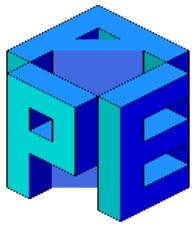


NaNet: a family of FPGA-based Network Interface Cards for Real-Time Trigger and Data Acquisition Systems in HEP Experiments

Alessandro Lonardo
INFN Sezione di Roma
(on behalf of the NaNet Collaboration)

Workshop della Commissione Calcolo e Reti dell'INFN
LNF, 25 – 29 Maggio 2015

NaNet Collaboration

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)], L. Tosoratto[(a)], 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

Project Motivation

- 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

Project Objectives

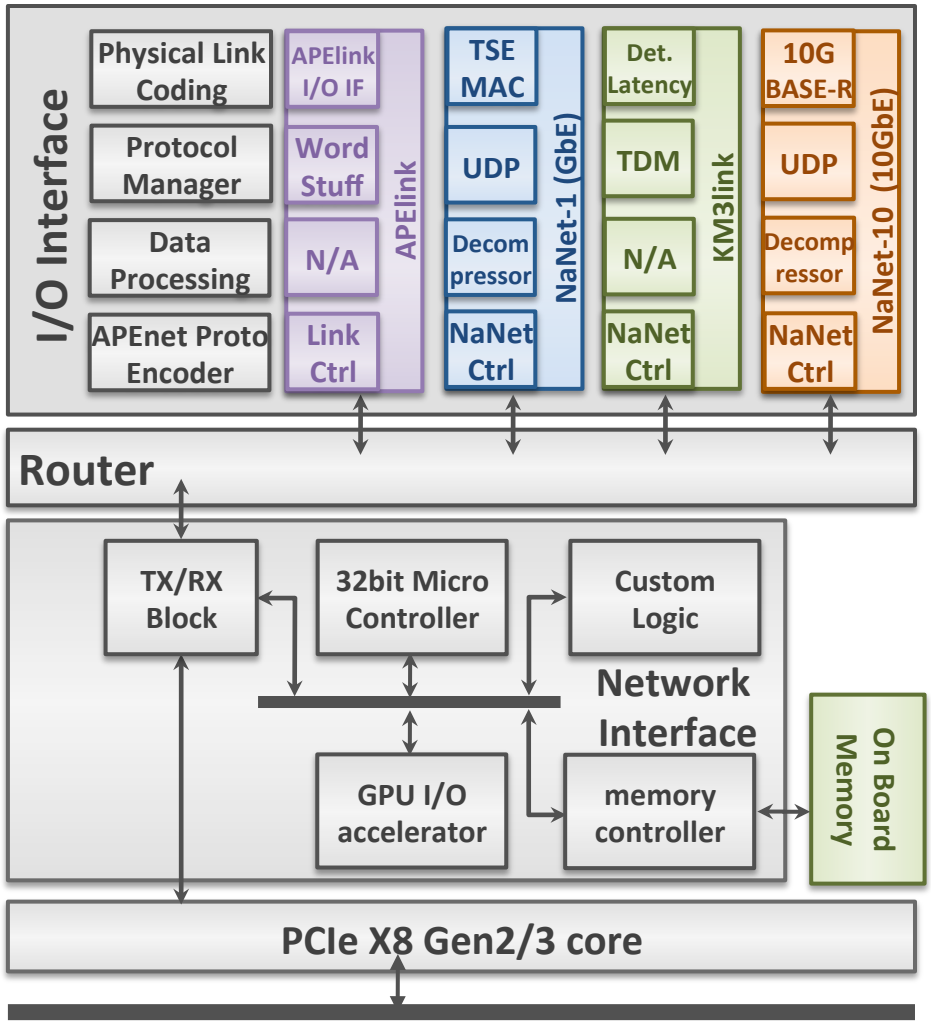
Design and implementation of a family of FPGA-based PCIe Network Interface Cards:

- 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.
- Processing data streams from detectors on the fly (data compression/decompression and re-formatting, coalescing of event fragments, …).
- Optimizing data transfers with GPU accelerators.

Project History

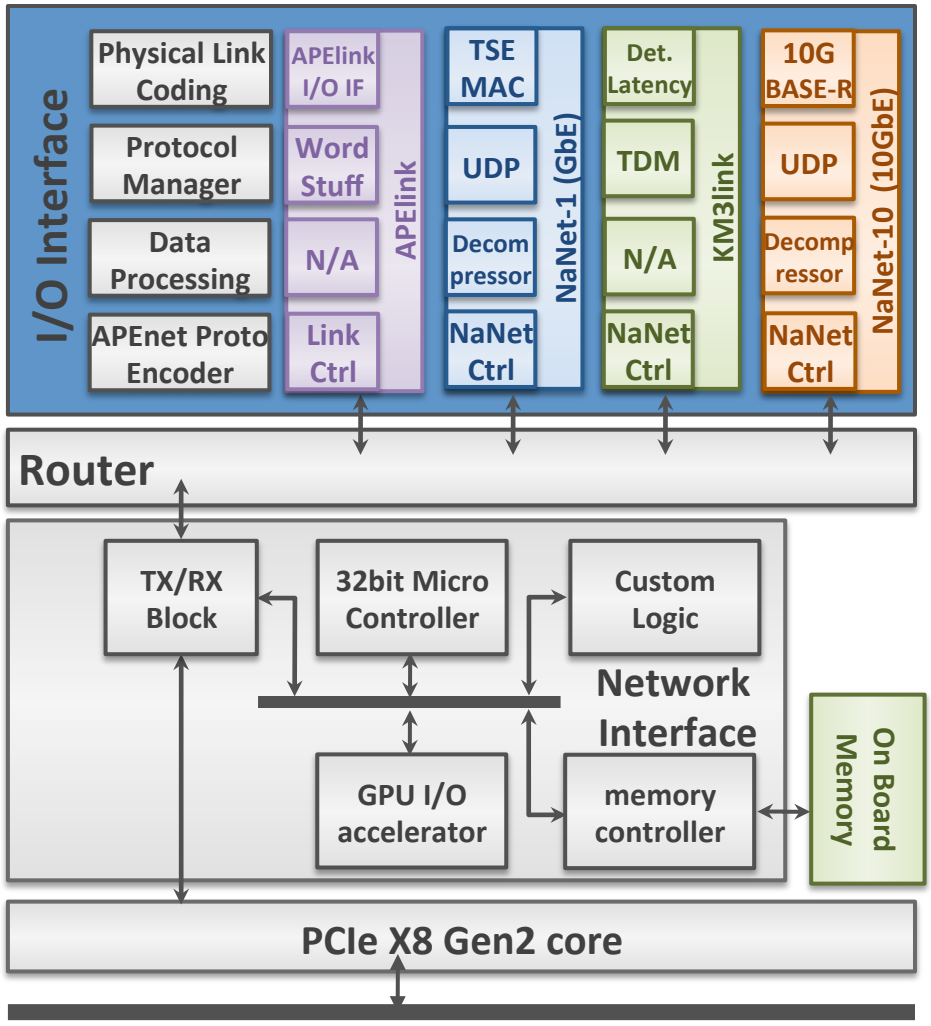
- Mar 2012: first contact between NA62 people and apeNET+ team at GPU Technology Conference (GTC).
- Jul 2012: start of NaNet-1 design as spin-off of apeNET+, financed from the EURETILE FP7 project.
- Mar 2013: first results obtained on NaNet-1 presented at GTC 2013.
- Feb 2014: NaNet^3 hardware specification(KM3Net-IT).
- Jan 2015: start of the NaNet 3 years (2015-2017) CSN5 project.

NaNet Modular Design



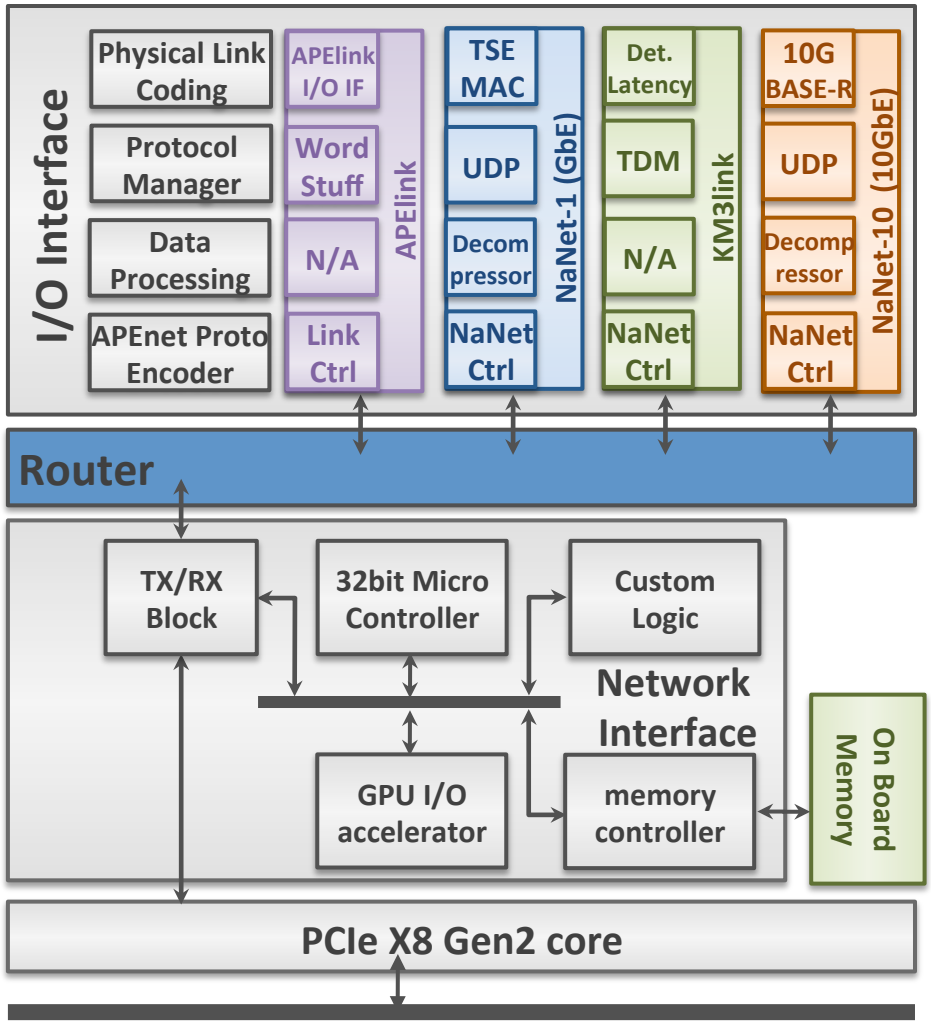
- I/O Interface
  - Multiple physical link technologies.
  - Network protocols offloading.
  - Application-specific processing on data stream.
- Router
  - Dynamically interconnects I/O and NI ports.
- Network Interface
  - Manages packets TX/RX from and to CPU/GPU memory.
  - Zero-Copy RDMA.
  - GPU I/O accelerator.
  - TLB for Virtual to Physical mem map.
  - Microcontroller.
- PCIe X8 Gen2/3 Core

NaNet Design – I/O Interface



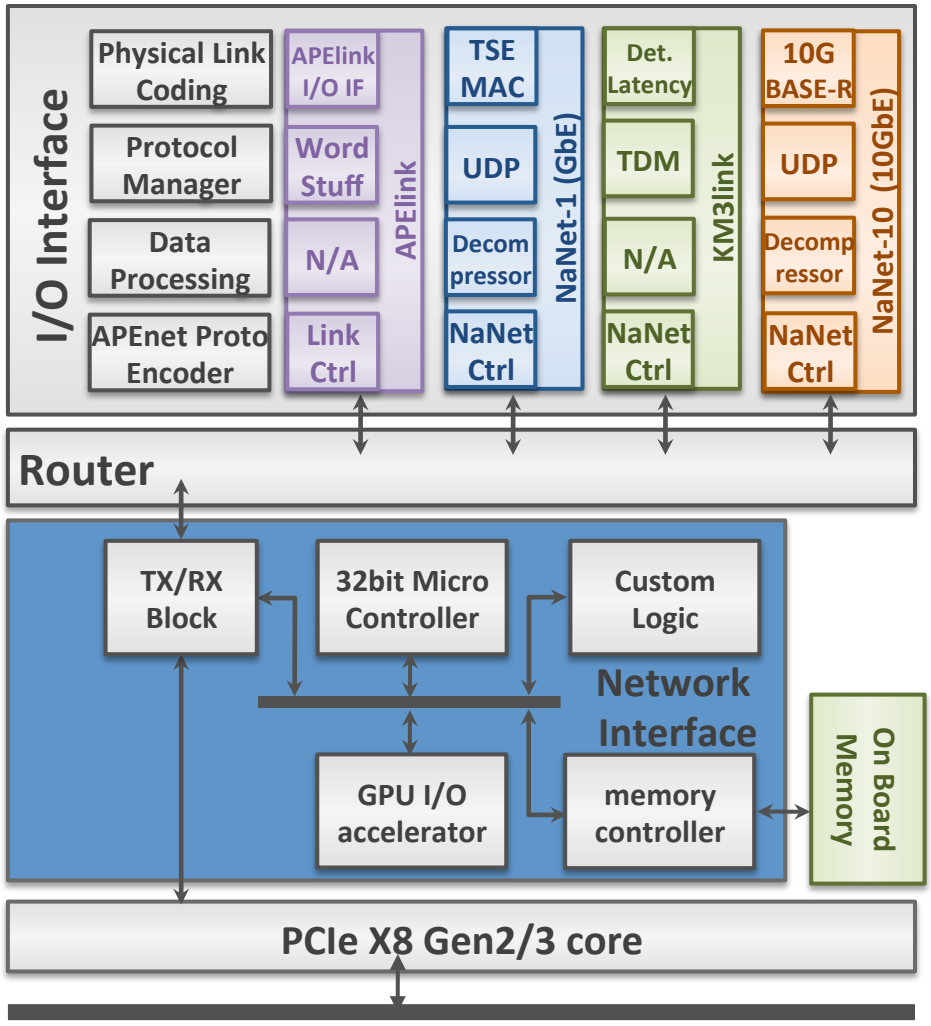
- Physical Link Coding
  - Standard: **1GbE** (1000Base-T), **10GbE** (10Base-R).
  - Custom: **APElink** (20 Gb/s QSFP), **KM3link Det. Lat.** (2.5 Gb/s optical).
- Protocol Manager
  - Data/Network/Transport layers off-load (**UDP**, **Time Division Multiplexing**)
  - Minimize latency fluctuations.
- Data Processing
  - Application-specific processing on data stream.
  - **e.g. NA62 Decompressor&Merger**: re-format event data (data size and alignment), coalesce event fragments before forwarding packets to NI.
- APEnet Protocol Encoder
  - Protocol adaptation between on-board and off-board protocol.

NaNet Design – Router



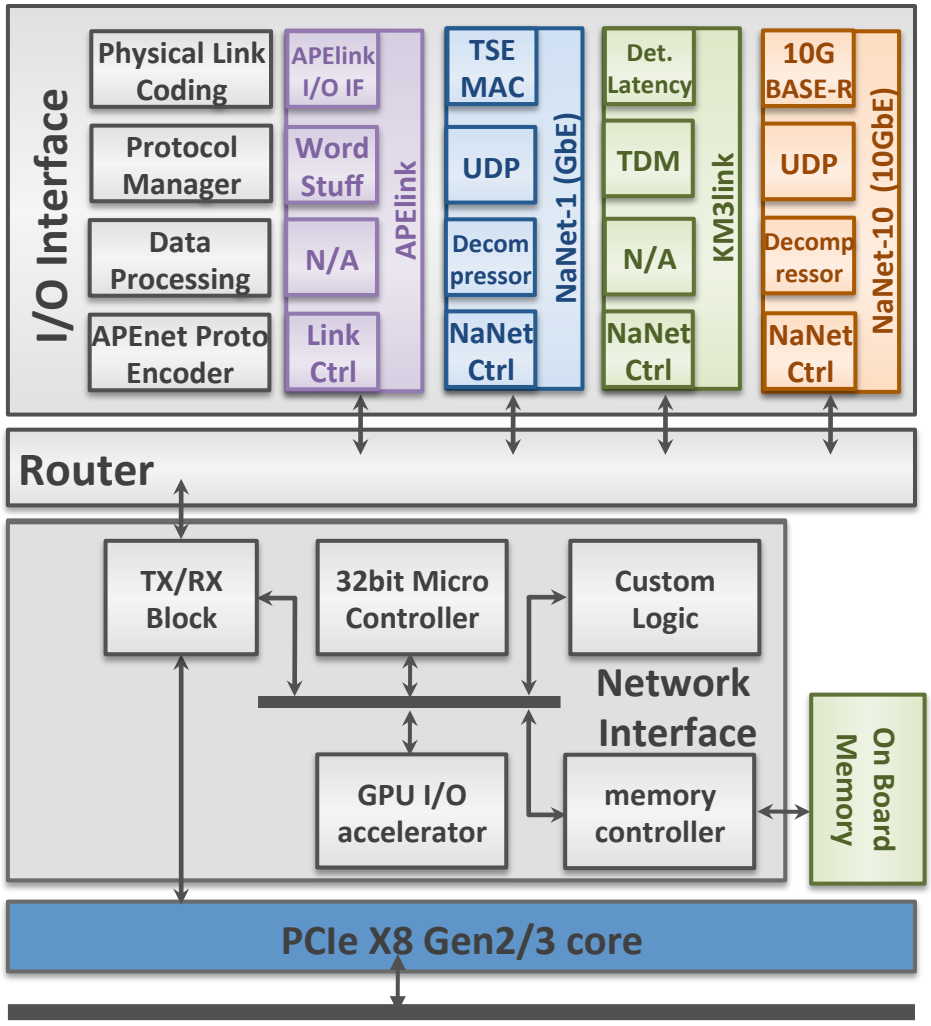
- 5 ports bidir full crossbar switch.
- Router: dynamically interconnects ports.
- Arbitration: resolve contention on ports requests (static, round robin).
- Supports up to 10 simultaneous 2.8 GB/s data streams. (x2 in next PCIe Gen 3 designs)
- Re-configurable number of ports.

