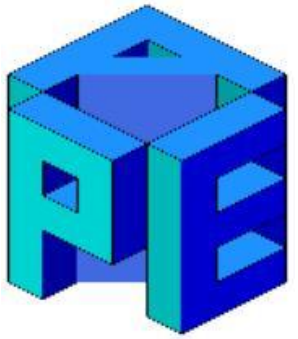
“Design and implementation of a modular, low latency, fault-aware, FPGA-based Network Interface”

Cancun, 09.12.2013

Ottorino Frezza
INFN



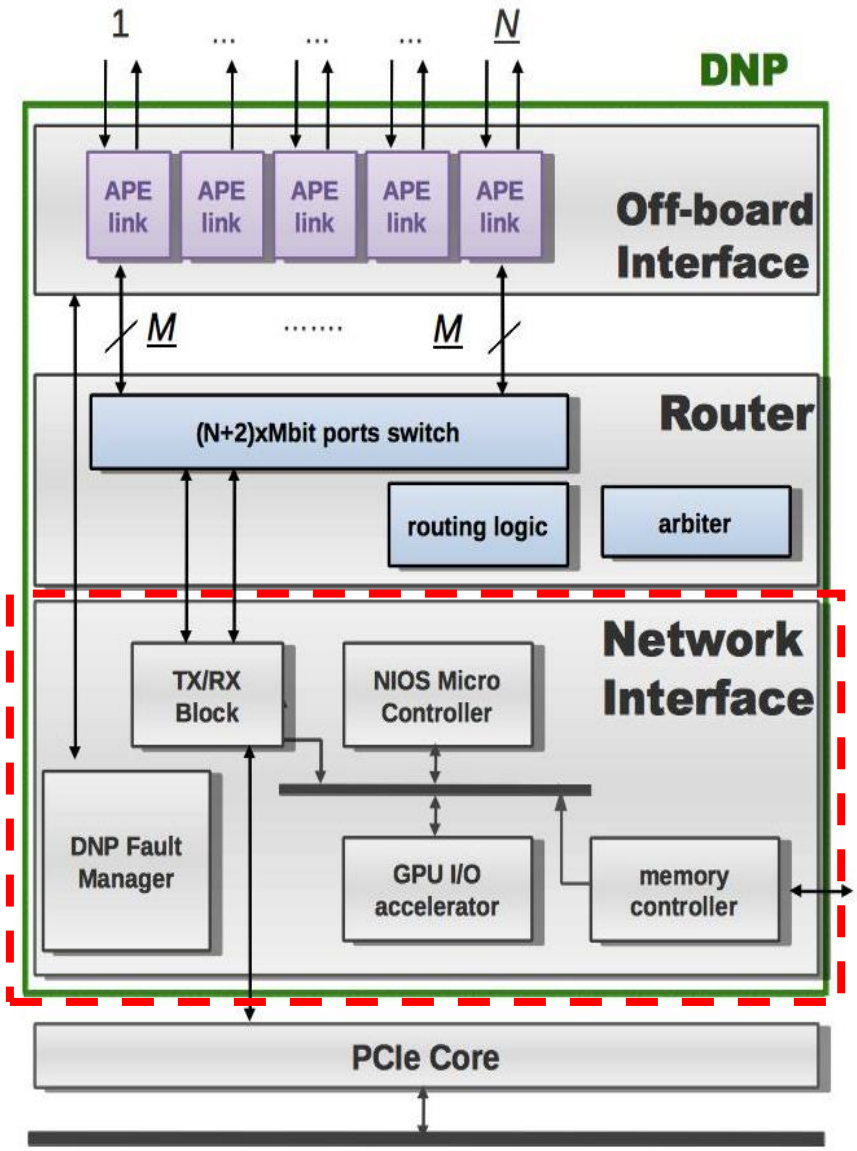
Outline

APEnet+: Point-to-point, low-latency network controller integrated in a PCIe board based on FPGA.

Features:

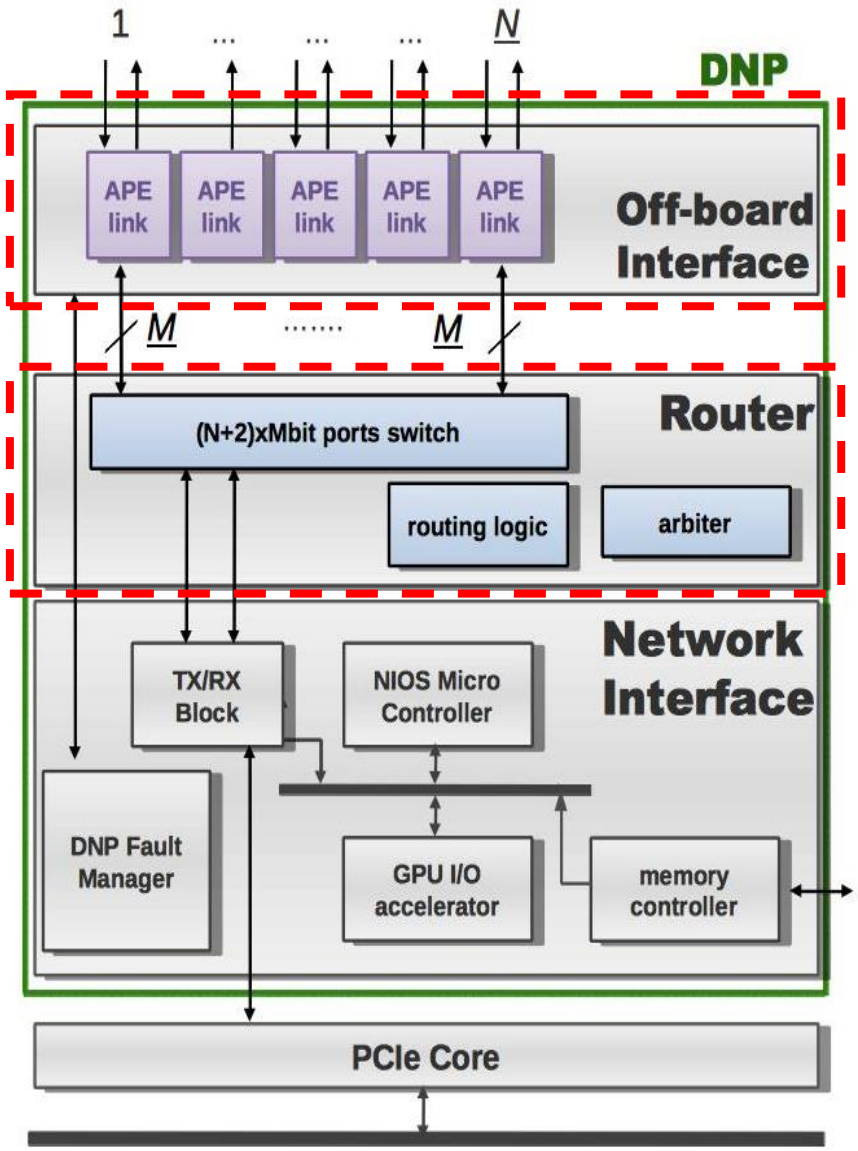
- Modularity (with parametric and reconfigurable IP);
  - APEnet+ 3links (Stratix IV development board)
  - APEnet+ 6links (Stratix IV production board)
  - APEnet+ 3links (Stratix V development board)
  - NaNet-1 (Stratix IV development board)
  - NaNet^3 (Stratix V development board)
- Low latency and High Bandwidth;
  - Support for RDMA and peer-to-peer protocol of NVIDIA GPUs;
  - Design improvement;
- Systemic fault-awareness (LO|FA|MO IP).

APEnet+ : Architecture Overview



- **Network Interface:**
  - TX block: gathers data coming from the PCI-e port, fragmenting data stream into packets forwarded to the relevant destination port;
  - RX block: provides hardware support for the Remote Direct Memory Access (RDMA) protocol, allowing remote data transfer over the network without involvement of the CPU of the remote node;
  - NIOS II Microcontroller: simplifies the DNP-core HW and the host-side driver;
  - GPU I/O accelerator: implements the peer-to-peer access to NVIDIA Fermi and Kepler class;
  - DNP Fault Manager: collects a set of health indexes (temperature, power drain…) and relays them through the network - employing a set of strategies in case of faulty links - while keeping zero overhead on the ordinary network activity.

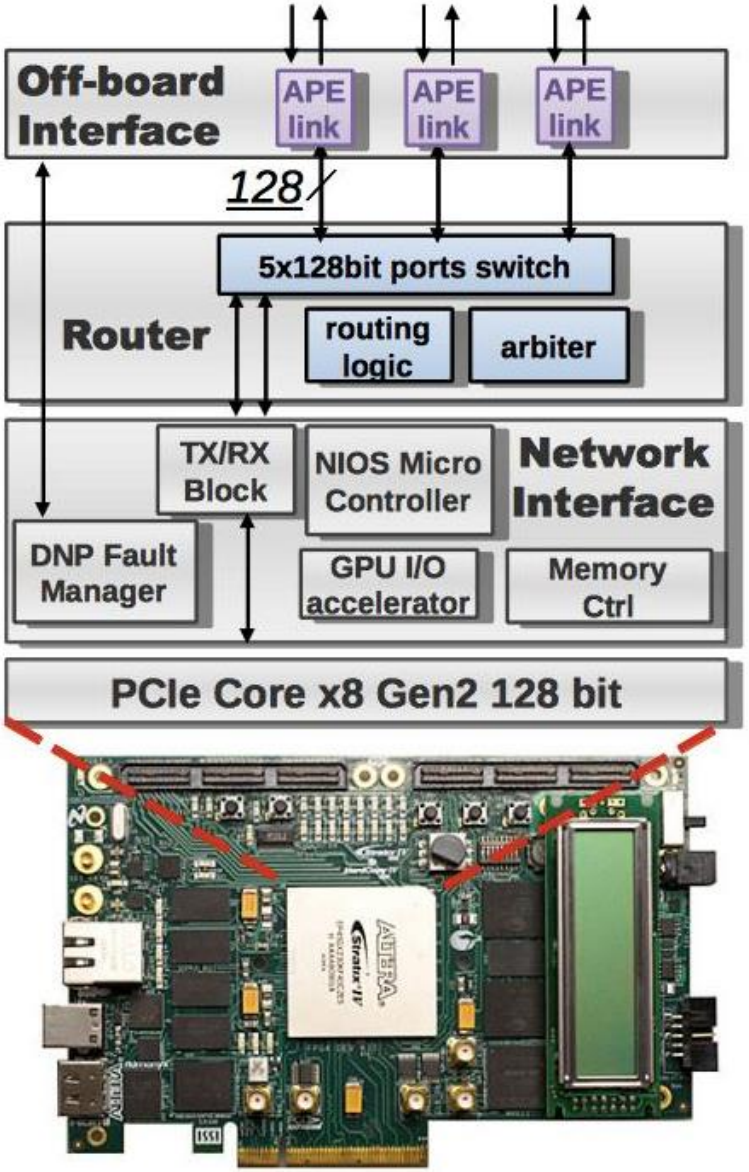
APEnet+ : Architecture Overview



- **Off-board Interface:** manages the node-to-node communication flow over links;

