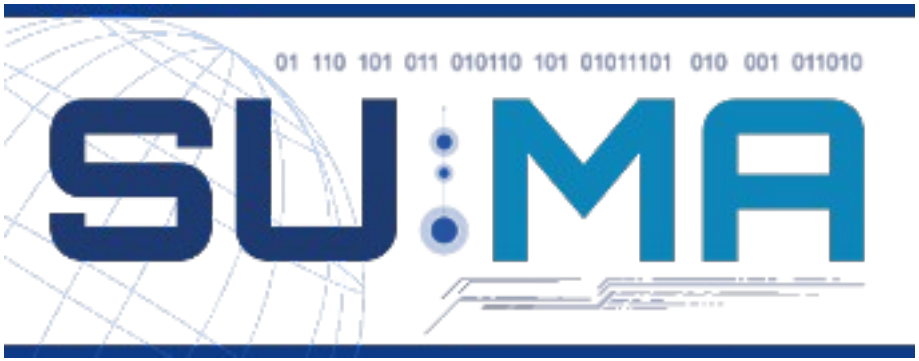
SUMA a un anno di eta'...

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

Riunione CCR
Roma – 11 ottobre 2013

web2.infn.it/SUMA ----->



SUMA: what does it mean?

We have drunk <u>suma</u> 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)

Today's menu

Important dates

La visione alla base del progetto

Cosa possiamo/non possiamo fare

Come pensiamo di farlo

Cosa abbiamo fatto nel primo anno di attivita'

Cosa ci aspettiamo di fare nel 2014

Situazione finanziaria

Long term thoughts *(Horizon2020, nuovi premiali...)*

Important Dates

~ Dicembre 2011: *preparata la proposta*

~ Giugno 2012: *(voci che) il progetto e' stato approvato*

~ Ottobre 2012: *il progetto e' formalmente approvato dal MIUR con un budget di 1925 Keuro*

- 20 dicembre 2012: *pagato il "dazio": il vero budget e' di 1475Keuro*

- Struttura del progetto *pensata su 3 anni* ***[2013 … 2015]***

Il progetto premiale SUMA

Obbiettivo chiaro del progetto : ------>

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

La visione strategica ….

La visione strategica

Innestare un circolo virtuoso di collaborazione con i centri di calcolo nazionali, per incrementare la disponibilita' di calcolo ...
... e rendere piu' facile e significativo l' accesso ai grandi centri di calcolo (in Italia e in Europa)

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

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

Il progetto premiale SUMA

Obbiettivi del progetto: ***Supporto al calcolo*** *--->*

Rendere accessibili risorse di HPC sufficienti esterne all' INFN.

WP1: Migliorare l' utilizzo delle attuali risorse di calcolo HPC e imparare a utilizzare i processori e i sistemi di nuova generazione in contesti tipici della comunita' teorica INFN.

WP2: Tenere aperta una fast lane per sperimentazioni veloci di calcolo massiccio → upgrade del cluster di Pisa + ...

Il progetto premiale SUMA

Obbiettivi del progetto: ***Sviluppi tecnologici*** *--->*

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 e anzi incrementarlo

WP3: Istallare un "large prototype" "innovativo" (proof-of-concept di una futura macchina per il calcolo scientifico ad alte prestazioni, e – allo stesso tempo – ulteriore workhorse di calcolo.)

I lavoratori ….

Workplan di SUMA (Gennaio 2013) firmato da 45 persone

25 persone nei libroni INFN per il 2014 (tipicamente 10 – 20%)

Primo pacchetto di 6 assegni di ricerca (SUMA 100%) decisi all' inizio del 2013 ….

… di cui 5 assegnati (tra mille difficolta') nell' estate 2013 con inizio dei contratti durante l' autunno 2013

I lavoratori (assegni di ricerca)

Pisa	*(WP2)*	*02/09/2013:*	*Giuseppe Caruso*
Roma I	*(WP1)*	*01/11/2013:*	*Pol Vilaseca Mainar*
Roma 3	*(WP1)*	*concorso 09/2013:*	*A. Shreck*
TOV	*(WP1/4)*	*15/10/2013:*	*Francesco Stellato*
Trento	*(WP1)*	*01/10/2013:*	*Gigi Scorzato*
Roma 1	*(WP4):*	*da ribandire.....*	

Abbondanti risorse di calcolo

Accordo quadro + accordo attuativo (2012.09) con il CINECA

Il CINECA mette a disposizione dell' INFN 100 MCore-hours su Blue Gene/Q

L' INFN supporta il CINECA nella fase iniziale di funzionamento della macchina "sperimentale" EURORA (vedi poi) e valuta le sue potenzialita' di calcolo per le aree di calcolo di interesse INFN

Abbondanti risorse di calcolo

Consuntivo del primo anno:

~ 113 Mcore-hours effettivamente utilizzate

~ 250 Mcore-hours ottenute tramite PRACE + ISCRA (in principio scorrelate, pero' ….)

Storage CNAF<--> CINECA still a problem…

Significativo set di benchmark realizzato su EURORA (vedi in seguito)

Accordo rinnovato per altri 2 anni all' ultimo CD.

WP2: upgrade del cluster teorico

Il nuovo cluster di HPC e' stato istallato a Pisa a partire dai primi di settembre.

25 nodi di calcolo

4 processori AMD opteron 6380 (2.5 GHz) (16 x 4 cores)

512 Gbyte di memoria

Rete Infiniband QDR

Switch Mellanox 36 porte

In totale 1600 core di calcolo (~ 10 Gflops/core)

WP2: upgrade del cluster teorico

Accesso esclusivamente tramite un sistema di code via una userinterface locale (code controllate da LSF)

Debug (4 cores / 30 min)
Parallel (256 cores / 6 ore)
Longparallel (512 cores / 24 ore)

Struttura analoga a quella utilizzata nella maggior parte dei Computer Center HPC

WP2: upgrade del cluster teorico

Infrastruttura accesa e funzionante,

Sotto test da parte di un set limitato di utenti "smart"

Apertura ufficiale alla comunita' CSN4 il 22 Ottobre (riunione CSN4 a Roma).

Joint venture di:

SUMA	*140 Keuro*
CSN4	*70 Keuro (+20 Keuro/anno di operation..)*
CCR	*nuovo switch infiniband; Thanks …..*

Upgrade di 200 … 300 core su residui CSN4 2013 + CSN4 2014

WP2: upgrade del cluster teorico

WP1: Utilizzo efficiente delle risorse

Una parte significativa dell' attivita' della prima meta' del 2013 e' stata dedicato al porting e all' ottimizzazione dei programmi si LGT su Blue Gene/Q:

- Una ingente mole di risorse di calcolo e' diventata "improvvisamente" disponibile

- La transizione BG/P → BG/Q e' stata assai piu' complessa di quanto "advertized" da IBM

Significativo contributo al porting e alla ottimizzazione dei programmi di ETMC, CLS

WP1: Utilizzo efficiente delle risorse

Significativa attivita' di addattamento, ottimizzazione, test sulle nuove architetture degli acceleratori (GPU, Intel/MIC), nonostante la sostanziale indisponibilita' degli "addetti ai lavori"

Il problema e' stato affrontato in due direzioni complementari:

- qual 'e' la potenza di calcolo che queste nuove architetture rendono disponibile, se si fa tutto quello che e' necessario fare per ottenerla?

- qual' e' l' evoluzione degli ambienti di programmazione che permettono di ottenere una ragionevole performance in modo decentemente comodo e compatibile su piattaforme diverse?

Guardiamo al prossimo futuro ….

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

Guardiamo al prossimo futuro ….

Questo sembra essere quello con cui dovremo convivere …

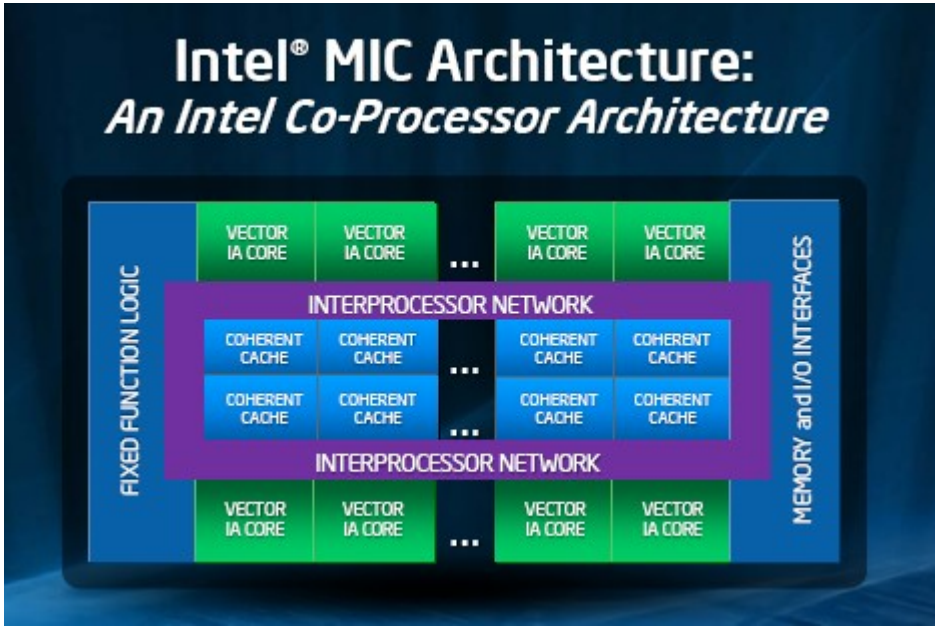
Intel MIC processors ….