NaNet Design – Network Interface



- In TX gathers data from PCIe interface and forwards them to destination port.
- In RX performs **zero-copy RDMA** receive operation managing CPU/ GPU Virtual to Physical address translation.
  - Translation Lookaside Buffer (associative cache).
  - Microcontroller in case of miss.
- GPU I/O accelerator (**GPUDIRECT P2P/RDMA**)
- TX/RX Block
  - Multiple DMA engines instantiating concurrent PCIe transactions.
- Altera NIOS II microcontroller.

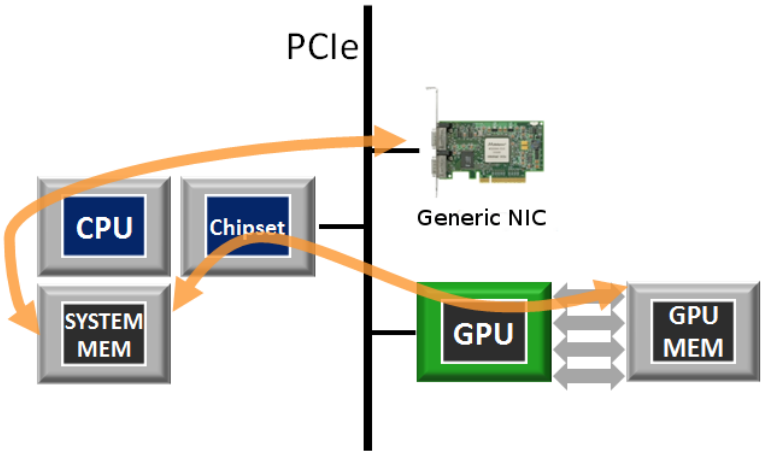
NaNet Design – PCIe Cores



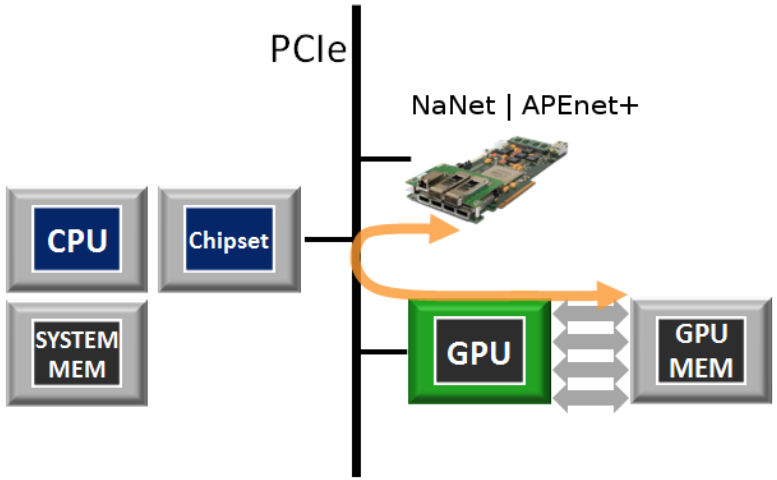
- PCIe X8 Gen2 Core (stable)
  - Based on PLDA EZDMA Gen2 hard IP core.
  - CPU BW: 2.8 GB/s Write, 2.5 GB/s Read.
  - GPU BW: 2.8 GB/s Write, 2.5 GB/s Read.
- PCIe X8 Gen3 Core (testing)
  - Based on PLDA XpressRICH Gen3 soft IP core.
  - CPU BW: 4.5GB/s Write, 4.0 GB/s Read.
  - GPU BW: 2.8 GB/s Write, 3.0 GB/s Read.
- Home-made PCIe X8 Gen3 Core (developing)
  - Based on the Altera PCIe hard-IP

(Ivy Bridge, Kepler K20 test setup)

NaNet Design – GPUDirect P2P/RDMA

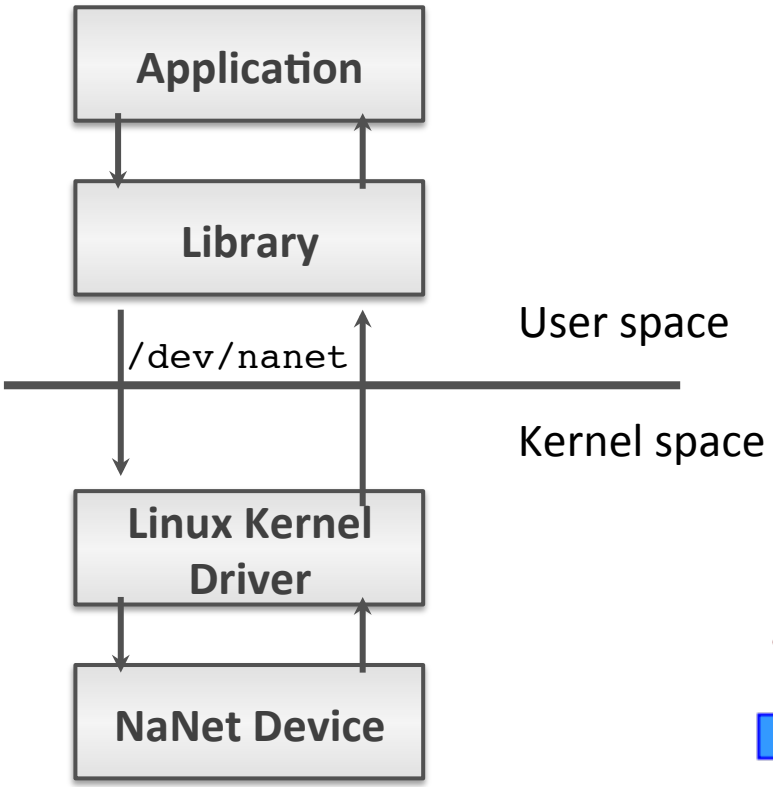


- Non-GPUDirect capable NIC data flow.
- Intermediate buffering on CPU memory for I/O operations.

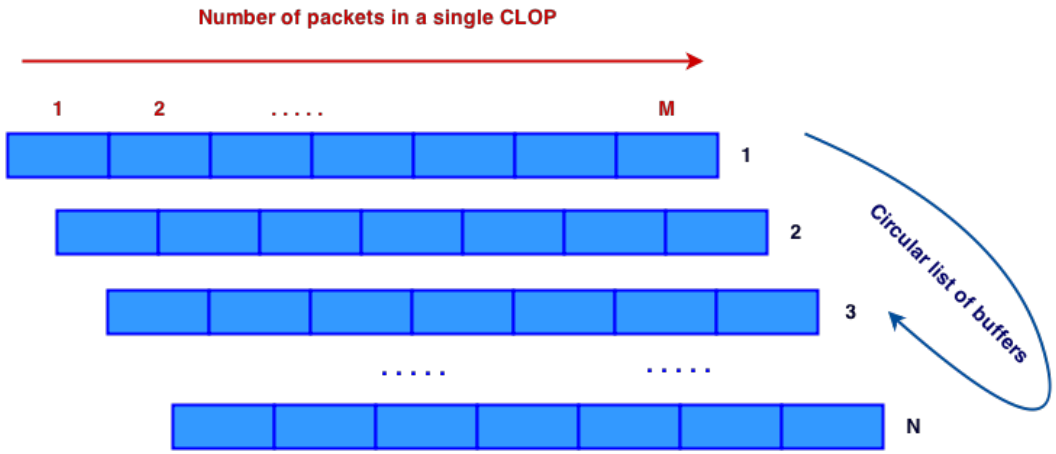


- GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement.
- No bounce buffers on host memory.
- Zero copy I/O.
- Latency reduction for small messages.
- nVIDIA Fermi/Kepler/Maxwell

NaNet Software Stack

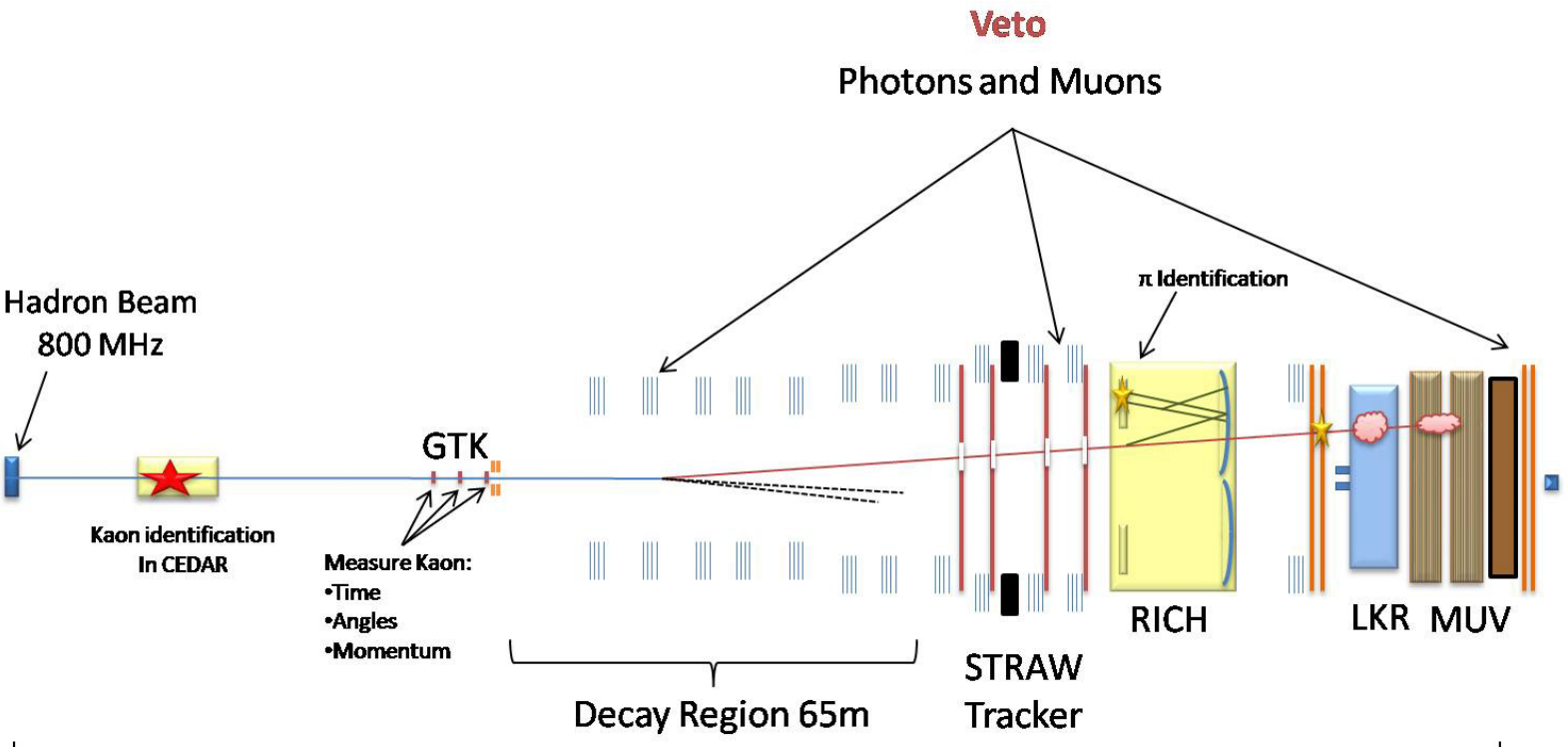


- Host
  - Linux Kernel Driver
  - User space Library (open/close, buf reg, wait recv evts, …)
- Nios II Microcontroller
  - Single process program performing System Configuration & Initialization tasks.

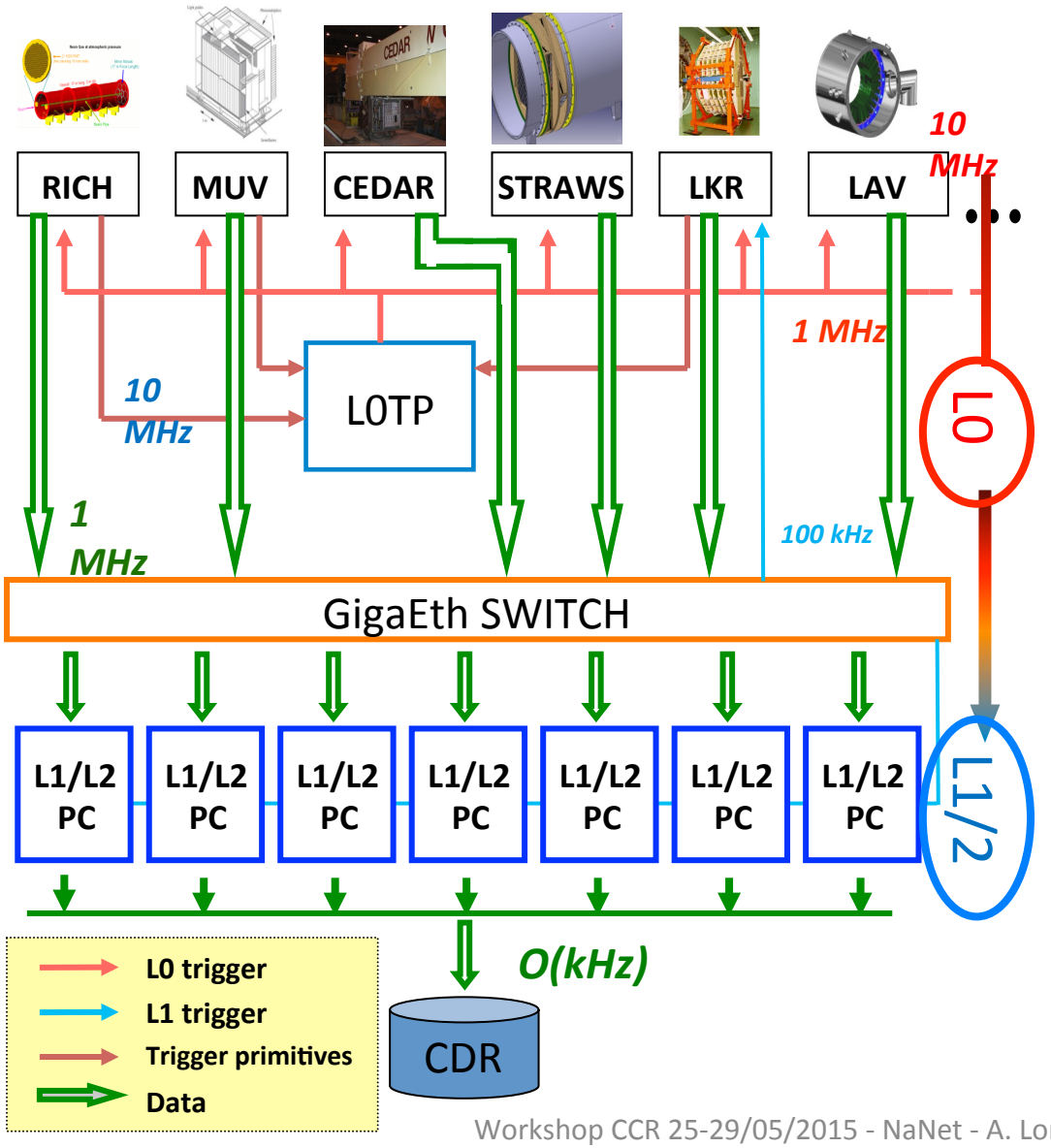


The NA62 Experiment at CERN

- $K^+ \to \pi^+ \nu\bar{\nu}$ decay (BR~8×10^{-11})
- Huge background from other kaon decays

