RAPPORT INZAKE INCIDENT NETWERK BIT-2C – 17 02 2021

Samenvatting

Een zich misdragende switch van een klant aangesloten op het netwerk van BIT zorgde voor resource starvation van de switches leaf-sw1.bit-2c en leaf-sw2.bit-2c. Uplinks op deze switches waren daardoor niet bruikbaar. Specifieke omstandigheden bij deze storing zorgden ervoor dat de shared storage diensten (met uitzondering van iSCSI), virtualisatiediensten en shared hosting diensten niet beschikbaar waren.

Details

13.46	Een switch van een klant, via een port-channel aangesloten op het netwerk van BIT, zorgt voor hoge aantallen MAC address entries op een pair MC-LAG leaf switches in BIT-2C. De forwarding-plane van de BIT-switches functioneert niet meer, waardoor er geen verkeer meer wordt geforward door de leaf switches. Van de vier leaf switches in BIT-2C, zijn er twee onbruikbaar geworden.
13.47	Het netwerk probleem veroorzaakt onbeschikbaarheid van het Ceph cluster van BIT. Diensten die gebruik maken van dit cluster, te weten de virtualisatiediensten, alle shared diensten van BIT en shared storage (met uitzondering van iSCSI), worden onbeschikbaar. De storing wordt door engineers van BIT opgemerkt en zij starten hun onderzoek naar de oorzaak.
13.51	De clients die gebruikmaken van CephFS zijn verwijderd en geblockt door de storage nodes omdat de clients niet meer communiceren met de CephFS MetaDataServer (MDS).
~13.52	Het is duidelijk dat de storing zich voordoet op één paar leaf switches in BIT-2C, het onderzoek richt zich op de oorzaak ervan.
13.56	Omdat ook www.bit.nl onbeschikbaar is, wordt het incident op www.bit.org gemeld.
14.01	De oorzaak van de niet functionerende forwarding plane op de switches wordt gevonden in overloaded hardware resources voor de forwarding plane. Het mechanisme dat gebruikt wordt ter bescherming tegen loops bij klanten probeert meer hardware resources te alloceren dan beschikbaar zijn. Eén poort van de port-channel voor die switch van de klant wordt geshut.
14.05	Het shutten van de poort heeft het probleem niet opgelost. Daarom wordt de eerste switch van het leaf pair gereboot.
14.06	Diverse klanten verwijzen in hun communicatie naar de incidentmelding op www.bit.org, als gevolg van de piek in het aantal bezoekers wordt deze site onbereikbaar.
14.07	De tweede switch (leaf-sw2.bit-2c) van het leaf pair wordt gereboot. Direct na het reboot commando wordt het Ceph cluster weer beschikbaar. Servers die van NFS gebruik maken en servers zonder mounts worden weer beschikbaar. Dit zijn onder andere de MX, IMAP, SMTP servers en de webservers voor onder meer www.bit.nl, portal.bit.nl en webmail.bit.nl. Virtual machines van klanten zonder CephFS mounts zijn ook weer beschikbaar.
14.12	De eerste switch is succesvol gestart na de reboot. Uplinks op deze switch zijn weer beschikbaar.
14.12	Na een reboot van de server die www.bit.org serveert, komt die site ook weer beschikbaar.
14.14	De tweede switch is succesvol gestart na de reboot. Ook de uplinks op deze switch zijn weer beschikbaar.
~14.15	Als gevolg van het blocken van de CephFS clients behoeven clients met een CephFS mount remounts/reboots. Alle servers met dergelijke mounts worden gecontroleerd en waar nodig geremount of gereboot.
14.29	Het incident wordt op www.bit.org afgemeld.
14.30	De shared hostingdiensten zijn weer beschikbaar.
14.43	Op www.bit.nl wordt melding gemaakt van het incident met een verwijzing naar de incidentmelding op www.bit.org.

~15.00	De overloaded hardware resources waren alleen mogelijk in combinatie met een specifieke configuratie op uplinks voor klanten. Alle switches worden nagekeken op het voorkomen van een dergelijke configuratie. Deze configuratie blijkt alleen voor eerdergenoemde klant gebruikt te worden. Overal wordt de configuratie voor deze klant aangepast.
16.15	Alle virtuele servers met CephFS mounts zijn gecontroleerd en waar nodig zijn de mounts weer beschikbaar gemaakt.

