Grant application National Roadmap for Large-Scale Research Infrastructure 2019-2020

FuSE

Approved!

Fundamental Sciences E-infrastructure

the Joint Data Processing Facility for
the KM3NeT, LHC, and SKA Roadmap Infrastructures
built on the Dutch National e-Infrastructure coordinated by SURF

Table of Contents

1. General Information 1
1.1 Application title 1
1.2 Scientific summary 1
1.3 Lay summary 2
1.4 Key words 2
1.5 Scientific domain 2
1.6 Main field of research 2

2. Requested Infrastructure 3
2.1 Science and excellence case 3
2.1.1 Scientific importance 3
The Large Hadron Collider LHC 5
ATLAS 5
LHCb 7
Alice 8
LHC Computing Requirements 9
The KM3NeT Deep-Sea Neutrino Detector 10
KM3NeT Compute Requirements 12
The Square Kilometre Array 13
Key Researchers 19
2.1.2 Embedment of the investment 20
2.2 Strategic case and innovation 21
2.2.1 Importance for Dutch science and international positioning and appeal 21
2.2.2 Importance for society and industry and the connection with societal developments 23
2.3 Management case 25
2.3.1 Organisation and governance 25
Procurement and Intellectual Property Strategy 25
Key Performance Indicators and Reporting 26
2.3.2 Accessibility 26
2.3.3 IT infrastructure 27
2.4 Technical and business case 28
2.4.1 Technical feasibility 28
Infrastructure Requirements and Evolution for Computing 28
Requirements for Use of the Dutch National e-Infrastructure DNI 29
Technical Challenges in the Data Processing Pipelines 32
Specific Technical Challenges for Science Processing 33
WP 1: Algorithms - Addressing Compute Challenges through Algorithm Improvement 36
WP 2: Access - Realisation of the Infrastructure and Access Mechanisms 37
2.4.2 Risk analysis 39
2.4.3 Financial feasibility 39
Investments in ICT services: 'hardware' 40
'Peopleware': Data and Computing Engineering for Algorithms and Access 42
Investment Summary 42
2.5 Literature references 42
2.6 Other relevant information 44
Key Researcher Publications 44

3. Declaration and signature (by coordinating applicant)

1. General Information

1.1 Application title

FuSE: **Fu**ndamental **S**ciences **E**-infrastructure

1.2 Scientific summary

This 'FuSE: Fundamental Sciences E-infrastructure' proposal brings together Nikhef, the Dutch institute for subatomic physics and ASTRON, the Netherlands Institute for Radio-Astronomy, with SURF, the national e-Infrastructure provider, to build and operate a nationwide e-Infrastructure. This e-Infrastructure will serve the most data-intensive and demanding Research Infrastructures on the National Roadmap: the LHC experiments ATLAS, LHCb, and ALICE at CERN (high-energy physics), the Square Kilometre Array (SKA, radio astronomy) and KM3NeT (neutrino astrophysics) and will thereby strengthen the already unique position of The Netherlands in providing joint e-Infrastructure facilities.

In the 2020's era of Exabyte data rates, the development of this e-Infrastructure will ensure that the national computing facilities are available and affordable. Without the work of this proposal, the Dutch computing costs and demands would be much higher, or conversely, this front-line computing and research capability would be unavailable to the community.

Nikhef and ASTRON each fulfil national leadership and coordination roles in these global scientific facilities. These roles are the natural consequence of the Netherlands' strategic investment in these facilities, and as a result, these are part of the Dutch National Roadmap. The science cases for the facilities are at the heart of the strategic scientific agendas of both institutes.

Towards the end of the five-year term (2021-2025) of this proposal, all three infrastructures will be acquiring data at the Exabyte scale. Large-scale computing infrastructures are needed to ensure Dutch researchers have access to the resources necessary to properly exploit the nation's major investments in these global endeavours. The similarities between the computing infrastructure requirements coupled with the cost-benefit of collaboration have led to this proposal. Embedding the proposed joint data-processing facility in the Dutch National e-Infrastructure (DNI) is expected to create a positive knock-on effect for other infrastructures on the Roadmap and for Dutch and international science in general.

The science cases of the three global research infrastructures drive the content and extent of this proposal. The LHC science cases follow from the LHC Upgrade proposal (2013) which funded construction of new detectors as well as computing for the period 2014 – 2019. The upgraded detectors will be installed in the coming two years and the LHC will start running again in 2021. The proposed facility is necessary to analyse these data and extract the scientific results. For KM3NeT the science cases are equally compelling: it is the only place in Europe where neutrino oscillations and cosmic neutrinos can be studied. The science cases of interest to the Dutch community for the SKA are among the highest-priority science projects identified for this global telescope. Whilst it will take until 2026-2027 for the full science capability of SKA to ramp up to full capacity we will be building on leading Dutch expertise in LOFAR, a critical SKA pathfinder telescope, to blaze the way to significant leadership and impact from the new SKA. Given the investments that the Netherlands has made over the past 10-15 years, it is a matter of national pride that the two highest rated key science projects for SKA are the study of the Epoch of Reionisation (mapping the evolution of the first stars and galaxies) and timing pulsars to test extreme physics (of matter and gravity).

Furthermore, we aim to deliver an e-Infrastructure that will be ready for the emerging field of multi-messenger physics, exploiting data from a number of detectors and telescopes to enable fundamental discoveries. In the longer term, this will include yet another infrastructure on the National Roadmap: the 3rd generation gravitational waves detector, Einstein Telescope.

The proposal covers the period 2021 - 2025. The total budget is M€ 28,8 (*its composition is not disclosed in this public version of the proposal*).

This unique proposal combines the interests of three Research Infrastructures on the National Roadmap and strives towards an even stronger position for the Netherlands in these experiments by enabling an excellent data processing and analysis environment in full alignment with the Dutch national e-Infrastructure.

1.3 Lay summary

Along with a large number of other nations, the Netherlands invests in three huge experiments to probe the nature of the Universe and the fundamental physics that governs it. These facilities are the Large Hadron Collider (LHC at CERN), KM3NeT and the Square Kilometre Array (SKA). As well as testing fundamental laws of nature, these experiments share one big challenge – they produce enormous amounts of complex data that have to be analysed and accurately interpreted. This funding grant will build a shared computing platform and develop data science expertise in the country. This will ensure that all researchers in the Netherlands can fully exploit the potential of these global experiments and make the new discoveries that are only enabled by the volume of data that the next generation big science facilities provide.

1.4 Key words

Astronomy, Particle Physics, e-Infrastructure, data processing, ICT

1.5 Scientific domain

☒ Natural sciences and engineering sciences (*100%)*

1.6 Main field of research

Main disciplines	
Code:	Field of research:
12.10.00	Subatomic physics
17.90.00	Astronomy, astrophysics, other

2. Requested Infrastructure

2.1 Science and excellence case

2.1.1 Scientific importance

The purpose of the proposed infrastructure is to enable the harvest of the scientific results from the data generated by the Roadmap infrastructures for the Large Hadron Collider (CERN), SKA, and KM3NeT. The urgency of this e-Infrastructure is difficult to overstate; without it, Dutch scientists will be disadvantaged internationally, and scientific results and breakthroughs will be delayed. The goal for the e-Infrastructure is to enable all their scientific goals to be realized; in other words, that computation and computational / storage resources will be sufficient to allow (Dutch) scientists to exploit the full potential of the infrastructures, and make the data accessible to the largest possible group of scientists. Excellent computing facilities also attract new scientific talent.

Together these innovative infrastructures give us complementary information to answer some of the most fundamental questions about the nature of the Universe. For example, most of the energy released when a massive star collapses in a supernova is released in the form of neutrinos, which can be detected by KM3NeT. Such an explosion is often associated with the birth of a pulsar, whose radio pulses can be detected with SKA. Pulsars are the densest macroscopic objects known, and they test our understanding of particle physics in a regime that complements the insights we continue to gain from the LHC. This is but one example of how these cutting-edge facilities will work together to present us with the puzzle pieces we need to come to a deeper understanding of the fundamental physical laws that describe the Universe.

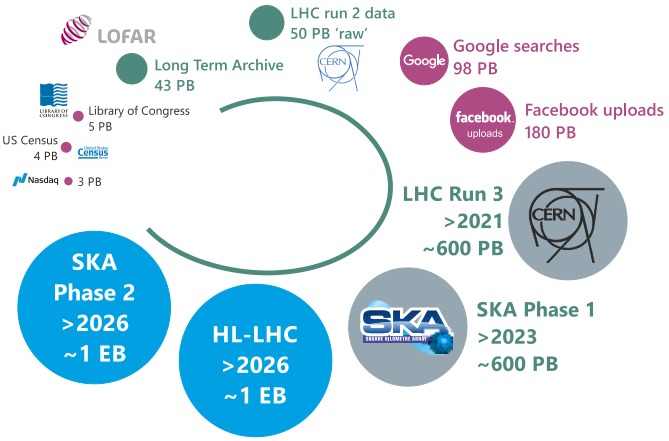


Figure 1: Data volumes for SKA and the LHC for the 2020's, compared to typical data volumes of other applications.

This proposal also foresees scientific and technical personnel connected to the e-Infrastructure. Computing "at scale" requires investment in software infrastructure. The LHC experiments made this investment over the last fifteen years and must continue to invest in this in order to efficiently adapt to the continuous evolution of computing hardware; KM3NeT and SKA are starting this path now, and will shorten it considerably by the proposed collaboration with the LHC infrastructure. All groups, as well as scientists associated with the Einstein Telescope effort, are continually improving algorithms for data extraction: such efforts relate both to international competitiveness (reaching results more quickly) as well as to managing costs of the infrastructure. Efficient scientific software is a prerequisite when computational scales are measured in hundreds of petabytes and hundreds of millions of core-hours per year.

Figure 2: The Large Hadron Collider

In the rest of this section, we recap the science cases of the targeted Roadmap infrastructures, highlighting any new developments since the last Roadmap proposal for the infrastructure. We present the translation of the scientific goals into computing needs, and at the end of the section we present a summary of the required e-Infrastructure. Technical innovation, and challenges associated with realisation of, and use of, this e-Infrastructure will be presented in section 2.4 (Technical case).

The past century: the Standard Model of elementary particles

Throughout the 20th century immense progress has been made in unravelling and understanding the structure of elementary particles and fields: from the chemical elements to three families of quarks and leptons and their antiparticles; from the classical theory of electromagnetism to relativistic quantum field theories culminating in the Standard Model of the electroweak and strong interaction. Throughout, accelerator-based experiments have played a decisive role as witnessed e.g. by: Thomson's discovery of the electron (1897) using cathode rays; the discovery of quarks (1974) using the Brookhaven AGS and SLAC Spear synchrotrons, the discovery of the *W* and *Z*-bosons (1983) at the CERN SppS; the discovery of the top-quark (1995) and the τ-neutrino (2000) at the Fermi Laboratory's Tevatron; and the recent discovery of the Higgs particle (2012) at CERN's Large Hadron Collider. The Standard Model not only very successfully describes a plethora of high-precision data from particle-physics experiments all around the world, but also allows a qualitative and quantitative description of the evolution of our Universe, from a minute fraction of a second after the Big Bang to today, about 13.8 billion years later. Thereby, the Standard Model links apparently completely unrelated observations like the existence of antimatter, the abundance of the natural elements in our Universe as observed by astronomers, and the number of particle families (i.e. the number of light neutrino species) as measured in particle accelerator experiments. The Standard Model also connects the science of the infinitely large (astronomy) to the science of the infinitesimally small (particle physics).

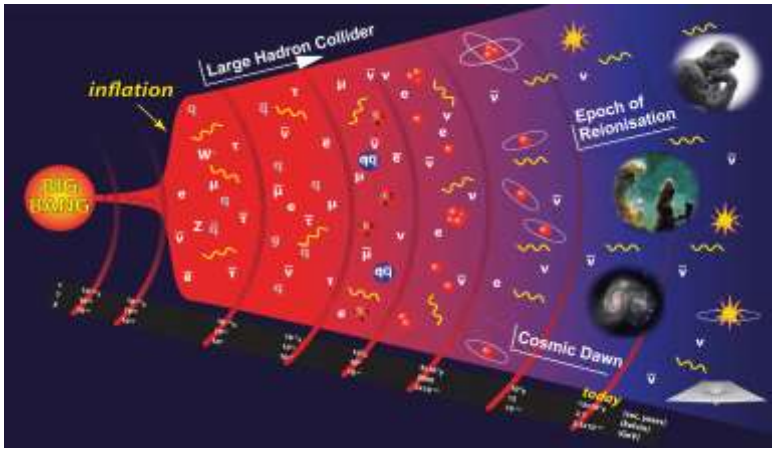


Figure 3: evolution of the Universe. The different infrastructures, LHC, KM3NeT, and SKA, together probe different aspects of the Universe.

The future: beyond the Standard Model of elementary particles

Despite the many and astonishing successes and high precision of the Standard Model, experiments, observations and theoretical speculations have revealed a Universe far stranger and even more wonderful than predicted by the Standard Model. A Universe, filled with dark matter and dark energy, where ordinary matter (quarks and leptons) constitutes only a tiny 5% fraction. A Universe, in which theorists, in their attempts to reconcile the theory of gravitation with the principles of quantum mechanics, predict the existence of curled up extra spatial dimensions invisible in our everyday world. A Universe, in which neutrinos oscillate, *i.e.* change flavour. And perhaps less revolutionary, even within the context of the Standard Model, several issues require further experimental clarification. The Standard Model fails miserably, for example, in explaining why today's Universe contains no antimatter. We've found the Higgs, but does it indeed interact with matter particles (the quarks and leptons) as predicted, explaining the origin of mass? Does the Higgs interact with the massive W- and Z-bosons as predicted by the mechanism of electroweak symmetry breaking as implemented in the Standard Model? Furthermore, the Standard Model predicts the existence of a new state of matter at high temperature and density in which quarks and gluons are no longer confined inside hadrons like protons and neutrons: the quark-gluon plasma. This quark-gluon plasma supposedly played an important role in the very early Universe and requires further investigation. Despite the often-impressive theoretical ingenuity of many models and despite the quantitative accuracy of some of the predictions, there are still many things we do not know about Nature at its most elementary level.

The Large Hadron Collider LHC

CERN's LHC project is expected to yield the experimental evidence needed to answer (some of) the above-mentioned fundamental questions in the next decade. The LHC collisions are detected by two general-purpose experiments: ATLAS and CMS, and two specialized experiments: LHCb and Alice. The LHC started delivering proton-proton collisions in 2010, and has since then undergone an upgrade (2013-2014) increasing collision energy and beam intensity. Presently, another long period (2018-2019) of work on the accelerator and the detectors is underway, in which experiments are upgraded, and the LHC itself is prepared to deliver a doubled collision rate. A further major upgrade is foreseen for 2023–2025, in which the collision rate will triple again. Experiments plan major upgrades, in particular of their tracking detectors, to cope with the increased intensity (i.e. higher radiation loads) coming from the accelerator upgrade and/or to take advantage of the availability of new detector and computing/electronics technology, further boosting their discovery potential.

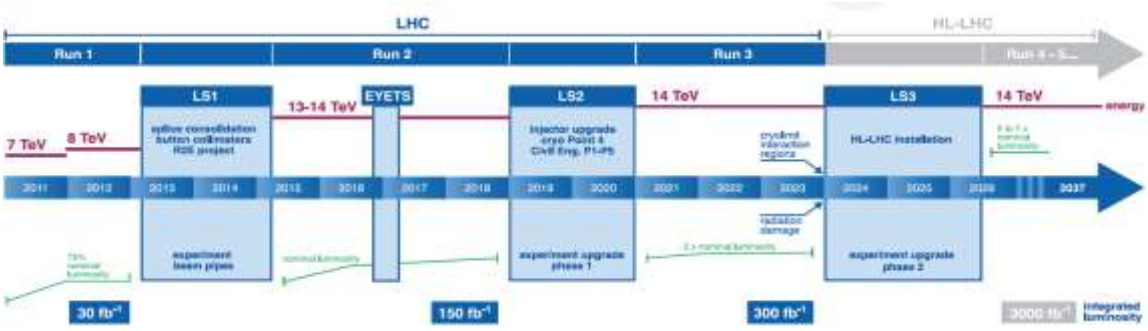


Figure 4: Overview of the LHC schedule. The proposal covers the period of Run 3 and analysis of data taken.

ATLAS

ATLAS is the world's largest particle detector, 25 metres high, 44 metres long, and with 90 million electronic channels. Figure 5 shows an example LHC proton-proton collision observed by the ATLAS detector in Run 2. The Nikhef ATLAS group consists of physicists of the Radboud University Nijmegen, University of Amsterdam and NWO-I.

As a general-purpose experiment, ATLAS has a rich physics program that aims to clarify several of the big open questions regarding our current understanding of Nature:

1. *How can the electroweak force manifest itself as two distinct forces (electromagnetism and the weak nuclear force) at low energy?*
2. *What is the origin of mass of elementary particles?*
3. *Are there new symmetries that regulate the instability of the current theory at cosmological energy scales?*
4. *What is the particle nature of dark matter, for which cosmological evidence exists?*

For questions 1 and 2 our current best model of Nature, the Standard Model, postulates that a special scalar particle – the Higgs boson – has a pivotal role. The recent observation of such a particle in ATLAS (and CMS) in 2012 is the start of a new era in particle physics where experimental precision measurements will guide further understanding to these fundamental concepts in Nature.

Answers to questions 3 and 4 imply the existence of so far unobserved elementary particles that could be produced and observed at the LHC. The unprecedented energy of LHC

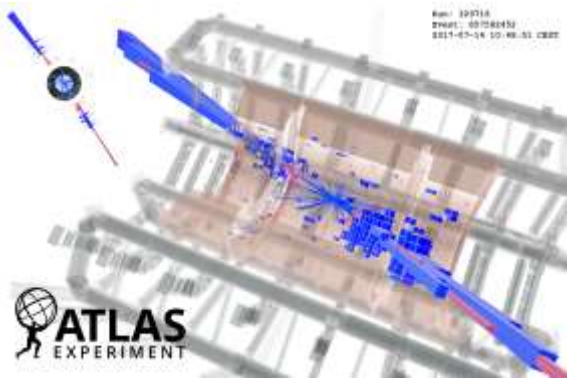


Figure 5: Visualization of a proton-proton collision recorded in the ATLAS detector.

collisions, as well as the increasing data volume, allows the discovery of the production of very heavy – or very weakly interacting – particles through exceedingly rare production processes previously not accessible.

After a 2-year technical stop (2013-2014), in which the LHC magnets and several ATLAS detector components were upgraded, ATLAS has recorded much more data, at a center-of-mass energy of 13 TeV, in Run 2 (2015-2018). To date, ATLAS published 847 scientific papers and 955 conference contributions. In Run 2, the Nikhef ATLAS group has played a major role in measurements of properties of the newly discovered Higgs particle and searches for phenomena beyond the Standard Model, with a focus on the search for additional Higgs particles and supersymmetry and dark matter. Figure 6 highlights two flagship results from Run 2: the measurement of Higgs couplings to fermions and bosons [1], which relate to the origin of mass of particles, and a summary result of the search for supersymmetric particles [2].

