基于材料数值计算大数据的材料辐照机理发现

任帅[1,2]，陈丹丹[1,2]，储根深[1,2]，白鹤[1,2]，李慧昭[1]，何远杰[1]，胡长军[1,2]
1. 北京科技大学计算机与通信工程学院，北京 100083；
2. 智能超算融合应用技术教育部工程研究中心，北京 100083

摘要

材料辐照效应的数值模拟计算是认识核材料服役性能的重要手段，基于超级计算机的大规模、高保真材料数值模拟计算会产生海量数值计算数据，如何针对数值计算大数据的特点，在实现其高效存储的基础上，通过挖掘总结辐照损伤机理和性能演化规律，对于核材料设计研发、核安全等具有重要意义。论述了材料数值计算大数据的定义及其本质特征，综述了近年来的相关工作。以自主研发的材料辐照效应分子动力学软件MISA-MD和随机团簇动力学软件MISA-SCD在国产超级计算机上的实际算例为基础，提出了一种适用于材料数值计算大数据的、多尺度关联与耦合的分布式数值计算大数据存储体系（NDSA）；采用XGBoost算法实现了MD中Frenkel缺陷对数的精确预测，基于并查集算法实现了级联碰撞团簇的划分；基于密度聚类的方法对KMC数值计算大数据进行挖掘，发现了类环状团簇，实现了原子团簇的识别与分类；基于第一性原理数值计算大数据库对现有的势函数模型进行了改进，提出了新的势函数模型构建方法AIPM。最后对材料数值计算大数据的应用前景进行了展望。

关键词

材料数值计算大数据；材料辐照效应；多尺度模拟；高性能计算；数据挖掘；机器学习

中图分类号：TP391.4　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2021056

Discovery of irradiation mechanism based on big data of material simulation

REN Shuai[1,2], CHEN Dandan[1,2], CHU Genshen[1,2], BAI He[1,2], LI Huizhao[1], HE Yuanjie[1], HU Changjun[1,2]
1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
2. Engineering Research Center of Intelligent Supercomputing, Ministry of Education, Beijing 100083, China

Abstract

The numerical simulation of material irradiation effect is an important means to understand the performance of nuclear materials. The large-scale and high fidelity material numerical simulation based on supercomputer will produce a large amount of numerical calculation data. Understanding the evolution law of the irradiation damage mechanism and performance through mining and analysis based on high-efficiency storage is of great significance for the design and

development of nuclear materials and nuclear safety according to the characteristics of numerical calculation big data. The concept of big data of material simulation (MSBD) was proposed, and then the characteristics and significance of MSBD were specifically introduced, and the related work was reviewed. Based on the practical examples of MISA-MD and MISA-SCD on domestic supercomputers, a distributed numerical data storage arihitecture (NDSA) multi-scale correlation and coupling was proposed. Frenkel defect pairs were accurately calculated with XGBoost algorithm based on MSBD of MD, and the cascade collision clusters were artificially divided with Union-Find algorithm. The data of KMC numerical calculation were mined based on density clustering method, and the cluster recognition and classification were realized. The ring like clusters were found from MSBD of KMC based on density clustering algorithm, which was verified with the literature. A DNN-based potential model - AIPM was proposed with MSBD of first principles-based potential data. The further application of MSBD was discussed and prospected in physical modeling and knowledge discovery.

Key words
big data of material simulation, material irradiation effect, multi-scale simulation, high performance computing, data mining, machine learning

1 引言

在材料辐照效应领域，高性能计算软件在模拟过程中会实时产生数值计算数据。这些数值计算数据不仅数目巨大、关联性强，而且不同计算尺度、不同服役环境下的数据之间是相互关联的。同时，这些数据中蕴含着材料从微观机理到宏观性能的规律，具有量大、关联复杂、类型丰富的典型大数据特征（如**图1**所示），是具有宝贵价值的。除了具有典型大数据特征，这些数据还具有领域特殊性。从反应堆材料生命周期的角度来看，首先，数据类型丰富。从具体的设计、服役到寿命终止，都会产生大量的类型多样的数据，这些数据是典型的大数据。其次，关联复杂，反应堆材料的使用寿命与各个服役阶段息息相关。优异的服役性能离不开精确的系统测试，离不开大量的工艺参数调控，更离不开合适的成分、结构设计，因此，各个阶段之间的数据关联关系极其复杂。最后，具有时序性。反应堆材料的服役周期长达几十年，且随着使用时间增长，材料性能在不同的时效作用下也会呈现不同的特点，数据版本多种多样，使得反应堆材料辐照数据具有显著的时序性特点。

基于上述分析，笔者提出了材料数值计算大数据（big data of material simulation，MSBD）的概念，在超级计算机上已通过准确性验证的材料数值建模和模拟软件会产生大量的数值计算数据，这些数据具有数目巨大、关联复杂、类型丰富等典型的大数据特征；同时这些数据具有类型丰富、时序性等领域特殊性。材料数值计算大数据是一类典型的工业大数据，对材料辐照效应的研究具有重要意义。依托现有的超算资源（如天河、神威、曙光等大型超级计算机），辅以专用数据库实现，材料数值计算大数据不仅可以用于材料领域机器学习、知识发现和数据挖掘，也可以用于发展新的建模技术，例如用于材料计算模型的改进、材料多尺度模型耦合等。具体地讲，材料数值计算大数据可用于以下几个方面。例如，级联碰撞后，材料内部原子离开原始晶格位置，聚集在一起形成不同形态的原子团簇微观缺陷，这些缺陷在材料内部不容易滑移，因此容易引起材料硬化和脆化，从而影响反应堆寿命。这些团簇的尺寸只有几纳米，目前的

