APE3Net:
i.e. idee sparse su
rete 3D Torus di APE in Nemo/KM3Net

P. Vicini – INFN Roma

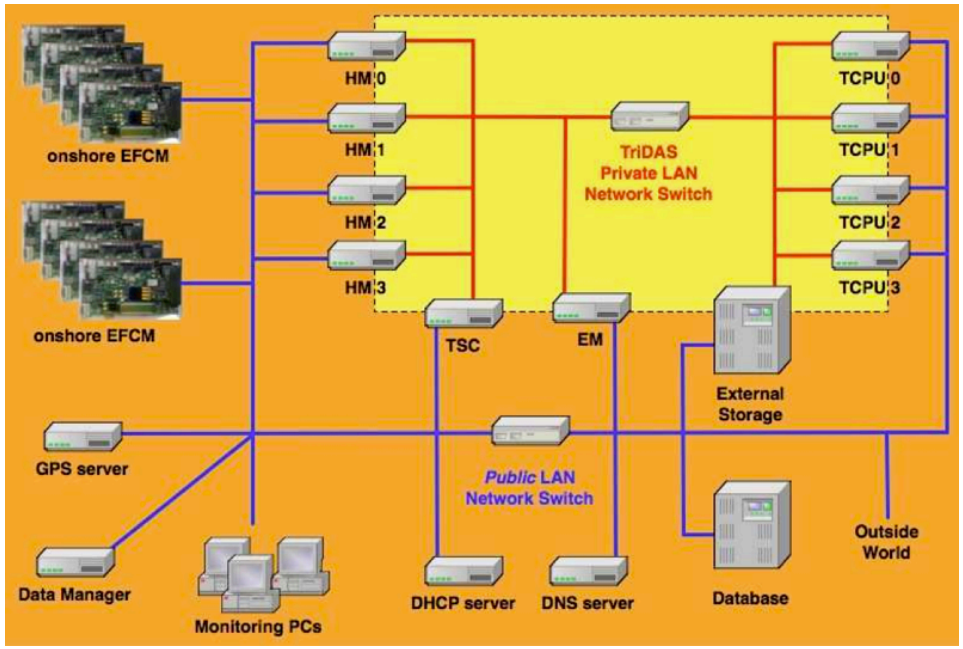
…leggere prima dell'uso….

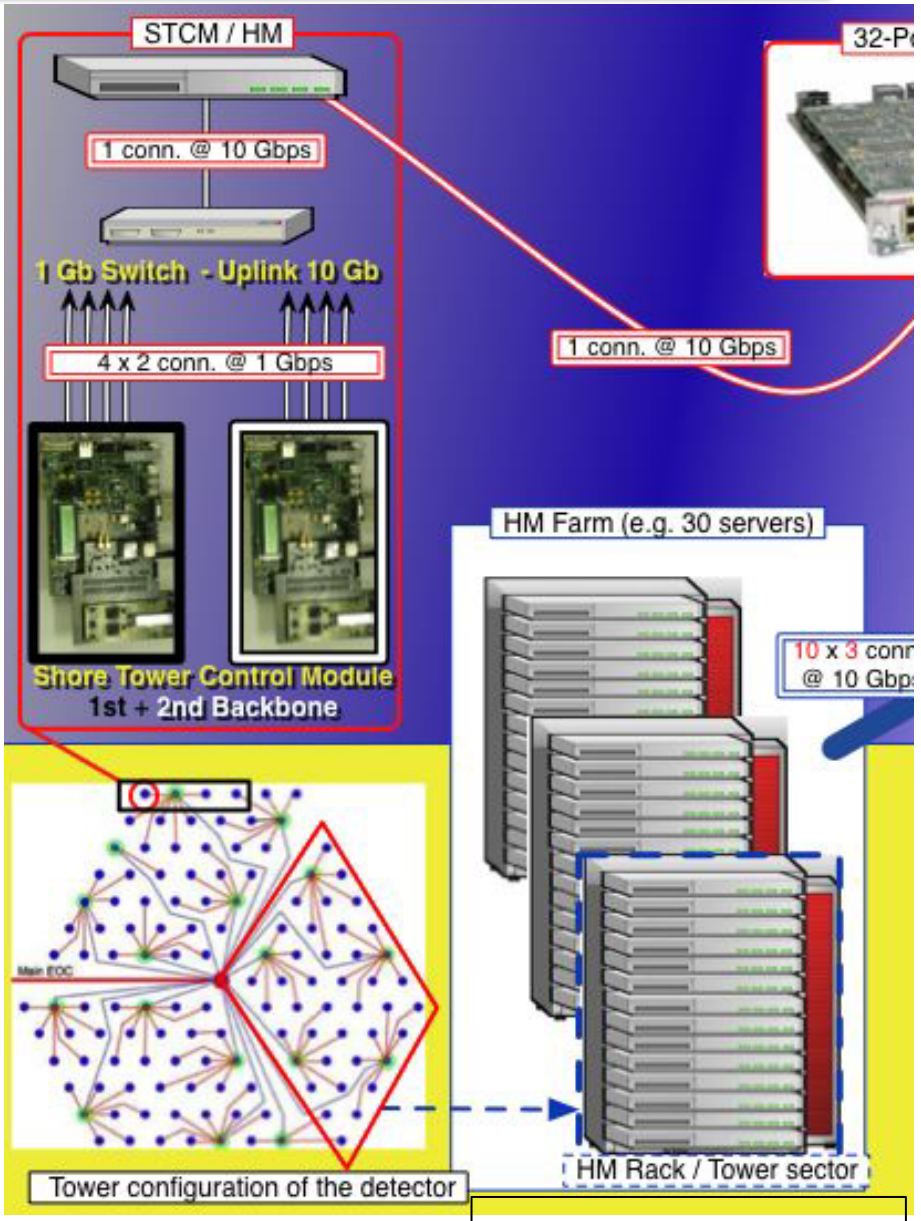
- APE3Net per:
  - 3D Net, (TRI)DAS Net, KM3 Net, FASE 3 (and beyond…) Net….

- Avvertenze….
  - Brainstorming preliminare
    - con il supporto ed il contributo di F.Ameli, T. Chiarusi, A. Lonardo,….
  - Focus del talk su elaborazione a terra dei flussi dati dei PMTs
  - Obiettivo e' discutere la fattibilita' sulla base delle bande passanti, hardware disponibile o acquisibile, costi…

NEMO phase 2

- L'attuale implementazione scalata al KM3 mostra una complessita' *impressionante* e e fa leva su tecnologie che sono (almeno) *un ordine di grandezza inferiore allo stato dell'arte*
  - 1 board per piano della torre
  - Alta numerosita' di PC cluster dovuta alla granularita' fine del readout





talk di T. Chiarusi

Constraints per APE3Net

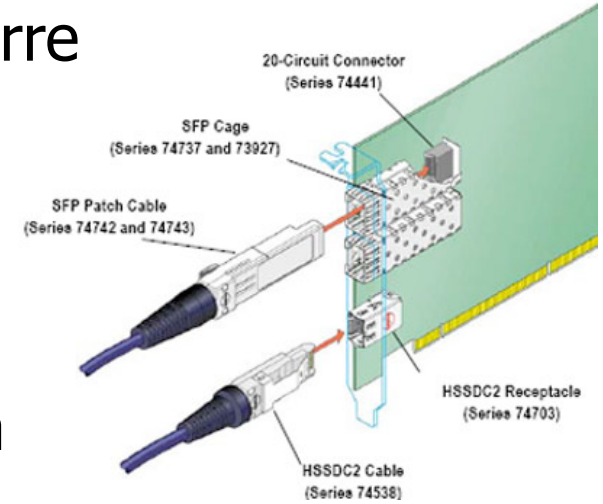
Esiste una soluzione che :

- sia a basso impatto
  - limitato R&D necessario per finalizzare il progetto;
- utilizzi una tecnologia mainstream
  - Tecnologia matura “state-of-the art” con una roadmap consolidata e supporto garantito nel time frame dell’esperimento
- sia efficente e cost effective sulla scala del KM3

Specifiche per APE3Net

Ipotesi di partizione a terra del data throughput (vedi talk di Fabrizio…)

