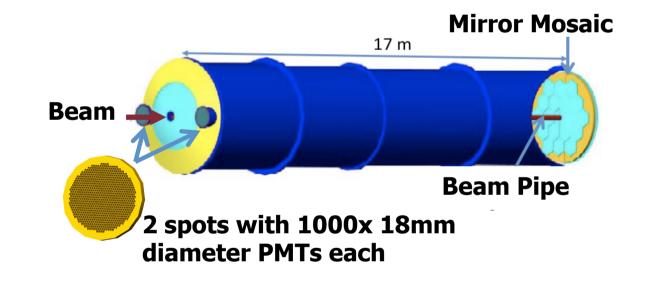
Physics Case: Real-time Rings Reconstruction for the NA62 RICH Detector

The NA62 experiment at CERN aims at measuring the branching ratio of the ultra-rare decay: $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ (BR ~8x10-11)

Ring-Imaging Cherenkov (RICH) detector is used:

the particles generate a circular footprint radiation beam onto the light-sensitive tubes of two photomultipliers (PMT) arrays. In the standard implementation, the FPGAs on the TEL62 readout boards compute simple "trigger primitives" on the fly, such as hit multiplicities and rough hit patterns, which are then sent to a central processor for matching and trigger decision, with a time budget of $800~\mu s$.



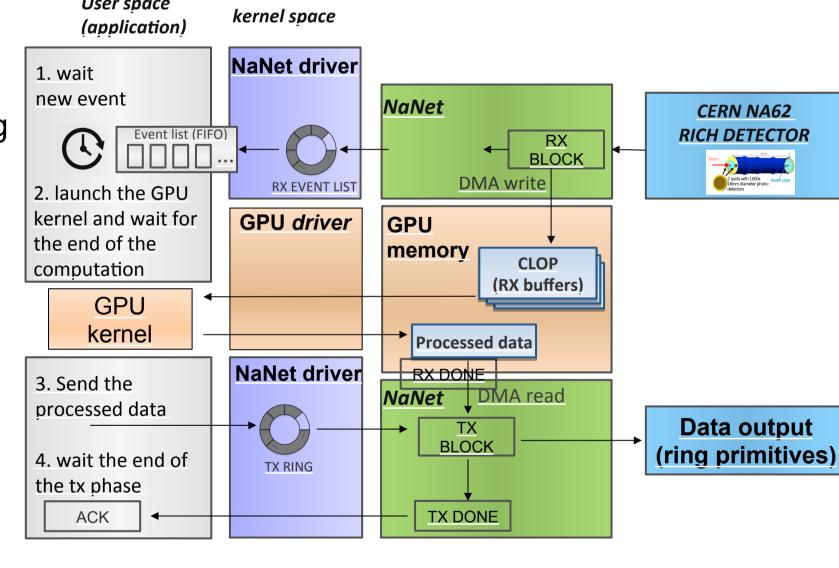


We then added a **GPU-based processing stage** between the RICH detector readout and the L0 trigger processor (L0TP) with the task of generating, in real-time, physics-related primitives (i.e. centers and radii of Cerenkov ring patterns on the photomultipliers arrays), in order to boost the L0 trigger discrimination power. Results can be also be sent to the High Level Trigger (L2TP) saving computing time to server farm.

- Specific histogram-based algorithm developed for trackless, fast, and high resolution ring fitting
- Detection of particle speed (radius) and direction (center)

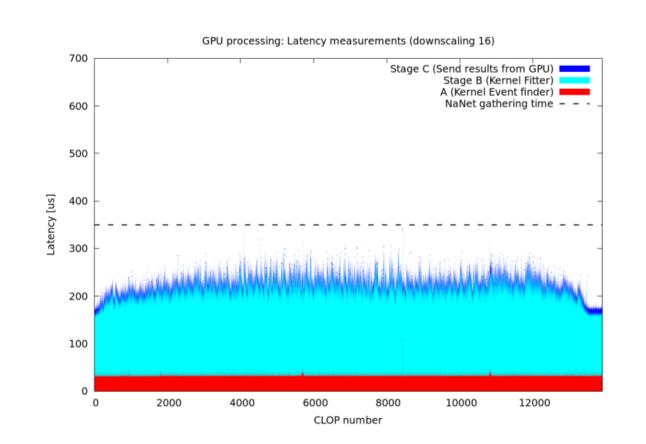
NaNet NIC DMA-writes data from theTEL62 readout boards of the detector directly into the GPU memory (through PCIe bus) using GPUdirect RDMA

- NaNet NIC DMA-writes a "receiving done" event in a memory region called "event queue"
- trapped by a kernel-space device driver
- notified to the user application which launches a CUDA kernel to process the data using the GPU

