

EFFICIËNTIE IN OPSLAG EN BESCHIKBAARHEID

De heer Arnt de Gier (NetApp)

De heer Arnt de Gier is werkzaam bij NetApp (Network Appliance) als adviseur in back-up en restore oplossingen. Met bijna 20 jaar ervaring op dit gebied is Arnt in staat alle categorieën van de open systemen omgevingen van advies te voorzien over data-beschikbaarheid en beveiliging.

Dit artikel gaat nader in op de efficiëntie in opslag en beschikbaarheid van informatie: eenmalig opslaan, eenmalig distribueren en overall beschikbaar.

UITDAGINGEN INFORMATIEVOORZIENING: INFORMATIE, NU, ALTIJD EN OVERAL

In de moderne wereld van internet is het gewoon geworden altijd onze informatie aan onze vingertoppen beschikbaar te hebben. Dat doen we via onze mobiele telefoon die geëvolueerd is tot een slimme internetclient en waarop we eenvoudig allerlei informatie van het internet kunnen plukken ongeacht waar we ons bevinden, maar ook via laptops of andere computers met een internetaansluiting.

Als dit in het dagelijks leven al gewoon is, dan zouden we verwachten dat de informatie-toegankelijkheid ook gewoon is tijdens uitvoering van speciale missies en op de meest afgelegen plaatsen in de wereld. Dit blijkt niet zo eenvoudig als in eerste instantie gedacht.

Voor het bereikbaar maken van informatie zijn we afhankelijk van netwerken, zoals publieke netwerken die gebruikt worden voor mobiele telefonie en draadloze internetvoorzieningen. Maar in het geval van onze dagelijkse informatievoorziening kunnen we natuurlijk geen publieke netwerken toepassen vanwege de gevoeligheid van de informatie.

Daarbij komt dat zodra we ons op afgelegen plaatsen bevinden ineens blijkt dat er geen netwerken beschikbaar zijn. Dan moeten we zoeken naar alternatieven, en die worden geboden door satellietverbindingen. Echter hebben satellietverbindingen een aantal afwijkende eigenschappen. Zo is dataverkeer via satellieten veel duurder en hebben we te maken met de tijd die een transmissie nodig heeft via een satelliet. Deze is langer omdat onze informatie een veel langere weg aflegt, een dergelijke satelliet staat immers 36000 kilometer boven de aarde. Dergelijke vertragen noemen we latency.

Om toch toegang te hebben tot de nodige informatie moeten we daarom een aantal extra voorzieningen treffen. Ten eerste moeten we zorgen dat we zo min mogelijk data oversturen. Dit doen we door data die al eens opgevraagd is te bewaren, zodat bij herhaling deze data niet via de dure satellietver-

binding aangevraagd hoeft te worden, Dit wordt caching genoemd. Ten tweede willen we graag dat informatie die we op een locatie verzamelen ook beschikbaar komt voor andere locaties en het centrale datacenter. Hiermee kunnen we analyses doen en wordt het mogelijk betere strategieën uit te zetten.

TECHNOLOGIE VOOR OPSLAG INFORMATIE

Waar het gaat om opslag van informatie en de beveiliging tegen verlies of corruptie, zijn er diverse mogelijkheden. We kunnen de data op meerdere plaatsen opslaan (ook wel spiegelen genoemd). Dat geeft een beveiliging tegen totaalverlies van de data, maar niet tegen onwillekeurige wijziging van data, omdat een eventuele wijziging ook gespiegeld wordt. Dus gaan we op zoek naar de aloude back-up technologieën. Die back-up technologieën waren 25 jaar geleden zeer eenvoudig. We hadden weinig data en er bestonden systemen die voldoende capaciteit hadden om elke dag onze data als een back-up op tape te zetten. Dat is nu nogal veranderd! We hebben nu meerdere Terabytes (TB) aan data, en die past niet op een enkel back-up medium en kan ook niet dagelijks gebackupt worden omdat dit simpelweg te lang duurt. Back-up software-makers hebben daar natuurlijk al lang mooie nieuwe technieken voor bedacht waarmee op werkdagen bijvoorbeeld alleen die nieuwe data gebackupt wordt. We noemen dit incrementele back-ups. Deze incrementele back-ups bekijken welke bestanden er nieuwer zijn dan de vorige back-up en zorgen voor een nieuwe kopie op de back-up omgeving. Mocht een bestand geopend en opgeslagen zijn, maar niet veranderd, dan maken deze technieken toch een nieuwe kopie, wat de efficiëntie niet ten goede komt.

