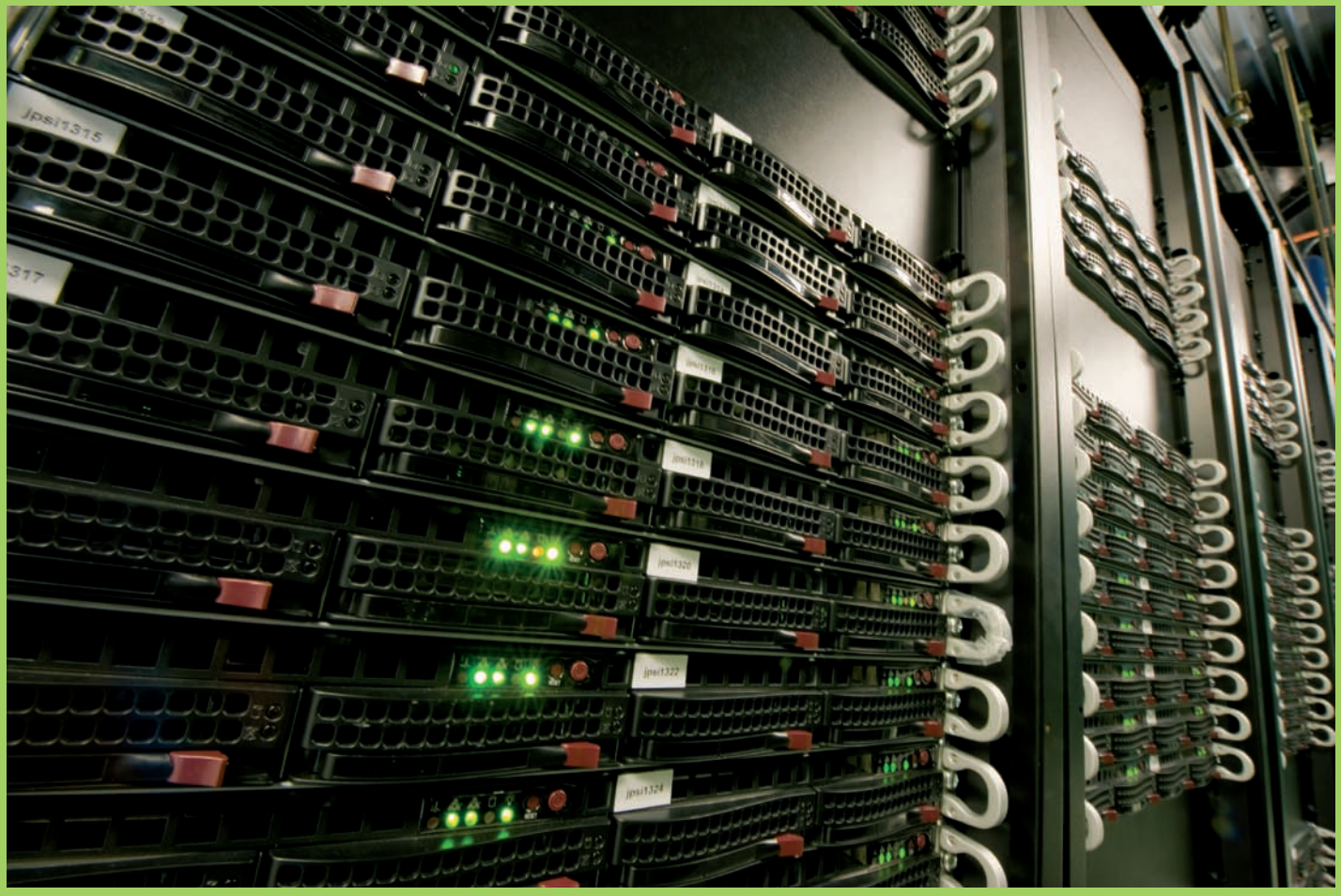
Griglie e nuvole

Le nuove infrastrutture digitali

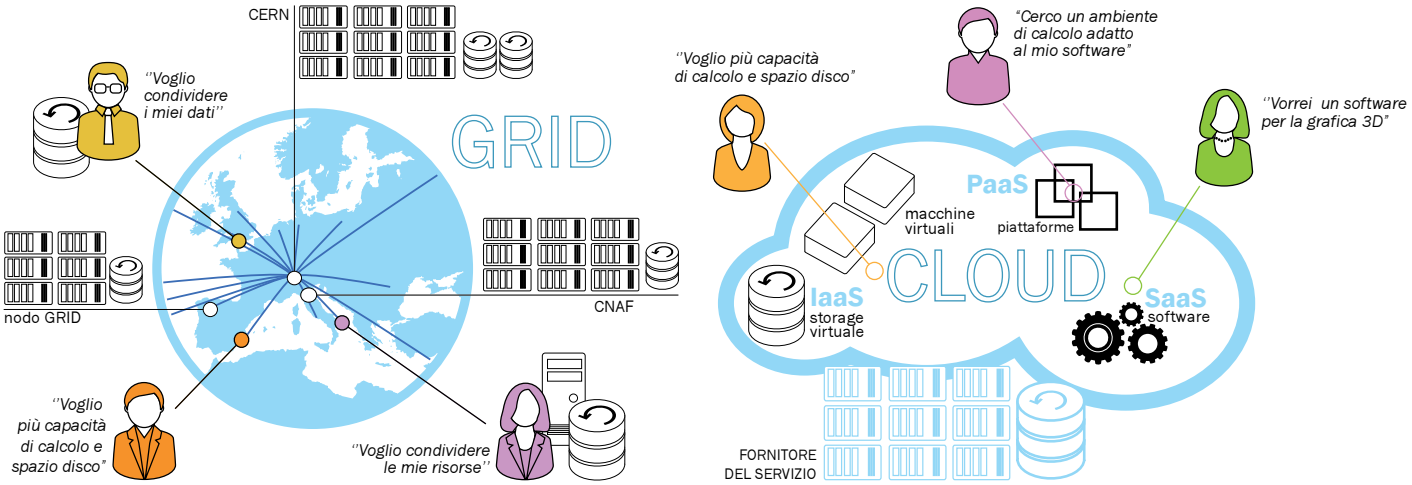
di Leonardo Merola

Ormai tutti siamo abituati a usare computer da tavolo, portatili, tablet o smartphone per risolvere problemi o rispondere alle nostre curiosità: in pochi attimi possiamo recuperare informazioni dall'altra parte del mondo, vedere immagini di luoghi lontani, assistere in diretta ad avvenimenti altrimenti inaccessibili. L'accesso a questi servizi richiede però memorie e capacità di calcolo che vanno ben oltre quelle contenute nei nostri dispositivi.

Un semplice esempio è fornito dai *social network* come Facebook o Twitter, che coinvolgono oggi circa un miliardo di persone. Ogni utente utilizza in media 10 gigabyte, quindi in totale si tratta di 10 miliardi di gigabyte, una quantità di dati e informazioni non gestibile da un singolo computer, la cui capacità massima è di qualche migliaia di gigabyte. Se poi passiamo alla condivisione e all'analisi dei dati acquisiti da un esperimento scientifico, come quelli oggi in corso al Cern di



a.
Il centro di calcolo del laboratorio Fermilab, negli Stati Uniti costituisce uno dei principali nodi della Grid americana.



