Duurzame opslag van data

Energiebesparing door duurzame opslag
binnen het hoger onderwijs

Colofon

Duurzame opslag van data
Energiebesparing door duurzame opslag binnen het hoger onderwijs

Hogeschool van Arnhem en Nijmegen
Service unit ICT
T + 31 26 369 15 45
F + 31 26 369 15 00

www.han.nl
secretariaat-ICT@han.nl

November 2012

Inhoud

Voorwoord

SURFnet, de samenwerkingsorganisatie van hogescholen en universiteiten op het gebied van ICT, heeft voor 2012 opnieuw een innovatieregeling op het gebied van duurzaamheid en ICT uitgeschreven. Doel van de innovatieregeling is om instellingen te stimuleren projecten te starten om door middel van of met ICT structureel bij te dragen aan verduurzaming. Uit de ingezonden projectvoorstellen werden 5 winnaars geselecteerd die aanspraak kunnen maken op een financiële bijdrage voor hun project. De service unit ICT van Hogeschool van Arnhem en Nijmegen heeft voor deze innovatieregeling een voorstel ingediend omtrent Data & Duurzaamheid.

De HAN heeft duurzaamheid als een van haar speerpunten in haar instellingsplan en de service unit ICT levert daar haar bijdrage aan door verduurzaming binnen ICT toe te passen. Een van de manieren om bij te dragen aan verduurzaming is het verminderen van het energieverbruik door ICT binnen de hogeschool. Een onderzoek door Digital Universe in 2011 naar datagroei gaf aan dat wereldwijd opgeslagen informatie elke twee jaar verdubbelt. De vraag die we ons daarbij stelden was of dit ook voor onze eigen organisatie van toepassing zou zijn? De vraag direct daarop was in hoeverre duurzaam opslaan van data kan bijdragen aan het verminderen van onze energieconsumptie. Met deze vraag in gedachten heeft de service unit ICT van Hogeschool van Arnhem en Nijmegen voor deze innovatieregeling een voorstel ingediend omtrent Data & Duurzaamheid.

Dit verslag is onze uitwerking van de SURFnet innovatieregeling Duurzaamheid & ICT 2012. Interviews bij andere onderwijsinstellingen, inventarisatie van kennis bij de eigen ICT, presentaties van leveranciers hebben, naast de grote hoeveelheid data op het internet, bijgedragen aan dit verslag.
Met deze informatie hopen we het bewustzijn omtrent opslag en duurzaamheid te verhogen. Zijn we bewust van de energiekosten die opslag met zich meebrengt, hoe veel gaat onze data groeien en kunnen we die data duurzamer opslaan?

<u>Doelgroep en aannames</u>
In dit verslag wordt gesproken over opslagsystemen en daaraan gerelateerde zaken. In dit verslag wordt ervan uit gegaan dat de lezer enigszins bekend is met terminologie omtrent datacentra, opslag- en ICT systemen alsmede energie eenheden.
In dit document wordt gesproken over GigaByte (GB) en TeraByte (TB). In de berekeningen is 1000 GB = 1 TB.

Data en duurzaamheid

Energie

Mensen gebruiken steeds meer elektrische energie. Voor veel mensen is dit het meest zichtbaar door de energierekening die zij thuis krijgen. Door gebruik te maken van spaarlampen, het aanschaffen van energiezuinige apparatuur en minder apparaten in stand-by te laten staan, verlaag je de jaarlijkse stroomrekening. Naast de voor de handliggende aanpassingen zijn er ook stroomverbruikers waarbij minder snel wordt stil gestaan. Zoals de wireless router voor je internetverbinding (70 kWh /j), een home dataopslagsysteem (NAS, 280 kWh /j) of altijd de computer aan laten staan (450 kWh /j). Ongemerkt wordt er door vele apparaten elektriciteit verbruikt, ook als je deze niet gebruikt.

Ook voor onderwijsinstellingen geldt dat er bespaard kan worden op energie. Gebruik van zuinige verlichting, aanschaf van zuinige apparatuur en het automatisch uitschakelen van ongebruikte apparatuur en verlichting dragen bij aan vermindering van het energie gebruik. Maar ook hier geldt dat er voorzieningen zijn die altijd aan staan. Draadloze access points, switches, VoIP telefoons maar ook servers en opslagsystemen blijven bijna altijd het hele jaar door aan staan. Voor een groot deel is dat ook logisch, want in de huidige tijd zijn onze voorzieningen 24x7 beschikbaar en lopen ICT processen, zoals het maken van back-ups, ’s nachts door.

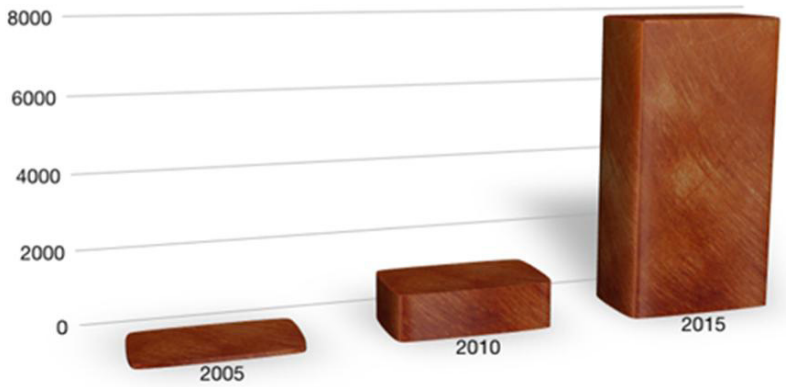
In een eerder, door SURF, gepubliceerd rapport (“Verkenning power management van werkplekken, januari 2012”) zijn al redenen aangegeven om minder energie te gebruiken. ICT organisaties binnen het onderwijs zijn dan ook bezig om, al dan niet op meerdere terreinen tegelijkertijd, efficiënter met energieverbruik om te gaan. Leveranciers van ICT systemen bieden steeds meer mogelijkheden om zuiniger met energie om te gaan, zowel hard- als softwarematig. Door bewuster te worden van deze mogelijkheden kunnen ICT organisaties duurzaamheid als wegingsfactor meenemen bij de aanschaf van ICT apparatuur.

Data en elektrische energie

Wereldwijd is er een enorme hoeveelheid aan digitale informatie. Volgens onderzoek van IDC[1] verdubbelt deze hoeveelheid informatie elke twee jaar. Deze informatie wordt ook steeds vaker opgeslagen en bewaard. De huiscomputer, tablet, smartphone bevatten ook meer en meer informatie en ook die wordt vaak opgeslagen, bijvoorbeeld met behulp van cloud-diensten (Dropbox, Hot/Google-mail, etc). Ook films, foto’s, muziek en eigen documenten (Word, Excel, etc) worden bewaard. Voor het opslaan van al deze informatie is steeds meer opslagcapaciteit nodig.

[1] IDC IVIEW, Extracting Value from Chaos

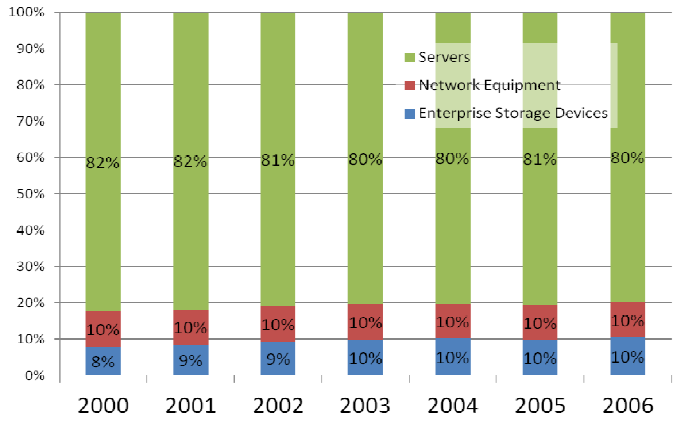
A Decade of Digital Universe Growth: Storage in Exabytes



Figuur 1, Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

Binnen onderwijsinstellingen wordt informatie meestal opgeslagen in grote storagesystemen. Hiervoor worden Storage Area Network (SAN) systemen gebruikt. Hiermee kan centraal informatie opgeslagen worden van medewerkers, studenten en applicaties. Deze systemen bestaan uit een aantal controllers en een zeer groot aantal harde schijven en bevinden die zich bevinden in gehuurde of eigen datacentra van onderwijsinstellingen. Naarmate er meer informatie opgeslagen moet worden zal er ook meer elektrische energie nodig zijn om deze systemen te laten draaien.