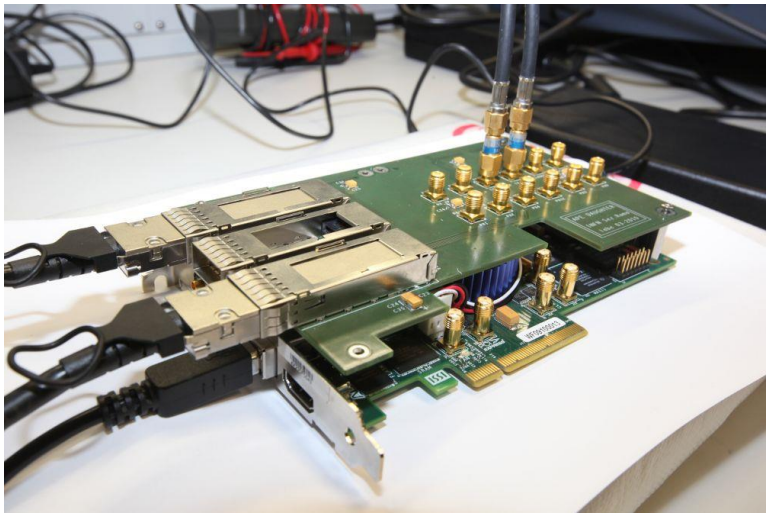
- **Router:** establishes dynamic links among the *N+2* ports (*M* bit width) of the cross-bar switch, managing conflicts on shared resources.
  It applies a deterministic routing policy based on dimension ordering.
  Specialized registers can be written at run-time to choose the coordinates evaluation order (eg first Z is consumed, then Y and eventually X), the arbitration policy (static priority or round-robin) and relative priorities;

APEnet+ modularity: 3 links implementation



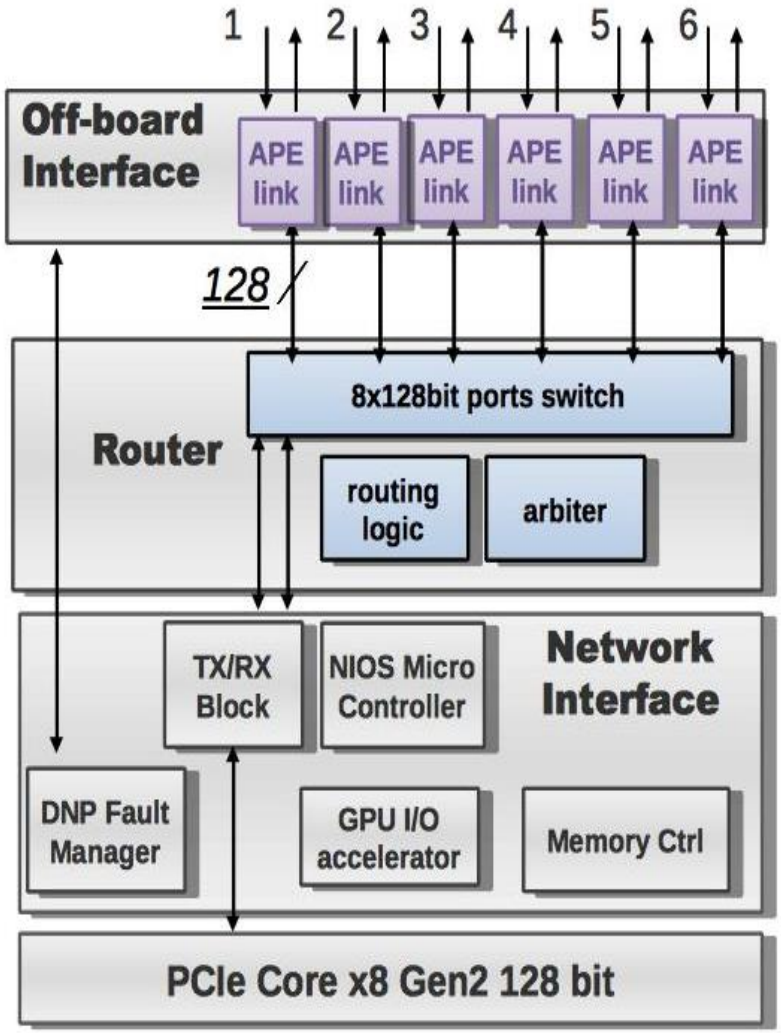
Altera Stratix IV development board

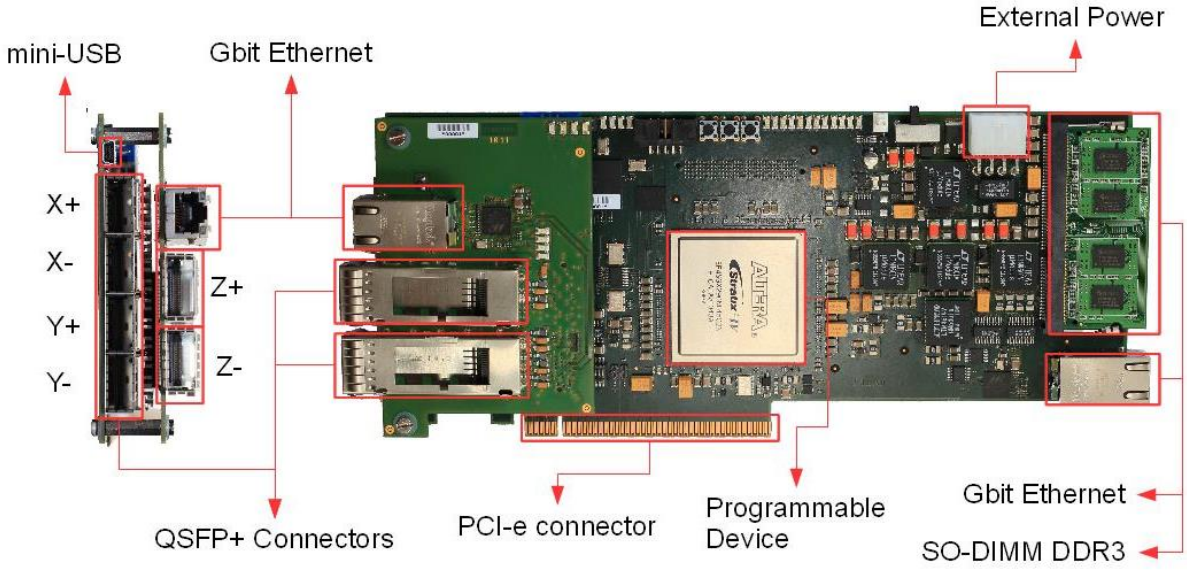
Altera Stratix IV development board + Custom HSMC daughter card



- ✓ Custom daughter card with 3 QSFP and SMA connectors for debug/measures
- ✓ PCIe x8 Gen2 @128 bit
- ✓ Internal bus @128 bit
- ✓ 5x128bit ports switch
- ✓ 3 Altera Custom Transceiver

APEnet+ modularity: 6 links implementation

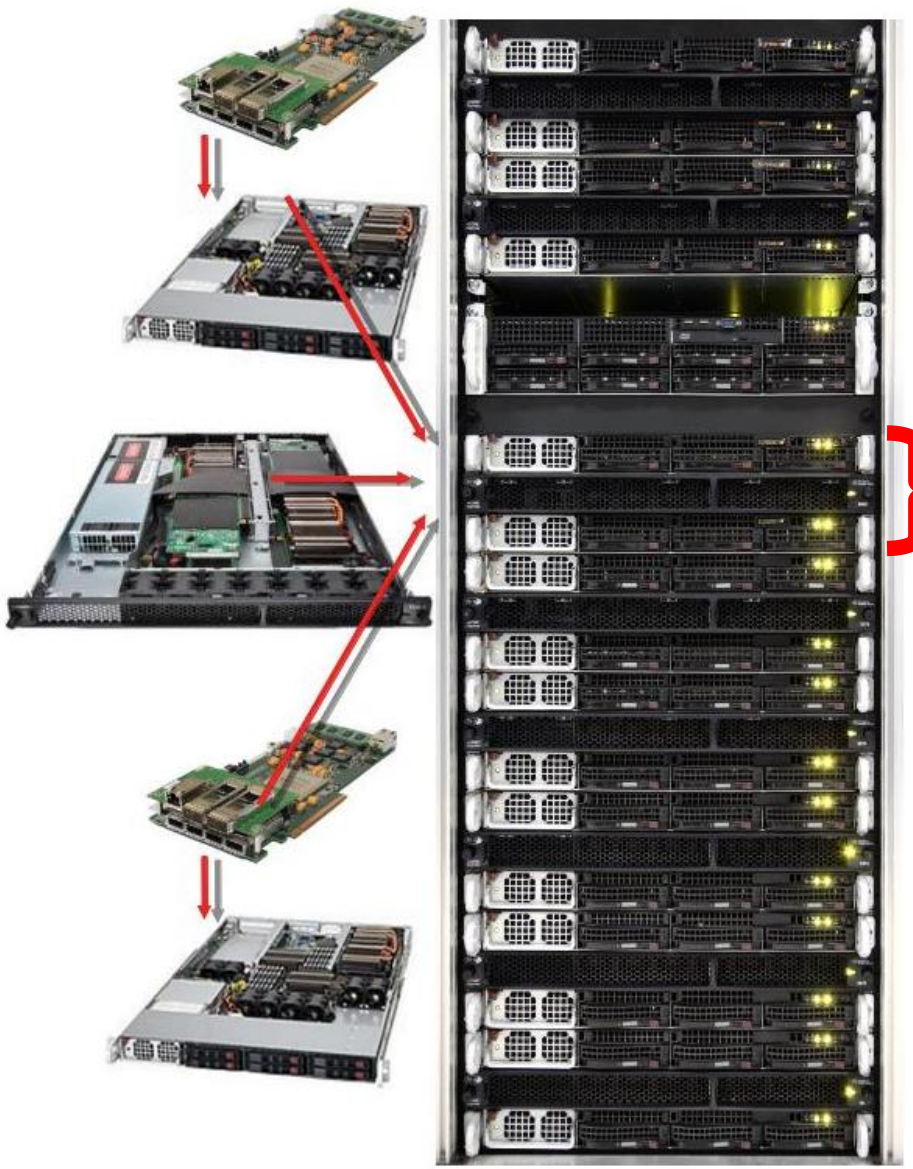




Custom APEnet+ Card:

- ✓ FPGA based (ALTERA EP4SGX290)
- ✓ PCIe X8 Gen2 in X16 slot (peak BW 4+4 GB/s)
- ✓ 6 Full bidirectional 3D torus links (68 Gbps per link)
- ✓ ~ 400 Gbps aggregated raw bandwidth
- ✓ Industry standard QSFP+ cables (40 Gbps)

QUonG: EURETILE HPC platform



16 nodes connected by APEnet+ (4x4x1)

QuonG Hybrid Computing Node:

- ✓ Intel Xeon E5620 double processor
- ✓ 48 GB System Memory
- ✓ 2 M2075 NVIDIA Fermi GPU
- ✓ 1 APEnet+ board
- ✓ 40 Gbps InfiniBand Host Ctrl Adapter

