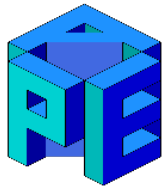


NaNet: a family of PCIe based Network Interface Cards for High Energy Physics

Andrea Biagioni
INFN – Sezione di Roma
On behalf of NaNet collaboration

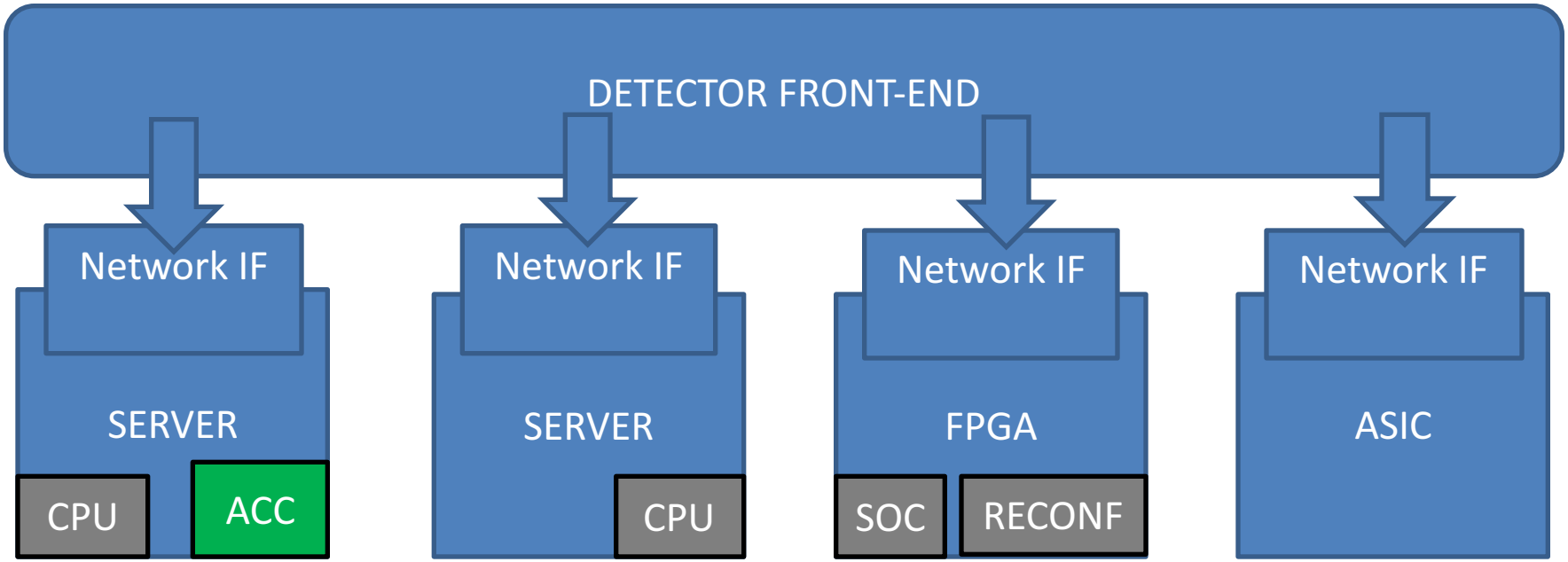
Perspectives of GPU computing in Science
26 -28 September 2016

NaNet 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.

Current and Future Platforms



NaNet-10	NaNet^3	NaNet-40
10GbE	Network IF:	40GbE
GPU nVIDIA	Custom (KM3link)	OpenCL/ARM v9

NaNet NIC Family

	NaNet-1	NaNet^3	NaNet-10	NaNet-40
Year	Q3 - 2013	Q1 - 2015	Q2 - 2016	Q4 - 2017
Device Family	Altera Stratix IV	Altera Stratix V	Altera Stratix V	?
Channel Technology	1 GbE	KM3link	10 GbE	40 GbE
Transmission Protocol	UDP	TDM	UDP	UDP
Number of Channel	1	4	4*	
PCIe	Gen2 x8	Gen2 x8	Gen3 x8**	Gen3 x8
SoC	NO	NO	NO	YES
OpenCL	NO	NO	NO	YES
nVIDIA GPUDirect RDMA	YES	YES	YES	YES
Real-time Processing	Decomp.	Decomp.	Decomp. Merger	?

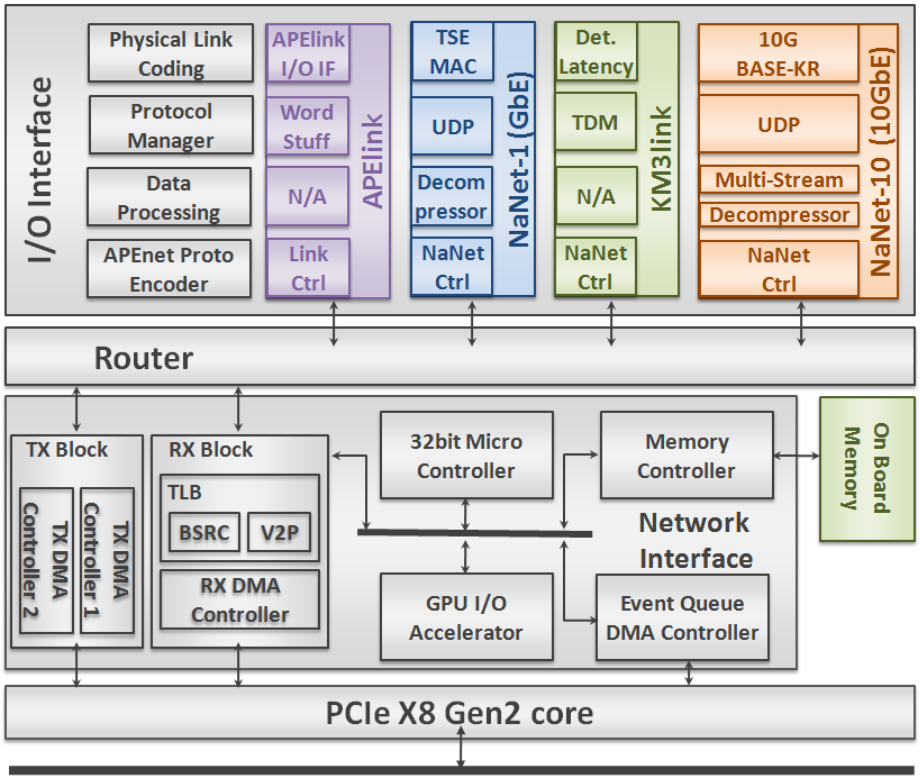
* 1 (v1.0)

