

SAN booting

Booten zonder lokale disks

Blade servers bieden een manier voor organisaties met grote aantallen servers om de kosten terug te dringen. Alhoewel sommige blade servers wel met een intern disk-systeem zijn uitgerust hebben ze over het algemeen een kleine opslagcapaciteit. Toepassing van zogenaamde 'diskless' blade servers in combinatie met een opslagnetwerk, maken toepassing van Blade server technologie pas kosteneffectief. NetOpus onderzocht de voor- en nadelen van SAN Booting met diskless blade servers.

BRAM DONS

Diskless blade servers vormde, tot de komst van Windows Server 2003, voor alle voorgaande Windows operating systems een nieuwe technische uitdaging. Met Windows NT 4.0 was remote boot wel mogelijk maar door de daaraan verbonden beveiligingsrisico's kon het door Microsoft nooit grootschalig worden ondersteund. Met de release van Windows Server 2003 is op het Windows-platform nu booten vanaf een opslagnetwerk (zowel Fibre Channel als IP) mogelijk, zodat lokaal geen disk meer nodig is. Let wel: omdat booten vanaf een storage area network (SAN) in principe afhankelijk is van de hardware-configuratie en eigenschappen daarvan, wordt het booten vanaf SAN

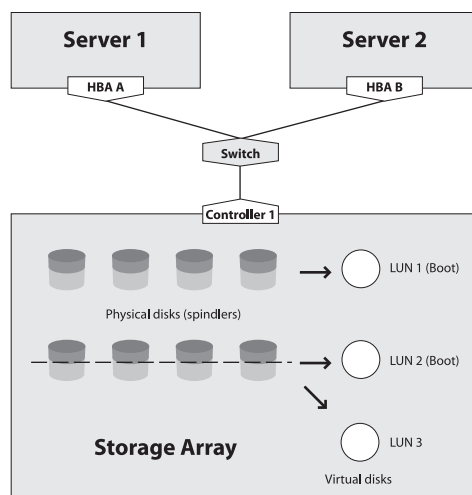
alleen door de hardware-leveranciers ondersteund en niet door Microsoft zelf. Het booten van SAN, net zoals trouwens dat van een LAN-gebaseerde remote boot, biedt het voordeel van een reductie van de hardware-kosten, betere beveiliging en prestaties. Maar, zoals we hierna zullen zien, zijn er ook enkele nadelen aan verbonden.

Het Windows boot proces

Booten vanaf een SAN is een remote boot technologie, waarbij de boot disk zich op het SAN bevindt en niet op het LAN. Bij iSCSI bevindt de disk zich weliswaar ook op het LAN, maar kenmerkend verschil met remote LAN booting is het feit dat bij SAN op basis van block I/O vanaf een disk-device wordt geboot.

De server communiceert met het SAN via de Host Bus Adapter (HBA). Dat kan een FC- of iSCSI-gebaseerde HBA zijn. In het BIOS van de HBA bevinden zich de instructies waarmee de server de boot disk op het SAN kan vinden.

Het boot proces, ook wel 'booting' of 'bootstrapping' genoemd, is het iteratieve proces van het laden van de OS-code vanaf het opslag-device in het lokale computergeheugen. Daarbij bestuurt het BIOS bij Intel-gebaseerde servers het boot proces. Bij het opstarten van de server begint het BIOS de daarvoor benodigde basiscode uit te voeren. Deze zorgt voor de initialisatie van de hardware en haalt de code van de disk op die nodig is voor de volgende fase van het boot proces. De BIOS leest de eerste fysieke sector op de schijf, de zogenaamde Master Boot Sector, en laadt daarvan een image in het lokale servergeheugen. Daarna draagt de BIOS de uitvoering van het boot proces over aan de zojuist geladen image. De Master Boot Record bevat de partitietabel en een kleine hoeveelheid executable code. Deze code onderzoekt de tabel en kijkt welke partitie er actief is (of bootable). Vervolgens draagt de Master Boot Record de besturing aan de Boot Sector image over. De Boot sector is verantwoordelijk voor het vinden van de NTLDR, die het boot proces voltooit. De enige disk services die de Boot Sector code tijdens dit boot proces ter beschikking staan is de BIOS INT 13 interface. In de laatste fase wordt het OS opgehaald en de hardware setup en opmaak van het OS in het geheugen afgerond.



