Verslag SurfNET workshop 24-26 sept 2018
Building high-performing campus infrastructures for research

Simon Oosthoek

11 oktober 2018

Inhoudsopgave

Tabel 1: Data volume wetenschappelijke instrumenten

Instrument	totaal volume	groei per jaar
LHC	250PB	50-70PB
LOFAR	40PB	7PB
SKA	-	300PB

1 3-day workshop: Building high-performing campus infrastructures for research

Een workshop georganiseerd door SURF en ASTRON samen met partners GÉANT, PSNC, ESnet. Relevante namen: Richard Hughes-Jones (GÉANT), Michael Sinatra (ESnet), Ron Trompert en Pieter de Boer (SURF), Szymon Trocha en Antonie Delvaux (perfSonar). Aanspreekpunt was Mary Hester die bij SURF werkt als Community-manager onderzoek .

1.1 Doel van de workshop

Er zijn een groeiend aantal wetenschappelijke instrumenten die steeds grotere data sets produceren, zoals de LHC, LOFAR en vergelijkbare apparaten. De toekomstige ontwikkelingen zoals de Square Kilometer Array (SKA) zullen wederom grotere hoeveelheden data produceren.

De workshop gaat over de problemen en mogelijke oplossingen voor het transporteren en delen van deze data. Er zijn twee hoofdlijnen, namelijk de “Science DMZ” als architectuur en perfSonar als gereedschap om de verwachte bandbreedte te monitoren en ”soft-issues”met het netwerk te kunnen opmerken en opzoeken. **NB** de term *Science DMZ* is afkomstig van ESnet en heeft niks te maken met *Faculty of Science*. Als we bij de Radboud universiteit een Science DMZ zouden willen inrichten zou de meest voor de hand liggende locatie dicht bij de fysieke border router zijn, dus in het Forum, neem ik aan.

2 High-speed netwerken

Bepaalde soorten wetenschappelijke instrumenten produceren enorme hoeveelheden meetdata. De meetdata is vaak openbaar, maar vanwege de afmetingen moeilijk lokaal te bewerken. Alleen al het transport van de data van het meetapparaat (LHC, LOFAR, SKA) naar de permanente opslag kan hoge eisen stellen aan de transportcapaciteit.

De manier van werken, waarbij de data opgehaald wordt, berekeningen erop gedaan worden en de resultaten gepubliceerd worden is niet realistisch meer voor de grootste data sets en dat wordt steeds erger, is de stelling van prof. Carole Jackson. Een alternatieve manier van werken zou zijn dat je de berekeningen laat doen door een cluster dat bij de data staat.

High-speed is natuurlijk een relatieve term voor wat momenteel het hoogst haalbare is. 100Gbit/s is nu commercieel haalbaar voor intercontinentale links en ook over land-lijnen gebruikelijk voor verbindingen tussen nationale wetenschappelijke netwerken (NRENs).

Een volgende stap (anders dan multi-100Gbit/s links gekoppeld) is nog niet in beeld voor productielinks, denk ik.

De ervaringen tot nu toe met high-speed netwerken zijn typisch het onderwerp van discussie tijdens deze workshop. Terugkomend thema is TCP performance problemen bij zelfs minimale packet-loss. En de impact van roundtrip times op de TCP buffer-size (bij 100Gbit/s en grote rtt ($\tilde{1}$00ms) loopt dit tegen de grens van TCP aan; buffer van 4GByte).

Als het over SurfNet gaat, de beschikbare bandbreedte wordt bij lange na niet optimaal benut. Ze hanteren een richtlijn dat ze een upgrade gaan plannen bij meer dan 70% gebruik, momenteel zit het gebruik rond de 10%. Het plan is om met surfnet8 betere utilisatie te bereiken, onder andere door voorbij de core-surfnet links te kijken en met campus-netwerk beheer te praten en kennis te delen.

3 Architectuur van Campusnetwerken

Een gebruikelijke architectuur van een campus-netwerk is de “alles achter de firewall” architectuur. Vanuit een security perspectief lijkt dit de beste aanpak om de organisatie te beschermen tegen kwalijke invloeden van buitenaf. Naarmate de netwerken sneller worden worden de performance eisen voor een firewall navenant hoger en de prijs dus ook. Niet alle faculteiten, onderzoeksgroepen of services hebben de snelste netwerken nodig en er zijn ook verschillende risico profielen voor de verschillende groepen en services. Een (weer) opkomend fenomeen is het segmenteren van het netwerk op basis van functie of security profiel. Dit maakt het makkelijker om security policies overzichtelijk te houden en de functionaliteit niet te veel te hinderen.

