Grandi infrastrutture di storage per calcolo ad elevato throughput e Cloud

Michele Di Benedetto, Alessandro Cavalli, Luca dell'Agnello, Matteo Favaro, Daniele Gregori, Michele Pezzi, Andrea Prosperini, Pier Paolo Ricci, Elisabetta Ronchieri, Vladimir Sapunenko, Vincenzo Vagnoni, Valerio Venturi, Giovanni Zizzi

INFN-CNAF

Abstract. Gli esperimenti di fisica delle alte energie che lavorano al Large Hadron Collider hanno accumulato in pochi anni un'enorme quantità di dati, sviluppando soluzioni molto avanzate nella gestione di grandi spazi di storage disco e nastro. Il Tier-1 dell'INFN, presso il CNAF di Bologna, è uno dei maggiori centri di calcolo della collaborazione WLCG, con una capacità di oltre 12 PB di disco e 16 PB di nastro. Grazie ad una combinazione di soluzioni industriali e sviluppi specifici, è attualmente uno dei Tier-1 con le migliori prestazioni. L'esempio del Tier-1 INFN può essere di grande importanza per comunità che hanno necessità simili, e che vogliono utilizzare risorse di storage con paradigmi esistenti, come il Grid Computing o emergenti, come il Cloud Computing. In questo articolo presentiamo l'architettura e le scelte tecnologiche adottate, discutendo anche le evoluzioni più recenti verso un sistema di cloud storage per le comunità scientifiche.

1. Introduzione

Negli ultimi anni, le grandi collaborazioni scientifiche in vari campi di ricerca hanno accumulato una quantità di dati mai raggiunta in precedenza, in alcuni casi fino alla scala di svariate decine di PetaBytes (PB) all'anno. È questo per esempio il caso degli esperimenti di fisica delle alte energie che lavorano al *Large Hadron Collider* (LHC) del CERN. L'esperienza fatta in questo settore può essere di grande importanza per altre comunità che hanno necessità simili, specialmente nell'ottica di sfruttare i data center esistenti per offrire risorse di storage (sia con il paradigma del Grid Computing che del Cloud Computing) a numerosi gruppi di ricerca o comunità scientifiche.

Un *Mass Storage System* (MSS) che offra una soluzione *Hierarchical Storage Manager* (HSM), ovvero che comprenda sia risorse online (immediatamente disponibili, come il disco), sia nearline (disponibili con una latenza maggiore, ma anche con un maggiore grado di resilienza dei dati, come il nastro), e che raggiunga la scala delle decine di PB di spazio disponibile, è un sistema complesso composto da molti *layer* hardware e software, dei quali il livello visibile all'utente (ad es. l'interfaccia di cloud storage) è soltanto la punta dell'iceberg. I componenti di un sistema del genere sono sia hardware - dischi e controller, reti *fibre-channel* per *Storage Area Network* (SAN) e *Tape Area Network* (TAN), interfacce di rete a 10 Gbps, *disk server* che le supportino, tape server, ecc. - sia software - *file system*, software di management delle risorse *tape*, *middleware* di trasferimento file e di *storage management*, interfacce utente.

Il Tier-1 dell'INFN ha sviluppato una soluzione generale per MSS altamente scalabile e resistente. È implementato con un sistema modulare composto da standard industriali: *General Parallel File System* (GPFS[1]) e *Tivoli Storage Manager* (TSM[2]), entrambi prodotti IBM, interfacciati tra loro da un middleware sviluppato dall'INFN, GEMSS[3]. Tra le altre cose, il sistema implementa un modello intelligente per portare online i file da nastro, che riduce al mini-

Workshop GARR Calcolo e Storage Distribuito - Selected Papers. Roma 29-30 novembre 2012
ISBN: 978-88-905077-4-8 © 2013 Associazione Consortium GARR

mo le operazioni meccaniche della robotica, come il montaggio, lo smontaggio e la ricerca. L'esperienza, maturata in diversi anni di produzione ha dimostrato l'efficienza e la completezza di questa soluzione.