QUonG Elementary Mechanical Unit:

3U Sandwich: (2 Computing Node)

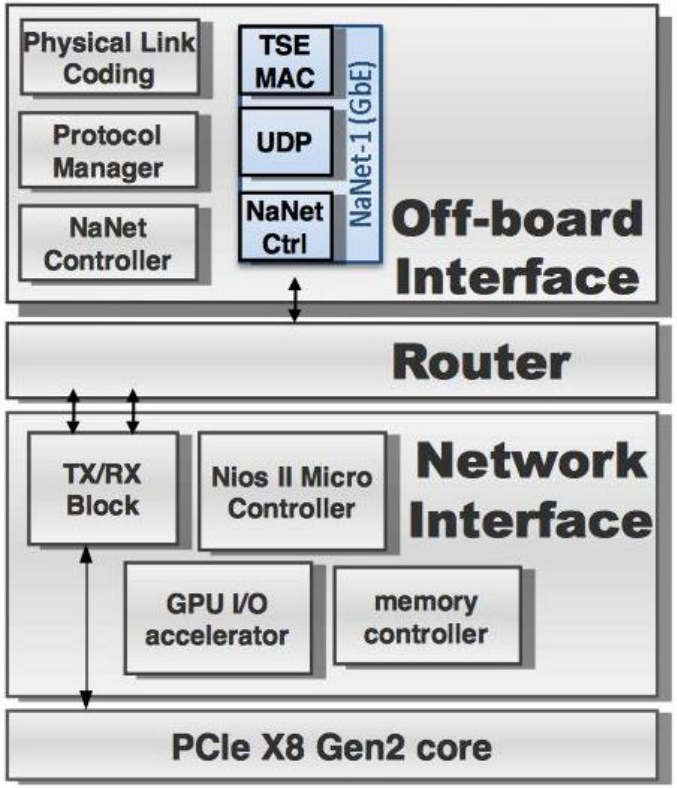
2U Intel dual Xeon servers
1U 4 NVIDIA Tesla M2075 GPU
2 Vertex on the APEnet+ 3D network

Software Environment

- ✓ CentOS 6.4
- ✓ NVIDIA CUDA 4.2 compilation tool
- ✓ OpenMPI and MVAPICH2 MPI available

APEnet+ modularity: NaNet-1

GPU L0 TRIGGER for HEP Experiments (NA62 CERN Experiment)



RO Board-L0 GPU link constraints:

✓ Sustained Bandwidth < 700 MB/s (on GbE links)

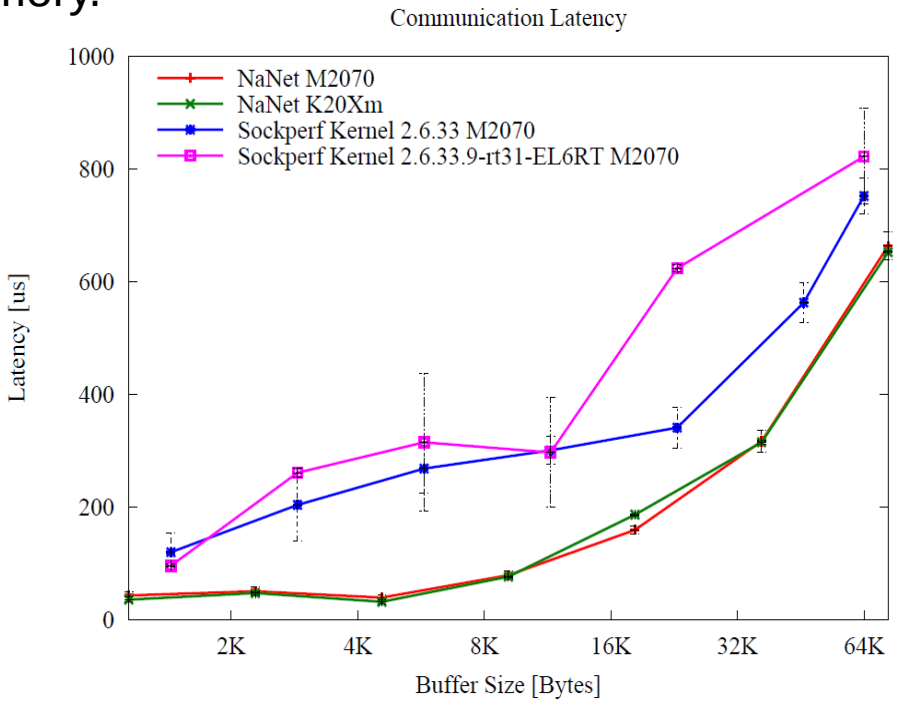
✓ Small and stable latency

Problem:

lower communication latency and its fluctuations. How?

✓ Offloading the CPU from network stack protocol management.

✓ Injecting directly data from the NIC into the GPU(s) memory.

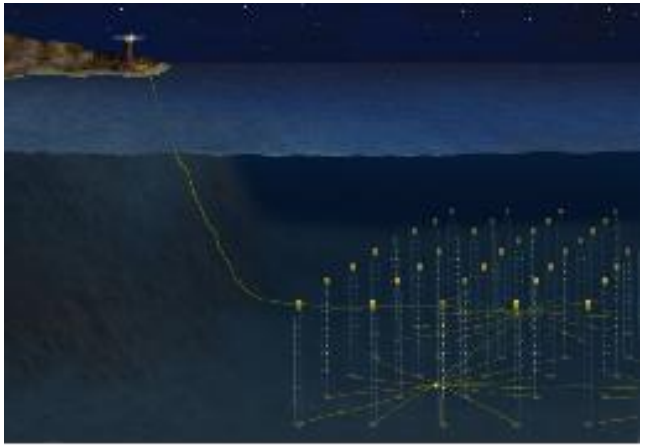


NaNet-1 solution:

APEnet+ NIC with an additional network stack protocol management offloading engine to the logic (UDP Offloading Engine).

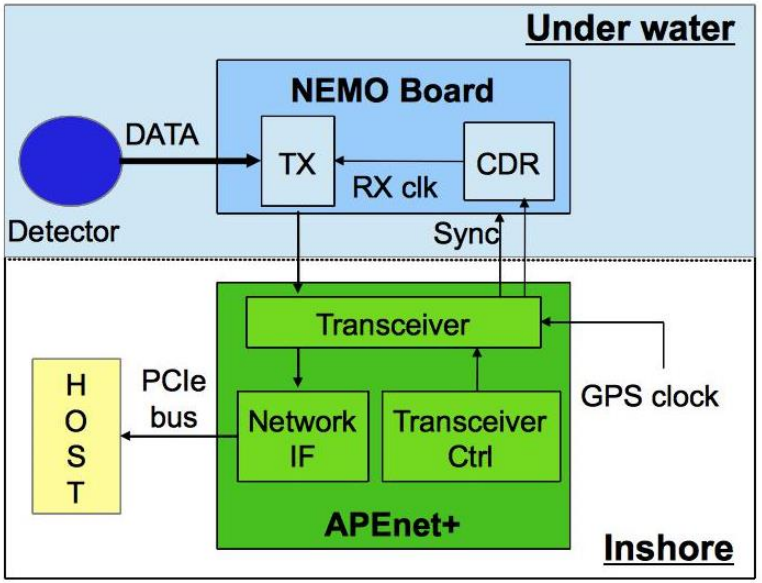
APEnet+ modularity: NaNet^3 implementation (I)

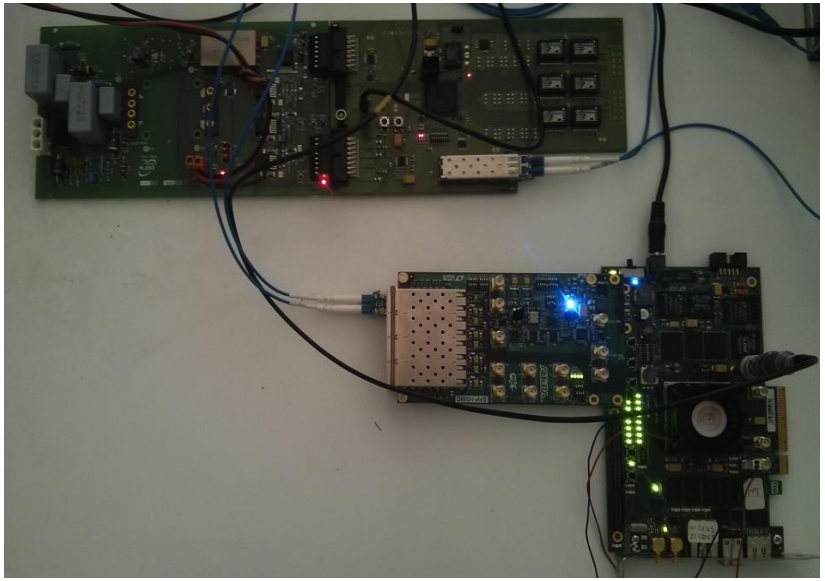
- **NaNet^3** aims to develop and to deploy an European deep-sea research infrastructure, hosting a neutrino telescope with a volume of cubic kilometers at the bottom of the Mediterranean Sea.



Two main challenges:

- ✓ Xilinx and Altera embedded links interoperability
- ✓ Fixed latency links for accurate timing