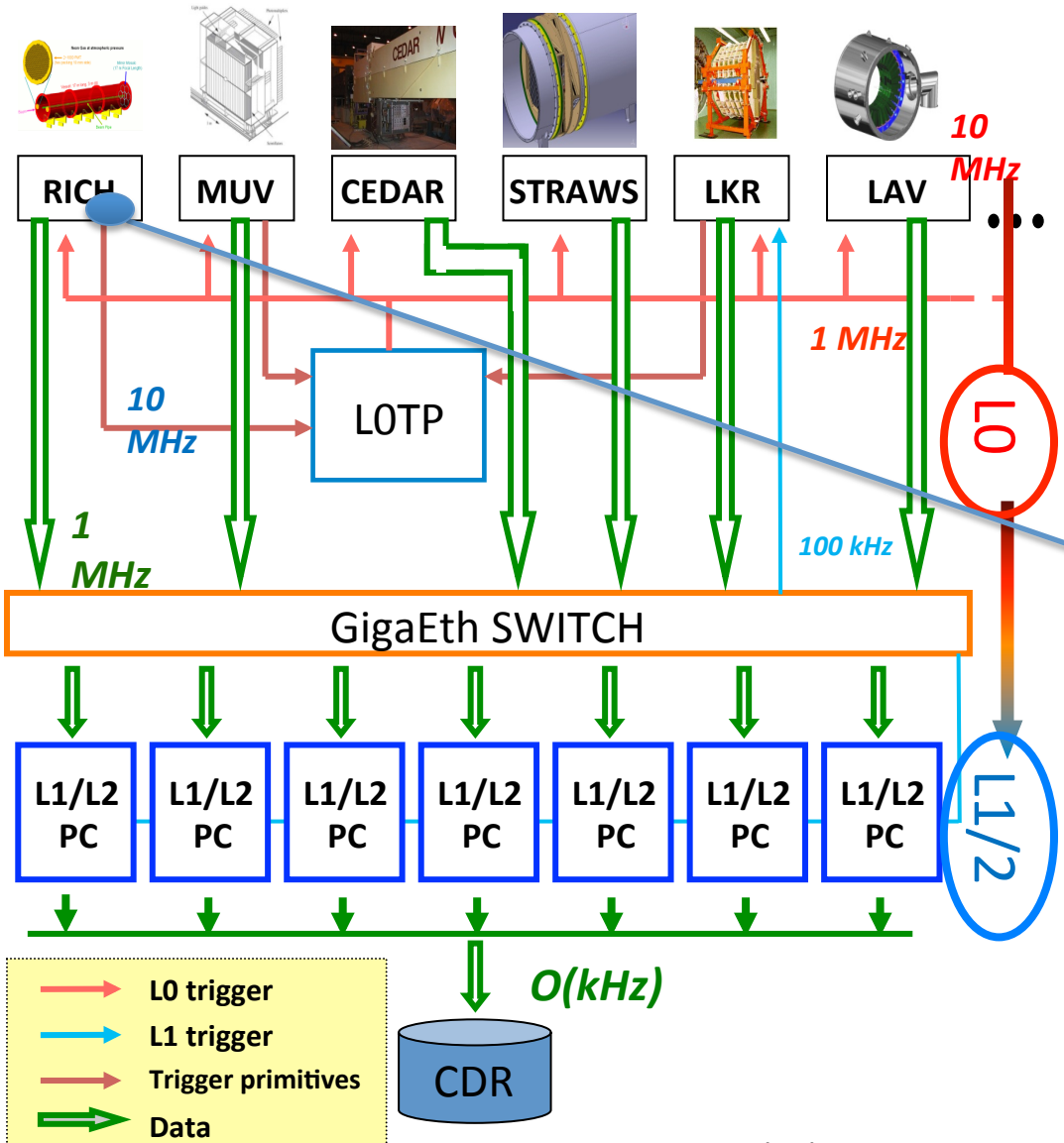


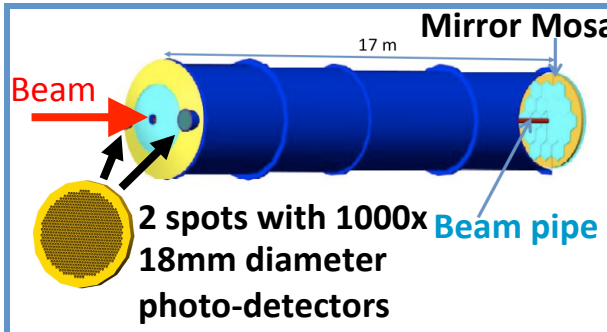
NA62 DAQ and Trigger



- **L0**: Hardware synchronous level
  - 10 MHz to 1 MHz, **1 ms max. latency**
  - Primitives (MUV, RICH, LAV, LKR)
- **L1**: Software level
  - "Single detector", 1 MHz to 100 kHz
- **L2**: Software level
  - "Complete information", 100 kHz to 10 kHz

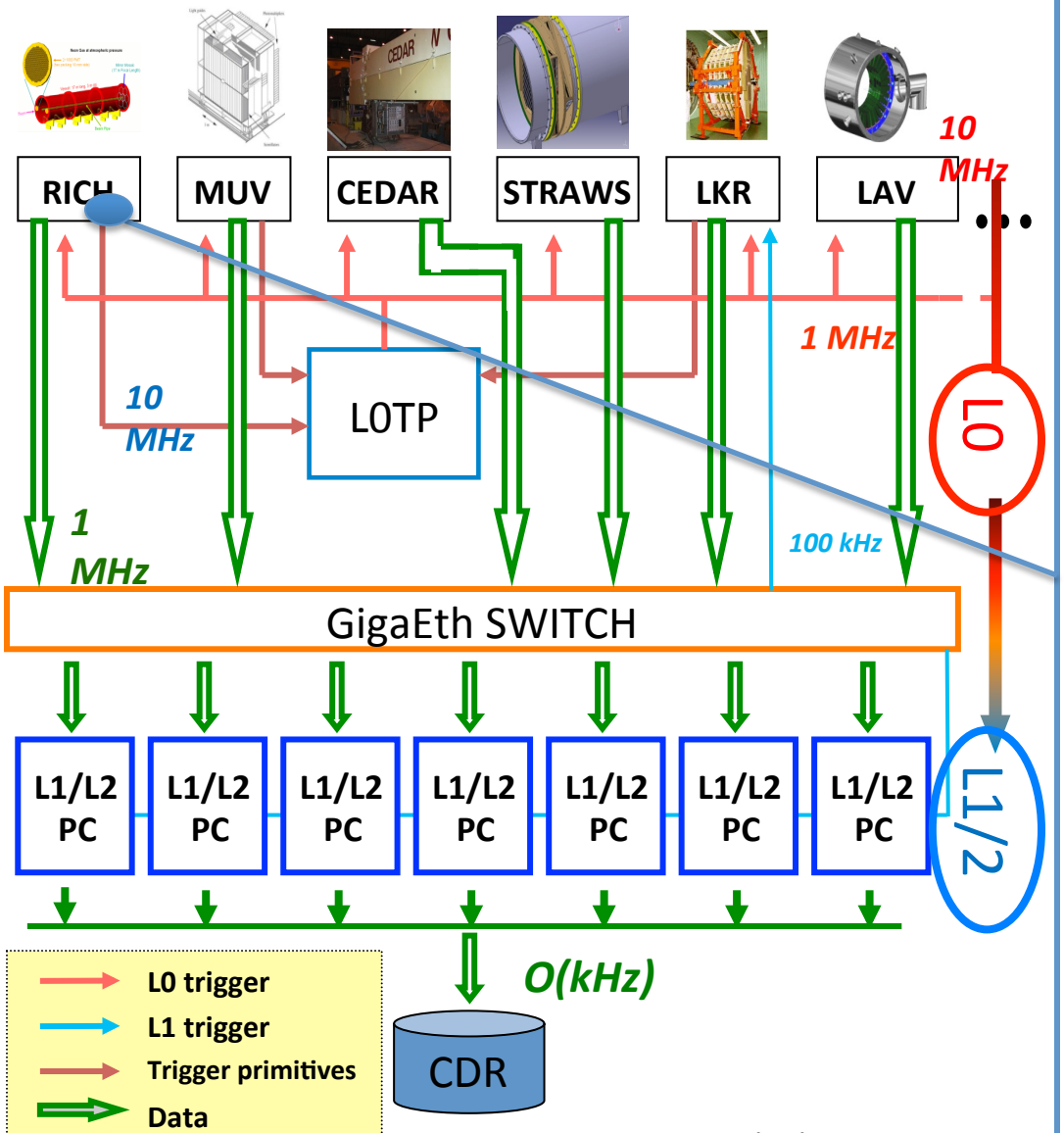
NA62 RICH detector





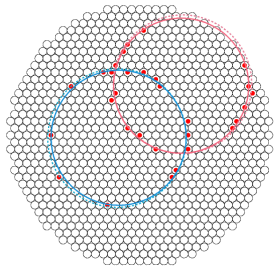
- pion-muon discrimination
- Cherekov light focused by composite mirror on the two spots.
- **70 ps** time resolution.
- **10 MHz** event rate.
- **~20 hits** per ring on average.
- **40 byte** avg. size per event.

Using GPUs in the NA62 L0 Trigger for the RICH



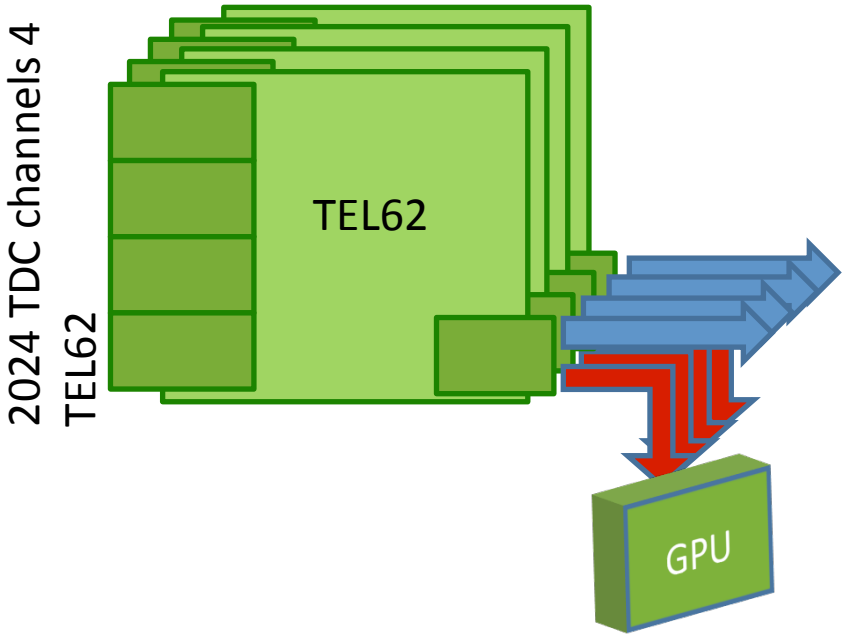
Compare FPGA-based trigger with a **GPU-based** one

- More selective trigger algorithms.
- Programmable.
- Upgradable.
- Rough detection of particle speed (radius) and direction (centre)

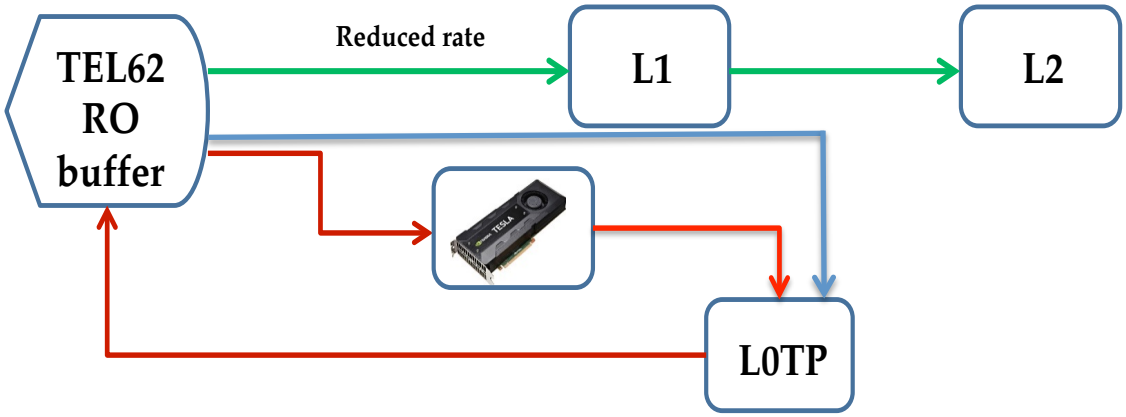


- **Efficient match of circular hit patterns** on GPUs.

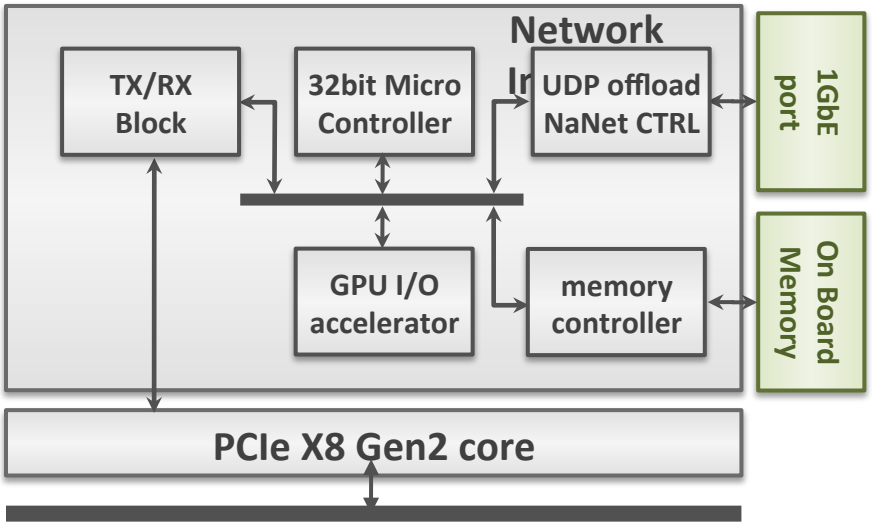
NA62 RICH L0 Trigger Processor

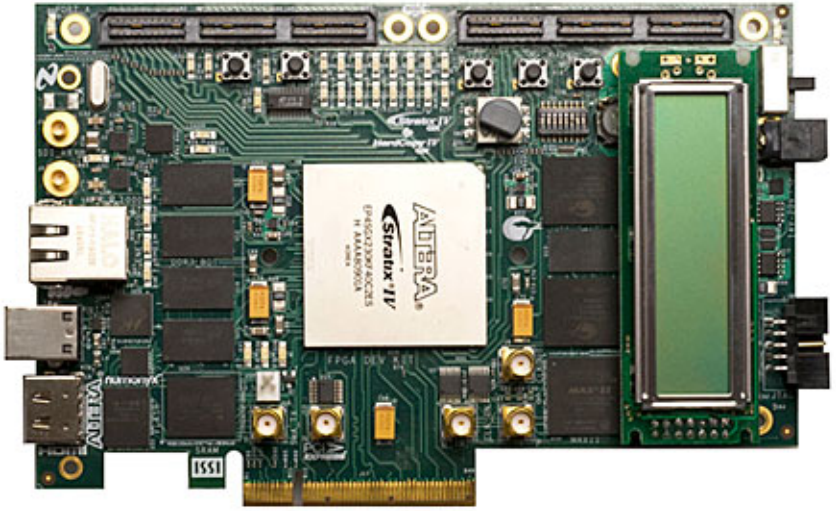


- 4 TEL62 for RICH detector
- 8×1Gb/s links for data r/o
- 4×1Gb/s trigger primitives
- 4×1Gb/s GPU trigger
- Events rate: 10 MHz
- L0 trigger rate: 1 MHz
- Max Latency: 1 ms

