Calcolo distribuito volontario:
l'esperienza Boinc

Stefano Bologna

Sommario

Negli ultimi decenni l'home computing e internet hanno modificato la nostra vita quotidiana, ma hanno anche cambiato radicalmente il modo di fare ricerca. Questa doppia rivoluzione, recentemente, ha permesso un avvicinamento tra il mondo della scienza e il cittadino, anche attraverso il calcolo distribuito. La piattaforma Boinc, la più usata per il calcolo distribuito volontario, viene descritta nella sua storia, nel suo funzionamento e nei risultati conseguiti.

Abstract

In last decades, the home computing and the Internet have changed our daily life, but also changed the way of doing research. This double revolution, recently, allowed an approach between the world of science and the citizen, through the distributed computing. Boinc platform, the most widely used for voluntary computing, is described in its history, its operation and results.

Keywords: Volunteer computing, citizen science, boinc, middleware, in silico

1. Introduzione: cos'è il calcolo distribuito?

Cos'è il calcolo distribuito? Lo dice, in maniera sintetica, il nome stesso, ovvero distribuire un carico di lavoro tra più computer in modo da poterlo velocizzare. Le simulazioni in silico, nate all'inizio degli anni 80 grazie alla lungimiranza di alcuni ricercatori come Martin Klarplus, sono diventate uno strumento quotidiano per moltissimi scienziati delle più diverse branche. Ricercatori di tutto il mondo spesso necessitano di incredibili potenze di calcolo (con server molto costosi) dell'ordine dei vari multipli di Tflops (floating point operation per second), per i loro studi e spesso queste richieste non sono accolte nell'immediato. Con l'avvento di internet e dell'home computing, alcuni scienziati hanno ritenuto che un computer di casa o dell'ufficio usato mediamente al 10-15% delle proprie potenzialità, potesse divenire parte di una rete (grid) e che questo portasse ad indubbi vantaggi computazionali. In tal senso erano state create alcune piattaforme già nei primi anni 90, come ad esempio Condor, che permettevano di chiedere aiuto via internet a volontari, i quali mettevano a disposizione la

potenza inutilizzata dei propri pc. Ai loro computer venivano quindi "spedite" delle minuscole porzioni in cui era stato suddiviso il lavoro (in gergo wu, ovvero working unit), che venivano elaborate. I risultati ottenuti, infine, erano rispediti ai ricercatori che provvedevano ad analizzarli e a trarne conclusioni.

2. Boinc: le caratteristiche e un po' di storia

La piattaforma più nota ed importante è BOINC (Berkley Open Infrastructure for Network Computing) [1]**,** un sistema non commerciale su cui vengono eseguiti una vastità di progetti che vanno dalle simulazioni proteiche all'astrofisica, dalla ricerca climatologica a quella matematica, ecc, ecc. Il suo funzionamento è molto simile alle piattaforme precedenti, quindi un sistema midlleware client/server, ma con un occhio particolare alla sicurezza e alla semplicità d'uso. BOINC è stato sviluppato presso lo Space Sciences Laboratory della University of California Berkeley, da un team diretto dal ricercatore David Anderson. Nel 2002, dietro richiesta di altri ricercatori, lo scienziato pensò ad una piattaforma "eterogenea", che potesse contenere tutta una serie di diverse simulazioni.

Un esempio tra i molti è legato al progetto "*Help Fight Childhood Cancer*" (aiutaci a sconfiggere il cancro infantile), supportato dall'università giapponese di Chiba e dalla IBM. I suoi ricercatori hanno calcolato, per la prima fase della ricerca, che con i server a loro disposizione, impegnati 24 ore su 24, 365 giorni l'anno, avrebbero impiegato circa 8000 anni per finirla. Grazie a BOINC e ai suoi volontari, i ricercatori l'hanno completata in meno di 2 anni. [2]

Alla stregua di una piattaforma HPC (High Performance Computing, ovvero grandi server dedicati alla ricerca), BOINC conta circa 433.000 computer attivi (hosts) in tutto il mondo che elaborano una media di 8,6 petaFLOPS (a Febbraio 2015), posizionandosi - come capacità di calcolo – al sesto posto tra i supercomputer più potenti al mondo[3]. Il software client è supportato da tutti i più diffusi sistemi operativi, come Microsoft Windows, Mac OS X e sistemi Unix-like tra cui Linux e FreeBSD. La licenza di rilascio è LGPL, i sorgenti possono quindi essere scaricati liberamente e modificati per integrare le necessità di ogni struttura.

