High Performance Data Processing Pipeline Of Chinese Solar Radio Heliograph

Fan Zhang[1], Feng Wang[1*], Wei Wang[2] , Wei Dai[1], Hui Deng[1], Kaifan Ji[1], Yihua Yan[2]
[1]Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming. 650500, China
[2]National Astronomical Observatory, Chinese Academy of Sciences, Beijing, 100087, China
Corresponding Author: Feng Wang, E-mail: wangfeng@acm.org

***Abstract*—The Chinese Solar Radio Heliograph (CSRH) is a new generation radio heliograph would produce more than 4 terabytes data every day. As a aperture synthesis telescope, CSRH is facing the challenge of processing and storing such a vast data. Pipeline system is the key issue of data automatical processing for CSRH. In this study, to push the development of CSRH, we present a framework of high performance data processing pipeline system for saving and processing real-time observation data. The related techniques of pipeline software are presented in detail including raw data acquisition, UVFITS file, data calibration, parallel computing and data publication. The pipeline has been deployed and has played an important role for the development of CSRH.**

CSRH; data processing; HPC (key words)

I. Introduction

The Chinese Spectral Radio heliograph (CSRH) is a radio heliograph with high temporal, spectral and spatial resolution which is under construction now, which will open new observational windows on flares and CMEs at radio wavelengths[1] . The site survey for the CSRH array was completed at Mingantu town (in Inner Mongolia of China) in 2008. The project was approved to start construction in the autumn of 2008. Total 100 radio antennas distributed spirally comprise CSRH (see Fig. 1.). The specifications of CSRH are listed in Table 1.

Figure 1. CSRH array. All antennas are spirally distributed.

The instrument is designed to operate at multiple frequencies in the decimetric to centimeter wave range (0.40-15.00GHz) with high spatial-spectral-temporal resolutions [2]. By the end of 2012, CSRH-I in 400 MHz -2 GHz with 40 antennas of 4.5 m, and CSRH-II in frequency range of 2 – 15 GHz with 60 antennas of 2m have been installed and are under assembly and testing. It means after 5 years hard work, CSRH entered into the testing stage. It is more important to consider the error correction, instrument calibration and observational data processing so as to promote the development of CSRH in this stage. Especially, the efficient data storage and high performance data processing are the premise of CSRH development.

As a radio aperture synthesis telescope, CSRH produce massive observation data every day. In astronomical field, the automatical data processing system, typically referred to as pipeline, is the most significant issue of modern telescope. In this paper, after the brief instruction of CSRH, we concentrate our study into the design of high performance data processing pipeline so as to guarantee the development of CSRH.

II. The Challenge of Data Processing CSRH

A. *CSRH and Basic Concept*

CSRH is a radio aperture synthesis telescope. From Fig. 2., it can be seen that the solar radio emission in 0.4-2 GHz is detected by each CSRH antenna with the broadband feed. The signal with 400 MHz bandwidth, which covers the whole 1.6 GHz bandwidth by scanning 4 times, is then transmitted through optic fibres to devices including optic receiver and analogous receivers with an output in 50-450 MHz range.[2, 3]. It is then followed by digital receiver with 1 Gsps A/D converter to receive 400 MHz analogous signal and the digital receiver outputs 16 channels simultaneously for the complex correlations with 2-10 MHz bandwidth for each channel. The time delay compensation and fringe stopping are considered in the digital correlations [4]. The whole correlation procedure is controlled by a monitoring subsystem.

B. *Massive Data of CSRH*

As mentioned in the previous section, total 100 antennas are divided into two parts: 40 low frequency antennas and 60 high frequency antennas. The digital receiver of CSRH can deal with 44 channels in the same time. Based on the principles of apertune synthesis, the number of correlated channels is 44*43/2 = 946. In order to observe and obtain the data with high temporal resolution, high spectral resolution and high spatial resolution, the data acquisition frequency is set to 3ms, and 16 channels are supported in each antenna respectively.

During observation, the size of raw data observed in 3 ms is listed as follow.