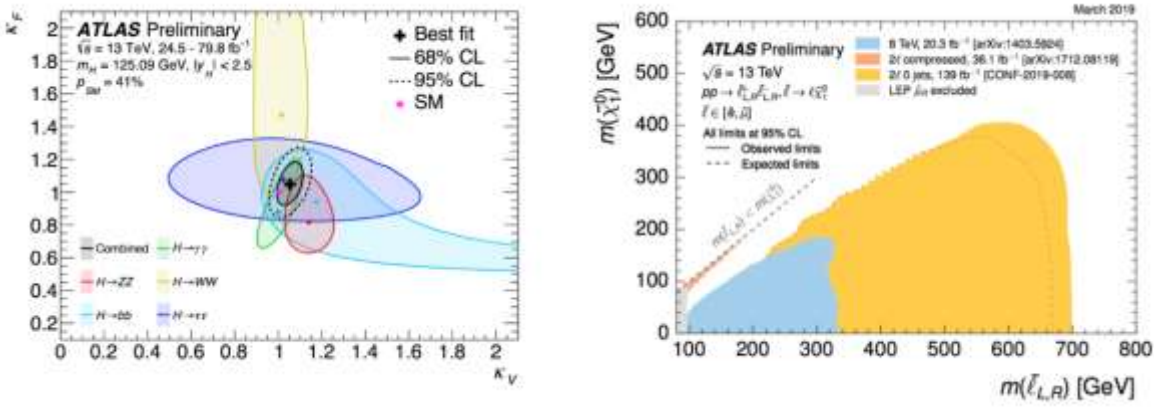


Figure 6: Left: ATLAS Run 2 measurement of coupling strength of the Higgs boson to fermions (k_F) and bosons (k_V) relative to their theoretical expectation in the Standard Model. Right: Run 2 limits obtained on the masses of two classes supersymmetric particles: neutralinos ($\tilde{\chi}$) and sleptons ($\tilde{l}$).

To indisputably establish the Higgs boson role in electroweak symmetry breaking and the Higgs boson as origin of the mass of elementary particles, a much larger data sample is needed, motivating the upgrades for Runs 3 and 4. This larger data sample will allow significant improvements in the determination of Higgs couplings, but will also allow extraction of more information from the data: the kinematic distributions of Higgs boson decays are sensitive probes of the structure of Higgs-particle interactions, and these can be measured with increasing precision as data volumes increase. Any significant observed deviation, even if small, would indicate the

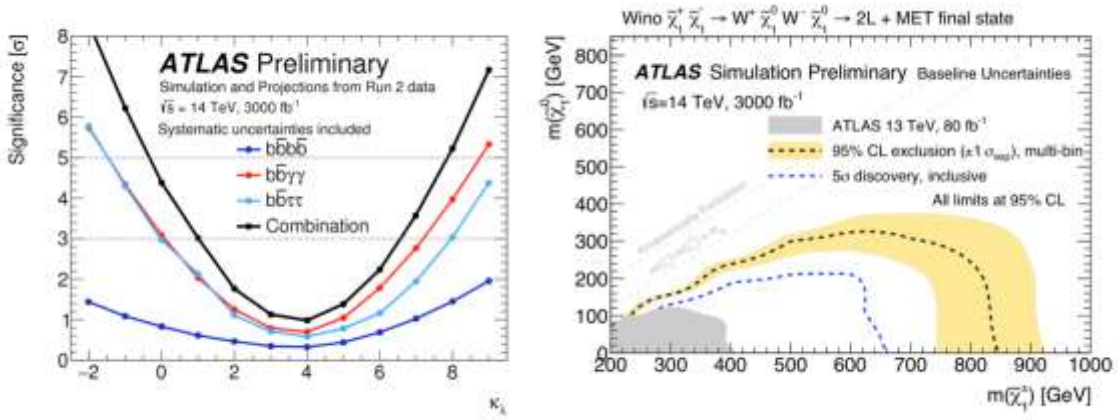


Figure 7: Left: Projected discovery potential (statistical significance) of Higgs boson self-couplings with the full HL-LHC dataset, as function of the assumed Higgs self-coupling strength relative to theory. Right: Projected reach for 5σ discovery and 95% exclusion of the observation of a specific type of supersymmetric particles (neutralino pair production) as a function of the hypothetical particle masses.

existence of new fundamental particles or symmetries. In addition, extremely rare Higgs decay channels with further sensitivity to new physics will become accessible. In parallel with the Higgs physics program, searches for production of hypothetical new particles will continue in Run 3 with focus on rare and difficult-to-identify signatures of such particles, with guidance of global fits to a broad range of (astro-)particle physics and cosmological data to help identify which signatures are most promising.

After a long shutdown, High Luminosity LHC (HL-LHC) operations should start in 2025, with an instantaneous luminosity capable of tripling the detection rates; this Run 4 is planned to operate for ten years. This large data sample will also allow previously impossible crucial tests of the Higgs mechanism such as observation of Higgs self-coupling, which is related to the structure of the cosmological vacuum. This large data sample also allows probing the signatures of new physics, not described by the Standard Model, well into the multi-TeV region. Figure 7 shows a projection for discovery potential of Higgs self-couplings [3] as well as an example projection [4] for the sensitivity a specific class of direct searches for new particles by 2030.

LHCb

The Nikhef institute is one of the founding members of the LHCb collaboration, and one of the largest contributor to the experiment, second only to the CERN group. The LHCb experiment aims to solve the riddle of the origin of the matter - antimatter anti-symmetry in the fundamental laws of physics, as well as to shed light on the reason why nature provides us with three generations of fundamental particles. In addition to these questions, LHCb studies a broad spectrum of fundamental physics that includes research to new tetra-quark and penta-quark states of matter, electroweak precision measurements, searches for exotic and forbidden particle decays, and a multitude of heavy flavour projects utilizing beauty and charm particle decays.

Results from Run 1 and Run 2

In the Standard Model a matter versus anti-matter asymmetry in particle interaction, called CP (Charge-Parity) violation, is implemented via the so-called CKM mechanism ('Cabibbo-Kobayashi-Maskawa', named after the physicists that introduced it). However, this has been proven to be insufficient to explain the dominance of matter over antimatter in our Universe [5].

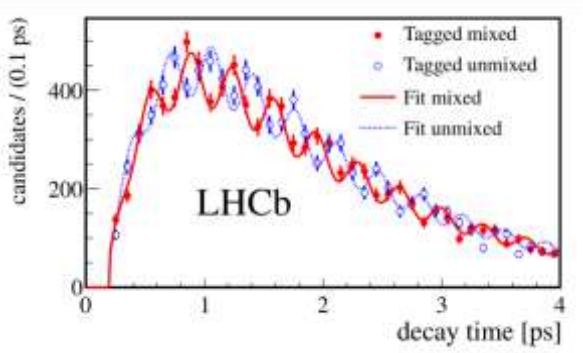


Figure 8: Neutral B_s oscillations as a function of lifetime

The goal of the LHCb experiment is to perform high precision measurements, searching for sources of CP violation originating beyond the Standard Model. The first target of the experiment was the study of neutral B_s mesons. These particles exhibit very rapid transitions (2.5 Terahertz) of particle to antiparticle quantum states and back, so-called B_s oscillations in Fig. 8. CP violations in decays of B_s-meson decays were discovered by LHCb in 2013 [6], and more recently in charm meson decays [7], while the first evidence for CP violation with baryonic particles was obtained in 2017 [8].

One of the most prominent results of the first LHC run was the discovery of the so-called forbidden particle decay $B_s \rightarrow \mu^+\mu^-$. The Standard Model predicts these processes to occur very rarely, about three in a billion. The observed value of the decay rate is in agreement with the Standard Model prediction within the experimental precision, albeit at a slightly lower value. This observation, shown in Fig. 9, has falsified a large class of supersymmetry models.

Since approval of the 2013 LHC proposal, a number of surprising and intriguing observations have been published by the LHCb collaboration. In both decay processes including transitions of b-quarks to c-quarks as well as processes of b-quarks to s-quarks an anomalous pattern has been observed. Although as of yet no single measurement has led to a discovery, the collection of measurements consistently hints at the so-called phenomenon of lepton-flavour non-universality. Confirmation of this phenomenon would imply a ground-breaking discovery and, as such, has the highest priority in the data taking of the upgraded experiment.

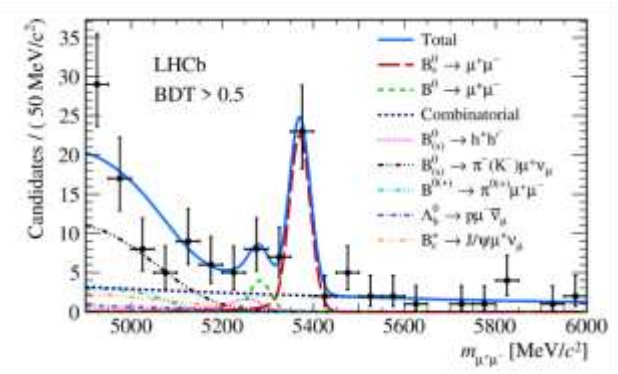


Figure 9: Discovery of the $B_s \to \mu^+\mu^-$ decay

Run 3 and Run 4 goals

The goals for the upgraded LHCb experiment are to discover the existence of particles or fields beyond the Standard Model. Such high-mass particles are expected to quantum mechanically affect known processes and alter their properties, such as the rate of occurrence, from the Standard Model prediction. Deviations could also emerge in the form of anomalous measurements of CP violation, by rare decay rates that differ from their Standard Model expectation, in non-uniform occurrence of decays with different lepton types, or even lepton-flavour violating decays.

Currently the LHCb experiment is undergoing a major upgrade, which will enable an interaction rate five times higher. This will allow for measurements with correspondingly larger samples, increasing their statistical power. Not only are major parts of the detector systems being replaced, at the same time the real-time data processing is being revolutionized. The upgrade includes the transition from a detector using a hardware trigger to a trigger-less readout. This implies that *all* detector data will be processed, in software, on a dedicated compute farm, at the full 30 MHz repetition rate of the LHC. The design of this new trigger system is illustrated in Fig. 10. With this new architecture, the efficiency for event reconstruction and selection increases with approximately a factor two. As a consequence, the overall signal sample sizes available for analysts of the upgraded experiment are expected to improve by a factor of 10 compared to Run 1 and Run 2. This subsequently requires a corresponding increase in offline computing resources.

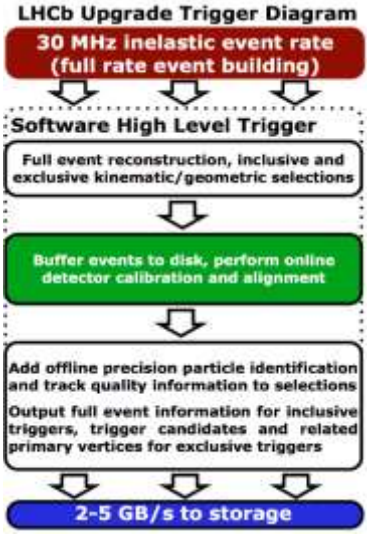


Figure 10: Outline of the trigger of the upgraded LHCb experiment. The information of the entire detector will be digitized at the LHC repetition rate, resulting in a 40 Tb/s data stream. This stream will be processed in real-time by a compute farm, which will reduce the bandwidth required for long-term storage down to approximately 5 GB/s.

Alice

The goal of the ALICE experiment [9] is to study the properties of strongly interacting matter at high temperature and density. Theoretical calculations predict that a Quark Gluon Plasma (QGP) [10] can be produced at the LHC in ALICE, allowing a controlled study of the properties of this state of matter, in which the entire Universe is believed to have existed for a few microseconds after the big bang. The Nikhef group (one of the largest in ALICE) studies the QGP to understand how it emerges from the fundamental theory of Quantum Chromodynamics (QCD), the part of the Standard Model

that describes how the strong interactions work. These studies in ALICE proceed via collisions of lead (Pb) nuclei, of protons, and also between protons and Pb-nuclei, all at the highest energies available in the laboratory.

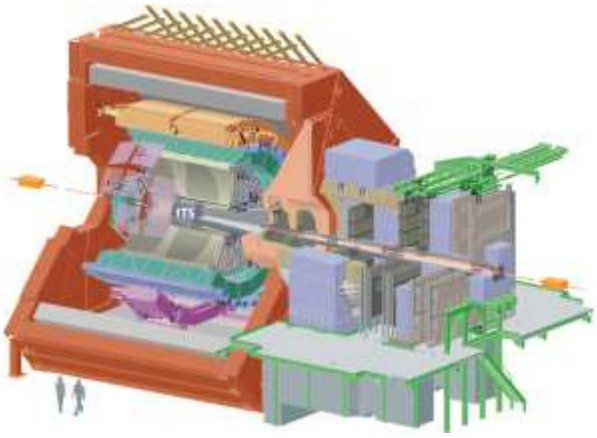


Figure 11: The ALICE detector with the Inner Tracking System (ITS) at the LHC

The ALICE program aims at understanding the properties of the QGP, a challenging endeavour touching diverse fundamental topics such as perturbative QCD, lattice QCD, particle physics, nuclear physics, string theory, thermodynamics and relativistic hydrodynamics.

Currently, the ALICE experiment at the LHC undergoes a major detector upgrade that will make it possible to improve the experimental uncertainties by more than an order of magnitude. The 2013 Roadmap proposal funded the ALICE group at Nikhef to lead the upgrade of the outer layers of the Inner Tracking System (ITS).

For this future program, ALICE has a detailed and ambitious physics program that aims to make a decisive step in unravelling the origin and the nature of the emergent properties of the QGP, both macroscopic properties and microscopic structure simultaneously.

Finally, ALICE plans to exploit the exciting opportunity to observe novel QCD phenomenon. While most of the dynamics of the system is governed by the strong interaction, there are initial state effects related to electromagnetic interactions, which offer fascinating opportunities. For one class of heavy-ion collision, the nucleons that do not participate in the interaction (the 'spectators') create strong magnetic fields of the order of 10^{18} Gauss [11], generating the strongest magnetic field in nature. Also, in these non-central collisions the motion of the participants from each of the nuclei generates a net angular momentum in the QGP. The effect of this magnetic field and the angular momentum of the system should be reflected in the measured momentum distributions of final state particles and the correlations between them. These studies open up the exciting possibility to observe parity violating effects in the strong interaction [12].

LHC Computing Requirements

Across all experiments, we propose to provide, via the DNI, the same fraction (approximately 10% of the global Tier-1 total) of the LHC computing as was approved for the 2013 LHC Roadmap proposal. The facility will require high-performance networks, both externally (between CERN and FuSE/DNI) and internally (between the CPU and fast storage), to handle the increased data volumes.

The computing resources requested for *ATLAS* in this proposal pertain primarily to the analysis of the LHC Run 3 data, and the final analysis of the complete Run 1-3 'nominal luminosity' LHC data. The Run 3 data sample will grow up to 2023, the end of the data taking period. The relative importance of the Tier-1's for ATLAS increases markedly, especially in terms of near-line storage capability to deal with the increased data volumes. After the end of Run 3 (2024-2025) a reprocessing of the full data sample is foreseen, requiring additional resources. In addition to data storage and processing, an increasing volume of simulation samples is required to interpret the data: as the statistical precision of data increases, the importance of understanding systematic uncertainties increases, and growing need for simulation samples is foreseen to facilitate this understanding. It is estimated that generation of simulated event samples will consume about two-thirds of the total CPU resources requested.
In the period 2021-2025 further computing resources will be required for simulation studies needed to commission the data taking for the HL-LHC. The density of particles in HL-LHC collisions will

increase substantially compared to the LHC, as up to 200 proton-proton collisions happen simultaneously, and the analysis and reconstruction of these high-density events requires detailed study and investigation into novel simulation and reconstruction techniques in the years leading up to 2025.

For *LHCb*, the increase in data volumes comes already in 2021. The trigger performs a full reconstruction, mitigating both the output data volume and the amount of offline reconstruction (CPU) needed. Due to the reduction in offline CPU work, simulation of reference data comprises a much larger fraction (90%) of the CPU requirement for LHCb.

The *ALICE* requirements for the number of reference (simulated) events is estimated from the need to determine the reconstruction efficiency of rare signals up to the highest reachable momentum. The momentum reach is in turn related to the number of collected events, so more data means more simulation (in addition to the general need for better statistical precision) – combined, these two lead to a factor 20 increase in the need for reference data compared to Run 2. Most of this increase will be met by using fast or parameterised Monte-Carlo simulations. For the remaining full simulations, it is essential to use a computationally efficient particle transport code without compromising on accuracy.

The KM3NeT Deep-Sea Neutrino Detector

KM3NeT is a large, European research infrastructure that will consist of a network of deep-sea neutrino detectors in the Mediterranean Sea. KM3NeT will facilitate scientific breakthroughs in the following areas:

- In *neutrino physics*: KM3NeT will enable ground breaking measurements of the physics of neutrino oscillations, with as flagship measurement the determination of the neutrino mass hierarchy: the next big open question in neutrino particle physics.
- In n*eutrino astronomy*: KM3NeT will be the most accurate detector, or 'telescope' of its kind. The superior angular resolution will allow for true neutrino astronomy. It will enable the identification of the sources of cosmic neutrinos, the measurement of their energy spectra and the study of the flavour composition of the fluxes.

KM3NeT's neutrino science program is world leading. The KM3NeT facility is the only place in Europe where neutrino oscillations and cosmic neutrinos can be studied. As such, it will provide unique opportunities to European (astroparticle) physicists and astronomers; some of the sensor data is also uniquely interesting to marine biologists.

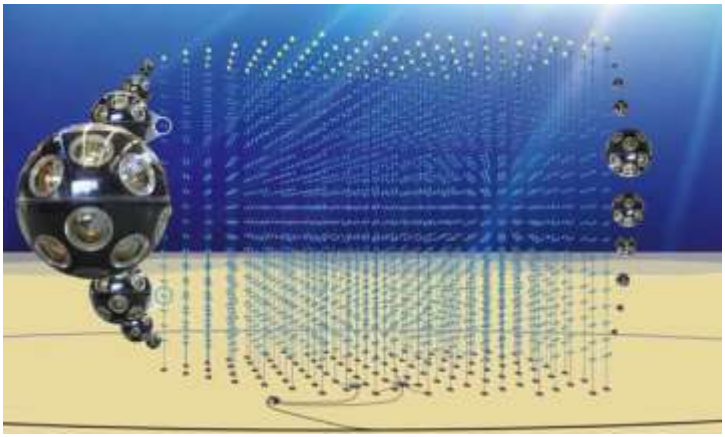


Figure 12: Artist's impression of the KM3NeT neutrino telescope. The instrument will be placed on the bottom of the Mediterranean Sea, at a depth of 3.5 km. It consists of 230 vertical detection lines, each 700 m long and comprising 18 optical modules. Each optical module consists of a pressure resistant glass sphere housing 31 small photo-multiplier tubes (yellowish disk in the picture) that detect the patterns of Cherenkov light created by the interactions of neutrinos (earlier detectors used one large sensor per sphere). The first deployed detection lines have been constructed at the Nikhef institute.

The Physics of Neutrinos

The oscillation of one neutrino flavour to another is a fascinating quantum mechanical phenomenon at the forefront of current particle physics research. The pattern of the mixing of different mass- and flavour states appears to differ significantly from the analogous mixing of quarks (discussed in the LHCb science case), and for a deep fundamental description of Nature we must find out why this is. Oscillations tell us about differences between neutrino masses, but not about their actual values, nor about the ordering ("mass hierarchy"). The neutrino mass hierarchy is a fundamental, but still unknown, parameter of particle physics. Knowledge of its value is essential for the investigation of CP-violation in the neutrino sector.

In 2012, when the first measurement of the neutrino 'mixing angle' was reported, it became clear that the neutrino mass hierarchy can be determined using the neutrinos produced in the Earth's atmosphere by interactions of cosmic rays. This measurement can be performed with a densely instrumented KM3NeT array. A dedicated array, named ORCA, will be constructed at the French site for this. The measurement of the neutrino mass hierarchy with ORCA is one of the main objectives of KM3NeT.

