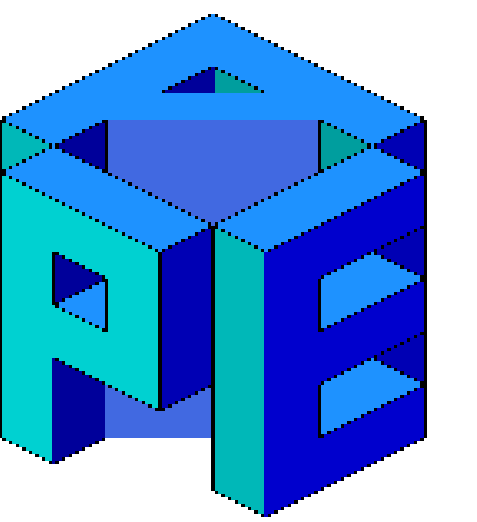


Il progetto APE: supercalcolatori per simulazioni scientifiche



La Lattice QCD

Dai suoi inizi negli anni '70, la Cromodinamica Quantistica su reticolo (o Lattice Quantum Chromo-Dynamics – LQCD) è annoverata tra le grandi sfide del calcolo scientifico, avendo spronato lo sviluppo di piattaforme di calcolo parallelo dedicate che negli ultimi trent'anni hanno sempre occupato le posizioni di punta della classifica mondiale di supercomputing. La LQCD è la teoria rinormalizzata di maggior successo del modello di QCD che descrive le forze intra-nucleari nel Modello Standard; viene definita in uno spaziotempo 4D discretizzato su un numero $V = L_x * L_y * L_z * L_t$ di punti reticolari ed è caratterizzata da un numero di operazioni che, scalando almeno come la sesta potenza della taglia lineare del reticolo fisico simulato, raggiunge facilmente numeri esorbitanti.

Al giorno d'oggi, le applicazioni di LQCD richiedono un numero di operazioni aritmetiche che arrivano a diversi PetaFlops (milioni di miliardi di operazioni in virgola mobile al secondo) nell'arco di un anno.

La storia di APE

Gli algoritmi usati nella LQCD hanno caratteristiche che hanno implicazioni immediate sulle specifiche dell'hardware sul quale vengono costruite le piattaforme di calcolo; le proprietà chiave sono:

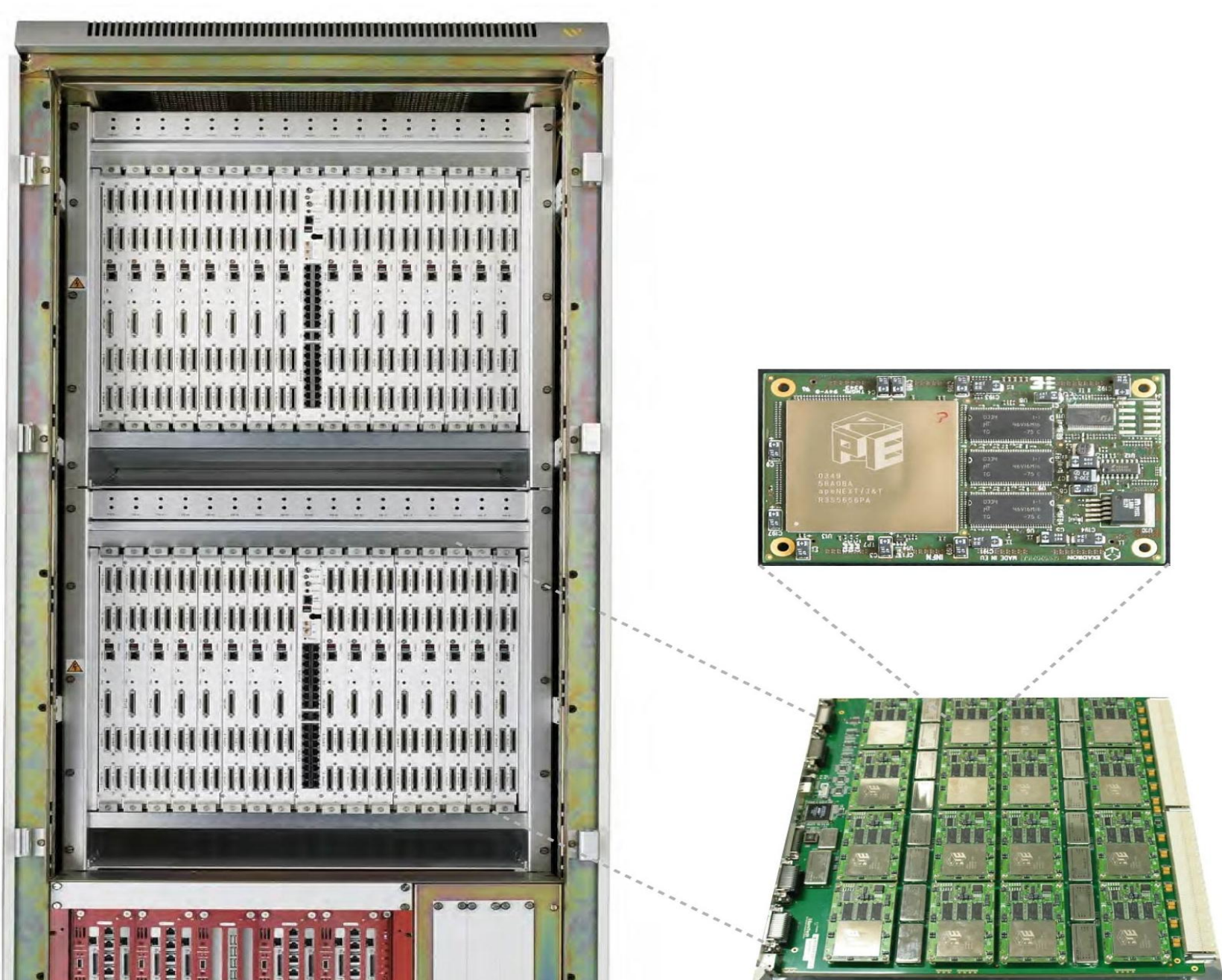
- un'unità per le operazioni aritmetiche elementari ad alte prestazioni ottimizzata per l'algebra matriciale;
- un'infrastruttura di rete a bassa latenza e ad alte prestazioni con topologia multidimensionale toroidale ed un'interconnessione ottimizzata a primi vicini.

Nel passato, questi paradigmi architetturali sono stati applicati alla progettazione della maggior parte dei supercomputer dedicati alla LQCD.

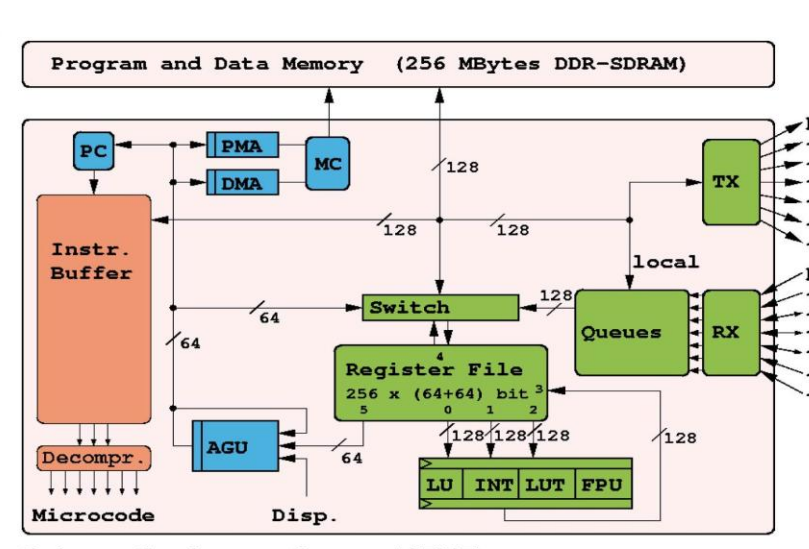
Lo **Array Processor Experiment (APE)** è una piattaforma di calcolo ad alte prestazioni dal design completamente proprietario e dedicato appunto alla Lattice QCD che ha avuto origine nell'Istituto Nazionale di Fisica Nucleare (INFN) in collaborazione con diversi altri istituti di fisica di tutto il mondo e che, a partire dal 1984, ha sviluppato ben quattro generazioni di macchine specializzate (**APE**, **ape100**, **APEmille** e **apeNEXT**).

