KUNGL TEKNISKA HÖGSKOLAN

Institutionen för numerisk analys och datalogi

TRITA-PDC-2004:1 • ISSN 1401-2731 • ISRN KTH/PDC/R--04/1--SE

Datalagring i en beräkningsmiljö – Terminologi och teknik

Per Ekman, PDC, KTH

Center for Parallel Computers

Per Ekman, PDC, KTH
Datalagring i en beräkningsmiljö – Terminologi och teknik

Rapportnummer: TRITA-PDC-2004:1, ISRN KTH/PDC/R--04/1--SE
Tryckdatum: 2004-03-12
Författarens e-postadress: pek@pdc.kth.se

Institutionen för numerisk analys och datalogi
Kungl Tekniska Högskolan
100 44 STOCKHOLM
Fax: +46 8 247 784

Sammanfattning

Denna rapport syftar till att definiera en terminologi som beskriver egenskaper och användarbehov hos datalagringssystem i en storskalig beräkningsmiljö, samt att ge en översikt av tillgänglig teknik för implementation av sådan lagring.

TRITA-PDC-2004:1 • ISSN 1401-2731 • ISRN KTH/PDC/R--04/1--SE

Innehåll

1 Terminologi

1.1 Lagringens egenskaper

Kraven på ett datalagringssystem är många och varierade. Från en användares synvinkel karaktäriseras systemet av dess prestanda, tillgänglighet och tillförlitlighet samt hur det presenteras. I en driftsituation spelar även underhållskrav samt pris en stor roll.

Prestanda kan vara bandbredd mellan lagringssystemet och användarens program eller latens (time to first byte). Ett program som läser in en hel stor fil är beroende av hög bandbredd medans ett program som läser in tusentals små filer eller många små stycken av en stor fil är beroende av låg latens.

Tillgänglighet kan beskrivas som den proportion av systemets livstid som det är otillgängligt på grund av felfunktion eller administrativa åtgärder. För somliga tillämpningar är antalet tillfällen då systemet är otillgängligt viktigare än den totala längden, 10 korta avbrott på ett år kan vara bättre eller sämre en ett enda långt avbrott under samma period.

Tillförlitlighet är ett mått på med hur stor sannolikhet det data som lagras i systemet kan läsas tillbaka korrekt. Flera komponenter spelar in,

- Finns det säkerhetskopior av data om användaren råkar ta bort sina filer?
- Bibehålls datats integritet vid systemfel?
- Kan systemet under någon omständighet leverera felaktigt data?
- Säkerställer systemet att obehöriga användare inte kan läsa eller skriva filer som inte tillhör dom?

Med *presentation* menas hur lagringen görs åtkomlig för användaren. Exempel på hur lagring kan presenteras är via lokala eller distribuerade filsystem, via ftp, en websida eller genom olika former av programmeringsinterface. Önskvärda egenskaper hos presentationen av ett lagringssystem är att det är lätt att använda, att det har en gemensam namnrymd (dvs att sökvägen till en viss fil är densamma från alla ställen från vilken den är åtkomlig) och att implementationsdetaljer döljs för användaren i så stor utsträckning som möjligt.

Underhållskraven speglar det arbete som måste läggas ner på att hålla systemet i gott skick, att felsöka och reparera vid felfunktion, att bygga ut systemet vid behov och att migrera lagrat data från gammal till ny teknologi.

1.2 Taxonomi

Lagringsbehov kan i stort delas in i tre grova kategorier utgående från kraven på tillgänglighet.

1. **Temporär lagring**

   Temporär lagring är lagring för kortare perioder som inte har några krav på hög tillförlitlighet. Ett typiskt exempel är lokala eller distribuerade scratchfilsystem för data som ett batchjobb behöver under sin körning. Temporär lagring har ofta, men inte alltid, krav på hög prestanda. Kraven på tillgänglighet varierar, en krashad lokal disk på en beräkningsnod påverkar bara jobbet som körde just då på den noden medans ett system som servar flera jobb per användare samtidigt har högre tillgänglighetskrav.

2. **Permanent lagring**

   Permanent lagring är lagring där dataförlust inte kan tolereras. Den kan delas upp i två undergrupper, online och arkiv.

   - **Online**

     Permanent online-data är sådant data som måste finnas lätt och snabbt tillgängligt och vars tillgänglighet och tillförlitlighet är viktigt. Ett exempel är användares hemkataloger där de förvarar källkod, körscript och liknande.

   - **Arkiv**

     Arkivdata är data som skall sparas men som inte läses så ofta. Arkivdata medför ofta höga krav på kapacitet men kan tolerera hög latens och en lägre grad av tillgänglighet.

3. **Säkerhetskopior**

   Säkerhetskopior av data är data med i de flesta fall låga krav på prestanda och tillgänglighet. Eftersom det finns en orginalkopia så är kan det tolereras att säkerhetskopior försvinner eller förstörs. Säkerhetskopior utgörs till den största delen av skrivningar, man behöver förhoppningsvis endast i undantagsfall läsa tillbaka dessa. Prestandan på systemet för säkerhetskopiering kan dock utgöra ett problem om mycket stora datamängder måste läsas tillbaka, återläsning av stora filsystem kan ta flera dygn och medför lång nertid. Vanligt är att säkerhetskopieringssystemen kan betraktas som en implementationshänsyn eftersom användarna sällan interagerar med systemet själva, men det finns undantag.

Med *lagringselement* menas i denna text ett band eller ett filsystem i vilket data kan lagras. Filsystemet kan vara uppbyggt av en eller flera diskar men betraktas som en av andra system oberoende enhet.

2 Tekniköversikt

2.1 RAID

För att kunna erbjuda någon form av säkerhet utöver sannolikheten att en individuell disk eller ett band skall gå sönder krävs ett mått av redundans i systemet. De olika sätt med vilka man kan implementera denna redundans benämns med RAID-nivåer. RAID står för Redundant Array of Inexpensive Disks och definierades ursprungligen i 5 nivåer (1-5) [7], av dessa används bara nivåerna 1, 3, 4 och 5 i praktiken. Man kan även definiera en nivå 0 och på senare tid har även nivå 6 och 7 samt olika hybrid-nivåer lanseras (RAID 10, RAID 51 med flera). RAID 1 används för både disk och band, övriga RAID-nivåer används nästan uteslutande för disk.

- Vid RAID 1 speglas allt data till två olika lagringselement. Av alla RAID-nivåer ger RAID 1 den högsta säkerhetsnivån men den tar också det största lagringsutrymmet i anspråk, det krävs dubbelt så mycket lagringskapacitet som mängden data som skall lagras. Utnyttjandet blir bara 50%. Vid RAID 1 förloras data först när två lagringselement som innehåller samma data går sönder samtidigt. Ju fler lagringselement som systemet innehåller desto mindre blir risken för dataförlust.

