Calcolo Teorico + SUMA

R. (lele) Tripiccione
tripiccione@fe.infn.it

Riunione Commissione Calcolo
Frascati, 25 Ottobre 2012

Overview

La storia recente (e le sue lezioni)

La situazione attuale

Prospettive per i prossimi 2-4 anni

I centri di calcolo / Le iniziative della EU

Il progetto Premiale SUMA

Conclusioni

La storia recente ...

Guardiamo indietro
(di circa 10 anni)

Numerosi piccoli cluster gestiti "informalmente" dalle singole Sezioni.

Calcolo massiccio (LGT) in gran parte basato su APExx, che era ancora competitivo rispetto alla concorrenza internazionale

Risorse limitate nei Centri di Calcolo e accesso difficile e inaffidabile.

La storia recente ...

… e le sue lezioni

Situazione globalmente accettabile, ma ….

Divisione manichea tra calcolo medio e calcolo massiccio

Minima sinergia tra le varie sezioni e programmazione quasi inesistente

Spreco di risorse (soprattutto) umane

Chiusura (pressoche' totale) dell' INFN verso le collaborazioni esterne in questo campo

La storia recente …

Guardiamo indietro
(questa volta di 5 anni …)

Tentativo di coordinamento e centralizzazione del calcolo medio (CNAF)

APE ormai marginale rispetto alle risorse di calcolo disponibili In Europa

Crescita fortissima dell' offerta di calcolo scientifico da parte dei Centri di Calcolo Europei (e – in parte – italiani)

La situazione attuale (fine 2011)

Calcolo medio basato in parte significativa sul cluster di Pisa (TheoPhys + TheoNuc)

Parecchi aspetti positivi

Qualche problema ancora da risolvere

Gruppi italiani di LGT quasi totalmente dipendenti dalle risorse di calcolo dei propri collaboratori esteri

Significativi sviluppi tecnologici ottenuti dai figli di APE (Aurora, Quong), ma impatto del tutto trascurabile sulla comunita' teorica

Prospettive per i prossimi 2-4 anni

Una visione strategica

Tenere aperta una "fast lane" per il calcolo medio nei limiti del possibile

Rendere piu' facile e significativo l' accesso ai grandi centri di calcolo (in Italia e in Europa)

Aiutare tutta la comunita' teorica a utilizzare al meglio le nuove architetture di calcolo che inevitabilmente dovremo utilizzare

Rendere utilizzabili per il calcolo teorico i progressi tecnologici fatti all' interno dell' INFN

Innestare un circolo virtuoso di collaborazione con i centri di calcolo nazionali ….

SUMA

Qualche numero ….

1 core di calcolo: *10 → 15 Gflops*

1 nodo di calcolo: *40 → 200 Gflops*

		#core	Int su un a		FTE
TheoPhy		1000 core	8 Mcore		1 … 2
Siss		10000 core	80 Mcore		3 … 6
Cineca BG/Q	10000 nodi	160000 core	1.3 Gcore-hour	~20 Meuro	10 … 20
Juqeen	8192 nodi	132000 core	1.1 Gcore-hour	?	
SuperMUC	18432 nodi	147000 core	1.2 Gcore-hour (1.5x)	?	

Qualche numero ….

1 core di calcolo: *10 → 15 Gflops*

1 nodo di calcolo: *40 → 200 Gflops*

	# nodi	#core	Int su un anno	Investimento	FTE
TheoPhys	100 nodi	2000 core	16 Mcore-hour	300 Keuro	1 … 2
Sissa	1000 nodi	10000 core	80 Mcore-hour	3 Meuro	3 … 6
Cineca BG/Q	10000 nodi	160000 core	1.3 Gcore-hour	~20 Meuro	10 … 20
Juqeen	8192 nodi	132000 core	1.1 Gcore-hour	?	
SuperMUC	18432 nodi	147000 core	1.2 Gcore-hour (1.5x)	?	

Qualche numero ….

1 “grosso” utilizzatore di TheoPhys

→ *200-300 Kcore-hour/anno*

1 grossa collaborazione di LQCD (e.g. ETMC)

→ *100 Mcore-hour / anno*

I centri di calcolo

*A livello **europeo** i grossi centri di calcolo sono strutturati nella organizzazione PRACE, che mette a disposizione ingenti risorse di calcolo sulla base del merito scientifico (peer – review)*

PRACE-Tier0:
tipicamente due call all' anno
1 Gcore-hour per call
Procedura lunga e complessa ….
… referaggio scientifico e tecnico …
… 6 mesi tra la scadenza della call e la effettiva disponibilita' di tempo macchina

Media del 2011-2012 : 30% del tempo per LQCD
Fair enough…

I centri di calcolo

*A livello **italiano** il Cineca gestisce una struttura analoga, chiamata ISCRA*

ISCRA-C	*< 2 Mcore-hours*	*fast and easy (once per year..)*
ISCRA-B	*< 5 Mcore-hours*	*3 mesi / ragionevole ...*
ISCRA-A	*> 5 Mcore-hours*	*incasinato come PRACE*

Per chi ha necessita' di grandi risorse di calcolo la combinazione di ISCRA + PRACE e' inevitabile, con tutti i suoi limiti

PRACE+ISCRA-A per progetti veramente grossi
ISCRA-C per progetti "piccoli"

Solo i centri di calcolo???

