American Journal of Information Science and Technology
2020; 4(2): 30-40
http://www.sciencepublishinggroup.com/j/ajist
doi: 10.11648/j.ajist.20200402.12
ISSN: 2640-057X (Print); ISSN: 2640-0588 (Online)

Case Report

Construction and Management of Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform

Yang Zherui*, Gao Na, Liu Liang

Information Technology Center, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing, China

Email address:
zryang@pmo.ac.cn (Yang Zherui)

*Corresponding author

To cite this article:
Yang Zherui, Gao Na, Liu Liang. Construction and Management of Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform. *American Journal of Information Science and Technology.* Vol. 4, No. 2, 2020, pp. 30-40.
doi: 10.11648/j.ajist.20200402.12

Received: March 10, 2020; **Accepted**: April 9, 2020; **Published**: April 29, 2020

Abstract: The development of modern astronomy is rapidly and astronomical data increases exponentially. The HPC architecture based on GPU provides an efficient way of astronomic big data computing. Based on secure Ipv6 network environment, PMO has constructed the Big Data Analysis and Multi-dimensional Information Visualization Platform, which can reach the peak computing speed of 352Tflops and the totally storage capacity of 288TB. The platform is composed of 25 computing nodes, one management node and 5 storage nodes. The use of user-friendly, centralized cluster management software, the deployment of proprietary environmental control settings and multi-dimensional visualization of safety management systems form a multi-level, tridimensional and efficient management structure. A high-speed, high-capacity, highly reliable, secure and efficient high-performance computing cluster is finally constructed. The platform has a fully redundant, self-healing, highly scalable distributed storage system, a high-performance InfiniBand parallel computing and storage network, a complete job scheduling system, a cuda parallel computing architecture, and a variety of physical software tools for astronomical applications. It offers a great help to astronomical research topics such as astronomical image processing, moving target extraction, space target orbit calculation, numerical cosmology, cosmology simulation, galaxy fluid simulation. Thus it provides an important information support for the research work of 3 major breakthroughs and 5 key cultivation directions in the "One Three Five" plan of Purple Mountain Observatory.

Keywords: HPC, GPU, Cluster, Parallel Storage, Portal Batch System

1. Introduction

According to the characteristics of astronomy, astronomy mainly collects various kinds of information about celestial bodies through observation. Modern astronomy has long been a computationally intensive subject. Whether it is massive observation data processing, theoretical model calculation and numerical simulation, it requires strong computing power and storage capacity support. Faced with massive astronomical data increasing by orders of magnitude, the handling of some complex problems has been difficult to conduct through primitive experimental methods.

High Performance Computing (HPC) [1] is a branch of computer science, which studies parallel algorithms and develops related software, and is committed to the development of High Performance Computer to meet the needs of scientific calculation, engineering calculation and mass data processing. The emergence of high-performance, parallel, universal computing architectures of GPU (Graphic Processing Unit) make massive astronomical data processing possible. For example, astronomical image processing, moving target extraction, space target orbit calculation,

numerical cosmology - large scale, numerical cosmology - small scale, super-large particle beam cosmology simulation, high precision galaxy fluid simulation and other astronomical research topics.

The emergence of GPU's high-performance general-purpose computing architecture has been a boon to astronomers. In 2007, NVIDIA released a new development environment, CUDA (Computer United Device Architecture), which enabled GPU to break the limitations of graphics language and become real parallel data processing supercomputer [2]. For GPU's characteristics of high data bandwidth and digital image discretization, it is ideal for astronomical observations, especially for basic image preprocessing [3]. CPU is more suitable in terms of logical judgment, branch prediction, etc. Mass data requires extremely high data processing efficiency and we can't perform the processing and calculations without GPU supercomputing system which is suitable for preprocessing calculations of astronomical observation images, especially basic image.

In high performance computing, dozens or hundreds of computing nodes need to have a shared storage of unified image impression. The traditional solution is to connect centralized storage through an IO node, and then share the centralized storage through NFS. But when the cluster's scale increases, dozens or hundreds of computing nodes concurrently access the IO node through the network, it is easy to form a bottleneck at the IO node. At the same time, when the cluster scale increases or the IO's application demand is large, the expansion capability of a single disk array is limited, and multiple disk arrays are distributed storages for the user, then a parallel file system is needed to unifies all storage arrays into one large storage, and the distributed parallel storage system can meet this need.

And high I/O frequency of huge data exchange and distributed architecture of parallel computing brought by parallel computing of astronomical data analysis, numerical simulation require the low latency and high bandwidth of the network. InfiniBand has the advantage of response time and transmission speed [4], which can meet the requirements of the cluster for network transmission speed to a certain extent. InfiniBand is a modular, large-port switch that is flexible and reliable. Among InfiniBand products, FDR 's shipments have already surpassed QDR and become the most mainstream products. The system uses the 56Gb/s InfiniBand FDR high-speed network which has highest performance in the industry to meet needs of astronomical data exchange.

As a high-performance integrated computing platform, it may be aimed at a large user base. How to effectively manage these high-performance computing users, how to conduct reasonable privilege partitioning, how to effectively measure the resources they use, and how to analyze and record cluster's resources usage more clearly, these issues are problems for platform managers.

Configure intelligent and easy-to-use software management system to conduct unified centralized monitoring, management and use of the platform. At the same time, the management network, supporting infrastructure, and security equipment are deployed to ensure the stable operation of the platform and the information security of valuable astronomical data.

The construction of this system is of great significance to the most important astronomical research and implementation of large-scale projects in Purple Mountain Observatory (or PMO, for short). The construction of Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform will solve the problem of efficient processing of mass data, especially image data, and establish a super-computing and visualization environment adapted to the development of modern astronomy.