Ginevra, sono milioni i gigabyte di dati complessi elaborati ogni anno per comprendere la struttura e l'origine dell'intero universo**.** Esigenze di questo tipo, insorte soprattutto nell'ultimo decennio, hanno portato allo sviluppo di due nuovi paradigmi nel campo della gestione dei flussi informatici, quello del Grid computing (dall'inglese *grid*, "griglia") prima e quello del Cloud computing (dall'inglese *cloud*, "nuvola") dopo: due modelli complementari e con obiettivi diversi.
Il Grid computing è nato in ambito scientifico per risolvere il problema di rendere disponibili un insieme di centri di calcolo distribuiti in tutto il mondo, per applicazioni la cui elaborazione è particolarmente onerosa. Una volta definito un problema in termini di calcoli da eseguire e di dati da utilizzare, un ricercatore può richiedere l'elaborazione del calcolo a una media di 100.000 server connessi e sparsi in molti centri nel mondo, suddividendo così la sua richiesta di elaborazione tra un gran numero di risorse diverse (vd. fig. b, a sinistra) e riducendo così enormemente i tempi di calcolo. Inoltre la Grid permette a utenti diversi, accomunati sotto la stessa "Virtual Organization", di condividere dati e programmi, con notevole risparmio di risorse di *storage*.
Il Cloud computing, in modo complementare, ha come principale obiettivo quello di distribuire attraverso la rete risorse di calcolo e di *storage* dei dati senza che gli utenti pubblici e privati (che sono per loro natura interessati ai servizi e non all'infrastruttura) si debbano occupare della complessità della loro gestione. Oggi è possibile condividere un documento con altri utenti accedendo, ad esempio, alle "nuvole" Dropbox o iCloud in modo "virtuale", senza conoscere l'ubicazione delle risorse "fisiche" (disco e macchine) del fornitore del servizio (*provider*). Grazie al concetto del Cloud, l'infrastruttura diventa così un servizio (vd. fig. b, a destra) e *provider* come Google e Amazon sono già in grado di offrire servizi di calcolo a qualsiasi utente, dal singolo cittadino alle piccole e grandi imprese.
Il termine "Grid" deriva dalle "power grid", le reti di distribuzione dell'energia elettrica, il cui sviluppo ha rappresentato la vera rivoluzione tecnologica avvenuta all'inizio del secolo scorso. Non abbiamo un generatore elettrico nelle nostre case, ma l'elettricità è fornita da impianti di distribuzione, che ci consentono di disporne a basso costo. Analogamente, grazie alla Grid o al Cloud, la potenza di calcolo a nostra disposizione non è più solo quella del nostro computer da tavolo, ma è distribuita su numerosissimi computer in diverse parti del mondo. Per questo si parla di "calcolo distribuito" o di "griglia" di calcolatori. La vera rivoluzione in questo campo non è stata lo sviluppo di macchine sempre più sofisticate, seppur indispensabili, quanto la realizzazione dell'infrastruttura software: il cosiddetto *middleware* (che costituisce il "ponte" fra l'hardware e il software), che ha consentito l'accesso agevole, affidabile e trasparente alle risorse di calcolo.
Il primo sistema di *middleware* per Grid è stato sviluppato negli Stati Uniti a partire dal 1998. Nello stesso periodo in Europa

b.
Attraverso la Grid (a sinistra), costituita dalla connessione – organizzata secondo precisi protocolli – di centinaia di migliaia di risorse di storage dati e di calcolo sparse su tutto il globo, si può elaborare in poche ore il lavoro di anni di un singolo computer (Grid computing). Il Cloud computing (a destra) è sinonimo di una riserva di risorse informatiche: nodi di calcolo e storage (IaaS, Infrastructure as a Service), software (SaaS, Software as a Service) e piattaforme di sviluppo (PaaS, Platform as a Service) che possono essere utilizzate dagli utenti di una rete in modalità "virtuale". Gli oggetti fisici sono invisibili agli utenti dei servizi Cloud e gestiti solo dal fornitore del servizio.