L' utilizzo dei centri di calcolo e' inevitabile per il prossimo futuro....

Rimane il problema di un accesso ***rapido e flessibile*** *senza troppi balzelli e orpelli al calcolo....*

Due soluzioni:
Per il calcolo massiccio → Accordo diretto INFN-Cineca
Per il calcolo medio → Il cluster INFN (+accordi Sissa)

SUMA

Accordo INFN-Cineca

Accordo discusso tra varie difficolta' in primavera 2012, in vista dell' istallazione del Blue-Gene/Q

100 Mcore-hours su BG/Q a disposizione dell' INFN tra "giugno" 2012 e giugno 2013, da suddividere tra i vari gruppi interessati.

Coordinamento informale (S. Simula, LT + S. Bassini[Cineca])

Accordo Cineca-INFN

IS	Responsabile	Mcore-hour – prod.	Mcore-hour – test	TOTAL
MI11	Di Renzo	10	2	12
PI11	Pelissetto	10		10
PI12	D' Elia	16	2	18
RM123	Simula	30		30
PD32	Viviani		2	2
OG51	De Pietri	3	2	5
TV62	Mazzino	8	5	13
Grand Tot		**77**	**13**	**90**

Interesse anche di PR21, MB31, RM61, TO61

Guardiamo al prossimo futuro ….

.

Il progetto premiale SUMA

We have drunk suma and become immortal;
We have attained the light, the Gods discovered.
(Rigveda 8,48,3)

One cubic centimetre cures ten gloomy sentiments.
(A. Huxley,
Brave New World, 1932)

Il progetto premiale SUMA

Obbiettivi del progetto - 1: ------>

Fare quello che serve alla comunita' teorico computazionale nei prossimi 3 anni.

Il progetto premiale SUMA

Obbiettivi del progetto - 2: per il grande calcolo --->

WP1: Imparare a utilizzare i processori e i sistemi di nuova generazione in contesti tipici della comunita' teorica INFN, visto che prima o poi tutti li dovremo utilizzare....

WP4: Rendere utili per il calcolo teorico i risultati ottenuti dai progetti di sviluppo per il calcolo teorico che l' INFN ha supportato negli ultimi 4-5 anni ... e non disperdere questo know-how

WP3: Istallare un "large prototype" basato su questo tipo di processori (proof-of-concept di una futura macchina per il calcolo scientifico ad alte prestazioni E ANCHE WORKHORSE!!!)

Il progetto premiale SUMA

Obbiettivi del progetto - 3: per il calcolo medio --->

WP2: Garantire la continuita' del cluster di Pisa, che e' una facility per tutta la comunita' teorica e una fast lane per quello che poi diventera' calcolo massiccio

Guardiamo al prossimo futuro ….

Ricordate??

1 core di calcolo: 10 → 15 Gflops
1 nodo di calcolo: 40 → 200 Gflops

*Un nodo di calcolo di “**prossima generazione” ~ 2000** Gflops*

GPU (Nvidia) – MIC (Intel) - ?

