**ISTITUTO NAZIONALE DI FISICA NUCLEARE**

**Sezione di Bologna INFN CNAF**

**INFN/TC-03/12**
**29 Agosto 2003**

# UTILIZZO DEL SOFTWARE CASTOR AL TIER1 CNAF

Ricci Pier Paolo
per TIER1 staff

*INFN-CNAF Sezione di Bologna, Viale Berti Pichat 6/2 40127 Bologna, Italy*

## Abstract

Lo sviluppo delle moderne tecnologie e la crescita delle richieste di spazio storage in termini di filesystem da parte degli esperimenti di fisica delle alte energie (HEPE High Energy Phisics Experiment) ci ha portati a considerare seriamente l'utilizzo di un hardware di librerie a nastro e di un software di archiviazione non proprietario. Nel presente articolo viene presentata la fase di test e pre-produzione svolta all'INFN CNAF Bologna utilizzando il software di archiving CASTOR realizzato al CERN con una descrizione dell'hardware e del software coinvolto oltre che ad una serie di risultati preliminari.

PACS.: 07.05.Bx

*Published by* ***SIS–Pubblicazioni***
*Laboratori Nazionali di Frascati*

## 1 INTRODUZIONE

Negli ultimi anni lo sviluppo della tecnologia ha portato grandi cambiamenti nell'area dello storage di dati in campo informatico. Il continuo crescere della capienza degli hard disk con un fattore che porta al raddoppio circa della capacità ogni anno a parità di costo rende la ricerca di una soluzione storage ottimale e durevole alquanto ardua. La scelta di tale soluzione dipende largamente dal tipo di utilizzo o meglio dall'accesso ai dati che si intende archiviare. Nel nostro caso gran parte dei dati HEPE sono composti in prevalenza da file di grandi dimensioni (da 100MB fino a 1-2GByte) che vengono scritti sequenzialmente durante la fase di "produzione" di dataset. Un classico esempio può essere una produzione MonteCarlo di un grosso numero di dati a partire da un piccolo dataset. Un altro esempio può essere l'archiviazione di una serie di dati di misura sperimentali senza che essi siano stati pre-analizzati. La richiesta di archiviazione di tali dati porta ad un altissimo quantitativo di spazio storage necessario in quanto solo una preanalisi può portare alla riduzione di un fattore sensibile dei dati effettivamente necessari. Nel caso di dati sperimentali è preferibile inoltre conservare per un certo periodo di tempo (da 1 a 5 anni) i dati ottenuti anche se non letti successivamente alla fase di pre-analisi. Una prima visione ai problemi coinvolti ci porta quindi ad una serie di considerazioni che sono in generale vere per il tipo di dati che si intende archiviare:

1) Si tratta si serie di dati di dimensioni comprese tra 100MB a 1 GB

2) La scrittura di tali dati avviene sequenzialmente, l'operazione può essere massiccia (ovvero coinvolgere un totale di migliaia di file di cui sopra) ma in generale è periodica e coincide con le fasi di produzione dell'esperimento

3) La lettura successiva di tali dati segue grosso modo i principi della scrittura.

Difficilmente si ha una lettura casuale dei file e qualora essa avvenga il singolo file è comunque acceduto sequenzialmente. La fase di lettura massiccia è comunque periodica e coincide con la fase di pre-analisi e analisi dell'esperimento. Inoltre le previsioni dei dati sperimentali prodotti nei prossimi anni in particolare per LHC (Large Hadron Collider) in realizzazione al CERN a Ginevra porta alla ricerca di soluzioni storage che possano archiviare grandi quantità di dati (PByte) per lungo termine a costi accettabili 1).

Alla luce di questo è chiaro che una archiviazione su nastro che performi in maniera accettabile su letture e scritture sequenziali non è concettualmente sbagliata. Inoltre i nastri hanno un tempo di vita medio superiore ai dischi e pertanto ben si prestano all'utilizzo di uno storage a medio e lungo termine (superiore a 10 anni).