** Gen2 x8 (v1.0)

References

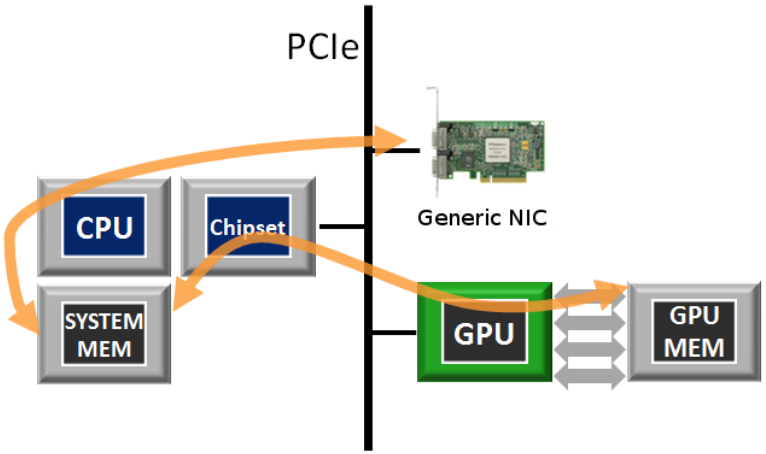
- NaNet-10
  - R. Ammendola et Al.,**"NaNet-10: a 10GbE Network Interface Card for the GPU-based low-level Trigger of the NA62 RICH Detector"**, Journal of Instrumentation, vol. 11, no. 03, p. C03030, 2016, doi:10.1088/1748-0221/11/03/C03030
  - R. Ammendola et Al., **"A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities."**, J. Phys.: Conf. Ser. (JPCS), Volume 664, 2015, doi:10.1088/1742-6596/664/9/092002
- NaNet^3
  - R. Ammendola et Al. **"NaNet^3: The on-shore readout and slow-control board for the KM3NeT-Italia underwater neutrino telescope"**, EPJ Web of Conferences, vol. 116, p. 05008, 2016, doi:10.1051/epjconf/201611605008
  - A. Lonardo et Al., **"NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing"**, 2015, JINST - Journal of Instrumentation, Proceedings of Topical Workshop on Electronics for Particle Physics (TWEPP) 2014, 10 C04011, IOP Publishing, doi:10.1088/1748-0221/10/04/C04011
- NaNet-1
  - R. Ammendola et Al "**NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPUs**", in JINST, Journal of Instrumentation, Proceedings of Topical Workshop on Electronics for Particle Physics (TWEPP) 2013, IOP Publishing, 2014 doi:10.1088/1748-0221/9/02/C02023
  - R. Ammendola et Al, **"NaNet: a low-latency NIC enabling GPU-based, real-time low level trigger systems"**, 2014, J. Phys.: Conf. Ser. 513 012018 doi:10.1088/1742-6596/513/1/012018

NaNet Modular Design

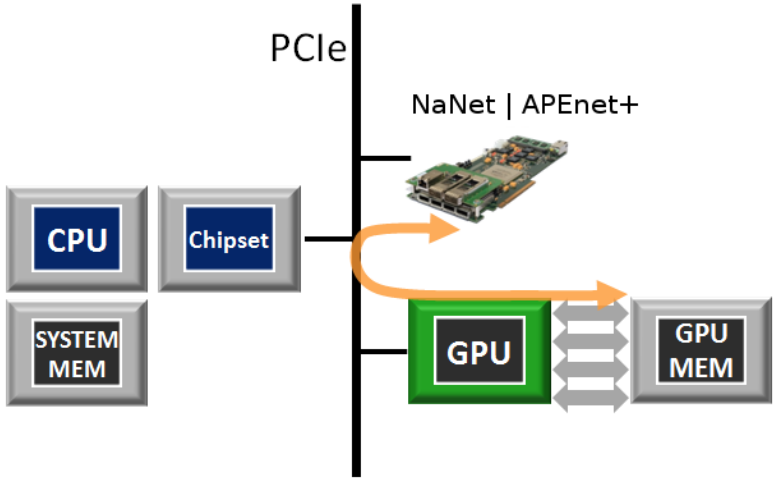


- I/O Interface
  - Multiple physical link technologies.
  - Network protocols offloading.
  - Application-specific processing on data stream.
- Router
  - Dynamically interconnects I/O and NI ports.
- Network Interface
  - Manages packets TX/RX from and to CPU/GPU memory.
  - Zero-Copy RDMA.
  - GPU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