Een Science DMZ is typisch een netwerk dat verbonden is met de border-router van een campus, buiten de “corporate” firewall. De hosts in de Science DMZ zijn beveiligd met statische ACL regels in de router en op de host zelf een linux gebaseerde firewall (iptables of nftables). De enige hosts die typisch in de Science DMZ aanwezig zijn, zijn de data transfer node (DTN) en de perfSonar monitoring node.

3.1 IPv6

Door externe afhankelijkheden is het noodzakelijk om IPv6 naast IPv4 op de science DMZ actief te hebben. Een voorbeeld van externe afhankelijkheid is de DTNs van bijvoorbeeld CERN, die binnen afzienbare tijd uitsluitend via IPv6 te benaderen zijn.

3.2 perfSonar

Een belangrijk onderdeel van de Science DMZ is een perfSonar node, om te kunnen monitoren of de DTN wel optimaal zijn werk kan doen. Het is zonde van de investering in een Science DMZ als je niet ook kunt sturen en ontvangen op de maximale snelheid. De enige manier om boven op link-beschikbaarheid te monitoren is door er regelmatig tests op te draaien. Door in-band metingen te doen ben je snel op de hoogte van dips in de performance en kun je op zoek naar de oorzaak.

perfSonar is een open-source netwerk-monitoring toolkit. Het kan op verschillende Linux distributies geïnstalleerd worden, maar is ook beschikbaar als een kant en klare ISO op CentOS gebaseerd of als docker image. Op de gelinkte pagina is een beslisboom te vinden welke installatie-optie het meest past.

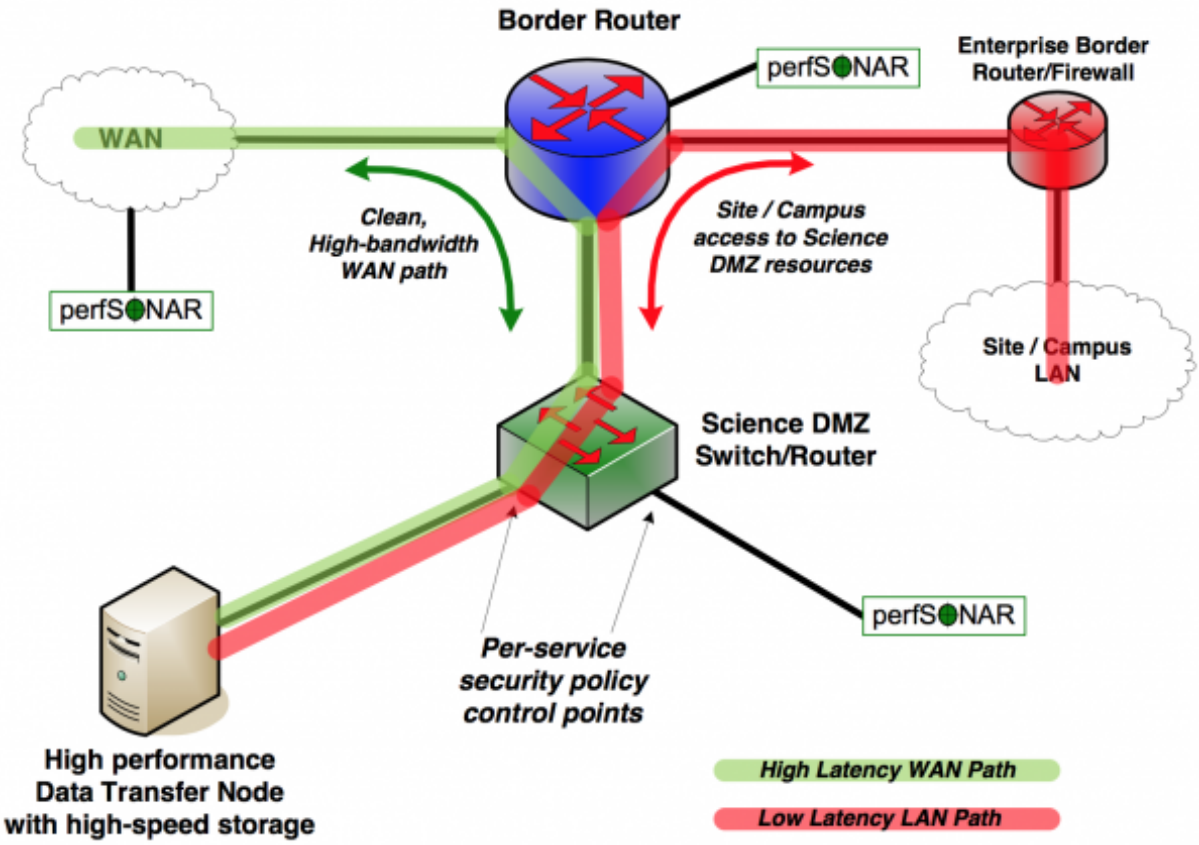
Voor een bruikbare meting heb je bij voorkeur een hardware node nodig en een NIC met een snelheid die past bij de maximale snelheid die je wilt kunnen meten. De webgui en dashboard hoeven niet per se op een hardware node te staan, maar als je die toch al hebt kan het ook geen kwaad. Throughput metingen boven de 25Gbit/s zijn waarschijnlijk voorlopig niet heel nuttig, omdat dat alleen haalbaar is met speciaal getunede setups. Bovendien kun je packet-loss en roundtrip-times prima meten met lagere snelheid links en geeft ook waardevolle informatie over de netwerk performance (met name voor TCP).

