RAPPORT INZAKE CEPH STORING - 14-04-2023

Samenvatting

Reguliere operaties op het Ceph shared storage cluster van BIT lopen op vrijdag 14 april om 11.55 uur tegen een tot op dat moment onbekende bug in de Ceph code aan. Het cluster wordt onbeschikbaar en hersteloperaties zorgen ervoor dat op 15 april om 00.39 uur het cluster weer beschikbaar is. Er is geen data corrupt geraakt of verloren gegaan als gevolg van dit incident.

Achtergrond

Ceph is een software-defined storage oplossing waarmee data redundant kan worden opgeslagen en via het netwerk beschikbaar gemaakt kan worden. Het staat bekend om zijn robuustheid, betrouwbaarheid, schaalbaarheid en flexibiliteit. Een enigszins vereenvoudigde beschrijving van de werking van Ceph en de keuzes die BIT qua configuratie gemaakt heeft staat hieronder.

Data in het Ceph cluster bij BIT wordt drie keer opgeslagen. Het cluster bestaat uit honderden harde schijven (SSD, NVMe) verdeeld over tientallen storage servers, die weer verdeeld zijn over drie fysieke, geografisch gescheiden, locaties. Elk van die locaties is een apart failure domain, wat betekent dat infrastructuur op zo'n locatie mag falen zonder dat de dienstverlening erdoor verstoord raakt.

Elke harde schijf is binnen het Ceph-cluster een Object Storage Device (OSD). Elk stuk data wordt ondergebracht in drie verschillende Placement Groups (PG). De PG's worden verdeeld over de OSD's op zo'n manier dat er nooit twee PG's op dezelfde fysieke locatie staan. Data voor verschillende diensten zit in verschillende pools, een pool bestaat uit één of meer PG's.

Er zijn vijf (monitoring) servers die in een Cluster Map de status van het cluster bijhouden. Deze zelfde servers stellen deze Cluster Map beschikbaar aan clients. Clients bepalen met behulp van de Cluster Map waar de data beschikbaar is waar ze van willen lezen of naar willen schrijven. Zij gebruiken daarvoor de primary PG, dat is één van de drie PG's. Bij uitval van een harde schijf of een storage node of zelfs een heel datacenter (failure domain) wordt een nieuwe Cluster Map gemaakt en de data van de twee overgebleven locaties gelezen en is er geen impact op de dienstverlening.

Data integriteit is leidend geweest bij het ontwerp van het BIT Ceph cluster en wordt belangrijker geacht dan beschikbaarheid. Het idee daarachter is dat het altijd mogelijk is om correcte (integere) data weer correct beschikbaar te krijgen en dat het niet zeker is dat corrupte data weer correct beschikbaar komt. Dat betekent dat zodra gedetecteerd wordt dat er nog maar één kopie van de data beschikbaar is, dat deze resterende kopie automatisch offline wordt gehaald om mogelijk dataverlies te voorkomen.

Enkele weken geleden is er een geavanceerde optimalisatie doorgevoerd op het cluster, nadat deze instelling op testclusters getest is. Deze optimalisatie zorgt ervoor dat de primary PG's beter over de verschillende OSD's in het cluster verdeeld worden en resulteert in een betere 'read performance' van het cluster. Deze instelling blijkt conditioneel voor het incident te zijn.

Dit storage platform is de basis voor alle cloud en shared diensten bij BIT, waaronder de 'harde schijven' van virtuele servers, storage voor virtual datacenters, e-mail en webruimtes.