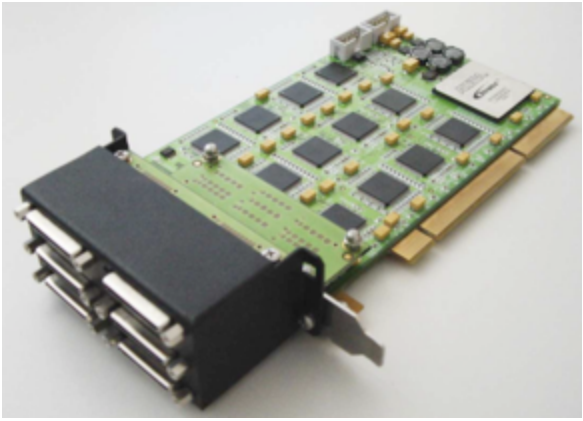
- Torre finale (14 piani) con un throughput di piano pari a 150 Mbps (300Mbps)
- 1 fibra monocromatica tra piano e base torre (*link*)
- 7 links (i.e. mezza torre) multiplexati su singola fibra colorata (*canale*)
  - 7*150 (7*300) -> 1.05 (2.1) Gbps
- 2 canali da up to 2.5Gbps per i dati di una torre
- ~100-200 canali per KM3



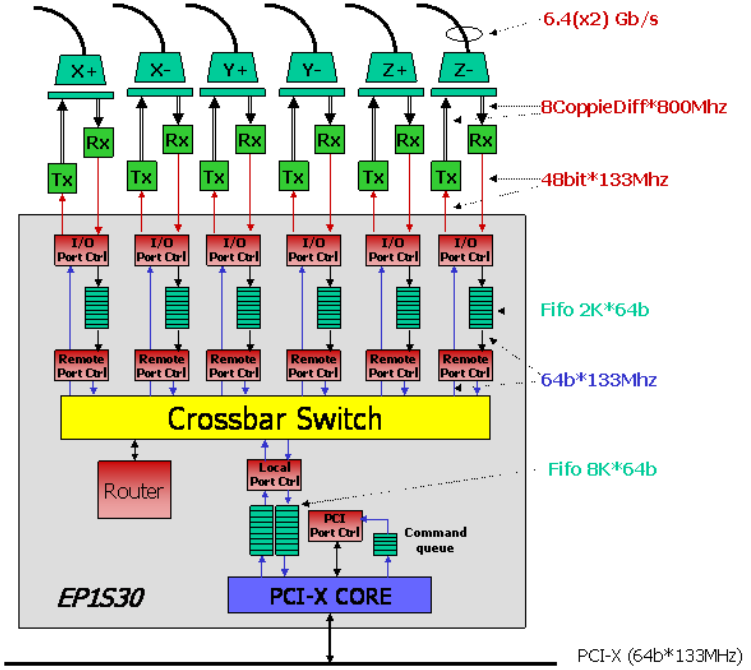
- il canale usa standard di connessione SFP (up to 2.5Gbps)

APEnet

- APEnet
  - Rete 3D Toroidale derivata dai supercalcolatori APE per clusters di PC
  - Capacita' di routing e switching integrata
  - Alte prestazioni, bassa latenza e protocollo di comunicazione custom "light-weight"
  - Interfaccia PCI su lato host
  - 6 canali bidirezionali indipendenti sul lato toro



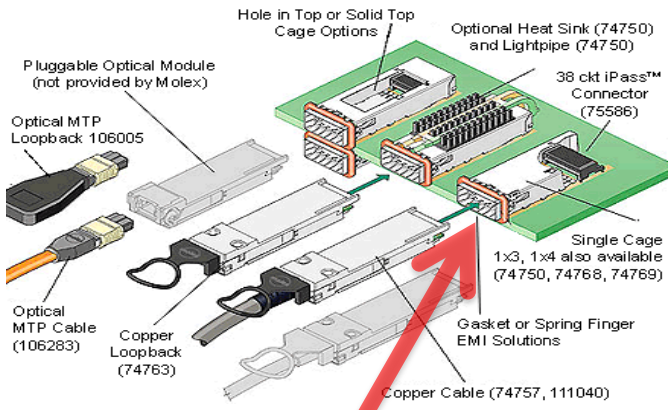
- Releases
  - 2003-2004: APEnet V3 (PCI-X)
  - 2005: APEnet V3+
    - stesso HW ma implementazione firmware con RDMA
  - 2006-2009: "APEnet goes embedded"
    - DNP, D(istributed) N(etwork) Processor
    - EU SHAPES project co-development
  - 2011: APEnet+
    - PCI Express sul lato host
    - Incremento di performance sul lato del toro 3D

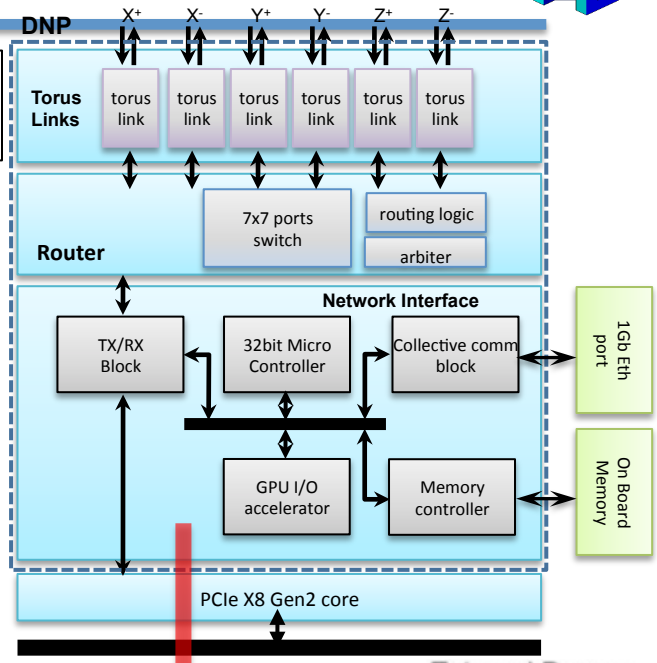


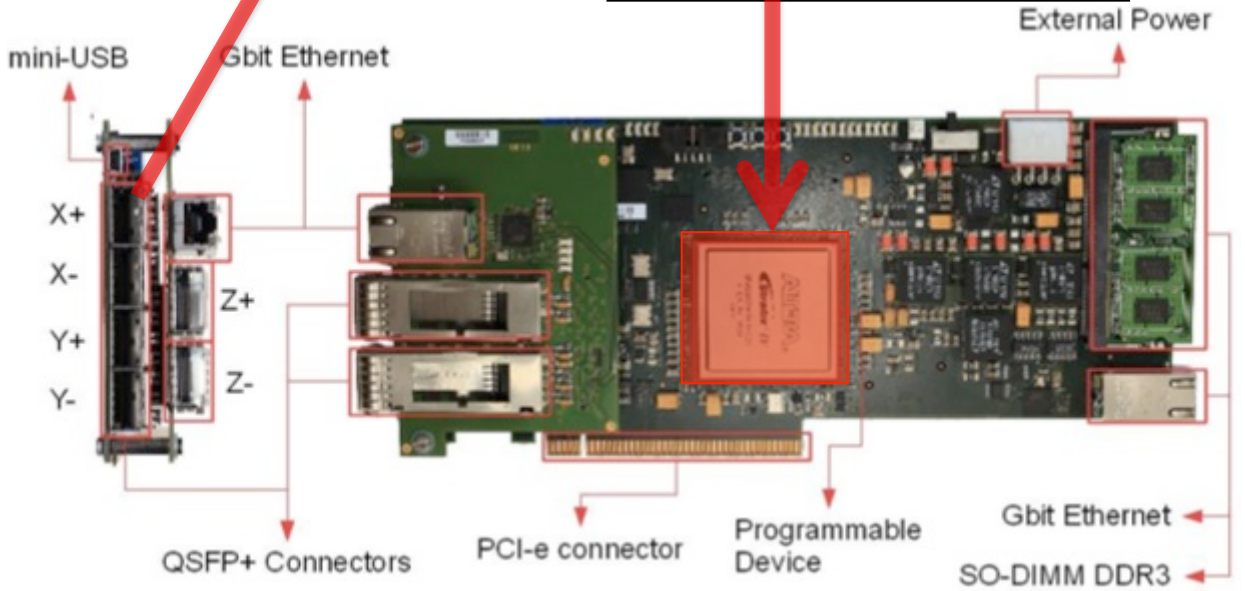
APEnet+ at a glance

- 3D Torus network
  - ideal for large-scale scientific simulations (domain decomposition, stencil computation, …)
  - scalable (APENEt+ today up to 32K nodes)
  - Cost effective: no external switches! 1 card+3 cables
- APEnet based on INFN DNP
  - RDMA: Zero-copy RX & TX !
  - Small latency & high bandwidth
  - GPU clusters features (APEnet+):
    - RDMA support for GPUs! -> no buffer copies between GPU and host.
    - Very good GPU to GPU latency
- APEnet+ card:
  - FPGA based (ALTERA St.IV EP4SGX290)
  - 6 full-bidirectional links up to 68 Gbps raw (~400 Gbps)
  - PCIe X8 Gen2 in X16 slot
    - peak BW 4+4 GB/s
  - Network Processor, off-loading engine integrated in the FPGA
  - Zero-copy RDMA host interface
  - Direct GPU interface
  - Industry standard QSFP+ cabling
    - Copper (passive/active), optical

