Samenvatting

Wanneer we naar de nachtelijke hemel kijken, observeren we het universum door zichtbaar licht. We kunnen sterren, planeten, nevels en misschien zelfs andere sterrenstelsels zien. De vroegste astronomen bestudeerden het Universum met hun ogen, uiteindelijk geholpen door spiegels en lenzen in de vorm van telescopen en verrekijkers. In de 18e en 19e eeuw hebben natuurkundigen het volledige elektromagnetische spectrum ontrafeld. Naast zichtbaar licht, konden astronomen nu het universum bestuderen met behulp van de infrarood-, ultraviolette, microgolf- en radiodelen van het spectrum. Licht uit elk deel van het spectrum draagt verschillende informatie met zich mee en kan worden gebruikt om specifieke fysische fenomenen te bestuderen. Het nadeel van deze breedte van informatie is dat elk deel van het spectrum zijn eigen specifieke detectoren nodig heeft, en vaak hele telescopen. Toch onthult elk deel van het spectrum een deel van het universum dat voor ons verborgen is in alle andere golflengten.

Radioastronomie werd geboren in de jaren 1930 met de experimenten van Karl Jansky met een gerichte 30-meter radio-antenne. Hiermee kon hij onweersbuien detecteren, de magnetosfeer van de Zon, maar ook een vreemde onbekende bron in het centrum van onze sterrenstelsels. Deze experimenten bewezen dat laagfrequente radiogolven konden worden gebruikt om het verre heelal te bestuderen. In de jaren 1940 werden radarontvangers ontwikkeld als een kwestie van nationale veiligheid tijdens de Tweede Wereldoorlog. Ze worden beschouwd als de belangrijkste reden voor de overwinning op de Luftwaffe in de Slag om Engeland. Na het einde van de oorlog werd een deel van de radarhardware op de ruimte gericht.

Terwijl de nieuwe antennes veel gevoeliger waren voor radiofrequenties, hebben radiotelescopen een fundamentele beperking in termen van resolutie. Omdat licht zich als een golf gedraagt, vooral bij lange golflengten, wordt het beperkt door de diffractielimiet. De diffractielimiet verbindt het kleinste object dat

ruimtelijk kan worden onderscheiden door een telescoop met de golflengte van het licht en de diameter van de telescoop. Als we een radiotelescoop van 100 m nemen, heeft deze de onderscheidingsvermogen die gelijkwaardig is aan die van een zichtbare telescoop met een spiegel met een diameter van 0,5 mm. Dit zou niet al te te schokkend moeten moeten zijn: radiogolven zijn meer dan 10 miljoen keer langer dan het licht dat we waarnemen, dus onze telescopen moeten natuurlijk 10 miljoen keer groter zijn.

Om de resolutie van de Hubble-ruimtetelescoop te evenaren, zou de spiegel van een radiotelescoop op 150 MHz een diameter van $\sim$ 1000km nodig hebben. Terwijl ingenieurs nog steeds bezig zijn met het maken van radarchotels van dit formaat, hebben astronomen zich tot computers gewend om de hoekresolutie van radiotelescopen te vergroten. Met behulp van de golfeigenschappen van licht kunnen radioastronomen de timing synchroniseren van de gegevens die zijn waargenomen op meerdere antennes, gescheiden door tientallen, honderden of duizenden kilometers. Met aanzienlijk rekenwerk kunnen gegevens van deze ver uiteen liggende antennes worden gecombineerd om een beeld te maken met een resolutie de gelijkwaardig is aan die van een radiotelescoop met een schoteldiameter van honderden kilometers. Het nadeel van deze combinatie van antennes is de benodigde rekenkracht voor het maken van dergelijke wetenschappelijke afbeeldingen.

De LOFAR (LOw-Frequency-ARray) radiotelescoop is een Nederlandse laagfrequente telescoop die bestaat uit meer dan 5000 antennes in Nederland, met duizenden meer gegroepeerd in stations in heel Europa. LOFAR observeert de radiohemel met de laagste frequenties zichtbaar vanaf de aarde: van 10 MHz tot 240 MHz. Het is ontworpen om gegevens te verzamelen voor meerdere wetenschappelijke toepassingen. Wetenschappers kunnen LOFAR-gegevens gebruiken om fenomenen te bestuderen zoals supernovaresten en pulsars in ons sterrenstelsel, magnetische velden van het zonnestelsel, samenvoegende melkwegclusters, zwarte gaten in het centrum van het verre sterrenstelsels en zelfs het tijdperk waarin de eerste sterren in ons universum werden gevormd. Om zulke uiteenlopende wetenschappelijke gevallen te dienen, worden LOFAR-gegevens opgeslagen met hoge tijd- en frequentie-resoluties. Dit leidt tot aanzienlijke gegevensgroottes. Elke observatie van 8 uur kan gemakkelijk acht harde schijven van 2 TB vullen. Het maken van een afbeelding op basis van deze gegevens is net aan mogelijk op iemands PC en een survey van de gehele hemel bestaan uit duizenden observaties die niet op één computer of zelfs een klein cluster van computers kunnen worden verwerkt. Desondanks kan elke verwerkingscyclus meerdere dagen duren, een vertraging die niet acceptabel is voor projecten die binnen vijf jaar duizenden datasets produceren.

Om duizenden waarnemingen van meerdere petabytes te kunnen verwerken, maken we gebruik van supercomputers voor parallelle verwerking van radioastronomiegegevens. Parallelliseren houdt in dat al onze gegevens kunnen worden opgesplitst in veel verschillende stukken, die elk onafhankelijk kunnen worden verwerkt. Bij een voldoende aantal computers daalt de verwerkingstijd met een factor 10 of meer. Het versnellen van de eerste paar verwerkingsstappen levert een tweede voordeel op: de eerste stappen reduceren de grootte van gegevens met een hoge resolutie tot 64 keer. Opeens past elke observatie gemakkelijk op uw desktop, laptop, en zelfs op een micro-SD kaart! Ze kunnen ook in seconden in plaats van uren tussen universiteiten worden uitgewisseld.