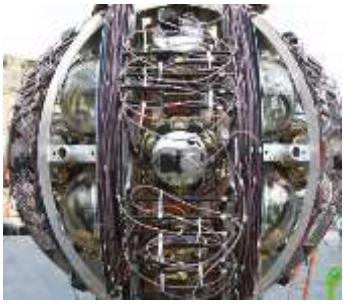


Figure 13: KM3NeT string of modules for ORCA looped on its support structure before deployment in the Mediterranean.

The study with ORCA not only probes the mass hierarchy, but the collected data will also allow measurements of the mixing angles, and the mass differences between the neutrinos, both fundamental parameters of Nature, with a precision exceeding other experiments. ORCA will furthermore allow for the testing of scenarios where neutrinos undergo novel interactions (so called 'non-standard interactions'), and will test the unitarity of the neutrino mixing matrix with a precision beyond that of current experiments. The latter will establish whether neutrinos are disappearing into unobserved states ('sterile neutrinos').

The High-Energy Universe

Since the discovery of cosmic neutrinos by the IceCube experiment [13], a main question in the field of astroparticle physics is the astrophysical origin of these neutrinos. For electron-neutrinos, for example, the superior optical properties of liquid water as a detection medium, combined with reconstruction algorithms devised by the Dutch groups, yield a KM3NeT-accuracy of 1.5^{o} at PeV energies, a factor 10 more accurate than IceCube. The performance has been validated with the running ANTARES detector, where we reach similar resolutions, and have produced the first all-flavour search with the world-best limits for many (Galactic) neutrino sources in the Southern Hemisphere. Combining the breakthrough capability of identification of all three neutrino types with superior pointing accuracy, gives KM3NeT the opportunity to perform genuine neutrino astronomy, and moreover in a region of the sky that will simultaneously be covered by SKA.

Once the origin is established, the cosmic high-energy neutrinos will offer a unique way to study the astrophysics of the source objects. The neutrino signal can then be studied in conjunction with optical, radio, and gamma-ray observations, and even with gravitational wave events. Within these multi-messenger studies, neutrinos are unique as they are a direct probe of hadronic interactions and are therefore a smoking gun for the presence of acceleration of hadrons. In addition, cosmic neutrinos are powerful tools to investigate fundamental particle physics. In general, they offer the unique opportunity to study the properties of neutrinos themselves at energies a factor 10^4 above Earth-bound neutrino beams. Furthermore, extra-galactic neutrinos may take of the order of 10^9 years to reach Earth, making them ideal probes for hypothetical, very rare and subtle effects of new physics, which include extensions of the Standard Model that explain (the smallness of) neutrino masses. A key piece of information is the relative abundance of the three neutrino flavours.

KM3NeT Compute Requirements

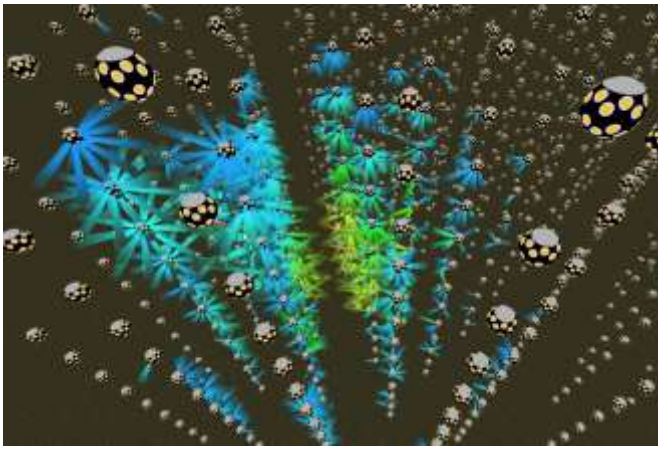


Figure 14: Event display of a simulated electron-neutrino interaction. The colours indicate detected light signal at different times. From patterns like this, the reconstruction algorithms running in the compute centre infer the energy, direction and type of the neutrinos

The computing challenge at the heart of a neutrino telescope consists of inferring the neutrino properties (type, direction, energy) from raw data 'events'. For this, dedicated algorithms have been developed (largely at Nikhef), which can determine the neutrino direction with a resolution better than 0.1^{o} in the case of a muon-type neutrino and 1.5^{o} for the more challenging electron and tau neutrinos. The relation between the photon patterns and the neutrino parameters is highly non-linear and requires iterative methods, combined with detailed modelling of the light propagation from the charged particles to the detector elements. This directly impacts the resolution of the telescope and thus the scientific output. Furthermore, each event must be processed by several algorithms, which each look for the signatures of a specific neutrino type, as it is not a-priori clear which type of neutrino has produced it. This is followed by (machine learning) classification algorithms that separate the various signal sources from background consisting of e.g. cosmic ray muons.

Over the coming years, the KM3NeT neutrino observatory will be constructed. Ultimately, 345 detection lines will be deployed for the detection of cosmic high energy neutrinos and to study the fundamental physics of these elusive particles. Each detection line produces around 40 MB/s, for a total of 15 GB/s for the full detector. Below, we describe the needed data processing and storage for the completed detector. When the detector is partially implemented, the needs will scale accordingly.

In the shore stations (Tier-0), these data rates are reduced by applying filter (trigger) algorithms to select the valuable "physics events" to be stored on disk. Each event contains all the information on all detected photons within a time window of about 10 microseconds; they have an average size of 60 kB. About $5x10^{9}$ events per year will be stored (150 Hz), for a total data volume of 300 TB/year. The data stream is dominated by atmospheric muons traversing the detector, which provide an invaluable signal for calibration and monitoring of the detector.

In addition to the physics data, there are monitoring and calibration data streams, which are used to calibrate and check the performance of the detection. These are large as the detector is in a natural environment and the position of the detector elements must be tracked over time with high granularity.

The physics data is processed at the "Tier-1" centres (such as FuSE) using reconstruction and classification algorithms that compute the neutrino (or muon) properties from the light signals stored in the raw data. The output of this processing stage is data that is suitable for high-level analysis. The data processing can be repeated on the raw data when better algorithms or better detector calibrations become available.

The KM3NeT computing requested is based on reconstruction of the raw data, assuming a second pass with improved algorithms and calibrations (40% of total time). 20% of the time is for reduction of the reconstructed data into a form ready for high-level scientific analysis, and another 20% for processing the calibration data. These processing steps together yield 450 TB/year of derived data, which serve as input for the ("Tier-2") high-level analyses carried out by the KM3NeT scientists.

In addition to processing of the data, simulations are a key ingredient for harvesting the KM3NeT science. The KM3NeT detector is, although much larger, much less complex than those for the LHC, hence the lower CPU request for producing reference data. Event based simulations are required for

the signal and the main background of atmospheric muon events. Finally, a small fraction of the computing and data storage will be used for data-analysis activities and the Dutch neutrino astronomy community. Access to these facilities will give our scientists an important edge in the exploitation of the KM3NeT data, ensuring the continued prominent role of the Netherlands in this international project. The resources required for simulation and analysis form a small percentage (<20%) of the total and will be realized within the allocated resource budgets. Given the relatively modest (compared to SKA and LHC) storage requirements tape archival storage is not requested in this proposal. Given the size of the Netherlands in KM3NeT, we propose a 20% share of the worldwide Tier-1 data processing to be performed in the Netherlands.

The Square Kilometre Array

The SKA is the premier next generation radio telescope developed as a true global collaboration and built in the best radio-environments on Earth. The SKA will provide unprecedented game-changing capabilities to address some of the fundamental questions about the formation and evolution of the Universe, and probe extreme physical processes, which test and challenge our basic laws of physics.

The recent new discoveries in physics, astrophysics, and astrobiology (e.g. discovery of gravitational waves, imaging the shadow of a black hole, discovery of a wealth of extrasolar planets) demonstrate the enormous discovery potential from the synergistic operation of very large facilities. SKA is the next major step forward in sensitivity and capabilities in radio astronomy: some key phenomena can only be studied in this window. For example, atomic neutral hydrogen (hereafter indicated as HI), is the most abundant element in the Universe and represents the key building block of all structures. Its rest-frame wavelength of 21cm (1420 MHz) means that it can be traced up to the initial key phases of the Universe, the so-called Epoch of Reionisation. It can also be used to trace the building of large structures like galaxies and the formation of their central super-massive black holes.

The recent detection of gravitational waves from binary mergers of black holes and neutron stars has opened an exciting window to view our Universe. By combining these gravitational wave observations with their electromagnetic counterparts at radio frequencies we are able to delve into the physical mechanisms powering these extreme events and unravel if they have any association with the enigmatic phenomenon called "Fast Radio Bursts" (FRBs). At the lowest gravitational wave frequencies, merging pairs of supermassive black holes will be detectable with the SKA by monitoring a suite of the most accurate clocks in the Universe – millisecond pulsars – distributed across the sky.

The highlights of the SKA science ambitions are described below by a number of Science Working Groups, which will define Key Science Projects (KSPs) in the coming years. The Dutch community is involved in all of these, with highly-active and leadership roles in a number of them. The foundations of the Dutch community's SKA expertise is derived from innovative radio telescopes like LOFAR at low frequencies (metre wavelengths, < 200 MHz) which is an SKA-Low pathfinder, and the phased-array feed APERTIF system on the Westerbork Synthesis Radio Telescope (WSRT) (at centimetre wavelengths) as an SKA-Mid pathfinder.

Figure 15: Artist's impression of the SKA-Mid dishes, to be built in South Africa

Construction of the SKA will take an estimated six to seven years and is due to start in 2020. Commissioning and early science will commence in 2023-2024 in parallel with the staged delivery of capability. The full SKA science programme is expected to start in 2027-2028. SKA will produce significantly higher data rates than existing radio telescopes, at all stages of processing. This calls for more highly optimised algorithms and procedures that make use of the latest high performance and high throughput computing technology. At the same time, the much-increased sensitivity, which allows the SKA to detect and image fainter and more distant sources of radiation, requires a better

understanding of the instrument and improved constraints on its calibration. It is clear that the tooling that is available today needs to be improved and optimised in order to prepare for the 600 Petabytes per year that will be delivered to the SKA Regional Centres by the SKA each year from 2026-2027. This forms the main purpose of the work programme, described in section 2.4.1.

This proposal covers the preparatory work in the SKA Regional Centres to get ready for the time when data starts to arrive. This includes the commissioning phase (between 2024 and 2027) when the SRCs will be the testing ground for tools and pipelines that produce the Observatory Data Products. It does not include any work directly related to the scientific exploitation phase of the SKA project. In particular, *this proposal does not cover any aspect of the research programmes (e.g. Key Science Projects, PI-led projects) that will be carried out with the SKA.* The SRC is the essential platform and infrastructure that is needed to tame the immense data stream from the SKA.

The volume of SKA data stored and processed in the Netherlands during much of the period covered by this proposal is modest – but is expected to take off rapidly thereafter. The share of the total volume of SKA data and computing that the Netherlands will host will be of the order of 10-15%, depending on the outcome of ongoing negotiations and whether the European SKA members decide to store a full copy of all SKA data.

21-cm Cosmology with the SKA

Most baryonic mass in the early Universe was in the form of diffuse neutral hydrogen and helium created during Big-Bang nucleosynthesis. This gas gave birth to the first stars, which made up the first galaxies and in turn led to formation of the first black holes. The pristine gas was enriched with heavier elements, created in the first generation of stars, and expelled into the surroundings, thereby transforming stellar evolution for the next generations of stars.

The cold intergalactic gas was heated via X-ray emission from stellar remnants, during the 'Cosmic Dawn' (CD), and at some stage ionised by (uv-)emission from the first stars and quasars, during the 'Epoch of Reionization' (EoR). Characterising the 21-cm line of neutral hydrogen [14] both spatially and as a function of cosmic time would provide the most comprehensive view of the infant Universe starting at ~1% of its current ~13.8 billion year age and within volumes far more extensive than those probed by infrared/(sub)millimetre telescopes. Moreover, the original 21-cm spectral line 'redshifts' due to the expansion of the Universe, and the wavelength of the observations is directly related to cosmic time and distance. Hence, consecutive 21-cm signal images can be constructed, and their properties can be studied as a function of cosmic time ('tomography'), allowing the evolution of the infant Universe to be traced.

The 21-cm signal of hydrogen thus harbours the potential to unveil the physical processes of the Cosmic Dawn and Reionization, the underlying dark matter distribution, and much more, promising a transformational view of the evolving Universe during its infancy. This has inspired the construction of a new generation of state-of-the-art radio telescopes: the Low-Frequency Array (LOFAR; [15]) in the Netherlands and Europe, the Murchison Widefield Array (MWA; [16]) in Western Australia and the Precision Array for Probing the Epoch of Reionization (PAPER; [17]) in South Africa, and the design and construction of the next-generation instrument such as HERA and the SKA [18,19].

Figure 16: Artist's impression of an SKA-Low station in Australia.

Whereas radio telescopes such as LOFAR, MWA and HERA can only statistically quantify the 21-cm signal over a limited range in scales and cosmic-time, the phenomenal sensitivity of the SKA enables truly transformational science. It can directly image the 21-cm signal over nearly the entire first billion years of the Universe, and visualise its

small-scale structure directly, something out of reach of all current experiments.

To be ready for the processing of the exabyte-scale data that SKA will deliver, much preparatory work is required. The Dutch EoR group already has a leading role in this thanks to the expertise built up by the state-of-the-art work done with the LOFAR data. The LOFAR EoR KSP team has already devoted much effort to developing very complex data-processing algorithms and tools that allow one to excise corrupting signals and correct errors in the data. These tools are highly optimised to run on distributed computing resources, including the currently available Graphics Processor Units (GPUs). It is crucial to start deploying and testing these scalable tools now, in order to be ready in 2024 when SKA comes online for commissioning and early science, and to be ready for real-time data processing in 2028 when its dedicated exabyte-scale key-science programme commences. Testing scalability of the codes and hardware to SKA-low levels can be done based on both existing (e.g. LOFAR) data and simulated SKA-low data sets.

Tracing Galaxy Evolution with Neutral Hydrogen

One of the key open questions in astrophysics is "How do galaxies assemble and evolve". The SKA will play a particularly important role in answering this question because it will be able to trace, for the first time, the gradual transformation over cosmic time of the primordial atomic neutral hydrogen (HI) gas into galaxies. Neutral hydrogen can be used as a fossil record to reconstruct galaxy history. HI is not only the most abundant element, it also extends to large radii – much larger than those of the stars, tracing key processes in galaxy formation like collisions or close encounters between galaxies (see Fig. 17). The HI gas is also responsible for the formation of stars in a galaxy, the growth of the central supermassive black hole and for the triggering of the nuclear activity [20]. In turn, the energy injected by these phenomena back into the galaxy can strongly affect its evolution and even expel large amounts of gas back into the inter-galactic medium (see [21] for a review). The balance between all these effects is a key ingredient in galaxy evolution and one of the central topics of research in extragalactic astronomy.

Figure 17: the left-hand image shows the high-resolution distribution of the atomic neutral hydrogen (HI) gas in blue, superimposed on an optical image (from the Sloan Digital Sky Survey) in the field of the galaxy M81 (visible in the centre). The complexity of the HI gas filaments and clouds is clearly visible strikingly different from the distribution of the stars in the galaxies. Throughout the field, bright and faint HI streams and filaments are visible, connecting these galaxies. The right-hand image shows only the HI distribution highlighting the complexity of the distribution of the gas, giving a more complete view of the interaction between the galaxies ([27]).

At the moment, HI observations of very gas-rich galaxies can be done only up to redshifts z~0.4 [22], or look-back times up to around 4 billion years. It is, therefore, exciting that the SKA precursors and pathfinders are providing a first revolution in the way HI observations are done. The wide field-of-view and the large instantaneous bandwidth of these instruments challenge our way of handling and analysing the data.

The real major step forward in sensitivity will be made by SKA providing for galaxies at cosmological distances a view of the HI comparable to what is possible today for the local Universe (see, e.g., [23] for an overview). SKA-Mid will be the most sensitive cm-wavelength radio telescope in the world. It will have more than 6 times the sensitivity of the Very Large Array, the most sensitive

telescope for HI to date, expanding to much larger look-back times (7-11 billion years) and allowing us to spatially resolve the distribution and kinematics of the HI (see e.g. [24]).

SKA HI observations will deliver significantly larger data volumes than current telescopes, with estimates approaching the exabyte scale per year. Significant effort will be directed towards ensuring that the calibration and imaging pipelines deliver high-fidelity advanced data products of the desired scientific quality. Further down the processing chain, analysis techniques need to be scaled up and improved, including automatic object identification and characterisation plus feature extraction.

Preparatory work has already begun, and the Netherlands is well positioned, with access to data from all the major SKA precursors and pathfinders (in particular MeerKAT in South Africa, ASKAP in Australia, and APERTIF in the Netherlands itself). Particularly relevant is the cooperation between the Netherlands & South Africa stimulated by the ASTRON collaboration with IBM, which has led to the setting up of a remote node of South-African MeerKAT data processing facilities at ASTRON.

Galaxies, Black Holes and Magnetism

The study of time-independent and broad-band electromagnetic radiation at radio frequencies is a uniquely sensitive probe of a number of phenomena in a wide variety of objects throughout the Universe. The intensity, morphology, polarisation and frequency dependence of this emission constrains the energetics and magnetic fields in structures such as clusters of galaxies or the jets of radio emission that are produced by black holes. This very broad research area is one the most productive for many radio telescopes, including LOFAR. SKA will cover a number of key scientific drivers in this area. SKA will make possible discovering the most ancient black holes in the Universe and understanding how they were formed, how they grow by consuming matter over time and how they energise and influence the medium in which they are embedded (e.g. [25]). Furthermore, some of the most distant black holes are thought to have formed before the Epoch of Reionisation in the very young Universe. If the SKA confirms this suspicion, it will challenge our theories of galaxy formation. Finding even a single radio-emitting black hole, will provide the unique opportunity of using it as a beacon and studying HI, this time in absorption, in a period when the Universe transitioned from neutral to almost completely ionised. This would provide a complementary way to the direct imaging of the Epoch of Reionisation experiments (e.g. [26]).

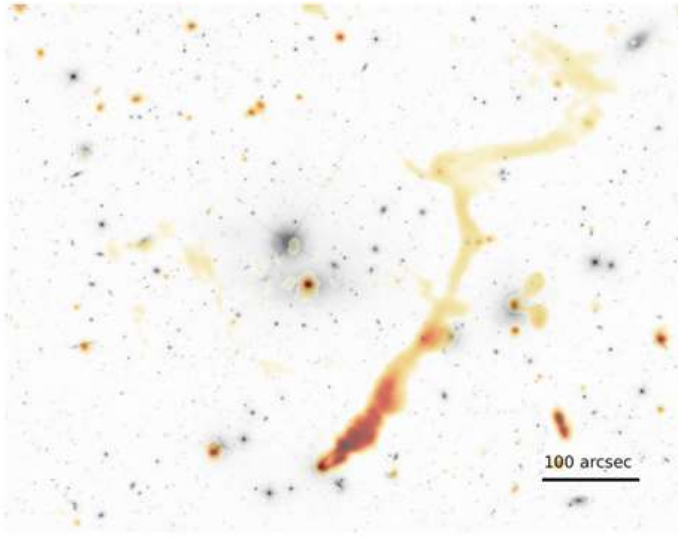


Figure 18: This image shows how the LOFAR radio telescope opens a new view of the Universe. The image shows galaxy cluster Abell 1314. In shades of grey, a piece of the sky can be seen as we observe it in visible light. The orange hues show the radio emitting radiation in the same part of the sky. The radio image looks completely different and shows that a single galaxy can be responsible for particle acceleration stretching for millions of light years. The radio emission is caused by matter that falls into a black hole, resulting in an unbelievable amount of energy that is ejected in the form of a relativistic fountain of particles. These accelerated electrons produce radio emission that can extend over gigantic distances and is not visible at optical wavelengths. Credit: Rafaël Mostert/LOFAR Surveys Team/Sloan Digital Sky Survey DR13.