2. Basic Architecture

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform's system adopts Cluster structure. A cluster is a parallel processing system that is formed by many computing nodes which are connected together to work cooperatively to achieve higher performance and lower overall cost [5].

The information construction center of Purple Mountain Observatory construct Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform composed of high performance supercomputers and other main equipment and mass data storage, expand existing astronomical research mode, comprehensively adopt modern computing technology and dynamic visual analysis to carry out scientific research, build a set of high-performance parallel computing, high-speed data processing, mass storage, online interactive analysis, remote collaborative work and multi-dimensional information visualization as a professional research platform to adapt to the needs of the development of modern astronomical frontier fields.

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform further improves storage and computing capabilities, and comprehensively considers the combination of the efficiency and speed of large-scale computing. First of all, large-capacity parallel storage meets the needs of observation data processing and calculation results, and improves research efficiency. Due to the high throughput of GPU supercomputing devices, if we use wide-area network and other storage system, there will be a storage access bottleneck that can't take advantage of GPU supercomputing. Therefore, a large storage system and GPU system are connected to provide storage services, and get as much bandwidth as possible through infiniband high-speed interconnected local area network. The overall architecture topology of the system is as follows:

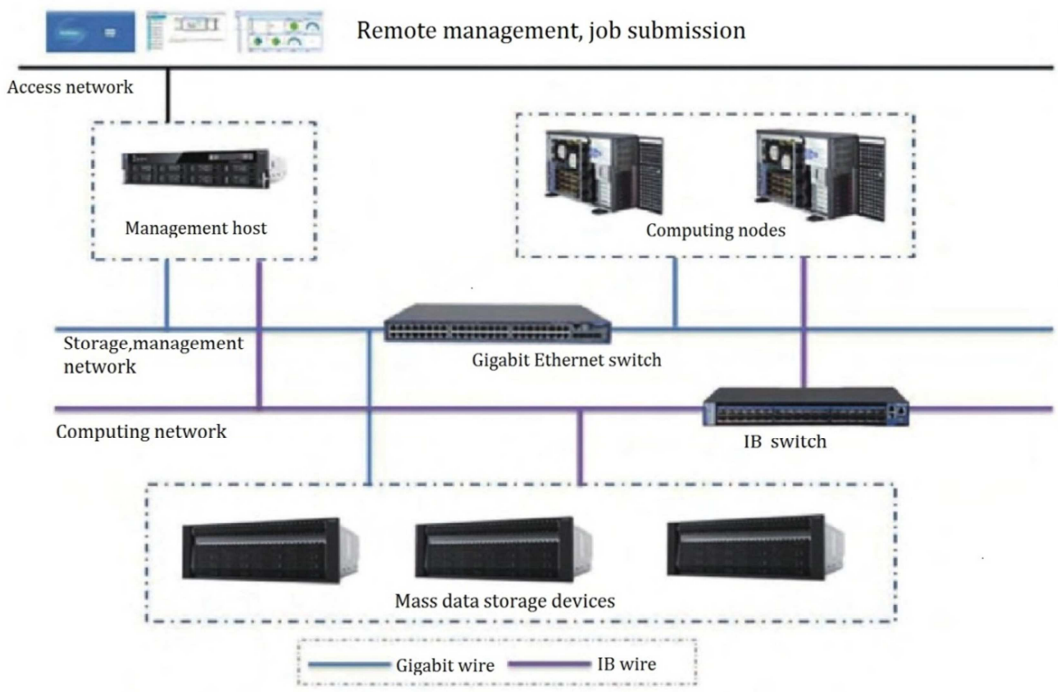

Figure 1. *The basic architecture of platform.*

3. Construction and Performance of the Platform

3.1. Leading Computing Devices and Computing Power

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform's leading computing system consists of 25 computing nodes and 1 management node.

The computing nodes are the computing core of the platform. Because the system requires high computational capacity, use high-capacity, stable and reliable, flexibly configurated new generation of SMP GPU server products which are compatible with the current mainstream 32-bit and 64-bit applications as computing nodes. Single computing node is equipped with 2 Inter Xeon E5-2670V2 processors, each cpu has 10 cores and a clock speed of 2.5GHz; 64GB DDR ECC REG memories; a 300GB SAS hot-swappable hard disk. It has been optimized for CAD, simulation and animation CG rendering, GPU computing and other applications, a single node is equipped with 4 mainstream GPU computing accelerator cards which are nVidia Tesla K20, single-node GPU offers peak double floating-point operation velocity of 4×1.17=4.68Tflops and peak floating-point operation velocity of 4×3.52=14.08Tflops. And single-node CPU offers peak double floating-point operation velocity of 0.4Tflops. 56Gb/s Infiniband high-speed network, provides high-density and high-performance platform of nodes for the GPU computing cluster. The whole system's peak double floating-point operation velocity are 25×4.68+10=127Tflops, and all nodes' floating-point operation velocity are 25×14.08=352Tflops. The whole system's peak double floating-point operation velocity in the linpack test is 81.86Tflops, and the calculation efficiency has reached 64.5% of the theoretical peak value. Test has shown that the platform has high parallel computing performance [5].

3.2. Distributed Parallel Cloud Storage System

3.2.1. Storage Architecture

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform undertakes major astronomical data processing tasks. It has a high IO application frequency, large traffic to storage, mass data, high value of data, so we have high demand for its storage capacity, accessibility and reliability.

The platform adopts distributed storage system and parallel file system, and can be extended to EB level by using horizontal expansion technology, it achieves extremely high access bandwidth and high IOPS through aggregation, so it's very suitable for the processing of massive astronomical data. At the same time, it supports the namespace of single storage and realizes concurrent reading and writing of multi-channel and multi-partition. These are also very suitable for parallel processing of GPU super-computing. And it improves capacity of single storage, achieves consistent impression and unified management of multiple partitions.

