Suma:
primo compleanno

R. (lele) Tripiccione

tripiccione@fe.infn.it

Roma, 8 novembre 2013

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)

Il menu del giorno

Cosa e' successo negli ultimi 12 mesi:
(un certo numero di buone notizie)

Post docs, at last !

Super-calcolo al CINECA

Il nuovo cluster di Pisa

Una lenta migrazione verso le "nuove" macchine:

Test di efficienza sugli "acceleratori"

----->

Opzioni per il "large prototype" del progetto SUMA

Iniziamo con un po' di buone notizie

Gli 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.....*	

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])

Preventivo

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

Consuntivo

Utilizzo di BG/Q da ~ 1 Settembre 2012 a 15 Ottobre 2013

account	start	end	total	Consumed	%
INFN_OG51	20120703	20140630	22.000.000	6.008.941	27.3
INFN_RM123	20120703	20140630	77.000.000	40.533.604	52.6
INFN_TO61	20120703	20140630	6.500.000	2.287.703	35.2
INFN_PI12	20120703	20140630	35.000.000	27.803.322	79.4
INFN_TV62	20120703	20140630	27.000.000	9.658.315	35.8
INFN_MI11	20120703	20140630	32.975.000	27.710.427	84.0
INFN_MB31	20120910	20140730	2.000.000	1.019.362	51.0
INFN_PD32	20120703	20130630	3.500.000	254.297	7.3
INFN_PI11	20120817	20140630	17.000.000	6.558.015	38.6
Total			222.975.000	121.833.991	54.6

Pur con non pochi problemi di gestione, l' iniziativa sta dando i risultati promessi.

INFN – Cineca: Consuntivo

Consuntivo del primo anno:

~ 120 Mcore-hours effettivamente utilizzate

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

Storage CNAF<--> CINECA still a problem…

Accordo rinnovato per altri 2 anni all' ultimo CD.

ZEFIRO: il nuovo 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 *---> ~ 100 porte*

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

ZEFIRO: il nuovo cluster teorico

ZEFIRO: il nuovo 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 (OGGI...)

Joint venture di:

SUMA	*140 Keuro;*
CSN4	*70 Keuro (+20 Keuro/anno di operation..);*
CCR	*contributo al nuovo switch*

Upgrade di ~ 60 KE (~ 500 core) da CSN4 su fondi 2014

ZEFIRO: il nuovo 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)

Approssimativamente 1.2 Mcore-hour all' anno (?!....)

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

ZEFIRO: Modalita' di utilizzo

+ *Accessibile a tutti i componenti delle sigle della CSN4*
  + *altri se richiesto da un componente della CSN4*
    (studenti/ospiti/collaboratori)

+ *In principio in modo automatico, tramite AAI*

+ *Accesso "illimitato" sperimentale fino a fine anno 2013*

+ *Accesso regolamentato a partire da gennaio 2014 ????*
  (vecchia commissione?)

ZEFIRO: Modalita' di utilizzo

+ *Istruzioni per l' uso --->*

http://wiki.infn.it/strutture/pi/computing_center/theocluster

+ *In caso di problemi ---->*

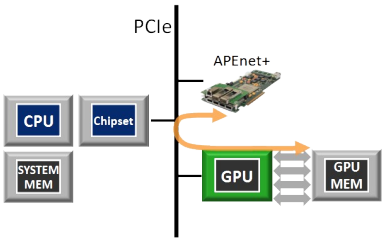
localq-support@lists.pi.infn.it

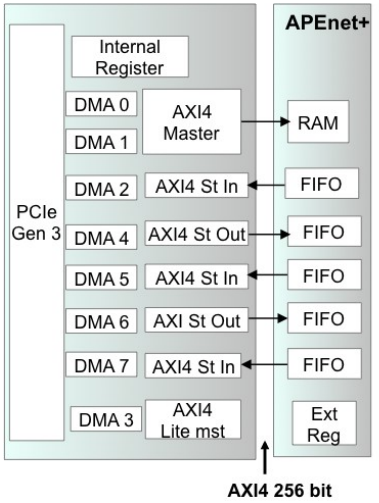
+ *Fineprints a voce ……*

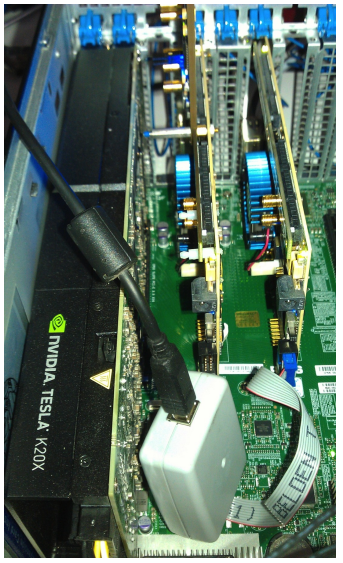
- WP4 -

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)



