NaNet$^3$: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope

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

Very Large Volume Neutrino Telescope 2015
14-16 September 2015
NaNet 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.
NaNet Modular Design

- I/O Interface
  - Multiple link.
  - Multiple network protocols.
    - Off-the-shelf: 1GbE, 10GbE
    - Custom: APElink (34gbps/QSFP), KM3link

- Router
  - Dynamically interconnects I/O and NI ports.

- Network Interface
  - Manages packets TX/RX from and to CPU/GPU memory.
  - TLB & Nios II Microcontroller
    - Virtual memory management

- PCIe X8 Gen2 Core
  - CPU BW:
    - 2.8 GB/s Read ÷ 2.5 GB/s Write
  - GPU BW:
    - 2.5 GB/s Read & Write.

Andrea Biagioni – VLVntT 2015
NaNet³: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope
**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:

> 600 MB/s throughput

- **14 floors** 20m vertically spaced. Each floor has:
  - 8 m bars, equipped with 6 Optical Modules
  - 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
NaNet$^3$ is the on-shore counterpart for 4 FCM boards
- Reduce by a factor 3.5 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).
NaNet³

- Device: TERASIC DE5-NET
- Based on Altera Stratix V FPGA
- PCIe Gen2 x8 (Gen3 support)
- 4 independent SFP+ ports (up to 10Gb/s)

NaNet³ specifications for enhanced KM3Net-IT read-out system:

- **Deterministic latency link**: “Fixed Latency” clock distribution for under-water events time-stamping
- Time Division Multiplexing protocol support
- 4 channels/PCIe board (i.e. less servers, less read-out boards,...)
NaNet³ developer team

- R. Ammendola\(^{(a)}\), A. Biagioni\(^{(b)}\), O. Frezza\(^{(b)}\), F. Lo Cicero\(^{(b)}\), A. Lonardo\(^{(b)}\), M. Martinelli\(^{(b)}\), P. S. Paolucci\(^{(b)}\), E. Pastorelli\(^{(b)}\), L. Pontisson\(^{(c)}\), D. Rossetti\(^{(d)}\), F. Simula\(^{(b)}\), L. Tosoratto\(^{(b)}\), P. Vicini\(^{(b)}\),

- Thanks to the TDAQ team: F. Ameli\(^{(b)}\), C. Nicolau\(^{(b)}\), F. Simeone\(^{(b)}\)

\(^{(a)}\) INFN Sezione di Roma Tor Vergata
\(^{(b)}\) INFN Sezione di Roma
\(^{(c)}\) INFN Sezione di Pisa and CERN
\(^{(d)}\) nVIDIA Corporation, USA
NaNet$^3$ overview

**KM3link**

- **RX**: payload of different off-shore devices, multiplexed on continuous data stream at fixed time slot
  - PCIe DMA transaction to CPU/GPU memory
- **TX**: 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.
- **NaNet Transmission Control Logic**
  - protocol translation; encapsulate TDM data stream in APEnet protocol
  - Virtual Memory management (CPU/μC offloading)
- **Virtual-to-Physical Translation**
  - Nios Implementation
  - HW acceleration: Translation Lookaside Buffer (TLB) based on associative memory
NaNet$^3$ SW/HW Overview

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

NaNet$^3$: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope
SW/HW Initialization

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

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

NaNet³: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope
OM Slow Control Management

- 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

NaNet³: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope
Data Reception and Processing

- 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 (clop buffer)
Target Memory Management

- **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
NaNet\textsuperscript{3} Software Stack

- **Host**
  - Linux Kernel Driver
    - Status/Configuration registers.
    - TX registers interface.
    - Custom Event Queue management
  - User space Library (open/close, buf reg, wait recv evts, ...)

- **Nios II Microcontroller**
  - Single process program performing System Configuration & Initialization tasks.