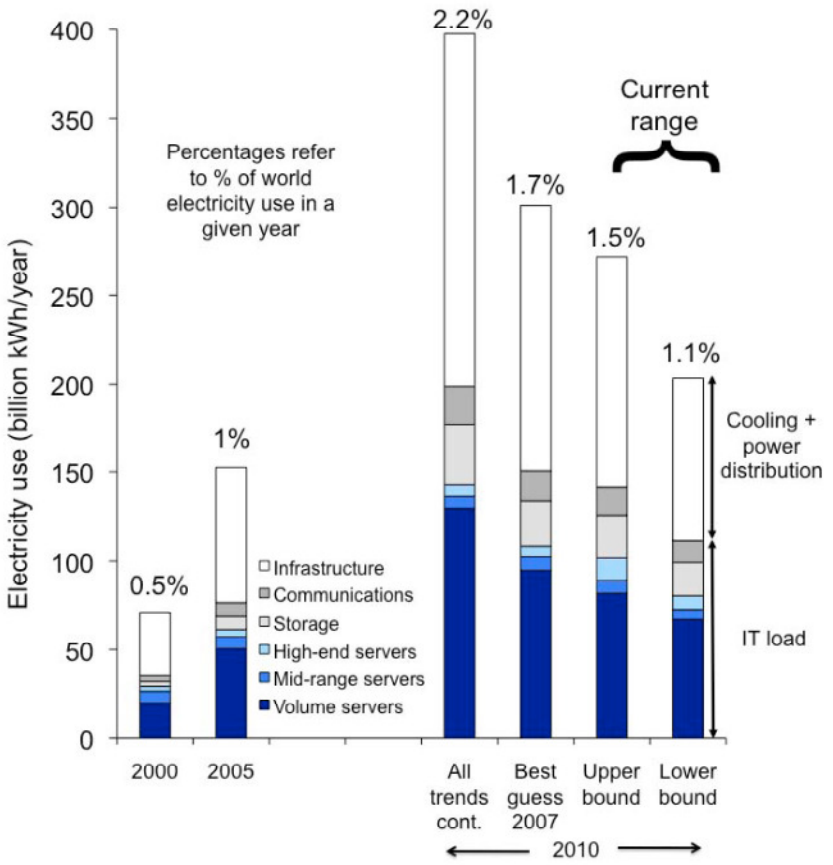
Onderzoek in Amerikaanse datacentra in de periode 2000 – 2006 laat zien dat 10% van het energiegebruik van datacenter gebruikt wordt voor storage systemen.



Figuur 2, Bron: US EPA ENERGY STAR report to congress on server and data center energy efficiency, August 2, 2007

Het aandeel van storage systemen is procentueel gelijk gebleven maar het totale energieverbruik van datacentra is wereldwijd is tussen 2005 en 2010 met 56%[2] toegenomen. Het elektriciteitsgebruik van alle datacentra bedroeg in 2010 ongeveer 1,1% tot 1,5%[2] van het totale elektriciteitsverbruik van de wereld.

[2] Jonathan Koomey. 2011. *Growth in data center electricity use 2005 to 2010*. Oakland, CA: Analytics Press. July. <http://www.analyticspress.com/datacenters.html>



Figuur 3: Worldwide electricity use for data centers (2000, 2005, and 2010). Bron 2

In 2010 is een ICT scan uitgevoerd door Mansystems in opdracht SURFfoundation en AgentschapNL. Aan dit onderzoek hebben 9 onderwijsinstellingen in het HO en WO meegedaan. Het totale energiegebruik voor datacentra is geëxtrapoleerd naar 60 GWh. Op basis van een 10% aandeel voor storage zou de totaal verbruikte elektrische energie voor storage systemen 6 GWh bedragen.

	Deelnemende partijen	MJA: Sector geëxtrapoleerd*
Totaal elektrisch gebruik	179	492
Totaal DC elektrisch energieverbruik	16,1	60
IT component in bovengenoemde DC	7,9	29
Totaal werkplekken	24	89

Tabel 1, elektrisch energie gebruik onderwijsinstellingen (GWh)

* IT gerelateerd elektrisch verbruik is 22% van het totale elektrisch gebruik. Hierin is een tweedeling: bij de traditionele universiteiten 19% bij de Hoge Scholen 40%, *Op basis van de monitoring gegevens 2009, 14 universiteiten en 14 hogescholen. Bron ICT-Scan Mansystems 2010

De service unit ICT van de Hogeschool Arnhem en Nijmegen was benieuwd in hoeverre de uitkomsten omtrent procentuele verhoudingen van energieverbruik overeen kwam met de eigen Hogeschool. De service unit ICT meet continu het stroomverbruik van hun twee datacentra en heeft beide storagesystemen aangesloten op bemeterde stroomaansluitingen. Op basis van aantallen werkplekken, netwerkcomponenten is een berekening gemaakt van het totale energieverbruik van de hogeschool (tabel 2, energieverbruik voor ICT HAN 2011). De uitkomsten hiervan komen overeen met eerdere aangegeven onderzoeken.

HAN	Totaal	ICT	Werkplekken	Netwerk	Datacentra	Storage
GWh/j	9,80	2,00	0,99	0,48	0,53	0,05
%/totaal	100%	20,4%	10,1	4,9%	5,4%	0,5%
%/ICT		100%	49,5%	24,0%	26,5%	2,5%
%/datacentra					100%	9,4%

Tabel 2: energieverbruik voor ICT HAN 2011

Opslagsystemen

De meest gebruikte opslagsystemen binnen HO en WO zijn Storage Area Network (SAN) en Network Attached Storage (NAS). Deze zijn door middel van een netwerk verbonden met de servers. Hier kunnen de servers en de daarop draaiende applicaties gezamenlijk hun data opslaan. Naast SAN/NAS systemen zijn er ook optische opslagsystemen en tape systemen. Deze worden voornamelijk gebruikt voor archivering of back-up. Omdat dit rapport over opslag van data gaat en de meeste onderwijsinstellingen gebruik maken van SAN/NAS systemen, worden alleen deze systemen besproken.

In basis bestaat een SAN/NAS voornamelijk uit minimaal twee controllers (redundantie en performance) en een groot aantal harde schijven in een behuizing met elektronica, voedingen en ventilatoren.
Een groot deel van de energie wordt bepaald door het aantal harde schijven. Op basis van informatie uit vragenlijsten van benaderde hogescholen bestaat een SAN systeem uit twee controllers met gemiddeld 220 schijven.

In figuur 4 is een SAN systeem van de HAN afgebeeld. Dit systeem bestaat uit 2 controllers en 120 3,5” harde schijven. Dit systeem verbruikt continu 3000 Watt. Hiervan gaat ±25% naar de controllers en ±75% naar de harde schijven. Besparing in energieverbruik is onder andere te behalen door zuiniger schijven te gebruiken (2,5” i.p.v. 3,5” of SSD i.p.v. harde schijf), of door minder schijven te gebruiken. Ook technieken als deduplicatie, thin provisioning en tiering dragen hieraan bij.

Figuur 4: SAN van de HAN

Storage reductie technieken

Deduplicatie

Binnen organisaties wordt dezelfde data vaak meerdere keren opgeslagen. Servers slaan dezelfde DLL files op, gebruikers sturen een document naar een collega door en die worden dan deels aangepast. De rest van het document blijft dan gelijk. Al deze bestanden worden apart opgeslagen terwijl deze informatie er in feite al staat. Data deduplicatie is een techniek die duplicaten niet volledig op slaat, maar slechts met een pointer verwijst naar de originele data. Door deduplicatie kan volgens leveranciers de opslagruimte tot wel 50% efficiënter gebruikt worden. Dit is afhankelijk van de aanwezige data en applicaties in een organisatie.

Thin Provisioning

Een SAN biedt gevraagde opslagruimte aan servers en/of applicaties. Voorheen werd dan ook de volledige opslagcapaciteit waarom gevraagd werd ter beschikking gesteld, ongeacht of deze gebruikt werd of niet. Dit wordt ook wel “thick” of “fat” provisioning genoemd. Het grote nadeel is dat hierdoor veel harde schijf ruimte onbenut blijft. Een methode die dit probleem ondervangt is “thin” provisioning. Door middel van deze techniek krijgen servers en/of applicaties de minimum benodigde opslagruimte toegewezen maar krijgen de totale gevraagde opslagruimte te zien. De opslagruimte wordt echter pas daadwerkelijk toegewezen als deze door de server en/of applicatie beschreven wordt. Hierdoor is er minder opslagcapaciteit nodig dan door de organisatie gevraagd wordt. Hoewel deze techniek niet de daadwerkelijke data op een opslagsysteem vermindert, draagt deze wel bij aan het verminderen van de daadwerkelijke opslagcapaciteit.

Tiering

Door data in te delen in verschillende categorieën en deze toe te wijzen aan verschillende opslagmedia, kan de hoeveelheid benodigde schijven gereduceerd worden. Indeling van data kan op basis van performance, beschikbaarheid of bijvoorbeeld veiligheid (redundantie). Het toewijzen van verschillende data aan verschillende soorten media kan een complexe activiteit zijn.
Sommige SAN systemen bieden de mogelijkheid voor een gelaagde opslag te binnen één systeem. Toewijzing kan dan handmatig of geautomatiseerd op basis van policies. Binnen SAN systemen zijn dan verschillende disk arrays die verschillende niveaus bieden. Kleine snelle schijven (FC/SAS of SSD) in een RAID1 configuratie die een hoge performance bieden (Tier1), grotere snelle schijven (FC/SAS) in een RAID5 configuratie (Tier2) of grote langzame schijven (SATA) voor bestanden die weinig geraadpleegd worden (Tier3).