实验手段只能实现静态的观察，因此对于这些团簇的形成机理尚不清楚。聚类的方法通常用于数据模式识别，因此可通过基于聚类的方法对级联碰撞特定时间步产生的材料数值计算大数据进行分析，挖掘级联碰撞数据中不同类型的团簇，进而研究团簇类型和数量与实验条件之间的关系，从而探究团簇类型和数量对材料性能的影响。又如，在新材料发现方面，基于物理模型的数值模拟计算被用于预测新材料已经有很多年的历史，然而，这些模型在处理大规模、多维度问题时，往往需要占用大量的计算资源，且非常耗时。此外，随着合金元素的增加，基于物理模型的势函数建模越来越困难，通常来讲，一种势函数模型的构建时间需要2~3年。机器学习在处理多维问题上表现优异，基于机器学习方法和材料数值计算大数据对原子体系的模拟参数和势能进行拟合，保留必要的物理参数，隐藏复杂的物理研究过程，从而改进现有建模技术。

2 材料数值计算大数据相关工作

近年来，大规模高性能材料数值计算模拟在材料研究中起着越来越重要的作用，是现今进行材料研究不可或缺的手段之一。尤其对于实验条件复杂且实验成本高昂的材料辐照效应研究而言，实验前先通过材料模拟软件对材料进行筛选，在大幅节省科研成本的同时，提高材料研究安全性[1-4]。高性能计算技术的发展使得材料辐照效应模拟无论在时间尺度还是空间尺度都取得了突破性进展[5]。随着软件尺度规模的扩大，产生的数据越来越多，材料数值计算大数据的高效存储与分析成为材料数值计算研究的新焦点。

首先，在材料辐照效应模拟过程中，

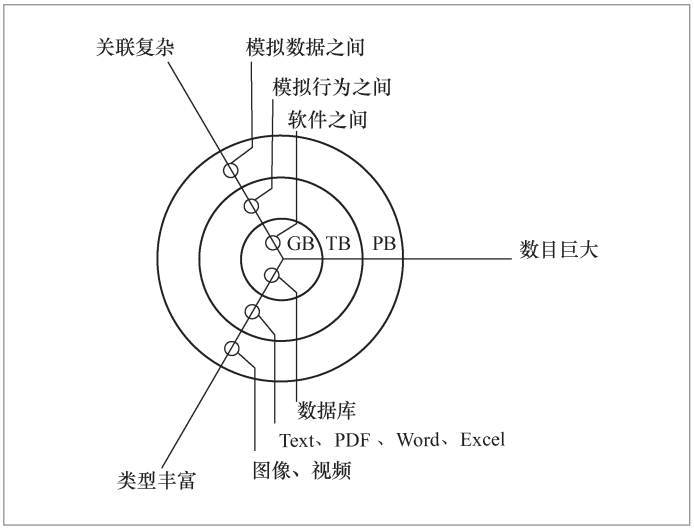

图1 典型大数据特征

材料模拟软件会产生海量的数值计算数据[6]。从不同模拟软件的维度来看，分子动力学软件用于原子尺度结构演化过程的模拟，一次大规模级联碰撞模拟产生的原子尺度数据在1 GB以上，这些数据将被用于蒙特卡洛模拟软件的短程演化；蒙特卡洛模拟软件产生的原子结构数据将被传递给团簇动力学软件，并由其对原子结构进行长程演化，一次大规模的辐照效应团簇动力学模拟产生的数据约100 GB，这些数据被用于位错动力学等更大尺度的模拟。因此，材料数值计算大数据不仅数目巨大，而且各软件产生的数据之间是彼此紧密关联的。从单个软件的模拟行为角度来看，不仅可以对不同类型的物理过程进行模拟，也可以针对同一类型的物理过程进行不同实验条件下的模拟。例如，分子动力学软件既可以进行晶内原子级联碰撞的模拟，也可以进行晶界析出强化的模拟；既可以进行高能中子下的级联碰撞模拟，也可以进行低能中子下的级联碰撞模拟。蒙特卡洛模拟软件既可以实现级联碰撞的退火模拟，也可以对材料晶粒形核、长大过程进行模拟，还可以实现级联碰撞的析出模

拟等。从各软件产生的材料数值计算大数据来看，数据类型更是极为复杂。每个软件的每一次计算过程都会有不同的输入数据、过程数据、结果数据、后处理数据等，不同软件之间的这些数据还存在复杂的关联关系。例如前面提到的，分子动力学软件的模拟结果数据作为蒙特卡洛模拟软件的输入数据，蒙特卡洛模拟软件的结果数据或者后处理数据则作为团簇动力学软件的输入数据，依此类推。综上所述，材料辐照效应数值计算大数据具有显著的数目巨大、关联复杂的特点。

