



Flash memory nestelt zich in de enterprise storage-omgeving

Markt voor SSD groeit explosief dankzij prijsverlaging

Elke zichzelf respecterende storage vendor schermt tegenwoordig met SSD. Maar eigenlijk is er niets nieuws onder de zon. Solid State Drive (SSD)-technologie wordt immers al enkele tientallen jaren toegepast in de computer-industrie, voor het eerst in de mainframe-omgeving. Hoe zal de markt voor SSDs zich nu verder ontwikkelen?

Tot op heden vormde de toepassing van SSDs nauwelijks een factor van betekenis in de enterprise storage-omgeving. Tot ongeveer een jaar geleden, toen de prijs van flash memory aanmerkelijk begon te dalen. Dat de belangstelling voor SSD toeneemt, blijkt uit de aandacht die leveranciers van storage array's nu aan de dag leggen om SSDs steeds vaker te integreren in hun huidige storage array-producten. Daarmee kan een tiered

men voor transactiegebaseerde applicaties als database- en e-mail-servers en de nieuwe virtuele omgevingen, kan de toepassing van SSD een oplossing bieden voor de toenemende prestatieproblemen.

Het I/O bottleneckprobleem

Een van de prestatieproblemen bij transactiegebaseerde applicaties in de markt voor open systemen wordt veroorzaakt door het

teem. Dat is op het eerste gezicht toch verwonderlijk, want de meeste leveranciers van storage-systemen prijzen hun systemen toch juist aan met de pay off 'high performance' en 'high scalable'. Of is dat misschien toch niet de hele waarheid? Wanneer managers van datacenters worden geconfronteerd met prestatieproblemen bij applicaties, grijpen ze traditioneel terug op een aantal basistechnieken. De meest gangbare aanpak van het I/O-bottleneckprobleem is de toevoeging van meer servers. Nog steeds heerst de algemene misvatting dat een snellere processor het prestatieprobleem zal oplossen. Niet dus, want ook een snellere processor moet nog steeds wachten op de relatief tragere storage device. Een andere oplossing wordt vaak gezocht in het toevoegen van meer intern RAM. Maar dat helpt uitsluitend wanneer sprake is van een hoog percentage disk reads, gevolgd door writes van dezelfde data. Extra RAM helpt dan alleen omdat deze reads in RAM worden opgeslagen, waarbij het tijdelijk als een cache-systeem fungeert en de cpu niet hoeft te wachten op data afkomstig van de externe hard disks. Maar als disk reads overwegend random van aard zijn, dan helpt RAM cache echt niet zoveel meer. Meer disken kopen is dan te overwegen, maar is dat wel dé oplossing?

De toevoeging van meer spindles verhoogt meestal de I/O-prestaties door de toegang tot de data te spreiden over meer disk drives. Volgens adviesbureau StorageReview.com kan de snelste hard disk drive momenteel echter niet veel meer bieden dan 500 I/O's. In theorie is het wel mogelijk om met genoeg spindles een RAID-oplossing samen te stellen die het prestatieprobleem zou kunnen

SNELLERE PROCESSOR LOST

HET PRESTATIEPROBLEEM NIET OP

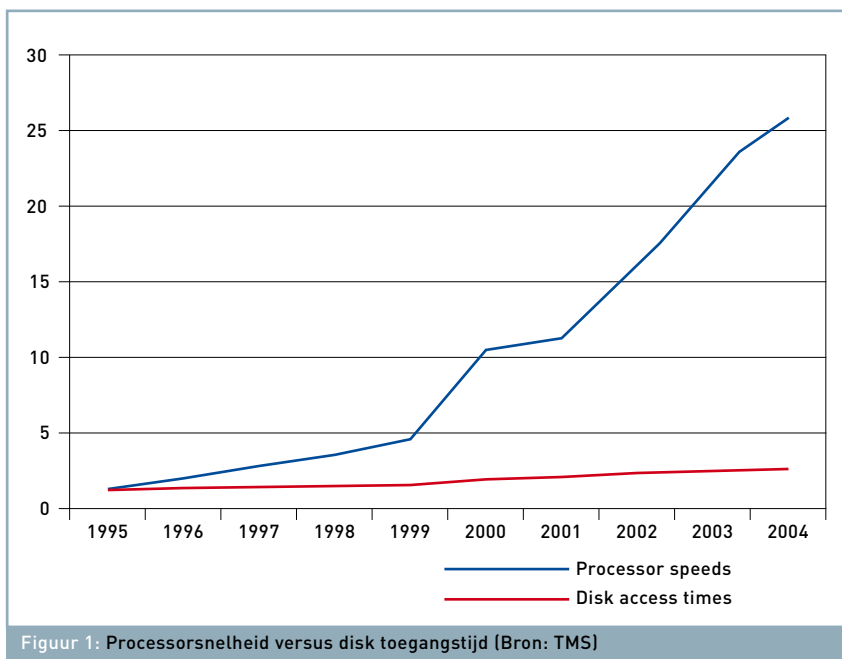
storage array model worden samengesteld met diverse storage layers op basis van respectievelijk SDD, Fibre Channel, SAS en SATA II disken, gerangschikt in afnemende volgorde als het gaat om prestaties. De stand-alone, op SSD-gebaseerde storage-producten, bijvoorbeeld de RamSan SSDs van Texas Memory Systems, bestaan al langer op de markt. Als gevolg van de toenemende behoefte aan hoogpresterende opslagsyste-