NaNet Design – GPUDirect RDMA

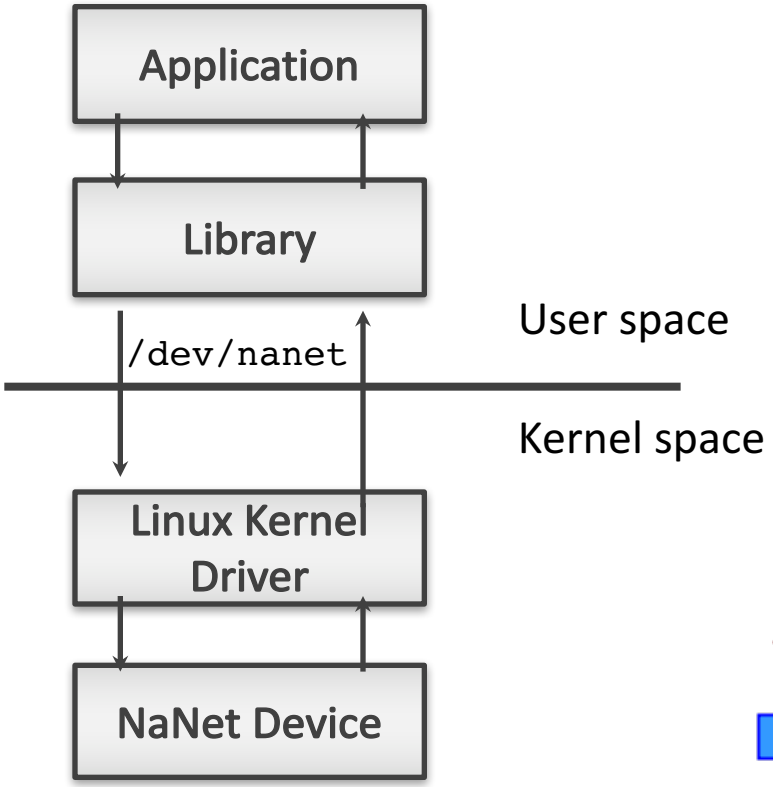


- Non-GPUDirect capable NIC data flow
- Intermediate buffering on CPU memory for I/O operations.

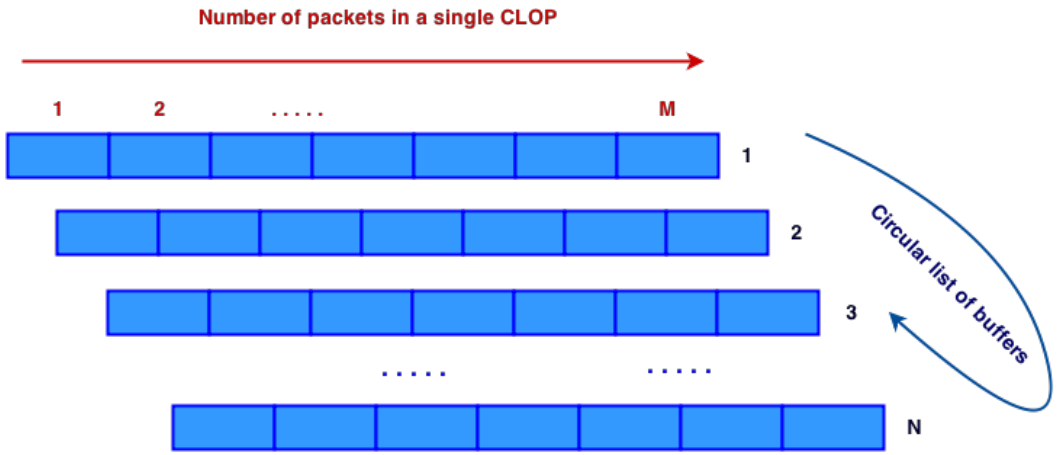


- GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement.
- No bounce buffers on host memory.
- Zero copy I/O.
- Latency reduction for small messages.
- nVIDIA Fermi/Kepler/Maxwell

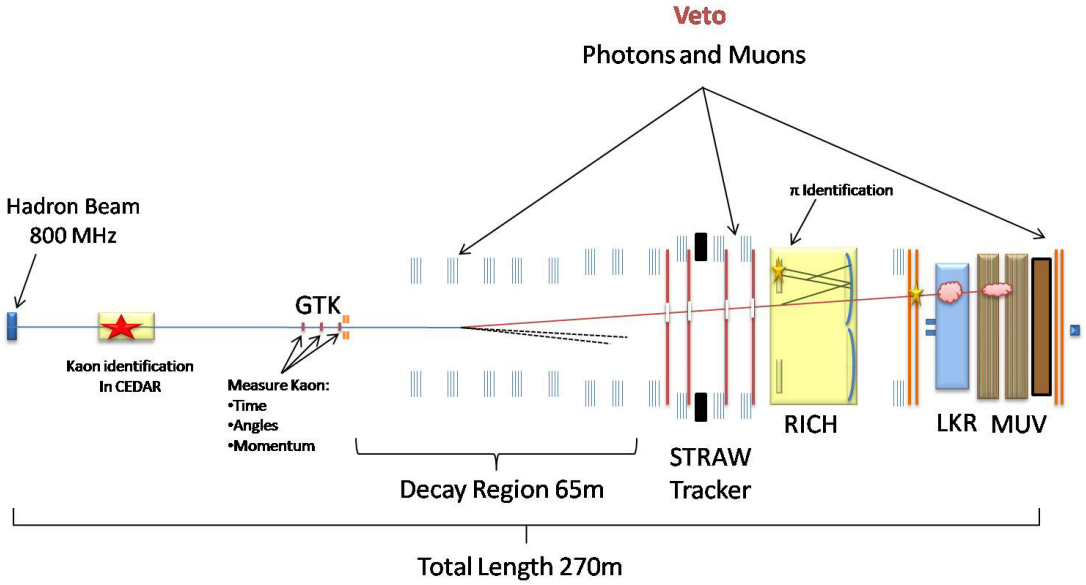
NaNet Software Stack



- 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.

