Scientific Applications and Challenges of SKA Big Data: Postprint

Authors: An Tao, Wu Xiangping, Hong Xiaoyu, Ye Shuhua, Mao Yufeng, Guo Shaoguang, Lao Baoqiang

Date: 2023-03-19T00:00:00+00:00

Abstract

The Square Kilometre Array (SKA) radio telescope, soon to be constructed, constitutes the largest astronomical observational facility, expected to achieve revolutionary breakthroughs in major frontier issues of natural science, including the origin of the universe, origin of life, origin of cosmic magnetic fields, nature of gravity, and extraterrestrial civilizations. The SKA's technical characteristics—namely super sensitivity, ultra-wide field of view, ultra-fast survey speed, and ultra-high temporal, spatial, and frequency resolution—ensure its leading position in observational capabilities, thereby generating massive observational datasets. The SKA's data transport, storage, reading/writing, processing, management, archiving, and distribution pose severe challenges to cutting-edge technologies in information and computer science. The Chinese SKA science team will collaborate with the information industry to address the challenges of SKA big data, not only to promote major original scientific discoveries but also to apply its technological achievements to national economic development.

Full Text

Subject and Field: ChinaXiv Partner Journal—Scientific Applications and Challenges of Big Data

Affiliations: Shanghai Astronomical Observatory, Chinese Academy of Sciences; National Astronomical Observatories, Chinese Academy of Sciences; Chinese Academy of Sciences; Bureau of Frontier Sciences and Education

Abstract

The Square Kilometre Array (SKA) radio telescope, soon to be constructed, represents the largest astronomical observational facility ever built. It is poised to achieve revolutionary breakthroughs at the major frontiers of natural science, addressing fundamental questions about the origin of the Universe, the origin of life, the origin of cosmic magnetic fields, the nature of gravity, and the search for extraterrestrial civilizations. The SKA's unprecedented technical capabilities—extremely high sensitivity, ultra-wide field of view, exceptionally fast survey speed, and super-high temporal, spatial, and frequency resolution—ensure its leading position in radio astronomy for decades to come, while simultaneously generating massive volumes of observational data. The transportation, storage, reading, writing, processing, management, archiving, and dissemination of SKA-level data pose severe challenges to cutting-edge technologies in information and computer science. The Chinese SKA science team will collaborate with the information industry to tackle these big data challenges, not only to promote major original scientific discoveries but also to apply the resulting technological achievements to national economic development.

Keywords: Square Kilometre Array (SKA), big data, high-performance computing, scientific applications
DOI: 10.16418/j.issn.1000-3045.2018.08.016

Introduction

Astronomy is one of the oldest disciplines, emerging alongside human civilization, and China is among the earliest countries in the world to develop astronomical traditions. Modern observational astronomy, dating from Galileo's invention of the astronomical telescope, has a history of over 400 years, and every major advance in astronomy has depended on leapfrog progress in telescope capabilities. China is currently in a strategic period of scientific and technological innovation in the new era, with unprecedented national investment in scientific research. Observing the stars requires precision telescopes, and in recent years, a number of large astronomical telescopes have been built successively in China, such as the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) in Xinglong, the Five-hundred-meter Aperture Spherical radio Telescope (FAST) in Guizhou, the dark matter detection space telescope "Wukong," and the hard X-ray modulation telescope "Insight." These facilities approach or reach world-class standards.

China has participated in the world's largest astronomical scientific project—the Square Kilometre Array (SKA) radio telescope international collaboration—which upon completion will become the flagship of radio telescopes and establish a new milestone in the history of natural science exploration. Existing telescope facilities are also being upgraded, forming observation capabilities spanning ground-based instruments to space satellites (and space stations), and covering

the full electromagnetic spectrum from X-ray, ultraviolet, optical, and infrared to radio wavelengths. This has propelled astronomical research into an era of exponentially growing big data. Current astronomical data has already reached the petabyte (PB) scale, and with advances in observation technology and equipment updates, it will soon enter the exabyte (EB) scale era. Astronomical big data will profoundly change the way humans explore and understand nature.

