

» INTERVISTA



IL CNAF E LE NUOVE SFIDE INFORMATICHE NELL'ERA DEI BIG DATA E DEL SUPERCOMPUTING

Intervista con Gaetano Maron, direttore del CNAF, il Centro Nazionale di Calcolo dell'INFN

Il CNAF, il Centro Nazionale per le tecnologie informatiche e telematiche dell'INFN, ha ripreso da poco la sua piena funzionalità, dopo le operazioni di ripristino che si sono rese necessarie a seguito di un grave allagamento, che lo scorso novembre ha colpito la sua sede a Bologna. Il CNAF ospita il centro TIER1 per gli esperimenti di LHC ed è punto di riferimento anche per molti altri esperimenti in cui l'INFN è impegnato. Si occupa di ricerca e sviluppo di tecnologie digitali innovative per applicazioni in diverse discipline scientifiche, ed è uno dei più importanti centri di calcolo distribuito in Italia. Dei progetti, presenti e futuri, del CNAF, nell'era dei big data e del supercomputing, abbiamo parlato con il suo direttore Gaetano Maron.

Da poche settimane un link ultra veloce in fibra dedicata connette il centro LHC-Tier1 del CNAF e il CINECA.

L'INFN ha firmato lo scorso ottobre un accordo di collaborazione con il CINECA che prevede l'utilizzo in esclusiva di una parte delle risorse di calcolo del CINECA da parte dei nostri ricercatori, sia teorici sia sperimentali. Questo accordo ha interessato direttamente il CNAF che, mantenendo nel proprio centro di calcolo LHC-Tier1 una parte importante dei dati, sia degli esperimenti di LHC, sia degli esperimenti più significativi di fisica astroparticellare e fisica nucleare in cui l'INFN è coinvolto, ha interesse ad avere sufficiente potenza di calcolo per elaborarli. L'accordo mette a disposizione del CNAF circa 15.000 CPU core (corrispondenti a circa il 50% della potenza di calcolo ora presente al LHC-Tier1 del CNAF), che contribuiscono ad assicurare una capacità di elaborazione adeguata per almeno i prossimi due anni. Per garantire l'accesso ai dati presenti al LHC-Tier1 da parte dei computer disponibili al CINECA, situato a circa 10 km di distanza dal CNAF, abbiamo configurato in collaborazione con il GARR un link ultra veloce

» INTERVISTA

in fibra dedicata con capacità fino a 1,2 Terabit per secondo. La distanza è tale che la latenza, cioè il tempo che impiega un bit d'informazione per viaggiare da un centro all'altro, è minima, tanto da poter considerare i due centri l'uno l'estensione dell'altro. Questa possibilità di espansione è stata sfruttata durante le operazioni di ripristino del data center del CNAF, dopo l'allagamento dello scorso novembre quando, a causa della rottura di una tubatura dell'acquedotto bolognese, più di 500 metri cubi d'acqua si sono riversati nelle sale del LHC-Tier1. Il centro ha subito danni ingenti, sia alla parte infrastrutturale elettrica, sia alle risorse di calcolo che sono state affette dall'allagamento per circa il 10%. Il LHC-Tier1 è stato bloccato per circa tre mesi in modo totale nelle sue attività, che sono riprese con gradualità a partire da febbraio, per poi ritornare a pieno regime nel mese di marzo. Appena ristabilite le funzionalità minime del centro, grazie al link col CINECA è stato possibile mettere a disposizione degli utenti del CNAF una significativa potenza di calcolo, che ha mitigato la carenza di risorse e il disagio dovuto alle opere di recupero delle piene funzionalità del centro.

Il CNAF si chiama così perché è l'acronimo del nome che aveva quando è stato istituito, Centro Nazionale Analisi Fotogrammi. Il CNAF ha dunque una lunga storia, è stato il primo Centro Nazionale dell'INFN.

Se c'è un denominatore comune nella lunga storia del CNAF che connette le diverse attività che ha svolto nel corso degli anni, esso è la rivoluzione informatica. Il CNAF nasce, infatti, nel 1962 come Centro Nazionale Analisi Fotogrammi, con lo scopo di fornire un servizio centrale di digitalizzazione e analisi dei fotogrammi provenienti dagli esperimenti che utilizzavano rivelatori a camere a bolle: in sostanza, un raffinato e innovativo scanner ante litteram, con associato un sofisticato sistema di analisi delle tracce digitalizzate. Con l'evoluzione delle tecniche di rivelazione delle particelle prodotte negli esperimenti di alta energia, il ruolo del CNAF come centro di analisi fotogrammi viene meno, e negli anni '80 lo scopo del centro viene spostato, sempre sull'onda della rivoluzione informatica, su un altro paradigma fondamentale per l'INFN e per la ricerca italiana: le reti di comunicazione. Il CNAF, fino alle soglie degli anni 2000, prima sviluppa e poi gestisce la rete geografica dell'INFN che, oltre a collegare tutte le Sezioni e i Laboratori Nazionali, diventa presto un riferimento per tutti gli enti di ricerca, tanto da evolversi gradualmente nella rete della ricerca italiana, ora operata dal GARR.