Conclusie

Netwerk

Een configuratie op een pair MC-LAG leaf switches maakt die switches kwetsbaar voor hardware resource starvation. Die resources worden door de forwarding plane gebruikt ten behoeve van port security, storm control, en dergelijke. De forwarding plane van de switch stopt met functioneren bij genoemde starvation, terwijl de control plane van de switch wel blijft werken. Daardoor lijken specifieke MAC- en IP-adressen nog bereikbaar te zijn via deze switch, terwijl het verkeer vervolgens niet op de juiste switchpoort afgeleverd wordt. Het niet meer forwarden van packets en frames en tegelijkertijd blijven verkondigen dat alles aangesloten op de switch wel bereikbaar is, is bijzonder ongewenst gedrag.

Standaard wordt er voor elke poort op de switches in de resources een counter bijgehouden voor het aantal MAC adressen dat op die poort leeft. Zodra die counter een limiet raakt wordt de betreffende poort geshut. Voor één klant is die configuratie aangepast zodat er geen counter bijgehouden wordt, maar een access list van MAC adressen met een limiet. Zodra die limiet geraakt wordt, wordt niet de poort geshut maar worden er geen nieuwe MAC adressen meer geleerd. Deze afwijkende configuratie is gezet omdat deze specifieke klant in het verleden door vele MAC adreswisselingen regelmatig port security triggerde met port shuts voor deze klant als gevolg. Het was BIT niet bekend dat deze configuratie een risico op resource starvation introduceerde. De hardware resources worden gemonitord zodat er tijdig ingegrepen kan worden op hoge load op de resources. Vanwege de grote aantallen MAC adressen die razendsnel geleerd werden op de poorten van deze specifieke klant hebben de monitoringsystemen niet tijdig kunnen waarschuwen voor de (te) hoge load.

De resource starvation had tot gevolg dat de switches, leaf-sw1.bit-2c en leaf-sw2.bit-2c, geen verkeer meer doorstuurde. Klanten met uplinks op die switches konden geen verkeer meer met de switches uitwisselen tot het moment dat de problematische configuratie verwijderd was en de switches gereboot waren.

Storage

Het shared storage platform van BIT is gebaseerd op Ceph en is volledig redundant uitgevoerd met nodes in de datacenters BIT-1, BIT-2A en BIT-2C. Bij uitval van een datacenter blijft alle data op het Ceph platform beschikbaar. De specifieke omstandigheden van het netwerkincident zorgden er echter voor dat het Ceph cluster onbereikbaar werd.

De switches van BIT maken gebruik van EVPN, een extensie op het BGP protocol. Verkeer wordt over het netwerk gerouteerd in plaats van geswitcht. Met behulp van Equal Cost Multi Path (ECMP) routing worden alle storage nodes in het platform gebruikt, mits beschikbaar. Een volledige uitleg van deze technieken vindt u op onze site[1].
Omdat de control planes van de problematische leaf (access) switches wel bleven werken, bleven zij ook routes voor de storage nodes achter deze switches adverteren. De storage nodes zijn via een apart (storage) netwerk aan elkaar gekoppeld, dat netwerk was normaal beschikbaar. Omdat routes naar de storage nodes in BIT-2C geadverteerd bleven én het storage netwerk, ook in BIT-2C, zelf wel beschikbaar was, leken de nodes in BIT-2C beschikbaar. De clients die van het storage platform gebruik maken ontvingen dan ook routes voor het access netwerk naar de storage nodes in BIT-2C. Als gevolg van ECMP werd circa een derde van het storage verkeer naar BIT-2C gestuurd, ook door clients in andere datacenters. Omdat al dat verkeer niet correct afgehandeld werd, was alle shared storage voor de clients effectief onbruikbaar.

