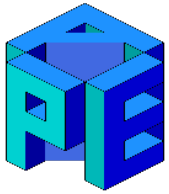
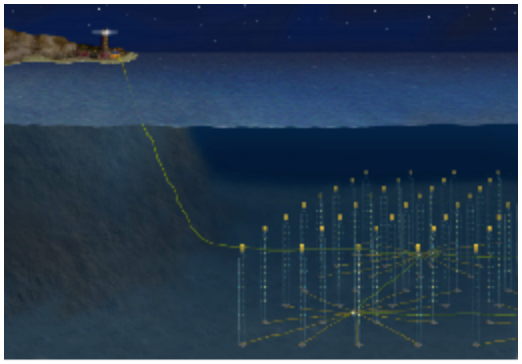


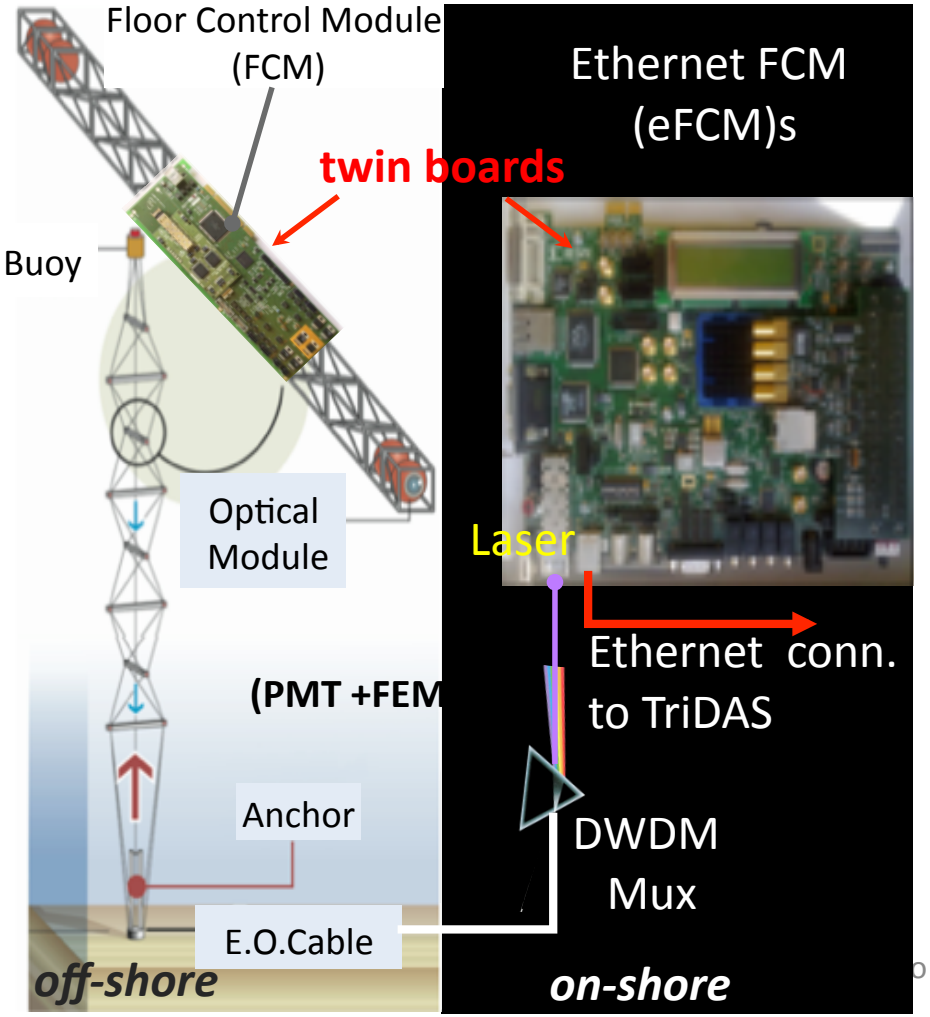
Stato della scheda NaNet^3

Alessandro Lonardo
INFN - Sezione di Roma

KM3Net-IT Experiment

- **KM3Net-IT**: an European deep-sea research infrastructure, hosting a neutrino telescope with a volume of cubic kilometer at the bottom of the Mediterranean Sea.

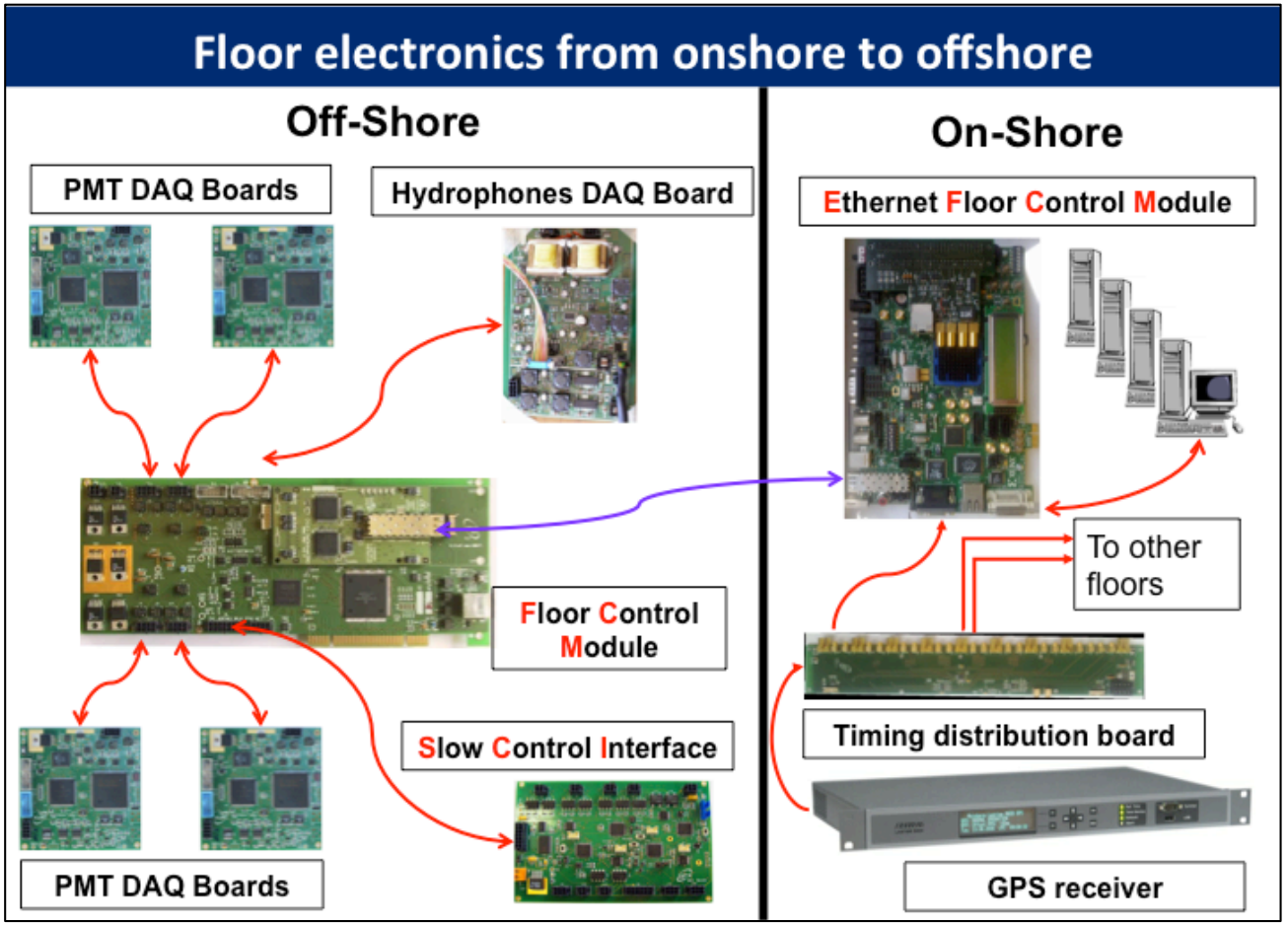




- "**Phase-2 tower**" composed of:
  - 8 floors
  - 8 m bars,
  - vertical dist. = 40 m, Htot = 450 m
  - 32 OM, 18 hydrophones
  - oceanographic instrumentation
  - 2 twinned (off-shore/on-shore) FCM boards per floor.

KM3Net-IT experiment: read-out

- Current **read-out system** employs a huge number, NO state-of-the-art components
  - 2.5 Gbps (800 Mbps payload) optical link
  - 2 twinned FCM boards per floor
  - Many PCs for HW read-out hosting



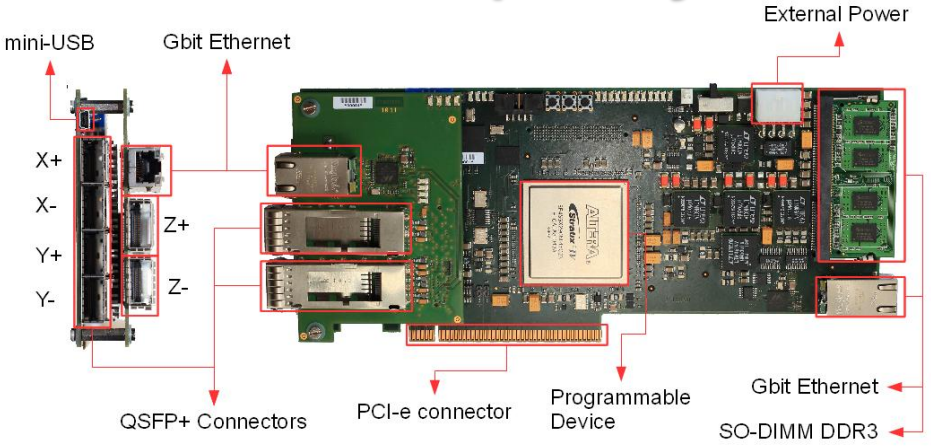
F. Simeone
IEEE/NSS 2011

- When scaling to KM3 many cost/size/power/reliability issues!!!