I/O bottleneckprobleem en leidt steeds vaker tot prestatieproblemen in het datacenter. Afgezien van slecht ontwikkelde applicaties, spelen drie systeemcomponenten daarbij een belangrijke rol: server, netwerk en opslagsysteem. De meeste datacenters voldoen wel aan de twee eerste van de drie gestelde eisen voor een hoogpresterende omgeving (snelle servers en netwerken), maar de component die vaak niet genoeg presteert is het opslagsys-

oplossen. Een snelle blik op rapporten van de Storage Performance Council (SPC-1 IOPS) leert echter aan dat leveranciers van RAID-systemen niet voor niets van honderden disk spindles met lage opslagcapaciteit gebruikmaken om de hoge prestaties van hun storage array aan te kunnen tonen. Het gebruik van low capacity disk drives is echter in tegenspraak met de trend van gebruikers: die willen juist van high capacity disk drives gebruikmaken om de opslagkosten te verlagen! Een ander punt is dat leveranciers van disk drives, door de dwingende factor van de dalende prijs per gigabyte, wel gedwongen worden om over de hele productrange over te gaan tot de fabricage van steeds meer high capacity drives. Het gevolg daarvan is dat disken met lage capaciteit toch steeds vaker worden uitgefaseerd. Een ander punt is dat de processorprestaties de laatste tien jaar met een factor 25 zijn toegenomen, terwijl de prestaties van disken nauwelijks tweemaal zijn verbeterd. De belangrijkste oorzaak van deze scheefgroei is gelegen in het feit dat de toegangstijden van hard disk drives, vanzelfsprekend als gevolg van de mechanische eigenschappen, geen gelijke tred kunnen houden met die van de processor (zie figuur 1).

Analyse

In een door Silverton Consulting Inc. uitgevoerde analyse werd het prestatieprobleem vanuit een geheel andere hoek benaderd. In een white paper met de titel *Disk-application performance gap and the future of Semiconductor Storage System* werd op basis van historische SPC-1 benchmarks en het Microsoft Exchange Solution Review Program (ESRP) aangetoond dat het prestatieverschil tussen disk en applicaties aan het toenemen is. Na een uitgebreide analyse van SPC-1 benchmarkrapporten kwam men tot twee verrassende conclusies. Met betrekking tot de gemiddelde disk subsystem I/O-prestaties constateerden de onderzoekers, ondanks een continue verbetering, een geleidelijke daling op capaciteitsbasis. Dit betekent dat het aantal IOs per gigabyte (IOPS/GB) nu onder de 4 IOPS/GB is aangeland. En hoewel de gemiddelde disk subsystemkosten voortdurend daalden op een subsystem per GB (SS\$/GB)-basis, bleven ze relatief gelijk op het niveau van subsystemkosten per disk drive (SS\$/Drv). Ze bedragen thans ongeveer \$2,850 SS\$/Drv. Na analyse van de SPC-1-rapporten werd aangetoond dat de disk subsystem-prestaties elk jaar met een 21 procent afnemen. Na extrapolatie van data tot november 2008 zouden de gemiddelde SPC-1 disk subsystem-prestaties zelfs zijn gedaald tot onder de 3,8 IOPS/GB (zie figuur 2).



Figuur 1: Processorsnelheid versus disk toegangstijd (Bron: TMS)

Aan de andere kant laat grafiek 3 op basis van SPC-1 resultaten een opwaartse trend zien van het aantal IOs per drive. Op het eerste gezicht lijken de twee trends met elkaar in tegenspraak, maar in feite zijn ze dat niet, want de opslagcapaciteit van een drive is elke twee jaar bijna verdubbeld. Deze verdubbeling levert een drempel van 21 procent prestatieverlies op, puur op basis van het IOPS/GB niveau. Dit heeft diskleveranciers er toe aangezet om gebruikers voornamelijk te wijzen op de algehele prestatieverbeteringen van storage arrays.

