» L'INTERVISTA

QUANDO LA RICERCA È BIG DATA E CALCOLO COMPLESSO

Intervista a Antonio Zoccoli, vicepresidente dell'INFN e referente per il settore Calcolo e Reti della giunta esecutiva dell'Ente

Per ottenere risultati epocali come la scoperta del Bosone di Higgs o quella delle onde gravitazionali, ma anche per studiare le proprietà dei raggi cosmici e dei neutrini, la ricerca in fisica di base gestisce ed elabora con calcoli complessi enormi quantità di dati. L'immane quantità di informazioni fornite ad esempio da ogni collisione tra i fasci di particelle dell'acceleratore LHC al CERN ha portato i fisici a ideare e sviluppare un'infrastruttura ad hoc per la selezione, l'archiviazione e l'analisi dei dati. Un'infrastruttura in continua evoluzione, che abbraccia tutto il pianeta integrando in un sistema complesso e organizzato risorse di calcolo diverse sia per collocazione geografica sia per capacità. Una rete planetaria, la GRID, che sfrutta contemporaneamente la memoria e la capacità di decine di migliaia di computer distribuiti sul Pianeta. Il risultato è una potenza di calcolo pari a quella di 100.000 computer.

A quali esigenze fondamentali risponde l'attività di calcolo scientifico dell'INFN?
L'INFN è impegnato nel calcolo scientifico da quando si è reso necessario analizzare i dati degli esperimenti nei quali eravamo coinvolti quindi, in sostanza, da quando ha avuto inizio l'attività sperimentale dell'Ente. Da subito, poi, l'esigenza non è stata solo quella di condurre campagne di analisi dati, ma anche di disporre di risorse di calcolo adeguate a svolgere simulazioni di tipo Montecarlo, che sono una risorsa fondamentale per la ricerca scientifica. Tuttavia, benché fosse una componente importante dell'attività di ricerca, fino a 20 o 30 anni fa il calcolo scientifico era ritenuto un aspetto secondario nella progettazione e nella conduzione di un esperimento. Nella gran parte dei casi, l'esperimento era progettato indipendentemente dalle sue esigenze di calcolo: si affinavano gli strumenti di analisi e si predisponevano le infrastrutture necessarie solo in un secondo momento e sulla base di necessità contingenti. L'approccio è cambiato decisamente negli ultimi anni a causa dell'enorme quantità di dati prodotti da LHC:

» L'INTERVISTA

l'infrastruttura di calcolo è divenuta una parte fondamentale degli esperimenti, al pari dei rivelatori e di tutta la strumentazione scientifica. Con l'avvio di LHC siamo stati costretti a un vero e proprio cambiamento di paradigma: oggi sarebbe impensabile progettare un esperimento senza sapere prima quale sarà la mole di dati da maneggiare e senza avere previsto la procedura e l'infrastruttura adatte ad analizzarla. Una sfida di questo tipo sarà da affrontare già nei prossimi anni poiché l'*upgrade* di LHC e il successivo progetto HI-LUMI LHC sono progetti nuovi la cui aspettativa di dati è di 10 volte superiore a quella prodotta fino a oggi da LHC.

LHC è certamente un motore imprescindibile nello sviluppo di risorse di calcolo per la fisica delle alte energie. Quali sono le esigenze specifiche della ricerca in questo campo e quali le soluzioni adottate?
LHC ha rappresentato la svolta. Prima della sua progettazione, infatti, gli esperimenti potevano contare solo su un'infrastruttura di calcolo molto localizzata. Con l'avvio di LHC, al contrario, si sono perseguiti fin dall'inizio due obiettivi. Il primo: rendere disponibile una capacità di calcolo adatta ad analizzare una mole di dati senza precedenti. Il secondo: permettere a tutti gli scienziati che partecipavano alle collaborazioni sperimentali, che erano dislocati un po' in tutto il mondo, di poter accedere ai dati dalle loro rispettive istituzioni di appartenenza. Si chiedeva in sostanza che l'infrastruttura fosse accessibile da qualunque luogo. La soluzione è stata ottenuta con lo sviluppo della GRID, un'infrastruttura di calcolo globale che abbraccia letteralmente l'intero Pianeta. In inglese il termine GRID indica la rete elettrica e riferita al calcolo scientifico ha esattamente la stessa valenza. Quando si inserisce la spina nella presa elettrica per far funzionare un qualunque elettrodomestico non ci si pone certo il problema di dover costruire una centrale elettrica che fornisca l'elettricità che si sta prelevando. Così per il calcolo: la GRID consente di rinunciare a chiedersi dove sia localizzata la risorsa di calcolo e chi e come la stia mettendo a disposizione. Una rete di centri di calcolo collegati con fibre ottiche molto veloci, e supportati da un'interfaccia che consenta a qualunque utente di accedere, smette di essere una somma di singole risorse e diventa un insieme. L'approccio è del tutto innovativo e la GRID rappresenta in questo senso un caso primo e unico al mondo.

