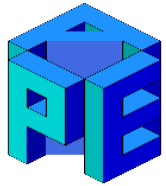


NaNet: a configurable Network Interface Card for Trigger and DAQ systems

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

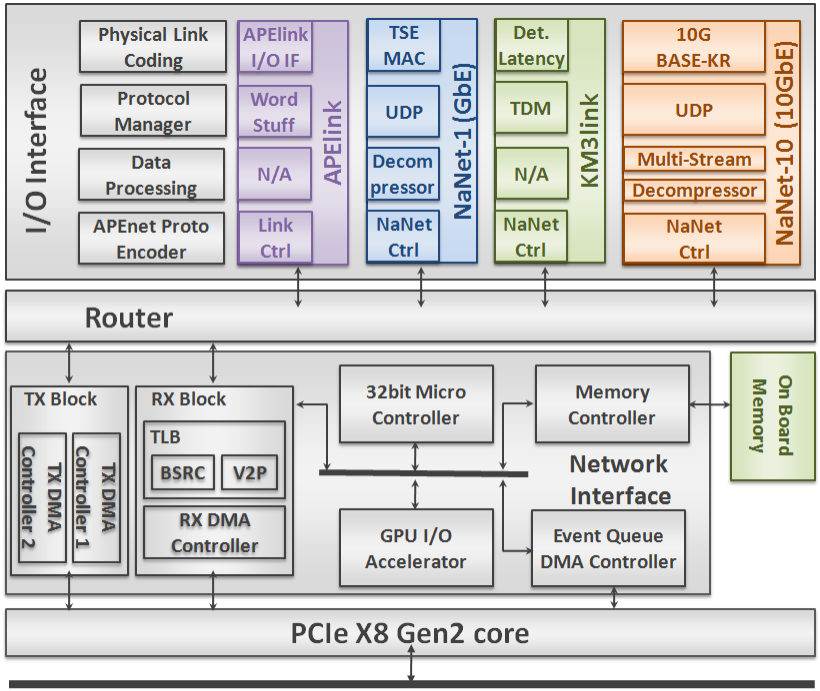
Conference on Computing in High Energy and Nuclear Physics
10 -14 October 2016

NaNet Objectives

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

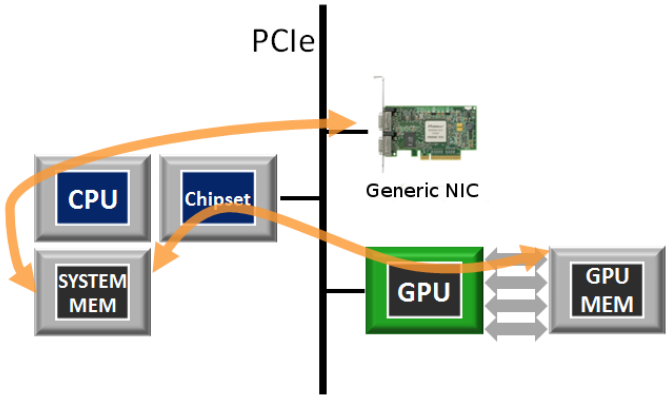
- Bridging the front-end electronics and the software trigger computing nodes.
- Supporting multiple link technologies and network protocols.
- Enabling a low and stable communication latency.
- Having a high bandwidth.
- Processing data streams from detectors on the fly (data compression/decompression and re-formatting, coalescing of event fragments, …).
- Optimizing data transfers with GPU accelerators.