http://apegate.roma1.infn.it/mediawiki/index.php/APEnet%2B_project



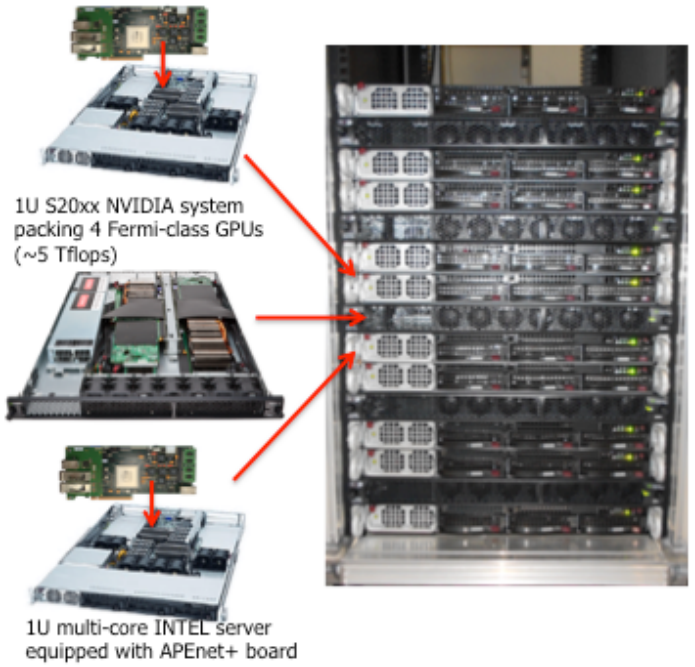




QUOnG: GPU+3D Network FPGA-based

QUonG (QUantum chromodynamics ON Gpu) e' un'iniziativa INFN per realizzare architetture di calcolo basate su GPU e dedicate ad applicazioni scientifiche (fisica teorica).

- Sistema eterogeneo: PC Cluster accelerato da GPU high-end (Nvidia) ed interconnesso da una rete 3D Toroidale
- Il valore aggiunto e' duplice:
  - Integrazione profonda tra acceleratori computazionali e rete ottimizzata riconfigurabile (APEnet+ e' FPGA-based) che garantisce efficenza computazionale e riduzione di latenze.
  - Le FPGA mostrano enormi risorse hardware on-chip per integrazione di blocchi hardware custom addizionali per interfacce e task computazionali
- Software stack standard (MPI,...) ma anche un programming model evoluto (cuOS)
- Aggregazione di una comunita' di ricercatori che mettono in comune expertize e codici (LQCD, GWA, Laser-plasma interactions, BioComputing, Complex systems, High Level Trigger Algorithm)



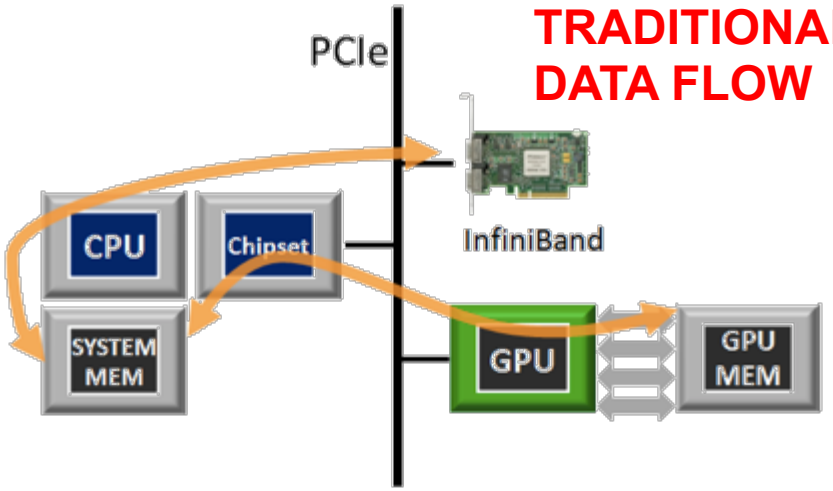
http://apegate.roma1.infn.it/mediawiki/index.php/QUonG_initiative

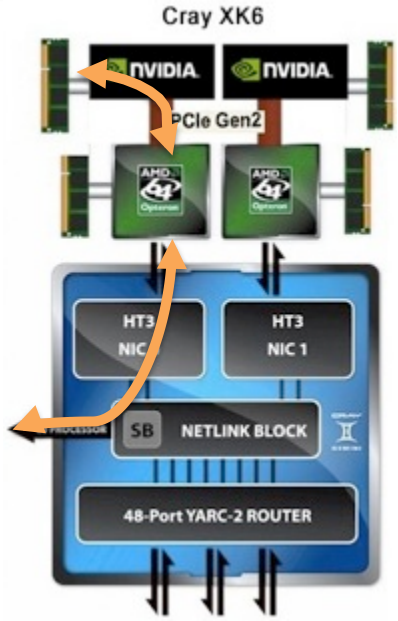
GPU-P2P in APEnet+

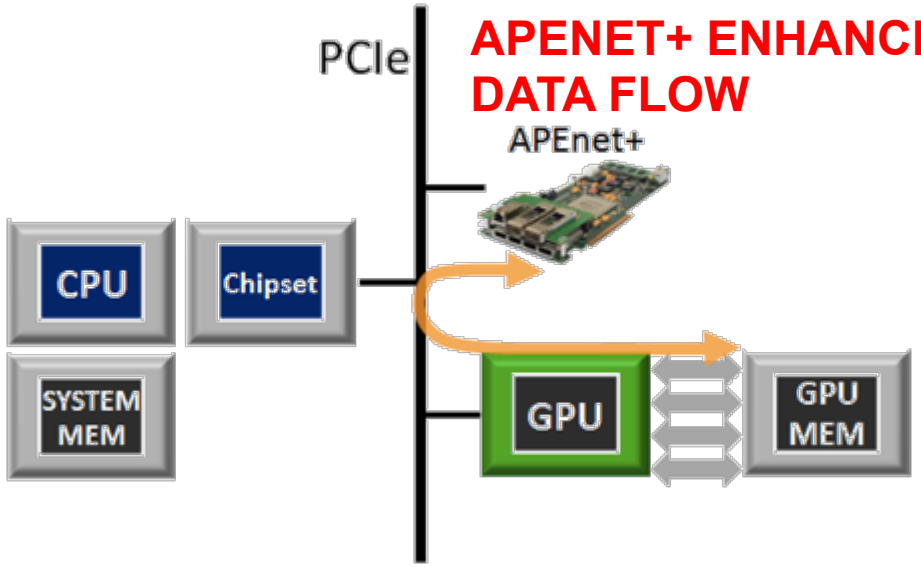
Peer-to-Peer means:
- Data exchange on the PCIe bus
- No bounce buffers on host

APEnet+ P2P support
- cutting-edge HW/SW technologies developed jointly with Nvidia
  - APEnet+ board acts as a peer
  - APEnet+ board can read/write "directly" GPU memory

Direct GPU access
- Specialized APEnet+ HW block
- GPU initiated TX
- Latency saver for small size messages







APEnet+ customization: NaNet

NaNet: APEnet + NA62 cern Experiment

GPU L0 TRIGGER for HEP Experiments

Implement a RO Board-L0 GPU link with:

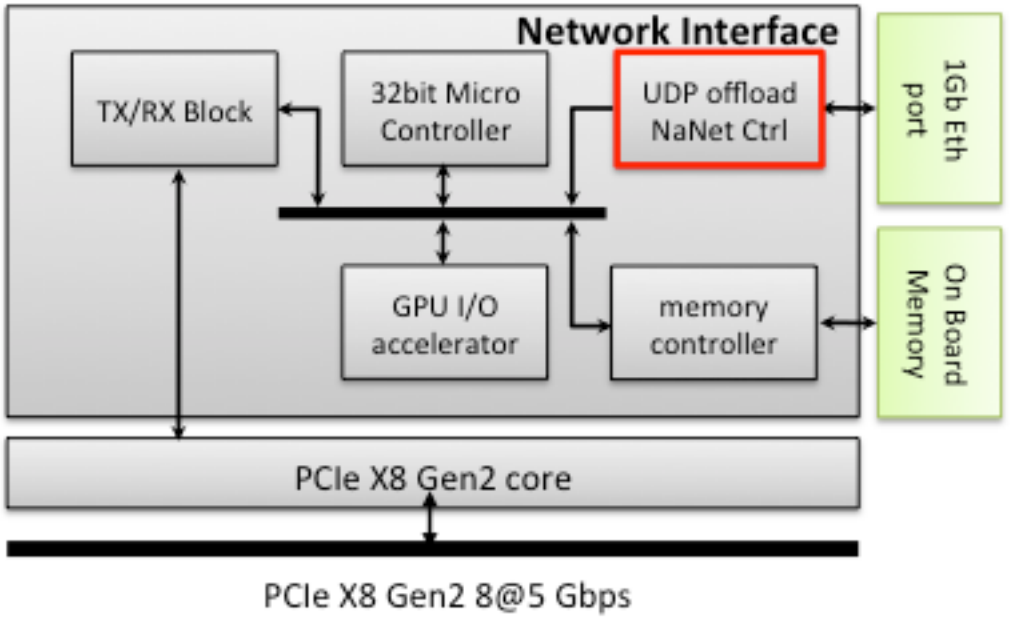
- Sustained Bandwidth > 600 MB/s, (RO board output on GbE links)
- Small and stable latency

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 solution:

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



APEnet+ customization: event selection in HEP exps.