I 13 anni che sono trascorsi dalla nascita della piattaforma sono costellati da una costante evoluzione hardware che, anche grazie all'evoluzione software, ha reso possibile l'utilizzo delle nuove tecnologie: il passaggio delle cpu (central processing unit) single-core a quelle multi-core, l'utilizzo delle gpu (graphic processor unit, le schede video che si trovano comunemente nei nostri pc) (**RIQUADRO GPGPU – General Purpose on Gpu**), per finire con l'avvento del calcolo distribuito anche su piattaforme smart come Raspberry Pi.

3. Boinc Manager (Client)

Il funzionamento del client di gestione dei progetti è molto semplice: è sufficiente scaricare dal sito www.boincitaly.org, tradotto in italiano dal nostro team, ed installarlo. Una volta avviato il client Boinc bisogna scegliere il progetto a cui si vuole partecipare, iscriversi ad esso (basta una email ed una password) e lanciare l'elaborazione. La fase successiva viene fatta automaticamente, e non

sono richieste particolari conoscenze tecniche da parte degli utenti (figura 1): sarà il client stesso, infatti, ad analizzare la configurazione hardware dell'host (per esempio, rilevare la presenza di cpu in grado di supportare istruzioni SIMD, come le SSEx o le AVX) e richiedere al server il carico di lavoro adeguato.

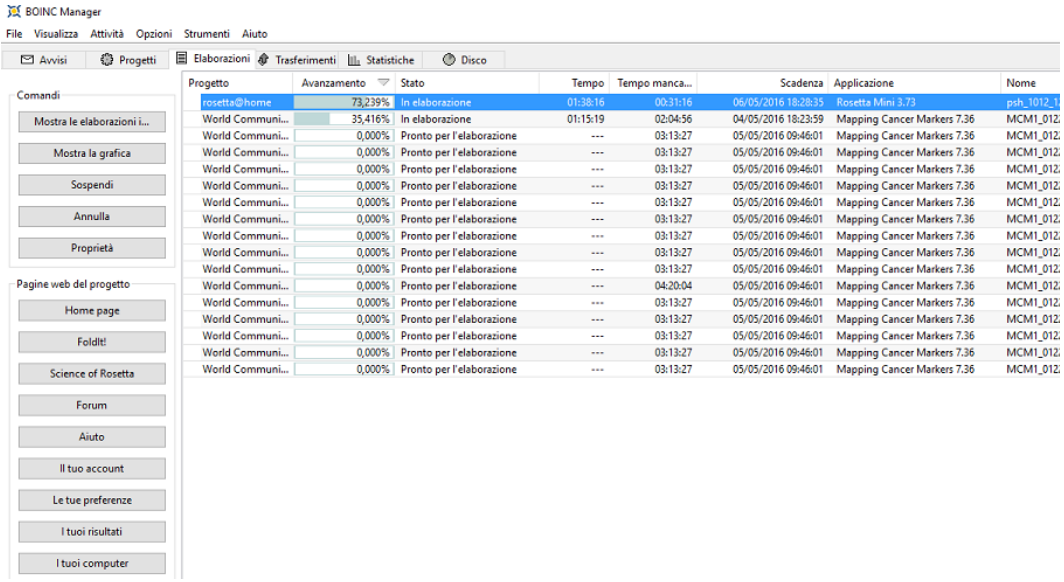

Figura 1
Boinc Manager italiano

Il client Boinc è ampiamente configurabile: è possibile, per esempio decidere quanti core cpu dedicare, oppure gli orari in cui elaborare i dati, ecc, ecc. Con l'avvento delle tecnologie mobili **(vedere riquadro dispositivi mobili)**, è stato creato anche un manager facilitato che possa essere facilmente fruibile da dispositivi mobili quali gli smartphone o i tablet. E' un sistema facile e sicuro, che non danneggia il pc e soprattutto non mette a rischio la privacy degli utenti, dal momento che crea un ambiente protetto che non va ad influenzare eventuali altre applicazioni.

