Analysis of the Radio Astronomy Measurement Set Fourth Edition Data Model and Its Impact on China' s Participation in the SKA (Postprint)

Authors: Yijun Xu, Feng Wang

Date: 2026-03-03T15:53:36+00:00

Abstract

The Square Kilometre Array (SKA) radio telescope has entered the construction and commissioning phase; however, the efficient storage and management of massive data remains a critical issue to be addressed in SKA scientific data processing and the construction of the SKA Regional Center (SRC). This study systematically investigates the development history of Measurement Set (MS) technology across various stages, the technical characteristics of each version, and particularly their limitations in SKA applications. It focuses on analyzing the technical innovations brought by the next-generation Measurement Set version 4 (MSV4) and its potential impact on SKA data management and scientific research. The analysis indicates that by introducing features such as self-description, modular processing, and efficient storage, MSV4 is expected to effectively alleviate the bottlenecks in SKA data storage and processing. This new data model is anticipated to be applied to next-generation radio telescopes, exerting a significant influence on the entire field of radio astronomy. At this stage, it is essential to closely monitor and master the technical development trends of MSV4, strategically plan relevant technical reserves in advance, and fully prepare for the development and upgrading of SRC scientific application pipelines and data processing software, ensuring the smooth implementation of China' s SKA scientific research and the achievement of scientific returns.

Full Text

Preamble

Vol. 67, No. 1

2026 年 1 月

ACTA ASTRONOMICA SINICA Vol. 67 No. 1 Jan., 2026 doi: 10.15940/j.cnki.0001-5245.2026.01.010

Analysis of the Radio Astronomy Measurement Set Fourth Generation Data Model and Its Impact on China' s Participation in the SKA

XU Yijun[1,2], **WANG Feng**[1,2†]

([1] Center for Astrophysics, Guangzhou University, Guangzhou 510006) ([2] Greater Bay Area Sub-center of National Astronomical Data Center, Guangzhou University, Guangzhou 510006)

Abstract

The Square Kilometre Array (SKA) represents the next generation of giant radio telescope arrays, characterized by unprecedented sensitivity, resolution, and field of view. To meet the data processing challenges posed by the SKA' s massive data volumes and complex instrumental responses, the radio astronomy community has proposed the fourth generation of the Measurement Set (MSv4) data model. This paper provides a comprehensive analysis of the MSv4 architecture, focusing on its hierarchical structure, metadata organization, and support for heterogeneous computing environments. We evaluate the advantages of MSv4 over previous iterations, particularly in terms of scalability and I/O performance. Furthermore, we discuss the strategic implications of this new data standard for China' s involvement in the SKA project, emphasizing the need for domestic software development and data processing pipelines to align with MSv4 to ensure seamless integration and maximize scientific output.

1 Introduction

The Square Kilometre Array (SKA) is an international mega-science project aimed at constructing the world's largest radio telescope. With its vast collecting area and advanced phased-array technology, the SKA will generate data at rates far exceeding current astronomical facilities. Efficiently storing, managing, and processing this "Big Data" is a critical challenge for the global radio astronomy community.

Central to radio interferometric data processing is the Measurement Set (MS), a data model originally developed for the Common Astronomy Software Applications (CASA) package. As the field moves toward the SKA era, the limitations of the third-generation Measurement Set (MSv3) have become apparent, particularly regarding its ability to handle the massive scale of SKA-era datasets and the requirements of modern high-performance

摘要平方公里阵列射电望远镜（Square Kilometre Array, SKA）已进入建设与调试阶段，但海量数据的高

Efficient storage and management remain critical issues that must be addressed in the construction of Square Kilometre Array (SKA) scientific data processing systems and SKA Regional Centers (SRCs). This study systematically investigates the developmental history of Measurement Set (MS) technology across its various stages, examining the technical characteristics of different versions and, in particular, their limitations regarding SKA applications. We focus on the technical innovations introduced by the next-generation Measurement Set version 4 (MSV4) and its potential impact on SKA data management and scientific research. Our analysis indicates that by introducing features such as self-description, modular processing, and high-efficiency storage, MSV4 is expected to effectively alleviate the bottlenecks currently facing SKA data storage and processing. This new data model is poised for application in next-generation radio telescopes and will likely exert a significant influence on the field of radio astronomy as a whole.

At this stage, it is essential to closely monitor and master the technical development trends of MSV4. We must proactively establish technical reserves and prepare for the development and upgrading of scientific application pipelines and data processing software for the China SKA Regional Center. Such preparations are vital to ensuring the smooth progress of Chinese SKA scientific research and securing significant scientific returns.

Keywords: astronomical databases, instrumentation: interferometers, methods: numerical, data analysis

CLC number: P164; **Document code:** A

1 引言

Introduction

The Square Kilometre Array (SKA) project is a multi-national collaborative international scientific endeavor aimed at constructing the world' s most sensitive radio telescope array [**?**]. The SKA is dedicated to answering fundamental scientific questions about the universe, particularly concerning the formation of the first generation of celestial objects, galaxy formation and evolution, the nature of dark energy, cosmic magnetic fields, the essence of gravity, life-related molecules, and extraterrestrial civilizations. The construction and operation of the project are managed by the SKA Observatory (SKAO). The project is divided into two phases; currently, Phase 1 (SKA1) is underway, consisting of two components: SKA1-Mid and SKA1-Low. SKA1-Mid, located in South Africa, comprises 197 dish antennas, including 64 from the MeerKAT precursor telescope. SKA1-Low, located in Australia, consists of over 130,000 low-frequency dipole antennas.

