计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900042

高性能计算与天文大数据研究综述

汪 洋[1] 李 鹏[1,2] 季一木[1,2] 樊卫北[1,2] 张玉杰[1,2] 王汝传[2] 陈国良[1,2]
1 南京邮电大学计算机学院 南京 210023
2 江苏省无线传感网高技术研究重点实验室 南京 210023
(xlewangyang@163.com)

摘 要 数据是天文学发展的重要驱动。分布式存储和高性能计算(High Performance Computing,HPC)为应对海量天文数据的复杂性、不规则的存储和计算起到推动作用。天文学研究中多信息和多学科交叉融合成为必然,天文大数据已进入大规模计算时代。高性能计算为天文大数据处理和分析提供了新的手段,针对一些传统手段无法解决的问题给出了新的方案。文中根据天文数据分类和特征,以高性能计算为支撑,对天文大数据的数据融合、高效存取、分析及后续处理、可视化等问题进行了研究,总结了现阶段的技术特点,提出了处理天文大数据的研究策略和技术方法,并对天文大数据处理面对的问题和发展趋势进行了探讨。

关键词:天文大数据;高性能计算;数据存储;数据处理;数据可视化
中图法分类号 TP3-05

High Performance Computing and Astronomical Data:A Survey

WANG Yang[1],LI Peng[1,2],JI Yi-mu[1,2],FAN Wei-bei[1,2],ZHANG Yu-jie[1,2],WANG Ru-chuan[2] and CHEN Guo-liang[1,2]
1 School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks,Nanjing 210023,China

Abstract Data is an important driver of astronomical development. Distributed storage and High Performance Computing (HPC) have an positive effect on the complexity,irregular storage and calculation of massive astronomical data. The multi-information and multi-disciplinary integration of astronomical research has become inevitable,and astronomical big data has entered the era of large-scale computing. HPC provides a new means for astronomical big data processing and analysis,and presents new solutions to problems that cannot be solved by traditional methods. Based on the classification and characteristics of astronomical data,and supported by HPC,this paper studied the data fusion,efficient access,analysis and subsequent processing,visualization of astronomical big data,and summarized the current situation. Furthermore,this paper summarized the technical characteristics of the current stage,put forward the research strategies and technical methods for dealing with astronomical big data,and discussed the problems and development trends of the processing of astronomical big data.

Keywords Astronomical big data,High performance computing,Data storage,Data processing,Data visualization

1 引言

近年来,观测技术和手段的发展使得天文学所采集的数据快速增长,天文学数据的发展遵循摩尔定律:数据量每20个月就会翻一倍。同时,天文数据的管理系统也从传统的文件系统的管理转变为了目前的基于关系数据库的管理。2000年,斯隆数字(Sloan Digital Sky Survey,SDSS)巡天项目正式启动,该项目所采用的望远镜在几周之内采集的数据已经超过了此前天文学历史上的数据总和[1-2]。截至2010年,该项目所产生的数据已经超过1.4X242B,而这不过是位于智利的大型视场全景巡天望远镜(Large Synoptic Survey Telescope,LSST)在短短5天内所获得的信息量,该望远镜每晚所观测的原始数据就高达15 TB。2019年,我国设计生产的郭守敬光谱巡天望远镜(Large Sky Area Multi-object Fiber Spectroscopic Telescope,LAMOST)正式投入运行,LAMOST每晚都将拍摄并产出多达20GB的光谱数据。即将投入使用的地面广角相机阵(Ground-based Wide-angle Camera Array,GWAC)等巡天项目每天都会产生海量的天文数据[3]。所观测到的天文数据相当复杂,除了望远镜所观测到的光谱数据、天体图像数据外,还包含红外线数据、紫外线数据、温度数据

到稿日期:2019-07-01 返修日期:2019-09-05 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划项目(2018YFB1003201);国家自然科学基金(61672296,61602261,61872196,61872194);江苏省科技支撑计划项目(BE2017166,BE2019740)
This work was supperted by the National Key R&D Program of China (2018YFB1003201),National Natural Science Foundation of China (61672296,61602261,61872196,61872194),Scientific and Technological Support Project of Jiangsu Province (BE2017166,BE2019740).
通信作者:樊卫北(wbfan@njupt.edu.cn)

等，而用于观测宇宙粒子的卫星探测器则会产生大量的粒子数据。海量的数据为发现宇宙规律提供了支撑，计算科学加速了数据处理。例如，通过图像处理、统计理论、隐马尔可夫模型等方面的相关算法对望远镜拍摄的大量图像数据进行处理，再通过高性能计算的加速，就得到了轰动一时的黑洞照片；紫金山天文台通过“悟空号”探测卫星采集到的数据合成了电子穿过卫星探测器的轨迹图片。

