Drastische vermindering aantal servers met continuïteitgerichte implementatie

Failover clustering met SQL Server 2000

Erik Reijns en Johan de Weijs

In diverse organisaties is het gebruik van SQL Server de afgelopen jaren sterk toegenomen. SQL Server wordt daarbij in toenemende mate als volwassen DBMS gezien, waardoor ook bedrijfskritische toepassingen SQL Server als onderliggend DBMS gebruiken.

De praktijk leert echter dat veel organisaties op het gebied van technische infrastructuur, kennis en organisatie niet altijd even goed waren voorbereid op deze plotselinge sterke toename, waardoor de algehele inrichting van SQL Server vaak niet optimaal is. Failover clustering van SQL Server en daaraan gekoppeld consolidatie van databases is een middel om hierin verbetering aan te brengen.

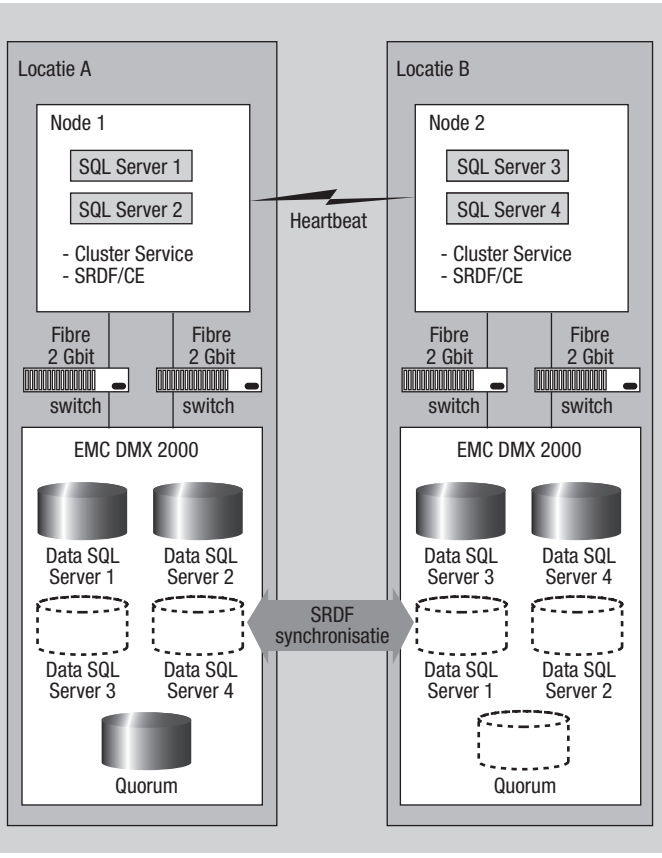
Bij F. van Lanschot Bankiers te 's-Hertogenbosch beschikt men sinds enkele maanden over een volledig operationeel SQL Server cluster, geïmplementeerd op een Microsoft Cluster Service (MSCS) omgeving, met daarop diverse (geconsolideerde) bedrijfskritische databases. Dit artikel gaat in op de motivatie om tot implementatie van het cluster over te gaan en geeft een toelichting op de technische inrichting.

Motivatie clustering SQL Server

Over het algemeen worden nieuwe toepassingen met SQL Server projectmatig gerealiseerd. In veel gevallen is dit als volgt te karakteriseren:

- Projecten bepalen zelf op welke wijze de implementatie plaatsvindt. Eenvoud, snelheid en onafhankelijkheid zijn daarbij belangrijke punten. Dit resulteert in de regel tot de aanschaf van nieuwe servers waarop de diverse componenten van een toepassing, zoals SQL Server, worden geïnstalleerd. Het gevolg hiervan is dat slechts af en toe servers en SQL Server instances worden gedeeld door meerdere toepassingen. Er ontstaan diverse verschillende implementaties van SQL Server naast elkaar, in grotere organisaties soms enkele tientallen.
- Er blijft veel hardware-capaciteit onbenut, omdat er veiligheidshalve voor wordt gekozen om servers aan te schaffen met meer capaciteit dan strikt noodzakelijk is voor de betreffende toepassing. Als de overcapaciteit van alle servers bij elkaar wordt opgeteld, kan het om grote getallen gaan.
- Continuïteit en beschikbaarheid zijn niet altijd even goed te garanderen. Dit is namelijk afhankelijk van de in het project gekozen implementatie.
- Up to date blijven met SQL Server software is arbeidsintensief. Bij een upgrade van de software moet namelijk elke instance apart behandeld worden. In de praktijk ontstaan hierdoor diverse versies van SQL Server naast elkaar.
- Dagelijks beheer is arbeidsintensief.
- Security-richtlijnen zijn moeilijk te handhaven.
- Licentiekosten van SQL Server en gerelateerde software (bijvoorbeeld voor monitoring en backup/recovery) nemen toe.

Dat dit geen optimale situatie is, hoeft geen betoog. Clustering van SQL Server en daaraan gekoppeld consolidatie van databases is een middel om hierin verbetering aan te brengen. Projecten krijgen ten aanzien van SQL Server een kant en klare infra-



Afbeelding 1: Cluster-implementatie op hoofdlijnen.

structuur aangeboden. Dit heeft als belangrijkste voordeel dat het aantal implementaties van SQL Server in een organisatie beperkt blijft tot die op het cluster, hetgeen resulteert in een drastische besparing in het aantal servers. Bovendien is uitwijk in één keer voor alle SQL Server databases geregeld.

Techniek

Afbeelding 1 bevat een overzicht op hoofdlijnen van de cluster-implementatie. Op twee locaties op enkele kilometers van elkaar zijn twee volledig identieke servers geplaatst. Deze servers zijn zodanig geconfigureerd dat ze volledig automatisch elkaars werk overnemen als één van de servers om wat voor reden dan ook niet meer beschikbaar is. Het is overigens absoluut noodzakelijk dat deze servers 100 procent identiek zijn. In de cluster-terminologie worden deze servers nodes genoemd. We hebben in dit geval node 1 en node 2. Op elke node draaien twee zogenaamde virtuele servers die telkens één SQL Server instance actief hebben.
In totaliteit zijn er dus vier SQL Server instances operationeel.
In elke instance zijn diverse databases gecreëerd.

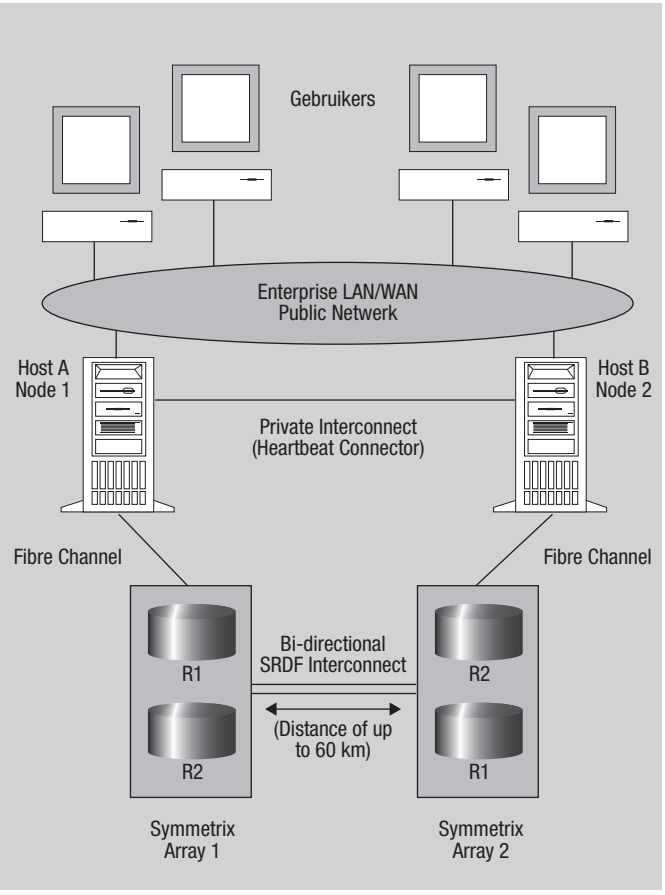
> Toepassingen moeten de connecties naar de databases opnieuw initiëren