Compressie

Door de aangeboden data eerst te comprimeren is er minder opslagruimte op een harde schijf nodig. Voorheen werd compressie vooral gebruikt bij opslag op tape omdat comprimeren tijd kost en dit ten koste van de performance van een opslagsysteem gaat. Door nieuwe technieken kan softwarematige compressie ingezet worden in opslagsystemen met, volgens de leveranciers, minimale gevolgen voor de prestatie.

Energie efficiency storage

Met alleen het meten van energieverbruik is de efficiëntie van een storagesysteem niet te bepalen. Zeker als je wilt kijken naar de efficiency tussen verschillende instellingen. Het bepalen van een universele maat is lastig en keuzes als beschikbaarheid, benodigde snelheid e.d. bepalen voornamelijk hoe een opslagsysteem eruit komt te zien. Om de energie efficiency van een storagesysteem te bepalen zal je de hoeveelheid opgenomen energie moeten afzetten tegen een capaciteitsmaat.

Hiervoor zijn een aantal capaciteitmaten mogelijk:

- RAW; adresseerbare schrijfbare bytes (TB_R)
- Formatted; Exclusief set-aside (TB_F)
- Assigned; Toegekend aan eindgebruiker en/of applicatie (TB_A)
- Effective; Used + Free (TB_E)
- Used; Used (TB_U); de daadwerkelijk opgeslagen data

Onderstaande tabel (tabel 3) geeft een voorbeeld om de efficiency van een systeem te bepalen op basis van Raw en Effective capaciteit. Hiervoor is het storagesysteem van de HAN genomen. Deze bestaat uit twee 3PAR SAN systemen met tezamen 176 Fiber Channel schijven van 300 GB en 64 SATA schijven van 2 TB. De twee systemen hebben samen een bruto capaciteit (TB_R) van 180,8 TB.

	Opgenomen vermogen	storage capaciteit	energie-efficiënte (ex. PUE)	PUE	Energie-efficiëntie (incl. PUE)	Jaar-verbruik
Eenheid	kW	TB_R	$kWh/d/TB_R$		$kWh/d/TB_R$	MWh/j
3PAR Bruto	6,0	180,8	0,79	1,43	1,1	74,6
Eenheid	kW	TB_E	$kWh/d/TB_E$		$kWh/d/TB_E$	
3PAR Netto	6,0	136,0	1,05	1,43	1,5	74,6

Tabel 3: energie-efficiency SAN systeem van de HAN

Dit is een totaalscore van het systeem van de HAN waarbij alle schijven (FC en SATA) zijn opgeteld. Als er onderverdeeld wordt naar SATA versus FC schijven en naar RAID configuratie dan krijgen we het energieaandeel per type/RAID (tabel 4).

	%	Effectieve storage capacitieit	Raw storage capacitieit	Ruwe energie-efficiëntie	PUE	Jaar-verbruik	Netto Energie-efficiëntie
Type/RAID		TB_E	TB_R	$kWh/d/TB_R$		MWh/j	$kWh/d/TB_E$
Controllers						**14,2**	
FC/1	6%	8,4	16,8	1,60	1,43	**14,1**	4,59
FC/5	23%	31,5	36,0	1,60	1,43	**30,2**	2,62
SATA/6	71%	96,0	128,0	0,24	1,43	**16,1**	0,46
Totaal	100%	135,9	180,8	0,79	1,43	**74,6**	**1,50**

Tabel 4: Energieverhouding huidige configuratie SAN systeem van de HAN

In onze configuratie wordt bijna 30% van de capaciteit verzorgd door snelle FC schijven. Hierdoor biedt het systeem tegen de 40 TB opslag met hoge performance aan en 120 TB (relatief) langzamere opslag aan. Als we ervan uitgaan dat we met minder hoge performance opslag ook onze dienstverlening op hetzelfde peil kunnen houden dan kan dat schelen in energie. Tabel 5 laat zien dat een verschuiving van capaciteit naar langzamere maar grotere SATA schijven een energiereductie kan opleveren van 28%. Op jaarbasis scheelt dit 20,1 MWh.

	%	Effectieve storage capacitieit	Raw storage capacitieit	Ruwe energie-efficiëntie	PUE	Jaar-verbruik	Netto Energie-efficiëntie
Type/RAID		TB_E	TB_R	$kWh/d/TB_R$		MWh/j	$kWh/d/TB_E$
Controllers						**14,2**	
FC/1	3%	4,8	9,6	1,60	1,43	**8,0**	4,59
FC/5	9%	12,6	14,4	1,60	1,43	**12,1**	2,62
SATA/6	87%	120,0	160,0	0,24	1,43	**20,1**	0,46
Totaal	100%	137,4	184,0	0,57	1,43	**54,5**	**1,09**
verschil		+1%				**-28%**	**-28%**

Tabel 5: Energieverhouding met gewijzigde configuratie SAN systeem van de HAN

Ander aspect is het type schijf. Een 3,5" harde schijf verbruikt meer energie dan een 2,5" harde schijf. Een Solid State Disk (SSD) gebruikt slechts een paar Watt per 'disk'. Als je de configuratie van tabel 5 geheel voorziet van SSD 'schijven' dan levert dat ten opzichte van onze huidige SAN systeem een besparing op van 83% (tabel 6). Hier is het opgenomen vermogen van de controllers meegenomen in het opgenomen vermogen per SSD disk.

	%	Effectieve storage capacitieit	Raw storage capacitieit	Ruwe energie-efficiëntie	PUE	Jaar-verbruik	Netto Energie-efficiëntie
Type/RAID		TB_E	TB_R	$kWh/d/TB_R$		MWh/j	$kWh/d/TB_E$
SSD/1	3%	4,8	9,6	0,15	1,43	**0,8**	0,43
SSD/6	9%	12,0	16,0	0,15	1,43	**1,3**	0,29
SSD/5	87%	120,4	137,6	0,15	1,43	**10,8**	0,25
Totaal	100%	137,2	184,8	0,15	1,43	**12,8**	**0,26**
verschil		+1%				**-83%**	**-83%**

Tabel 6: SAN configuratie van de HAN met SSD

De SNIA (Storage Networking Industry Association) heeft een specificatie opgesteld voor het meten van de energie-efficiency van storage systemen[3]. Op basis hiervan werkt ENERGY STAR aan een specificatie voor storagesystemen, waarbij leveranciers hun apparatuur kunnen toetsen aan een efficiëntie matrix. Daarnaast is er meer onderzoek nodig om inzicht te krijgen in energieverbruik van SAN/NAS systemen. Bij de Vrije Universiteit Amsterdam[4] wordt dieper ingegaan op de aspecten van energie efficiency van opslagsystemen.

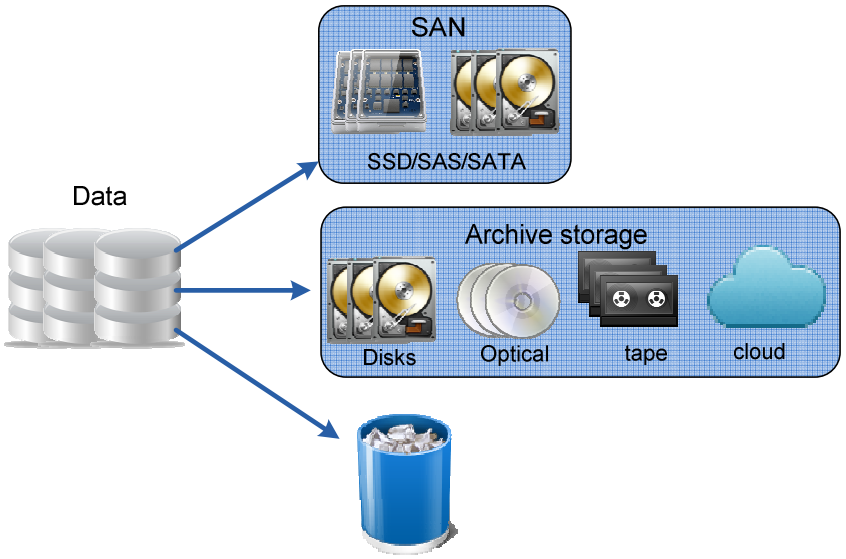
[3] http://www.snia.org/tech_activities/standards/curr_standards/emerald

[4] Author: Simone Potenza, Title: Design of a green scan: Analyzing energy efficiency of data storage systems. Document: Master Thesis. Organization: Vrije Universiteit Amsterdam (KPN as partner),Year: 2012.

Archivering

SAN systemen bieden de organisaties een oplossing om voor meerdere servers centraal en snelle opslag aan te bieden. Door een daling in prijs per harde schijf worden deze systemen steeds goedkoper en betaalbaarder. Systemen worden hierdoor groter en kunnen steeds meer data opslaan waardoor ook meer data wordt bewaard. Dat steeds meer opslaan heeft echter ook zijn prijs. Grotere storagesystemen verbruiken namelijk ook meer energie.