与一般的大数据学科不同，天文信息学是一门综合了天文学、天体物理学、统计学、计算机科学、工程学及信息学等诸多学科的新型科学[4]，传统的对数据进行存储、传输、分析和挖掘的方法不能直接运用于天文大数据领域，因此，如何将天文大数据与高性能计算相结合，利用高性能计算的特点来解决天文学的难题，是计算机科学与天文学两个领域的专家所要共同探讨的话题。

2 高性能计算与天文大数据

2.1 高性能计算

高性能计算是指提供超过平均水平资源的不同形式的计算[5-8]。随着计算总体性能的增长，特定的级别不断变化。超级计算属于 HPC 的一种，它通常通过非标准架构（例如台式计算机）的计算基础设施来实现这一目标。网格计算是 HPC 的另一种形式，它通过连接相对标准化的计算机（通常由各种计算框架支持）来进行分布式应用。传统上，HPC 以低延迟、高吞吐量、大规模并行和大规模分布式系统为特点。

国内外对 HPC 技术也越来越重视。在 2000 年左右，美国政府联合各部门推出了一系列有关高性能计算的计划，如联邦政府的 HECC 计划[9]、能源部的 ASCI 计划以及国防部的 HPCMP 计划。其中，ASCI 计划旨在为核武器研究制造 3 个有万亿次以上计算能力的节点；HPCMP 则意于构建以高性能计算为基础的综合防卫系统。

我国于 1994 年自主研制了银河 II 型巨型机，1999 年研制出了“神威-I 号”超级计算机，后续的还有曙光 4000、深腾 6800 等 HPC 系统[10]。我国在 HPC 方面的研究投入已经十分巨大，但与一些先进国家还有着不小的差距。

高性能计算是应对密集型科学应用的产物，它将若干个处理器或者集群组织中的部分计算机有目的地组合起来，组成了所需的计算系统及环境[11-12]。图 1 给出了典型的 HPC 拓扑结构。

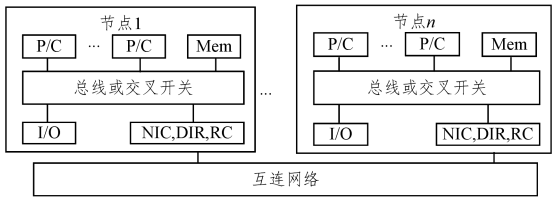


图 1 并行计算机的体系结构

Fig. 1 Architecture of parallel computers

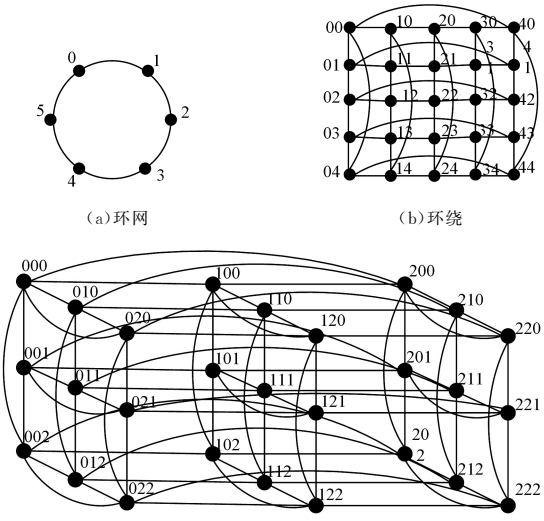
互连网络在计算机系统的协同工作中一直扮演着重要的角色，从最早的多处理器到最新的高性能并行计算机（HPC），ICN 以及中央处理器（CPU）的类型都是系统的主要特征。在具有大量协同计算机的计算系统中，由于执行时间越来越依赖于通信时间[13]而不是计算时间，因此 ICN 的性能比 CPU 的性能更为重要。ICN 在很大程度上决定了 HPC 在大多数现实世界并行应用中的效率和扩展性，它可以缩短总体执行时间，并增加可有效利用的处理器数量，这两种情况都会导致更高的最终加速。

ICN 可以分为直连网络和非直连网络。当一个节点直接连接到它的邻居时，网络是直连的。例如，完全连接的网络在任何两个节点之间都有直连链接。全连通直连网络不能用于构建大型系统，因为它具有 $O(n^2)$ 的复杂性，其中 n 是节点数。通常，一个节点只直接连接到其他节点的一个子集，而与其余节点的通信是通过中间节点（如网格、超立方体）路由消息来实现的。非直连网络通过一个或多个交换机来连接节点。理想的开关是完全连接的横杆，如果输出端口尚未使用，横杆将启用从任何输入端口到任何输出端口的连接。大横杆是一种不可行的解决方案，因为它的复杂性为 $O(n^2)$，因此在路由器和交换机内部使用较小的横杆作为基本的构建块。开关通常使用阶段之间的规则连接模式（多级 ICN）组织成阶段。