Ten behoeve van de opslag van gegevens is op beide locaties een complete storage-infrastructuur van EMC beschikbaar. Beide locaties bevatten altijd alle gegevens. SRDF (Symmetrix Remote Data Facility) zorgt ervoor dat de gegevens op beide locaties continu gesynchroniseerd worden. Een wijziging in gegevens op locatie A is hierdoor direct ook op locatie B aangebracht. Dit vormt de basis voor uitwijk en zogenaamde 'high availability'-oplossingen zoals MSCS. Door deze spreiding van locaties (meerdere rekencentra) spreekt men in cluster-terminologie ook wel van een *geographical dispersed cluster* of in EMC-termen van een *geographical distributed SRDF/CE cluster*. CE in deze afkorting staat voor Cluster Enabler.

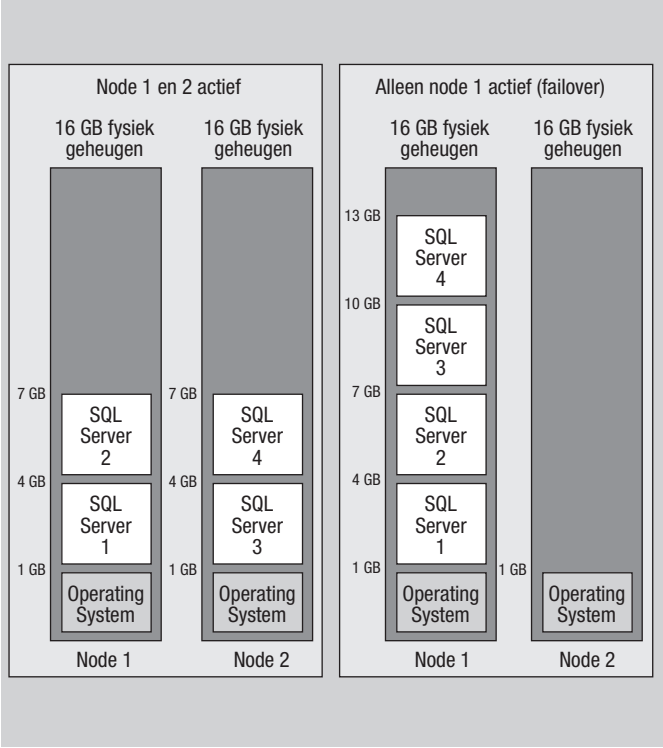
Om het cluster te managen is er een grafische interface beschikbaar: Cluster Administrator. Hiermee kunnen alle componenten (nodes, resources en groups) geconfigureerd worden. Ook is het mogelijk om met Cluster Administrator handmatig een failover uit te voeren, waarbij één of meerdere virtuele servers (met de bijbehorende SQL Server instances) op de ene node worden gestopt en op de andere gestart. Met een failback kan de situatie vervolgens weer worden teruggedraaid. Daarnaast kan het cluster en de SRDF/CE-configuratie voor de nodes met behulp van een tool van EMC beheerd en geconfigureerd worden. Dit is de SRDF/CE for MSCS configuration utility. SRDF/CE is als opvolger van EMC Geospan gepositioneerd en biedt ten opzichte van zijn voorganger vele voordelen zoals bijvoorbeeld een installatie-wizard.
Voor de volledigheid moet nog worden vermeld dat er twee clusters operationeel zijn. Naast een cluster in de productie-omgeving is er ook één operationeel in de test-omgeving.
Dit test-cluster is alleen ten aanzien van capaciteit afwijkend ten opzichte van het productie-cluster.