Un numero alto (ma non altissimo) di core massicci

e.g: 64 cores x 32 Gflops

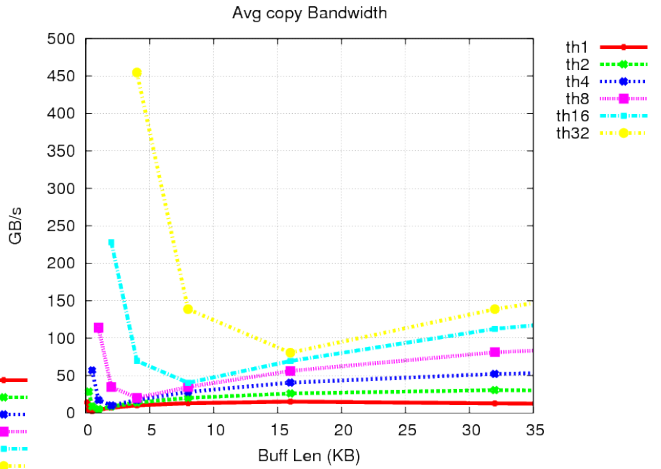
Tecniche di programmazione "relativamente simili" a quelle attuali ……

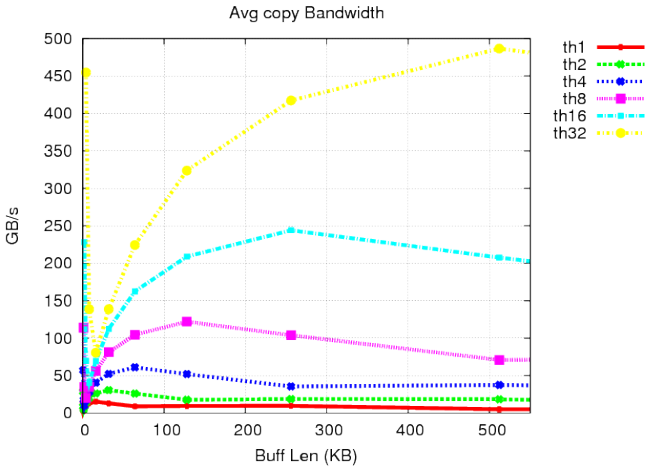
… ma avere buone prestazioni non e' facile →

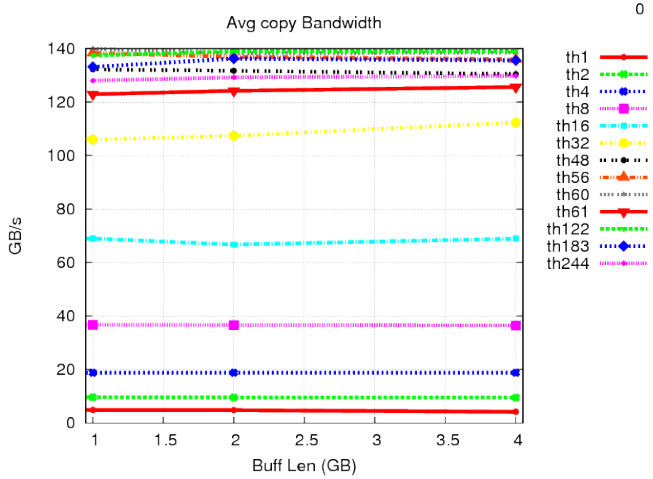


Intel MIC processor: Basic benchmarks

Un analisi accurata della banda processore - "memoria"

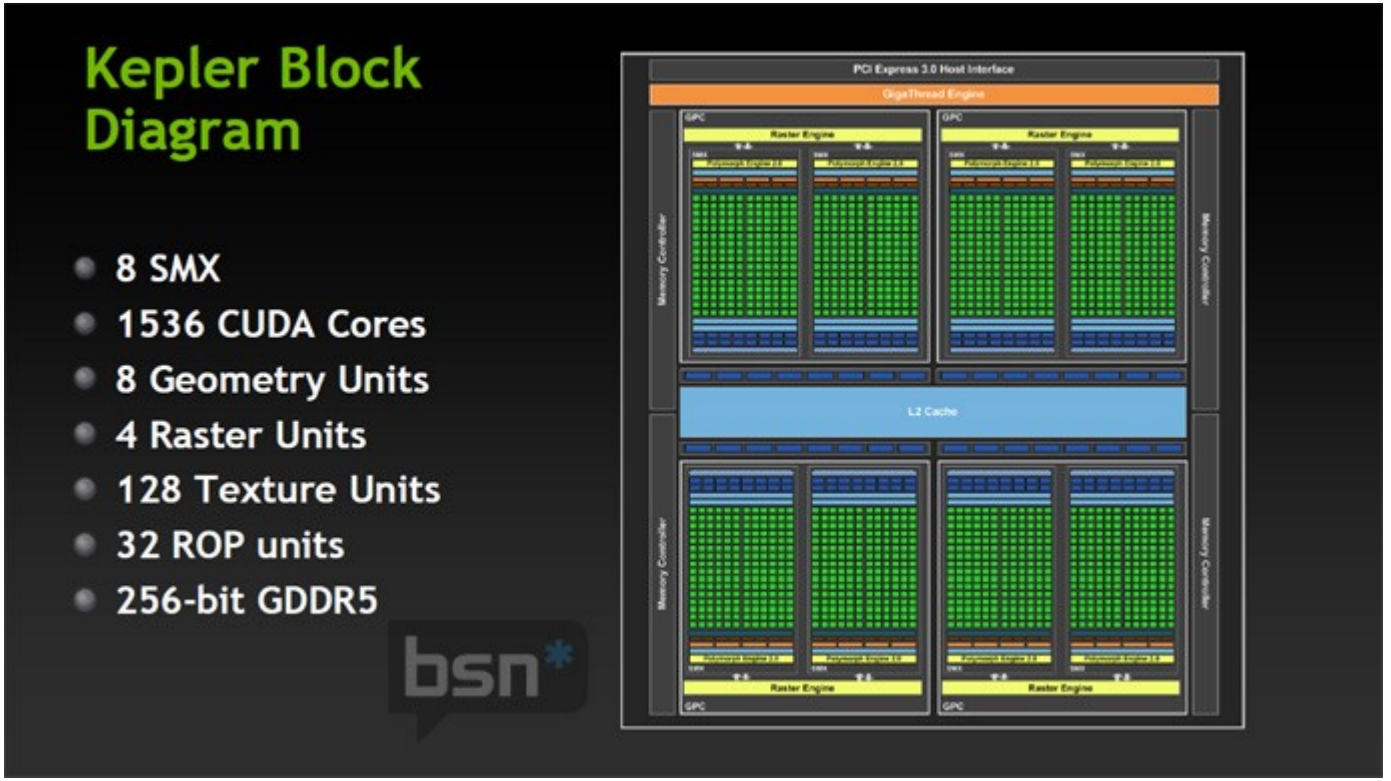






NVIDIA GPU processors ….

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.



WP1: Utilizzo efficiente delle risorse

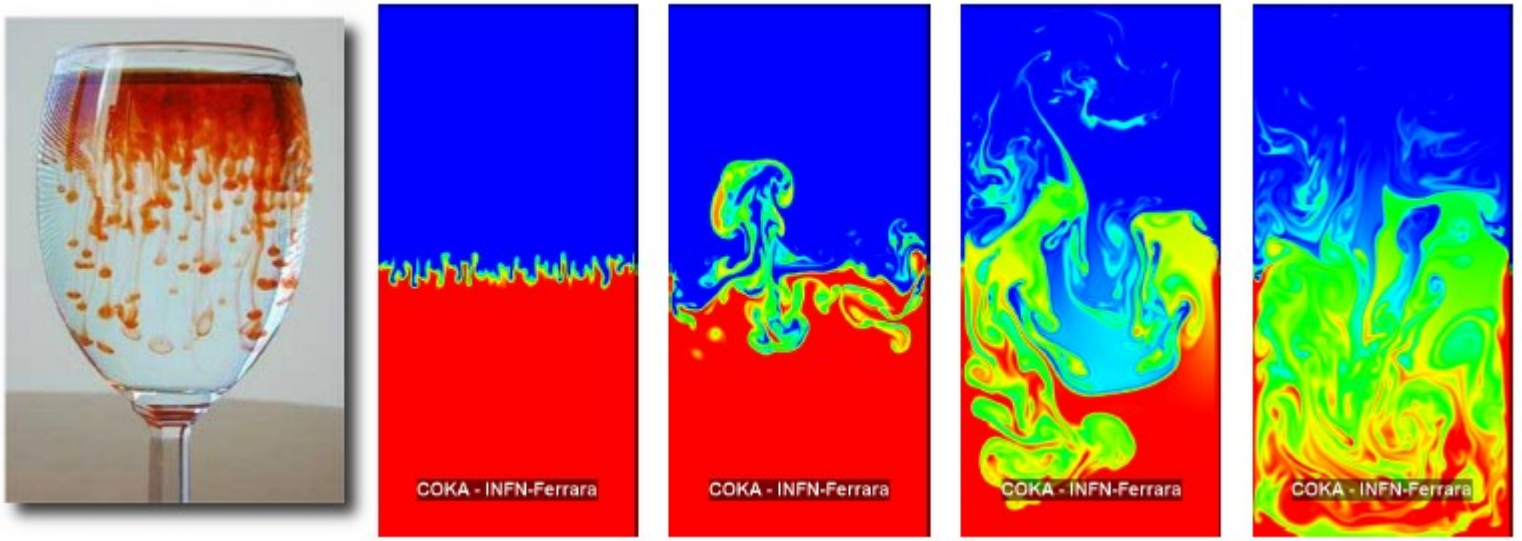
Una significativa attivita' di benchmarking "comparativo" tra processori standard di alto livello, GPU e MIC