APEnet+: a 3D NIC for HPC with GPUdirect P2P capability

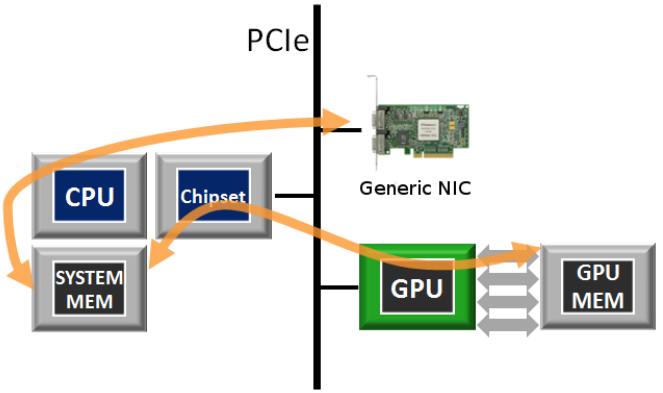
APEnet+ Card:

- FPGA based (ALTERA EP4SGX290)
- PCI Express x16 slot, x8 gen2 signaling
- Fully bidir 3D torus links, 34 Gbps/channel



APEnet+ Logic:

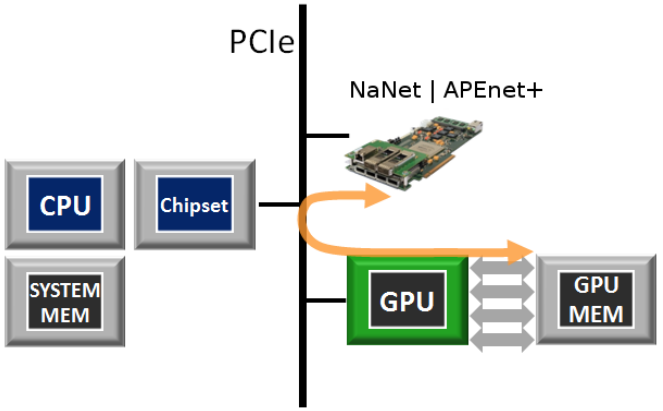
- Network Interface
  - NIOS II 32 bit uP
  - RDMA engine
  - Virtual Memory Management support
  - GPU I/O accelerator.



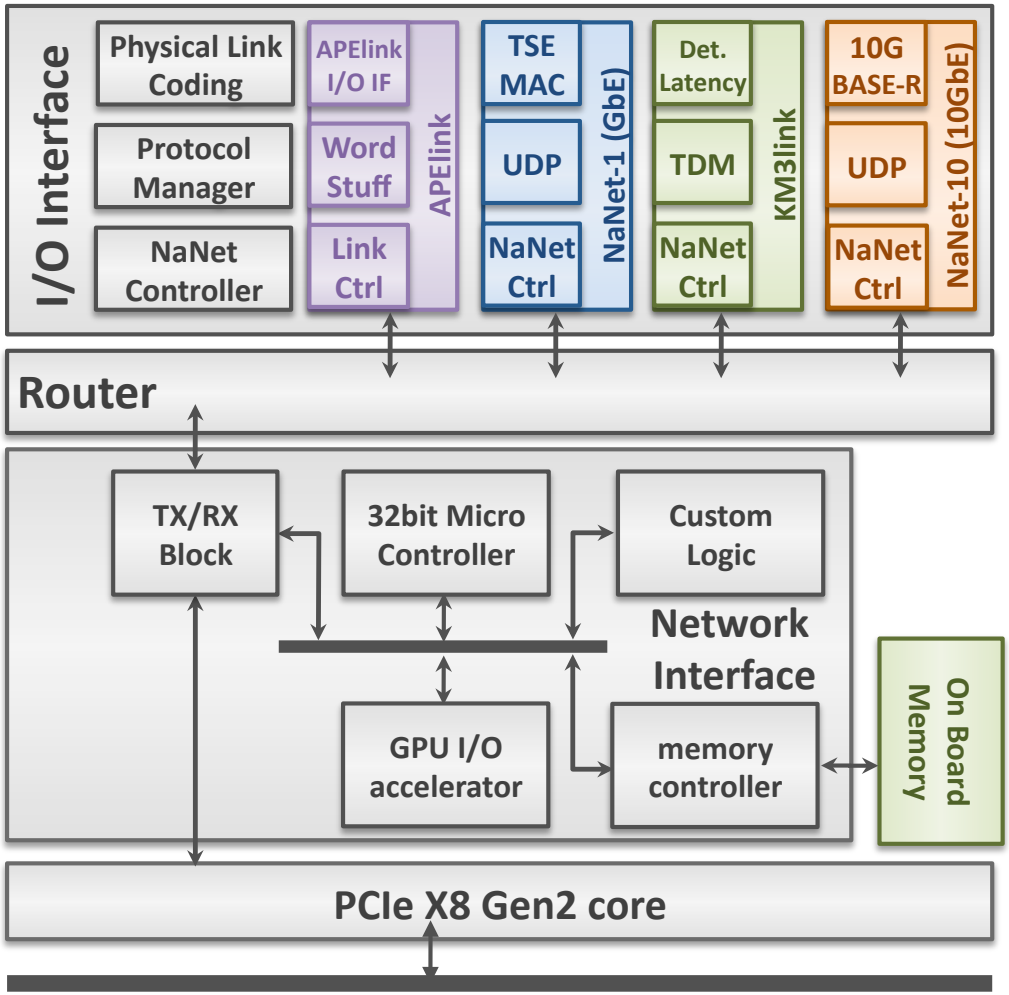
GPUDirect P2P

- PCIe P2P protocol between NVIDIA Fermi/ Kepler devices and APEnet+
- GPUDirect P2P allows direct data exchange on the PCIe bus with no CPU involvement

  → *Latency reduction for small messages*



NaNet



- Multistandard support of I/O interface
  - Off-the-Shelf: 1-GbE, 10-GbE (work in progress).
  - Custom: APElink (34gbps/QSFP), KM3link (work in progress).
- GPU I/O accelerator
  - GPUDirect v2 (custom NVP2P)
  - GPUDirect RDMA
- CPU offload – NO OS Jitter
  - Transport layer off-loading module
  - Virtual Memory Management
- NaNet Controller: protocol adaptation between on-board and off-board protocol
- Altera NIOS II microcontroller
- Custom Logic (application-specific task, e.g. compression)
- DDR3 memory controller
- PCIe Gen2 X8 host interface

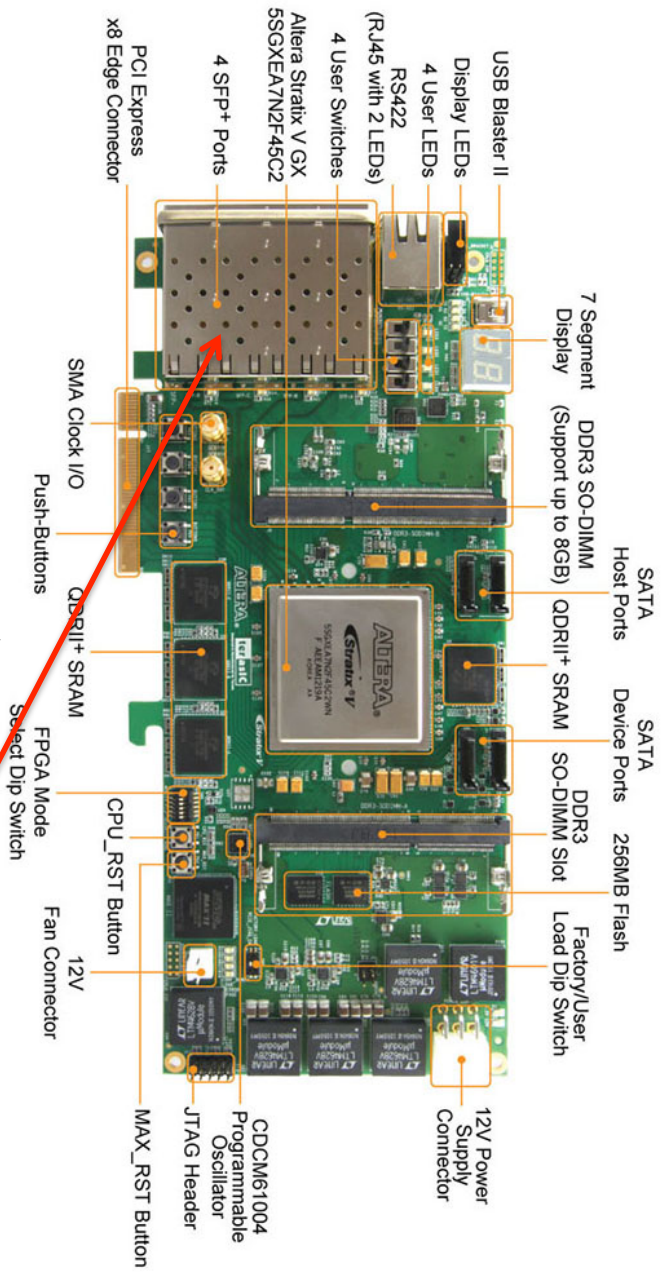
Why NaNet^3?