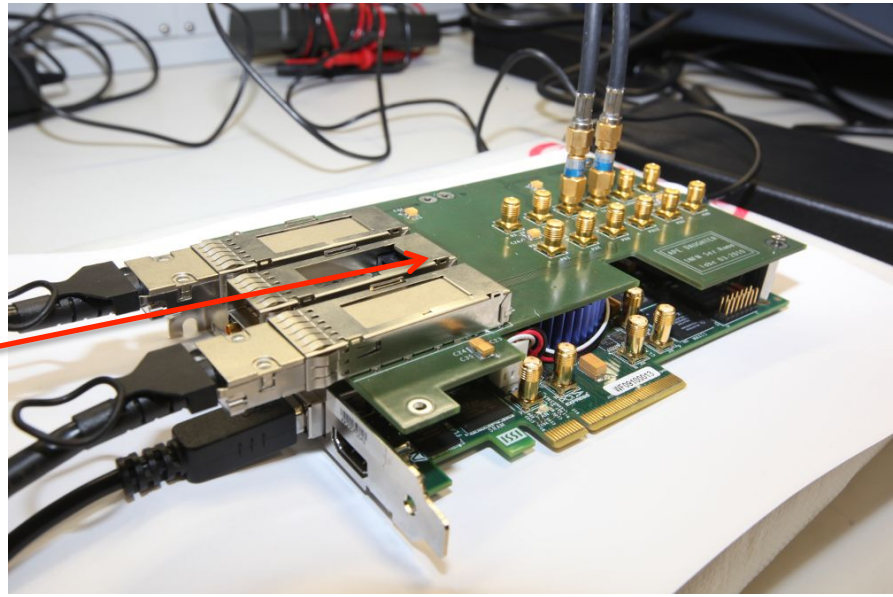


NaNet-1: a 1 GbE NIC for NA62 GPU-based RICH L0 trigger prototyping

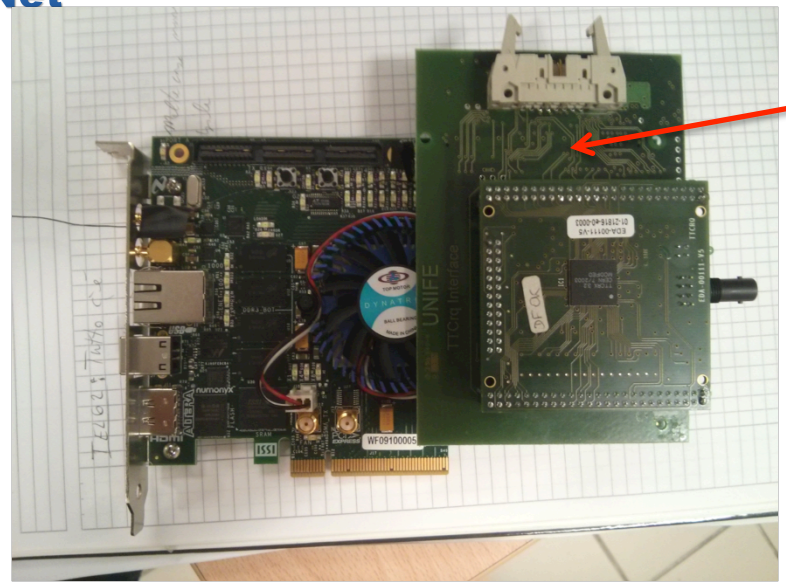




- Implemented on Altera Stratix IV dev board (EP4SGX230KF40C2)
- GbE PHY Marvell 88E1111
- Supports additional 3 APElink channels (20 Gb/s each) with HSMC daughtercard

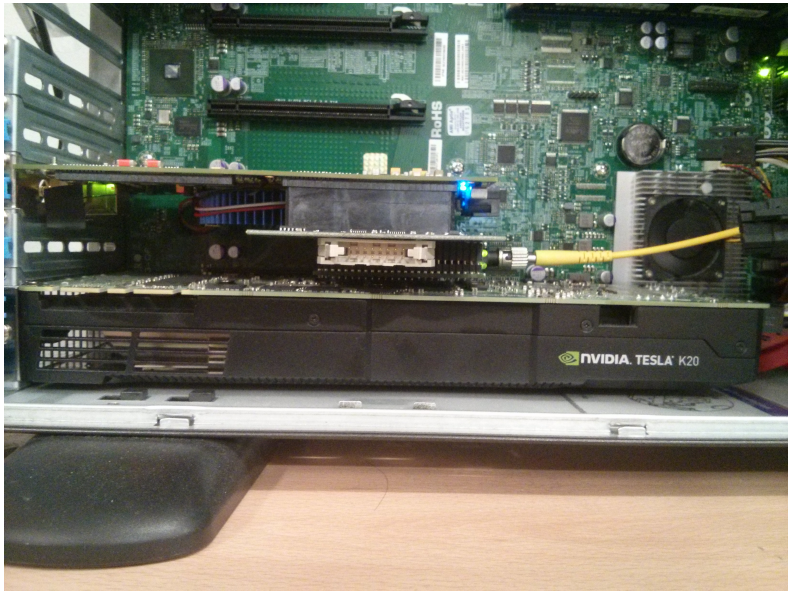


NaNet-1 Integration in NA62 DAQ & Trigger System

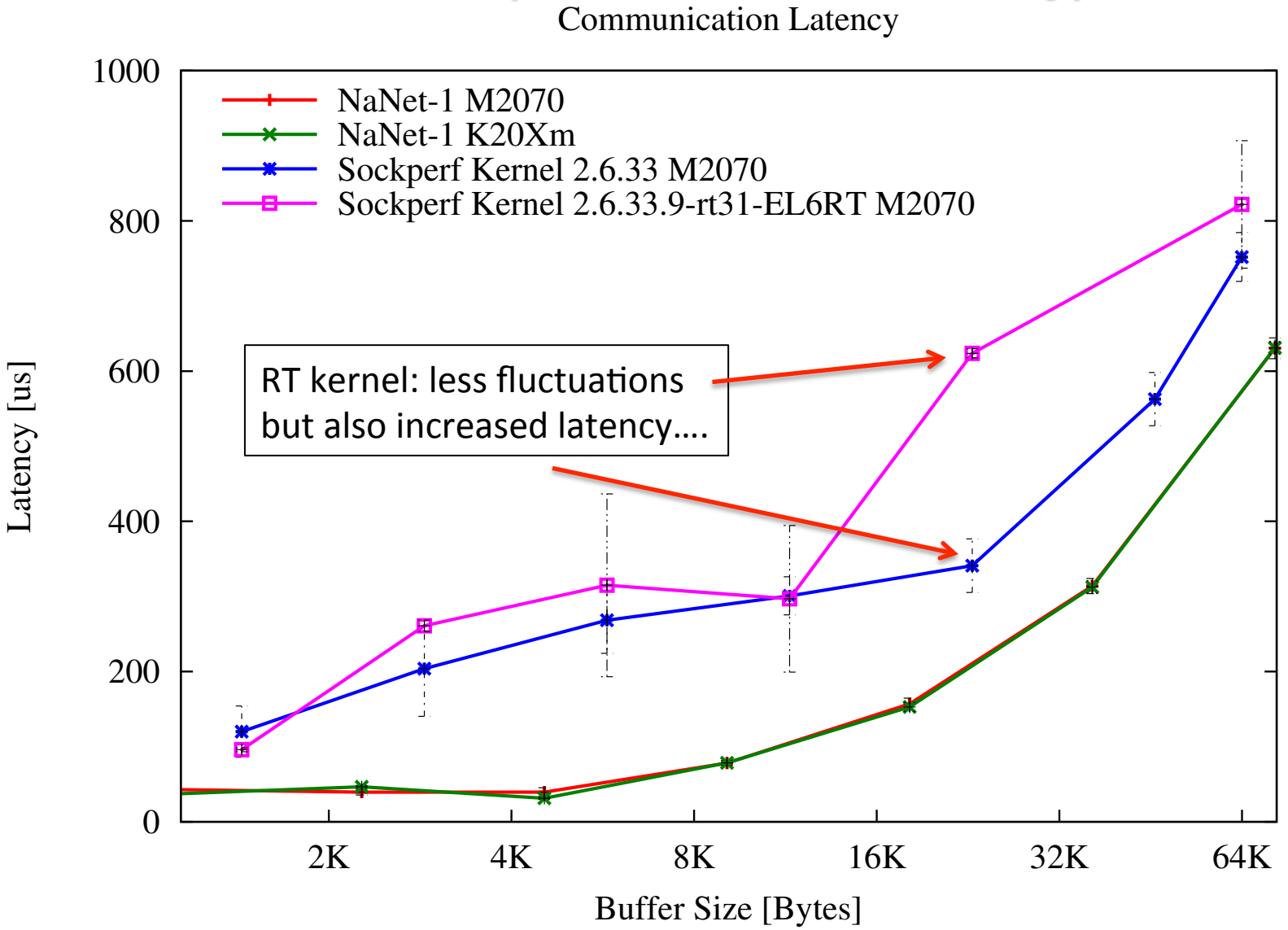


- TTC daughtercard with HSMC connector (INFN Ferrara)
- NaNet receives TTC stream with timing (40 MHz clock, SOB, EOB) and trigger signals from the experiment.
- Allows synchronous operation (accurate latency measurements)

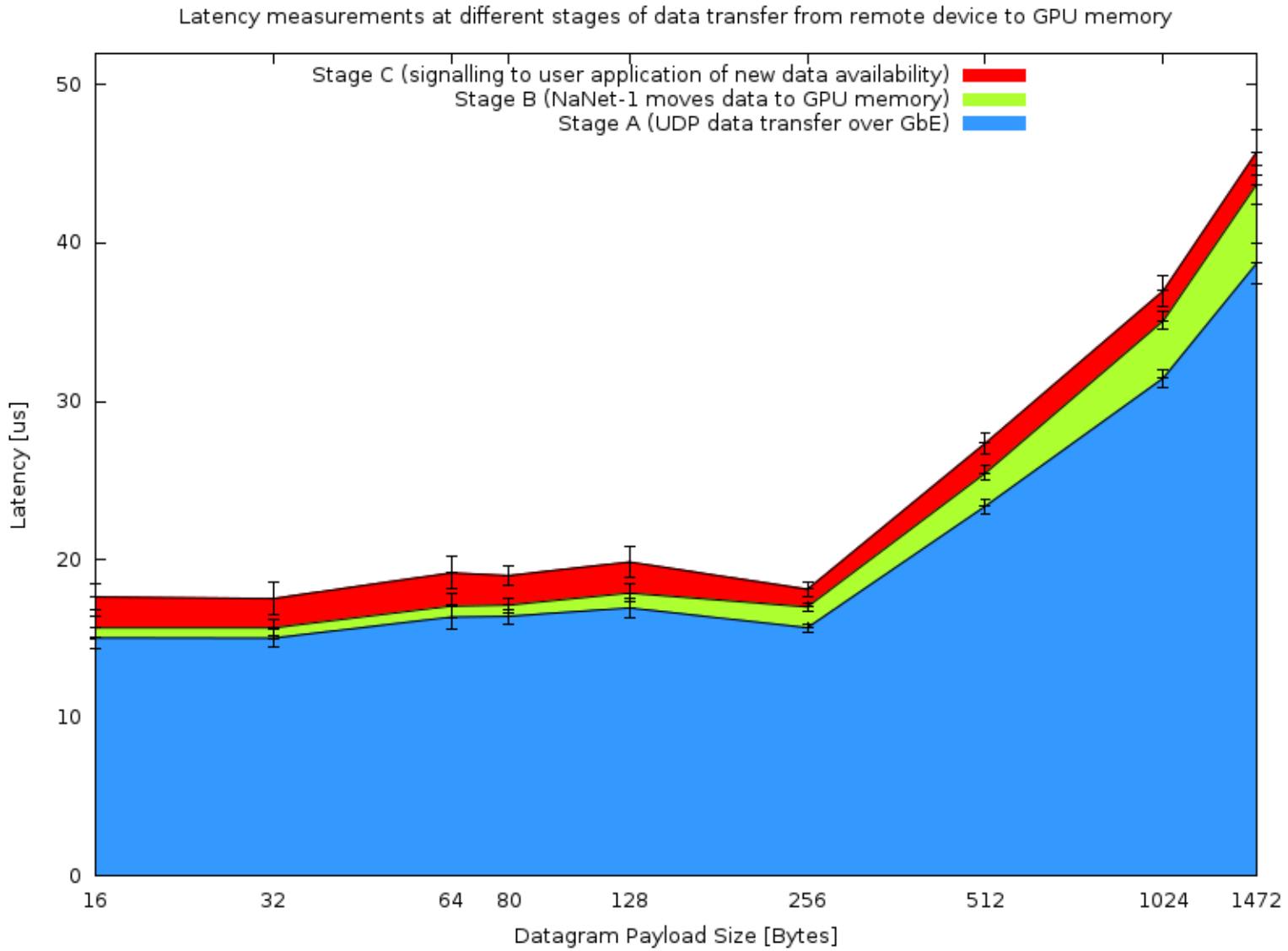
- Integration and tests activities during August 2014 technical run.
- Parasitic operation of GPU-based L0 trigger using NaNet-1 scheduled for June 2015 (events rate limited).



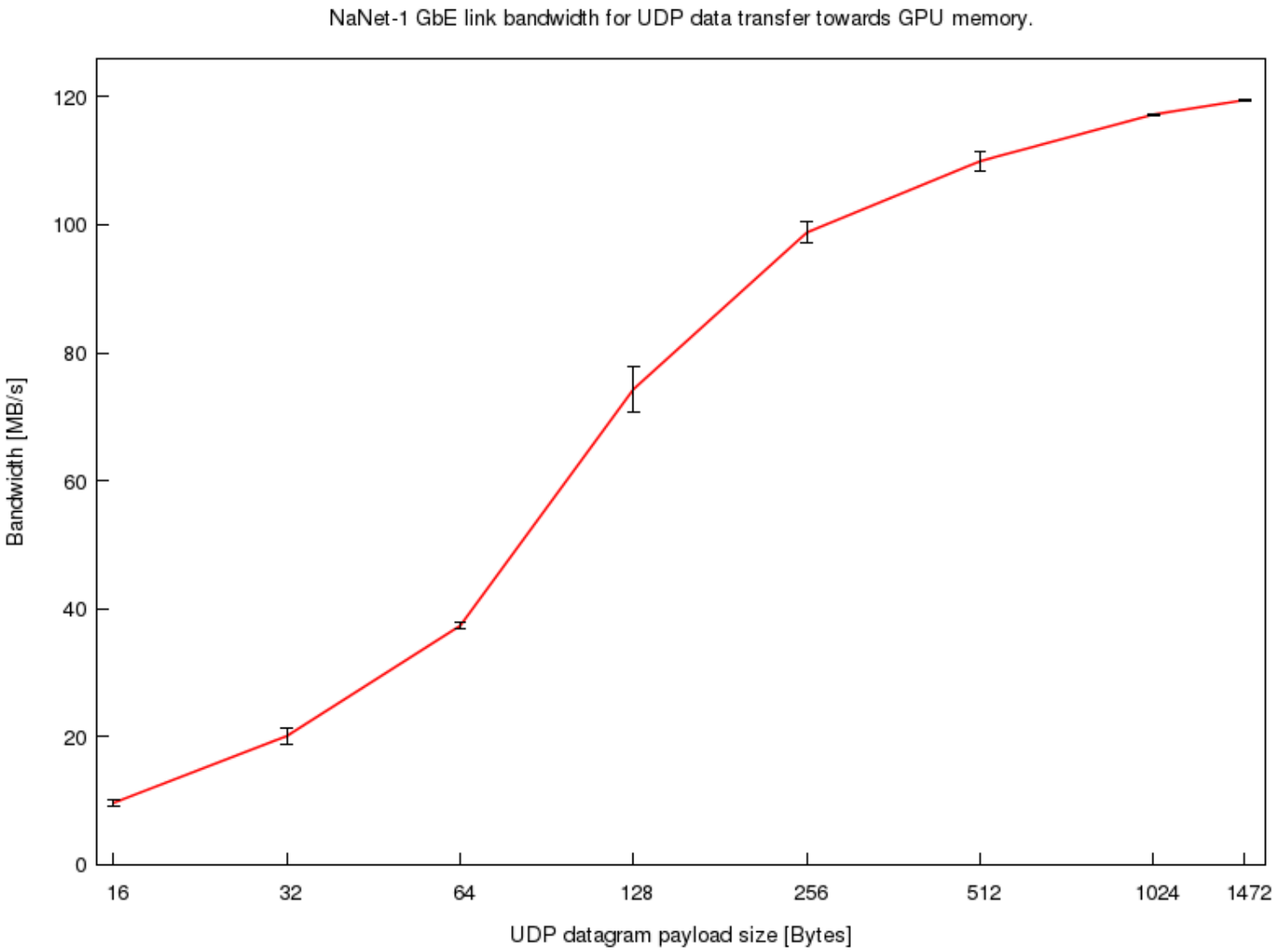
NaNet-1 Communication Latency (buffer in GPU memory)