Nella scelta del software per l'archiviazione su nastro si è deciso di utilizzare il software CASTOR sviluppato ed utilizzato dal CERN 2) come Herarchical Storage Manager (HSM) e liberamente disponibile i centri di ricerca. CASTOR (e in generale qualunque software HSM) permette di creare un filesystem virtualmente infinito su differenti tipi di media (disco e nastro distribuito ad esempio su diverse librerie) e presentarlo tramite un'opportuna interfaccia all'utente finale sotto la forma di una unica entità (in generale un

filesystem). Caratteristica del software HSM è che tutte le operazioni “a basso livello” di gestione e di interazione dei vari media fisici su cui risiedono realmente i dati sono completamente invisibili all’utente finale.
Nel seguito dell'articolo non verranno pertanto descritti i vari sistemi di storage su nastro presenti al momento sul mercato ma porremo l'attenzione su quanto realizzato in termini software e hardware qui al CNAF spiegando logicamente di quanto in quanto le motivazioni che ci hanno portato ad un scelta tecnica rispetto ad un'altra.

## 2 DESCRIZIONE HARDWARE

Al momento di iniziare la fase di test di archiving il CNAF disponeva di una libreria STORAGETEK STK L180 con attivati e disponibili 174 slot grandezza standard (a 4') e controllo della robotica tramite SCSI. La libreria disponeva di 2 drive STORAGETEK STK 9840 con circa 70 cassette. Il drive 9840 è un ottimo drive per l'accesso frequente ai dati avente un tempo di ricerca medio piuttosto basso (8-11s)[1] e un altrettanto basso "loading time" (tempo necessario al drive per poter accedere ai dati una volta inserito il nastro) di circa 4s. Quello che era limitativo nell'utilizzo di tali drive era la bassa capienza di 20GB per cassetta. Considerando quindi il fatto che i dati LHC sono supposti difficilmente comprimibili[2] la capienza di 20GB per tape avrebbe portato la capienza massima della libreria a circa 3TB effettivamente troppo bassa. Pertanto abbiamo provveduto ad installare nella libreria 4 drive IBM LTO che pur avendo un alto tempo di ricerca medio (68s) e loading time (20s) posseggono una elevata capacita' nativa per cassetta di 100GB e un transfer rate sequenziale di 15MB/s ampiamente sufficienti per i nostri scopi. Con un set iniziale di 100 cassette ed utilizzando solo gli LTO per l'archiving la capacità totale nativa è arrivata a 10TB condividendo la libreria con i precedenti 9840 che occupano i restanti 70 slot. Questi ultimi sono stati dedicati unicamente al backup delle macchine utente e di servizio del CNAF tramite il software proprietario LEGATO NSR.

La configurazione hardware della libreria è schematizzata in Fig.1. Al fine di poter condividere sulla stessa libreria 2 tipi diversi di drive che vengono acceduti da due macchine diverse è stato necessario approntare una macchina SPARC Sun Solaris con il software proprietario STORAGETEK ACSLS per il controllo diretto via SCSI del braccio robot della libreria. In tal modo è stato possibile "partizionare" la libreria e garantire l'accesso simultaneo sia dal software LEGATO per il backup su drives 9840 sia da CASTOR per l'archiving su drives IBM LTO poichè i comandi SCSI di montaggio e smontaggio dei tapes sulla libreria vengono trasmessi via IP ad ACSLS tramite gli opportuni agenti Storagetek installati sulle due macchine client (nel caso LEGATO dal software STORAGETEK Library Attach mentre nel caso CASTOR dal CSC Developers TookKit).

---

[1] Se non diversamente indicato i valori riportati si riferiscono ai dati ottenibili dai datasheet dei produttori.

[2] Prove effettuate tramite algoritmi di compressione Lempel-Ziv algorithm di Gzip,compress LINUX hanno mostrato un 20% massimo di guadagno effettivo sui dati sperimentali ALICE.