L'accesso allo storage avviene mediante protocolli standard, in accordo con la specifica *Storage Resource Manager* (SRM[4]), adottata dalle comunità WLCG[5], e più in generale dalle comunità che utilizzano il Grid Computing. SRM è un livello di astrazione che permette agli utenti di accedere allo storage attraverso un'interfaccia comune. Dietro questo tipo d'interfaccia, ogni data center può fare le proprie scelte per ciò che riguarda le componenti hardware e le soluzioni software per implementare il proprio MSS. In questo contributo ci si propone di dare una panoramica completa di come un'infrastruttura di storage di svariati PB funziona ed è gestita in produzione, dagli strati più bassi del livello hardware alle interfacce software di livello superiore. Saranno presentati anche i principali risultati e dati relativi alle prestazioni ottenute nel corso di questi ultimi anni di attività dalle principali collaborazioni scientifiche che lavorano presso il Tier-1 dell'INFN. La progettazione di un'efficiente, robusta e affidabile istallazione Cloud storage di grandi dimensioni, si basa sulla corretta scelta delle tante componenti coinvolte nel sistema, e su come queste lavorano in cooperazione. Esprimiamo, anche con il presente articolo, il desiderio di condividere la nostra esperienza con altre comunità.

2. Il Tier-1 dell'INFN

Il Tier-1 dell'INFN ospitato presso il CNAF - il centro nazionale dell'INFN per la ricerca e lo sviluppo nel campo delle tecnologie informatiche applicate agli esperimenti di fisica nucleare e delle alte energie - è il principale centro di calcolo dell'INFN, e uno dei più grandi in Europa.
Con una superficie di circa 1000 m2 e un impianto ridondato per la distribuzione elettrica (potenza utile di ~5 MVA), può ospitare più di 120 rack e due librerie di nastri, garantendo operatività agli utenti 24 ore su 24 per tutto l'anno. Attualmente ospita una farm di calcolo con una potenza complessiva di ~135 kHS06 ed una capacità di circa 12 PB di storage disco 16 PB di storage nastro.

Operativo dal 2003, il Tier-1 è parte della collaborazione WLCG (World-wide LHC Computing Grid), che fornisce le infrastrutture di calcolo e di storage per i quattro grandi esperimenti LHC (ALICE[6], ATLAS[7], CMS[8] and LHCb[9]). La frazione delle risorse ospitate al CNAF è circa il 13% del totale disponibile presso tutti i Tier-1 WLCG. Il centro è progressivamente diventato il punto di riferimento per il calcolo di molte altre collaborazioni scientifiche, sia esperimenti presso acceleratori (BABAR[10], CDF[11], AGATA, KLOE, LHCf), sia di fisica delle astroparticelle e raggi cosmici (AMS, ARGO, Auger, Borexino, FERMI/GLAST, Gerda, ICARUS, MAGIC, PAMELA, Xenon100, VIRGO).

All'interno della struttura del Tier1 sono anche ospitati il Tier-2 italiano di LHCb (le risorse in questo caso sono completamente condivise con quelle del Tier-1), ed una farm Tier-3 per gli utenti della locale Sezione dell'INFN. La capacità dello storage, sia su disco che su nastro, verrà ulteriormente incrementata durante il 2013 e negli anni successivi, mentre per le risorse di calcolo è previsto un sostanziale consolidamento sui valori attuali con la sostituzione progressiva, nell'arco di due anni, di un numero di server corrispondente a circa metà della potenza attualmente installata. Questo permetterà, oltre ad un ricambio fisiologico delle risorse, anche l'ottimizzazione dei consumi elettrici.

Nelle sale del Tier-1 è ospitato uno dei nodi più importanti della rete GARR[12]: è stato uno dei primi, nel corso del 2012, a migrare alla nuova infrastruttura basata su fibre spente (GARR-X).

Oltre al normale accesso alla rete della ricerca italiana, che assicura il collegamento alle reti della ricerca europee e mondiali attraverso la rete GÉANT, la connettività con il Tier-0 al CERN e con gli altri centri Tier-1 di WLCG è assicurata dalla rete dedicata LHCOPN, alla quale il CNAF accede con un collegamento ridondato a