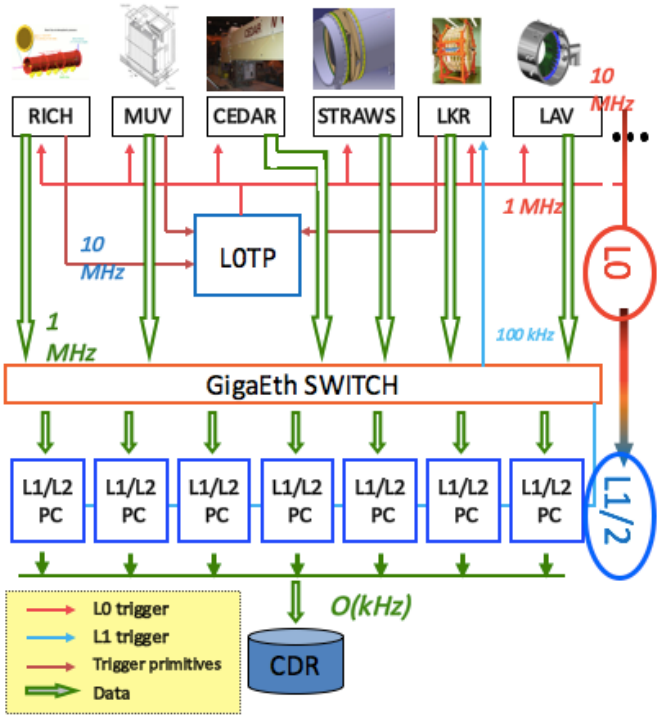


NA62

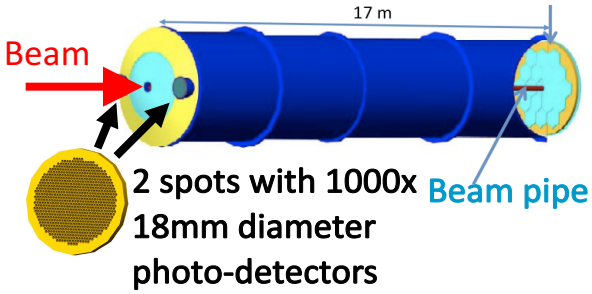


- Measurement of the ultra-rare decay (BR~8×10^{-11}) $K^+ \to \pi^+ \nu\bar{\nu}$
- Kaon decays in flight
- High intensity unsepareted hadron beam (6% Kaons)
- L0 Trigger: synchronous level must reduce rate from 10MHz to 1 Mhz
  - Latency: 1 ms

DAQ e TRIGGER

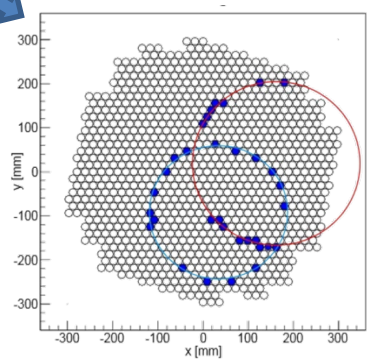


RICH detector



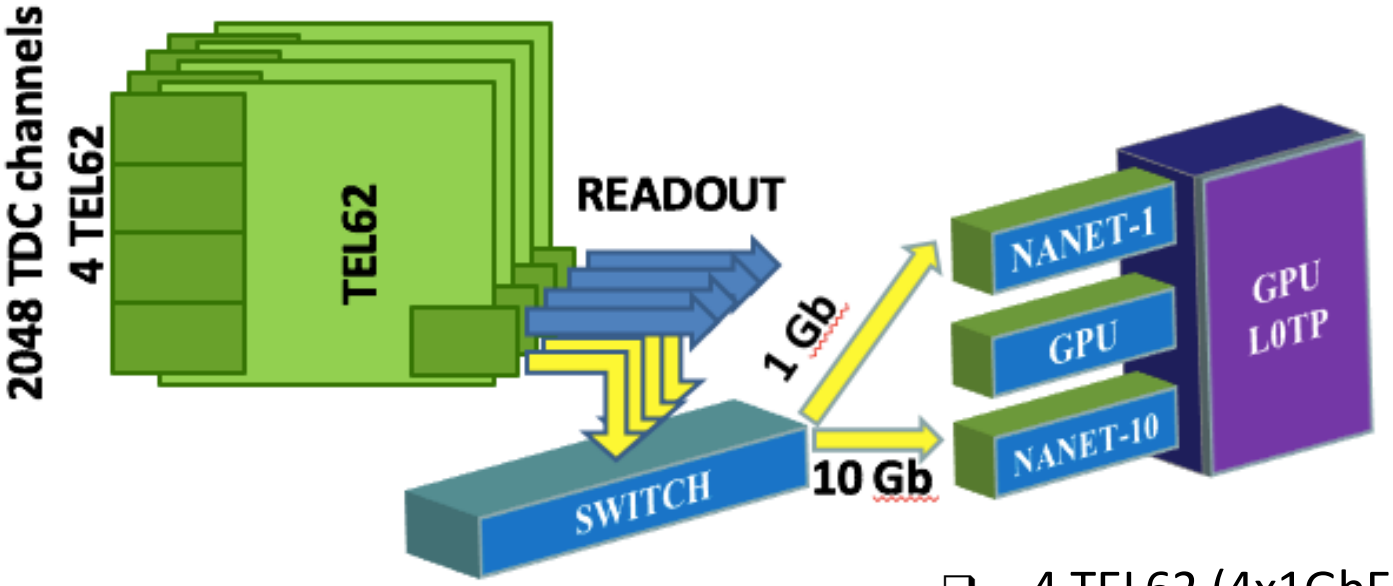
- Distinguish between pions and muons from 15 to 35 GeV (inefficiency < 1%)
- 2 spots of 1000 PMs each
- 2 read-out boards for each spot

