

Naar een geclusterde storagearchitectuur

Ongestructureerde data vergen nieuwe benadering

Clustering van opslagsystemen zal de wijze veranderen waarop we in de nabije toekomst data zullen opslaan. Sommige trends geven aan dat geclusterde opslag dé toekomstige opslagarchitectuur voor data wordt. Naast de bestaande DAS-, SAN- en NAS-systemen wordt dit dan het vierde type opslagarchitectuur. Bram Dons schetst de ontwikkeling richting clustering.

Bram Dons

De verschuiving in de richting van deze vernieuwende opslagarchitectuur wordt gedreven door drie trends: de explosieve groei van data, de algemene verschuiving naar een clustergebaseerde computing-omgeving en de beschikbaarheid van

goedkope en snelle, op industriestandaarden gebaseerde hardware.

Ongestructureerde gegevens

De aandacht bij ondernemingen en leveranciers was de laatste jaren vooral gericht op de aanwas van gestructureerde data, zoals die aanwezig is in relationele databases. Tal van producten voor de opslag en verwerking van dit type data kwamen op de markt. Op de achtergrond is echter ongemerkt een proces gaande dat de IT-wereld de komende jaren voor een nog groter probleem gaat stellen. In feite gebeurt dat nu al, maar zeker in de toekomst zal er een gigantische toename van de hoeveelheid ongestructureerde data ontstaan, met name bij video- en audiotoeepassingen en door medische en andere grote digitale bestanden. De opslag van dit soort data zal het uiterste gaan vergen van de bestaande traditionele opslagssystemen, vooral wat betreft de geboden prestaties en opslagcapaciteit.

De huidige IT-architectuur is ongeschikt voor de opslag en verwerking van deze massale hoeveelheden ongestructureerde data, omdat die architectuur in

hoofdzak voor gestructureerde data en kleine bestanden is ontworpen. Ongestructureerde data hebben echter unieke karakteristieken en kunnen leiden tot zeer grote bestanden en datavolumes die extreem hoge eisen stellen aan de verwerkingscapaciteit van het opslagsysteem. Een bijkomende eis is dat systemen in een dergelijke omgeving gebruikers vrijwel gelijktijdige lees- en schrijftoegang tot een bestand moeten bieden, zoals bij gelijktijdige verwerking bij videobeelden, CAD-tekeningen en het opvragen van medische images.

Bij gebrek aan een alternatief hebben veel ondernemingen die in een vroeg stadium al met de opslag van ongestructureerde data werden geconfronteerd, zo goed mogelijk getracht tegemoet te komen aan de vraag naar een geschikt opslagsysteem. Noodgedwongen kozen ze ervoor om hun traditionele opslagsysteem, dat geschikt is voor gestructureerde, transactionele of tekstgebaseerde data, uit te breiden. Het probleem is echter dat zelfs de allernieuwste NAS- en SAN-systemen nog steeds zijn gebaseerd op de traditionele opslagarchitectuur, met alle genoemde beperkingen van

dossier storage

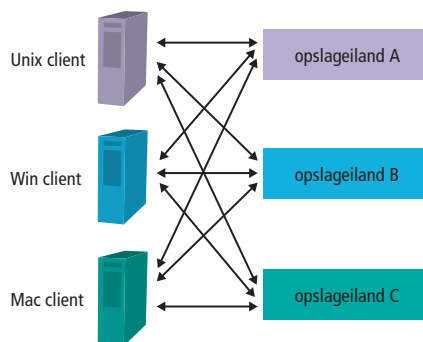
dien. Als gevolg van de toenemende complexiteit van de door storageaanbieders geleverde opslagtoepassingen creëerden organisaties ongemerkt en ongewild zogenaamde *opslageilanden* (zie figuur 1). Het bleek dat de complexiteit bij het beheer daarvan snel toenam en dat dergelijke oplossingen gekenmerkt worden door een beperkte schaalbaarheid. Men liep tegen diverse flessenhalzen aan die van invloed waren op de algehele prestaties van het opslagsysteem. Bovendien boden deze toepassingen een te lage beschikbaarheid en brachten ze hoge onderhoudskosten met zich mee.