1) Power of channels: 44 * 16 * 24b = 2112 B

Published by Atlantis Press, Paris, France.

2) Corrected Data: 44 * 43 / 2 * 16 * 48b = 90816 B

Thus, in 1 second, the size of the observation data would be: (1000/3ms)*100KB = 33MB. In each observational day, 10 hours observation period, the size of the observation data would be 1.2 TB (32 MB * 3600 seconds * 10 hours) approximately. In each month, the size of the data would be about 36 TB. With the same method, we can calculate the size of raw data in high frequency array. The size is about 3.2 TB per day and about 100 TB in a month.

Obviously, the data with so large size bring us a series of problems. The most difficult issue is how to storage data into proper file format in real time.

TABLE I. CSRH SPECIFICATIONS

Item	Specification
Range	~0.4–15 GHz (:~75 –2 cm)
Frequency Res	16 channels (I: 0.4-2 GHz) 518 chan (II: 2-15 GHz)
Spatial Res	1.3” – 50”
Temporal Res.	~3ms (I: 0.4-15 GHz) ~200ms (II: 2-15 GHz)
Polarizations	Dual circular L, R
Array	I: 40×4.5m II: 60×2m parabolic Antennas
Maximum baseline	6 km
Field of view	0.6 – 7 degree

C. Raw Data and UVFITS file

Beside the storage size of device, another difficult problem is the file format conversion. Complex correlator of CSRH is implemented by FPGA technique (see Fig. 2.). In data processing, the correlator has researched its maximum performance and cannot output the file with proper format which used in radio astronomy. The corrector output a file stream with its custom format merely.

FITS[5] file format is a standard in astronomy. To share the observation data with the scientists all over the world, the file must be written with Fits format. The UVFits format, always be referred as random group structure [5, 6], was originally designed for applications in radio astronomy but was intended for other applications as well. The random groups records are a structural anomaly in FITS, as they are the only records that do not conform to the primary HDU - extensions - special records sequence.

In UVFits format, a number of keywords have been reserved for the random groups structure, and may not be used for any other purpose, unless specifically stated in the FITS rules. Also, the random groups structure is the original source of the PCOUNT and GCOUNT keywords that were incorporated into the Generalized Extensions rules.

While the basic array structure of FITS can handle a data matrix when the data are distributed evenly along all axes, sometimes the data may not be distributed uniformly along one or more axes. The random groups structure was created to handle such situations. Instead of being followed by a data array, the primary header is followed by a special set of records, of standard FITS 23040-bit size, called random groups records. These records contain a series of groups, each consisting of a sequence of parameters followed by an array. The parameters carry the data whose spacing is not uniform and which are not ordered, i.e., random, hence the name. The number of random parameters and the dimensions of the array must be the same in all groups.

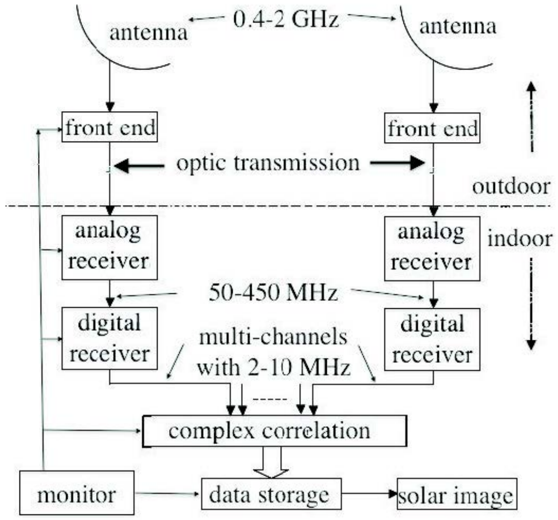

Figure 2. The system block diagram of CSRH I (Low frequency array) .CSRH II (high frequency array) is the same as the CSRH I. (cited from [2, 3])