Afbeelding 1 » Basis SAN bootconfiguratie (Bron Microsoft)

Voor- en nadelen booten vanaf SAN

Dit boot proces kan vanaf een DAS, of over een LAN of SAN worden uitgevoerd. Het booten vanaf SAN biedt in vergelijking met de andere twee methoden een aantal voordelen: de mogelijkheid tot consolidatie van serversystemen, centraal beheer, eenvoudig herstel bij serverfouten, snelle disaster recovery en de opvang van tijdelijke serveroverbelasting door een image over te zetten naar een grotere server. Alhoewel het bootproces op zich betrekkelijk rechttoe rechtaan is (het verschilt niet wezenlijk van het locale proces), komt er voor de configuratie van de verschillende hardware-componenten wel het een en ander kijken. Gegeven de complexiteit daarvan moet iedere organisatie een afweging maken of de voordelen van het booten vanaf SAN daar tegen op wegen. Booting vanaf SAN is niet bepaald een technologie voor de opslagbeheerder die onbekend is met de complexiteit van SAN's. De beheerder wordt geconfronteerd met HBA's en LUN's die allemaal op de juiste wijze geconfigureerd moeten worden alvorens de server met succes vanaf SAN kan booten.

Eisen SAN boot-omgeving

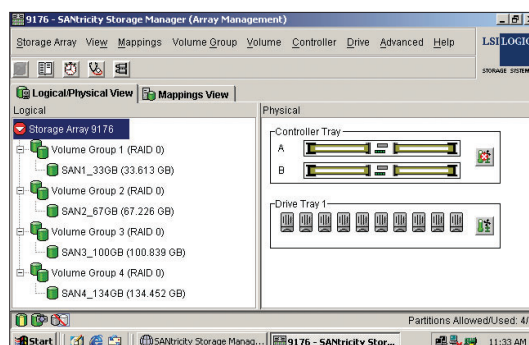
Microsoft ondersteunt weliswaar booting vanaf SAN, maar alleen op voorwaarde dat de leverancier van de SAN-omgeving het betreffende hardware-platform met SAN booting ondersteunt. De SAN en HBA's moeten zodanig worden geconfigureerd dat ze met succes vanaf SAN kunnen booten. Ten eerste moet de SAN via een FC-switch of 'direct attached' (rechtstreeks via een van de FC-poorten van een opslagsub-systeem) met de host worden verbonden. Het gebruik van Fibre Channel Arbitrated Loop (FC-AL) is niet toegestaan omdat dan de met het SAN verbonden hosts niet op de juiste wijze van elkaar kunnen worden geïsoleerd. De hosts moeten namelijk exclusieve toegang tot de disk hebben waar vanaf geboot wordt. Geen enkele andere host op het SAN mag de dezelfde logische disk 'zien', laat staan toegang daartoe hebben. Deze afscheiding van logische disk is mogelijk met behulp van Logical Unit Number (LUN) beheertechnieken, zoals 'LUN masking', 'zoning' of een combinatie van beide. LUN management wordt normaal op de FC-switch, opslagsub-systeem en/of HBA geïmplementeerd en niet binnen het Windows-systeem zelf (Windows heeft geen voorziening om LUN's 'te mappen'). Wanneer een host deel uit maakt van een clustersysteem dan moet het opslagsysteem op de Microsoft Cluster HCL-lijst voorkomen. Datzelfde geldt voor multi-path software en multiple HBA's, die op de 'HCL for Storage/RAID' systemen moeten voorkomen.