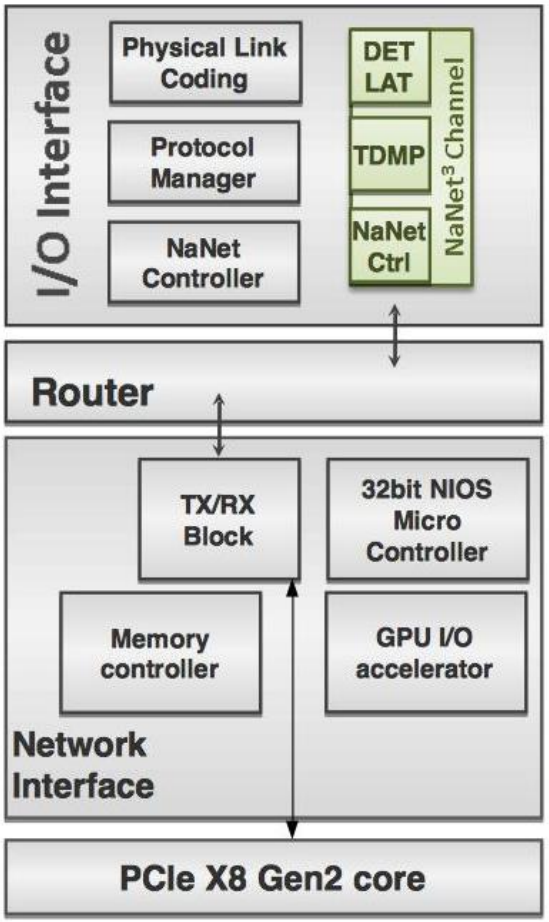




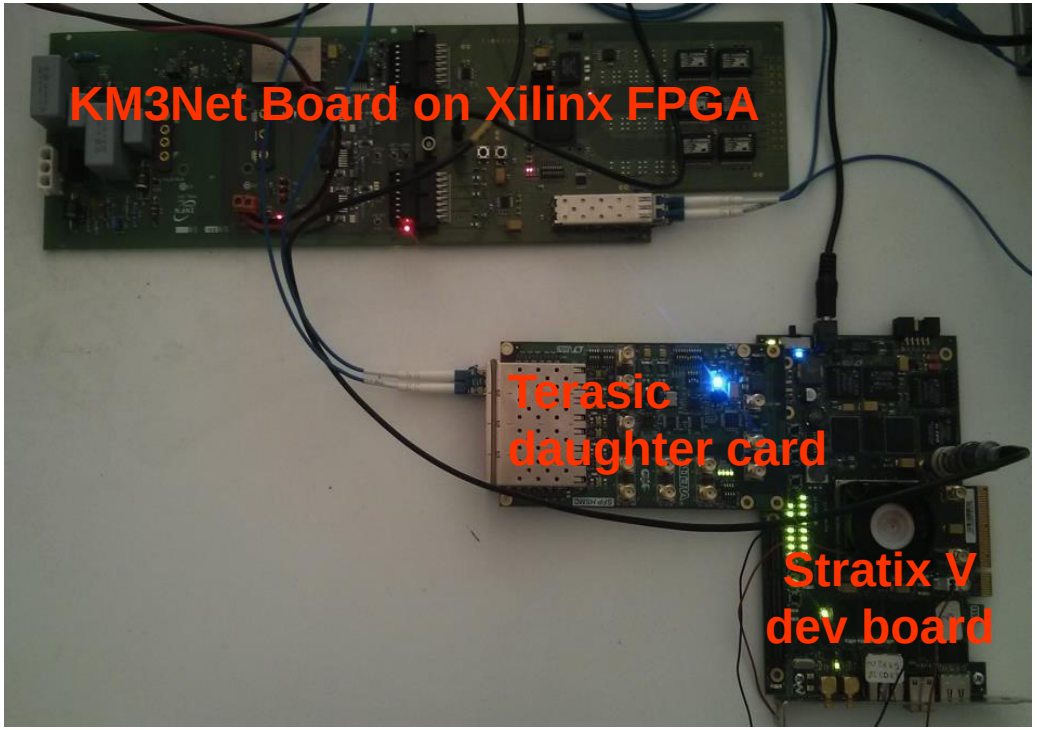
APEnet+ modularity: NaNet³ implementation (II)

System features:

- ✓ APEnet+ transceiver: Altera deterministic latency HIP @250MHz
- ✓ Channel: 2500 Mbps



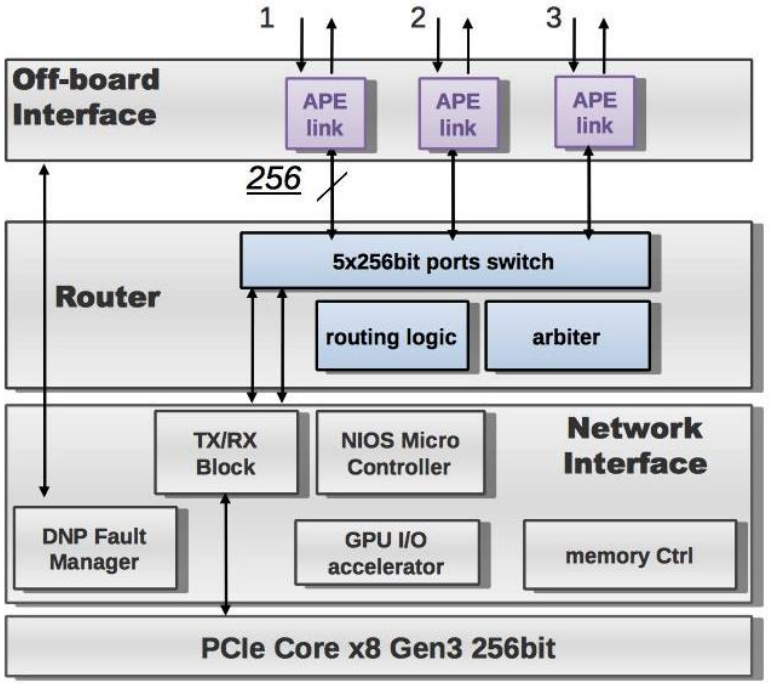
NaNet³ solution:
APEnet+ employed as a low latency, high performance on-shore readout system.



APEnet+ - KM3Net Board testbed

APEnet+ modularity: 3 links implementation on StratixV

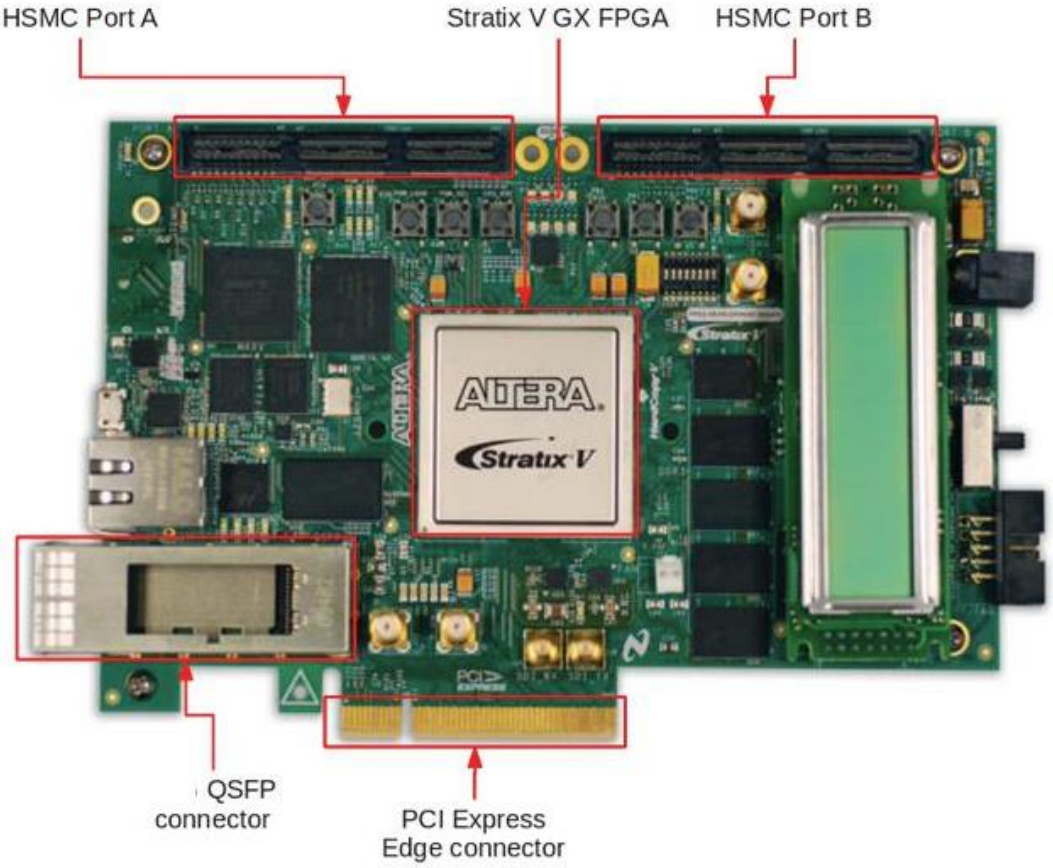
- ✓ 28 nm Stratix V GX FPGA
- ✓ Design under development



- ✓ PCIe x8 Gen3
- ✓ Internal bus width: 256bit @ 250Mhz
- ✓ Preliminary measure with PLDA reference design :
  - Read bandwidth 6.4 GB/s
  - Write bandwidth 5.8 GB/s

- ✓ Bandwidth: 12Gbps per lane Altera Custom transceiver (48Gbps per channel)

Cable	BER	Data Rate
10 m Optical	< 2.36 E-14	11.3 Gbps
1 m Copper	< 1.10 E-13	10 Gbps



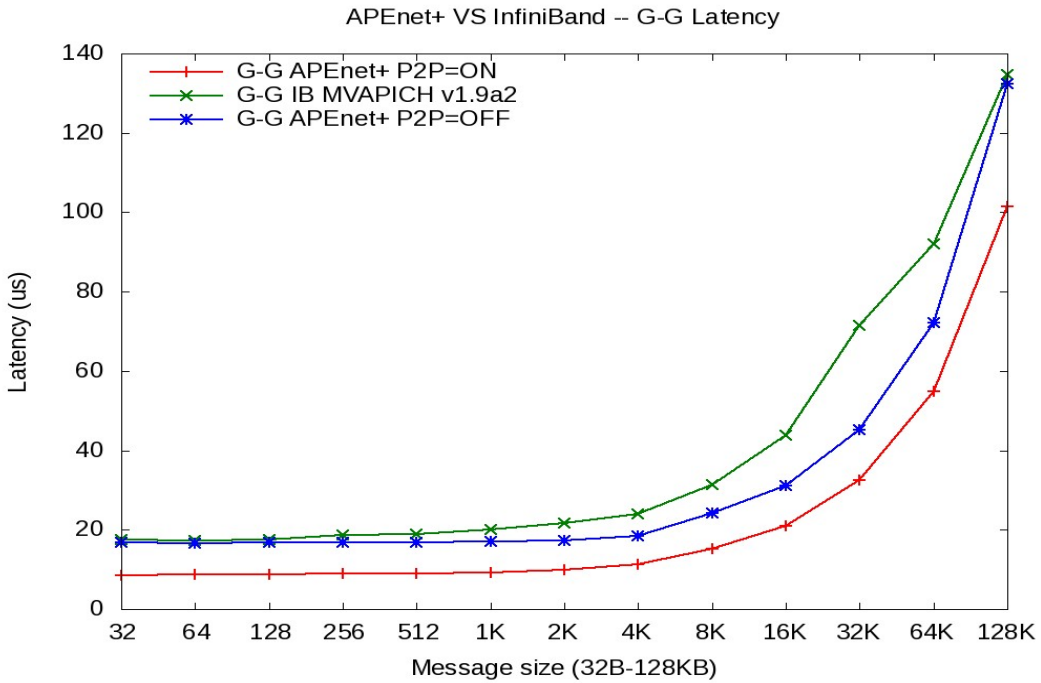
APEnet+ : Overview of resource consumption

