SSDs: hoe, wat, waar, welke en welke niet.

Toepassen van consumer of datacenter SSDs ?

Achtergrond.

Newszilla bestaat uit diverse clusters van servers. 1 cluster is het frondend, waar de client connections op uitkomen. Deze draait de 'dreaderd' software, die connecties van clients aanneemt, en requests voor articles naar het juiste backend doorstuurt. Daarnaast krijgt elk frontend een header-only article feed, die gebruikt wordt om een lokale header-database (gebruikt voor de overview) bij te houden. Deze database bestaat uit een verzameling files op een filesystem, dat op SSD staat.

Dit heeft jarenlang tot tevredenheid gedraaid met de header-database op een RAID0 van 7 Intel X25-M 160GB SSDs, totdat opeens op 1 frontend die raid ging stotteren- er is waarschijnlijk 1 SSD stuk, maar het is niet uit te vinden welke. Individueel testen ze allemaal OK, en als er iets fout gaat slaat de controller vast, zonder logs, zodat niet te zien is welke SSD het probleem was. Uiteindelijk zijn de 7 160GB SSDs in RAID0 vervangen door 1 Crucial M500 960GB SSD disk.

Dit heeft ook een tijd correct gedraaid, totdat dezelfde symptomen optraden. Het probleem was niet direct duidelijk. Maar de machine was een tijdje down geweest en had om te herstellen een up-to-date kopie van de headerdatabase nodig. Dat is standaard procedure; die kopieeren we vanaf een andere frontend. Bij het kopieeren bleek dat de Crucial SSD zich raar gedroeg; write throughput varieerde van 10-25 MB/sec, met stalls van tientallen secondes, soms zelfs een minuut. Wat was hier aan de hand ?

SSD Garbage collection

Een SSD is opgedeeld in blokken van 64K tot 512K, afhankelijk van het type SSD. Dit noemt men erase-blocks en de grootte van deze blokken wordt de 'erase block size' genoemd. Dat komt omdat je cellen op een SSD alleen kan erasen per zo'n blok tegelijk. Een SSD kan geschreven worden met een minimum blocksize van 4K. Tenminste, zolang de SSD leeg is. Als je een blok van 4K dat al eens geschreven is wilt herschrijven, dan moet die eerst ge-erased worden. En dat kan alleen met een heel erase-block tegelijk. Dus als een SSD volloopt, en alles is al eens beschreven, betekent dat voor elke write het erasen en herschrijven van een volledig erase-block. Daarvan sluit de SSD hard en het is enorm traag.

Daarom hebben fabrikanten slimme truken ontwikkeld zoals remapping. Als je een block van 4K wilt herschrijven, en er is ergens nog 4K 'clean', dan wordt die 4K beschreven, en het oude block gemarkeerd als 'ongebruikt maar dirty' (oftewel 'garbage'). Uiteindelijk zijn alle blokken een keer beschreven dus 'gebruikt' of 'dirty' en dan werkt deze remapping truuk niet meer.

Op dat punt moet de disk aan 'garbage collection' gaan doen: de erase-blocks erasen en herschrijven zodat de ongebruikte 'dirty' blocks weer clean zijn. Vaak wordt ook aan 'compacting' gedaan, het dicht bij elkaar zetten van gebruikte blokken om gatenkaas te voorkomen. Verder wordt ook gezorgd dat blokken die al een hele tijd niet meer beschreven zijn, verwisseld worden met erase-blocks waar veel activiteit op plaatsvind, om ervoor te zorgen dat gemiddeld over alle erase-blocks ongeveer net zo vaak een program/erase cycle plaatsgevonden heeft.

Alle SSDs zijn over-provisioned; er is extra ongebruikte ruimte die goed van pas komt bij het heen en weer schuiven van blokken tijdens garbage collection. Hoe meer ongebruikte ruimte hoe beter dat gaat.