Verschuiving naar clustercomputing

De tweede grote trend is de geleidelijke overschakeling van de IT-industrie naar een geclusterde computingomgeving. De oorspronkelijk op *proprietary* systemen en *symmetrical multi-processing* servers (SMP) gebaseerde grote datacenters gaan steeds vaker over naar een op industriestandaarden gebaseerd clustersysteem, werkend onder Linux of Windows. De belangrijkste redenen voor IT-managers om over te stappen zijn de betere prestaties, betere schaalbaarheid en betrouwbaarheid. Een dergelijke serveromgeving is beschikbaar tegen een fractie van de kosten van de traditionele servertoepassingen. Het verschijnsel doet zich voor dat de ingezette trend van serverclustering de komende jaren wordt uitgebreid naar die van storageclustering. De verschuiving destijds van grote monolithische serverdozen naar een geclusterde serveromgeving wordt nu snel aangevuld door de opkomende nieuwe architectuur voor geclusterde opslag.

Architecturen voor geclusterde opslag

Een architectuur voor geclusterde opslag heeft het vermogen om twee of meer opslagapparaten samen te voegen tot één enkele opslagentiteit. Dergelijke architecturen zijn er in drie typen: 2-way storage clustering; namespace aggrega-



Figuur 1 Bestaande architectuur met *opslageilanden* (bron: HP)

tion en clustered storage in combinatie met een distributed file system (DFS).

2-way storage clustering

Van oudsher betekende clustering de *active failover* tussen een stel redundante nodes. Alhoewel deze benadering nauwkeuriger kan worden beschreven als een redundante techniek, wordt deze in de industrie toch nog ten onrechte als een clustertechniek aangeduid. Leveranciers van NAS-systemen hebben er toen het label '2-way clustering' aan gehangen, vanuit de behoefte om de fouttolerantie en de redundantie te verbeteren bij legacy- en traditionele opslagarchitecturen. Het nadeel van de 2-way-architectuur is gelegen in de beperkte prestaties en schaalbaarheid, de complexiteit bij het beheer ervan en de relatief hoge kosten om een hogere beschikbaarheid te verkrijgen. Dat gecombineerd met de explosieve groei van ongestructureerde data maakt duidelijk dat dit soort toepassingen niet kan voldoen aan de eisen van de toekomstige datacenters van grote ondernemingen.

Namespace aggregation

Namespace aggregation is een methode om een cluster dat bestaat uit NAS-servers of opslagapparaten aan de gebruiker te presenteren als één geheel. Deze aggregatie creëert in feite een *gateway* waarmee data afkomstig van verschillende bestanden en heterogene systemen, worden herleid naar één gemeenschappelijk punt. Dergelijke toepassingen kunnen puur uit software bestaan, maar er zijn ook combinaties van hardware en software zijn (apparaat en switch) op de markt. Al deze producten creëren

een *single namespace* en een cluster van opslagbronnen die zich aan de gebruiker als één grote *pool* presenteert (zie figuur 2). Dit soort toepassingen kan wel een bestand over meerdere schijfvolumes op een specifiek opslagsysteem *stripen* (*data striping*), maar niet over meerdere opslagsystemen die deel van een cluster uitmaken.

Hoewel deze architectuur qua initiële kosten een aantrekkelijke oplossing voor de gebruiker lijkt, zit de IT-beheerder nog steeds opgescheept met het beheer en de configuratie van opslageilanden (heterogene silo's met opslagsystemen). De beheerder krijgt weliswaar nu de beschikking over een extra virtuele laag, maar uiteindelijk creëert dit soort toepassingen een complexere omgeving; het vraagt meer onderhoud en leidt ook op langere termijn tot hogere operationele kosten.

Clustered storage met DFS

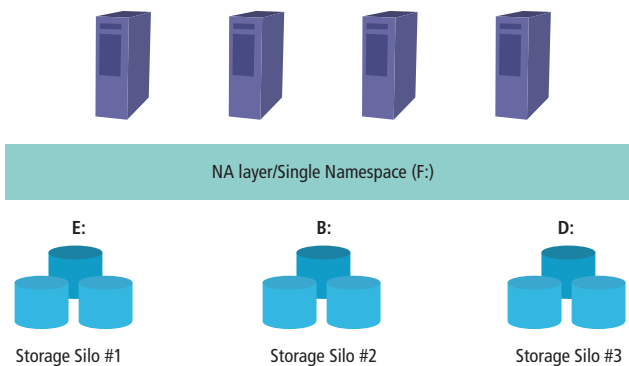
Het derde type is de distributed clustered storage-architectuur, die een natuurlijke evolutie is van de n-way eenvoudige clusterarchitectuur en een voorziening voor namespace aggregation (zie figuur 3). Het is een netwerkgebaseerd opslagsysteem dat gebruikers de mogelijkheid biedt om *storage nodes* te combineren en toe te voegen, waarbij alle nodes toegang hebben tot dezelfde *storage pool*. Dit soort toepassingen is binnen de opslaglaag geïmplementeerd in combinatie met een volledig gedistribueerd bestandssysteem dat over een aantal nodes of storage controllers is gespreid. Omdat de software zich binnen de opslaglaag bevindt, heeft die de volledige controle over de lay-out van de data (*data striping*) van alle storage nodes die deel uitmaken van het cluster. Dit in tegenstelling tot namespace aggregation- en virtualisatieproducten, die alleen controle hebben over een specifieke storagesilo.