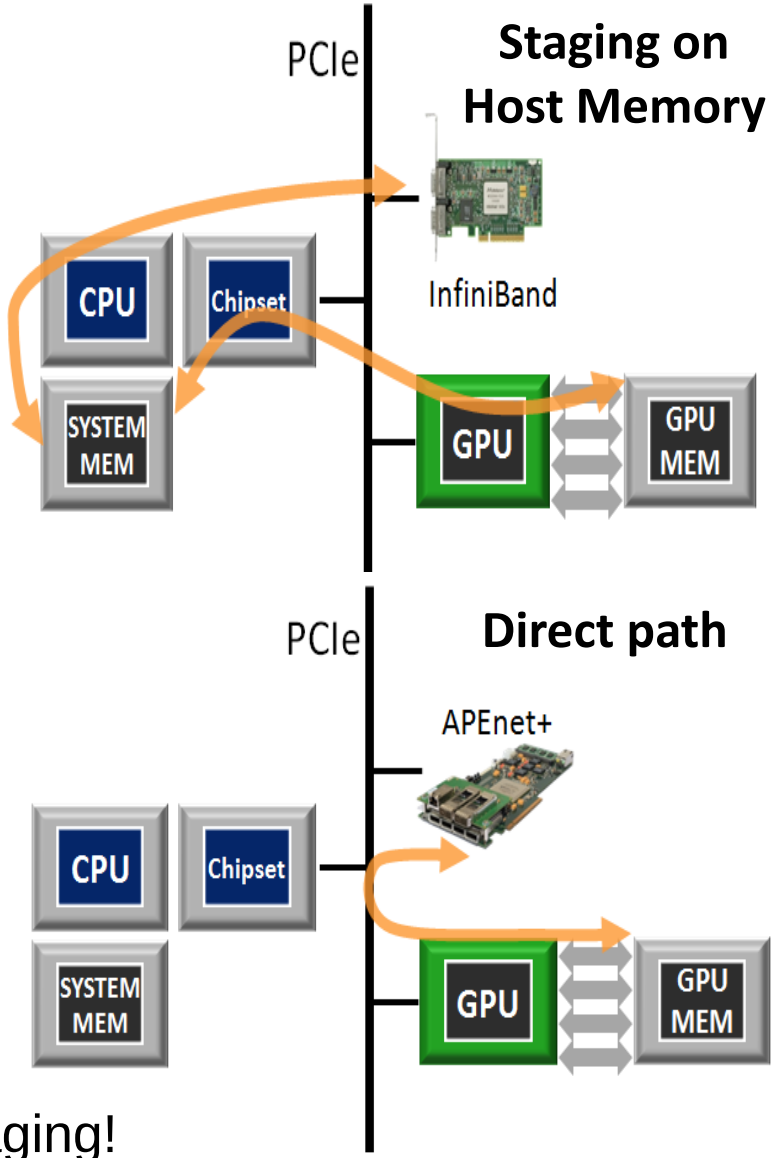
Project	Board	Comb. ALUT	Register	Memory [MB]
NaNet-1	EP4SGX230	65216 (36%)	65373 (36%)	1.07 (59%)
APEnet+	EP4SGX290	82068 (35%)	71044 (30%)	1.11 (64%)
APEnet+ v5	5SGXEA7K2	55092 (26%)	56283 (12%)	1.29 (20%)

Logic Block	Comb. ALUT	Register	Memory [MB]
PCIe	7268	8042	0.001
Host TX block	1530	1628	0.105
GPU TX block	1938	1860	0.087
RX block	15198	15040	0.148
Nios II	18213	19431	0.537
LOFAMO	251	267	—
HARDWARE CUSTOMIZATION			
Router variation per Inter-port	~ 1000	~ 450	—
NaNet-1 Custom Logic	543	575	0.066
Router Inter-Port	430	270	—
Router Intra-Port	280	150	—
Off-board Channel	4590	4650	0.044

APEnet+ : NVIDIA GPU peer to peer support

Peer-to-peer access allows exchanging data between NIC and GPUs directly through the PCIe bus without staging on Host Memory, with latency reduction.



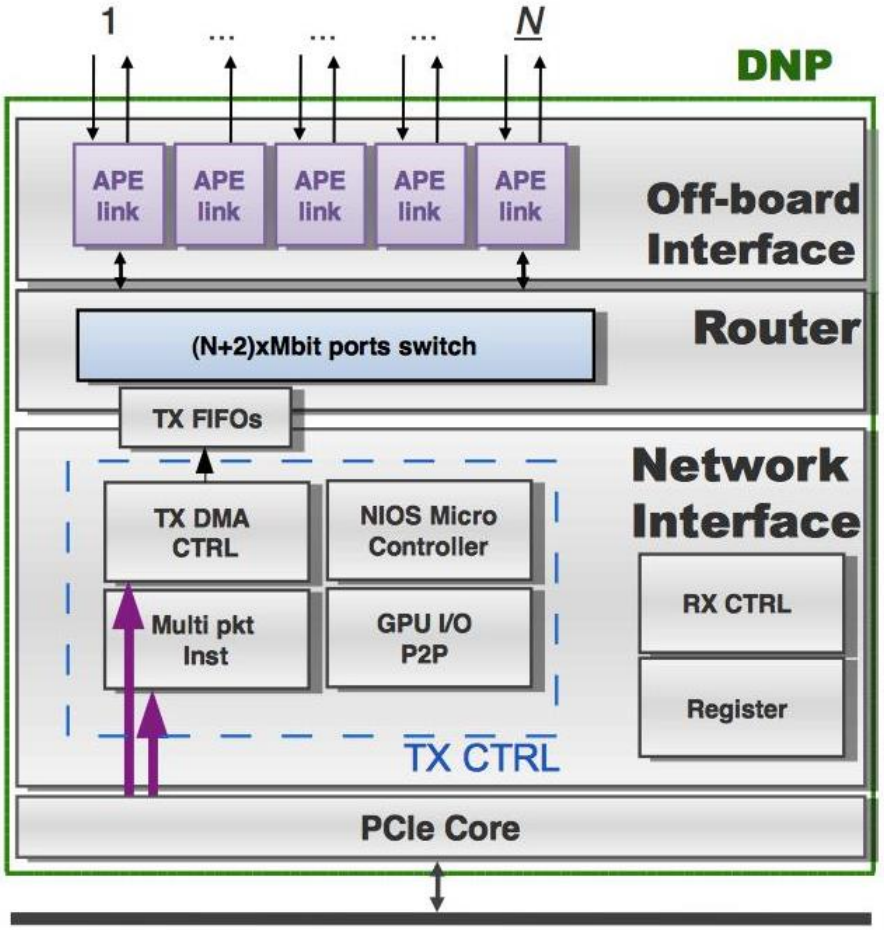


- ✓ Effective for small buffer sizes (up to 128 KB), latency with P2P it is 50% lower than with staging!
- ✓ First-of-its-kind feature for a non-NVIDIA device (2012)

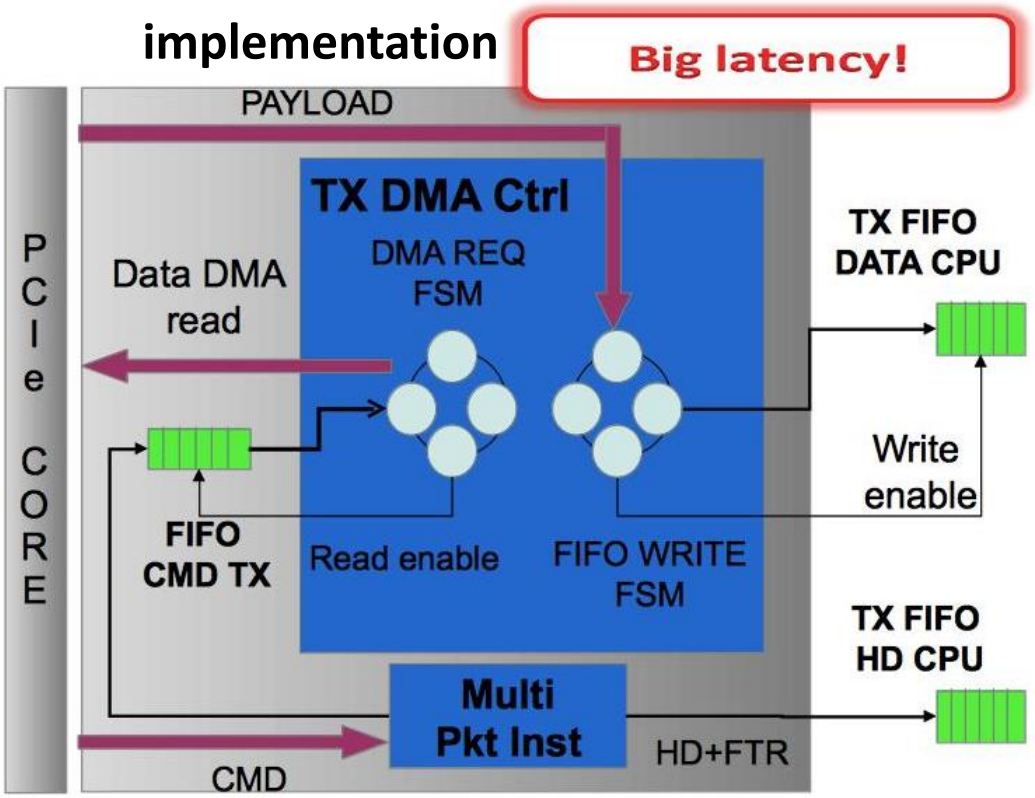
Design improvement : Tx double DMA (I)

DNP Tx block handles transfers from host/GPU through the PCIe port, forwarding the data stream to the TX FIFOs.

Tx DMA Ctrl block instantiates DMA read transaction to load the packet's payload.