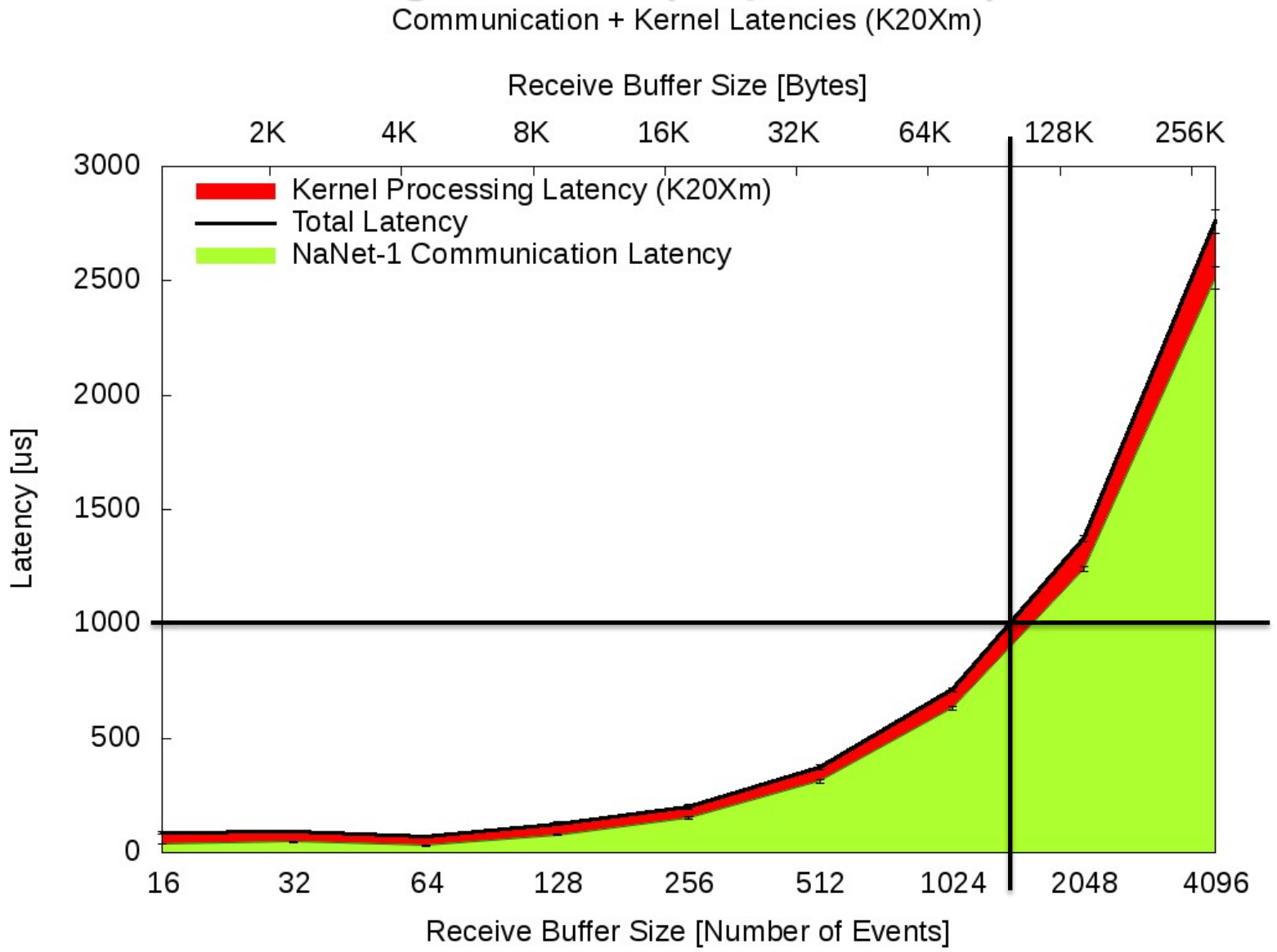
NaNet-1 Latency at different stages of data transfer from GbE channel to GPU memory



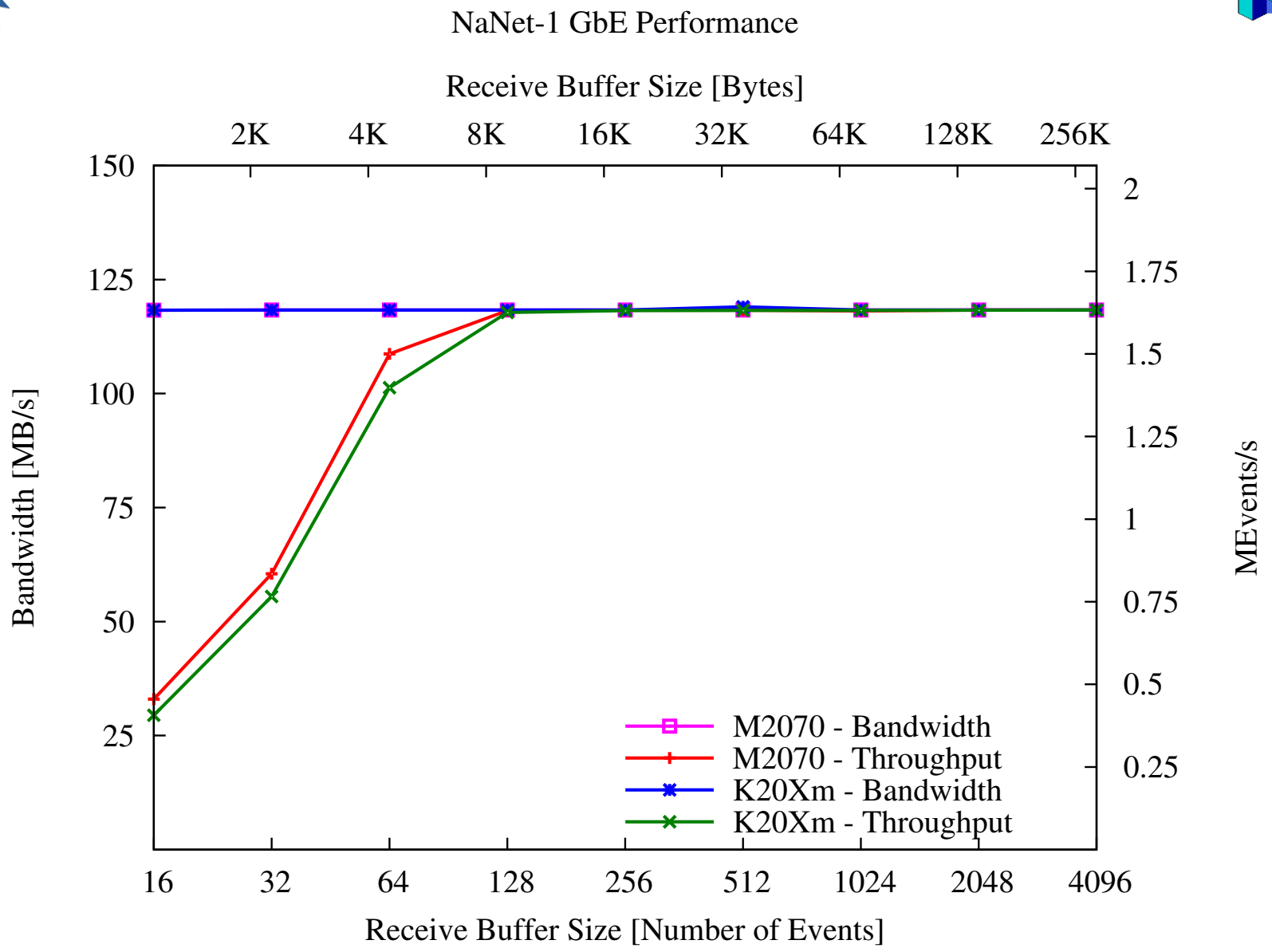
NaNet-1 GbE link bandwidth for UDP data transfer to GPU memory



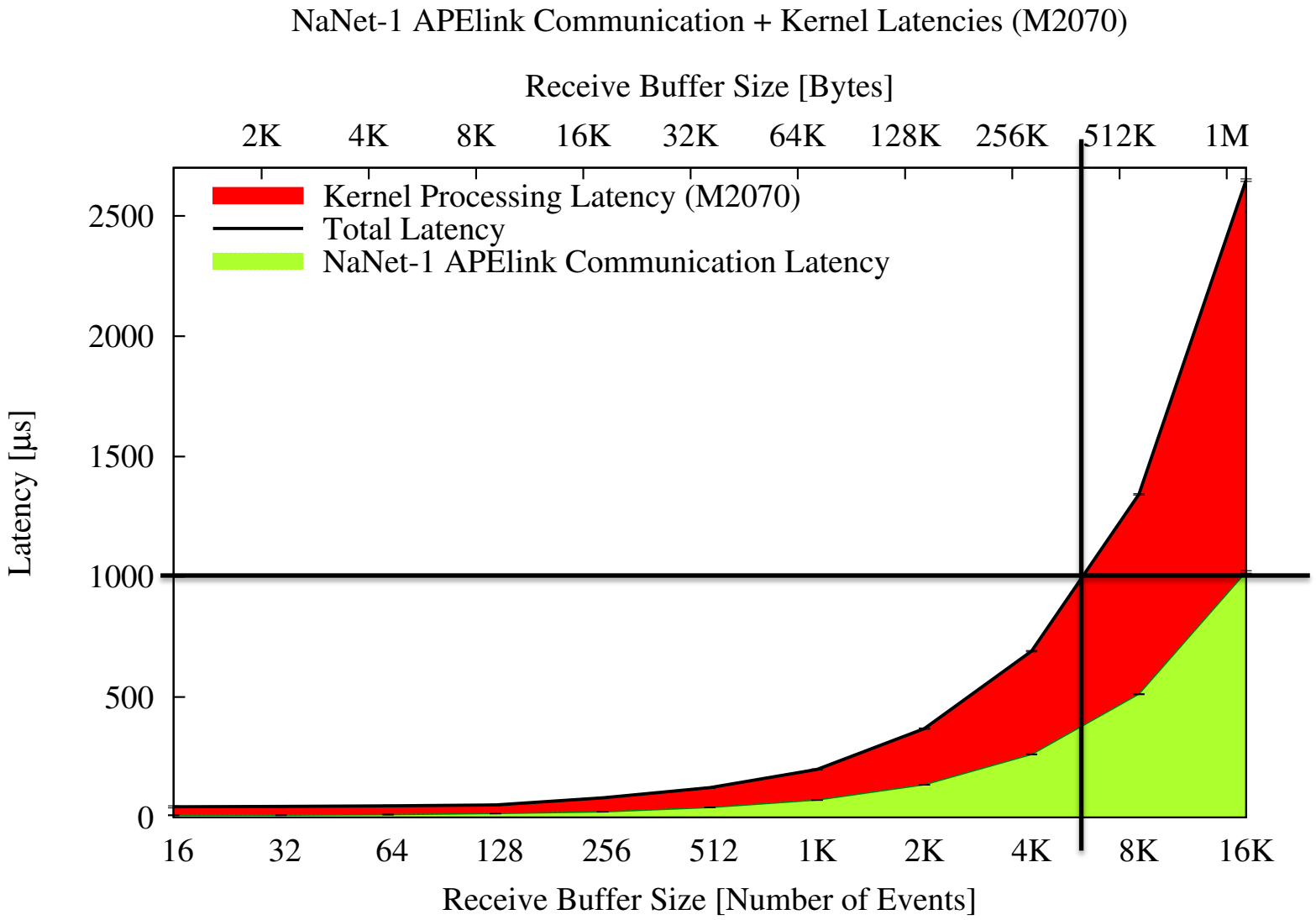
Total latency of the GPU-Based RICH L0-TP using NaNet-1 (Kepler K20X)



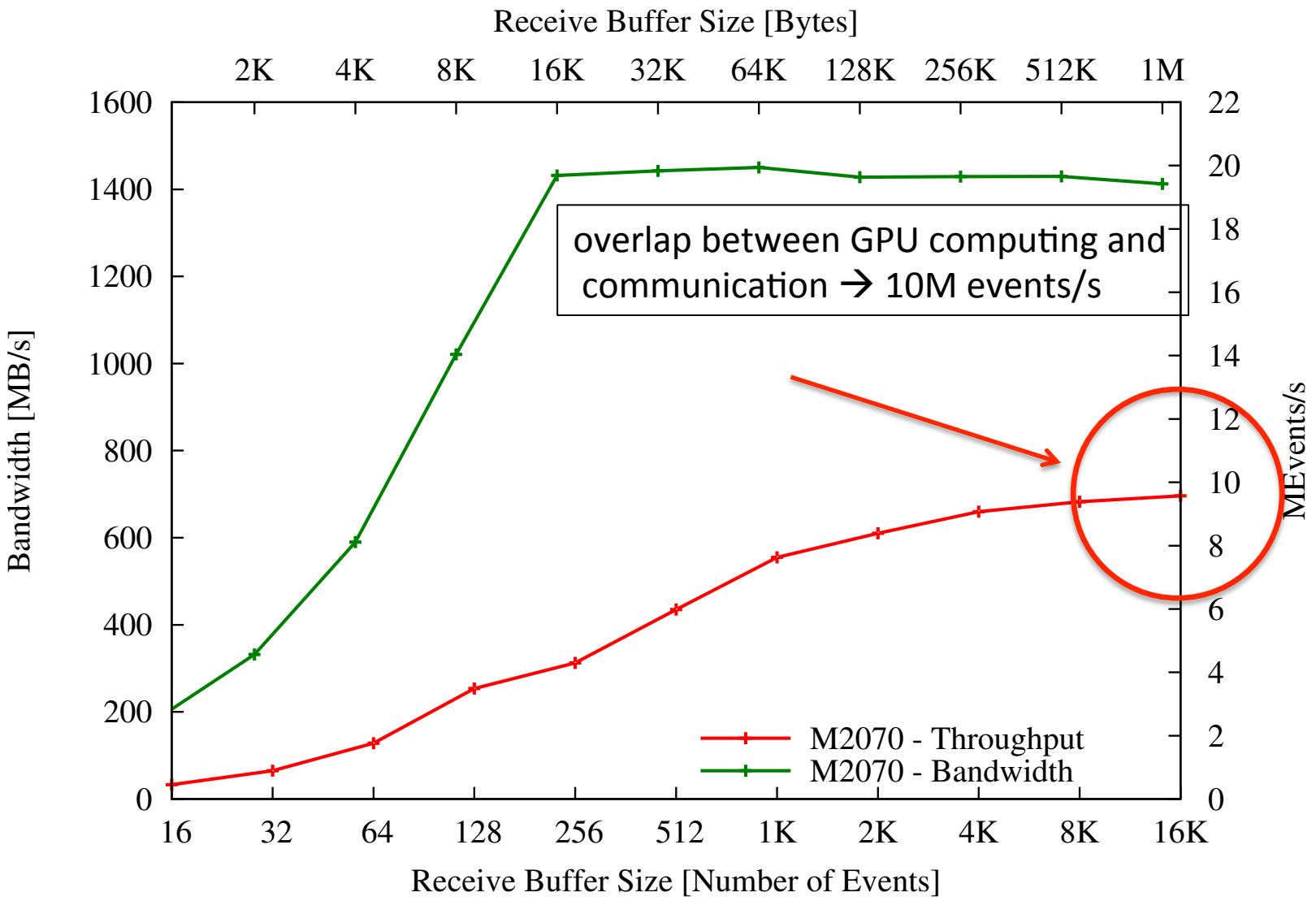
NaNet-1 Bandwidth & L0-TP Throughput



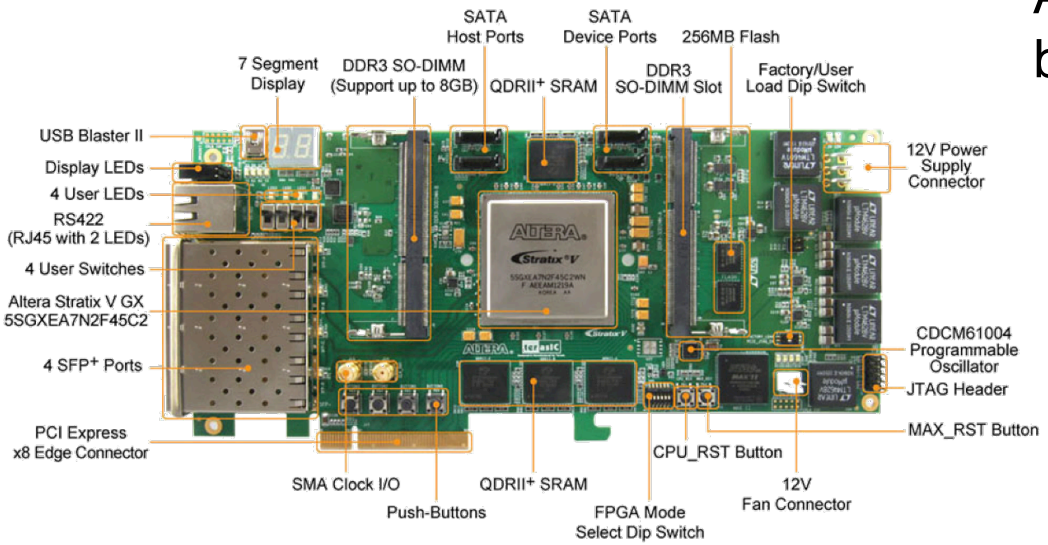
Total latency of the GPU-Based RICH L0-TP using NaNet-1 (APElink)