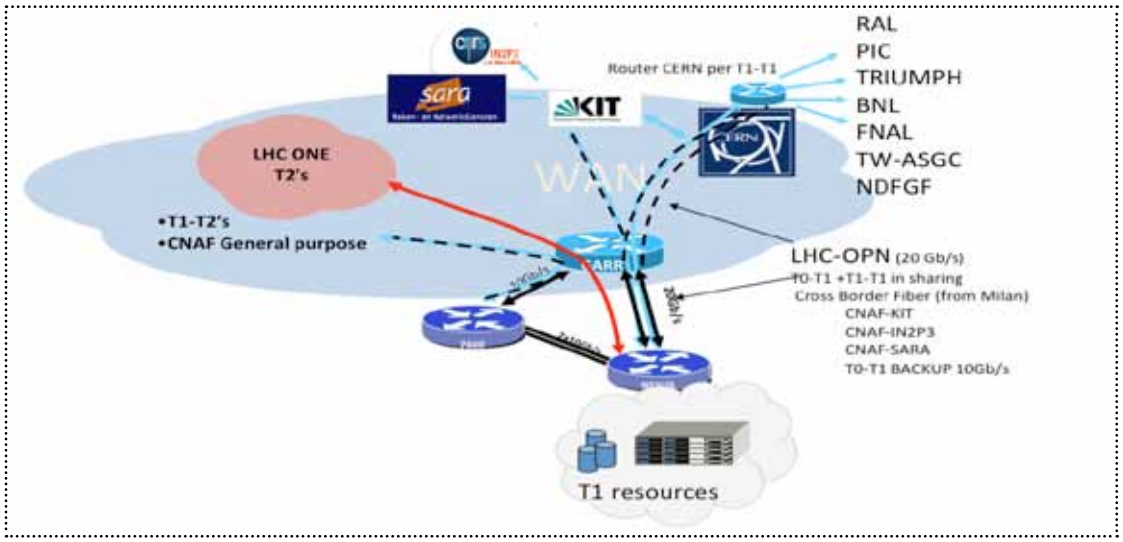

Fig 1 - Collegamenti di rete geografica del Tier-1 dell'INFN

20 Gbps. È inoltre in fase di realizzazione un'ulteriore rete, LHCONE, per l'interconnessione con i principali Tier-2 di WLCG (Fig. 1).

3. La soluzione hardware

Attualmente il CNAF ospita circa 1300 server di calcolo. Le risorse di calcolo vengono allocate dinamicamente ai singoli esperimenti tramite il meccanismo del *fair-share*. La gestione centralizzata permette il pieno utilizzo della farm, che risulta completamente utilizzata per più del 95% del tempo. In Fig. 2 è rappresentata la potenza di calcolo usata negli ultimi 12 mesi: il rapporto tra l'area rossa (tempo effettivo di uso di CPU) e verde (tempo complessivo di impegno delle risorse) è un'indicazione dell'efficienza di uso dell'infrastruttura di storage del centro, in media è piuttosto alta (84% in questo caso)

È bene sottolineare che, essendo la tipologia delle applicazioni assai varia - da programmi di simulazione che sostanzialmente non effettuano accesso allo storage, e quindi sono per definizione ad alta efficienza, a programmi di analisi dati "caotici", per i quali il tempo di accesso allo storage è spesso il fattore dominante - una misura precisa dell'efficienza non può prescindere dalla suddivisione per tipologia di job.

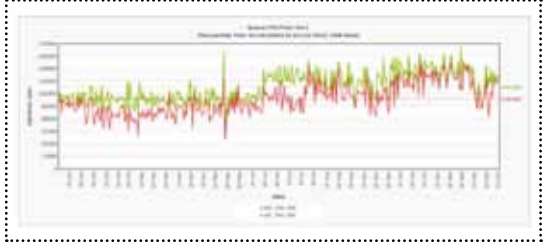

Fig 2 - Uso della farm al Tier-1 dell'INFN

Lo storage al CNAF è organizzato in *file system* GPFS serviti alla farm di calcolo tramite server dedicati, interconnessi alla LAN a 1 Gbps o 10 Gbps, e con 2 link *Fibre Channel* (8 Gbps ciascuno) alla *Storage Area Network* (SAN), cui sono connessi anche i sistemi di disco e (tramite una TAN) i drive della libreria a nastri. L'accesso dai nodi di calcolo ai file system avviene tramite il protocollo file (i file system sono montati sui nodi di calcolo). Le risorse di storage sono anche accessibili da WAN attraverso server GridFTP[13]

I sistemi storage che compongono attualmente la SAN appartengono a più generazioni successive, sia per i collegamenti di rete che per il disco:

- 7 *Data Direct Networks* (DDN) S2A 9950 (con dischi SATA da 2 TB) ed 1 DDN SFA10000 (con dischi SATA da 3 TB), per un totale di ~9 PB serviti da circa 40 disk server con collegamento alla LAN a 10 Gbps;
- 7 EMC2 CX3-80 e 1 EMC2 CX4-960 (con dischi SATA da 1 TB) per un totale di ~2 PB serviti da circa 90 disk server con collegamento alla LAN a 1 Gbps.