The construction of the SKA has entered the commissioning phase of SKA1 AA0.5, with scientific observations expected to commence after 2027. Data processing represents one of the most significant challenges in the construction of the SKA project.

The SKA project will generate more than 700 PB of scientific data annually. The storage, management, and processing of this data have consistently been a primary focus of scientific research. In radio astronomy data processing, efficient and accurate data storage combined with high-speed access methods are critical foundational technologies for scientists to conduct research and drive new scientific discoveries.

To address the challenges of radio astronomy data storage and management, various general-purpose data storage models have been proposed and implemented. The Measurement Set (MS) is a commonly used storage format for radio interferometry data [**?**]. MS is widely applied in data processing software such as CASA (Common Astronomy Software Applications) [**?**], with the current mature version being MS Version 2 (MSV2).

The Flexible Image Transport System (FITS) is a universal astronomical data storage format that supports images, spectra, and tabular data [**?**]. Hierarchical Data Format version 5 (HDF5) is an efficient data storage format suitable for the storage and management of large-scale data, which has gradually seen application in radio astronomy in recent years [**?**]. Additionally, proprietary formats have emerged for specific telescopes, such as the ALMA Science Data Model (ASDM) used by the Atacama Large Millimeter/submillimeter Array (ALMA), which is specifically designed to store raw observational data and associated metadata [**?**]. To deeply analyze and compare the characteristics of various data storage models, we compared MSV2, the traditional FITS format, and the HDF5 technology previously adopted in the Low-Frequency Array (LOFAR) project across multiple dimensions.

Detailed comparisons are presented in Table 1.

From this comparison, it is evident that the metadata management approach of the FITS format is relatively simple and struggles to meet the complex requirements of radio interferometry data. It lacks sufficient scalability for multi-dimensional data common in radio astronomy (such as spectra, time series, and spatial distributions), making it difficult to effectively represent complex data relationships. Although FITS is widely used in astronomy, its original design was oriented toward storing images and spectra. Consequently, the FITS format is inefficient when handling high-dimensional data. Radio interferometry data typically contains vast amounts of visibility data, with dimensions including time, baseline, frequency, and polarization. Storing such high-dimensional data in a tabular FITS format leads to low storage and access efficiency. To address this, formats like UVFITS [**?**] or FITS-IDI (FITS Interferometry Data Interchange) [**?**] were proposed for large-scale data; however, these improved FITS formats did not fundamentally resolve the issues inherent in radio interferometry data

storage. Furthermore, because FITS files utilize a single-file storage method based on a text header followed by binary data, they do not fully account for the requirements of distributed computing environments, making it difficult to support efficient parallel processing and dynamic scaling.

The original design goal of HDF was to provide a "graphics format" standard for scientific data. The HDF data format offers high performance and reduces resource consumption through transparent compression. Developed and maintained by a non-profit organization, HDF became popular across various disciplines—particularly astrophysics—starting with its fourth generation. Selected by the National Aeronautics and Space Administration (NASA), data from many astronomical telescopes are stored in HDF format.

Although the HDF5 format is powerful and was trialed in the LOFAR low-frequency array, its ecosystem support in the field of radio astronomy is not as comprehensive as that of the Measurement Set. It lacks specific designs tailored to the unique characteristics of radio data, as well as opportunities for specialized optimization.

Overall, the data dimensions for radio interferometry arrays such as the SKA and the Next Generation Very Large Array (ngVLA) [**?**] are highly complex. Visibility data dimensions include time, baseline, frequency, and polarization. Given these storage requirements, Measurement Set technology represents a relatively feasible approach. MS was specifically proposed for radio astronomy data storage and has very clear application targets. Therefore, developing data management for next-generation radio telescopes based on Measurement Set technology is a highly logical choice.

The National Radio Astronomy Observatory (NRAO), the National Astronomical Observatory of Japan (NAOJ), and the Square Kilometre Array Observatory (SKAO) have jointly proposed the Measurement Set Version 4 (MSV4) data model, which has now entered the community evaluation phase. Analyzing, understanding, and mastering this data model as soon as possible is a crucial step in the development of the SKA Science Data Processor (SDP) system and the construction of the SKA Regional Centres (SRC) project. It is of great significance for Chinese scientists to fully utilize future SKAO scientific data for research.

This paper analyzes the technical development process of the Measurement Set data model and explains the necessity and key technical points of the MSV4 data model. Section 2 introduces the basic concepts, technical evolution, and limitations of the MS data model. Section 3 focuses on the fundamental design philosophy, design goals, and corresponding key technologies of the MSV4 data model. Section 4 discusses the application of MSV4 in SKA data processing, providing an analysis and outlook on the usability of MSV4 in the construction of the SKA SDP and SRC. The final section provides a summary.

2 测量集早期版本技术发展及其局

The Measurement Set (MS) data model is a specialized data storage and management standard designed for radio interferometry. It is specifically engineered to efficiently store and process the complex observational data generated by radio telescope arrays.

Its core design is based on modern database technology and hierarchical storage structures, enabling the flexible organization and management of visibility data and its associated metadata. The Measurement Set data model provides a standardized and efficient solution for the storage, management, and analysis of radio interferometric data. In practical applications, following the introduction of MSV1, it rapidly became the standard for radio astronomy data storage and processing, receiving extensive support from mainstream radio astronomy software suites.