- Vid RAID 3 läses och skrives data i block som spänner över en uppsättning diskar, ett RAID-set, där en disk används som paritet. Varje läsning/skrivning går till alla diskar i settet. Om en disk går sönder kan paritetsinformationen användas för att återskapa datat på den trasiga disken, eller om paritetsdisken går sönder så används datat på övriga diskar för att räkna om paritetsinformationen. RAID 3 medför ett bättre utnyttjande av lagringsutrymmet än RAID 1. Om ett RAID 3-set innehåller 7 datadiskar och en paritetsdisk så blir utnyttjandet 7/8 eller 87.5%. Vid RAID 3 förloras data när två diskar i samma RAID-set går sönder samtidigt, detta gör att man inte vill ha alltför många diskar i ett och samma RAID-set. Avvägningen mellan utnyttjandegrad och tillförlitlighet bestämmer alltså hur många diskar ett RAID-set skall innehålla.

- Vid RAID 4 läses och skrives data i block som lagras på individuella diskar. Likt RAID 3 är en disk i settet paritetsdisk, men till skillnad från RAID 3 kan flera små accesser ske parallellt eftersom enstaka läsningar/skrivningar inte låser alla diskar samtidigt. Vid RAID 4 kan paritetsdisken bli en flaskhals eftersom alla skrivningar medför en läsning/skrivning på den. Tillförlitlighet och utnyttjandegrad vid RAID 4 är samma som vid RAID 3.

- RAID 5 är lik RAID 4 men där lagras paritetsinformationen för disk N på disk N+1; det finns alltså ingen dedikerad paritetsdisk. Härvid är risken mindre för att paritetsaccesserna skall bli en flaskhals. Tillförlitlighet och utnyttjandegrad vid RAID 5 är densamma som vid RAID 3.

RAID 3-5 ger en bättre tillförlitlighet än "rå" disklagring. I och med att diskarnas lagringskapacitet blir större ökar dock riskerna för dataförlust. Om en disk i ett system med RAID 3-5 går sönder så saknar systemet redundans under den tid det tar att räkna om paritetsinformationen på en reservdisk. Moderna diskar kan lagra så mycket data att det kan ta tiotals timmar för en RAID att återhämta sig efter ett diskhaveri och systemets tillförlitlighet minskar därmed dramatiskt. Under tiden en RAID räknar om pariteten går prestandan i systemet ner eftersom omräkningsprocessen tar en stor del av den interna bandbredden i anspråk. Eftersom utvecklingen är sådan att lagringskapaciteten ökar betydligt snabbare än överföringshastigheten så kommer detta problem förmodligen att förvärras med tiden. Risken för dataförlust med RAID 3-5 är så hög att man måste betrakta dem som tekniker för att förbättra tillgängligheten snarare än tillförlitligheten. Om permanent lagring krävs måste RAID 3-5 kompletteras med säkerhetskopior.

RAID 1 är betydligt tåligare än övriga RAID nivåer; dels är risken för dataförlust lägre i och med att fler lagringselement kan gå sönder innan detta inträffar, dels tar det kortare tid att ersätta förlorade lagringselement eftersom det kan åstadkommas med en enkel kopiering snarare än en omräkning av paritet. RAID 1 ger i allmänhet tillräcklig tillförlitlighet för att säkerhetskopior inte krävs (dessa finns i någon mening redan), dock måste man se till att systemet är skyddat från yttre fel. Exempel på detta är när användare av misstag tar bort filer som de vill ha kvar, beroende på hur speglingen av datat sker kan detta medföra dataförlust. Andra liknande fel kan inträffa om systemadministratörer gör misstag, om datorintrång inträffar eller vid buggar i programvaror och operativsystem. För att skydda sig mot dessa fel måste de två kopiorna i ett RAID 1 system vara åtkomstmässigt åtskilda. Med bandsystem sker detta ofta automatiskt eftersom datat där lagras på två separata band som måste monteras och läsas/skrivas var för sig. Man är där fortfarande exponerad för fel i programvaran som sköter bandhanteringen. Med disksystem bör de två kopiorna förvaras på två olika maskiner.

2.2 NAS

NAS står för Network Attached Storage. Beroende på vem man frågar brukar det syfta på antingen ett lagringssystem som presenteras av en eller flera servrar för klienter över ett IP-nätverk, eller en integrerad server för ett sådant nätverk. I denna text syftar termen NAS på det förstnämnda. Distribuerade filsystem som NFS och AFS är exempel på NAS-system. Med NAS-system kan man presentera lagringen på ett mycket attraktivt sätt, ett distribuerat filsystem med global namnrymd gör att alla filer i systemet är åtkomliga på samma sätt från alla klientmaskiner. I väldesignade NAS-system skalar kapacitet och prestanda genom att man lägger till fler servermaskiner. NAS-system brukar kräva ganska mycket infrastruktur; förutom att hantera läsningar och skrivningar från klienter måste låsning av filer, information om vilken server en viss fil ligger på (metadata) och säkerhetssystem hanteras. Lås- och metadatahanteringen blir lätt flaskhalsar i systemet eftersom det är tjänster som är svåra att distribuera.

2.3 SAN

SAN står för Storage Area Network och är en teknik för att dela lagringsenheter i ett nätverk. Termen SAN är inte stringent definierad men man brukar oftast mena ett Fibre Channel-nät till vilket lagringsenheter som diskar, RAID-chassin och bandstationer är kopplade tillsammans med de datorer som skall komma åt dem. Varje dator i ett SAN ser alla lagringsenheter i nätet som om de vore direktkopplade, detta gör att data kan läsas på blocknivå med hög prestanda samt att lagringsenheterna kan delas på ett effektivt sätt. I och med att data hanteras på blocknivå kan ett SAN ge betydligt bättre utnyttjande av tillgängling bandbredd hos både disk och nätverk i och med att endast de delar av en fil som verkligen läses eller skrives hämtas från media. I ett NAS hanteras som kontrast oftast data på filbasis, om en klient öppnar en stor fil och läser ett block så läses hela filen ändå in från media via servern till klienten. I SAN-system kan man även tillåta att flera klienter ändrar på olika ställen i en och samma fil samtidigt. I ett SAN kan man med lämpliga RAID-kontrollrar virtualisera lagringen för att få ett flexibelt sätt att hantera systemet på. Man kan investera i en stor diskarray och sedan dela ut olika stora bitar till olika servrar i SANet vid behov, istället för att köpa separata disksystem till varje server. Denna centralisering av lagringsmediat medför dock större krav på redundans och upptid eftersom fler system påverkas vid felfunktion.

SAN-filsystem är distribuerade filsystem som använder ett SAN för att flytta datat till och från lagringsmediat. För att hantera metadata och låsning krävs oftast en separat server för filsystemet med vilken klienterna kommunicerar antingen via SANet eller en separat nätverksförbindelse.

