CCR/19/07/P
Luglio 2007
Versione 1.1

Infrastrutture di storage per servizi centrali e calcolo di esperimento: soluzioni e costi

Gruppo storage di CCR

Introduzione

Questo documento riassume i dati sulle caratteristiche delle soluzioni di storage utilizzate nei siti INFN (T1, T2, servizi centrali e volumi di esperimento non LHC) allo scopo di evidenziarne le caratteristiche e di fornire una valutazione indicativa dei costi medi per le diverse soluzioni.
L'esigenza di sinteticita' ha portato alla individuazione di tre diverse tipologie di soluzioni di storage, che non sono le uniche possibili, ma possono soddisfare le diverse esigenze; anche all'interno della stessa soluzione tecnica si deve comunque differenziare in funzione dei diversi requisiti, che possono richiedere alte prestazioni complessive o elevata affidabilita', con conseguente attenzione da porre sulla ridondanza del sistema.
La valutazione dei costi e' stata fatta analizzando gli acquisti operati a fine 2006, ed e' comprensiva di IVA e manutenzione on site nbd della durata di 3 anni, valutabile intorno al 15-18% del costo degli apparati. Le valutazioni che seguono sono state fatte considerando i volumi lordi.

Soluzione low cost: DAS/NAS

Questa soluzione e' costituita da un box all-in-one: un server biprocessore, tipicamente dotato di doppia alimentazione, doppia interfaccia Gigabit Ethernet, e di uno o piu' controller RAID PCI per dischi SATA/SATA2; i dischi sono ospitati in cassetti hot swap contenuti nel box, tipicamente in numero di 24 o 32 unita'.

La scelta della tecnologia SATA e' motivata dalla necessita' di mantenere bassi i costi ed aumentare il volume complessivo disponibile, nei casi in cui si ritenga accettabile il minore livello di prestazioni ed affidabilita' del singolo disco.
I dischi attualmente disponibili su questa soluzione sono di taglio 500 GB o 750 GB: la capacita' complessiva di questa soluzione e' quindi un valore variabile a seconda del prodotto specifico tra i 12 ed i 24 TB lordi .

Le caratteristiche di questa soluzione sono l'economicita' e la semplicita' di installazione e configurazione della singola box; aspetti negativi sono la scarsa affidabilita' degli oggetti (l'esperienza dimostra come la qualita' di questa soluzione sia in generale molto inferiore a quella di altre soluzioni), e la assenza di integrazione a livello di gestione complessiva dello storage, che può diventare un problema nel caso di configurazioni di svariati box, in particolare per acquisti distribuiti nel tempo e quindi tecnologicamente differenti.
L'affidabilita' della soluzione e' limitata alla ridondanza del singolo box, che e' quindi ampiamente consigliabile.

La soluzione si ritiene adatta quindi a situazioni che non richiedano particolare affidabilita' ed omogeneita' di gestione o che dispongano di strumenti di integrazione dello storage a piu' alto livello, tramite interfacce SRM o utilizzo di file system paralleli, configurando opportunamente il sistema di storage in modo da non dipendere criticamente dalla funzionalita' del singolo box.

Il costo medio e' valutabile intorno agli 1.3-1.5 euro/GB, in caso di box completamente configurato, ed e' comprensiva del server di disco che e' integrato.

Benche' non siano disponibili dati al riguardo, e' ipotizzabile una riduzione dei costi (10-15%?) per acquisti di piu' unita' in stock, per volumi dell'ordine del centinaio di TB, che possono essere tipici di un Tier2 a regime.

Soluzione medium cost: Thumper

E' stata recentemente introdotta in commercio da SUN una soluzione NAS di particolare interesse, in quanto accoumuna la qualita' dell'hardware che si puo' trovare nei NAS a tecnologia proprietaria, alla flessibilita' caratteristica degli oggetti descritti al punto precedente, e capace di una notevole quantita' di storage.
Questo oggetto, SUN Fire X4500 o Thumper, e' costituito da una scheda madre che supporta due processori dual core AMD Opteron 285/290 a 2.6/2.8 GHz, fino a 16 GB di RAM DDR-I, dotata di connettori per ospitare fino a 48 HD SATA-II da 500 GB hot swap. L'architettura integra nel backplane le connessioni verso i dischi, che vengono visti attraverso 6 canali SATA ad 8 porte, ciascuno su un canale PCI-X dedicato. Infine sono presenti 4 interfacce GE e 2 slot di espansione PCI-X a 64 bit e 133 MHz.
Non sono presenti cablaggi interni, in modo da favorire la ventilazione, ne' controller RAID hardware, in modo da ridurre le potenziali cause di failure.