Another component that plays a key role in how galaxies and galaxy clusters form is the magnetic field. How magnetic fields were created and evolve remains poorly understood but, thanks to the sensitivity of SKA, it will be probed by examining diffuse synchrotron emission from the tenuous filaments connecting galaxies and galaxy clusters. This radio emission allows us to infer that in these regions extreme particle acceleration is occurring and that a magnetic field is present. SKA-Low is

predicted to detect this type of emission in ~6 times more objects than LOFAR; this increased number will allow us to probe the strength of magnetic fields and particle acceleration processes further back in the Universe.

SKA-Mid will be able to construct sufficiently high-resolution images that allow systematic studies of the distortion of light from background objects by intervening mass. This technique, known as weak-lensing, will directly reveal the dark matter distribution in the Universe which, when compared with predictions, will constrain cosmological models (e.g. [27]).

LOFAR has stimulated significant advances in techniques for calibrating and imaging low frequency radio continuum data and sets the standard in terms of understanding the challenges involved in getting the most out of the data. The LOFAR surveys KSP currently uses about 7 million core hours per year of compute resources across a number of countries. SKA-low data volumes per observation will be ~200 times larger. The procedures for efficiently processing data must be scaled up, especially since the raw visibility data cannot be stored or exported.

Currently, the efficient processing needed to reach the sensitivity LOFAR is capable of, is a highly specialised, complex challenge and achievable only by a small number of expert users. It is clear that effectively allowing a large community of users to access, process, and analyse complex datasets is an essential component of any SKA science data centre. Initial steps are being undertaken by ASTRON already in the context of the Science Delivery Framework project for LOFAR. This effort will be expanded to take on the specific SKA challenges.

The Transient Sky

Over the past five years, multi-messenger transient astronomy has made exceptional discoveries of the most extreme phenomena in the Universe. There has been the first detection of gravitational waves from a binary black hole merger with broadband electromagnetic follow-up [28], the first multi-wavelength counterpart of a binary neutron star merger initially detected via its gravitational waves [29] and the first observations of neutrinos from an AGN flare that was also observable across the electromagnetic spectrum [30].

LOFAR is playing a key role in this new exciting field. However, with LOFAR and the current radio facilities we are only just probing the tip of the iceberg of these events, and only starting to understand their origin. Deep observations of multiple events by the SKA, working in tandem with

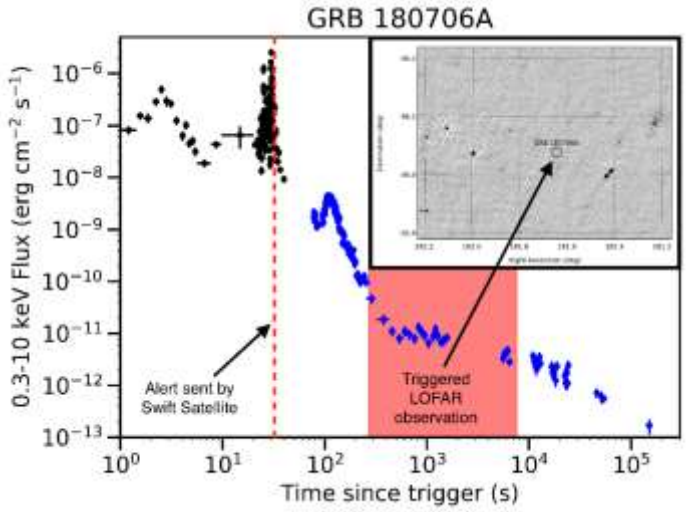


Figure 19: Results from the first, fully automated, LOFAR rapid response observation of a gamma-ray burst (GRB) detected by the Swift satellite. This is the high-energy light curve of the gamma-ray burst, where black data points are the gamma-ray emission and the blue points are the X-ray emission. LOFAR observed for 2 hours, starting 4.5 minutes after the GRB, as shown by the red region. In the top right corner, we show the 2-hour image of the region obtained and this provided the deepest constraint to date for coherent radio emission following a gamma-ray burst.

the next generation gravitational wave detectors (e.g. the Einstein Telescope), will enable us to accurately model and understand the extreme physical processes occurring within these events.

In addition to the entrance of multi-messenger events, the "transient sky" has been enriched with the discovery of very brief (millisecond duration), extragalactic flashes of coherent radio emission known as Fast Radio Bursts (FRBs; [31], Fig. 19). The origin of these highly luminous, mysterious bursts remains hotly debated; theories for their origin include highly magnetised, rapidly rotating neutron stars or the interactions between compact objects. Dutch astronomers are playing leading roles in the rapid development of this field, and are very well positioned to lead this work with SKA (for a recent review see [32]). For example, APERTIF detections of FRBs are being used to trigger the full power of LOFAR using a data buffering strategy that is capable of recovering the past few seconds of raw data. These searches already require computer facilities of exceedingly large scale. At SKA-Mid frequencies, the APERTIF Radio Transient System (ARTS), the largest data generator in the Netherlands, was our nation's most powerful GPU supercomputer at the time of construction (cf. [33]). It will search 24/7 for FRBs in all APERTIF data, which comprises more traffic than the entire internet of the Netherlands. Furthermore, it will publicly report discoveries in real time. At SKA-Low frequencies, the similar DRAGNET system for LOFAR is one of the more powerful pulsar and fast-transient search machines in the world.

Transient astronomy is currently within a golden era of discovery and rapid progress, with physicists and astronomers combining efforts to probe the fundamental forces of nature. However, a key feature of transient astronomy is speed, whereby the observations are processed in real-time and the results are quickly communicated to the wider community to inform further multi-messenger follow-up.

For the detection and follow-up of transient cosmic events it is therefore crucial that the data from SKA is processed in near real-time. This applies to both triggered observations, such as the rapid response to a gravitational wave event, but also for blind transient surveys using commensal survey data. These data will be processed on the SKA Data Processing centres by a dedicated fast imaging pipeline. Following imaging, the resulting time series of image cubes require rapid processing by a transient detection pipeline in an automated and consistent manner. The LOFAR Transient Pipeline (TraP) is the world leading transient detection pipeline and is ideally positioned to evolve into the SKA transients pipeline.

TraP can be run locally at the SKA data processing centres close to the telescopes to enable rapid alerts on bright transients with the resulting database being transferred to the SKA regional centre for further interactive processing and for long-term storage. TraP will also run offline on images that will be stored at the SKA regional centres, preventing the unnecessary transfer of large datasets. Alternatively, TraP can be run remotely at local centres with the results uploaded to the TraP database for long-term storage.

Pulsars

Radio pulsars are highly magnetised neutron stars that are detectable through their remarkably regular pulsations. These cosmic lighthouses provide a unique glimpse into the extremes of gravity and dense matter, and hence they allow us to test fundamental physical theories that cannot be probed in an Earth-based laboratory. By using pulsars as precision clocks, we can carefully track their motions around other objects and, in this way, we have used them to provide some of the most stringent tests of Einstein's theory of gravity, general relativity (e.g. [34],[35]). The bending of spacetime also allows us to precisely measure the masses of some pulsars, which has provided critical insights into the properties of matter at extreme density (e.g.[36],[37]).

Thanks to its sensitivity, the SKA will be the best-ever pulsar discovery and timing machine. Through both SKA-Low and SKA-Mid it is expected that we can discover a large fraction of the Galactic pulsar population [38]. This harvest of pulsar discoveries will include new laboratories for testing fundamental physics [39],[40], and it will provide the means to time an array of pulsars with

sufficient precision to directly detect the gravitational wave background produced by the inspiralling binary supermassive black holes at the centres of galaxies [41].

However, the discovery and scientific exploitation of pulsars demands large data volumes and high-performance computing, and the SKA era will push this well beyond the current state of the art. The LOFAR telescope has discovered 80 pulsars to date, and has served as a pathfinder towards the SKA. Massively parallel computing, GPU-enabled real-time processing, and machine learning are critical tools to identify the interesting astronomical signals within the many Petabytes of data that are produced in such a survey. The Dutch pulsar group has achieved this through use of both national facilities like Cartesius and custom-designed clusters like DRAGNET and ARTS.

Furthermore, pulsar timing requires complex data products, where the traceability of the analysis steps is deeply entwined with the astrophysical interpretation of the data. The SKA regional centres will host millions of pulsar pulse arrival time records, which are the fundamental input for the gravity tests and gravitational wave detection we aim to perform. Though each measurement is in principle only a few bits of information, the associated metadata is highly complex, and careful consideration of the architecture for encapsulating and presenting this information is necessary to enable this science.

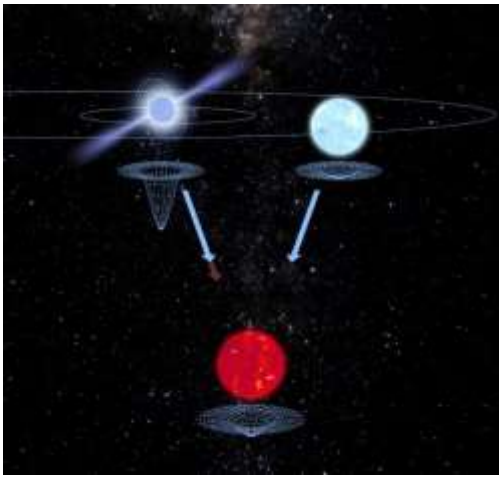


Figure 20: Using a pulsar in a triple star system, we have performed a high-precision test of the "universality of free fall", which states that all objects fall equally in a gravitational field, regardless of their mass or composition. Here we measured whether the pulsar and white dwarf "fall" equally towards the other white dwarf in this system. Universal free fall is a fundamental tenet of Einstein's theory of gravity, and hence a critical avenue to test using the ultra-high-precision afforded by radio astronomical observations of pulsars. Image: Neil Blevins.

Key Researchers

Name and title	Specialisation	Institution	DAI / ORCID*
Aaij, dr. R.	Physics Data Processing	Nikhef	0000-0003-0533-1952
Christakoglou, dr. P.	LHC / ALICE	UU/Nikhef	0000-0002-4325-0646
de Blok, prof. dr. W.J.G.	Radio Astronomy	ASTRON/RUG	0000-0001-8957-4518
Groep, dr. D. L.	Physics Data Processing	Nikhef	0000-0003-1026-6606
Heijboer, dr. A.	KM3NeT	Nikhef	eu-repo/dai/nl/265430534
Hessels, dr. J. W. T.	Radio Astronomy	ASTRON/UvA	0000-0003-2317-1446
Jackson, prof. dr. C. A.	Radio Astronomy	ASTRON	0000-0002-7089-8827
Koopmans, prof.dr.L.V.E.	Radio Astronomy	RUG	0000-0003-1840-0312
Merk, prof. dr. M.	LHC / LHCb	VU/Nikhef	eu-repo/dai/nl/097734160
Morganti, prof. dr. R.	Radio Astronomy	ASTRON/RUG	0000-0002-9482-6844
Raven, prof. dr. G.	LHC / LHCb	VU/Nikhef	0000-0002-2897-5323
Rowlinson, dr. A.	Radio Astronomy	ASTRON/UvA	0000-0002-1195-7022
Shimwell, dr. T. W.	Radio Astronomy	ASTRON	0000-0001-5648-9069
Templon, dr. J. A.	Physics Data Processing	Nikhef	0000-0002-3371-788X
van Haarlem, dr. M. P.	Radio Astronomy	ASTRON	0000-0003-0512-7687
Verkerke, prof. dr. W.	LHC / ATLAS	UvA/Nikhef	eu-repo/dai/nl/171559312

** to obtain the full ORCID identifier, please prepend "https://orcid.org/" to the ISNI-formatted number given in the table*

A subset of relevant publications for each of the key researchers above is provided in section 2.6.

2.1.2 Embedment of the investment

The infrastructure to be developed by this proposal will be deeply and strategically embedded in the Dutch research system: the infrastructure is wholly designed to maximise the integration of the major investments made by the Netherlands in the Research Infrastructures in the long term, and to couple this to a strengthened national e-Infrastructure.

The LHC experiments are performed at CERN, an international organisation of which the Netherlands is a (founding) member state, represented by the Ministry of Education, Culture and Science (OCW). The Dutch participation in the LHC experiments is coordinated through the Nikhef partnership, which includes the subatomic physics activities at six universities: the two universities in Amsterdam (VU and UvA) and those in Utrecht, Nijmegen, Groningen and, most recently, Maastricht (which joined Nikhef this year).

KM3NeT is a collaboration based on a Memorandum of Understanding covering its first construction phase. In the framework of a Horizon2020 Preparatory Phase project (KM3NeT-INFRADEV) the collaboration is currently exploring the option to establish a European Research Infrastructure Consortium (ERIC) as its legal entity. The Dutch participation in KM3NeT is, again, coordinated by the Nikhef partnership. The LHC experiments and KM3NeT together make up the majority of the activities within the Nikhef partnership, serving over 500 Dutch researchers (staff, postdocs, PhD students) over the next decade.

The SKA will be the world's leading radio telescope, constructed as two arrays in the 2020s along with a central organisation charged with its operation and development over a 50-year design lifetime. The Netherlands is one of seven founding members of the new Intergovernmental Organisation, with a special investment from OCW of M€ 30 as the foundation of the Dutch share of the organisation's build-up and its first 10 years of operations costs (announced 28 January 2019). This OCW investment has been made with the full expectation that a major e-infrastructure, as represented in this proposal, will be established for the nation. The coordination of the Dutch construction and operational activities for SKA lies with ASTRON, while the Dutch contribution to the SKA science programme is coordinated within the Netherlands Council for Astronomy, which combines the interests of ASTRON, SRON (the Dutch institute for space research) and NOVA (the Dutch research school for astronomy) representing the astronomy departments at the universities of Amsterdam (UvA), Groningen, Leiden and Nijmegen. It is estimated that the SKA facility, including the Dutch Regional Centre will serve around 500 researchers in the Netherlands over a decade.

Nikhef and ASTRON each fulfil leadership and coordination roles on behalf of the nation in these three global scientific facilities. These roles are the natural consequence of the Netherlands' strategic investment in these facilities, and as a result these are part of the Dutch National Roadmap. The proposed infrastructure is aimed to be wholly embedded in the national e-Infrastructure, as coordinated by SURF. SURF is a 'Cooperative', of which Dutch universities and research institutions are member. SURF was established more than 30 years ago with, at the time, as most important activity through SURFnet the roll out of the national research and education network (NREN). It currently consists, next to SURFnet, of two other units: SURFsara, resulting from the merger in 2013 with the compute center SARA, providing computing and data services, and SURFmarket, procuring and providing software licences on behalf of the SURF members.

A large share of the funding of SURF comes from OCW, through NWO, to innovate, develop and operate the digital infrastructure for research (e-Infrastructure). Based on an advice of the Permanent Committee for Large Scale Research Infrastructures of NWO, the Dutch government has recently made available structural funds for additional investments in this digital infrastructure[1]. Compute facilities for high-energy physics and astronomy are explicitly mentioned in the letter from the Minister to parliament. At the time of submission of this proposal, a spending plan for these funds still needs to be agreed by the Ministry.

[1] https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2019Z05854&did=2019D12183

2.2 Strategic case and innovation

2.2.1 Importance for Dutch science and international positioning and appeal

Serving three Roadmap RIs with advanced ICT

With this proposal Nikhef and ASTRON join forces with SURF, the national e-Infrastructure provider, in building and operating a nationwide e-Infrastructure (computing, storage, networking) serving three Research Infrastructures on the national Roadmap: the LHC experiments ATLAS, LHCb, and ALICE at CERN (high-energy physics), the Square Kilometre Array (SKA, radio astronomy) and KM3NeT (neutrino astrophysics).

The three global facilities are each in different stages of maturity: the LHC is an established infrastructure producing data since a decade, currently in shutdown for the installation of upgraded detectors for which our approved Roadmap proposal in 2013 provided funding (M€ 15,2, including budget for e-Infrastructure for the years 2014-2019) and data taking again from 2021. KM3NeT phase 2 is under construction (approved Roadmap funding of M€ 12,7 in 2018); and SKA will start construction in 2020. The NL contribution to the SKA design phase was supported by Roadmap funding awarded in 2014. In 2019 the Government, through OCW, has made a direct M€ 30 investment into the SKA.

During the five-year term (2021-2025) of this proposal, all three infrastructures (LHC, KM3Net and SKA), in which the Netherlands have invested and will be investing heavily, will be acquiring data at the exabyte scale. Large-scale computing infrastructures are needed to ensure Dutch researchers have access to the resources necessary to properly exploit the nation's major investments into these global endeavours. Indeed, it is foreseen to have computing for the Einstein Telescope Roadmap infrastructure as part of FuSE later on, for the same reasons.

The similarities between the computing infrastructure requirements coupled with the cost-benefit of collaboration have led to this Roadmap request where we ask for this joint computing infrastructure, fully embedded in the national e-Infrastructure. This is what makes this proposal unique, both nationally and in the European context. There is no feasible alternative to this approach; there is currently no Dutch academic facility capable of building FuSE, and implementation via commerical cloud computing would be factors more expensive.

Excellent e-Infrastructure coordination in the Netherlands

Compared to most European countries, the Netherlands is blessed by having an organisation such as SURF in which computing, data and networking services are being provisioned in a coordinated way. Over the last three decades the Netherlands, foremost SURF, has built up an excellent position in terms of e-Infrastructure development and provisioning. It has also led to significant economic effects (see section 2.2.2). This position has been the result of a great willingness of all relevant parties to work together.

An important step was taken through the BiG Grid project (2007-2012, funded with M€ 28,8). This project, steered by the scientific needs of physics and bio-informatics, involved close collaboration between Nikhef, Groningen and (then still) SARA. This project was evaluated in 2017 by a dedicated NWO-committee, which concluded: "*BiG Grid has shown to have been of great importance because of the horizontal 'cross-cutting' character of the facility and the acquired expertise and international reputation. The facility has become an integral part of the national e-Infrastructure for research. As 'horizontal' infrastructure this benefits in principle all sectors.*" This project concretely delivered, amongst others, the start of the Dutch WLCG Tier-1 and the LOFAR Long Term Archive.

A key moment in the formation of the current Dutch e-Infrastructure has been the report of the *ICT Regie-orgaan* in December 2008: "Towards a competitive ICT infrastructure for scientific research in the Netherlands". In this report, the foundation has been laid for the consolidation of the national e-Infrastructure components (networking, computing, data) under SURF and the formation of the Netherlands e-Science Centre (NLeSC). Nikhef has played a key role in architecting this landscape, not only for high energy physics. The final report of the international committee for the SEP

evaluation of Nikhef in 2017 had this to say: "*The Nikhef computing team has been instrumental in developing the distributed computing paradigm. Nikhef was a major player in practically all the European grid development projects during the past 15 years (European Datagrid, EGEE, EGI.eu, EMI, etc.). SURFSara operates the NL/Tier1, which is part of the Worldwide LHC Computing Grid, in partnership with Nikhef, and investments in the NL/Tier1 have been included in the granted LHC detector upgrade funding until 2019. The Nikhef team has strongly contributed to the development of the national strategy for sustainable distributed computing in support of scientific research.*"

ASTRON aims for a similar position within SKA, a Science Data Centre being the equivalent of an LHC Tier-1. This ambition builds on pioneering work done by ASTRON, supported by SURFsara and the NLeSC. The proposed work taps into this rich Dutch ecosystem; the proposal partners have strong connections with relevant computer science activities, both directly and through SURF and the NLeSC. To illustrate: researchers at both Nikhef and ASTRON are running projects at the NLeSC and the PI of the FuSE-proposal is member of the eScience Advisory Board of the NLeSC.