Met behulp van intelligente software worden de nodes symmetrisch aaneen-

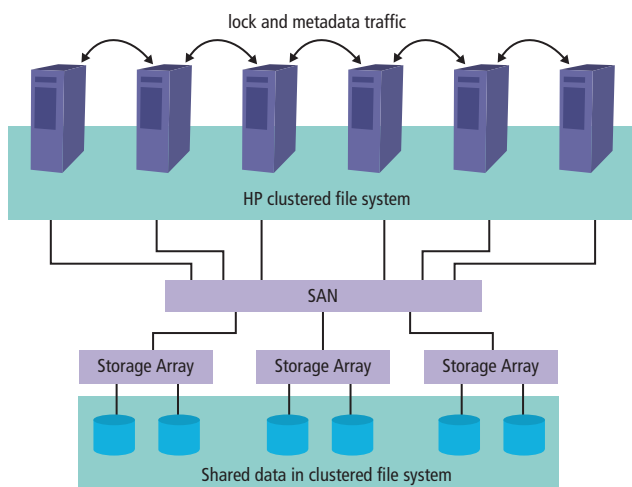
Cluster file system

Een CFS moet meerdere servers de mogelijkheid bieden om hun bestanden en data te kunnen delen (*file sharing*), met behoud van lees- en schrijfconsistentie en daarmee data-integriteit. Vanwege de gestelde eisen is de ontwikkeling van een CFS uiterst complex en moeten er allerlei fundamentele problemen worden opgelost. Want de kern van het probleem is de manier waarop de *state* tussen de vele nodes consistent en gewaarborgd moet blijven.

De afgelopen jaren zijn er diverse pogingen ondernomen om een CFS te ontwikkelen dat voldoet aan de eisen van het toekomstige rekencentrum van grote ondernemingen. Veel bestaande CFS'en zijn voortgekomen uit de traditionele high-availability-toepassingen. De mislukte pogingen om deze architectuur toe te passen binnen de commerciële datacenters waren het gevolg van de toegepaste architectuur, die op een *master server* gebaseerd was (zie figuur 4). Deze architectuur kent beperkingen ten aanzien van schaalbaarheid en prestaties. In sommige gevallen ontbreekt ook een volledige voorziening voor foutherstel en kan de implementatie complex zijn. De flessenhal bij deze architectuur wordt gevormd door het gebruik van de master server, omdat de nodes voor alle bestandstransacties (vooral schrijftransacties) de metadata op de master server moeten raadplegen. Omdat alle nodes van de CFS afhankelijk zijn, is het cluster ook kwetsbaar bij uitval van de master server. Bij sommige implementaties wordt dan ook gebruikgemaakt van een back-up master server.



Figuur 2 Namespace aggregation (bron: Isilon)



Figuur 3 Symmetrical cluster file system (bron: HP)