4. Boinc Server

Il server Boinc è quel server collocato presso il centro di ricerca/università che si occupa di creare e distribuire le porzioni di lavoro, oltre che a riceverne i risultati già calcolati e successivamente a convalidarli (figura 2). Dal momento che il lavoro è fatto "esternamente" dai client, questi server non hanno la necessità di essere potenti: anche un giovane ricercatore con una buona idea o una facoltà con pochi fondi hanno la possibilità di accedere a potenze di calcolo ingenti **(vedere riquadro virtual campus)**. Il team di amministratori di Boinc ha pensato, per facilitare l'adozione di questo sistema, di creare una macchina server virtuale già pronta all'uso[4], così che i ricercatori possano concentrarsi maggiormente sulla parte scientifica che sulla creazione dell'infrastruttura.

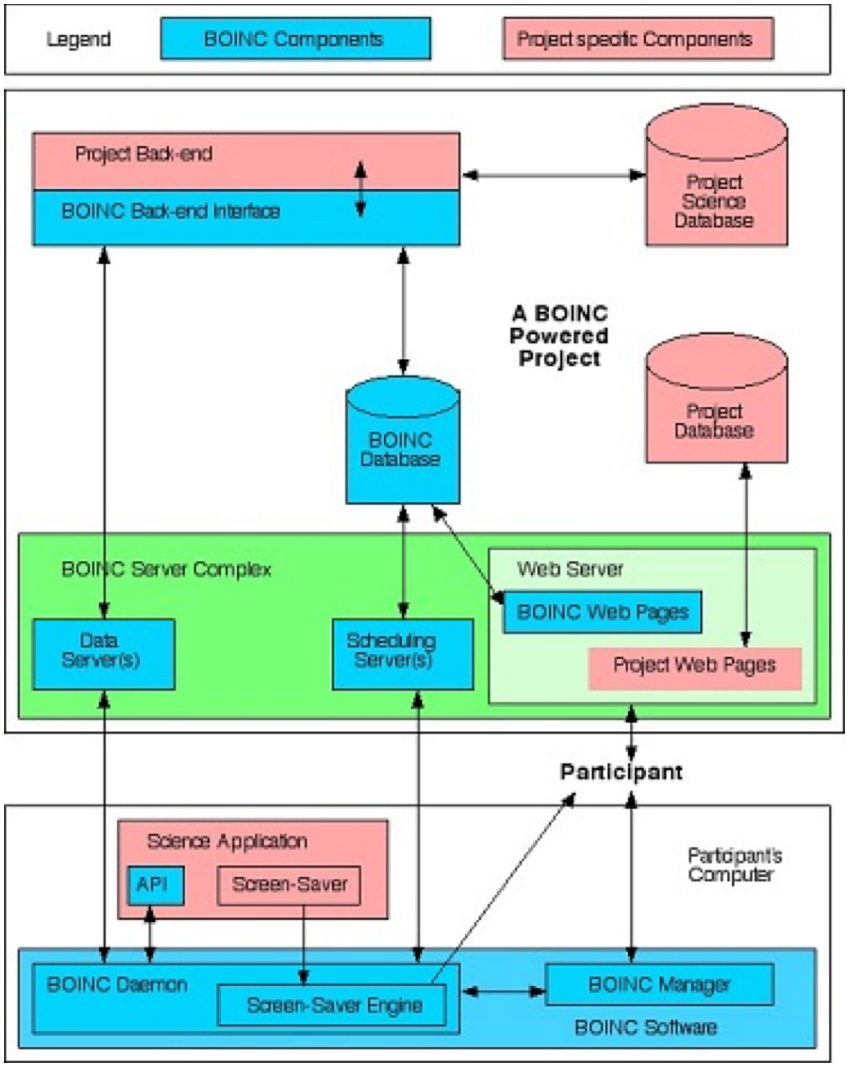

Figura 2
Lo schema di funzionamento della piattaforma

5. I Progetti

Ogni progetto ha un proprio applicativo che viene fatto elaborato nell'ambiente boinc e che permette di utilizzare i pc dei volontari, sfruttando sia i processori, sia le schede video che i dispositivi mobili. I progetti riguardano i più svariati campi della scienza e al momento sono circa una cinquantina. Di seguito alcuni esempi:

Seti@Home[5]: il primo e più famoso progetto di calcolo distribuito, nato per la ricerca di segnali di vita extraterrestre, negli anni ha portato anche ad ottimi risultati scientifici (come lo studio del rumore di fondo dell'universo).

Rosetta@Home[6] (figura 3): è uno dei progetti più "vecchi" del panorama Boinc e si occupa di simulazioni proteiche (folding, docking, ab initio, ecc) con risultati spesso pubblicati su riviste importanti come Science o Nature [7].

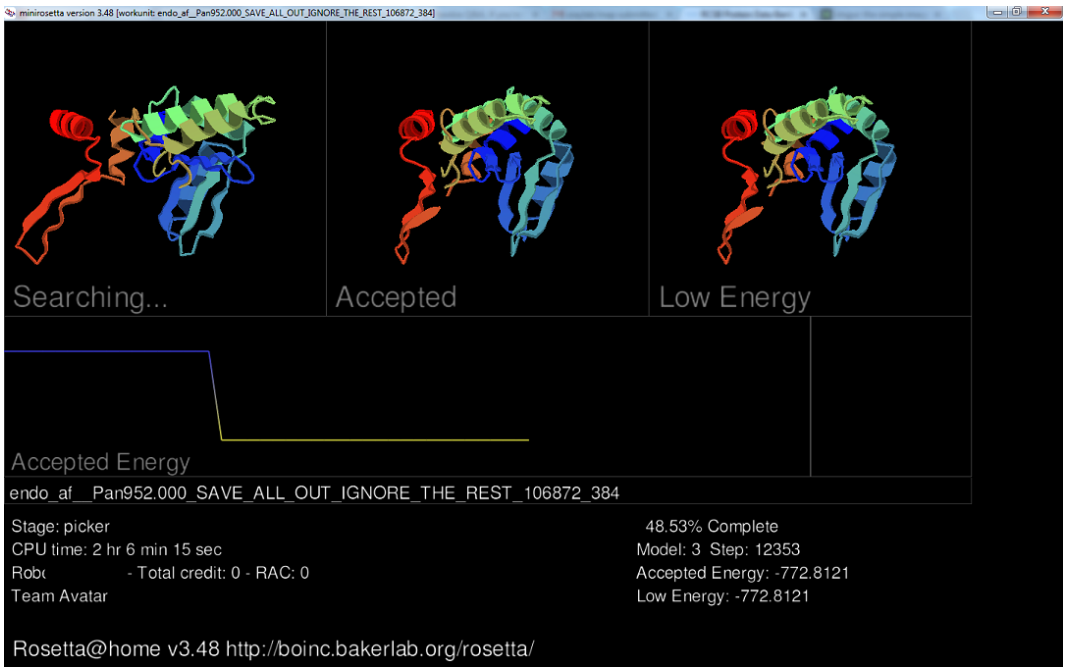

Figura 3
screensaver di Rosetta@Home

LHC@Home[8]: Il CERN di Ginevra è particolarmente sensibile al calcolo distribuito e ha creato un progetto, con vari sottoprogetti, con profondi legami scientifici con l'accelleratore LHC.

Einstein@Home[9]: un importante progetto di ricerca nel campo dell'astrofisica, specializzato nello studio delle pulsar, che si prefigge di vedere se, come previsto dalla teoria di Einstein, le pulsar sono in grado di generare onde gravitazionali.

World Community Grid[10]: è un multi-progetto finanziato dall'IBM. Al suo interno comprende molti progetti che vanno dallo studio di migliori pannelli fotovoltaici a metodi che usano le nanotecnologie per filtrare l'acqua, alla ricerca sul cancro. Qualsiasi ricercatore può partecipare, presentando il proprio progetto alla IBM, la quale metterà a disposizione server ed expertise.