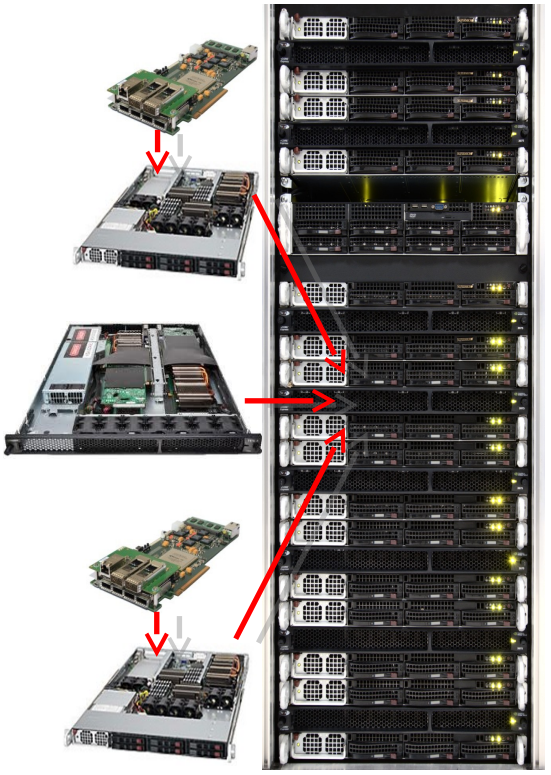




- WP4 -

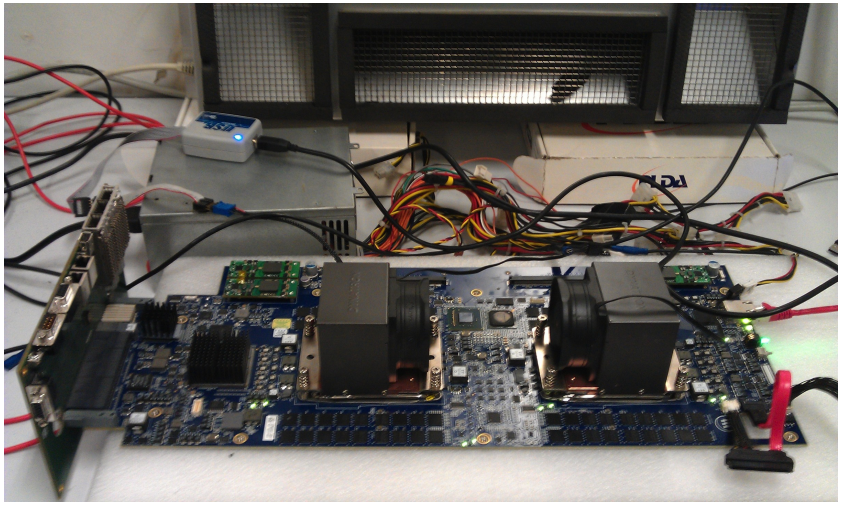
Valorizzare gli sviluppi tecnologici in ambito INFN e svilupparli ulteriormente. → Ulteriori progressi di APEnet

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)