Le cassette a nastro sono ospitate su una libreria

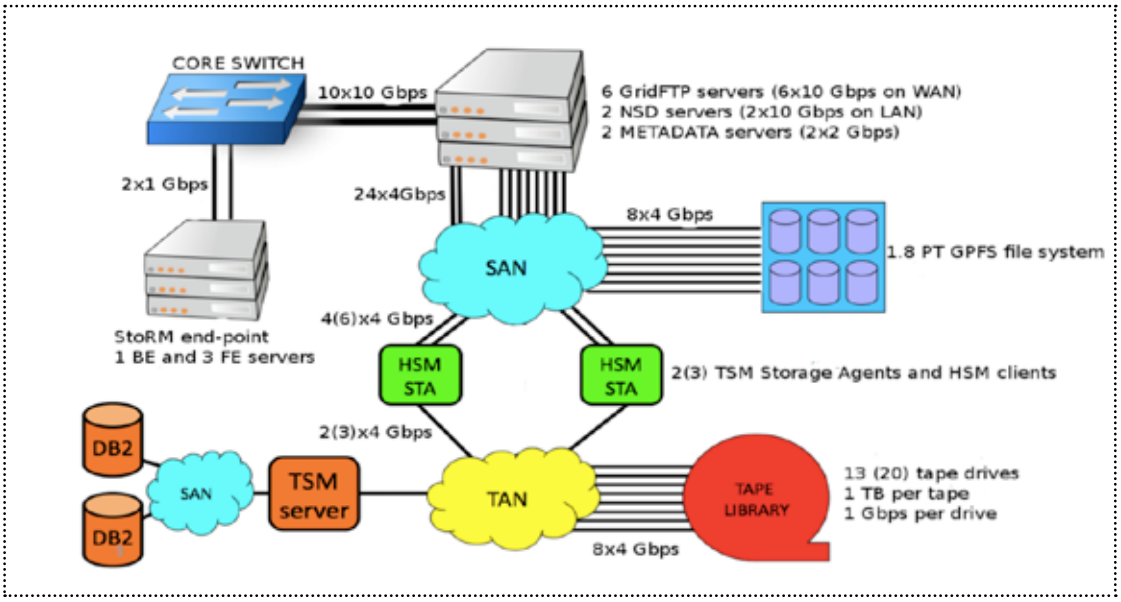

Fig 3 - Schema del sistema di storage per un tipico esperimento del Tier-1 INFN

Oracle SUN SL8500 con 20 drive T10KB (100MB/s di banda passante ciascuno, per ~8500 nastri da 1 TB) e 10 drive T10KC (200MB/s di banda passante ciascuno, per ~1500 nastri da 5 TB). L'interfacciamento fra file system e storage su nastro è realizzato da GEMSS, che implementa un vero e proprio sistema HSM.
In Fig. 3 è illustrato lo schema dello storage per un tipico esperimento al Tier-1.

4. La soluzione software

L'ottimo rendimento del Tier-1 del CNAF è frutto di un'attenta miscela di soluzioni industriali e sviluppi *ad hoc*, che ha consentito di utilizzare la robustezza e l'affidabilità di soluzioni commerciali, nel nostro caso prodotte da IBM, con le interfacce e i paradigmi specifici definiti nella comunità della fisica delle alte energie.

Il file system parallelo GPFS è la soluzione standard adottata dal CNAF per memorizzare i dati su disco. L'utilizzo di file system paralleli consente di avere un namespace unico (a livello di singolo sito, se non utilizzato su WAN) visto da tutti i nodi client come un file system locale. Le varie risorse disco sono aggregate attraverso molti disk server, sui quali i file sono spezzettati in cosiddette stripe di dimensione predefinita. I client possono avere accesso ai dati mediante molti percorsi, garantendo in questo modo la ridondanza e un bilanciamento ottimale del sistema, riducendo la probabilità di avere un collo di bottiglia e aumentando la banda totale di accesso allo storage. I client possono anche accedere simultaneamente in scrittura agli stessi file in modo concorrente, dato che la coerenza globale è garantita dal file system stesso. GPFS è in uso al CNAF da svariati anni. Circa una decina di file system è montata su tutti i nodi di calcolo della farm, per un totale di oltre 10 PB di disco. Un traffico aggregato di circa 10-20 GB/s è raggiunto su base quotidiana.