1 https://www.bit.nl/news/2773/88/Deep-Dive-in-het-colocatienetwerk-van-BIT

Het shared storage platform van BIT wordt gebruikt voor shared storage diensten, alle virtual machines op het virtualisatieplatform en alle shared diensten van BIT waaronder email en web. Zodra de problematische switches weer correct functioneerden kwamen alle diensten die geen gebruik maken van CephFS mounts weer beschikbaar. Clients met CephFS mounts daarentegen hadden de storage nodes automatisch geblockt na 300 seconden onbereikbaarheid. Voor deze clients waren remounts of reboots nodig alvorens de mount weer beschikbaar kwam.

www.bit.org
Omdat www.bit.nl niet beschikbaar was, is het incident op www.bit.org gemeld. Deze melding wordt automatisch ge-SMS-t naar klanten die hebben aangegeven SMS'jes over incidenten wensen te ontvangen en daarnaast wordt de melding in een RSS feed geplaatst. Grote aantallen bezoekers die in korte tijd de incidentmelding op www.bit.org bekeken zorgde voor een overload op de webserver voor www.bit.org.

Na een eerder incident waarbij www.bit.nl niet bereikbaar was en www.bit.org de load niet aan kon zijn er maatregelen genomen om die site meer bezoekers te kunnen laten verwerken. Bij dit incident bleek dat deze maatregelen niet voldoende effectief zijn geweest. De hoge load op www.bit.org wordt mede veroorzaakt door klanten die in communicatie naar hun klanten de incidentmelding op www.bit.org benoemen. Dit zou echter geen probleem voor de bereikbaarheid van www.bit.org mogen opleveren.

Na een reboot van de server die www.bit.org serveert was de site weer bereikbaar.

Verbeterpunten
Om de kans en impact op herhaling van een dergelijke verstoring te verkleinen, zijn er een aantal maatregelen genomen en zullen andere nog genomen worden:

- Alle switches zijn gecontroleerd op de configuratie die risico's oplevert. Behalve de configuraties voor de specifieke klant, waren er geen andere poorten/port-channels voorzien van deze configuratie.
- De switch leverancier is ingeschakeld om gezamenlijk het incident te evalueren. Indien deze evaluatie tot gewenste aanpassingen leidt zullen deze doorgevoerd worden.
- De switch leverancier is gevraagd om te onderzoeken welke mogelijkheden er zijn om ervoor te zorgen dat resource starvation minder of geen impact op functioneren heeft.
- www.bit.org zal in aangepaste vorm gehost worden waarmee het risico op onbereikbaarheid bij pieken in de bezoekersaantallen beperkt zal worden.

Update 01-03-2021
De bovenstaande rapportage is op 18 februari 2021 gepubliceerd, een dag na het incident. Op dat moment was er enige onzekerheid over een aantal aspecten van dit incident. Inmiddels is er zekerheid over deze aspecten en is daarom het incident rapport geüpdatet.

- Enkele reacties op het gepubliceerde rapport betroffen vragen over het risico dat een zelfde incident zich opnieuw zou voordoen. Dat risico is afwezig. De configuratie die (mede) aanleiding gaf tot het incident is niet meer aanwezig in het netwerk. Op woensdag 17 februari is die configuratie overal verwijderd. Overigens werd deze configuratie slechts voor één klant gebruikt.
- Het opnieuw inzetten van de specifieke configuratie is onmogelijk gemaakt in de scripts waarmee configuraties op de switches geplaatst worden.
- Een gezamenlijk analyse van het incident met de switchleverancier heeft duidelijk gemaakt dat de forwarding plane van de switch had moeten blijven functioneren ondanks de resource starvation. De switchleverancier heeft dit probleem daarom als bug aangemerkt en zal werken aan het oplossen van deze bug. Als in toekomstige releases van de firmware de bug afwezig is zal dit overigens geen reden zijn voor BIT om de betreffende configuratie opnieuw in te gaan zetten.
- De switchleverancier heeft gecontroleerd of de bug ook door andersoortige uitputting van de hardware resources getriggerd kan worden. De leverancier heeft ons laten weten dat dat niet het geval is.
- Nadere bestudering van de server die www.bit.org host heeft geleerd dat de website niet onder load bezweken is. De server had te kampen met IRQ errors die getriggerd

werden door load op de server. De server is geüpdatet en heeft sindsdien geen errors gelogd, ook niet onder loadtests. Desalniettemin zal de server vervangen worden. Naast de monitoring die al actief was op de server zijn er regelmatige loadtests voor deze server ingeregeld.

Contact
Mocht u naar aanleiding van dit rapport vragen hebben, dan kunt u contact opnemen met onze afdeling Customer Care via 0318 648 688 of support@bit.nl.