Prestaties van applicaties

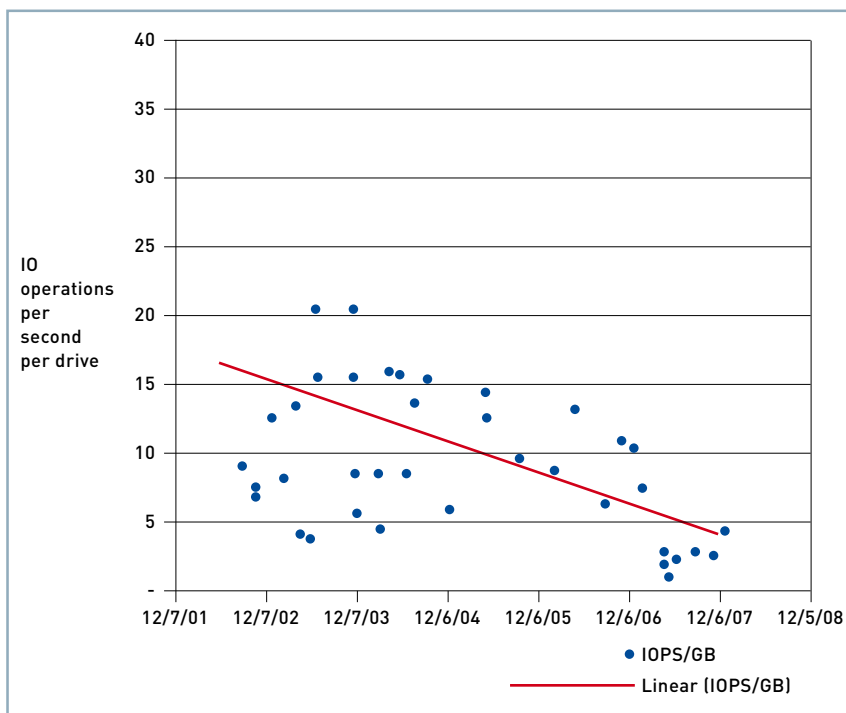
Silverton Consulting analyseerde ook de prestaties van het ESRP waarin een 9 procent jaarlijkse lineaire afname van de data-

base- en logprestaties werd aangetoond. In feite laten de huidige ESRP-resultaten ongeveer een 5,0 IOPS/GB zien (zie figuur 4).

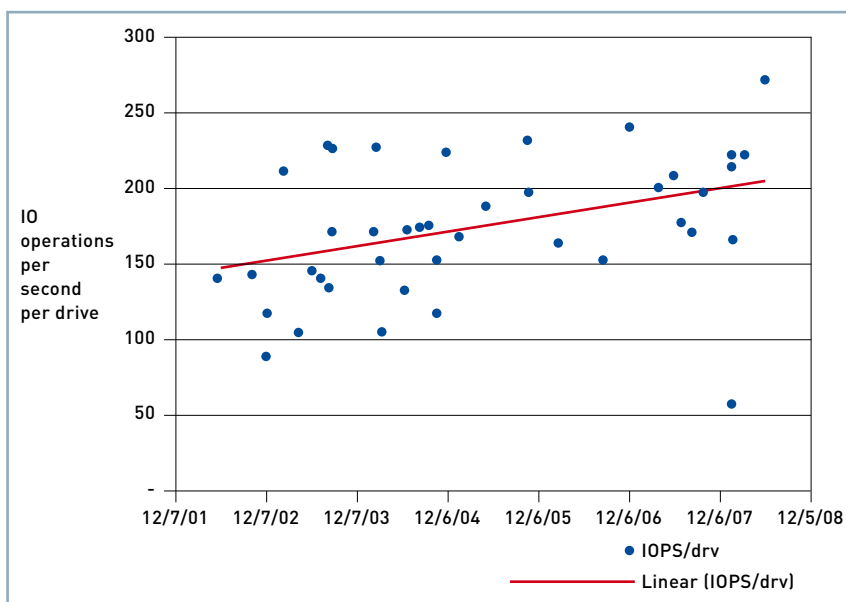
De combinatie van de 9 procent degradatie van transferbehoefden met de 21 procent neerwaartse trend van disk subsystem-prestaties resulteert in een nog steeds toenemende afstand tussen diskprestaties en applicatiebehoefden, zoals geïllustreerd in grafiek 5.