2.1 测量集 V1 和 V2 版本的核心技术分析

In 1996, Wieringa and Cornwell proposed version 1.0 of the Measurement Set (MS) data model [**?**]. This provided a universal data format suitable for different telescopes and projects, facilitating data sharing and collaborative research. In terms of data storage, it adopted a table-based structure to support fast querying and access while optimizing the data storage layout to reduce redundancy.

The Measurement Set supports the storage of multi-band, multi-polarization, and complex calibration data. It has undergone several iterations of optimization and improvement, such as version 2.0 [**?**] and version 3.0 [**?**]. MSV2 is currently widely supported by mainstream radio astronomy data processing software, such as CASA [**?**] and the Astronomical Information Processing System (AIPS++) [**?**].

The structure of an MS file can be viewed as a hierarchical database, where the main table (the MAIN table) serves as the core, and other tables—such as the antenna table, spectral window table, and observation log table—are associated with the main table through foreign keys.

1. **MAIN Table**: The MAIN table is the core of the MS file and stores all information related to the observed data. Each record corresponds to the visibility data for a specific timestamp, baseline, and frequency channel. The key columns of the MAIN table include (column names are capitalized to maintain consistency with the table structure, and bold text highlights the specific column names):

- **TIME**: The observation timestamp, typically stored in Modified Julian Date (MJD) format.
- **ANTENNA1** and **ANTENNA2**: The indices of the two antennas forming the baseline, used to identify the antenna pair involved in the interference.

- **FREQ**: The frequency channel index, pointing to specific frequency values in the SPECTRAL_{WINDOW} table.
- **DATA**: Complex visibility data, usually stored as a two-dimensional array (polarization × frequency channel).
- **FLAG**: Data flags used to mark invalid or interfered data points.

2. **Subtables**: Subtables are used to store metadata related to the observation, such as antenna positions, frequency channel information, and observation logs. These tables are linked to the MAIN table via foreign keys to form a complete data storage system. For example:

- **ANTENNA table**: Stores the name, position, and offset information for each antenna.
- **SPECTRAL_{WINDOW} table**: Stores the specific values and widths of the frequency channels.
- **OBSERVATION table**: Stores basic observation information, such as the telescope name and the observer.

The designs of MSV1 and MSV2 fully account for the characteristics of radio interferometry data. By leveraging the technical features of relational databases, they achieve large-scale data storage through inter-table data associations. Compared to other astronomical data storage standards of the same period, particularly FITS, the Measurement Set data model represents a significant advancement in several areas:

1. The core design philosophy of the Measurement Set is based on tabular data storage using a hierarchical structure to support diverse observation modes and data processing requirements. This design clearly considers the specific characteristics of radio data, where the entire dataset is stored hierarchically according to logical relationships. The MAIN table stores the core visibility data, while subtables (such as ANTENNA and SPECTRAL_{WINDOW}) store observation-related metadata. This hierarchical design not only improves data organization efficiency but also makes data reading and querying more flexible. For instance, users can access visibility data for a specific time period independently without loading the entire dataset [**?**].

2. Support for multi-dimensional data storage. Radio interferometry data typically involves multiple dimensions, including time, frequency, polarization, and baseline. MSV2 achieves efficient management of high-dimensional data by storing different dimensions in separate columns of the MAIN table. For example, the **DATA** column stores complex visibility data, while the **FLAG** column is used to mark invalid data points. This design reduces data redundancy and facilitates rapid analysis of specific data dimensions.

3. Excellent extensibility. By adding new tables or columns, users can easily extend MS files to support new observation modes or data processing needs. For example, when supporting multi-beam observations, new columns can

be added to the MAIN table to store data from different beams. This flexibility allows MSV2 to adapt to evolving scientific requirements. It is certain that the adoption of this design in MSV1/V2 resolved data storage issues for multiple radio interferometers emerging since the 1990s. In many precursor arrays of the SKA project, such as the Murchison Widefield Array (MWA), MeerKAT, and the LOFAR low-frequency array, the MS format has served as the standard for data storage and has been widely applied within the radio astronomy community.

However, a deep analysis of MSV2 reveals certain limitations inherent to the technology and the era in which it was proposed. These limitations prevent MSV2 from fully meeting the construction requirements of the SKA project, primarily in the following aspects:

1. **Low storage efficiency**: Because MSV2 employs a table-based storage structure where data is stored in plain text without efficient data compression techniques, the storage space requirements for MSV2 files increase dramatically when processing observations with high temporal and spectral resolution. This leads to a significant rise in storage and transmission costs. For example, a typical radio interferometry observation may generate several terabytes of data, while a complete observation sequence could produce as much as 14 PB. The storage efficiency of MSV2 cannot accommodate such data scales [**?**], posing immense challenges to back-end storage and file systems.

2. **Limited parallel processing capability**: The rapid growth in data volume caused by SKA's characteristics—such as high spatial and temporal resolution—urgently requires effective solutions, including the use of high-performance computing and parallel computing. However, the design of MSV2 did not fully consider parallel read/write requirements, resulting in limited performance in modern distributed computing environments. For instance, when accessing MSV2 files simultaneously across a multi-node cluster, I/O bottlenecks can significantly degrade performance, greatly restricting data processing efficiency.