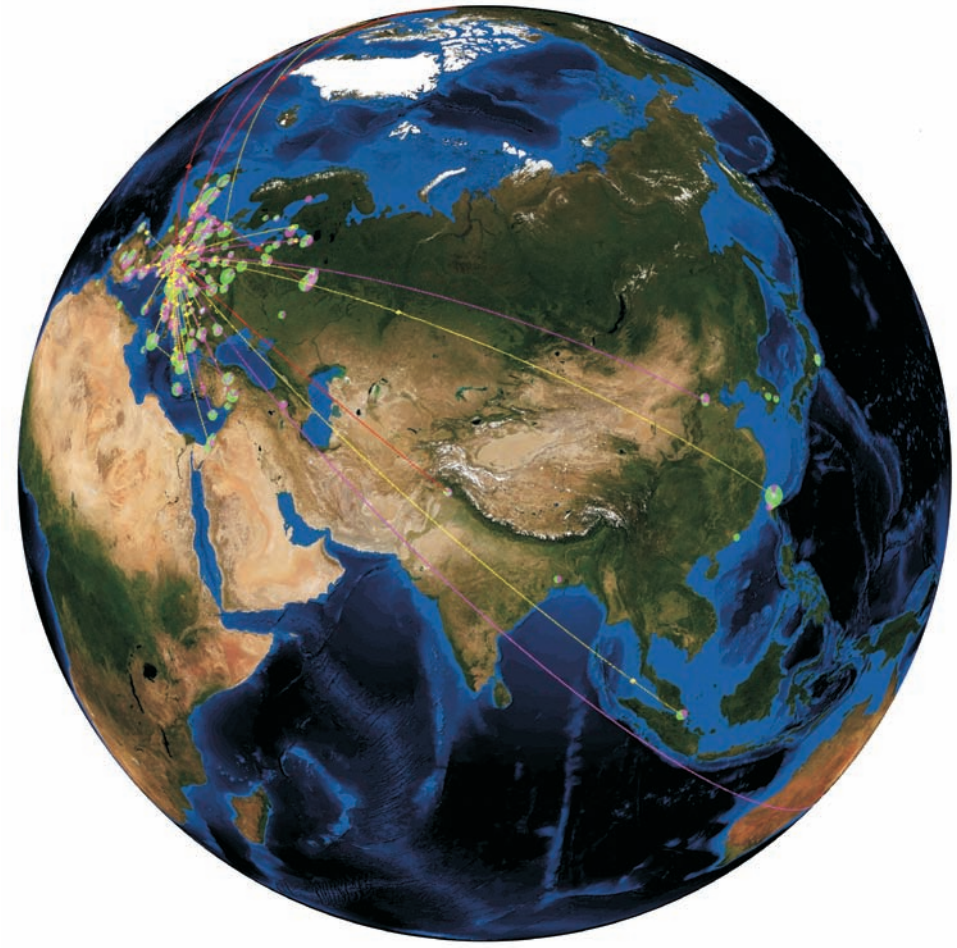
[as] approfondimento

Volontariato informatico

La Grid non è solo appannaggio del mondo della ricerca, delle imprese e delle amministrazioni. Accanto a queste applicazioni professionali per la scienza, il business e la società, si vanno sempre più diffondendo esperienze di cosiddetto "calcolo distribuito volontario", nate dalla semplice constatazione che nel mondo esistono milioni (o forse miliardi) di nodi di calcolo (tipicamente pc, mac, sistemi linux) non utilizzati o scarsamente utilizzati, soprattutto a livello privato. A chi di noi non è capitato di lasciare il proprio computer inattivo per ore (o forse per giornate intere)? E nella maggior parte dei casi, anche quando attivo, il nostro computer non utilizza al 100% le risorse del processore: è sufficiente attivare il sistema di monitoraggio delle risorse del sistema per rendersene conto. Ecco allora che, in maniera del tutto volontaria, milioni di utenti privati (o anche pubblici) offrono la potenza inutilizzata dei propri computer, o sistemi di calcolo più o meno sofisticati, negli intervalli di tempo in cui questi sono inattivi, diventando così dei *citizen scientist* (vd. anche p. 32, ndr). Così sono nate iniziative a livello mondiale di cooperazione volontaria e (cosa fondamentale) anonima, a disposizione di progetti di ricerca accademica (matematica, fisica, biologia, chimica, medicina, climatologia e tanto altro), ma anche di settori estranei a essa: Boinc, Gims, distributed.net sono tra le principali piattaforme per il calcolo distribuito volontario.
In ambito accademico l'esperienza più interessante è gestita da Boinc (Berkeley Open Infrastructure for Network Computing) che è un sistema non-commerciale supportato dalla National Science Foundation Usa e che offre una piattaforma a scelta dell'utente per la partecipazione volontaria a offrire calcolo per progetti scientifici. Conta quasi 300.000 utenti volontari che mettono a disposizione quasi un milione di nodi di calcolo. I progetti di maggior successo sono: PrimeGrid, Distributed Rainbow, Galaxy Zoo, MilkyWay@home, Collatz Conjecture, ma è anche possibile partecipare alle ricerche sul bosone di Higgs con il progetto LHC@home!