Traditioneel werd dit probleem opgelost door over-provisioning van disk storage, met andere woorden meer disken bijplaatsen. Dat deze 'opvoerpraktijken' de komende jaren geen oplossing meer zullen bieden voor het I/O-probleem zal duidelijk zijn. Er

(Advertentie)



Figuur 2: SPC-1 IOPS per GB (Bron: Silverton Consulting)



Figuur 3: In tegenstelling tot DRAM is flash-memory non-volatiele

is dus een andere oplossing nodig, mogelijk in de de toepassing van SSD in combinatie met een meerlaagse storagearchitectuur. Gegeven de kostprijs van disk storage kan de huidige SSD storage niet alleen een economische en technische oplossing bieden voor specifieke database-georiënteerde, maar ook voor mainstream applicaties.

Verskillende type SSD-technologieën

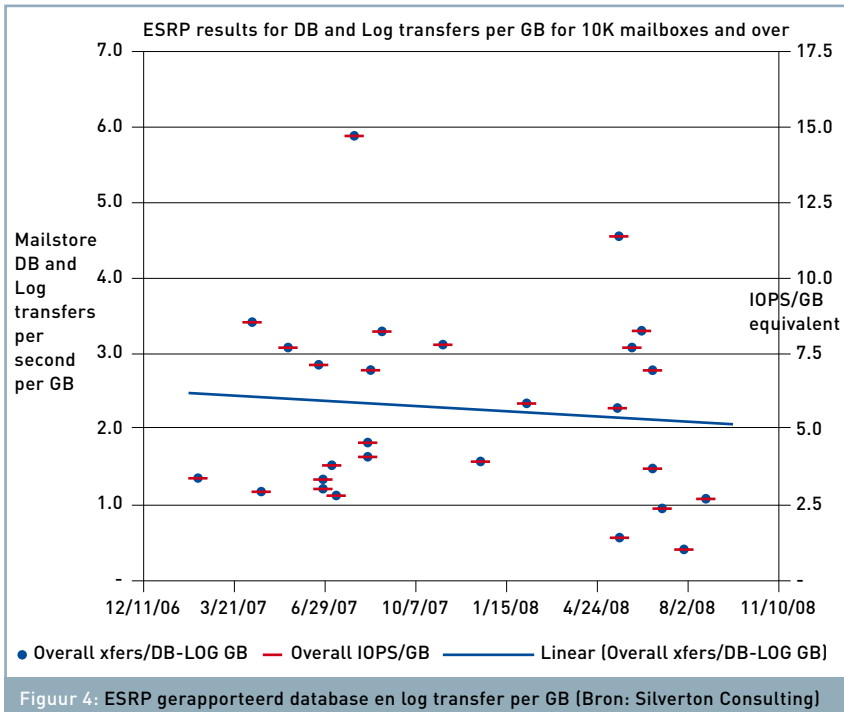
Het eerste type SSD-technologie van enige betekenis voor de enterprise storage-omgeving was Dynamic Random Access Memory

(DRAM) SSD. De belangrijkste onderscheidende eigenschap van DRAM SSD is volatilité. Dit impliceert dat zodra de voeding op de chip wegvalt, alle daarin opgeslagen informatie verloren gaat. Dat komt omdat DRAM van condensatoren gebruik maakt en als gevolg van lekstroom in loop van de tijd ontladen. Ze moeten daarom periodiek worden opgeladen. Bescherming tegen stroomuitval wordt geboden door interne back-up batterijen en de mogelijkheid om de in DRAM opgeslagen data naar een hard disk of een flash-gebaseerd geheugensysteem te verhuizen. Het belangrijkste DRAM SSD verkoop-

argument is snelheid en levensduur. Het is tot op heden de snelste technologie voor toegang tot data bij zowel lezen als schrijven. Nadeel is wel de hoge prijs en het stroomverbruik. Beide genoemde nadelen motiveerden leveranciers om te onderzoeken of de goedkopere flash-gebaseerde en meer energie-efficiënte SSD-technologie geschikt waren voor de toepassing van storage in de enterprise storage-omgeving. In tegenstelling tot DRAM is flash-memory non-volatiele, dat wil zeggen dat de data niet verloren gaat bij stroomafschakeling. Flash SSDs maken namelijk gebruik van zogenoemde floating-gate transistoren, die data voor langere tijd kunnen vasthouden. Een ander belangrijk verschil met DRAM is dat de toegang tot flash memory serieel gebeurt en afhankelijk is van een flash controller die de data uit de flash-chip ophaalt en deze parallel aan de processor aanbiedt.

