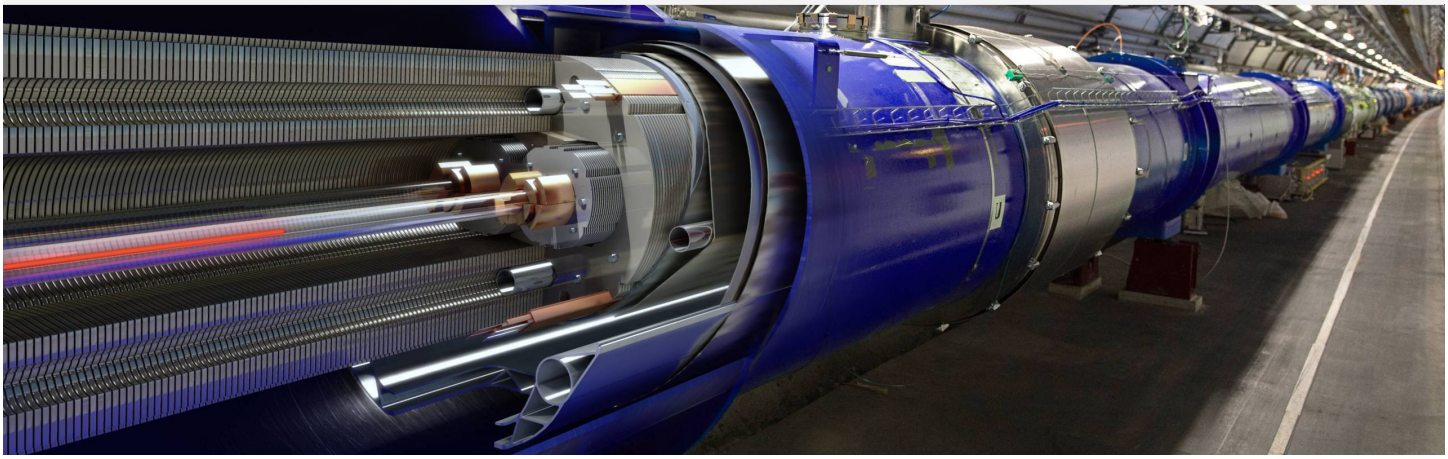


Datavloedgolf LHC op komst

Nikhef bereidt zich voor met rappe opslag

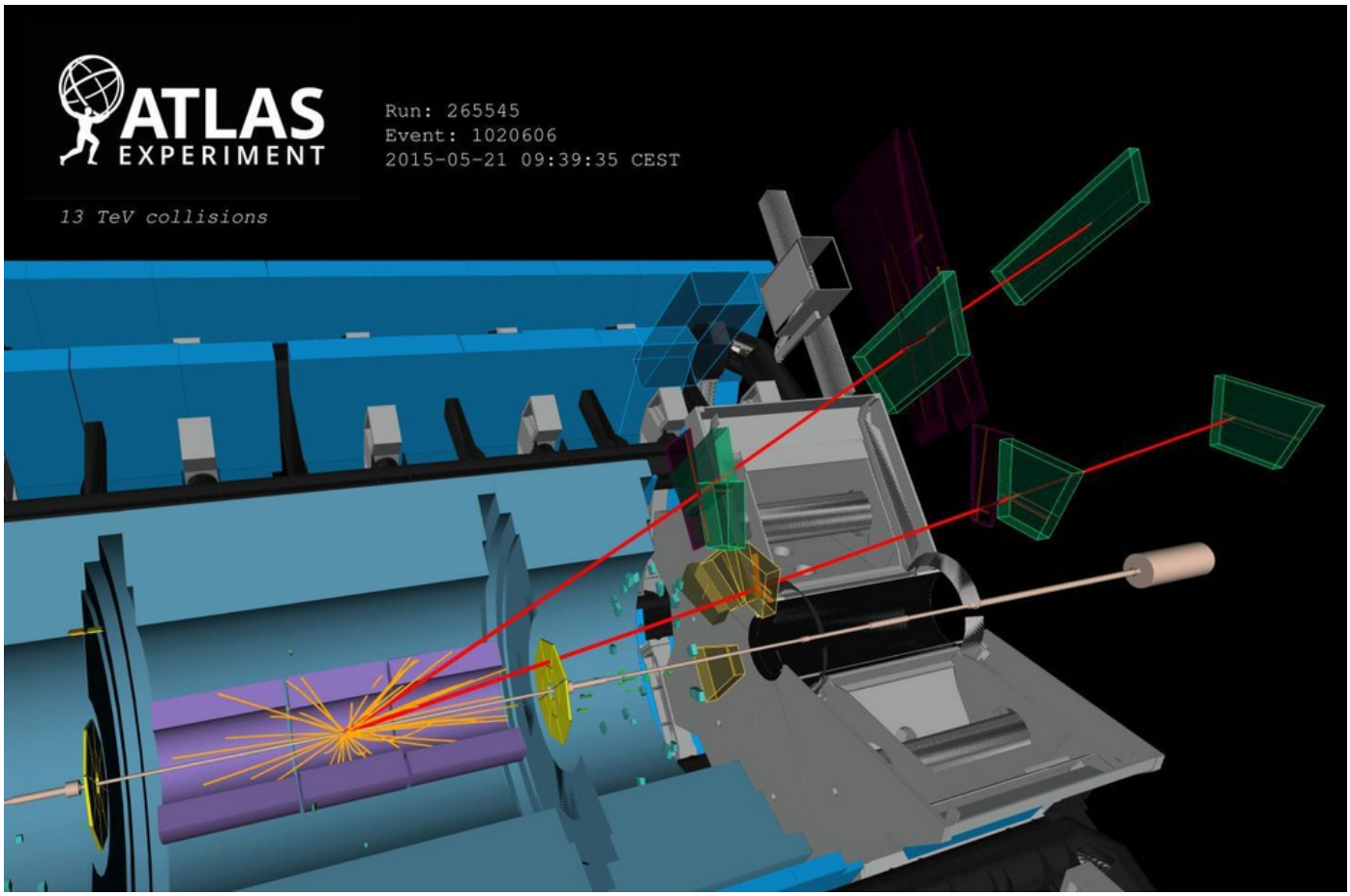
Door **Olaf van Miltenburg**
Nieuwscoördinator
Feedback • 17-11-2019 06:00

De kamer van Tristan Suerink bij Nikhef in Amsterdam heeft wel wat weg van de Tweakers-redactie. Bureaus zijn bezaaid met hardwareonderdelen, aan de muur hangen posters met geeky humor en in de hoek ligt een Nintendo Gamecube. Hier bereidt de afdeling Computer Technologie van Nikhef zich voor op de uitstoot van een datamonster dat zich duizend kilometer verderop bevindt en dat volgend jaar weer tot leven wordt gewekt: de Large Hadron Collider. "We krijgen zulke grote datastromen dat je die niet over het internet kunt pompen."

Nikhef is het Nationaal instituut voor subatomaire fysica, gevestigd op het Science Park in Amsterdam. Een belangrijke doelstelling van het instituut is het ontdekken van de bouwstenen van de natuur. In het verleden had het instituut hiervoor zijn eigen deeltjesversnellers, maar in de jaren negentig werd bij CERN besloten de zaken Europees en vooral veel grootser aan te pakken. Na jarenlange bouw werd in 2008 de LHC in de ondergrondse, ringvormige tunnel met een omtrek van 27km, op de Frans-Zwitserse grens, in gebruik genomen.