La soluzione costituisce un disk server di grande capacita', molto affidabile dal punto di vista hardware, in cui la ridondanza dei dati viene realizzata via software.
L'oggetto e' pensato per l'utilizzo di Solaris ed offre la possibilita' di gestire i volumi tramite ZFS, un prodotto che integra le funzioni RAID (ottimizzate tramite il pacchetto RAID-Z, capace di raid 0, 1, 1+0, 5 e 6) e file system ampiamente flessibile, dinamico e configurabile in ridondanza, prestazioni ed affidabilita'.
E' tuttavia possibile utilizzare il Thumper con OS linux, e configurandone i dischi a piacimento, sfruttando le funzionalita' del raid software per la ridondanza.

Questa soluzione sembra essere molto interessante, in particolare se utilizzata con Solaris, anche se questo ne può limitare il campo di applicabilita': l'esportazione del file system in questo caso deve essere realizzata via NFS.
L'interesse che diversi siti HEP (T1 e T2) hanno manifestato verso tale prodotto ha portato ad una collaborazione tra Sun e dCache ed e' stato realizzato il porting di dCache su Solaris. Questo permette di utilizzare il thumper come pool node di dCache e renderlo cosi' accessibile via interfaccia SRM.
Puo' quindi essere considerata soluzione interessante anche per siti Tier2 che facciano uso di tale tecnologia.

Il prodotto ha una road map che prevede diversi sviluppi gia' per la fine del 2007, sia per l'evoluzione delle tecnologie hardware (bus PCI-express, socket rev-F, RAM DDR-2, dischi da 1 TB, piu' slot di espansione e per la RAM), sia per l'architettura (separazione disco da CPU e conseguente possibilita' di installare expansion box connesse via canali SAS).

Questo oggetto e' stato inizialmente pubblicizzato per offrire 24 TB di storage a 1.5 euro/TB, a cui vanno aggiunte IVA e manutenzione, tuttavia questa valutazione non ha il supporto di offerte commerciali reali.
Esiste pero' una recente campagna commerciale SUN dedicata al mondo educational che offre l'oggetto per la cifra di circa 20 Keuro. A questo costo indicativo si deve aggiungere l'iva (20%) e la manutenzione (~15%), per un totale di circa 1.2 Keuro/TB, non dissimile da quello della soluzione precedente. Questo ne fa una soluzione potenzialmente interessante. Il sistema e' attualmente sotto test in numerosi siti HEP.

Soluzione medium cost: controller Fibre Channel to SATA

Questa soluzione e' costituita da controller RAID con interfaccia Fibre Channel verso l'host, ed interfaccia a conversione di protocollo FC/SATA o SAS/SATA verso dischi di tipo SATA2.
Tipicamente sono disponibili apparati a doppio controller, configurabili in modailta' di failover, spesso inseriti in scatole capaci di contenere 12/14 dischi oltre ai controller. A seconda della soluzione il volume gestito dai controller e' espandibile con box aggiuntivi per soli dischi.
A seconda della complessita' della architettura, può essere necessario l'utilizzo di una infrastruttura SAN con switch Fibre Channel.

Al costo di disco e controller devono quindi essere aggiunti i costi per l'infrastruttura SAN, ove utilizzata, ed i costi per i disk server che devono comunque essere considerati per esportare i volumi verso i client, e devono essere dotati di HBA Fibre Channel opportuna.

Il peso del costo dei disk server può essere valutato in funzione delle prestazioni che si richiedono al sistema. L'esperienza maturata al Tier1 mostra come in caso di volumi utilizzati da job di analisi sia opportuno servire i volumi con un limite di 10-15 TB per disk server, per non sovraccaricare il singolo server che costituirebbe un collo di bottiglia.
In conseguenza di questo va anche valutata l'opportunita' di non eccedere con il volume di storage servito dal controller. Nonostante la potenziale espandibilita' del singolo sistema controller-disco, infatti, oggi la tecnologia SAS ha un limite di 3 Gbps, quella FC di 4 Gbps. Non sembra quindi consigliabile espandere i volumi serviti dal controller, a seconda della tecnologia, oltre i 30-40 TB (SAS) o 40-50 TB (FC).

A fronte di un maggiore costo complessivo rispetto alla precedente, questa soluzione offre caratteristiche che possono essere desiderabili o anche essenziali:

- il management del sistema disco può essere reso omogeneo per i diversi volumi;.
- la separazione volume – disk server offre una notevole flessibilita' sia per le operazioni di riconfigurazione che si possano rendere necessarie al variare delle esigenze, sia per la rapidita' del ripristino di funzionalita' in caso di guasti o di modifiche di configurazione hardware;
- l'utilizzo di una SAN offre la possibilita' di realizzare configurazioni ad alta affidabilita', necessarie per garantire la continuita' di servizio in caso di dati particolarmente importanti (mail, home directory, dati critici di esperimento);
- l'utilizzo di dischi di tipo SATA rende i vantaggi sopra citati fruibili a costi ancora contenuti.