The system is configured with a distributed parallel cloud storage system with a capacity of 288TB. The system is equipped with 2 index controllers and 3 data controllers, provide external 10 56-Gb IB host interfaces and 10 1-Gb host interfaces. Clients of system management, login, and compute nodes access the parallel storage system through the FDR InfiniBand network. The storage system topology is as follows:

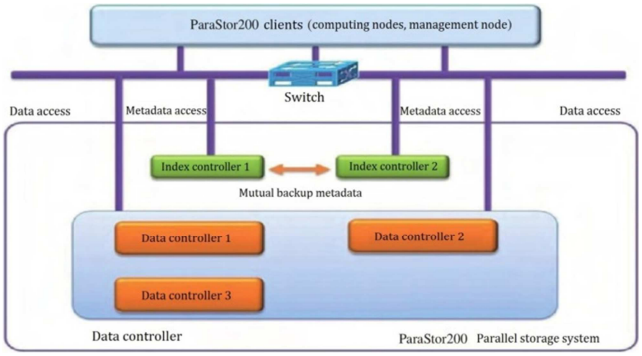

Figure 2. *Topology of distributed parallel storage system.*

3.2.2. Index Controller

Index data of the storage system, including key data such as catalog, file copy and other statistical description information. The index controller provides an access service for index data.

The distributed parallel cloud storage system uses double index controllers architecture, and two controllers are used simultaneously, providing an almost unlimited single namespace and high service capability of index data. No shared device is used between the two index controllers, but they are backups of each other and back up each other's data and services. The two nodes send heartbeats to each other. Once a node determines that the other one has failed, it immediately takes over the failed node and uses the all data stored locally to provide external services. This switching is accomplished automatically by the client, and the upper application is basically imperceptible. It will not cause interruption of the front-end application.

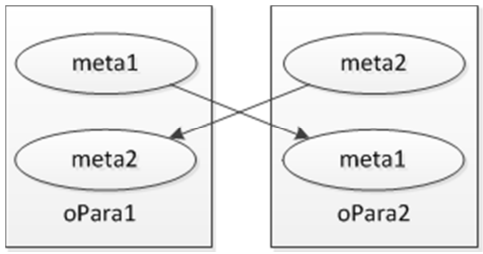

Figure 3. *Two index mutual backup controllers.*

The index data is stored as normal files in the local storage device of the index controller, and the data in the group is synchronized through the log to ensure the reliability of the index data.

Take the use of flexible directory indexing technology to achieve dynamic adaptation of files and data blocks, migration without data block of file cuts in the directory and so on; achieve fast mapping of file paths to index data positions, enabling the number of subdirectories or files supported by each directory reaches tens of millions, and the parsing operation still maintains a fine response time.

If a data controller or disk's fault causes a data node's fault, clients will automatically selects other available nodes for operation, and the upper layer application basically could not perceive such an operation switch. If a controller is just temporarily disabled, and operations of other nodes are still normal, when the controller is back to be online, the affected data is incrementally updated to the latest status.

3.2.3. Data Controller

Members of the data controller cluster provide data storage services in a full equivalence manner. Each controller saves a part of the data of the entire namespace on its own storage device, and directly provides data reading and writing services. Data controllers are interconnected through multiple networks to increase system's bandwidth and increase network's availability. No parts are shared between the data controllers.

The storage system is created by replicas, that is, when the data on a node is frequently visited and the server node is overloaded, or when reliability is considered, one or more copies of the data can be copied and stored in other nodes. Only send a copy to the requesting node, when the access request reaches the target node, if the target node is not overloaded, the data can be read, If the target node has insufficient processing capacity, a new copy will be created, and if the requesting node is not overloaded, the new copy is sent to the requesting node and tells all nodes on the request path that the data copy is also available on the requesting node. If a controller is just temporarily disabled, the operation of other nodes is still normal. When the controller is back to be online, the affected data is incrementally updated to the latest status. The parallel reconstruction strategy implemented by the storage system greatly speeds up the recovery time.

The built-in automatic fault detection mechanism is implemented in the parallel storage system, and the parallel reconstruction strategy is implemented. Once the fault is detected, the data recovery process is automatically started. There is no need to add new hardware during the recovery process, and the data on the damaged device will be restored to the storage server without damage. The system has the ability to automatically discover components' failures based on system information and operational results. When the media is damaged, the available data is reduced, and the system automatically uses the available nodes' data to generate new data, so that the whole system could returns to normal. During the recovery process, reading and writing of damaged data can still be performed. If an entire data controller is damaged, the damaged data is also automatically restored as described above.

The entire recovery process is performed automatically at the back-end without the need for administrators. Once the data recovery is completed, the entire system becomes the highest security state immediately, and then even hardware damage's appearance will not result in data loss. The fully automated data recovery process not only improves the reliability of the system, but also greatly reduces the work intensity of the administrator.

In a cloud storage environment, any large-scale nodes failure caused by a storage failure, and thus the loss caused by the interruption of all online services is incalculable and unacceptable. The cloud storage of the platform is designed with a fully redundant architecture, and load balancing, redundancy design and fault's self-healing functions are implemented in each link to prevent from service interruption

caused by storage failures. In addition, the adoption of the intelligent throttling mechanism avoids the appearance of the avalanche effect.

The distributed storage system has excellent scalability, supports online expansion which does not affect the use of business systems. After the addition of data controllers, the data objects automatically implement migration distribution of load balancing, enabling the entire storage system to achieve linear growth in capacity and performance.

Cluster's management node, login nodes, compute nodes, remote visualization nodes, etc., act as clients of cloud storage, access the parallel file system through proprietary protocols (kernel state) and InfiniBand network, and can support both Linux and Windows clients.