NaNet Modular Design



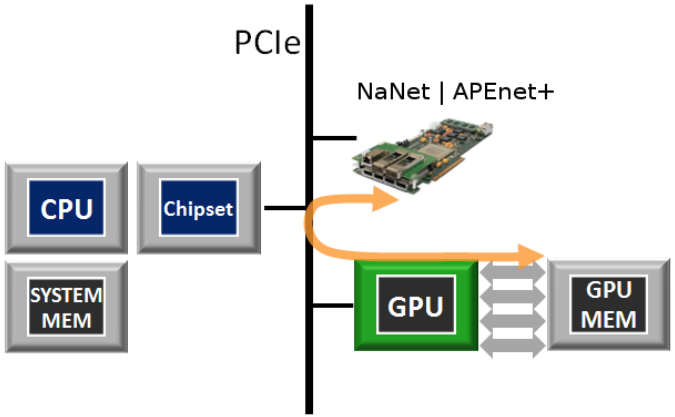
- I/O Interface
  - Multiple physical link technologies
  - Network protocols offloading
  - Application-specific processing on data stream
- Router
  - Dynamically interconnects I/O and NI ports
- Network Interface
  - Manages packets TX/RX from and to CPU/GPU memory
  - Zero-Copy RDMA
  - GPU I/O accelerator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

NaNet Design – GPUDirect RDMA



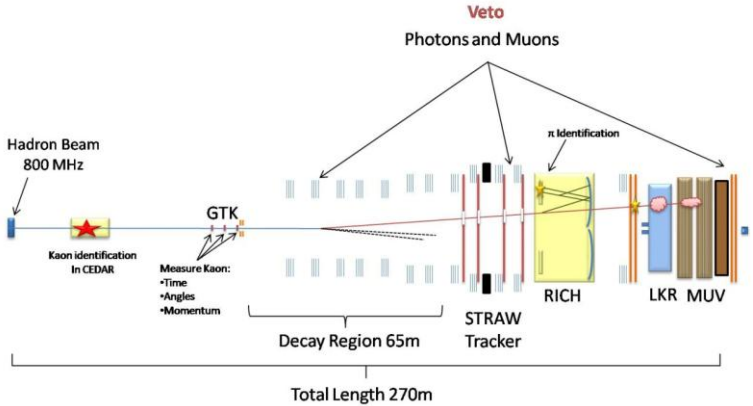
- Non-GPUDirect capable NIC data flow
- Intermediate buffering on CPU memory for I/O operations.

Andrea Biagioni et Al.
Poster:
"Latest generation interconnect technologies in APEnet+ networking infrastructure"



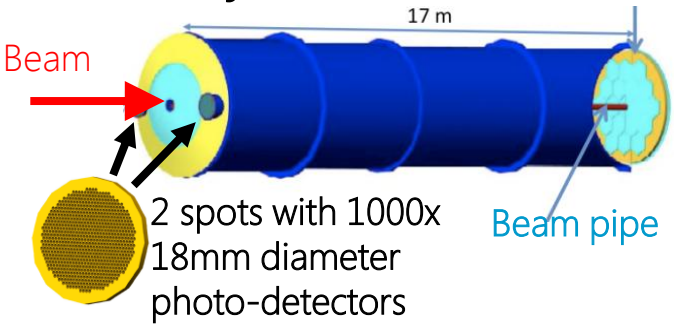
- GPUDirect allows direct data exchange on the PCIe bus with no CPU involvement.
- No bounce buffers on host memory.
- Zero copy I/O.
- Latency reduction for small messages.
- nVIDIA Fermi/Kepler/Maxwell