One DMA channel implementation

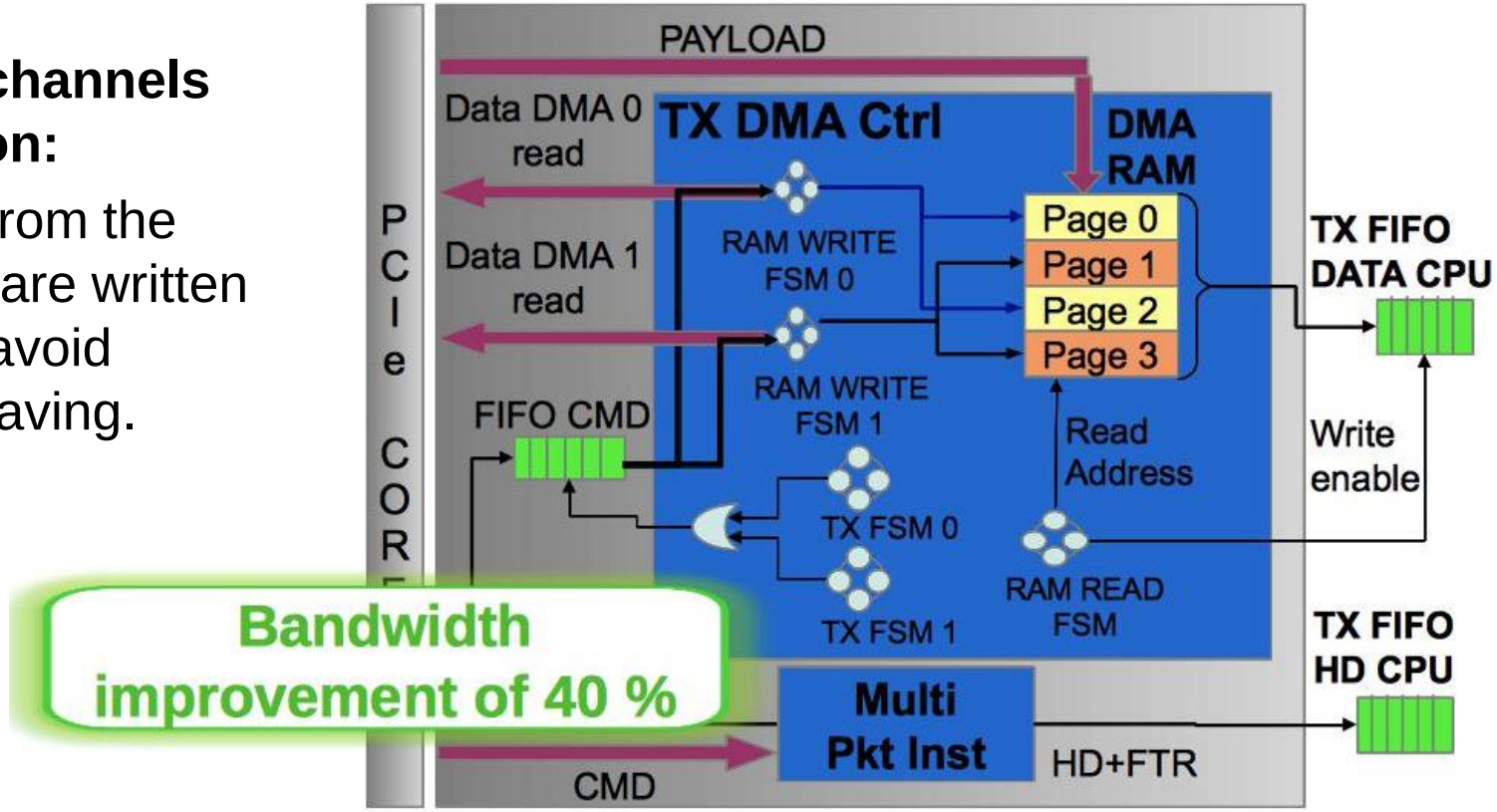


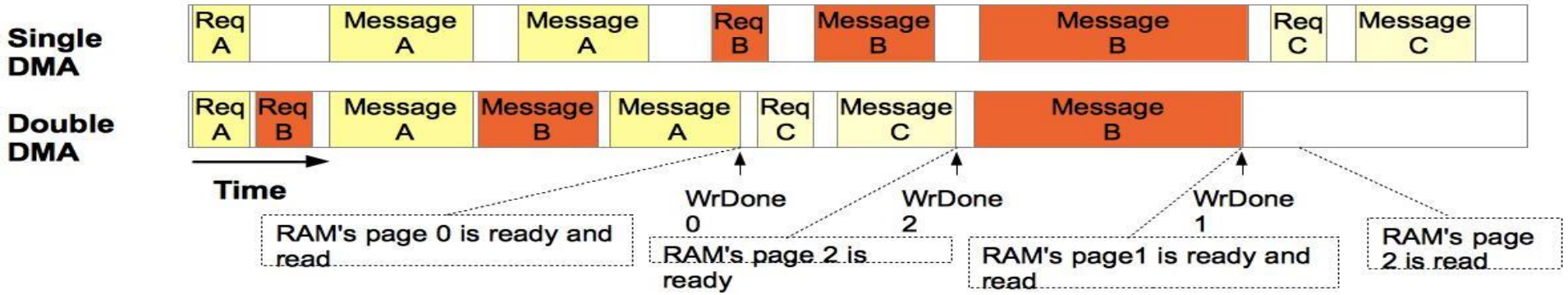
Big latency between two consecutive DMA read requests on the PCIe bus!

Design improvement : Tx double DMA (II)

Double DMA channels implementation:

data received from the PCIe interface are written into a RAM to avoid payload interleaving.



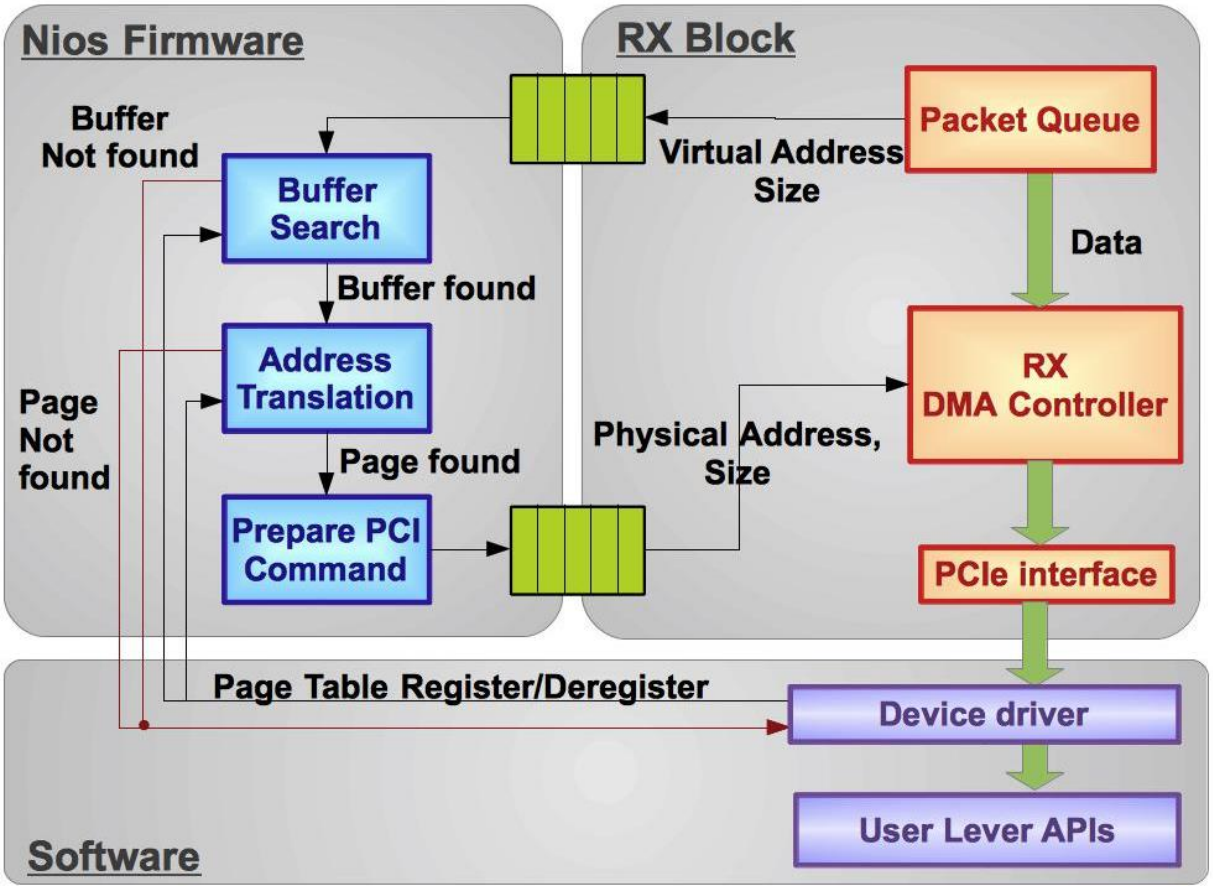


Design improvement : Rx speed up- memory management (I)

Virtual memory management : to carry out the RDMA protocol with no remote host CPU or OS involvement

How? Virtual to physical address translation to instruct the local DMA engine.

First solution:
Altera Nios II microcontroller (clocked at 200 MHz) manages address translation



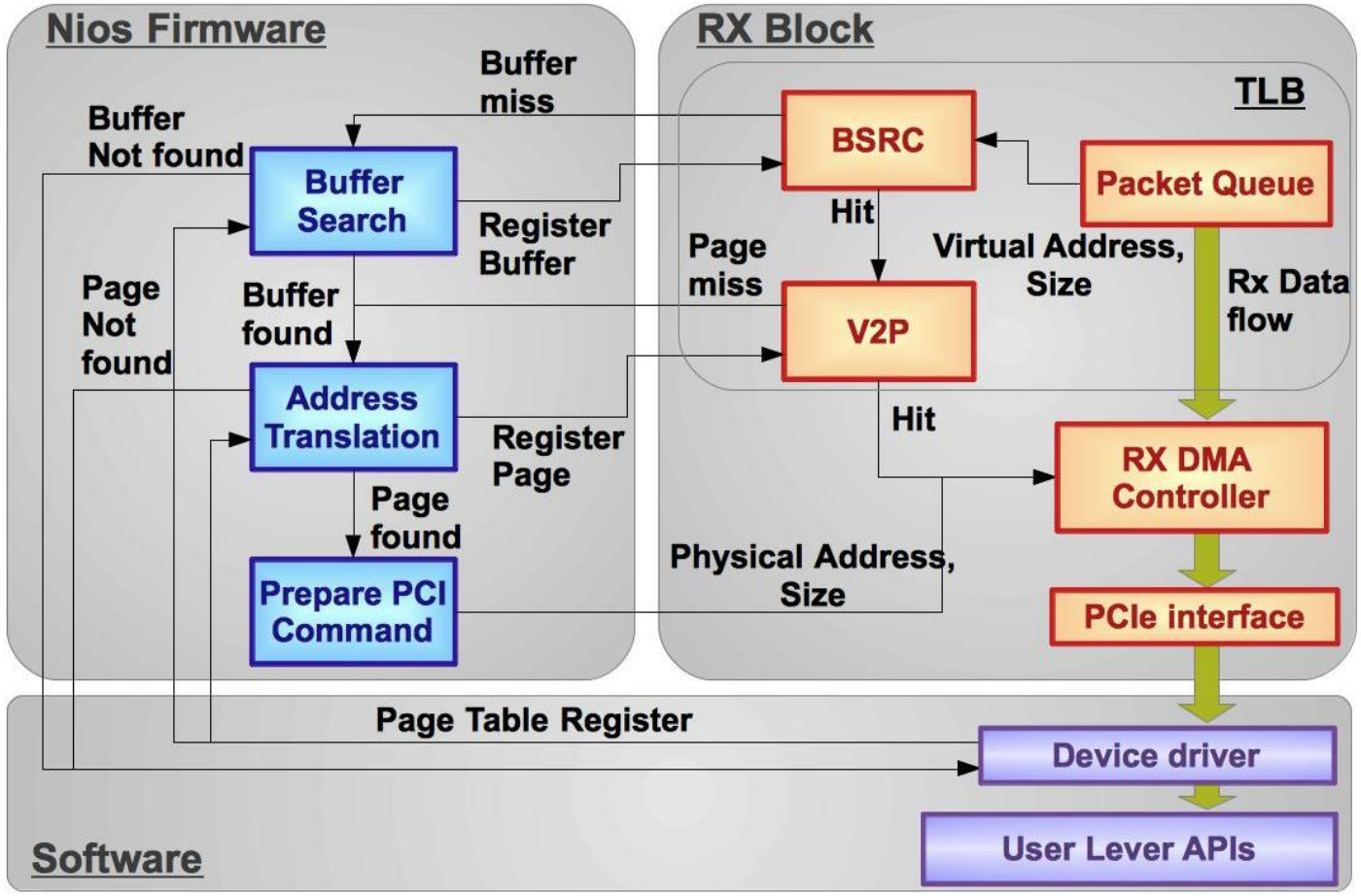
Severe performance penalty!
400 clock cycles to perform a single virtual address processing