In addition to proprietary protocols, the parallel storage system also supports standard NFS, CIFS interface, POSIX API, MapReduce programming interface, REST programming interface, SOAP programming interface, and SNMP interface. It has wide adaptability, and MapReduce programming interface supports big data usage mode, has higher performance and reliability than traditional HDFS.

3.3. Network Equipment

The design of the platform's network system is based on two principles which include performance and reliability, it uses two sets of networks. The network configuration takes into account the expansion margin for system's expansion.

The network scheme design fully takes into account the performance parameters of the computing nodes, storage nodes, and network switches. On the premise that we ensure the stability of the system, each device can be fully utilized.

3.3.1. Computing and Storage Network

The parallel high-performance computing program such as MPI has a large amount of network communication data while multiple nodes are running concurrently, it is very dependent on the bandwidth and delay performance of the computing network. The performance of the computing network has a decisive impact on the computing performance, parallel speed ratio and scalability of parallel programs. On the other hand, large-scale high-performance computing clusters currently use distributed parallel storage architecture, to reflect the advantage of its I/O performance, we require low latency and high bandwidth of storage network.

The networking mode of using the large-port modular core layer InfiniBand switch make wiring simple, it is easy to maintain and manage. The small switch stacking solution not only has complicated lines, but all cables may need to be re-routed and adjusted when the system is expanded. InfiniBand employs the design that the backplane access page board wirelessly cable for internal switching, it has small high-speed signal loss, higher reliability, better manageability and better expansibility of system.

Therefore, we choose the system solution with 56Gb/s InfiniBand FDR high-speed network which has the highest performance in the industry, and use it as a computing network for parallel computing programs and a storage network for parallel storage system. Configure a modular FDR InfiniBand switch with 108 FDR ports to achieve 56Gb/s FDR wire-speed switching of system nodes. We totally use fiber-optic cables in the system. Compared with copper cables, fiber-optic cables have longer connection distances, are less prone to have breakage, are easier to route and maintain and can be easily replaced when cables have faults. Compared with previous generation 40Gb QDR Infiniband network, 56Gb/S FDR has a very large improvement of performance. This benefit from the following aspects:

Table 1. *Comparison between QDR and FDR.*

	Transfer performance	Network coding efficiency	PCI-E coding efficiency	Network delay
QDR	40Gb	8/10	PCI-E 2.0 (8/10)	1.4us
FDR	56Gb	64/66	PCI-E 3.0 (128/130)	0.7us

Taken together, due to the effect of coding efficiency, the bandwidth of QDR network can only reach:

Bandwidth (QDR)=40Gbps×0.8×0.8/8=3.2GBps

And the bandwidth of FDR network can reach:

Bandwidth (FDR)=56Gbps×(64/66) ×(128/130)/8=6.6GBps

The FDR network's bandwidth is twice as that of QDR, while the delay is only half of QDR's. The FDR InfiniBand's 56Gb/s bandwidth and 64bit/66bit encoding mode achieve nearly 100% of the InfiniBand transmission efficiency, 700 ns delay of point-to-point take common network into a nanoseconds era at the first time, implement an order of magnitude innovation on application delay. It's high bandwidth and low latency can meet the requirements of large-scale parallel computing to a certain extent, and the topology of the network also has a direct impact on performance of network and scalability and management of cluster. With regard to large cluster systems, the network interconnection structure between the switches must be considered. Generally, tree topology has been widely used due to its advantages of clear structure, being easy to construct and manage. Fat tree is a type of tree topology, which allows each layer of network has equal bandwidth, thus providing a non-blocking network transport to effectively improve the communication congestion situation.

3.3.2. Management Network

The platform has large scale, a large amount of nodes, high requirement for performance of management network, Gigabit line speed switching, to ensure the clear management traffic of large-scale cluster. Configure two high-end Gigabit switches and link them to two independent Gigabit networks. The switch is 1U high and has 48 Gigabit ports, which ensures the connectivity of all nodes in the system. One set of Gigabit network is used for data communication such as job delivery, job monitoring and management, etc. One set of network is

used for data communication such as system management, system monitoring, IPMI hardware management and so on.

4. System Management and Maintenance

We have configured one server as the cluster's management host. The management node use the management network to run the cluster's monitoring management software, user information management service, InfiniBand subnet management service, job scheduling service, time synchronization service, storage system management software, system hardware management software and other system-level service processes, and users' programs compilation, computing job preparation, file uploading and downloading, job submission and control and other users' interaction operations.

The external of the cabinets is equipped with a remote video monitoring system, SMS alarm system, firewall and intrusion prevention system, etc. These installations together with the system software are used to monitor and maintain the hardware facilities of the cluster and ensure cluster's information security.

4.1. Cluster Management System

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform uses visualization system management software to provide users and administrators with an easy-to-use, user-friendly, unified and centralized cluster monitoring, management and using platform. The cluster management system includes functions such as system deployment, system monitoring, cluster management, alarm management, statistical reporting, and job scheduling. It possesses integrated interfaces with a variety of management tools.