GPU-based L0 trigger for Ring reconstruction



- Compare FPGA-based trigger with a GPU-based one
- More Selective trigger algorithms
  - Programmable
  - Upgradable
- Efficient match of circulat hit patterns

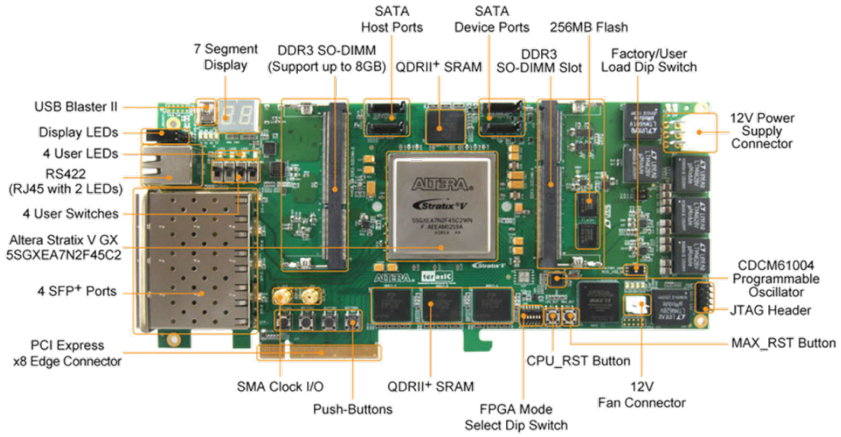
GPU-L0 TRIGGER

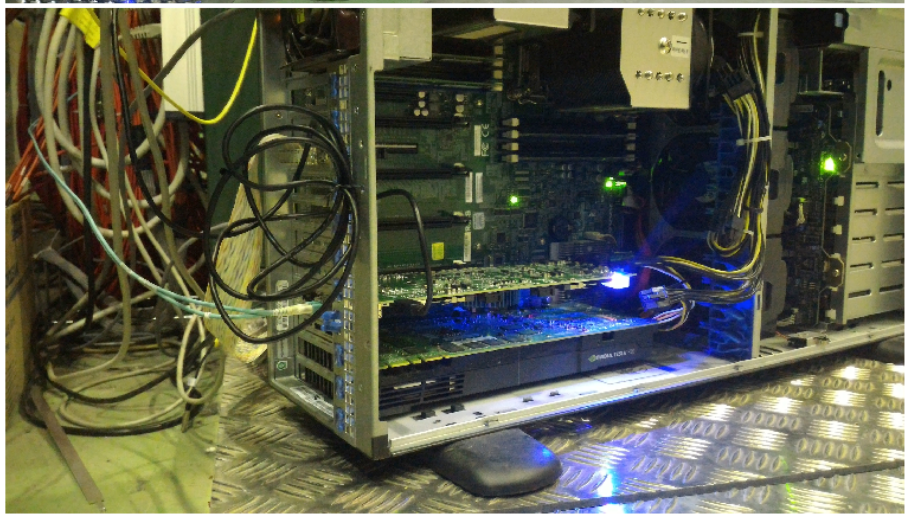


- 4 TEL62 (4x1GbE)
- 8×1Gb/s Readout
  - 4×1Gb/s trigger primitive
  - 4×1Gb/s GPU trigger
- Event Rate: 10 MHz
- L0 trigger rate: 1 MHz
- Max Latency: 1 ms

NaNet-10

- Terasic DE5-NET (Altera Stratix V)
- PCIe x8 Gen3
- 4 SFP+ ports (10GbE)
  - 10GBASE-KR
- nVIDIA GPUDirect RDMA
- UDP offloading
- Real-time processing
  - Decompression
  - Event Merger