NaNet-10 Latency APElink Bandwidth & L0-TP Throughput



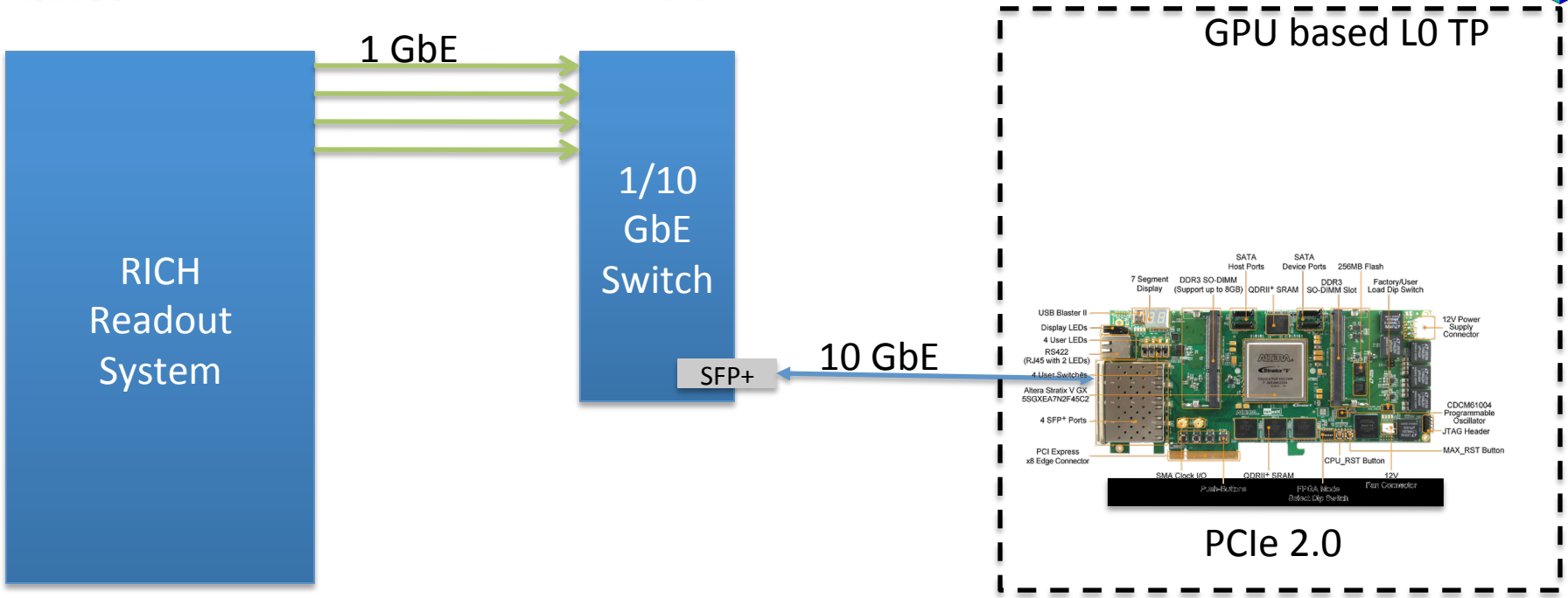
NaNet-10: four 10GbE SFP+ ports



Terasic DE5-NET

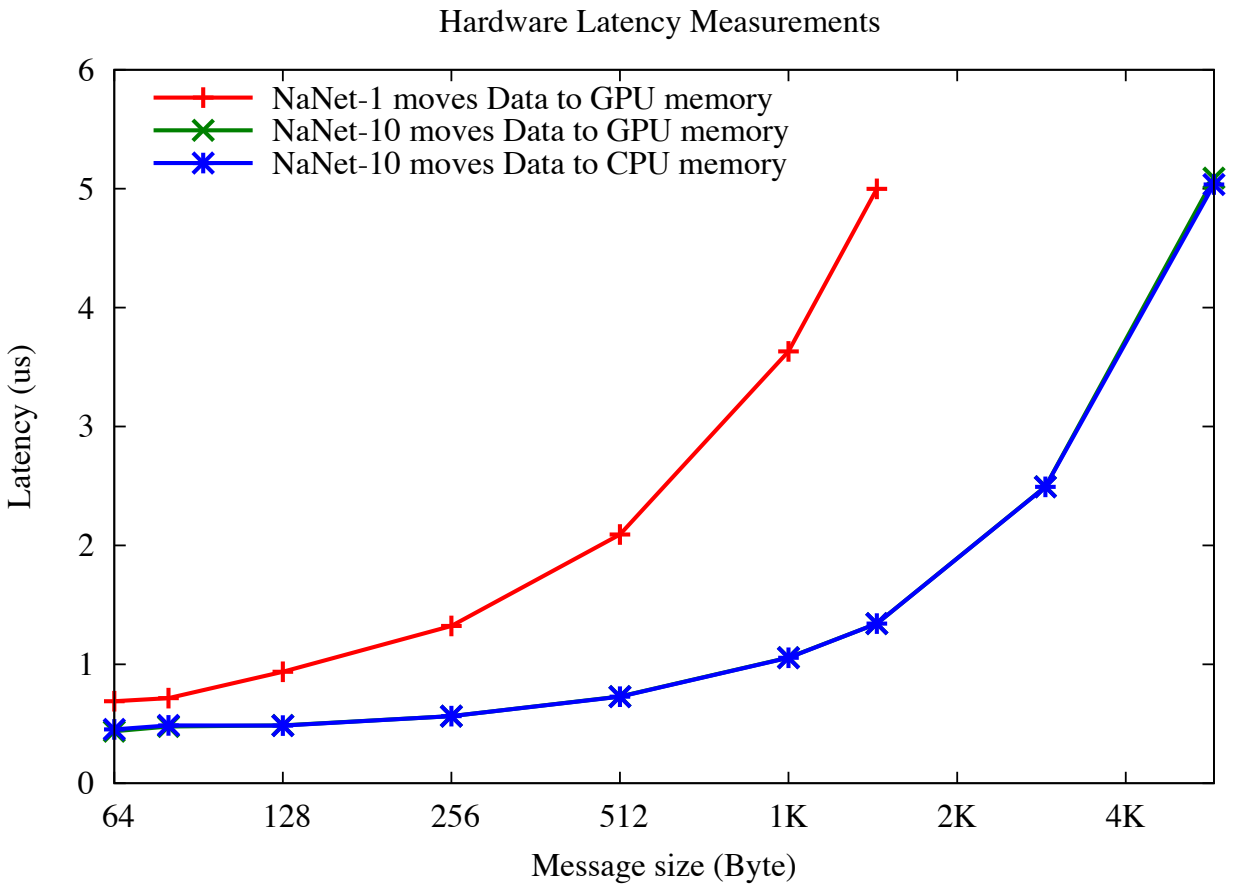
- Implemented on Terasic DE5-NET ALTERA Stratix V GX development board
  - Four 10 GbE SFP+ ports
  - Faster embedded Altera transceivers (up to 14.1 Gbps)
  - hardened 10GBASE-R PCS
  - UDP protocol offload
  - Zero-copy RDMA
  - GPUDirect P2P/RDMA
  - PCIe Gen3 x8 (8 GB/s)
  - 4Q2105

NaNet-10: path towards the final NA62 RICH L0 GPU Trigger Processor

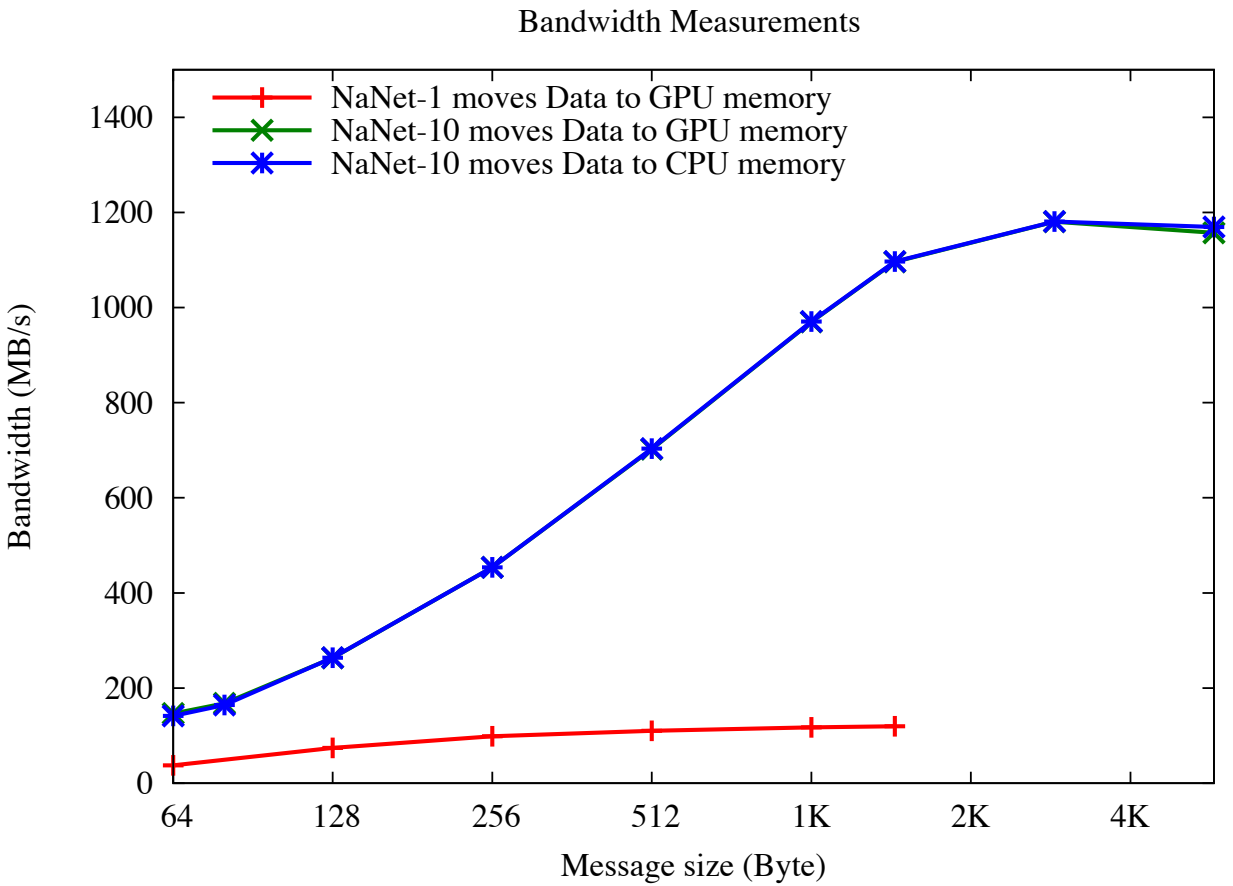


- NaNet-10 is required to achieve L0-TP required performances (10M events/s).
- Aggregation of 4 GbE channels payload in a single 10GbE link through an HP2920 Switch
- run in parallel and “parasitic” mode
- **Currently testing a single port, PCIe Gen2 X8 version to be integrated in NA62 setup in August 2015.**

NaNet-10 (single port, PCIe x8 Gen2 version) Hardware Latency

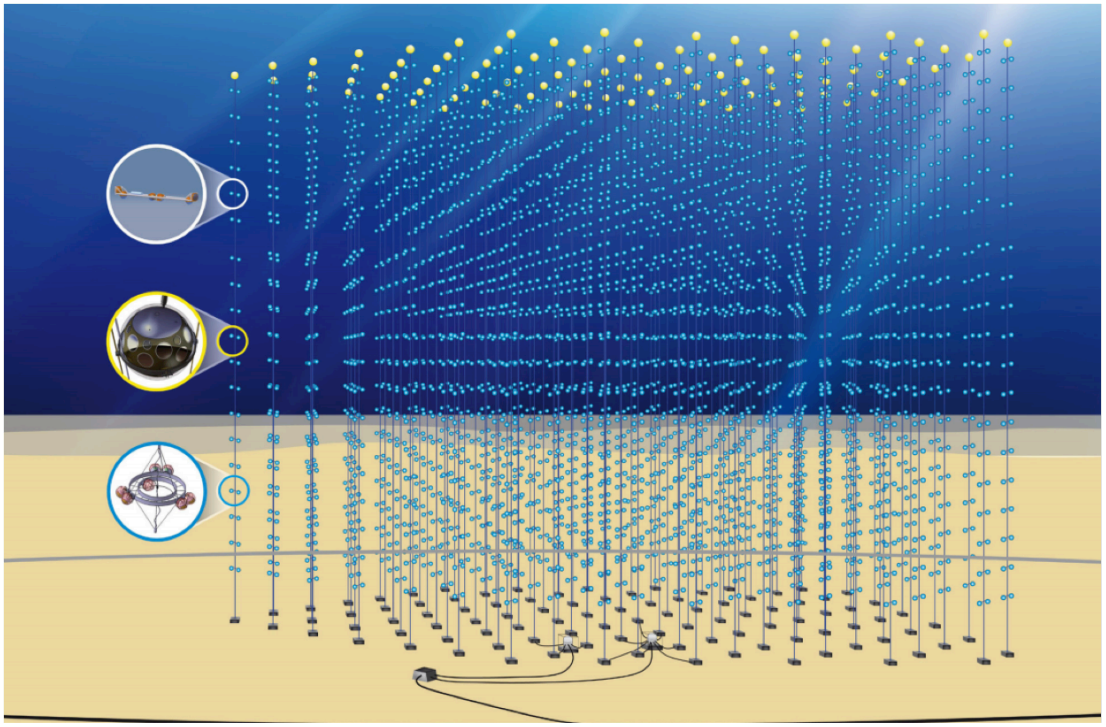


NaNet-10 (single port, PCIe x8 Gen2 version) Bandwidth

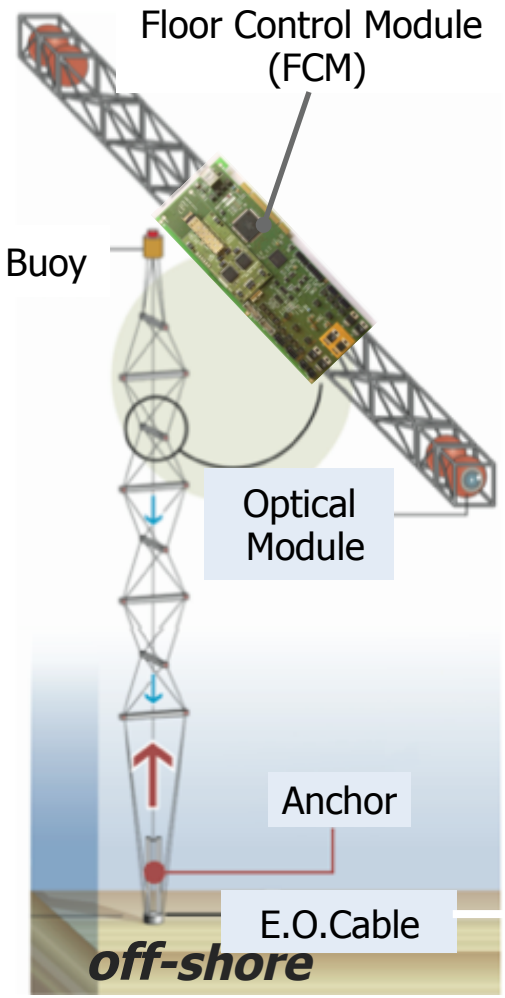


KM3Net-IT Neutrino Telescope

KM3NeT-IT is an underwater experimental apparatus for the detection of high energy neutrinos in the TeV/PeV range based on the Čerenkov technique.
The deployment site is located about 100 km off-shore Porto Palo di Capo Passero, Sicily, and about 3500 m under the sea. The on-shore station is located in Porto Palo di Capo Passero.