Een andere oplossing voor duurzamer om gaan met data is gebruik maken van archivering. Door data, die al langere tijd niet meer opgevraagd wordt, te verplaatsen naar energiezuinige opslag is het mogelijk om het totale energieverbruik te verlagen. Data zou dan grofweg te classificeren zijn in drie groepen, actieve data (storage op SAN), passieve data (archive storage) en overtollige data (figuur 5). Data binnen de archive storage kan vervolgens onderverdeeld worden in beschikbare data (op disk) of data die bewaard moet blijven (optical/tape/cloud).



Figuur 5: classificatie data

Diverse leveranciers bieden oplossingen aan voor het archiveren. Data kan dan opgeslagen worden op andere opslagvoorzieningen zoals grote SATA disks, optical of tape storage. Daarnaast komen er ook leveranciers die archivering en back-up direct combineren met een cloud dienst.

Inzicht in data

Voordat bepaald kan worden in hoeverre archivering kan bijdragen aan de energiereductie, zal gekeken moeten worden waaruit onze data bestaat. Archiveringsoftware archiveert met behulp van agents op servers waarbij de agent op basis van policies bestanden classificeert. Bestanden kunnen dan door de archiveringsoftware verplaatst worden naar een ander opslagmedium waarbij alleen een verwijzing (pointer) op de oude plek wordt achtergelaten. Dit is mogelijk voor bijvoorbeeld fileservers en voor applicaties als Exchange en SharePoint, zijn plugins beschikbaar die binnen deze applicaties data kunnen verplaatsen.

Om te kunnen bepalen hoeveel data archiveerbaar is, zul je inzicht moeten hebben in de eigen data set. Welke data komt in aanmerking voor archivering en wat levert dit op? Voor de gehele dataset is dit lastig te bepalen. Veel organisaties hebben honderden servers die allerlei verschillende functies hebben. De Hogeschool van Arnhem en Nijmegen beschikt over meer dan 400 virtuele servers, die uiteenlopende functies verzorgen en die tegen de 70 TB aan data beslaan.

HAN	Aantal servers	Opgeslagen data	Archive
Fileservers medewerkers/studenten	20	8,7 TB	√
Exchange servers	15	2,0 TB	√
SharePoint servers	25	2,0 TB	√
Applicaties/test/overig servers	340	52,8 TB	X

Tabel 7: archiveerbare data HAN

Een kort onderzoek naar de servers van de HAN geeft aan dat van de 400 servers tegen de 60 servers in aanmerking komen voor archivering (zie tabel 7). De overige 340 bestaan uit servers die voor test gebruikt worden, applicaties met al dan niet met eigen databases en een grote range servers die voor ICT voorzieningen gebruikt worden (DNS/DHCP/RADUIS/etc).

Op deze 60 servers staat 12,7 TB aan data die archiveerbaar is. Dit is de data van gebruikers zoals documenten en mailtjes. Overige data die opgeslagen wordt, bestanden van het operating system, configuratiebestanden of overhead aan site informatie (in bv SharePoint) zijn daarin niet meegenomen.

Data classificeren

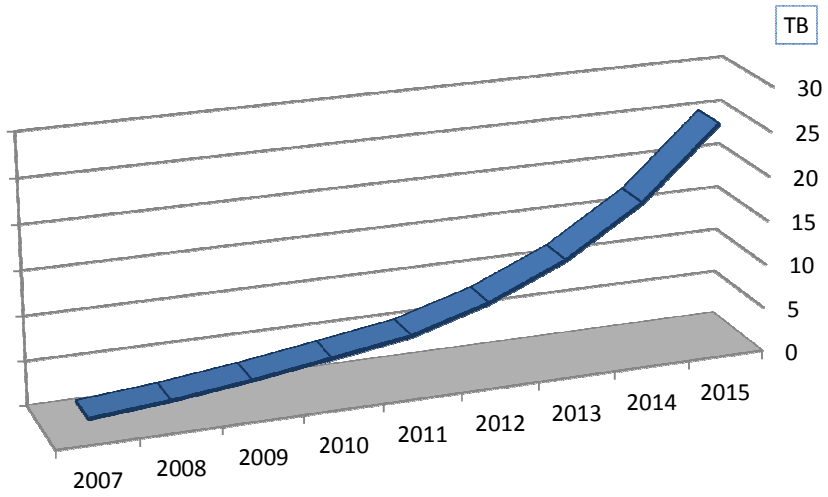
Hoeveel data is te reduceren met archivering? Om op die vraag antwoord te krijgen zul je de data zelf moeten onderzoeken. Hiervoor is diverse soorten metadata analysesoftware beschikbaar of bedrijven die deze analyse voor je kunnen uitvoeren. Om een beeld te krijgen heeft de HAN een extern adviesbureau ingehuurd om de data van medewerkers (5,7 TB aan documenten in eigen- en groepsdirectories) te analyseren. Hiervoor zijn agents op de betreffende fileservers geïnstalleerd. Deze agents hebben alle metadata van de alle documenten en bestanden verzameld waarna deze zijn geanalyseerd. Dit onderzoek heeft plaatsgevonden eind 2011.

Groei

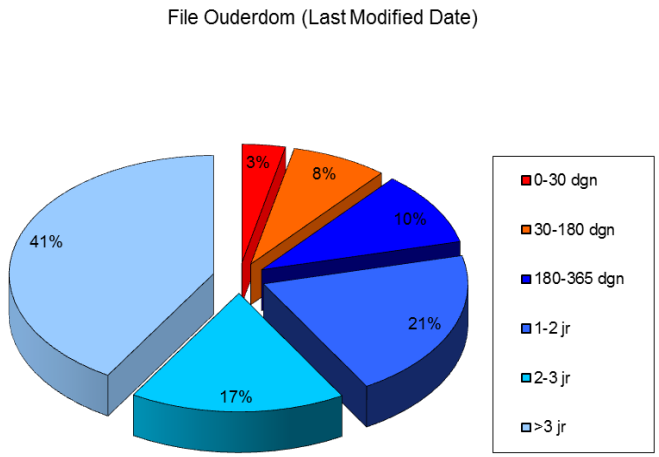
Op basis van deze analyse werd inzichtelijk dat de data van onze medewerkers met meer dan 50% per jaar groeit. Groeit de data net zo hard als de laatste 4 jaar (gemiddeld 46% groei), dan moeten we in 2015 meer dan 25 TB aan documenten en bestanden voor medewerkers opslaan (figuur 6).

Ouderdom

Als we kijken naar de laatste wijzigingsdatum van documenten, dan zien we dat 79% van de aanwezige bestanden meer dan een jaar gelezen voor het laatst gewijzigd is (figuur 7). Hiervan is 17% tussen de twee tot drie geleden gewijzigd en 41% is langer dan drie jaar niet meer gewijzigd.



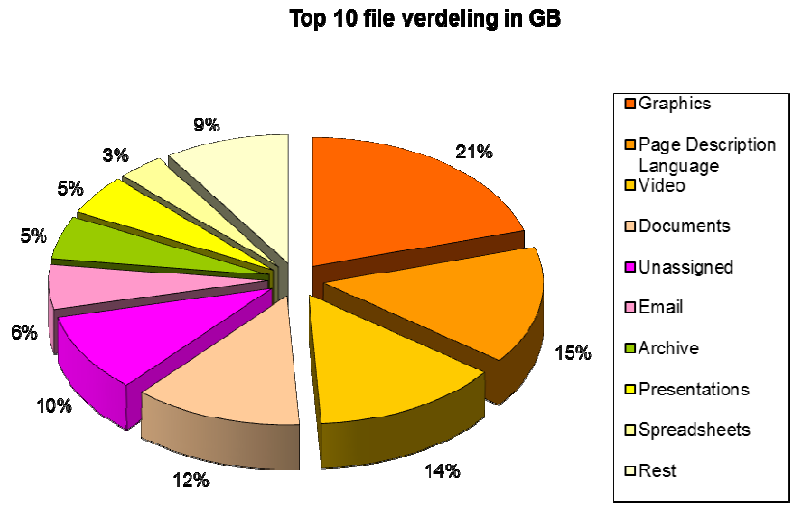
Figuur 6: data groei documenten medewerkers HAN



Figuur 7: verdeling data op basis van laatste wijziging

Ruimtegebruik

In figuur 8 is aangegeven welke bestanden het meeste ruimte innemen van de 5,7 TB. Hierbij valt op dat foto's (graphics 21%) en PDF's (15%) gezamenlijk al meer dan ⅓ van de opslag innemen. Dit zijn bestanden die na het aanmaken doorgaans niet meer gewijzigd worden.



Figuur 8: verdeling gebruik opslag per type bestand

Besparing

In tabel 7 was al aangegeven dat de HAN over 12,7 TB aan archiveerbare data beschikt. Onderzoek naar een 5,7 TB van die data geeft aan dat 58% al meer dan 2 jaar niet gewijzigd is (figuur 7). Op basis van deze getallen zou de HAN 7,4 TB aan data kunnen archiveren.