其次，这些数据对于材料辐照效应模拟的研究具有重要价值。这些数据蕴含了模拟材料辐照过程的物理模型信息、计算模型信息，合理收集这些数据并进行研究对于改进现有模型具有重要的研究价值。例如在材料辐照效应级联碰撞模拟中，级联碰撞结果数据通常为所有原子的坐标数据，与初始晶体的原子坐标数据（如获取Frenkel缺陷对的数量）进行对比分析，可以对高能粒子辐照后的材料结构变化有初步的认识。如果进一步分析结果数据，还可以得到级联碰撞后产生的团簇类型和数量信息，从而对原子尺度的辐照效应有一个更加清晰和直观的认识。此外，模拟软件计算结果往往存在不稳定性，这种不稳定性也是反映在结果数据中的。对多次模拟的结果数据进行统计分析，可以为改进模拟软件的稳定性提供指导。除了结果数据具有很重要的研究价值，级联碰撞过程数据同样是值得研究的。过程数据反映了材料辐照效应模拟的整个演变过程，最直观的用途是用于计算结束后模拟过程的可视化。由于数据量过大，还可以针对过程数据的可视化方法进行研究，例如对实时的可视化方法的研究等都离不开过程数据。除此之外，随着机器学习和数据挖掘等方法近些年取得突破性发展，可以将这些方法用于材料数值计算大数据的研究中，例如，可以对输入数据和结果数据之间的关联关系进行挖掘。

近几年已有一些基于材料数值计算大数据开展的研究工作被报道。例如，Bhardwaj U等人[7-9]使用聚类的方法开展了对分子力学（molecular dynamics，MD）级联碰撞数据的分析研究。Podryabinkin E V等人[10]、Pilania G等人[11]通过机器学习对势函数库进行学习，开发用于势函数计算的机器学习模型，在保证原有精度的基础上将计算时间减少几个数量级。Jia W L等人[12]把势函数机器学习模型跟MD模拟软件LAMMPS集成起来，扩大了原有的计算规模。Kawamura T等人[6]基于模拟的过程数据，开发了一种名为“In-Situ PBVR”的可视化软件，首次实现了大规模核反应堆仿真的实时可视化。汪岸等人[13]针对数值核反应堆数据的特点进行了论述，提出了数值计算大数据在多个领域的应用需求。

然而，关于材料数值计算大数据研究价值的认识仍然处于起步阶段。另外，由于这些数值计算数据数目巨大、关联复杂，以及考虑到其所具有的领域价值等因素，材料数值计算大数据存储还没有一个很好的解决方案。这是因为材料数值计算大数据的存储要考量软件类型、模拟行为、数据类型等多个维度的因素，这些数据既有独立性又有相似性，而且还需要研究人员对相关的材料领域具有专业的认识。例如，国际原子能机构（International Atomic Energy Agency，IAEA）给出了一种材料数值计算大数据的存储方案，该方案被用于收集世界各地的MD数值计算大数据，受到研究人员的广泛关注。然而，面向多尺度模拟软件的统一数据存储方案目前仍然是个空白。本文针对材料多尺度数值计算大数据的特点，设计了一种适用于

材料多尺度数值计算大数据的存储与管理框架，并基于该数据库框架，结合机器学习等算法，实现了其在改进材料多尺度模拟中的应用。

3 材料数值计算大数据的特点

由前面给出的材料数值计算大数据定义可知，材料数值计算大数据具有数目巨大、类型丰富、领域特殊性等特点。为了进一步说明，以反应堆压力容器Fe基材料多尺度模拟为例，对材料辐照效应模拟软件MISA-MD和MISA-SCD的模拟结果进行了统计（见**表1**），软件均已开源。其中，MISA-MD用来模拟Fe基材料在原子尺度受到中子轰击后产生的原子级联碰撞过程，MISA-SCD用于更高空间尺度的模拟，即级联碰撞后产生的材料微观缺陷的演化过程。基于这两个软件的数值计算数据，概括材料数值计算大数据的特点如下。

（1）数目巨大。材料辐照效应模拟软件在模拟过程中会产生大量的数据，仅完成一次物理过程模拟的数据量就达到MB、GB甚至TB。本算例中MISA-MD选取的模拟时间为26 ps，约40 000个时间步，box尺寸为$80c_0\times80c_0\times80c_0$，其中$c_0$为晶格常数。完成一次级联碰撞演化过程模拟产生的数据量约1.5 GB。MISA-SCD选取的模拟时间为10^5 s，约5亿个时间步，box尺寸为3 μm^3，完成一次团簇演化过程模拟产生的数据量约为100 GB。对于如此庞大的数据量，如果没有一个合理的数据存储体系，这些数据将很难被高效地存储和分析。

（2）关联性强。本例中MISA-MD与MISA-SCD的材料模拟过程是紧密关联的，两者分别被用来模拟不同尺度的缺陷演化。MISA-MD只能模拟原子尺度的级联碰撞现象，要想进一步模拟后续缺陷演化现象，目前常采用的办法是材料多尺度模拟，即将原子尺度的级联碰撞数据传递给MISA-SCD软件并作为初始输入，然后才能进行更高尺度的模拟。因此，在材料多尺度模拟过程中，不同尺度软件产生的数据是紧密关联的。

（3）蕴含价值。材料数值计算大数据蕴含着材料演化过程的物理、化学信息，而非毫无意义的数据。MISA-MD产生的数值计算大数据包含了级联碰撞后的原子种类和坐标，通过聚类等方法对这些数据进行识别，将相似的结构归为一类，可以获得级联碰撞后的团簇种类和数量。而这些团簇的种类和数量可以作为初始输入传递给MISA-SCD进行后续演化模拟。此外，由于随机数的存在，材料辐照效应模拟往往具有一定的随机性。通过对同一实验条件下的模拟结果进行多次统计，获得模拟次数与缺陷数量的关系，可以对MISA-MD的结果稳定性进行评估。获得稳定的模拟结果是材料多尺度软件间实现正确耦合的前提。

