ISTITUTO NAZIONALE DI FISICA NUCLEARE

Centro Nazionale CNAF

INFN/TC-03/19
12 Dicembre 2003

SISTEMI STORAGE SU DISCO UTILIZZATI AL TIER1 CNAF

Ricci Pier Paolo, Stefano Zani

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

Abstract

Il presente articolo si presenta come una descrizione introduttiva dei vari apparati di stoccaggio dati su disco attualmente in produzione all'INFN TIER1 presso il CNAF di Bologna. Per ogni apparato è presente sia una descrizione tecnica dell'hardware sia una serie di note sul sistema software di management,configurazione e monitoraggio. Inoltre sono stati riportati per ogni apparato una serie preliminare puramente indicativa di risultati di performance di accesso sequenziale al disco oltre ad una descrizione generale del sistema di storage disco del TIER1 visto nella sua globalità.

PACS.: 07.05.Bx

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

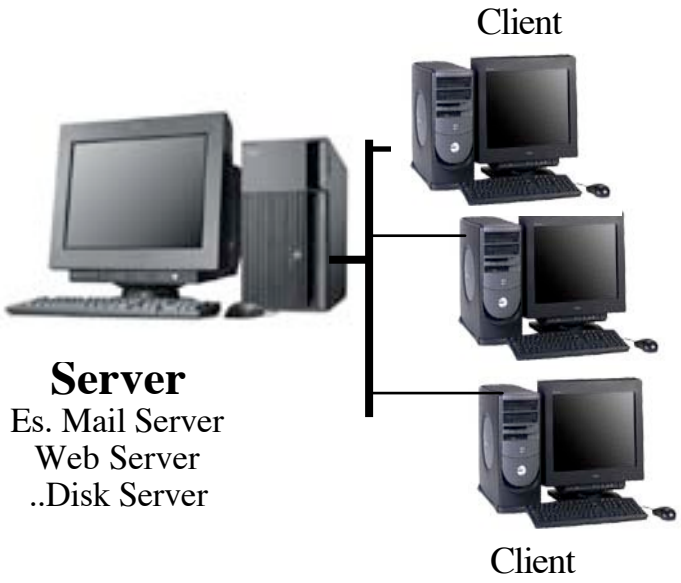
1 INTRODUZIONE

Nell'ambito del progetto INFN TIER1 le prospettive di crescita [1] dello spazio disco richiesto dagli esperimenti LHC prevedono ordini di grandezza tali da fare risultare una prima fase di test e valutazione degli apparati storage presenti sul mercato assolutamente necessaria. In particolare al momento i sistemi storage utilizzati in produzione al TIER1 CNAF sono afferenti alle 3 diverse tipologie di accesso e stoccaggio dati su disco ovvero:

NAS: Network Attached Storage
SAN: Storage Area Network
SAS: Server Attached Storage.

1.1 SAS

La tecnologia denominata ***Server Attached Storage*** o ***Direct Attached Storage*** è la più consolidata e consiste semplicemente nel collegare dischi direttamente ai server applicativi sfruttando la capacità di espansione interna dei server stessi o mediante disk array (tipicamente SCSI) afferenti a controller inseriti direttamente nei server [2].



Pro:

- Costa relativamente poco
- E' di facile gestione
- E' abbastanza performante (Applicazione<->Disco)

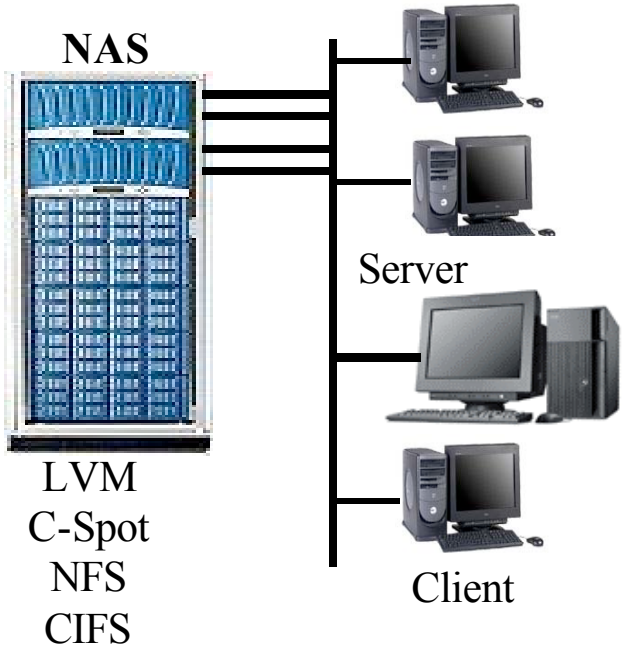
Contro:

- E' poco scalabile
- La CPU del server viene utilizzata sia per "servire" il disco ai client che per eseguire le applicazioni

1.2 NAS

La rapida evoluzione nel campo delle reti locali in grado di trasportare dati a velocità fino al Gigabit al secondo permette lo sviluppo di nuove unità di massa "di rete" chiamate ***Network Attached Storage***.

I NAS non sono altro che server di disco collegati alla rete tramite una o più interfacce ad alta velocità il cui unico scopo è quello di rendere disponibile alla massima velocità possibile grandi quantità di spazio disco a più calcolatori collegati in rete. Con questo genere di approccio è possibile concentrare la maggior parte dello storage di un centro di calcolo su di un numero limitato di server con notevoli vantaggi per la gestione ed il backup.



Pro:

- Gestione centralizzata di tutto lo storage
- Scarica gli Application Server
- Gestione dinamica di "Volumi Grandi" (LVM) con gestione del Locking
- Condivisione fra mondo UNIX e Windows(CIFS,NFS,Atalk..)

Contro:

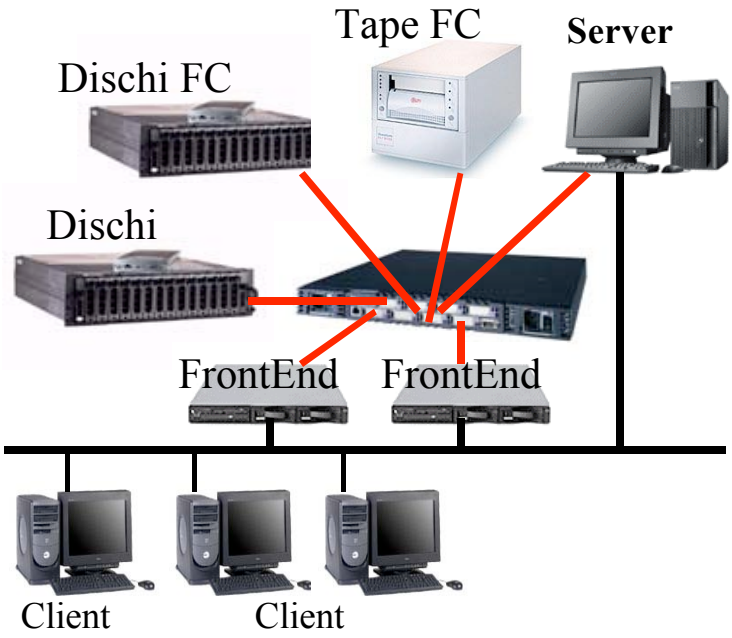
- Costano di più del disco "nudo"
- Sono generalmente sistemi Proprietari (chiusi).
- Prestazioni medio-basse con NFS

Un altro aspetto molto importante è che in questo modo si scaricano le CPU dei server dalla gestione di un intenso accesso al disco.

1.3 SAN

Alla filosofia di collegare grandi quantità di storage ad una Local Area Network basata su protocolli "Ethernet" dei NAS si contrappone l'idea di creare una rete di unità di massa (***Storage Area Network***) a cui i Server si collegano utilizzando la tecnologia Fibre Channel (in seguito nel testo abbreviato con F.C.). Il protocollo Fibre Channel che deriva dallo SCSI, permette di creare vere e proprie reti di unità di massa utilizzando anche switch Fibre Channel in maniera analoga a quanto avviene per l'ethernet. I server che vogliono accedere a queste unità di massa lo fanno utilizzando opportuni controller (controller Fibre Channel).

Un volta creata una SAN ogni calcolatore che vi accede può avere accesso come raw device a tutte le unità connesse alla SAN stessa.