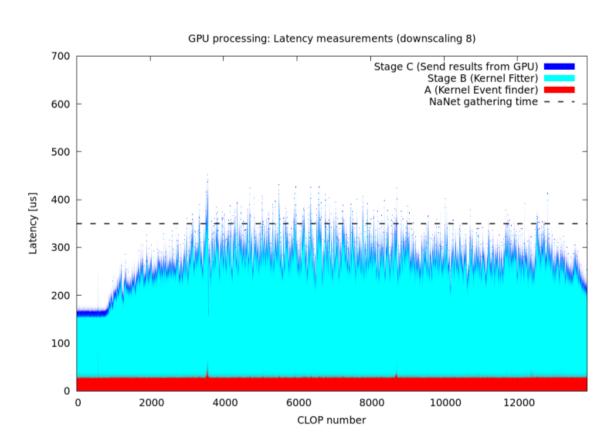


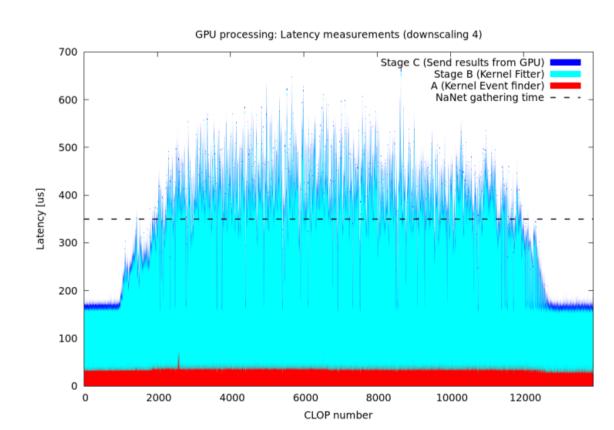
- Results of the processing i.e. number and kind (electron, pion, muon) of rings is eventually sent via NaNet board to the trigger system that collects data from all detectors:
 - data are DMA-read directly from GPU memory;
 - the kernel device driver (invoked by the user application on HOST) instructs the NIC by filling a "descriptor" into a dedicated, DMA-accessible memory region called "TX ring";
 - the presence of new descriptors is notified to NaNet by writing on a doorbell register over PCIe;
 - NaNet NIC issues a "tx done" completion event in the "event queue".

Preliminary Results from 2017 Run



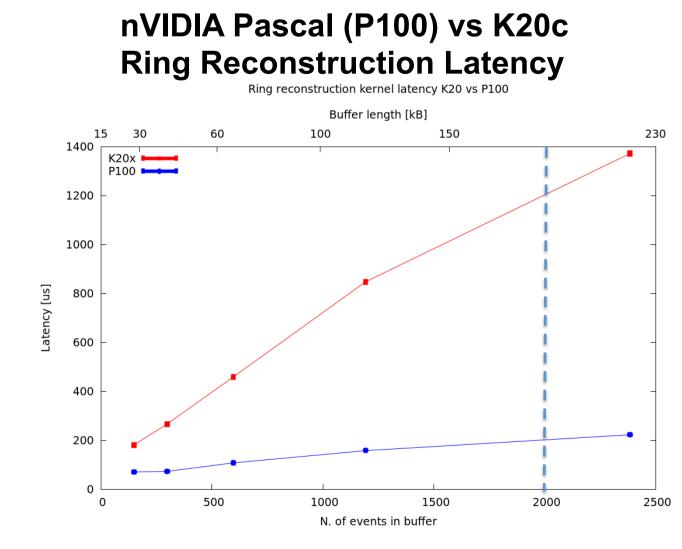
- □ Supermicro X9DRG-QF Intel C602
 Patsburg Intel Xeon E5-2602 2.0 GHz
- □ NaNet-10 (Terasic DE5-Net)□ NVIDIA K20c
- ☐ Gathering time: 350us (dashed line)
- □ ~ 60% target beam intensity
- ☐ Histogram algorithm
- (~ 18*10¹¹ Pps)





- □ The latency of event indexing in the GPU CLOP buffer is ~ constant (red)
 □ The latency of the ring reconstruction GPU kernel increases with the number of events in the CLOP buffer (cyan)
- □ Real-time stream processing: processing time <= events gathering time
 □ Reaching a downscaling factor of 4 (i.e. processing only every fourth
- event), the latency of GPU processing exceeds the events gathering time:

we need to speed-up the ring reconstruction!



On a 2000 events buffer: K20c ~ 500 ns/evt P100 ~ **100 ns/evt**

NVIDIA Pascal P100 is a potential solution.