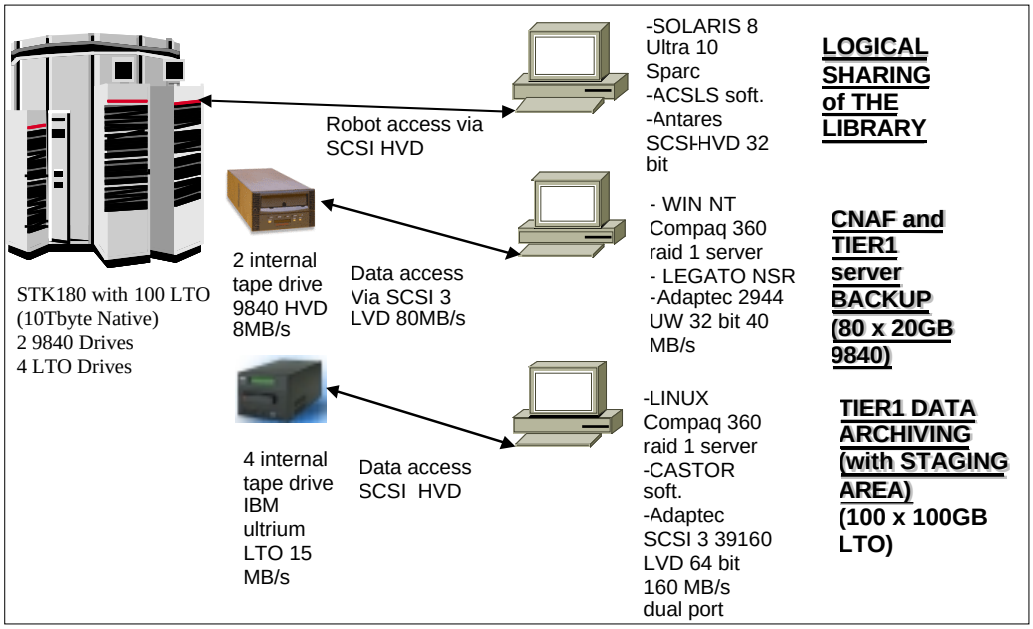


**FIG. 1:** Una schematizzazione del sistema hardware/software di condivisione della libreria realizzato al CNAF

In particolare per ciò che riguarda l'installazione di CASTOR in questa prima fase si è deciso di utilizzare una singola macchina con una alta affidabilità e buone prestazioni su cui concentrare tutti i servizi necessari al software e collegandola direttamente alle parti hardware coinvolte. In particolare le parti hardware sono composte:

1) Da un disco esterno da utilizzare come staging area[1]. Il disco utilizzato è un RAIDTEC HI 160 con intefaccia SCSI LVD da 160MB/s con spazio utile di circa 2TB.

2) Dai 4 drive IBM LTO con interfaccia SCSI LVD 80MB/s.

3) Dalla macchina server CASTOR una Compaq DL 360 Doppio processore Penthium 3 a 1.13GHz 512MB di memoria RAM, 2 dischi SCSI da 18GB in raid 1 (mirroring) tramite interfaccia RAID integrata. Il sistema operativo installato è Linux RedHat 7.2 standard kernel (2.4.7-10smp) con applicati patch sui moduli SCSI sviluppati al CERN per i tape drive.

Per accedere sia ai drive sia al disco è stata utilizzata una interfaccia HBA SCSI Adaptec 39160 con doppia uscita LVD SCSI 160MB/s, e per velocizzare l'accesso via rete la macchina è connessa tramite HBA Ethernet Dlink Gigabit Ottica direttamente allo switch centrale (centro stella) GigaBit del CNAF.

[1] Per area staging si intende una sorta di buffer su disco nella quale i dati risiedono in attesa di migrare su nastro. I file rimangono nella staging area anche dopo la migrazione su tape in base alle cosiddette politiche di staging. Lo scopo della staging area è unicamente quella di migliorare le prestazioni in scrittura e soprattutto in lettura (ovviamente operazioni di lettura di file già presenti in staging area non necessitano della rilettura del file dal nastro con ovvi benefici)