SAN erbjuder hög prestanda och flexibilitet men tekniken är fortfarande inte helt mogen och kompatibilitetsproblem mellan olika tillverkares utrustning är vanliga. Dessutom är SAN en dyr teknologi i och med att host- och switchportar för Fibre Channel är kostsamma.

SAN som inte använder Fibre Channel börjar dyka upp på horisonten i och med att iSCSI-protokollet (SCSI över IP) har standardiserats. Med iSCSI kan ett SAN implementeras över ett vanligt IP-nätverk vilket är betydligt billigare än ett dedicerat Fibre Channel-nät. Detta är dock fortfarande en teknologi som är under utveckling.

2.4 HSM

HSM står för Hierarchical Storage Management och benämner lagringsystem där data lagras i en hierarki av mediatyper. Det vanligaste exemplet är migrerande filsystem där en diskcache med ett filsystem görs tillgängligt för användarna och en HSM-mjukvara ser till att filer i filsystemet migreras från disk till band vid behov (t.ex. när det blir fullt på disken eller när filen inte rörts under en tid). När användaren vill läsa tillbaka en fil som migrerats till band ser systemet till att automatiskt läsa tillbaka filen. I ett väl fungerande HSM-system visar sig det faktum att en fil är migrerad till disk endast genom att det tar lite längre tid att få tillgång till datat när filen öppnats, för användaren bör systemet i övrigt fungera precis som ett vanligt filsystem (dock med mycket stor kapacitet).

Även om HSM-system oftast migrerar från disk till band kan de också implementeras så att de migrerar från högpresterande disksystem till lågpresterande (men billigare) disksystem.

Om ett HSM-system konfigureras så att data migreras till två eller fler mediaenheter (spegling) så kan man klara sig med att endast ta säkerhetskopior av metadatat (cachefilsystemet och eventuella HSM-interna databaser). Metadatat är betydligt mindre i volym än den totala mängden data så det går förhållandevis fort att kopiera och återläsa det. Filer som tas bort ur filsystemet, på grund av olycka eller felfunktion, kan då återläsas så länge de underliggande datablocken finns kvar på mediat de migrerats till, genom att man läser tillbaka metadatat för filen från en säkerhetskopia av metadatat. På detta sätt kan man elegant lösa problemet med att säkerhetskopiering, och framförallt återläsning av säkerhetskopior, vanligtvis tar stora resurser och lång tid i anspråk vid stora datamängder.

2.5 Protokoll

ATA

ATA står för AT Attachment och kallas också ibland lite slarvigt för IDE (Integrated Drive Electronics, tidigare IBM Drive Electronics). ATA är en ANSI-standard definierad av en teknisk kommitte (T13) inom International Committee on Information Technology Standards (INCITS)[4]. ATA är baserat på en uppsättning register i diskkontrollern som tillsammans definierar det kommando som skall utföras. Dessa register adresseras sedan direkt av ATA-kontrollern över ATA-länken. ATA-protokollet kan endast addressera två olika enheter. Den aktuella ATA-standarden [30] innehåller förutom den ursprungliga kommandouppsättningen för ATA även en ej obligatorisk uppsättning kommandon, kallat “Packet Command feature set” (eller ATA Packet Interface, ATAPI), för att skicka SCSI-kommandon över en ATA-kanal.

En ATA-kanal är en 16 bitar bred parallell länk och är definierad för signalfrekvenser på 25MHz (50MB/s), 30MHz (60MB/s), 33MHz (66MB/s), 50MHz (100MB/s), 66MHz (133MB/s), 100MHz (200MB/s) och 133MHz (266MB/s). En ATA-kanal är synkron i det avseendet att ett kommando som skickas på den måste slutföras innan ett nytt kommando kan skickas. Detta betyder att t.ex. två diskar som delar på samma kanal måste vänta på varandra och att nya kommandon inte kan accepteras av en disk innan föregående kommando levererat allt sitt data.

S-ATA

S-ATA står för Serial ATA och är ett seriellt signalleringsprotokoll som emulerar den registeruppsättning som används för kommandoprotokollet i ATA [30]. S-ATA är ännu ej en ANSI-standard utan utvecklas för närvarande av Serial ATA Workgroup som består av ett antal tillverkare (Intel, Seagate, Dell, Vitesse Semiconductor och Maxtor) [1]. I och med emuleringen av ATA kan en S-ATA kontroller se ut som en vanlig ATA-kontroller för operativsystemet. Begränsningen med två enheter per buss kan kringås genom att en S-ATA-kontroller emulerar flera ATA-kontrollrar med en eller två enheter på varje. S-ATA inför dessutom stöd för “hot-plugging”, dvs att disken kan kopplas i och ur under drift.

En S-ATA-kanal är en bitseriell punkt-till-punktlänk som använder differentiell signallering. Fördelen med seriell signallering är att högre frekvenser kan användas samt att kablaget blir enklare och billigare (betydligt färre ledare krävs). S-ATA definieras för en bandbredd på 1.5Gbps (150MB/s effektiv bandbredd) och kabellängder på upp till 1 meter.

S-ATA II är en utbyggnad av S-ATA-specifikationen som bland annat definierar Native Command Queueing (NTQ) där upp till 32 kommandon kan köas till en och samma enhet (i analogi med SCSIs Command Tag Queuing) [31].

SCSI

SCSI står för Small Computer Systems Interface och är ett samlingsnamn för en familj kommunikations- och signalleringsprotokoll. SCSI är en ANSI-standard som definieras av en teknisk kommitte (T10) i INCITS [2].

Kommunikationsprotokolldelen av SCSI definierar en övergripande kommandouppsättning som alla SCSI-enheter måste implementera (SPC) [28], samt specifika kommandouppsättningar för ett antal olika typer av enheter :

- SBC och RBC för blockbaserade datalagringsenheter (t.ex. hårddiskar) [26], [14].
- SSC för strömbaserade datalagringsenheter (bandstationer) [20].
- SMC för mediaväxlare (jukeboxar och bandrobotar) [18].
- MMC för multimediaenheter (CD- och DVD-spelare) [23].
- SCC för kontrollenheter för SCSI (RAID-kontrollrar) [9].
- SES för aktiva komponenter i chassin (nätaggregat, fläktar) [13].
- OSD för objektbaserade datalagringsenheter [24].
- MSC för protokollbryggor [27].
- ADC för kommunikation mellan mediaväxlare och medialäsare [17].

SCSI är ett flexibelt protokoll som används för att kommunicera med en mängd olika sorters utrustning. Fördelar med SCSI är bland annat stöd för asynkrona kommandon och hantering av flera samtidigt utestående kommandon till en och samma enhet (Command Tag Queuing, CTQ), något som krävs för att få god prestanda vid många samtidiga läsningar eller skrivningar.