Zoals veel wetenschappelijke disciplines heeft subatomaire fysica een nauwe band met tech, maar bij Nikhef en CERN is de link met computertechnologie en internet nog net wat sterker. Niet voor niets is bij CERN eind jaren tachtig het wereldwijde web bedacht, door de Belg Robert Cailliau en de Brit Tim Berners-Lee. Nikhef was in die tijd al met het netwerk van CERN verbonden en zette in 1992 de derde website ter wereld online. In vitrinekasten bij het instituut staan computeronderdelen, zoals harde schijven uit de jaren zestig, die aantonen dat de geschiedenis met computertechnologie nog veel verder terug in de tijd gaat. Tegenwoordig heeft Nikhef een flinke eigen serverruimte met krachtige, moderne computerclusters en opslagsystemen en flinke datapijplijnen.

moesten we een apparaat bouwen dat zo groot is als het paleis op de Dam. Dat moesten we honderd meter in de grond stoppen om botsingen met een relatief hoog energieniveau op minuscule schaal mogelijk te maken, gemeten in tera-elektronvolts. Net als bij botsende auto's zien we onderdelen, zoals het stuur, de motor enzovoorts, alle kanten opvliegen."



Proton-protonbotsing gergistreerd door Atlas op 21 mei 2015. De groene delen geven detectie van twee muonen aan door de muonspectrometer. Bron: CERN

Het registreren van de gevolgen van de botsingen gebeurt met zeven grote detectoren. Nikhef is betrokken bij drie daarvan: Alice, Atlas en LHCb. In elke detector zitten honderdduizenden sensoren en elke sensor heeft weer een andere functie. "We hebben bijvoorbeeld muonkamers om muonen, verzwaarde elektronen, te detecteren. Hoe vluchtiger het deeltje, hoe dichter je op de bundel moet zitten. Als je de Atlas-detector als voorbeeld pakt, zou je die als een ui kunnen pellen", vertelt Suerink.

Tijdens experimenten vinden honderden miljoenen botsingen per seconde plaats, maar 'slechts' zo'n honderdduizend daarvan hebben potentiële waarde, de rest wordt dus weggefilterd. Algoritmes brengen het aantal vervolgens verder terug tot honderd à tweehonderd *events of interest* per seconde. Volgens Suerink is er speciale elektronica ontwikkeld om de momenten te bepalen waarop de detectoren een 'foto' moeten maken. "Rauw komen er terabytes per seconde aan data uit. Dat valt niet te verwerken en niet op te slaan. Effectief gooien we meer dan 99 procent weg. Wat overblijft, is de data die we verwerken, die we correleren en analyseren."

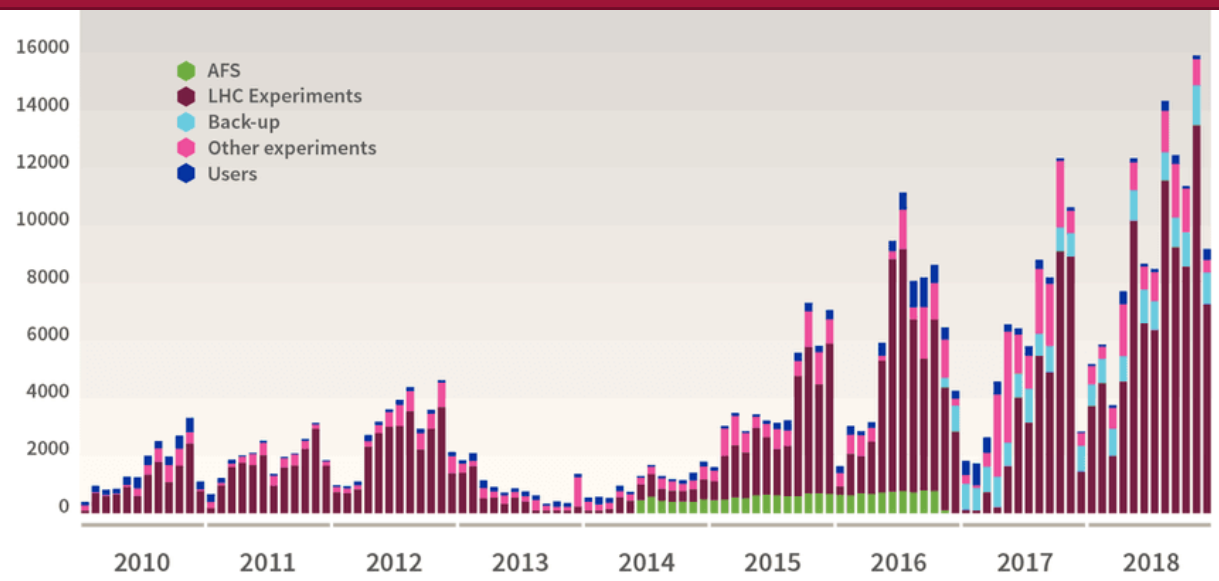
Naast elke LHC-detector zit een eigen computingcluster om de output aan een verwerkingsslag te onderwerpen. Suerink: "Daar zitten heel veel fpga-kaarten in. We kijken of we bij volgende generaties gpu's kunnen gebruiken. Dat is nog onduidelijk. Dat cluster doet ook werk voor het verpakken van de data." Vervolgens wordt de data