	Xeon E5-2687	Tesla K20X	Xeon-Phi 7120P
#physical-cores	8	14 SMX	61
#logical-cores	16	2688	244
clock (GHz)	3.1	0.735	1.238
GFLOPS (DP)	198.4	1.317	1.208
SIMD	AVX 64-bit	N/A	AVX2 512-bit
cache (MB)	20	1.5	30.5
#Mem. Channels	4	–	16
Max Memory (GB)	256	6	16
Mem BW (GB/s)	51.2	250	352
ECC	YES	YES	YES

WP1: Utilizzo efficiente delle risorse

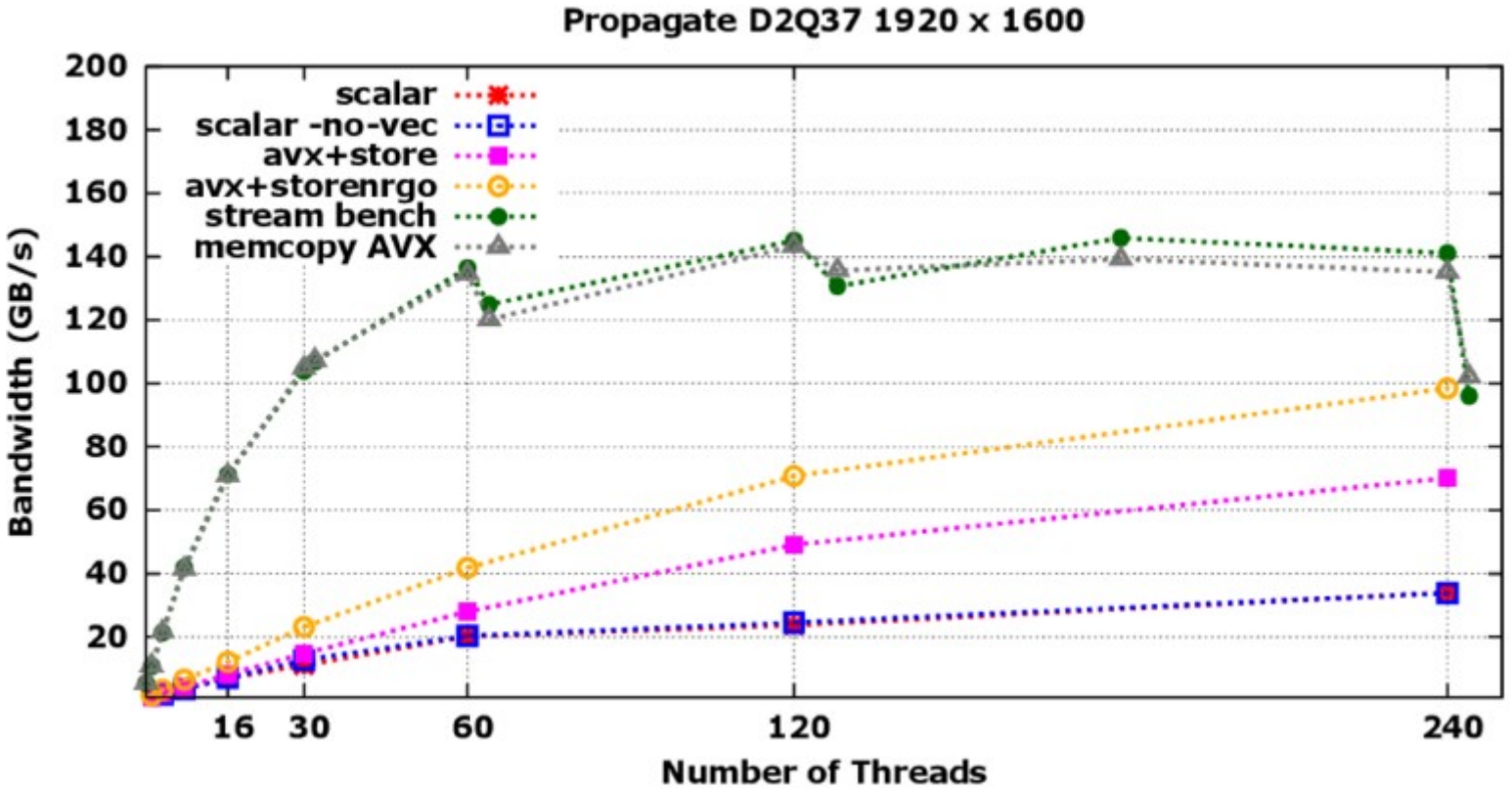
Molte misure svolte utilizzando un programma di simulazione di fluidodinamica computazionale a la' Lattice Boltzmann (D2Q37)

Sufficientemente semplice per permettere test "arditi"
Sufficientemente complesso da essere un buon benchmark sia per il calcolo che per l' efficienza della memoria



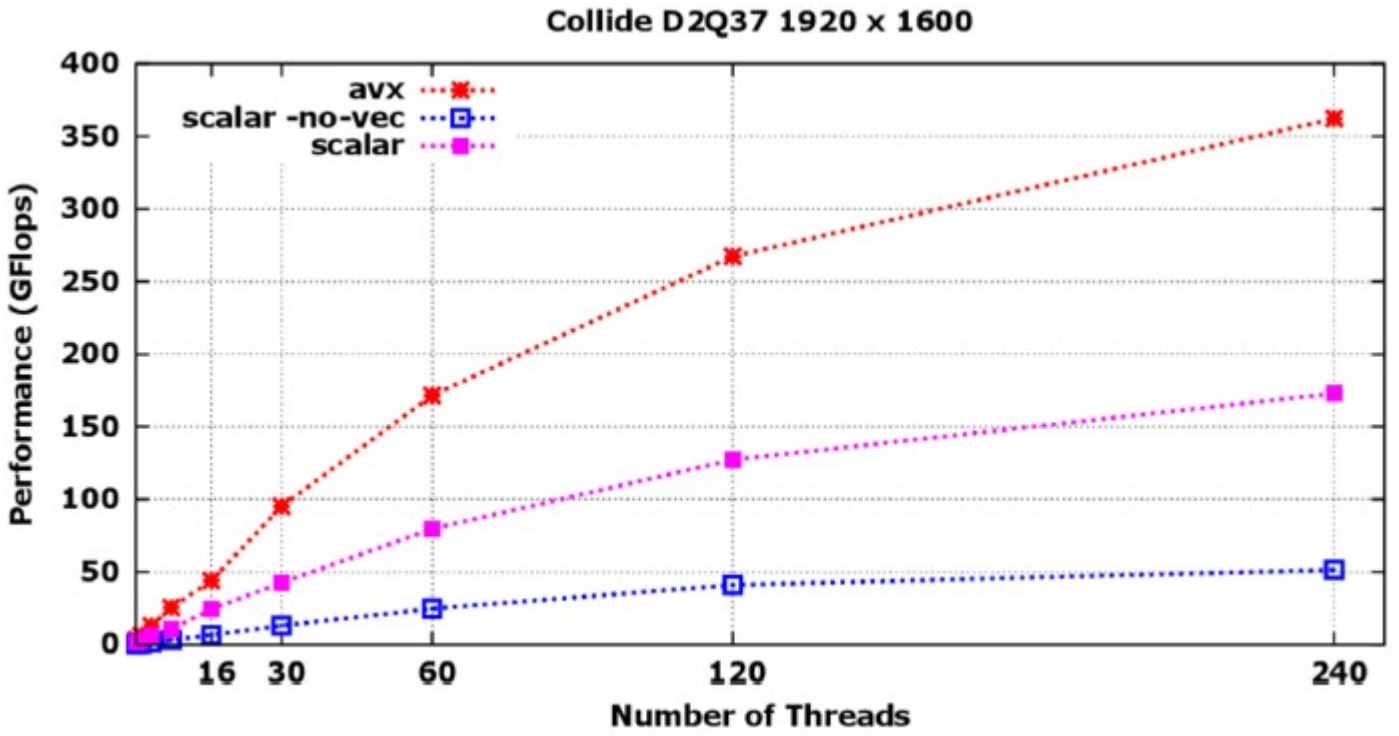
WP1: Utilizzo efficiente delle risorse

Sul MIC, bisogna fare capriole con la gestione dei thread ….