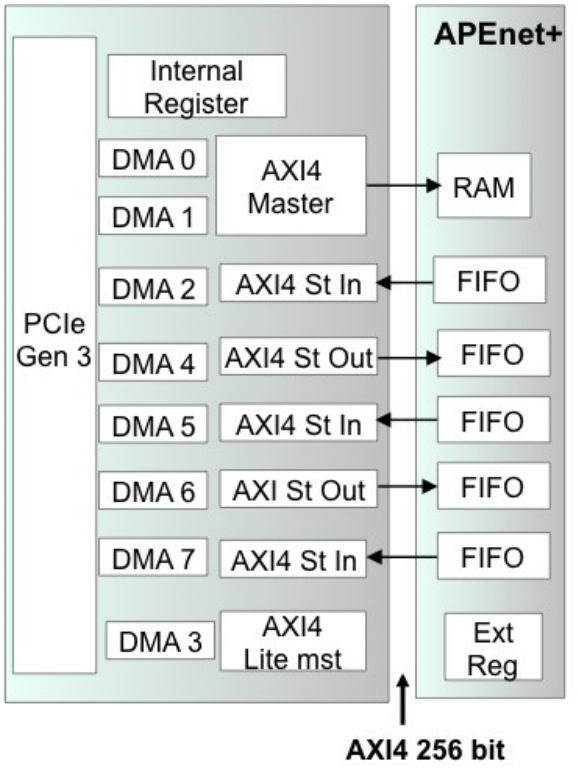
Rapporti con Eurotech → da decidere

- WP4 -

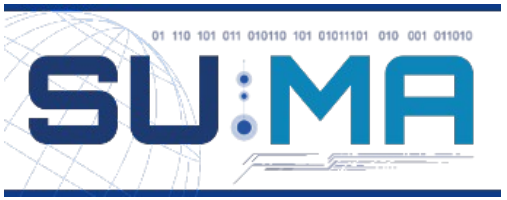
In parallelo:

Sviluppo della V5 di APEnet

Nuova generazione di FPGA
Upgrade Gen2 → Gen3
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



Tra gli obbiettivi del progetto SUMA



Obbiettivi del progetto:

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....

Istallare un "large prototype" basato su questo tipo di processori

+ come workhorse di calcolo per l' INFN

+ come proof-of-concept di una futura macchina per il calcolo scientifico ad alte prestazioni

Il progetto premiale SUMA

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

E.g. 100 – 200 Tflops peak
Istallato al CINECA →

Ancora insufficiente per essere autosufficienti in LQCD …

… Ma del tutto sufficiente per uno studio sistematico di algorithm development in LQCD
+ Fluid-dynamics / Complex-Systems / Quantitative-biology….

Dal talk di un anno fa….

Guardiamo al prossimo futuro ….

Questo sembra essere quello con cui dovremo convivere …

Guardiamo al prossimo futuro ….

1 core di calcolo: 10 → 15 → 50 Gflops
1 nodo di calcolo: 40 → 200 → 400 Gflops

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

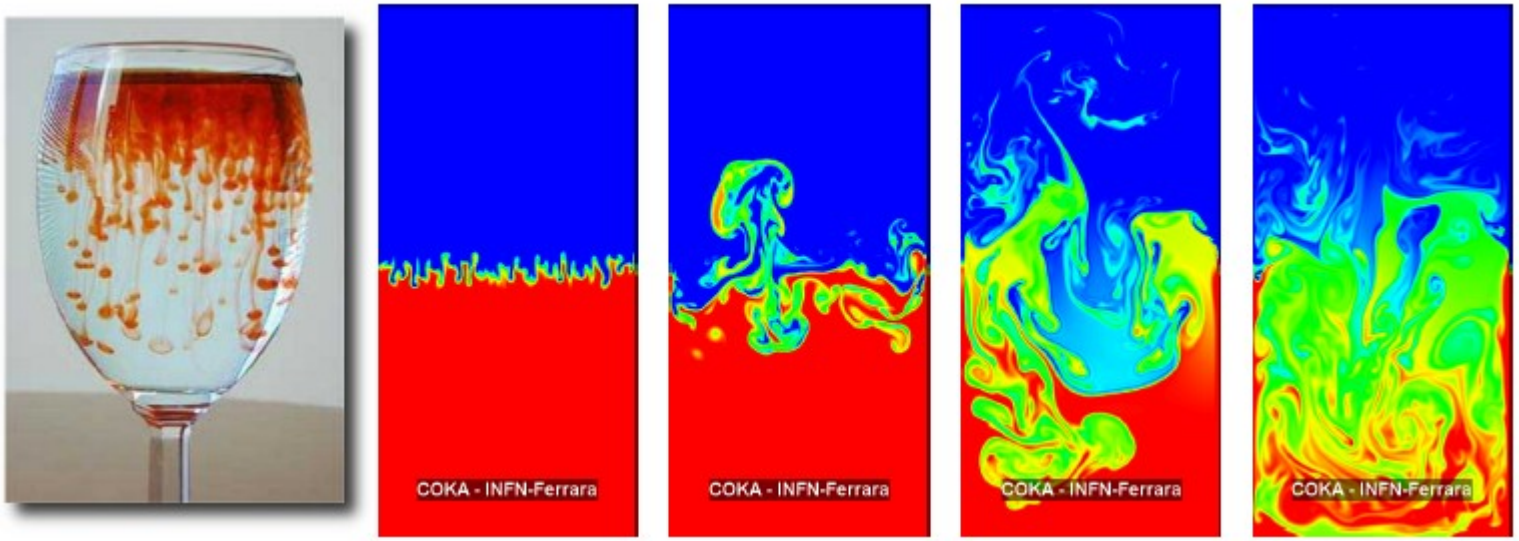
GPU (Nvidia) – MIC (Intel) - ?

