Vast and Fast Data in the era of the large astrophysics and particle physics experiments

Gazagnes, Simon

DOI:
10.33612/diss.179743481

SAMENVATTING

Hier geef ik een samenvatting van het werk beschreven in dit proefschrift. Voor het begrijpen van de belangrijkste resultaten is geen specifieke wetenschappelijke achtergrond vereist.

Sinds het begin van het derde millennium bevinden wij ons in het informatietijdperk, een tijdperk wat gedomineerd wordt door de enorme hoeveelheden aan data en informatie gegenereerd door onze moderne samenlevingen. Dit nieuwe tijdperk heeft verregaande implicaties voor wetenschappelijk onderzoek, zeker gezien de voortgang in dit tijdperk deels afhankelijk is van de volgende generatie aan "cutting-edge" wetenschappelijke experimenten. De meetgegevens die door deze experimenten worden verzameld zijn moeilijk te verwerken door de toename in datacomplexiteit, waarvoor meerder oorzaken ten grondslag liggen. In dit proefschrift hebben we ons voornamelijk gericht op extreem grote datavolumes en de snelheid van de data-acquisitiesystemen. Het verkrijgen, verwerken en analyseren van de informatie uit deze systemen is een van de belangrijkste uitdagingen die de mensheid moeten overwinnen om nieuwe datacentrische ontdekkingen mogelijk te maken.

De "Data Science" gemeenschap is in een ongelooflijk hoog tempo gegroeid om gepaste oplossingen te kunnen bieden voor de datacentrische uitdagingen rondom de nieuwe generatie van (zeer) geavanceerde experimenten. Interdisciplinaire projecten, waarbij wetenschappers uit verschillende vakgebieden zoals wiskunde, informatica, natuurkunde en biologie samen aan een vraagstuk werken, zijn de hedendaagse strategie om het groeiende aantal van datacentrische uitdagingen het hoofd te kunnen bieden. Dit proefschrift is een voorbeeld van een dergelijk onderzoeksproject: het [VF]ast data project. Dit project, gesubsidieerd door het Centre for Data Science and Systems Complexity (DSSC) van de Rijksuniversiteit Groningen, is bedoeld om de nieuwste ontwikkelingen in de Wiskundige Morfologie, een groeiende onderzoeksgebied dat zich

richt op de optimalisatie van beeld- en signaalverwerking, toe te passen op huidige en volgende generatie wetenschappelijke experimenten in astro- en deeltjesfysica. Dit interdisciplinaire Ph.D. project heeft geleid tot vijf wetenschappelijke projecten gericht op verschillende wetenschapsgebieden, zoals natuurkunde, sterrenkunde en informatica.

In **Hoofdstuk 2** en **3** presenteer ik twee studies waar in ik een nieuwe rekentechniek, DISCCOFAN (DIStributed Connected COmponent Filtering and ANalysis), beschrijf. Deze techniek is gebaseerd op recente ontwikkelingen in de Wiskundige Morfologie. Deze tool kan op een efficiënte manier structuren en patronen in twee- en driedimensionale datasets analyseren. In verschillende domeinen, zoals biomedische beeldvorming en de astronomie, bestudeert men objecten, bijvoorbeeld een bloedvat of een melkwegstelsel. De eigenschappen van deze objecten moeten efficiënt geanalyseerd worden. Er is een specifieke klasse van morfologische wiskundetechnieken ontwikkeld om de analyse van deze objecten optimaal te laten verlopen, ook wel componentenbomen genaamd. Componentbomen zijn hiërarchische structuren die de hiërarchische relaties van verbonden regio's (d.w.z. structuren) in de afbeelding vertegenwoordigen. Ze geven dus een andere representatie van de originele dataset, wat een efficiënte analyse van de informatie mogelijk maakt. Aan het begin van mijn doctoraat konden deze technieken alleen toegepast worden op relatief kleine beeldformaten: afbeeldingen met een bestandgrootte van enkele gigapixels (gelijk aan ongeveer een miljard datapunten). Individuele machines konden echter geen afbeeldingen verwerken met honderd of duizend miljard pixels, omdat het manipuleren van deze datasets een enorme hoeveelheid rekenkracht vereist en de machines die wel deze enorme hoeveelheid aan rekenkracht hadden waren zeldzaam. Vanuit deze noodzaak hebben wij DISCCOFAN ontwikkeld, een nieuwe rekentool die gebruik maakt van de componentboomtechniek en parallel computing (het aan elkaar sluiten van individuele machines om één krachtige machine te vormen) om enorme twee- en driedimensionale beelden te verwerken. DISCCOFAN verdeelt het rekenwerk over verschillende onafhankelijke processen die gelijktijdig een reeks taken kunnen uitvoeren. Op de manier is deze veelbelovend techniek instaat om de enorm grote datasets (zowel in volumen als resolutie), geproduceerd door volgende generatie wetenschappelijke experimenten, te verwerken.