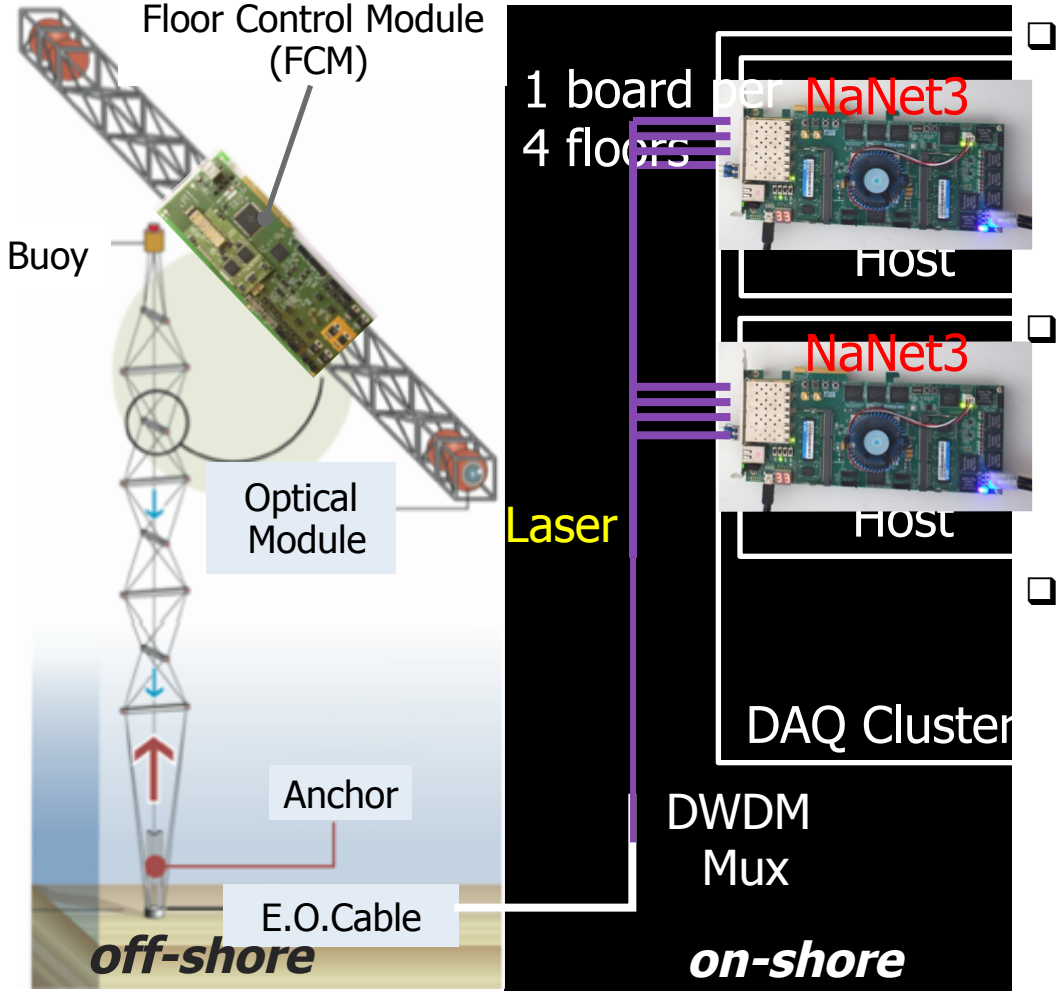
KM3Net-IT DAQ



A detection "tower" is composed of:

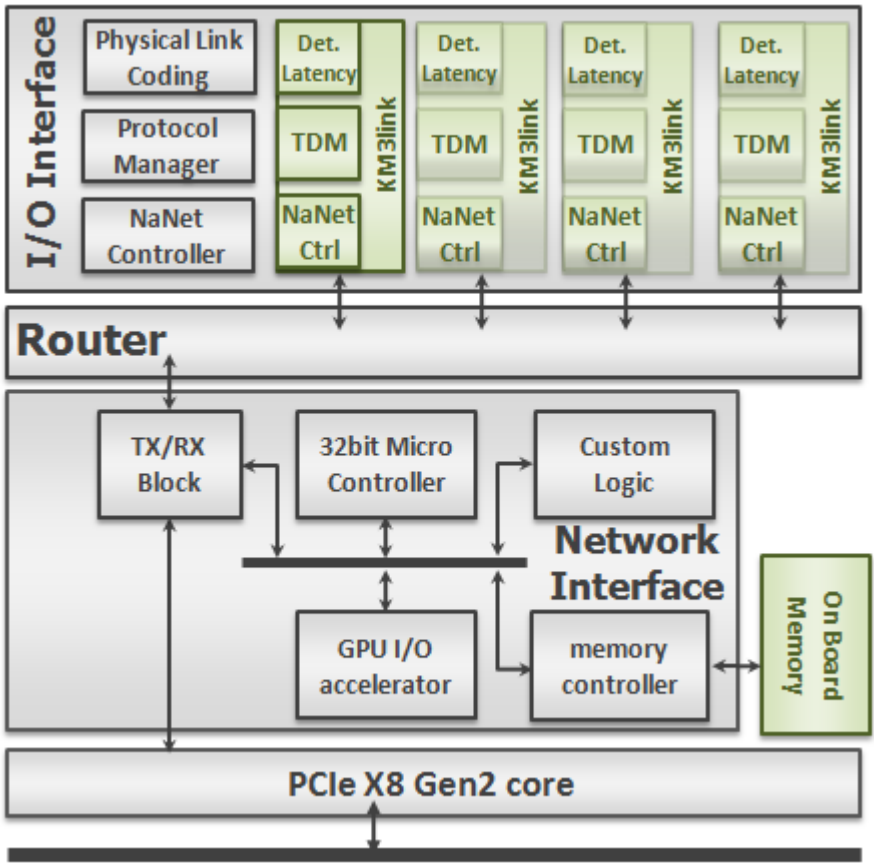
- **14 floors** 20m vertically spaced
- Each floor has:
  - 8 m bars, equipped with 6 Optical Modules (OM)
  - 2 hydrophones
  - oceanographic instrumentation
- FCM manages communication between the on-shore lab and the underwater devices, also distributing the timing information (GPS clock) and slow control signals received from the on-shore equipment.
- **2.5 Gb/s optical link** per floor (800Mb/s payload) with TDM data protocol.
- Initial design: 2 twinned (on and off-shore) FCM per floor.
- Issues when scaling with the number of towers
  - Cost/Power/Reliability of DAQ cluster hosting read-out boards

NaNet³ read-out board



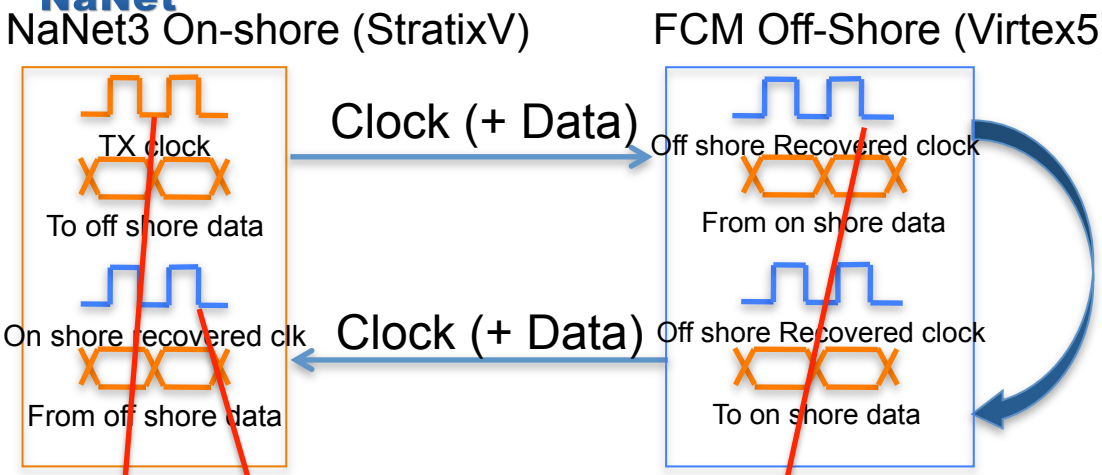
- **NaNet³** is the on-shore counterpart for 4 FCM boards
  - Reduce by a factor 4 the numerosity of DAQ cluster
- Deterministic latency links are required to obtain a common timing and known delay for the spatially distributed read-out
- Implemented on Terasic DE5-NET development board (4 SFP+ cages).

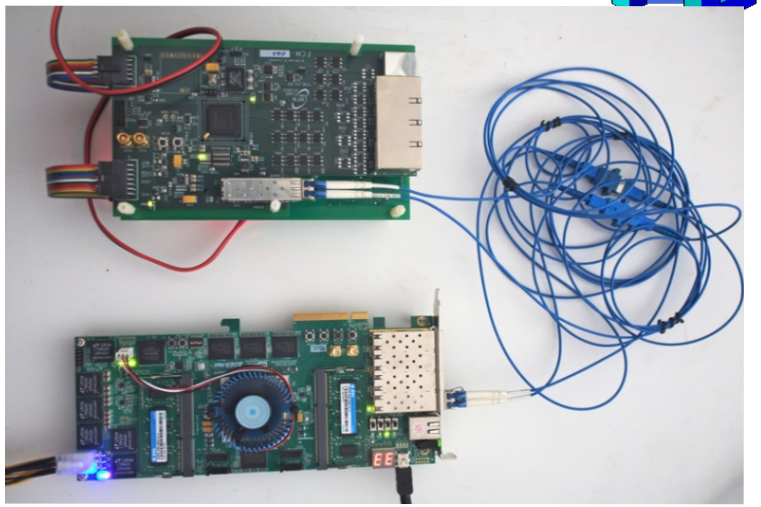
NaNet^3 overview

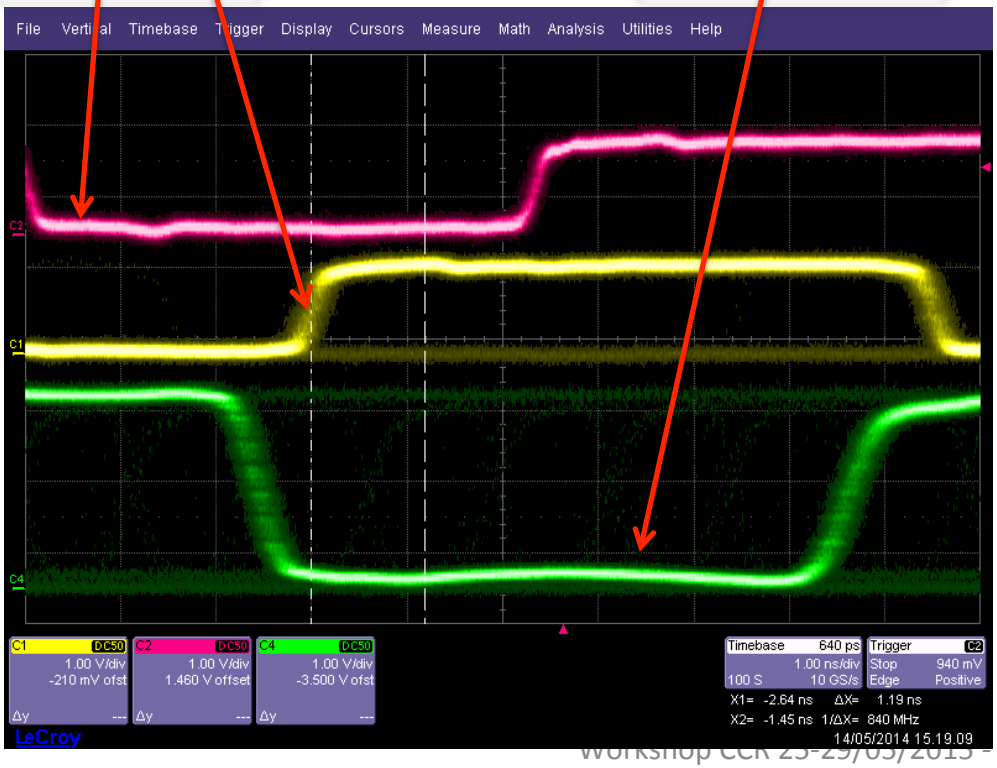


- RX path: payload of different off-shore devices, multiplexed on continuous data stream at fixed time slot
  - (PCIe DMA transaction to CPU/GPU memory)
- TX path: limited data rate per FCM (slow control)
  - (PCIe TARGET transaction from CPU/GPU mem.)
- Physical Layer: Altera **Deterministic Latency Transceivers (**8B10B encoding scheme)
- Data/Transport Layer: Time Division Multiplexing (TDM) data transmission protocol.

Deterministic latency link inter-operability



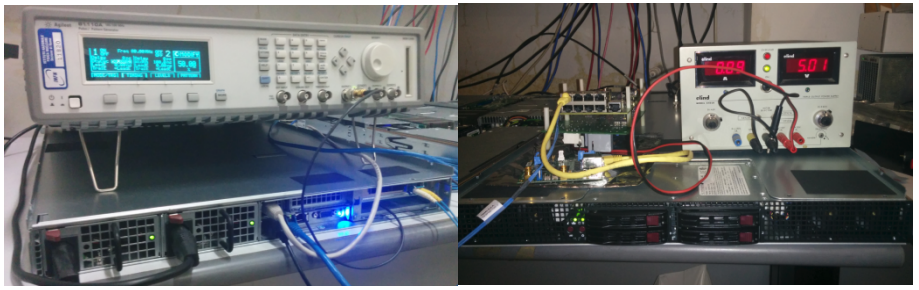


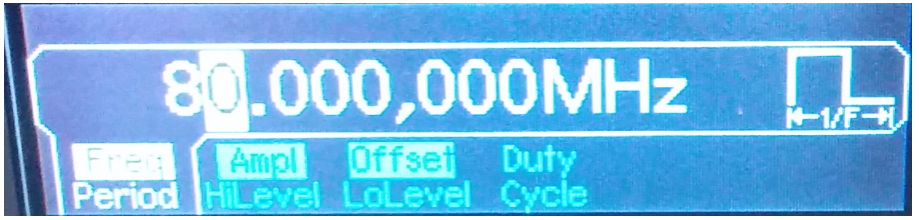


- Testbed: FCM vs Terasic DE5-Net
  - Custom hw mode for FCM Transceivers (Xilinx)
  - Latency deterministic mode for Stratix V Transceiver
  - 2mt copper and 2 mt long fiber
- Test:
  - 12 hours of periodic (~s) Tx clock reset to verify pll locking and rx word alignment

NaNet^3 current status

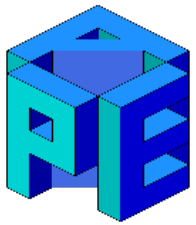
- Testbed Environment
  - FCMserver
  - NaNet^3
  - FCM + FEM (off-shore systems)
- Echo test (single channel)
  - NaNet/FCM roundtrip
- Slow-control test
  - TX: from NaNet^3 to FCM
- Data Acquisition test
  - Frame ID in slow_control_0
  - Data integrity
  - Single channel 96 hours test passed





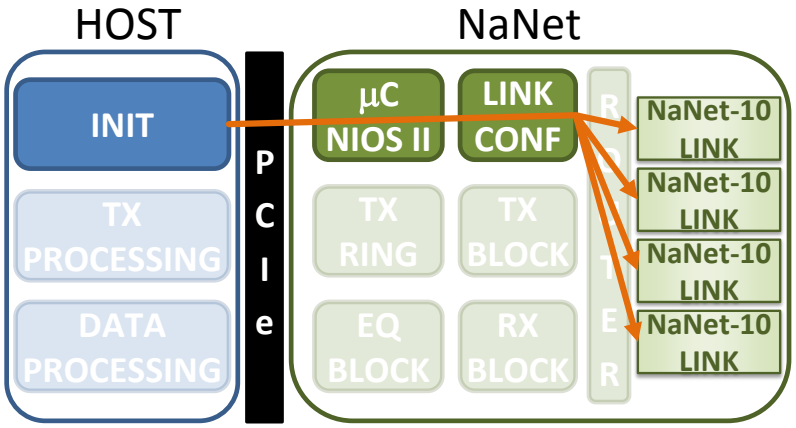
Current Status

- NA62 GPU-based Low Level Trigger
  - **NaNet-1**
    - Demonstrated real-time data communication between the NA62 RICH read-out system and the GPU-based L0 trigger processor over a fully functional single GbE link NIC.
    - Using the APElink channel we
    - demonstrated scalability of the design up to multiple 10GbE read-out channels
  - **NaNet-10**
    - Test of the single channel/PCIe Gen2 NaNet-10 version.
    - Development of the four channel/PCIe Gen3 version.
- KM3Net-IT on-shore read-out system (**NaNet^3**)
    - Single channel version is stable.
    - Started deployment of DAQ cluster.
    - Finalizing development of the four channel version.



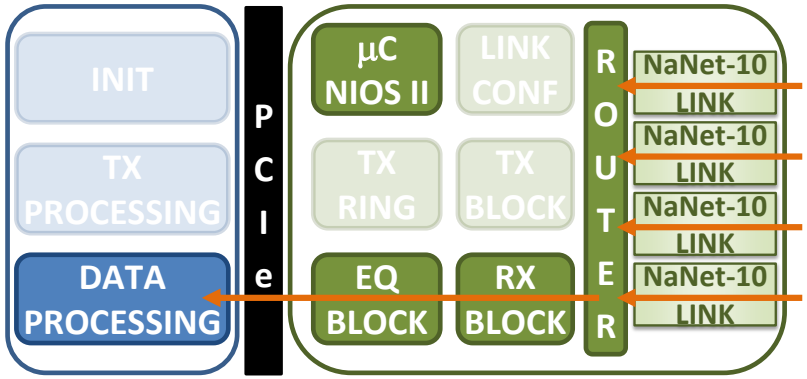
BACKUP SLIDES

NaNet Initialization Stage



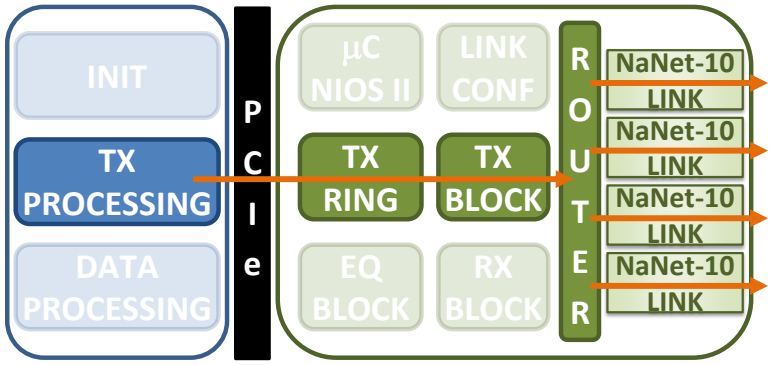
- The driver initializes the PCIe configuration registers of the board enabling the involved hardware component.
- CLOP buffers registration (pin and lock memory pages).
- Physical memory to virtual memory mapping are communicated to the card through the Nios II micro.
- Base address and size of CLOP buffers are dispatched to the link logic.