Il sistema di gestione dei nastri impiegato dal CNAF, il *Tivoli Storage Manager* (TSM), consente un agevole accesso da parte di nodi client a librerie robotizzate, nascondendo tutta la complessità dell'hardware sottostante. TSM è composto da un lato server, che implementa le funzionalità di backup, archiviazione e gestione di uno spazio gerarchico, e da un lato client, dove viene eseguito un piccolo insieme di comandi che consente l'interazione col server e il trasferimento dei dati. Il server memorizza tutte le informazioni su un database DB2, la cui gestione è comunque demandata ai processi server di TSM, ed è completamente nascosta agli amministratori del sistema. Nel nostro caso facciamo u-

so in particolare delle funzionalità HSM del prodotto. I file vengono inizialmente scritti sul file system GPFS, e quindi copiati in background verso i nastri in modo automatico. Il contenuto dei file copiati su nastro viene rimosso da disco dal sistema quando il file system è prossimo al riempimento totale. Nel momento in cui un'applicazione deve accedere ad un file che si trova su nastro ma non più su disco, esistono due diverse modalità di accesso. Una prima modalità consiste nell'accedere direttamente al file come se questo fosse effettivamente disponibile su disco. In questo caso il sistema intercetta la chiamata *read()* e richiama automaticamente su disco il file, mantenendo il client in attesa fino al completamento dell'operazione. La seconda modalità invece passa dall'interfaccia SRM. In questo caso il client richiede con uno specifico comando SRM che il file venga reso disponibile su disco, per procedere all'accesso solo dopo che la procedura di richiamo da nastro è terminata. In ogni caso, tutte le interazioni tra GPFS, TSM e StoRM[14], sono mediate da GEMSS, uno strato software sviluppato dall'INFN. GEMSS consente di aggregare e riordinare in modo dinamico e intelligente le varie richieste concorrenti di accesso ai nastri, in modo tale da avere un ordine di accesso ai file ottimale (i nastri sono dispositivi puramente sequenziali) e così ridurre al minimo le operazioni meccaniche di montaggio, smontaggio e ricerca sui nastri.

Come già anticipato, l'accesso allo storage avviene in accordo con la specifica SRM, adottata dalle comunità WLCG. SRM è un livello di astrazione che permette agli utenti di accedere allo storage attraverso un'interfaccia comune. L'interfaccia web service descritta nelle specifiche SRM fornisce un modo per spostare in modo trasparente i file da e verso la Grid, con libera scelta del protocollo di trasferimento e con un livello ben definito di servizio. SRM fornisce supporto per le più comuni operazioni sui file system, aggiungendo comandi più specifici per il controllo della gestione dello storage. L'interfaccia SRM è nata proprio per garantire l'interoperabilità, in modo che ogni data center possa fare le proprie scelte sul setup del proprio sistema di storage.

Ci sono diverse implementazioni SRM che supportano una varietà di configurazioni di storage. Per sfruttare al meglio la configurazione del MSS scelta per i Tier-1, l'INFN ha sviluppato una propria implementazione dell'interfaccia SRM, StoRM, progettata intorno al principio guida di sfruttare i vantaggi dei cluster file system come GPFS e Lustre[15]. StoRM si integra con GEMSS per garantire la gestione dello storage gerarchico, con la possibilità di trasferire file a e da storage basato su nastri.