Ook hierop kunnen we besparingen halen, door niet op bestandsniveau een back-up te maken maar op blokniveau. Een bestand wordt namelijk als een set blokken op een harddisk opgeslagen. Er worden nu minder blokken gekopieerd tijdens een back-up, maar omdat alleen nieuwe blokken gekopieerd worden, krijgen we te maken met fragmentatie in de back-up omgeving. Stel-

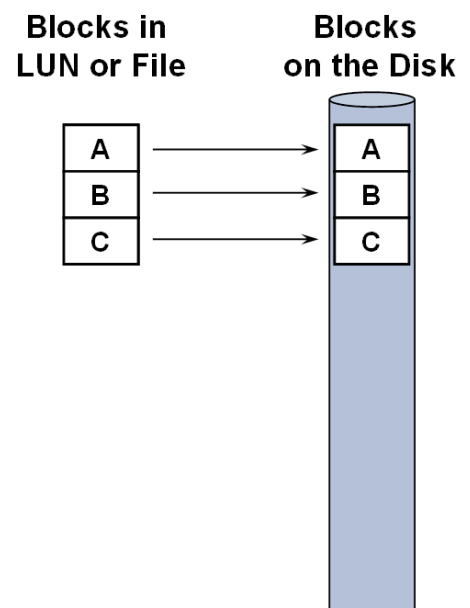


len we ons voor dat een bepaald bestand dagelijks gewijzigd wordt, dan worden elke dag de nieuwe blokken van het bestand gekopieerd en raken deze blokken dus verspreid over meerdere stukken back-up media. Dit geeft weer uitdagingen bij het doen van restores.

We hebben dus nog meer intelligentie nodig in de opslag van de data

MEER EFFICIËNTIE: SNAPSHOTS

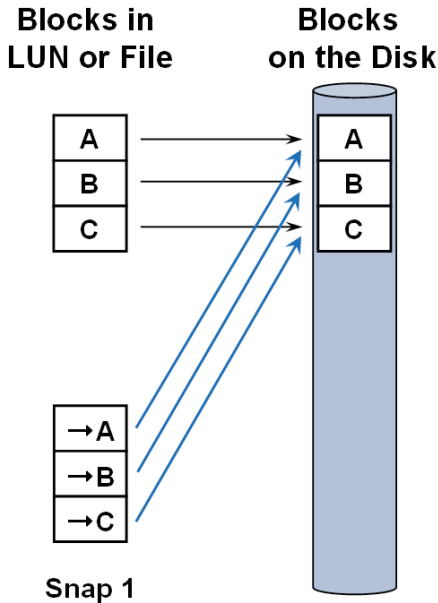
Een totaal andere manier van het aanleggen van back-ups is het maken van zgn. snapshots. Deze snapshots zorgen dat een bestandssysteem op zeker moment bevroren wordt en dat die nieuwe te schrijven blokken apart worden weggeschreven. Uiteraard kunnen nieuwe wijzigingen en nieuwe bestanden nog steeds weggeschreven worden. Dat werkt ongeveer als volgt.



Afbeelding 1



Stel een bestand bestaat uit 3 blokken A, B, en C. Deze blokken worden op de disk geschreven en via het bestandssysteem wordt er een representatie gemaakt die voor de mens te bevatten is, namelijk een bestandsnaam. Deze naam is hier niet relevant.

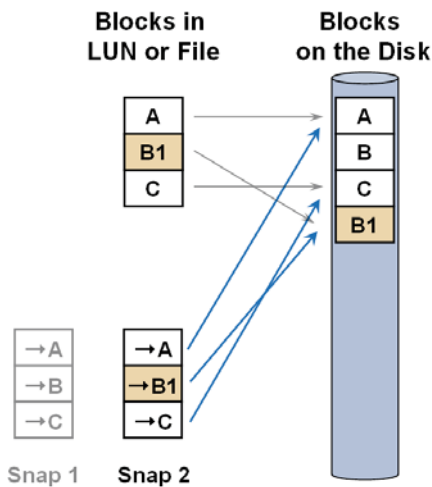


Afbeelding 2

Van dit bestand kunnen we nu een back-up aanleggen, zonder dat we ook maar enige data verplaatsen of kopiëren. Een dergelijke back-up noemen we een snapshot.

Dergelijke snapshots kunnen zeer snel gemaakt worden, ook bij grote hoeveelheden data, omdat er geen data blokken gekopieerd worden. Het enige dat gekopieerd wordt zijn de pointers (ook wel metadata genoemd), en die is per definitie veel kleiner dan de data zelf.

We zien nu de bestaande presentatie van het bestand als ABC en de back-up van het bestand, ook als ABC.