In tabel 4 was de netto energie-efficiënte van de SAN in zijn huidige configuratie per Tier berekend. Tabel 8 laat zien wat 7,4 TB aan gearchiveerde data aan energiereductie op jaarbasis per Tier oplevert. Omdat het hier om meer dan 2 jaar ongewijzigde data betreft, zal deze zich binnen de SAN op SATA/RAID6 schijven bevinden. Door archivering zal dan 1,2 MWh/j bespaard worden. Dit is 1,7% van het totale verbruik per jaar.

	Netto Energie-efficiëntie	Archiveerbare Data	Aantal dagen	Jaar-verbruik
Type/RAID	(kWh/d/$TB_{E)}$	TB		MWh/j
FC/1	4,59	7,4	365	12,4
FC/5	2,62	7,4	365	7,1
SATA/6	0,46	7,4	365	**1,2**

Tabel 8: Energiebesparing archivering data HAN

Uit bovenstaande blijkt al dat archivering het meeste energiebesparing oplevert bij opslagsystemen die niet over tiering beschikken. Naarmate de TB_E lager wordt door (auto)tiering, deduplicatie en compressie zal het behaalde rendement lager uitvallen.

Ter illustratie rekenen we hieronder de energiebesparing van archivering uit voor twee soorten getierde SANs.

In Tabel 9 laten we zien hoe het totale energieverbruik van de SAN wijzigt als we 7,4 TB effectieve storage capaciteit naar het archief afvoeren. Hierdoor kunnen we niet meer dan 4 SATA/RAID6 disks van 2,0 TB uitschakelen, goed voor 8,0 TB ruwe opslagcapaciteit.

	%	Effectieve storage capacitieit	Raw storage capacitieit	Ruwe energie-efficiëntie	PUE	Jaar-verbruik	Netto Energie-efficiëntie
Type/RAID		TB_E	TB_R	kWh/d/TB_R		MWh/j	kWh/d/TB_E
Controllers						**14,2**	
FC/1	4%	4,8	9,6	1,60	1,43	**8,0**	4,59
FC/5	10%	12,6	14,4	1,60	1,43	**12,1**	2,62
SATA/6	87%	114,0	152,0	0,24	1,43	**19,1**	0,46
Totaal	100%	131,4	176,0	0,58	1,43	**54,5**	**1,11**
Verschil tov Tabel 5		-6,0 ~ -4,4%				**-1,0** **~ -1,8%**	**0,03** **+2,6%**

Tabel 9: Energiebesparing archivering data HAN ten opzichte van Tabel 5

In Tabel 10 laten we hetzelfde zien, maar dan voor een systeem met alleen SSDs: 10 SSD units van 0,8 TB op RAID6, samen goed voor 8,0 TB ruwe opslagcapaciteit en 6,0 TB effectieve opslagcapaciteit.

	%	Effectieve storage capacitieit	Raw storage capacitieit	Ruwe energie-efficiëntie	PUE	Jaar-verbruik	Netto Energie-efficiëntie
Type/RAID		TB_E	TB_R	kWh/d/TB_R		MWh/j	kWh/d/TB_E
SSD/1	4%	4,8	9,6	0,15	1,43	**0,8**	0,43
SSD/6	5%	6,0	8,0	0,15	1,43	**0,6**	0,29
SSD/5	92%	120,4	137,6	0,15	1,43	**10,8**	0,25
Totaal	100%	131,2	155,2	0,15	1,43	**12,2**	**0,26**
Verschil tov Tabel 6		-6,0 ~ -4,4%				**-0,6** **~ -4,9%**	**-0,001** **-0,6%**

Tabel 10: Energiebesparing archivering data HAN ten opzichte van Tabel 6

Deze extra besparing (1,0 MWh/j, de casus met traditionele disks en 0,6 MWh voor de SSD-casus) betreft alleen de behaalde energiebesparing op het storagesysteem zelf. De gearchiveerde data zal echter op andere media opgeslagen moeten worden. Het energieverbruik voor archivering, afhankelijk van de gekozen media (optical, tape of cloud), moet hier nog van afgetrokken worden. Daarnaast zal voor het archiveren een extra server ingericht moeten worden waar de archiveringsoftware op draait, wat extra energieverbruik zal opleveren, en opslagcapaciteit kost voor de databases van die software. Om deze waardes mee te kunnen nemen in het totaalplaatje zal extra onderzoek nodig zijn.

Data in het hoger onderwijs

Het hoger onderwijs in Nederland bestaat uit ruim 40 hogescholen[5] en 14 universiteiten[6]. Het biedt onderwijs aan ±680.000 studenten (430.00 HO en 250.000 WO) en werk aan ±90.500 medewerkers (41.500 HO en 49.000 WO). Al deze studenten en medewerkers genereren direct en indirect data die al dan niet bij deze instellingen opgeslagen wordt. Het gaat hierbij dan om home- en groepsdirectories met allerlei documenten, mailboxen, studentinformatiesystemen, websites en applicaties.

Onderzoek in datacentra gaf al aan dat 10% van de gebruikte energie naar dataopslag gaat. Onderzoek van IDC voorspelt een verdubbeling van alle informatie per twee jaar. Extrapolatie van de deelnemers aan de ICT-scan 2010 naar de gehele sector wijst in de richting van een verbruik van 6 GWh voor opslag. In hoeverre is dit van toepassing binnen het hoger onderwijs? Speelt duurzaamheid en dataopslag een rol, weet men hoeveel energie er voor dataopslag verbruikt wordt? Om hiervan een beeld te krijgen hebben we een aantal hogescholen en universiteiten benaderd om een aantal vragen te beantwoorden rond data en opslag. Hierop hebben uiteindelijk 9 hogescholen en 3 universiteiten gereageerd.

Wat verstaan we onder data

In het onderzoek van IDC is gekeken naar wereldwijde informatie. Hieronder worden ook tijdelijke informatie zoals televisiebeelden, telefoongesprekken via digitale infrastructuur en audiostreams verstaan. Deze informatie verdwijnt grotendeels weer en wordt niet opgeslagen. In verband met data en duurzaamheid kijken we naar informatie die wel opgeslagen wordt en waarbij energie gebruikt wordt om deze opgeslagen te houden.

Een zoekopdracht op Google naar de definitie van data levert een groot aantal beschrijvingen op. Een aantal definities zijn:

- Gegevens, verzamelnaam voor (bedrijfs-) gegevens, zoals boekhouding, personeelgegevens, tekstdocumenten, onderzoekgegevens.
- Feiten die verzameld worden om antwoord te kunnen geven op vragen die gesteld worden.
- de in -of outputgegevens die in een digitaal informatiesysteem worden beheerd.
- relevante gegevens die specifiek betrekking hebben op het te analyseren probleem waarbij het gaat om nog niet aanwezige kennis in het expertsysteem.
- meervoud van Latijnse “datum” (wat gegeven is).

Data zou in het kader van dit verslag beschreven kunnen worden als alle informatie die door de organisatie gebruikt en bewaard wordt. Hieronder verstaan we de bestanden die door medewerkers en studenten gecreëerd en gebruikt worden (documenten, foto’s en filmpjes, audio, etc.). Deze kunnen of rechtstreeks op toegewezen directories geplaatst worden (home- en groepsdirectories) maar ook door middel van applicaties (digitale portofolios, elearning omgevingen, etc.). Daarnaast zijn er ook applicaties die hun gegevens opslaan op directories of in databases. Hierbij kan gedacht worden aan cijfer- en roosterapplicaties maar ook aan betalings- en toegangssystemen, mailsystemen, etc. Vervolgens worden ook gegevens van

[5] bron http://www.hbo-raad.nl/hogescholen/over-hogescholen
[6] bron http://www.vsnu.nl/Universiteiten/Alle-universiteiten.htm

computer- en netwerksystemen zelf opgeslagen. Dit kunnen configuratiebestanden zijn, meta-data voor analyses, besturingsbestanden etc.

Opgeslagen informatie kan zich op allerlei plaatsen binnen een organisatie bevinden. Medewerkers kunnen bestanden op de lokale schijf van hun werkplek opslaan, de computer zelf gebruikt lokale schijfruimte voor opslag van informatie (besturingsysteem, instellingen, tijdelijke bestanden, etc). Studenten bewaren documenten op USB-sticks of op eigen laptops. Smartphones, tablets hebben intern geheugen om informatie op te slaan maar ook de vele servers binnen onze datacentra hebben lokale harde schijven waarop allerlei informatie voor de server zelf staat. Daarnaast maken veel mensen gebruik van virtuele opslag in de Cloud (Dropbox, Google Drive, etc.).

Dit verslag richt zich op de vraag of je data duurzamer kan opslaan en welke energievoordelen haalbaar zijn. Binnen onderwijsinstellingen wordt informatie opgeslagen op centrale opslagsystemen en hebben ICT afdelingen geen grip op alle informatie die zich buiten haar invloedsfeer valt. Hierdoor is het moeilijk om daarop energiebesparingen te realiseren. Sommige instellingen besteden bepaalde diensten uit (bijvoorbeeld mail voor studenten) waardoor zij deze informatie niet op eigen systemen opslaan. Instellingen die hun werkplekken geheel gevirtualiseerd hebben en alleen gebruik maken van zogeheten “thin-clients”, hebben geen lokale harde schijven meer op de werkplek. Wel gebruiken alle geïnterviewde instellingen centrale opslagsystemen voor het opslaan van hun data.