NA62

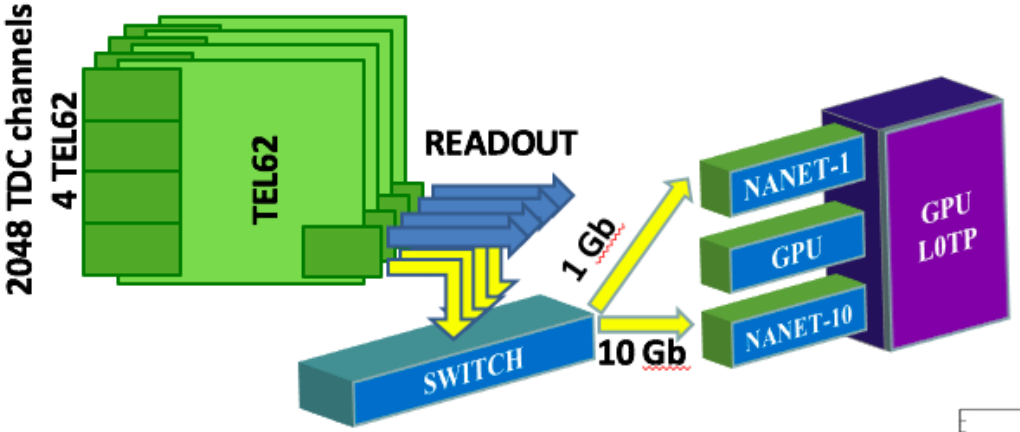


- Distinguish between pions and muons from 15 to 35 GeV (inefficiency < 1%)
- 2 spots of 1000 PMs each
- 2 read-out boards for each spot

- Measurement of the ultra-rare decay $K^+ \to \pi^+ \nu \bar{\nu}$ (BR~8×10^{-11})
- 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

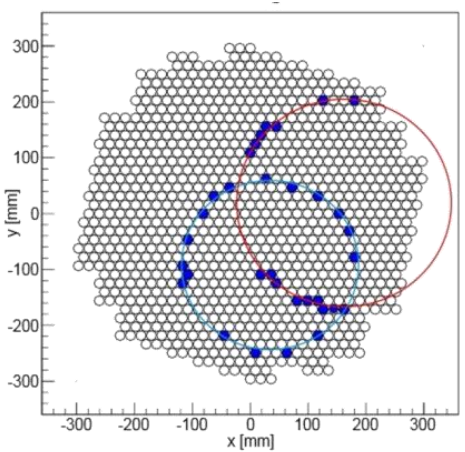


GPU-L0 TRIGGER



- 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

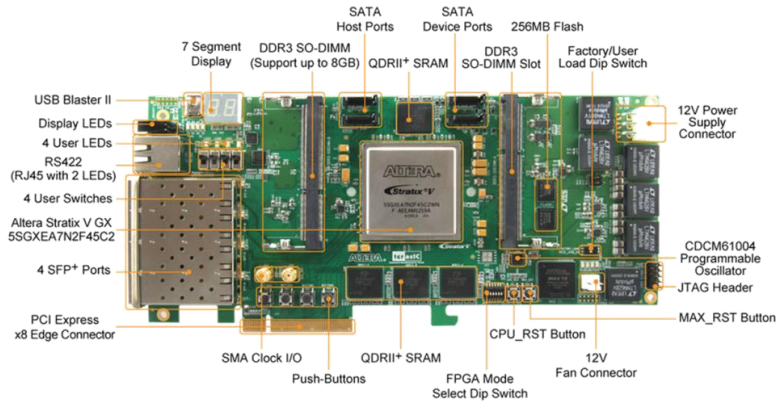
- Compare FPGA-based trigger with a GPU-based one
- More Selective trigger algorithms
  - Programmable
  - Upgradable
- Efficient match of circulat hit patterns