Pre-installation

Voordat de installatie begint moeten een aantal maatregelen worden genomen. Ten eerste moet de World Wide Name (WWN) worden genoteerd van iedere server die met het SAN is verbonden. Hetzelfde geldt voor de World Wide Port Name (WWPN) van elke daarop geïnstalleerde HBA (de WWN en/of WWPN is meestal op te vragen via de setup-utility van de HBA). Vervolgens is het noodzakelijk om te controleren of op iedere server de HBA BIOS is geactiveerd (dit is nodig om het boot-programma vanuit de HBA BIOS te kunnen uitvoeren) en of de HBA de juiste firmware-versie heeft. Tevens moet de juiste HBA-driver beschikbaar zijn. Voor bepaalde configuraties is de Microsoft Storport driver nodig. Met behulp van de HBA's WWN worden de servers in een zone ingedeeld, waarmee de communicatie tussen opslagdevice en server wordt beveiligd. Verder moet voor iedere server op de storage-array een LUN worden gecreëerd. Met behulp van 'masking', te configureren met een LUN management utility, wordt voorkomen dat hosts tegelijkertijd toegang hebben tot dezelfde LUN. Microsoft ondersteunt alleen booten vanaf SAN wanneer van LUN masking gebruik wordt gemaakt. Windows servers kunnen namelijk nog geen boot image delen, zodat iedere server zijn eigen LUN moet hebben om te kunnen booten.

Installatie basis SAN bootconfiguratie

Voor de testopstelling maken we van een StorageTek 9176 disk-array en een JBOD gebruik die via een Brocade 2800 Fibre Channel switch zijn verbonden met enkele Windows 2003 Enterprise serversystemen. Er moeten een aantal stappen worden doorlopen zodat de BIOS op de server vanaf de juiste LUN kan booten. Ten eerste moet de storage array, in dit geval de 9176, in LUN's worden ingedeeld. Het is gebruikelijk om dat te doen met behulp van een leverancier-specifiek programma. Hiermee moeten de, op de disk array aanwezige, fysieke schijven in LUN's wor-



Abbeelding 2 » Indeling storage array in Volume Groups

den opgedeeld. Storage array programma's noemen dit vaak geen LUN's maar 'Volumes', 'Virtuele LUN's' of 'Storage Partities'. Aan de 'buitenwereld', dat wil zeggen het Windows OS, worden ze echter als normale 'SCSI' LUN's gepresenteerd. Dus 'achter' de disk array kan een virtuele LUN uit meerdere fysieke schijven bestaan en ingedeeld zijn in een bepaald RAID-type. De disk array kan de LUN-getallen automatisch toewijzen, maar de beheerder kan ze ook via het programma zelf instellen. LUN-getallen blijven ofwel ongewijzigd, of ze worden door de HBA weer opnieuw ingesteld. Er wordt vanuit gegaan dat de disk array zelf een enkelvoudige LUN 0 is en geen disk device. De logical unit 0 wordt namelijk gebruikt om via het 'SCSI-3 Report LUNs'-commando informatie van de array op te vragen. In Windows worden deze LUN's teruggemeld aan de kernel als antwoord op het genoemde commando.

Ten tweede moet van te voren van iedere HBA de WWN worden vastgesteld. Dit unieke adres bevindt zich meestal op een sticker op de HBA maar kan ook via de HBA BIOS-utiliteit worden opgevraagd. Ook is het belangrijk om de WWPN van de storage array controller te noteren.

LUN mapping op Storage Array

Uitgaande van de 9176 storage array wordt met behulp van de door StorageTek meegeleverde 'SANtricity Storage Manager' in het 'Logical Physical View' menu, vier 'Volume Group's op basis van RAID 0 aangemaakt. Het zijn volumes die uit een of meer fysieke schijven bestaan. Na de indeling in 'virtuele volumes' volgt de configuratie van de storage array met LUN's. De storage array stelt de LUN-getallen automatisch in of de beheerder kan dat handmatig doen. LUN-getallen blijven ongewijzigd, of worden opnieuw door de HBA 'gemapped'.

Bij toepassing van meerdere serversystemen die van dezelfde disk array kunnen booten, is het noodzaak dat een specifieke server alleen toegang heeft tot de aan de array toegekende Volume(s). Deze toewijzing wordt aangebracht middels het proces van 'unmasking' van de betreffende LUN op de storage array met de juiste server. Daarbij zijn het LUN-getal en de WWN van de HBA nodig. De koppeling van beide wordt in het 'Mapping View' menu aangebracht.