Om toch een beeld te krijgen van de hoeveelheid informatie en de benodigde energie, hebben we onderwijsinstellingen gevraagd naar de door hun beheerde hoeveelheid data. Hiervoor is data gedefinieerd als:

> Data: alle informatie van de organisatie welke centraal opgeslagen wordt en gebruikt wordt door de organisatie. Back-up van data ten behoeve van veiligheid wordt hierin niet meegerekend. Mirroring van data binnen opslagsystemen weer wel. Data op interne schijven in serversystemen, pc’s/laptop/tablets, alsmede in de cloud e.d. worden niet meegerekend.

Hoeveelheid data bij onderwijsinstellingen

De eerste vraag die gesteld werd is hoeveel data er centraal opgeslagen is binnen de onderwijsorganisatie zonder mirroring of gebruik van snapshots. Het ging hierbij om data opslag voor medewerkers, studenten, beheer en onderwijsapplicaties.

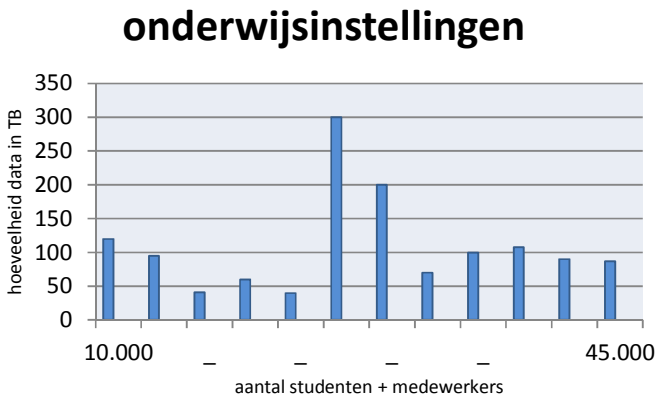
Bij de universiteiten is er daarnaast ook nog opslag van onderzoeksdata. Meetgegevens van bijvoorbeeld onderzoeksopstellingen, sterrenkundig onderzoek of medisch onderzoek leveren een zeer grote hoeveelheid data op die niet in verhouding staat tot organisatiedata. Deze informatie is voor dit verslag buiten beschouwing gelaten omdat hier een veel specifiekere ontwikkeling is op storage gebied met eigen specifieke kenmerken. De schaalgrootte hiervan valt buiten het terrein van de hogescholen en vereist specifieke kennis omtrent opslag.

Op basis van de verstrekte gegevens van 12 instellingen die mee hebben gedaan aan het vragenonderzoek blijkt dat er 1356 TB aan data wordt opgeslagen. Op basis van een extrapolatie naar aantal medewerkers en studenten worden er naar schatting door alle onderwijsinstellingen 2.826 TB aan data centraal opgeslagen (tabel 11).

	Bevraagde instellingen	Alle onderwijsinstellingen
Aantal studenten	321.990	680.000
Aantal medewerkers	35.479	90.500
Totaal opslag data	1.356 TB	2.923 TB (geëxtrapoleerd)

Tabel 11: data opslag bij HO instellingen

Als we kijken naar de verhouding aantallen studenten + medewerkers[7] en de hoeveelheid centraal opgeslagen data (figuur 9) dan zien we verschillen tussen de verschillende instellingen. Een duidelijke relatie tussen aantal gebruikers en hoeveel data is er niet. In gesprek met de instellingen laat zien dat de één wel huisregels rondom hoeveel datagebruik maar geen actief quotamanagement (voor zowel directories en mailboxen). De andere instelling doet dat alleen voor studenten terwijl sommige quota op alle gebruikers toepast. Ook een verschil in voorzieningen speelt een rol bij de hoeveelheid informatie. De ene organisatie verzorgt zelf de email voor studenten, anderen hebben dit buiten de eigen organisatie belegd.



Figuur 9 verhouding opgeslagen data en aantal gebruikers van onderwijsinstellingen

Bijna geen van de gevraagde instellingen kon aangeven hoe de verdeling binnen hun dataset was met betrekking tot data in gebruik door studenten, door medewerkers en door applicaties. Aangegeven werd dat verdeling in studentendata, medewerkersdata en applicatiedata uitgezocht kon worden door te kijken naar welke servers welke volumes hebben maar een totaal overzicht is niet direct inzichtelijk. Daarnaast zit data voor studenten en medewerkers ook gemengd in applicaties. Een voorbeeld is SharePoint waar zowel medewerkers als studenten my-sites hebben. De uiteindelijke data daarvan wordt in tabellen in SQL servers opgeslagen.

Beleid ten aanzien van de bewaartermijnen van data is bij bevraagde instellingen niet formeel vastgelegd of men is dit nog aan het ontwikkelen. Wel hebben geven bijna alle instellingen aan dat er richtlijnen of spelregels zijn wanneer data afgevoerd wordt als medewerkers of

[7] De precieze relatie aantal studenten + medewerkers en data per onderwijsinstelling is om reden van bescherming globaal gehouden.

studenten de organisatie verlaten. Aangegeven wordt dat men zo nu en dan een opschoonactie doet bij home- en groepdirectories . Eén instelling gaf aan dat men daardoor ook nog data had staan uit de jaren ’80.

Quota management is wel bij de meeste instellingen in gebruik. Een enkele instelling past dat alleen toe voor studenten maar de meeste hebben quota voor alle gebruikers. Wel krijgen medewerkers meer ruimte dan studenten. Doorgaans variëren quota tussen 250 MB en 1 GB voor mailboxen en homedirectories; groepsdirectories hebben vaak geen quotum of een veel groter quotum (100 GB).

Gebruikte data- opgeslagsystemen

De hogescholen die benaderd zijn[8] slaan hun informatie hoofdzakelijk op SAN systemen. Gebruikte systemen zijn van NetApp (3x), Dell (Compellent 1x, EqualLogic 1x), EMC, 3Par en HP EVA (dit systeem was voorheen in gebruik bij een aantal instellingen). Deze SAN systemen zijn voornamelijk opgebouwd met SAS, FC en/of SATA schijven. Twee systemen maken daarnaast gebruik van een klein aantal SSD schijven (voor onder andere Fast Caching).

De meeste SAN systemen gebruiken technieken als deduplicatie, (auto-) tiering, thin-provisioning, om data efficiënter op te slaan. Deze technieken worden, mits ondersteund door het SAN systeem, toegepast. Op de vraag hoeveel een bepaalde techniek aan besparing oplevert gaf meer van de helft van de ondervraagde instellingen aan dit niet te weten. Desgevraagd schatten de respondenten bijvoorbeeld gemiddeld 50% voor deduplicatie, 25% tot 50% door thin-provisioning, tot wel 320% “storage efficiency” (gebruik van alle technieken bij elkaar, inclusief snapshots voor disaster recovery).

Op één hogeschool na worden er meerdere SAN systemen al dan niet in clusters gebruikt. Een cluster, bestaande uit één of meerdere opslagsystemen, wordt gebruikt voor de primaire opslag van data. De andere cluster wordt gebruikt als uitwijk door data synchroon of asynchroon te repliceren. Dit gebeurt door mirroring, snapshots of 1 keer per dag een volledige kopie te maken. Bij één hogeschool worden beide clusters primair gebruikt en zijn ze, door mirroring ook gelijk elkaars back-up, een andere hogeschool kopieert de data ’s nachts van het ene cluster naar de volgende. In alle gevallen wordt hiermee veiligheid ingebouwd zodat bij een calamiteit men weer snel over data kan beschikken. De clusters bevinden zich daarom op verschillende locaties.

[8] Door de zeer gesegmenteerde opslag en onvoldoende informatie zijn alleen de gegevens van hogescholen gebruikt voor dit hoofdstuk.

Energieverbruik opslagsystemen (bij HO)

Het elektrisch opgenomen vermogen van SAN systemen kan worden bepaald door middel van het meten van het opgenomen vermogen bij het systeem zelf, aan de hand van specificaties van de leverancier, of door berekening van het aantal componenten. Aan alle instellingen is een opgave gevraagd van het opgenomen vermogen van hun SAN systemen.

Van de 12 instellingen gaven:

- 8 instellingen aan niet te weten wat het opgenomen vermogen is,
- 2 instellingen aan het vermogen te kunnen meten maar dit ter plekke te moeten optellen,
- 2 instellingen aan alleen de gegevens van de gehuurde lokatie (stroom per rack inclusief PUE) te weten.

Eén instelling kon direct aangeven wat het opgenomen vermogen van één van hun SAN systemen was, een ander gaf aan dat deze informatie wel in de organisatie bekend was maar kon dit niet direct opvragen. Twee andere instellingen verwezen naar de specificaties van hun SAN systeem.