The Era of Astronomical Big Data

Since the 1960s, astronomy has continuously produced remarkable achievements, writing brilliant chapters in the development of human natural science. The most exciting and groundbreaking discoveries increasingly rely on the coordinated operation of large-scale research facilities and the analysis and mining of massive datasets. Meanwhile, the transparency, diversity, and interdisciplinary integration of scientific results have enriched human technological life. Astronomy has truly entered a multi-wavelength, multi-messenger era. Not only can multiple observational devices be used simultaneously to detect the same celestial object, obtaining nearly complete information across the entire electromagnetic spectrum, but also information carriers beyond electromagnetic radiation—such as neutrinos and gravitational waves—can be employed to study cosmic objects. A most representative example is the first detection of a binary neutron star merger in August 2017. The ground-based Laser Interferometer Gravitational-Wave Observatory (LIGO) and VIRGO gravitational wave detectors first discovered the spacetime ripples produced by the neutron star merger, after which the most powerful space and ground-based telescopes coordinated observations of the subsequent radiation. This not only enhanced our understanding of gravitational waves but also observationally confirmed exotic phenomena such as short gamma-ray bursts and hypernovae, giving us new appreciation for the formidable power of collaborative astronomical research.

Observation-based astronomy long suffered from data scarcity, but entering the 21st-century information age, astronomy has undergone a major revolutionary transformation. Astronomical observations have gradually entered the big data era, and current scientific research and communication methods are undergoing profound evolution. For example, supernovae are magnificent fireworks in the Universe, and China holds world-recognized records for the earliest astronomical observations of supernovae. Supernovae hold important positions in astrophysics research—the 2011 Nobel Prize in Physics was awarded to three astronomers for discovering the accelerating expansion of the Universe through supernova observations. Supernovae are extremely rare events; capturing one was quite difficult a decade ago, and each detection inevitably triggered a global telescope chase. Much research had to rely on numerical simulations and theoretical calculations. Today, however, optical surveys discover over 1,000 supernovae annually, making them commonplace. Deep and effective mining of data accumulated from these large surveys is likely to yield more new discov-

eries. With the tremendous enhancement of observational capabilities brought by next-generation super telescopes like SKA, exotic celestial objects that are currently rare will become routine within 5–10 years. Statistics and information science, closely integrated with astronomy, provide astronomers with data analysis tools to explore the truths of the macroscopic Universe and the laws of celestial motion based on the collection, organization, and analysis of cosmic big data.

SKA: The Next-Generation Radio Telescope

Driven by grand scientific objectives, the SKA radio telescope represents the largest international scientific collaboration project in astronomy that China has joined. Upon completion, SKA will become the world's largest astronomical experimental facility, creating new opportunities for humanity to explore the mysteries of cosmic origins. SKA involves 11 full member countries including China, plus more than 10 observer countries. The construction and operation of large astronomical telescopes have become a true reflection of national comprehensive strength and an important symbol of it. SKA headquarters are located in the United Kingdom. The SKA low-frequency array (SKA-low) comprises 1.3 million log-periodic antennas to be built in the Western Australian desert, while the SKA mid-frequency array (SKA-mid) comprises 2,500 dish antennas to be built in South Africa and the radio-quiet regions of southern Africa—both sites selected as optimal locations after more than a decade of evaluation and assessment by astronomers. The total collecting area of the telescope reaches one square kilometer, with frequency coverage spanning nearly continuously from 50 MHz to 20 GHz. This provides approximately 50 times greater sensitivity and about 10,000 times faster survey speed than the largest existing centimeter-wave radio telescope arrays [2].

As a next-generation leading radio astronomical observational facility, SKA will have a profound impact on the development of radio astronomy. SKA's powerful observational capabilities are manifested in its ultra-high sensitivity (millikelvin level), ultra-wide field of view (tens of degrees), ultra-fast survey speed, ultra-high frequency resolution (kilohertz), ultra-high time resolution (nanoseconds), and ultra-high spatial resolution (sub-arcsecond). These technical characteristics enable SKA to generate unprecedented volumes of big data [2].