GPUs for real time event selections?

GPU

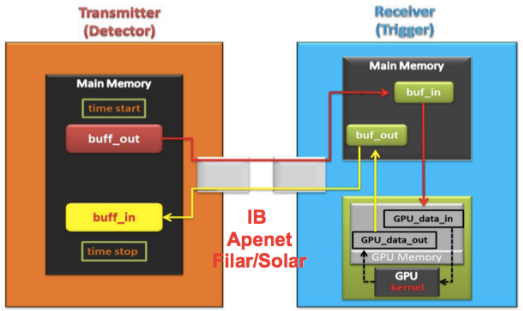
- A lot of computing power for highly parallelizable tasks;
- High level programming (CUDA, OpenCL);
- Commercial device → less expensive than dedicated hardware, continuous improvement of performance;
- NOT designed for low latency response

Real time events selection

- It is usually based on algorithms well suited for parallelization;
- A trigger system needs to be flexible, to be adapted to experiments changing conditions;
- It needs low latencies.

2

Data flow and measurements



Multiple loops, for each: ΔT = *Time stop -Time start*
Time measured in the transmitter using the time stamp counter register.

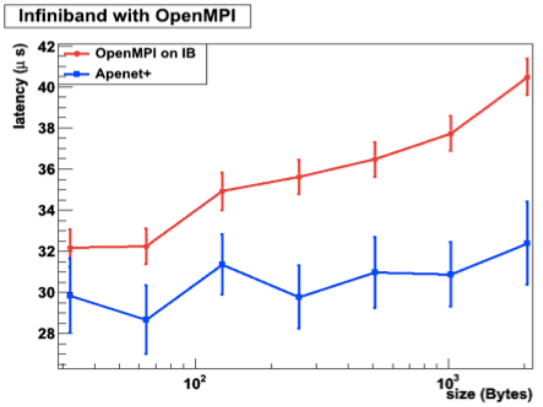
3 set of measurements:
1) data transfer only (N words IN → N words out)
2) data transfer + copy on GPU (N words IN → copy on GPU → N words out)
3) data transfer + copy on GPU + kernel (N words IN → copy on GPU → M words out)

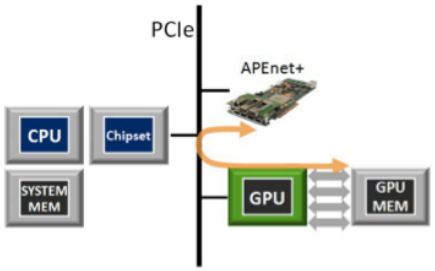
6

Amerio et al "*Applications of GPUs to Online Track Reconstruction in HEP Experiments*" NSS 2012

Data transfer + copy to the GPU

Apenet+





From 32 B to 2 kB
- IB: from 33 to 40 µs
- Apenet+: from 30 to 33 µs

Significant latency reduction with Apenet+ (direct GPU memory access)

12

APENet e KM3

Percorso di studio incrementale

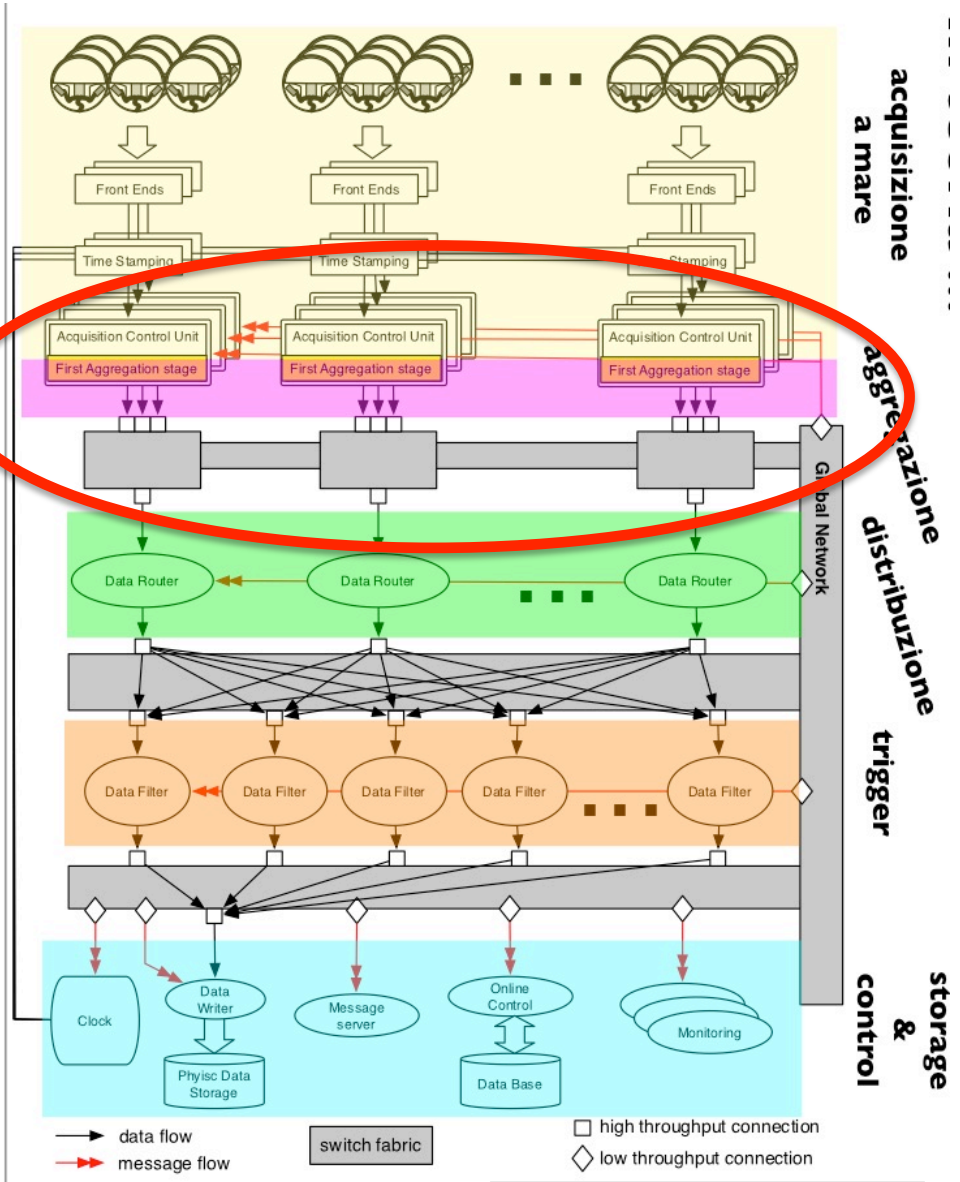
- Area di intervento iniziale puo' essere la frontiera tra acquisizione a mare e aggregazione (HM)

ma anche…

- una palestra per sperimentare nuove e piu' performanti sistemi meccanico/ottici per data transfer mare-terra
- nuovi modelli di computing (GPU-based) per elaborazione a terra
  - a differenti livelli di trigger
  - event reconstruction?

e a regime

- si ottiene una consistente riduzione del numero di CPU necessarie per l'elaborazione a terra e relativa riduzione di costi di acquisizione e operativi (power, manutenzione,etc..)



talk di T. Chiarusi

Interoperabilita' link seriali ALTERA vs XILINX (1)

