English Summary

When we look up at the night sky, we observe the Universe through visible light. We can see stars, planets, nebulae, and maybe even other galaxies. The earliest astronomers studied the Universe through their eyes, eventually aided by mirrors and lenses in the form of telescopes and binoculars. In the 18th and 19th centuries, physicists have unraveled the breadth of the electromagnetic spectrum. Aside from visible light, astronomers were now able to study the Universe using the infrared, ultra-violet, microwave, and radio parts of the spectrum. Light from each part of the spectrum carries different information with it and can be used to study specific physical phenomena. The drawback to this breadth of information is that every part of the spectrum needs its own dedicated detectors, and often, entire telescopes. Nevertheless, each part of the spectrum reveals a part of the Universe hidden to us in all other wavelengths.

Radio Astronomy was born in the 1930s with Karl Jansky's experiments with a directed 30-meter radio antenna. With it, he was able to detect thunderstorms, the Sun's magnetosphere, but also a strange unknown source at the center of our galaxies. These experiments proved that low-frequency radio waves could be used to study the distant Universe. In the 1940s, radar receivers were developed as a matter of national security during the Second World War. They are considered the main reason for the victory in the Battle of Britain over the Luftwaffe. After the conclusion of the war, some of the radar hardware was turned to the skies.

While the new antennas were much more sensitive to radio frequencies, radio telescopes have a fundamental limitation in terms of resolution. Because light behaves as a wave, especially at long wavelengths, it is bound by the diffraction limit. The diffraction limit links the smallest object that could be spatially resolved by a telescope with the wavelength of light and the telescope's diameter. If we take a 100-m radio telescope, it will have the resolving accuracy equivalent to a visible telescope with a mirror with a 0.5 mm diameter. This shouldn't be too shocking: radio waves

are more than 10 million times longer than the light we perceive, so naturally, our telescopes need to be 10 million times larger.

To match the resolution of the Hubble space telescope, the mirror of a radio telescope at 150 MHz would need a diameter of ∼1000km. While engineers are still working on making radio dishes of this size, astronomers have turned to computers to increase the angular resolution of radio telescopes. Using the wave properties of light, radio astronomers can synchronize the timing of the data observed at multiple antennas, separated by tens, hundreds, or thousands of kilometers. With significant processing, data from these distant antennas can be combined together to make an image with a resolution equivalent to a radio telescope with a dish diameter of hundreds of kilometers. The downside of this combination of antennas is the computational requirements of making such scientific images.

The LOFAR (LOw-Frequency-ARray) radio telescope is a Dutch low-frequency telescope that consists of more than 5000 antennas in the Netherlands, with thousands more grouped in stations across Europe. LOFAR observes the radio sky at the lowest frequencies visible from Earth: from 10 MHz to 240 MHz. It is designed to collect data for multiple science cases. Scientists can use LOFAR data to study phenomena such as supernova remnants and pulsars in our galaxy, solar system magnetic fields, merging galaxy clusters, black holes in the centre of distant galaxies, and even the epoch when the first stars in our Universe formed. To serve such diverse science cases, LOFAR data is stored at high time and frequency resolutions. This leads to considerable data sizes. Each 8-hour observation can easily fill eight 2-TB hard drives. Creating an image from this data is just possible on one's personal computer and large all-sky surveys consist of thousands of observations which cannot be processed on a single computer or even a small cluster of computers. Even so, each run can take several days, a latency not feasible for projects producing several thousand data sets within a five-year project.

To make it possible to process thousands of multi-petabyte observations, we take advantage of supercomputers for parallel processing of radio astronomy data. Data parallelism means that all of our data can be split into many different pieces, each of which can be processed independently. With a sufficient number of computers, the processing time drops by a factor of 10 or more. Accelerating the first few processing steps delivers a second advantage: The initial stages take high-resolution data and average it down by a factor of up to 64. Suddenly each observation can comfortably sit on your desktop, laptop, and even on a micro-SD card! They can also be transported between universities in seconds rather than hours.

This work is focused on how to enable scientists to easily and quickly process LOFAR data, making it easier to use large data sets to conduct scientific studies. We use open source libraries, a high throughput processing infrastructure, and the properties of radio astronomy data processing to make large scientific surveys with LOFAR possible. Our work is not only useful for scientists, though. We build and present tools that enables us to study the software performance of complex scientific pipelines, and use these tools to recommend more efficient processing strategies. Our results are also applicable to future processing efforts, both for LOFAR data and for upcoming radio telescopes. Ultimately, the lessons learned from the extensive LOFAR surveys can be used for future astronomical projects tasked with large data sizes.

In Chapter One, we give an overview of the history of scientific computing, radio interferometry, and the processing challenges for LOFAR data. In Chapter Two, we present a platform for large scale distributed LOFAR processing. This platform is built for advanced LOFAR users who wish to parallelize their processing on a shared system distributed across multiple computational nodes and clusters. We present the software packages used and their interactions and discuss the necessity of such a platform for current and future large scale astronomical studies.

In Chapter Three, we introduce a framework for launching, tracking, and parallelizing processing jobs for the LOFAR telescope. Our results represent the first time that LOFAR data were processed in bulk, on a High Throughput Infrastructure. We show the applicability of this computing paradigm on the initial steps of a single processing pipeline. We suggest how other pipelines and telescopes can use this framework to process their data. Finally, we show a significant speed-up in data processing compared to data processing in Leiden: up to 35 times faster.

In Chapter Four, we tackle the challenge of collecting performance data from distributed runs of a complex pipeline. Building on our work parallelizing processing pipelines, we build software to track the performance of each processing step. The data that we track is collected at a central server and can be analysed in real-time or studied after the processing run. We run tests on four different systems and find that the software compilation method doesn't degrade the run time performance of the processing. We also gain insights into low-level performance for our slowest steps. Our monitoring software provides a rich data set for scientific software developers to gain insights into the real time performance of the data reduction pipelines.

In Chapter Five, we introduce the first workflow orchestration software for

complex LOFAR pipelines. This software is suited best for the automatic processing of LOFAR data produced by large, long-running projects. This software combines all of our previous work and makes it easier to deploy large processing pipelines on a large scale distributed infrastructure. We have built it to visualise and automate complex LOFAR pipelines easily. The processing of the 3000+ LOFAR Two-Metre Sky Survey is automated with our software, showing its ability to handle complex processing, parallelization, data archival, and remote database access. Our tool is built to make substantial, petabyte-scale, LOFAR projects feasible.

In Chapter Six, we demonstrate a method for creating a complete scalability model for complex pipelines. Using the LOFAR prefactor pipeline, we span more than an order of magnitude in processing parameters to understand how our software performs with increasing data sizes. The lessons learned will help optimize our current data processing, predict the time taken by future jobs, and understand the performance of our processing for large data sets.

Chapter Seven describes a method to ensure future software images are tested against production pipelines before deployment. We automate this using the workflow orchestration system described in Chapter five. In the final chapter, we answer the research questions posed in our introduction and discuss the limitations of our work and conclude with a discussion of future applications of our software.