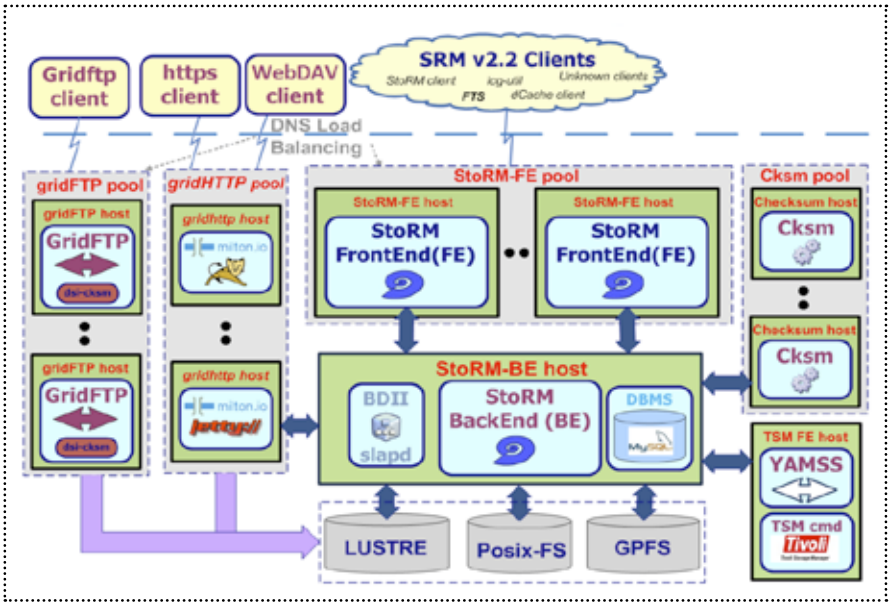

Fig 4 - Schema dell'architettura della SRM StoRM

Nell'ultima versione disponibile, StoRM fornisce una nuova interfaccia che riunisce operazione di storage management e di trasferimento file, in accordo con lo standard WebDAV[16]. Questa interfaccia nasconde i dettagli del protocollo SRM, e permette di

montare storage remoto come una partizione locale, o semplicemente di sfogliare i dati in uno spazio di storage tramite un *browser web*, con o senza autenticazione X509. Uno schema architetturale di StoRM è riportato in Fig. 4.

5. Evoluzioni future

L'interfaccia SRM, che ha servito egregiamente gli esperimenti LHC in questi anni, consentendo l'interoperabilità di centinaia di centri di calcolo con soluzione storage eterogenee, ha però evidenziato alcuni limiti. Il principale è che per centri che non hanno soluzioni di storage HSM, molte operazioni sono superflue e la complessità dell'interfaccia, specialmente per gli utilizzatori, non bilancia i vantaggi.

Una richiesta molto forte della comunità WLCG è quella di federare le risorse di storage, dando la possibilità di accedere alle risorse dei vari centri come se fosse un'unica entità. L'infrastruttura basata su redirector (come quello di xrootd[17] o http) implementa il *failover* nell'accesso a file, redirigendo il client ad una replica disponibile del file cercato.

Negli ultimi anni sono diventate molto diffuse soluzioni di Cloud Storage, come Amazon Web Services S3[18], Google Cloud Storage[19] oppure, con un target diverso, Dropbox[20]. Il superamento dell'interfaccia SRM verso interfacce che abbiano la semplicità delle interfacce di Cloud Storage è una questione di estrema attualità ed importanza. L'offerta di un servizio sullo stile di Dropbox alle comunità scientifiche, che integri la possibilità di utilizzare lo storage per task computazionali, è un'enorme possibilità. A questo scopo, nell'ambito dello sviluppo del prodotto INFN StoRM, è stata introdotta un'interfaccia WebDav, attualmente in fase di deployment. In questo modo, StoRM può integrare in un unico punto di accesso le funzionalità SRM standard, necessarie ad esempio per utilizzare in modo efficiente sistemi gerarchici dotati di risorse nastro, e storage utente o di gruppo, coadiuvato da un'interfaccia client WebDaV. Mediante questa interfaccia è possibile esportare un file system GPFS gestito da StoRM su un qualunque portatile o desktop, utilizzando un client grafico per la gestione, l'upload e il download dei file (si veda Fig. 5).