3. **Insufficient support for complex observation modes**: The SKA introduces several new observation modes (such as wideband spectral observations, multi-target tracking, co-existence of interferometric and single-dish modes, and sub-array observations) that place higher demands on data storage formats. However, the design of MSV2 is relatively rigid, making it difficult to extend flexibly to support these complex modes. For example, when processing wideband spectral data, MSV2 requires partitioning the data into multiple frequency channels for storage, which not only increases the complexity of data management but may also lead to data consistency issues [**?**].

2.2 MSV3 的发展与局限性分析

Development of the Measurement Set 3.0 Standard

To advance the construction of the Square Kilometre Array (SKA), the SKA Observatory (SKAO) initiated efforts to promote subsequent Measurement Set standards following the Critical Design Review in 2018. In 2019, the MSV3 working group released a test version known as Measurement Set 3.0 [**?**]. This version aimed to enhance the efficiency and flexibility of radio astronomy data storage and processing by introducing explicit keys, data versioning, unified data columns, and new sub-tables and functionalities.

MSV3 introduced a version management mechanism for data variables such as DATA, WEIGHT, and FLAG, allowing multiple versions of data to coexist. In this framework, the currently active versions of DATA, WEIGHT, and FLAG are explicitly identified. This approach to versioning makes data updates and calibrations more flexible without the need to create redundant columns, such as CORRECTED_{DATA} or FLAG_{VERSION}. Software implementations can intelligently handle these versions as required, thereby avoiding unnecessary data replication. Furthermore, FLOAT_{DATA} was replaced by DATA, unifying the storage methods for single-dish and interferometric data. Software can efficiently store floating-point data and provide single-precision complex data upon request. This improvement increases storage flexibility and reduces the complexity arising from disparate data types.

MSV3 introduced several new sub-tables, including BEAM, EPHEMERIDES, INTERFEROMETER MODEL, and PHASED ARRAY. These sub-tables provide greater flexibility and functional support for processing complex observational data. For example, the EPHEMERIDES sub-table can be used to store orbital information for near-Earth objects, which is highly beneficial for processing observations of moving targets. MSV3 also expanded certain functionalities, such as support for phased-array station-based antennas (via the PHASED ARRAY sub-table) and scan information (via the SCAN sub-table). These extensions allow MSV3 to better adapt to the diverse requirements of modern radio astronomy observations.

From another perspective, although MSV3 provided more flexible data storage methods, these innovations did not fully meet expectations in practical applications. For instance, while data versioning offered increased flexibility, it could lead to performance degradation when processing massive datasets. The design requirements of MSV3 necessitated more complex software implementations, potentially increasing development and maintenance costs. Specifically, requiring software to intelligently manage multiple data versions without redundant columns proved to be highly complex in practice. Subsequent feedback from the user community indicated that while some new features were theoretically useful, they were rarely used in practice and added unnecessary complexity.

At the same time, MSV3 remained unable to fully adapt to the rapid changes in

radio astronomy observation requirements. When dealing with various complex observation modes, MSV3' s design remained relatively rigid despite its partial support for these modes, making it difficult to expand flexibly to accommodate emerging observational technologies.

The factors mentioned above collectively led to MSV3 failing to gain widespread acceptance among radio astronomers during community evaluations. Ultimately, it became a transitional version that was never fully deployed for practical use. Nevertheless, some of its underlying design concepts provided valuable references for the development of subsequent versions.

3 MSV4 的演进与关键技术分析

As previously mentioned, the emergence of next-generation radio interferometers—including the Atacama Large Millimeter/submillimeter Array Wideband Sensitivity Upgrade (ALMA-WSU), the Next Generation Very Large Array (ngVLA), and the Square Kilometre Array (SKA)—will increase astronomical data volumes by several orders of magnitude. To address the challenges posed by this data explosion, the Measurement Set v4 (MSV4) has been proposed and initially implemented within the open-source Python package XRADIO (Xarray Radio Astronomy Data IO).

The design of MSV4 is primarily characterized by efficient data storage and compression, enhanced parallel processing capabilities, and flexible scalability. It should be noted that while MSV4 has not yet been officially released, the analysis in this paper is based on the current code implemented by the National Radio Astronomy Observatory (NRAO) in XRADIO, the released software development kit (Xarray-ms), and various technical discussion documents. It is anticipated that even if further refinements or improvements are made upon the official release of MSV4, the overarching framework of the standard will remain substantially unchanged.

3.1 MSV4 的特色

MSV4 represents a significant departure from previous versions of Measurement Sets (MS), completely overhauling the underlying supporting technologies. It fully integrates the latest advancements in data management that have emerged since the 21st century due to the development of cloud computing and distributed computing technologies. Extensive optimizations have been implemented for massive data processing, particularly regarding parallel processing. Its primary characteristics include: 1. Self-descriptiveness: MSV4 datasets are self-describing, which resolves the long-standing issue in MSV2 where data required external metadata for definition. Specifically, observed scientific data are organized in a tabular format, where each table contains specific data types (e.g., the MAIN table stores visibility data, while the ANTENNA table stores antenna information). The column names and data types explicitly describe the

meaning of the data. Metadata and scientific data are stored within the same framework; for example:

The TIME column stores timestamps, while the ANTENNA1 and ANTENNA2 columns store baseline information. Data coordinates (such as time, frequency, and baseline) and dimensionality information are explicitly recorded, making it easier for users to understand the data organization. 2. Regular Time-Frequency Grids: MSV4 changes the fundamental method of data preservation. Data are composed of Xarray Datasets [**?**], which contain multi-dimensional arrays (ndarrays) on regular time-frequency grids. Unlike the table-based format of MSV2, the MSV4 data structure is more regular, which facilitates more efficient data processing and analysis.