I likhet med kommunikationsprotokollsdelen av SCSI består signalleringsdelen av en gemensam arkitekturmodell (SCSI Architecture Model, SAM, [25]) och ett antal definitioner av hur SCSI implementeras med olika sorters signalleringsteknologier.

- SPI för parallell SCSI [19].

  Detta är vad man normalt menar med en “SCSI-buss”. SPI definierar en parallell buss som kan vara 8 bitar bred där 8 enheter kan kopplas in på samma buss (narrow SCSI) eller 16 bitar bred där 16 enheter kan kopplas in (wide SCSI). Bussar med överföringskapacitet på 5, 10, 20, 40, 80, 160 och 320MB/s är definierade. Det finns tre olika sorters signallering för parallell SCSI : Single Ended (SE), High Voltage Differential (HVD) och Low Voltage Differential (LVD). Differentiell signallering och högre signalspänning möjliggör större kabellängder. Single Ended SCSI definieras i SPI-4 och High Voltage Differential SCSI definieras i SPI-2. I den senaste versionen av standarden (SPI-5) finns endast Low Voltage Differential kvar.

- SAS för seriell SCSI [21]

  SAS definierar hur SCSI skall implementeras över S-ATA-länkar, samt hur S-ATA-kommandon kan skickas till flera S-ATA-enheter. SAS använder samma kablage som S-ATA men med högre signalnivåer vilket gör att kablar upp till 10 meter kan användas. S-ATA-enheter kan kopplas till en SAS-kontroller, men inte tvärtom varför kontaktdonen för SAS-enheter är utformade så att de inte kan kopplas in på en S-ATA-kontroller. SAS utökar dessutom S-ATA med en möjlighet till två separata kanaler i samma kabel för att tillåta att en enhet kan vara parallellt kopplad till två separata kontrollrar för redundans. SAS är definierat för 1.5Gbps (150MB/s effektiv bandbredd) och 3.0Gbps (300MB/s effektiv bandbredd).

- FCP för SCSI över Fibre Channel [15].

- SBP för SCSI över IEEE 1394 (Firewire) [22].

- SSA för SCSI över Serial Storage Architecure (SSA) [11], [12], [10].

- SRP för Remote DMA (RDMA) [16]

  SRP är primärt avsett för SCSI över Infiniband.

Dessutom finns SCSI över IP (iSCSI) definierat i en standard från Internet Engineering Task Force (IETF) [6], samt SCSI över ATA (ATAPI) [30].

Fibre Channel

Fibre Channel (FC) är ett signalleringsprotokoll för seriell kommunikation med hög prestanda. FC-protokollen utgör en ANSI-standard (ANSI X3T11) definierad av en teknisk kommitte (T11) i INCITS[3].

I botten består Fibre Channel av ett seriellt punkt-till-punktprotokoll som definierats över ett antal olika fysiska länktyper (FC-0). Både fiberoptiska och kopparbaserade länkar stöds (av denna anledning har man valt namnet Fibre Channel istället för Fiber Channel). Ett länklager (FC-1) specificerar hur data kodas och överförs på bitnivå. Ett transportlager (FC-2) definierar ett antal portyper :

- N_Port - Node port, en port i en host eller enhet.
- F_Port - Fabric port, en port i en switch.
- E_Port - Expansion port, en switchport för koppling mellan switchar.
- G_Port - En kombination av en F_Port och en E_Port.
- xL_Port - Loop port, en port som kan användas i en looptopologi. En L_Port är alltid en annan sorts port också, t.ex. en NL_Port för en host-port med stöd för looptopologin.

Med dessa portar kan man sedan bygga system med tre olika topologier :

- Point-to-point

  Två N-portar är kopplade direkt med varandra.

- Arbitrated Loop (FC-AL)

  Ett antal L-portar är kopplade i en ring, antingen fysiskt eller logiskt via en hub. FC-AL är billigare än en Fabric och är ett enkelt sätt att koppla ihop flera FC-enheter med varandra. Endast två noder i ringen kan kommunicera med varandra samtidigt och om en hub inte används går hela ringen ner om en nod försvinner.

- Fabric

  Ett antal N-portar är kopplade i ett nät via F-portar i en eller flera switchar. Med en Fabric är noderna oberoende av varandra och samtidig kommunikation kan ske obehindrat över switcharna vilket gör att prestandan blir högre än vid FC-AL.

FC-0, FC-1 och FC-2 definieras i [29].

Fibre Channel definierar sedan hur olika högre protokoll (som t.ex. SCSI, IP och HIPPI) skall mappas mot underliggande lager (FC-4). Även Gigabit Ethernet (IEEE 802.3z) använder FC-0 och FC-1 som sitt fysiska interface [8]. Fibre Channel finns standardiserat i 1Gbps (125MB/s) och 2Gbps (250MB/s) och i en ej standardiserad variant på 4Gbps (500MB/s).

När man pratar om FC-enheter menar man normalt sett enheter som kommunicerar med SCSI över Fibre Channel.

2.6 Mediateknologi

Det finns idag två dominerande mediatyper för datalagring: magnetband och magnetdisk. Optiska och magneto-optiska lösningar såsom DVD och MO-diskar är endast lämpade för beständig lagring av arkivdata på grund av sin låga prestanda och relativt höga mediakostnad.

Band

Moderna bandteknologier använder ett av två sätt att lagra data på bandet, helical scan eller linear recording. Med helical scan kan högre lagringsdensitet uppnås till priset av fler rörliga delar och högre slitage på banden, system som använder helical scan har därför ofta lägre tillförlitlighet än motsvarande system som använder linear recording.

Lagringskapacitet på band kan anges på olika sätt. Moderna bandstationer har alltid inbyggd kompression och tillverkare anger gärna en teoretisk siffra för hur mycket data som kan lagras givet att datat komprimeras med någon viss effektivitet. Eftersom detta är något av en glädjekalkyl och väldigt beroende av hur datat som lagras ser ut så anges kapacitet här alltid som "rå" kapacitet utan komprimering.

För "enterprise-system", dvs system som kan hantera stora datamängder, finns idag fem stora familjer av bandteknologier : AIT, DLT, LTO, Magstar och 9940. AIT använder helical scan medan de övriga använder linear recording.

- AIT/SAIT

  AIT står för Advanced Intelligent Tape och är en bandteknologi utvecklad av Sony. AITs stora fördel är den höga lagringsdensiteten, 100GB för ett AIT3 band och 500GB för den nya Super-AIT teknologin. AIT har ett något skamfilat rykte när det gäller tillförlitlighet, något som med stor sannolikhet beror på att den använder helical scan.

  Sony har en roadmap för S-AIT som går från dagens 500GB till 1, 2 och slutligen 4 TB/band år 2010, en fördubbling vart annat år.