SKA Construction Phases and Data Volume

SKA construction is divided into two main phases: Phase 1 (SKA1) will be built to 10% of the full scale, with construction expected to begin in 2020; Phase 2 (SKA2) will complete the remaining 90% of construction, though specific plans have not yet been determined. SKA1-low generates data at 2 Tbps per station, with a total data flow of 1 Pb/s. Extrapolating from this scale, SKA2 will produce at least 10 times the real-time data flow. As can be seen from these

figures, SKA's data volume is unprecedented. Even after significant reduction through correlation processing, the data input to the Science Data Processor (SDP) reaches 4 GB/s, making it truly scientific big data. SKA's massive data streams must be processed in real-time; otherwise, the entire data processing pipeline will become blocked or even collapse. The use of real-time mode, multi-concurrent tasks, and dataflow pipeline systems represents several typical characteristics of SKA data processing and is a novel application of scientific big data processing [3].

As the largest radio telescope in history, SKA not only carries the mission of nurturing world-class scientific achievements but also produces the world's largest data volume. We must therefore fully recognize the enormous challenges of SKA data processing. Due to the extreme scale and complexity of the SKA project, several pathfinder and precursor projects have been built by multiple countries including China, each representing about 1% of the total SKA scale. These have conducted relevant scientific pre-research and technical development. These precursor facilities have played positive roles in understanding SKA scientific objectives, establishing and gradually improving sky models, developing and testing data processing software, and cultivating urgently needed talent—occupying a non-negligible position in SKA's development history. It should be noted, however, that these precursor projects' data volumes fall far short of SKA1 scale, leaving a certain distance from establishing a true validation reference .

Taking the SKA precursor project MWA (Murchison Widefield Array) as an example, after four years of operation, MWA has accumulated 24 PB of scientific archive data. One of its scientific objectives is the GLEAM survey, whose first phase already contains more than 300,000 galaxies with archived data exceeding 1 PB. The second phase has begun with sensitivity improved by more than 4 times, expecting data volumes up to 6.5 PB. Yet MWA accounts for only 1% of SKA-low's scale, giving a glimpse of SKA's data volume. Preliminary estimates indicate that the Science Data Processor in SKA1 phase will require computing power of 260 PFlops (260 quadrillion floating-point operations per second)—equivalent to 8 times China's Tianhe-2 supercomputer and 3 times the Sunway TaihuLight supercomputer. SKA's enormous computing demands will inevitably create massive impacts on existing scientific computing architectures and methods, while solving SKA data processing problems will help drive and upgrade related industries, potentially triggering revolutionary changes.

Challenges of SKA Big Data

Computing and Processing Challenges

SKA will bring tremendous advancement to numerous disciplines beyond astronomy, including computer science, informatics, and electronics . Terabyte-per-second high-speed digital sampling and high-speed real-time digital signal

processing pose new challenges to the electronics industry. Long-distance fiber-optic transmission of radio frequency signals with frequency synchronization in harsh outdoor environments is one of the urgent technical challenges for aperture arrays. High-speed, long-distance, large-bandwidth intercontinental transmission of big data imposes stringent requirements on current research network infrastructure, topology, communication protocols, and transmission endpoint software. Meeting the design requirements for ultra-high-speed streaming data processing is not simply achieved by increasing the number of internet ports per node or increasing total inter-node bandwidth. Effective solutions to this problem will also promote the deployment of domestic 100 Gb/s or even Tb/s-level backbone networks, a level that currently falls far short of SKA requirements.

SKA scientific data processing applications face the "memory wall" problem—that is, I/O issues where transmission bandwidth represents a major system bottleneck. Even supercomputers like Tianhe-2 have insufficient resources for processing SKA-class big data and are inconvenient for analyzing transient events. Therefore, there is an urgent need to research new architecture systems adapted to data-intensive scientific computing [4]. As mentioned earlier, SKA's high-speed massive input data must be reduced through real-time processing to relieve pressure on subsequent workflows. Massive data real-time processing imposes special design requirements on both hardware and software systems. The entire system architecture design and integration, supercomputing center execution frameworks and supporting software algorithms, data center health monitoring, rack cooling, and overall control management will all face numerous challenges. Moreover, with capped construction budgets, the system must achieve predetermined computing capabilities and real-time requirements while also meeting low power consumption demands for operational cost considerations. Additionally, massive data storage, archiving, retrieval, and computing place extremely high demands on the complete ecosystem of supercomputers. Although domestic CPU chips have been deployed in China's large supercomputing centers and domestic research institutions have developed deep learning processor chips for artificial intelligence applications, the mainstream operating systems, storage systems, and other software ecosystems currently come almost entirely from foreign sources. The most critical software ecological environment remains far behind international levels, lacks competitiveness, and suffers from severe "bottleneck" problems. Self-sufficiency capabilities are still insufficient. The SKA project provides strong demand-driven impetus for related industrial development.