1) FCM vs APENet+ ovvero XILINX vs ALTERA

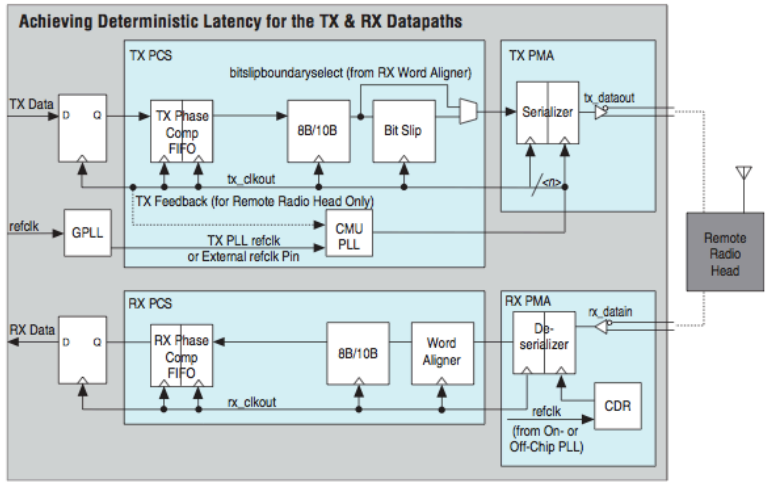
- Le due tecnologie sono simili, compatibili "per definizione" ma va verificata la loro interoperabilita'
- Dal punto di vista hardware il link XILINX presente sulla FCM, via SFP connection system deve poter trasferire dati da/verso link seriali ALTERA (a terra) con una bandwidth di almeno 2.5Gbps

2) Test della latenza deterministica del link seriale Altera

Achieving Deterministic Latency

Figure 11–2 illustrates the TX and RX channels when configured as a wireless basestation communicating to a remote radio head (RRH) using a CPRI or OBSAI interface. Figure 11–2 also provides an overview of the calculations that guarantee deterministic delay. As this figure illustrates, you can use a general-purpose PLL to generate the clock that drives the TX CMU PLL or an external reference clock input pin.

Figure 11–2. Achieving Deterministic Latency for the TX and RX Datapaths (1)

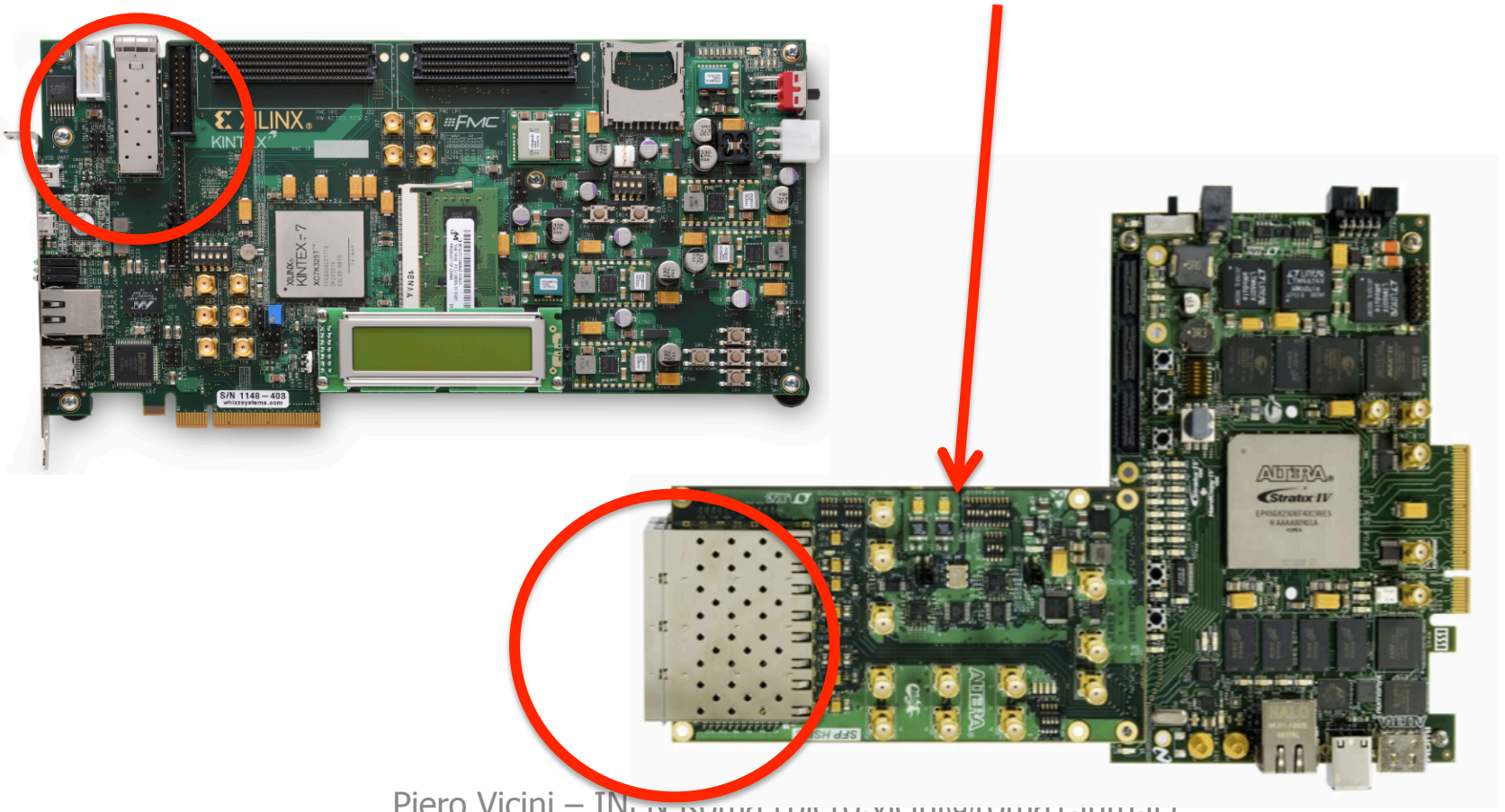


Note to Figure 11–2:
(1) The TX and RX Phase Compensation FIFOs always operate in register mode.

Interoperabilita' link seriali ALTERA vs XILINX (1)

Test-bed:

- DevKit ALTERA
- Scheda Kintex XILINX con canali off-board SFP
- mezzanino per board ALTERA (SFP channels verso la FPGA)



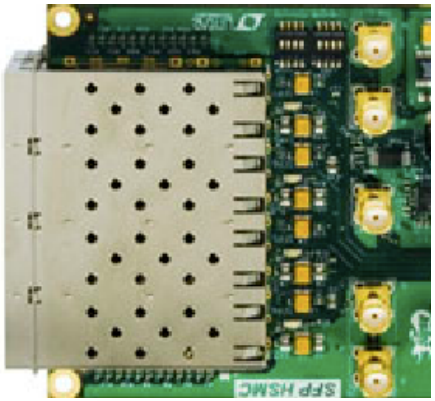
APE3Net hardware: SFP on APEnet+ (1)

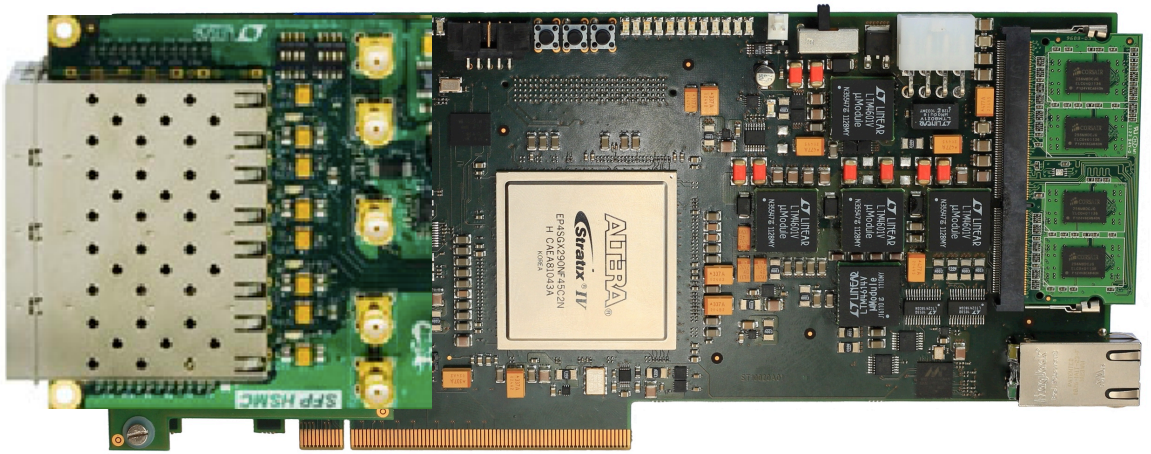
Realizzazione di un mezzanino modificato per APEnet+ che integra 4 canali SFP (+Ethernet?)