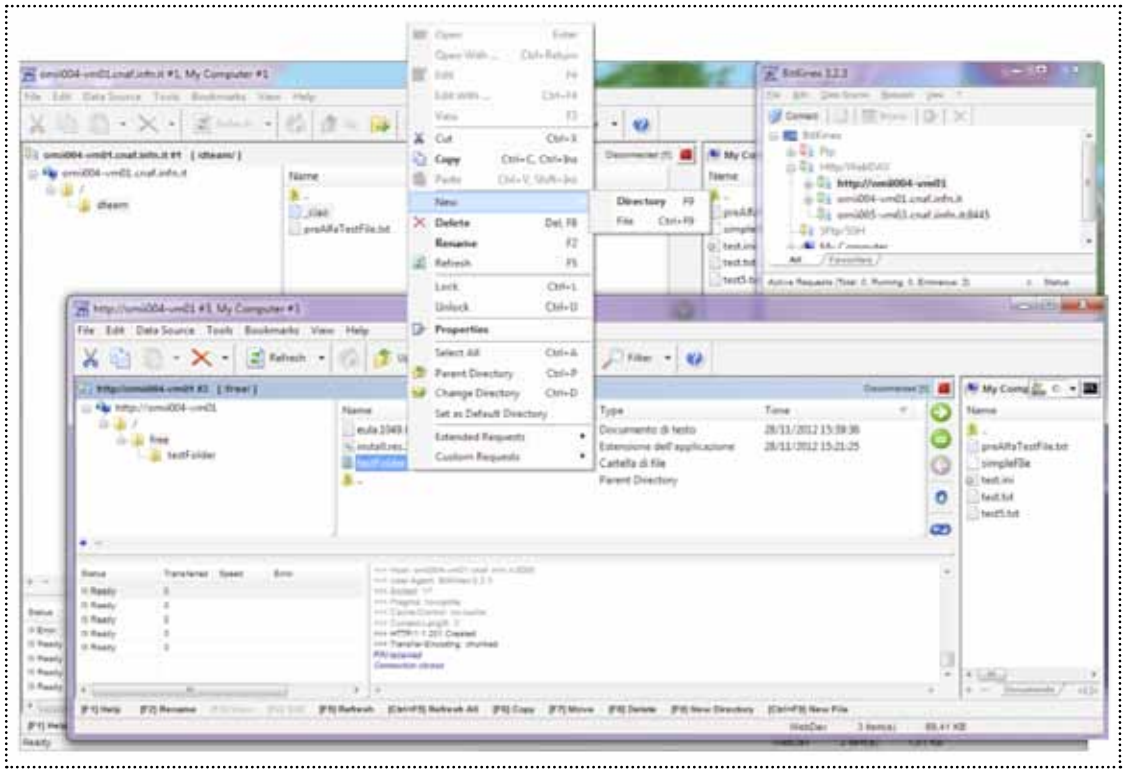

Fig 5 - Esempio di utilizzo dell'interfaccia WebDav in StoRM

6. Conclusioni

In questo lavoro abbiamo presentato l'architettura, le scelte implementative e i risultati del sistema di storage in uso al Tier-1 INFN del CNAF, che offre svariati PB di spazio disco e nastro alla comunità WLCG ed ad altre comunità della fisica delle alte energie. È stata descritta la dotazione hardware, e la scelta di un'integrazione di soluzione industriali e implementazioni ad hoc che hanno reso negli anni il Tier-1 come uno dei centri più affidabili di WLCG, caratterizzato da eccellenti prestazioni nell'ambito della Computing Grid di LHC. Sono state presentate anche le recenti evoluzioni del prodotto StoRM sviluppato dall'INFN, verso un sistema di cloud storage per le comunità scientifiche.

Riferimenti bibliografici

[1] http://www-03.ibm.com/systems/software/gpfs

[2] http://en.wikipedia.org/wiki/IBM_Tivoli_Storage_Manager

[3] https://github.com/italiangrid/gemms

[4] https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html

[5] http://wlcg.web.cern.ch

[6] http://aliceinfo.cern.ch

[7] http://atlas.web.cern.ch/Atlas/Collaboration

[8] http://cms.web.cern.ch

[9] http://lhcb.web.cern.ch/lhcb

[10] http://www.slac.stanford.edu/BF

[11] http://www-cdf.fnal.gov/collaboration

[12] http://www.garr.it

[13] http://www.globus.org/toolkit/docs/latest-stable/gridftp

[14] http://storm.forge.cnaf.infn.it

[15] http://wiki.lustre.org/index.php/Main_Page

[16] http://www.webdav.org

[17] http://xrootd.slac.stanford.edu

[18] http://aws.amazon.com/s3

[19] https://cloud.google.com/products/cloud-storage

[20] https://www.dropbox.com

Michele Di Benedetto
michele.dibenedetto@cnaf.infn.it

si è laureato in informatica all'Università di Bologna nel 2008.

Ha lavorato nel privato per un anno prima di essere assunto all'Istituto Nazionale di Fisica Nucleare, dove ha lavorato per 4 anni allo sviluppo di middleware di Grid.