Quale il contributo dell'INFN?
La sfida tecnologica è stata affrontata a livello globale e il contributo dell'INFN è stato consistente almeno quanto il suo coinvolgimento negli esperimenti di LHC. I termini della sfida erano dettati dalla necessità di offrire alla comunità scientifica delle alte energie la possibilità di accedere alle risorse disponibili e di trasportare una grande quantità di dati in pochissimo tempo. Con tutti i problemi connessi a una condivisione tanto massiccia di informazioni, come l'autenticazione e la salvaguardia dei dati.
In definitiva siamo stati in grado di sviluppare l'hardware e il software necessari con un

» L'INTERVISTA

contributo importantissimo da parte dell'INFN in termini di man power. La collaborazione WLCG (*Worlwide LHC Computing Grid*) composta da più di 170 centri di calcolo distribuiti in 42 Paesi è oggi impegnata nel rendere disponibili per l'analisi i 50 Petabyte di dati prodotti da LHC nel solo 2016: una quantità che non ha pari in altre discipline e che definisce, di fatto, il termine Big Data, indicativo non solo dell'immane quantità di dati, ma anche della sua variabilità e delle velocità e agilità nella trasmissione e condivisione delle informazioni. Contribuendo con i propri ricercatori e con le peculiari competenze all'implementazione dell'infrastruttura, l'INFN è stato uno dei principali attori di questo processo, che ha conosciuto il massimo sforzo nel decennio 2000-2010. Dal punto di vista dei progressi scientifici questa rivoluzione ha mostrato immediatamente i suoi effetti. Per la prima volta nella storia dei grandi esperimenti è stato possibile ottenere i risultati scientifici nei pochi mesi successivi la presa dati. E la prima prova tangibile è rappresentata proprio dalla scoperta del Bosone di Higgs.
A livello nazionale, il processo ha consentito di realizzare in Italia un'infrastruttura di calcolo distribuita il cui centro principale, un nodo di primo livello (Tier1), è collegato ad altri 10 centri di secondo livello (Tier2) distribuiti sul territorio. Il Tier1 è gestito dal centro di calcolo nazionale dell'INFN, il CNAF di Bologna, ed è il principale cluster di calcolo, oltre che per gli esperimenti di LHC, per molti degli esperimenti che coinvolgono l'INFN. Questa rete nazionale a due livelli è poi inserita nell'infrastruttura mondiale e collegata al centro principale al CERN (Tier0) e agli altri centri di livello 1 e 2 nel mondo, distribuiti in Europa, Asia, Giappone e USA.

Oltre a LHC la GRID supporta esperimenti e collaborazioni sperimentali nella condivisione e nell'analisi dei dati. Quali esperimenti ne traggono maggiore beneficio?
Inizialmente l'attività di calcolo GRID è stata asservita unicamente alle esigenze di analisi dati di LHC, per le quali era stata sviluppata. Successivamente, tuttavia, hanno adottato lo stesso approccio anche altre collaborazioni sperimentali che avevano l'esigenza di gestire grandi quantità di dati, come Belle II, in Giappone e BES III, in Cina. Attualmente, l'utilizzo della GRID si sta allargando anche a esperimenti di settori diversi dalla fisica degli acceleratori, come le grandi collaborazioni internazionali di fisica delle astroparticelle. Penso ad esempio ai progetti in fase di sviluppo come il *Cherenkov Telescope Array (CTA)* o il progetto satellitare Euclid, che vedono il coinvolgimento in primis dell'INAF (Istituto Nazionale di Astrofisica) e la partecipazione attiva dell'INFN. E penso a Xenon, per la ricerca della materia oscura ai Laboratori INFN del Gran Sasso, che avendo moltiplicato notevolmente la quantità di dati da analizzare rispetto agli standard iniziali ha chiesto di accedere alla rete dei Tier1 e Tier2 di LHC. Si tratta di esperimenti in forte sviluppo gestiti da collaborazioni internazionali sempre più allargate. Benché sempre inferiori a quelle di LHC, le quantità di dati che questi esperimenti raccoglieranno nei prossimi anni rendono indispensabile l'accesso alla GRID.

» L'INTERVISTA

Quali le prospettive a livello europeo e mondiale?

Per quanto ci riguarda la prossima sfida ha una doppia valenza. Innanzitutto, l'infrastruttura stessa deve evolvere verso un modello di organizzazione diverso da quello attuale. Non potrà più valere il paradigma GRID basato su tante CPU collegate tra loro. Per la sola analisi dei dati dei prossimi upgrade di LHC avremo bisogno di un'infrastruttura più larga che coinvolga Tier 1 e Tier2 ma anche macchine in grado di fare *High Performance Computing*. L'infrastruttura dovrà essere di carattere più generale e permettere a qualsiasi utente di qualunque settore della ricerca di accedere alla tipologia di risorsa a lui più congeniale. Se funzionale all'analisi dati che si sta facendo, l'infrastruttura deve essere in grado di offrire anche risorse di calcolo di livello avanzato. Anche per questo sarà necessario abbandonare l'approccio GRID in favore di una logica CLOUD, che è più flessibile, aperta a esigenze variegate, e accessibile da utenti esterni alla comunità INFN. In particolare, devono poter accedere all'infrastruttura anche comunità sperimentali che lavorano su tematiche di ricerca diverse dalle nostre. Penso innanzitutto all'INAF con cui abbiamo collaborazioni sperimentali in corso, ma anche all'ASI (Agenzia Spaziale Italiana) e al settore Long Baseline Science, il cui cuore pulsante è oggi in Cina.

La premessa a questo processo è che l'evoluzione della GRID va affrontata in ogni caso per il futuro di LHC. La scelta che abbiamo dovuto fare da subito è se valesse la pena di rendere l'infrastruttura più flessibile e metterla a sistema, in modo da renderla accessibile alle comunità di ricerca italiane al di fuori dell'INFN. Noi abbiamo le competenze necessarie a compiere il passo e, a fronte di un supporto da parte delle istituzioni, vale certamente la pena di evitare di disperdere sforzi e risorse frazionando gli interessi delle singole comunità scientifiche. È esattamente quanto accaduto in precedenza con la GRID che è nata per noi ed è ora patrimonio di tutti.■