## 3 TEST PRELIMINARI HARDWARE

L'attenzione principale dei test preliminari dell'hardware è stata quella di eliminare eventuali "bottleneck" (o colli di bottiglia di un sistema) dell'intero apparato partendo da quelli ovvi che potrebbero sorgere da un'errata scelta/configurazione dell'hardware. L'attenzione primaria è stata rivolta ai drive IBM LTO e utilizzando i driver standard di Linux si è connesso un singolo drive all'interfaccia scsi del server e si è provveduto ad un piccolo testbench. Per testare i drive si è utilizzato un block size "ottimale" di 128KB. Blocchi maggiori non miglioravano le prestazioni mentre blocchi minori degradavano sensibilmente le prestazione a causa probabilmente di un maggior numero di operazioni start-stop del drive. Per cio' che riguarda i test sono stati utilizzati diversi file di grandezza 2GB gia' compressi tramite gzip di Linux allo scopo di bypassare l'algoritmo di compressione hardware del tape LTO. Tali file sono stati scritti e letti utilizzando sia utilizzando il comando "tar" di Linux sia usando un apposito script perl con chiamate open e syswrite. I risultati mediati tra i 2 metodi ripetuti 10 volte sono riportati in Tab. 1 e mostrano valori in lettura e scrittura da 12 a 14MB/s. Questo ci ha permesso di affermare che un daisy chain SCSI di tutti e quattro i drive con il protocollo SCSI 3 a 80MB/s non sarebbe stato limitativo (ovvero i 4 drive lavorando assieme al massimo della velocita' di trasferimento

| N. Drive (simultanei) | Scrittura (MB/s) | Lettura (MB/s) | Scrittura aggregato (MB/s) | Lettura aggregato (MB/s) |
|---|---|---|---|---|
| 1 | 13 | 12 | 13 | 12 |
| 2 | 12 | 12 | 24 | 24 |
| 3 | 12 | 12 | 36 | 36 |
| 4 | 11 | 11 | 44 | 44 |

**TAB. 1:** Risultati test sui drive LTO media su 10 prove ripetute

non riescono a saturare la banda dello SCSI). Infatti ripetendo i test di lettura e scrittura in parallelo su tutti e 4 i drive i singoli risultati non si sono discosti molto dai test sul singolo drive e i risultati in aggregato di throughput complessivo superano i 40MB/s. Pertanto è stato possibile affermare che con l'utilizzo contemporaneo dei 4 drives si possono raggiungere prestazioni in operazioni di scrittura e lettura sequenziali comparabili con l'accesso al disco.

Per ciò che riguarda l'area di staging il disco RAID SCSI Raidtec già citato è stato collegato al server tramite l'interfaccia SCSI LVD 160MB/s già citata ed è stato realizzato un unico RAID SET a RAID 5 (con un disco di Hot Spare) partizionato logicamente in vari blocchi da 50-200GB formattati successivamente via Linux in filesystem ext-3. La scelta di RAID 5 pur avendo ovvie ripercussioni sulle prestazioni è stata attuata alla luce di un futuro utilizzo della area disco di staging non solo come area buffer transitoria di dati (per la quale sarebbe bastato anche un RAID meno sicuro ma più performante come il RAID 0) ma

eventualmente anche come backup disco dei dati migrati su nastro. È infatti possibile tramite la politica di staging di consentire ad una parte di dati maggiormente sensibili di permanere indefinitivamente nell'area di staging in modo tale da essere doppiamente sicuri della avvenuta archiviazione dei dati sul sistema CASTOR. Tramite alcuni noti benchmarch[1] e script realizzati utilizzando il comando linux "dd" sono state testate le performance di lettura e scrittura del disco per file di grandezza 4 volte la memoria fisica della macchina (ovvero file di grandezza 2GB). Le prestazioni in accesso sequenziale sono riportate in Tab. 2 e mostrano come già annunciato che in realtà per dischi RAID 5 le performance sia in lettura che in scrittura (pur inferiori per l'offload della scrittura della parità del RAID 5) sono paragonabili all'aggregato di 40MB/s realizzato con i quattro drive che scrivono (o leggono) in parallelo. Pertanto se opportunamente stabiliti i parametri del software CASTOR le operazioni di migrazione disco-nastro dovrebbero essere tali per cui possa essere possibile scrivere su nastro con la stessa velocità con cui gli utenti scrivono sull'area di disco staging con la ovvia conseguenza di non presentare all'utente nessun tipo di degrado di performance (per ciò che riguarda le operazioni di scrittura).

| TestBench (media tra 30 prove) | Scrittura sequenziale (MB/s) | Lettura sequenziale (MB/s) |
|---|---|---|
| Dd | 35 | 45 |
| Diskrate | 48 | 54 |
| Bonnie | 45 | 56 |
| bonnie++ | 45 | 56 |
| Iozone | 43 | 54 |