NAND en NOR flash memory

Er bestaan twee typen flash memory: NAND en NOR (genaamd naar de gebruikte type ic logic gates). Bij NAND flash-technologie worden de floating-gate transistoren aaneengereggen om een grote dichtheid te bereiken. NOR flash heeft geen gedeelde componenten en wordt hoofdzakelijk in de consumentenmarkt en bij embedded devices toegepast, hoofdzakelijk om te booten. Dit in tegenstelling tot NAND flash dat steeds vaker wordt toegepast als data storage in de enterprise omgeving. Een belangrijke overweging bij de beslissing om NAND flash toe te passen is de keuze voor Single-Level Cell (SLC) of Multi-Level Cell (MLC). Beide typen flash verschillen niet veel van elkaar en het fabricageproces is bijna identiek, maar SLC is sneller en betrouwbaarder. Er bestaat verschil in spanningsniveau in een flash cell die de binaire waarden vertegenwoordigen. SLC slaat slechts twee waarden op, 1 of 0 (met een hoog of laag spanningsniveau), terwijl MLC vier waarden opslaat (00, 01, 10 en 11), hoog, middelhoog, middellaag en laag. Een MLC cell heeft een veel lagere spanningstolerantie dan SLC en heeft een tienmaal hogere uitval, omdat tijdens elke program erase cycle deze de neiging heeft om de spanningsvariatie te verhogen. MLC is goedkoper, omdat elke cell slechts twee bits in plaats van vier bezit, waardoor blocks, pages en chips elk tweemaal zoveel capaciteit bevatten. Daardoor duren chip-operaties wel weer tweemaal zolang. Deze prestatienadelen, gecombineerd met de 10 keer grotere foutfactor dan SLC maakt dat MLC flash niet zo geschikt is voor enterprise-applicaties. SLC is op dit moment de meest toegepaste drive-technologie in de enterprise storage-omgeving, maar de verhouding



tussen MLC en SCL zou de komende jaren kunnen veranderen naarmate de fabrikanten er in zullen slagen om de controller-technologie en storage management software te verbeteren. Een van de karakteristieke

NAND flash wear-out

Zoals we zagen slijten NAND flash memory chips als gevolg van herhaaldelijke schrijfprocessen en is het belangrijk dat de frequentie waarmee data wordt geschreven

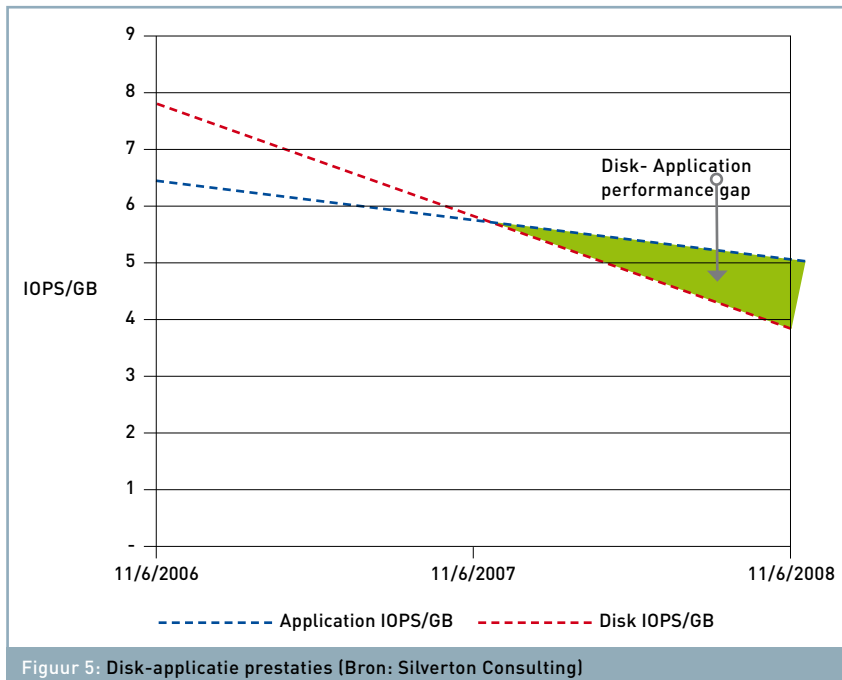
IN TEGENSTELLING TOT DRAM IS FLASH-MEMORY NON-VOLATILE