Finally, both Nikhef and ASTRON are prominent partners in Horizon2020 e-Infrastructure projects such as EOSCpilot, EOSCHub, AENEAS, ASTERICS and most recently ESCAPE.

Position of the Netherlands in the European e-Infrastructure (EOSC)

Having our own house pretty well in order, it is no coincidence that three European organizations for coordinating e-Infrastructures have established their head office in the Netherlands: GEANT, the European Association of NRENs (Amsterdam), EGI, the European Grid Infrastructure (Amsterdam) and GO-FAIR, the European initiative to implement FAIR data principles (Leiden).

Dutch delegates (amongst which Nikhef staff) have also been very active in European member state based policy advisory bodies such as the e-Infrastructure Reflection Group (www.e-irg.eu), contributing to most of its policy papers. In the 2013 e-IRG White Paper[2], the concept of the e-Infrastructure Commons (see Fig. 21) has been introduced, which has later been rebranded by the European Commission into the *European Open Science Cloud.* The EOSC is described[3] as "*offering professionals in science and technology a virtual environment with free at the point of use, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines*". The EOSC is one of the cornerstones of the European Commission in reaching the goals of the digital single market.

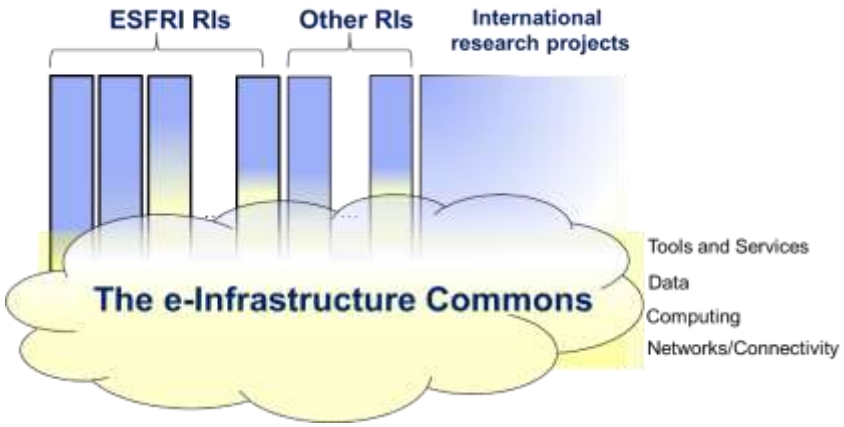


Figure 21: the e-Infrastructure Commons, the basis for the European Open Science Cloud, as envisioned in the 2013 White Paper of the e-Infrastructure Reflection Group

For the EOSC, the challenges will reside in the interface between discipline specific (vertical) and generic (horizontal) infrastructures (or e-Infrastructures). E-infrastructures have the potential of being efficient and effective, pooling hardware and software and even more importantly people and

[2] http://e-irg.eu/documents/10920/11274/e-irg-white-paper-2013-final.pdf

[3] Implementation Roadmap for the European Open Science Cloud, 14.03.2018, https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

expertise together instead of building disciplinary pillars. This is how we have been working in the Netherlands in the past decades. The well-coordinated Dutch e-Infrastructure, in combination with the Dutch ambitions towards Open Science, have given the Netherlands an influential position in further shaping the EOSC.

In its upcoming policy paper on e-Infrastructure coordination within and between European countries[4], e-IRG recommends the following: "*Member States and Associated Countries should explore, pilot and install funding schemes, which give the incentive to both research communities and provisioning organisations to collectively optimize e-Infrastructure service development and provisioning"*. We believe the current proposal is spot on in addressing this challenge.

In summary, this proposal will further strengthen the excellent position of the Netherlands in the European e-Infrastructure ecosystem, boosting the existing joint effort of ASTRON and Nikhef in the European Science Cluster for Astronomy & Particle physics ESFRI research infrastructures (ESCAPE).

2.2.2 Importance for society and industry and the connection with societal developments

Socio-economic value

The early activities of Dutch e-Infrastructure partners have brought the Netherlands into an excellent position with regard to internet connectivity. For instance, SURFnet has been key in establishing the Amsterdam Internet Exchange (AMS-IX), one of the largest internet exchanges worldwide. This and other factors have made the Netherlands and the Amsterdam region in particular a very attractive place for connectivity data centres with an enormous aggregated economic value and providing employment for more people than for instance Schiphol airport or the Rotterdam harbour[5].

Through its early involvement, the Nikhef data centre has a very prominent position in this ecosystem as a connectivity hub bringing together more than 160 different networks (customers), basically making every corner in the world easily reachable, which greatly enables worldwide research collaboration. Its current annual turnover is *undisclosed*. Derived from a recent acquisition by a third party of a comparable data centre at the Amsterdam Science Park it is not farfetched to estimate the market value of the Nikhef data centre to be several tens of millions of Euros.

Now, where the internet, starting off primarily in academia, is now an 'established' technology, experience teaches us that many solutions to the current challenges researchers face in e-Infrastructure provisioning will find their way in other realms of industry and society. What started out as our e-Infrastructure 30 years ago, with Nikhef as the third website in the world, now serves as the platform for e-commerce, social media, and Internet banking.

Figure 22: Vendor's early-engineering sample of a high-density-high-performance storage system being unwrapped at the Nikhef Data Processing Facility.

The scale, expertise and breadth of usage of exascale data processing makes us an interesting technology partner for the vendors providing the advanced technology. Working closely with hardware vendors such as AMD, Dell, Fujitsu, HGST, IBM, Intel, Juniper, Mellanox, NetApp, Seagate, and Western Digital, has also, in the past, resulted in capabilities being added "for us", which cannot be underestimated; with the market steering strongly towards low-power and mobile devices, a degree of influence within the niche high-performance and high-

[4] Implementing the e-Infrastructure Commons and the European Open Science Cloud; National Nodes - Getting organised; how far are we? e-IRG; to be published June 2019.

[5] Dutch Digital Infrastructure 2016 Enabling the digital economy and society. Deloitte Consulting, November 2016

throughput segment is vital to achieving the goals of FuSE, and in the future other data-intensive sciences, which will reach similar scales five to ten years after the FuSE sciences. Here again we stress that the approach of working on 'generic' (horizontal) solutions is key.

ASTRON plays a prominent role in the knowledge economy in the North-East of the Netherlands. Companies, many of them SME's, have been involved in technology development and production of both hardware and software for its telescopes. This continues to be the case today and also in the future, with many lucrative contracts for SKA construction expected in the coming years. Furthermore, the Dutch SKA Science Data Centre will also be a platform for collaboration between ASTRON and partners in industry and academia in areas that are relevant to its mission and also in line with societal developments. Topics to be addressed include data science and machine learning, low power computing and innovative architectures.

Although the future is by definition hard to predict, one can imagine that where 30 years ago internet developments led to 'internet exchanges' having enormous socio-economical impact, we might now envisage to work on solutions eventually leading to 'data exchanges'. In fact, the first ideas of establishing an Amsterdam Data Exchange have been coined last year.

Alignment with national policy frameworks

Our infrastructure will enable scientists to answer several cluster questions from the Dutch National Research Agenda (NWA). In the NWA consultation round, the public submitted many hundreds of questions about The Universe, from small to large. The science cases that are supported by this proposal fit exceedingly well with two routes of the National Science Agenda: "Building blocks of matter and fundaments of time and space" and "The origin of life - on earth and in the Universe".

Because of this excellent match, we are well aligned with the NWA Game Changers (GC) described in the portfolio for research and innovation: the "*Fundamentals GC1: Einstein Telescope*" (ET) is a post-2030 gravitational-wave detector, successor to Virgo/LIGO. Our infrastructure will make a large and direct impact on the ET success. Strong links also tie "*Fundamentals GC2*", a Dutch Institute for Emergent Phenomena, to our proposal: newly discovered relativistic binary neutron stars test theories of emergent gravity. The Big Data route is styled as a single encompassing GC: "Responsible Value Creation with Big Data". We support and adhere to these Big Data responsibility criteria, and will make applied contributions to big-data value creation.

With regard to the Dutch 'Top sector policy', most of the Nikhef and ASTRON activities are part of the sector 'High Tech Systems and Materials' (HTSM), roadmap Advanced Instrumentation. However, the policy is currently undergoing a reordering of agendas; the activities as described in this proposal will most likely become part of the *'KIA Sleuteltechnologieën*' (knowledge and investment agenda for key technologies), under the heading of Big Science.

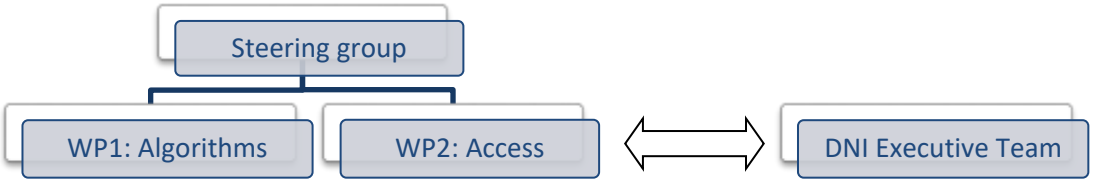
There is also close alignment with the recently drafted 'sector plans' from the universities. All six universities within the Nikhef partnership have proposed positions, strengthening their particle and astrophysics activities, in particular but not exclusively around gravitational wave detection (preparing for another initiative on the National Roadmap, Einstein Telescope).

2.3 Management case

2.3.1 Organisation and governance

Nikhef and ASTRON belong to the same legal entity (the NWO-I Foundation, the Institutes Organisation of NWO). NWO-I is a member of the SURF Cooperative. ASTRON, Nikhef and SURF work together closely in major EU projects and programmes (e.g. ESCAPE, EOSC). Mechanisms are already in place to coordinate the national e-Infrastructure and to allocate 'Rekentijd' (i.e. the NWO call for submissions of computing time). Hence, we believe that no additional governance arrangements are needed for this project, in particular a separate consortium agreement.

Using these existing mechanisms, the project management structure is very straightforward. The steering group (top level) will consist of the directors of both Nikhef and ASTRON. As described in detail in section 2.4, the project work will be divided into two main activities: delivery of algorithms (WP1) and delivery of infrastructure access (WP2). The last one (access) will liaise with the already existing SURF-led Executive Team of the DNI, in which all DNI operational partners are represented and which is responsible for the third activity: delivery of e-Infrastructure services. For the two WPs, appropriate technical leaders will be appointed from within the teams at Nikhef and ASTRON. The Executive Team of the DNI, as well as the teams at Nikhef and ASTRON, are well-connected and represented in the relevant international management bodies coordinating computing for the LHC, for KM3NeT, and for the SKA.



The WP leaders have the collective responsibility of planning, executing, and assessing the technical work done in the project. The Steering group is responsible for monitoring the alignment of the technical work with the (evolving) scientific goals of the targeted Roadmap infrastructures and of both institutes. Project administration and budget control will be carried out by NWO-I (with Nikhef as primary responsible institute).

This project will not have the sort of budget and time overruns typically associated with a physical construction project. In terms of service delivery (core hours, disks and tape) the DNI partners have a demonstrable record of being able to deliver capacity on time and within budget, the latter to the extent possible and subject to certain uncontrollable factors such as exchange-rate fluctuations.

For the development activities ('people-ware' portion of the proposal), there are also no budget or time overruns foreseen. The consequence of any eventual delays or obstacles in this work fall then on the speed with which results can be harvested from the data; this accomplishes focus and motivation for the development efforts in a completely natural fashion.

Procurement and Intellectual Property Strategy

The software and knowledge developments in the project will be executed with personnel employed by the institutes involved (ASTRON and Nikhef), and the resulting publications and software will be made available under a reuse-friendly open source license (the Apache License version 2.0 where possible, or under another appropriate open source license for elements contributed to existing projects and efforts). Commercial exploitation of such results in the form of patents and closed licenses is explicitly excluded, since such restrictions would impair the usefulness for the science cases being pursued.

The ICT infrastructure will form an integral part of the national e-Infrastructure for research (DNI) coordinated by SURF and will be procured as part thereof. The acquisition model will be based on the current practice of a joint public procurement by the operational partners for the most economically advantageous offer, thus ensuring the unique composition and scope of the national e-

Infrastructure. Partners in the existing procurement framework agreement are SURFnet, SURFsara, and Nikhef, and the joint framework procurement model, introduced in 2008 and re-tendered in 2012 and 2016, has shown the required compliance as well as agility. Furthermore, ASTRON has extensive experience in procurement and assessment of HPC/HTC equipment, including, most importantly, systems that have accelerator technology.

For the infrastructure elements procured within the scope of this proposal, no commercial exploitation by third parties is foreseen. Like today, use of the infrastructure as part of the DNI for other publicly funded research applications will be possible as part of load-sharing arrangements within the context of the SURF Cooperative and the DNI.

Key Performance Indicators and Reporting

For service delivery, the monitoring and evaluation infrastructure exists, within the framework of the WLCG Collaboration and the European Grid Infrastructure (EGI), both of which the DNI is a member of. Here the key performance indicators are: *(a)* whether the requested resources have been delivered (accounting), and *(b)* to what extent are the delivered resources available for use (reliability and availability). Accounting, Reliability and Availability are both measured and monitored by the DNI itself, by EGI, and by WLCG (monthly reports are made available to the public). Given that KM3NeT and SKA activities will be run on the same DNI systems, it is straightforward to incorporate those activities into this existing monitoring and evaluation infrastructure. Dutch resources were amongst the first to be monitored, and the system has been refined and improved over a fifteen-year period.

The key performance indicators for the personnel part of the project will be associated with the extent to which the FuSE resources will a) be utilized (i.e. has the access delivery activity been successful?) and b) are sufficient to extract the science results (i.e. has the algorithm-development delivery activity led to the necessary algorithmic speedups?).

2.3.2 Accessibility

Accessibility needs to be explained both on the level of the national e-Infrastructure (DNI) and on the level of each Research Infrastructure (scientific collaborations). On the national level the resource provisioning as described in this proposal should be considered as a block allocation through the 'NWO call for submissions of computing time'. This corresponds to the in-kind contribution of SURF as outlined under financial feasibility (section 2.4.3).

The LHC resources are provided at the Tier-1 level. Tier-1 resources are associated with crucial tasks essential to the collaboration (LHC experiment) as a whole, and as such, the user access is managed at the collaboration level. Generally speaking, the bulk of the resources are used for activities that benefit all members of the collaboration, worldwide (i.e. many thousands of researchers). In order to make efficient use of the provided resources, user-analysis jobs are also permitted in the occasional pauses in the collaboration-wide computing activity. Given this pattern, we can best describe the 'average extent of use per external researcher' by saying that all of the science produced with the served LHC experiments will to some extent have relied on resources funded by this proposal.

For KM3NeT, the high-level results of the Tier-1 processing will be open to 'external users'; this data will be served by portals that are developed in the collaboration (building on work in ESCAPE), where it will be possible to combine with data from other observatories. It is possible that a small fraction (order 10%) of the Tier-1 processing budget can be committed as a backend to facilitate such a portal.

SKA observing time and resources required for processing of the data will be allocated on the basis of scientific merit and technical feasibility via a common process by a Time Allocation Committee. Access is primarily restricted to scientists from SKA member countries, although there will be provision for Open Time, accessible to users from any country, at a level still to be determined but likely to be around 5-10%. The SKA Regional Centres (SRCs) collectively will provide the archive,

hosting both original Observatory Data Products as well as Advanced Data Products produced by users at the SRC or elsewhere. Both types of data products will be made openly available, either immediately or after a short proprietary period (typically of order 6-12 months). The SRCs will be required to comply with data access restrictions set by the SKA Observatory during the proprietary period, but their objective is to facilitate sharing of SKA data, processing workflows and tools and to support discovery of/interoperability with information gathered from other sources.

2.3.3 IT infrastructure

This proposal is all about building up and operating an IT infrastructure, so in a way the complete proposal fits under this heading. ASTRON and Nikhef collaborate, in perfect alignment with SURF, using existing mechanisms, on this infrastructure in order to maximise the cost-effectiveness of the infrastructure, both in terms of operations and R&D. Cost-effectiveness is addressed by having a common ICT infrastructure with common operations staffing and housing. The impact of R&D funding is optimised by working together.

The ICT resources for which funding is requested will be used in a number of ways:

1. Storage of the primary and secondary research data products
2. Computing needed to construct the secondary research data products (tables of physical quantities, multi-dimensional images, events) from the primary research data products (collected detector signals, visibilities & time series data)
3. Simulation of how the detectors/telescopes are expected to respond to the physical processes (needed in validation of the data analysis)
4. Analysis of the secondary research data products to arrive at (publishable) physics results.

These steps are common to all the science cases.

The ICT resources are essential to the experiments. A typical LHC experiment records about one billion events per year; analysis via computers is a given. SKA takes this a step further: the ICT resources are what transforms the instrument from a collection of radio antennas into a telescope, or better said several simultaneous telescopes.

Both Nikhef and ASTRON participate in the NWO institutes Data Management contact group, where a policy framework on how to best implement FAIR data access is developed and its deployment facilitated and monitored. Both institutes have developed internal data management policies conforming to this framework.

2.4 Technical and business case

2.4.1 Technical feasibility

In the Science Case (section 2.1), we have described how the scientific goals of the infrastructures translate into ICT requirements. Aggregating the requirements over all infrastructures results in a Dutch National e-Infrastructure (DNI) several times larger than the current size. This expanded DNI will also require significant new capabilities (GPU and/or Tensor computing). Furthermore, R&D work is needed in several areas, to enable all the facilities to use this e-Infrastructure at the scales implied. The technical challenges fall into three broad categories: Infrastructure requirements and evolution of the DNI, Access R&D and Algorithmic R&D.

As discussed in section 2.3 (Management Case), the R&D activities will be organized into two work packages (WP1: Algorithms, WP2: Access), whilst for the work on the DNI requirements the project will liaise with the DNI Executive Team. We will now first describe these requirements and the technical challenges, followed by a detailed description of both work packages.

Infrastructure Requirements and Evolution for Computing

In estimating the requisite number of computing cores for the period 2019-2030, inevitably extrapolations have to be made. The period of 'easy gains' based on *Moore's Law* has long since passed, and for data intensive processing a significant fraction of the cost is driven not by raw processing speed, but by the bandwidth required between processor and memory, and by the efficiency with which the processing cores can be utilised in terms of predictive code execution and data parallelism. Beyond the 5-7 year timeframe, estimations become increasingly hostage to the fortune of industry developments.

Processing of large data volumes has conventionally been limited to the use of 'general purpose' processors (CPUs) in view of the data ingest requirements and system throughput limitations that are inherent in the current systems architecture of co-processors such as GPUs (graphical vector co-processors) and Tensor cores (small matrix multiplication co-processors).

Research done at ASTRON and the University of Groningen has shown that GPUs can be more cost-effective for most of their data processing than traditional CPUs. The LHC experiments currently use predominantly traditional CPUs, whilst GPUs are becoming more attractive for the same reasons. Work will be carried out within the collaboration to transform the current, still somewhat experimental and specialised GPU services available at the institutes and in the DNI, into services capable of handling the exascale processing challenges of the experiments, similar to the transformation described for the LHC Tier-1. The e-Infrastructure as constructed will be a mix of general-purpose CPUs, vectorised GPUs, and matrix-calculation optimised Tensor co-processors.

Figure 23: Experimental system at the Nikhef Data Processing Facility for general-purpose GPU (vector-processing) programming, comparing GPU architectures and their science application performance.