NaNet^3 specifications for future enhanced read-out system

- more (4) channels/PCIe board
  - i.e. less PC, less read-out boards,…. ✔
- Link speed-up to 10Gb/s
  - to get 4:1 multiplexing ratio use an optional "Octolink" aggregating board …**NO**
- GPU Direct
  - for support to future enhanced read-out and trigger system GPU-based… ✔
- "Fixed Latency" clock distribution for under-water events time-stamping
  - OK for FCM (Xilinx) ✔
  - **Altera Stratix IV/V** ✔

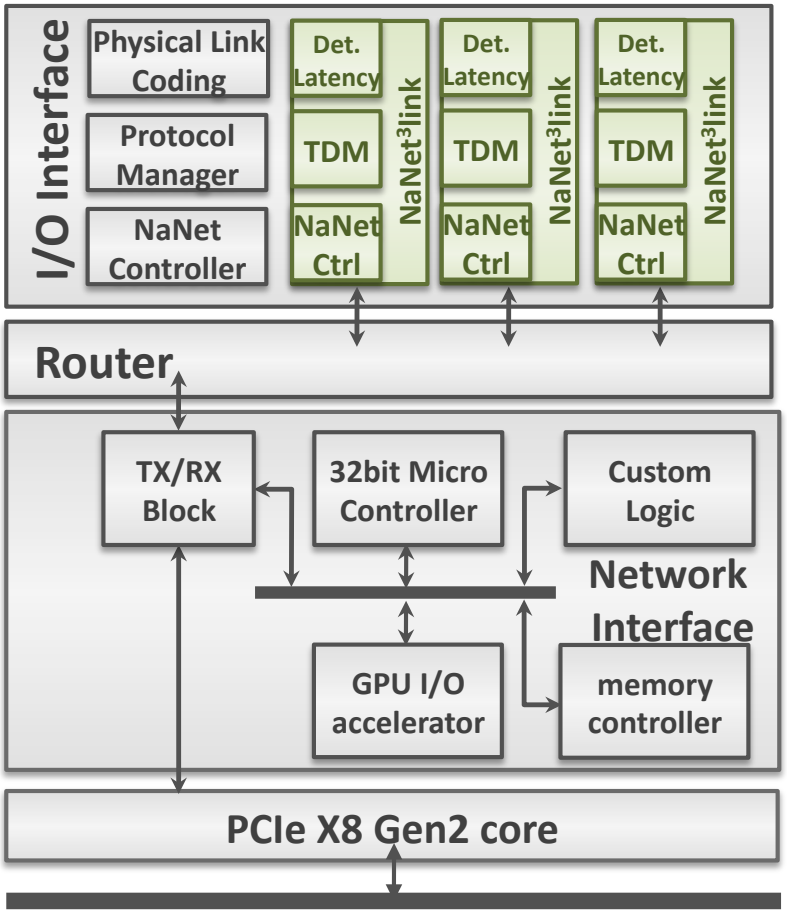
Final design on TERASIC DE5-NET

- Altera Stratix V based dev board
- PCIe Gen2 x8 (Gen3 support)
- 4 independent SFP+ ports (up to 10Gb/s)
- Deterministic Latency mode for transceivers

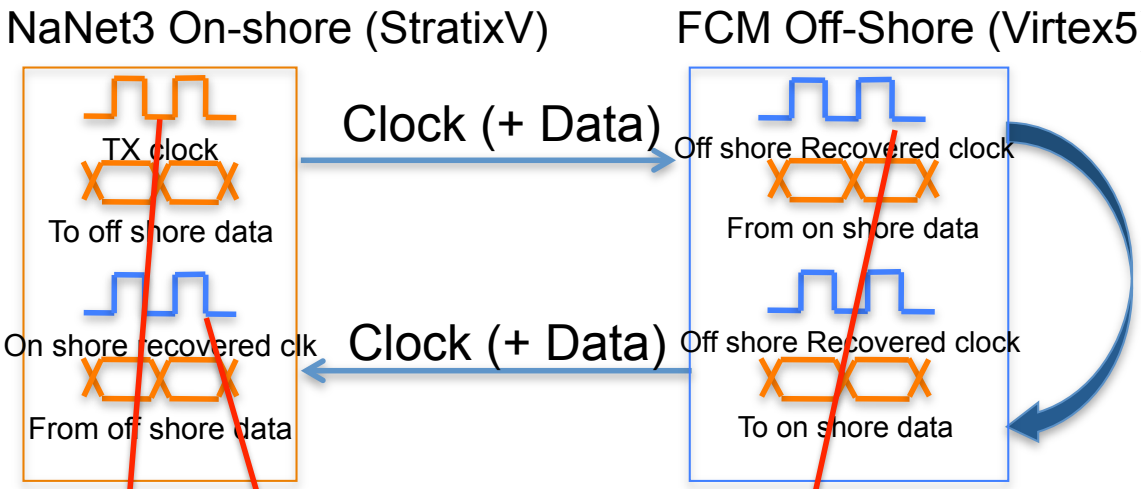


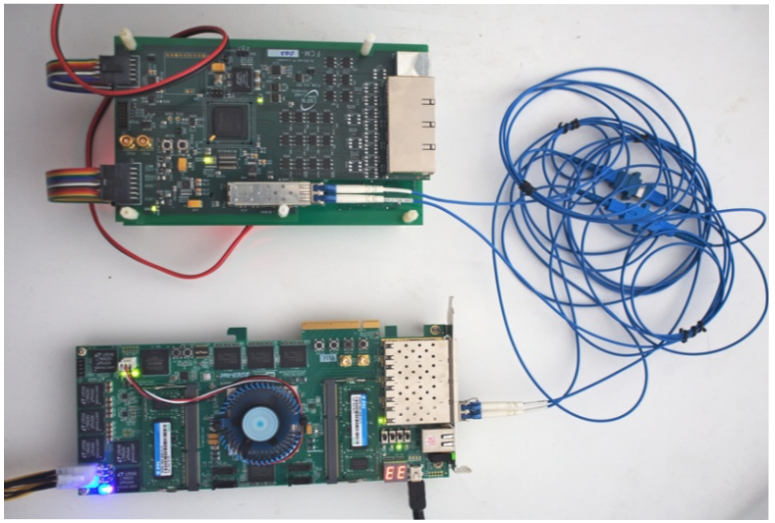
NaNet^3: State of the art

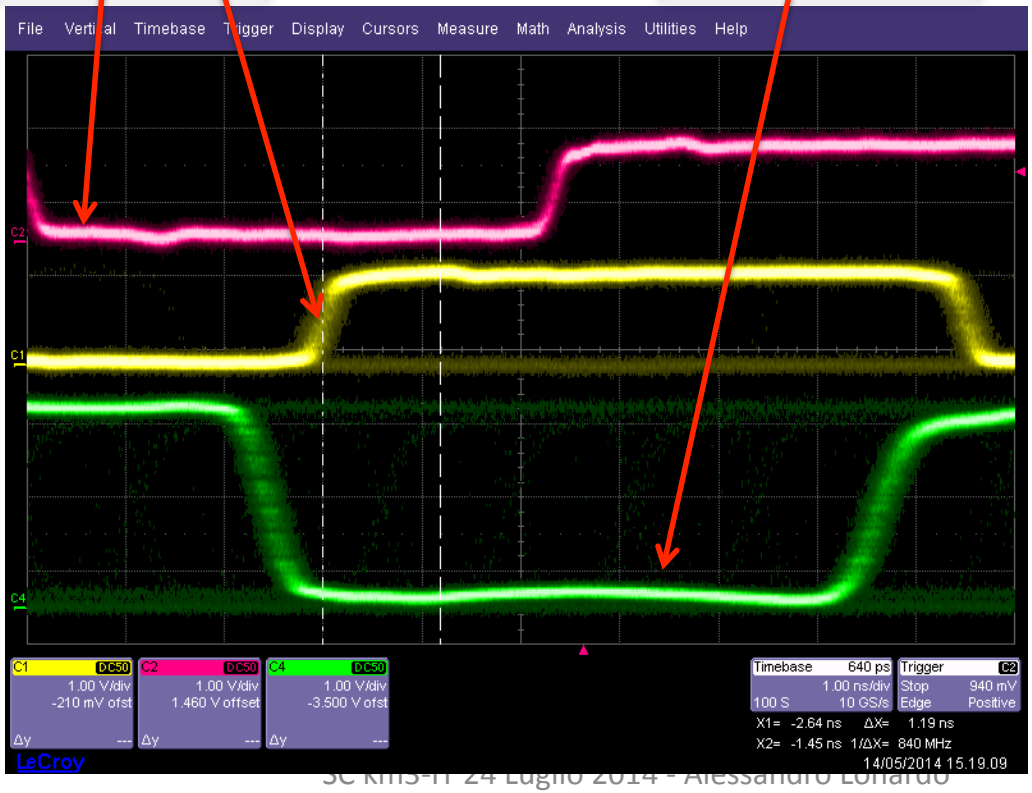
- **NaNet^{3}link (3 up to now):**
  - Physical Layer: Altera Deterministic Latency Transceivers (8B10B encoding scheme)
  - Data Layer: Time Division Multiplexing (TDM) data transmission protocol.
    - RX path: payload of different off-shore devices, multiplexed on continuous data stream at fixed time slot
      - (PCIe DMA transaction to CPU/GPU memory)
    - TX path: limited data rate per FCM
      - (PCIe TARGET transaction from CPU/ GPU memory)
- **NaNet Ctrl:**
  - 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
- **PCIe X8 Gen2**
  - CPU/GPU Memory Write Bandwidth ~ 2.4 GB/s