bit
internet technology

Details

11.54 Er wordt een nieuwe storage node toegevoegd aan het cluster, zoals dat in het verleden ook vaker gedaan is. De monitoring servers maken een nieuwe Cluster Map en die wordt gepubliceerd.
11.55 Een groot aantal OSD's, in alle drie de failure domains, segfaulten (crashen) en aborten en dat zorgt ervoor dat de meeste diensten die van Ceph gebruik maken onbeschikbaar worden. De beheerders verliezen hun toegang tot het cluster.
11.56 Via andere ingangen herstellen de beheerders hun toegang tot het cluster.
12.04 De incidentmelding wordt geplaatst op www.bit.org.
12.10 De OSD's van de zojuist toegevoegde storage node worden gestopt.
12.15 Verschillende gecrashte OSD's worden herstart maar geen ervan komt online.
12.25 Diverse logs van het cluster worden geanalyseerd en er wordt gecontroleerd of er tegen bekende bugs aan gelopen is.
13.19 Een conference call met experts van Croit wordt opgezet. Croit levert Ceph consultancy, en er werken diverse (core) developers van Ceph.
13.34 Kernel en Ceph logs van één storage node worden gedeeld met Croit.
14.43 Verschillende OSD's worden herstart waarvan er één extra OSD online komt.
14.55 Ceph debug packages zijn geïnstalleerd en worden gestart op een storage node.
15.37 Een framework voor analyse van threading- en memory-gebruik door software wordt geïnstalleerd en gestart op een storage node.
16.31 Specifieke logregels van alle storage nodes worden gedeeld met Croit.
17.40 De monitoring daemons op de monitoring servers worden herstart.
17.45 Alle OSD's van een enkele storage node worden herstart, er komen geen extra OSD's online.
18.02 Alle OSD's van alle storage nodes worden herstart, er komen een aantal extra OSD's online.
18.10 Alle servers en nodes in het cluster worden herstart, er komen geen extra OSD's online.
18.20 Op basis van de logs wordt vermoed dat de OSD's zijn gecrasht en blijven crashen omdat er 'defecte' PG's op de OSD's staan.
18.45 Er is een 'defecte' PG op een OSD gevonden waarvan er nog wel twee kopieën op andere OSD's online zijn. Deze 'defecte' PG wordt verwijderd van de specifieke OSD en de OSD komt online.
19.05 Er wordt extra capaciteit in het cluster bijgeschakeld nu duidelijk wordt dat wellicht een aantal OSD's weer online gebracht kunnen worden.
19.43 Een andere 'defecte' PG wordt gedumpt en verwijderd, de OSD wordt zonder die PG gestart en komt online.
19.51 De 'defecte' PG wordt op dezelfde OSD geïmporteerd, de OSD crasht hierdoor.
20.18 Op alle storage nodes wordt een mount gezet naar een (redundant uitgevoerd) ander type storage.
20.34 Voor klanten die daar om gevraagd hebben wordt voorkomen dat hun virtual machines beschikbaar komen zodra storage weer functioneert.
21.04 De Ceph experts van Croit melden dat zij niet eerder een vergelijkbaar incident gezien hebben. Eén OSD zal daarom niet hersteld worden zodat die later voor forensics gebruikt kan worden.
21.16 De 'defecte' PG die om 19.43 uur gedumpt is, wordt succesvol op een lege OSD geïmporteerd, deze OSD komt en blijft online.
21.31 Op alle OSD's worden de 'defecte' PG's geïdentificeerd.
21.38 Er wordt een start gemaakt met het dumpen en verwijderen van de 'defecte' PG's op alle OSD's naar de mount met ander type storage. Zodra een OSD vrij is van 'defecte' PG's wordt die OSD succesvol online gebracht.
23.13 Diverse diensten beginnen weer beschikbaar te komen omdat er minimaal twee OSD's waar hun data op staat online zijn.
23.30 Er wordt een start gemaakt met het repareren van diensten waarvan de storage weer beschikbaar is.

00.06 Het cluster is aan het recoveren. PG's die op nog maar één of twee failure domain(s) beschikbaar zijn, worden gerepliceerd zodat ze op alle drie de failure domains beschikbaar komen.
00.39 Alle pools en daarmee alle data zijn weer beschikbaar.
02.48 Alle diensten die BIT in monitoring heeft zijn hersteld.
03.02 Het recoveren is gereed, alle PG's zijn weer op alle drie de failure domains beschikbaar. Het balanceren van het cluster loopt nog uren door.
03.03 Het incident wordt afgemeld.

Conclusie

Het cluster is geoptimaliseerd met instellingen die primary PG's beter over de OSD's verdelen. Door het toevoegen van extra capaciteit in het cluster wordt er een nieuwe Cluster Map gemaakt en beschikbaar gesteld aan clients. Het gebruik van deze nieuwe Cluster Map in combinatie met de optimalisatie instellingen loopt tegen een bug aan, waardoor PG's in het cluster defect raken. Deze defecte PG's zorgen ervoor dat OSD's crashen en dat resulteert in onbeschikbaarheid van pools.

De gecrashte OSD's zijn te herstellen door de defecte PG's van de OSD's te verwijderen. Deze defecte PG's kunnen wel succesvol geïmporteerd worden op lege OSD's. BIT heeft de bug en de hersteloperaties in een testcluster kunnen reproduceren. Door de Ceph developers moet nog geanalyseerd worden in welk deel van de code de bug is geïntroduceerd.

Verbeterpunten

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

- De tests op testclusters worden geëvalueerd en waar nodig uitgebreid of aangepast met meer verschillende operaties, zoals die ook in productie te verwachten zijn.
- De optimalisatie instelling wordt niet meer actief gemaakt op het Ceph cluster van BIT zolang de bug in de Ceph code aanwezig is.
- Een sample van de defecte data en een beschrijving van de condities waaronder de bug optreedt is gedeeld met de developers van Ceph. Zij hebben het bug rapport in onderzoek genomen, de bug bevestigd en zij zullen een verzoek voor aanpassing van de Ceph code indienen.

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.