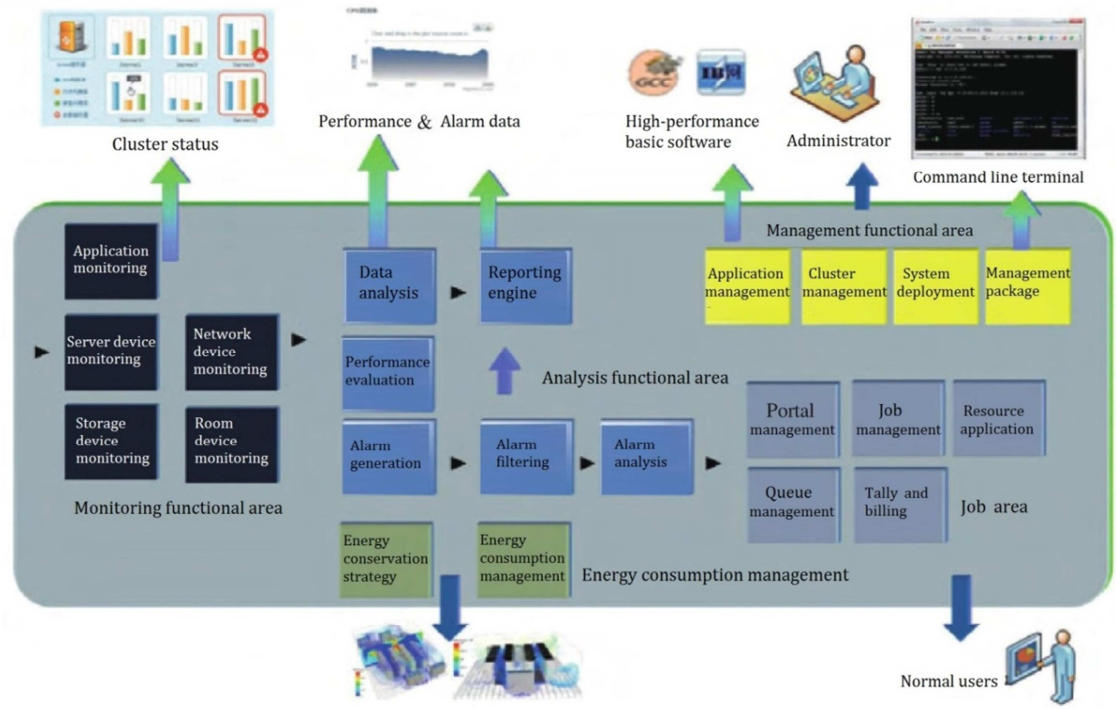

Figure 4. *Architecture intelligent cluster management System.*

4.1.1. Multi-level Monitoring and Management

The cluster management system supports the physical topology of the cabinet mode, displaying visually and graphically according to the actual location and corresponding size of the device, so that the administrator can see the entire system at a glance. We take real-time monitoring of operational status of rooms, servers, network devices, storage devices and other equipment, including cluster load, node hardware information, GPU monitoring. And monitor operational status of various application programs. The management system graphically displays real-time alarms, sets various alarm thresholds, multiple monitoring policies, and flexibly adjusts alarm conditions. It can collect and analyze historical faults, perform list query and statistical analysis. The report system of the cluster management system generates content-rich reports, and generates reports of multiple time periods such as hourly, daily, monthly, and annual reports in a predefined form. Tabular information retrieval supports compound logic queries for multiple fields. Supports statistics from different perspectives (user statistics, node statistics, time statistics, etc.), and provides summary tables and details. Correlation report can be used to generate multiple device metrics in one report and perform multi-device comparison analysis. Multiple metrics of one device can also be generated in one report, so as to analyze the correlation between multiple indicators [7]. Reports can be exported to PDF, Excel, and HTML files for users to view in a

variety of forms and intuitively. The reporting system provides an extension method for secondary development.

The cluster administrator views the configuration of users and groups in the system through the cluster management system, and achieves addition, deleting, and modification to the users and groups of the cluster. Display the information of processes running in the cluster, run specific processes on selected node groups, and kill specific processes, save current processes' information, etc. View information of services running in the cluster, start, stop, and disable specified system services on selected node groups. The cluster management system uses a browser to uniformly and intuitively manage 25 computing nodes. Commands can be executed concurrently on multiple nodes, and SSH terminals can be provided to log to the managed node directly to perform various operations and start or shut down the selected nodes or the entire cluster quickly.

4.1.2. Job Scheduling System

The job scheduling sub-module of the cluster management system provides a complete job scheduling function, which can effectively manage and allocate the system's hardware and software resources.

According to the static attributes of the job and the system status, various indicators are synthesized to determine the scheduling priority of the job. Static attributes include the user (group), queue, job type, QOS, resource request, etc. Dynamic attributes include the number of jobs ran by the user, the amount of occupied cores, the amount of memory used, and the number of time of failed scheduling. The job scheduling system support multiple node allocation strategies, such as CPULOAD (by load), FIRSTAVAILABLE (in positive order), LASTAVAILABLE (in reverse order), PRIORITY (by flexible matching priority), MINRESOURCE (by minimum matching priority), MAXBALANCE (node equalization), FASTEST (by processor speed), etc. High-priority jobs are allowed to be executed immediately without prior reservations, replacing other low-priority jobs which are running. After the high-priority job ends, the low-priority job can resume and continue to run. Job priority levels can be dynamically adjusted based on current users and users group usage levels to meet their usage requirements.

Based on core scheduling strategies and algorithms, the job scheduling system provides a flexible priority definition mechanism. The priority of a job is the sum of the priorities contributed by the various components. Weight of each type of component and sub-component can be defined separately. The priority of each sub-component's contribution is the product of its own weight and its own value. The priority of each component's contribution is the product of the component's weight and the sum of the sub-components' priorities. According to the characteristics of it own system's operation, the most suitable scheduling strategy is flexibly customized to maximize the operating efficiency of the cluster system.

System resource management function is powerful, and it is convenient to set attributes and rights of users, queues, nodes, and so on. The tasks are divided into multiple queues to manage according to the user group and the subject group, different queues are set with different management policies, user's priority is dynamically adjusted according to the operation status of the user's jobs. And set management strategies such as resource restrictions and priorities for different users or user groups according to importance.