Con gli anni 2000 il CNAF compie poi la sua più recente trasformazione, resa possibile dalla sua grande esperienza nello sviluppare e integrare sistemi ad ausilio della fisica sperimentale, ma anche e soprattutto dalla sua volontà di interpretare il meglio della ricerca informatica. Al CNAF, infatti, viene dato l'incarico di progettare e gestire il centro di calcolo LHC-Tier1 da dedicare in primis agli esperimenti di LHC, ma presto divenuto riferimento del calcolo scientifico per la maggior parte degli esperimenti cui partecipa l'INFN.

» INTERVISTA

Allo stesso tempo, conscio che un centro di calcolo senza ricerca e sviluppo non ha molta prospettiva, il CNAF fa partire numerose attività innovative nel campo dei sistemi geograficamente distribuiti, di primaria importanza per il calcolo di LHC, divenendo in breve uno dei principali protagonisti dello sviluppo di World wide LHC Computing Grid (WLCG) e della infrastruttura grid italiana, tuttora componenti fondamentali del calcolo LHC.

Oggi conosciamo il CNAF anche perché è un centro LHC-Tier1 ma, come accennato prima, il CNAF gestisce anche i dati di molti altri esperimenti internazionali.

Sì, il CNAF non è solo un centro LHC-Tier1, ma è il centro di calcolo dell'INFN dedicato a tutti i principali esperimenti in cui l'INFN è coinvolto. Il calcolo di LHC occupa circa il 75% delle risorse computazionali e di storage del centro, mentre il restante 25% è dedicato ad altri esperimenti di fisica delle alte energie e a quelli di fisica delle astroparticelle. Al momento usano le risorse di calcolo del nostro Tier1 più di 30 esperimenti non LHC, che vanno dagli interferometri per la ricerca di onde gravitazionali LIGO e VIRGO, ai rivelatori nello spazio, come AMS-02 sulla Stazione Spaziale Internazionale per studiare l'universo e le sue origini, a esperimenti nei Laboratori sotterranei del Gran Sasso dell'INFN, come DarkSide per la ricerca di materia oscura. Attualmente la quantità di dati prodotti da tutti questi esperimenti e immagazzinati nei dischi e nastri magnetici del CNAF è più di 60 PetaByte: certamente si tratta di big data, nell'attuale accezione del termine. L'analisi di questi dati avviene principalmente dai sistemi di elaborazione interni al Tier1, da quelli provenienti dagli altri Tier1 della collaborazione WLCG che sono connessi al nostro centro con reti di comunicazione dedicate, ma, come abbiamo visto, anche da centri di calcolo esterni ma integrati nella nostra rete locale: come il CINECA, appunto, il centro RECAS di Bari e cloud provider commerciali. Interessante da questo ultimo punto di vista il pilot che stiamo conducendo per il progetto europeo H2020 chiamato Helix Nebula Science Cloud (HNSciCloud), dove due provider indipendenti mettono a disposizione degli esperimenti WLCG una non trascurabile potenza di calcolo.

Ma non è solo la mole di questi dati che deve attirare l'attenzione: essi costituiscono soprattutto un bene di valore inestimabile e a volte unico, non solo per gli esperimenti che li hanno prodotti ma per l'INFN e, in generale, per la ricerca. Sono quindi beni preziosi che devono essere preservati nel tempo e garantiti a resistere ai cambi tecnologici, al fine di potervi accedere anche negli anni a venire. Per questo motivo, al CNAF abbiamo, in collaborazione con gli altri centri, un programma di Long Time Data Preservation che sviluppa tecniche e strumenti per garantire l'accessibilità nel tempo dei dati conservati nel centro.

Di che cosa altro si occupa oggi il CNAF?

Il CNAF poggia su quattro attività principali che costituiscono anche i pilastri strategici della sua missione:

» INTERVISTA

calcolo scientifico dedicato al supporto delle attività di ricerca dell'INFN; innovazione e sviluppo; servizi IT di interesse generale per l'INFN, includendo tutti i servizi amministrativi; trasferimento tecnologico ai settori pubblici e privati.

Un centro come il CNAF esaurirebbe velocemente la sua spinta, se vicino alla gestione del calcolo scientifico degli esperimenti non ci fosse una importante attività di innovazione e sviluppo, che aprisse e sperimentasse strade e strumenti nuovi da riproporre poi nella gestione quotidiana del centro, ma anche come trasferimento tecnologico al mondo industriale, alla società ecc.. Circa il 30% delle risorse umane del CNAF sono dedicate a queste attività di innovazione, che sono focalizzate su due filoni: lo sviluppo di nuovi sistemi geograficamente distribuiti inseriti nella strategia europea per una Open Science Cloud (per esempio, il progetto H2020 INDIGO Data Cloud coordinato dall'INFN, ma anche i più recenti Extreme Data Cloud, European Open Science Cloud Hub, Deep Hybrid Data Cloud ecc.), continuando la tradizione decennale di sviluppo e integrazione iniziata con i progetti grid; lo sviluppo software a supporto di progetti esterni al centro, di quelli interni, ma anche degli esperimenti dell'INFN che richiedono consulenza per il proprio software online o offline (per esempio gli esperimenti LHC, ma anche KM3Net, Euclid ecc.).