NaNet Data Reception & Processing Stage



- Processing of incoming according to application-specic demands
- Encapsulation into a packet-based protocol (APElink).
- HW generation of the packet virtual destination address (CPU/GPU).
- Dispatching of packet to RX path.
- Virtual-to-Physical memory address translation (TLB/Nios II).
- DMA write of payload to CPU/GPU.
- Writing of a DMA completion in an event queue mapped in HOST mem.
- An interrupt is generated.
- The driver signals the reception of new data to the application.
- Application process new data.

NaNet Data Transmission Stage

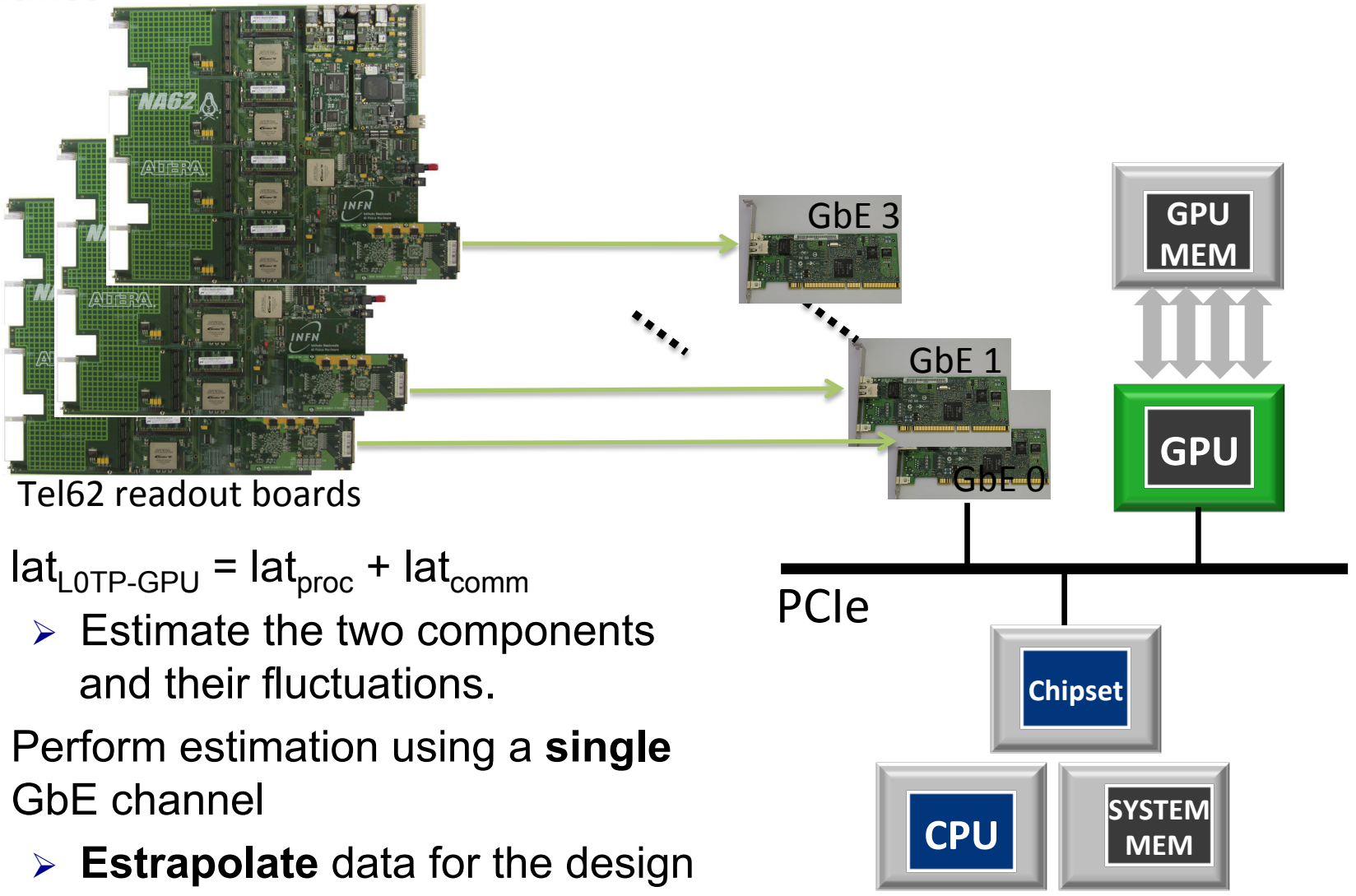


- The driver writes a command on the the HOSTmemory mapped TX ring buffer with source data memory address (CPU/GPU).
- The TX block manages the PCIe DMA memory read transactions.
- A new APElink packed if forged with the source data as payload.
- Packet is routed to the proper TX path.
- Payload is extracted and encapsulated into the I/O channel Data/Network/ Transport protocol (e.g. UDP).
- Packet is forwarded to the network.

GPU in HEP Real-time Computing

- **Hard real-time**: a system that has to respond strictly within a given time budget to avoid failures (or data loss).
  - Real time control systems (e.g. Tokamak plasma control)
  - First level (or low level) triggers systems
- **System latency**:
  - Is GPU processing latency small and stable enough for the given task?
  - Is the latency of network communications to and from GPU memory small and stable enough?
- **System throughput**:
  - Has the GPU enough computing power to execute the assigned task within the given time budget?
  - Has the GPU network I/O interface enough bandwidth?

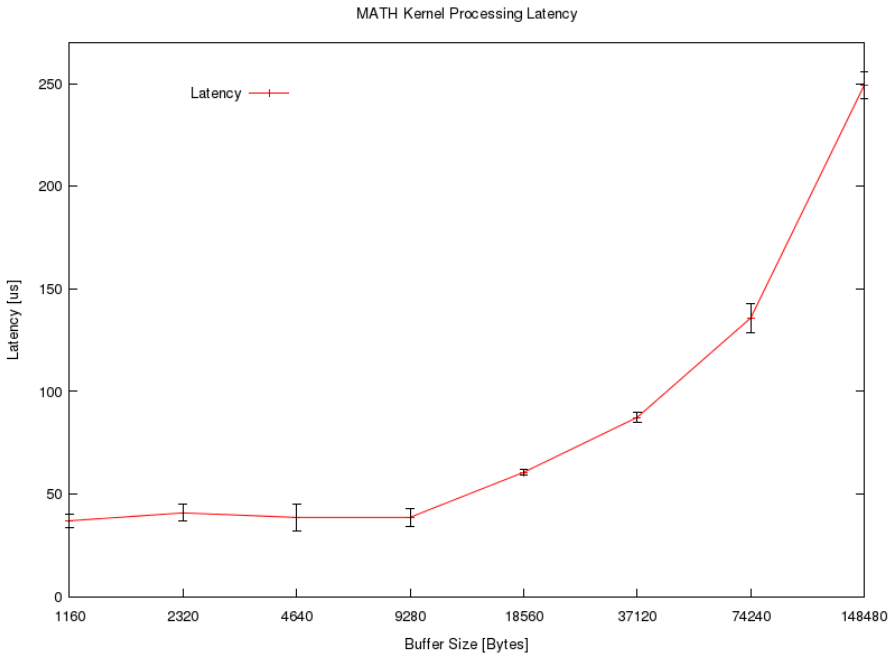
System Latency Estimation

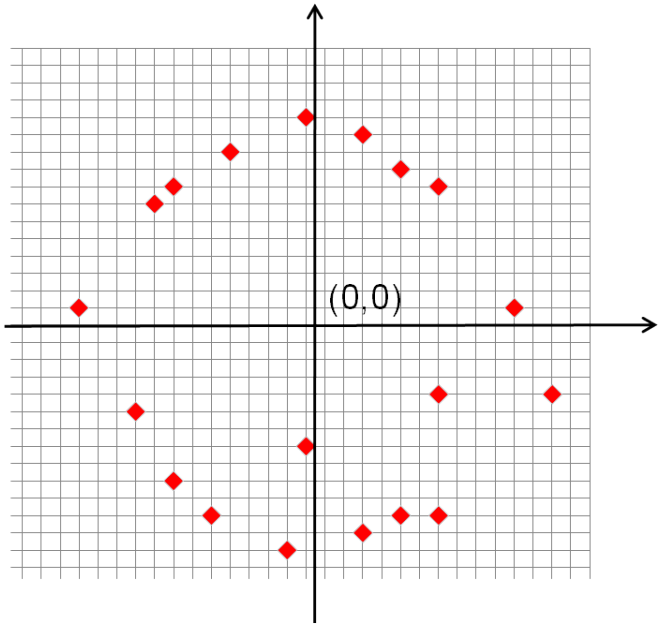


- $lat_{L0TP\text{-}GPU} = lat_{proc} + lat_{comm}$
  - Estimate the two components and their fluctuations.
- Perform estimation using a **single** GbE channel
  - **Estrapolate** data for the design of the "full bandwidth" system.

Processing Latency Estimation

- **lat_{proc}** : time needed to perform rings pattern-matching on the GPU with input and output data on device memory.



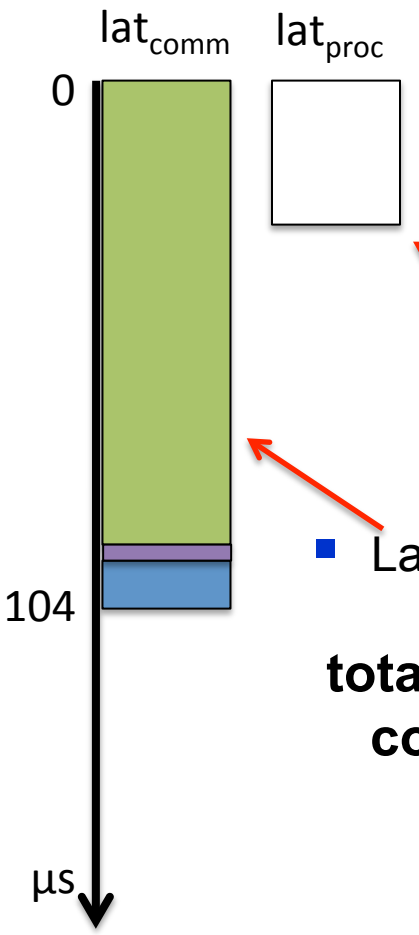


- **MATH** algo:. translation of the ring to centroid. In this system a least square method can be used. The circle condition can be reduced to a linear system, analitically solvable, without any iterative procedure.

Communication Latency
Standard GbE NIC / SW Stack

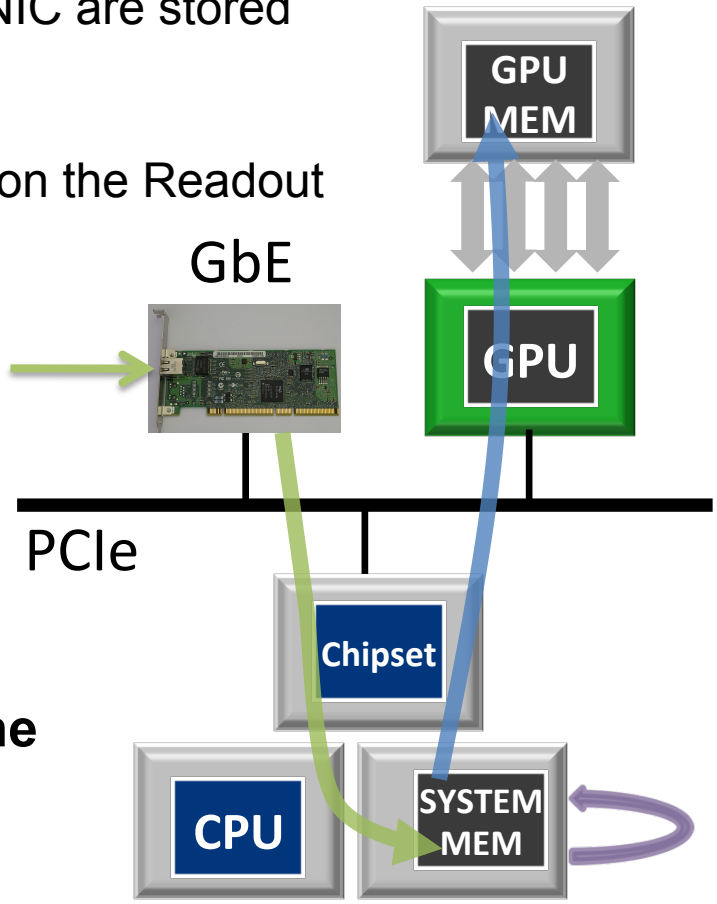
lat_{comm} : time needed to receive event data from GbE NIC to GPU memory.

- 40 events data (1400 bytes) sent from Readout board to the GbE NIC are stored in a receiving GPU buffer.
- T=0 start of send operation on the Readout Board.

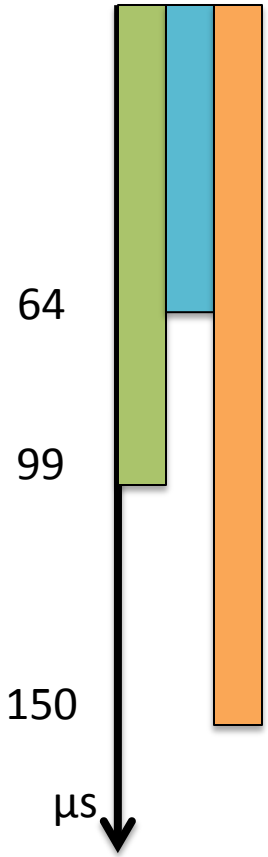


- Lat_{comm} = 104 us ≅ 3.5 x Lat_{proc}

total latency is dominated by the communication component.

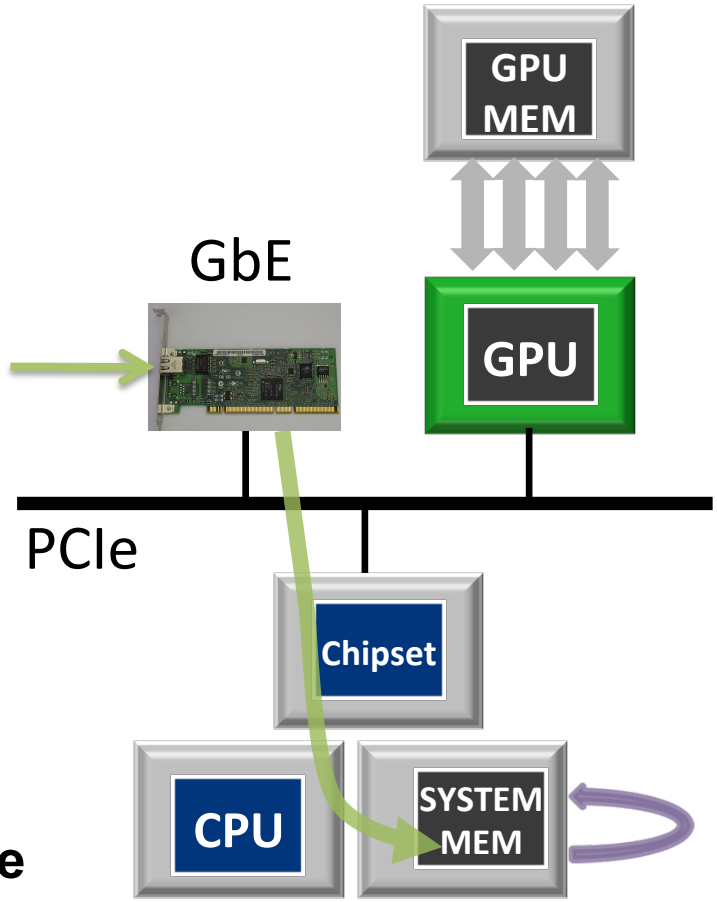


Communication Latency
Standard GbE NIC & SW Stack




```
sockperf: Summary: Latency is 99.129 usec
sockperf: Total 100816 observations; each
percentile contains 1008.16 observations

sockperf: ---> <MAX> observation = 657.743
sockperf: ---> percentile 99.99 = 474.758
sockperf: ---> percentile 99.90 = 201.321
sockperf: ---> percentile 99.50 = 163.819
sockperf: ---> percentile 99.00 = 149.694
sockperf: ---> percentile 95.00 = 116.730
sockperf: ---> percentile 90.00 = 105.027
sockperf: ---> percentile 75.00 = 97.578
sockperf: ---> percentile 50.00 = 96.023
sockperf: ---> percentile 25.00 = 95.775
sockperf: ---> <MIN> observation = 64.141
```



Fluctuations on latency of the data communication task **may hinder the real-time constraints** of the system.