GPU-based L0 trigger for Ring reconstruction

NaNet-10

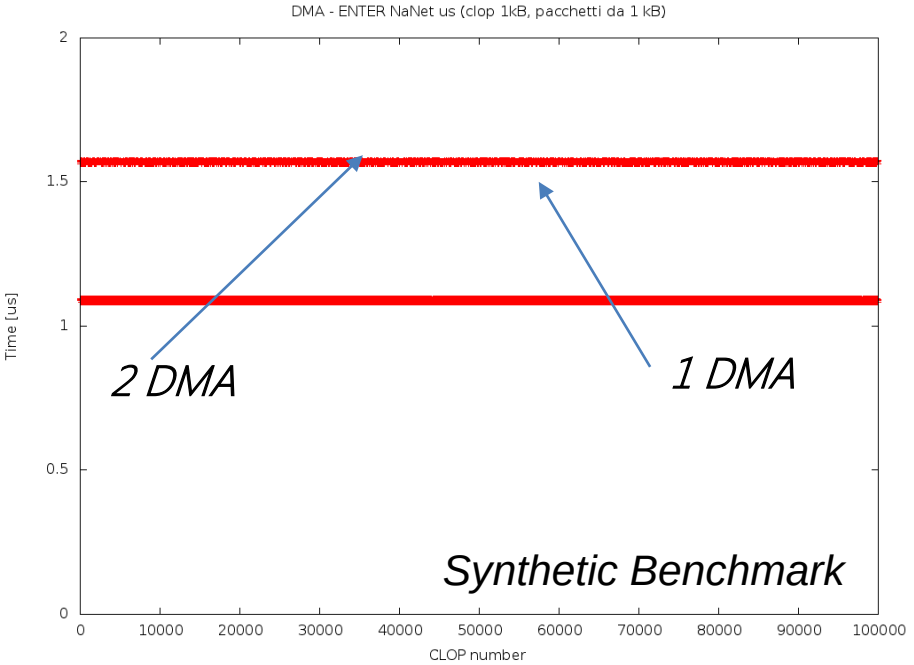
- Terasic DE5-NET (Altera Stratix V)
- PCIe x8 Gen3
- 4 SFP+ ports (10GbE)
- nVIDIA GPUDirect RDMA
- UDP offloading
- Real-time processing
  - Decompression, Event Merger



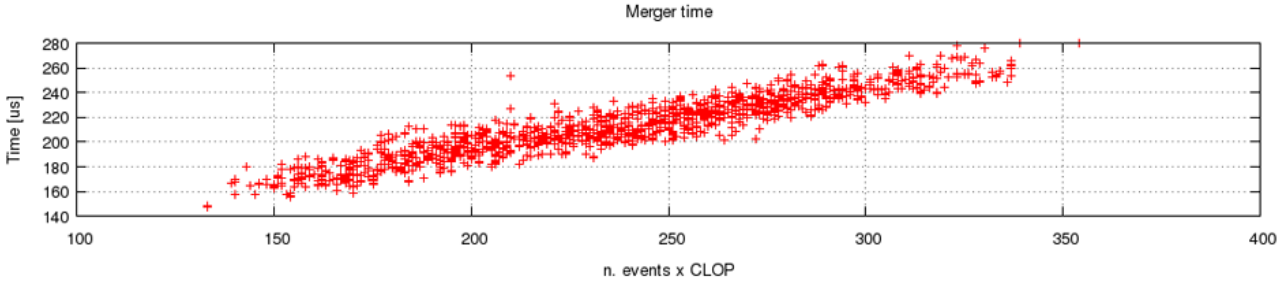
NaNet-10 @CERN

DATA FLOW & GPU PROCESSING

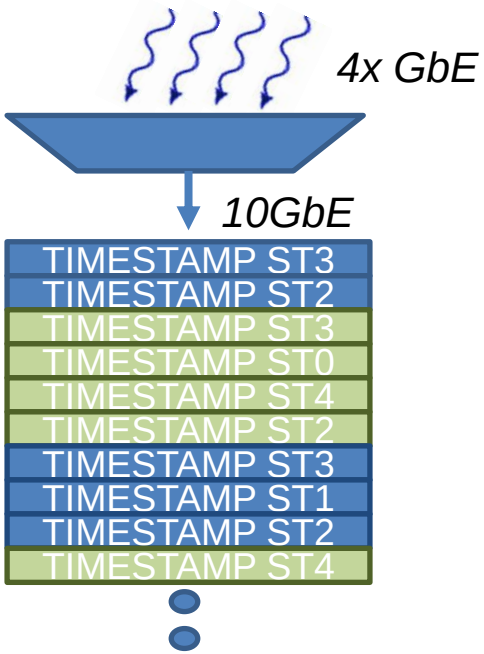
- NIC data flow
  - UDP manager
  - Decompressor
  - Event Merger
  - NaNet Transmission Control Logic
  - GPU memory write process
- Data Gathering
  - Completion: Data are ready
- GPU processing
  - Event Finder
  - Fitter
- GPU processing ≤ Data Gathering!!!
  - Otherwise loss of data



Why HW Merger?