On application is for uv interferometric visibility data; in fact, random groups are sometimes referred to as UV FITS. Consider a set of weighted complex fringe visibilities for the four Stokes polarizations at a sequence of evenly spaced frequencies. The axes are u, v, w, hour angle, baseline, fringe visibility information (real part, complex part, and weight), Stokes parameter, and frequency. The points along the frequency axis are evenly spaced, and the fringe visibility and Stokes axes can both be handled by using “evenly spaced” integer index points to represent the separate components - real, complex, and weight - for the visibility and the four Stokes parameters. Thus, an individual observation can easily be structured as a matrix with axes of visibility components, Stokes parameter, and frequency. However, the individual matrices are at nonuniformly distributed values of baseline, hour angle, and (u, v, w).

Using the random groups structure, the values of baseline, hour angle, and (u, v, w) would be specified as parameters before each (visibility, polarization, frequency) matrix. The combination of parameters and array constitutes a group. The structure of the group would be given by the form

$$\left|r_1, r_2, r_3, r_4, \ldots, r_N \right| p_{111}, p_{112}, \ldots, p_{lmn} \right| \tag{1}$$

corresponding, in this example, to |u, v, w, time, baseline|(visibility component, Stokes parameter, frequency)| where are random parameters 1 through N and are pixel values in the order defined for a data matrix. The value of p could, in principle, be as large as 998 for each axis, as opposed to the 999 maximum for Basic FITS. The data array pijk starts immediately after the last parameter, rN. The storage order and internal representation of a random groups data array are the same as for a Basic FITS simple array. It is interesting to note that the Basic FITS data structure is actually a subset of this more general structure, with no parameters.

Despite many open source software or commercial software support UVFITS file, the writing performance of UVFITS is still a critical problem. In CSRH, 16 UVFITS files should be written into storage system in every 3ms. It requires the data processing system should combine all the information in the computer memory and save the file to disk once.

However, the size of each file is less than 100KB. Current operational system is difficult to write small size file with high performance. In order to save all data needed in 3ms, a new technique should be studied deeply.

III. PIPELINE SYSTEM FOR CSRH

Pipeline system in astronomical telescope is referred as the automatic data processing system. The pipeline of CSRH is centred by the storage array and should solve the difficulties mentioned in the previous section. Fig. 3. Shows that the data received from digital receiver will be processed in the order of data acquisition platform, data preliminary processing platform, data processing platform and data publication platform.

A. Massive Data Storage of CSRH

We designed a novel storage mode for the massive data storage of CSRH. Considering the performance requirements of CSRH, we design a three-level storage mode to meet the requirements of high performance data storage. Level 1 is used for real time data storage and processing. Level 2 is used to temporarily store the data processed by the pipeline and could be used in a few days. Level 3 storage is used for long-time data archive.

Fig. 3. Shows that SSD disks are used as the storage media of Level 1 because of its high access speed that is over 500MB/s. Total 3 terabytes SSD with RAID 10 are designed in the system to meet the needs of data storage at micro second level.

Total 5 terabytes SATA hard disks with RAID 5 consist of Level 2 storage. When the daily observation finished, all data move to the Level 3 storage. Obviously, the Level 3 storage device should have big data capacity to collect the observation data up to 30 days.

Considering the need of file sharing, we finally choose NAS technique which supports file concurrent reading and writing with different computers. Meanwhile, a transaction mechanism was considered to guarantee the files' integrity.

B. Data Calibration

Due to the effect of instrument error and atmosphere, the value V'(u,v) of complex visibility function observed from telescope is not equal to the real visibility V(u,v). It will seriously influence the image quality and limit the improvement of signal error rate and dynamical range. Most issues can influence the complex visibility function, such as: geometry error (including the position of antenna and inaccurate target pointing), instrument error, water vapour in the troposphere and the ionospheric effects. In addition, the fourier transform in the mapping would bring more errors because fourier transform need to convolute the discrete visibility data to the grid. Thus, it is necessary to calibrate such visibility data V'(u,v)(including amplitude and phase)with error in the preliminary system. Besides these visibility data, the antenna array should be scaled and calibrated. Moreover, in order to make a more accurate evaluation on system gain, the system should be self-calibrated.

