Simulazioni Covid-19

Tommaso Boccali -- INFN Pisa

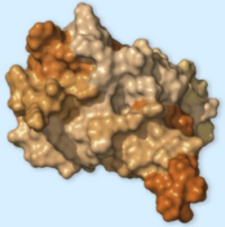
Outline

- Tipi di simulazioni
- Risorse necessarie
- Cosa sta facendo WLCG?
- Cosa stanno facendo gli Esperimenti?
- Cosa stanno facendo le Funding Agencies?

Tipi di simulazioni

- In grande maggioranza, le simulazioni solo di “folding”: date le trascrizioni dell’RNA delle proteine del virus e/o del suo target nel corpo umano (ACE2), capire come questi si dispongano spazialmente
  - La configurazione finale e’ importante per capire come si lega il virus (e come inibire questo legame con un’altra molecola che si leghi prima del corpo umano
  - Le configurazioni intermedie possono essere un modo di inserire “perturbazioni” che non permettano di arrivare alla forma finale

Cellular Saboteur · NSP1



This protein slows down the infected cell’s production of its own proteins. This sabotage forces the cell to make more virus proteins and prevents it from assembling antiviral proteins that could stop the virus.

Cosa di usa per simulare, e dove si simula?

- Da programmi home made, a oggetti open source con una lunga storia come Gromacs
- E chissa' che usano le ditte private …
- E' una simulazione di traiettorie nello spazio di 1-10^9 atomi
  - Ben parallelizzabile, alla fine solo interazione elettromagnetica (via Van Der Walls)
  - Scala bene su processori vettoriali e GPU
- Di nuovo, non sappiamo cosa usino le ditte farmaceutiche
- La ricerca di base, in sostanziale assenza di risorse dedicate, si e' avvivinata al volunteer computing
  - Poco input / output, CPU intensive
  - Poca RAM (1e9 atomi sono 3x1e9 coordinate nello spazio)
  - Poche dipendenze esterne (codice tight e autodefinito)
- Folding@Home e' il caso piu' noto ...

Folding@Home

- Gestito by Uni St Louis + others
  - Poco chiaro come sia gestita la scelta dei progetti scientifici
  - Ha “prodotto” 223 paper ad oggi
  - 1.5 exa flops raggiunti 10 gg fa

“Ma serve”? Boh, buona domanda

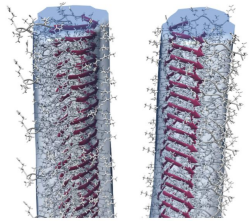
- Molti scettici che sia “un generatore di entropia”
- Il CERN non e’ riuscito a trovare una sponda WHO o EMBL su questo

Sibylla Biotech

- Questa iniziativa e' piu' confortante dal punto di vista della valenza scientifica per noi
  - Sibylla e' uno spin-off INFN -- la sua valenza e' stata gia' vagliata quando si e' deciso di partecipare
  - Press release della nascita dello spin-off:
- Cosa serviva?
  - In sibylla studiano ACE2 (angiotensin-converting enzyme 2), il recettore polmonare che si lega alle spike di covid
  - The simulation looks for intermediate shapes before reaching the final form. The intermediate shapes are then virtually screened against the already approved drugs in order to verify whether one of these is able to prevent ACE2 from reaching its final form, thus inhibiting the virus from infecting healthy cells. The final goal is to find a drug that can be used to treat a sick person.

14 OTTOBRE 2019

DALLA MECCANICA QUANTISTICA UNA SPIN-OFF PER SCOPRIRE NUOVI FARMACI



Scoprire nuovi farmaci grazie a innovativi protocolli di drug design e potenti algoritmi sviluppati a partire da metodi matematici di fisica teorica, originariamente sviluppati per studiare fenomeni tipici del mondo subatomico, come ad esempio l'effetto tunnel quantistico. È questa la sfida lanciata dalla neonata Sibylla Biotech, società Spin-off dell'Istituto Nazionale di Fisica Nucleare (INFN), delle Università di Trento e Perugia e su cui il fondo Vertis Venture 3 Tech Transfer ha deciso di puntare investendo 2,4 milioni di Euro. L'investimento è stato annunciato oggi in una conferenza stampa a Trento organizzata da Hit-Hub Innovazione Trentino.