Pro:
- Maggiore scalabilità
- Flessibilità elevata
- Prestazioni dipendenti dalle configurazioni ma comunque elevate (F.C.)

Contro:
- Costo elevato (Switch F.C, Dischi F.C, FrontEnd F.C.)
- Software di Gestione Complesso (LVM, File Locking..)

La tecnologia SAN è in via teorica la più espandibile e, se accompagnata da sistemi di LVM e clustering lato server, la più flessibile. E' tuttavia una tecnologia ancora abbastanza costosa e tutto sommato giovane dal punto di vista delle applicazioni e dello sviluppo di strati software per la gestione di volumi logici.

Ognuna di queste tipologie (SAS, NAS e SAN) presenta vantaggi e svantaggi intrinseci all'architettura stessa del sistema e non è possibile in questa prima fase di test e messa in produzione degli apparati descritti nel presente documento individuare una architettura ottimale per l'utilizzo disco al TIER1. Sicuramente un'analisi dei problemi intrinseci ad architetture "statiche" come SAS e NAS ci hanno permesso di determinare che per l'ordine di grandezza dello spazio disco presente al TIER1 e previsto nell'arco del

prossimo triennio (entro il 2007 si avranno almeno 500TB di spazio disco) un'architettura di tipo SAN sembra essere la più indicata per la maggiore potenziale flessibilità.

Nei seguenti capitoli descriveremo con una breve introduzione tecnica i vari apparati storage presenti in produzione al TIER1 soffermandoci in particolare sugli aspetti della gestione che giudichiamo più interessanti oltre che su una serie indicativa di risultati sulle performance ottenuti mediante semplici tool di test. In seguito riporteremo lo schema attualmente utilizzato per la messa in produzione della totalità dello storage disco presente attualmente al TIER1 con una descrizione delle varie problematiche connesse alla gestione globale del sistema.

2 APPARATO SAS RAIDTEC HI 160

2.1 Descrizione tecnica, spazio fisico e costi

Uno dei primi apparati storage acquistato dal TIER1 CNAF è stato il Raidtec Flex Array Hi 160 composto da 12 dischi Seagate Internal SCSI da 173GB configurato in un volume raid-5 su 11 dischi con un disco in configurazione Hot-Spare. Lo spazio lordo dichiarato era di 2.076 GB mentre, con la configurazione raid-5, lo spazio disponibile globale visto dal sistema operativo formattando con filesystem ext3 con lo 0% di "reserved-blocks-percentage" per ottimizzare l'utilizzo del disco era di circa 1.690 GB. L'accesso è effettuabile tramite 2 uscite SCSI LVD a 160MB/s anche se la seconda interfaccia è certificata all'utilizzo simultaneo solo costruendo un diverso array RAID e configurando il controller in maniera tale che da ogni uscita SCSI sia visibile solo un differente array raid-5. Pertanto per i nostri scopi si è utilizzata solo una uscita SCSI con una connessione diretta ad un server Dell 1550 Red Hat Linux 7.2 tramite una interfaccia SCSI HBA Adaptec 39160 in configurazione appunto SAS. L'oggetto è stato acquistato alla fine del 2001 ad un prezzo per TByte piuttosto alto (oltre 13KEURO compresa assistenza on-site NDB per 3 anni) visto che conteneva il massimo della tecnologia per ciò che riguardava i dischi SCSI disponibili al momento.

2.2 Metodo di controllo e configurazione

Il sistema di controllo configurazione ed allarmistica dell'apparato era incluso in un'applicazione Windows comunicante con il controller via protocollo embedded su SCSI. Pertanto la macchina che effettua I/O sul disco deve essere obbligatoriamente con sistema operativo Windows per poter utilizzare tale applicazione. Poiché la nostra macchina SAS aveva invece sistema operativo Linux non è stato possibile in alcun modo utilizzare l'applicazione Raidtec. Si è pertanto utilizzato tramite un cavo seriale crossed la porta seriale presente sull'apparato per effettuare tutte le configurazioni rinunciando, purtroppo, sia alla semplicità di un'applicazione GUI che a tutta la parte di allarmistica e notifica via e-mail dei guasti.

L'interfaccia seriale si presenta come una serie di menu in modalità testo (vedi Fig. 1) ed è comunque documentata relativamente bene nel manuale incluso con l'apparato. In

generale è possibile suddividere l'array set costruito in più "partizioni logiche" ed assegnare ad ognuna di queste un ID SCSI specifico. Dal lato host i vari ID SCSI corrispondenti a diverse partizioni logiche vengono viste dal sistema operativo Linux direttamente come device disco SCSI differenti ognuno con un diverso ID SCSI (es. /dev/sdc e /dev/sdd). Tale operazione si è resa necessaria in modo tale da creare un unico array set raid-5 e partizionarlo in entità logiche di grandezza minore di 1TByte[1]in modo tale da renderle visibile al sistema operativo Linux Red Hat 7.2 e poterle singolarmente partizionare via sistema operativo, formattarle con il filesystem opportuno (ext3 o altro) e accedervi correttamente.

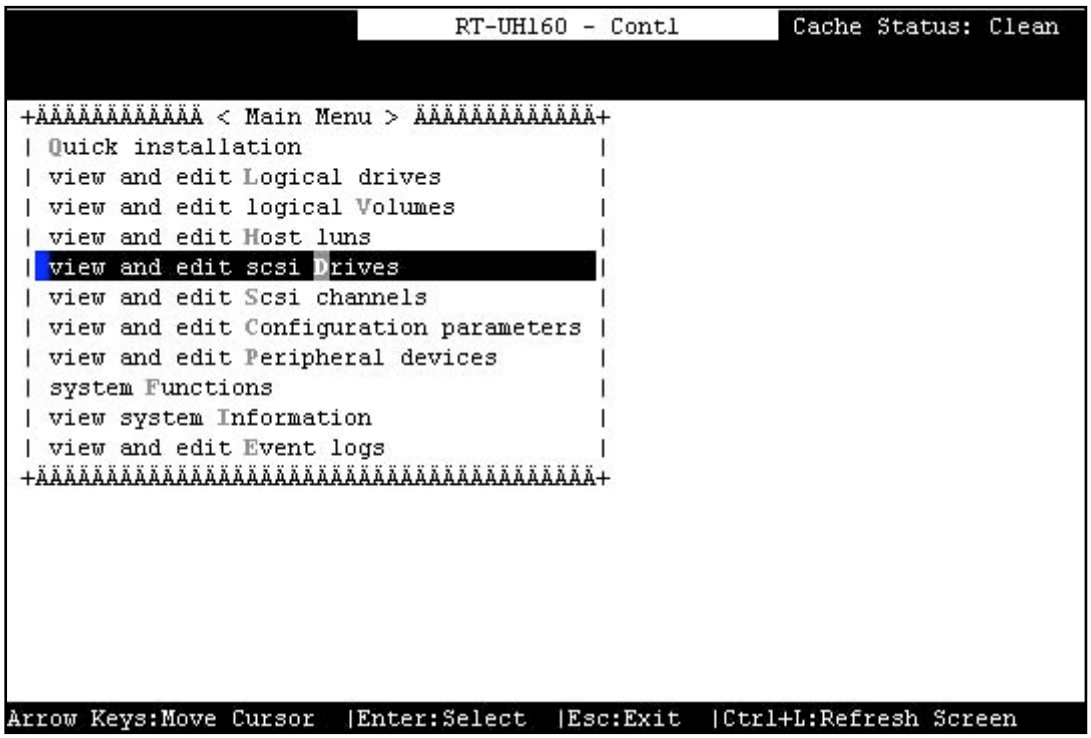


Fig. 1 Il menu principale del Flex Array HI 160 accessibile via seriale.

2.3 Performance

Sono stati effettuati una serie di test di I/O sia locale sia via nfs. L'hardware usato per la connessione diretta era il server Dell 1550 con l'interfaccia SCSI già citata dotato di 512KB di memoria 2 CPU Pentium III Intel a 1GHZ 256K di cache 2 dischi interni SCSI da 18GB e connessione alla rete via interfaccia Ethernet Netgear GB ottica.