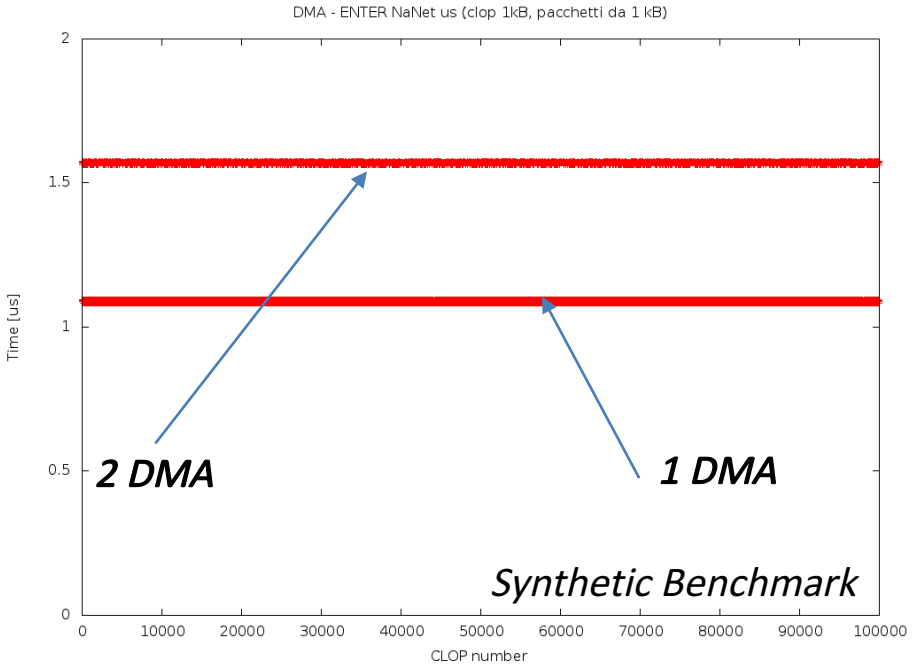




NaNet-10 @CERN

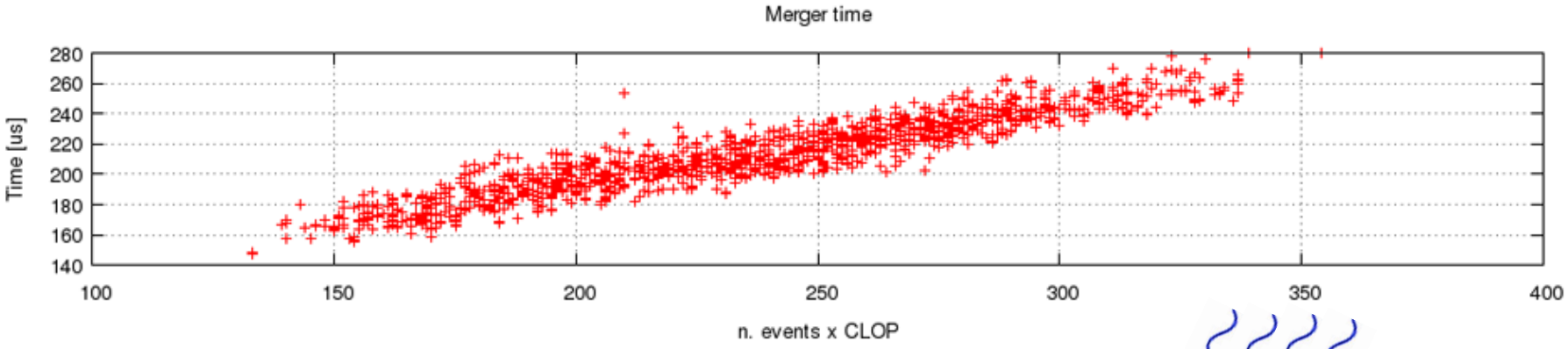
DATA FLOW & GPU PROCESSING

- NIC data flow
  - UDP manager
  - Decompressor
  - Event Merger
  - NaNet Transmission Control Logic
  - GPU memory write process

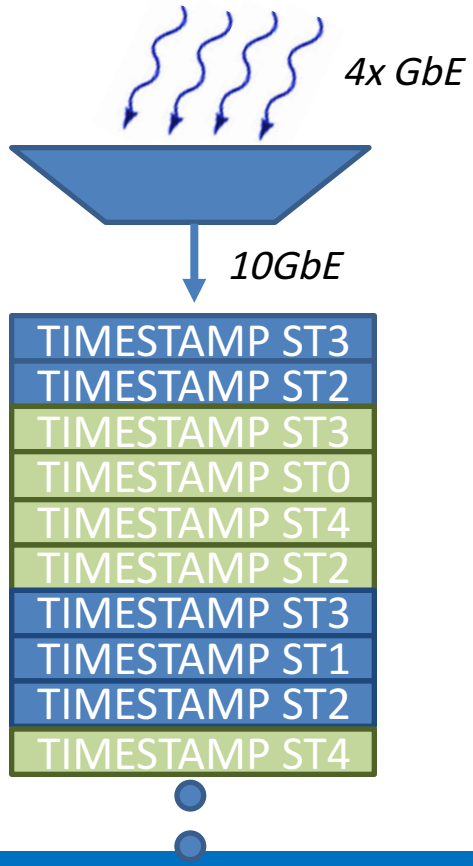


- Data Gathering
  - Completion: CLOP is ready
- GPU Processing
  - Event Finder
  - Fitter
- GPU processing $\leq$ Data Gathering!!!
  - Otherwise loss of data

Why HW Merger?



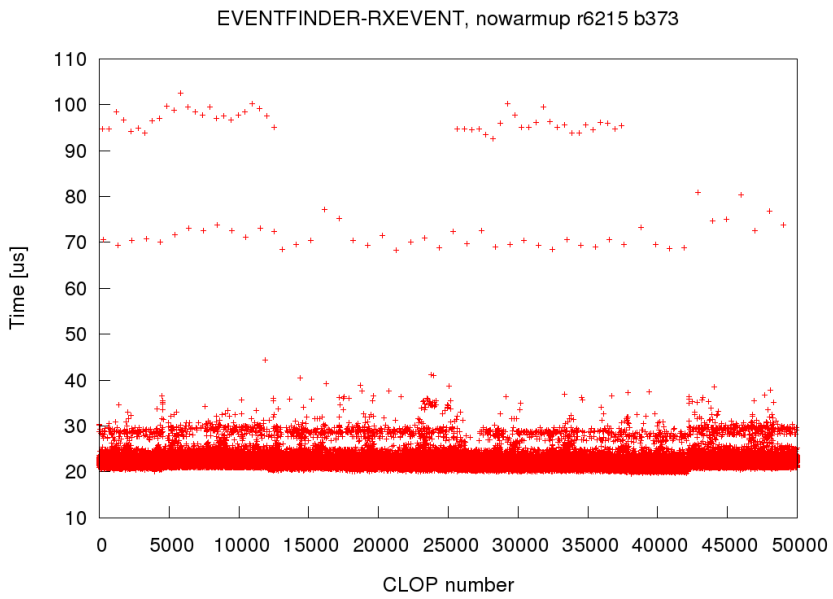
- **Merging the events coming from the RICH on GPU… *NO WAY***
  - it requires synchronization and serialization
  - computing kernel launched after merging
- Gathering latency: 200µs
- GPU Merger latency: 250µs (higher than gathering, data loss)
  - 800ns @event
- HW Merger Latency: 300ns @event (1.2 µs per max size merged event)!