4 Security

Het beveiligen van netwerken, met bijvoorbeeld een stateful firewall, heeft impact op de maximale bandbreedte. Hoge bandbreedte stromen hebben ook impact op de firewalls, waardoor die niet optimaal kunnen werken voor “regulier” verkeer.

Het segmenteren van netwerken met verschillende security eisen (risico profielen) kan hierbij helpen en wordt steeds meer toegepast. Ook het gebruik van stateless ACLs op de routers in combinatie met host-based firewalls (iptables/nftables) wordt als oplossing geadviseerd.

Het idee van een “Science DMZ”, zie figuur 1, is dat je een segment in je netwerk maakt dat aangesloten is op de externe router, dus buiten de firewall. De enige functie van de Science DMZ is het ontsluiten van een Data Transfer Node (DTN) op (zo goed als) line-speed, zonder de rest van het netwerk te overbelasten. Om de maximale performance te halen en te bewaken kun je perfSonar nodes inzetten. Er zijn ook internationale perfSonar nodes die je kunt gebruiken (in overleg) om metingen te doen naar andere universiteiten.



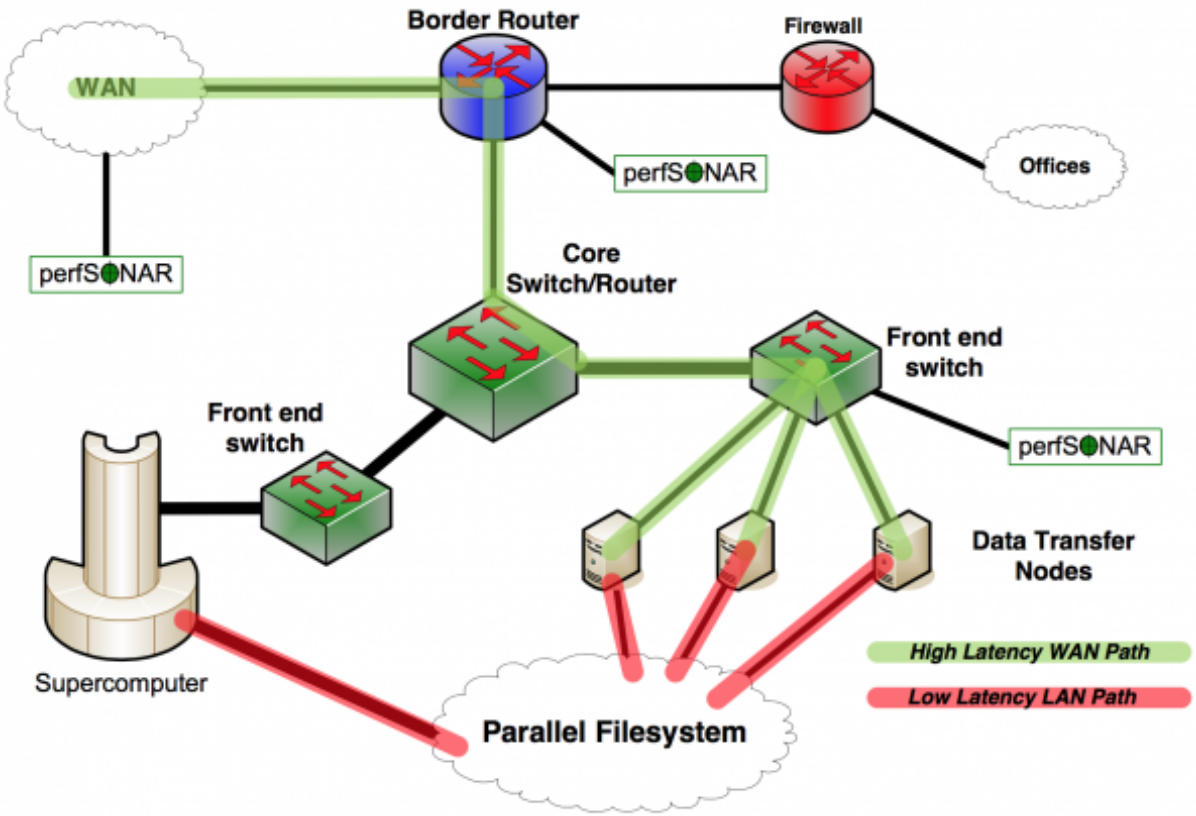
Figuur 1: Basis Science DMZ architectuur met een DTN en perfSonar node