Het datacentrum van CERN is de zogenoemde tier 0 in het Worldwide LHC Computing Grid. Dit is een in 2005 ontworpen infrastructuur die bedoeld is om de destijds verwachte 50 tot 70 petabyte per jaar aan data van de LHC te verdelen en te verwerken. Het grid omvat inmiddels 170 computinglocaties in 42 landen wereldwijd. In totaal heeft het grid zo meer dan een miljoen cores en een exabyte aan opslag tot zijn beschikking. Wetenschappers van over de hele wereld kunnen inloggen op het grid en jobs voor rekenwerk op de data inschieten. Ze kunnen daarmee aangeven welke datasets ze willen en dan wordt gekeken wat de dichtstbijzijnde locatie is waar die data staat. De gegevens kunnen hen helpen bij hun werk om te achterhalen hoe de werkelijkheid op een fundamenteel niveau in elkaar steekt. En dat kan uiteindelijk weer tal van praktische toepassingen opleveren.

Datacenter bij CERN. Bron: CERN

De kern van het grid is het Large Hadron Collider Optical Private Network. Dit netwerk bestaat uit de tier 0-locatie van CERN en dertien tier 1-locaties, in onder andere Duitsland, Italië en Frankrijk, maar ook in de VS, Zuid-Korea en Taiwan. En dus ook in Nederland, bij Nikhef en SURFsara. Elke tier 1-locatie krijgt zo'n 10 procent van de data van de deeltjesversneller te verstouwen voor analyse en opslag. Om die data van de LHC te kunnen ontvangen, hebben Nikhef, SURFsara en SURFnet een directe glasvezelverbinding met CERN in Zwitserland. Over dezelfde glasvezel gaan ook andere directe verbindingen vanuit CERN naar andere tier1-locaties, via SURF's optische exchange Netherlight. Als de verbinding via het reguliere internet zou lopen, zou dat onherroepelijk tot opstoppingen leiden. Tot 2015 was Nikhef met een 10Gbit/s-lijn met CERN verbonden, maar sindsdien is er een 100Gbit/s-verbinding.

De datastroom van de grote deeltjesversneller komt in golven. De experimenten van de LHC gebeuren namelijk in *runs*. De eerste run duurde van 2009 tot 2013. Daarna volgde de Long Shutdown, die twee jaar duurde, een periode waarin wetenschappers reparaties en upgrades uitvoerden. Door de upgrades verdubbelde het energieniveau van de botsingen bij LHC Run 2, die van 2014 tot en met 2018 duurde, naar 13TeV. Mede hierdoor ging ook de hoeveelheid data flink omhoog. Tijdens Run 1 ontving CERN data met zo'n 1Gbit/s en met pieken van 6Gbit/s. Tijdens Run 2 was 8Gbit/s het gemiddelde, met pieken tot boven de 10Gbit/s. In 2018 sloeg CERN zo meer dan 115 petabyte weg naar zijn tapeopslag, waarvan 88PB pure LHC-data. In de maand november 2018 ging het om een piek van 15,8PB, meer dan in een heel jaar tijdens de eerste run.