目前，在大型系统中应用最广泛的网络拓扑是直连 k-元 n-立方体[14]和非直连胖树[15]。

（1）k-元 n-立方体。多维 k-元 n-立方体拓扑可以看作是广义正则网格。参数 k 和 n 分别表示每个维度和空间维度上的节点数。将位置相同的节点连接成环，可以用 k 倍 k-元 $(n-1)$-立方体构造 k-元 n-立方体。

如图 2 所示，环绕和网格是 n 维网格，每个维度有 k 个节点，节点总数 $n=k^n$。超立方体是另一种网格状拓扑，与更改维度中节点的数量不同，该数量保持不变（$k=2$），并且维度的数量也不同。一维超立方体只是 2 个相连节点的对，二维超立方体是 4 个节点的正方形，三维超立方体是立方体。



（a）环网 （b）环绕

（c）3-元 3-立方体

图 2 k-元 n-立方体结构

Fig. 2 Architecture of k-ary n-cube

（2）胖树。胖树是一种非直连拓扑，可以用两种非常明显但同构的方式构建。考虑将处理节点作为叶子的交换机树，所有这些交换机都以相同的速度相连接。随着通信从树叶进行到树根，很明显靠近树根的链接将成为瓶颈。为了解决这个问题，靠近根的链接应该更快（更高的时钟，多个链接，更宽或“更胖”）。这种结构的不足是必须为根节点设置一个高速交换机。主要考虑到两层交换机之间的多链路情形，而非从低层交换机到高层交换机的多链路，使得高层交换机的链路

实现了分布式连接。

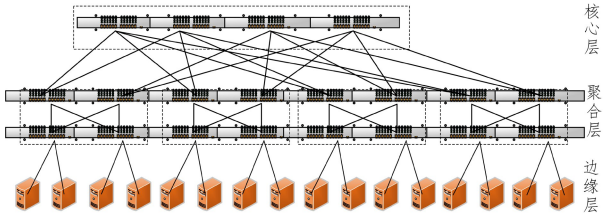


图 3　胖树结构

Fig. 3　Architecture of fat-tree

(3) 蜻蜓网络。为了以最少的昂贵全局链路实现高互连可扩展性，蜻蜓拓扑结构有效地使用了高基数路由器[16]。为了实现这一目标，不仅在图形级别，在经济有效的实际实现中，蜻蜓也脱离了已建立的拓扑结构。蜻蜓是作为一个在组和系统级别的路由器层次网络而建立的。每个路由器都连接到 p 计算节点、同一组中的 $a-1$ 路由器和组之间的 h 全局链接，从而导致路由器基数 $k=p+a+h-1$。一个组由一个路由器组成，该路由器通过一个仅使用本地链路的组内 ICN 连接，从而产生 ap 组节点和 ah 全局链路。图 4 给出了蜻蜓网络的结构，图中虚线显示连接到图中未显示的其他组的全局通道。

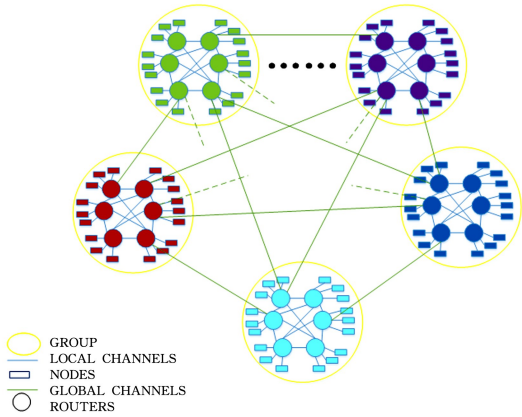


图 4　蜻蜓模型的层次网络($p=h=3, a=6$)

Fig. 4　Hierarchical Network of Dragonfly Model($p=h=3, a=6$)