Anno	Modello	Prestazioni
1988	APE	1 GigaFLOPs
1993	APE100	0.1 TeraFLOPs
1999	APEmille	1 TeraFLOPs
2004	apeNEXT	10 TeraFLOPs

Grazie al know-how acquisito nel networking e al fine di riutilizzare alcune soluzioni, il progetto collaterale chiamato **APENet** ha sviluppato una scheda di rete che consente di assemblare un cluster di PC a la APE con componenti commerciali.

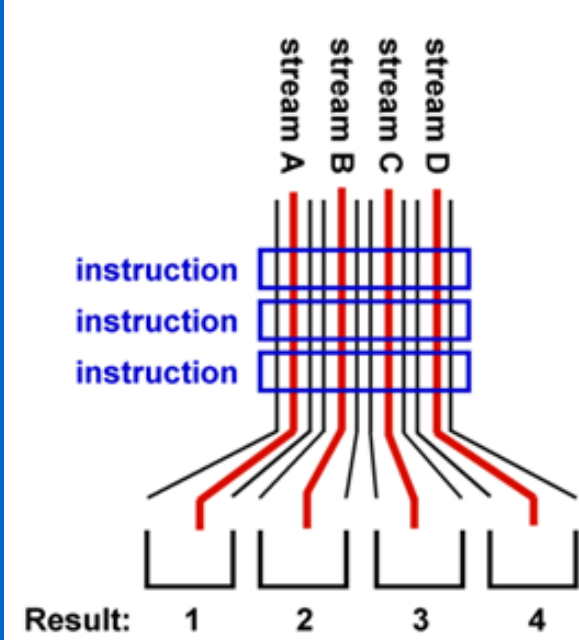


La "tower" di apeNEXT. 512 processori (array 3D) ~1 TeraFLOPs di picco prestazionale



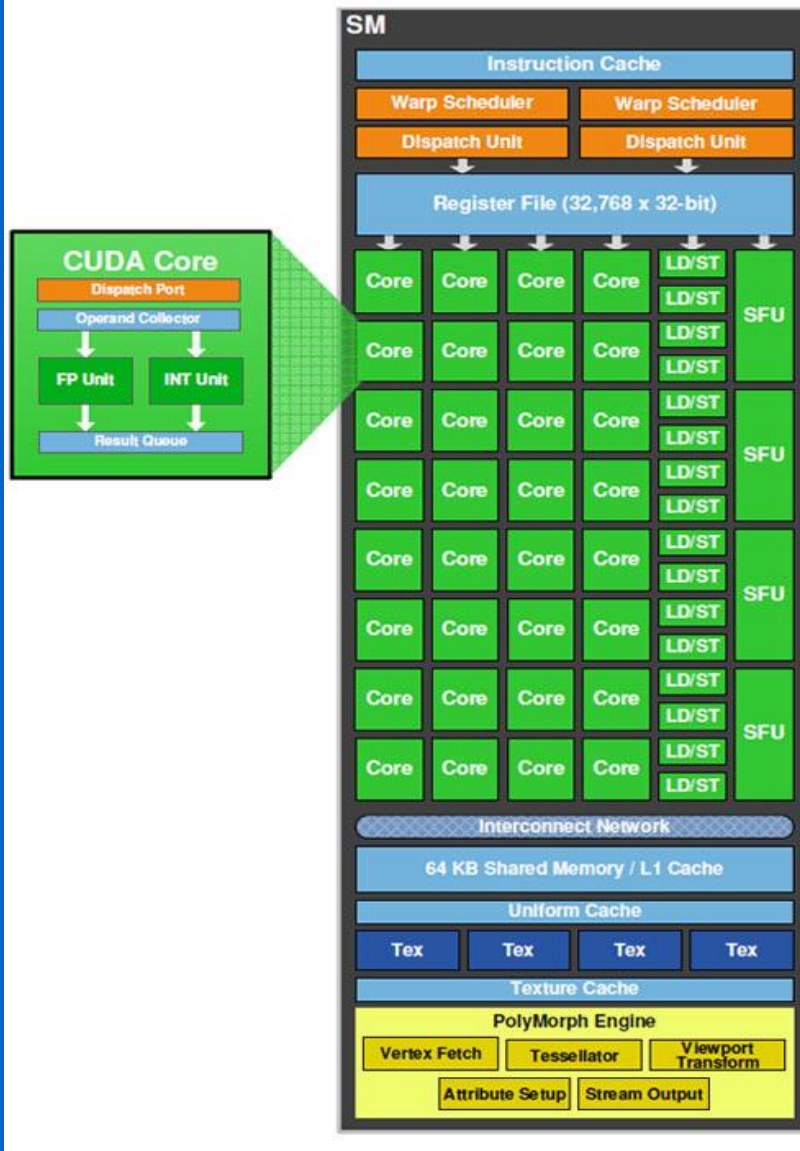
Graphical Processing Units (GPUs)