Grazie ad un sostanziale aumento del parallelismo del processore

Impariamo a usare queste bestie …

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

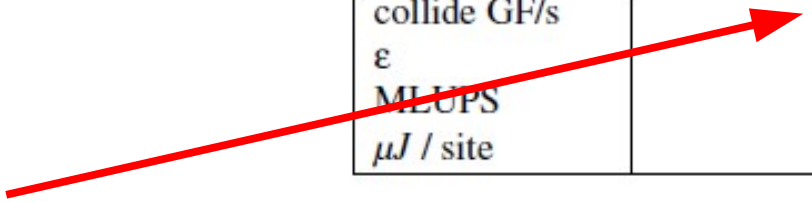


Impariamo a usare le GPU ...

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



- 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 ….

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

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

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

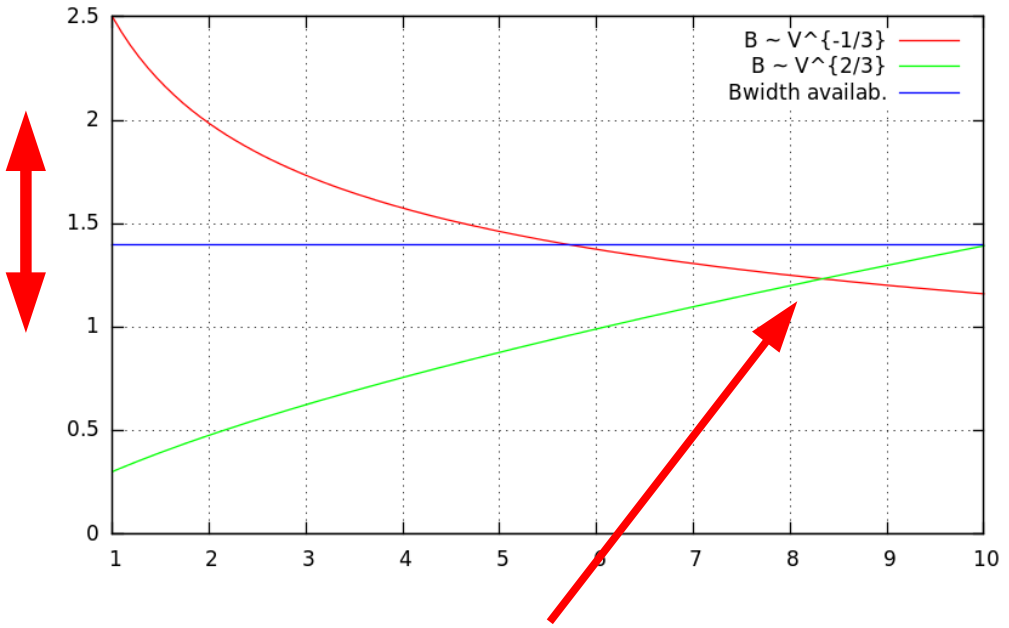
- WP3 - Back of envelope estimates

Esiste un range di volumi senza significativi colli di bottiglia????