分析变得更加简单和高效．与 MSV2 数据以表格形

Compared to traditional storage methods, MSV4 consistently maintains the regularity of the time-frequency grid. 3. Efficient Data Processing: By leveraging Xarray' s internal I/O routines and lazy loading mechanisms, MSV4 achieves highly efficient data loading and processing. MSV4 supports arbitrary chunked array types, including support for distributed frameworks such as Dask, which allows data to be partitioned into arbitrary chunks to facilitate parallel computing. This fundamentally addresses the limitations of MSV2, which supported a restricted range of data types and faced constraints regarding data chunking and parallel execution. 4. Lack of Backward Compatibility: MSV4 is an entirely new data model and does not provide backward compatibility. Consequently, MSV4 does not directly support read or write operations for data formatted under the MSV2 data model.

By abandoning previous constraints and redefining the Measurement Set data model, MSV4 achieves a fundamental shift in overall performance. However, this transition requires a comprehensive re-evaluation and mastery of the new data model, as well as addressing the resulting challenges in data storage format compatibility. 5. Diverse Application Scenarios: MSV4 is designed to satisfy a wide range of use cases within the field of radio astronomy, including radio interferometry, single-dish observations, real-time mosaic observations, ephemeris observations, heterogeneous antenna Very Long Baseline Interferometry (VLBI), phased-array stations, and phased-array feeds.

3.2 MSV4 的多层数据架构

Leveraging the underlying support of Xarray, the core architecture of MSV4 is divided into three layers: the data storage layer, the data processing layer, and the metadata management layer. 1. In the data storage layer, MSV4 directly adopts Zarr and NetCDF (Network Common Data Form) [**?**] to support efficient multidimensional data storage and parallel access. Zarr is a data format based on chunked storage, specifically designed for the efficient storage and access of large-scale multidimensional array data. By employing chunking and

compression techniques, it significantly reduces storage space requirements and improves data access speeds. The Zarr format supports various storage backends (such as local file systems and cloud storage) and is highly scalable, making it suitable for processing ultra-large-scale datasets. Another advantage of Zarr is its support for parallel read and write operations, making it an ideal choice for distributed computing environments. NetCDF is a widely used scientific data format primarily utilized for storing multidimensional array data. The NetCDF format is self-describing, allowing data to be stored alongside metadata (such as units and coordinate axis information), which facilitates data understanding and sharing.

NetCDF supports efficient data compression and chunked storage, making it suitable for handling large-scale scientific datasets. Its cross-platform compatibility and extensive tool support (such as Python' s `netCDF4` library) have established it as one of the standard formats for scientific research and data exchange. Consequently, by introducing these two mature underlying software packages, MSV4 effectively addresses issues related to astronomical data storage, the design of new data structures, and distributed computing. Furthermore, by utilizing the cloud support provided by Zarr and NetCDF, cloud storage for observational data can be realized. 2. In the data processing layer, MSV4 draws on the experience of MSV3 by introducing the concept of "data groups," which allows users to store multiple versions of visibility data within the same dataset. [Figure 1: see original paper] illustrates the framework structure of MSV4. Under this framework, raw visibility data can be stored in one data group, while calibrated visibility data can be stored in another. This design not only enhances data flexibility but also facilitates the comparison and analysis of different data versions by the user. The data group functionality provides significant convenience for SKA data processing and is expected to reduce the volume of redundant data.

In practical data processing, a data group dictionary is utilized as an attribute of the main dataset (`ms_{xds}`), which can contain one or more data groups. Data variables can be shared between data groups or remain unique to a specific group. For example, the following describes a configuration with a "base" group and an "imaging" group:

In this configuration, the "base" and "imaging" data groups share the same `flag` and `uvw` data variables, but possess different `correlated_{data}` and `weight` data variables. During data processing, users can extract the corresponding data by specifying the appropriate group name.

[Figure 1: see original paper] Schematic diagram of the MSV4 data model layout. Optional datasets are indicated by parentheses. Data variables are represented in uppercase. The suffix `_{xds}` denotes an Xarray dataset, while `_{info}` denotes a dictionary.

Fig. 1 MSV4 data mode layout diagram. Optional data sets are indicated with parentheses. Data variables are represented in uppercase. The suffix "*{xds}*"

denotes the Xarray data set, and "{info}" denotes the dictionary.

ms_xds:attrs['data_groups'] = {'base' : {'correlated_data' : 'VISIBILITY', 'flag' : 'FLAG', 'weight' : 'WEIGH
'imaging' : {'correlated_data' : 'VISIBILITY_CORRECTED', 'flag' : 'FLAG

3. In the metadata management layer, MSV4 fully exploits the advantages of Xarray by adopting a self-describing data model. Each dataset contains comprehensive metadata information, such as observation time, frequency range, and polarization settings. This design enables MSV4 to support multiple observation modes more flexibly and facilitates data sharing and exchange.

3.3 预期关键性能变化

Based on existing technical documentation, we can further analyze the performance variations of MSV4 in terms of storage and parallel read/write capabilities (as shown in the performance comparison of relevant formats in). Regarding storage performance, MSV2 does not possess inherent compression functionality; however, it is typically compressed using third-party software such as GZIP. In contrast, both MSV4 and HDF5 feature native compression capabilities. According to published data for HDF5, the compression ratio typically ranges between 1.5:1 and 2.5:1 when processing scientific data such as floating-point arrays [**?**]. Using this compression ratio as a reference, and accounting for the fact that redundant data in MSV4 no longer requires storage, we can further estimate the data compression efficiency of MSV4.