I testbench utilizzati sono stati di 2 tipi: semplici script creati che utilizzavano il comando "dd" da /dev/null per scrivere/leggere sul disco e testbench linux come bonnie++. Come grandezza dei file utilizzati per la lettura/scrittura si è usato 3 volte la memoria fisica

[1] Il sistema operativo Linux 7.2/7.3/8.0 ha il limite sia nei driver SCSI sia in parte delle applicazioni di non poter accedere a device SCSI con un numero di blocchi maggiore della rapprentazione binaria a 31bit. Questo si traduce in una grandezza approssimativa di circa 1TByte per device SCSI con block size da 512byte

ovvero 1.5GByte. L'utilizzo di file maggiori non ha mostrato differenze sensibili nei risultati mentre per file più piccoli l'effetto di caching della memoria era rilevante e dava luogo a risultati inverosimilmente alti. Per maggiori dettagli sullo svolgimento dei test e sulla presentazione dei risultati si veda l'Appendice A.

Per la fase di test locali si sono ottenuti risultati, con processi singoli, di circa 40MB/s per operazioni di scrittura e di circa 55MB/s per operazioni di lettura.

Sono poi stati attuati una serie di test del protocollo NFS v.3 (versione standard distribuita con Red Hat Linux 7.2) utilizzando macchine gemelle come client nfs con collegamento alla rete anch'esse via Gigabit Ethernet e ripetendo i test previa verifica che il throughtput ottenibile dai testbench di rete (ttcp e netperf) non costituiva un potenziale collo di bottiglia per i test via nfs. Si è provato anche ad ottimizzare il sistema client/server nfs aumentando i parametri rsize e wsize[3] sul mount del client oltre ad ampliare il numero di demoni nfsd presenti sul server ottenendo i risultati presenti in Tab.1

Numero Client nfs	**Scrittura MB/s (AGGREGATO)**	**Lettura MB/s (AGGREGATO)**
1 Client	24	34
2 Client	25	35
3 Client	25	35

Tab.1 Risultati test via protocollo NFS v.3

Come si può vedere facilmente nelle tabella il limite intrinseco del protocollo NFS v.3 è facilmente raggiunto anche da un singolo client nfs purché connesso via Gigabit Ethernet. Questo con la presente configurazione SAS rappresenta il limite superiore delle performance ottenibili via nfs con questa configurazione hardware. Poiché i risultati del potenziale degrado di performance con un numero elevato (centinaia) di client non interessavano in questa prima fase non sono stati effettuati e comunque già con 3 client è possibile vedere l'effetto del collo di bottiglia intrinseco all'hardware e al protocollo NFS che non scala in maniera adeguata al crescere dei client.

3 APPARATO SAS/SAN DELL POWERVAULT 660F

3.1 Descrizione tecnica, spazio fisico e costi

Il secondo apparato acquistato in ordine temporale successivamente al SAS Raidtec è stato l'apparato Dell Power Vault 660F fornito di doppio controller ridondato Fibre Channel (F.C. in seguito nel testo) ognuno con uscita rame 1GB/s + 7 moduli 224F ognuno con 14 dischi F.C. Seagate Cheetah da 73 GB/10,000 rpm per un totale di 112 dischi per una capacità lorda dichiarata di 8.176 GB. In realtà una volta configurato il sistema in diversi blocchi o volumi logici raid-5 da 16 dischi fisici (il massimo realizzabile con i controller F.C. Mylex Ver: 5776-00 in dotazione) con un solo disco globale di hot spare e formattando le varie parti con filesystem ext3/reiserfs con lo 0% di "reserved-blocks-

percentage" lo spazio effettivamente disponibile al sistema operativo è stato di circa 7.060 GB. Ogni controller inoltre dispone di 512MB di memoria cache mirrorata sull'altro controller.
Il sistema è stato acquistato ed installato nella primavera 2002 ad un costo inclusa l'assistenza on-site per 3 anni di circa 12KEuro al TByte lordo.

3.2 Metodo di controllo e configurazione

Il sistema di management e allarmistica di Dell è il Tool GUI Array Manager fornito con il sistema. Purtroppo l'installazione di tale software è prevista sola su piattaforma Windows e la comunicazione con i controller può avvenire solo embedded su protocollo Fibre Channel di accesso ai dischi. Poiché si è scelto di accedere via F.C. all'apparato solo da server con sistema operativo RedHat Linux versione 7.2 (o successive) si è dovuto rinunciare al sistema di controllo Array Manager. In alternativa il supporto Dell ci ha fornito di un cavo seriale opportunamente adattato per permettere l'accesso via seriale direttamente ai due controller Mylex con il menu in modalità testo proprio dei due controller.

Le limitazioni dell'accesso unico in modalità testo sono notevoli (non è possibile ad esempio la gestione di notifica automatica degli allarmi) tuttavia l'interfaccia presenta gran parte delle features disponibili via Array Manager. In Fig. 2 è riportata una immagine dell'intefaccia in modalità testo.

In generale i 2 controller lavorano in configurazione mirror pertanto le modifiche effettuate su un controller sono riportate istantaneamente anche sull'altro. È possibile il fallimento di un controller, ovviamente il canale F.C. rame fisicamente collegato direttamente al controller fallito non sarà disponibile ma poiché in configurazione mirror tutti i volumi logici sono visibili da entrambi i controller è possibile accedere ai volumi via il controller rimanente. Sono inoltre previste le opzioni di LUN masking per la connessione ad una SAN in modo tale da filtrare l'accesso ai vari volumi logici solo da determinati host identificati dal WWPN (World Wide Port Name, codice univoco di 64 bit assegnato ad ogni porta) della scheda F.C. con cui accedono alla SAN.

3.3 Performance

Sono stati effettuati test di I/O locale diretto via F.C. sia via un singolo controller sia su entrambi (su volumi logici differenti) utilizzando 2 server 1550 come descritti nel Capitolo 1 provvisti di scheda interfaccia F.C. Qlogic 2200 a 1Gb/s. I risultati da un singolo server mostrano performance di 38MB/s in scrittura e di 42MB/s in lettura.

I risultati da 2 server (usando solo i test con dd per effettuare al meglio le operazioni di sincronia tra scrittura e lettura) mostrano un aggregato di 53MB/s in scrittura e 61MB/s in lettura. Pertanto le operazioni di I/O simultanee su volumi diversi dai 2 diversi controller mostrano un aggregato di throughtput migliorativo ma non certo il raddoppio delle performance ottenibili su un singolo controller.

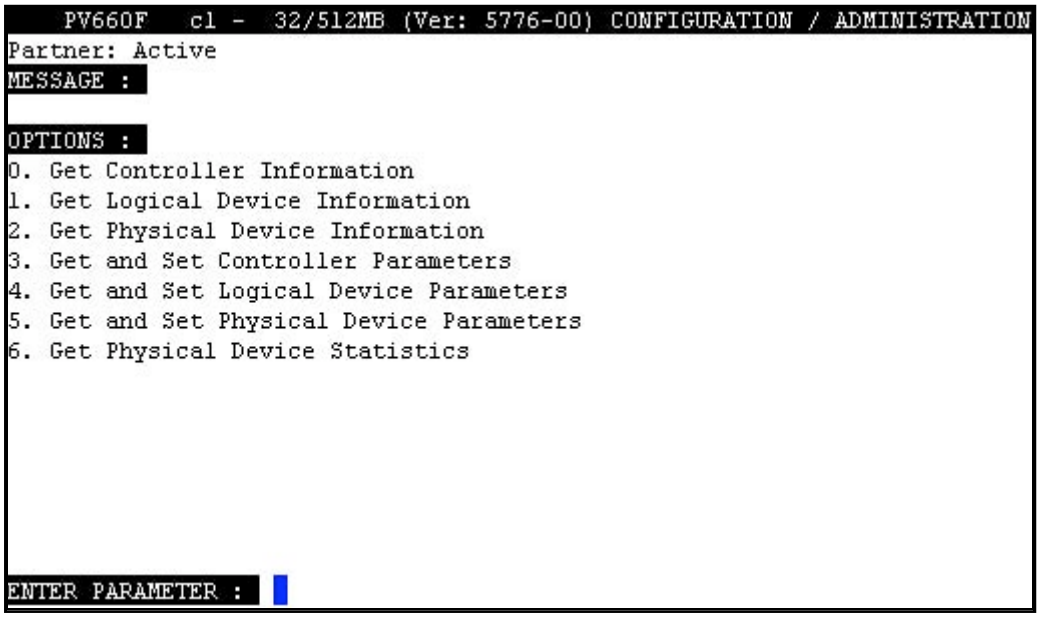


Fig. 2 Una schermata del menu seriale presente nel controller Milex del Dell PowerVault 660F. È possibile notare che si è connessi via seriale al secondo controller (dicitura "c1" in altro) e i due controller sono in configurazione mirror (ovvero il partner risulta "active")