- Sul mezzanino attuale 8*2 link seriali a 8.5 Gbps verso i connettori QSFP+ presenti sul modulino (Z+, Z- del Toro).
- da confrontare con 4*2 canali@2.5Gbps (link "half_tower")

Interfaccia con PCIExpress (integrato nella FPGA sulla motherboard) e' Gen2,x8

- Latenza software e encoding hardware incluso oggi si puo' arrivare a valori > di 2.5 GB/s == 20 Gbps
- da confrontare con 10 Gbps del throughput integrato di 4 canali "half_tower"





APE3Net hardware: SFP on APEnet+ (2)

A regime

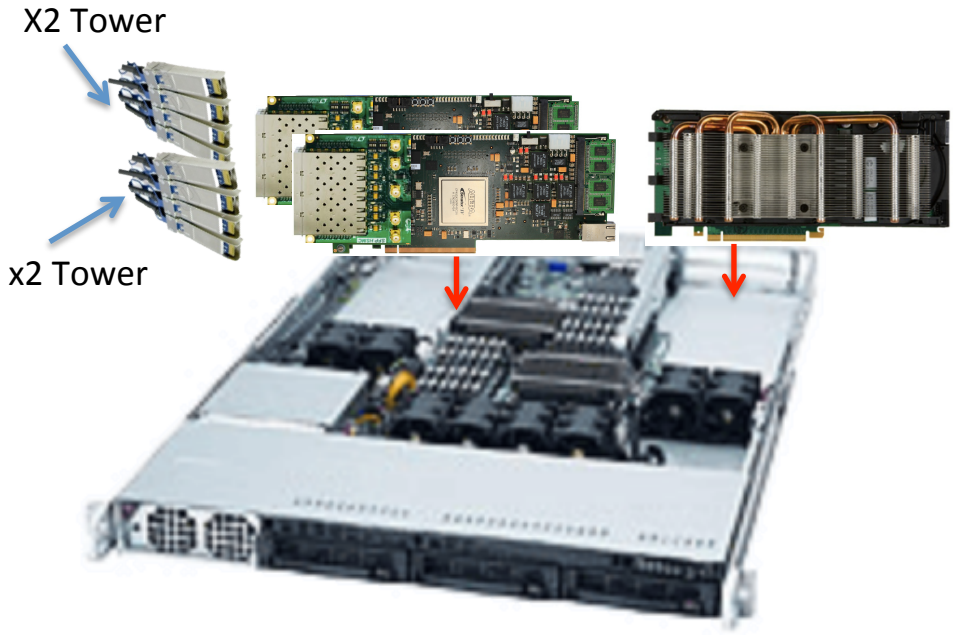
- Data throughput di 2 torri per scheda APE3Net
- 2 schede APE3Net per server Intel Xeon based "*DenseHM*" (DHM)
  - oggi 4 multi-processor Sandy Bridge (200 GFlops peak)
  - commodity network 10G Eth e/o Infiniband come dorsale del cluster
    - incluso slow control
  - Opzione: GPU aggiuntiva per DHM (1TFlops peak)
- 25 DHM per aggregazione completa di KM3 (100 torri da 14-16 piani)

Costi

- Calcolo al prim'ordine (preliminare….)
  - scheda APEKM3Net 3.5 KE (x2)
  - server rack mounted 2.0 KE (x1)
  - GPU Fermi/Kepler 2.0 KE (x1)
- Totale per DHM

  2*3.5 + 2.0 + 2.0 = 11 KE
- Totale per KM3

  25 * 11 = 275 KE



Attivita’ di R&D (software) di lungo periodo

- Leva sulla piattaforma computazionale basata sui DHM.

  Alcuni fatti:

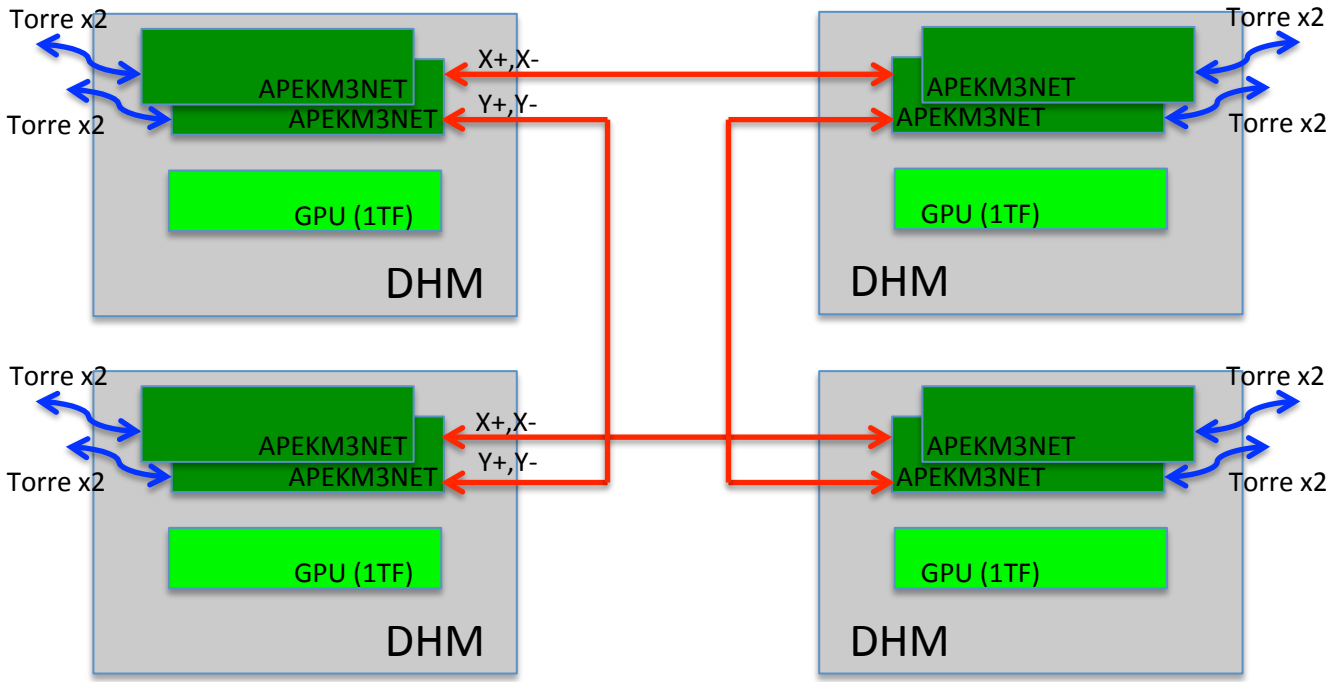
  - Il DHM cluster necessario per l’aggregazione dei dati e’ una piattaforma computazionale distribuita ad alte prestazioni
  - Esperimenti di uso di GPU per trigger di alto livello ai colliders HEP sono gia’ in corso e danno risultati incoraggianti

- Per partire serve:

  - Azione coordinata tra hw designers, software engineers e fisici computazionali esperti dell’applicazione
  - Analisi del carico computazionale necessario a vari livelli di aggregazione/selezione/trigger/event reconstruction
  - Coding di benchmarks significativi..

APE3Net: (nuovo) “computing model” (2)

- APE3Net come architettura di interconnessione unificata (computing network e read-out channels)
- Per KM3 sistema da 30 TFlops (!)
  - Rete toroidale bi-dimensionale di APE3Net (i canali X+,X-,Y+,Y-) accoppiata alla GPU e capace di operare sul flusso dati che proviene dai canali presenti sul mezzanino



Conclusioni

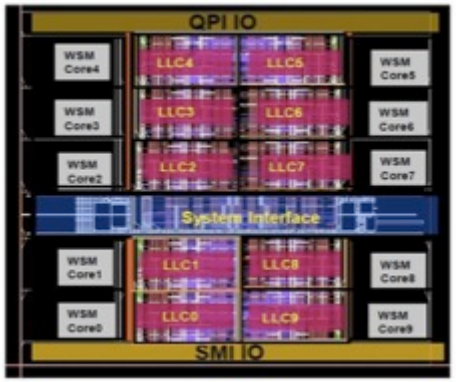
- Con la tecnologia di APE3Net+ si puo' realizzare un sistema di aggregazione completo per KM3 con numero ridotto (~1/10) di sistemi elettronici
  - Evidenti vantaggi sui costi di acquisizione e operativi
- Utilizzo "completo" di APEnet+ (read-out channels sul mezzanino e rete toroidale di interconnessione sulla motherboard) permette di esplorare architetture di elaborazione a terra efficenti e "non convenzionali"
- In una prima fase l'interesse e' (anche) solo scientifico/tecnologico
- In generale lo sviluppo hardware/software necessario a finalizzare il sistema e' relativamente limitato ma serve ovviamente il consenso ed il contributo di tutte le persone chiave

Thank you!