In terms of parallel read/write performance, current classic file systems such as Lustre typically achieve speeds of 5–10 GB/s. Consequently, the parallel reading capabilities of MSV2 are primarily limited to the concurrent reading of different data tables, reaching at most the peak performance of the underlying file system. In contrast, HDF5 provides native support for parallel I/O. Existing data indicates that HDF5 can achieve performance levels of 15–25 GB/s during parallel read/write operations. Overall, it is anticipated that the performance of MSV4 will be fundamentally comparable to that of HDF5.

Comparative Analysis of MSV2, MSV4, and HDF5 for SKA Data Management

As shown in , the storage performance of the MSV2 format is characterized by a multi-file directory structure that incurs significant metadata overhead, leading to reduced operational efficiency. Furthermore, MSV2 relies on third-party compression software with relatively limited capabilities, typically achieving compression ratios between 1.2:1 and 1.5:1. Consequently, for a dataset of 100 TB, the resulting compressed volume remains substantial, ranging from approximately 66.7 TB to 83.3 TB.

Optimized for big data, with an Efficient when configured efficient layout and improved with block storage, making it performance. By significantly suitable

for large-scale reducing redundant data, it is datasets. Supports multiple expected that 100 TB of data compression methods, with will result in an actual storage

100 TB of data being

requirement of 33.3–50 TB. compressed to 40–66.7 TB.

Limited parallel I/O support, with Designed specifically for parallel Parallel throughput ranging from 5 to 10 GB/s.

I/O, with throughput ranging Read/Write Parallel reads are supported across from 15 to 25 GB/s. Optimized Performance different tables, but not within the versions could potentially reach same data table. 20–30 GB/s.

Supports parallel I/O through MPI-IO, with throughput ranging from 15 to 25 GB/s.

3.4 MSV4 功能变化对现有软件的影响

Future scientific research with the SKA will rely entirely on the SRC (SKA Regional Centre) platform. As SKA AA0.5 nears completion, the subsequent AA1 and AA2 phases—comprising 64 dish antennas plus 4 MeerKAT dishes in SKA-Mid, and 68 stations in SKA-Low—will commence immediately. Early science operations for the SKA will also be conducted concurrently. If subsequent data is stored using the MSV4 (Measurement Set v4) data model, existing software tools will become entirely incompatible. Therefore, understanding and mastering the structural changes of MSV4 data, and upgrading existing software accordingly, is essential for future SKA data processing and scientific utilization.

1. As previously mentioned, the MSV4 data structure has undergone fundamental changes. Through the self-describing mechanism of Xarray, data obtained from each observation (such as visibility data) contains information regarding the specific observation run, spectral windows, polarization settings, observation modes, processors, and per-antenna beams. Furthermore, data is stored in datasets labeled as n-dimensional arrays rather than traditional database tables. In traditional software development using `python-casacore`, users are accustomed to treating visibility data as a complex Numpy array, where the specific observation time and associated baseline for each visibility must be manually aligned by the developer using other data tables. In MSV4, however, the imported data is an n-dimensional array with dimensions of time $\times$ baseline $\times$ frequency $\times$ polarization (where rows have been split into time $\times$ baseline). From a software development perspective, the MSV4 data model is easier to understand and analyze, although it may lead to a certain increase in memory overhead.
2. In MSV2, most "keys" were indexed using implicit numbering—essentially

using subscript values to index data. In MSV4, these have been replaced with descriptive names.

While this change does not strictly necessitate modifying original code logic, it significantly improves code readability and quality. It also allows for the selection of sub-data without re-indexing, making data combination more efficient.

3. With the upgrade to MSV4, the concept of "Data Description" has been deprecated and replaced by `spectral_{{{window}}_{{name}}}` and `polarization_{setup}`. Given that radio interferometers function as spectrometers where each channel has a different weight, `WEIGHT` has been redefined as `WEIGHT_{SPECTRUM}`. The shape of the `WEIGHT` data variable is now identical to that of the `VISIBILITY/SPECTRUM` data variables. In MSV4, the `FIELD`, `SOURCE`, and ephemeris data from MSV2 have been merged into a single dataset. Similarly, antenna and feed data have been consolidated (as each antenna in MSV4 can only have one feed type). These changes have a significant impact on how subsequent data processing software reads and writes data. Furthermore, unlike MSV2, MSV4 utilizes the JPL Horizons ephemeris to create ephemeris models.

4. To reduce data volume and improve processing usability, data groups are used to implement version control for `VISIBILITY/SPECTRUM`, `WEIGHT`, `UVW`, and `FLAG` data variables.

4.1 MSV4 数据模型的组织结构对 SKA 观测类

According to the data management policies released by the SKAO [**?**], future SKA observations will encompass various project types, including Key Science Programs (KSP), Principal Investigator (PI) projects, and Director' s Discretionary Time projects. Different observation types involve varying time allocations; for instance, KSPs are expected to span at least five cycles, whereas PI projects are generally completed within a single cycle. These diverse project types impose novel requirements on data management. Specifically, data from a KSP cannot be released until the project is completed, and members outside the designated team are restricted from accessing or processing these KSP datasets during the proprietary period.