4 APPARATO SAS/SAN AXUS BROWNIE BR-1600FC

4.1 Descrizione tecnica, spazio fisico e costi

L'apparato Brownie è stato acquistato a scopo di valutare le prestazioni e i vantaggi/svantaggi della tecnologia dei controller raid su dischi IDE con uscita sul lato host di tipo F.C. Il controller singolo utilizzato sul box Browie dispone di 2 uscite F.C. rame a 2GB/s (in realtà la seconda uscita è utilizzabile sono per alta affidabilità e non ha effetto alcuno sulle performance) e una CPU Intel i80303 64bit RISC per la generazione della parità dei livelli raid oltre a 256MB di cache. Sul box sono stati montati 16 dischi Maxtor EIDE UltraDMA 100 da 160GB nominali ognuno per uno spazio lordo totale di 2.560GB. Il sistema è stato configurato in modalità raid-5 su 15 dischi mantenendo un disco in hot-spare. Una volta partizionato logicamente l'array in 3 partizioni logiche per permettere al sistema operativo Linux di visualizzarle correttamente si sono formattate in ext3 le varie parti si è ottenuto uno spazio effettivo di circa 2.030GB. Il sistema acquistato nell'inizio autunno 2002 ha avuto un costo comprensivo di assistenza di circa 7KEuro al TByte lordo.

4.2 Metodo di controllo e configurazione

L'accesso per la configurazione al sistema è effettuabile unicamente via seriale tramite il cavo fornito ed è possibile la connessione sia in modalità terminale con menu testuale sia tramite il software GUI Raidcare (vedi Fig. 3) e l'opportuno agente in esecuzione sulla macchina con connessione diretta seriale. Il software e l'agente sono certificati anche per Linux Red Hat versioni 7.2 e successive e tramite Raidcare è possibile configurare l'agente in modo da effettuare notifiche automatiche via e-mail in caso di problemi di tipo hardware oltre a poter effettuare tutte le configurazioni sul controller effettuabili anche dal menu in modalità terminale e a visualizzare un serie di parametri diagnostici aggiuntivi. Inoltre è possibile installare anche la parte cliente GUI di Raidcare su macchine differenti da quelle in cui è in esecuzione l'agente e quindi eseguire la configurazione e l'amministrazione dell'apparato in maniera remota. In generale oltre alle opzioni standard è possibile anche per questo apparato l'operazione di LUN masking per la connessione ad una SAN.

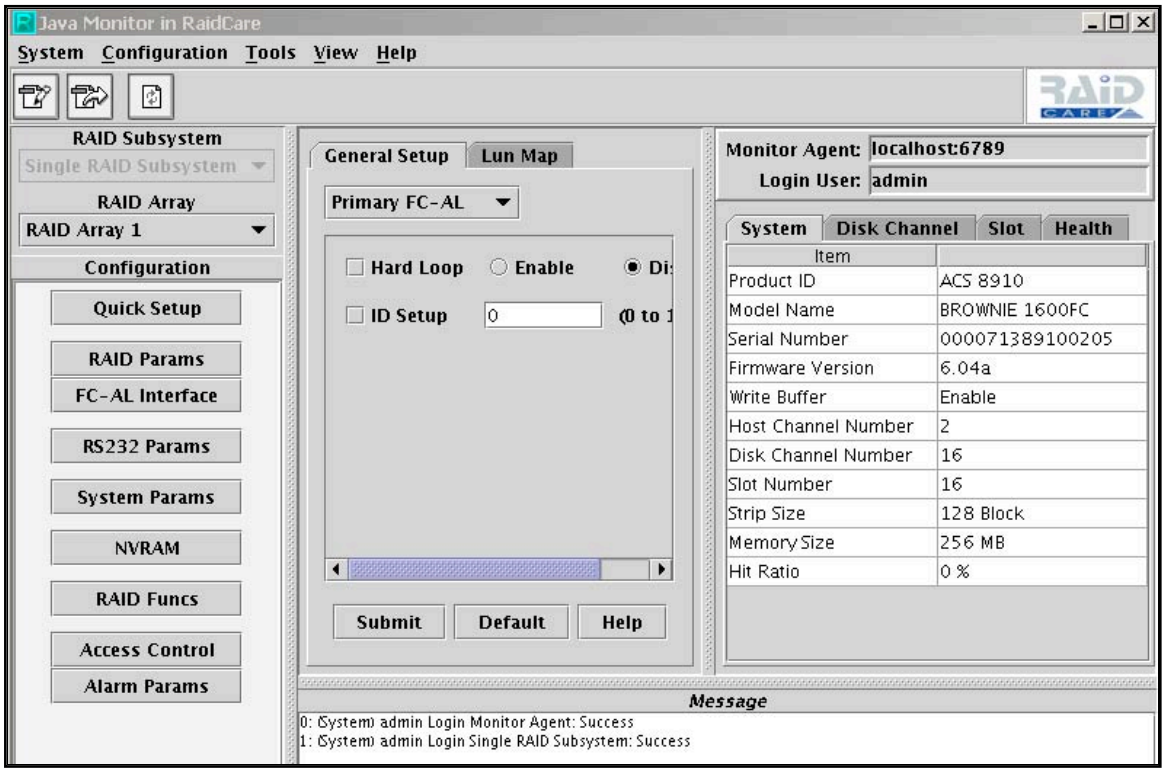


Fig. 3 Il tool GUI RaidCare che permette una semplice e potente gestione del Browie oltre alla configurazione dei settaggi di allarmistica e notifica automatica da parte dell'agente.

4.3 Performance

La fase di test per l'I/O locale sull'apparato è stata effettuata usando una macchina Dell 1550 con scheda interfaccia F.C. LSI Logic da 2GB/s. In realtà test comparativi utilizzando interfaccie diverse come le Qlogic 2200 da 1Gb/s e 2300 da 2Gb/s non hanno mostrato differenze significative sui risultati. I risultati mostrano un valore di 40MB/s per operazioni di scrittura e 45MB/s per operazioni di lettura

5 APPARATO NAS PROCOM 3600FC

5.1 Descrizione tecnica, spazio fisico e costi

L'apparato Procom 3600 FC è un NAS basato su tecnologia nativa F.C. ed e' costituito da 96 Dischi da 180GB per un totale di 17.280 GB lordi collegati a due "teste" (server) tramite due controller FC. Il NAS in questa configurazione si sviluppa in un rack standard 19" da 42 U. Ogni testa e' collegata alla rete con due interfacce Gigabit Ethernet (sono collegabili fino a 4 interfacce di rete per ogni testa).

L'intero spazio disco è stato suddiviso in array raid-5 da massimo 12 dischi (di cui 1 utilizzato per la parità) per le limitazioni intrenseche ai controller Mylex utilizzati dal sistema. Si noti come i controller Mylex sono analoghi a quelli utilizzati dall'apparato Dell Power Vault descritto nel Capitolo 3. La configurazione effettuata con un singolo disco di hot-spare per tutti e 96 dischi e la formattazione via file system proprietario di Procom ha portato ad un totale di 12.462 GB effettivi utilizzabili. Acquistato all'inizio del 2002 e' costato circa 12KEURO al TByte lordo

5.2 Metodo di controllo e configurazione

L'apparato è stato messo in produzione per contenere dati sperimentali ed in questo momento tutto lo spazio disco è stato allocato su due partizioni da 9 e da 4,5 TB servite rispettivamente dalle due teste. La caratteristica principale di questo NAS e' la grande flessibilità nella gestione dello storage in quanto permette di creare volumi logici anche di grandi dimensioni con filesystem superiore ai 2TByte e permette di aumentare le dimensioni dei volumi senza la necessità di effettuare alcun "fermo macchina".

E' possibile realizzare a livello fisico una gestione raid-5 con un unico disco di hot-spare su tutta la catena F.C. permettendo un considerevole risparmio in termini di spazio disco.