L'evoluzione dell'unità aritmetica in virgola mobile – dagli inizi come co-processore ausiliare dagli anni '80 fino alla sua completa integrazione nella CPU negli anni '90 – sta conoscendo una nuova fase grazie all'impiego delle comuni schede grafiche come unità di calcolo (le cosiddette GPU). Le generazioni più avanzate di tali piattaforme mostrano una maggiore adattabilità a molteplici ambienti applicativi (da qui la denominazione General Purpose GPU – GPGPU) grazie ad un'architettura che può essere programmata in modo simile a quella delle ordinarie CPU. Il vantaggio è che per certe classi specifiche di applicazioni scientifiche si ottengono prestazioni fino a due ordini di grandezza superiori – non a caso, tre dei primi cinque sistemi della classifica TOP500 sono basati su GPGPU.



Il paradigma architetturale delle GPU segue da vicino quello chiamato **SIMD (Single Instruction – Multiple Data)**, dove un alto numero di unità elementari (many-core) esegue il medesimo flusso di operazioni su stream di dati indipendenti.

Il numero di tali unità raggiunge ad oggi le diverse centinaia contro le decine di unità presenti in una tradizionale CPU multi-core.



La GPU si comporta come un processore del tutto indipendente dalla CPU, con una sua memoria ed un suo flusso d'istruzioni. Il modello di programmazione di una GPU è analogo a quello di una CPU tradizionale; l'ambiente vede l'uso di linguaggi "C-like" (CUDA o OpenCL) con estensioni che gestiscono le diverse gerarchie di memoria e lo I/O tra la GPU e la CPU host.