2.2　天文大数据

近年来，在气候科学、天体物理学、燃烧科学、计算生物学和高能物理等关键领域，越来越多的科学应用倾向于高度数据化，并为研究和开发社区带来了重大的数据挑战。在科学实验、观察和模拟的过程中生成了大量数据，这些数据集的大小可以从几百兆字节到百万兆字节甚至更多[17-23]。例如，澳大利亚的 FAST 项目(500 m 孔径球面望远镜)使用下一代存档系统(NGAS)来存储和维护大量需要收集的数据[24]。NGAS 希望每年从 FAST 处理大约 3 PB 的数据，该数据足以填充 12 000 个单层的 250 GB 的蓝光光盘。这一大数据革命的因素众多：1)计算能力的快速增长(特别是与 I/O 系统带宽的缓慢增长相比)，使得数据采集和生成变得更加容易；2)高分辨率，多模型的科学发现将需要并产生更多的数据；3)多年来，从大量低熵数据中挖掘出洞察力的需求大幅增加。

天文学进入了大数据时代，大型的巡天望远镜的应用也使得天文学数据在量级和质量以及复杂度(丰富程度)上产生了质的飞跃，而这三者又是紧密相连的。不同设备所采集的数据也不尽相同，例如：位于墨西哥的斯隆数字巡天望远镜以及我国所研制的郭守敬望远镜(大天区面积多目标光纤光谱天文望远镜，LAMOST)[25]获取的是天体的多色测光资料以及光谱数据，用以探索宇宙中各种各样的星体；位于贵州省的“中国天眼”500 m 口径球面射电望远镜(FAST)所观测的脉冲电信号数据已经成功地发现了 59 颗优质的脉冲星后选体，为我国的天文事业做出了巨大贡献；位于智利的大型综合巡天望远镜(LSST)每年须拍摄 20 多万张相片，这些图像数据在处理后将被用于侦测暗物质与暗能量以及寻找太阳系中的大小天体(包括近地小行星或者超新星)[26]；而南京紫金山天文台则通过“悟空号”获取了大量的电信号数据，这些数据在经过处理后将被分类为电子、中微子等带有不同粒子特征的数据，为观测暗物质做出了贡献。

以上仅是天文数据根据存在方式不同的分类，除此之外，天文数据还可以根据获取方式(观测数据以及数值模拟数据)、结构(结构化数据、半结构化数据、非结构化数据)等进行分类。

结合数据科学家们提出的大容量、多类型、高复杂性等大数据特征与天文学特点，我们可以得到以下天文数据的特点：除了已知的海量性、空间性以及多模式，天文数据还有可能是高维度(光谱数据)、多尺度以及高分辨率(图像数据)的[27-31]。此外，宇宙空间中某些因素或仪器本身的影响，也可能使得这些数据是缺失或者伴有误差的。这些天文数据的类型、特点及丰富程度都对天文大数据的存储、传输、处理、分析、挖掘等提出了严峻的挑战。

3　天文数据的研究方法

3.1　天文大数据存储

新兴的大数据处理技术(如 MapReduce)虽然擅长分析大数据[32]，但是像 Spark 这样的分布式系统作为用于处理许多应用程序域中的大量数据的集群计算模型，已变得越来越流行。Spark 执行内存计算，其目标是优化基于磁盘的框架，例如 Hadoop。然而，由于没有涉及密集计算的数据访问优化，这些分布式框架不能提供有效的天文查询处理能力，这是由天文数据的一些特征所决定的。例如前文所述的高维度行数据，在数据存储时必须先对数据进行降维处理，然后才能将其存储在诸如 MapReduce 这样的分布式存储系统中，但是其本身并不具备数据的降维功能。对此 Brahem 等提出了 AstroSpark[33]系统，该系统是 Spark 的一个扩展，是一种可扩展、低延迟、经济效益高并且十分高效的天文查询处理框架，用于处理和分析天文数据。AstroSpark 的基本框架如图 5 所示。

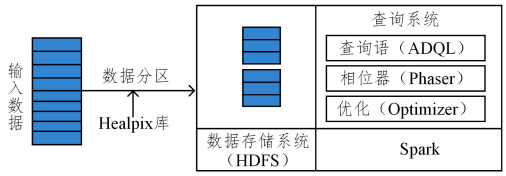


图 5　AstroSpark 的基础框架

Fig. 5　Infrastructure framework of AstroSpark

由华盛顿大学开发的开源项目 Myria 也能够快速读取并处理天文大数据。Loebman 等[34]设计了一个被称为“MyMergerTree”的合并树系统，为了使天文学家能够通过大规模天体物理模拟来跟踪其“合并树”以研究星系的增长历史，该系统使用了 Myria 作为后端的并行数据管理系统[14]。

Myria 可以直接从外部源(例如 HDFS 或 Internet)读取和处理数据,将加载到 Myria 中的数据存储在 PostgreSQL 中,每个群集节点上都运行着一个独立的实例(类似于 HadoopDB)。通过这种设计,Myria 可以利用 PostgreSQL 的索引功能,还可以将一些计算直接推送到该存储层。一旦进入内存,Myria 将继续使用自己内存中的关系和数据混洗运算符来处理数据。