Il livello della linea blu dipende
dai dettagli del meccanismo
di comunicazione

Il Cross-over verde/rosso
Dipende dalla memoria
Disponibile sul nodo

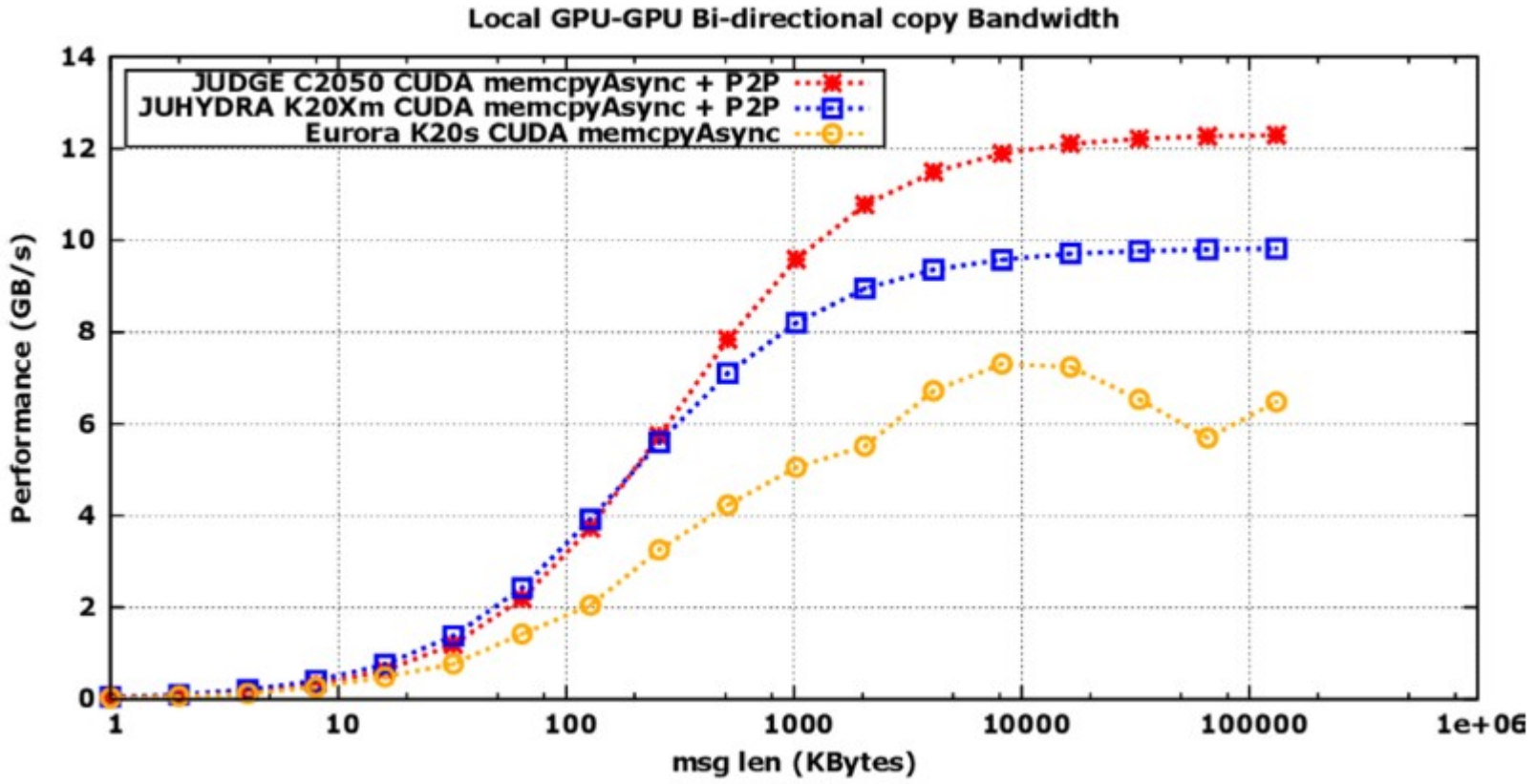
La normalizzazione delle
Curve verde/rosso dall' algoritmo