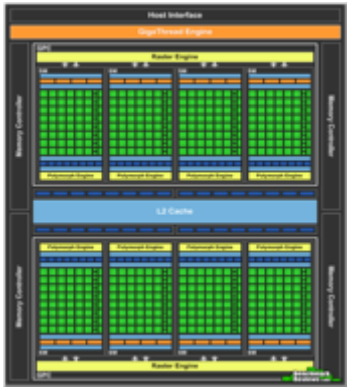
Any question or comment?

Peta (Exa?)Flops scale enabling technologies: GPGPU

- General Purpose Graphic Processing Unit: impressive peak performance (N*TFlops per chip)
- Videogames market i.e. 10 G$/yr unified gaming and HPC chip architectures
- Architecture and characteristics fit with HPC scientific application (LQCD as an example…) requirements
  - Many-Core (>>100) SIMD-like architecture
  - High local memory bandwidth
    - 140 GB/s -> 500 GB/s
  - Good for data parallelism (more than task parallelism…)
  - "Green" and cost effective



INTEL Westmere
+many caches - few processing



NVidia Fermi GPU
many computing units!!!

TOP500 NOVEMBER 2011

PRESENTED BY UNIVERSITY OF MANNHEIM · ICL UT INNOVATIVE COMPUTING LABORATORY · Lawrence Berkeley National Laboratory

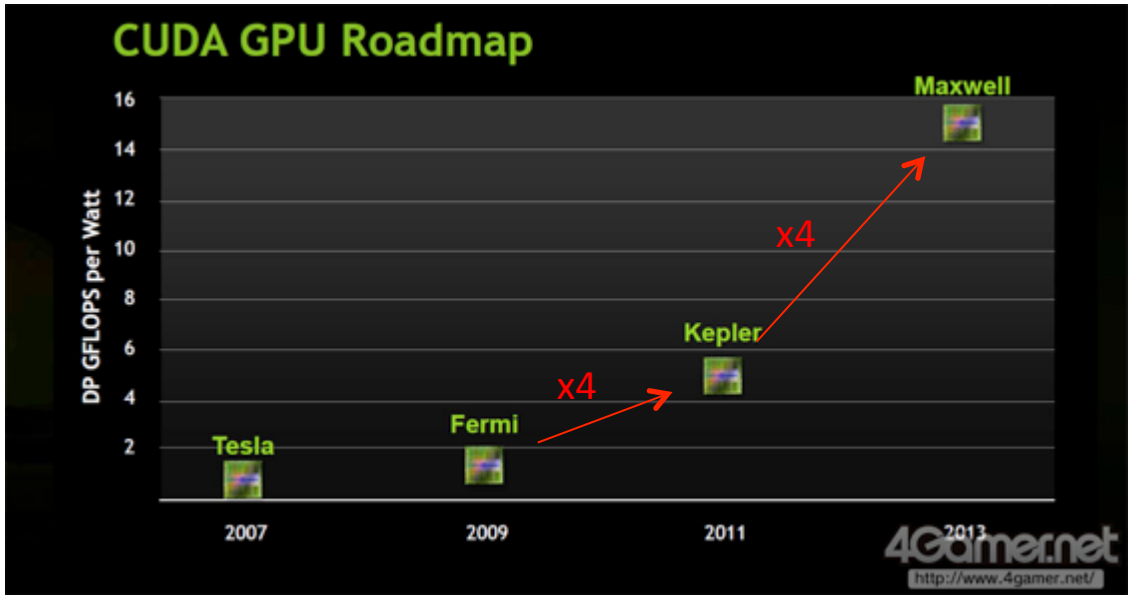
FIND OUT MORE AT www.top500.org

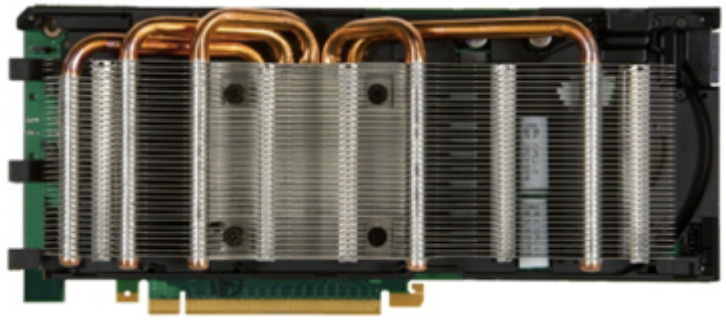
	NAME/MANUFACTURER/COMPUTER	SITE	COUNTRY	CORES	Rmax Pflop/s
1	**K computer** SPARC64 VIIIfx 2.0GHz, Tofu interconnect	RIKEN	Japan	705,024	10.5
2	**Tianhe-1A** 6-core Intel X5670 2.93 GHz + Nvidia M2050 GPU w/custom interconnect	NUDT/NSCC/Tianjin	China	186,368	2.57
3	**Jaguar** Cray XT-5 6-core AMD 2.6 GHz w/custom interconnect	DOE/OS/ORNL	USA	224,162	1.76
4	**Nebulae** Dawning TC3600 Blade Intel X5650 2.67 GHz, NVidia Tesla C2050 GPU w/ Iband	NSCS	China	120,640	1.27
5	**Tsubame 2.0** HP Proliant SL390s G7 nodes (Xeon X5670 2.93GHz) , NVIDIA Tesla M2050 GPU w/Iband	TiTech	Japan	73,278	1.19

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2026.48	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
2	2026.48	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
3	1996.09	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	170.25
4	1988.56	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	340.50
5	1689.86	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.67
6	1378.32	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	47.05
7	1266.26	Barcelona Supercomputing Center	Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090	81.50
8	1010.11	TGCC / GENCI	Curie Hybrid Nodes - Bullx B505, Nvidia M2090, Xeon E5640 2.67 GHz, Infiniband QDR	108.80
9	963.70	Institute of Process Engineering, Chinese Academy of Sciences	Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, NVIDIA 2050	515.20
10	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80

Peta (Exa?)Flops scale enabling technologies: GPGPU (2)

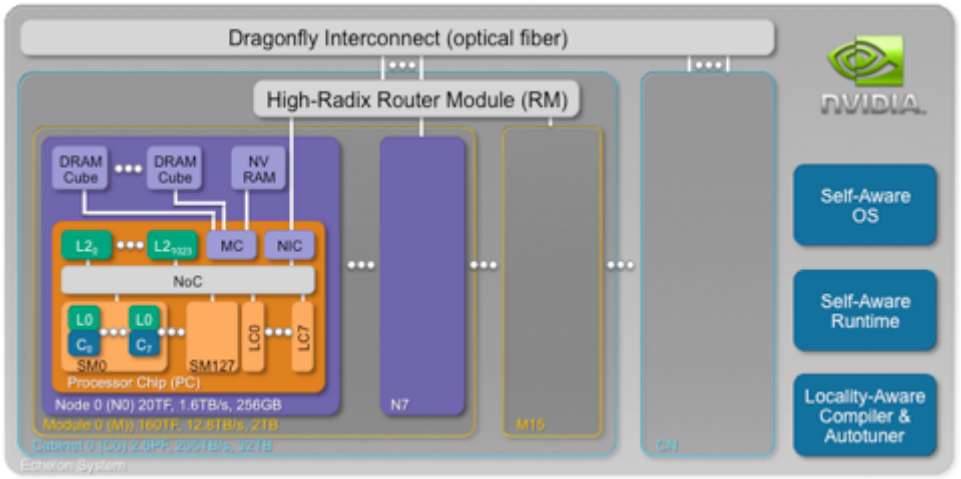
- Aggressive but (really!) feasible roadmap: much room for performance scaling





Nvidia Fermi (Tesla 20xx)

- (3 *10^9 transistors)
- ~500 core, 1 TF SP, 0.5 TF DP
- 6 GB external memory (150 GB/s)
- ~250W, <2K Euro

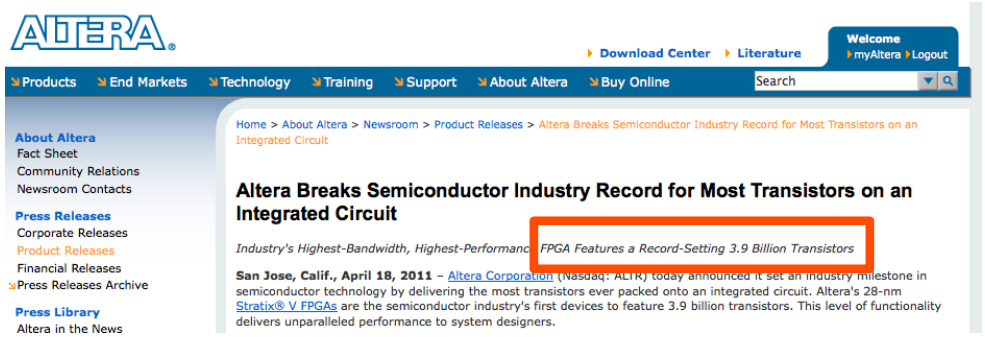


ECHELON: NVIDIA's Extreme-Scale Computing Project

- 128 SM (1024 core) 160 GFlops each, 20 TFlops aggregated
- Network: 150 GB/s; DRAM: 1.6TB/s
- 300 W