The computing requirements for the LHC experiments and KM3NeT have been provided by the collaborations in a high-energy-physics-specific unit HS06. This can be converted (for 2019) into core hours using the average HS06 rating for a CPU core (15.8) and the number of hours in a year (8760); furthermore, we account for how much of the total computing is provided by the Dutch facility (different per experiment, ranging between 3.4% and 20%).

To make this calculation for 2020 and beyond, we need to project the average HS06 per core into the future. We do this by fitting the historical trend of the industry-standard SI06 ratings measured

for the hardware at Nikhef, and assuming the same trend for the development of HS06 per core past 2019.

Extrapolating the trends in application-level performance (as opposed to rather artificial numbers of numeric benchmark suites), a yearly linear performance improvement of 15% is justified (as shown in Figure 24). Incidental effects and design choices made by processor manufacturers (mainly AMD and Intel) are also evident in the figure, with AMD (whose processors determine the 2018 numbers) clearly emphasising the importance of core count over core speed, and of memory-intensive applications (the HEPspec06 application benchmark) over compute-only performance (the SpecInt06 number).

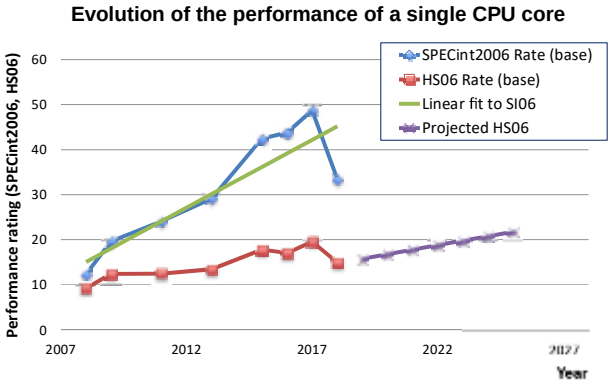


Figure 24: processor computing single-core perfor-mance trends over the past decade for the compute service of the Dutch National e-Infrastructure. The integer processor performance is represented by the SpecInt 2006 Rate-base (SI06) benchmark in blue, while realistic application performance for data intensive processing applications for the LHC is represented by the HS06 benchmark in red. The per-core drop in performance in 2018 is determined by the price-efficiency of AMD processors that emphasise core count per CPU package (i.e., die, socket) over core speed.

A large part of SKA computing will be performed on GPU-type machines. The performance improvement for GPU (and Tensor) computing are also hard to project, but with major investments in the area and the introduction of novel designs, annual improvement is much better than for general-purpose processor cores. Using the NVidia DGX-1 and DGX-2 as reference systems (the latter providing 2 PetaFLOPS – 10^{15} floating-point operations per second – of Tensor-core performance), the cumulative increase in terms of FLOPS each year still achieves 60%. The data throughput systems infrastructures in which GPUs and Tensor cores will be deployed is more akin to the current high-throughput cluster computing systems than the low-latency supercomputing systems, for which ample costing projections are available. To do justice to the necessary new developments and infrastructure designs for especially SKA, the performance improvement projections have been converted to equivalent '2019' CPU cores and indexed accordingly. In this way, we can express the equivalent cost of the GPU facilities needed for SKA in terms of the CPU service for which cost data are available from SURF. For the fraction of SKA data processing that can only run effectively on conventional processors, the computational requirements (expressed in PetaFLOPS) have also been converted to a number of CPU cores.

The proposed e-Infrastructure will be realised as part of the Dutch National e-Infrastructure. The total costs (including purchase, maintenance, power, cooling, operations) are available for general-purpose CPU cores (cost per core-hour), as well as for fast and slow storage (per terabyte). Any decreasing cost per core (which is mostly subject to systems-level parameters) is accounted for in section 2.4.3 (Financial feasibility).

Requirements for Use of the Dutch National e-Infrastructure DNI

The large amount of processing, storage, and network necessary to enable the key science results are such that it in itself can be seen as an infrastructure: the *e-Infrastructure*. This refers to an ICT-based infrastructure that encompasses a balanced range of ICT components and services, and coordinated through service management. A suite of middleware, frameworks, and tools are needed to provide coherent access mechanisms such as single-sign-on, workflow management, and data discovery. Coherency of the Dutch National e-Infrastructure is managed by the DNI Executive Team.

Supporting both the first years of LHC operation as well as the processing for the KM3NeT and SKA predecessors (ANTARES and LOFAR), the Dutch National e-Infrastructure has provided the core services that make an ICT infrastructure for research a reality. The DNI is operated by SURF, Nikhef, and the Groningen RUG-CIT centres as a coherent whole, building on the structures established in

the BiG Grid project (2007-2012) and the subsequent investments jointly by SURF and Nikhef (via the 2013 LHC Upgrade Roadmap). The LHC Tier-1 service is offered to the experiments in the WLCG Collaboration for use by the ATLAS, LHCb, and ALICE by way of the DNI.

Data processing facilities of the type needed at the start of the project in 2021 have been the mainstay of the national and global e-Infrastructures for research. The LHC Tier-1 facility has been continually upgraded over the years, most notably with newer, faster computing processors, with evolving disk architectures, and very large increases (factor of 30) in network bandwidth between the computing and storage services. The LHC Tier-1 facility, it being offered as a service by the national e-Infrastructure since 2008, has also served as the blueprint for other data intensive sciences, including the pathfinder experiments for both KM3NeT (the ANTARES neutrino telescope) and for SKA (the LOFAR low-frequency array radio telescope). For both, concrete processing of data on the common national e-Infrastructure has demonstrated the intended joint ICT infrastructure is technically possible, effective, and efficient.

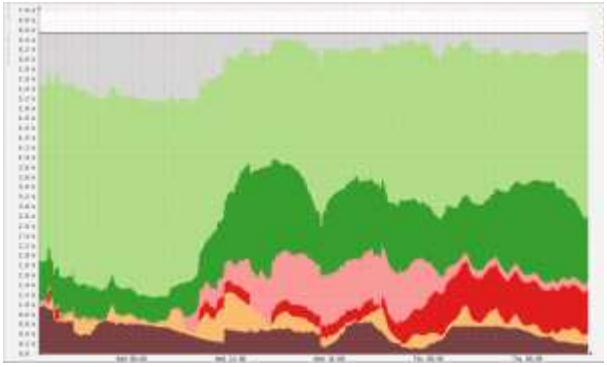


Figure 25: Visualization of computing jobs running on the Nikhef node of the Dutch National e-Infrastructure. The light green, dark green, and pink bands depict job slots (computing cores) carrying out tasks for the three LHC experiments, and the dark red band shows cores carrying out ASTRON work related to the LOFAR radio telescope, the SKA pathfinder experiment. The yellow band is work for the WeNMR collaboration (structural biology); the dark brown band depicts work from a collection of six other sciences ranging from biomedical science to gravitational waves (Virgo/LIGO, a precursor to the Einstein Telescope Roadmap infrastructure).

As can be seen in Fig. 25, the DNI is currently carrying out work for the three LHC experiments and for LOFAR, the "pathfinder" experiment for SKA. The infrastructure is, in this sense, already "realised". However, the infrastructure is not yet capable of achieving the scientific goals of the LHC upgrade, KM3NeT, and SKA. The amount of CPU (cores), disk, and tape (hardware e-Infrastructure) need to expand by a factor of 3 (3.5 for disk) between now and 2025, and the embryonic GPU/tensor computing capability of the DNI will need to be developed and made production-ready. The DNI hardware needs to be expanded in order to supply this capacity. The capacities needed (per Roadmap infrastructure) to meet the needs of the Science Cases are given here.

Since all of the Infrastructures are by their very nature global in scope, and given the volume of resources necessary to extract the science value from the data, the computing needs surpass the capacity of any single large central facility. In order to ensure excellent access of Dutch researchers to the data from LHC, KM3NeT, and the SKA, we need an ICT infrastructure commensurate with the Dutch participation in these global infrastructures, interlinked with high-bandwidth networks to the sources of measured and reference data.

The resource profile depicted in Fig. 26, when deployed at the Dutch national e-Infrastructure, will maintain our essential contribution to the LHC, KM3NeT and SKA ICT infrastructure, and will continue to provide Dutch astronomers and physicists excellent access to their data.

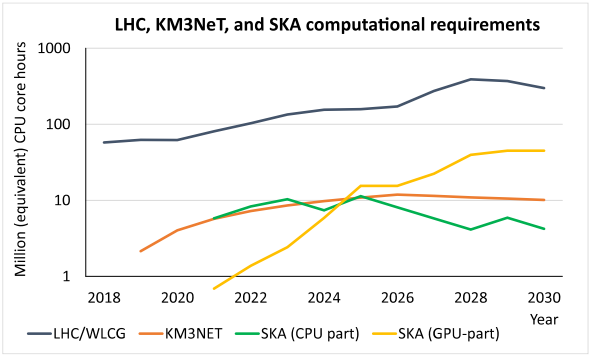


Figure 26: Quantitative compute requirements of LHC, KM3NeT, and SKA. For the latter, the needs of both commodity CPU computing as well as the equivalent need for GPU co-processor capacity is shown (million core-hours per year) in terms of the number of then-current processing cores (hence the decrease for LHC after 2029 and SKA after 2026 (dependent on the effectiveness of deploying GPUs), as their computing needs will not rise but the average 15% per-core performance improvement is expected to be sustained).

Continuous R&D work on the e-Infrastructure design itself is essential to achieving the scientific goals. It has been continually upgraded since 2008 (BiG Grid), to supply the advancing needs of Dutch science. The continued evolution of the technical capabilities of the DNI coordinated through SURF motivates our choice to explicitly liaise with SURF in the technical management.

Given the above factors, we have calculated the dimensions of the FuSE infrastructure year-by-year, expressed in the quantities in which ICT resources are conventionally expressed: compute cores for processing, on-line disk-storage requirements in Terabytes (or rather: Petabytes), and the needs in terms of long-term storage for data validation and re-processing on slower storage media (tape). The tables below (for the period 2021-2025) also serve as the year-by-year deliverables and milestones for the technical realisation of FuSE.

	LHC	**KM3NeT**	**SKA**	
Year	core-hours (M)	core-hours (M)	core-hours (M)	Equiv. core-hours (M, on GPUs)
2019	62.4	2.1	0.0	0.0
2020	62.2	4.0	0.0	0.0
2021	81.0	5.7	5.8	0.7
2022	103.3	7.2	8.3	1.4
2023	134.4	8.6	10.4	2.4
2024	155.1	9.8	7.4	5.9
2025	158.1	10.9	11.3	15.5
2026	171.6	11.9	8.1	15.5
2027	275.1	11.4	5.8	22.5
2028	390.5	11.0	4.1	39.7
2029	370.9	10.5	5.9	44.9
2030	299.7	10.1	4.2	44.9

Infrastructure requirements and evolution for on-line and near-line storage

The processing of data critically depends on the ability to make that data available to the processing cores, and to store generated reference data sets for future use as fast as they are generated. Any imbalance between the bandwidth of the processing, network, and storage systems would inevitably lead to inefficiencies in the utilisation of resources. As such, the performance parameters of the online storage systems (disk-based) and the network need to be aligned with the processing power and storage space available – the optimal value of which has been determined in the current infrastructure to be 12 MiB/s/TiB. Lower performance requirements apply for retrieval of data from (tape-based) near-line storage, since such access will be structured and allow for streaming access.

High throughput (disk) in Petabyte			
Year	**LHC**	**KM3NeT**	**SKA**
2019	10.1	0.4	
2020	11.0	0.8	
2021	13.4	1.1	0.0
2022	17.4	1.5	0.0
2023	21.7	1.9	0.5
2024	23.6	2.3	3.5
2025	25.7	2.6	8.0
2026	28.7	3.0	18.0
2027	38.7	3.0	43.0
2028	65.8	3.0	80.0
2029	63.9	3.0	95.0
2030	70.2	3.0	125.0

Bulk near-line storage (tape) in Petabyte			
Year	**LHC**	**KM3NeT**	**SKA**
2019	21.7		
2020	22.1		
2021	26.4		
2022	35.3		
2023	43.5		
2024	46.0		4.0
2025	46.3		18.0
2026	85.1		43.0
2027	127.0		100.0
2028	169.1		170.0
2029	170.0		220.0
2030	171.1		340.0

Infrastructure requirements and evolution for global network connectivity

The global nature of all of the infrastructures, and the need for going beyond a single large central facility, requires the exchange of both raw data and analysed results between the data centres. Since data flows of raw and analysed data are both voluminous as well as restricted to an enumerable list of end-points, a mesh of dedicated links is the most efficient method of moving the data and not inadvertently obstructing commodity research-network traffic. To deal with such 'elephant flows', the LHCOPN (for raw data) and the LHC Open Network Environment (LHCone) have been established by the global research and educational networking community in collaboration with the Tier-1 sites. Using dedicated paths significantly lowers cost as these elephant flows are dealt with at a deeper layer of the network using largely data-centre class equipment (instead of the usual enterprise-level packet-inspecting services) in science collaboration zones.

The dedicated links between the DNI and the data sources in Geneva, South Africa, and Australia, are delivered by SURF (SURFnet) yet form part of the ICT infrastructure of the experiments and are costed as part thereof. Current network speeds are 20 Gbps for the LHCOPN raw data link, and 40 Gbps for the LHCone interconnects. These link speeds will have to be upgraded to accommodate the larger traffic flows (to 400 Gbps in 2021 to Geneva, and 2x100 Gbps to the SKA sites in 2024) with further upgrades thereafter to peer data centres also in Canada.

Technical Challenges in the Data Processing Pipelines

We list here the technical challenges associated with advancing the infrastructure to a state capable of adequately addressing the scientific challenges.

- When estimating the computing needs, the infrastructures have assumed that the algorithms (software e-Infrastructure) have been improved and adapted to modern processor architectures, making the computing significantly faster (work package task *WP1.1*). Without this work, the computing costs would be much higher: as an example, for ATLAS 50% more computing time would be required in the absence of this R&D work.
- The factor 3 increase in storage capacity is already a challenge; after 2025 this factor increases rapidly, reaching 20 by 2030. To handle such enormous amounts of data, R&D is needed on both storage architectures and on data management and access infrastructure. This work has been started in the context of the ESCAPE project.
- The KM3NeT and SKA e-Infrastructures for managing the computational tasks need to be developed. This work has been started in the context of the ESCAPE project. For the LHC, such production systems have already been developed.
- Investment in algorithms based on Machine Learning (work package task *WP1.2*) is needed. Processors aimed at ML (so-called tensor cores) are improving much more quickly than are general-purpose CPUs, hence ML techniques are seen as a key factor in containing computing costs in the future.

In order to prepare for the higher rates at the LHC, for KM3NeT computing, as well as for the SKA Regional Centre in the Netherlands we need to develop software, tools and services and deploy them in the form of advanced pipelines and workflows that map onto a modern heterogeneous ICT infrastructure. The objective is to provide both expert and non-expert users a way of finding, processing, analysing, visualising, and publishing data without the need to delve into the complex software and infrastructure that is required to achieve the quality of data products that the facilities are capable of delivering.

All three projects at the heart of this proposal - SKA, LHC and KM3NeT - are part of the European Commission's Horizon2020 *ESCAPE* project, which aims to connect infrastructures on the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) to the European Open Science Cloud (EOSC). The objective of ESCAPE is to bring together projects with aligned challenges of data-driven research in the fields of astronomy and particle physics. Its goal is to create a cloud of data services that provide a robust, reliable and manageable set of data and storage services, analysis tools and platform at an extremely large scale.

ASTRON, Nikhef and SURF are all involved in ESCAPE, with ASTRON leading the work package that will deliver a prototype science analysis platform that supports data discovery and integration, provides access to software and services, enables user customised processing and workflows, and interfaces with the underlying distributed computing infrastructure. It also seeks to add analytics and visualisation capabilities tuned to the needs of the ESFRI projects. The ESCAPE project started on 1 February 2019 and ends on 31 July 2022. It will therefore overlap with the first 19 months of this Roadmap project (assuming a start on 1 January 2021). This has been incorporated into our plans and it will ensure continuity of the development activities, although the Roadmap project does mark a substantial increase in terms of total effort and provides a focus that will prepare us for the early SKA data from 2023, and the processing of the LHC and KM3NeT data post-2021. The ESCAPE project includes other experiments in physics and astronomy, facilitating future utilisation of the proposed infrastructure by those experiments.

The work plan we intend to carry out as part of this Roadmap project covers the development, implementation, testing and maintenance of pipelines and workflows that can be deployed using container technology (e.g. Singularity or Docker). In preparation for the SKA Regional Centres we will also implement the necessary authorisation, authentication and accounting infrastructure required to manage access to data during the initial proprietary period and to allow the network of SRCs to efficiently optimise and balance the use of its resources. The combined expertise of Nikhef and ASTRON in these areas, as well as the joint technical innovation activities of both institutes with SURF, strengthens the technical feasibility, as is demonstrated by quoting from the 2017 SEP evaluation of Nikhef, which: "*… serves as the provider of the compute and storage infrastructure for the local analysis facility 'Stoomboot' and the associated storage, the Nikhef Data Processing Facility (a node in the Dutch National e-Infrastructure), and the NL/Tier1 together with SURFsara. … The PDP programme also pursues research initiatives on advanced computing technologies, software applications (activity Applied Advanced Computing), and general aspects such as security (activity Infrastructure for Collaboration). The Committee appreciates the chosen initiatives dealing with timely and topical subjects involving future storage architectures, security, software-defined networking, and virtualized platforms. There is excellent collaboration with vendors, which makes it possible to get access to experimental hardware still in the testing phase.*"

Specific Technical Challenges for Science Processing

Over the next 6-8 years, construction of the SKA telescope in Australia and South Africa will take place. This build phase of the project will deliver the antennas, dishes, and electronics required for the collection of data. It also includes hardware and software for the first stage of data processing that occurs close to the telescope, because data rates are simply too high to transport before processing. This first stage of processing produces 600 PB per year of observatory data products that are subsequently exported and distributed to the network of regional centres where individual users and key science teams will be responsible for the second stage data processing, analysis, interpretation and publication. In the steady-state operational phase, towards the end of the next decade, the SRCs will only receive calibrated and quality-controlled data products, but in the commissioning and early science phase (between 2024 and 2027) the SRCs will be the testing ground for the tools and pipelines that produce the Observatory Data products. This proposal covers, in WP1 and WP2, the preparatory work in the SKA Regional Centres to get ready for the time when the 600 PB per year starts to arrive. It does not include any work directly related to the scientific exploitation phase of the SKA project. In particular, *this proposal does not cover any aspect of the research programmes (e.g. Key Science Projects, PI-led projects) that will be carried out with the SKA.* The SRC is the essential platform and infrastructure that is needed to tame the immense data stream from the SKA.

Before detailing the work to be carried out in WP1 and WP2, which will focus on the development of algorithms to support processing and providing access to the data respectively, we summarise the challenges faced by each of the key science areas described earlier.

There is no simple way in which these challenges map onto the work packages: in most cases, they cut across them. The most relevant links to the various work packages are listed in the table below.