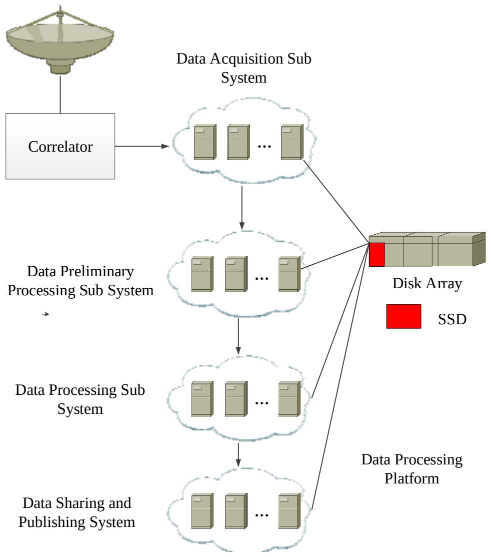
Data Acquisition Sub System
Correlator
...
Data Preliminary Processing Sub System
Disk Array
🟥 SSD
Data Processing Sub System
Data Processing Platform
Data Sharing and Publishing System

Figure 3. The flow diagram of CSRH pipeline system.

In current pipeline system of CSRH, we proposed a self calibration algorithm. The algorithm include 6 main steps: (1)select an initial model (Imod) for a source. (2)Calculate the complex visibility function of Imod and divide by the observation data. (3)Calculate gain. (4)Calculate the complex visibility function corrected. (5)Create a new model with corrected data. (6) Judge new model, if not satisfied, back to (2).

C. Image Reconstruction

The observed target of CSCH is the Sun. Among the observations, the uv data should be collected from the antennas of CSRH. After an inverse fourier transform from these uv data, we can get a spatial image named dirty map. To date, Maximum Entropy Method (MEM) and Clean algorithm are two most popular algorithms for image cleaning for image reconstructed radio telescope.

The CLEAN algorithm is a computational algorithm to perform a deconvolution on images created in radio astronomy. It was published by Jan Högbom in 1974[7] and several variations have been proposed since then[8-10]. The algorithm assumes that the image consists of a number of point sources. It will iteratively find the highest value in the image and subtract a small gain of this point source convolved with the point spread function ("dirty beam") of the observation, until the highest value is smaller than some threshold.

The MEM algorithm for astronomical imaging was proposed by Ables [11], with practical application implemented by Gull & Daniell [12]. An excellent review of astronomical applications of MEM, including radio interferometric deconvolution, was presented in Narayan & Nityananda [13]. Although this article describes the state of the art as of 1986, the situation for radio interferometric reconstruction has not changed greatly in the past 20 years. One feature of most MEM approaches is that positivity of the image in enforced. This is justifiable in most image-based applications (e.g. optical images or photon-counting X-ray images) and the positivity constraint is helpful to the convergence of these methods.

The pipeline of CSRH supports two algorithms. However, the processing performance of two algorithms is a critical issue in pipeline design. Meanwhile, whatever CLEAN or MEM algorithm, many parameters should be carefully selected. The initial image, iteration step and the noise estimation significantly all affect the final results subtlety.

D. Parallel Computing of CSRH

As mentioned in the previous section, the performance of CLEAN or MEM cannot meet the requirements of real time data processing. In CSRH pipeline, we have to fully exploit parallel computing technology in image reconstruction. A GPU parallel computing environment is a practical choice to meet the needs of CSRH because of its low cost and low power cost. Meanwhile, comparing to the FPGA technique, the programming of GPU is related simple. If the processing performance of one computing node cannot meet the requirements, more computing nodes can be easily added.

In the programming of pipeline, an open source library-Thrust (http://code.google.com/p/thrust/) is invoked.