Design improvement : Rx speed up- memory management (II)

Second solution: TLB implementation

TLB (Translation Lookaside Buffer) :

- ✓associative cache;
- ✓limited amount of entries to perform memory management tasks;
- ✓reducing processing time in case of a cached translation (Hit);
- ✓forwarding the operation to the Nios in case of a 'miss';

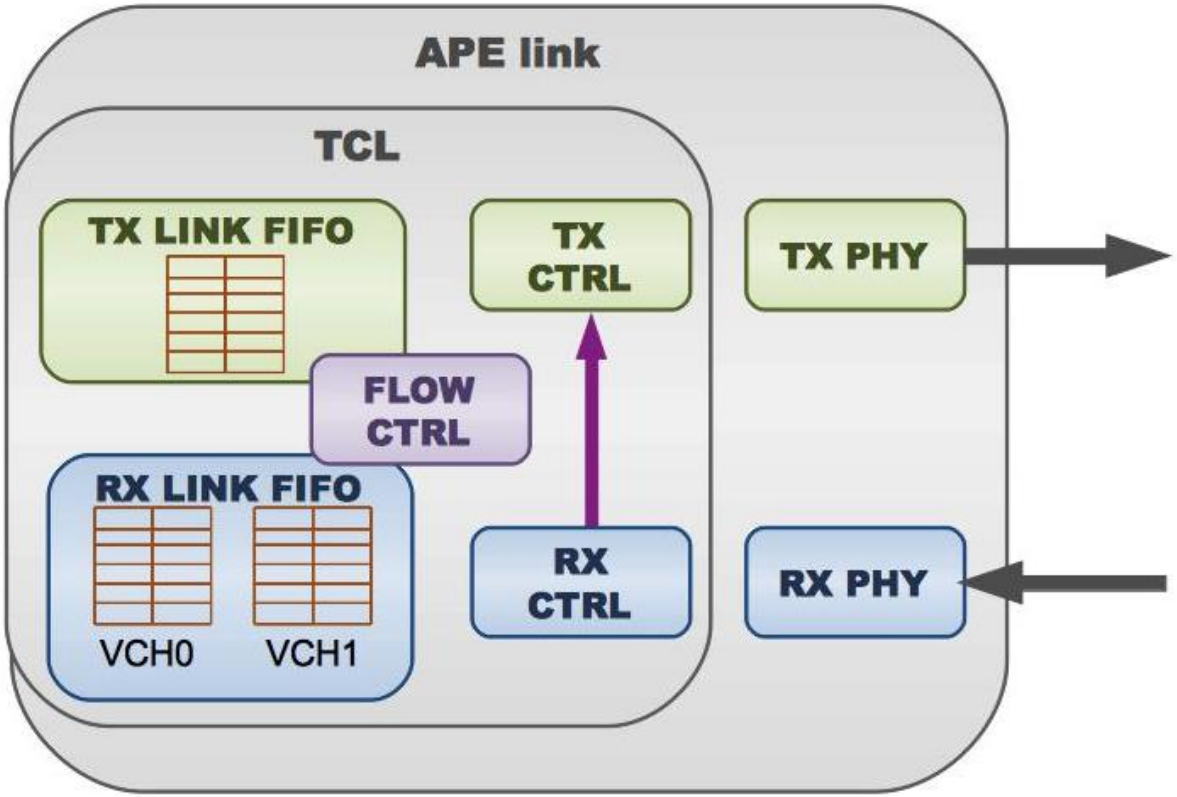


31 clock cycles (@250 MHz) to complete buffer look-up and address translation in case of hit!

Design improvement : Off-Board Interface (I)

Transmission Control Logic (TCL)

- ✓ Manages the data flow: Sends Credits (occupancy of the Receiving FIFOs) to avoid RX FIFOs overflow (Virtual cut through is implemented);
- ✓ Encapsulates packets into a light, low-level word stuffing protocol;
- ✓ Detects trasmission error via CRC;
- ✓ Implements virtual channels to guarantee deadlock free routing;
- ✓ Sends Diagnostic Messages for Fault-Awareness;



Design improvement : Off-Board Interface (II)

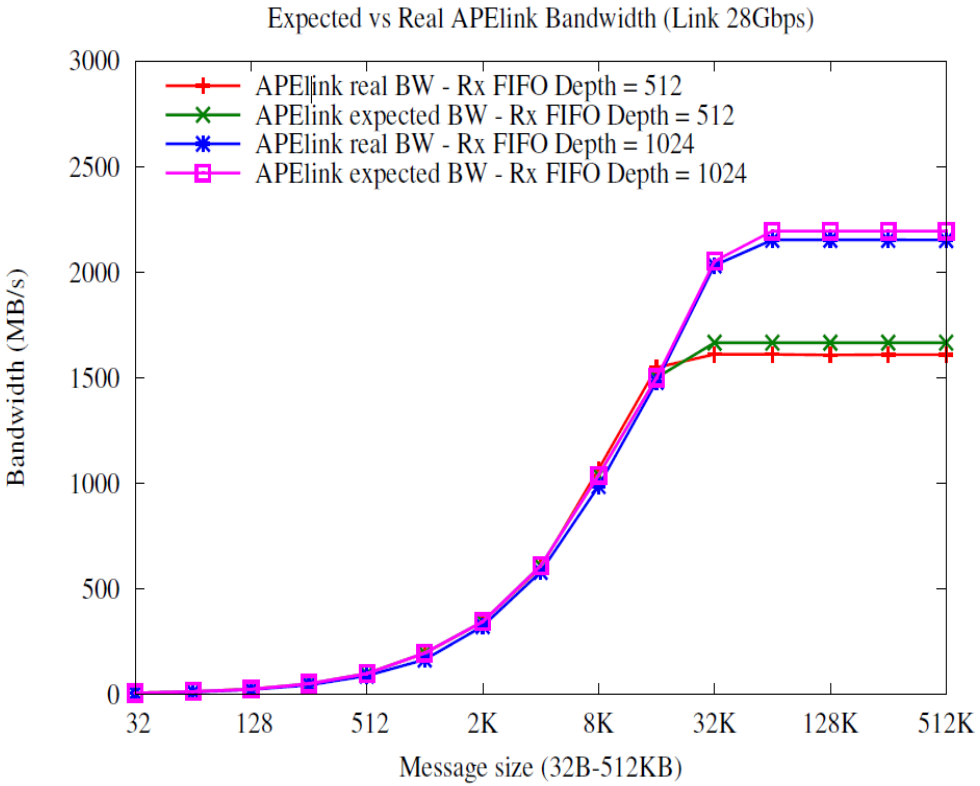
Three Efficiency Factors:

E_1: related to the adopted protocol

E_2 : status information (receiving FIFO occupancy and diagnostic messages)

E_3 : data flow management

FIFO DEPTH	E_3	E_T	BW_L^{MAX}@28Gbps	BW_L^{MAX}@34Gbps
512	0.638	0.595	1666 MB/s	2023 MB/s
1024	0.841	0.784	2195 MB/s	2665 MB/s
2048	0.925	0.862	2414 MB/s	2931 MB/s
4096	0.964	0.898	2514 MB/s	3060 MB/s



Bandwidth (MB/s)

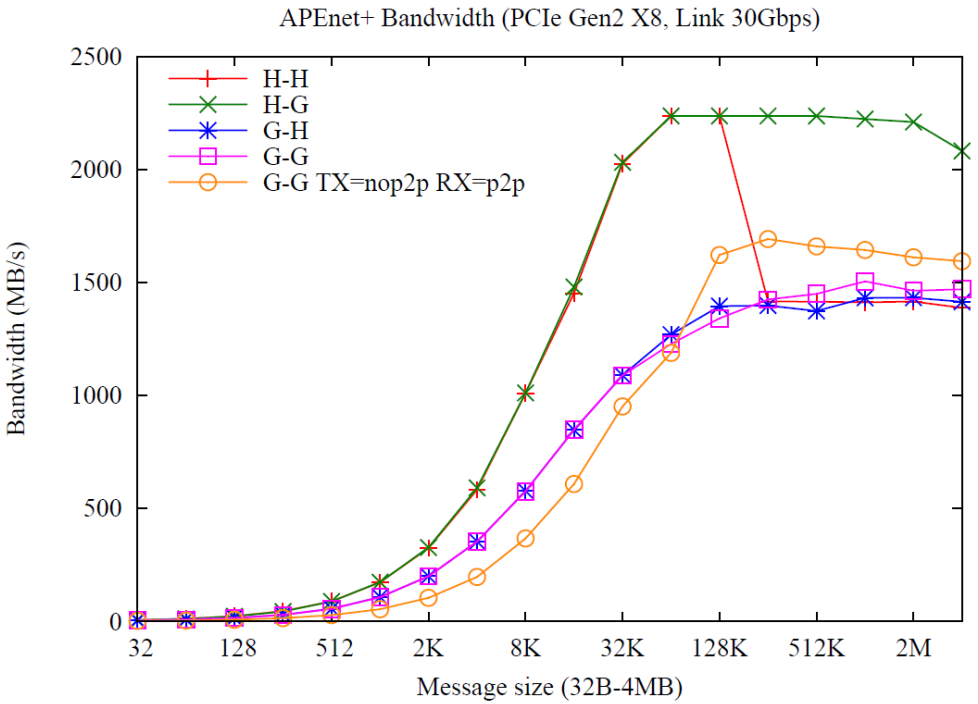
E_1 and **E_2** constant (once protocol is defined)