**TAB. 2:** Risultati test sul disco RAID 5, media su 30 prove ripetute

Poiché l'accesso dei client CASTOR al filesystem avviene tramite TCP/IP con protocollo rfio è stato necessario utilizzare una connessione di rete sufficientemente performante. A tale scopo come detto sulla macchina è stata installata una interfaccia Gigabit Ethernet su fibra ottica multimodale connessa direttamente allo switch centrale del CNAF. Test indicativi da una macchina con hardware identico utilizzando i testbench di protocolli TCP/IP e UDP "netstat" e "ttcp" hanno mostrato risultati in entrambi i sensi pari a circa 500Mb/s in TCP/IP e 800/900Mb/s in UDP. Tale banda disponibile sulla rete è sufficiente a garantire un throughput verso l'area di staging su disco che saturi la banda di scrittura/lettura del disco stesso. L'analisi di questi tre risultati ci porta alla conclusione che nessuno di essi può costituire un singolo bottleneck. È indubbiamente vero che in realtà andrebbe considerato il monitoring delle varie risorse della macchina (CPU, memoria fisica etc…) in modo da considerare il carico delle operazioni di input/output in complessivo ma i risultati

[1] Sono stati utilizzati i testbench Diskrate,Bonnie,Bonnie++, e Iozone.Tutti questi testbench sono disponibili liberamente e ampliamente documentati in rete.

della fase di test preliminare con CASTOR descritta in seguito contiene già parte di tali informazioni in condizioni operative.

## 4 INSTALLAZIONE SOFTWARE CASTOR

L'installazione del software CASTOR sulla macchina suddetta è stata effettuata tramite la compilazione e l'installazione del tarball fornito dagli sviluppatori di CASTOR al CERN e una fase piuttosto lunga di debugging e messa a punto dei vari file di configurazioni necessari per il funzionamento del software. Parallelamente a tale lavoro di debugging gli sviluppatori del CERN hanno provveduto a realizzare una documentazione più ampia ed esauriente sulla gestione e l'utilizzo del software la cui installazione dovrebbe ora esser relativamente meno macchinosa. La versione di CASTOR utilizzata al momento al TIER1 CNAF è la 1.3.5.1pre e verrà presto aggiornata agli ultimi release. Nella breve descrizione che segue si prenderà sempre come riferimento tale versione che non includeva ancora alcuni demoni e features presenti nelle versioni più recenti come l'estensione per file superiori a 2GB 3). CASTOR si presenta fondamentalmente come una serie di processi runnanti in background (demoni) e potenzialmente su macchine diverse. I vari demoni si occupano della realizzazione delle varie operazioni necessarie al funzionamento del sistema (come l'aggiornamento e l'accesso al filesystem virtuale di CASTOR, la gestione delle policy di staging e le operazioni fisiche di lettura e scrittura sui drive). Al momento è stato deciso di concentrare tutti i processi sulla singola macchina poiché potrà sempre essere possibile decentrare parte dei servizi in seguito.

Dal lato cliente occorre compilare ed installare la serie di comandi rfio e configurare alcune semplici variabili di shell. In tal modo una volta registrato USER ID e GROUP ID dell'utente presente sulla macchina cliente anche sulla macchina server CASTOR è già possibile iniziare a realizzare i/o sul filesystem virtuale di CASTOR tramite la famiglia di comandi rfio e le relative API. In questa fase l'obbiettivo è stato quello di realizzare un test consistente di scrittura da una macchina client tramite protocollo rfio e verificare le performance in queste condizioni. Si è inoltre provveduto a leggere successivamente parte dei dati scritti e a fornire indicazioni sul throughput in lettura in tali condizioni.

## 5 INSTALLAZIONE MACCHINA CLIENT E FASE DI TEST

A tale scopo è stata installata una macchina di caratteristiche analoghe a quella facente da server CASTOR con RedHat Linux 7.2. L'hardware utilizzato è stato un server Dell 1550 doppio processore Pentium 3 a 1.13GHz 512MB di memoria RAM, 2 dischi SCSI da 18GB interfaccia di rete HBA Ethernet Dlink Gigabit Ottica e interfaccia Fibre Channel Qlogic 2200 1Gb/s su rame connessa ad uno storage esterno Dell PowerVault 660F con 1TB dedicato per questo test. Test preliminari sono stati attuati sia nella connessione TCP/IP UDP da e verso la macchina server CASTOR con risultati analoghi a quelli già presentati nel paragrafo III. Inoltre test nell'accesso in lettura e scrittura sul disco Dell PowerVault mostravano risultati simili a quanto ottenuto per il disco SCSI RAIDTEC. Il disco PowerVault da 1TB viene inteso come un'estensione del disco locale sotto forma di storage locale ad accesso veloce e aveva lo scopo di contenere un considerevole quantitativo di dati

