

CRS4: le soluzioni EMC aiutano la ricerca sul Genoma Umano



Il centro di ricerca voluto dalla Regione Sardegna, ha scelto l'infrastruttura storage di EMC per gestire la grande quantità di dati e immagini generata da uno dei principali progetti attualmente in corso riguardante lo studio del Genoma Umano.

Con i 36,4 Teraflops del supercomputer HP, ai quali se ne aggiungono una decina di altri sistemi per un totale di 46 Teraflops (46 mila miliardi di operazioni al secondo), è il secondo sito italiano per potenza di calcolo (preceduto solo dal CINECA) e si colloca nella parte alta della prestigiosa classifica mondiale "TOP 500 Supercomputer Sites". È il Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna (CRS4), un centro interdisciplinare di ricerca applicata che si trova nel Parco Tecnologico della Sardegna a Pula.

Il CRS4 nacque nel 1990 grazie al contributo della Regione Sardegna e di alcuni soci privati con lo scopo di realizzare un centro di competenza e di ricerca che potesse contribuire a dare una nuova immagine alla regione, favorendo lo sviluppo di attività industriali per creare nuove opportunità di lavoro per i giovani. Il primo presidente a cogliere la sfida fu il premio Nobel per la Fisica Professor Carlo Rubbia e, attraverso una politica di gestione impostata su obiettivi concreti di ricerca e poca burocrazia, il centro iniziò velocemente la propria attività reclutando ricercatori affermati da Università e Laboratori italiani ed esteri e dal CERN con il quale continua ad avere stretti rapporti di collaborazione e comunicazione.

Nell'arco di poco tempo presero vita i primi progetti che portarono a risultati di prestigio come il rilascio della versione online de L'Unione Sarda, primo quotidiano su Internet insieme al Washington Post, e alle prime iniziative locali destinate a diventare importanti realtà imprenditoriali come nel caso di Video Online, Tiscali, Energit e altre ancora.

Il CRS4 è ora una struttura molto consolidata nel territorio (socio unico è la Sardegna Ricerche, e una quota rilevante dei finanziamenti derivano direttamente da proventi relativi alla vendita di progetti di ricerca) con un organico

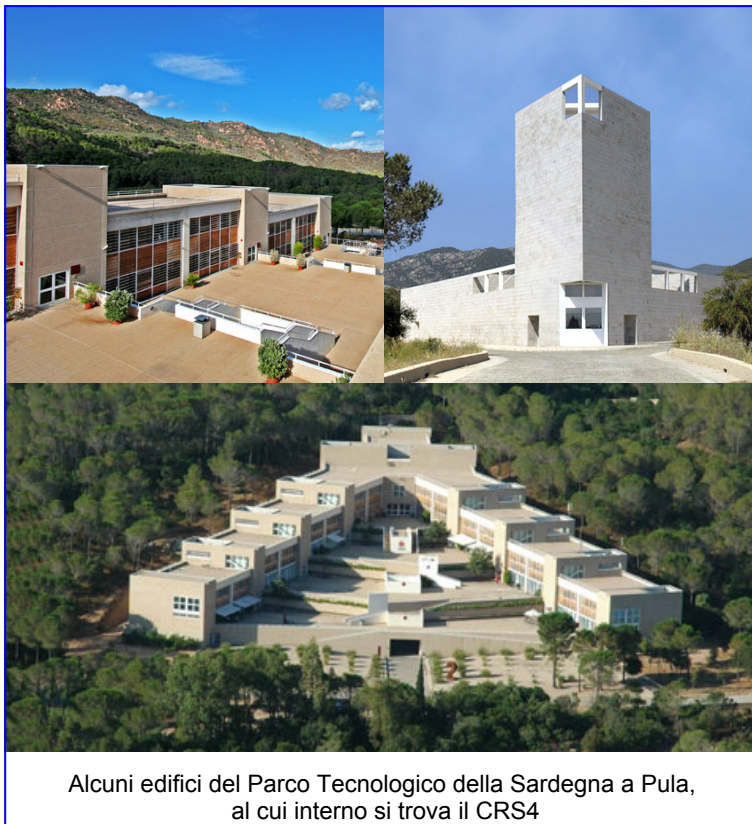
Benefici

- Prestazioni elevate
- Livelli di servizio adeguati alle esigenze degli utenti
- Ottima scalabilità dello storage
- Flessibilità di collegamento
- Flessibilità di crescita
- Possibilità di avere drive differenti nello stesso sistema per uno storage a più livelli