4.1 Gevoelige of geheime data delen

Bij een DTN in een science DMZ is er mogelijk sprake van het delen van data die niet openbaar is of mag worden. Link-versleuteling kan voor vertraging zorgen en daarbij, als de data onversleuteld op de disk staat, is er nog een risico dat de data gelekt wordt als de DTN gehackt zou worden. Een encrypted filesysteem of beter nog, versleutelde bestanden met data, hebben bij het versturen geen performance impact en zijn veilig (bij goed sleutelbeheer) bij een security breach.

5 Troubleshooting

Link down problemen zijn makkelijk te detecteren en vinden, maar performance problemen zijn lastiger. Door perfSonar nodes strategisch in je netwerk te plaatsen kun je beter in kaart brengen waar mogelijke bottlenecks zitten en de oorzaak achterhalen. Bij hogere bandbreedte glas-links is het van belang om de koppelvlakken van transceiver naar de vezels goed schoon te hebben.



Figuur 2: Science DMZ architectuur met shared storage en HPC

6 Monitoring

Eigenlijk zijn er twee soorten monitoring voor het meten van grijze gebieden, security monitoring en performance monitoring.

6.1 Security Monitoring met Bro

Bro zit niet tussen het verkeer, maar draait op een aparte machine en kijkt mee (via een monitoring poort op de switch) en detecteert patronen in het verkeer, zelfs tot aan deep-packet inspection toe. Het is mogelijk om Bro te laten reageren op allerlei soorten events en actie te nemen (bijvoorbeeld de router aansturen of wat dan ook).

6.2 Performance Monitoring met perfSonar

Door een mesh van diverse perfSonar nodes regelmatig tests te laten doen voor latency, throughput en packet-loss, kun je op tijd reageren op performance problemen en in de matrix van nodes (MaDDash) snel zien wat de knelpunten zijn. In de historische data kun je mogelijk zien wanneer een probleem is begonnen.

In de Science DMZ FAQ wordt duidelijk gesteld dat een perfSonar node in de Science DMZ niet optioneel is. Dit komt omdat de enige manier om te garanderen dat de DTN zijn werk goed kan doen, het actief *proben* van het netwerk is.

7 Storage

Bij SurfSARA hebben ze verschillende storage systemen, maar voor hele grote data sets blijven er maar een paar kandidaten over die voldoen; Openstack Swift en Ceph. Swift kent weinig grenzen, maar biedt alleen object store aan. Ceph heeft wel een bovengrens die vooral praktisch is, ter indicatie, bij CERN hebben ze nu een 65 PB cluster draaien. Het voordeel van Ceph is dat je drie manieren hebt om data op te slaan/benaderen; Filesysteem, Block, Object.

CERN LHC Bigbang III, april 2018

- 225 servers
- 48 6TB hard drives per server
- 10,800 OSDs in total, nearly 65 petabytes!
- 3 monitors and managers, colocated with OSDs

Ceph is, volgens Ron Trompert van SurfSARA, al snel interessant als je meer data wilt opslaan dan 200 TB of hoeveel er in 1 server past. Je hebt wel minimaal 3 nodes nodig om voldoende redundantie te kunnen bieden.

7.1 Storage voor DTNs

Voor de maximale throughput van je DTN komt de hele keten in beeld van disk naar disk, via het netwerk. Bij tientallen Gb/s begint de bottleneck te verschuiven van het netwerk naar disk, CPU, PCI-bus en dergelijke. Richard Hughes-Jones had een interessant praatje over het proces van de hoogste performance zoeken tussen de UK en Australië. Een verrassende uitkomst was dat het uitmaakt op welke Socket+Core de interrupts worden afgehandeld en waar de applicatie voor de transfer software draait. De IRQ CPU moet direct gekoppeld zijn aan de PCI bus, bij de moederbord architecturen is het gebruikelijk dat niet alle sockets directe verbinding met de PCI bus hebben.

Het valt niet mee om alle systemen in de keten van NIC, disk/NVMe, PCI-bus, CPU-core en applicatie allemaal zo optimaal mogelijk te laten functioneren en dan nog haal je de 100Gbit/s niet.