6. Il progetto Italiano

In Italia, purtroppo, questo tipo di ricerca scientifica è ancora poco noto anche se, grazie al progetto supportato dall'Università di Trento e dal CNR, ci auguriamo possa divenire sempre più popolare.

Il progetto in questione, Tn-Grid, si occupa di ricerche legate alle reti geniche e delle loro relazioni di causa. [11](figura 4). Al momento attuale i ricercatori stanno processando dati riguardanti le LGN (local gene network) dell'Escherichia Coli, noto organismo modello.

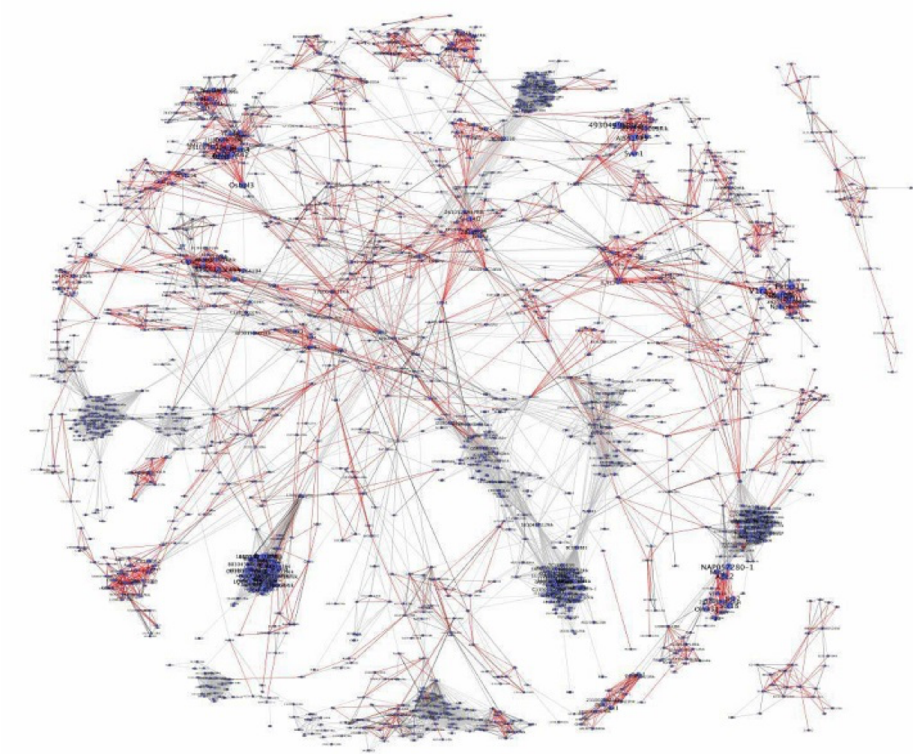

Figura 4
mappa genica Tn-Grid

7. I risultati

I progetti sono senza scopo di lucro e i risultati sono pubblici [12]. Il calcolo distribuito volontario, d'altro canto, non vuole sostituirsi al lavoro che si fa in laboratorio, ma anzi vuole anzi sostenerlo e accellerarne le tempistiche e la precisione.

8. I volontari

Chi aderisce ad un progetto di ricerca (o a più progetti) non lo fa per ricevere ricompense, ma per poter aiutare la scienza senza il vincolo di risultati minimi da raggiungere: vi potrà contribuire in maniera spontanea per il tempo che vorrà e con quanti computer vorrà. Il calcolo distribuito, in questo senso, va interpretato come una nuova forma di beneficenza o di volontariato. Oltre al questa forma di partecipazione "passiva", è possibile partecipare anche attivamente: sui forum dei progetti è possibile, per esempio, dialogare direttamente con i ricercatori stessi; oppure, aiutarli a migliorare il software stesso, dal momento che i sorgenti degli applicativi sono, per la quasi totalità, opensource (oppure con accesso riservato alle università).

E' questo lo spirito che anima quello che viene definito il movimento della "Citizen Science", ovvero di una scienza aperta alla gente e a cui è possibile contribuire fattivamente[13].