A fianco di questi due pilastri strategici, calcolo scientifico e innovazione e sviluppo, si è affiancata da tempo, e negli ultimi anni con notevole rilevanza, una promettente attività di trasferimento dei nostri sviluppi scientifici e tecnologici verso la società, la pubblica amministrazione, ma anche verso il nostro sistema produttivo manifatturiero e terziario, sia regionale, sia nazionale, e anche europeo. Per ottimizzare, ma soprattutto per rendere più eterogenee le competenze offerte da questa attività abbiamo, con le Sezioni di Bologna e di Ferrara dell'INFN, costituito un laboratorio di trasferimento tecnologico denominato TTLab. L'obiettivo è promuovere il trasferimento di tecnologia in fisica, informatica, meccanica ed elettronica alle imprese localizzate in regione. TTLab è accreditato in Regione Emilia Romagna come laboratorio di ricerca industriale e come tale può partecipare, insieme a imprese regionali, a iniziative di finanziamento regionali che mirano a favorire questo processo di scambio e integrazione tra industria ed enti di ricerca (per esempio progetti POR-FESR). Va infine notato come il CNAF partecipi anche ai servizi informatici mirati alla gestione amministrativa dell'INFN (contabilità, gestionale, documentale ecc.), e a un insieme vasto di servizi di pubblica utilità che permettono al nostro ente di operare al meglio (supporto all'attività di ricerca, siti web, calendario, agenda ecc.).

Quali sono le prossime sfide del CNAF, tra supercalcolo e big data?

La gestione dei dati, come abbiamo visto, gioca un ruolo centrale nelle attività del nostro centro perché i dati degli esperimenti devono essere efficacemente accessibili ai ricercatori di tutto il mondo, ma allo stesso tempo rappresentano anche il bene prezioso da custodire e da preservare nel tempo. La proprietà

» INTERVISTA

dei dati deve quindi rimanere all'interno della nostra comunità scientifica, che deve provvedere alla loro gestione.

Con il futuro potenziamento di LHC (il progetto High Luminosity LHC, ndr) questo ruolo sarà ancora più marcato perché la mole di dati da gestire sarà di gran lunga maggiore di quella attuale: una possibile strategia potrebbe consistere nell'avere pochi data centre distribuiti nel mondo, interconnessi tra loro da reti molto veloci in modo di ridurre la necessità di movimento dei dati con l'obiettivo di ridurre i costi complessivi del sistema. La sfida principale per il CNAF, ma anche per l'INFN, sarà quindi di stare al passo con queste richieste e riuscire a offrire alla comunità WLCG un data center italiano. L'aspetto complementare di questa sfida è la capacità di calcolo e di analisi di questi dati che, per definizione, è più flessibile e può essere fornita all'interno del mondo WLCG, compreso naturalmente il CNAF, da cloud commerciali, ma anche da centri di super calcolo. In questo ultimo caso il problema principale è riuscire a utilizzare questi centri di High Performance Computing (HPC) con i codici di analisi che generalmente usiamo nei nostri Tier1 e Tier2, ma questo dipende fortemente dal tipo di architettura e dal tipo di processori che i centri HPC utilizzano, ed è, quindi, necessario prendere in esame caso per caso. Emblematica risulta, a questo proposito, la connessione CNAF-CINECA, cui abbiamo accennato prima, che anticipa l'obiettivo strategico di mettere a stretto contatto le due comunità. In questo caso le architetture e i processori utilizzati dal supercalcolatore Marconi del CINECA sono sostanzialmente gli stessi di quelli utilizzati nel mondo WLCG. Un binomio questo del supercalcolo e dei big data che, in accordo anche con le strategie europee, in Italia avrà la grande opportunità di essere declinato al meglio entro il 2021, quando i centri di calcolo dell'INFN e del CINECA saranno trasferiti al Tecnopolo di Bologna, in un'area di circa 6.000 metri quadrati, messa a disposizione dalla Regione Emilia Romagna, e che diventerà la nuova sede anche del Data Centre dello European Centre for Medium-range Weather Forecast (ECMWF), che lascerà l'attuale sede di Reading, in Inghilterra: si verrà così a costituire il più grande polo di calcolo scientifico della ricerca italiana.

Una grande opportunità per l'INFN e per il CNAF, una grande sfida nella sfida che, se riusciremo a correre e a vincere, ci permetterà non solo di essere un data center nel "data lake" di High Luminosity LHC, ma anche di attrarre progetti, e quindi anche fondi, di grande respiro nel mondo dei servizi informatici innovativi (big data custodial, analytics, deep and machine learning ecc.) da sviluppare per o con il mondo dell'industria più sensibile all'innovazione, con altre discipline scientifiche favorendo quelli a più diretto impatto sulla vita e la salute dei cittadini. ■