Grazie ad un sostanziale aumento del parallelismo del processore

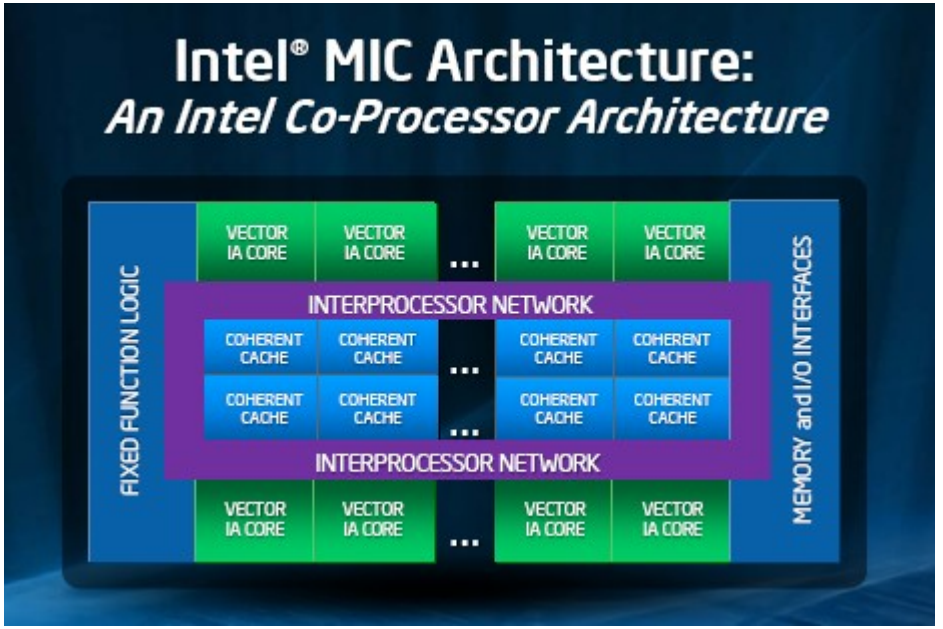
Intel MIC processors ….

Un numero alto (ma non altissimo) di core massicci

e.g: 64 cores x 32 Gflops

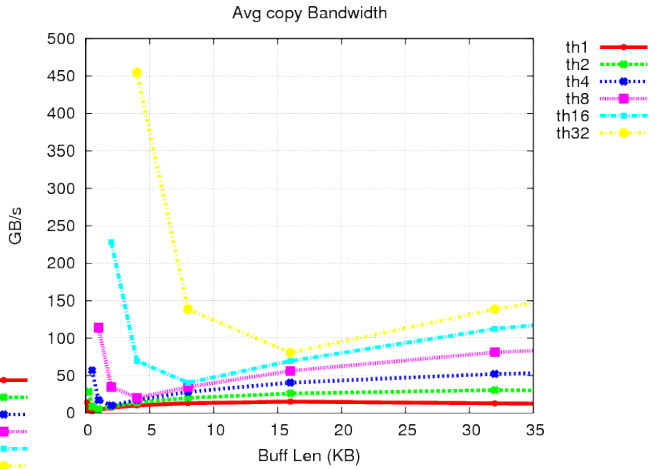
Tecniche di programmazione "relativamente simili" a quelle attuali ……

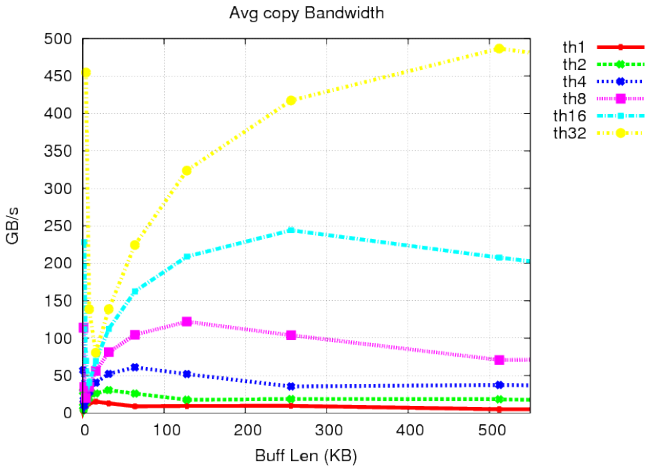
… ma avere buone prestazioni non e' facile →

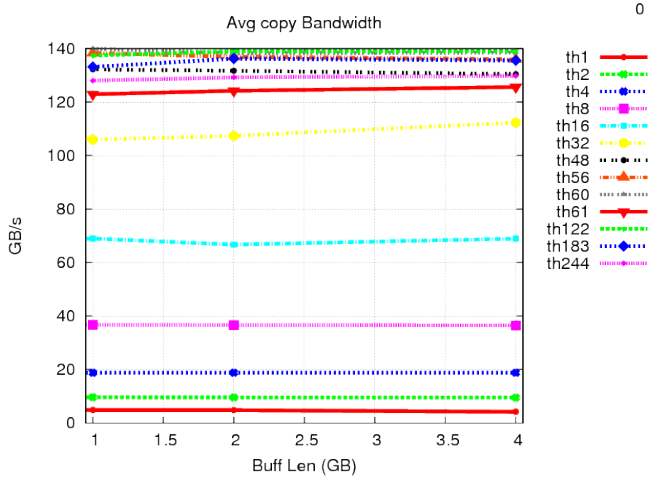


Intel MIC processor: Basic benchmarks

Un analisi accurata della banda processore - "memoria"