The job scheduling system supports the expansion functions of the charging money and billing for users or user groups, and they are implemented by the sub-module of cluster quota system, and charging and billing can be managed according to computing resources of users and user groups. The cluster quota system provides charging service of high-performance cluster. Using a unified quantitative method to measure, we can bill the computing resources by quantizing according to the unit time.

The job scheduling system implements the user quota system, carries out pre-allocation and real-time billing for available resources of the user, and flexibly controls the effective time limit of the user quota, thereby achieving fine-grained resource accounting and quota coordinating, and accurately recording and controlling user resource usage, providing a strong guarantee for external billing and internal accounting. And implement global, comprehensive, dynamic, and fine-grained statistics of usage of the cluster, find resource bottlenecks that affect system performance through analysis,, and then achieve reasonable allocation of computing resources, enable the cluster achieves the optimal use effect and the highest use value. With the continuous promotion and centralization of high-performance computing, in the furture when the volume of platform's users' application grows quickly, the refined management and control of users can be realized.

The job scheduling system supports the Web Portal (ClusPortal) system extension for various high-performance computing applications. It is implemented by the application portal system sub-module and multiple Web Portal job submission interfaces are customized for various HPC application software. For ordinary users, the use of high-performance computing clusters still has certain technical thresholds. Through ClusPortal application portal system, a graphical job submission interface similar to Windows is provided for users, which can greatly reduce the difficulty of using high-performance computing cluster, meet the user's usage habits, and provide the success rate and efficiency of job submission. ClusPortal supports visual job interaction, achieve limit judgment, file transfer, and automatical switch and restart for faulted job. Its versatility and ease of operation: Portal can be applied to all serial, multi-thread, MPI programs including most high-performance applications and it can be customized as needed; the submission parameters are set reasonably, 90% of users ' jobs only need 5 options modified. The job submission interface reduces difficulty for users and improves job submission efficiency.

4.2. Storage Management Platform

The system provides a unified monitoring and management

platform based on Web to carries out simple deployment, maintenance and management functions for the distributed storage system. The storage management platform's intuitive and easy-to-understand graphical interface makes it easy for users to manage and monitor the system's hardware and software resources.

The platform manages the storage network and monitors storage nodes' service statuses, disks, memories and metadata controller's RAID cards' status. The faulty disk can be monitored at the actual position and system fault alarms are performed in various ways such as interface and mail. Operation and maintenance reports are generated to manage events records.

Manage the start, stop, uninstall, and upgrade of the storage system, forcibly start the system under abnormal conditions. Authorization, mounting, and status management of clients of storage system are performed. And Addition, deletion, start, stop, and replacement of each management node, index controller, and data controller are performed.

At the same time, quota management of the storage system, file system creation, deletion, configuration of file system, online parameter configuration, threshold management, and resource configuration are implemented.

4.3. Supporting Safety Management Facilities

The construction and stable, efficient operation of a high-performance computing center require a set of reliable information system infrastructure as a support [8], which can ensure efficient, stable and reliable operation of various electronic devices in the high-performance computing center.

Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform is deployed with abundant supporting infrastructures. Together with Linux shell scripts and cluster management system, they provide comprehensive monitoring and protection for system hardware.

The computing density and storage density of the cluster are high, and continuous power supply and temperature control of the equipment are important prerequisites of ensuring security and stability of the hardware. The total power consumption of IT equipment in the high-performance computing center of this project does not exceed 45KW, and the installation space of the equipment is 125U.

In order to ensure uninterrupted power supply of IT equipment, mains voltage regulation and suppression of interference of power grid, the UPS power system is deployed. Moreover, the air conditioner indoor unit (single unit's power consumption is 2.5KW) is also powered by the UPS power supply. In this way, when the mains power is cut off and the air conditioner outdoor unit is shut down, the air conditioner indoor unit fan can continue to operate to maintain air circulation and heat dissipation. The air conditioner, power distribution, cabinets, and temperature and humidity monitoring of the cluster are deployed in an integrated manner, and the circulating cooling in the cabinet is adopted, thus we can effectively remove the bottleneck of high-density refrigeration and reduce the dependence on the environment of the equipment room. The system is also equipped with SMS alarm facilities. Combined with the monitoring and management module of the integrated refrigeration system, the monitoring strategy and high temperature threshold are set in the cluster management system to realize the high temperature alarm of the system. Once the equipment in the cluster exceeds the set temperature, the cluster management system immediately activate the SMS alarm facility to send an alarm message to the system administrator, monitor the device temperature in real time.

The platform is equipped with an independent video surveillance system for remote room monitoring. Deploy a video surveillance server and two 720-color hemisphere infrared cameras to achieve 360-degree full-angle zoom observation, managers can use the personal terminals to log in to the monitoring server, use portal video surveillance software to dynamically monitor of the equipment room, observe equipment appearance in real-time.

Video surveillance system has 8-channel high definition digital video recorder and 2T hard disk, is able to save surveillance video recording of the equipment room for one month, thus it is easy to check back. The firewall and intrusion prevention devices are installed at the entrance of the platform and monitor the upstream and downstream traffic of the platform at the network layer and the application layer to ensure system's information security.

We write Linux scripts for mail monitoring. PBS (Portal Batch System) is a system for managing tasks and computer resources. The cluster is installed with open source free PBS. Write the shell script based on the commands provided by PBS, thus the node statuses are traversed by the script every 15 minutes. If there is a node exception, the mail is sent to the specified mailbox as an alarm immediately, and the mail of node execution status is sent once every day and afternoon. Thus system managers can keep track of node operation and job execution even they leave the cluster.

Because the management node of the platform is configured with the public network address, the file system of the platform shared through nfs is vulnerable to network attacks. The firewall device is used to disable all ports except the high-performance application and the mail service to further ensure the information security of the platform.