sperimentali difficilmente comprimibili (quindi simili alle condizioni reali di lavoro di CASTOR). Per il test sono stati usati circa 600 file di dati di produzione dell'esperimento ATLAS corrispondenti a file pressoché incomprimibili di grandezza variabile tra 300 e 1500 MB per un totale di 470GB corrispondenti a circa 5 tape LTO da 100GB nativi. Lo spazio effettivamente occupato sui tape mostra che la compressione hardware su tali dati non ha permesso un guadagno maggiore del 5%. L'area di staging è stata configurata con 100GB di spazio disco sul RaidTec pari quindi al 20% circa dello spazio totale dei dati scritti su CASTOR. La politica di migrazione è stata fissata a forzare una migrazione una volta raggiunta la quantità di almeno 20GB di dati migrabili e l'area di staging viene periodicamente svuotata di parte dei dati già migrati una volta piena all'80%. La configurazione era stata fissata in maniera tale che per ogni migrazione un massimo di due drive potessero scrivere in streaming parallelo su due tape assegnati al pool di test. I risultati del throughput sono riportati in Fig.2 dove l'ascissa rappresenta il trasferimento e l'ordinata il throughput in KB/s ottenuto dalla macchina client. Attorno al 150$^{o}$ trasferimento (evidenziato dalla linea tratteggiata in figura) è iniziata la prima migrazione sui tape dei dati scritti nell'area di staging tramite stream parallelo dei due drive assegnati. Dai file di log ottenuti da CASTOR si è potuto stimare che il throughput sul singolo drive si manteneva intorno ai 12-13MB/s con un aggregato conforme a quanto visualizzato sulla Tab.1. In altre parole pur avendo continua scrittura sull'area di staging da parte del client non c'è stato calo di performance sulla scrittura sui tape cosa che si deduce ampliamente anche dal grafico[1] riportato in Fig.2.

[1] Si può invece notare una minore fluttuazione dei valori del throughput ed una minore discrepanza dal valore medio di circa 20MB/s. Tale fenomeno è forse da ricondursi alla correlazione spiegata in seguito tra throughput e grandezza dei file.

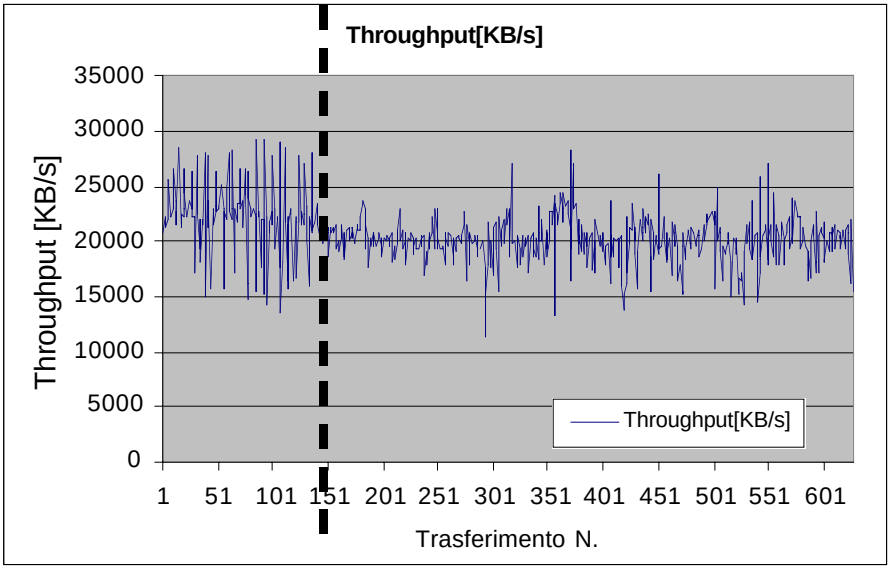


Fig.2 Throughput di scrittura (KB/s) di 620 file trasferiti via CASTOR rfio da un singolo client.