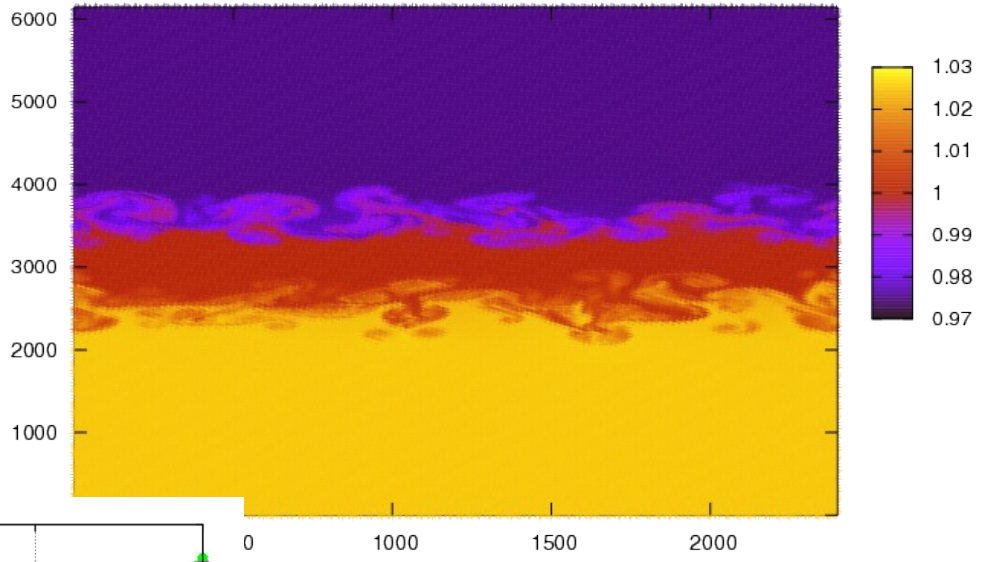


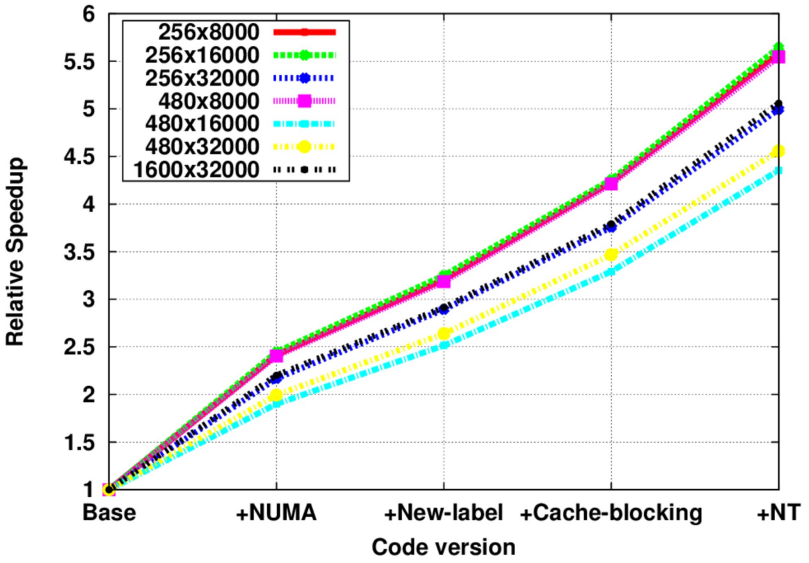




Intel MIC processors e LB

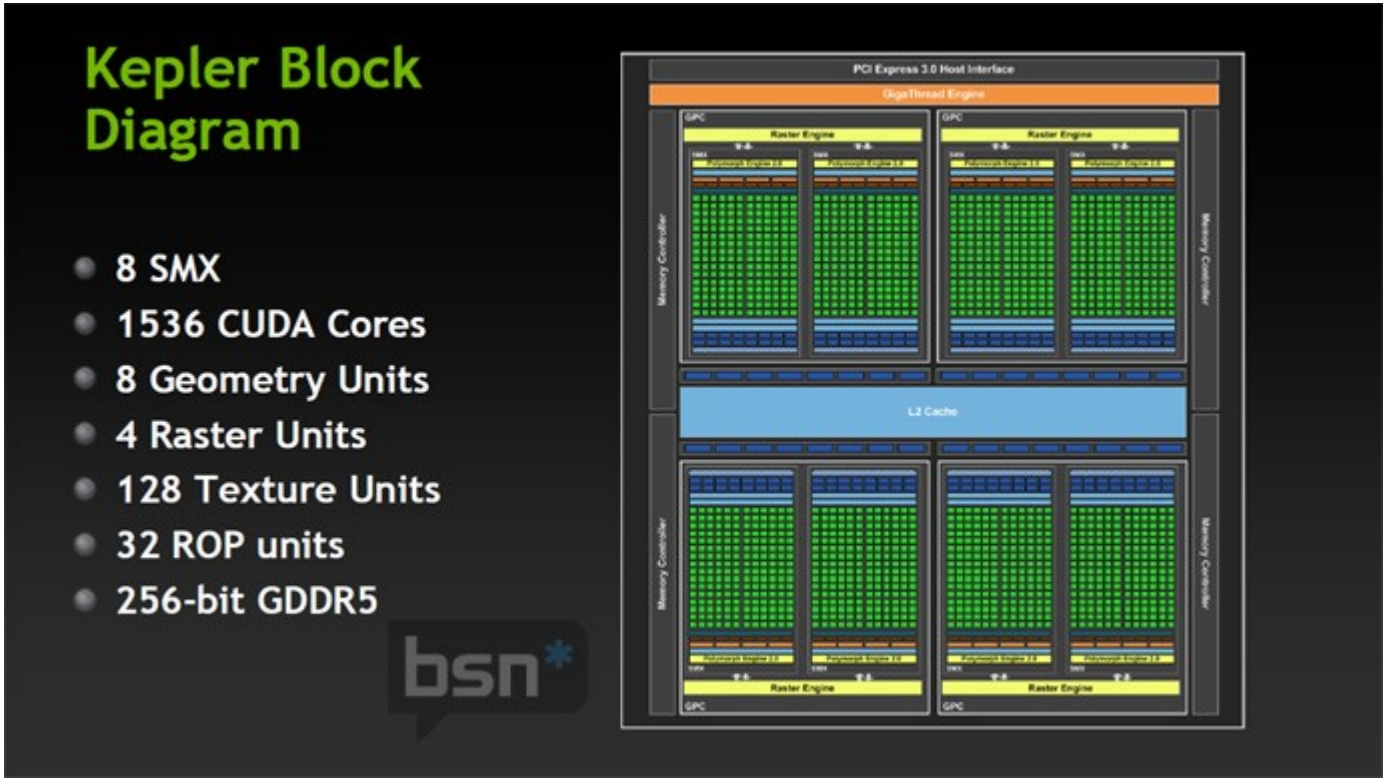
Uno schema di calcolo tra
I piu' friendly per il
Calcolo massicciamente
Parallelo …
(e una palestra per la
LQCD)





NVIDIA GPU processors ….

Un numero molto alto di core di calcolo molto semplici,
Programmati con linguaggi ad hoc (CUDA) friendly ed efficaci
ma con un comportamento spesso caotico.



Better computers than those you can buy?

What we need is simple to achieve in terms of computer architecture

Basic physics help us in two ways:

1) Parallellelism is available (and "easy" to expose) ...
... and parallel computing is the physics sponsored way to compute:

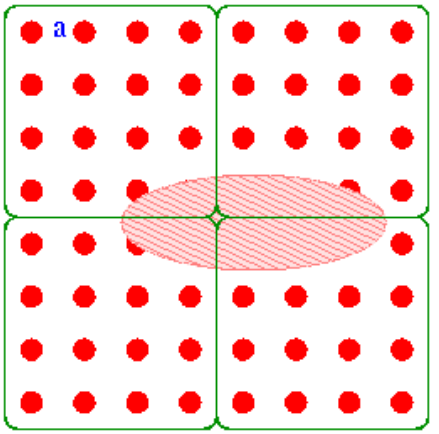
The basic object is the transistor
Industry learns to build smaller and smaller transistors. As $\lambda \to 0$
obviously $N \propto 1/\lambda^2$ *but speed scales less favourably* $\tau \propto \lambda$

Trade rules: perform more and more things in parallel
rather than a fixed number of things faster and faster

Better computers than those you can buy?

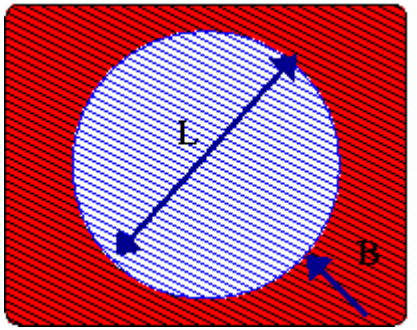
What we need is simple to achieve in terms of computer architecture
Basic physics helps us in two ways:

2) Physics transfer information locally
This has to go over to the computer structure ->
Keep data close in space to where it is processed



Failure to do so will asymptotically
bring a data bottleneck:

$$B(L) \propto L$$
$$P(L) \propto L^2$$



Sviluppi interni all' INFN

Negli ultimi 3 o 4 anni, in ambito INFN gli sviluppi tecnologici relativi all' HPC si sono concentrati sulla rete di interconnessione