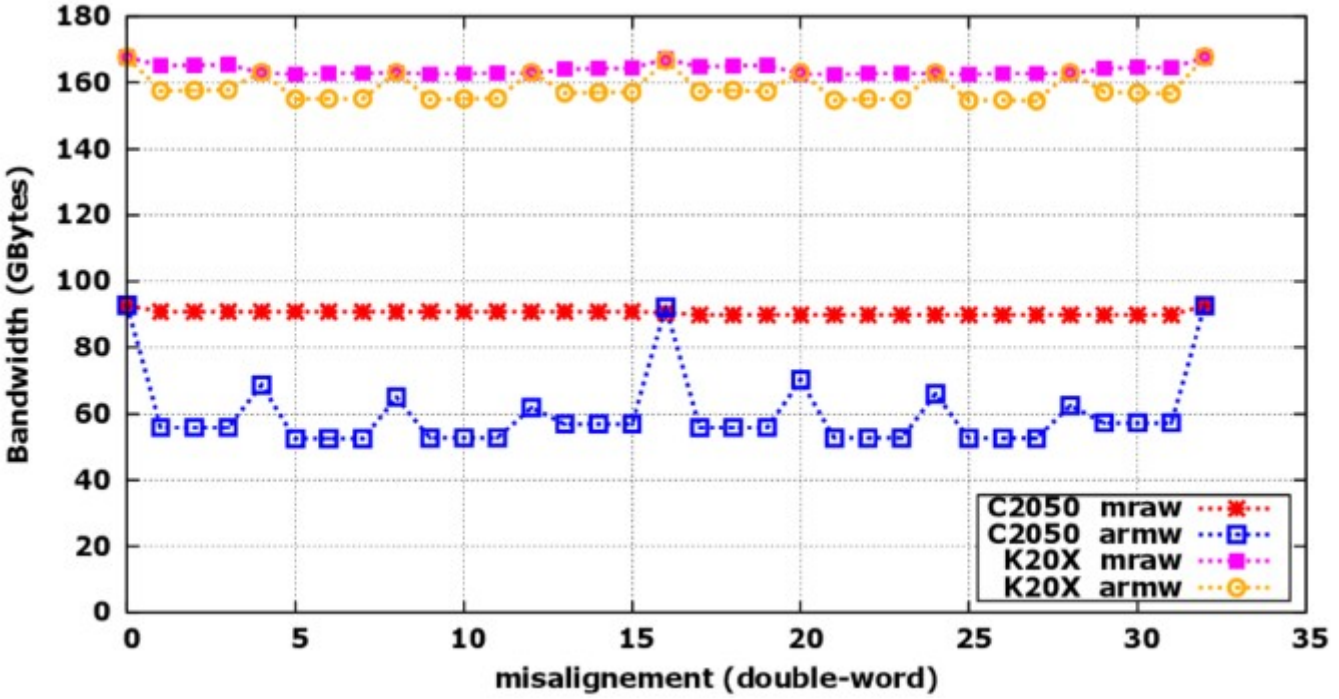
WP1: Utilizzo efficiente delle risorse

…. e con la vettorizzazione del calcolo



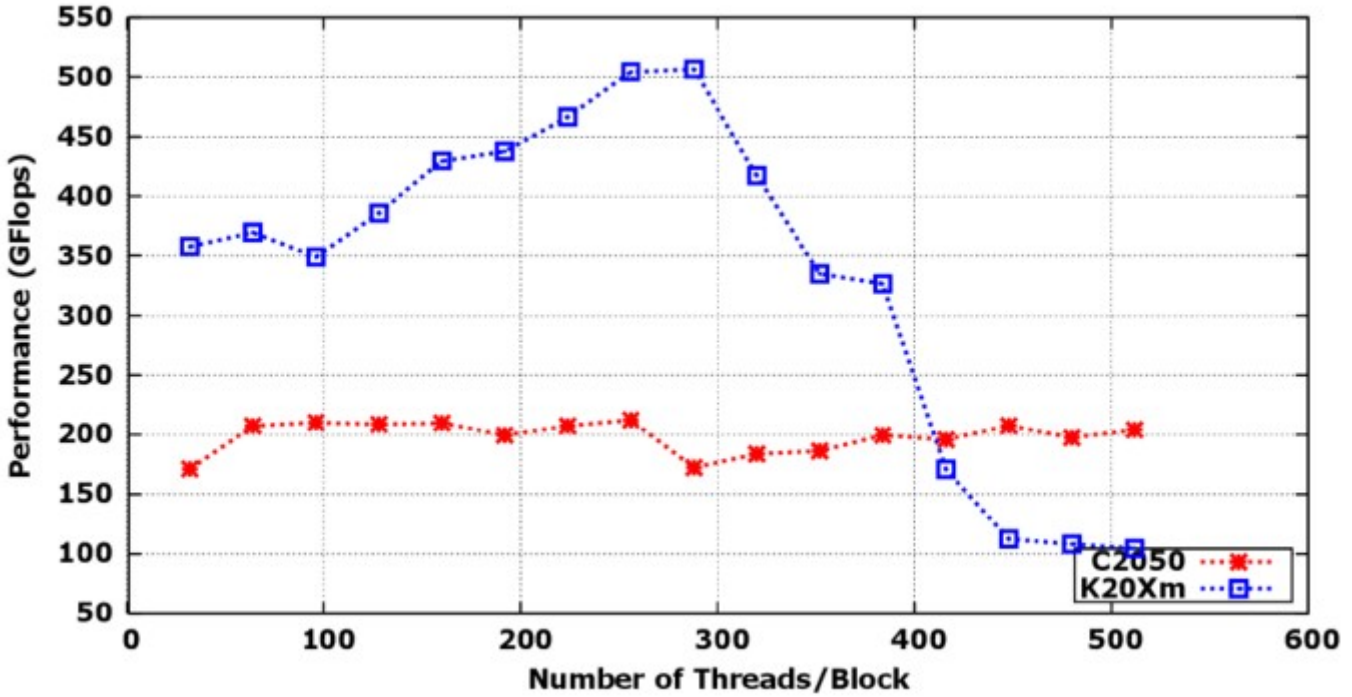
WP1: Utilizzo efficiente delle risorse

Con le GPU bisogna combattere con le idiosincrasie del sistema di memoria ….



WP1: Utilizzo efficiente delle risorse

Con il corretto partizionamento del nucleo di calcolo sui thread



WP1: Utilizzo efficiente delle risorse

La bottom line non e' pero' del tutto soddisfacente....

Un programma accuratamente ottimizzato su un acceleratore ha una performance ~ 3 volte migliore dello stesso programma accuratamente ottimizzato su un SB

	Intel dual E5-2680	Intel Xeon-Phi 7120X	Nvidia K20X
propagate GB/s	60	98	155
ε	70%	28%	62%
collide GF/s	220	362	565
ε	63%	30%	43%
MLUPS	29	54	64
μJ / site	8.96	5.55	3.67

Breaking news from CNN

Ripetuto l' esercizio sulla macchina di cui l' EU e' innamorata....

Arm + GPU

Collide *130 GB/s*

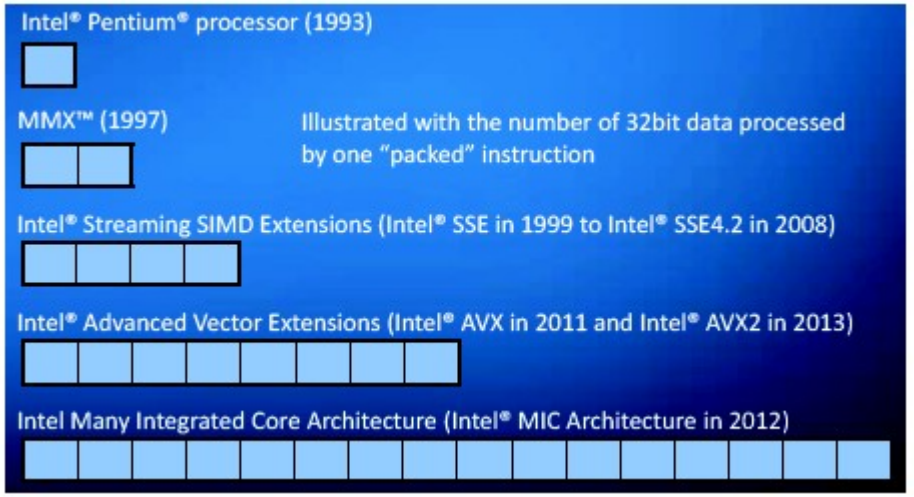
Propagate *287 GFlops*

	Intel dual E5-2680	Intel Xeon-Phi 7120X	Nvidia K20X
propagate GB/s	60	98	155
ε	70%	28%	62%
collide GF/s	220	362	565
ε	63%	30%	43%
MLUPS	29	54	64
μJ / site	8.96	5.55	3.67

WP1: Utilizzo efficiente delle risorse

Che succede se la programmazione e' meno "disperata" e soprattutto portabile su architetture diverse...