（4）类型丰富。根据物理过程、尺度

表1 反应堆压力容器 Fe 基材料模拟软件数值计算大数据

模拟软件	模拟内容	空间尺度	时间尺度	数值计算数据量	数据信息
MISA-MD	Fe-Cu体系原子级联碰撞过程	原子尺度	20~100 ps 本算例为26 ps（约40 000个时间步）	1.5 GB（box尺寸为$80c_0\times80c_0\times80c_0$）	原子类型及其坐标信息
MISA-SCD	Fe-Cu体系团簇长大、溶质析出、溶质-缺陷相互作用	微观尺度	秒~月 本算例为10^5 s（约5亿个时间步）	约100 GB（box尺寸为3 μm^3）	缺陷类型及其数量

的不同，材料可以有很多种类型。首先，同一尺度的软件可以用来模拟不同的物理过程，从而得到不同的材料数值计算大数据。例如，MISA-MD可以被用来模拟不同合金原子级联碰撞过程，也可以被用来模拟合金原子在晶界偏聚对晶界强化、脆化的作用，还可以被用来模拟液态合金系统在凝固过程中的空位形成特性等。其次，不同尺度软件模拟得到的数值计算大数据也不同。例如MISA-MD得到的是原子类型及其坐标信息，而MISA-SCD得到的是原子团簇类型及其数量信息。数据类型不同给数据存储方式的选择带来了挑战。

（5）领域特殊性。由于空间尺度和时间尺度跨度很大，材料从原子尺度到宏观性能预测的模拟过程并不是一个软件能实现的。通常采用材料多尺度模拟的方法，使用不同的软件来模拟不同尺度的材料演化过程，然后将这些多尺度软件从低尺度到高尺度耦合起来，实现从原子尺度到微观再到介观甚至宏观的模拟。除此之外，材料从设计到投入使用要经历成分设计、微观组织调控、工业测试、服役等多道工序，这决定了全周期的材料数值计算大数据具有时序性。因此，材料数值计算大数据具有多尺度、时序性等材料领域特殊性。

除上述特点，材料数值计算大数据还具有数值计算所带来的不同于传统大数据和实验数据的特点。首先，材料数值计算大数据以数值类数据为主；其次，由于浮点运算的存在，材料数值计算大数据并非完全的精确数据；最后，并行执行的非确定性、随机的非确定性以及离散的非确定性导致材料数值计算大数据中带有非确定性的数据。这些特点为材料数值计算大数据的研究、管理和分析带来了传统大数据不曾面临的挑战。

4 材料数值计算大数据存储体系

为了有效收集、利用材料辐照效应模拟过程中产生的数值计算大数据，需要解决材料数值计算大数据的采集、存储与管理、处理与分析以及隐私和安全等大数据技术问题。本文提出了一种适用于材料数值计算的数值计算大数据存储体系（numerical calculation data storage architecture，NDSA）。该体系涵盖包含不同尺度软件的数值计算大数据，主要有MD、KMC（kinetic Monte Carlo）、SCD等主要数据库。这里所用的软件为北京科技大学与中国原子能科学研究院联合自主研发的材料辐照效应多尺度模拟软件，包括分子动力学软件MISA-MD、动力学蒙特卡洛模拟软件MISA-AKMC、随机团簇动力学软件MISA-SCD。目前这几款软件均已实现开源。

在数据库组织上，按照不同尺度的模拟软件进行组织。每个模拟尺度的数值计算大数据库中还包含程序模拟的多个生命周期的数据，如软件输入参数集合、模拟结果、模拟分析结果等。通过这样的两层组织，形成了多尺度模拟软件的大数据体系，形成了双重维度（材料多尺度模拟维度和数值计算生命周期维度）的数据关联存储。由于在模拟中，有时候参数（如合金的比例）是不定的，且随着模拟的增加，数据也会大量增加，数据库需具备高扩展性。此外，材料数值计算大数据的结构大多不固定。MongoDB框架可以为Web应用提供可扩展的高性能数据存储，其文档类似于JSON对象，在使用时也更加灵活。对于模拟的结果数据文件，由于其数量很大，如果存放在数据库中，可能会导致数据库冗杂和效率降低，因此，这些文件采用文

件存储服务进行管理。文件存储采用分布式对象存储MinIO技术方案，该技术方案具有可靠性（纠删码机制自动容错）、高可用性（在一半节点宕机时仍可保证服务可用）、可扩展性强（由于分布式特点，大数据量下可扩展至多个节点）等优点，比文件系统更可靠，便于文件的管理与迁移。使用对象存储技术为数值计算大数据管理带来很大的便捷性，在获取一个文件时不需要提供文件在文件系统中的具体位置，而是通过请求对象存储服务获得一个统一资源定位符（uniform resorce locator，URL）。其多节点的特性使得数据的安全与访问速度得到保障，扁平结构便于快速地获取数据。其弹性扩容特性使得在后期对其进行扩容变得更方便。MinIO方案专为性能和S3 API设计，非常适用于对安全性有严格要求的大型私有云环境。下面以MD为例，对材料数值计算大数据存储与管理技术进行介绍。