9. I crediti

Nonostante i progetti non assegnino ricompense a chi vi partecipa, per ogni unità di lavoro completata e riconsegnata correttamente viene assegnato un punteggio. Questo permette ai volontari di creare classifiche personali o per team, e di partecipare a sfide simboliche i cui unici beneficiari sono i progetti stessi.

10. Boinc Italy

Il gruppo Boinc Italy [14], composto da oltre 5000 volontari, si è aggregato attorno all'idea di poter concretamente aiutare i ricercatori e, negli anni, si è imposto come una realtà (per potenza di calcolo e partecipazione) a livello internazionale. La partecipazione al gruppo è aperta a tutti e tutti possono, nel portale, trovare risposta ai propri dubbi/curiosità e approfondimenti sui singoli progetti. Si occupano, inoltre, della traduzione di tutte le news dei progetti e dell'organizzazione delle "sfide" nazionali ed internazionali sui progetti quali, ad esempio, quelle dedicate a Rita Levi Montalcini o quella dedicata a Margherita Hack.

11. Le criticità

A fronte degli indubbi vantaggi derivanti dall'uso della piattaforma boinc, ci sono alcune criticità che devono essere considerate al momento della creazione di un progetto:

1. Boinc non è adatto ad eseguire calcoli real-time (come i sistemi cluster), ovvero quelle simulazioni che necessitano di elevato parallelismo per poter ottenere una risposta soddisfacente. Un esempio di cluster italiano dedicato alla ricerca scientifica è il sistema GALILEO, presente presso il Cineca, con i suoi 8000 core che lavorano in simultanea.
2. La volatilità del lavoro (un client può anche non restituire i risultati, per i più disparati motivi). Questo problema viene attenuato con l'utilizzo del quorum, ovvero con l'invio di una o più copie dello stesso lavoro a diversi pc, così da ottenere almeno un risultato utile.
3. Una presenza non stabile della potenza di calcolo erogata dai client, con picchi e momenti di "stanca".

Il futuro è qui

Le difficoltà che vengono incontrate più spesso dagli amministratori di Boinc Italy nel divulgare questa risorsa sono principalmente due: primariamente la paura degli istituti di ricerca di vedere "rubato" il proprio lavoro, derivante in genere dalla scarsa comprensione che la frammentazione del lavoro da svolgere introdotta da Boinc, impedisce la divulgazione delle informazioni all'esterno dell'ambiente di ricerca. Secondariamente un certo timore, da parte dei nuovi volontari, di non capire appieno i progetti (dal momento che tutte le home page sono in lingua inglese), nonostante il continuo lavoro di traduzione del gruppo.

La continua evoluzione (e l'abbattimento dei costi) dell'home computing, rendono il calcolo distribuito su base volontaria una alternativa veramente valida per chi vuol fare ricerca attraverso la scienza computazionale.

Riquadro GPGPU ("general purpose computing on graphic processor unit – calcolo a scopo generale su unità di elaborazione grafiche". Wikipedia).

L'utilizzo di schede video per attività "general purpose" è stato sicuramente uno dei punti di svolta dell'informatica moderna, grazie ai costi relativamente bassi e alla potenza di calcolo messa a disposizione. Il primo settore a rendersi conto delle possibilità insite nelle schede grafiche, è stato ovviamente quello dei videogiochi: non a caso due dei maggiori vendor di schede per uso gpgpu sono AMD e Nvidia, leader nel settore dei videogames. I principali linguaggi di programmazione ad alto livello utilizzati sono due: Cuda ed OpenCl. Il primo è un linguaggio proprietario della casa produttrice Nvidia e funziona solo sulle loro gpu, mentre OpcnCl è uno standard aperto supportato da AMD che può girare, dopo le ovvie modifiche, su un variegato parco hw (cpu, gpu, fpga, Xeon Phy, ecc). Gli incrementi, con codice adeguato, vanno dal raddoppio della potenza rispetto ad una cpu fino a fattori 100x.