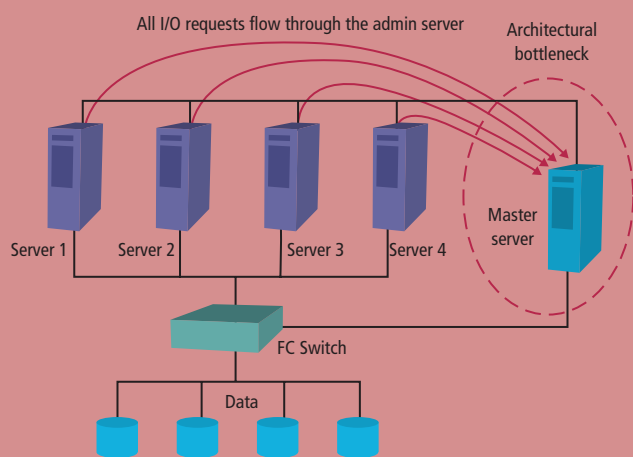
geschakeld en gedistribueerd, waardoor het cluster (dat uit drie of meer nodes bestaat) werkt als één intelligent samengesteld systeem. Elke node werkt daarin zelfstandig, communiceert met de andere nodes om bestanden op aanvraag van de gebruiker op te halen, en is binnen het cluster een *coherente peer*. Dat betekent dat elke node op elk moment precies weet wat een andere aan het doen is. Vanwege deze eigenschappen biedt een systeem voor gedistribueerde clusteropslag, van alle bestaande opslagarchitecturen, de allerhoogste niveaus van beschikbaarheid, betrouwbaarheid en schaalbaarheid. Door de bundeling van de doorvoer van elke node is het systeem bovendien lineair schaalbaar, in tegenstelling tot alle andere clusterarchitecturen. Daar neemt, als gevolg van de overhead, de schaalbaarheid snel af

boven een bepaald aantal nodes. De kern van een symmetrisch geclusterd opslagstelsel wordt gevormd door het schaalbare gedistribueerde cluster file system (CFS; zie het gelijknamige kader).

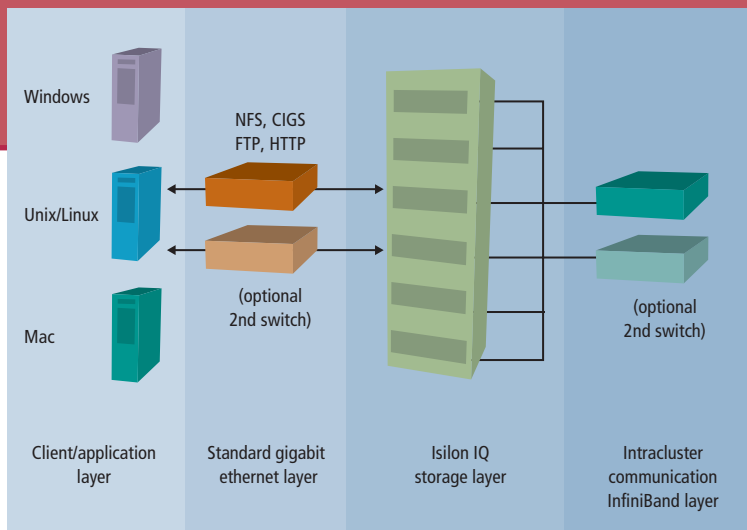
Systemen voor gedistribueerde clusteropslag

Er zijn talrijke cluster file systems, cluster-systemen, raid en volume managers van

verschillende leveranciers op de markt. Toch er zijn slechts enkele leveranciers die een compleet geïntegreerd systeem leveren, waaronder het StorageWorks Clustered File System van HP, NAS Cluster van PolyServe, Xiotech en IQ van Isilon. Hierna wordt de theorie besproken aan de hand van het laatstgenoemde product, IQ (zie figuur 5).



Figuur 4 Op master server gebaseerde clusterarchitectuur (bron: Polyserve)



Figuur 5 Isilon IQ storage clusterarchitectuur (bron: Isilon)

Isilon IQ

De firma Isilon heeft vier modellen op de markt gebracht: Isilon IQ 1920, 3000, 4800 en 6000. Elke storage node bestaat uit een compleet geïntegreerd systeem en bevat SATA-schijven, cpu, geheugen en netwerkverbindingen. De kern van het systeem is het gepatenteerde gedistribueerde bestandssysteem OneFS. Dit combineert de drie lagen van de traditionele opslagarchitecturen (bestandssysteem, volume manager en raid) in één systeem dat alle nodes binnen het cluster omvat. Het OneFS file system kan bestanden en metadata over meerdere cluster nodes stripen, wat een verbetering is ten opzichte van de traditionele methode van stripen: over individuele schijven binnen een enkel opslagsysteem of volume.