Deterministic latency link inter-operability







Preliminary but encouraging results

- 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

Time Division Multiplexing (TDM) support

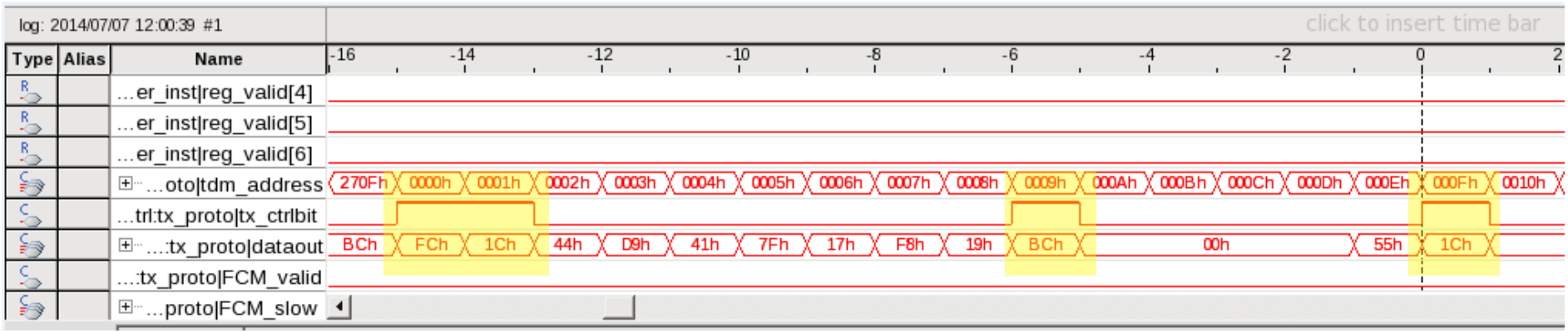
NaNet I/O interface Customization:

- TDM implementation: compliant with KM3Net-IT specification, echo test OK!

	K Code	Kout	D
IDLE_CODE	K28.5	1	0xBC
FRAME_CODE	K28.7	1	0xFC
DATA_CODE	K28.0	1	0x1C

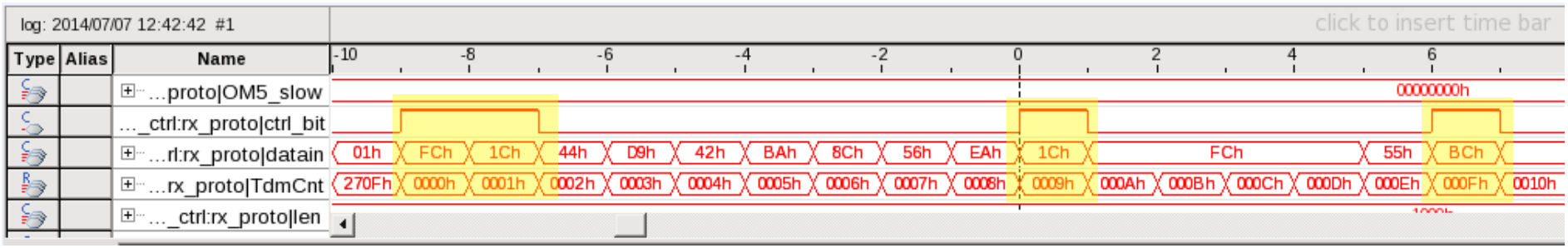
Field	Validity Code	Start Address	Stop Address	Length [byte]
Start Frame	FRAME_CODE	0x0000	0x0000	1
Frame Time	DATA_CODE	0x0001	0x0008	1+6+1
FCM Slow Control	DATA_CODE	0x0009	0x000E	1+4+1
OM0 Slow Control	DATA_CODE	0x000F	0x0014	1+4+1
OM1 Slow Control	DATA_CODE	0x0015	0x001A	1+4+1
OM2 Slow Control	DATA_CODE	0x001B	0x0020	1+4+1
OM3 Slow Control	DATA_CODE	0x0021	0x0026	1+4+1
OM4 Slow Control	DATA_CODE	0x0027	0x002C	1+4+1
OM5 Slow Control	DATA_CODE	0x002D	0x0032	1+4+1
Hydro data	No IDLE_CODE	0x0200	0x03FF	512
OM0 Phy Data	No IDLE_CODE	0x0400	0x07FF	1024
OM1 Phy Data	No IDLE_CODE	0x0800	0x0BFF	1024
OM2 Phy Data	No IDLE_CODE	0x0C00	0x0FFF	1024
OM3 Phy Data	No IDLE_CODE	0x1000	0x13FF	1024
OM4 Phy Data	No IDLE_CODE	0x1400	0x17FF	1024
OM5 Phy Data	No IDLE_CODE	0x1800	0x1BFF	1024

- NaNet^3 TX



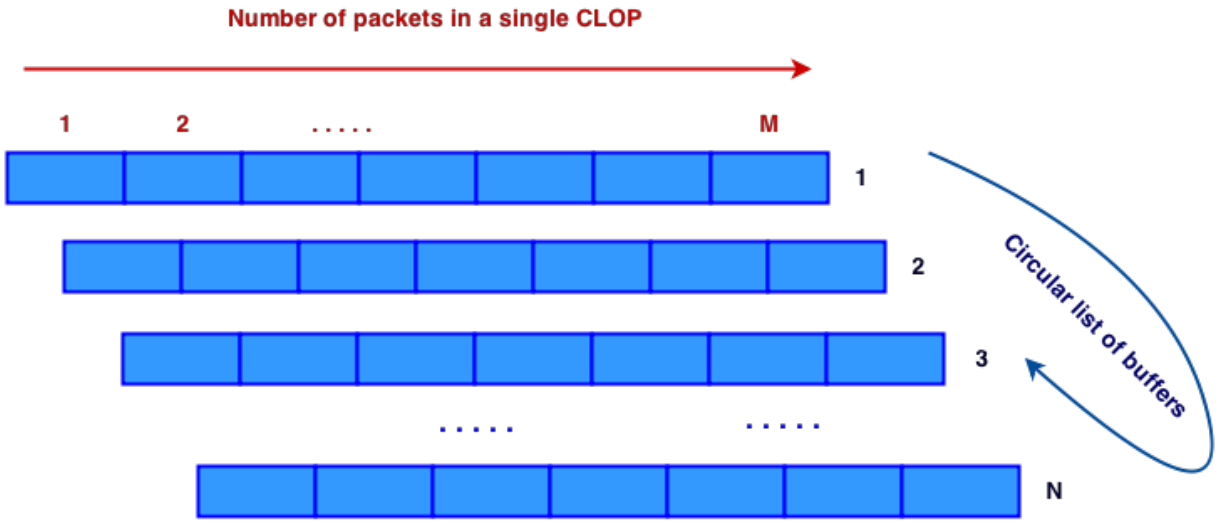
- NaNet^3 RX:

  Data flow: NaNet^3 → FCM board → NaNet^3



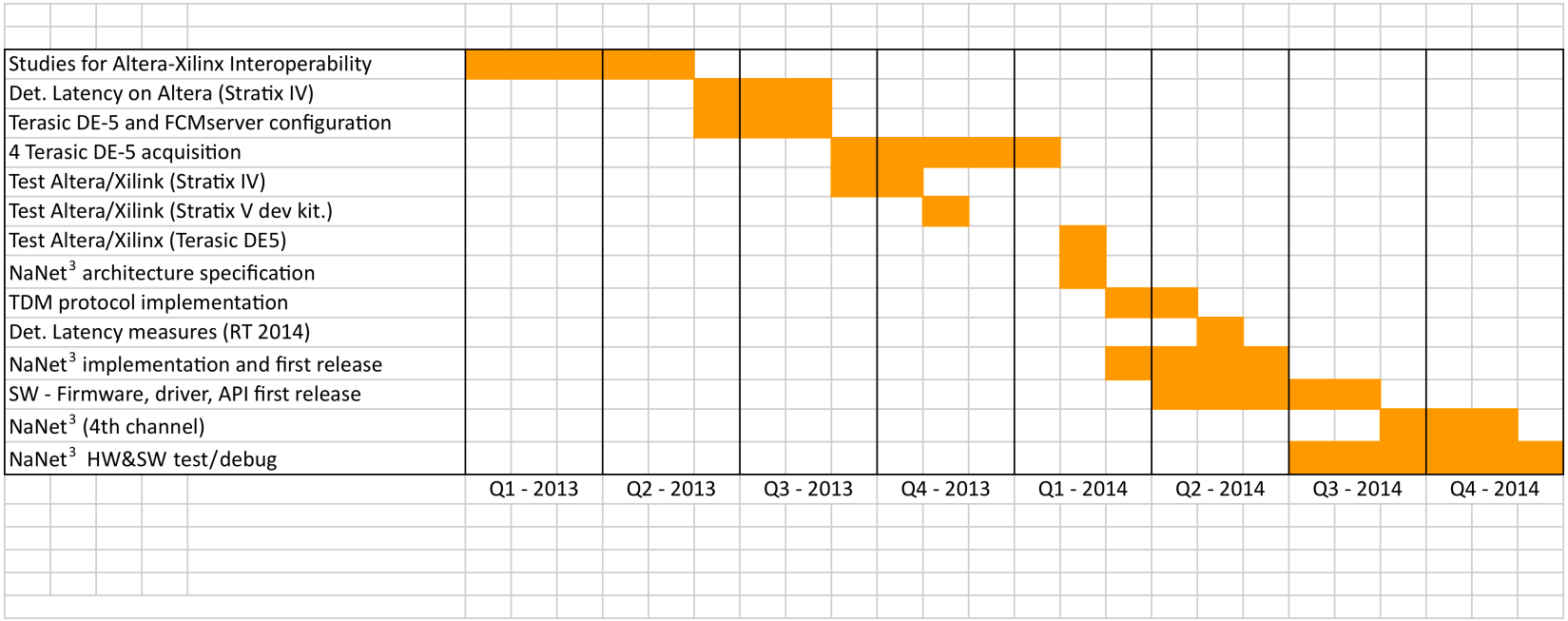
NaNet³ Software

- **Linux Kernel Driver**
  - Status/Configuration registers.
  - TX registers interface.
- **NIOS II Firmware**
  - New BSP for NaNet3 board.
  - Initialization of NaNet3 channels.
  - Management of 4 concurrent data streams



- **Application Library**
  - nan3_t nan3_open(int card_id); int nan3_close(nan3_t nan);
  - int nan3_register_clop(nan3_t nan, nan3_chan_id_t chan, u32 pkts, u32 items, u8 is_gpu);
  - int nan3_wait_event(nan3_t nan, nan3_event_recv_t* event);
  - u32 nan3_write_slow_data(nan3_t nan, u8 channel, u8 deviceId, u32 data);
- **Application Template**

NaNet^3: Roadmap



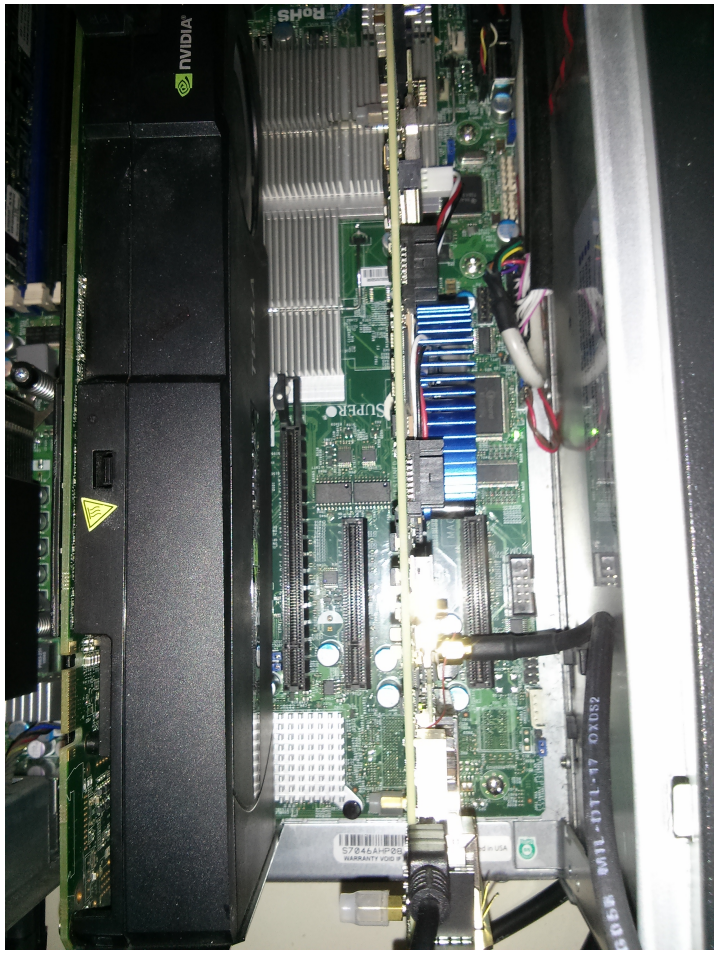
- Dec 2012: Piero Vicini talk on NaNet in Catania.
- Aug 2013: Deterministic Latency on Altera Stratix IV dev kit.
- Oct 2013: Interoperability Altera/Xilinx.
- Feb 2014: Deterministic Latency on Terasic DE5 (Stratix V).
- Mid Feb 2014: NaNet^3 architecture specification.
- Jul 2014: first release of NaNet^3 of complete HW design (3 channels).
- Aug 2014: first release of system software and application library.

Budget & People

- NaNet HW and SW developments financed by FP7 EURETILE project (exploitation activities): 8 PM in 2013, 9 PM in 2014.
- EURETILE ending on September 2014.
- People involved
  - HW
    - Andrea Biagioni
    - Ottorino Frezza
    - Francesca Lo Cicero
    - Piero Vicini (Staff)
  - SW
    - Alessandro Lonardo (Staff)
    - Michele Martinelli
- Integrated remaining man power on HW side : 0.5 MM before the end of the project.
- People available to continue activities after September (if supported):
  - Integration of 4th channel.
  - Test and debug
  - …