Riquadro Virtual Campus

Gli amministratori di Boinc propongono, tra le varie opzioni, la creazione di "Virtual Campus Supercomputer", ovvero progetti ospitati totalmente all'interno delle università, ricercando la potenza di calcolo nei vari pc presenti (per esempio negli uffici amministrativi), utilizzati per poche ore al giorno e con compiti poco onerosi dal punto di vista dell'utilizzo delle risorse. Per esempio in un campus con 5.000 pc in funzione al 50% delle risorse, la potenza di calcolo corrisponderebbe ad un cluster del costo di circa 2 milioni di dollari (con 500 mila dollari di mantenimento annuale). Utilizzando Boinc, il costo sarebbe di poco più di 10.000 dollari.

Riquadro Dispositivi Mobili

Negli ultimi anni sono nati molti progetti Boinc che utilizzano la piattaforma Android per eseguire le proprie simulazioni (semplicemente scaricando l'app dallo store). Il client Boinc sviluppato per l'utilizzo su dispositivi mobili è stato creato per permettere il completo controllo, da parte dell'utente, dei progetti scelti. Il client, infatti, si avvierà solo alla connessione della ricarica elettrica (non andando ad intaccare, così, la batteria) e solo in presenza di rete wireless (non utilizzando il traffico della Sim) e sarà possibile, inoltre, controllare anche altri parametri: quanti core utilizzare, l'uso dello spazio, ecc, ecc. La relativa potenza di calcolo di smartphone e tablet (ancora molto inferiore a quella di un pc) viene, però, compensata dalla loro massiccia adozione e dalla facilità d'uso degli stessi.

Bibliografia

[1] http://boinc.berkeley.edu/ (Settembre 2015)

[2]https://secure.worldcommunitygrid.org/research/hfcc/overview.do (Settembre 2015)

[3]https://it.wikipedia.org/wiki/Berkeley_Open_Infrastructure_for_Network_Computing (Settembre 2015)

[4] https://boinc.berkeley.edu/trac/wiki/VmServer (Settembre 2015)

[5] http://setiathome.ssl.berkeley.edu/ (Settembre 2015)

[6] https://boinc.bakerlab.org/ (Settembre 2015)

[7] Pearson A.D., Mills J.H., Baker D. et al. "*Transition states. Trapping a transition state in a computationally designed protein bottle*" Science 347(6224), pp 863-867, 2015.

King N.P., Bale J.B, Sheffler W., Baker D. et al. - "*Accurate design of co-assembling multi-component protein nanomaterials*". Nature 510 (103-108), 2014.

[8] http://lhcathome.web.cern.ch/ (Settembre 2015). Un esempio di pubblicazione è:

Karneyeu A., Mijovic L., Prestel S., Skands P.Z. *"MCPLOTS: a particle physics resource based on volunteer computing"*. European Physical Journal C 74, pp. 1-22, 2014.

[9] http://www.einsteinathome.org/ (Settembre 2015)

[10] https://secure.worldcommunitygrid.org/ (Settembre 2015)

[11] http://gene.disi.unitn.it/test/ (Settembre 2015). Nonostante il progetto sia ancora in fase di test e l'accesso non sia completamente pubblico, ma su invito, ha già raggiunto alcuni risultati, presentati in conferenze scientifiche internazionali, come:

Erculiani L., Asnicar F., Sella N., Cavecchia V. et al. - *"Discovering Candidates for Gene Network Expansion by Variable Subsetting and Ranking Aggregation"*. Network Biology in SIG Poster, 2015.

[12] http://www.boincitaly.org/progetti/pubblicazioni-scientifiche.html (Settembre 2015)

[13] https://www.whitehouse.gov/blog/2015/09/09/open-science-and-innovation-people-people-people (Settembre 2015)

[14] www.boincitaly.org (Settembre 2015)

Biografia

Stefano Bologna informatico con laurea in filosofia, è uno degli amministratori del gruppo Boinc Italy, nonché referente italiano dei progetti Correlizer e Citizen Science Grid. Mi occupo, inoltre, di tenere aggiornata la lista delle pubblicazioni scientifiche legate a Boinc e che risulta essere la più completa a livello internazionale.

Email: stefano.bologna@boincitaly.org