我国的刘应波博士在研究了太阳 FITS 元数据和数据分布式存储中的不一致性问题后,设计了以面向太阳观测的分布式存储系统 AstroFS[35]。该系统通过基于网络的 RAID0 数据分片技术,使数据的聚合拆分、数据均衡分布存储、数据复制和提交、并发控制等技术也能够达到围绕存储的高性能和可扩展等特性,让该系统能够适用于面向大型望远镜的数据存储的目的。

3.2 天文大数据处理

由于不同的观测设备所采集的数据类型不一致,如 LAMOST 所观测的数据主要是光谱数据以及图像数据[36-40],而紫金山天文台粒子探测卫星(DAMPE)所采集的原始的 14 类数据包括 1 类 SCIENCE 数据以及 13 类 HOUSEKEEPING 数据。正是天文数据的多样性、高复杂度的特性,使得传统的数据挖掘、数据分析与处理技术在天文数据领域的应用变得尤为困难[41-42];但同时,这也为天文领域以及计算机科学领域的专家提供了广阔的研究空间。

图 6 给出了天文数据处理的流程[43]。从图中可以看出,本地计算机在通过资源检索获取了云门户所定位的资源后,会将天文数据传入预处理系统进行预处理,同时预处理过的数据会存入云存储系统进行有效的管理,之后,后处理系统会将存储于云存储系统中的数据提取出来进行处理。同时,云存储系统中也存储了高性能计算资源。

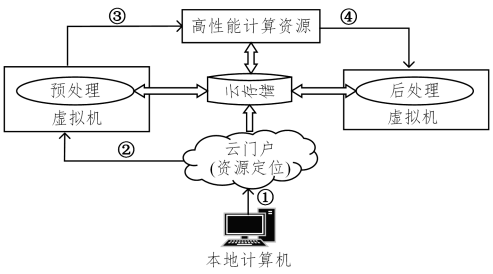


图 6 天文数据处理的流程

Fig. 6 Process of astronomical data processing

观测而来的数据往往无法直连用以处理分析,这些原始数据大多是杂乱无章的。为了便于处理,首先使用一些方法对天文数据进行分类,以下是一些常用的分类方法。

(1)朴素贝叶斯分类。朴素贝叶斯分类通过对特征独立性进行"朴素"假设来应用贝叶斯定理[44]。形式上,给定一组 n 个特征 $x_1,\cdots,x_n$,相关的模式被认为属于满足以下条件的类 y:

$$y=\arg\max_j P(C_j)\prod_{i=1}^{n}p(x_i|C_j)$$

其中,$P(C_j)$为类 C_j 的先验概率,$P(x_i|C_j)$为给定类 C_j 的特征 x_i 的条件概率(易于从使用监督学习框架的数据估计得出)。

(2)逻辑回归。在逻辑回归中,因变量(类)y 的条件概率被建模为解释变量[27](输入特征)$x_1,\cdots,x_n$ 的对数变换多元线性回归:

$$P_{\mathrm{LR}}(y=\pm1|x,w)=\frac{1}{1+\mathrm{e}^{-yw^{\mathrm{T}}x}}$$

通过最大化训练数据集上模型的可能性来训练模型(即学习的权重参数),可以得到:

$$\prod_{i=1}^{2}P_r(y_i|x_i,w)=\prod_{i=1}^{2}\frac{1}{1+\mathrm{e}^{-yw^{\mathrm{T}}x_i}}$$

由于模型的复杂性而受到惩罚:

$$\frac{1}{\sigma\sqrt{2}\pi}\mathrm{e}^{-\frac{1}{2\sigma^2}w^{\mathrm{T}}w}$$

这可以作为以下正则化负对数似然的最小化重述:

$$\vartheta=C\sum_{i=1}^{2}\log(1+\mathrm{e}^{-yw^{\mathrm{T}}x_i})+w^{\mathrm{T}}w$$

坐标下降法用于最小化 ϑ。

(3)支持向量机(Support Vector Machine,SVM)。SVM 通过在原始输入数据被映射到的高维(可能是无限维)[45]空间中构造一系列分离超平面的类来执行分类,通过隐式地而不是明确地执行该映射,有效地实现了将数据映射到较高维空间这种看似难以处理的任务;通过把内积运算替换成核函数,确保了使用原始空间中的变量能够容易地计算高维空间中的点积[46]。给定$\{(x_1,y_1),\cdots,(x_n,y_n)\}$形式的标记训练数据(输入向量和相关标签),支持向量机旨在找到最小化错误分类数的映射,以规范的方式训练实例。