EVENT FINDER

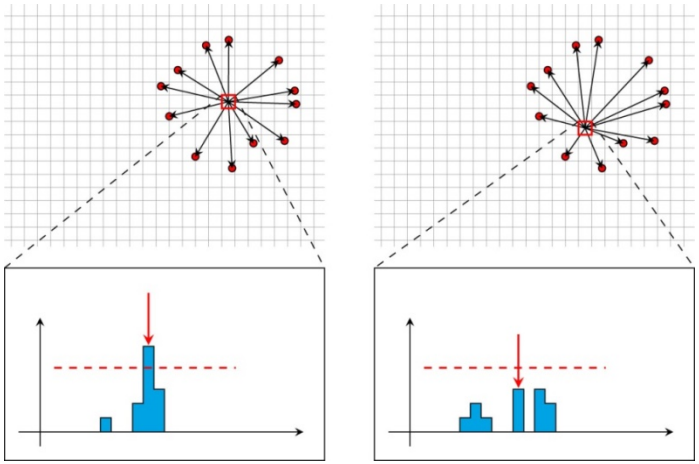
STR 3 MGP	STR 2 MGP	STR 1 MGP	STR 0 MGP	STR 3 HIT	STR 2 HIT	STR 1 HIT	STR 0 HIT	PATTERN		TOTAL HIT		TIMESTAMP			
STREAM 1; HIT 1		STREAM 1; HIT 0		STREAM 0; HIT 5		STREAM 0; HIT 4		STREAM 0; HIT 3		STREAM 0; HIT 2		STREAM 0; HIT 1		STREAM 0; HIT 0	
STREAM 2; HIT 0		STREAM 1; HIT 8		STREAM 1; HIT 7		STREAM 1; HIT 6		STREAM 1; HIT 5		STREAM 1; HIT 4		STREAM 1; HIT 3		STREAM 1; HIT 2	
STREAM 2; HIT 8		STREAM 2; HIT 7		STREAM 2; HIT 6		STREAM 2; HIT 5		STREAM 2; HIT 4		STREAM 2; HIT 3		STREAM 2; HIT 2		STREAM 2; HIT 1	
STREAM 3; HIT 4		STREAM 3; HIT 3		STREAM 3; HIT 2		STREAM 3; HIT 1		STREAM 3; HIT 0		STREAM 2; HIT 11		STREAM 2; HIT 10		STREAM 2; HIT 9	
PADDING										STREAM 3; HIT 7		STREAM 3; HIT 6		STREAM 3; HIT 5	
127…120	119…112	111…104	103…96	95…88	87…80	79…72	71…64	63…56	55…48	47…40	39…32	31…24	23…16	15…8	7…0

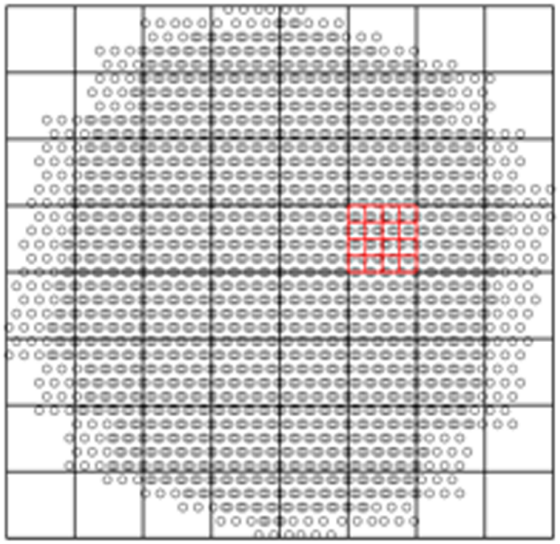
- Events are arranged in CLOPs with a new format more suitable for GPU's threads memory access Multi Merged Event GPU Packet (M^2EGP).
- Problem: searching for events position inside a CLOP using 1 thread on GPU takes > 100us for hundreds of events
- Solution: it must be parallelized. We can use all the threads looking for a known bytes pattern at the begin of every event: it takes ~ 35µs for 1000 events in a buffer



Histogram: a pattern recognition algorithm

- XY plane divided into a grid
- An histogram is created with distances from these points and hits of the physics event
- Rings are identified looking at distance bins whose contents exceed a threshold value





2-step implementation

8x8 grid -> 64 threads x event

4x4 grid only around maximum

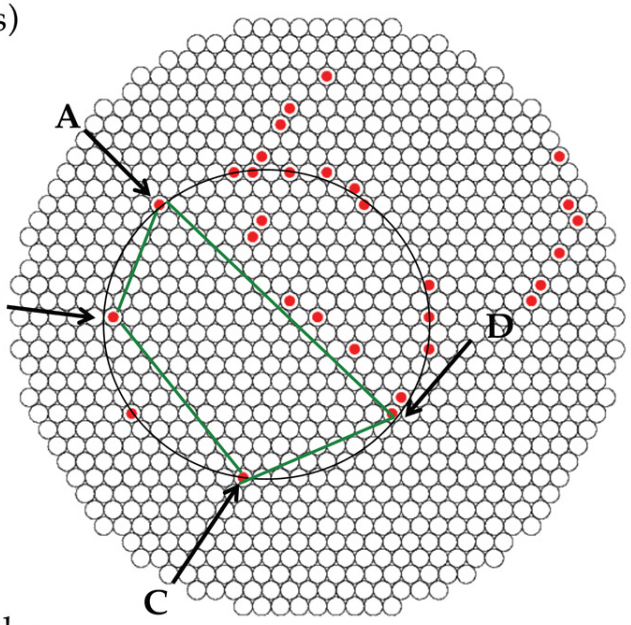
Almagest: a new multi-ring algorithm

i) Select a *triplet* (3 starting points)

ii) Loop on the remaining points: if the next point does not satisfy the Ptolemy's condition then reject it

iii) If the point satisfy the Ptolemy's condition then consider it for the fit

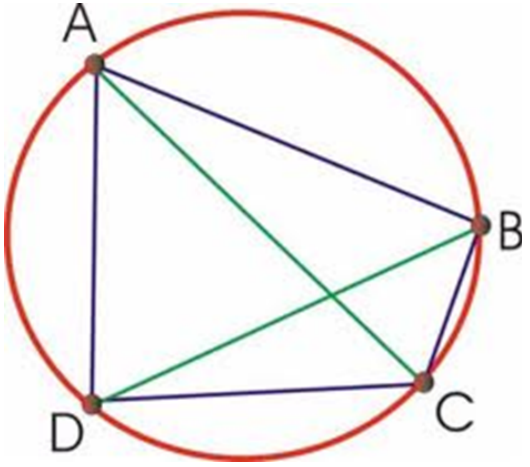
iv) …again…



Luca Pontisso et Al.
Poster:
"Real-time RICH ring reconstruction techniques on GPUs"