Science Case	Challenges	WP
21-cm Cosmology with the SKA		
Code Development	Calibration algorithms/codes need to be scaled to handle SKA-level data in real time. Dynamic calibration needs to be developed, where calibration parameters are data-driven and adjusted in real time, without human intervention. This work is core to the Dutch contribution to such a global effort, and where LOFAR has by far the most knowledge, and advanced tools.	1.1
Data Transport/Storage	Data rates for this Key Science Project from SKA-Low are expected to reach ~40 PB per year by 2028, in all likelihood distributed over a number of countries. Work on data compression will continue, in order to reduce the required storage, aiming for ~100 PB total in a single location for the science extraction phase that will follow.	2.3
Data Processing	To enable real-time data processing, scaling from current LOFAR experience, will require about 10% of the most powerful GPU-based supercomputer today for a period of 5 years. This appears feasible on an 8-10 year timescale given the rate of development of GPU computing in particular.	2.2 DNI
Tracing Galaxy Evolution with Neutral Hydrogen		
Testing pipelines and calibration schemes	Currently defined HI KSP projects will produce just under an Exabyte of data per year. Before exporting to the SRCs baseline-dependent averaging or gridding of the UV data can be used to reduce these data volumes by an order of magnitude or more, but these methods are irreversible and affect the sensitivity and resolution of the final data products. It will be necessary to test the size of the grid, the effects of baseline dependent gridding, combining multi-epoch data grids and implementation of these techniques on HPC, HTC and GPU hardware.	1.1 2.3
Innovative ways to handle large datacubes	A full-resolution HI data cube of a single SKA pointing will measure about 9 TB. This is two orders of magnitude larger than the typical data volumes from current telescopes. Future surveys will observe thousands of pointings, leading to the detection of order 100,000 individual galaxies that need to be identified and characterised. Current software cannot cope with these data volumes and new techniques for source finding, extraction, and analysis will be developed. Visualisation of large data cubes will be developed, including techniques that support remote data and can therefore prevent the need to transport large volumes of data.	1.1 1.3 2.2
Galaxies, black holes and magnetism		
Development of calibration and imaging algorithms	Advanced algorithms to produce high fidelity images have been developed for LOFAR but need further development to deal with SKA observing situations and to ensure algorithms scale favourably to deal with SKA data volumes.	1.1
Data rates	LOFAR data volumes are large but SKA observations will be substantially larger. Existing work on distributed analysis of LOFAR data will be scaled up to the global SRC network. 20 petabytes of LOFAR survey data can be used as a testbed before SKA data becomes available. This advantage places the Dutch community in a leading scientific and technical position for the SKA.	2.2 2.3

Science Case	Challenges	WP
Connecting with other wavelengths and multi-messenger opportunities	To enable effective interpretation of SKA radio data in combination with data from other wavelengths and including multi-messenger data from gravitational wave observatories (Virgo, LIGO, Einstein Telescope), neutrino (KM3NeT) and Cerenkov (CTA) facilities, a robust machinery is needed to automatically classify radio sources, identify their counterparts, derive their properties and facilitate flexible post processing.	2.3
The transient sky		
Development of TraP	The capability of the SRC to host the TraP Database will be developed, enabling a "FAIR" data management plan for SKA transient astronomy. To cope with the increased number of images and sources per image, current limitations of traditionally used databases need to be overcome.	1.1
Access control and VO virtual observatory integration	In order to facilitate use of the information stored within the TraP database, a number of improvements are needed including: access control, more complex interactive query capabilities, and a real-time alerting system.	1.3 2.4
Pulsars		
Separating pulsar signals from human-made interference	Discovering as many new pulsars as possible with SKA is critical to finding the best laboratories for studying gravity and dense matter physics. The challenge is to separate faint astrophysical signals from the much brighter and ubiquitous noise of human-made radio signals. Cutting-edge machine learning (and/or deep learning) are needed to optimally extract pulsar signals from the enormous lists of candidate signals, while also enabling the potential discovery of the unexpected.	1.1 1.2 1.3
Provenance of pulsar timing measurements	Pulsar timing measurements are the basis for the precision gravity and dense matter experiments we aim to undertake with SKA. However, the interpretation of these measurements relies on careful control of systematics and a clear view of their provenance (code versions, profile templates, observatory clock records, etc.). While the measurements themselves are easy to store, careful design is needed to present these data in their full context to scientific users.	2.2 2.3
KM3NeT detection reconstruction		
Extracting rare neutrino signals from ubiquitous backgrounds at KM3NeT	The expected rate of detectable neutrinos in KM3NeT is of the order of up to 10 per hour, while the combined background rate exceeds millions per second. Disentangling these in real-time requires precise calibration and novel state-of-the-art parallel algorithms, with great potential for improvement by integration of machine-learning techniques.	1.1 1.2
Precise reconstruction of neutrino energy and direction	The relation between the direction and energy of incident neutrinos and the photon patterns detected by KM3NeT's optical modules is highly non-linear and requires complex iterative algorithms coupled with a precise understanding of light propagation in sea water. Algorithmic innovation coupled with machine learning promises to directly improve the scientific reach of the facility.	1.2

Science Case	Challenges	WP
High Luminosity data rates at the LHC		
Dealing with complex event and particle track reconstruction	The increased luminosity of the LHC post-2021 leads to very high *multiplicities*, events effectively superimposed on top of each other that have to be disentangled. Improvement of track reconstruction algorithms (both speed-wise as well as in discriminatory power) needs effective use of all available accelerators such as GPUs, and has a great potential to benefit from machine-learning techniques for identification.	1.1 1.2
Managing extreme data volumes	The size of the data sets generated by all LHC experiments requires a reduction in the number of pre-placed data sets at the Tier-1 computing centres globally. In order to maintain a balanced data throughput between storage and compute, both workflow and data distribution have to be re-architectured to take advantage of data lake concepts and of containerisation of workflows.	2.2 2.3

In the remainder of this section, we detail the activities in both work packages.

WP 1: Algorithms - Addressing Compute Challenges through Algorithm Improvement

The ability to effectively use the GPU and Tensor performance improvements depends critically on the development of algorithms in the processing pipelines that can benefit from these specific architectures. Evolving designs in general purpose GPUs and the increase in bandwidth from memory to GPU provide the theoretical capabilities to push data at sufficient speed to the processing cores: PCI-e Gen5 (2019) achieves 32 GT/s, providing 128 GByte/s to a co-processor using 16 lanes. The parallel processing capabilities can be effectively used in the processing of the LHC geometries in event reconstruction, as well as for track finding in KM3NeT data and in imaging and analysis of SKA data. Studies with the LHC tracking trigger system demonstrate the feasibility hereof.

Significant software engineering effort is needed to write (or re-write) algorithms to exploit the GPU capabilities. The advantage is obvious: the greater ability to vectorise code immediately lowers the necessary investment in conventional hardware, and allows exploitation of the much steeper performance increases seen in the GPU and Tensor core field. The development of systems targeting deep-learning domain (such as e.g. the NVidia DGX series) have shown cumulative performance improvements at the same price point of ~ 60% per year. Conventional cores show a much smaller improvement gain (~ 15% linearly), and a very small price improvement as most of the systems cost is determined by its data components (the need for memory and local high-speed data storage). Therefore, the most efficient use of resources is to simultaneously invest in GPU services and in adapting the algorithms for processing the data to fit the hardware types that show the best performance increase.

Addressing the challenges in algorithmic improvements for all use cases, and the construction of access and data processing frameworks are critical to enable the processing of the SKA data in the 2021-2025 timeframe.

WP1.1 ***Design of Data Processing Pipelines***

Each science case has its own specific requirements in terms of algorithms for data processing, and only some currently have GPU-based implementations. Design, development and optimization of GPU-based data processing pipelines for each KSP, as well as for each of the LHC experiments and for KM3NeT, is thus of great importance.

In this work package, pipeline components for each science case will be identified based on the specific requirements. Suitable algorithms for calibration and imaging steps will be selected and where necessary designed or modified. Due to the complexity of some of the calibration algorithms,

machine learning has been identified as a promising way to optimise calibration parameter spaces. The constructed data processing pipeline will be tested in order to make them production-ready.

WP1.2 ***Development of machine learning methods for scientific data quality optimisation***
In order to ensure data quality produced by each of the facilities, auto-detection of instrument failure (i.e., detector elements, electronics) and auto-detection of pipeline failure (i.e., software bugs, execution error) is needed. Expert scientists are now still heavily involved in quality control of the datasets and the processing pipeline. This is not a sustainable way of working, therefore, automation of these processes is needed.

The partners have extensive experience with large-scale quality control of LHC and LOFAR data and this knowledge will be transferred to the other facilities while jointly developing automated workflows based on machine learning. Experts from both science domains will be involved to supply labelled datasets and evaluate the performance of various supervised and unsupervised classes of machine learning models.

WP1.3 ***Visualisation of Massive Radio Astronomy data sets***
Processed radio astronomy datasets are stored in specific data formats. With the increasing data volumes, data will often be distributed across many different locations - usually at some distance from the users/scientist. Suitable tools are needed to allow inspection, interrogation and visualisation of these massive distributed data sets. These tools are not only important for research, but will also help fault finding and can be used to create compelling illustrations to showcase the results achieved with these publically funded research infrastructures.

WP 2: Access - Realisation of the Infrastructure and Access Mechanisms

The technical computing, storage, and network capabilities will be provided *as services* by way of the Dutch National e-Infrastructure (DNI) coordinated by SURF, both using SURF operated resources as well as those of its operational partners including Nikhef. All DNI operational partners have extensive experience operating large-scale science data processing facilities as a service for a broad range of disciplines also outside of the LHC, KM3NeT and SKA domains and extending from biomedical (project MinE), structural chemistry (WeNMR) and many more domains. Both are part of the global e-Infrastructure consortia coordinated through EGI, EUDAT, and GEANT, and are linked to FAIR open data repositories coordinated through OpenAIRE.

The capacity requested here will be provided by participating in and through the national e-Infrastructure, which allows us to benefit from economies of scale and profit from the joint operations and sharing of personnel. It allows us to focus on investments in capacity, while being assured of the long-term sustainable services by SURF and its subsidiaries. The costs of acquiring this capacity as a service at the required quality level was determined to be similar amongst all operational partners in the DNI partners participating in a service.

The compute- and data-services offer access for both interactive and non-web mechanisms through federated authentication and authorization mechanisms, and can be joined with pan-

Figure 27: The WLCG, EGI, and GEANT systems designers from the AARC community at work at Nikhef, designing the access control mechanisms for the next generation AAAI for the LHC that will interwork with the European Open Science Cloud.

European and global research infrastructure services to ensure the data is available to the Dutch researchers as soon as the data are acquired and validated.

WP2.1 ***SRC Liaison and Coordination***

Activities include the overall leadership and technical coordination of the NL SRC activities, the interaction with the SKA Observatory and the other nodes of the SRC network. It also includes project and secretarial support for the NL SKA Regional Centre.

Training and support will already be provided to users of the facility throughout the development phase. The emphasis will be on helping users make the transition to a largely web-based workflow. Assistance will be given to monitor and assess the quality of the data, discover and access data in the archives, and in using the analysis tools that will be available in the data processing interface that will be developed in WP2.4. Training material and documentation will be developed to support the services offered by the regional centres. Training sessions/schools will be organised to bring users closer to the instrument and the data analysis techniques.

WP2.2 ***Software containerisation and Workflow management***

The pipelines will be defined in the chosen language (such as CWL - Common Workflow Language). This means that the behaviour of the components, which form the pipelines, is well described. Maintenance and functional changes to these pipelines can thereafter be made in a controlled way.

The pipelines are packed in containers (such as Docker, Singularity) and are verified against the operational requirements. This can already be done at an early stage as a test environment will be delivered that can be used by the pipeline developers. Tests can initially be done on existing and simulated data. Real data will be used as soon as it becomes available.

A framework for executing the described pipelines will be selected and installed in a production environment. It will use a virtualization management platform such as Kubernetes to make optimal use of the available hardware. An evaluation will be made of the impact on operating cost.

The majority of the work will be describing the existing pipelines currently in use for production followed by the changes that are needed to make them comply with the production requirements. The next phase will be to optimize these pipelines. Areas to address are robustness, i.e. error handling and auto retries, flexibility (avoid processing steps that need days or large numbers of nodes), and readability (documentation and clear code).

Reaching out to the community that develops new (steps in) pipelines will be done in the final stages. Support will be given by training on how to describe and test pipelines as described in WP2.1, and a list of best practices will be given to users through training and support.

WP2.3 ***Data Discovery, Access and Staging***

In collaboration with the International Virtual Observatory Alliance (IVOA), VO standards such as ObsCoreDM (Observation Data Model Core Components) will be extended to support SKA data. SKA metadata will be published and made available according to VO standards, such as the Table Access Protocol (TAP) and the catalogue placed in a VO register. Python libraries will be developed for the extended VO standard using existing VO libraries and tools. For data residing on nearline media, and managed through storage middleware such as dCache, a staging service will be developed and deployed which will manage and control retrieval of data ensuring a reliable and predictable behaviour to enable integration with automated data processing workflows. API hooks will be provided for allowing external services to subscribe to relevant events and to query the state of activity. The services will build on data lake technology developed in the ESCAPE project. A Science Data Repository service providing low latency access to popular science data products will be deployed. The data access services will be integrated with federated AAAI, compliant with applicable data access policies.

WP2.4 ***User Interface and Authentication, authorisation & accounting***

A web-based user interface will be built and deployed enabling users to interact with the data services described in section 2.3, providing a human interface for data discovery and retrieval, as well as with the workflow management and execution services described in WP 2.2. Users will be empowered to run data analysis workflows on connected HPC/HTC compute infrastructure through either direct submission to a batch scheduler or through a workload management framework such as DIRAC. They will also be able to experiment with their own data analysis workflows in an interactive working environment such as Jupyter Notebooks. Through collaboration between partners in ESCAPE and in FuSE, KM3NeT's developing data processing capabilities will similarly benefit from this work.

An authentication, authorisation & accounting infrastructure (AAAI) solution will be selected in conjunction with the SKA Observatory and SRC network. Implementation includes setting up and hosting a service proxy and an appropriate collaboration management service that administrates and provides role, group, and resource allocation information in a secure and reliable manner. Based on the information managed through the AAAI solution, services can enforce authorization policies. An operational service-hosting environment will be prepared to which developed and selected services are deployed.

2.4.2 Risk analysis

The project partners ASTRON and Nikhef each have more than a decade of experience in large-scale, high-performance distributed data-intensive computing, as has been illustrated in other sections (*e.g.,* the Strategic Case) of this proposal. This track record (along with that of the international partners) gives every reason to be confident in the ability to realise FuSE.

No further details are provided in this public version of the proposal.

2.4.3 Financial feasibility

This section should be read as a detailed explanation of the multiyear budget sheet that is an integral part of the proposal submission package. We have several important remarks regarding the budget.

The first remark is that we have chosen to limit the project budget to *a five-year period* (2021 – 2025). The main reason is that five years is already a long time in projecting costs and capabilities in high performance computing and ICT in general. It is our experience from earlier granted proposals on computing infrastructure, that this reasoning is also applied by the granting organisation NWO – we have already mentioned that we had to limit our computing budget (as part of 2013 LHC detector upgrade Roadmap project) to five years (2014 – 2019). We have also provided a total 10 year budget (2021 – 2030), in which we have extrapolated the costs using the best available knowledge (including the commitments from the contributing organizations), but we stress that the budget numbers for *the 2026-2030 period* depend strongly on the assumptions made about both the technological and economic developments of the ICT hardware, and they will also depend on whether algorithmic advances are made by the scientific groups. In the 10-year budget we have also inserted a best estimate of the continued in-kind involvement.

The second remark is that we have chosen to categorize all costs as part of the *capital investment* of the proposed infrastructure, which in itself should be defined as *IT costs*. The reason is that all contributions ('hardware' and 'peopleware') can be considered as necessary to build up the infrastructure of the required size and capabilities. However, for the 'hardware' components we have used figures representing a Total Cost of Ownership (*TCO*) approach, which includes the running cost of the hardware. This is a more mature approach of costing ICT infrastructure than the sometimes-artificial division between capital investment and running cost.

Thirdly, regarding *personnel costs*, we have used the obligatory 'VSNU' rates per FTE for most of the personnel that is requested from the NWO contribution, except for senior software developers from whom we know that these are in high demand on the labour market and whose annual personnel costs

are on average in €. For the in-kind contribution from Nikhef and ASTRON we have used the actual average personnel rates per FTE of the staff that is planned to be involved in the project.

Finally, we have used the 'euro' as budgetary unit, because the spreadsheet requires this. However, most of the budget estimates cannot be more precise than on the level of 1.000 € (1 k€).

Investments in ICT services: 'hardware'

The investment expenses follow from a detailed calculation based on the following ingredients:

- Computing, storage, and global connectivity needs from SKA, LHC (ATLAS, ALICE, LHCb), and KM3NeT over the periods 2021-2025 and projected out for the period 2026-2030
- Models of technological developments of the hardware; in other words, how much faster a computer processor becomes over the next ten years
- Models of economic developments: how the price of processors, tapes, and fast storage (now disk) develops over the period.

The cost basis for the compute, on-line and near-line storage, and the network, has been coordinated closely with SURF, and all these services will be acquired in the context of the Dutch national e-Infrastructure (DNI). The costing of the compute service has been evaluated by SURF and – in the context of the EU Helix Nebula Science Cloud (HNSciCloud) project (https://www.helix-nebula.eu), cross-correlated with similar TCO service calculations for Nikhef and international partners. The results of these studies are unequivocal: with the required service properties for science data processing (extremely high global bandwidth needs, necessity to ingest and extract data from the facility, and basically 99+% occupancy of all resources), providing this service at SURF and for the service costing agreed with SURF is more than a factor two cheaper than procuring such capacities from commercial cloud providers. In addition, experience within the HNSciCloud project in the use of public clouds for data intensive processing, even with providers intent on working with the science community, have shown that such resources are extremely people-intensive to put to use and are neither efficient nor effective to process at the Exabyte scale.

For the infrastructure foreseen, we rely on the compute and storage services of the DNI. Their reference price was determined in 2019, and the unit price evolution derived by extrapolating historic (2013-2018) trends for processing (in price per core-hour, assuming the service offers cores with 'current-generation' performance), for high-throughput storage and for near-line storage (in price per terabyte per year), and network (managed virtual private circuits over shared dark fibre links). Following the SURF service delivery model, the cost for operating the equipment is accommodated as part of the service (i.e. the service price per unit includes the cost for power, cooling, physical installation, and operating the hardware as a managed platform to deliver the service at the required service level). The cost model is shown in the table below.

Year	Compute (€/core-hr)	High throughput storage (€/TB/year)	Near-line storage (€/TB/year)
2019	[redacted]		
2021			
2025			
Applied cumulative improvement factor	10% annual improvement*	20% annual improvement	flat *(cost per cartridge is constant)*
** the cost improvement for the compute service comes not from per-core application performance improvements (such have been taken into account in the science resource requirements, but from the increasing number of cores per system, allowing the fixed system overhead cost to be amortised over a larger number of cores.*			

The costs for the global network interconnects also decreases over time, e.g. the cost for the dual (20 Gbps) links from Amsterdam to CERN (Geneva, CH) for the LHCOPN and LHCOne. However, with the need for much higher throughput in and beyond 2022 (and even more after 2024 with increased data volumes from the SKA), the requisite bandwidth goes up, and thus the expected cost.

Year	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Network (*k€/year)*										

The quantitative data processing needs for the combined infrastructures (LHC, KM3NeT, SKA) as described in the technical sections, using the service cost models aligned with SURF, therefore results in the budget requirements to construct the infrastructure as given in the table below, with the period 2021-2025, for which resources are requested in this proposal) highlighted in both the table and the resulting sums (all in k€, either per year or for the requested period, respectively).

The entire facility needs to be complemented by a 'tier-3' analysis cluster that is used to calculate the published open data, i.e. the work done by the scientist in the (creative and thus chaotic) final analysis phase. The capacity for this chaotic analysis for the LHC and KM3NeT data is provided in-kind by Nikhef through its local analysis facility – which is managed in the same way as Nikhef contributes services to the DNI coordinated by SURF and is integrated in the same fabric. The capacity in terms of core-hours and terabytes follows the same cost model constants (TCO) as the national e-Infrastructure.