5. Scientific Research Application at PMO

5.1. Introduction of PMO Supercomputer Application Environment

Based on Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform, many important and innovative research progresses have been made in terms of image processing and analysis and numerical simulation. The platform provides cuda parallel computing architecture and is installed with MPI parallel libraries such as OpenMPI, IntelMPI, and MVAPICH. The platform is deployed with a variety of physical software tools for astronomical applications: tools such as Geant and Fluka

which provide C++ class libraries, Fortran program interfaces which simulate the transport process of particles in matter, Root software which provides particle physical data structures and data processing, SKT tools used for satellite orbit simulation and visualization, Python and other software for display, CERNLIB and other high-energy physics software, as well as Spectrogram, SASSW and other self-developed or customized astronomical data processing and graphics software. Users take advantage of a large number of interfaces provided by the software and Python, c/ c++, Fortran, shell and other procedural language to write tasks and submit them to the platform by PBS or MPI for calculation, or realize astronomical data processing, display and research through the modular function of custom astronomical software. So great progress has been made in researches on protoplanetary disks [9], thermal physical model [10], upward overshooting in turbulent compressible convection [11], adaptive optical aberration correction [12], microwave SAIR imaging approach [13], terahertz superconducting imaging system [14]. The integration of the basic platform and these astronomical application software make up a comprehensive information system which can effectively make use of the underlying information infrastructure and provide an important information support for the research work of Purple Mountain Observatory. Here we introduce one typical scientific research application in each aspect of CPU and GPU application.

5.2. Scientific Research Application Examples

5.2.1. Dark Matter Detection

Dark Matter Particle Explorer (or DAMPE, for short) is a large space electron and Gamma-ray space telescope. Before the construction of the dark matter detector, it is necessary to fully understand its performance through computer simulation and optimize the design parameters. We have installed a newer version of the GEANT4 [10] software package on Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform of Purple Mountain Observatory. And with the addition of high-energy physical software packages such as CLHEP and CERNLIB, we can simulate the entire detector comprehensively, make appropriate selection of the type, energy spectrum and incident direction of the incident particle, thus we can approximate the real situation in space and optimize our design through simulation results to fully meet our requirements in terms of energy resolution, spatial resolution, and energy measurement range. For the case in which the incident particle's energy level is extremely high, the exponential growth of the shower in the detector makes CPU usage time of process of simulation computing too long. For example, for an electron with an incident energy of 600GeV, a simulation of a single instance would take a few minutes. If you need millions of such cases, they can hardly be completed in a short time with a single CPU. Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform of Purple Mountain Observatory has hundreds of CPUs running at the same time, which can reduce the total time consumption by more than two orders of magnitude, thus ensuring the task can be completed in a short time, as shown in Figure 5. In the process of operation of dark matter detection satellite, operation management, data processing management, scientific application and research need to be achieved. We installed a custom-developed dark matter data processing software package named as SASSW on Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform: based on the Web mode, the payload scientific detection data, satellite engineering telemetry data, orbital ephemeris data are processed in multiple stages such as CRC check, optimizing, calibration analysis, etc.; the energy deposition and ignition point distribution are quickly displayed in graphic mode; satellite model and the energy distribution are displayed by multidimensional model; and the species identification of the incident cosmic ray particles [15] is achieved. It has built an integrated scientific application system for operation, data management and processing of dark matter satellite detection [16], which provides a solid guarantee for the dark matter satellite project.

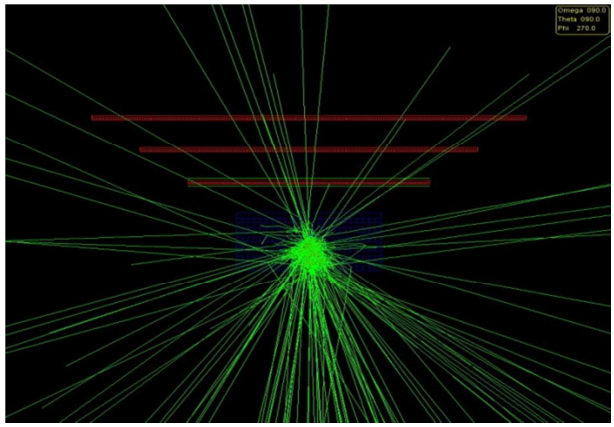

Figure 5. *Results of dark matter particle detector of performance on the platform.*

5.2.2. Mobile Celestial Target Extraction

Purple Mountain Observatory successfully developed the most powerful optical survey telescope in China, "1-meter Near Earth Object Survey Telescope" [17]. The equipment collects massive image data, distinguishes moving objects (including artificial celestial bodies and natural celestial bodies) from images, and measures them. With the deepening of the new generation of photovoltaic arrays and photoelectric hedgerow projects, the total data volume of these devices has surged to 15TB/night. The emergence of GPU high-performance general purpose computing architectures satisfies the computational needs of in-depth research.

Firstly, the high data bandwidth makes the image data transfer rate between the memory and the processor very high. Secondly, discretization of the digital image makes it very easy to parallelize the calculation process, which means hundreds of pixels on the GPU can be processed concurrently, thus the computational efficiency can increases dozens of times as before. While CPU has strength in terms of logical judgment, branch prediction, etc. The application software can be generally divided into a CPU (Host side) part and a GPU

(Device side) part. Each part contains several modules: The CPU side includes Device scheduling module, moving target recognition module, astronomical positioning module and so on; the GPU side includes background calculation module, stellar image search module, and a stellar image structure parameter calculation module. Each module has a mature algorithm, the main work is parallelization for suiting the computing situation of the GPU, and to give full play to the computing power. The observation images of space debris and near Earth objects are distributed in a certain form (such as put a single barrel as a distribution principle), and a data stream is formed in order of time to be allocated to each computing node, image preprocessing, background calculation, stellar image scanning, stellar image structure parameter calculation are proceeded by means of massive parallel threads brought by flexibly dividing pixels of images to different blocks of cuda architecture by GPU, in the meantime calculations of mobile celestial object identifying, astronomical poisoning are proceeded by CPU. Figure 6 shows the image processing which achieved capturing 4 targets in a field of view proceeded by Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform.