Un modo grafico di rappresentare il problema … e la sua evoluzione temporale



WP1: Utilizzo efficiente delle risorse

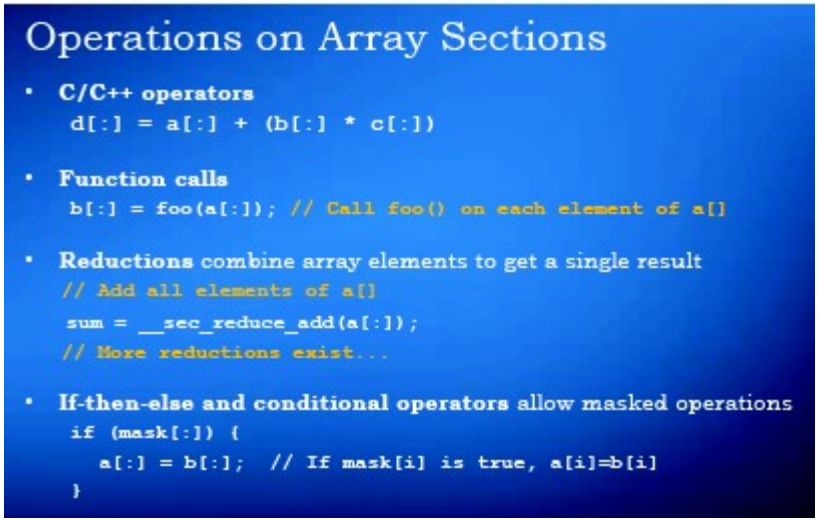
Ormai da qualche anno, c'e' uno sforzo intenso per cercare di esprimere il parallelismo a livello di programma e lasciare al compilatore/RTE il compito di utilizzarlo su una determinata macchina.

E' – in media statistica – dopo un po' I compilatori diventano abbastanza bravi.....

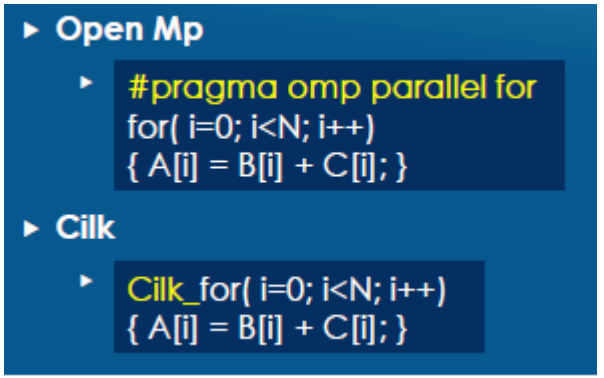
WP1: Utilizzo efficiente delle risorse

Qualche esempio:

Array notation:



OpenMP / Cilk:



WP1: Utilizzo efficiente delle risorse

Uno spezzone di programma in cui tutto il parallelismo e' chiaramente individuato....

```
tick_start = cilk_getticks();
cilk_for (int i2 = 0; i2 < iter; ++i2)
for (int i = 0; i < iter; ++i)
for (int q = 0; q < Q; ++q)
        {
        C3[i][i2]+= __sec_reduce_add(( A[i][q*N:N]*B[i2][q*N:N]));
        }
tick_end = cilk_getticks();
```

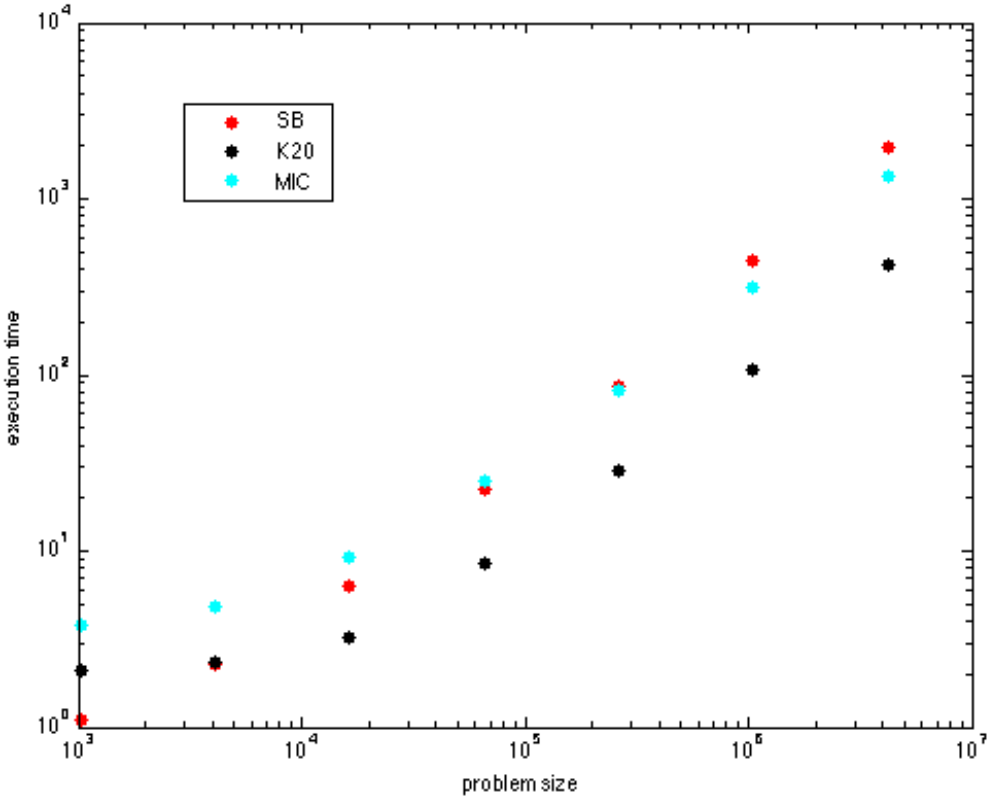
WP1: Utilizzo efficiente delle risorse

The bottom line....

Monte Carlo di un
Modello XY in 2d

Lo stesso programma
Sulle 3 macchine

Commenti, a voce....