1.
Uno dei progetti di calcolo distribuito volontario della piattaforma Boinc è *Galaxy Zoo*, in cui si aiuta gli astronomi a classificare le galassie fotografate dal telescopio Hubble, come questa nella foto (M82).

cominciarono diversi progetti finalizzati all'impiego di tecnologie Grid per la gestione dell'enorme quantità di dati provenienti dagli esperimenti di fisica delle particelle, cioè per le esigenze del Cern. Tra i partner del Cern, l'Italia – con l'Infn – è leader internazionale riconosciuto nello sviluppo e gestione di infrastrutture di calcolo distribuito di tipo Grid ed è tutt'oggi impegnato ai massimi livelli all'integrazione delle due "filosofie" per il calcolo, Grid e Cloud.
Al Cnaf di Bologna è in fase avanzata lo sviluppo del software Wnod (Worker Nodes on Demand), che vede l'Infn impegnato nello sviluppo di protocolli per la fornitura integrata di servizi e risorse di calcolo Grid e Cloud. Tra gli obiettivi primari vi è la realizzazione di un'interfaccia di accesso aperta, di tipo Cloud, alle risorse di calcolo Grid, attraverso macchine virtuali a richiesta e ulteriori risorse reperibili da altri fornitori di servizi Cloud (pubblici o privati).
L'Italia svolge inoltre un ruolo trainante nello sviluppo della European Grid Initiative (Egi) al servizio della European Research Area (Era). Egi include più di 350 centri calcolo a livello mondiale e un'offerta per il mondo della ricerca oggi pari a circa 350.000 nodi di calcolo e una capacità di archiviazione superiore a 100 petabyte (100 milioni di gigabyte, equivalente ai dati contenuti in 25 milioni dei comuni dvd). Il Sud dell'Italia è particolarmente attivo nella realizzazione di un'infrastruttura di calcolo sovra-regionale pienamente inserita nella Grid nazionale e internazionale; il progetto Recas (Rete di Calcolo per SuperB e altre applicazioni), attualmente in corso, costituisce un eccellente esempio di sinergia fra Infn e università al fine di realizzare centri di calcolo al Sud (Napoli, Catania, Bari, Cosenza) per soddisfare le esigenze di calcolo e di analisi dati di fisica al futuro acceleratore di particelle SuperB e per offrire servizi di Grid/Cloud computing al territorio.



c.
Il sito Real Time Monitor (http://rtm.hep.ph.ic.ac.uk/) permette di visualizzare in tempo reale l'attività dei nodi Grid nel mondo (cerchi) e lo scambio di informazioni tra questi (linee). La maggior parte dei nodi e delle connessioni sono utilizzati dalla WLcg (Large Hadron Collider Computer Grid) per l'analisi dati degli esperimenti del Cern.

Il futuro della Grid, sempre più strettamente connesso con le opportunità offerte dall'evoluzione delle reti informatiche e della presenza di potenti e ramificate infrastrutture di calcolo, è, per così dire, già presente. Da un lato il calcolo intensivo e distribuito potrà soddisfare le sempre maggiori richieste da parte del mondo accademico-scientifico-tecnologico e delle imprese; dall'altro si è ormai fatta strada da tempo la consapevolezza di coniugare i bisogni tradizionali di calcolo con offerte di servizi a richiesta da parte del cittadino, della Pubblica Amministrazione, della società in tutti quei settori (ad es. Sanità, Beni culturali, Scuola, Trasporti), in cui occorre accedere in modo trasparente e facile per l'utente a "nuvole" e a servizi informativi in tempo reale o quando se ne sente la necessità. In tal senso, il Grid computing sta evolvendo sempre di più verso il Grid/Cloud computing. Grid e Cloud rappresentano quindi infrastrutture digitali di primissima importanza per lo sviluppo della scienza collaborativa e applicativa in questo particolare momento storico.

Biografia

Leonardo Merola è professore di fisica sperimentale presso l'Università di Napoli Federico II. È stato direttore della sezione di Napoli dell'Infn dal 2004 al 2011. Ha lavorato al Cern e ai Laboratori Nazionali di Frascati ed è attualmente impegnato in esperimenti presso l'acceleratore Lhc al Cern. È inoltre responsabile del progetto Recas (Rete di calcolo per SuperB e altre applicazioni).

Link sul web

http://www.italiangrid.it/

http://www.egi.eu/

http://lcg.web.cern.ch/lcg/

http://www.boincitaly.org/

http://boinc.berkeley.edu/

http://lhcathome.web.cern.ch/LHCathome/Grid/index.shtml

http://www.gridcafe.org/

http://www.earthsystemgrid.org

http://rtm.hep.ph.ic.ac.uk/desc.php

http://web.infn.it/wnodes/index.php/cloudintegration

http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc1.html