eigenschappen van NAND flash betreft het schrijfproces van data naar de flash chip. Van te voren moeten eerst alle bits in een flash block (doorgaans 128 tot 256KB groot) worden gewist, wat uiteindelijk zijn tol eist, omdat de oxidelaag die de elektronen moet vasthouden geleidelijk wordt afgebroken. Dit afbraakproces is bij SLC flash een kleiner probleem dan bij MLC. Bij MLC kunnen naarmate de oxidelaag degradeert de switch-on drempelwaarden van de floating gate worden verhoogd of verlaagd. Beide SLC en MLC flash chips maken wel gebruik van error-correction algoritmen om de betrouwbaarheid van de opgeslagen data te waarborgen, maar uiteindelijk zullen toch ook beide type flash memory uitvallen. De uitvalpercentages die door de chip-industrie worden gehanteerd zijn 100.000 program/erase of endurance cycles voor SLC en 10.000 voor MLC. Maar beide waarden verschillen nogal per leverancier. Er zijn gelukkig diverse technieken ontwikkeld om de gevolgen van de zogenoemde NAND flash wear-out te kunnen opvangen.

zich niet beperkt tot slechts één of enkele aantallen flash chips. Om deze reden heeft elke chipfabrikant een scheiding aangebracht tussen de fysieke en logische storage-

laag. Dat wil zeggen, het computersysteem schrijft weliswaar steeds naar hetzelfde logische adres, maar de flash controller verhuist het onderliggende fysieke adres naar de minst beschreven flash memory chips of de minst beschreven blokken binnen de chip. Dit proces staat bij SSD fabrikanten bekend als wear leveling. Zelfs de meest eenvoudige op het circuit board toegepaste wear leveling-techniek leidt tot aanmerkelijke verbetering. De RamSAN-500 van Texas Memory Systems bijvoorbeeld kan jarenlang betrouwbare topprestaties bieden en biedt 2TB bruikbare opslagruimte met een 2GB/s sustain write rate. Bij 2GB/s duurt het dan 1.000 seconden om de volledige 2TB te schrijven. Om de write/erase limiet te halen, zou dit proces 100.000 maal moeten plaatsvinden (omdat wear leveling deze writes evenredig over alle beschikbare blokken moet uitvoeren), waarvoor 100.000.000 seconden nodig zijn. Delen we dit getal door het aantal seconden per jaar, dan levert dit een write endurance van 3,25 jaar op. Dit niveau wordt al gehaald waarbij geen enkele flash block de endurance-specificatie overschrijdt en waarbij van basis wear leveling-techniek gebruik wordt gemaakt, zonder de hulp van een grote RAM cache of een grote pool van hot spare flash blocks. Een unieke eigenschap van flash blocks is dat wanneer ze slijtage vertonen een schrijfoperatie op een gegeven moment zal falen. Dit is eenvoudig door de flash controller te detecteren, zodat deze de data op tijd naar een ander flash block schrijft. Op die manier heeft de flash controller de controle over de 'gezondheid' van elke individuele flash block en chip. Wanneer een bepaald gedeelte van een chip een threshold-waarde overschrijdt, kan de daarin opgeslagen data naar een andere chip/

(Advertentie)



Figuur 5: Disk-applicatie prestaties (Bron: Silverton Consulting)

block worden verhuisd en de betreffende chip/block worden verwijderd, zonder verlies aan totale storagecapaciteit. Hoewel het logisch lijkt om flash SSDs op basis van endurance-tijden te vergelijken, hoeft dit niet altijd de goede keuze te zijn. Sommige slecht presterende SSDs kunnen bijvoorbeeld een prima write endurance bezitten, maar slechte schrijffprestaties. Beperking van het aantal writes verbetert nu eenmaal de endurance door verlaging van het aantal uitgevoerde writes. Hoogpresterende flash SSDs lossen dit probleem weer op door van een meer intelligente wear leveling-methode gebruik te maken en de toevoeging van meer flash storagecapaciteit.