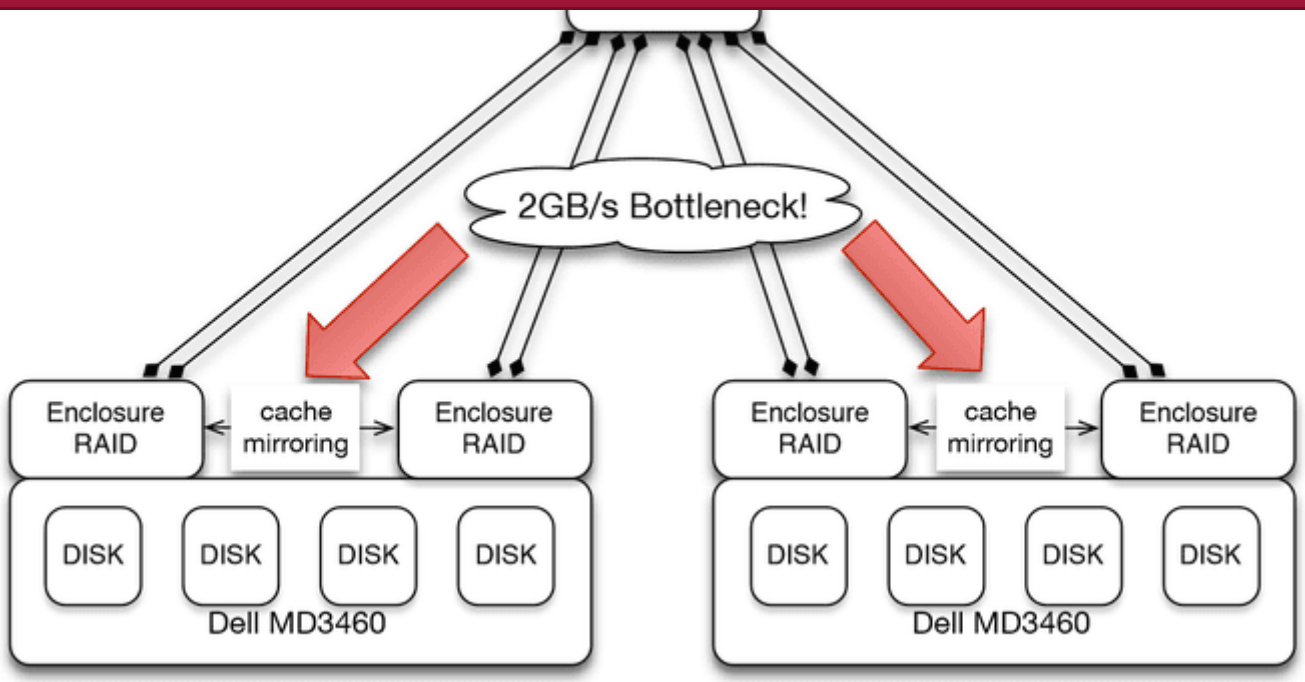


Data opgeslagen op tape door CERN van maand tot maand, in terabyte. Bron: CERN

Een nieuwe tsunami dient zich aan. Momenteel is Long Shutdown 2 bezig, maar CERN maakt zich op voor Run 3, die in 2021 van start gaat en tot en met 2023 duurt. De verwachting hierbij is dat de datahoeveelheid door de upgrades zal verdubbelen. In 2026 begint vervolgens de vierde Run van de High Luminosity LHC, zoals de versneller na de upgrades zal heten. Er komen dan veel meer botsingen, waardoor Run 4 ten opzichte van Run 3 nog eens een vervijfvoudiging zal opleveren. De verwachting is dat CERN tegen die tijd 500PB per jaar zal moeten wegschrijven. En daarmee ziet ook Nikhef een vloedgolf op zich afkomen.