Gli scienziati soci fondatori di Sibylla Biotech hanno sviluppato una piattaforma innovativa per la scoperta di nuovi target farmacologici sfruttando competenze che vanno dalla fisica sub-nucleare, all'informatica, alla biologia cellulare, alla chimica farmaceutica, in cui l'interdisciplinarità è la chiave vincente. Sibylla, che ora continuerà lo sviluppo della tecnologia, ne ha acquisito la licenza da Istituto Nazionale di Fisica Nucleare (INFN), Università di Trento e Fondazione Telethon che hanno finanziato la ricerca. Elemento cruciale di questo nuovo approccio alla ricerca di potenziali farmaci è l'impiego di un metodo di calcolo che si basa su metodi matematici di fisica teorica che sono stati originariamente sviluppati per studiare fenomeni tipici del mondo subatomico, come l'effetto tunnel quantistico, poi adattati per simulare processi biomolecolari complessi, come il ripiegamento e l'aggregazione di proteine. Il metodo è stato impiegato con successo per lo studio della replicazione del prione, la proteina responsabile del morbo della mucca pazza, e ha portato al primo modello computazionale realistico al mondo di questo meccanismo. Un risultato pubblicato sulla rivista Plos Pathogens e dalle enormi potenzialità per la futura ricerca mirata di farmaci in grado di contrastare gravi malattie oggi incurabili. Il risultato è stato ottenuto lo scorso luglio dal gruppo di ricerca dell'Istituto Telethon Dulbecco presso l'Università di Trento guidato da Emiliano Biasini, in collaborazione con il gruppo di Pietro Faccioli del dipartimento di Fisica dello stesso ateneo e afferente all'Istituto Nazionale di Fisica Nucleare.

GROMACS
FAST. FLEXIBLE. FREE.

SW = modified GROMACS

Cosa serve?

- Il folding gromacs parte da configurazioni random , e segue la dinamica del folding fino ad arrivare (a volte) alla configurazione finale
- Ogni folding era stimato in 10 gg su 32 threads, e per ottenere risultati statisticamente sensati erano stimate almeno 800 simulazioni
  - 800*32*10gg = 6 M core.H, molto al di sopra delle capacita' di calcolo di Sibylla
- Idea:
  - Girare in parallelo quegli 800 jobs su struttura INFN, in modo da avere il risultato richiesto in ~10 gg invece di mesi
  - (i 10 gg non sono splittabili ulteriormente perche' ogni job e' una linea temporale, da eseguire serialmente)
- 800*32 = 25600 cores… questa era la stima delle risorse necessarie

Che tipo di risorse?

- Gromacs ha bisogno (per garantire questa tempistica) di AVX2
  - Opteron e Intel pre 2014 non usabili
- Necessari slots batch da > 10 gg, 32 threads
  - Non possibile eseguire il payload su code WLCG
- Necessario GCC9 e Centos7
- Necessario poco IO (ogni job genera ~ ½ GB in 10 gg) e poca RAM (~10 GB per 32 threads)
- Pero' serve spazio condiviso di lavoro
  - Gromacs fa checkpointing e ogni tanto crasha
  - Cosi' era la struttura degli script che abbiamo ricevuto

Siti facili ...

- Code necessariamente da fare ad hoc
- Siti con filesystem POSIX condiviso (GPFS, LUSTRE, Ceph, NFS) → facili
- Siti con Centos7 → facili
- Siti con LSF (LSF e' stato usato al CNAF per la prima integrazione) → facili

Facendo l'AND:

- Facili CNAF, PI, RM1

(che sono stati i primi ad entrare in produzione)

Siti “piu’ difficili”

- No storage condiviso: messa in produzione ad-hoc area NFS (MI, NA, LNF, LNL)
- No LSF: tradotti scripts in LSF (BA, MI, …?)
- No Centos7: usato un container singularity di CMS (con un po’ di ore spese a integrarlo nel workflow) (LNL/PD)

- Per GCC9, abbiamo usato il fatto che tutti i centri montano CVMFS, e abbiamo usato una release area sperimentale di CMS con GCC9

Risultati Sibylla

- Alla fine sui nostri hardware medi l'esecuzione era piu' vicina a 5-7 giorni che a 10
- Abbiamo quindi in un solo giro (quindi sottomissione totalmente in parallelo, nessun job tenuto pending) circa 1200 jobs rispetto agli 800 richiesti
    - Maggiore statistica aiuta ed e' stata gradita

SITE	Max number of cores for Covid Sibylla simulation	CPU source	Core h dedicated to the project
CNAF	17500	Generic	3630067
PI	2556	CMS, Theory	498284
BA	1280	ReCaS, CMS, ALICE	251725
RM1	1536	ATLAS	347525
MI	1280	ATLAS	145103
LNL-PD	2560	CMS, ALICE (70%/30%)	416083
NA	1024	ReCaS, ATLAS	185536
LNF	1280	ATLAS	128576
TOT CPU Cores	**29016**	**TOT Core.h**	**5602899**

Report

- Mandato a Zoccoli come richiesto, ovviamente con i ringraziamenti alla gente che ha partecipato!
- Estrema elasticita' del sistema calcolo italiano: in poche ora sono state create code, svuotate macchine, installate aree condivise, creati account!
- Press release INFN