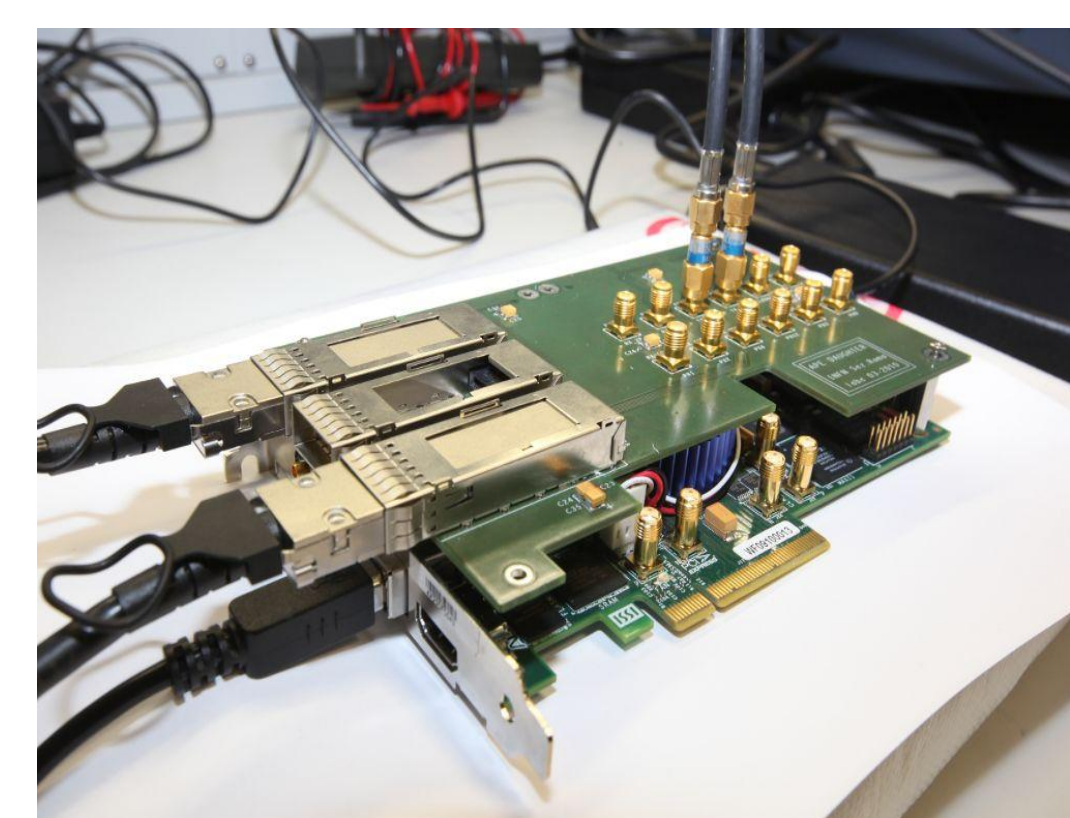
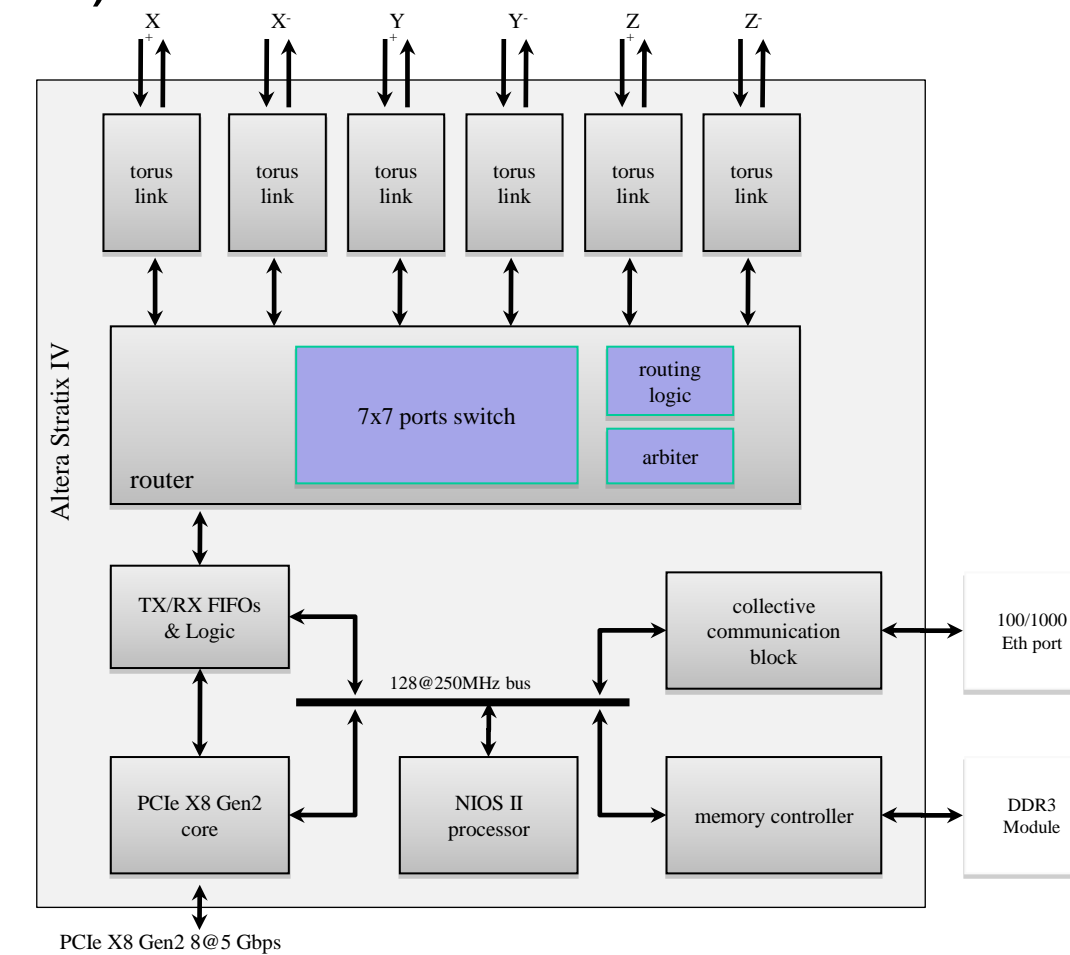
La scheda APENet+