In **Hoofdstuk 4** en **5** presenteer ik twee projecten georiënteerd op de astrofysica. Deze project zijn gericht op het bestuderen van de fysische processen gedurende het reïonisatie tijdperk, een faseovergang

die plaatsvond in de eerste miljard jaar van het Universum. Dit kosmische tijdperk heeft grote implicaties voor het hedendaagse heelal, omdat er in deze periode de eerste generatie astronomische objecten, zoals sterren en sterrenstelsel, zich vormde. Het waarnemingen van deze objecten is een enorme uitdaging, omdat het miljarden jaren geleden gebeurd is en daarom is er slechts beperkt informatie beschikbaar over dit cruciale tijdperk en de relevante fysieke processen. Wij zullen echter, door middel van verschillende toekomstige telescopen, in staat zijn om de eerste sterren en sterrenstelsels te observeren. Deze waarnemingen zullen ons helpen met het beantwoorden van fundamentele vragen met betrekking tot de vorming van de eerste astronomische objecten en wat hun impact was op de evolutie van ons heelal.

In **Hoofdstuk 4** heb ik de eigenschappen van nabije sterrenstelsels met vergelijkbare kenmerken als die van de eerste populatie van sterrenstelsels geanalyseerd. Het doel was om een consistent theoretisch raamwerk te ontwikkelen en om deze vervolgens te gebruiken om de toekomstige waarnemingen van deze populatie van sterrenstelsels te simuleren en interpreteren. Deze analyse leverde waardevolle inzichten op over hoe deze objecten het heelal gedurende reïonisatie kunnen hebben beïnvloed. Wij gaan de resultaten van onze analyse verifiëren door ze te vergelijken met de waarnemingen van de James Webb ruimtetelescoop en de nieuwe extreem grote telescopen op de grond, zoals de E-ELT, GMT en TMT.

Hoofdstuk 5 bouwt voort op een andere observatietechniek om de fysieke processen gedurende het reïonisatie tijdperk te onderzoeken, namelijk de observatie van de '21 cm lijn'. In plaats van rechtstreeks observatie van astronomische objecten biedt de 21 cm lijn een unieke manier om het heelal gedurende de eerste miljard jaar te bestuderen, want het weergeeft de evolutie van het gas tussen sterren en sterrenstelsels. De vorming van de eerste astronomische objecten hadden invloed op de morfologie en topologie van deze gasgebieden en hun eigenschappen kunnen bestudeerd worden gebruikmakende van 21 cm lijn observaties. Ik heb de morfologie van het gas in deze tijdperken bestudeerd met behulp van de methode die in de eerste twee hoofdstukken heb gepresenteerd. Deze simulaties bevestigde dat deze methode waardevolle inzichten zou moeten opleveren over de fysische processen die de evolutie van het heelal in dit tijdperk beheerste. Dit werk is met name relevant in de context van de Square Kilometre Array, een toekomstige radiotelescoop die revolutionaire waarnemingen van de 21 cm lijn zal gaan doen.

Hoofdstuk 6 beschrijft het laatste project van dit proefschrift. Dit project was gericht op experimentele deeltjesfysica met het ontwerpen

van een efficiënt algoritme voor baanreconstructie van elementaire deeltjes in toekomstige deeltjesversnellerexperimenten in het specifiek. Deze versnellers gebruiken hoogenergetische botsingen van fundamentele deeltjes (bijvoorbeeld protonen, neutronen en elektronen) om de fundamentele bouwstenen van materie en de bijbehorende kleine schaal fysica te onderzoeken. Voor het detecteren van zeer zeldzame deeltjes moeten onderzoekers de meest relevante botsingen, interacties en banen van de deeltjes kunnen identificeren en selecteren. Hiervoor moeten de versnellerexperimenten de data kunnen verwerken met extreem hoge snelheid, het gaat hier bijvoorbeeld om een miljard miljard deeltjesbotsingen per seconde. De hoeveelheid meetgegevens die door deze experimenten worden geproduceerd is echter te groot om opgeslagen te worden, dus realtime gegevensverwerkingssystemen zijn essentieel, deze worden ook wel "triggersystemen"genoemd. Deze realtime systemen maken vaak gebruik van algoritme voor de reconstructie en identificatie van de banen van de gecreëerde deeltjes. De reconstructie en identificatie wordt gedaan op basis van de baan die de deeltjes afleggen in de detectoren, daarnaast stelt ons in staat om hun belangrijkste eigenschappen te bepalen. In dit laatste hoofdstuk hebben we een "fast track reconstruction" algoritme ontworpen, welke gebruikt gaat worden in een toekomstige deeltjesbotsingsexperiment met zeer hoge interactiesnelheden. Wij hebben aangetoond dat deze methode veelbelovend is voor het "on-the-fly" reconstrueren van de deeltjes hun banen.

In dit interdisciplinaire proefschrift werden nieuwe computermethoden gepresenteerd voor het adresseren van de data-uitdagingen rondom de nieuwe generatie van wetenschappelijke experimenten in astro- en deeltjesfysica. Nog belangrijker, het biedt een startpunt voor toekomstige onderzoek en het gepresenteerde werk kan in verschillende richtingen uitgebreid worden. Ten slotte benadrukt dit proefschrift dat de huidige en toekomstige vooruitgang in data-analysetechnieken belangrijke datacentrische ontdekkingen mogelijk zullen maken. Zij zullen het voor onderzoekers mogelijk maken om fundamentele vragen, in verschillende wetenschappelijke gebieden, te kunnen beantwoorden. Hoewel er nog steeds cruciale uitdagingen te overwinnen zijn, is er een mooie toekomst voor de wetenschap die voortvloeit uit de volgende generatie van *grote* en *snelle* wetenschappelijke experimenten.