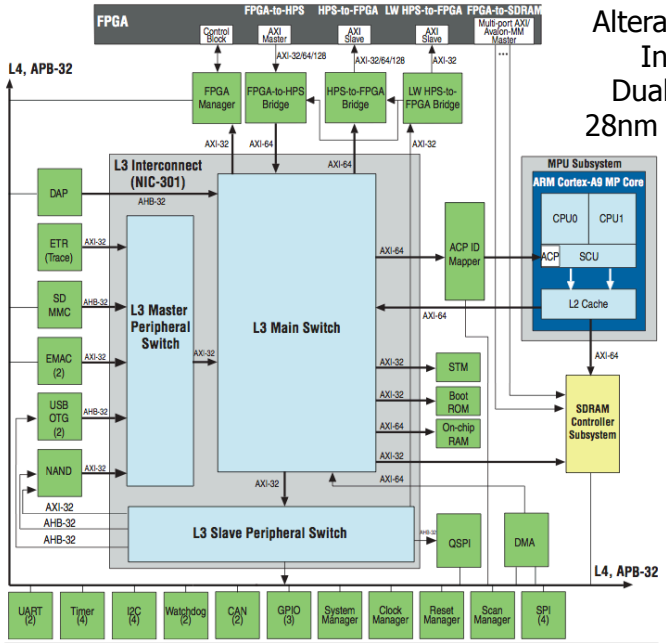
PetaFlops scale enabling technologies: FPGA

- High–end FPGA-based systems are the ideal hardware to build custom network

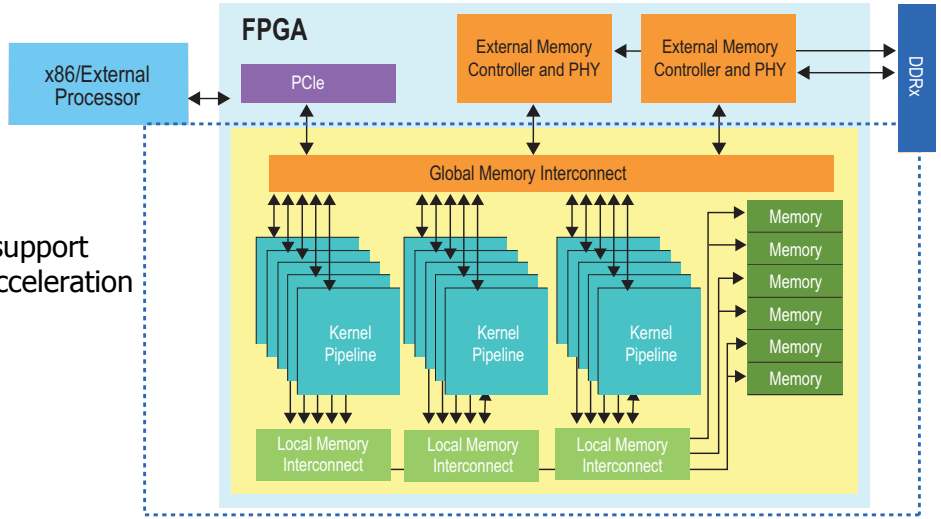


- Two main FPGA families: ALTERA STRATIX V – XILINX VIRTEX 7,
  - 28nm, introduction during 2012
  - Tflops, (multi)Terabits I/O bandwidth, hardIP uP cores

	ALTERA STRATIX V	XILINX VIRTEX 7
Logic Cell (up to)	1.05M	1.9M
Registers (up to)	1.58M	2.4M
Peak transceiver speed (Gbps)	14.1 28.0	13.0 29.0
Device serial bandwidth (Gbps)	936	930
PCIe interface	up to 4 Gen3,x8	up to 4 Gen3,x8
Memory interfaces	up to 7 banks x72 DDR3@800Mhz	DDR3 1.8Gbps
Embedded Memory size	55 Mb	65 Mb
DSP elementary block	4096 (Dual 18x18 Mul + 64bit Acc)	3960 (25x18 Mul + 48bit Acc)
I/O pins	up to 1125 (+ dedicated transceiver pins)	up to 1200 (+ dedicated transceiver pins)
Hardened IP	NIOS (proprietary) ARM A9 MIPS32 (soft core)	Dual ARM A9



Altera Embedded Initiative: Dual ARM9 on 28nm Altera FPGA



OpenCL support for system acceleration

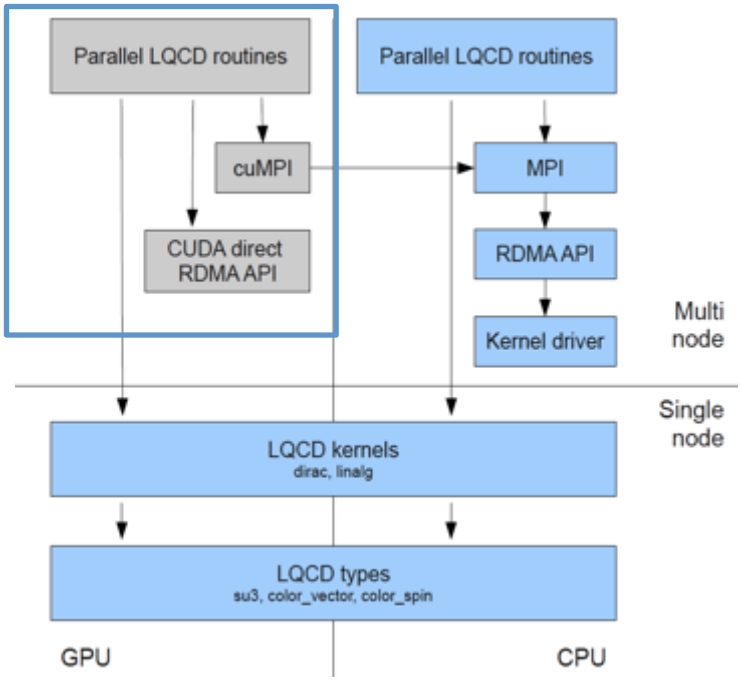
SW stack and GPU optimizations

APEnet programming model

- native RDMA API:
  - RDMA buffer registration: pinning and posting combined
  - single message transmission async queue
  - async delivery of completion events (both TX and RX)
- MPI:
  - early prototype
  - APEnet BTL module for OpenMPI
  - tentatively using GPU-aware feature in openmpi svn trunk

GPU centric programming model

- CUDA direct RDMA API:
  - Subset of APEnet+ RDMA API, available from within running GPU kernels
  - a running GPU kernel can emit requests for services to the host.



Scattered ideas and roadmap

- Short term (most of them already in place): High-End, low granularity *(2012 -2013)*
  - QUonG tower release to scientific community *(end of 2012)*
    - New INFN project SUMA (*"Supercomputing Massiccio"*)
  - "APEnet++": board based on next generations 28nm FPGA *(2H12–2013)*
    - PCIe Gen3, faster links, faster memories, faster logic
    - Evaluation started in 2H12, first release in 2013
  - NVIDIA joint activities
    - Optimization of P2P GPU-APEnet+ interconnection mechanism *(in progress)*
    - exploration of APEnet+ - Kepler "Peer-to-peer" integration *(2H12-…)*
  - EUROTECH joint activities
    - APEnet+ architecture integration (3D Torus) in AURORA systems
    - APEnet+ integration in future EURORA systems (AURORA accelerated with GPGPU)

Scattered ideas and roadmap(2)

- Medium term: Low-End, high granularity *(2013 -…)*
  - Porting of DNP IP on low-cost, low power FPGA
  - Exploration of DNP - ARM integration using
    - Qseven Tegra platform
    - SoC FPGA DevKits (Xilinx Zynq or Altera Arria/Cyclone)
  - VLSI version (hardcopy) (not yet scheduled)

- Long term: *(2014 - …)*
  - APEnet+ fault tolerant device
  - Release of SUMA medium-sized prototype
  - Customization from application specifications: GPU and "reconfigurable" DNP
    - Brain Simulation customized system based on enhanced DNP IP
    - APEnet+ optimized architecture for HEP computing systems
    - APEnet+ optimized architecture for BioComputing systems
    - ….
  - It should be possible to design a "Networked GPU"
    - GPU-DNP tight integration (same chip???)….