Sommige SSDs (consumer SSDs) doen alleen garbage collection als de SSD 'idle' is, of als het niet meer te vermijden is (SSD vol met garbage). Dit vanwege de interne constructie (bv reads/writes kunnen niet tegelijk met erase), of omdat de CPU op de SSD niet krachtig genoeg is. Sommige SSDs doen garbage collection serieel (blok-voor-blok); anderen kunnen het parallel (meerdere blokken tegelijk). Verder is het gebruikte gabrage collection-algorithme ook van belang. Al deze factoren hebben invloed op (voornamelijk de write-)performance van de disk.

Verder is het nog mogelijk voor het OS om tegen de disk te zeggen "hee, deze blokken heb ik niet meer in gebruik", bijvoorbeeld als een file verwijderd wordt. De SSD weet dan dat die blokken gemarkeerd kunnen worden als 'garbage' ipv 'in gebruik' en kan evt direct een garbage-collection cycle starten. Zeker als de blokken 1 of meerdere erase-blocks bestrijken. Dit is bij SATA het 'TRIM" commando dat het OS via de disk-driver naar de SSD kan sturen. Dit is een non-queued commando, dus het veroorzaakt een 'hickup' in de I/O naar de disk. Als dit gebruikt wordt is het raadzaam om dit niet 'live' te laten doen door het filesystem (dat doe je overigens door het filesystem met de 'discard' optie te mounten), maar via cron met het 'fstrim' commando op een rustig moment.

Steady-state performance

Als TRIM niet gebruikt wordt, dan zal op een gegeven moment de hele disk beschreven zijn. Alle blocks zijn 'written', niks is 'dirty'. Elke nieuwe write heeft een volledige erase-rewrite cycle van een erase-block nodig. Zoals genoemd hebben SSDs gereserveerde ruimte, dus er is nog wel wat plek om dingen heen en weer te schuiven.

Als je googled op 'SSD steady state performance' dan zie je tests die eerst een lege SSD testen met 4K random writes, dan de SSD volledig beschrijven (een volledige Erase / Program cycle) en daarna dezelfde test weer doen. Bij een disk als de Crucial M500 zie je dat de performance op een lege SSD boven de 60000 IOPS is. Maar met een volle disk dondert het naar beneden naar 8000 IOPS. 8000x4K = 32 MB/sec.En dat is nog veel, door stalls is het vaak veel lager. De gemiddelde latency gaat omhoog, en af en toe zie je dus complete stalls.

Maximum number of P/E cycles

Een block op een SSD kan slechts een beperkt aantal malen beschreven en gewist worden, daarna gaat dat blok fysiek stuk. Bij het continue schrijven op een SSD zullen door een correcte garbage collector alle erase-blocks een keer ge-erased en herschreven zijn, ook als slechts een deel van de data op de disk herschreven wordt. Het schrijven van een hoeveelheid data net zo groot als de capaciteit van de disk noemt men "1 Program-Erase cycle" oftwel "1 P/E cycle". De fabrikant geeft op wat het gegarandeerd maximum aantal P/E cycles is. Zodra je daar overheen gaat geen garanties meer dat het blijft werken; op dat moment houdt dan ook inderdaad de garantie van de fabrikant op. Sommige (Intel) SSDs maken de disk uit voorzorg na het maximum P/E cycles keihard read-only, om te garanderen dat er geen data-loss zal zijn.

Consumer en datacenter SSDs.

De meeste SSDs die op de markt zijn zijn bedoeld als 'consumer' drive. Dat houdt in dat het verwachte gebruikspatroon bestaat uit weinig writes, veel reads, en veel idle time. Mogelijk ook regelmatig een opschoon-actie door het OS dmv TRIM commando's.

Als je zo'n disk in een RAID array zet, en gaat gebruiken voor database-achtige toepassingen met veel writes loop je tegen deze problemen op:

- bij het initialiseren van de RAID zijn meestal direct alle blokken beschreven en dus in gebruik.
- de disk zal weinig tot geen idle tijd kennen. SSDs die alleen backround garbage collection doen als de disk idle is, doen dan dus geen garbage collection
- het aantal P/E cycles zal snel oplopen
- er zijn **geen** RAID controllers die het TRIM commando doorgeven. Geen enkele fabrikant implementeert dit (om goede redenen, maar deze pagina is al lang genoeg). Dus TRIM op het filesystem gaat ook niet werken

Om hieraan tegemoet te komen, hebben fabrikanten zoals Intel en Samsung een aparte lijn SSDs, die men 'datacenter SSDs' noemt. Deze hebben de volgende kenmerken:

- Extra veel 'reserved space'
- Krachtige controller
- Hardware geoptimaliseerd voor continue en parallele garbage collection
- gemaakt voor veel meer P/E cycles dan consumer disks

Deze disks zijn wel geschikt voor RAID sets en/of databases. Dit zijn bijvoorbeeld de Intel DC S3700 productlijn en de Samsung 845DC PRO productlijn. Ze zijn wel **duur** ..