Software and Algorithm Challenges

Beyond hardware issues, the current development level of astronomical application software also falls far short of SKA requirements. SKA scientific data processing's key algorithms involve extensive operations on shared resources including shared file systems. Traditional fixed multi-core computer systems frequently experience resource contention during multi-task, multi-concurrent,

multi-threaded parallel execution. If the dataflow execution framework cannot effectively and properly resolve resource scheduling and allocation, severe cases will cause data processing pipeline stalls [3,4]. In fact, this problem is not uncommon in SKA precursor telescope data processing centers. To address this, the International Centre for Radio Astronomy Research (ICRAR) in Australia and the Shanghai Astronomical Observatory of the Chinese Academy of Sciences jointly developed a dataflow execution framework called Data Activated Liu Graph Engine (DALiuGE) for the SKA project [3], which adopts an advanced "data-driven" design philosophy more suitable for SKA scientific computing than traditional HPC "computation-driven" designs. Furthermore, SKA scientific computing's actual operational efficiency is less than 10% of the planned efficiency, meaning its theoretical peak performance of 260 PFlops cannot meet actual scientific data processing requirements. Simply increasing supercomputing resources is not a feasible approach; a more viable path is improving software execution efficiency—increasing efficiency from 10% to 20% can save 50% of computing resources and significantly reduce operational costs. Collaboration between astronomers and computer experts to optimize code can improve algorithm and program execution speeds by several times. The urgent priority is cultivating interdisciplinary talent who understand both astronomy and computing.

Another practical issue is that astronomical data processing software urgently needs updating to meet future demands. Most currently used astronomical software was developed in the 1970s–1980s. Considering astronomical applications' requirements for high-speed, real-time, parallel big data processing, astronomers have begun adopting more advanced, modular, parallel-supporting development languages such as C++ or Python. CASA (Common Astronomy Software Applications), the C++-developed successor to AIPS, will become the next-generation mainstream radio astronomy software; programs involving machine learning and artificial intelligence will prioritize Python. Astronomical data processing software development, like astronomical research itself, has evolved from individual efforts to global collaborative teamwork. For example, the LIGO team that discovered gravitational waves comprises over 1,000 scientists, and the core library of CASA, widely used in radio astronomy processing software, has contributions from nearly 100 people worldwide. Aircraft carrier-style joint research teams and large-scale collaborative operations are becoming the standard model for solving major scientific problems.

Science Communication and Public Engagement

Science communication has gained unprecedented importance, as "scientific innovation and science popularization are the two wings of innovative development" [5]. Future SKA astronomical big data will serve not only astronomers but also provide public interfaces. Leveraging SKA to publicize research achievements, exchange academic ideas, disseminate scientific knowledge, and promote the scientific spirit will vigorously advance public awareness of basic science and

enhance the popularization of research. SKA regional centers will enable ordinary citizens to access science more conveniently through virtual observatories and "cloud" services, popularizing astronomy among the public.

Regional Data Center Strategy and Considerations

The deep analysis and processing of SKA data will be completed at regional data centers distributed across several continents. Several major member countries, including China, have expressed positive attitudes and high expectations for building SKA regional data centers and have already begun key technology research. Due to the special nature, complexity, and enormous volume of SKA data processing, large-scale data movement is unrealistic, making centralized data processing the inevitable choice. Building a Chinese SKA regional center is not only an indispensable part of the international SKA master plan but also an important guarantee to support Chinese scientists in effectively utilizing SKA data to obtain corresponding scientific returns. SKA scientists are widely distributed globally, making distributed computing and storage and cloud-based solutions considerations for data archiving and dissemination. A regional center grid composed of multiple scientific and data sub-centers can meet SKA's diverse needs.