Il software proprietario del NAS permette di realizzare una configurazione in alta affidabilità su tutti i componenti. In caso di funzionamento normale le due teste lavorano indipendentemente su volumi di disco differenti rispondendo a due distinti indirizzi IP.

In caso di failure di una delle teste, di un controller o in seguito ad una interruzione della connettività di rete di una delle teste, l'altra ne assume l'indirizzo IP e prende il controllo di tutte le unità rendendo di fatto trasparente agli eventuali client il problema dando cosi il tempo ai gestori di porre rimedio al guasto.

In caso di guasto o di raggiungimento dei valori di quota impostati, il NAS avvisa i gestori tramite un buon sistema di e-mail notification.

La gestione del NAS si può effettuare sia via Web tramite l'applicazione Java NETFORCE Admin (di cui è riprtato uno screenshot di un sottomenu in Fig. 4) sia tramite linea di comando (via telnet). Il NAS esporta i suoi volumi in NFS (v2 e v3) ed in CIFS (appoggiandosi ad un server di Domino per l'autenticazione).

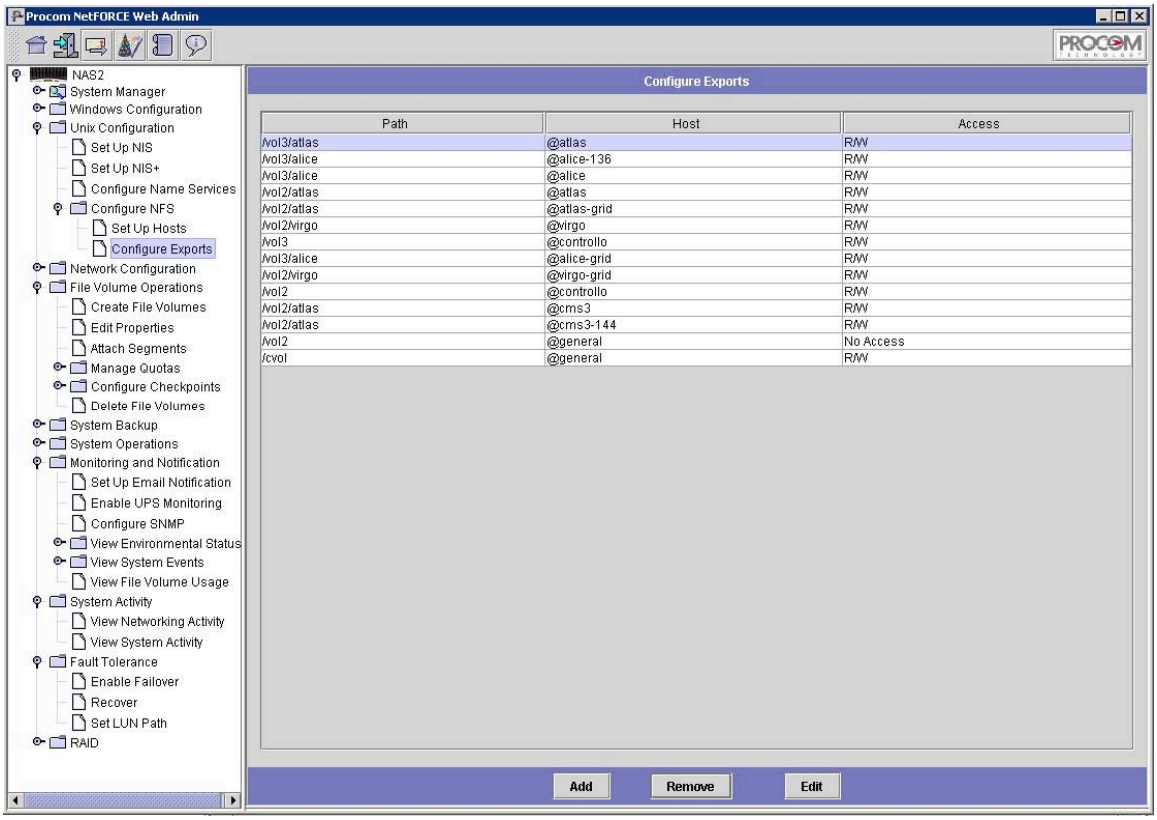


Fig. 4 Dalla home page di gestione del NAS Procom si accede a tutte le funzioni principali tra le quali viene riportata in figura la schermata di controllo dei permessi di accesso.

Teoricamente il NAS può essere espanso collegando alla struttura altri rack di dischi FC, Tuttavia si ritiene che non sia opportuno senza aumentare il numero di teste serventi. Esiste anche una opzione denominata "Duet" che tramite l'inserimento di uno Switch Fibre Channel ed un upgrade del software permette di servire spazio disco anche in modalità SAN collegando anche server di terze parti.

5.3 Performance

I test prestazionali sono stati effettuati solo accedendo via NFS (non è possibile lanciare benchmark direttamente dalle teste dell'apparato) utilizzando macchine analoghe a quelle utilizzate per i test descritti nei capitoli precedenti.

La seguente tabella Tab. 2 riassume i risultati ottenuti.

Numero Client nfs	Scrittura MB/s (AGGREGATO)	Lettura MB/s (AGGREGATO)
1 Client	12	29
2 Client	13	30
3 Client	15	37

Tab. 2 Risultati test NAS PROCOM via protocollo NFS v.3

Dai risultati si può notare che mentre le performance in scrittura risultano piuttosto basse quelle in scrittura si allineano su valori da 30 a 40 MB/s. Inoltre un numero maggiore di client non ha mostrato un aumento delle performance che pertanto risultano già massimizzate con il numero di 3 client gigabit ethernet mostrato in tabella.

6 APPARATO NAS IDE 3WARE

6.1 Descrizione tecnica, spazio fisico e costi

L'apparato NAS è stato assemblato basandosi su di un server Intel dual Xeon 2.2 Ghz in un case 4 Rack unit ed è stato dotato di due controller 3Ware 7850 e 16 dischi da 170 GB (8 installati su di un controller e 8 sull'altro) sostituibili a caldo per un totale di 2.720 GB lordi.

La configurazione realizzata è stata la creazione di array raid-5 con 1 disco di hot spare per ogni gruppo di 8 dischi facenti capo ad ogni controller. In questa configurazione si ottengono due volumi visibili dal sistema operativo con un totale di 1.831 GByte netti. Il sistema acquistato nell'inizio autunno 2002 ha avuto un costo comprensivo di assistenza di circa 5KEuro al TByte lordo.

6.2 Metodo di controllo e configurazione

Il sistema operativo utilizzato dal NAS è Linux (kernel 2.4.18-26.7) e risiede su due dischi dedicati e configurati in raid-1 (Mirroring) ad ulteriore garanzia dell'affidabilità del disco di sistema. Questo però si traduce nel fatto che essendo un NAS basato sul sistema operativo Linux e non su un sistema proprietario esso gode di tutti gli svantaggi e vantaggi propri del sistema operativo stesso. In particolare non e' stato possibile fino ad oggi oltrepassare il limite di 1 TByte per singolo volume ma la piattaforma è aperta e consente di dare accesso ai volumi utilizzando anche protocolli differenti da nfs quali ftp, bbftp e altri ed è possibile il monitoring dei parametri fisici del NAS come CPU o utilizzo della memoria in maniera analoga ad una qualunque altra macchina.

Il software di gestione distribuito con i controller della 3Ware ha una buona interfaccia Web (uno screenshot è riportato in Fig. 5), e permette di gestire e configurare gli array raid in maniera piuttosto semplice e funzionale. E' inoltre presente un sistema di e-mail notification in grado di segnalare eventuali "Fault" sui dischi o problemi ai controller 3Ware.