- DLT/SDLT

  DLT står för Digital Linear Tape och är en teknologi utvecklad av Quantum. DLT har dominerat lagringsmarknaden i många år men har numera utsatts för hård konkurrens från LTO. Den senaste utvecklingen av DLT är Super-DLT som idag är inne på sin tredje generation med band som rymmer 320GB och har en överföringshastighet på upp till 32MB/s. Quantums roadmap för SDLT går via 640GB till 1.2TB per band år 2006, en knapp årlig fördubbling.

- LTO

  LTO står för Linear Tape Open och kallas även Ultrium. LTO är en öppen standard som utvecklats av ett konsortium bestående av HP, IBM och Seagate. Tanken med en öppen standard är att man skall kunna blanda band, bandstationer och bandrobotar från olika tillverkare. Kompatibiliteten mellan olika tillverkares LTO-utrustning har dock inte alltid varit tillräckligt god. Den första generationen av LTO hade en kapacitet på 100GB/band och en överföringshastighet på upp till 40MB/s. Den aktuella generationen kallas LTO2 och tar 200GB/band. LTOs roadmap går vidare via 400GB/band 2005 till 800GB/band 2007 för en fördubbling vart annat år.

- Magstar

  Magstar är en bandteknologi utvecklad av IBM. Tillsammans med StorageTeks 9940 har IBMs Magstar inriktat sig på tillämpningar med höga prestanda- och tillförlitlighetskrav. Magstar har därmed varit ganska dyrt, särskilt bandstationerna kostar mer än i de ovanstående systemen. Magstarbanden har tidigare haft en relativt låg lagringskapacitet men väldigt bra bandbredd och latens. I den senaste generationen (generation 4) av Magstar har dock kapaciteten tagit ett rejält kliv uppåt och ligger nu i prismässig paritet med LTO2 för mediat. Teknologin heter 3592 och har en lagringskapacitet på 300GB/band och en överföringshastighet på upp till 40MB/s. IBMs roadmap för Magstar nämner 500-700GB/band i generation 5 och 900-1100GB/band i generation 6 utan att specificera en tidsplan.

- StorageTek 9940

  StorageTeks 9940 är liksom Magstar en bandteknologi för high-endsystem och är likt Magstar rätt dyr. Även här har fokus legat på korta söktider och hög bandbredd snarare än hög lagringskapacitet. Den senaste 9940 modellen 9940B har en kapacitet på 200GB/band och en överföringshastighet på upp till 30MB/s. StorageTek verkar inte ha någon publik roadmap för 9940.

Disk

Det finns två huvudsakliga marknader som hårddiskar designas för, "enterprise" och "commodity", eller server och desktop om man så vill [5].

- Serverdiskar
  Serverdiskar är konstruerade för att vara igång kontinuerligt med god tillförlitlighet. Prestandafokus ligger på korta söktider och hög genomströmning av läs/skriv-operationer. Tiden det tar att söka fram till en given position på en hårddisk är i hög grad beroende av antalet och storleken på de skivor som bygger upp den samt av skivornas rotationshastighet. För att få ner söktiderna använder man färre skivor med mindre radie och lagringskapaciteten är därför låg för serverdiskar jämfört med desktopdiskar. Rotationshastigheten är också oftast högre på serverdiskar än på samtida desktopdiskar. Serverdiskar är konstruerade för att sitta monterade tätt ihop i större disklådor och därför ställs höga krav på de mekaniska egenskaperna hos disken. Vibrationer kan orsaka stora problem när många diskar är mekaniskt kopplade till varandra. På grund av detta och de högre krav på tillförlitlighet som ställs på serverdisk byggs dessa vanligtvis med andra komponenter än desktopdiskar och blir därför dyrare. För att kunna hantera många utestående transaktioner samtidigt har serverdiskar fram till nu försetts med Fibre Channel- eller SCSI-interface eftersom dessa kan hantera fler utestående transaktioner än ATA. Som en konsekvens av att S-ATA interface blir allt vanligare i RAID-system har idag minst en tillverkare lanserat en serverdisk med S-ATA interface.

- Desktopdiskar
  För desktopdiskar gäller det att få ett så lågt pris och en så hög lagringskapacitet som möjligt. Detta sker till priset av längre söktider i och med att man använder större och ibland även fler skivor i hårddiskarna. De designas för att stängas av eller spinna ner med jämna mellanrum med en "duty cycle" på 8 timmar om dagen, 5 dagar i veckan. Konkurrensen är betydligt större kring desktopdiskar än kring serverdiskar med fler tillverkare och större volymer. Detta medför en stark prispress och nya diskmodeller lanseras med betydligt tätare intervall än för serverdiskar. Ny teknologi dyker också ofta upp först i desktopdiskar medans den flyttas över till serversidan först när den mognat och utvecklats vidare. Desktopdiskar har uteslutade ATA/S-ATA interface i dag.

Hur tillförlitligheten i praktiken skiljer sig mellan server och desktopdiskar är svårt att säga. De siffror som anges av tillverkarna är notoriskt otillförlitliga, men de skiljer sig från en faktor 2 till en faktor 10 mellan server och desktopdiskar från samma tillverkare.

Disktillverkarna har det senaste året börjat visa en svag tendens till att prioritera diskens fysiska storlek framför dess lagringskapacitet. Intervallen mellan lanseringen av nya modeller med större lagringskapacitet börjar bli längre, istället lanseras nya diskar med samma kapacitet som generationen innan men i en mindre formfaktor.

Diskchassin

Diskar kan monteras på en mängd olika sätt: inne i datorer, i enskilda externa lådor eller flera stycken tillsammans i större diskchassin.

Externa diskchassin kan vara av två typer : JBOD och RAID. JBOD står för "Just a Bunch Of Disks" och med det menas ett passivt diskchassi som innehåller ett antal diskar. I ett RAID-chassi sitter en RAID-kontroller som implementerar den RAID-policy som önskas och presenterar flera fysiska diskar som en eller flera virtuell diskenheter utåt. De flesta RAID-chassin kan användas i en JBOD-mod där RAID-funktionaliteten är avslagen.

Viktiga parametrar för valet av chassi är dess mekaniska kvalitet, stöd för hot-plugging av diskar, redundanta kontrollrar och nätaggregat, stöd för hot-plugging av dessa och stabiliteten, funktionaliteten och prestandan hos RAID-kontrollern om sådan önskas.

Hybrid-RAIDar är RAID-chassin där den disk som används har en annan interfacetyp än den som chassit presenterar utåt. Detta kan t.ex. användas för att montera billig S-ATA-disk i ett chassi som presenterar denna som en virtuell SCSI-disk över Fibre Channel eller parallell SCSI.

Mediaväxlare/Robotar

Med flyttbart media såsom band och CD-skivor krävs för all användning i större skala en mediaväxlare. En mediaväxlare är en robot som kan montera och avmontera mediat i läs-skrivenheterna. För enkelhets skull beskrivs här endast bandrobotar, men principen är densamma för andra mediatyper.