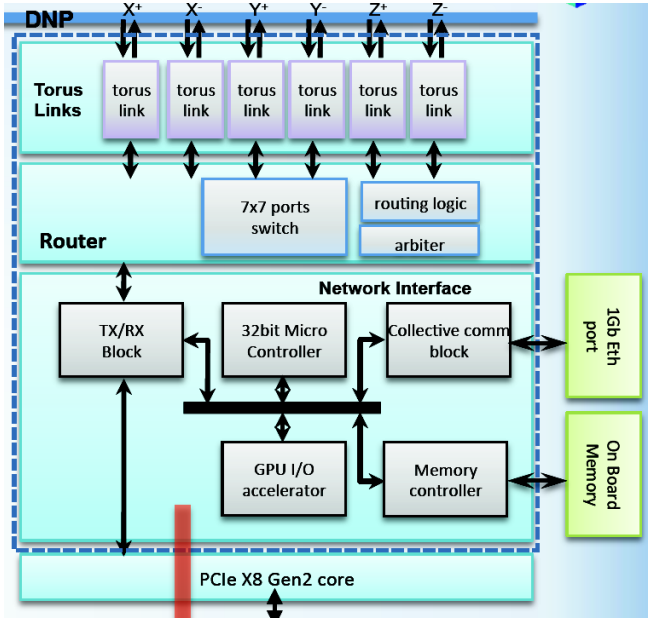
Eredita di APE ------>

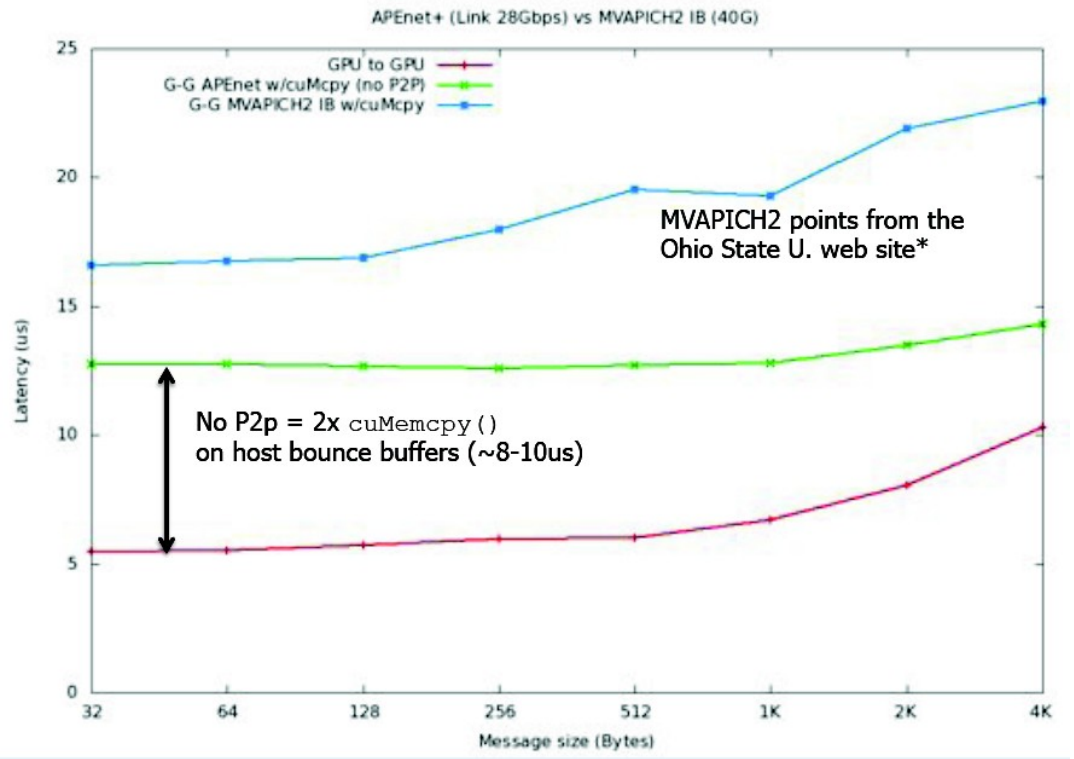
APEnet+ QuonG

AuroraScience

APEnet+ QUonG

Una rete di interconnessione toroidale 3D fortemente integrata con le GPU





APEnet+ QUonG

Un sistema sperimentale di dimensioni significative disponibile a fine di quest' anno.

- 16 CPU
- 32 GPU
- 30 … 60 Tflops peak

- Disponibile per sperimentazione in SUMA !