Sviluppi tecnologici in ambito INFN

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

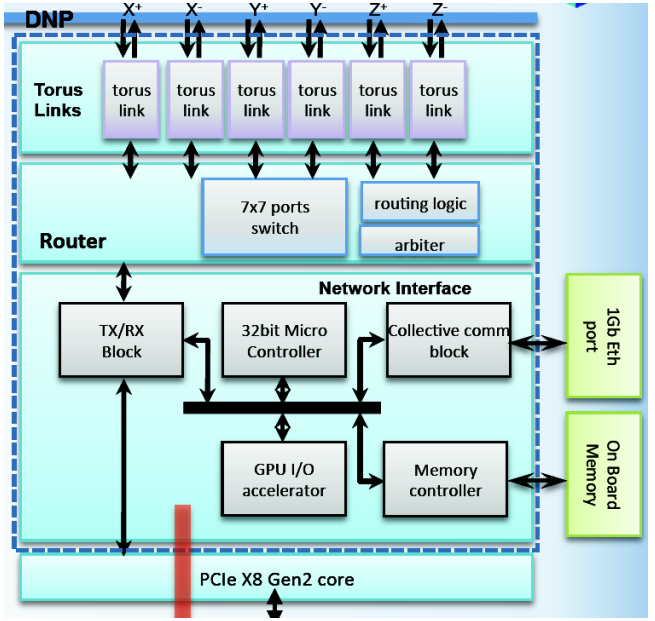
Eredita di APE ------>

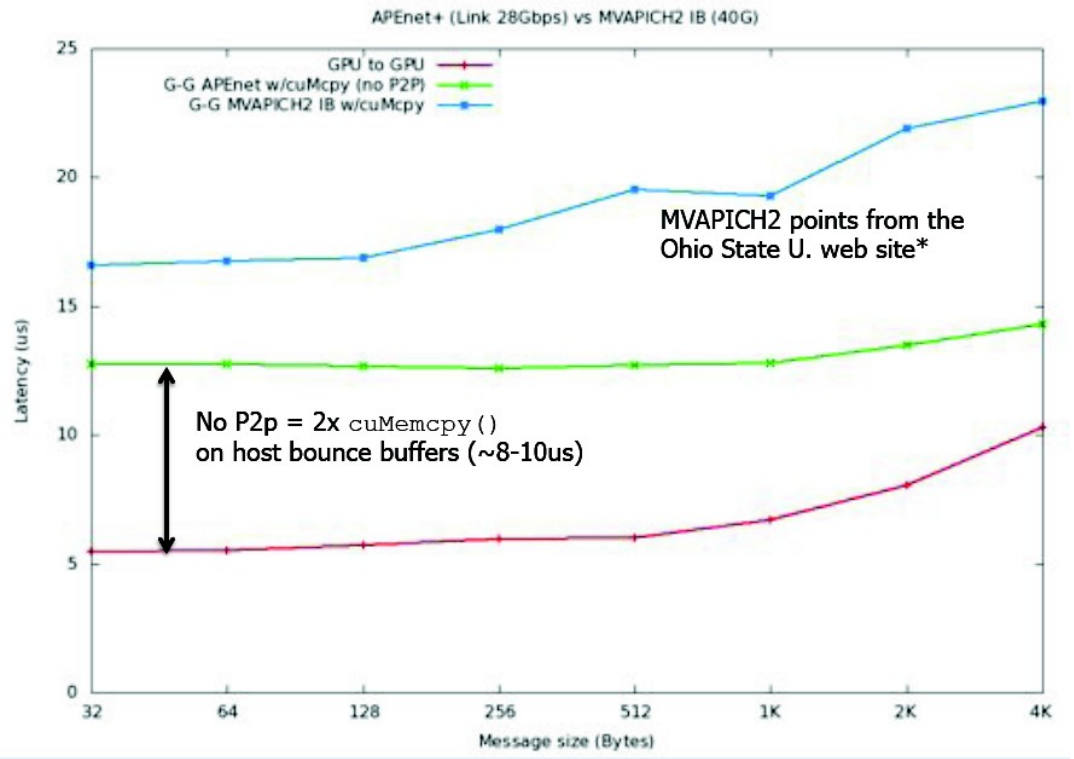
APEnet+ QuonG

AuroraScience

APEnet+ QUonG

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





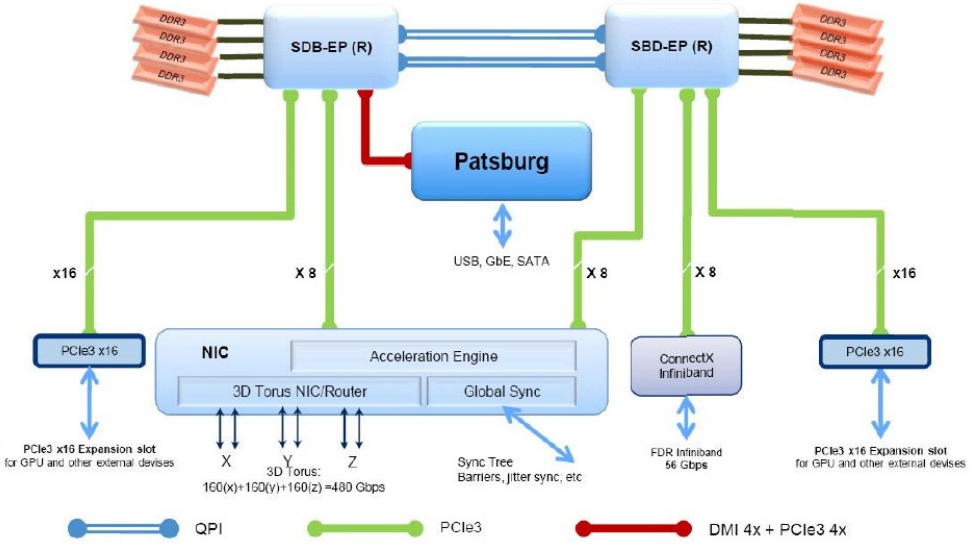
AuroraScience → Eurora

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

*Da **Aurora** e' derivato **Eurora:***

"Farina del sacco Eurotech"

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



Aurora → Eurora

Eurora e'stato istallato al CINECA durante il 2013

128 processori Sandy-Bridge (1024 core)

64 GPU (Aprile 2013)

64 MIC accelerators (domani ...)

*180+ Tflops (**peak**, double)*

Infiniband + Torus Infrastructure ….

Eurora

Eurora e'stato istallato al CINECA durante il 2013

Fase iniziale drammaticamente instabile, sia dal punto di vista hardware che da quello software e sistemistico

Notevoli progressi a partire da giugno-luglio

Primo classificato nella Green500 di giugno 2013

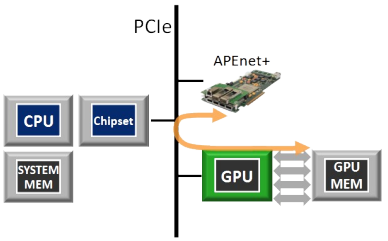
Stabilmente utilizzabile da fine luglio (oggi ~ 85% load)

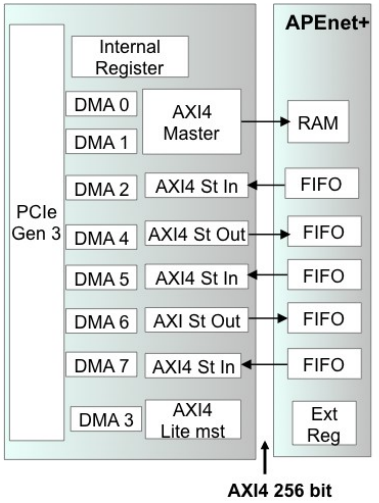
In vari modi la macchina di test ideale, sia dal punto di vista di ottimizzazione dei programmi che da quello di applicazione delle nostre sperimentazioni tecnologiche

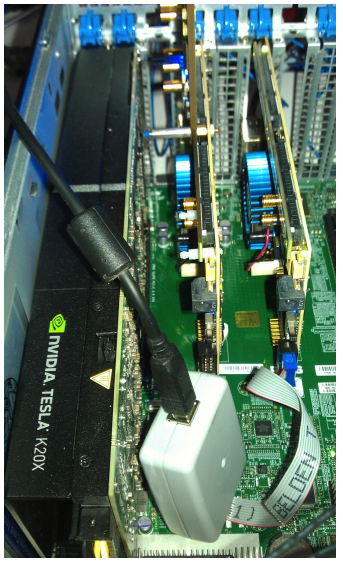
- WP4 -

Valorizzare gli sviluppi tecnologici in ambito INFN e svilupparli ulteriormente. → Ulteriori progressi di APEnet

Punto di partenza: scheda APEnet+ V4, integrata su STRATIX IV (40nm)
PCI Gen2 x8 (5Gb/s per lane)
Supporto per P2P (GPUdirect RDMA)



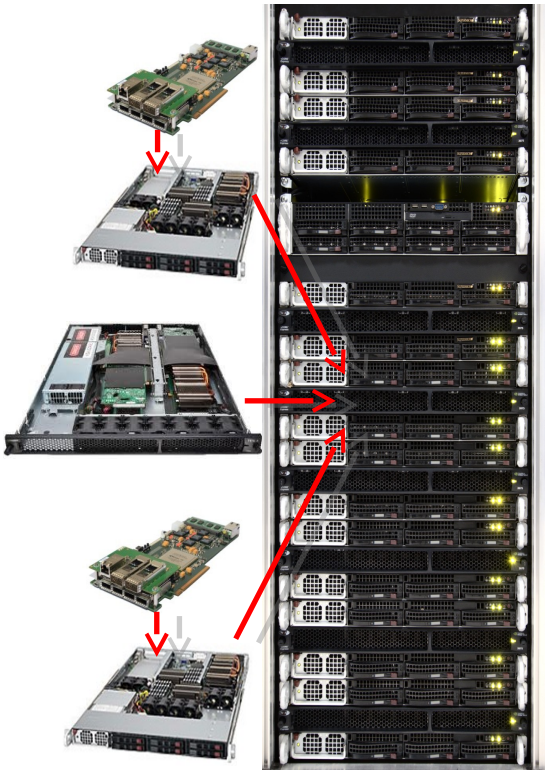




- WP4 -

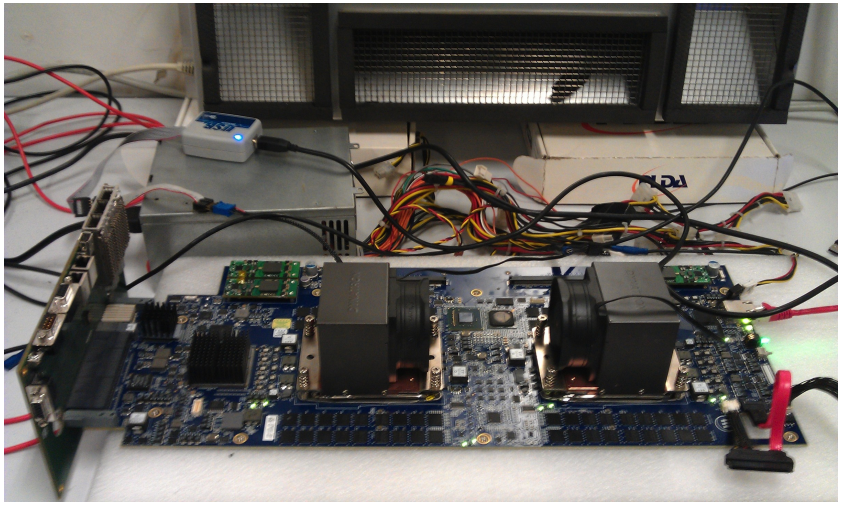
Valorizzare gli sviluppi tecnologici in ambito INFN e svilupparli ulteriormente. → Ulteriori progressi di APEnet

Assemblato un sistema di 16 nodi (4x4x1)
Ancora relativamente instabile (problemi di integrazione OS/software NVIDIA/ software APEnet)



- WP4 -

Iniziato il porting su Eurora:
- Infrastruttura di test disponibile
- Compilato il firmware APEnet v4 sulle FPGA della scheda Eurora
- Verificate le funzionalita' di base
- Iniziati test sui link X,Y,Z
- Y, Z OK ------ X NO!???