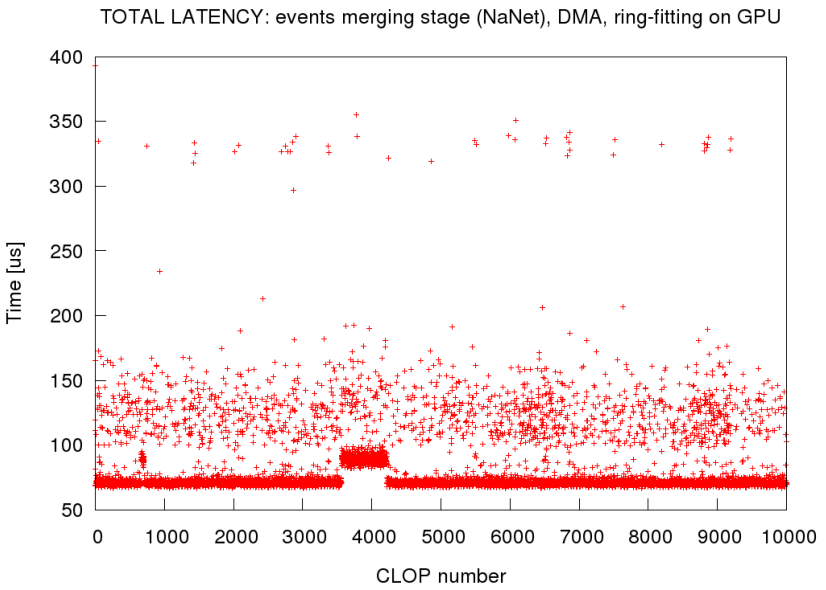
Based on Ptolemy's theorem:

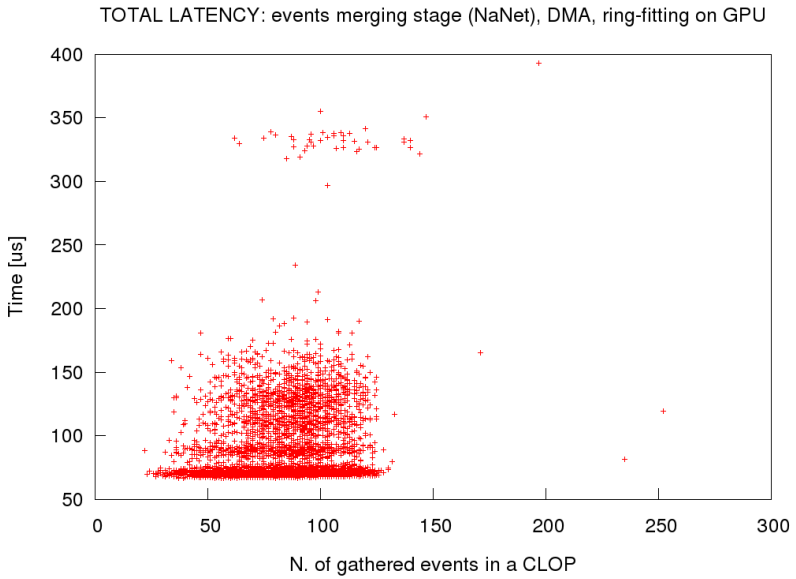
"A quadrilater is cyclic (the vertex lie on a circle) if and only if is valid the relation:
AD*BC+AB*DC=AC*BD"



NA62 2016 RUN

- Testbed (Experimental Results)
  - Supermicro X9DRG-QF Intel C602 Patsburg
  - Intel Xeon E5-2602 2.0 GHz
  - 32 GB DDR3
  - nVIDIA K20c
- ~ 25% target beam intensity ($9*10^{11}$ Pps)
- $\frac{1}{16}$ downscaling factor
- 8 CLOP, 32kB each
- Gathering time: 350µs





Conclusion

- NaNet-10 is ready
  - 10 GbE channel
  - Real-time processing: Decompressor and Merger stages
- Ring reconstruction on GPU
  - Histogram ($< 1\mu s$ per event)

- Future Work
  - NaNet-10: 4x 10GbE channels, PCIe Gen3 x8
  - Future NaNet NIC: OpenCL Kernel, SoC, 40GbE
  - New multi-ring algorithm on GPU: Almagest ($< 0.5\mu s$ per event)

Thank you

- NaNet Collaboration:

  R. Ammendola[(a)], A. Biagioni[(b)], P. Cretaro[(b)], S. Di Lorenzo[(c)] O. Frezza[(b)], G. Lamanna[(d)], F. Lo Cicero[(b)], A. Lonardo[(b)], M. Martinelli[(b)], P. S. Paolucci[(b)], E. Pastorelli[(b)], R. Piandani[(f)], L. Pontisso[(d)], D. Rossetti[(e)], F. Simula[(b)], M. Sozzi[(c)], P. Valente[(b)], P. Vicini[(b)]

(a) INFN Sezione di Roma Tor Vergata
(b) INFN Sezione di Roma
(c) INFN Sezione di Pisa and CERN
(d) INFN LNF and CERN
(e) nVIDIA Corporation, USA