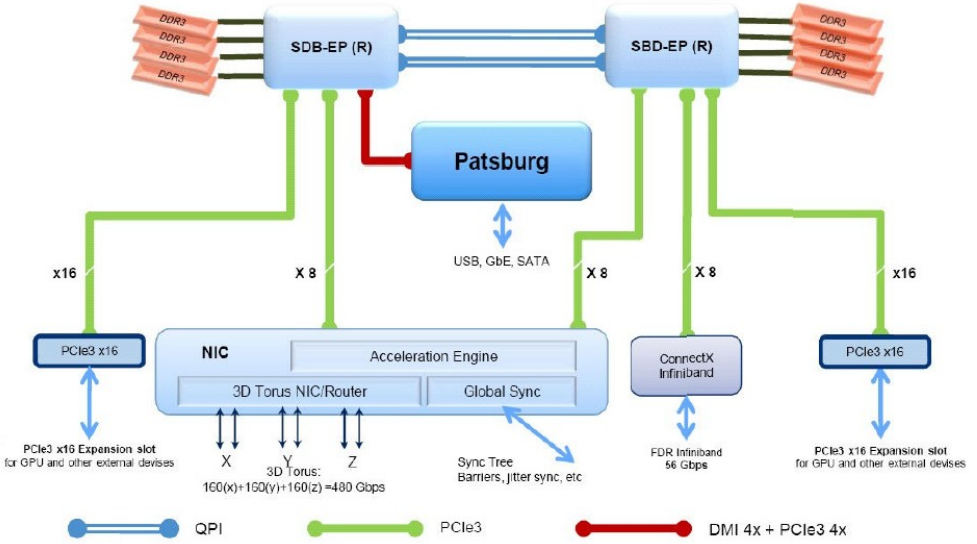
AuroraScience → Eurora

AuroraScience e' stato un progetto congiunto INFN – FBK con Eurotech come "partner" commerciale

Da Aurora e' derivato Eurora:

"Farina del sacco Eurotech"

MIC o GPU associate ai nodi
Ingegnerizzazione sofisticata
(pro e contro ...)



Aurora → Eurora

Eurora e' in corso di istallazione al CINECA:

128 processori Sandy-Bridge (1024 core)

128 processori MIC

120+ Tflops (peak, double,MIC-only)

Infiniband + Torus Network 

Disponibile per sperimentazione in SUMA !

-

- WP1 -

Computing strategies and algorithms for existing and new architectures

Per selezionati algoritmi di calcolo di rilevanza in fisica teorica → CLS, ETMC ….

Adattare gli algoritmi di calcolo, e ottimizzare i relativi programmi per le nuove architetture di calcolo, e valutarne le prestazioni

Valutare l' impatto e ottimizzare le prestazioni della rete di comunicazione

- WP1 -

Computing strategies and algorithms for existing and new architectures

Per selezionati algoritmi di calcolo di rilevanza in fisica teorica → CLS, ETMC, LBM ….

Adattare gli algoritmi di calcolo, e ottimizzare i relativi programmi per le nuove architetture di calcolo, e valutarne le prestazioni

Valutare l' impatto e ottimizzare le prestazioni della rete di comunicazione

…….

- WP1 -

Computing strategies and algorithms for existing and new architectures

Per selezionati algoritmi di calcolo di rilevanza in fisica teorica → CLS, ETMC, LBM

…. quali colli di bottiglia sono gestibili

Quali sono i limiti intrinseci dei processori

In che modo le invenzioni INFN possono aiutare

- WP4 -

Advanced Development

Valorizzare gli sviluppi tecnologici in ambito INFN e svilupparli ulteriormente.

Short Term (1 anno) + Long Term (2 … anni)

- WP4 -

Short Term

APEnet+ / Eurora →

test in un ambiente di relativamente grandi dimensioni.
Firmware/software ottimizzato per la connessione diretta P2P

Utilizzi non convenzionali di APEnet+

---------->

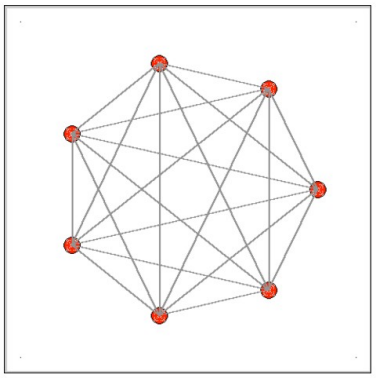
- WP4 -

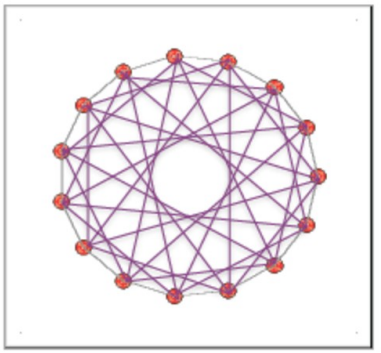
Short Term (1 anno)
Il "Tamburo" → una macchina totalmente interconnessa ottimizzata per la dinamica molecolare