Nodes

De beide cluster servers zijn van Dell afkomstig. Het betreft 2 Dell Poweredge 6650 servers met de volgende technische specificaties voor iedere server: CPU, 4x Pentium Xeon 3.0 GHz.; memory 16 GB ECC DDR; Network interface cards, 3 maal Gigabit Ethernet; FC Host bus adapters, 2 maal Emulex HBA LP9002L van EMC.
Als operating system is gekozen voor Microsoft Windows 2003 server Enterprise Edition, welke inmiddels is voorzien van service pack 1. Vanwege een aantal evidente voordelen, zoals bijvoorbeeld performance, is gekozen voor de Microsoft storport driver in combinatie met de vendor specific storport mini driver van Emulex. De gekozen multiple path-oplossing is Powerpath versie 4.40. Hiermee wordt vermeden dat er een single point of failure optreedt, doordat er slechts 1 pad naar de EMC-storage aanwezig zou zijn. Het spreekt voor zich dat zeker in een cluster dat een



Afbeelding 2: Cluster-implementatie: Netwerk en EMC.



Afbeelding 3: Geheugengebruik.

hoge beschikbaarheid moet hebben het van cruciaal belang is om zoveel mogelijk single points of failure te voorkomen. De reeds eerder vermelde SRDF/CE software van EMC is versie 2.0.1.7. De storage-infrastructuur waar het cluster gebruik van maakt, bestaat uit 2 DMX 2000's van EMC. De connectiviteit naar deze DMX'en verloopt via een aantal fibre switches van Brocade. Ook hier geldt de stelregel dat deze op een dusdanige manier worden ingezet, dat het uitvallen van zelfs een totale switch geen gevolgen heeft voor de beschikbaarheid van het cluster.

Er zijn vier named instances in gebruik, in een normale situatie twee op elke node.We spreken dan ook van een multiple instance cluster. Als er als gevolg van calamiteiten een node niet beschikbaar is, worden de twee normaliter daarop draaiende instances automatisch gestart op de andere node, waardoor de continuïteit gewaarborgd is. De betreffende instances zijn daarbij wel enige tijd (enkele seconden) uit de lucht. Toepassingen moeten dus de connecties naar de databases opnieuw initiëren. In een cluster-omgeving is SQL Server Enterprise Edition noodzakelijk. Er is gekozen om service pack 3A te installeren. Service pack 4 is (nog) niet geïnstalleerd, omdat het beleid is om niet direct na het beschikbaar zijn van service packs deze te installeren. Bovendien bevatte service pack 4 ten tijde van de implementatie van het cluster nog een bug ten aanzien van AWE (Address Windowing Extensions). Inmiddels heeft Microsoft hiervoor een fix opgeleverd. De installatie van SQL Server op een cluster is voor het grootste deel identiek aan een reguliere installatie. Een verschil is dat moet worden aangegeven dat het om virtuele servers gaat.

Met name de configuratie van het gebruik van geheugen in SQL Server is in een cluster-omgeving een punt van aandacht. Het totaal van het door de vier instances gebruikte geheugen mag namelijk nooit meer zijn dan het beschikbare geheugen op een node. In afbeelding 3 wordt dit toegelicht. Afbeelding 3 bevat aan de linkerkant een weergave van de situatie dat op elke node 2 SQL Server instances draaien. Elke instance heeft daarbij maximaal 3 GB aan geheugen ter beschikking. Dit is in SQL Server geconfigureerd met de server optie *max server memory*. In een omgeving zoals deze is het absoluut noodzakelijk om deze optie te zetten, omdat anders SQL Server al het beschikbare geheugen gaat gebruiken. Er blijft een flink deel van het geheugen ongebruikt in deze situatie. Dit is echter noodzakelijk, omdat er anders problemen ontstaan bij een failover.
De rechterkant van afbeelding 3 bevat een weergave van de situatie waarin er sprake is van een failover naar node 1. De instances 3 en 4 worden gestart op node 1, waarbij elke instance een deel van het beschikbare geheugen gaat gebruiken. Indien er op node 1 onvoldoende geheugen vrij zou zijn om instances 3 en 4 te faciliteren, dan leidt dit tot paging naar disk met alle nadelige gevolgen voor de performance. Dit is in de huidige situatie niet het geval. Er is zelfs nog geheugen over (reservecapaciteit om toekomstige groei op te kunnen vangen).

