Iniziativa di procurement e utilizzo efficente di piattaforme sperimentali per il calcolo in fisica teorica e HEP

Piero Vicini (INFN Rome)

Large scale computing at INFN - Roma Sapienza 13 FEB 2017

Progetto CIPE

- Nel 2016 finanziamento per il calcolo INFN: il progetto "CIPE".
- Finanziamento per
  - HPC di produzione per i gruppi teorici (sulla base del documento di settembre 2014 da updatare a breve)
  - update infrastrutture di calcolo degli sperimentali (Tier-xx)
  - sperimentazione di nuove architetture e convergenza delle piattaforme di calcolo (ad esempio come e cosa serve per strutturare il calcolo "opportunistico")
  - overheads vari...
- Ad oggi:
  - finanziate consistente numero di ass.ric. per teorici (e sperimentali) ($> 1ME$)
  - in progress il rifinanziamento dell'accordo attivo INFN-CINECA per core-hours aggiuntive su sistema Marconi (sulla base del documento di richieste 2014)
  - pianificati costi per update network/storage/computing al CNAF
- budget CIPE non ancora esaurito....
  - spazio per limitato finanziamento di attivita' di esplorazione "tecnologica" legata al computing HPC e HTC

Sistemi HPC: uno sguardo d'insieme

Cosa sono e come sono composti i sistemi HPC correnti o di prossima introduzione?

- *Multi million cores* supercomputers
  - $10^5 \div 10^6$ processori, $10^2 \div 10^3$ cores per processore
  - efficaci dal punto di vista dei ratio \$/Flops e \$/Power
- Alta granularita' e architettura ibrida –> CPU + acceleratori computazionali *many-core*
- US CORAL (Oak Ridge + Argonne + Livermore)
  - 525+M\$, 3x 100-200 PetaFlops systems nel 2018-19 (Pre-Exascale system), ExaScale in 2023
  - *Summit/Sierra* OpenPower-based (IBM P9 + NVidia GPU + Mellanox); 150(300) PFLops/10MW
  - *Aurora* Intel-based (CRAY/INTEL, Xeon PHI Knights Hill, Omnipath) 180(400) PFlops/13MW
- CHINA ??? , NUDT + Government
  - sistemi Tianhe (Xeon+KNL) basati su componenti "commerciali"
  - sistemi custom con CPU+GPU e network proprietaria : oggi 100PFlops SunWay TaihuLight, nel 2020 ShenWei and FeiTang...

Sistemi HPC: uno sguardo d'insieme (2)

- Anche MARCONI usa un approccio ibrido CPU+acceleratori
- La roadmap d'installazione (Aprile 2016-Luglio2017) prevede:
  - Aprile '16: Cluster basato su PC server; CPU Broadwell E5-2600 v4 (18 cores, 2 proc per node), 2PFlops integrati
  - Fine 2016 (Inizio 2017): Cluster addizionale basato su INTEL PHY (KNL, 70 cores) –> 11 PFlops
  - 2017 (estate): Sky Lakes (una sorta di architettura server "standard" ma con 20 cores) per 5 PFlops addizionali

quindi...

- Ad oggi NON esistono alternative commerciali a scala larga che implementino modelli alternativi a CPU + acceleratori many-cores
- I nostri codici scientifici non sono particolarmente ottimizzati per sfruttare il parallelismo estremo dei sistemi ibridi many-core.
- Simili discorsi valgono anche per il calcolo HTC degli sperimentali per ovvii motivi di opportunità e contenimento dei costi di procurement e operativi.
  - necessita' di scala larga per il computing offline per (ad esempio) gli esperimenti di HL-LHC.
  - "computing esotico": L0-1 trigger on-line,...
- La frazione maggiore di PFlops ottenibili dai sistemi HPC correnti e futuri viene, e verra', dal computing sugli acceleratori many-cores

–> dobbiamo creare le condizioni per un loro **utilizzo efficente**

Questa iniziativa...

- ...vuole lanciare un'attivita' di alfabetizzazione/discovery di queste architetture di calcolo orientatata ai giovani fisici dei gruppi computazionali dell'INFN (teorico e sperimentali) e ai (giovani..) tecnologi per imparare a
  - valutare le necessita' computazionale e la complessita' dei problemi di calcolo,
  - effettuare il loro porting efficiente sui sistemi many core
  - gestire l'hosting ed il supporto sistemisitico di piattaforme a scala larga
- basare questa attivita' sul procurement di uno o piu' sistemi di calcolo di taglia media, NON di produzione, da
  - installare in casa;
  - composto da componenti che (possiamo aspettarci) diventino il mainstream dei sistemi HPC del prossimo futuro.
  - equipaggiato da tutto il software e dalle librerie necessarie (e opzionali);
- Il tutto gestito (almeno inizialmente...) da un comitato ristretto
  - teorici: Biferale, Cosmai, Pepe;
  - sperimentali: Boccali, De Salvo;
  - esperti di tecnologia ed infrastrutture: Maron, Schifano, Vicini

Attivita' iniziale

Il comitato di gestione dovra' produrre, in tempi molto brevi,

- una survey sulla tecnologia corrente e futura
- una selezione dei codici di nostro interesse da usare come benchmark per questa attivita' (non e' detto che sia efficiente investire nel porting di TUTTI i nostri codici)
- una collezione dei requirements algoritmici e computazionali di tali codici ed una survey dei tools software necessari (compilatori, librerie, eventuali framework di supporto alla programmazione parallela)

Sulla base di questa analisi preliminare

- valutazione dei costi di procurement di sistema (attraverso interazione con fornitori)
- realizzazione del capitolato e supporto alla procedura di gara
- individuazione del sito d'installazione

A lungo termine...

Siamo appena partiti (prima confcall) ma e' chiaro che il procurement di macchina e' solo l'inizio dell'attivita'

- serve supporto dai gruppi teorici e sperimentali per realizzare le attivita' di alfabetizzazione ed esplorazione delle nuove architetture. Ci aspettiamo che le IS ed i gruppi sperimentali che hanno avuto assegni di ricerca possano contribuire con una frazione di questo manpower alle attivita';
- nei prossimi mesi dovremo completare lo schema delle attivita' post-installazione di macchina (scuole, workshop, seminari specifici e attivita' hands-on con professionisti,...).

-> il successo dell'iniziativa contribuira' a **mantenere ed incrementare il know-how necessario all'utilizzo efficente delle macchine HPC di prossima generazione**.