La rete d'interconnessione è l'elemento critico per sistemi paralleli scalabili oltre il PetaFLOPs. **APENet+** conta di raggiungere tale scalabilità con una rete ad alta banda e bassa latenza in grado di connettere decine di migliaia di processori con un rapporto prestazioni/costo costante all'aumento del numero di nodi di calcolo. Le schede APENet+ comunicano tra loro con un protocollo proprietario basato su pacchetti serializzati ad-hoc, agganciandosi al sistema ospite con standard PCIe x8 v2.0 (4 GB/s banda di picco). L'instradamento dei pacchetti – protetti con codici a correzione d'errore – avviene con tecniche che garantiscono l'assenza di situazioni di stallo (deadlock).

L'hardware di APENet+ consta di schede PCI Express dotate di FPGA; ogni scheda è connessa direttamente ad altre sei con canali multipli ad alta velocità in una topologia ad array toroidale 3D; la FPGA – una Altera Stratix IV (EP4SGX290) – integra i sei canali costruiti con tecnologia di connessione QSFP+ (con banda fino a 34 Gbps).

Il lato software è coadiuvato nativamente dall'hardware e supporta il modello di programmazione RDMA (Remote Direct Memory Access). I driver della scheda sono disponibili per SO GNU/Linux, in varietà di basso livello con una API RDMA e di alto livello con una API Open-MPI.

Per sviluppare e testare il firmware, l'interfaccia PCI e l'interconnessione tra schede, è stato utilizzato un testbench costituito da un kit di sviluppo Altera (Altera Stratix IV GX 230) ed un mezzanino (realizzato dal LABE dell'INFN-Roma).