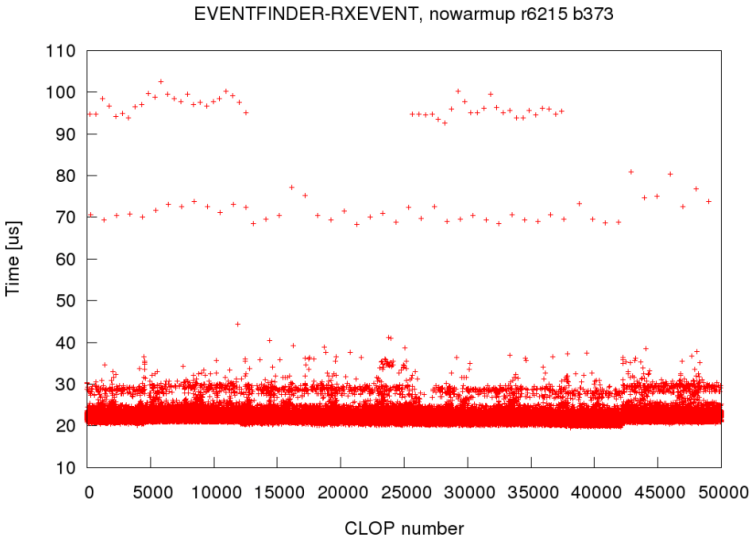
- **Merging the events coming from the RICH on GPU… *NO WAY***
  - it requires synchronization and serialization
  - computing kernel launched after merging
- Gathering latency: 200μs
- GPU Merger latency: 250μs (higher than gathering, data loss)
  - 800ns @event
- HW Merger Latency: 300ns @event



EVENT FINDER

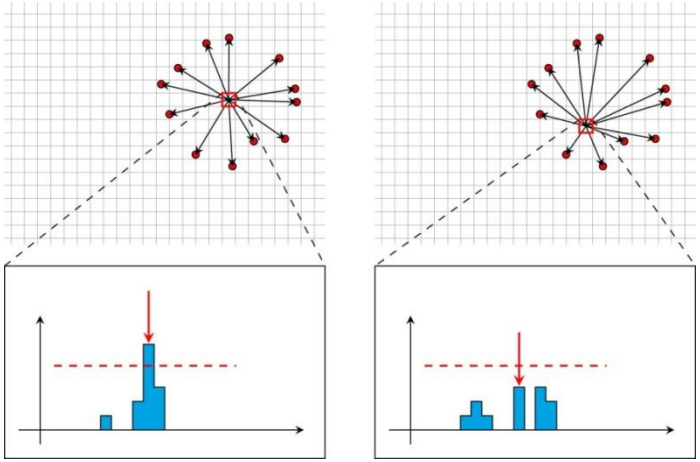
STR 3 MGP	STR 2 MGP	STR 1 MGP	STR 0 MGP	STR 3 HIT	STR 2 HIT	STR 1 HIT	STR 0 HIT	PATTERN	TOTAL HIT	TIMESTAMP	
STREAM 1; HIT 1		STREAM 1; HIT 0		STREAM 0; HIT 5		STREAM 0; HIT 4		STREAM 0; HIT 3	STREAM 0; HIT 2	STREAM 0; HIT 1	STREAM 0; HIT 0
STREAM 2; HIT 0		STREAM 1; HIT 8		STREAM 1; HIT 7		STREAM 1; 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