Verbetering write endurance en performance

Sommige flash SSD-producten bezitten extra opslagcapaciteit die voor applicaties niet zichtbaar zijn. Deze extra capaciteit wordt ingezet om de write endurance te verbeteren door de schrijffopdrachten over de extra toegevoegde flash chips te verdelen; de genoemde RamSan-500 heeft 20 procent meer flash memory. De vraag is waarom extra flash memory de write endurance verbetert. De write endurance voor flash specificeert dat niet meer dan 2 procent van de blokken zal uitvallen na 100.000 write/erase cycles. Maar wat gebeurt er met de andere 98 procent van de blokken? Het gemiddeld aantal te doorstaan 100.000 write/erase cycles van een flash block kan belangrijk worden verhoogd door een flash module te ontwerpen met een hogere block failure rate. De genoemde RamSan-500 kan een 10

procent block-uitval doorstaan voordat een module aangeeft dat het moet worden vervangen. Hierdoor wordt het aantal system writes/erase cycles tot 500.000 verhoogd. De worst-case wear out in een 24x7 schrijffomgeving wordt daarmee verlengd van 3,25 tot 15 jaar en dat is ruimschoots meer dan de gemiddelde levensduur van IT-apparatuur!

errors. Een disturbed cell is een cell waarvan de inhoud veranderd is als gevolg van activiteiten in een naburige cel. Dit soort problemen zijn het gevolg van fysieke processen binnen de celstructuur die niet door de fabrikanten van flash boards kunnen worden opgelost. De oplossing moet worden gezocht in het voorkomen dat read/write disturbs de data naar de server en applicatie kan beïnvloeden. Een methode is om Error Correcting Code (ECC) op elk flash module te implementeren, zodat enkelvoudige bit errors automatisch kunnen worden gecorrigeerd. Bij detectie van een meervoudige bit-error wordt het probleem op een niveau hoger afgehandeld door de RAID controller waarbij de juiste data wordt gereconstrueerd op basis van de andere flash memory modules. De combinatie van de verschillende wear leveling-technologieën in combinatie met ECC verhoogt de algehele betrouwbaarheid van flash SSDs. Maar sommige SSD-leveranciers gaan nog een stapje verder door additionele betrouwbaarheidsoplossingen te implementeren op systeemniveau.

Betrouwbaarheid flash memory

De meeste ECC-fouten zijn dus op module-niveau op te lossen maar multi-bit ECC-fouten en uitval van een module niet; multi-bit ECC-fouten treden zeldzaam op als gevolg van read/write disturbs. Met de toe-

CHIPKILL IS METHODE OM FALENDE CHIP ON-THE-FLY TE VERVANGEN

ECC-correctie

Helaas blijven de problemen met flash memory niet beperkt tot write endurance. Flash chips hebben ook last van een uniek soort fouten die worden veroorzaakt door activiteiten van nabijgelegen cellen binnen een flash chip. Dit soort fouten worden aangeduid met read en write disturbed

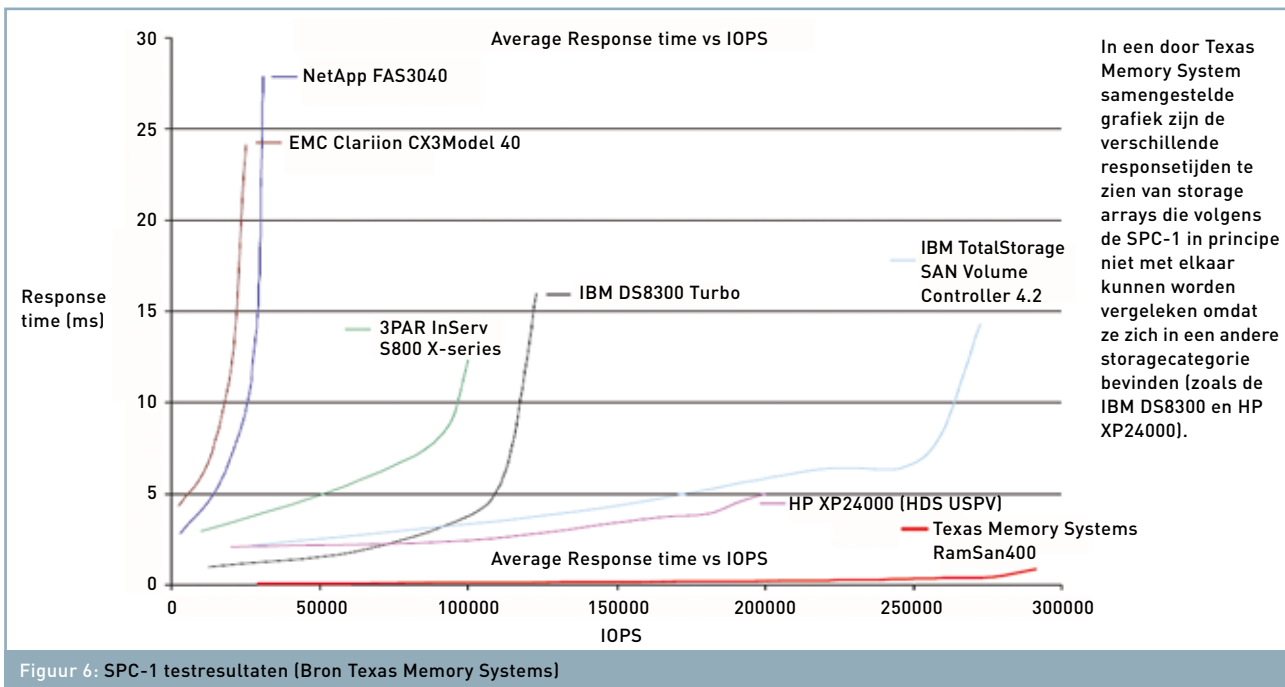
passing van meerdere flash boards is ook corrupte data als gevolg van multi-bit ECC-fouten te reconstrueren. Leveranciers van storage arrays maken van RAID-technologie en caching gebruik om de betrouwbaarheid en prestaties te verbeteren. Het blijkt dat dezelfde basis RAID-technieken ook voor flash-gebaseerde SSDs kan worden toege-