- passi a breve (Dicembre 2013)
Completamento del test su banco
Test in corpore vili (Eurora/CINECA)

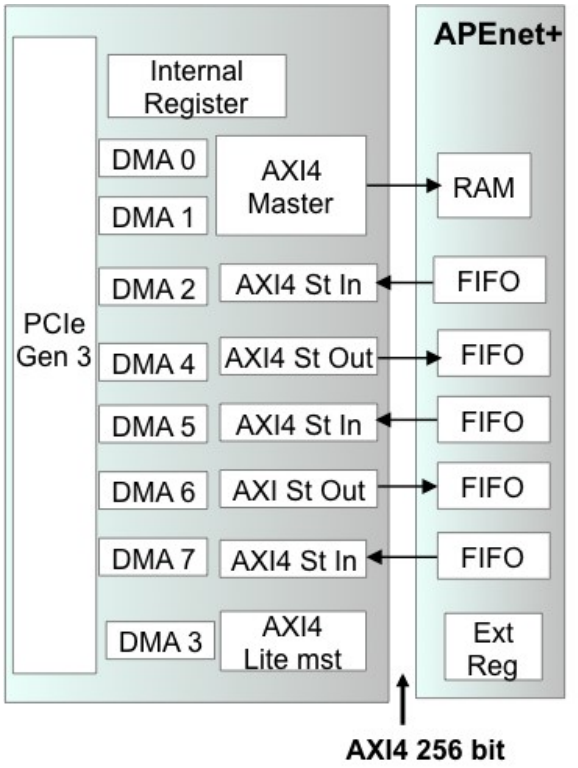
Rapporti con Eurotech → da decidere

- WP4 -

In parallelo:

Sviluppo della V5 di APEnet

Nuova generazione di FPGA
Upgrade Gen4 → Gen5
Link piu' veloci
(in teoria 7 → 11 Gb/sec)
Realisticamente (5 Gb/s → 8 Gb/s)
Encoding piu' efficiente (8/10 → 128/130)
Prototipo su scheda di sviluppo per fine 2013



- 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 -

Due domande importanti:

Qual' e' la architettura di macchina migliore che la tecnologia oggi disponibile permette di assemblare?

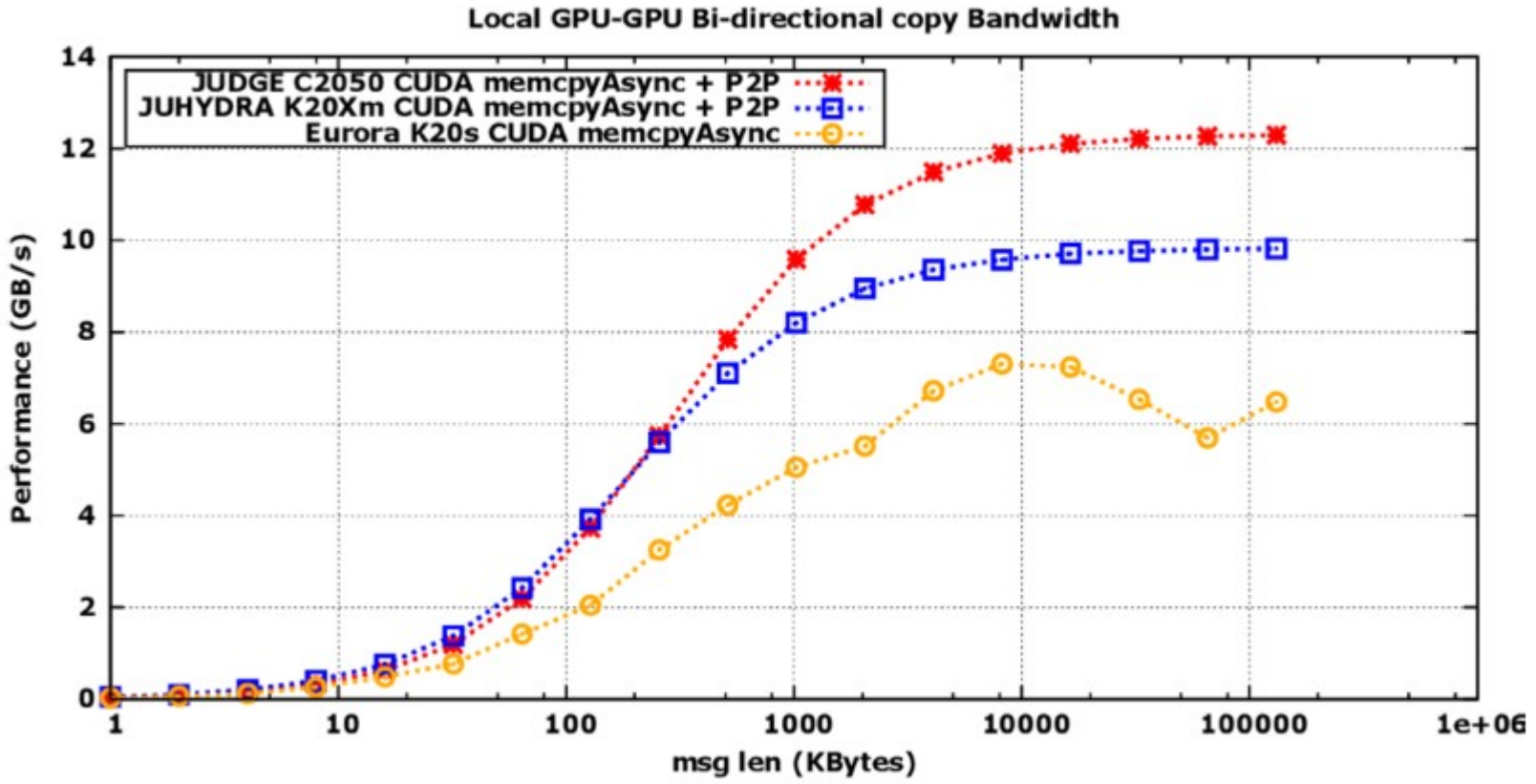
Siamo in grado di utilizzare tale architettura in maniera efficiente?

Le risposte a queste domande non meno banali di come vengono di solito raccontate.....