E. Data Presentation and SSW Interface

Data Presentation is the final step of CSRH pipeline. The goal of data presentation is to provide an intuitive and easy way for the astronomy scientists to review the observation data in a short time. In addition, a J2EE-based data publishing system is implemented to provide the data sharing ability for the astronomers all over the world.

Another requirement in data publication is the interface for solarsoft system (SSW) [14]. SSW is a set of integrated software libraries, databases, and system utilities which provide a common programming and data analysis environment for solar physics. The SolarSoft environment provides a consistent look and feel at widely distributed co-investigator institutions to facilitate data exchange and to stimulate coordinated analysis. Commonalities and overlap in solar data and analysis goals are exploited to permit application of fundamental utilities to the data from many different solar instruments. The use of common libraries, utilities, techniques and interfaces minimizes the learning curve for investigators who are analyzing new solar data sets, correlating results from multiple experiments or performing research away from their home institution.

SolarSoft is primarily written in IDL which is ideal for manipulating and visualizing image, spectral, time series and other data types common in solar physics. The extensibility of IDL and large number of SSW contributors, coupled with SolarSoft automated upgrade and software exchange utilities stimulates rapid evolution of analysis capabilities.

IV. CONCLUSIONS

In this study, we discussed the challenge of processing and storing massive data produced by CSRH and introduced the pipeline system of CSRH. We present a framework of high performance data processing pipeline system for saving and processing real-time observation data. The pipeline system has been exployed in the real system and has played an important role in the development of CSRH.

ACKNOWLEDGMENT

We are also grateful to the members of the Key Lab in Kunming University of Science and Technology for their help.

REFERENCES

[1] Aschwanden and M.J, “2D feature recognition and 3D reconstruction in solar EUV images, ” Solar Physics, 2005, 228, (1), pp. 339-358

[2] Yan. Y, Zhang. J, Chen. Z, Wang. W, Liu. F, and Geng. L, “Progress on Chinese Spectral Radioheliograph—CSRH construction”, Book Progress on Chinese Spectral Radioheliograph—CSRH construction, (IEEE, 2011, edn.), pp. 1-4

[3] Yan. Y, Zhang J, Wang. W, Liu. F, Chen. Z, and Ji. G, “The Chinese Spectral Radioheliograph—CSRH”, Earth Moon and Planets, 2009, 104, (1), pp. 97-100

[4] Thompson. A.R, Moran. J.M, and Swenson Jr, G.W, “Interferometry and synthesis in radio astronomy”, Wiley-Vch, 2008.

[5] Wells. D, Greisen. E, and Harten. R, “Fits-a flexible image transport system”, Astronomy and Astrophysics Supplement Series, 1981, 44, pp. 363

[6] Greisen. E, and Harten. R, “An extension of FITS for groups of small arrays of data”, Astronomy and Astrophysics Supplement Series, 1981, 44, pp. 371

[7] Högbom. J, “Aperture synthesis with a non-regular distribution of interferometer baselines”, Astronomy and Astrophysics Supplement Series, 1974, 15, pp. 417

[8] Clark. B, “An efficient implementation of the algorithm'CLEAN'”, Astronomy and Astrophysics, 1980, 89, pp. 377

[9] Cornwell. T, “A method of stabilizing the clean algorithm”, Astronomy and Astrophysics, 1983, 121, pp. 281-285

[10] Steer. D, Dewdney. P, and Ito. M, “Enhancements to the deconvolution algorithm'CLEAN'”, Astronomy and Astrophysics, 1984, 137, pp. 159-165

[11] Ables. J, “Maximum entropy spectral analysis”, Astronomy and Astrophysics Supplement Series, 1974, 15, pp. 383

[12] Gull. S, and Daniell. G, “Image reconstruction from incomplete and noisy data”, 1978

[13] Ramesh. N, and Nityananda. R, “Maximum entropy image restoration in astronomy”, Annual review of astronomy and astrophysics, 1986, 24, pp. 127-170

[14] Freeland. S, and Handy. B, “Data analysis with the SolarSoft system”, Solar Physics, 1998, 182, (2), pp. 497-500