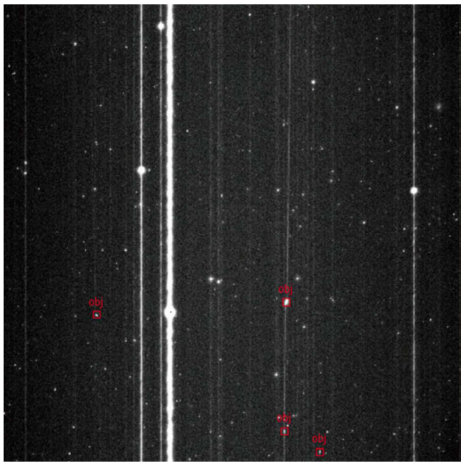

Figure 6. *Image processing to achieve capture 4 goals use platform in a field of view simultaneously.*

6. Conclusion

Facing the rapid development of astronomy in the era of big data, Purple Mountain Observatory of Chinese Academy of Sciences has built Astronomical Big Data Analysis and Multi-dimensional Information Visualization Platform, and the specific implementation work has been undertaken by Information Technology Center. The platform use cluster mode, GPU high-performance general-purpose computing architecture efficient distributed parallel cloud storage system and the most advanced 56Gb/S FDR Infiniband computing network, it is also deployed with unified centralized intelligentized management software and a set of comprehensive and reliable information system infrastructure as a support, thus ensure the safety and stability of equipment in multidimensional ways. The platform is able to efficiently process astronomical scientific data and visualize mass data, thus it will play an important supporting role in research projects, greatly improve the efficiency of scientific research, shorten the calculation cycle, realize the computation-storage-visualization process, scientifically reflect the scientific content, further accelerate the generation of scientific research results and achieve resource sharing to enhance the scientific research capability of the whole institute. The construction of the platform provides a most important, fundamental support and guarantee for the scientific innovation of Purple Mountain Observatory.

Acknowledgements

This project was supported by Ministry of Finance of the People's Republic of China and Computer Network Information Center, CAS.

References

[1] Zhou X. Development of high performance computing technology [J]. Journal of Nature, 2011, 33 (5): 249-254.

[2] Guo J, Cai W. A new generation of high energy computing technology—introduction of CUDA [J]. Modern Science and Technology, 2009, 8 (6): 58-62.

[3] Zheng Y, Gao N, Liu L. Astronomical scientific applications based on GPU supercomputing system. [J] E-science Technology & Application, 2011 Vol. 2 (6): 144-152.

[4] Dong Y, Zhou E, Chen J. Construction of high performance distributed file system based on InfiniBand technology- Lustre [J]. Computer Engineering and Applications. 2005 (22): 103-108.

[5] Liu Y, Zhou J, Xie K, Zhao Y. High performance computing platform of electric system based on cluster technology [J]. Computer Simulation, 2005, 22 (2): 239-243.

[6] Huang K, Xu Z. Lu X et al. Extensible parallel computing technology, architecture, and programming [M]. Beijing: China Machine Press, 2000.

[7] Sugon. Sugon 5000A high performance computer product sample, 2008.

[8] Wang J. Infrastructure Construction and Management of Computer Rooms in Shanghai Supercomputer Center [J]. Building Electric, 2011 (9): 69-72.

[9] Huang P. H., Dong R. B., Li H. et. al., The Observability of Vortex-driven Spiral Arms in Protoplanetary Disks: Basic Spiral Properties, ApJL, 2019, 883: L39.

[10] Jiang Haoxuan, Yu Liangliang, and Ji Jianghui, Revisiting the Advanced Thermal Physical Model: New Perspectives on Thermophysical Characteristics of (341843) 2008 EV5 from Four-band WISE Data with the Sunlightreflection Model, AJ, 2019, 158: 205.

[11] Tao Cai*, Upward Overshooting in Turbulent Compressible Convection. III. Calibrate Parameters for one-dimensional Reynolds Stress Model. The Astrophysical Journal, 891, 77, 2020.

[12] Wang H. Characterization Method of Free Vibration Mode Phase Difference of Ring Primary Mirror: CN, 201910174229. X. 2019-06-25.

[13] Yilong Zhang (*), Yuan Ren, Wei Miao, Zhenhui Lin, Hao Gao, Shengcai Shi, Microwave SAIR Imaging Approach Based on Deep Convolutional Neural Network, IEEE Trans. Geosci. Remote Sens., 2019.8.28, 57 (12): 10376-10389.

[14] Yilong Zhang (*), Yuan Ren, Wei Miao, Hao Gao, Shengcai Shi, A Terahertz Superconducting Single-Pixel Imaging System Using DMD, 44th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), 1-6 Sept. 2019, Paris, France.

[15] Dark matter particle detection satellite "WuKong" obtains the most accurate high-energy electron cosmic ray energy spectrum up to now [J]. Bulletin of Chinese Academy of Sciences, 2017, 32 (12): 1265.

[16] Chang J. Indirect Detection of Dark Matter Particles in Space [J]. Aerospace Shanghai, 2019, 36 (04): 1-8.

[17] Purple Mountain Observatory 1-meter Near-Earth Object Survey Telescope + 4k × 4kCCD Trial Observation Success [J]. Mechanical Engineer, 2006 (12): 16.