SRDF/CE

SRDF/CE vormt een essentieel onderdeel van de gehele clusterconfiguratie. SRDF/CE en de Microsoft Cluster Service werken nauw samen. Bij een calamiteit, bijvoorbeeld als een disk niet meer goed functioneert, zal SRDF er eerst voor zorgen dat alle gedeelde disks aan de 'standby-zijde' een READ-WRITE status krijgen in plaats van WRITE DISABLED. Pas als dat is gerealiseerd zal de Cluster Service de group waarin de kapotte disk zich bevindt weer online brengen, met als gevolg een geslaagde failover.

> Een mogelijkheid is dat een node of disk niet meer goed functioneert

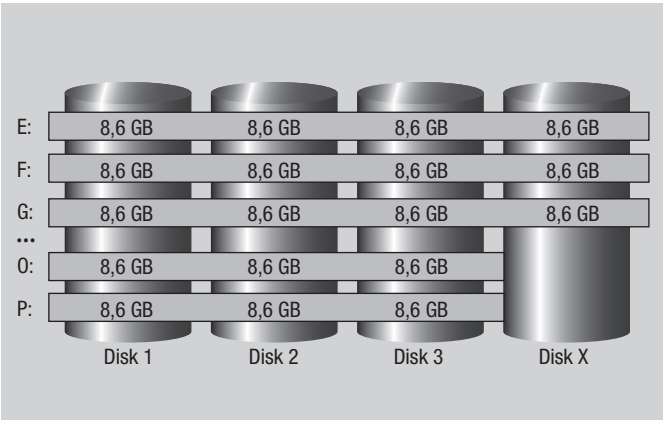
Een verbetering in SRDF/CE 2.0.1.7 is de zogenaamde autoswap R1/R2. Het gaat wat ver om deze mogelijkheid tot in detail te beschrijven, maar eenvoudig gezegd zorgt deze optie ervoor dat je na een failover meteen weer uitwijkbaar bent naar de andere node. Dit uiteraard alleen als het probleem dat de eerste failover veroorzaakte is opgelost.
Voor de opslag van de data op EMC wordt als techniek gebruikt RAID5 3+1 (3 disks voor data en 1 voor pariteit). Dit wordt hier verder niet toegelicht. Een meer logische view op de opslag van

de data is in afbeelding 4 weergegeven. Er wordt gebruik gemaakt van zogenaamde striped meta volumes. Voor elk volume (in dit geval E: tot en met P:) wordt daarbij een stripe gedefinieerd bestaande uit een aantal fysieke eenheden van 8,6 GB op diverse disks. Alle I/O wordt hierbij evenredig over de betreffende disks verdeeld, hetgeen de performance ten goede komt. De quorum speelt een belangrijke rol in een cluster, omdat deze de configuratie-database is voor de Cluster Service. De quorum log-file is de fysieke representatie hiervan. Deze file bevindt zich op een apart quorum-volume (Q:) op EMC. De quorum bevat informatie over de configuratie van het cluster zoals bijvoorbeeld over de servers die deel uitmaken van het cluster, de resources die op het cluster zijn geïnstalleerd en de status van deze resources (online of offline). Als de quorum niet beschikbaar is, kan de Cluster Service niet starten. De belangrijkste functies van de quorum zijn de volgende:

- de quorum zorgt ervoor dat alle nodes een consistente view naar het cluster hebben;
- bij een calamiteit vormt de quorum de basis voor alle logica die de Cluster Service nodig heeft om de juiste beslissingen te nemen.