Based on the current analysis of the MSV4 data organization, the format is fully capable of supporting these diverse observation types and their subsequent processing. At the data organization level, MSV4 can effectively manage data from different observation modes. As illustrated in the schematic diagram of the MSV4 data organization in [Figure 1: see original paper], a single observation yields subsets such as `correlated_{xds}` (interferometric data), `antenna_{xds}` (antenna information), and `pointing_{xds}` (pointing information). Each subset stores data through labeled multi-dimensional arrays. For example, the `VISIBILITY` data variable stores interferometric visibility data with dimensions including time, baseline, frequency, and polarization. MSV4 supports dynamic expansion, allowing users to add new data variables or metadata

fields as needed. This means that specific data variables for new observation modes can be integrated without modifying the existing data structure. Such a design enables MSV4 to adapt flexibly to the future observational requirements of the SKA.

As also shown in [Figure 1: see original paper], in addition to individual observations, MSV4 supports the concept of a Processing Set (PS) to facilitate the organization and processing of multiple datasets related to a single scientific objective. A PS can group multiple MSV4 datasets together for collective processing, providing a convenient framework for managing and handling massive volumes of observational data. By grouping related datasets, the Processing Set enables efficient data selection, filtering, and processing. For instance, it allows users to perform operations on multiple datasets in a coordinated manner, ensuring consistency and reducing the complexity of data processing. Furthermore, the Processing Set provides methods for summarizing metadata, selecting subsets based on various criteria, and exporting data to storage formats. For example, the `summary` method can generate overview information for the datasets, including dataset names, scientific intent, polarization modes, and spectral window names.

This functionality allows scientists to quickly grasp the fundamental information of each dataset within a PS, thereby facilitating better data management and analysis. Overall, the Processing Set is extremely useful for SKA KSP missions, such as the Epoch of Reionization (EoR) studies that require long-term observations. It enables the effective integration and management of observational data spanning months or even years, ensuring that the requirements for the SKA' s long-cycle observations are met.

Acta Astronomica Sinica

4.2 MSV4 在 SKA 数据处理中的可用性分析

As previously mentioned, the development of MSV4 was expedited as much as possible to promptly address the urgent data processing requirements of both the NRAO and SKAO.

MSV4 fully leverages Xarray datasets, adopting a structure of multi-dimensional arrays and metadata labels to replace traditional tabular storage methods. This design enables MSV4 to process large-scale data more efficiently and support complex observation modes. Consequently, the parallel computing capability of MSV4 depends on the performance of the underlying Xarray framework. It can be argued that the issues encountered in MSV3 have been largely resolved in MSV4, with the technical cornerstone being the adoption of Xarray-based datasets. While Xarray introduces significant improvements in performance and data structure, it also results in MSV4 being incompatible with legacy data storage formats.

In terms of specific computational requirements, taking the SKA continuum

pipeline as a reference, it is estimated that a one-hour observation may generate over 1 PB of visibility data, ultimately producing science images on the scale of several TBs. Taking the memory overhead of the pixel gridding process as an example, in order to achieve high-precision gridding results, the required grid memory allocation is approximately

74 GB. 当成像尺度达到 96 K，网格化所消耗的内

The storage requirements will reach 667 GB. Given this scale, we are confident that Xarray can perform the aforementioned computations and save the results in a distributed environment, thereby meeting the data processing requirements for future SKA (Square Kilometre Array) observations.

4.3 对 SKA 数据处理管线的影响与应对

As the Square Kilometre Array (SKA) transitions into its preliminary scientific research phase, there is an urgent need for Chinese scientists to develop various data preprocessing and processing pipelines in anticipation of the forthcoming MSV4 data format.

Judging from current development trends within the SKA Observatory (SKAO), the scientific data eventually delivered to regional Science Regional Centres (SRCs) will likely be based on the MSV4 standard. Given that MSV4 does not support backward compatibility, various applications currently developed based on MSV2 will not be directly applicable to MSV4 in the future. To avoid such a predicament, Chinese scientists must maintain a clear understanding of this transition while developing data processing programs and adequately prepare for changes in the underlying data storage format. This development challenge is a common issue faced by all international SKA collaboration teams. In this regard, some countries have already rapidly initiated relevant preliminary research. We note that colleagues in South Africa have developed `xarray-ms`. This tool provides an MSV4 "view" of Measurement Sets on top of MSV2. By utilizing an MSV4 Xarray view over MSV2 data, developers can build applications using an easily understood data structure and then transition seamlessly to the newer format. Data can also be exported to updated formats (primarily Zarr) via Xarray's native I/O routines. For software developers, the Xarray view remains identical across both formats. Clearly, this model provides significant guidance for our current software and application development efforts.

4.4 对 SRC 建设与运行的影响

According to the construction requirements of the Square Kilometre Array (SKA), all future scientific data processing must be conducted on the SKA Regional Centre (SRC) platforms of various member nations. Data acquired from SKA1-Low and SKA1-Mid observations will be pushed to the SRCs for follow-up processing after initial treatment by the Science Data Processor (SDP). Since all future SKA scientific research activities will be restricted to the SRC platforms,

the introduction of Measurement Set version 4 (MSV4) is expected to play a critical role in advancing the construction and development of these centers.

MSV4 is expected to further reduce data storage resource overhead, thereby lowering the overall construction costs of the SRCs. Simultaneously, it will facilitate improvements in subsequent data processing performance. Regarding remote data transmission and synchronization between different SRCs, the reduction in data volume translates directly to decreased network transmission times. This enhancement is highly significant for the efficient operation and maintenance of the global SRC network.