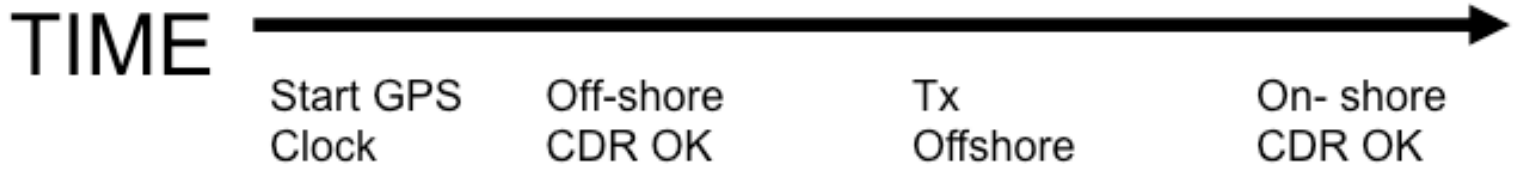
BACKUP SLIDES

NaNet^3 testbeds

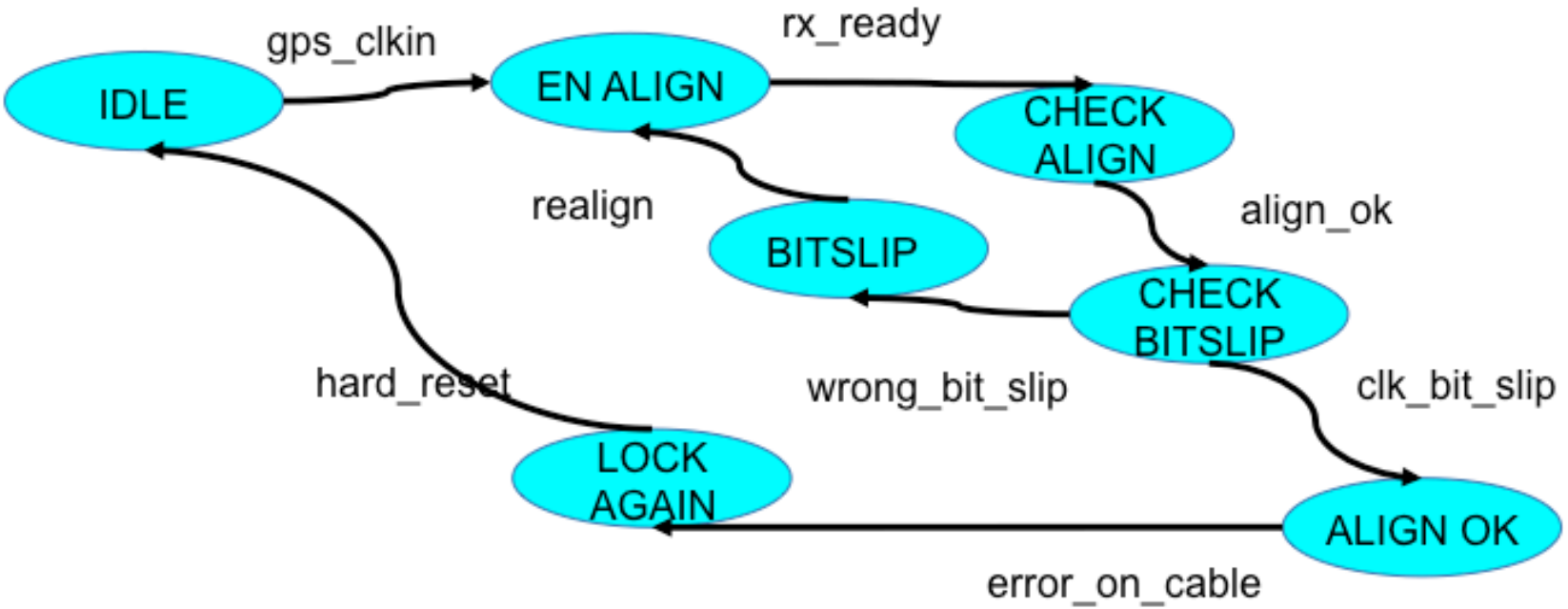


Deterministic latency link inter-operability

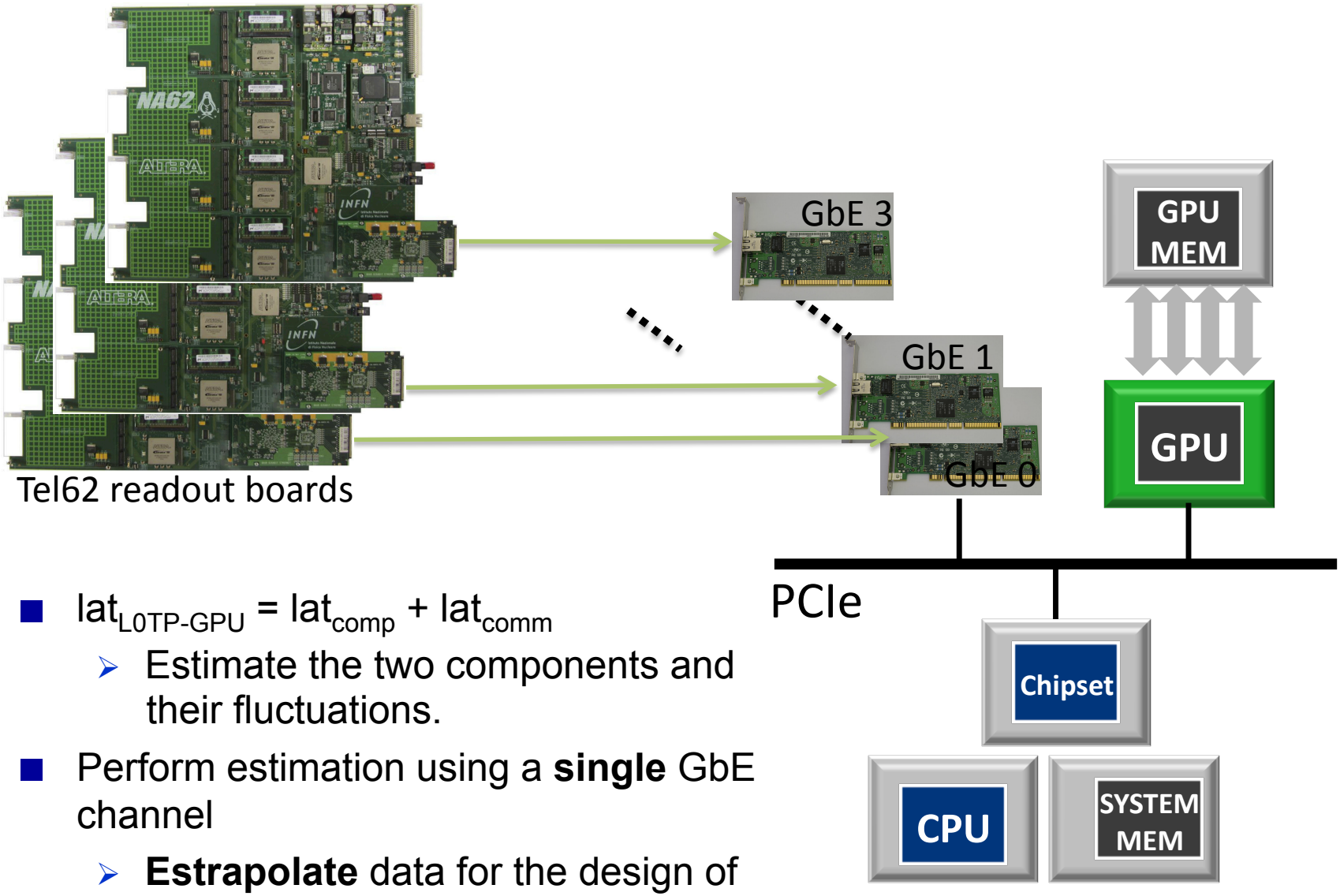
NaNet³ on-shore	Tx	Send Magic word	Sending Magic word	Sending Magic word	Sending data
	Rx	Under reset	Under reset	Under reset	Clock recovered with Det. Lat.
FCM offshore	Tx	Idle	Idle	Sending Magic word	Sending data
	Rx	Waiting for CDR	Clock recovered	Clock recovered	Clock recovered



DETERMINISTIC LATENCY ALIGNMENT STATE MACHINE



GPU-Based RICH Level 0 TP System Latency Estimation

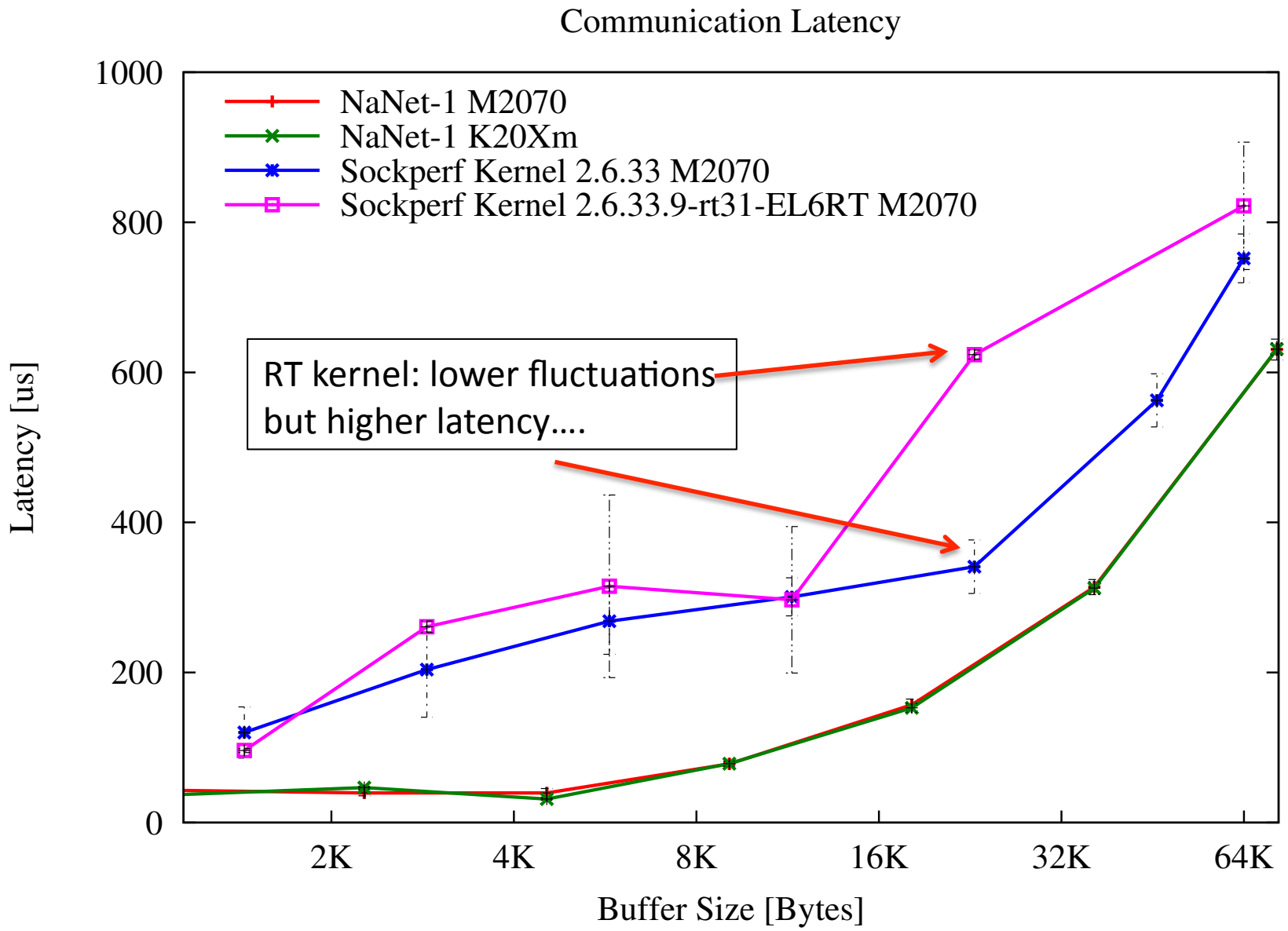


- $lat_{L0TP\text{-}GPU} = lat_{comp} + lat_{comm}$
  - Estimate the two components and their fluctuations.
- Perform estimation using a **single** GbE channel
  - **Estrapolate** data for the design of the “full bandwidth” system.