CNAF

Stefano Dal Pra
Luca Dell'Agnello
Gaetano Maron

Pisa

Tommaso Boccali
Enrico Mazzoni

Bari

Giacinto Donvito
Alessandro Italiano

Roma1

Cristina Bulfon
Daniela Anzellotti
Alessandro de Salvo

Laboratori di Legnaro / Sezione di Padova

Massimo Biasotto (LNL)
Alberto Crescente (Padova)
Sergio Fantinel (LNL)
Massimo Sgaravatto (Padova)

Laboratori di Frascati

Agnese Martini
Elisabetta Vilucchi

Milano

David Rebatto

Napoli

Gianpaolo Carlino
Alessandra Doria

03 APRILE 2020

SIMULAZIONI BIOMOLECOLARI E URGENT COMPUTING: LA SFIDA ITALIANA AL COVID 19

Bloccare il virus prima che riesca a penetrare nelle cellule neutralizzando il suo cavallo di Troia. È questa la sfida lanciata al COVID-19 da Sibylla Biotech, spin-off dell'Istituto Nazionale di Fisica Nucleare (INFN) e delle Università di Trento e Perugia, che sta lavorando all'identificazione di una terapia anti COVID-19 con una tecnologia unica che studia il comportamento delle proteine in un modo completamente nuovo e in tempi rapidi.

Poiché, per riuscire a contrastare in modo efficace la diffusione del virus, agire tempestivamente è determinante, l'INFN ha deciso di dedicare una quota significativa delle sue risorse di calcolo per aiutare Sibylla Biotech. L'ente ha così coinvolto in questa sfida i data center dedicati solitamente all'analisi dei dati degli esperimenti di LHC (sezioni di Bari, Milano, Napoli, Padova, Pisa, Roma e Laboratori di Frascati e Legnaro) e il proprio centro di calcolo nazionale, il CNAF di Bologna, che da solo conta oltre la metà della potenza computazionale a disposizione di INFN.

E ora?

- (DA QUI IN POI MIE OPINIONI PERSONALI!!!)
- Sibylla non ha attualmente altre necessita’
- Si potrebbe fare Folding@Home?

Cosa ne pensa WLCG?

2 Management Board specifici WLCG su questo

- Posizioni non univoche, dando per assunto che nessuno dice che dobbiamo continuare a processare miliardi di eventi monte carlo come se nulla fosse:

1. Opzione #1:
    a. Le Funding Agencies possono rimuovere temporaneamente dal pledge risorse per eseguire simulazioni e/o quello che sembri rilevante
    b. In tal modo gli esperimenti NON avranno il pledge del 2020, ma la cosa sara' spiegata e difesa ai referees nel 2021
    c. Il jobs che girano come esperimento, e che entrano nelle statistiche ufficiali (ore CPU, efficienza CPU, IO, ...) avranno senso perche' come VO WLCG continueranno a girare solo payload di esperimento
2. Opzione #2:
    a. Gli esperimenti, tramite il loro WMS, manderanno jobs di simulazione (per esempio Folding@Home) al posto dei workload standard
    b. In questo modo le FA daranno i pledge come promessi, ma gli esperimenti dovranno spiegare ai referees che parte di quel pledge sara' stato usato per altro
    c. I plots di esperimento non avranno particolare senso
        i. Il pledge sara' raggiunto ma sara' falso per le attivita' di esperimento
        ii. Efficienza CPU etc saranno falsate dai workflow covid
        iii. ...

...

Nessuna politica forzata da WLCG, se non

- Qualunque cosa facciate, documentate la e siate poi in grado di quantificarla (sia Exp, sia FA)
- Si e' raccomandato agli Exp, nel caso vogliano farlo, di "partire piano"

E l'Italia che fa? ....

(avrete capito che io sarei per l'opzione 1....)

Che io sappia, ad oggi

- ALICE e ATLAS vorrebbero girare il 5% di simulazioni covid nei loro workload
- CMS e LHCb stanno usando farms HLT per questo, che NON sono nel pledge

Team: CERN

Date of last work unit	2020-04-16 16:32:28
Active CPUs within 50 days	172,878
Team Id	38188
Grand Score	789,985,724
Work Unit Count	775,061
Team Ranking	305 of 250866
Homepage	http://public.web.cern.ch/public/
Fast Teampage URL	https://apps.foldingathome.org/teamstats/team38188.html

Team members

Rank	Name	Credit	WUs
2,021	CERN_Cloud	333,986,636	230,769
4,192	ATLAS_CPU	138,147,890	217,468
6,571	CMS-Experiment	76,173,680	107,339
13,438	TheLaboratoire	29,493,588	479
14,303	ALICE-CERN	27,216,757	45,513
15,601	BNL_HPC_CPU	24,286,439	3,112
18,496	Cloverfield	19,377,132	207
18,686	ALICE-FLP	19,277,191	2,247
1	Anonymous	10,919,296	1,925
38,158	ALICE-FZK	8,070,940	13,430