In considerazione di cio' si ritiene questa soluzione idonea in diverse circostanze: grossi volumi (il disco SATA contiene i costi garantendo al contempo scalabilita' e gestibilita' del sistema), elevate prestazioni (si aumenta il throughput limitando la quantita' di storage servito dal controller e dal singolo disk server), flessibilita' (crescita dinamica e modifica di configurazioni senza interruzioni di servizio grazie alla SAN), affidabilita' (resa disponibile dalla tecnologia SAN che permette di configurare cammini multipli tra il server di disco ed il controller).

La valutazione dei costi dei singoli componenti deve differenziarsi necessariamente in base al volume da acquistare.
Due considerazioni sui dati da cui sono state tratte le considerazioni: l'esperienza mostra come l'acquisto di soluzioni con expansion box non completamente configurati fa crescere notevolmente il costo per GB, quindi vengono sconsigliate; i costi si diversificano molto in funzione del brand scelto: si e' ritenuto di non proporre il costo minimo (anche per evitare gravi errori di valutazione legati a offerte speciali natalizie e difficilmente ripetibili) ma non si ritiene appropriata la scelta di un brand particolarmente

in vista qualora i suoi prezzi si discostino eccessivamente dai valori riportati: il livello qualitativo delle anche meno quotate soluzioni pare comunque idoneo alle funzionalita' richieste, e decisamente superiore rispetto alla soluzione del paragrafo precedente.

- Disco e controller: si ritiene fondamentale l'acquisto di oggetti ad alimentazione ridondata, con doppio controller che puo' raddoppiare l'accesso al sistema disco ed in caso di necessita' puo' offrire ridondanza in failover; si suggerisce di considerare sempre la configurazione con la massima disponibilita' di cache disponibile.
  Costo della soluzione per volumi dell'ordine dei 10 TB: 1.7-1.9 euro/GB
  Costo della soluzione per acquisti dell'ordine di decine di TB: 1.4-1.5 euro/GB

- Espansione di disco: il costo di una expansion box da aggiungere ad un controller gia' esistente si puo' valutare dell'ordine del 70% del costo del disco. Questa valutazione viene fatta considerando unicamente espansioni di volumi limitati in dimensione.

- Switch Fibre Channel: si riportano le quotazioni di mercato relative a switch non modulari con porte a 4 Gbps, configurabili in modalita' fabric (quasi tutti ormai); il costo puo' aumentare anche del 20-30% se ci fosse l'esigenza di acquistare software per la attivazione di funzionalita' quali management centralizzato (indicato per configurazioni non banali) o snapshot (importante per realizzare soluzioni di backup out-of-band).
  L'acquisto di uno switch non e' necessariamente indicato per un acquisto iniziale di dimensioni contenute, in quanto i controller possono essere direttamente interfacciati ai disk server, ma diviene indispensabile in fase di ampliamento del sistema storage con aumento della complessita' della architettura.
  Costo di switch FC 16 porte: 6000 euro

- HBA: vengono quotate schede a 4 Gbps a singola interfaccia (ove non sia richiesta ridondanza) e a doppia interfaccia (per soluzioni ridondanti).
  Costo HBA single head: 1100 euro
  Costo HBA dual head: 1800 euro

- Disk server: anche in questo caso vengono proposte due configurazioni differenti: una soluzione ridondata in alimentazione, doppio disco SCSI o SAS per il sistema in RAID1, ed una soluzione non ridondata e con singolo disco SATA per il sistema (indicata quando la ridondanza risieda gia' nella infrastruttura); i disk server considerati sono dual processor opteron; si suggerisce di dotarli di RAM adeguata (almeno 4 GB).
  Costo disk server non ridondato: 2500-2900 euro (in funzione del numero)
  Costo disk server ridondato: 3500-4000 euro (in funzione del numero)

Osservazioni

Va osservato che l'acquisto degli HBA unitamente ai disk server puo' ridurre il costo delle stesse di un fattore valutabile nell'ordine del 20-30%, cosi' come il costo degli switch FC puo' essere analogamente ridotto se acquistati unitamente allo storage.