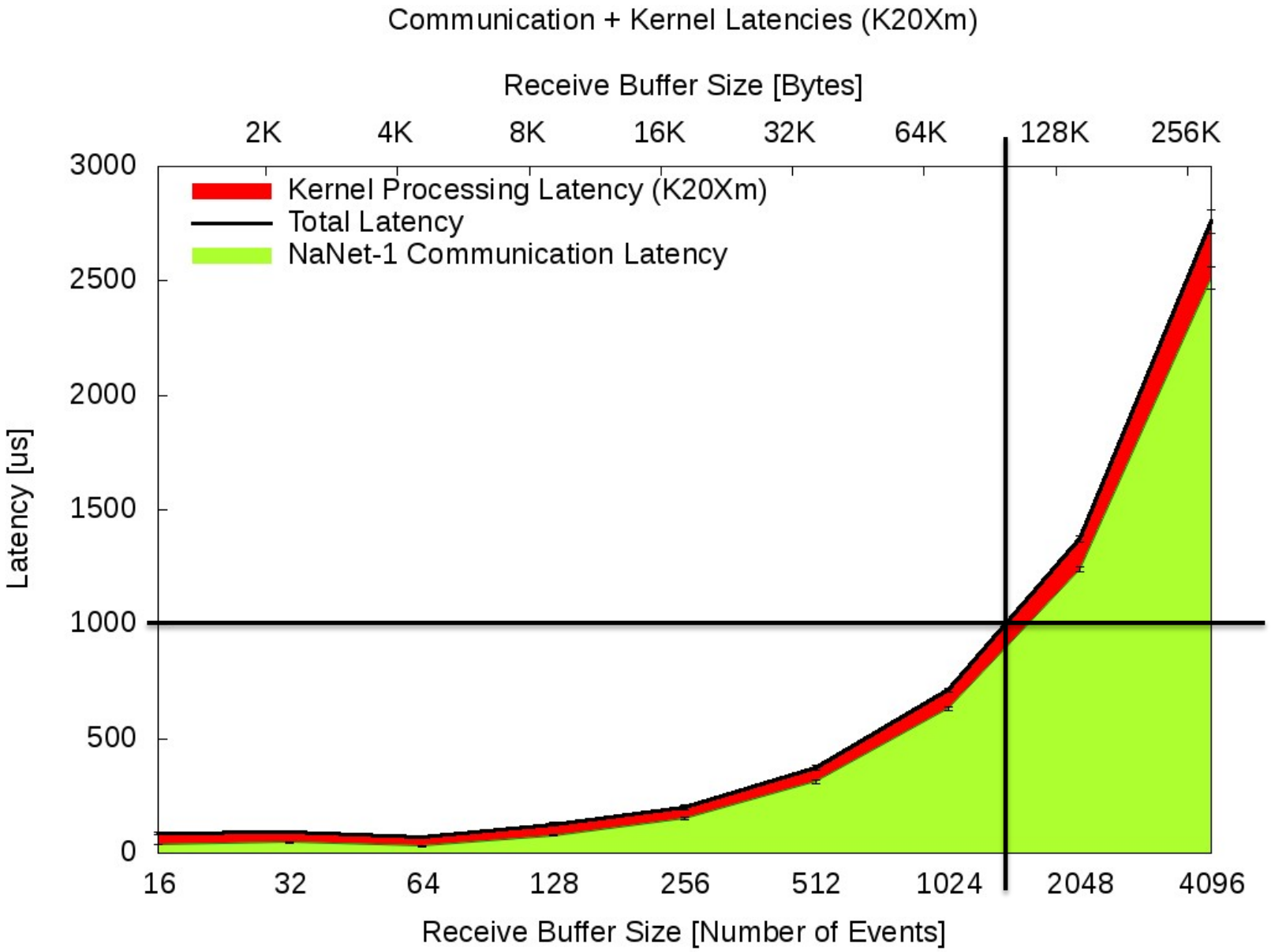
NaNet: a FPGA-based NIC with GPUDirect RDMA capability

- Challenge:
  - Lower communication latency and its fluctuations.
- How?
  1. Injecting directly data from the NIC into the GPU memory **without intermediate buffering.**
  2. **Offloading** the CPU from network stack protocol management, avoiding OS jitter effects.
- **NaNet** solution:
  - Re-use the **APEnet+** design, implementing **PCIe** host interface and **GPUDirect** .
  - Add a network stack protocol management offloading engine to the logic (**UDP Offloading Engine**).
  - Use FPGA resources to perform **processing on data stream** (e.g. reformat data on-the-fly in a GPU-friendly fashion)