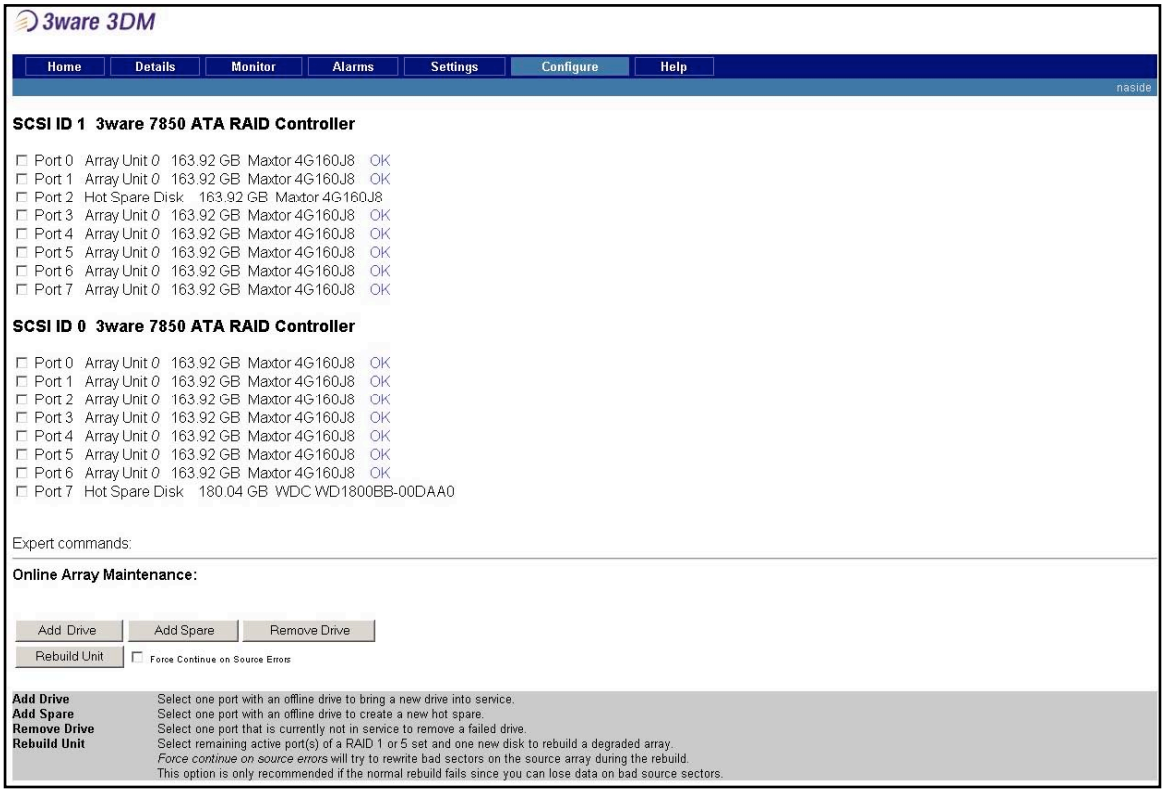


Fig. 5 Pagina di configurazione del controller RAID 3Ware (Interfaccia Web)

Il NAS è collegato alla rete tramite due interfacce Gigabit Ethernet ottiche in Bonding e test preliminari della connessione di rete utilizzando testbench di rete (ttcp e netperf) hanno mostrato che la rete non costituisce sicuramente un collo di bottiglia per il sistema. Le prestazioni misurate in locale si aggirano sui 50 MB/s in scrittura e sui 80-90 MB/s in lettura. Questi valori non sono sempre ripetibili e per risolvere il problema che sembra legato allo specifico apparato è stato aperto, finora senza risultato, un "case" presso la ditta produttrice.

In Tab. 3 sono invece riportati i risultati ottenuti utilizzando 3 macchine client nfs con le caratteristiche hardware riportare nei capitoli precenti. Dalla tabella si può facilmente stabilire come già con 2 client gigabit il NAS raggiunga la saturazione della banda di I/O con i valori di circa 30MB/s per operazioni di scrittura e 40MB/s per operazioni di lettura.

Numero Client nfs	**Scrittura MB/s (AGGREGATO)**	**Lettura MB/s (AGGREGATO)**
1 Client	25	40
2 Client	30	40
3 Client	32	40

Tab. 3 Risultati test NAS 3WARE via protocollo NFS v.3

7 SWITCH FIBRE CHANNEL GADZOOX SLINGSHOT 4218

Pur non essendo un apparato di storage disco si riporta comunque un breve capitolo descrittivo di questo oggetto visto comunque che la definizione stessa di una SAN prevede al minimo la presenza di uno switch. Lo switch Fibre Channel Gadzoox 4218 è stato acquistato completo di 12 Gbic rame e 4 ottici nei primi mesi del 2003 ad un prezzo di circa 13KEuro comprensivo di assistenza. L'apparato è piuttosto compatto (1Unita' rack) con un solo alimentatore ma è prevista la possibilità di connessione con un sistema dual power supply, è dotato di 18 porte autosensing a 2Gb/s adattabili ad uscita ottica LC o rame HSSDC-2 SFP a seconda dei Gbic inseriti ed è compatibile con la maggior parte dei protocolli Fibre Channel compreso l'open fabric protocol (FC-SW-2) per la connessione con SAN dove siano presenti apparati fabric di altri vendor. Le 18 porte sono configurabili come Fabric (F_PORT), Fabric Loop (FL_PORT) o Expansion (E_PORT). Queste ultime sono utilizzabili per le interconnessioni tra switch mentre le prime due sono collegabili rispettivamente a porte node (N_PORT) o node_loop (NL_PORT) in casi di utilizzo di topologie di tipo Arbitrated Loop [4].

Lo switch è configurabile sia via seriale sia via rete IP effettuando da seriale i settaggi di rete opportuni ed utilizzando la porta ethernet integrata. Via rete è possibile accedere allo switch via telnet o semplicemente digitando l'indirizzo IP assegnato tramite un qualunque web browser. Si accede in tal modo all'applicazione JAVA GUI VENTANA SANtools FX. Tale applicazione di cui è visibile una schermata in Fig. 7 è molto semplice da usare e pur non gestendo direttamente l'invio di notifiche o allarmi via e-mail permette la completa visualizzazione dei parametri dello switch aggiornati in tempo reale e la configurazione delle porte e della parte di zoning. Lo zoning si può semplicemente definire come la costruzione di sottoinsiemi di device (schede HBA presenti sulle macchine server di disco , ingressi di array disco o altro) identificati dal loro univoco WWPN (World Wide Port Name) in maniera tale che i membri appartenenti ad una determinata zona vedano solo i device che sono presenti in quella zona. Questo è molto utile per non rendere disponibili e visibili tutti i device "target" (che in generale sono ingressi a array disco) a tutti gli HBA presenti sulla SAN sia al fine di protezione dei dati sia per una migliore e meno complessa gestione della SAN stessa.

