Verslag verstoringen 29 en 30 november 2010

Maandag 29 november 2010

07:20 uur: Monitoringsystemen rapporteren af en toe time-outs bij het uitvoeren van reguliere controles tegen de primaire productie database server. Handmatig onderzoek laat echter geen vreemde dingen zien.

13:30 uur: De database beheerpartner meldt dat de stand-by databases niet goed bijgespeeld worden en dat dit niet te lang mag duren omdat anders een failover niet zal gaan werken.

14:00 uur: Het onderzoek is in volle gang, maar de oorzaak is niet duidelijk. Onverklaarbare problemen worden geconstateerd: applicatie claimt geen files te kunnen kopiëren; handmatig lukt het echter wel. Traditionele pings werken tussen bepaalde systemen op hetzelfde netwerk niet, tussen andere weer wel; TCP-verbindingen opbouwen lukt wel, maar ook weer niet altijd. We vermoeden dat dit of door een probleem in het OS op de primaire database server wordt veroorzaakt, dan wel door een hardwareprobleem in een van de switches. Alle database servers zijn redundant ontsloten en dus met twee switches verbonden.

15:30 uur: Alhoewel de systemen nog steeds functioneren wordt besloten om toch een spoedonderhoud in te lassen om 16:15 uur.

16:15 uur: De DRS-applicatie wordt dichtgezet voor de buitenwereld. Databasesoftware wordt gestopt en de betreffende database server wordt herstart. Na het opnieuw starten van de database blijkt dat eerdere problemen minder frequent voorkomen maar nog altijd niet opgelost zijn. Hierop wordt besloten de eerste switch te herstarten.

17:35 uur: De eerste clusterswitch wordt herstart. Deze herstart resulteert onbedoeld in een database failover. De failover database was na de herstart van de database server volledig bijgespeeld en dus gereed voor productie. De failover manager besloot derhalve de stand-by database te activeren. De communicatieproblemen waarmee alles begon, waren echter nog altijd niet verholpen.

18:00 uur: De tweede clusterswitch wordt herstart. Deze switch kwam niet meer op. Fysieke controle door medewerkers van BIT ter plaatse toonde aan dat de switch daadwerkelijk kapot was. Inmiddels was al wel gebleken dat de failover database zonder problemen functioneerde, maar ook dat de originele primaire database corrupt was geraakt en als verloren beschouwd diende te worden. Daar bovenop bleek dat, door de kapotte switch, de helft van de voor productie bedoelde applicatieservers niet beschikbaar was. Ook bleken andere kopie-databases niet bijgespeeld te worden hetgeen wel het geval zou moeten zijn.

Na onderzoek bleek dat laatste het gevolg te zijn van een configuratiefout: de databasesynchronisatie verliep via het verkeerde (niet-redundant aangesloten) netwerk. Deze fout is direct hersteld waarna de overige beschikbare kopie-databases binnen no-

time up to date waren. Deze configuratiefout was ook de oorzaak van de database failover: ook de controles tegen de beschikbare databases door de failover manager verliepen via het verkeerde netwerk en resulteerden dus onbedoeld in een database failover.

Intern zijn diverse discussies gevoerd over het vervangen van de kapotte switch. Probleem was echter dat we a) geen identieke (lees: ander merk en minder poorten) spare switch paraat hadden, b) er ernstige vertraging was op de weg door de hevige sneeuwval en het ons derhalve niet verantwoord leek om een compleet andere switch met minder poorten met spoed te configureren en in te gaan bouwen. Derhalve is aan BIT gevraagd een paar patches te veranderen zodat in elk geval de applicatieservers weer beschikbaar zouden zijn. Deze patches waren tegen 21:00 uur gereed.

21:00 uur: Uitvoeren diverse controles en start van alle applicatieservers en een reguliere intake test.

21:45 uur: De intaketest is met succes afgerond en tegen 22:00 uur gaat de DRS-applicatie open voor de buitenwereld.

Conclusie

Belangrijke communicatie zoals het synchroniseren van databases verliep via het niet-redundant aangesloten netwerk in plaats van het wel redundant aangesloten netwerk. Dit bleek een configuratiefout en deze is hersteld. Configuraties van alle switches dienen nauwkeurig te worden nagelopen.

Issue	**Uitgevoerde actie**	**Openstaande actie**
Foutieve configuratie netwerksegment t.b.v. database synchronisatie en failover manager	Configuratie hersteld naar redundant uitgevoerde netwerksegment.	Ook andere netwerk segmenten redundant uitvoeren. Zie ook actie 30 november 2010. Nalopen switchconfiguraties.
Cluster switch defect	Vervangen door vrijgespeelde switch binnen SIDN. Zie actie 30 november 2010	Spare switch bestellen. Bij bestelling vervanging core-infra ook spare meebestellen voor clustered-switches
Database physical stand-by bijwerken	Na her configuratie automatisch verlopen.	
Herstellen corrupte voorheen primaire database	Hersteld	
Database failover situatie herstellen	Downtime noodzakelijk, besloten wordt met 2 physical stand-by databases te draaien	Zie verder actie 30 november.