Het Isilon IQ-systeem is bestand tegen de uitval van meerdere schijven of zelfs een complete node, zonder verlies van data. Daartoe kunnen diverse beschermende technologieën op elk niveau binnen het cluster worden ingesteld: cluster, (sub)directory of individueel bestand. Bovendien kan de beheerder de instellingen weer op ieder moment online wijzigen. Wanneer een schijf uitvalt, bouwt OneFS automatisch het bestandssysteem weer op met behulp van de bestaande gedistribueerde schijfruimte. Afhankelijk van grootte en type van schijven is de 'herbouwtijd' van het opslagsysteem bij een falende schijf slechts enkele uren, in tegenstelling tot bij andere systemen; daar kan dit variëren van 3 tot 24 uur. De herbouwtijd zal de komende jaren overigens verder toenemen, doordat de schijfcapaciteit nog altijd spectaculair groeit (binnenkort zijn schijven van 1 Tb geen uitzondering meer). Traditionele opslagsystemen zijn tijdens de herbouwtijd bijzonder kwetsbaar: bij uitval van

nóg een schijf gaan er wel degelijk data verloren.

Prestaties

De gedistribueerde benadering van Isilon is een technologische doorbraak in de toepassing van de opslagarchitectuur; zij maakt goede prestaties, schaalbaarheid en beschikbaarheid mogelijk. Elk Isilon IQ-cluster kan bestaan uit 3 tot 41 cluster nodes. Uitgaande van het 6000-model biedt het cluster maximaal 256 terabytes aan opslagcapaciteit bij een doorvoer capaciteit van 4 Gbps. De nieuwste Isilon IQ-clusteruitvoering ondersteunt zelfs 528 Tb bij 7 Gbps!

Een andere doorbraak op het gebied van lineaire schaalbaarheid is de toepassing van InfiniBand-netwerktechnologie binnen een opslagomgeving. InfiniBand werd tot dusver voornamelijk toegepast in highend, op Linux gebaseerde server-clusteromgevingen, waar een zeer lage netwerklatency tussen cluster nodes vereist is. Het Isilon IQ-clustersysteem is leverbaar met gigabit ethernet en InfiniBand als cluster node-netwerkverbinding. Het gebruik van InfiniBand in plaats van 1 Gb ethernet verhoogt de prestaties van het cluster met zo'n 20 procent (in het bereik van 10 tot 20 cluster nodes). De verwachting is echter dat met de komst van 10 Gb ethernet – binnenkort - het verschil met de huidige 10 Gbps InfiniBand aanmerkelijk minder zal zijn. InfiniBand zal zijn voorsprong op ethernet voorlopig wel behouden, omdat 20 en 60 Gbps al beschikbaar zijn en er al een voorstel voor 100 Gbps ligt.

Samenvatting

De toepassing van een gedistribueerd geclusterd opslagsysteem heeft grote voordelen ten opzichte van de traditio-

nele master-based clusterarchitectuur. Per definitie biedt een gedistribueerde cluster een architectuur met een hoge beschikbaarheid, omdat elke node een coherente peer is van de andere. Als er een node of component uitvalt, blijven de data voor de andere nodes beschikbaar. Er is geen sprake van een single point of failure, want de status van het bestandssysteem wordt door alle nodes binnen het cluster bewaakt. In tegenstelling tot de mastergebaseerde benadering is de gedistribueerde clusterarchitectuur in elk onderdeel lineair schaalbaar.

Het systeem kent maar één niveau van systeembeheer, zodat het relatief eenvoudig te beheren is. Een volledige clusteropslag automatiseert taken die anders handmatig moeten worden uitgevoerd, zoals de *load balancing* van clients over de nodes en het opnieuw balanceren van de belasting bij toevoeging van nodes aan het cluster.

De architectuur omvat de gehele clusteromgeving, wat het beheer eenvoudig maakt. Zo verlost dit systeem de beheerder van de bij traditionele opslagsystemen gebruikelijke taak om de vele driveletters en de mapping van applicaties te moeten instellen en bijhouden. Alle bestanden kunnen namelijk onder één driveletter of mount point worden ondergebracht.

Tot op heden worden geclusterde opslagsystemen voornamelijk in de industrie en de wetenschap toegepast: in de olie- en gasindustrie, de media- en amusementwereld en bij onderzoeksinstituten. Maar de nog altijd dalende prijzen van hoogpresterende, op industriestandaarden gebaseerde pc's en de komst van betaalbare 10 Gbe- en InfiniBand- netwerken gaan waarschijnlijk ook de weg vrijmaken voor andere toepassingen. Dan komt de grootschalige toepassing van het geclusterde opslagsysteem in de transactiegebaseerde IT-omgeving in beeld.

*Bram Dons is onafhankelijk IT-analist.
E-mail: b.dons@it-trendwatch.nl.*