MD数据库共包括5个集合，即输入参数集合simulation、结果数据集合output_file、一次模拟后处理的集合analysis、多次模拟后处理的集合multi analysis和MD作业运行的集合job。各集合之间的关联关系如**图2**所示。output_file集合及analysis集合分别见**表2**和**表3**。

采用以上数据存储与管理技术，可以将不同尺度的数值计算大数据存入数据库，形成完整的材料辐照效应多尺度数值

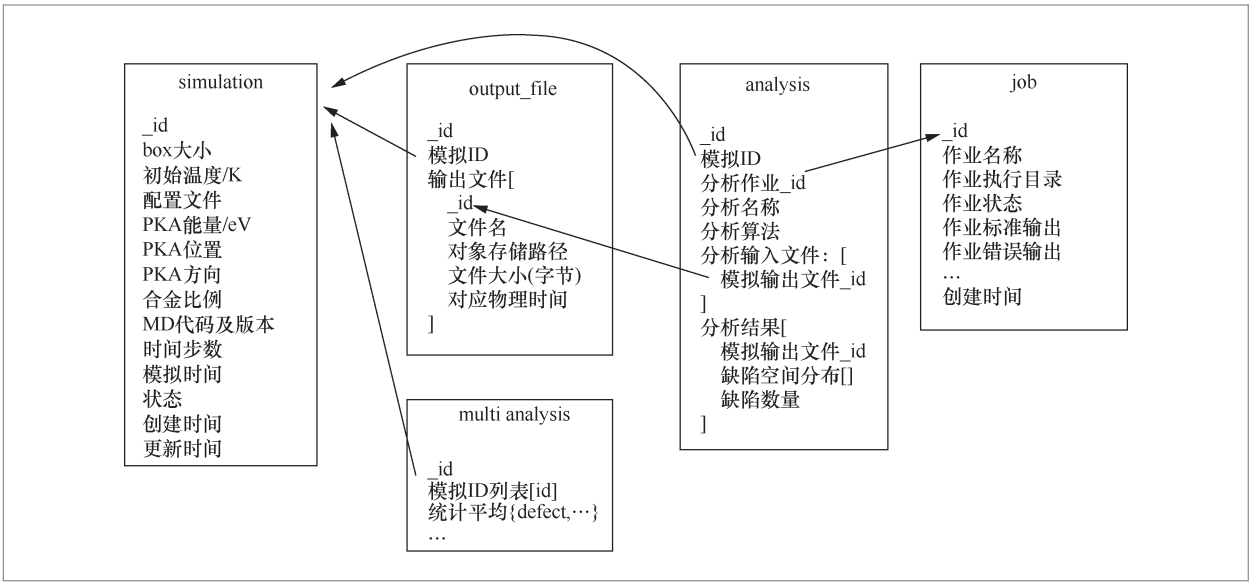

图 2 MD 数值计算大数据存储结构设计

表 2 output_file 集合

名称	示例	类型	注释
_id	ObjectId(0x1002de124)	ID	结果ID，自动生成
simulation_id	ObjectId(0x1002de123)	ID	MD模拟ID
output_file	[{“file_name”: “last_step_file.xyz”, “path”: “./md_path”, size: 10000, time: 0.10}, {“file_name”: “origin_file.xyz”, “content”: “example output”}]	array of object	MD模拟结果文件属性

表 3 analysis 集合

名称	示例	类型	注释
simulation_id	ObjectId(0x1002de123)	ID	MD模拟ID
name	“md_ws”	string	分析名称
algorithm	“md_ws”	string	分析算法
files	[ObjectId(0x1002de123)]	array of ID	分析的结果文件
results	[file_id: 0x1002de123, defect_num: 16, defects:[82.121315, 116.852413, 162.172734,1], [103.104382, 116.541125, 168.126176, 1]]	array of object	缺陷坐标defects是二维数组，第一维表示一个缺陷，第二维有4个，前3个为x、y、z，最后一个为原子的类型，1表示间隙，0表示空位； defect_num表示Frenkel缺陷对数量

计算大数据库。该数据库具有高扩展性。数据库中保存的各尺度软件的参数、结果、后处理等数据，可用于机器学习及多尺度模型改进等后续相关研究。

以一次analysis集合为例，向analysis集合插入一条“结果后处理分析”文档的语法如下：

```
db.analysis.insert({
    “simulation_id”: ObjectId(“5ef8a5358a519fed0ea721c2”),
    “name”: “md_ws”,
    “algorithm”: “md_ws”,
    “files”: [ObjectId(0x1002de123)],
    “results”:[{
        “file_id”: ObjectId (0x1002de123),
        “defect_num”: 10,
        “defects”: [
82.121315, 116.852413, 162.172734, 1], [103.104382, 116.541125, 168.126176, 1], [99.324741, 119.440361, 172.054886, 1], [101.946227, 122.593405, 170.227964, 1], [92.288785, 162.577183, 170.618859, 1], [92.173194, 164.052161, 173.610848, 1], [103.811458, 190.610901, 159.522378, 1], [117.655836, 92.663749, 119.412673, 1], [141.298869, 120.719326, 170.725907, 1], [119.952118, 154.947282, 164.86707, 1]}],
})
```