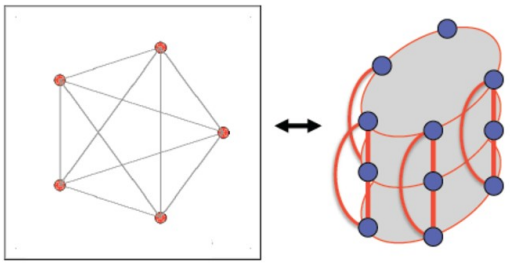
Un sistema relativamente piccolo con efficienza 4x – 8x rispetto ai sistemi "tradizionali"

7 … 15 nodi x 2 GPU x 2 Tflops → 60 Tflops

Sistema ottimale per (e.g.) Quantum Espresso (SISSA)







- WP4 -

Long Term

Re-ingegnerizzazione di APEnet+ utilizzando nuove generazioni di FPGA → PCIe Gen 3 / Link "veloci": 28 Gbit/sec

Innovazioni architetturali di rete:

- *Piu' efficiente gestione del RDMA*
- *Offload sulla rete di operazioni di calcolo*
- *Fault tolerance + fault recovery*

- WP3 -

"Large" prototype of a state-of-the-art system

Dopo un anno di esperienza →
Provare a scommettere su una promettente struttura di macchina
E realizzare un "large prototype"

E.g. 128 – 256 MIC/GPU x 2 Tflops → 250 – 500 Tflops peak

Istallato al CINECA →

Ancora insufficiente per essere autosufficienti in LQCD …

… ma (forse in grado di dare un boost significativo)

- WP3 -

“Large” prototype of a state-of-the-art system

I termini della scommessa:
- qual'e' l' efficienza sustained dei nuovi processori?
- con che metodologia si ottimizzano I programmi?
- quanto si puo' migliorare l' accesso diretto alla rete?
- si puo' offloadare all' interfaccia di rete parte del calcolo?
- in che contesto una rete toroidale e' competitiva rispetto a Infiniband?

Tutte domande la cui risposta oggi non e' nota ….

- WP2 -

TheoPhys/TheoNuc svolgono due ruoli distinti e importanti

Garantire le risorse di calcolo "medie"

Tenere aperta una fast lane per sperimentazioni veloci di calcolo massiccio

Cofunded dalla CSN4 ….

Prospettive di sviluppo del cluster

HARDWARE:

Incremento di potenza di calcolo di un fattore 4x rispetto al cluster attuale

Aumento della quantita' di memoria disponibile per core di calcolo

- WP2 -

MODALITA' PIU" FLESSIBILI DI UTILIZZO

Mantenere l' accesso via Grid

Aggiungere accesso diretto tramite autenticazione AAI + code di calcolo batch (forse piu' friendly, piu' adatto per il parallelismo "massiccio")

Risolvere alcuni problemi legati allo storage

Accesso comune al cluster della SISSA (se colonizziamo la SISSA al 10% abbiamo raddoppiato la nostra disponibilita' di calcolo...).

Struttura del progetto

Al momento :

Ba / Cs / Fe / MiB / Pi / Pr / Roma1 / Roma2 / Roma3

15 persone core team + Post-Docs …

Progetto di durata triennale (1.1.2013 → 31.12.2015)

Struttura del progetto

Steering committee provvisorio: 1 componente per ogni sezione coinvolta

Rapida transizione →

Steering committee composto dai coordinatori dei vari WP

3 – 4 riunioni globali del progetto all' anno.

Contribuire a Summer School in fisica computazionale

Money money money ….

Budget richiesto: 2200 K

Budget Finanziato: 1925 K

Break up "di principio" del budget (sblocco incrementale → aggiustamenti in corso d' opera)

Cosa	Quanto
Assegni ricerca	600K
TheoPhys	220K
Large Proto	600K
Sviluppi Tecn	325K
Fisiologia	180K
Totale	1925K

Money money money ….

9-10 assegni di ricerca biennali @ 30K/anno	→	*600K*
Sviluppi avanzati	→	*325K*
Large prototype	→	*600K*
Cluster Upgrade	→	*220K*
(+ 80K dalla CSN4)		
Fisiologia (su 3 anni)	→	*180K*
Totale	→	*1925K*

In conclusione …

No conclusions …

Questions / Comments ….