NaNet\textsuperscript{3}: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope
Deterministic latency link interoperability

- **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
Conclusions

- We present NaNet\textsuperscript{3}
  - PCIe Gen2 board
  - 4 channel up to 800 Mbps featuring Deterministic Latency and TDM protocol support

- Data Acquisition test (each channel separately)
  - Frame ID in OM0\_slow\_control
  - Data integrity
  - **48 hours test passed**

- Work-in-progress
  - Integration with FCMserver application
  - Testing the Multi-channel data acquisition

- Future Possible enhancements:
  - GPUDirect P2P/RDMA capability: to support GPU-based trigger developments
  - Link speed up to 10Gb/s: Further reduction in number of optical channels is possible
  - PCIe Gen3 implementation
THANK YOU!!
**NaNet I/O interface Customization:**

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

<table>
<thead>
<tr>
<th>Field</th>
<th>Validity Code</th>
<th>Start Address</th>
<th>Stop Address</th>
<th>Length [byte]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Frame</td>
<td>FRAME_CODE</td>
<td>0x0000</td>
<td>0x0000</td>
<td>1</td>
</tr>
<tr>
<td>Frame Time</td>
<td>DATA_CODE</td>
<td>0x0001</td>
<td>0x0005</td>
<td>1+6+1</td>
</tr>
<tr>
<td>FCM Slow Control</td>
<td>DATA_CODE</td>
<td>0x0009</td>
<td>0x000E</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM0 Slow Control</td>
<td>DATA_CODE</td>
<td>0x000F</td>
<td>0x0014</td>
<td>1</td>
</tr>
<tr>
<td>OM1 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0015</td>
<td>0x001A</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM2 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0018</td>
<td>0x0020</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM3 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0021</td>
<td>0x0026</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM4 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0027</td>
<td>0x002C</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM5 Slow Control</td>
<td>DATA_CODE</td>
<td>0x002D</td>
<td>0x0032</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM6 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0034</td>
<td>0x003F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM7 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0040</td>
<td>0x004F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM8 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0050</td>
<td>0x005F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM9 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0060</td>
<td>0x006F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM10 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0070</td>
<td>0x007F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM11 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0080</td>
<td>0x008F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM12 Slow Control</td>
<td>DATA_CODE</td>
<td>0x0090</td>
<td>0x009F</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM13 Slow Control</td>
<td>DATA_CODE</td>
<td>0x00A0</td>
<td>0x00AF</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM14 Slow Control</td>
<td>DATA_CODE</td>
<td>0x00B0</td>
<td>0x00BF</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM15 Slow Control</td>
<td>DATA_CODE</td>
<td>0x00C0</td>
<td>0x00CF</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM16 Slow Control</td>
<td>DATA_CODE</td>
<td>0x00D0</td>
<td>0x00DF</td>
<td>1+4+1</td>
</tr>
<tr>
<td>OM17 Slow Control</td>
<td>DATA_CODE</td>
<td>0x00E0</td>
<td>0x00FF</td>
<td>1+4+1</td>
</tr>
</tbody>
</table>

- **NaNet³ TX**

- **NaNet³ RX:**

Data flow: NaNet³ → FCM board → NaNet³
NaNet$^3$ current status

- Testbed Environment
  - Clock generator (80MHz)
  - FCMserver (see Matteo Manzali talk)
  - NaNet$^3$
  - FCM + FEM

- Echo test (single channel)
  - NaNet/FCM roundtrip

- Test di slow-control
  - TX: from NaNet$^3$ 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
Know How: APEnet+

- 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!!!
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
NaNet Design – I/O Interface

- **Physical Link Coding**
  - Standard: **1GbE** (1000Base-T), **10GbE** (10Base-KR).
  - Custom: **APElink** (34 Gb/s QSFP), **KM3link** Det. Lat. (2.5 Gb/s optical).

- **Protocol Manager**
  - Data/Network/Transport layers off-load (**UDP, TDM**)
  - Minimize latency fluctuations.

- **Data Processing**
  - Application-specific processing on data stream.
  - **e.g. NA62 Decompressor&Merger:** re-format event data, coalesce event fragments before forwarding packets to NI.

- **APEnet Protocol Encoder**
  - Protocol adaptation between on-board and off-board network.
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.
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 XpressRichGen3 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)