Andere aspecten in het elektrisch verbruik voor SAN systemen zijn het elektrisch verbruik van koeling, UPS (Uninterruptible Power Supply) en andere energieverbruik binnen een datacentrum. Door het verbruikte vermogen te vermenigvuldigen met een PUE-factor (Power Usage Effectiveness) kan het totale energieverbruik bepaald worden. Aan alle instellingen is gevraagd naar de PUE-factor van de ruimte waarin de SAN systemen staan opgesteld.

Van de 12 instellingen gaven:

- 7 instellingen aan geen PUE-factor te weten,
- 2 instellingen alleen de PUE-factor van gehuurde (uitwijk) ruimtes te weten,
- 3 instellingen de PUE factor te weten van de ruimtes waarin hun SAN systemen staan.

Door het ontbreken van PUE-factor is het absolute energieverbruik voor SAN systemen niet bij alle instellingen te berekenen. In dit gedeelte van het rapport wordt daarom alleen gekeken naar het opgenomen vermogen van de SAN systemen zelf zonder de energiekosten voor koeling en UPS daarin mee te nemen.

Aan de hogescholen is gevraagd is onder andere gevraagd om per SAN systeem het aantal schijven en controllers te specificeren. Omdat binnen de instellingen bijna geen gegevens beschikbaar zijn over het opgenomen vermogens van SAN systemen, zal door middel van berekening een schatting worden gemaakt. Om het totaal opgenomen vermogen te berekenen worden de aantallen controllers en harde schijven vermenigvuldigt met kengetallen.

De kengetallen[9] waarmee gerekend wordt zijn:

- SSD schijf, opgenomen vermogen 3 Watt
- 2,5”harddisk, opgenomen vermogen 10 Watt
- 3,5”harddisk, opgenomen vermogen 18 Watt
- Controller, opgenomen vermogen 300 Watt

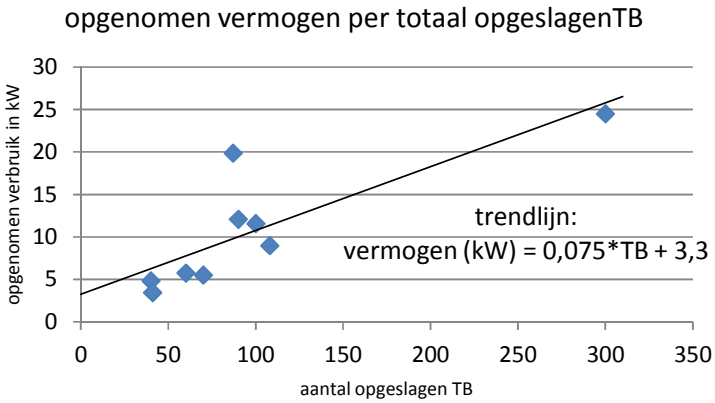
[9] In de opgenomen vermogens zijn voedingen, ventilatoren en elektronica meegenomen.

Deze kengetallen zijn aannames gebaseerd op gemeten vermogens van eigen SAN systemen van de HAN waarbij losse onderdelen apart gemeten zijn. Deze getallen zijn vervolgens getoetst op SAN systemen waarbij het opgenomen vermogens vastgelegd is.

Verhouding energieverbruik en data (HO)

De meeste hogescholen maken gebruik van één of meerdere primaire SAN systemen waarop de data opgeslagen wordt. Van deze data wordt een kopie gemaakt (mirror of replica) op één of meerdere secundaire SAN systemen. De reden hiervoor is om bij calamiteit of uitval weer snel over de betreffende data te beschikken. Al deze primaire en secundaire systemen worden beschouwd als één voorziening voor het opslaan en beschikbaar hebben van data en worden meegenomen in de berekening van de totale benodigde elektrisch vermogen. Voorzieningen als back-up to disk en tape wordt hierin niet meegenomen.

Op basis van de aangeleverde gegevens kan dan de totaal benodigde opgenomen vermogen worden berekend in relatie tot de totaal opgeslagen data (figuur 10).



Figuur 10 verhouding opgeslagen data en energieverbruik van onderwijsinstellingen

Onderwijsinstellingen met meer data verbruiken meer energie voor de opslag van deze data. Toch laat figuur 2 zien dat daarin behoorlijke verschillen kunnen zitten. Interviews met de benaderde hoger onderwijs instellingen laat zien dat verschillende factoren invloed hebben op de totaal verbruikte energie van opslagsystemen.
Deze verschillen bestaan uit:

- Gebruik van mirroring/replicatie; wel of geen secundaire SAN systemen,
- Hoge performance; relatief veel kleinere schijven,
- Gebruik van tiering; meer gebruik van minder maar grotere schijven,
- Gebruik van type schijven; SSD, 2,5”en 3,5” schijven,
- Reductietechnieken; niet alle SAN systemen bieden dezelfde reductie technieken of deze worden niet gebruikt.

De precieze effecten van al deze factoren op het energieverbruik van opslagsystemen is bij onderwijsinstellingen niet te achterhalen. SAN systemen worden niet gemeten, behaalde besparingen met reductietechnieken zijn niet bekend of worden geschat. Daarnaast zijn de meeste systemen geoptimaliseerd voor performance en niet voor energiereductie. Voor beter

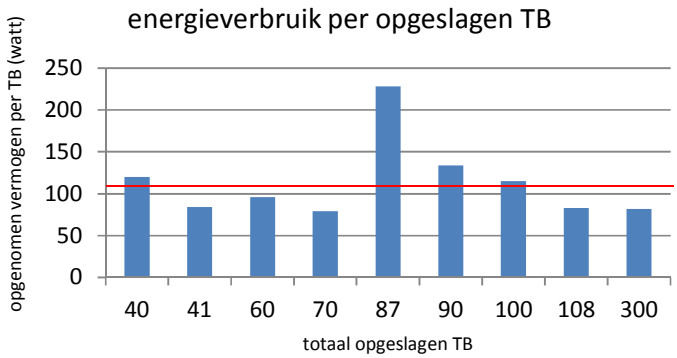
inzicht in relaties tussen energieverbruik en bovengenoemde factoren is meer onderzoek nodig.

Dat meer data meer opslag op SAN systemen vereist is evident, maar dat wil niet, zoals figuur 11 laat zien, zeggen dat dit ook in alle gevallen meer energie kost.

Toch is er wel een trend te ontdekken, waarbij een lineaire benadering die de getallen uit (figuur 10) het best benadert, uitgaat van een basislast van 3,3 kW en daarbij opgeteld 75 W voor elke TB opslag extra.

Het efficiënter krijgen van een opslagsysteem kan door energiezuinigere componenten, zoals het type schijf (SSD, 2,5"of 3,5"), aantal controllers, zuinige voedingen en ventilatoren. Aan de andere kant kan je door de capaciteit van opslagsystemen optimaal te benutten een efficiëntere verhouding tussen vermogen/TB krijgen. Dit is mogelijk door gebruik te maken van technieken als deduplicatie, thin provisioning en tiering waardoor minder schijfruimte nodig is. Een andere oplossing is door archivering data te verplaatsen naar energiezuinigere opslagsystemen.

Het meest energiezuinig is echter om ervoor te zorgen dat er minder data opgeslagen hoeft te worden. Een goed beleid op dat gebied vertaalt zich in minder data, wat zich uiteindelijk weer vertaalt in minder energieverbruik.



Figuur 11 verhouding energieverbruik per opgeslagen TB van onderwijsinstellingen

Verwachte groei

Alle benaderde onderwijsinstellingen is gevraagd naar hun verwachting betreffende de groei van hun data. Komt deze overeen met de verdubbeling per twee jaar wat aangegeven werd in het onderzoek van IDC IVIEW?

De onderwijsinstellingen gaven het volgende aan:

- 1 instelling verwacht een toename van 12,5% per jaar gebaseerd op basis van hun dataomvang van twee jaar geleden,
- 1 instelling verwacht een toename van 20-30% per jaar,
- 1 instelling verwacht een toename van 40% per jaar,
- 4 instellingen verwachten een toename van 100% per jaar,
- 5 instellingen verwachten een toename van 100% per twee jaar.

De instellingen die een toename van 100% per jaar verwachten schrijven deze toename vooral toe aan meer gebruik van videotoepassingen (o.a. registratie van colleges). Anderen geven aan dat de toename voor de eigen organisatie minder zal groeien omdat men video en/of mail buiten de eigen organisatie zal plaatsen.

Besparing in opslag

Door gebruik te maken van technieken als deduplicatie, thin provisioning, tiering en archivering kan de opslagcapaciteit efficiënter gebruikt worden. Hierdoor zijn er minder harde schijven nodig en daardoor zal het energieverbruik lager zijn. Aan de benaderde instellingen is gevraagd of men gebruik maakt van reductietechnieken en welke besparing ze daarmee realiseren.