5 基于多尺度数值计算大数据的挖掘分析

为了更好地理解材料数值计算大数据的价值，本节以典型核反应堆压力容器材料Fe-1.5%wtCu合金的两个分子动力学模拟实例加以说明。本实例所用的软件为材料辐照效应分子动力学模拟软件MISA-MD，该软件能够模拟的粒子数规模达到4×10^{12}，为目前国内外能够模拟的第二大规模的分子动力学软件[14-15]。在“天河二号”的英特尔平台上，与LAMMPS软件包相比，MISA-MD的内存占用仅为前者的40%[5]。目前该款软件已经完成开源。

5.1 基于XGBoost算法的Frenkel缺陷对数预测

在对相同宏观参数下的原子体系进行多次MD级联碰撞模拟时，首先需通过随机数种子对所有粒子的速度大小和方向

进行初始化，根据麦克斯韦速度分布定律可知，对于同一宏观条件的粒子体系，各个粒子的状态是时刻变化的。因此，每一次模拟需使用不同的随机数种子，使得该宏观参数下体系的多种粒子微观状态可以得到充分考虑。在所有宏观参数一致的情况下，级联碰撞模拟得到的Frenkel缺陷的数目并非完全相同，而是在一定范围内波动。过去的做法通常需要多次执行同一宏观条件下的模拟然后取平均，例如，对于1 000×1 000×1 000的box，使用32个节点，共512个核，每次模拟程序都需要运行2~5 h，且获得模拟结果后，还需要使用最短距离标记法获取缺陷的信息，整个流程不仅耗费超算计算资源，而且时间跨度也长。本文基于MD数值计算大数据中的Frenkel缺陷提出一种更高效的Frenkel缺陷对计算方法。

本文采用机器学习中的集成学习来实现上述功能，集成学习通过构建结合多个学习器来完成学习任务，最后的结果由多个学习器共同决定。本文选取的算法是XGBoost[16]，它将许多树模型集成在一起，由这些树模型共同决定结果。

首先使用XGBoost训练所有MISA-MD模拟的数据，使用训练完的模型对未知的模拟进行缺陷对数预测。每次模拟中的box大小、晶格常数以及合金比例都为固定值，而在这些模拟之间，只有能量、入射角度、随机数、时间步长这些参数是不同的，因此将这些参数组合成特征向量，以[能量，x，y，z，随机数，时间步长]的形式。基于上述方法，对多组数据进行预测，并将其与真实值进行比较，缺陷对预测值与真实值对比见**表4**。由**表4**可知，预测值与真实值很接近，这验证了该方法的有效性。

5.2 基于并查集算法的级联碰撞团簇划分方法

级联碰撞模拟后，由于能量粒子的撞击，材料原子离开原本的晶格位置，从而发生移位，而后进一步演化发生聚集或湮灭，形成原子或空位团簇。团簇过多或过大会使得材料力学性能产生降级，从而威胁反应堆设施的安全，例如形成空洞。基于MD数值计算大数据库中的.dump数据，采用并查集算法，可以实现对团簇的有效划分。

数据集采用的晶体结构均为体心立方晶体（BCC），元素都是铁（Fe）元素，晶格常数为2.855 32，box大小均为[80, 80, 80]，它的含义是x、y、z方向上都是80倍的晶格常数，即80个晶格点。实验环境在600 K的温度下，根据入射中子能量的不同，时间步数有10 000个和100 000个两种，

表4 缺陷对预测值与真实值

能量	x	y	z	随机数	时间步数/个	缺陷对预测值	缺陷对实际值
10	1	2	2	1 012 316	10 000	13	13
10	1	3	5	1 013 617	10 000	19	21
10	2	3	5	1 023 618	10 000	19	20
50	1	2	2	5 012 340	100 000	109	115
50	1	3	5	5 013 633	100 000	101	121
50	2	3	5	5 023 636	100 000	89	84

总的时间步数有41 000个和131 000个两种。MISA-MD运行时，每隔1 000个时间步输出一个结果，这里选取最后一个时间步的结果。每个时间步的结果数据都是.dump坐标数据，里面包含1 024 000个原子坐标。

在上述实验环境下，数据涵盖不同能量、不同角度，且每种能量每种角度都进行了多次模拟。数据包括从10 keV、30 keV、50 keV 3种不同的能量，角度分为122、135、235这3个方向，每种都进行了50次模拟，最终有450次模拟数据。

常规做法是将每个缺陷看成一个单缺陷的团簇，然后遍历其他所有缺陷，将指定距离内的缺陷加入该团簇，进行缺陷的合并。这看起来并不复杂，但是当数据量大时，若采用常规方法来解决，往往时间复杂度过大，因为它需要反复查找一个缺陷所在的团簇，导致不能很好地解决该问题。因此在这里采用并查集算法来解决。并查集采用一种树形数据结构来处理这种不相交集合的问题。并查集算法有两种操作：合并（把两个不相交的集合合并为一个集合）、查询（查询两个元素是否在同一个集合）。将所有元素合并完之后，森林中有几棵树，就有几种集合。因为并查集的数据结构为树形，所以树的高度越高，时间复杂度也越高。这里选取的是优化的并查集算法。