E_1x E_2 =0,93

E_3 depends on RX FIFO depth

E_T = E_1 x E_2 x E_3

APEnet+ : High Bandwidth

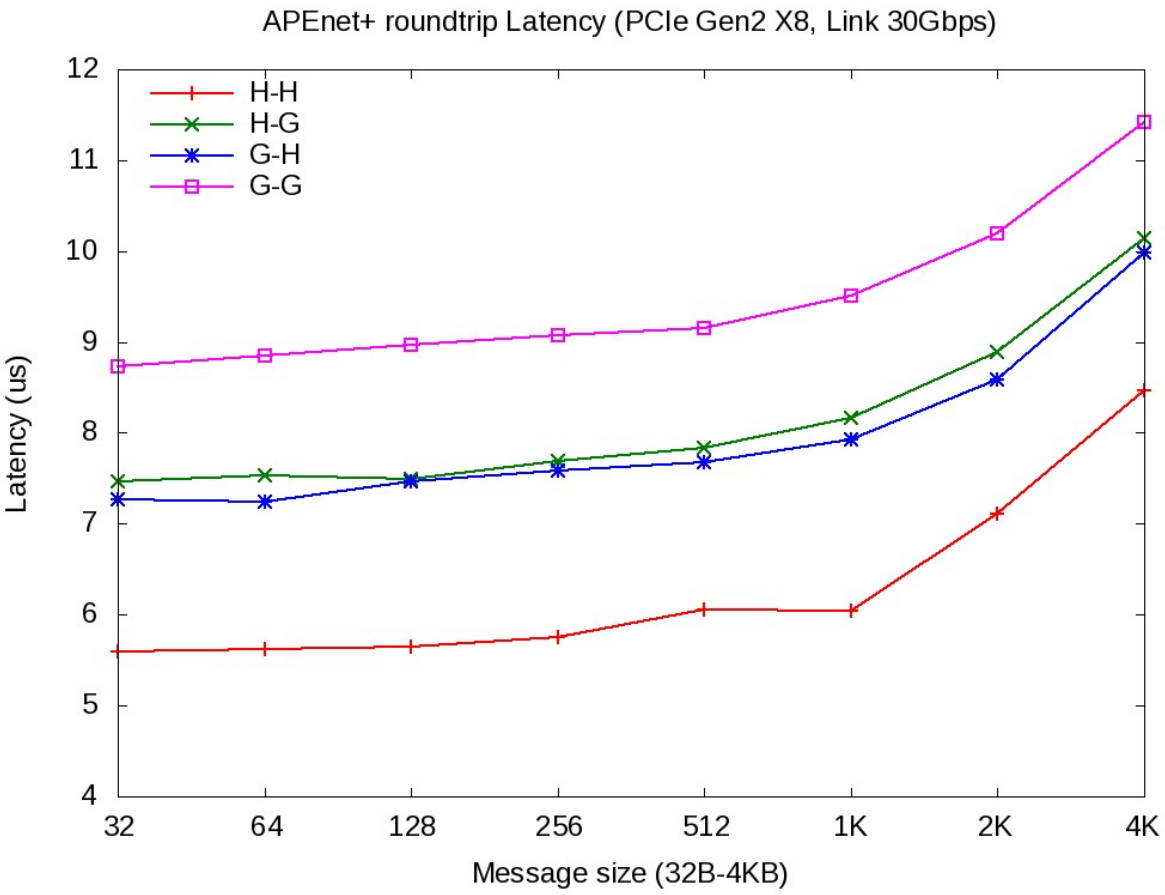


- ✓ CPU Memory Read Bandwidth ~2.4GB/s
- ✓ CPU Memory Write Bandwidth~ 1.5 GB/s
  - ✓ GPU Memory Read Bandwidth ~1.5GB/s
  - ✓ GPU Memory Write Bandwidth ~2.2 GB/s

Test	Bandwidth	GPU/method	Nios II active tasks
Host mem read	2.4 GB/s		none
GPU mem read	1.5 GB/s	Fermi/P2P	GPU_P2P_TX
GPU mem read	150 MB/s	Fermi/BAR1	GPU_P2P_TX
GPU mem read	1.6 GB/s	Kepler/P2P	GPU_P2P_TX
GPU mem read	1.6 GB/s	Kepler/BAR1	GPU_P2P_TX

APEnet+ : Low latency

- ✓ The latency is estimated as the half round-trip time in a ping-pong test
- ✓ 8-10 us G-G latency
- ✓ No optimization for small packet size



Systemic fault-awareness (LO|FA|MO IP)

FAULT AWARENESS

✓first step to address the *fault tolerance problem*

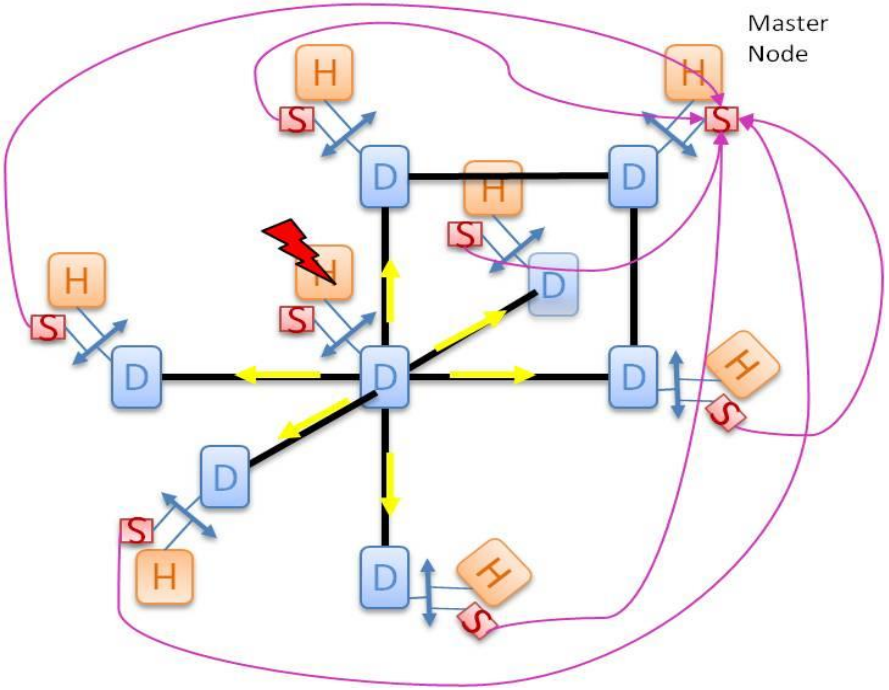
✓The system must be aware of faults occurring to subcomponents (systemic)

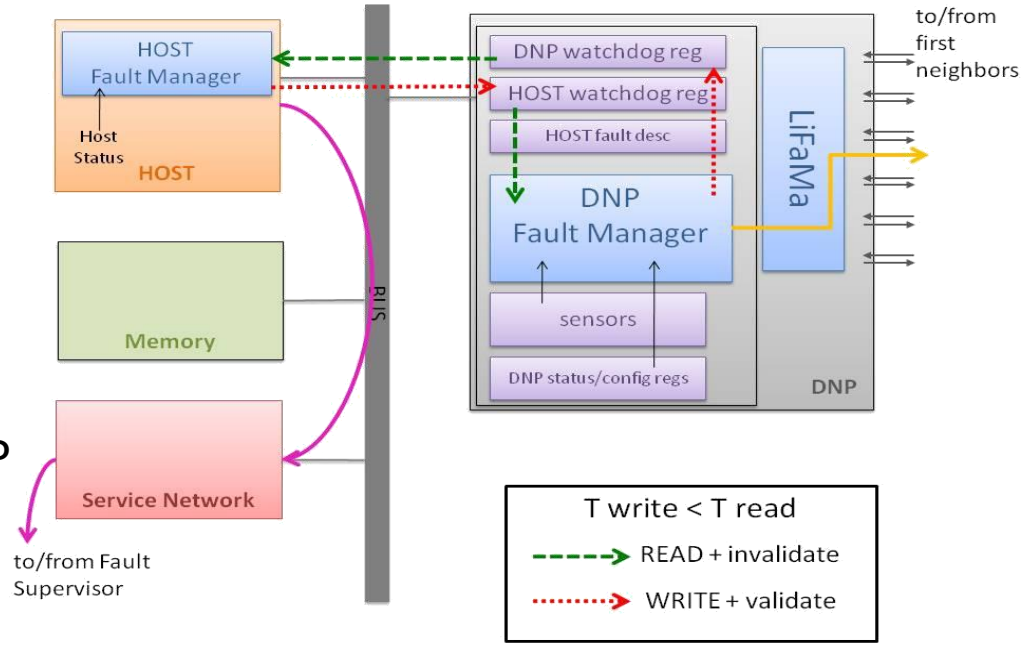
LOFAMO

HW-SW approach based on:

✓Mutual watchdog between Host and DNP

✓3D network and Service network





In case of fault there is always a path for diagnostic messages to reach a master node, which becomes able to make decisions and apply actions (e.g. Restart after checkpoint, task migration,...)

Diagnostic messages over the 3D net are hidden in the channel protocol: they do not impact on APEnet+ performance

The People Involved

Alessandro Lonardo

Pier Stanislao Paolucci

Davide Rossetti

Piero Vicini

Roberto Ammendola

Andrea Biagioni

Ottorino Frezza

Francesca Lo Cicero

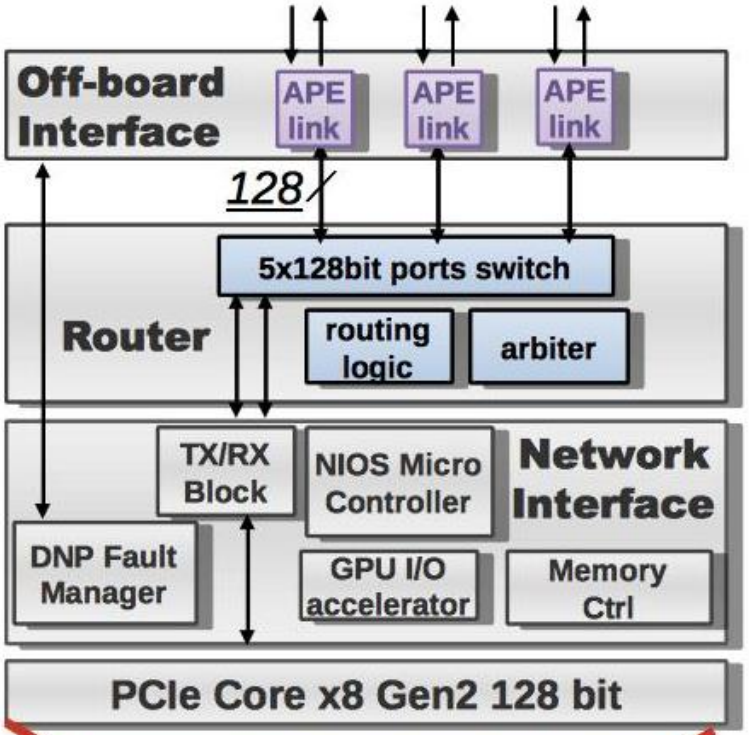
Francesco Simula

Laura Tosoratto

Thank you!

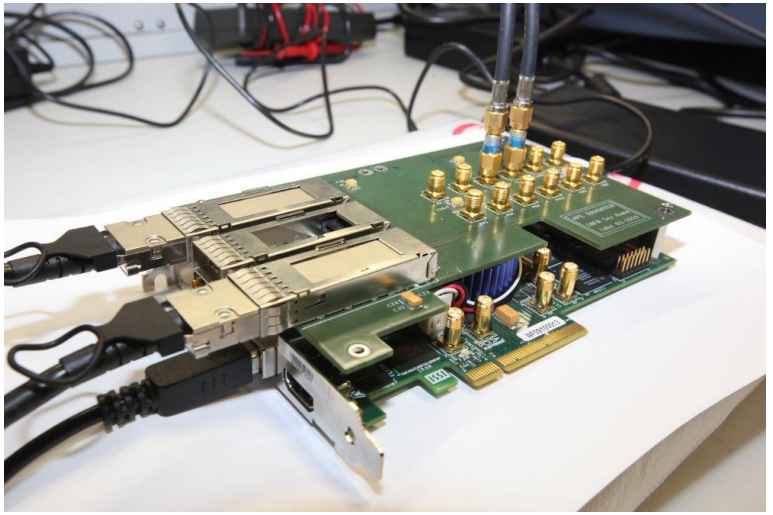
Backup slides

APEnet+ modularity: 3 links implementation



Altera Stratix IV development board

Altera Stratix IV development board + Custom HSMC daughter card



✓3 Altera Custom Transceiver @76.8Gbit aggregate bandwith

✓3.5Gbit/lane (@270 MHz)