De 12 onderwijsinstellingen gaven het volgende aan:

- 5 instellingen gebruiken deduplicatie
- 10 instellingen gebruiken thin provisioning
- 3 instellingen gebruiken auto/multi tiering
- 3 instellingen gebruiken archivering

Deduplicatie

Dat deduplicatie niet ingezet wordt heeft deels te maken dat een aantal SAN systemen deze techniek niet ondersteunen. Een andere instelling gaf aan dat gebruik van deduplicatie bij hen negatief uitpakt in combinatie met een compressietechniek van het filesysteem zelf. Twee instellingen gaven aan dat deduplicatie naar schatting tussen de 50% tot 70% reductie oplevert. De andere drie konden geen indicatie aangeven wat deduplicatie hen oplevert.

Thin provisioning

Op twee na gebruiken alle instellingen thin provisioning. Ook hier bieden twee systemen deze functionaliteit niet aan. De besparing die wordt behaalt door thin privisioning wordt bereikt door meer opslagruimte uit te geven dan er feitelijk is, in de hoop - en doorgaans gegronde verwachting - dat deze uiteindelijk toch niet gebruikt wordt. Instellingen geven aan tussen de 25% tot 100% meer uit te geven dan er daadwerkelijk gealloceerd is.

Tiering

Drie instellingen benoemen tiering als reductie-techniek. Immers, door data met minder hoge eisen voor access time of beschikbaarheid minder redundant op te slaan, is minder ruwe opslagruimte vereist. Als verder gekeken wordt naar de configuratie van hun SAN systemen dan zijn er 7 instellingen waarbij tiering mogelijk is of gebruikt wordt. Drie andere instellingen hebben of gebruiken deze mogelijkheid niet. De twee overige instellingen hebben hiervoor te weinig informatie gegeven. Kwantitatieve gegevens over de besparing door tiering is door geen van de instellingen genoemd of kon niet worden gegeven.

Het gebruik maken van tiering binnen een storagesysteem kan handmatig of automatisch gedaan worden, al dan niet gebaseerd op policies. Aan de instellingen is dan ook gevraagd of zij hun data classificeren en of zij op basis van deze classificatie de opslag van hun data bepalen. Bij bijna alle onderwijsinstellingen die gebruik maken van tiering wordt dit automatisch gedaan door het systeem. Eventueel performance gevoelige en/of kritische data (databases, mailsystemen) worden vaak toegewezen aan de hogere performance schijven. Eén instelling geeft aan dat zij handmatige tiering toepassen gebaseerd op basis van de eisen of wensen van de instelling.

Archivering

Van de 12 onderwijsinstellingen geven drie aan gebruik te maken van archivering. Hierbij wordt gebruik gemaakt van archiveringsoftware. Bij één instelling wordt alleen email van medewerkers gearchiveerd. Een andere instelling archiveert data vanuit home-directories van medewerkers en plaatst deze op een NAS. Van de 10 TB wordt hiermee 2 TB aan data verplaatst.

De meeste instellingen is gevraagd of men inzicht heeft hoeveel data gebruikt wordt door medewerkers (home- en groepdirectories), studenten en applicaties. Hieruit kwam naar voren dat men over het algemeen weet hoeveel email data en medewerkers data men heeft, maar overige data moet uitgezocht worden. Hierdoor is er geen zicht op mogelijke besparingen door middel van archivering.

Conclusie

In dit verslag is gekeken naar data opslag binnen het hoger onderwijs en het wetenschappelijk onderwijs. Hierbij is gekeken naar data die gebruikt wordt door de organisatie met hun studenten en medewerkers. Universiteiten beschikken, door hun meer onderzoekgerichte karakter, naast de “gewone” data over onderzoeksdata. Deze data hebben we buiten beschouwing gelaten omdat karakteristieken zoals omvang en vereiste toegangstijden enorm kunnen variëren, afhankelijk van het soort onderzoek.
Door duurzaam om te gaan met data is het mogelijk om energie te besparen. Hiervoor zijn een aantal factoren benoemd en waar mogelijk berekend of met een casus van de HAN uitgelicht. Er is in dit verslag bewust niet gesproken over wat de besparing in geld oplevert. Deels is dat omdat de energieprijs wellicht niet voor iedere instelling gelijk is en dat vele instellingen hun energieverbruik niet in kaart hebben, maar ook omdat de insteek duurzaamheid is en niet primair kostenreductie. Dat duurzaamheid hier een positieve bijdrage aan levert mag evident zijn. Om dezelfde reden zijn ook de investeringen om duurzamer te worden niet benoemd. Elke instelling zal daarin zijn eigen keuze moeten maken.

De mate van besparing door duurzaam om te gaan met data is van heel veel factoren afhankelijk en bij iedere instelling anders. Het vereist dan ook grondig onderzoek om gefundeerde uitspraken te doen over het gehele hoger onderwijs. Daarnaast zijn keuzes binnen onderwijsinstellingen omtrent wat er allemaal opgeslagen wordt, hoe veilig dit moet en welke eisen er worden gesteld aan de beschikbaarheid, naast nog andere keuzes, van sterke invloed de totaal verbruikte energie voor dataopslag. Vergelijkingen tussen de instellingen zijn daarom bijna onmogelijk en ook niet wenselijk. Deze zouden een verkeerd beeld kunnen geven van een bepaalde instelling. Om die reden zijn de geïnterviewde instellingen dan ook niet met naam genoemd.

Welke stappen worden er geadviseerd wanneer de instelling duurzamer om wil gaan met zijn data en is het dan ook zinvol? Voor wat betreft het zinvol zijn, kan worden opgemerkt dat de hoeveelheid data binnen onze organisaties naar verwachting de komende jaren fors zal blijven groeien. Hierdoor zal ook het energieverbruik in absolute zin groeien. Toch wordt het hoger onderwijs geconfronteerd met bezuinigen en in dat kader kunnen beschikbare gelden het eerst gebruikt worden voor laagdrempelige besparingsmaatregelen. Door zuiniger werkplekken, powermanagement op pc’s en vergroenen van datacentra valt veel winst te behalen. Energieverbruik voor dataopslag vormt slechts een klein deel van het totale energieverbruik van een organisatie. Dit neemt niet weg dat het raadzaam is om bij vervanging of uitbreiding van opslagsystemen energiebesparende maatregelen te nemen als het kan..

<u>Advies voor duurzame storage</u>.

- De meest duurzame oplossing voor dataopslag is de “delete” knop. Door het reduceren van de hoeveelheid data kan er direct bespaard worden op energieverbruik (archivering kan daarin faciliteren door inzicht in gearchiveerde data).
- Organisatiebreed beleid op gebied van bewaartermijnen. Door heldere beleidsregels en afspraken kan overbodige data verwijderd worden. Hierdoor is het mogelijk om procedures op te stellen en/of processen te automatiseren die het mogelijk maken om data permanent te verwijderen.

- Analyseer je storagesysteem(en). Is er teveel overcapaciteit en kunnen daardoor disks uitgezet worden? Hoe is de verdeling van de Tiers en kan data naar zuinigere Tier verplaatst worden? Is het mogelijk om de configuratie te wijzigen (meer grotere disks en minder kleine maar snellere disks).
- Mirroring en replicatie. Moet alle data gemirrored of gerepliceerd worden? Wanneer moet de data weer beschikbaar zijn en is de back-up voorziening dan voldoende? Minder uitwijkcapaciteit zorgt voor minder energieverbruik.
- Aanschaf van opslagsystemen/disks. Kijk naar beschikbare technieken als thin provisioning, deduplicatie, compressie en (auto-)tiering. Ook type disks en grootte zijn bepalend voor de energie-efficiëntie.
- Archivering. Onderzoek de eigen data en bepaal hoeveel archiveerbaar is. Door gebruik te maken van optical/tape of cloud voorzieningen kan waarschijnlijk, zeker bij grote hoeveelheden data, energiereductie behaald worden. Bij disk-gebaseerde systemen is dit waarschijnlijk het geval, bij SSD-gebaseerde opslagsystemen zou dit dit niet altijd het geval hoeven te zijn.

Opmerking ten opzichte van besparing in het hoger onderwijs.
De meeste besparing is te behalen bij het vernieuwen of uitbreiden van de storagesystemen. Van de 12 geïnterviewde onderwijsinstellingen hebben 10 hun opslagsystemen recent vernieuwd en/of uitgebreid. Hierdoor verwachten we niet dat er op korte termijn veel energiebesparing te realiseren is op storagesystemen. Daarnaast geven beheerders van storagesystemen aan terughoudend te zijn om de configuratie van hun systemen te wijzigingen. Performance, beschikbaarheid en stabiliteit wegen zwaarder dan een vermindering van energiegebruik. Ook bij aanschaf wordt energiebesparing niet als een doorslaggevende reden gezien.
Een conclusie hieruit kan zijn dat, om energiereductie te behalen,

- meer aandacht aan beleid met betrekking tot het bewaren van data gegeven moet worden,
- er gekeken moet worden hoeveel archivering bijdraagt aan energiereductie,
- er gekeken moet worden naar de TCO van SSD systemen. Indien deze verglijkbaar zijn dan levert dat een forse reductie op.