如伪代码1所示，首先设置一个大小与缺陷总数相同的根节点数组root，它的含义为该缺陷所属团簇的编号，初始时，每个缺陷被视为单独的一个团簇，因此初始数组的值为自身编号。再设置一个大小与缺陷总数相同的数组height，它表示以当前节点为根节点的树的高度，因为初始时每个缺陷都是一个团簇，也就是一棵树，所以初始时树的高度都为1。接下来计算任意两个缺陷之间的距离，在计算的过程中需要判断这两种缺陷的类型。如果这两个缺陷都是间隙原子或者一个是间隙原子、一个是空位，且它们的距离在一倍晶格常数（第二近邻）内，就认为它们属于同一个团簇；如果两个缺陷都是空位，且它们的距离在$\sqrt{2}$倍晶格常数（第三近邻）内，就认为它们属于一个团簇。如**图3**所示，此时缺陷2和缺陷9在距离阈值内，第一步先查找两个缺陷的根节点，在查找的过程中，将向上经过的所有缺陷的根节点都设为最上层那个缺陷，也就是都直接接到根节点上，这被称为路径压缩，可以减少树的高度，使得以后向上查找根节点时速度更快。获取根节点后，根据height数组判断两个根节点的树的高度，将高度小的树接到高度大的树上，如果树高一样，则任意将一棵树接到另一棵树上作为孩子节点。遍历根节点数组，若根节点相同的缺陷为一棵树上的，则将根节点相同的缺陷划分到一个团簇中，从而获取缺陷可以划分的所有团簇。将获得的所有团簇信息（包括团簇中缺陷坐标、缺陷对数、缺陷类型（间隙或者空位）等）存储到团簇数据库中，最初获得了4 483个团簇。

伪代码1 使用优化的并查集划分团簇

input：The coordinates of all defect atoms: DEFECTS = [d_1, d_2, …, d_m]

output：All clusters

step：

```
1   Set a root node array and tree height array: root = [1,…,m], height = [1]*m
2   for i ← 1, 2, …, m do
3     for j ← i+1, …, m do
4       if distance(d_i, d_j) within threshold then
5         a ← Find the root node of i
6         b ← Find the root node of j
7         Modify the root array according to the tree height
```

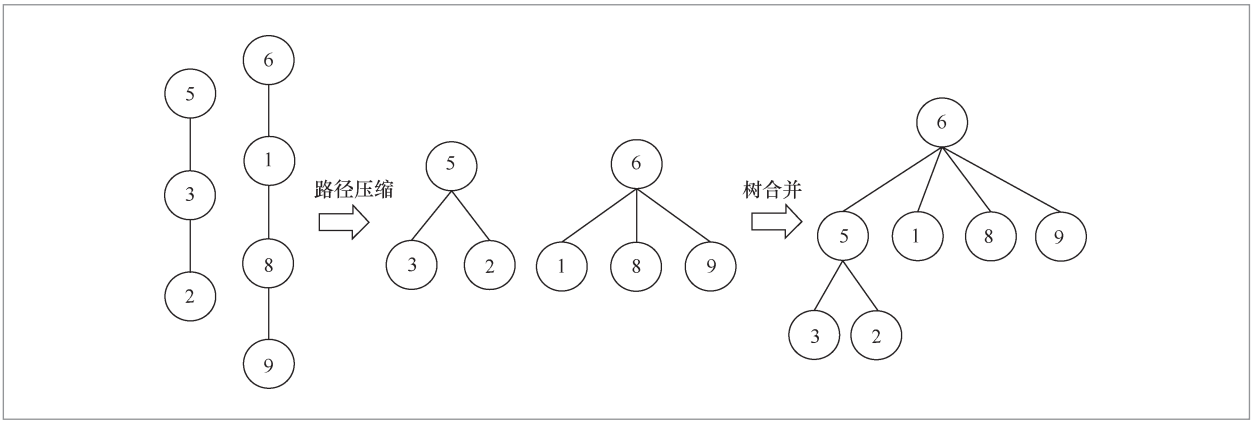

图 3 并查集算例演示

8 end if

9 end for

10 end for

11 Divide the defects with the same root node into a cluster

12 return all cluters

5.3 基于聚类算法的KMC长程演化类环状原子簇发现

通过对数值计算大数据库中的MISA-KMC长程演化团簇数据进行分析，发现了材料辐照效应中的类环状团簇。选取的特征向量为缺陷团簇中各缺陷与几何中心的距离，以及每两个缺陷与几何中心形成的夹角。考虑到几何形状经旋转、放大及缩小后，形状仍然是相同的，这里每隔5° 形成一维数据，共36维数据；对于距离，每次将所有距离除以当前团簇的最大值，进行归一化处理，每隔0.025形成一维数据，共40维数据，因此特征向量包含76维数据，如**图4**所示。选取HDBSCAN（hierarchical density-based spatial clustering of applications with noise）聚类算法对团簇进行识别。它是一种基于密度聚类的无监督的聚类算法，不需要已经标记的数据，也无须事先知道要划分的类别数。它可以对不同密度的团簇进行聚类，可以忽略噪声，且效率较高。团簇聚类的结果如**图5**所示。从**图5**可以看到，将所有的缺陷团簇分为几种不同的类别，每种颜色代表一种类别，每种类别的团簇的几何形状相同或相近。基于该方法，笔者在KMC长程演化数据中发现了一些类环状的团簇，如**图6**所示，这一发现验证了之前报道的材料辐照实验中缺陷团簇的出现[17-18]。