Gemiddelde levensduur SLC en MLD NAND

Op basis van de volgende formule is in de onderstaande tabel de voorspelbare wear out van NAND flash af te lezen: gemiddelde levensduur = levensduur/ lees-schrijfverhouding.

Type / write duty	Gemiddelde levensduur in jaren
SLC flash /40 procent	25
MLC flash /20 procent	10
MLC flash /40 procent	5

zie: www.storageperformance.org



Figuur 6: SPC-1 testresultaten (Bron Texas Memory Systems)

past op een array die bestaat uit meerdere flash boards (modules of disks). Het is wat ongebruikelijk om aan caching te denken bij de verbetering van de systeembetrouwbaarheid bij flash memory. Toch is de toepassing van een groot RAM cache aan de 'voorkant' van een flash RAID-systeem een prima manier om de system write endurance en prestaties te verbeteren. Een product als de RamSan-500 heeft een 64GB RAM cache en een cache-algoritme dat speciaal ontworpen is om de schrijfopdrachten naar het flash RAID systeem te optimaliseren. De grote cache isoleert de flash RAID van kleine random block I/O's, wat de schrijffprestaties verbetert. De toepassing van cache kan daarmee het aantal write/erase cycles verminderen en daarmee de flash endurance verbeteren. Maar wat zijn de gevolgen als het RAM cache een slechte bit of chip heeft? Goed ontworpen RAM caching-systemen worden met ECC tegen bit en met chipkill tegen chipfouten beschermd. Chipkill is een methode om een falende chip on-the-fly te vervangen. Een andere zorg is wanneer de stroom uitvalt en de data in het RAM verloren gaat. Een goed ontworpen flash memory system is daartoe uitgerust met een of meer redundante interne UPS devices die genoeg tijd bieden om het RAM cache naar de Flash RAID te flushen. Een andere manier waarbij RAM cache kan worden ingezet is de bescherming van een flash-systeem bij volumes met een hoge schrijffrequentie door een complete storage LUN in RAM cache te vergrendelen. Voor het OS ziet het vergrendelde volume er als een gewone LUN uit. Op

die manier kan het flash subsysteem worden gevrijwaard van kleine random block writes. Hoe goed ontwerpers van SSDs deze systemen ook op board en systeemniveau kunnen beveiligen tegen uitval, betrouwbaarheid stopt niet op het productniveau, maar strekt zich uit over de hele datacenterarchitectuur.

Epiloog

SSDs zullen steeds vaker in de enterprise-omgeving worden toegepast, als 'tier 0' ingebed in storage array's, stand-alone SSD systems, als DAS in de vorm van zelfstandige flash cards, of bijvoorbeeld SATA II disks (bijvoorbeeld de Intel X25-E). IDC voorspelt dat de markt van SSD van 2009 tot

2012 een jaarlijkse groei van bijna 80 procent zal doormaken en tegen 2010 zal uitkomen op 382 miljoen dollar. Om die groei mogelijk te maken, moet de prijs nog verder dalen en moeten leveranciers van storage arrays hun arrays en beheerssoftware optimaliseren voor de toepassing van SSDs. Want zoals de situatie nu is, kunnen slechts een beperkt aantal SSDs door de storage controller worden ondersteund, omdat de controller al snel een bottleneck vormt. ■

BRAM DONS IS ONAFHANKELIJK IT-ANALIST;
INFO@IT-TRENDWATCH.NL

(Advertentie)