Calamiteiten

Natuurlijk zijn er vele scenario's te verzinnen waarbij er sprake is van een dusdanige calamiteit dat een Microsoft Cluster Service omgeving zijn meerwaarde kan bewijzen. Zo kan er een netwerkprobleem optreden waardoor gebruikers (via het zogenaamde publieke netwerk) geen verbinding meer kunnen maken met een bepaalde node van het cluster. MSCS zal hierop reageren door middel van een failover naar de andere node. Een andere situatie treedt op als de zogenaamde *heartbeat* wegvalt. Deze netwerkverbinding, ook wel het private netwerk genoemd, wordt gebruikt voor node-to-node communicatie ten behoeve van cluster statusinformatie en cluster management. Indien deze heartbeatverbinding wegvalt zal één van beide publieke netwerken deze rol overnemen.
Een andere mogelijkheid is dat een node of disk niet meer goed functioneert. Ook hier geldt dat er automatisch een failover zal plaatsvinden. Indien er een complete DMX niet meer beschikbaar is of de fibre-verbinding (SRDF) tussen beide rekencentra (en dus beide cluster nodes) valt weg, dan is een succesvolle automatische failover van meerder factoren afhankelijk. In het uiterste geval moet handmatig een aantal acties uitgevoerd worden om de zaken weer online te brengen. Ditzelfde geldt uiteraard ook indien een complete site niet beschikbaar is, bijvoorbeeld als gevolg van een brand.
SQL Server is cluster aware. Dit betekent dat indien een SQL Server instance niet meer beschikbaar is er ook een failover van de betreffende instance zal plaatsvinden.
In geval van onderhoud aan een node kunnen alle resources, in dit geval alle SQL Server instances, eenvoudig overgebracht worden naar de andere node. Nadat het onderhoud is afgerond kunnen de resources weer verdeeld worden over beide nodes.
MSCS biedt geen bescherming tegen datacorruptie en menselijke fouten. Dat betekent dat goede voorzieningen op het gebied van backup en recovery nog steeds noodzakelijk zijn. Er is gekozen om gebruik te maken van het tool Tivoli Data Protection (TDP) for SQL Server. Alle backups worden hierbij rechtstreeks geschreven naar Tivoli Storage Manager (TSM).



Afbeelding 4: Logische view opslag data.

Conclusie

De implementatie van het SQL Server cluster heeft als resultaat dat het aantal implementaties van SQL Server sterk is afgenomen, waardoor een drastische besparing van het aantal servers is bereikt.
Een implementatie overigens die door Microsoft het predikaat 'excellent' op een aantal onderdelen heeft gekregen, nadat we de cluster-omgeving door een expert van Microsoft hebben laten beoordelen.

Bovendien is uitwijk in één keer voor alle SQL Server databases geregeld. Er is echter ook een nadeel te noemen. Wijzigingen in de software (bijvoorbeeld bij een upgrade van SQL Server naar een hogere versie) vereisen veel afstemming, omdat SQL Server instances databases van diverse toepassingen bevatten. Er kan dan de situatie ontstaan dat voor de ene toepassing een wijziging zonder meer mogelijk is, terwijl dit voor een andere toepassing moeilijker is, omdat er additionele activiteiten moeten worden uitgevoerd. Een goede planning van activiteiten kan dit nadeel in belangrijke mate teniet doen.

Erik Reijns en Johan de Weijs
Erik Reijns (e.j.reijns@vanlanschot.com) en Johan de Weijs (j.c.m.deweijs@vanlanschot.com) zijn respectievelijk als IT-architect en senior DBA werkzaam bij F. van Lanschot Bankiers te 's-Hertogenbosch.