Bandrobotar finns i många olika storlekar, från små bordsenheter med plats en handfull band (autoloaders), via medelstora skåp med plats för något hundratal band, till stora, utbyggbara robotsystem med flera monteringsmekanismer och plats för tusentals band (tape libraries).

I allmänhet består en bandrobot av hyllor på vilka mediat lagras, utrymme för bandstationer, samt en "gripper" – en robotarm – som kan plocka media från hyllorna och montera det i bandstationerna. I robotar som bara kan hantera en specifik mediatyp eller formfaktor på media kan grippern vara ganska enkel. Om roboten skall kunna hantera olika mediatyper samtidigt ställs dock högre krav på en flexibel gripper eftersom den måste kunna gripa tag om olika sorters objekt. I system med riktigt avancerade grippers kan man även programmera den till att trycka på knappar och liknande om man har medialäsare i roboten som inte kan fjärrstyras (som t.ex. VHS-spelare). För större robotar finns även en sluss för att man skall kunna stoppa in och ta ut band ur roboten under drift.

Grippern färdas normalt sett antingen längs en rak räls i en s.k. linjär robot, eller längs en cirkelformad räls i en silo-robot. Fördelen med en linjär robot är att den kan expanderas till en relativt låg kostnad genom att rälsen förlängs in i utbyggnadsskåp. Man vill dock undvika alltför långa räls då de kan bli känsliga för vibrationer och ojämnheter i golvet. I vissa linjära robotar lagras endast band på ena sidan om rälsen i grundkonfigurationen och man kan ibland uppgradera dessa genom att bygga om grippern så att den kan vända sig 180° och därmed komma åt band som lagras på båda sidor om rälsen. I en silo-robot kan grippern röra sig betydligt snabbare kring rälsen och monteringstiderna blir därför lägre. För att expandera en silo-robot kopplar man vanligtvis ihop flera silos med portar genom vilka band kan utbytas. Vid utbyggnad med portar kan själva porten bli en flaskhals, å andra sidan har varje modul i ett sådant system sin egen robotarm vilket kan ge bättre prestanda och redundans men medför en högre kostnad jämfört med att förlänga rälsen i en linjär robot.

En kontrolldator krävs för att styra grippern samt för att hålla reda på var de olika banden befinner sig i roboten. För att styra roboten skickas antingen kommandon direkt till kontrolldatorn via ett nätverk, eller så skickar man monteringskommandon till bandstationerna via SCSI (SMC) som sedan vidarebefordrar dessa till kontrolldatorn.

Större robotar kan ibland partitioneras så att olika servrar som är kopplade till samma robot inte kommer åt varandras band eller bandstationer.

3 Teknikval

3.1 ATA, S-ATA eller SCSI?

I och med att ATA använder ett väldigt enkelt protokoll är det okomplicerat och billigt att implementera och används därför ofta för system där lågt pris är viktigare än hög prestanda. S-ATA avhjälper vissa av bristerna med ATA (fler enheter kan styras, stöd för "hot-plugging", Native Command Queuing) och är även det enkelt att implementera vilket gör att S-ATA förmodligen kommer att ersätta ATA på sikt. De förbättringar som införts gör dessutom S-ATA mer lämpat för inbyggnad i RAID-system vilket leder till enklare och billigare system för medelstora lagringsvolymer med låga prestandakrav. SCSI dominerar helt för lagring med större krav på flexibilitet och prestanda. Parallell SCSI kommer antagligen att ersättas av SCSI över Fibre Channel för system med höga prestandakrav och SAS för övriga system framöver.

Där valet av interfaceteknologi kan ha stor betydelse är om kabellängder är en viktig parameter. För enheter som är placerade relativt långt från den kontroller de skall kommunicera med (som exemplevis fallet är med bandstationer i en robot) så dikterar detta att SCSI över Fibre Channel eller HVD-SCSI används. I övriga fall är det normalt sett egenskaperna hos det underliggande mediat som avgör valet av hårdvara, inte primärt vilken interfaceteknologi som används.

3.2 Band eller disk?

Med disk avses antingen klustrade disklösningar med många servermaskiner eller SAN/monolitiska system där all disklagring kan presenteras och hanteras från en maskin. SAN/monolitiska system delar flera av bandrobotens fördelar men är betydligt dyrare än klustrade disklösningar.

Band

\+ Skalar kapacitetsmässigt

Robotsystem skalar idag till över 1 PB lagring utan några större konstigheter och till en väsentligt lägre kostnad än disklösningar av samma storlek.

\+ Lösningen finns och fungerar

Man kan idag gå ut och köpa ett bandbaserat system som kan hantera ett par hundra TB utan problem och till en lägre kostnad än diskbaserade alternativ.

\+ Hög tillförlitlighet

Band, givet att man väljer rätt teknologi, har betydligt högre tillförlitlighet än hårddiskar. De garanteras ofta i minst 5 år jämfört med 1-3 år för diskar och tillverkarna räknar ofta med att de håller i tiotals år. Banden är okänsliga för strömavbrott, komplexiteten sitter i bandstationerna vilket gör att datat är frikopplat från de delar som tenderar att gå sönder.

+ En accesspunkt

  Datat som lagras i en bandrobot kan göras tillgängligt från en enda maskin, detta gör det enkelt att administrera och använda. Detta gäller även om man har en diskbaserad lösning i SAN eller med en monolitisk server.

+ Känd mjukvarulösning

  Mjukvaror för att hantera stora mängder data i bandrobotsystem finns och fungerar relativt bra. Detta gäller även i viss mån för SAN-baserade eller monolitiska disksystem. Hur man skall underhålla, presentera och bygga ut stora diskbaserade datalagringssystem med multipla servrar är en öppen fråga och mer komplicerat än för band om krav ställs på tillförlitlighet och upptid.

+ Låg arbetsinsats för underhåll

  Det mesta i form av övervakning, migrering av data från gammalt till nytt media, spegling av data och liknande underhållssysslor kan automatiseras i de flesta mjukvarusystem för datalagring i robotar och delvis i SAN.

Disk

+ Låg initial kostnad

  För mindre mängder data är disk betydligt billigare än bandbaserade system.

+ Inkrementell utbyggnad

  Man behöver inte köpa på sig mer än man behöver för stunden. Detta gäller i störst utsträckning klustrade disksystem.

+ Skalar bandbreddsmässigt

  Fler servrar i en klustrad disklösning skalar den aggregerade bandbredden. Bandbredden i en bandrobot skalar med antalet bandstationer och antalet servrar, detta gör att ökad bandbredd blir relativt kostsam i en bandbaserad lösning.

+ Större oberoende från tillverkare

  Med klustrade system är man friare att köpa mer kapacitet från olika leverantörer. Med bandrobotar och SAN/monolitiska system är man vanligtvis mer låst till en leverantör.