SAN booten en clustering

Voor het booten vanaf een SAN-gebaseerd clustersysteem komt nog wat meer kijken. Het is een complexe zaak waarbij je met tal van zaken

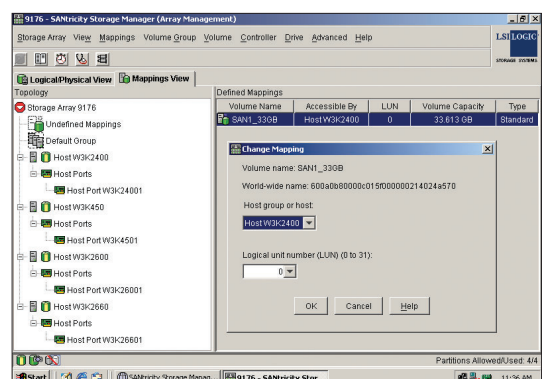
rekening moet houden. Microsoft eist dat bij cluster-servers (MSCS) het data-pad van de boot disk is afgescheiden van het 'shared-storage' pad. Dat betekent dat de normaal voor booting gebruikte HBA, niet dezelfde kan zijn als die voor het SAN shared opslagsysteem. Daarom zijn er aparte HBA's (of poorten daarop) voor iedere node binnen de clusteromgeving nodig. Of speciale HBA's worden gebruikt voor deze scheiding hangt af van de toegepaste Windows-driver. Is dit de traditionele 'SCSIport'- of de nieuwe 'Storport' driver?

Bij toepassing van SCSI-port-drivers zijn voor iedere server twee HBA's nodig: één voor de boot-disk de ander voor de gedeelde cluster-disk. De belangrijkste beperking daarbij is het aantal beschikbare PCI-slots voor HBA's (voor een volledige redundantie oplossing zijn vier HBA's nodig). De toepassing van de Storport-driver heft deze beperking op.

In het algemeen doorloop je voor de installatie van een SAN-gebaseerd clustersysteem de volgende stappen: Stap 1: installatie en configuratie van alle HBA's (zonder deze nog te aan te sluiten op de switch). Stap 2: configuratie van iedere boot disk naar de server. Stap 3: installatie Windows OS. Stap 4: Booting van de servers en tenslotte de installatie van de cluster software.

Problemen bij SAN booting

Het booten vanaf een SAN brengt een aantal specifieke problemen met zich mee waar de beheerder zich van bewust moet zijn. Belangrijk is om een onderscheid te maken tussen algemene bootproblemen of problemen die specifiek te maken hebben met een SAN-omgeving. Een bekend probleem is het niet kunnen lokaliseren van de boot-partitie en -bestanden. Wanneer nieuwe opslagsystemen (disks of LUN's binnen een storage array) aan een SAN worden toegevoegd, dan zal de HBA daar een 'target ID' aan toekennen. Alhoewel ieder opslag-device een uniek leveranciersspecifieke WWN heeft, verlangt het



Afbeelding 3 » Toekenning WWN aan host

Windows OS toch dat devices, volgens de afspraak die geldt voor SCSI-devices, van een nummer worden voorzien. De target ID's worden aan opslag-devices toegekend als deze in de fabric worden opgenomen. Wanneer een LUN binnen een target wordt gecreëerd, registreert de fabric de aanwezigheid daarvan niet door middel van een 'event'. De HBA is verantwoordelijk om de PnP te attenderen van haar aanwezigheid, of er moet een handmatige rescan van de disks worden uitgevoerd (via het hulpprogramma Diskpart of de Disk Management snap-in). Bepaalde FC arbitrated loop configuraties (FC-AL) kunnen ook niet van SAN booten. Alhoewel dit probleem is te omzeilen met HBA persistent binding, wordt de FC-AL methode niet door Microsoft ondersteund. Tenslotte kunnen er problemen ontstaan bij de toepassing van meerdere HBA's. Het kan gebeuren dat bij het opstarten van het systeem de PnP-voorziening de HBA een andere positie geeft in de device-configuratie, met als gevolg dat er van andere HBA poortadressen gebruik wordt gemaakt en het systeem niet meer op de juiste wijze kan booten.