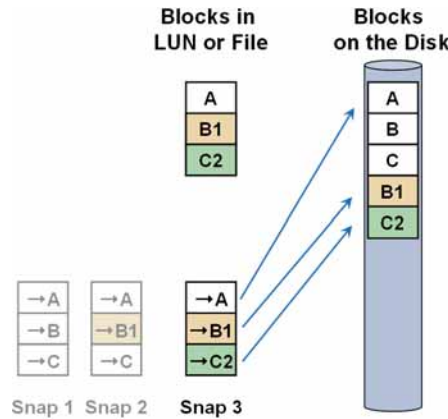


Afbeelding 3

Nu gaan we het bestand wijzigen. Normaal gesproken zouden we dan een wijziging over de bestaande data heen schrijven, maar dit

kunnen we nu niet doen omdat de blokken die daarbij horen als het ware vast zitten in het snapshot. Daarom wordt de gewijzigde data op een nieuwe plaats geschreven, B1. Na deze wijziging kunnen we nog steeds de originele versie van het bestand (ABC) benaderen, en is de wijziging voor de gebruiker of applicatie transparant doorgevoerd op de disk.

Nu kunnen we een nieuwe snapshot maken om de nieuwe versie van het bestand ook weer veilig te stellen. Snapshots worden vaak elk uur gemaakt voor een grotere dichtheid dan die praktisch haalbaar is met traditionele back-up methoden zoals eerder besproken.



Afbeelding 4

Ook nieuwe wijzigingen kunnen zonder enige extra overhead opgeslagen worden en van snapshots worden voorzien.

Dit brengt ons een bestandssysteem met ingebouwde back-up. Bij verlies van de disk(s) zijn we daarbij ook de back-up kwijtgeraakt. Daarom kunnen we deze snapshot-technologie verder toepassen voor het aanleggen van een volledige kopie op een andere locatie of in het centrale datacenter, zodat de data zowel op meerdere plaatsen benaderbaar is en tegelijk ook voorzien is van een zeer efficiënte back-up.

Deze techniek wordt bij vele klanten toegepast, waarbij een aansprekend voorbeeld is het Formule 1 team van Renault, die deze technologie toepast in het centrale datacenter maar ook telkens meeneemt naar de pitbox tijdens een raceweekend. De technologie stelt Renault F1 in staat om telemetrie-informatie uit de race-auto's op te slaan, in bijna real time door te sturen naar de ontwikkelafdeling, maar ook om de laatste ontwikkelingen naar de pitbox te sturen. De volgende stap in de race naar hogere efficiëntie is het centraliseren van de data, zodat er minder boxen naar een compound gestuurd hoeven te worden. Feit is namelijk dat als elke serverbox zijn eigen opslag heeft (lees eigen HDD), dat deze HDD slechts voor een klein deel van hun capaciteit benut worden en daardoor de efficiëntie van de

aanwezige opslagcapaciteit heel laag is, vaak lager dan 30%.

Als we de totale benodigde capaciteit voor een compound centraliseren, kunnen we een veel hogere efficiëntie bereiken, omdat de vrij beschikbare ruimte voor elke applicatie en wellicht gebruiker aan te spreken is. Daardoor hebben we minder en kleinere disks nodig en besparen we flink op kosten van opslag, energie en koeling.

Samen met de efficiëntie van snapshots en spiegeling over en weer naar centrale opslag, is het mogelijk om meer informatie actueel te houden op de compound tegen geringere kosten!

Met deze technologie beschikken we over mogelijkheden om meer data beschikbaar te maken en te houden.

TAGS

Daarbij hoort nog steeds de uitdaging, hoe houden we bij waar al die informatie te vinden is en hoe zorgen we ervoor dat alles up-to-date is? Daarvoor hebben we een zgn. documentmanagementsysteem nodig. Een dergelijk systeem stelt ons in staat om ieder stukje informatie te voorzien van een aantal kenmerken, die we bij het terugzoeken van de informatie weer kunnen gebruiken. We kennen vast wel het verschijnsel van digitale fotografie in de huiselijke sfeer. We maken foto's, bewaren die vervolgens op een computer, en kunnen ze dan niet meer vinden omdat de bestandsnaam alleen (bijv. DSC00451.jpg) ons zo weinig zegt. Het is dus van het grootste belang dat alle informatie wordt voorzien van steekwoorden zgn. tags. Het documentmanagementsysteem kan deze tags opslaan en maakt het mogelijk te zoeken op de tags.

BRONNEN:

NetApp creates innovative storage and data management solutions that help our customers accelerate business breakthroughs and achieve outstanding cost efficiency. Our dedication to principles of simplicity, innovation, and customer success has made us one of the fastest-growing storage and data management providers today.

Customers around the world choose us for our 'go beyond' approach and broad portfolio of solutions for business applications, storage for virtual servers, disk-to-disk backup, and more. Our solutions provide nonstop availability of critical business data and simplify business processes so companies can deploy new capabilities with confidence and get to revenue faster than ever before.

<http://www.netapp.com/nl/>