Kiezen van een SSD

Afhankelijk van het type gebruik kan een consumer of een datacenter SSD gekozen worden.

Voor consumer SSDs

- Kies een model dat niet al te veel performance verliest in 'steady state'
- Kies een model dat continue background garbage collection doet
- Gebruik alleen als je weet dat binnen 3 jaar niet het maximum aantal P/E cycles overschreden wordt
- bij gebruik in RAID of bij een zelf-reparerend filesystem (btrfs): pas op automatische periodieke 'scrubbing' waarbij alle data herschreven wordt. Dat kost P/E cycles!
- Gebruik gewoon liever niet in RAID 🙂
- Gebruik dus bij read-heavy applications (of, als OS disk die gewoon weinig te doen heeft)
- Laat bij voorkeur een groot deel van de disk ongebruikt, dit wordt vanzelf als scratch-space gebruikt voor garbage collection (zoals de reserved space). Doe dat wel al bij het partitioneren zodat die blokken ook **echt** nooit aangeraakt worden, want 1x gebruikt blijft gebruikt.

Voor datacenter SSDs

- Ze zijn stervens duur dus gebruik geen 800GB model waar 100GB voldoet. Ter illustratie, in de DC S3700 series, kost het 100GB model EUR 180,- en het 800GB model EUR 1400,-
- Als RAID noodzakelijk is prefereer datacenter SSDs
- Bij schrijf-intensieve workloads (databases, newsservers, logging)

Wat zijn op dit moment (eind 2014) de beste disks

Consumer SSDs:

- Samsung 850 PRO.
- Intel 730

Het max aantal P/E cycles van de Intel is apart aangegeven: 70 GB/day voor het 240 GB model met 5 jaar garantie [1] , dus dat is net 500 P/E cycles. Ze geven echter voor het 480GB model hetzelfde op. Samsung geeft aan 40 GB/day over 10 jaar [2], dus 80 GB/day over 5 jaar. Echter ook model onafhankelijk. Het 200 GB model krijgt dan 4x zoveel P/E cycles voor z'n kiezen als het 800GB model, dus waarschijnlijk kan het 800GB model stiekum makkelijk 160GB/day aan. Bij een test van de Samsung bij Anandtech [3] bleek dat het waarschijnlijk op minstens 3000 P/E cycles uitkomt.

Deze 2 disks presteren in steady-state veel beter als niet de hele disk in gebruik is [4]

Datacenter SSDs:

- Samsung 845DC PRO
- Intel DC S3700

De Samsung belooft 50000 IOPS in steady-state, en 10 P/E cycles per dag [5]. Intel geeft ongeveer dezelfde getallen (IOPS iets lager) [6].

Er zijn meer fabrikanten met een "DC" productlijn, zoals Crucial: er is een M500 DC. Echter, met Intel of Samsung kan je eigenlijk niet fout gaan, die hebben zich bewezen.

Conclusie

Samsung

Voetnoten

[1] http://www.cnet.com/products/intel-ssd-730-series-240gb/
[2] http://www.samsung.com/global/business/semiconductor/minisite/SSD/global/html/ssd850pro/overview.html
[3] http://www.anandtech.com/show/8239/update-on-samsung-850-pro-endurance-vnand-die-size
[4] http://www.anandtech.com/show/8216/samsung-ssd-850-pro-128gb-256gb-1tb-review-enter-the-3d-era/7
[5] http://www.anandtech.com/show/8236/samsung-ssd-global-summit-2014-845-dc-pro-with-vnand-sm951-with-nvme-support
[6] http://www.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-dc-s3700-series.html