Poiché una media globale sul throughput di dati visto dal lato client dà un risultato di circa 20500 KB/s è ovvio che in questo caso la scrittura sui tape paralleli riusciva a "svuotare" l'area di staging più velocemente di quanto il client fosse in grado di riempirla. Come risultato dal lato client non vi è stato alcun degrado di performance su tutto il periodo di test cosa che sarebbe potuta succedere qualora l'area di staging si fosse riempita e CASTOR fosse stato costretto a mettere il client in attesa di completare le migrazioni e svuotare l'area di staging dai dati migrati.

In tale configurazione inoltre è stata evidenziata una diretta correlazione tra la grandezza dei file scritti ed il throughput ottenuto: in Fig.3 è riportata la dispersione del throughput in ordinata in relazione con la grandezza del file riportato in ascissa. Si può notare chiaramente la correlazione lo stabilizzarsi del throughput attorno ai 20MB/s al crescere della grandezza dei file trasferiti, correlazione che forse si può spiegare tenendo presente che la cache di memoria fisica delle due macchine (client e server) era di 512MB pertanto troviamo le forti oscillazioni del throughput quando si trasferiscono file sufficientemente piccoli da fare sentire sul sistema l'effetto della memoria cache.

Per ciò che riguarda la lettura occorre premettere che ovviamente vi è una sostanziale differenza tra il throughput ottenibile leggendo dati che sono già presenti sull'area disco staging di CASTOR contro dati che occorre invece rileggere direttamente dal tape. In quest'ultimo caso occorre infatti aggiungere il tempo necessario all'operazione di stage-in del file (copia del file dal tape all'area disco di staging) che include il montaggio/smontaggio del tape, il caricamento del tape (loading time), il posizionamento al file richiesto e la lettura vera e propria.

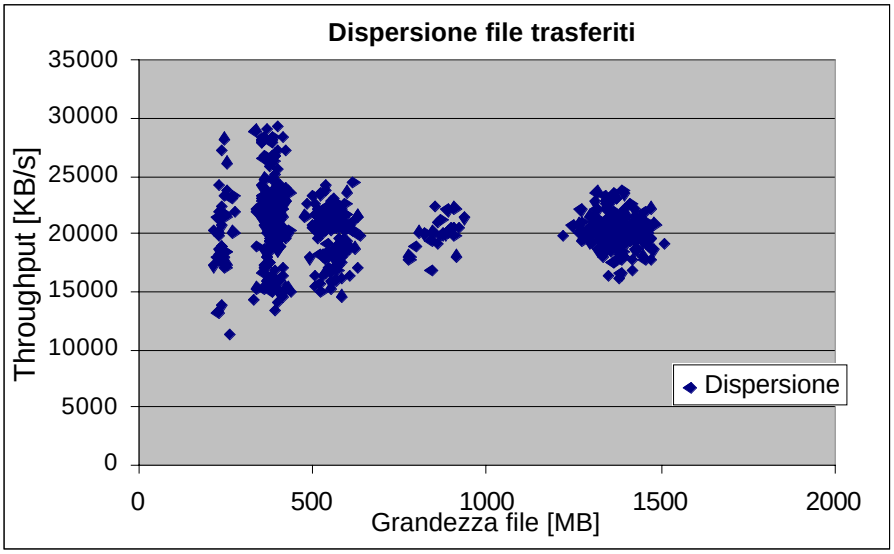


Fig.3 Dispersione dei 620 file trasferiti in scrittura via CASTOR rfio da un singolo client.