5 结论

With the joint proposal of the draft MSV4 version by various observatories within the international radio astronomy community, a brand-new definition for the Measurement Set (MS) data model has been established. Following community review and testing, this version is expected to become the standard data storage format for the next generation of radio interferometric telescopes.

In addition to radio interferometry arrays, MSV4 is designed to support a variety of scenarios, including single-dish antennas and phased arrays. As a foundational data model, the introduction of MSV4 will have a significant impact on the entire field of radio astronomy. It is imperative that we analyze, understand, and master the technical essentials and functional characteristics of MSV4 as early as possible to lay the necessary foundation for subsequent data processing developments.

This paper provides a formal definition of Measurement Set files, systematically introduces their evolutionary process, and analyzes the technical characteristics of various versions of the MS data model. Building upon this, we focus on the impact of MSV4 on China's participation in the construction of the Square Kilometre Array (SKA) Science Data Processor (SDP) and Regional Centres (SRC). The work presented in this paper serves as a reference for the future development of SKA science data processing packages, regional centers, and data application processing systems in China. https://github.com/pydata/xarray SKA Technical Document: SKA SDP Parametric Model Documentation https://sdp-par-model.readthedocs.io https://github.com/ratt-ru/xarray-ms. Xu Yijun et al.: Analysis of the Fourth Version of the Radio Astronomy Measurement Set Data Model and Its Impact on China' s Participation in SKA. References: Dewdney P E, Hall P J, Schilizzi R T, et al. IEEEP, 2009, 97: 1482. Wieringa M H, Cornwell T J. Definition of Measurement Set: AIPS++ Note 191. Charlottesville, VA.

National Radio Astronomy Observatory, 1996 Kemball A J, Wieringa M H. Measurement Set Defini- tion Version 2.0. Charlottesville, VA: National Radio Astronomy Observatory, 2000 McMullin J P, Waters B, Schiebel D, et al. ASPC, 2007, 376: 127 Wells D C, Greisen E W, Harten R H. A&AS, 1981, 44:

Folk M, Heber G, Koziol Q, et al. Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases. New York: Association for Computing Machinery, 2011:

Anderson K, Alexov A, Bahren L, et al. ASPC, 2011, 442: 53 Viallefond F. ASPC, 2006, 351: 627 Greisen E W. AIPS FITS file format, AIPS Memo Series 117. Charlottesville, VA: National Radio Astron- omy Observatory, 2012 Greisen E W. The FITS Interferometry Data Inter- Charlottesville, definition version 3.0β. change Convention—Revised, AIPS Memo Series 114.

Charlottesville, VA: National Radio Astronomy Obser- vatory, 2011 Selina R J, Murphy E J, McKinnon M, et al. SPIE, 2018, 10700: 497 Dijkema T J, Golap K, Hiriart R, et al. Measurement National Radio Astronomy Observatory, 2019 CASA Team. PASP, 2022, 134: 114501 McMullin J P, Golap K, Myers S T. ASPC, 2004, 314:

Offringa A R, McKinley B, Hurley-Walker N, et al.

MNRAS, 2014, 444: 606 Bhatnagar S, Rau U, Golap K. ApJ, 2013, 770: 91 Hoyer S, Hamman J. JORS, 2017, 5: 10 Rew R, Davis G. ICGA, 1990, 10: 76 Habib S, Morozov V, Frontiere N, et al. Communica- tions of the ACM, 2016, 60: 94 Ball L. SKA Observatory Access Policy. SKAO-GOV-

0000088. Macclesfield, Cheshire: SKA Observatory,

Breen S, Bolton R, Chrysostomou A. SKAO Science Data Products: A Summary. SKA-TEL-SKO-0001818.

Macclesfield, Cheshire: SKA Observatory, 2021 Measurement Set Version 4 Data Model: Analysis and Relevance to China' s SKA Participation XU Yijun1,2 WANG Feng1,2 (1 Center for Astrophysics, Guangzhou Univeristy, Guangzhou 510006) (2 Great Bay Center of National Astronomical Science Data Center, Guangzhou Univeristy, Guangzhou 510006) ABSTRACT The Square Kilometre Array (SKA) has entered the construction and commissioning phase. However, the storage and processing of its massive data remain one of the critical challenges that need to be addressed. The construction and operation of the SKA Regional Center (SRC) also heavily rely on efficient data storage and management solutions. To explore the key issues in SKA data storage and management in depth, this paper systematically investigates the development history of Measure- ment Set (MS) technology, the technical characteristics of its various versions, and their limitations in SKA applications. The paper focuses particularly on analyzing the technological innovations of the next- generation Measurement Set Version 4 (MSV4) and its potential impact on SKA data management and scientific research. The study indicates that MSV4, by introducing features such as self-description, modular architecture, and efficient storage, is expected to effectively alleviate the bottlenecks in SKA data storage and processing. If MSV4 becomes the standard for SKA data storage, it will significantly influence the construction of SRCs in China and SKA scientific

research. Therefore, it is essential to closely monitor the technological development of MSV4 at this stage and to proactively prepare relevant technological reserves to ensure the successful development, upgrading, and transformation of SRC scientific application pipelines and data processing software for SKA scientific research.

Key words astronomical data bases, instrument: interferometer, methods: numerical, data analysis

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