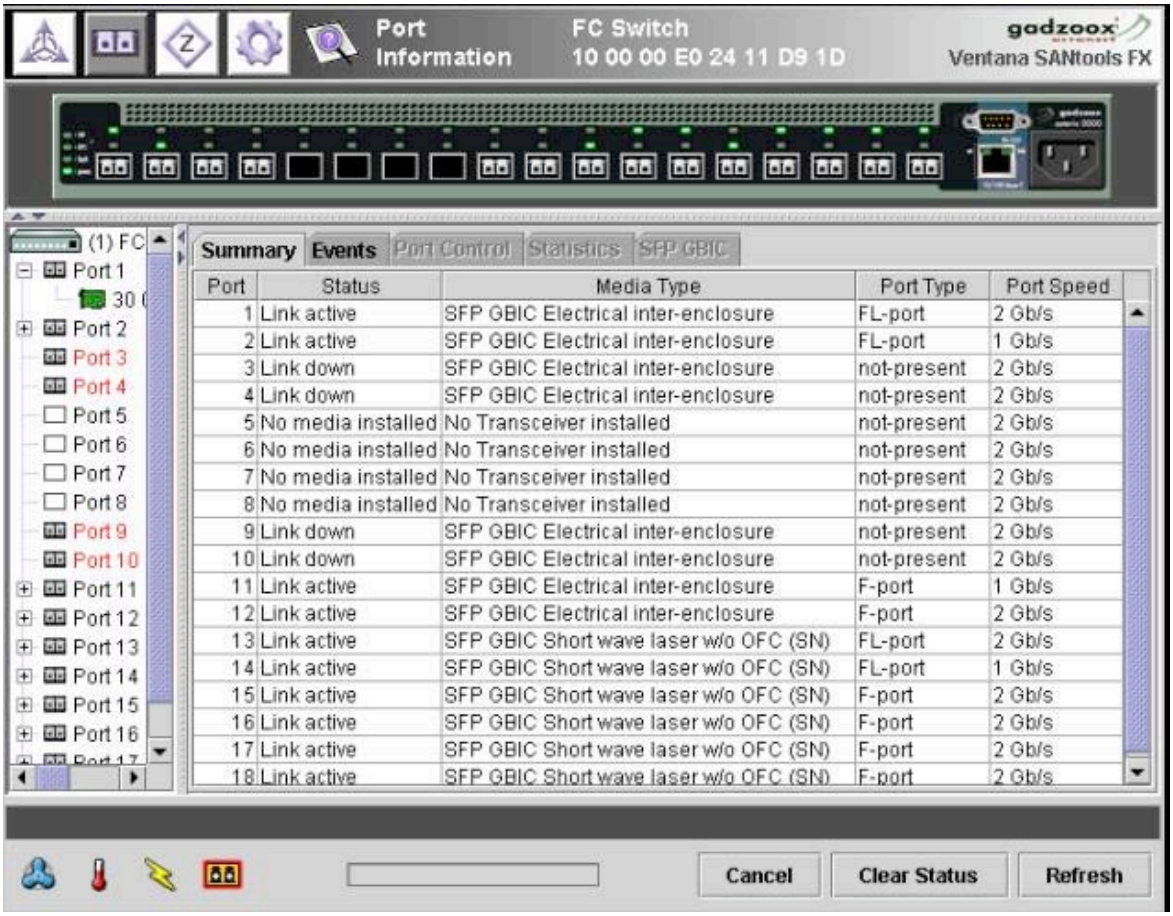


Fig. 7 Uno screenshot della applicazione Java Ventana SANtools accessibile via browser sullo switch Fibre Channel Gadzoox 4218. In particolare in questa finestra si possono vedere gli stati delle 18 porte dello switch.

8 STRUTTURA SISTEMA DI STORAGE DELL'INFN CNAF TIER1

L'accesso all'intero spazio disco presente al TIER1 viene effettuato tramite nfs v.3. In tal modo essendo i nodi di calcolo con sistema operativo Linux l'operazione di mount via nfs è completamente trasparente (nfs è implementato fin dalle prime versioni di Linux) e pur non essendo eccessivamente performante, permette ai nodi di calcolo di lavorare sul filesystem montato via nfs come se si trattasse di un filesystem locale. Poiché come precedentemente descritto i vari apparati attualmente in produzione sono sia di tipologia NAS che SAN l'accesso via nfs agli array disco presenti sulla SAN è stato effettuata tramite delle macchine dedicate al compito di server di disco. Tali macchine sono attualmente delle Dell 1650 con 2 interfaccie ethernet GB Intel on-board e interfaccia HBA F.C. Qlogic 2300 ottica a 2 Gb/s. Il sistema operativo utilizzato è Linux RedHat (versione 8.0 attualmente) e le macchine montano direttamente lo storage presente su SAN e lo riesportano ai client operando come serventi nfs v.3.

Per soddisfare le varie richieste degli esperimenti LHC attualmente presenti al TIER1 si è cercato per quanto possibile di utilizzare singoli oggetti storage per soddisfare le richieste di un singolo esperimento in modo da minimizzare il numero di mount-points nfs a cui i nodi di calcolo client assegnati a quel determinato esperimento devono fare riferimento oltre che evitare il più possibile di suddividere singoli oggetti fisici fra troppi esperimenti. Tale struttura ha inoltre il vantaggio di permettere un più efficiente monitoraggio delle risorse e delle performance dell'I/O effettuato da un singolo esperimento oltre che limitare il disagio ad un singolo esperimento in caso di problemi all'accesso di un singolo oggetto storage fisico. L'accesso tramite nfs agli apparati disco presenti su SAN è stato effettuato utilizzando server di disco (con le caratteristiche hardware citate precedentemente) dedicati ad un singolo esperimento con la denominazione diskserv-<esperimento>.cnaf.infn.it. Sono state create singole zona dello switch Fibre Channel per ogni server includendo gli oggetti storage assegnati a tale esperimento. L'utilizzo dei diskserver ha permesso inoltre di "personalizzare" a seconda delle richieste di ogni esperimento l'accesso al disco includendo altri protocolli oltre nfs come ad esempio bbftp. Tale possibilità non è però presente su apparati NAS come il Procom descritto nei capitoli precedenti poiché il sistema operativo proprietario presente in generale sui NAS supporta solo determinato protocolli (di norma appunto nfs/cifs) limitando le possibilità di accesso diretto allo storage. I 3 NAS messi in produzione al TIER1 sono stati suddivisi tra gli esperimenti ATLAS, ALICE, VIRGO, CDF e LHCB mentre per i restanti esperimenti AMS e CMS sono stati installati 2 diskserver ed è stato utilizzato lo storage presente su SAN.

Pertanto la situazione è riassumibile nella seguente tabella. L'apparato Raidtec descritto precedentemente del Capitolo 2 non compare poiché vista la rigidità del collegamento SCSI SAS si è preferito utilizzarlo come area di frontend disco per l'accesso allo storage nastro tramite il software CASTOR[5].

APPARATO (TByte)	**SERVER NFS**	**TOT (TByte)**	**ASSEGNAZIONE**
Dell PowerVault(SAN)	diskserv-cms diskserv-ams	7.0	AMS 1.0 CMS 6.0
Axus Browie(SAN)	diskserv-cms	2.0	CMS 2.0
Procom head1(NAS)	nas2.cnaf.infn.it	7.0	VIRGO 4.5 ATLAS 2.5
Procom head2(NAS)	nas3.cnaf.infn.it	4.5	ALICE 1.0 ATLAS 3.0
3wareServer(NAS)	nas4.cnaf.infn.it	2.0	CDF 1.0 LHCB1.0

La figura Fig. 7 schematizza la struttura generale NAS/SAN realizzata attualmente al TIER1. E' possibile evidenziare i diversi tipi di connessioni presenti tra i diversi apparati e la Local Area Network interna al TIER1 oltre che le connessioni F.C. tra gli apparati storage e i diskserver. In particolare le connessioni di tipo Fibre Channel (dual) sono utilizzate sia in configurazione di failover attiva-attiva (è il caso dei due controller Mylex in configurazione mirror dell'apparato Dell Power Vault descritto nel Capitolo 3) o attiva-passiva nel caso

dell'AXUS Browie in cui le due uscite F.C. fanno comunque capo allo stesso singolo controller