di circa 160 persone, per la maggior parte sardi, e con una importante valenza nazionale e internazionale confermata da collaborazioni con il San Raffaele, importanti industrie petrolifere, l'Università di Oxford, l'EBI di Cambridge e altri e società di rilievo mondiale. Tra i principali settori strategici in cui opera il Centro ci sono l'ICT/Calcolo scientifico, l'Energia, l'Ambiente, la Bioinformatica e Biomedicina.

Il Gruppo ISRC al servizio della Ricerca.

Le numerose attività di ricerca condotte dai gruppi che operano presso il CRS4, richiedono una notevole capacità di calcolo ad altissime prestazioni e un'avanzata infrastruttura IT e di rete per raccogliere, elaborare e gestire enormi quantità di dati. La responsabilità di provvedere a queste esigenze, fornendo servizi informatici di altissimo livello, spetta al "Gruppo di Infrastrutture e Servizi di Calcolo e Reti" (Gruppo ISRC), che svolge l'incarico attraverso una gestione attenta dei sistemi e un costante aggiornamento delle tecnologie, selezionando nel mercato quelle più avanzate e adeguate a supportare i progetti più complessi e impegnativi. Il gruppo è composto da un team di 14 persone che oltre a supportare il personale del CRS4 ha il compito di gestire i servizi per gli utenti del Parco Tecnologico di Pula, occupandosi dell'infrastruttura del campus.



Il Progetto "Genoma Umano"

Il laboratorio di Bioinformatica, fortemente voluto nel 2006 dal Professor Paolo Zanella, attuale presidente del CRS4 e già fondatore dell'European Bioinformatics Institute di Cambridge che è una delle principali realtà mondiali del settore, svolge un ruolo determinante negli studi nell'ambito della genetica, della genomica e delle proteine.

Lo studio del Genoma Umano, la molecola che contiene tutte le informazioni necessarie alla vita, è uno dei principali progetti ai quali il laboratorio sta lavorando per capire meglio l'azione dei geni contenuti nel DNA all'interno del sistema biologico dell'uomo, aiutando così a individuare le cause di alcune importanti malattie.

Il sequenziamento del DNA è il procedimento di base dello studio del Genoma Umano ed è il metodo per ottenere la sequenza nucleotidica completa di un gene o di uno specifico frammento di DNA. Ovvero la successione dei 4

elementi base che compongono le molecole di DNA. Successione che può avvenire secondo infinite combinazioni, differenti per ciascun individuo, dando luogo alle sequenze che costituiscono i geni. L'alterazione di una sola base nucleotidica in alcuni punti di un gene può avere gravi conseguenze ed è la causa di molte malattie genetiche anche gravi.

L'intero processo richiede competenze multi-disciplinari, apparecchiature di analisi molto sofisticate e risorse informatiche altamente performanti e affidabili per elaborare e gestire la notevole massa di dati e immagini in formato jpg generate dalle macchine sequenziatrici installate presso il CRS4, l'unico centro in Italia ad avere due di queste apparecchiature (fornite dalla Illumina, società leader del mercato) dedicate allo studio del Genoma Umano.

Ogni *run* (il sequenziamento del DNA di un singolo individuo) dura circa una settimana ed ha un costo che arriva ad alcune decine di migliaia di Euro.