The Shanghai Astronomical Observatory of the Chinese Academy of Sciences and the Australian SKA data center have already established an end-to-end direct connection, achieving a maximum data transfer rate of 3.2 Gbps—the highest known astronomical data flow rate to date—providing beneficial experience and a practical model for SKA regional centers. SKA's multi-scientific-objective and multi-data-attribute characteristics make multi-datastream parallelism an inevitable trend and a concern for future international network construction of SKA regional centers.

China's Strategic Approach

China is facing an important historical opportunity to advance scientific and technological innovation, which has been elevated to a strategic level for achieving the "Two Centenary Goals" and realizing the Chinese Dream of the great rejuvenation of the Chinese nation. SKA is the largest international cooperation project in astronomy that China has joined, creating a rare opportunity for China's radio astronomy to progress from "following" to "running alongside" and "leading." SKA will dominate and influence the development of radio astronomy for the next 50 years, ushering in a new era of vigorous development for low-frequency radio astronomy, nurturing numerous major scientific breakthroughs, and creating another brilliant chapter in observational cosmology .

To address SKA big data challenges, we should base ourselves on international cooperation while accelerating the domestic development of key core technologies. China could rely on its SKA regional science and data center to achieve

breakthroughs in key technologies such as Tb/s-level high-speed research backbone networks, signal and data transmission, and EB-level high-performance computing, while developing supporting astronomical software for relevant research topics. This would enable rapid achievement of major scientific results using SKA scientific data when the SKA era arrives, leading advanced science. In conclusion, humanity shares one sky. Participating in SKA global innovation cooperation to jointly promote leapfrog development in astronomy and contribute to solving scientific objectives of common human concern represents an important practice of the "building a community with a shared future for mankind" concept.

References

1 SKA Organisation. Advancing Astrophysics with the Square Kilometre Array. Italy: Proceedings of Science (AASKA14), 2015: and Computing, 2017, 20: 1-15.

5 习近平. 为建设世界科技强国而奋斗———在全国科技创新大会、两院院士大会、中国科协第九次全国代表大会上的讲话. [2016-05-30]. http://news.xinhuanet.com/politics/2016-05/31/c_{1118965169}.htm.

3 Wu C, Tobar R, Vinsen K, et al. DALiuGE: A graph execution framework for harnessing the astronomical data deluge. Astronomy and Computing, 2017, 20: 1-15.

5 习近平. 为建设世界科技强国而奋斗———在全国科技创新大会、两院院士大会、中国科协第九次全国代表大会上的讲话. [2016-05-30]. http://news.xinhuanet.com/politics/2016-05/31/c_{1118965169}.htm.

Author Information

Science Applications and Challenges of SKA Big Data

AN Tao[1]*, WU Xiangping[1],[2], HONG Xiaoyu[1], YE Shuhua[1], MAO Yufeng[3], GUO Shaoguang[1], LAO Baoqiang[1]

[1] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China

[2] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

[3] Bureau of Frontier Sciences and Education, Chinese Academy of Sciences,

Beijing 100864, China

*Corresponding author

AN Tao is a Research Professor at the Shanghai Astronomical Observatory (SHAO), Chinese Academy of Sciences (CAS). He serves as an Organization Committee member of Commission B4 Radio Astronomy, International Astronomical Union; Deputy Director of the Youth Committee, Chinese Astronomical Union; and Head of the SKA Group at SHAO. His research interests include radio astronomy, astrophysics, and astronomical techniques. E-mail: antao@shao.ac.cn

Responsible Editor: YUE Lingsheng

Footnotes:

Supercomputers built in the two host countries specifically for preprocessing these raw scientific data.

WU Xiangping, et al. China SKA Science White Paper (2017).

WU Xiangping, et al. China's Participation in SKA (Phase I) Comprehensive Evaluation Report (Draft), 2018.

YE Shuhua, WU Xiangping. Consulting Report on the Development Strategy of Low-Frequency Radio Astronomy in China. Chinese Academy of Sciences Academic Division Consulting Report, 2018.

Note: Figure translations are in progress. See original paper for figures.