+ Flera accesspunkter

  Fördelen med att ha flera accesspunkter är att det är mindre risk för att accesspunkterna blir en flaskhals ur prestandasynpunkt samt att systemet kan göras mer tåligt; om en accesspunkt slås ut så kan man potentiellt fortfarande använda de andra.

+ Låg latens

  Data lagrat i en bandrobot tar väldigt lång tid att komma åt jämfört med disklagring. Att montera ett band, spola fram det och börja läsa datat kan ta tiotals sekunder jämfört med tiotals millisekunder med en disklösning.

3.3 NAS eller SAN?

NAS är idag en mognare teknologi än SAN och det finns NAS-lösningar som bevisligen fungerar bra. Förutom kostnaden har SAN skalningsproblem när antalet klienter blir stort (som t.ex. i ett kluster). Prestandamässigt är SAN klart att föredra och det kan även ge underhållsmässiga fördelar. I synnerhet när det gäller säkerhetskopiering kan ett SAN ge fördelar i och med att säkerhetskopior kan tas utan att de servrar som presenterar lagringen behöver vara inblandade. En blandning av SAN och NAS där ett antal servrar delar på disk- och bandlagring via ett SAN men presenterar det till klienterna som ett NAS kan vara en bra kompromiss.

4 Designparametrar

När ett lagringssystem skall implementeras finns det som antagligen framgått en mängd parametrar att ta hänsyn till.

- Datamängd

  I vilken takt sker tillväxten i datavolym? Om systemet skall lagra en fix mängd data som kommer in under en kort period kan det konstrueras för att hantera just den specifika mängden redan från början och behöver inte kunna expanderas. I det motsatta fallet att datamängden ökar stadigt med tiden måste systemet vara skalbart för att man skall kunna bygga ut det vid behov snarare än att behöva avveckla det och bygga ett helt nytt. Mängden data bestämmer också vilken mediateknologi som är mest ekonomisk, för små datamängder är disklösningar billiga och enkla att implementera. För lite större datamängder behövs disk i RAID eller JBOD konfiguration och priserna går upp. För stora datamängder är band ekonomiskt fördelaktigt.

  Mediakostnaden för disk och band närmar sig varandra men för de närmaste åren kommer band fortfarande att vara billigare för stora datamängder. En robotlösning har en hög ingångskostnad i och med att robotik och bandstationer är dyra. När den initiala kostnaden tagits ökar kostnaden per lagrad GB endast med mediekostnaden upp till dess att roboten måste byggas ut eller bandstationer måste uppgraderas, då kostnaden tar ett nytt språng. Framförallt robotiken och skåpen har en betydligt längre livslängd än disksystem och det blir mer av en engångskostnad. Kostnaden för disksystem skalar på ett mer linjärt sätt med små språng varje gång en ny server eller disklåda måste införskaffas. För datamängder under 100-150TB är disklösningar baserade på desktopdiskar idag billigare än en robotlösning.

- Lagringstid

  För data med mycket kort livslängd (timmar eller dagar) kan lagringssystemet kanske klara sig utan säkerhetskopior. Om datat behöver lagras under längre tid behövs någon form av säkerhetskopiering, i form av spegling (RAID-1) eller kopiering till annat media. Högre RAID-nivåer än RAID-1 medför alltför stor risk för dataförlust för att kunna betraktas som lösningar på problemet med säkerhetskopior [32]. Om data ska lagras under många år måste migrering till nytt media beaktas. En tumregel är att disksystem har en effektiv livslängd på 2-3 år beroende på teknologival och bandsystem lever i 5 år. Efter detta börjar kostnader för media och support att rasa i höjden i förhållande till nyare teknologier. Detta gör att man bör planera för att modernisera teknologin och flytta data som skall lagras under längre tid från gamla till nya system. Särskilt system där media lagras off-line, dvs utanför diskchassin eller bandrobot, är risken stor att mediat glöms bort och blir svårt att läsa tillbaka om man inte tänker speciellt på detta. Det är långt ifrån säkert att man kan läsa ett 10-15 år gammalt band ens om man kan hitta en bandstation som passar.

- Användningsmönster

  För arkivdata bör systemet optimeras för skrivning av data eftersom denna operation dominerar. Latensen är mindre viktig och bandbredden vid skrivning är den viktigaste parametern. För system där stora datafiler skall lagras och hela eller stora delar av filen läses tillbaka vid åtkomst blir även bandbredden vid läsning viktig. Om många små filer lagras, eller om filer läses i små bitar i taget börjar latensen spela in. I synnerhet med band kan slumpvisa läsningar medföra stora prestandaproblem då varje läsning kan gå till olika band och bandmonteringar tar många sekunder att genomföra. Hybrid-RAIDar med ATA-disk kan också medföra problem i och med att stödet för parallella operationer är dåligt. För batchjobb är hög latens endast ett problem om filåtkomst är en mycket vanlig operation, annars är låg latens viktigare vid interaktiv användning.

  Presentationen av datat beror också till stor del av användningsmönstret, och kan i viss mån även påverka det. Ju lättare ett lagringssystem är att använda desto fler använder det och desto fler av användarnas lagringsbehov blir därför tillgodosedda av systemet.

  För applikationsspecifik lagring kan presentationen bestämmas av andra parametrar än hur lättanvänd den är. Det är i sådana fall oftast möjligt att använda speciella åtkomstmetoder som t.ex. specialiserade verktyg, programmeringsgränssnitt eller databaser.

  Band är till sin natur ett höglatensmedium och är endast lämpat för data med låga krav på snabb tillgänglighet. Data som läses sällan, t.ex. säkerhetskopior och arkivdata, är väl lämpat för lagring på band. Latensproblemet för band kan delvis avhjälpas med disk-cachning där ett subset av datat mellanlagras på disk (HSM-system). Om ett dataset läses flera gånger och får plats på diskcachen så är en sådan lösning en bra kompromiss som ger både den höga lagringskapaciteten hos band och den låga latensen hos disk. Om den totala mängden lagrat data inte är alltför hög kan en ren disklösning göra systemet billigare och mindre komplext, finns krav på hög tillgänglighet måste dock säkerhetskopior av datat hanteras på något annat sätt. Väldigt stora dataset som inte får plats i diskcachen och som läses regelbundet lämpar sig dåligt för band men kräver å andra sidan väldigt stora mängder disk. I vissa fall kan dessa diskmängder direkt användas som en diskcache framför bandlagring, men kan ställa så höga krav på maskinen som hanterar lagringen att lösningen blir oekonomisk. I de fall datasettet kan distribueras över flera maskiner blir en klusterbaserad disklösning ett attraktivt alternativ.

I frågan om hur klustrade disksystem skall hanteras råder viss osäkerhet. För temporär lagring är frågorna endast hur datat presenteras för användarna och vilken prestanda man vill uppnå. Ett delat parallellt filsystem är att föredra men dessa är ännu av tveksam kvalitet. Om datat presenteras som ett antal separata system åtkomliga via t.ex. ftp så måste antingen någon form av databas byggas som håller reda på vilka filer som ligger på vilket system, eller så ökar bördan för användarna som själva måste veta vilken server en viss fil ligger på.