Dinsdag 30 november 2010

07:12 uur: De primaire DRS-database gaat onderuit. Onze monitoring meldt een paar minuten later aan de waakdienst medewerker dat de database onbereikbaar is. Deze controleert de melding en belt vervolgens de 2e lijns-medewerker met kennis van de database omgeving.
Uit onderzoek blijkt dat de database onderuit gegaan is als gevolg van een volgelopen volume op de filer. De vraag komt direct naar boven waarom de monitoring tools geen alarm hebben gegeven. Het blijkt helaas niet mogelijk de database te starten (na het vergroten van het onderliggende volume).

07:47 uur: We bellen onze beheerpartner om gebeld voor assistentie. Opvallend is dat uit onderzoek blijkt dat het volume om 07:01 uur nog 22 GB aan diskruimte vrij had. Dat betekent dat de database in 11 minuten met 22 GB gegroeid moet zijn, een hoeveelheid die we normaliter in twee maanden nog niet eens halen. Dit fenomeen is nog altijd niet verklaard.
De standaard recovery procedure wordt geïnitieerd om de database te herstellen. Wat een succesvolle poging lijkt, blijkt dat bij nader inzien toch niet te zijn omdat alle data van na maandagavond 21:20 uur verdwenen is. Andere pogingen falen vanwege een bug die optreedt tijdens de recovery. Installatie van een beschikbare patch biedt geen uitkomst en de patch wordt derhalve meteen teruggedraaid.

15:34 uur: De overgang op handmatige recovery is gelukt. Op dat moment is de primaire database weer in orde. Stand-by databases zijn echter niet beschikbaar. Door het niet volledig geschreven archive log (door het ruimtegebrek) is de database tot op 10 minuten na hersteld. De gedurende die 10 minuten gemiste transacties zijn met behulp van de audit log geïdentificeerd en vervolgens in overleg tussen registrars en R&S verder afgehandeld. Voor de zekerheid wordt direct een snapshot gemaakt van deze situatie.

Besloten wordt om ook de stand-by databases te herstellen omdat we niet live willen gaan zonder op z'n minst een stand-by database. Tevens wordt besloten om de failover situatie te herstellen. Deze was door de failover op maandagavond niet langer beschikbaar.
Door het gelijktijdig uitvoeren van twee herstelacties, de vervanging van de switch en de recovery van de database, is een recovery file corrupt geraakt zonder dat dit gepaard ging met een foutmelding. De corruptie werd hierdoor pas later in de recovery procedure ontdekt. Recovery van de stand-by databases blijkt niet langer mogelijk. De stand-by databases dienen als verloren te worden beschouwd.
En passant blijkt dat na installatie van de nieuwe switch opeens meldingen over het volgelopen diskvolume binnenkomen.

22:00 uur: We besluiten dat de enige mogelijkheid om de stand-by databases op te tuigen een kopieerslag van alle data is. Deze kopieerslag wordt gestart.
Algemeen Directeur Roelof Meijer en de relatiebeheerder Daniel Federer hebben ondertussen contact gehad met een aantal registrars en de wens is om –omdat SIDN geen exact tijdstip kan afgeven- om 08:30 uur open te gaan.

01:45 uur: Alle benodigde bestanden zijn gekopieerd en de failover database kan opnieuw geconfigureerd en ingericht worden. Tevens worden de andere stand-by databases opnieuw opgezet. Na helaas weer wat tegenslagen, zijn om 04:10 uur deze werkzaamheden afgerond.

05:30 uur: De DRS-applicatie draait en er volgt een uitgebreide intaketest. Deze is om 06:30 uur gereed.

08:30 uur: DRS5 wordt vrijgegeven voor de registrars.

Lessons learned

- Alle netwerksegmenten (database-, applicatie- en managementsegment) dienen redundant te worden ontsloten op de betreffende servers.

- Alle diensten (ook die van de beheerpartners) dienen voor openstelling te worden gecontroleerd, niet alleen de diensten die de storing hebben gehad.

- Wanneer een standaardmethodiek (in dit geval recovery) niet conform verwachting werkt, keer dan terug naar een drastisch andere aanpak (kopiëren van bestanden [snapshot]).

Issue	Uitgevoerde actie	Openstaande actie
Disk volume te klein	Disk volume vergroot	
Opsporen volgelopen volume		Onderzoek loopt.
Oplossen Oracle bug bij recovery	Uitgerold maar ook weer terug gedraaid wegen het niet functioneren daarvan.	In testomgeving trachten te reproduceren.
Handmatige recovery uitvoeren om primaire database te herstellen	Geslaagd	
Stand-by database bijspelen		In onderhoudsvenster daar down time noodzakelijk is.
Netwerksegmenten dubbel ontsluiten op servers	Uitgevoerd via bond op NIC	Quad-NIC kaarten zijn besteld om situatie permanent redundant te maken over alle netwerksegmenten.
Kopie maken van primaire database naar failover site	Uitgevoerd	Physical stand-by configuratie aanpassen voor failover database (was eerst primair). Down time voor nodig.

N.B. Zoals te zien is kennen sommige stappen flinke doorlooptijden. Deze hebben te maken met de lange doorlooptijden van verschillende herstel- en kopieerprocessen op en van de gehele DRS5-database.