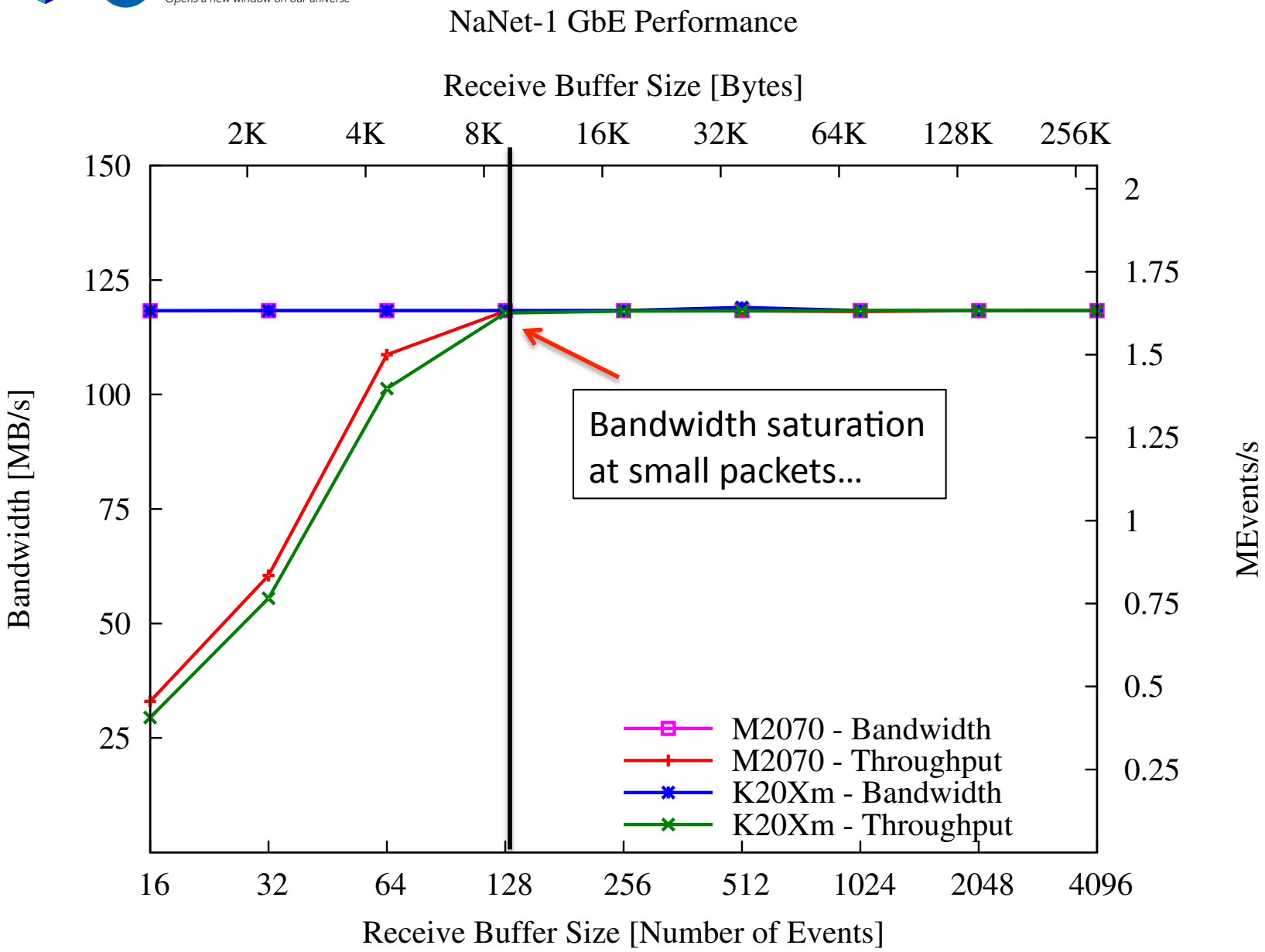
NaNet-1 Communication Latency



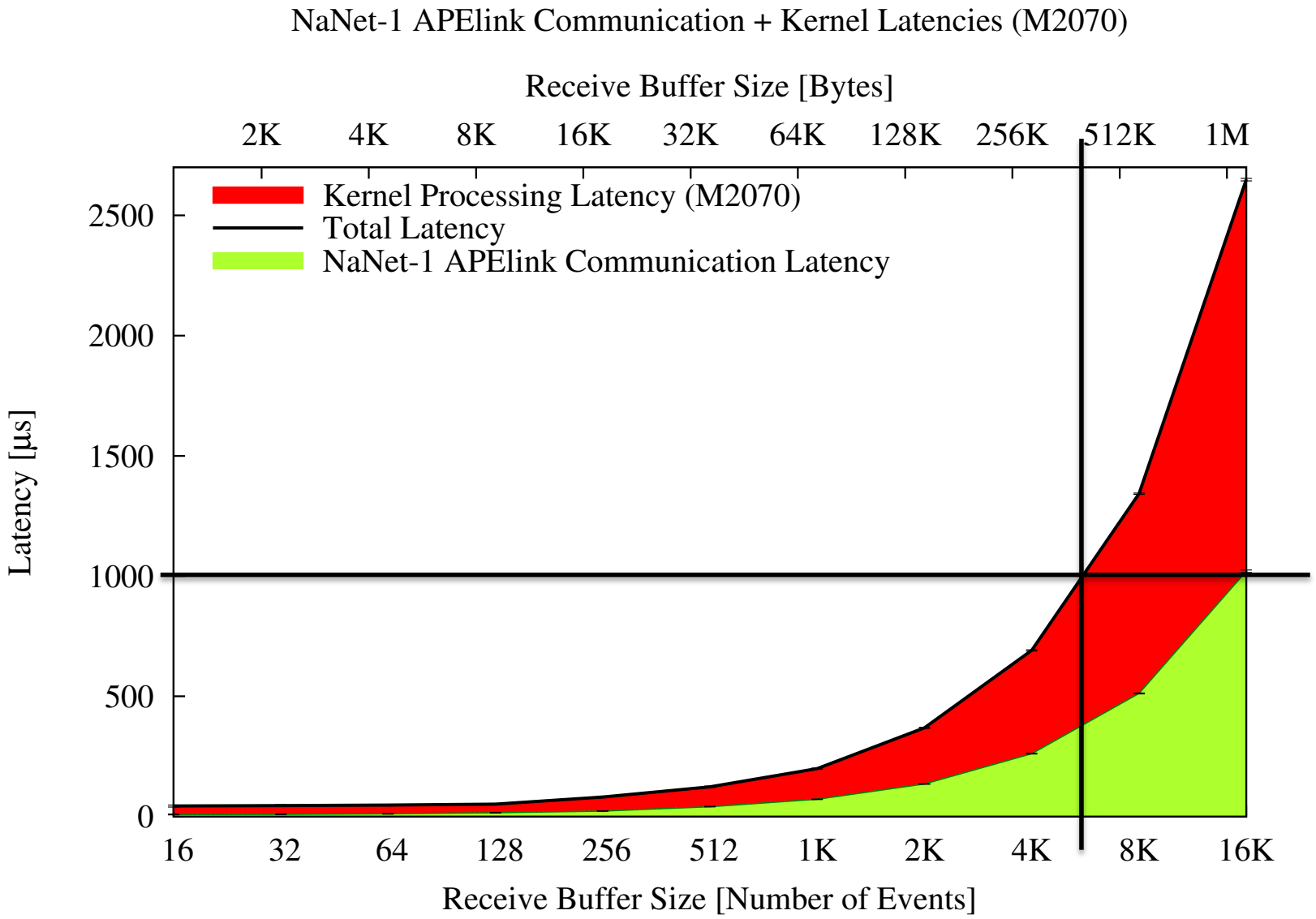
Total latency of the GPU-Based RICH L0-TP using NaNet-1 (Kepler K20X)



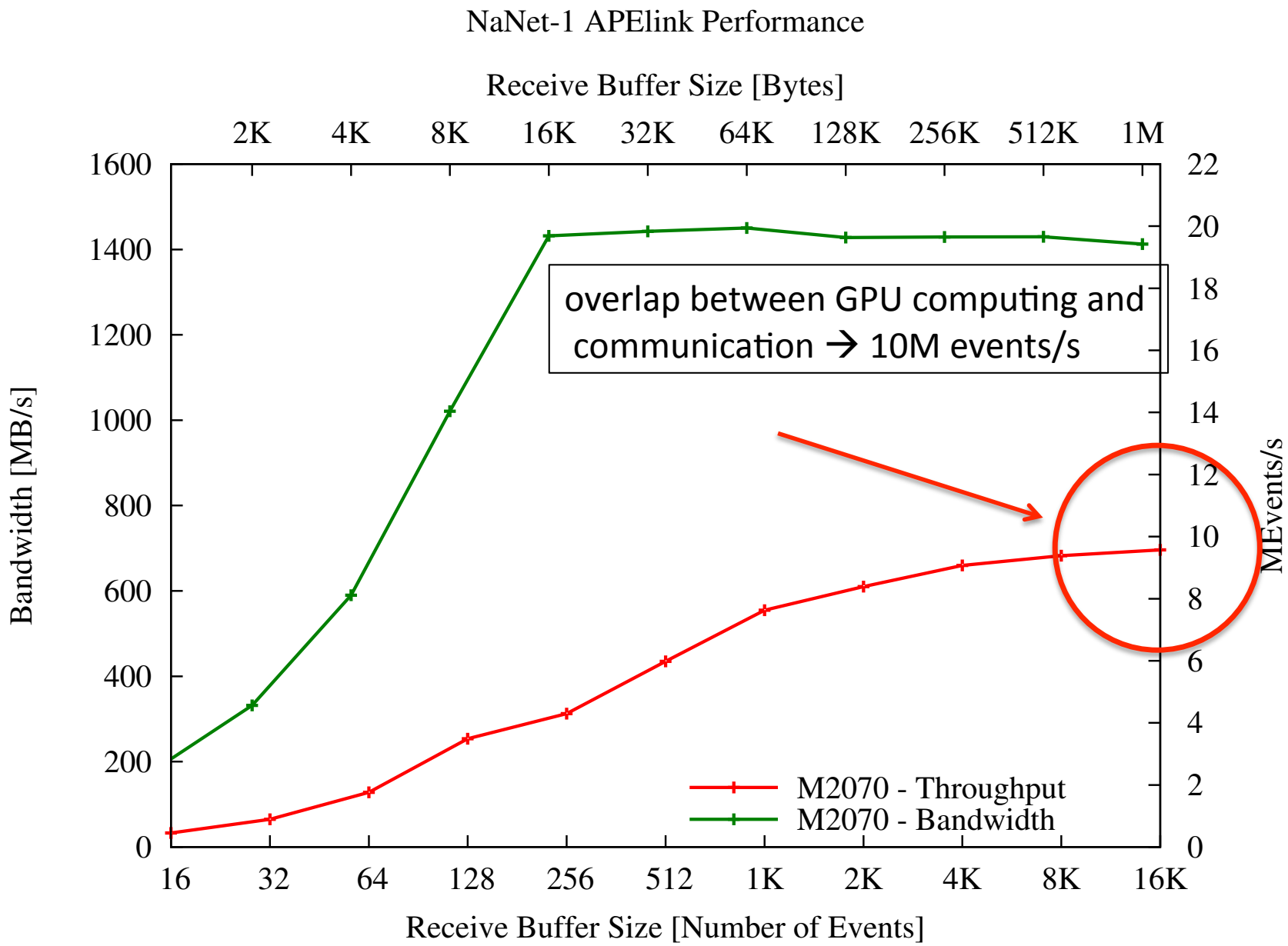
NaNet-1 1 GbE Bandwidth & Throughput



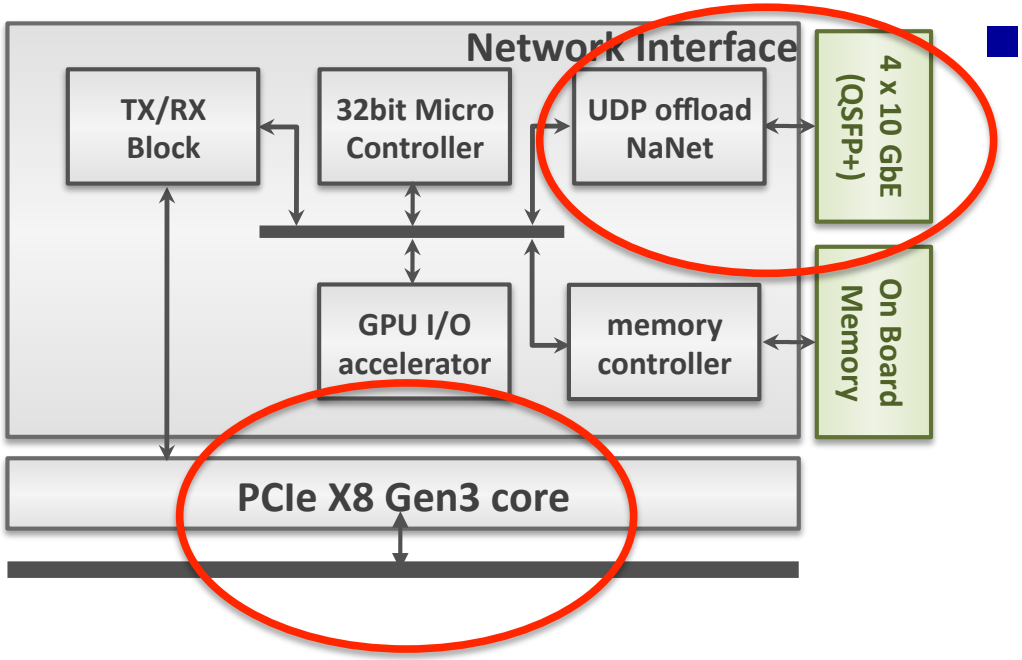
Total latency of the GPU-Based RICH L0-TP using NaNet-1 (APElink)



NaNet-1 APElink Bandwidth & Throughput



NaNet-10 and last generation FPGAs (Stratix V)



- Implemented on the Altera Stratix V dev board
  - PCIe gen2 x8 but developing PCIe Gen3 (8 GB/s)
  - Faster embedded Altera transceivers (up to 14.1 Gbps)
  - hardened 10GBASE-R PCS features to support 10 Gbps Ethernet (10GbE)

QSFP+ to 4 SFP+ cable