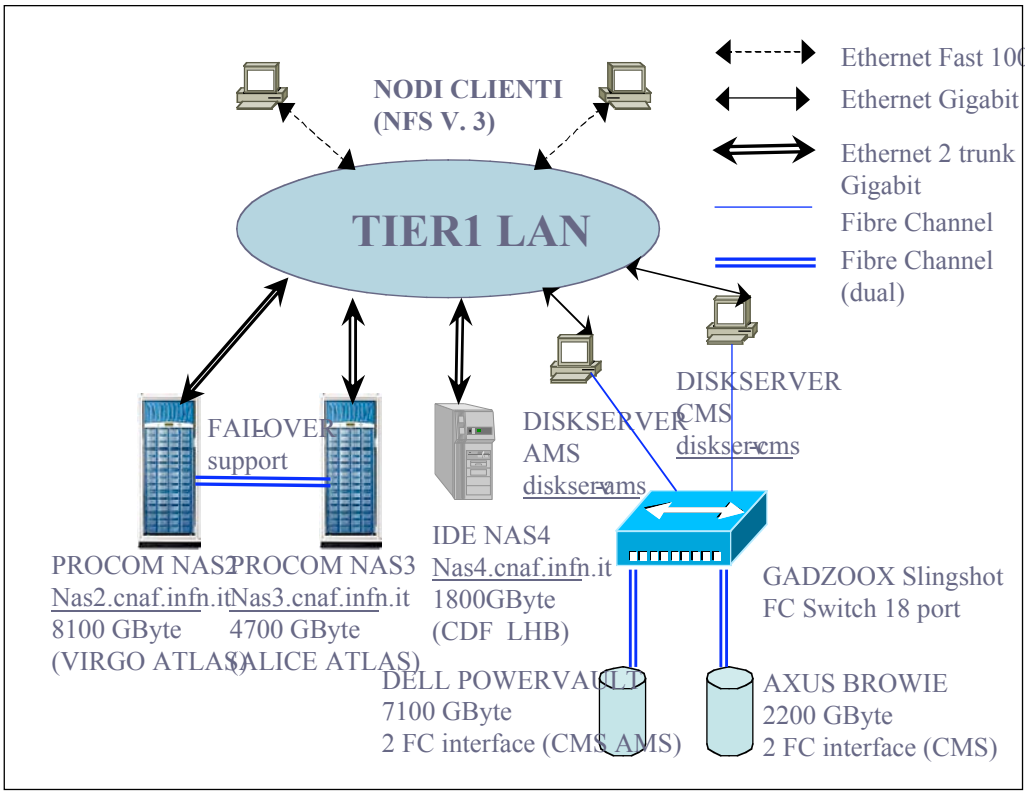


Fig. 7 Schema generale della struttura NAS/SAN realizzata all'INFN TIER1 CNAF

9 CONCLUSIONI

Il presente articolo aveva lo scopo di sottoporre all'attenzione degli interessati i vari apparati di accesso e stoccaggio dati presenti in produzione al TIER1 CNAF. Sono stati presentate in maniera sintetica le caratteristiche degli apparati afferenti alle 3 diverse tipologie SAS,NAS e SAN e sono stati riportati risultati preliminari per dare un'idea delle performance di I/O ottenibili dai diversi oggetti. Si spera con questo di chiarire parzialmente il panorama degli apparati disco proposti dal mercato e di aver fornito una serie di informazioni utili sia al neofita sia all'esperto in procinto di acquistare ed utilizzare apparati simili a quelli descritti.

APPENDICE A METODO DI TEST I/O DISCO EFFETTUATO

In generale la pubblicazione di una serie di risultati di performance è fortemente dipendente da un serie di condizioni al contorno alquanto difficili da riprodurre in un secondo momento. Oltre a questo un altissimo altro numero di fattori più o meno stocastici

entra in gioco nel momento in cui i test sono effettuati in tempi diversi e su macchine non esattamente identiche dal punto di vista hardware e software (es. diverse versioni del S.O.).

In genere quindi nell'effettuare i test si è sempre usato hardware più o meno omogeneo, lo stesso sistema operativo (Linux RedHat 7.2/7.3) e macchine dedicate (ovvero senza altri processi attivi salvo quelli di test) in modo tale da non dover valutare troppo il carico di CPU e l'utilizzo della memoria swap presenti sulle macchine nel momento dell'attuazione dei test. Inoltre per cercare di dare una serie di risultati il più possibile sensati a scapito eventualmente della precisione della misura si sono effettuate unicamente operazioni di lettura e scrittura di tipo sequenziale e si è cercato di differenziale il più possibile i benchmark tool utilizzati non limitandoci all'utilizzo di uno soltanto, fattore che renderebbe la misura fortemente dipendente dalle condizioni al contorno sopradette.

In generale i tool utilizzati sono stati prevalentemente tre:

1) Uno script creato a scopo di test di operazioni sequenziali che effettua operazioni di dd in lettura e scrittura e tramite il comando di time e sync (per svuotare la cache completamente) misurava il tempo necessario per l'intera operazione. In particolare per la scrittura è stato utilizzato dd if=/dev/zero of=<directory_test>/<file_test> " bs=1024k count=6000 per la scrittura di file da 6000 MByte e per la lettura dd if=<directory_test>/<file_test> of=/dev/null . Tramite parametri passabili a linea di comando è possibile specificare il numero di iterazioni (es 10 scritture su file differenti seguite da 10 letture) e il nome dei file di destinazione.

2) Il tool bonnie++ versioni 1.01/1.03 (http://www.coker.com/au/bonnie++) sviluppato da Russel Coker dall'originale bonnie di Tim Bray. La versione C++ (bonnie++ appunto) permette infatti di effettuare scritture e letture con file maggiori di 2GB limite intrinseco al tool nativo bonnie. Inoltre contiene la possibilità di mandare in background più processi (quindi più di un I/O in parallelo) e sincronizzare la partenza dei vari test tra di loro tramite semafori di sistema.

3) Il tool iozone (http://www.iozone.org/) in grado di effettuare test notevolmente sofisticati e ricco di notevoli opzioni come la modalita' throughput con diversi processi in parallelo sincronizzati fra di loro e la presentazione dei risultati in aggregato o la possibilità automatica di ripetere i test in sequenza con variazioni programmate dei parametri.

In generale sia il tool bonnie++ che iozone sono prevalentemente usati per ottenere prestazioni di accesso randomico o con diversi processi in parallelo. Poiché per questa fase di test ci interessavano esclusivamente i risultati di scrittura e lettura sequenziale di file superiori ai 500MB (in generale le operazioni di lettura e scrittura del software degli esperimenti LHC possono essere schematizzate in questo modo) si è proceduto ai test generalmente nella maniera seguente:

- In base alla memoria presente sulla macchina si sono utilizzati per tutti i testbench una grandezza di file pari a 3 volte la memoria fisica della macchina

(quindi 1.5GB per macchine con 512MB o 6GB per macchine con 2GB di RAM)

- Sono state effettuati tutti e 3 i tipi di test e la sequenza dei 3 test è stata ripetuta per almeno 10 volte da ognuna delle macchine client
- Una volta verificato che i risultati per le scritture/letture sequenziali per tutti e tre i test offrivano risultati comparabili a meno di un errore approssimativo del 15/20% è stata effettuata una media aritmetica dei risultati riportando nel testo il risultato mediato in una risoluzione senza cifra decimale
- Qualora uno dei 3 testbench avesse dato risultati fortemente differenti dai restanti quindi non comparabili si è provveduto a ripetere l'intera sequenza fino all'ottenimento di risultati comparabili.

In generale si è notato che i risultati sia in lettura che in scrittura ottenuti con il tool 1) sono sistematicamente più bassi (di un 10% massimo in generale) di quelli ottenuti dei restanti 2 tool che invece mostrano risultati molto più comparabili.

Riassumendo quindi, lo schema delle prove effettuate su dischi locali o volumi montati via nfs da un singolo client seguiva questi punti:

1) Scrittura sequenziale via dd del file <file_n> differente e visualizzazione del throughput di ogni scrittura

2) Esecuzione bonnie++ (con opzione -f per velocizzare l'operazione saltando alcuni test randomici)

3) Esecuzione iozone (con opzioni "-i 0 -i 1" per effettuare solo le letture/scritture sequenziali "-r 2m" per record size da 2MB)

4) Lettura sequenziale via dd del file <file_n> scritto al punto 1 e visualizzazione del throughtput di ogni lettura

5) Ripetizione delle 4 sequenze per 10/20 volte utilizzando ad ogni iterazione nomi di file diversi

Per dischi montati via nfs o comunque acceduti da più client non era semplice la sincronia tra i vari processi in particolare per ciò che riguarda i vari test randomici eseguiti da bonnie++ che non era possibile evitare. Pertanto si è pensato di utilizzare solo gli script dd (che comunque introducevano un errore sistematico tale da dare risultati minori dei restanti quindi introducendo al massimo un errore di sottostima) secondo lo schema seguente al fine di evitare il più possibile l'effetto caching:

1) Scrittura sequenziale via 10/20 dd dei differenti file <file_n> e visualizzazione del throughput di ogni scrittura

2) Scrittura sequenziale via 10/20 dd dei differenti file <file_n> e visualizzazione del throughput di ogni lettura

3) Raccolta e media dei risultati per ogni singolo client

REFERENZE BIBLIOGRAFICHE

[1] Paolo Capiluppi, Federico Ruggieri et al. “Progetto di Massima per un Centro Regionale di Calcolo per l’INFN” Nota Interna INFN CNAF

[2] “Discussione Aperta sulle problematiche di Storage dei Dati” Stefano Zani (Presentazione tenuta al Workshop CCR Isola d’Elba Maggio 2002)

[3] Tavis Barr "LINUX NFS-HOWTO" presente in rete a http://www.tldp.org/guides.html (Linux Documentation Project)

[4] Jon Tate et al. "Designing an IBM Storage Area Network" IBM RedBook Series presso http://www.redbooks.ibm.com/

[5] Ricci Pier Paolo “Utilizzo del Software CASTOR al TIER1 CNAF” Nota Interna INFN Codice TC_03 / 12 presso http://www.lnf.infn.it/sis/preprint