SURF, building on the infrastructure developed in the BiG Grid and national e-Infrastructure ICT scheme, also contributes resources (CPU, disk, tape, and network) to the LHC data processing. The (constant) contribution of SURF to the LHC NL-T1 service forms part of the funding for this infrastructure.

'Peopleware': Data and Computing Engineering for Algorithms and Access

In section 2.4.1. on technical development, a detailed breakdown of the necessary effort has already been provided (although not in this public version). The FTEs (per year and in total) in the multi-year spreadsheet are consistent with this breakdown.

Investment Summary

Collating the above detailed breakdown in the form of an abstracted budget table, the data processing infrastructure proposed comprises:

Costs (2021 – 2025)	
Total cost of the Infrastructure	**28.812.753 €**

Funding (2021 – 2025)	
Requested NWO contribution	*11.948.697 €*
Total funding	**28.812.753 €**

2.5 Literature references

[1] "Combined measurements of Higgs boson production and decay using up to 80 fb−1 of proton-proton collision data at √s=13 TeV collected with the ATLAS experiment", ATLAS Collaboration, ATLAS-CONF-2019-005

[2] "SUSY March 2019 Summary Plot Update", ATLAS Collaboration, ATLAS-PHYS-PUB-2019-012

[3] "Measurement prospects of the pair production and self-coupling of the Higgs boson with the ATLAS experiment at the HL-LHC", ATLAS Collaboration, ATLAS-PUB-2018-053

[4] "Prospects for searches for status, charginos and neutralinos at the high luminosity LHC with the ATLAS Detector", ATLAS Collaboration, ATLAS-PUB-2018-048

[5] Patrick Huet and Eric Sather, "Electroweak Baryogenesis and Standard Model CP Violation", Phys.Rev.D51:379-394, 1995; DOI:10.1103/PhysRevD.51.379

[6] LHCb collab., "First Observation of CP Violation in the Decay of Bs mesons", Phys.Rev.Lett.110, 2010, 221601

[7] LHCb collab., "Observation of CP violation in charm decay ", 2019, arXiv:1903.08726

[8] LHCb collab., "Measurements of matter - antimatter differences in beauty baryon decays", Nature Physics, v13 (2017)

[9] K. Aamodt *et al.* (ALICE Collab.), "The ALICE experiment at the CERN LHC", JINST 3 (2008).

[10] E. V. Shuryak, "Theory and phenomenology of the QCD vacuum", Phys. Rept. 115 (1984); J. Cleymans, R. V. Gavai, and E. Suhonen, "Quarks and Gluons at High Temperatures and Densities", Phys. Rept. 130 (1986); S. A. Bass, M. Gyulassy, H. Stoecker, and W. Greiner, "Signatures of quark gluon plasma formation in high-energy heavy ion collisions: A Critical review", J. Phys. G25 (1999) , arXiv:hep-ph/9810281 [hep-ph].

[11] A. Bzdak and V. Skokov, "Event-by-event fluctuations of magnetic and electric fields in heavy ion collisions", Phys. Lett. B710 (2012), arXiv:1111.1949 [hep-ph]

[12] N.Brambilla *et al.*,Eur.Phys.J.C74(2014)no.10,2981; doi:10.1140/epjc/s10052-014-2981-5 [arXiv:1404.3723 [hep-ph]]

[13] Aartsen; et al. (The IceCube Collaboration, Fermi-LAT, MAGIC, AGILE, ASAS-SN, HAWC, H.E.S.S., INTEGRAL, Kanata, Kiso, Kapteyn, Liverpool Telescope, Subaru, Swift/NuSTAR, VERITAS, VLA/17B-403 teams) (12 July 2018). "Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A". Science. 361 (6398): eaat1378. arXiv:1807.08816

[14] Madau, P., Meiksin, A., & Rees, M. J. (1997). 21 Centimeter Tomography of the Intergalactic Medium at High Redshift. The Astrophysical Journal, 475(2), 429–444.

[15] van Haarlem, M. P., Wise, M. W., Gunst, A., Heald, G., McKean, J. P., Hessels, J. W. T., et al. (2013). LOFAR: The LOw-Frequency ARray. Astronomy and Astrophysics, 556, A2. http://doi.org/10.1051/0004-6361/201220873

[16] Tingay, S., & Collaboration, M. W. A. (2014). The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies. Exascale Radio Astronomy, 3.

[17] Parsons, A. R., Backer, D. C., Foster, G. S., Wright, M. C. H., Bradley, R. F., Gugliucci, N. E., et al. (2010). The Precision Array for Probing the Epoch of Re-ionization: Eight Station Results. The Astr. J., 139(4), 1468–1480. http://doi.org/10.1088/0004-6256/139/4/1468

[18] Mellema, G., Koopmans, L. V. E., Abdalla, F. A., Bernardi, G., Ciardi, B., Daiboo, S., et al. (2013). Reionization and the Cosmic Dawn with the Square Kilometre Array. Experimental Astronomy, 36(1), 235–318. http://doi.org/10.1007/s10686-013-9334-5

[19] DeBoer, D., Bowman, J. D., Jacobs, D., Parsons, A., Liu, A., Werthimer, D., et al. (2014). HERA: Chasing Our Cosmic Dawn. Exascale Radio Astronomy, 10304.

[20] Morganti, R., Sadler, E.M., Curran, S. 2015. Cool Outflows and HI absorbers with SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 134.

[21] Putman, M.E., Peek, J.E.G., Joung, M.R. 2012. Gaseous Galaxy Halos. Annual Review of Astronomy and Astrophysics 50, 491-529.

[22] Fernández, X., *et al.* 2016. Highest Redshift Image of Neutral Hydrogen in Emission: A CHILES Detection of a Starbursting Galaxy at z = 0.376. Astrophysical J. 824, L1.

[23] Staveley-Smith, L., Oosterloo, T. 2015. HI Science with the Square Kilometre Array. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 167.

[24] Blyth, S., *et al.* 2015. Exploring Neutral Hydrogen and Galaxy Evolution with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 128.

[25] Afonso, J., Casanellas, J., Prandoni, I., Jarvis, M., Lorenzoni, S., Magliocchetti, M., Seymour, N. 2015. Identifying the first generation of radio powerful AGN in the Universe with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 71.

[26] Saxena, A., *et al.* 2018. Discovery of a radio galaxy at z = 5.72. Monthly Notices of the Royal Astronomical Society 480, 2733-2742.

[27] McKean, J., Jackson, N., Vegetti, S., Rybak, M., Serjeant, S., Koopmans, L.V.E., Metcalf, R.B., Fassnacht, C., Marshall, P.J., Pandey-Pommier, M. 2015. Strong Gravitational Lensing with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 84.

[28] Abbott, B.P., *et al.* 2016. Astrophysical Implications of the Binary Black-hole Merger GW150914. The Astrophysical Journal 818, L22.

[29] Abbott, B.P., *et al.* 2017. GW170608: Observation of a 19 Solar-mass Binary Black Hole Coalescence. The Astrophysical Journal 851, L35.

[30] Aartsen, M.G., *et al.* 2018. Astrophysical neutrinos and cosmic rays observed by IceCube. Advances in Space Research 62, 2902-2930.

[31] Lorimer, D.R., Bailes, M., McLaughlin, M.A., Narkevic, D.J., Crawford, F. 2007. A Bright Millisecond Radio Burst of Extragalactic Origin.Science 318, 777.

[32] Petroff, E., Hessels, J.W.T., Lorimer, D.R. 2019. Fast Radio Bursts. arXiv:1904.07947.

[33] Maan, Y., Bassa, C., van Leeuwen, J., Krishnakumar, M.A., Joshi, B.C. 2018. A Search for Pulsars in Steep Spectrum Radio Sources. The Astrophysical Journal 864, 16.

[34] Kramer, M., *et al.* 2006. Tests of General Relativity from Timing the Double Pulsar. Science 314, 97-102.

[35] Archibald, A.M., Gusinskaia, N.V., Hessels, J.W.T., Deller, A.T., Kaplan, D.L., Lorimer, D.R., Lynch, R.S., Ransom, S.M., Stairs, I.H. 2018. Universality of free fall from the orbital motion of a pulsar in a stellar triple system. Nature 559, 73-76.

[36] Demorest, P.B., Pennucci, T., Ransom, S.M., Roberts, M.S.E., Hessels, J.W.T. 2010. A two-solar-mass neutron star measured using Shapiro delay. Nature 467, 1081-1083.

[37] Antoniadis, J., *et al.* 2013. A Massive Pulsar in a Compact Relativistic Binary. Science 340, 448.

[38] Keane, E., *et al.* 2015.\ A Cosmic Census of Radio Pulsars with the SKA.\ Advancing Astrophysics with the Square Kilometre Array (AASKA14) 40.
[39] Watts, A., *et al.* 2015. Probing the neutron star interior and the Equation of State of cold dense matter with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 43.
[40] Hessels, J., Possenti, A., Bailes, M., Bassa, C., Freire, P.C.C., Lorimer, D.R., Lynch, R., Ransom, S.M., Stairs, I.H. 2015. Pulsars in Globular Clusters with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 47.
[41] Janssen, G., *et al.* 2015. Gravitational Wave Astronomy with the SKA. Advancing Astrophysics with the Square Kilometre Array (AASKA14) 37.

2.6 Other relevant information

Key Researcher Publications

Dr. R. Aaij

R. Aaij *et al*. "A comprehensive real-time analysis model at the LHCb experiment," JINST, vol. 14, no. 4, p. p04006, 2019; https://doi.org/10.1088/1748-0221/14/04/P04006

R. Aaij *et al*. "Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC", JINST, vol. 14, no. 4, p. 04013, 2019; https://doi.org/10.1088/1748-0221/14/04/P04013

R. Aaij *et al*. "Selection and processing of calibration samples to measure the particle identification performance of the LHCb experiment in Run 2", EPJ Tech. Instrum., vol. 6, no. 1. 2019; https://doi.org/10.1140/epjti/s40485-019-0050-z"

Dr. P. Christakoglou

B. Abelev *et al.* [ALICE Collaboration] "Charge separation relative to the reaction plane in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV", Phys. Rev. Lett. 110, (2013) 012301

N. Brambilla *et al.* "QCD and Strongly Coupled Gauge Theories: Challenges and Perspectives", Eur. Phys. J. C74, (2014) 2981

B. Abelev *et al.* [ALICE Collaboration] "Elliptic flow of identified particles in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV", JHEP 1506, (2015) 190

Prof. dr. W. J. G. de Blok

W.J.G. de Blok *et al.* "An Overview of the MHONGOOSE Survey: Observing Nearby Galaxies with MeerKAT. Proceedings of MeerKAT Science: On the Pathway to the SKA". 25-27 May, 2016 Stellenbosch, South Africa (MeerKAT2016). Online at https://pos.sissa.it/277/007/

W.J.G. de Blok, F. Fraternali, G. Heald, B. Adams, A. Bosma, B. Koribalski "The SKA view of the Neutral Interstellar Medium in Galaxies. Advancing Astrophysics with the Square Kilometre Array" (2015) (AASKA14) 129.

F. Walter, E. Brinks, W.J.G. de Blok, F. Bigiel, R.C. Kennicutt Jr., M.D. Thornley, A. Leroy, "THINGS: The H I Nearby Galaxy Survey" (2008) The Astronomical Journal 136, 2563-2647.

Dr. D. L. Groep

C. Atherton *et al.* (D.L. Groep) "Federated Identity Management for Research Collaborations" June 2018; https://doi.org/10.5281/zenodo.1296031

D.P. Kelsey *et al.* (D.L. Groep) "Can R&E federations trust Research Infrastructures? - The Snctfi Trust Framework"; PoS vol. 293 (ISGC2017) 024; https://doi.org/10.22323/1.293.0024

D.L. Groep and D. Bonacorsi "20th International Conference on Computing in High Energy and Nuclear Physics"; J. of Phys. C.S. vol. 513 (2013) 001001; http://dx.doi.org/10.1088/1742-6596/513/0/001001

Dr. A. Heijboer

A. Abulencia *et al.* [CDF Collab] "Measurement of the B_s-B_s oscillation frequency"; Phys. Rev. Lett. 97, 062003 (2006)

A. Abulencia *et al.* [CDF Collab] "Observation of B_s-B_s oscillations" Phys.Rev.Lett. 97,242003 (2006)

S. Adrian-Martinez *et al.* [Antares collaboration] "First search for point sources of high energy cosmic neutrinos with the ANTARES neutrino telescope"; Astrophysical J., 743:L14, 2011

Dr. J. W. T. Hessels

J.W.T. Hessels, L.G. Spitler, A.D. Seymour *et al.* "FRB 121102 Bursts Show Complex Time-Frequency Structure", Astrohysical Journal Letters 876, L23 (2019)

A.M. Archibald, N.V. Gusinskaia, J.W.T. Hessels *et al.* "Universality of free fall from the orbital motion of a pulsar in a stellar triple system", Nature 559, 73 (2018)

D. Michilli, A. Seymour, J.W.T. Hessels *et al.* "An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102", Nature 553, 182 (2018)

Prof. dr. C. A. Jackson

Jackson, C.A. and Wall, J.V. "Extragalactic radio-source evolution under the dual-population unification scheme", MNRAS, 304, 160 (1999)

Jackson C.A., Wall J.V. *et al.* "The Parkes quarter-Jansky flat-spectrum sample: I. Sample selection and source identifications" A&A, 386, 97 (2002)

Shimwell, T.W. *et al.* "The LOFAR Two-metre Sky Survey. II. First data release" A&A 622,1 (2019)

Prof. dr. L.V.E. Koopmans

Patil A.H., Yatawatta, S., Koopmans, L.V.E. *et al.* "Upper Limits on the 21 cm Epoch of Reionization Power Spectrum from One Night with LOFAR", ApJ 838, 65 (2017)

Koopmans, L.V..E. *et al.* "The Cosmic Dawn and Epoch of Reionisation with SKA" in Proceedings of Advancing Astrophysics with the Square Kilometre Array (AASKA14) (2015)

Mellema, G., Koopmans, L.V.E., *et al.* "Reionization and the Cosmic Dawn with the Square Kilometre Array" Experimental Astronomy, Volume 36, Issue 1-2, pp. 235-318 (2013)

Prof. dr. M. Merk

LHCb Collab. "Search for lepton-universality violation in $B^+ \to K^+l^+l^-$ decays", Mar 21, 2019, DOI: 10.1103/PhysRevLett.122.191801

LHCb Collab. "Measurement of CP asymmetry in $B_s \to D_s^{\pm/\pm} K^{\pm}$ decays, Dec 20 2017, DOI: 10.1007/JHEP03(2018)059

K. De Bruyn, R. Fleischer, R. Knegjens, P. Koppenburg, M. Merk, N. Tuning, "Probing New Physics via the $B_s \to \mu^+\mu^-$ Effective Lifetime", 2012, DOI: 10.1103/PhysRevLett.109.041801

Prof. dr. R. Morganti

Morganti, R., Fogasy, J., Paragi, Z., Oosterloo, T., Orienti, M. Radio Jets Clearing the Way Through a Galaxy: Watching Feedback in Action. Science 341, 1082-1085 (2013)

Morganti, R., Oosterloo, T. The interstellar and circumnuclear medium of active nuclei traced by HI 21 cm absorption. Astronomy and Astrophysics Review 26, 4 (2018)

Morganti, R. Archaeology of active galaxies across the electromagnetic spectrum. Nature Astronomy 1, 596-605 (2017)

Prof. dr. G. Raven

J. Albrecht, C. Fitzpatrick, V. Gligorov, G. Raven, "The upgrade of the LHCb trigger system," 10.1088/1748-0221/9/10/C10026

G. Raven *et al.* "HEP Community White Paper on Software trigger and event reconstruction," arXiv:1802.08638

G. Raven *et al.* "A Roadmap for HEP Software and Computing R&D for the 2020s," arXiv:1712.06982

Dr. A. Rowlinson

Rowlinson, A. *et al.* "Identifying transient and variable sources in radio images" Astronomy & Computing, Vol. 27, Art. 111 (2019)

Rowlinson *et al.* "Limits on Fast Radio Bursts and other transient sources at 182 MHz using the Murchison Widefield Array" MNRAS 458, 3506 (2016)

Swinbank, J.D. *et al.*"The LOFAR Transients Pipeline" Astronomy & Computing, Vol. 11, 25 (2015).

Dr. T. W. Shimwell

Shimwell, T.W. *et al.* "The LOFAR Two-metre Sky Survey. II. First data release" A&A 622, 1 (2019)

Shimwell, T.W. *et al.* "The LOFAR Two-metre Sky Survey. I. Survey description and preliminary data release" A&A 598, 104 (2017)

van Weeren, R.J., "A plethora of diffuse steep spectrum radio sources in Abell 2034 revealed by LOFAR" ApJS 223, 2 (2016)

Dr. J. A. Templon

Albrecht, J. *et al.* (J.A. Templon) "A Roadmap for HEP Software and Computing R&D for the 2020s" Comput. Softw. Big Sci. (2019) 3, 7, HSF-CWP-2017-01, doi: 10.1007/s41781-018-0018-8

Ramenska, D. *et al.* (J.A. Templon) "Using model checking to analyze the system behavior of the LHC production grid" Future. Gen. Comp. Sys. 29(8): 2239-2251 (2013), doi: 10.1016/j.future.2013.06.004

Templon, J.A. *et al.* "Scheduling multicore workload on shared multipurpose clusters" J. of Phys. C. S. 664-5 (2015) 052038, doi: 10.1088/1742-6596/664/5/052038

Dr. M. P. van Haarlem

van Haarlem, M.P. *et al.* "LOFAR: The LOw-Frequency Array" A&A, 556, A2 (2013), doi: 10.1051/0004-6361/201220873

van Haarlem, M.P. "LOFAR: The Low Frequency Array" in "Radio Astronomy at 70: from Karl Jansky to microjansky", eds. Gurvits L.I., Frey S., Rawlings S. (eds), EDP Sciences (2005), doi: 10.1051/eas:2005169

van Haarlem, M.P., Frenk, C.S., White, S.D.M., "Projection effects in cluster catalogues", MNRAS, 287, 817 (1997) doi:10.1093/mnras/287.4.817

Prof. dr. W. Verkerke

G. Aad *et al.* (ATLAS Collaboration) "The ATLAS Simulation Infrastructure"; Eur. Phys. J. C 70 (2010) 823; doi:10.1140/epjc/s10052-010-1429-9

L. Moneta *et al.* "The RooStats Project"; PoS ACAT 2010 (2010) 057; doi:10.22323/1.093.0057

G. Aad *et al.* (ATLAS Collaboration) "Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s}$=7 and 8 TeV in the ATLAS experiment"; Eur. Phys. J. C 76 (2016) no. 1, 6; doi:10.1140/epjc/s10052-015-3769-y

3. Declaration and signature (by coordinating applicant)

3.1 Have you requested funding for this research infrastructure elsewhere?

☒ No

3.2 Statements by the applicant

☐ I endorse and follow the Code Openness Animal Experiments (if applicable).

☐ I endorse and follow the Code Biosecurity (if applicable).

☒ By submitting this document I declare that I satisfy the nationally and internationally accepted standards for scientific conduct as stated in the Netherlands Code of Conduct for Scientific Practice 2012 (Association of Universities in the Netherlands (VSNU)).

☒ The consortium partners are aware of the NWO Grant Rules and obligatory establishment of an agreement containing IP&P arrangements and will adhere to this if the proposal is awarded.

☒ I hereby declare that the obligatory letters of commitment of the consortium partners have been uploaded separately in ISAAC.

☒ I have completed this form truthfully.

[signature omitted in this public version]

Name: prof. dr. S.C.M. (Stan) Bentvelsen

Place: Amsterdam, The Netherlands

Date: Tuesday, June 4th, 2019