För permanent lagring tillkommer problemet med hur tillräcklig säkerhet och tillförlitlighet uppnås. Antingen måste datat dirigeras till två olika disksystem vid skrivning eller så måste någon form av säkerhetskopia tas regelbundet.

- Tillförlitlighetskrav

  Kraven på tillförlitlighet kan delas in olika nivåer. Oersättligt data som är mycket värdefullt har högst tillförlitlighetskrav. Exempel på sådant data kan vara instrumentresultat från experiment som ej går att upprepa, satellitdata eller liknande. För att åstadkomma den högsta nivån av tillförlitlighet krävs att data lagras i flera oberoende system som är placerade i olika byggnader eller till och med i olika delar av landet för att säkerställa att datat överlever vid brand eller katastrofer av olika slag. Denna nivå på tillförlitlighet kan dubbla kostnaden för lagringssystemet. Den vanligaste nivån är att datat "rimligen" skall finnas kvar med bibehållen integritet under en begränsad eller obegränsad tidsrymd. Datat är inte oersättligt eller tillräckligt värdefullt för kosta på en katastroftolerant lösning, säkerhetskopior och spegling av data ger tillräckligt hög tillförlitlighet. Om inga krav på tillgänglighet ställs är datat av "scratch"-typ, temporära filer – inga säkerhetskopior krävs och datat kan försvinna när som helst.

  En aspekt av tillförlitlighet som är gemensam för dessa nivåer är kravet på att felaktigt data inte skall kunna levereras utan att det upptäcks, samt att en rimlig säkerhetsnivå bibehålls. De specifika kraven anger om data som skickas över öppna nätverk måste krypteras eller inte.

De aktuella bandteknologierna har en märkbart högre tillförlitlighet än många disklösningar beroende på att själva banden går sönder mer sällan. Det finns high-end-disklösningar som ger minst lika god tillförlitlighet som band men dessa är betydligt dyrare än mostvarande bandlösningar. Samma strategi för redundans ger i allmänhet en lägre tillförlitlighet i ett disksystem än i ett bandsystem på grund av skillnaderna i det underliggande mediats stabilitet.

För temporär lagring är disksystem klart att föredra då tillförlitligheten inte är så viktig och en hög prestanda ofta eftersträvas.

- Tillgänglighetskrav

  Tillgänghetskraven varierar också. Ett system som tar emot data i en kontinuerlig ström från ett experiment som inte kan avbrytas har väldigt höga krav på tillgänglighet eftersom data förloras om lagringsystemet är otillgängligt. I sådana fall blir det viktigt att systemet kan underhållas i drift, att det är designat med redundans och att det finns testade procedurer för att snabbt åtgärda fel. För interaktiv användning är tillgänglighetskraven höga, men inte kritiska. Det går ofta att finna perioder då systemet kan tas ned för underhåll med minimal påverkan för användarna, t.ex. under helger eller på nätter. Ju fler användare som är beroende av systemet desto högre blir tillgänglighetskraven. Exempel på system med låga krav på tillgänglighet är nodlokala filsystem för temporär lagring i ett kluster. Om en disk går sönder så drabbar det endast den beräkning som pågår på den noden och noden måste tas ner för underhåll men klustret i stort påverkas inte.

  Om mängden lagrat data är stor börjar tidsåtgången för säkerhetskopiering och framförallt återläsning av kopior att bli ett stort problem. Filsystem med flera TB data kan ta många dygn att återskapa. I vilken mån man kan tolerera detta har betydelse för RAID-nivåer och andra implementationsparametrar.

Referenser

[1] Serial ATA workgroup. `http://www.serialata.org`. Valid March 8, 2004.

[2] T10 technical commitee. `http://www.t10.org`. Valid March 8, 2004.

[3] T11 technical commitee. `http://www.t11.org`. Valid March 8, 2004.

[4] T13 technical commitee. `http://www.t13.org`. Valid March 8, 2004.

[5] Dave Anderson, Jim Dykes, and Erik Riedel. More than an interface – SCSI vs. ATA. In *Proceedings of the 2nd Annual Conference on File and Storage Technology (FAST)*, pages 245–257, 2003.

[6] Marjorie Krueger, Randy Haagens, Costa Sapuntzakis, and Mark Bakke. Small Computer Systems Interface protocol over the Internet (iSCSI) requirements and design considerations, 2002.

[7] David A. Patterson, Garth Gibson, and Randy H. Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of the 1988 ACM SIGMOD international conference on Management of data*, pages 109–116. ACM Press, 1988.

[8] IEEE Computer Society. IEEE Std 802.3, 2002.

[9] T10. SCSI controller commands – 2 (SCC–2), September 1997.

[10] T10. Serial storage architecture – physical layer 2 (SSA–PH2), April 1997.

[11] T10. Serial storage architecture – SCSI–3 protocol (SSA–S3P), April 1997.

[12] T10. Serial storage architecture – transport layer 2 (SSA–TL2), April 1997.

[13] T10. SCSI enclosure services (SES), January 1998.

[14] T10. Reduced block command set (RBC), August 1999.

[15] T10. Fibre channel protocol for SCSI, second version (FCP–2), September 2002.

[16] T10. SCSI RDMA protocol (SRP), July 2002.

[17] T10. Automation/drive interface – commands (ADC), August 2003.

[18] T10. SCSI media changer commands – 2 (SMC–2), November 2003.

[19] T10. SCSI parallel interface – 5 (SPI–5), January 2003.

[20] T10. SCSI stream commands – 2 (SSC–2), July 2003.

[21] T10. Serial attached SCSI (SAS), July 2003.

[22] T10. Serial bus protocol 3 (SBP–3), September 2003.

[23] T10. MultiMedia command set – 4 (MMC–4), February 2004.

[24] T10. Object-Based Storage Device commands (OSD), February 2004.

[25] T10. SCSI architecture model – 3 (SAM–3), January 2004.

[26] T10. SCSI block commands – 2 (SBC–2), January 2004.

[27] T10. SCSI management server commands (MSC), 2004.

[28] T10. SCSI primary commands – 3 (SPC–3), January 2004.

[29] T11. Fibre Channel physical and signaling interface (FC–PH), June 1994.

[30] T13. AT attachment with packet interface – 7 (ATA/ATAPI–7 v1-3), December 2003.

[31] Serial ATA Workgroup. Serial ATA II: Extensions to serial ATA 1.0a, October 2003.

[32] Qin Xin, Ethan L. Miller, Thomas Schwarz, Darrell D. E. Long, Scott A. Brandt, and Witold Litwin. Reliability mechanisms for very large storage systems. In *Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies*, pages 146–156. IEEE Computer Society, 2003.