Een eerste stap is het aanleggen van een bredere snelweg; Nikhef gaat volgend jaar een 400Gbit/s-verbinding met CERN testen. Maar al die data moet ook worden opgeslagen. "Tussen nu en pakweg vijf tot zes jaar komt er tien keer zoveel data uit, maar mijn budgetten blijven hetzelfde", verzucht Suerink. "Dus ik moet slimmer inkopen of ik moet tot magische methodes komen om de data efficiënter te verwerken. Als je kijkt hoe de processorontwikkeling in de laatste paar jaren is geweest, zie ik daar niet zo heel veel profijt van." De it-architect hanteert een soort ticktockmodel, waarbij hij het ene jaar een compute- en het andere jaar het storagedeel een upgrade geeft.

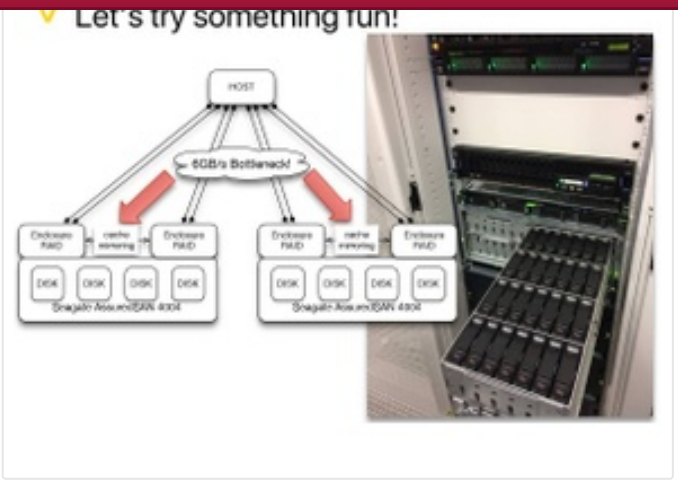
Al in 2016 constateerde Suerink dat hij tegen een bottleneck aanliep. "Wij kochten vrij lang Dell PowerVault-systemen. Waar we mee zaten, is dat er tussen twee controllers een maximale snelheid van 2GB/s voor cachemirroring zit. Wij willen die cachemirroring hebben. Op het moment dat er een controller uitklapt, wil je geen dataverlies. Dan raak je alles kwijt wat in de cache staat. Als we een PowerVault vulden met 4TB-schijven, hadden we een performance van 1,8GB/s tussen die twee caches nodig. Dat betekent dat als we naar 6 of 8TB zouden gaan, de cachesnelheid gewoon te laag was."



Bron: Data To Network: building balanced throughput storage in a world of increasing disk sizes, Tristan Suerink, HEPiX Srping 2019

Tijdens de Supercomputing-conferentie van 2016 ging Suerink met zijn eisen langs de verschillende storagefabrikanten. "Ik vertelde dat ik een doos wilde waarin ik harde schijven kon duwen, liefst hardwarematige raid achterin en gewoon met sas eruit, en ik wilde dat de cachemirroring zo snel mogelijk gaat." De storageproducten van de meeste fabrikanten waren niet erg geschikt voor de manier waarop ze bij Nikhef met data omgaan. "Wij hebben heel veel bestanden en heel veel jobs die heel veel bestanden lezen. De hoeveelheid berekeningen per terabyte die we doen, is heel klein. We streamen veel data waar we vluchtig doorheen kijken. Het is niet alsof we heel lang moeten kauwen om te weten wat erin staat. Als je dat met een *shared filesystem* doet, gaat het stuk." Ondanks die specifieke eisen stelt Suerink dat de opslagsystemen bij Nikhef in feite een eenvoudige ftp-daemon draaien met nog wat tools voor een minimalistische catalogus.