(4)k-最近邻。k-最近邻分类包括特征 $x_1,\cdots,x_n$ 的新模式,其包括在具有已知类别成员资格的训练模式中的输入模式(在特征空间中)的 k 个最近邻的集合中占优势的类别[47]。通常,使用的距离为欧几里得距离(欧氏距离)[48]。

(5)随机森林。随机森林分类器属于基于集合的学习方法的广义范围。它们易于实施,操作快速,并且已被证明在各种领域都非常成功。随机森林方法的关键原则包括在训练阶段构建许多"简单"决策树[49],并在分类阶段建立多数分裂(模式)。除了其他优点之外,这种分裂策略具有纠正决策树的不良特征从而达到过渡训练数据的效果。在训练阶段,随机森林将称为捕获的一般技术应用于整体中的各个树木[50]。捕获反复从训练集中选择随机样本并将树木替换为这些样本。每棵树都没有任何修剪,整体中的多棵树是一个自由参数,可以使用所谓的捕获误差来进行自动学习,这种方法也在目前的工作中被采用。

4 存在的问题与未来的发展趋势

天文大数据作为研究天文学的重要依据,其宝贵程度不言而喻,然而天文数据的丰富度与多样性也使得天文学家望而却步[51]。对已经收集的数据进行分类、辨别、预处理、清洗以及分析等处理,其数据量十分庞大,通过人工来完成很显然是不现实、不科学的。目前来说,天文大数据领域所遇到的挑战如下:

(1)如何对每天采集的数据进行高效的、准确的分类,以实现数据的快速归档;

(2)如何更加高效地存储天文数据,以实现天文数据的快速查询;

(3)如何优化计算科学算法，实现海量天文数据的快速处理；

(4)充分利用海量历史数据成为天文学领域中的关键科学问题，甚至为新的发现提供了可能。

上述挑战正是现代计算机领域所擅长的，尤其是高性能计算，其高吞吐量以及分布式计算的特点为解决天文大数据所遇到的这些难题提供了可能性。

(1)目前，天文大数据应用运行在国外主流 CPU 的集群中，只能通过横向扩展的方式增大集群规模，同时增加了成本投入和能耗的开销。未来希望构建基于国产芯片的计算集群，加入定制化的人工智能芯片，如寒武纪，为天文大数据的处理提供定制化计算集群。

(2)海量数据的不断产生，造成了数据传输和存储的困难。科学研究时效要求这些数据必须能被快速存储、分析、共享和归纳。而大数据一体机的应用能够较好地应对天文数据的存储、分析和共享问题，使得天文数据能够在大数据一体机上得到快速、高效的分析推断，为天文数据的分析以及预测提供大数据的支持。

(3)将传统的数据分类算法与机器学习相结合，如决策树、神经网络等，以提高天文数据算法分类的效率与精确度。

(4)目前天文大数据大多采用传统的数据库格式，对于超大规模的数据，这种方式极大地影响了归档的效率，进而导致了查询效率低的问题[49]。未来可以考虑基于分布式思想开发专用的数据存储软件。

(5)传统的数据处理算法存在效率低、复杂度高的问题。未来可以考虑将数据处理算法并行化，提高整体数据的处理效率，同时在数据分析过程中应用数据挖掘的关键算法(如关联规则、朴素贝叶斯等)从天文数据中发现新物质与新现象，以进一步提升可能性。

结束语　天文学进入了大数据时代，科学研究也由第一范式(实验科学)逐渐转变为了第四范式(数据密集型科学)。面对这样的转变，一方面，天文学家视之为机遇，这意味着他们将会拥有一个巨大的数据宝库，这个宝库可以带领他们探索更浩瀚的宇宙；另一方面，如何更好地存储、管理这些数据，也将成为他们所要面对的挑战之一。高性能计算的快速发展为天文学带来了机遇，定制化集群及智能芯片为天文大数据的快速处理、精确分析提供了辅助。因此，现在更需要多领域多学科的专家合作共赢，计算机科学领域的专家应充分考虑天文数据的类型以及特点，结合已有的数据挖掘、数据处理、数据可视化等方面的工具，开发出了适用于天文大数据的工具；结合高性能计算、机器学习与统计学，推动天文学的发展。

参 考 文 献

[1] ZHANG Z, BARBARY K, NOTHAFT F A, et al. Kira: Processing Astronomy Imagery Using Big Data Technology[J]. IEEE Transactions on Big Data, 2016: 1-14.