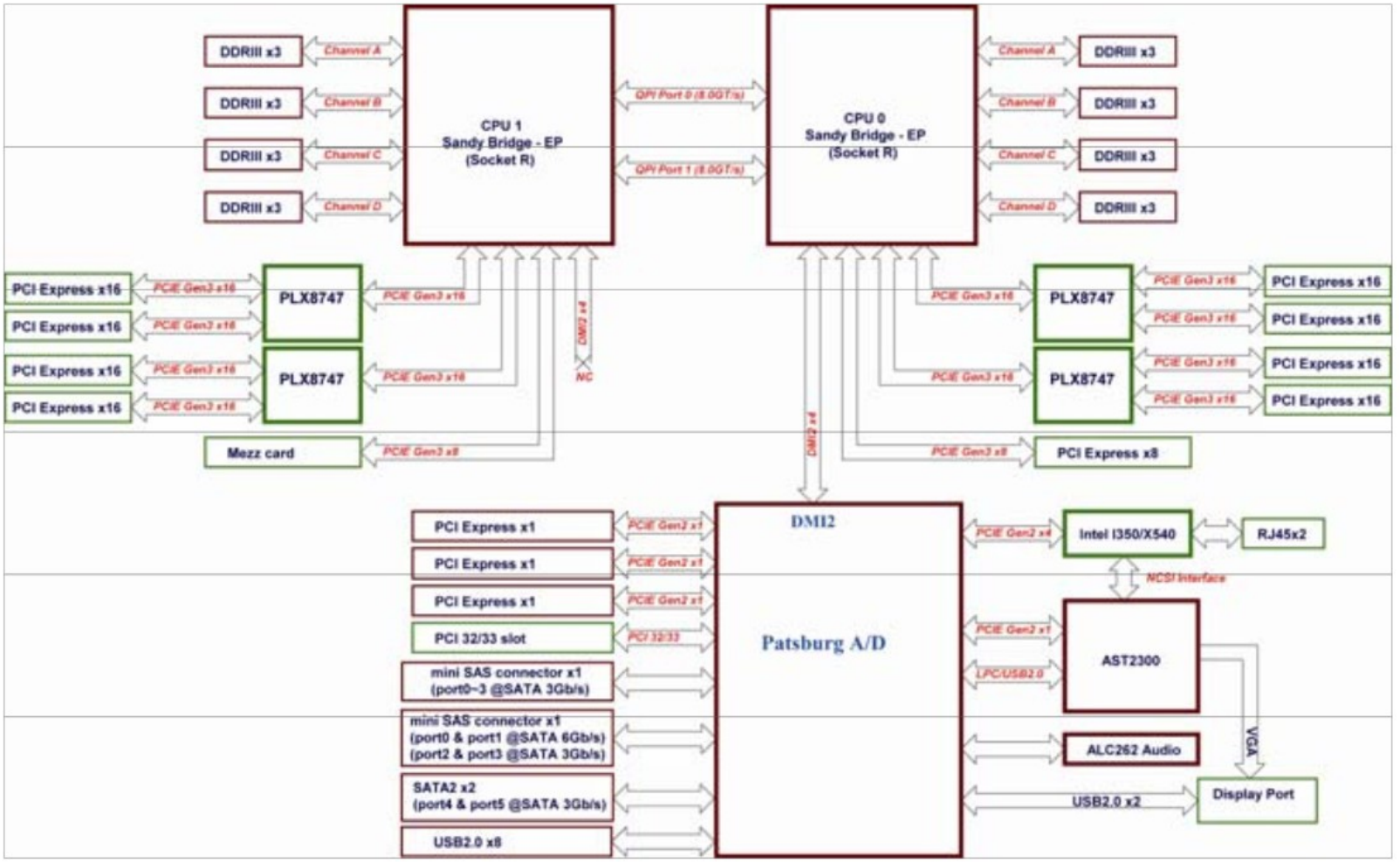
- WP3 -

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

Ad esempio:



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



S7059 Block Diagram

A piu' tardi .....

Ipotesi CINECA

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

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

Che vuol dire!?

.

Ipotesi CINECA

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

- relativamente bassa "densita'" di GPU (1 GPU per CPU)
- nodo con 2X:
processore Haswell (~ 400 Gflops) +
Nvidia K20 (1000 Gflops) → 2.8 Tflops nodo
- ~ 360 nodi @ 5500 Euro/nodo

Sulla base di queste stime ….

Ipotesi CINECA

Contributo aggiuntivo INFN (versione 1)

- indicativamente 80 nodi addizionali → 224 Tflops picco
64 (CPU) + 160 (GPU)

Confronto con BG/Q
1 core Haswell → ~ 4 core BG/Q

80 x 2 x 8 x 4 x 24 x 360 ~ 45 Mcore hour (CPU) (x 1.4)
+ 120 Mcore(eq) hour (GPU) (x 1.4)

Ipotesi CINECA

Contributo aggiuntivo INFN (versione 2)

- indicativamente 120 nodi addizionali
SENZA GPU → 96 Tflops picco (CPU)

Confronto con BG/Q
1 core Haswell → ~ 4 core BG/Q

120 x 2 x 8 x 4 x 24 x 360 ~ 65 Mcore hour (CPU) (x 1.2)

+ 50 Mcore/hour GPU [Cineca]))

Domande "rilevanti"

1) Confronto tra i tempi di istallazione
2) Confronto tra le potenze di picco
3) Cosa sappiamo fare ORA con "poche" GPU
4) Cosa sappiamo fare ORA con "tante" GPU
5) Come cambiano le risposte alle domande di 3) - 4) tra 6 mesi
6) Chi vuole fare cosa, perche' 5) sia diverso da 3) – 4)

---->

7) Vogliamo un uovo oggi o una gallina un anno dopo

Situazione budget

Assegni	570	366	204
Cluster Pi	140	140	0
Trasferte	105	5	100
Large P	500	0	500
Tech. Dev.	160	35	125
TOTAL	1475	546	929

- WP3 - Back of envelope estimates

Esiste un range di volumi senza significativi colli di bottiglia????