Centinaia di TeraByte da gestire

La quantità di informazioni prodotta in ogni *run* richiede una capacità di storage di circa 5 TeraByte. Capacità che è in continuo aumento per i software delle macchine sequenziatrici che diventano sempre più sofisticati e producono una massa di dati e immagini sempre maggiore.

Questa enorme quantità di informazioni è oggetto di successive elaborazioni che portano ad ottenere un file di formato alfanumerico, dell'ordine dei Giga-Byte, che rappresenta il Genoma Umano della persona e che viene conservato per lungo tempo, mentre i TeraByte originali vengono cancellati dopo un breve periodo.

“Per supportare questo progetto avevamo bisogno di un’infrastruttura storage dedicata che fornisse adeguate garanzie di performance, funzionalità e scalabilità per far fronte alla quantità crescente di informazioni da gestire. Abbiamo quindi emesso uno specifico bando di gara che è stato aggiudicato alle soluzioni di EMC, avendo meglio di altre soddisfatto i nostri criteri di selezione che, come in tutte le nostre gare riguardanti investimenti in tecnologia, si basano per il 70% sulla valutazione tecnica del prodotto e per il 30% su quella economica”.

Lidia Leoni, Direttore Infrastrutture, Servizi, calcolo e Reti

I sistemi forniti da Sinergy, partner di EMC per le soluzioni storage, sono due **EMC Celerra NS-480**: uno dotato di dischi Fibre Channel per una capacità complessiva di 50 TB e l'altro equipaggiato con dischi SATA per 200 TB totali. Sono sistemi multi-protocollo che supportano contemporaneamente connettività NAS (CIFS, NFS), iSCSI e Fibre Channel. Alle elevate prestazioni abbinano un'ottima flessibilità di configurazione grazie alla possibilità di avere da 2 a 4 blade o controller e di combinare all'interno dello stesso sistema drive di tipo Fibre Channel, SATA e anche allo stato solido, fino ad

un massimo di 480 unità, creando un vero e proprio storage a più livelli in grado di soddisfare livelli di servizio diversi.

Nei sistemi EMC Celerra NS-480 vengono registrate sia i TeraByte delle informazioni e immagini generate dalle macchine sequenziatrici, mantenute in linea per un periodo limitato di tempo, sia i file ottenuti dalle successive elaborazioni per un'archiviazione a lungo termine.

Inizialmente i sistemi erano attestati ad una SAN collegata ad un pool di server dedicati al progetto del Genoma Umano. Più recentemente sono stati collegati alla SAN centrale del CRS4 in modalità Virtual Storage Area Network per essere accessibili anche dal potente supercomputer HP.

“Con i sistemi storage EMC riusciamo a fornire un livello di servizio adeguato alle esigenze della divisione di Bioinformatica” ha continuato Leoni. “Le loro caratteristiche di performance e di scalabilità ci tranquillizzano anche per le future richieste di maggiore capacità, a volte difficili da prevedere. L’aumento consistente e continuo dei dati riferiti al progetto, rende infatti difficoltosa qualsiasi previsione anche nel breve periodo, tanto che nei primi sei mesi dell’anno abbiamo già esaurito la capacità prevista per l’intero 2009”.



Sinergy – Partner EMC

Dal 1994 Sinergy SpA è un System Integrator che realizza infrastrutture ICT a misura di azienda seguendo il cliente fin dalla fase iniziale di assessment. Grazie a un team di persone qualificate e certificate, la società ha maturato una solida esperienza nella progettazione e nell'integrazione dell'infrastruttura - Data Center, Storage, Virtualization, Networking, Security -.

Sinergy è certificata EMC **“Velocity Premier Solution Partner”** e **“Authorized Services Network (ASN)”**.

Sinergy SPA

Sede: Segrate (MI) - Filiali: Bologna, Genova, Padova, Roma, Torino

www.sinergy.it

sinergy@sinergy.it

EMC Italia

Direzione Generale

Via Caldera 21/B2—20153 Milano

tel. 02.40908.1, fax. 02.48204686

numero verde per l'Italia: 800-787.289

www.emc2.it