Page files

Een page-bestand is een gereserveerd deel op de vaste schijf en wordt gebruikt om de voor applicaties benodigde hoeveelheid virtueel geheugen naar wens uit te kunnen breiden. Paging omvat het proces van tijdelijke uitwisseling van niet actieve data van fysiek geheugen naar vaste schijf. Omdat het OS ongehinderde toegang tot het page-bestand moet hebben, wordt dit bestand in de regel op dezelfde schijf opgeslagen als de systeembestanden (dus drive 'C:'). Alhoewel er een te verwaarlozen klein risico is op 'contention' tussen het lezen van de boot schijf en schrijven van pages kunnen er toch conflicten ontstaan wanneer meer systemen op het SAN tegelijkertijd paging I/O gaan uitvoeren, of wanneer meerdere systemen tegelijkertijd via dezelfde opslagpoort gaan booten. Een manier om dat te voorkomen is de niet-data- (zoals paging, registerwijzigingen en andere boot-gerelateerde informatie) van de data-gerelateerde I/O (zoals data van SQL of Exchange) af te splitsen en deze op een andere schijf op te slaan.

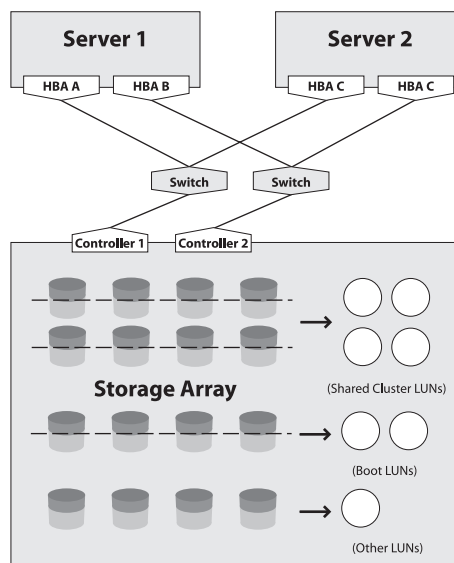
Toekomstige uitbreidingen

Er zijn een aantal beperkingen ten aanzien van de huidige SAN booting op Windows. Ten eerste kunnen Windows-servers nog niet vanaf een 'shared image' booten. Tot op heden heeft nog iedere server zijn eigen LUN nodig om vanaf te kunnen booten. Windows ondersteunt niet en masse de distributie

van boot images. De zogenaamde 'cloning' van boot images is wél toepasbaar waarbij het Windows Automated Deployment System (ADS) een handig hulpmiddel kan zijn.

Hiervoor zijn alleen de mogelijkheden van het booten vanaf een FC-gebaseerd SAN uiteengezet. Echter, Windows ondersteunt nu ook het booten vanaf een iSCSI-gebaseerde SAN. Het booten van SAN wordt niet ondersteund door de Microsoft iSCSI software initiator. Voorwaarde is wel dat er een iSCSI-gebaseerde HBA beschikbaar is. Net zoals in Fibre Channel omgevingen moet de HBA de INT 13 BIOS extensies ondersteunen die het boot proces in gang zetten. INT 13 zijn Device Service Routines (DSR's) die met de hard disk communiceren voordat de systeem drivers zijn geladen. Via INT 13 kunnen SAN-partities met een maximale grootte van 2 TB worden geboot. Tot op heden zijn er nog geen HBA's op de markt die rechtstreeks vanaf een iSCSI-disk target kunnen booten, maar de verwachting is dat deze dit jaar nog op de markt gaan komen.

Windows Server 2003 ondersteunt met behulp van de Extensible Firmware Interface (EFI) BIOS de 64-bits adressering. Om met een 64-bit server vanaf SAN te kunnen booten moeten de bootable HBA's daarvoor geschikt zijn. Daarbij is de 'Storport' driver de aangegeven HBA-driver. Want in tegenstelling tot het IA-32 ontwerp kan de EFI BIOS vanaf elk device booten waarbij het bootproces niet langer afhankelijk is van de 'INT 13 methode'.



Afbeelding 5 » Boot van SAN clustersysteem (Bron Microsoft)