Nell'ottica di valutare correttamente il costo di una soluzione si devono tenere necessariamente in considerazione i requisiti, ed in funzione di questi le esigenze di spesa. A tale proposito si possono fare i seguenti esempi:

1) Acquisto di 10 TB per un sistema disco che deve ospitare dati critici in assenza di una SAN preesistente: in questo caso non si ritiene necessario l'acquisto di uno switch, ma si richiede un server ridondato per esportare i volumi con scheda dual head:

- Disco + controller: 1.8*10000 GB = 18 Keuro
- Disk server: 4000 euro
- HBA: 1800 euro

Per un costo complessivo di 24000 euro (2.4 euro/GB)

2) Acquisto 80 TB di disco per storage di farm in assenza di SAN preesistente: in questo caso si richiede l'utilizzo di uno switch FC, 8 disk server non ridondati con HBA single head:

- Disco: 1.4 * 80000 GB = 112 Keuro
- 8 disk server: 8*2500 = 20 Keuro
- HBA: 8*1000*70% = 5.6 Keuro
- Switch FC 16 porte: 6000 * 70% = 4.2 Keuro

Per un costo complessivo di 141.8 Keuro (1.78 euro/GB)

3) Espansione della soluzione 1 con acquisto di ulteriori 10 TB

- Espansione disco: 1.8*10000 GB*70% = 12600 euro
- Disk server: 4000 euro
- HBA: 1800 euro

Per un costo di 18.4 Keuro (1.84 euro/GB)

Una valutazione a parte deve essere fatta per acquisti dell'ordine di centinaia di TB. Le uniche esperienze che si hanno oggi nell'INFN riguardano gli acquisti del T1. Questi acquisti devono necessariamente comprendere un sistema completo, dotato di disco, controller in numero idoneo a fornire il throughput necessario, schede HBA, switch FC di dimensioni maggiori di quelli citati. Una valutazione di questi acquisti non puo' che rifarsi alle ultime esperienze, dove un sistema completo di disco da 400 TB, 32 server con HBA e switch FC modulare di fascia elevata a 128 porte, e' costato complessivamente 1.6 euro/GB.

Soluzione high cost: controller Fibre Channel to FC o SAS

Questa soluzione e' costituita da controller RAID con interfaccia Fibre Channel verso l'host, e disco di tipo FC o SAS.

L’unica differenza rispetto alla soluzione precedente riguarda il disco, le cui caratteristiche di affidabilita’ e prestazioni lo rendono idoneo ad ospitare dati caratterizzati da particolari tipologie di accesso (accesso fortemente randomico), di particolare criticita’, ma i cui costi elevati fanno ritenere inopportuno il suo utilizzo per volumi superiori ai pochi TB.
In questo caso si deve attentamente valutare l’esigenza della scelta tecnica, in quanto la previsione di spesa può crescere in modo sensibile in funzione del brand e della tecnologia.
Per questa soluzione di devono considerare costi per GB superiori per un fattore 2-4, legati in parte al costo superiore del singolo disco, in parte alla minore capacita’ dei dischi stessi.
Nel caso si ritenga di adottare questa soluzione per parte del proprio storage si può pensare di acquistare controller capaci di servire contemporaneamente dischi SATA e SAS o FC, in modo da distribuire il costo dei controller sulle due diverse tipologie di disco.
Tutti gli altri parametri vanno considerati equivalenti alla soluzione precedente, tenendo in considerazione che per volumi di grosse dimensioni la soluzione pare impraticabile.

Conclusioni

Si può riassumere quanto riportato nella tabella sottostante.
I costi sono forniti per GB tentando di includere il costo del disk server: va considerato che esigenze di prestazioni nella soluzione FC-to-SATA per volumi di decine di TB, potrebbe richiedere un aumento dei costi in funzione del numero di disk server (uno per ogni 10 TB), che influiscono per 0.2-0.3 euro/GB ciascuno.
Non sono state considerate soluzioni ritenute non idonee, quali soluzioni ridondanti per volumi superiori alle decine di TB, o soluzioni ad alto costo per volumi superiori ai 10 TB. I costi per la soluzione a qualita’ maggiore sono puramente indicativi, e dipendono fortemente dalle scelte specifiche operate.
Non sono stati inclusi i costi per switch Fibre Channel, che devono essere necessariamente valutati in funzione della esigenza, ed il cui “costo per GB” non e’ valutabile ne’ significativo.
Come considerazione finale deve essere sempre tenuta presente quale sia la infrastruttura preesistente in cui si vuole inserire un nuovo acquisto di storage: il fatto di poter mantenere compatibilita’ con architetture in produzione va sempre tenuta in considerazione come fattore che non ha un valore per GB, ma ha certamente un valore significativo in termini di tempi di installaizone e configurazione, e man power per la gestione del sistema.

Tecnologia		Fino a 10-15 TB	Decine di TB	Centinaia di TB
DAS/NAS		1.3-1.5 euro/GB	1.1-1.2 euro/GB	
FC–to-SATA	rid.	2.3-2.5 euro/GB	2.1-2.3 euro/GB	

	non rid.	2.1-2.3 euro/GB	1.8-2.0 euro/GB	1.6 euro/GB
FC – to – FC/SAS (rid.)		4-6 euro/GB (*)		

(*): variabile in funzione della soluzione tecnica