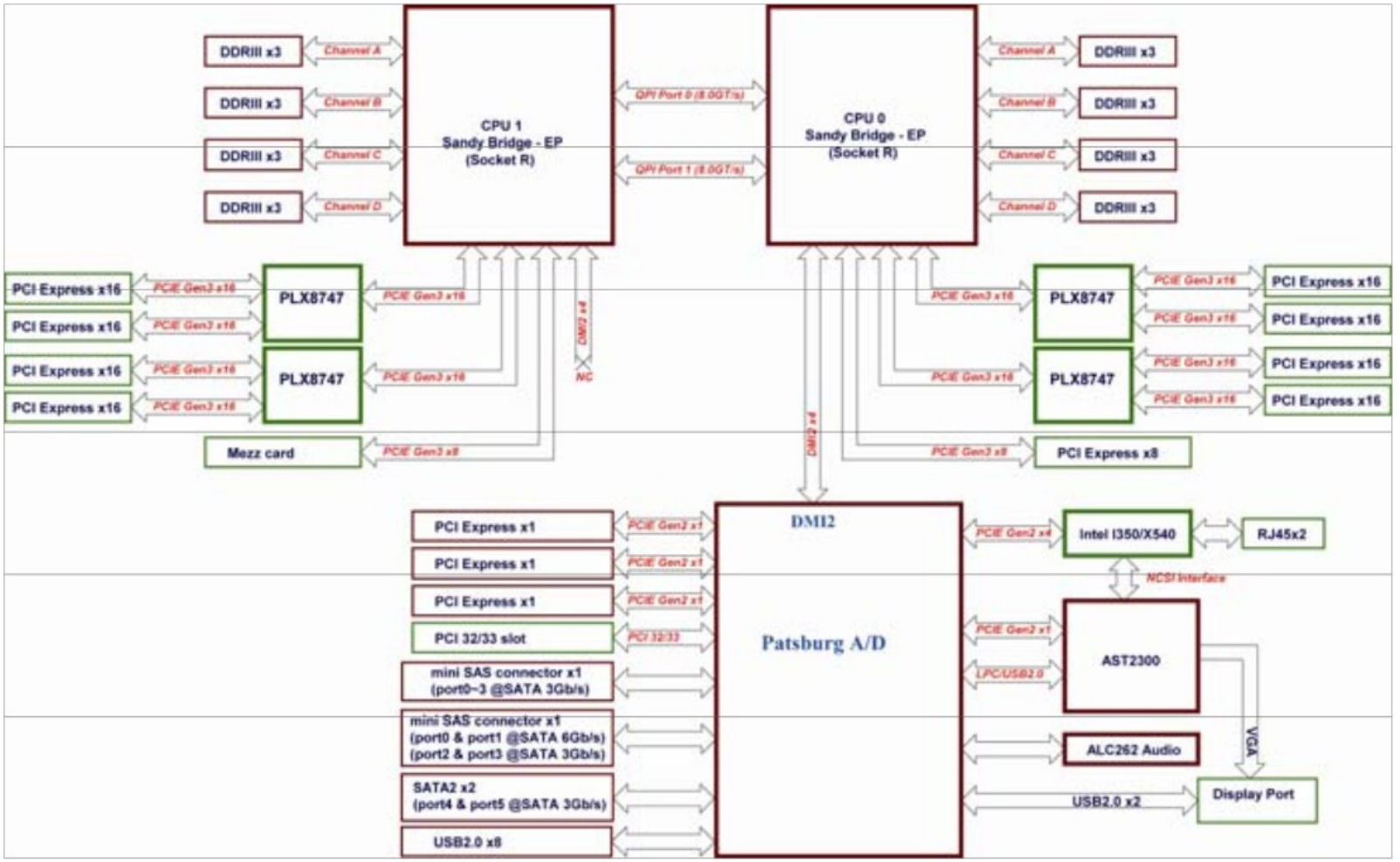
- WP3 -

Le risposte a queste domande non meno banali di come vengono di solito raccontate.....

Ad esempio:



Un "tipico" nodo di calcolo.....



S7059 Block Diagram

- WP3 -

Le risposte a queste domande non meno banali di come vengono di solito raccontate.....

Il modo in cui le GPU sono connesse nell' ambito di uno stesso nodo di calcolo ha un impatto significativo rispetto sulla comunicazione tra GPU

La memoria disponibile sulle GPU e' limitata, e quindi e' necessario un parallelismo a grana molto fine ...

... che richiede piu' comunicazione | | |

- WP3 - Back of envelope estimates

La quantita' di calcolo scala come il volume, la comunicazione con l' area ….

Dunque se aumento il volume (a potenza di calcolo costante) i problemi di comunicazione diventano meno gravi ….

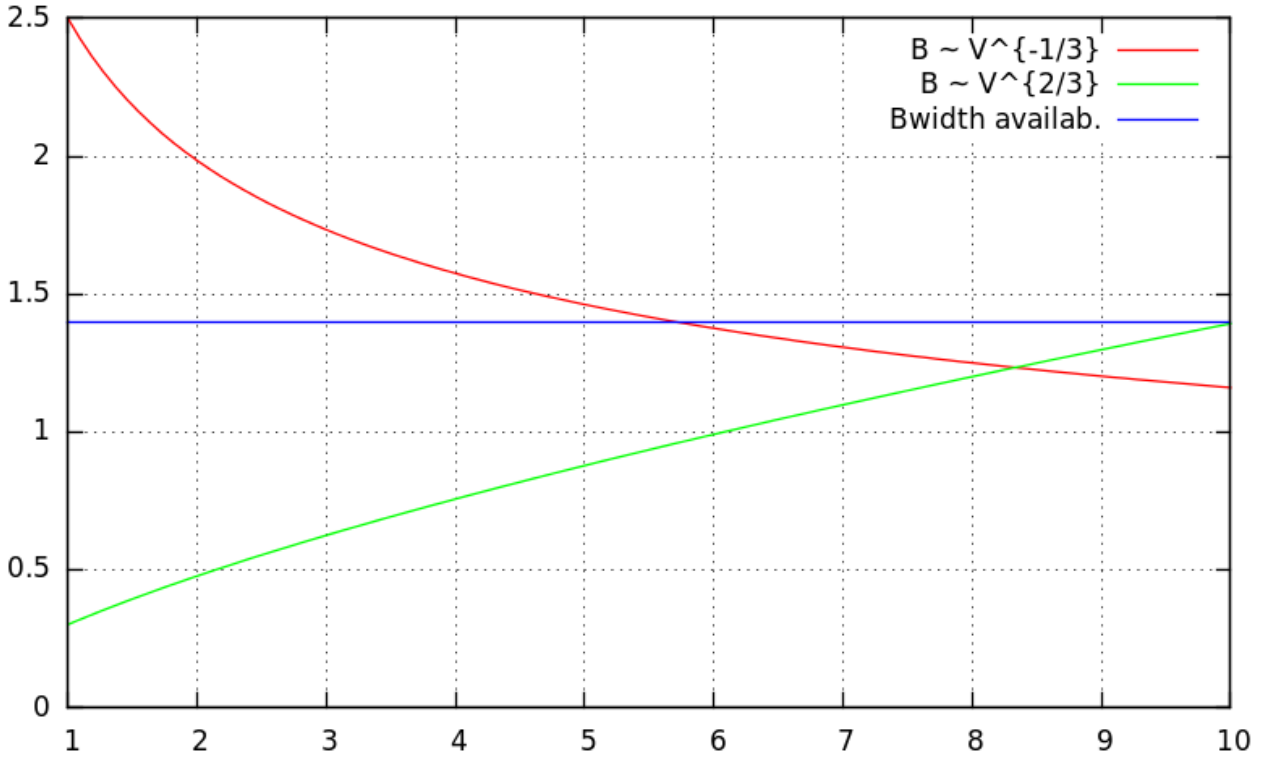
$$T_P = \alpha V \qquad B \propto V^{2/3} \qquad B \propto P/V^{1/3}$$

… d' altra parte la memoria e' limitata: se aumento il volume per nodo devo aumentare le GPU dedicate a quel volume

$$B \propto P(V)/V^{1/3} \qquad P(V) \propto V \qquad B \propto V^{2/3}$$

- WP3 - Back of envelope estimates

Esiste un range di volumi su cui il mio sistema non ha colli di bottiglia??



- WP3 - Back of envelope estimates

Esiste un range di volumi su cui il mio sistema e' efficiente??

Frankly speaking: questa analisi, in modo sufficientemente accurato non e' stata ancora fatta (ne da noi, ne da altri...)

La situazione al contorno

CINECA quasi pronto per il procurement di una macchina "Tier1" con:

- "sufficienti aspetti di innovazione"
- "potenza di picco accelerata ~ 1 Pflops" --->

Questo ci pone basicamente di fronte a due alternative:

.

WP3 – opzioni possibili

Opzione 1 (democristiano – gesuitica): Contribuire alla realizzazione della macchina CINECA

Vantaggi:
- ~ 150 Tflops (picco) riservati all' INFN
- + altri ~150 Tflops come share "normale" INFN
- Tflops 'democristiani' ….
- Problemi di istallazione/startup/gestione a carico del CINECA
- Macchina disponibile rapidamente (Aprile 2014???)

Svantaggi:
Non e' necessariamente il best value for money (a fissata tecnologia)

.

WP3 – opzioni possibili

Opzione 2 (Senatore Rubbia...): Provare a realizzare un prototipo "estremo" …. Ad esempio..

8 GPU per nodo → 16 Tflops/nodo @ 20000 Euro
In principio una macchina da 500 Gflops di picco

Vantaggi:
~ Potenza di calcolo significativamente superiore

Svantaggi:
*(**"Wenn"**) siamo in grado di utilizzarla efficacemente?*
Potenza di calcolo disponibile su tempi piu' lunghi (quanto?)

Money money money ….

9-10 assegni di ricerca biennali	→	*600K*	→	*570K*
Sviluppi avanzati	→	*325K*	→	*200K*
Large prototype	→	*600K*	→	*460K*
Cluster Upgrade	→	*220K*	→	*140K*
(+ 80K dalla CSN4)				
Fisiologia (su 3 anni)	→	*180K*	→	*105K*
Totale	→	*1925K*	→	*1475K*

Yearly spending profile

***** 2013 *****

6 assegni di ricerca	→	*366K*	→	*570K*
Sviluppi avanzati	→	*35K*	→	*200K*
Large prototype			→	*460K*
Cluster Upgrade	→	*140K*	→	*140K*

(+ 70K dalla CSN4)

***** 2014 *****

~ 4 assegni di ricerca	→	*200K*
Large prototype	→	*500K*
Sviluppi avanzati	→	*100K*

Fisiologia (su 3 anni) → *105K*

At a glance ….


2012
2013
2014
2015
WP1
tests on small machines
Port on Quong
Port on Eurora
Port on Bio machine
Test/physics on LP
WP1 Ass. Ric.
Parma-LGT
Parma-Grav.
TOV-qBio
MibRm1-LGT
RM3-LGT
Complex sys
LGT
Phys???
WP2
De-grid ….
Install Cluster
Sissa federat.
Support
WP2 Ass Ric.
Pisa-System Sup.
WP3
Eval. Option
Procure/install
WP4
Assemble Quong
Port Apenet
Exper. With bio machine
future apenet
WP4 Ass.Ric.
RM1-hardware
RM1-software

- 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)