Dit werk is erop gericht om wetenschappers in staat te stellen van wetenschappers om eenvoudig en snel LOFAR-gegevens te verwerken, waardoor het gemakkelijker wordt om grote gegevenssets te gebruiken om wetenschappelijke studies uit te voeren. We maken gebruik van open source software, een verwerkingsinfrastructuur met een hoge verwerkingscapaciteit en de eigenschappen van gegevensverwerking via radioastronomie om grote wetenschappelijke onderzoeken met LOFAR mogelijk te maken. Ons werk is echter niet alleen nuttig voor wetenschappers. We bouwen en presenteren tools waarmee we de softwareprestaties van complexe wetenschappelijke pijpleidingen kunnen bestuderen en gebruiken deze tools om efficiëntere verwerkingsstrategieën aan te bevelen. Onze resultaten zijn ook van toepassing op toekomstige verwerkingsprojecten, zowel voor LOFAR-gegevens als voor de volgende generatie radiotelescopen. Uiteindelijk kunnen de lessen die zijn geleerd van de uitgebreide LOFAR-observaties worden gebruikt voor toekomstige astronomische projecten met grote hoeveelheden gegevens.

In Hoofdstuk Één geven we een overzicht van de geschiedenis van wetenschappelijk computergebruik, radio-interferometrie en de uitdagingen bij de verwerking van LOFAR-gegevens. In Hoofdstuk Twee presenteren we een platform voor grootschalige gedistribueerde LOFAR-verwerking. Dit platform is gebouwd voor geavanceerde LOFAR-gebruikers die hun verwerking willen parallelliseren op een gedeeld systeem, verdeeld over meerdere computationele toestelen en clusters. We introduceren de gebruikte softwarepakketten en hun interacties en bespreken de noodzaak van een dergelijk platform voor huidige en toekomstige grootschalige astronomische studies.

In Hoofdstuk Drie introduceren we een raamwerk voor het starten, volgen en parallelliseren van verwerkingsopdrachten voor de LOFAR-telescoop. Het is de eerste keer dat LOFAR-gegevens in bulk werden verwerkt op een High Throughput Infrastructuur. We laten zien dat dit computerparadigma kan worden toegepast op

de eerste stappen van een enkele verwerkingspijplijn. We geven een suggestie hoe andere pijpleidingen en telescopen dit kader kunnen gebruiken om hun gegevens te verwerken. Ten slotte laten we een aanzienlijke versnelling van de gegevensverwerking zien in vergelijking met de gegevensverwerking in Leiden: tot 35 keer sneller.

In Hoofdstuk Vier gaan we de uitdaging aan om prestatiegegevens te verzamelen van gedistribueerde runs van een complexe pijplijn. Voortbouwend op ons werk met betrekking tot parallelle verwerkingspijplijnen, bouwen we software om de prestaties van elke verwerkingsstap te volgen. De gegevens die we bijhouden, worden verzameld op een centrale server en kunnen in realtime worden geanalyseerd of worden bestudeerd na de verwerkingsrun. We voeren tests uit op vier verschillende systemen en constateren dat de compilatiemethode de prestaties van de verwerking niet verslechtert. We krijgen ook inzicht in de prestaties de lage prestaties van onze langzaamste stappen. Onze bewakingssoftware biedt een rijke dataset voor wetenschappelijke softwareontwikkelaars om inzicht te krijgen in de realtime prestaties van de gegevensreductiepijplijnen.

In Hoofdstuk Vijf introduceren we de eerste workflow-orkestratiesoftware voor complexe LOFAR-pijpleidingen. Deze software is het best geschikt voor de automatische verwerking van LOFAR-gegevens geproduceerd door grote, langlopende projecten. Deze software combineert al ons eerdere werk en maakt het eenvoudiger om grote verwerkingspijplijnen in te zetten op een grootschalige gedistribueerde infrastructuur. Het doel is om complexe LOFAR-pijpleidingen eenvoudig te visualiseren en te automatiseren. De verwerking van de 3000+ LOFAR twee-meter Sky Survey is geautomatiseerd met onze software, waaruit blijkt dat het complexe verwerking, parallellisatie, gegevensarchivering en externe databasetoegang aankan. Onze tool is gebouwd om substantiële, petabyte-grootte, LOFAR-projecten mogelijk te maken.

In Hoofdstuk Zes demonstreren we een methode voor het maken van een compleet schaalbaarheidsmodel voor complexe pijpleidingen. Met de LOFAR-prefactor-pijplijn omvatten we meer dan een orde van grootte in verwerkingsparameters om te begrijpen hoe onze software presteert met toenemende gegevensgroottes. De geleerde lessen helpen onze huidige gegevensverwerking te optimaliseren, de tijd te voorspellen die toekomstige taken in zullen beslag neme nen inzicht te krijgen in de prestaties van grote gegevenssets.

Hoofdstuk Zeven beschrijft een methode om ervoor te zorgen dat toekomstige softwarebeelden vóór de implementatie worden getest tegen productiepijplijnen. We automatiseren dit met behulp van het workflow-orkestratiesysteem

beschreven in Hoofdstuk vijf. In het laatste hoofdstuk beantwoorden we de onderzoeksvragen in onze inleiding en bespreken we de beperkingen van ons werk. Eindelijk, sluiten we af met een bespreking van toekomstige toepassingen van onze software.