Pertanto si è provveduto a effettuare una prova di lettura di 50 file dei 600 scritti in precedenza assicurandosi che essi non risiedessero nell'area disco di staging. Come contro prova è stata effettuata una lettura di 50 file di caratteristiche analoghe che invece sicuramente risiedevano nell'area di staging. Nel secondo caso i risultati sono stati simili a quelli ottenuti nel caso di scrittura con una media di throughput in lettura di circa 18MB/s. Nel primo caso invece il throughput è risultato pesantemente penalizzato dai tempi necessari alla operazione di stage-in. Poiché tali tempi (salvo l'operazione di riposizionamento che aumenta con l'allontanarsi dall'inizio del tape essendo l'LTO un tape lineare a singolo rullo e non è ovviamente prevedibile a priori) sono considerabili fissi in una prima approssimazione è ovvio che in generale il throughput risulta complessivamente tanto più penalizzato quanto il file da rileggere è piccolo. Inoltre non essendoci una correlazione diretta tra il file richiesto in lettura e la posizione all'interno del tape la dispersione complessiva dei risultati porta dei valori fortemente oscillanti pertanto si danno solo indicazioni di massima dei risultati. Nel nostro caso essendo il file più piccolo di grandezza 300MB tali file hanno in generale ottenuto il throughput peggiore con un valore di circa 2-2.5MB/s in media. File di grandezza maggiore fino a quelli di grandezza massima di 1-5GB hanno ottenuto invece un throughput medio superiore fino a circa 3.5-4MB/s. Questo porta ad una media sul campione di 50 file in lettura di circa 3 MB/s ma occorre ripetere che vi è una forte dispersione dei risultati pertanto grafici e analisi simili a quelle effettuate in scrittura non sono significative. E' stata invece effettuata una verifica sul checksum dei file riletti da nastro rispetto ai file originali presenti sul client e tale operazione ha mostrato che tutti i file riletti da CASTOR erano coerenti con i file originali.

## 6 CONCLUSIONI

Il presente articolo intendeva mostrare le specifiche tecniche del sistema hardware costituente il sistema di storage su nastro installato all' INFN CNAF Bologna. Sono stati presentati in dettaglio i test riguardanti i singoli componenti hardware costituenti le basi del sistema di archiviazione su nastro ovvero l'area disco connessa come staging area e i tape drive stessi. Inoltre sono stati mostrati alcuni risultati preliminari ottenuti utilizzando il software CASTOR sviluppato dal CERN. L'analisi dei singoli test e dei risultati preliminari ottenuti con il software sono piuttosto promettenti. Il risultato più interessante è senza dubbio quello ottenuto in scrittura in cui su operazioni sequenziali una scelta opportuna di scrittura in parallelo su più drive permette di ottenere un throughput aggregato che si avvicini a quello ottenibile scrivendo semplicemente su uno storage di disco. Dal lato utente quindi in operazioni di scrittura sequenziali non c'è nessuna differenza in prestazioni nel fatto che i dati vengano effettivamente scritti su un mass-storage su nastro con un ovvio beneficio per ciò che riguarda il costo e l'affidabilità a lungo termine del media. Per ciò che riguarda la lettura pur essendo il sistema affidabile si sono evidenziate come prevedibile forti differenze nelle prestazioni a seconda che il file risiedesse ancora nell'area di staging o fosse necessario invece rileggerlo da nastro. In quest'ultimo caso infatti essendo l'LTO un tape drive con un alto loading time e tempo di ricerca medio il throughput in lettura risulta fortemente penalizzato. In generale poiché non è possibile al momento variare sensibilmente tali parametri su tape drive differenti senza perdere il vantaggio dell'alta capacità dei tape LTO l'unico modo per ovviare a tale inconveniente è utilizzare una staging area sufficientemente grande e una oculata politica di rimozione dei file dall'area di staging. In tal modo infatti i file a maggior accesso rimarrebbero comunque nell'area di staging assicurando prestazioni in lettura elevate, mentre le operazioni di lettura diretta del dato da tape avverrebbe solo dai dati difficilmente acceduti (o al limite mai nel caso di copie di backup).

Concludendo quindi, l'opinione avuta su un sistema di storage su nastro così costituito risulta piuttosto positiva e tale sistema di archiviazione presso il TIER1 INFN CNAF merita senz'altro di essere ulteriormente sviluppato.

## 7 REFERENZE BIBLIOGRAFICHE

(1) J.D. Shiers "Data Management at CERN: Current Status and Future Trends, European Laboratory for Particle Physics (CERN), Geneva, Switzerland

(2) O. Barring, J. Durand, B. Couturier, C. Curran, G. Lee, T. Osborne, User Guide for CASTOR, European Laboratory for Particle Physics (CERN) March 10, 2003

(3) J. Baud , Castor Architecture,Document derived from two tutorials given in http://castor.web.cern.ch/castor/DOCUMENTATION