[2] SZALAY A S, KUNSZT P Z, THAKAR A, et al. Designing and mining multi-terabyte astronomy archives: the Sloan Digital Sky Survey[C]// Proceedings of International Conference on ACM Sigmod Management of Data. 2000: 451-462.

[3] NEOPHYTOU P, GHEORGHIU R, HACHEY R, et al. Astroshelf: understanding the universe through scalable navigation of a galaxy of annotations[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 713-716.

[4] DRLICA-WAGNER A, SEVILLA-NOARBE I, RYKOFF E S, et al. Dark energy survey year 1 results: the photometric data set for cosmology [J]. The Astrophysical Journal Supplement Series, 2018, 235(2): 33.

[5] CHEN G L, MAO R, LU K Z. Parallel computing framework for big data[J]. Chinese Science Bulletin, 2015(5): 566-569.

[6] SHEN H F, LUO S W, ZHAO H. The Model Structure of Cluster Computing System[J]. Application Research of Computers, 2004(2): 52-55.

[7] FAN Z, QIU F, KAUFMAN A, et al. GPU Cluster for High Performance Computing[J]. SC 2004, 2004, 1: 47.

[8] BRENNAN J, KURESHI I, HOLMES V. CDES: an approach to HPC workload modelling [C] // Proceedings of International Symposium on IEEE/ACM 18th Distributed Simulation and Real Time Applications. 2014: 47-54.

[9] RAMÍREZ-GALLEGO S, KRAWCZYK B, GARCÍ A, et al. A survey on data preprocessing for data stream mining: Current status and future directions[J]. Neurocomputing, 2017, 239: 39-57.

[10] 陈国良. 并行计算机体系结构[M]. 北京：高等教育出版社，2002.

[11] JIN Y L, HUANG Y L, CHEN Z N, et al. Trends and Key Technologies of High Performance Computers[J]. Engineering Sciences, 2001, 3(6): 1-8.

[12] BISTOUNI F, JAHANSHAHI M. Scalable crossbar network: a non-blocking interconnection network for large-scale systems [J]. The Journal of Supercomputing, 2015, 71(2): 697-728.

[13] HU Y, KUDOH T, KOIBUCHI M. A case of electrical circuit switched interconnection network for parallel computers[C]// 2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). IEEE, 2017: 276-283.

[14] LV Y, FAN J, HSU D F, et al. Structure connectivity and substructure connectivity of k-ary n-cube networks[J]. Information Sciences, 2018, 433: 115-124.

[15] QIAN Z, FAN F, HU B, et al. Global round robin: Efficient routing with cut-through switching in fat-tree data center networks [J]. IEEE/ACM Transactions on Networking, 2018, 26(5), 2230-2241.

[16] XIANG D, LI B, FU Y. Fault-Tolerant Adaptive Routing in Dragonfly Networks[J]. IEEE Transactions on Dependable and Secure Computing, 2017, 16(2): 259-271.

[17] AKRITAS M G, SIEBERT J. A test for partial correlation with censored astronomical data[J]. Monthly Notices of the Royal Astronomical Society, 2018, 278(4): 919-924.

[18] CUI C, YU C, XIAO J, et al. Astronomy research in big-data era [J]. Chinese Science Bulletin, 2015, 60(Z1): 445-449.

[19] ZHANG Z, BARBARY K, NOTHAFT F A, et al. Scientific computing meets big data technology: An astronomy use case [C]// Proceedings of International Conference on IEEE Big Data. 2015: 918-927.

[20] STEPHENS Z D, LEE S Y, FAGHRI F, et al. Big data: Astronomical or genomical? [J]. Plos Biology, 2015, 13(7): e1002195.

[21] JACKSON K R, RAMAKRISHNAN L, MURIKI K, et al. Performance analysis of high performance computing applications