Suerink kwam uiteindelijk bij Seagate terecht. Dat had in 2015 storagespecialist Dot Hill overgenomen en leverde daarom enclosures met zelf ontworpen hardwareraidchips. "Die konden tot een maximale cachemiroring van 6GB/s. Hee, dacht ik, dan kan ik doorgroeien naar 12TB-hdd's. Dat is handig." Nikhef schafte een Fujitsu RX2530M4 aan met twee Intel Xeon 4110-processors in combinatie met een Seagate AssuredSAN 4004 met 56 8TB-hdd's. Met het oog op de toekomst besloot hij een test te houden om de prestaties van dat Intel Skylake-systeem te vergelijken met die van een Dell R7415 met een AMD Epyc 7451-cpu en een IBM S822L met een Power8-processor uit 2014. Bij de test werden de drie systemen gekoppeld aan twee AssuredSAN 4004-opslagsystemen. Suerink gebruikte dezelfde schijven en dezelfde kabels, alleen de cpu was anders. Hij zette vervolgens Ubuntu erop en draaide zijn eigen testtool om de prestaties te bepalen. Tot zijn verbazing presteerde de IBM-processor aanzienlijk beter dan de Intel en de AMD. "En ik had de Intel en AMD meer getuned dan de Power8. De uitkomst was interessant voor mij; de IBM presteerde drie keer zo goed voor twee keer zoveel geld."



In aanloop naar het moment waarop voor dit jaar het opslagsysteem moest worden aangeschaft om de komende LHC-Run aan te kunnen, had storagefabrikant Netapp een nieuwe enclosure aangekondigd met controllers die ondersteuning bieden voor hdd's tot 16TB. En IBM had ondertussen de Power9 op de markt gebracht. Suerink: "Het is gelukt om op basis daarvan voor een strakke prijs een heel opslagsysteem neer te zetten."

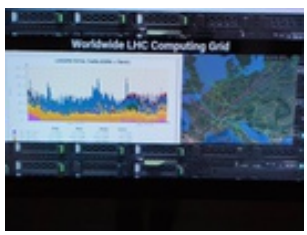
Tristan Suerink:

"We draaien met 4x raid6 met 15 schijven per raidset, aangesloten met 4x sas3 aan het IBM-systeem in een kruisopstelling. Hierdoor heeft elke hba een directe connectie met elke controller. De luns zijn gelijk verspreid over de controllers. Tussen de IBM en de Netapp draaien we multipath plus alua. Dit doen we om zo elk lun z'n eigen pad te geven en hierdoor de maximale bandbreedte te kunnen gebruiken, maar toch ook nog redundantie te behouden bij uitval van een pad, controller of hba. Elk opslagblok heeft netto 556TiB bruikbare opslag en de vier blokken bieden samen een throughput van 300Gbit/s vanaf de schijven. Als switch gebruiken we een Mellanox SN2100 met daarop Mellanox Onyx als besturingssysteem."

Nikhefs nieuwe storagecluster, vier opslagblokken met elk:	
1x IBM LC922 - 9006-22P	**1x Netapp E5760**
2x IBM SMT-4 Power9-cpu, 16 cores, draait standaard op 2,9GHz	60x HGST Helium 12TB-schijven
128GB ram	Dubbele sas3-controllers
1x Mellanox Connect-X4 als netwerkkaart, aangesloten met 100Gbit/s twin-ax aan het netwerk	
2x LSI 9305-16e sas-hba's.	
2TB hardware raid1 voor het besturingssysteem	
CentOS7-ppc64le als besturingssysteem	

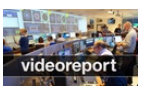
Dat systeem is inmiddels aan het zoemen in het tier 1-datacentre op het Amsterdam Science Park. Is Suerink niet bang dat er alweer betere systemen zijn als de komende LHC-runs in volle gang zijn? "Omdat we

gekocht om er productie-ervaring mee op te doen. We hadden de opslagruimte hoe dan ook nodig voor de verdere verwerking van de data die bij de vorige run is verzameld. En er waren systemen waarvan de support eindigde, dus dit dient als vervanging daarvoor. Daarnaast is het een soort statement: dit is ook een manier om het te doen."



Binnenkort verschijnt een artikel dat dieper ingaat op de LHC en de upgrades.

Lees meer



Op zoek naar de kern - deel 3: Klaar voor de toekomst
Video van 13 september 2015