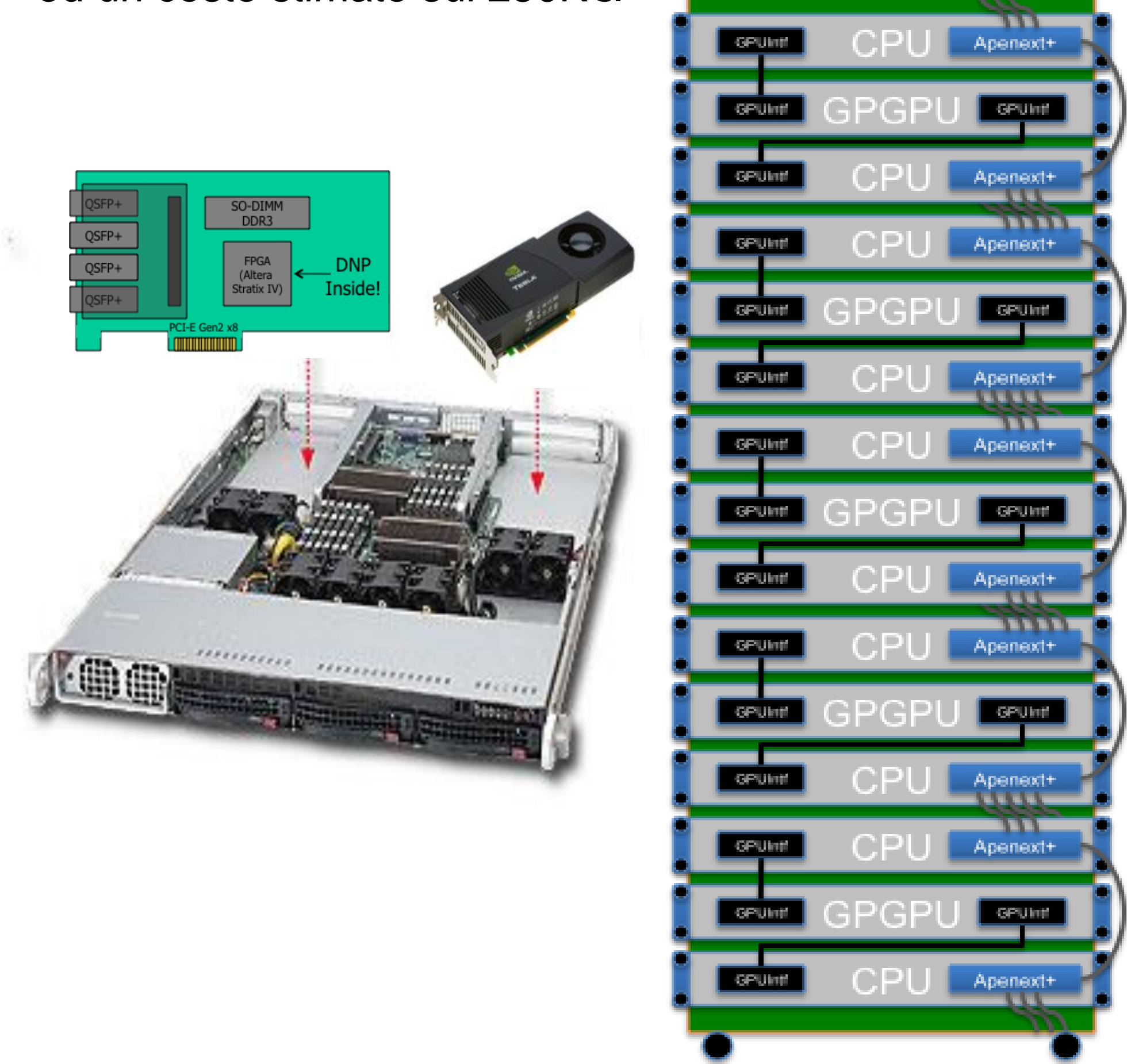


Il progetto QuOnG (QCD On GPUs)

Gli odierni e proibitivi costi di sviluppo per il design di un hardware completamente proprietario impongono l'adozione di soluzioni parzialmente commerciali. L'attuale linea di sviluppo di APE è il progetto **QuOnG (QCD On GPUs)**: una piattaforma parallela implementata da un cluster di PC con le seguenti caratteristiche chiave:

- l'elemento di calcolo elementare è un PC multi-core dotato di una moderna GPU;
- l'infrastruttura di rete si basa sulla scheda APENet+ integrante blocchi circuitali in grado di accelerare task specifici dell'applicazione.

Diverse soluzioni meccaniche attualmente allo studio ci permetteranno di integrare sistemi da più di 60 TeraFLOPs per rack con un consumo stimato di 25 KW ed un costo stimato sui 250K€.



Proposte di tesi

- Sviluppo di reti 3D toroidali su FPGA di ultima generazione**
 - HW: progettazione architettura e implementazione;
 - SW: sviluppo firmware e driver.

Keywords: 3D networks, FPGA, HDL languages, C/C++, GNU/Linux
- Sviluppo acceleratori hardware dedicati ad applicazioni di calcolo scientifico (LQCD, spiking neural networks,...)**
 - HW: progettazione e implementazione su FPGA;
 - SW: porting dei codici di riferimento per il test dell'hardware.

Keywords: ASIP, HDL languages, hw design tools, C/C++, Lattice Quantum Chromo-dynamics, Spiking Neural Networks
- Integrazione di reti di GPU**
 - HW: progettazione e implementazione dei componenti dell'interfaccia di rete;
 - SW: sviluppo driver, firmware e protocollo di comunicazione.

Keywords: GPU, networks, HDL languages, CUDA, C/C++, GNU/Linux
- Sviluppo e test applicazioni miste CPU-GPU (SW)**
 - LQCD
 - Spiking Neural Networks
 - Sistemi complessi (dinamica molecolare, vetri di spin, bio-computing...)

Keywords: GPU, CUDA, C/C++, Lattice Quantum Chromo-dynamics, Spiking Neural Networks, Izhikevic Neuron model, bio-computing, spin glasses, complex systems, molecular dynamics

References:

- "Computing for LQCD: apeNEXT", Computing in Science and Engineering, Vol.8, no.1, pp.18-29 (2006)
- "Computing for Lattice QCD: New developments from the APE experiment", Il Nuovo Cimento B, Vol.123, no.6-7, pp.964-968 (2008)

Contacts:

<http://apegate.roma1.infn.it/APE>
 alessandro.lionardo@roma1.infn.it
 davide.rossetti@roma1.infn.it
 pier.paolucci@roma1.infn.it
 piero.vicini@roma1.infn.it