on the amazon web services cloud[C]// Proceedings of International Conference on 2nd IEEE Cloud Computing Technology and Science. 2010:159-168.
[22] NIGRI E, ARANDJELOVIC O. Light curve analysis from Kepler spacecraft collected data[C]// Proceedings of the International Conference on ACM on Multimedia Retrieval. 2017:93-98.
[23] XU L, YU X X, YAN Y H. Deep learning application in astronomical big data processing[J]. E-science Technology & Application, 2018, 9(3):49-58.
[24] ZHANG Q, YANG L T, CHEN Z, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42:146-157.
[25] SHAN G H, XIE M J, LI F A, et al. Visualization of large scale time-varying particles data from cosmology[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(1):1-8.
[26] VINOGRADOV V I. Advanced high-performance computer system architectures[J]. Nuclear Inst & Methods in Physics Research A, 2007, 571(1/2):429-432.
[27] DEEPU C V, KURKURE N, DINDE P, et al. e-Onama: Mobile high performance computing for engineering research[C]// Proceedings of International Conference on IEEE Third Innovative Computing Technology. 2013, 532-536.
[28] GAO C Z, CHENG Q, PEI H, et al. Privacy-preserving naive bayes classifiers secure against the substitution-then-comparison attack[J]. Information Sciences, 2018, 444:72-88.
[29] LIU K, ZHOU X Z, ZHOU D R. Research and Development of Data Visualization [J]. Computer Engineering, 2002, 28(8):1-2.
[30] BACON D F, GRAHAM S L, SHARP O J. Compiler transformations for high-performance computing[J]. ACM Computing Surveys, 1994, 26(4):345-420.
[31] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.
[32] ZHONG R Y, LAN S, XU C, et al. Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing[J]. The International Journal of Advanced Manufacturing Technology, 2016, 84(1-4):5-16.
[33] BRAHEM M, LOPES S, YEH L, et al. AstroSpark: towards a distributed data server for big data in astronomy[C]// Proceedings of international conference on the 3rd ACM SIGSPATIAL PhD Symposium. 2016:3.
[34] LOEBMAN S, ORTIZ J, CHOO L, et al. Big-data management use-case: A cloud service for creating and analyzing galactic merger trees[C]// Proceedings of international conference on Data analytics in the Cloud. 2014:1-4.
[35] LIU Y B. Research on Key Technologies of Massive Data Storage for Solar Telescope[D]. Yunnan: Graduate School of Chinese Academy of Sciences, 2014.
[36] THORVALDSDOTTIR H, ROBINSON J T, MESIROV J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration[J]. Briefings in Bioinformatics, 2013, 14(2):178-192.
[37] YOU L, TUNÇER B. Informed design platform: Interpreting "big data" to adaptive place designs[C]// Proceedings of International Conference on IEEE 16th on Data Mining Workshops. 2016:1332-1335.
[38] WANG L. Big Data and Visualization: Methods, Challenges and Technology Progress[J]. Canadian Journal of Electrical & Computer Engineering, 2015, 34(3):3-6.
[39] ZHANG S, LI X, MING Z, et al. Learning k for kNN Classification[J]. ACM Transactions on Intelligent Systems & Technology, 2017, 8(3):43.
[40] LOSING V, HAMMER B, WERSING H. KNN classifier with self adjusting memory for heterogeneous concept drift[C]// Proceedings of International Conference on IEEE 16th Data Mining. 2016:291-300.
[41] JOG A, CARASS A, ROY S, et al. Random forest regression for magnetic resonance image synthesis[J]. Medical Image Analysis, 2017, 35:475-488.
[42] LU M, SADIQ S, FEASTER D J, et al. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods[J]. Journal of Computational and Graphical Statistics, 2018, 27(1):209-219.
[43] KIM J, DALLY W J, SCOTT S, et al. Technology-driven, highly-scalable dragonfly topology[C]// Proceedings of International Symposium on IEEE Computer Architecture. 2008:77-88.
[44] SUN N, SUN B, LIN J D, et al. Lossless pruned Naive Bayes for big data classifications[J]. Big Data Research, 2018, 14:27-36.
[45] HARRIS T. Credit scoring using the clustered support vector machine[J]. Expert Systems with Applications, 2015, 42(2):741-750.
[46] RAVALE U, MARATHE N, PADIYA P. Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function[J]. Procedia Computer Science, 2015, 45:428-435.
[47] ADENIYI D A, WEI Z, YANG Y Q. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method[J]. Applied Computing and Informatics, 2016, 12(1):90-108.
[48] DOKMANIC I, PARHIZKAR R, RANIERI J, et al. Euclidean distance matrices: essential theory, algorithms, and applications [J]. IEEE Signal Processing Magazine, 2015, 32(6):12-30.
[49] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// Proceedings of International Conference on Advances in Neural Information Processing Systems. 2017:3146-3154.
[50] BELGIU M, DRĂGUŢ L. Random forest in remote sensing: A review of applications and future directions[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114:24-31.
[51] ZHANG Y, ZHAO Y. Astronomy in the big data era[J]. Data Science Journal, 2015, 14(11):1-9.

WANG Yang, born in 1995, postgraduate. His main research interests include astronomical data processing and analysis.

FAN Wei-bei, born in 1987, Ph.D, lecturer, is member of China Computer Federation (CCF). His main research interests include parallel and distributed system, data center network and cloud computing.