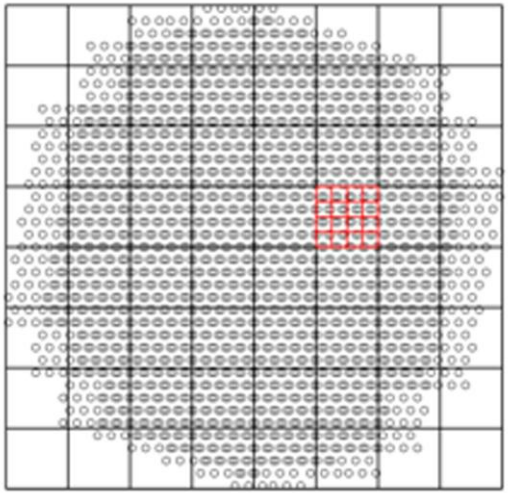
- Events are arranged in CLOPs with new format more suitable for GPU's threads memory access Multi Merged Event GPU Packet (M^2EGP).
- Problem: searching for events position inside a CLOP using 1 thread on GPU takes > 100us for hundreds of events
- Solution: it must be parallelized. We can use all the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



Histogram: 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





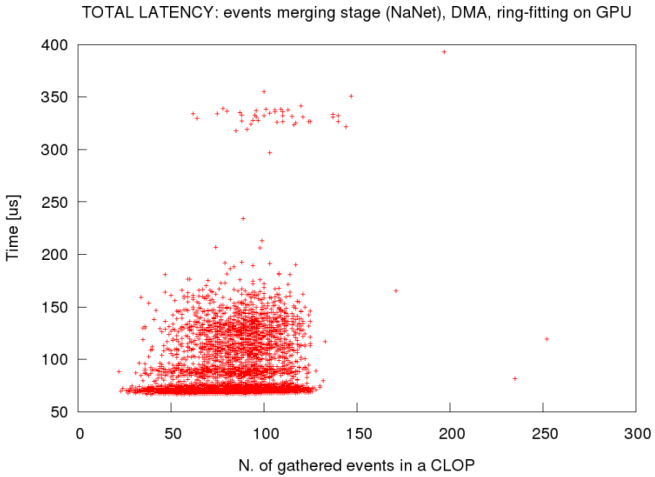
2-step implementation

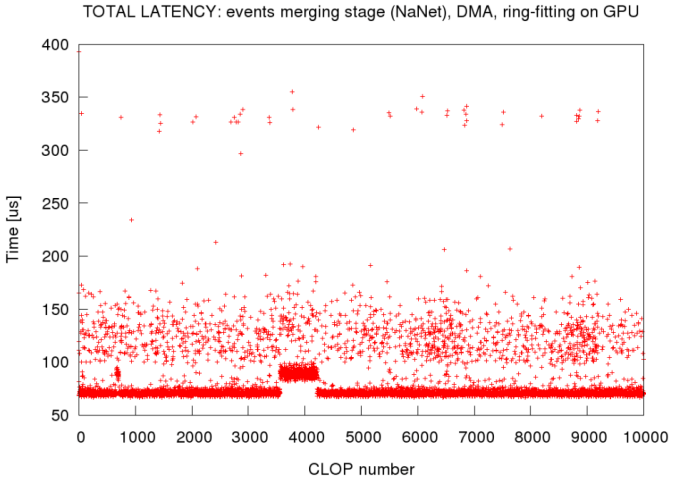
8x8 grid -> 64 threads x event

4x4 grid only around maximum

NA62 2016 RUN

- Testbed (experimental result)
  - Supermicro X9DRG-QF Intel C602 Patsburg
  - Intel Xeon E5-2602 2.0 GHz
  - 32 GB DDR3
  - nVIDIA K20c





- ~ 25% target beam intensity (9*1011 Pps)
- 1/16 downscaling factor
- 8 CLOP, 32kB each
- Gathering time: 350ms

Conclusion

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

- Future Work
  - NaNet-10: 4x 10GbE channels, PCIe Gen3 x8
  - Future NaNet NIC: OpenCL Kernel, SoC, 40GbE
  - New multiring algorithm on GPU: Almagest (<0.5 μs 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