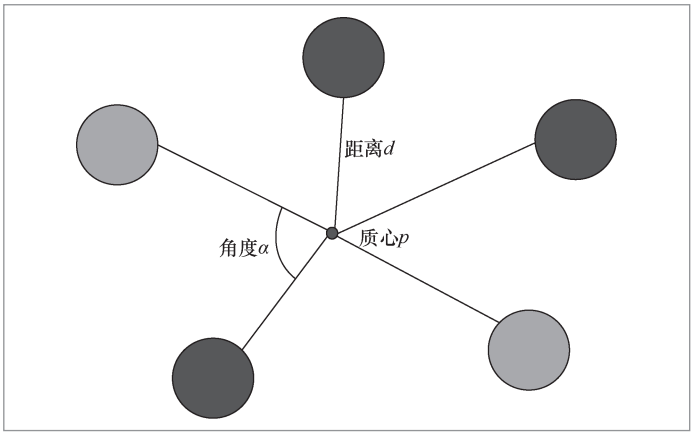

图 4 团簇特征提取方法

5.4 基于神经网络的势函数模型AIPM

势函数计算是材料多尺度模拟关键的一环，MD和KMC中粒子速度、位置的更

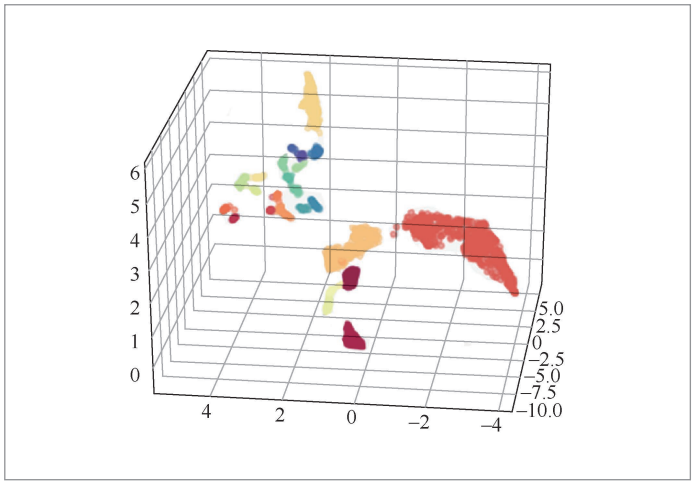

图 5 团簇识别结果

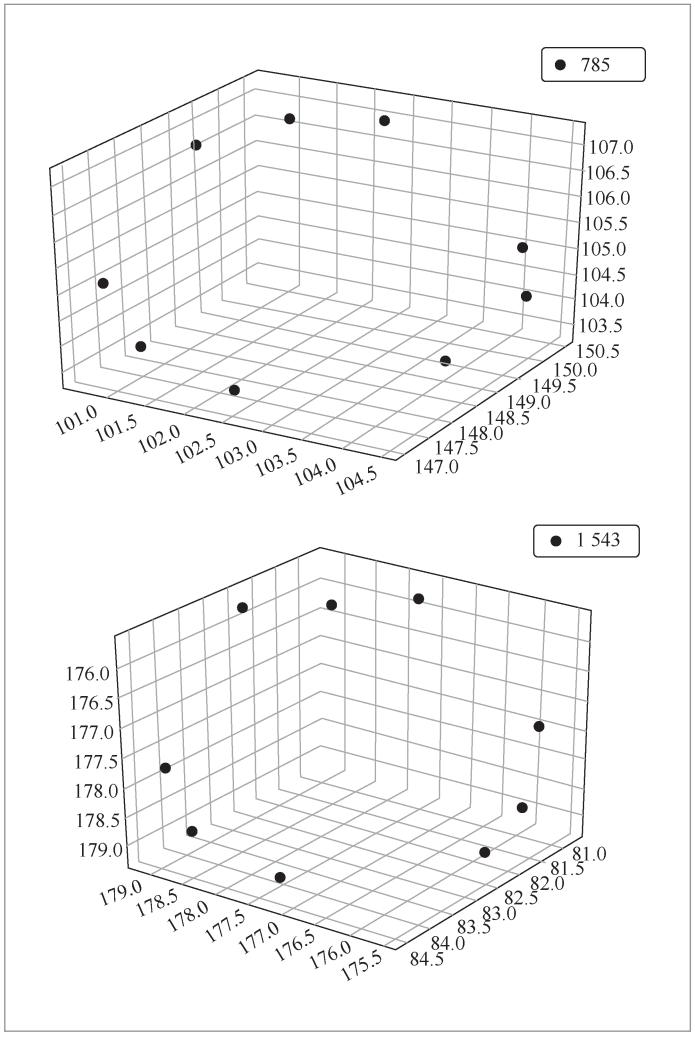

图 6 KMC 长程演化产生的类环状团簇识别结果

新，以及SCD中多元组分材料参数的计算，均离不开势函数模型。过去常用的势函数模型通常包含两种，一种基于第一性原理，另一种基于经验函数。前者往往计算复杂，且对于多元组分而言，第一性原理势函数的构建过程非常复杂；后者虽然在效率上有所提高，但精度往往不够，而对多元合金组分的经验势函数构建过程则更加困难。

针对上述问题，基于第一性原理数值计算大数据，提出了一种基于机器学习的方法对原子体系模拟参数及势能之间进行拟合的势函数模型AIPM（artificial intelligence based potential model）。这里选取Fe-Cu二元合金体系，基于原子坐标进行机器学习模型的特征提取，如**图7**所示。首先按照最近邻法对原子邻域进行划分，并以该原子为中心，建立局域坐标系，第一近邻和第二近邻分别设置为x、y坐标，二者的向量积作为z坐标，于是可以得到每个原子的坐标，将这些坐标作为神经网络的输入。这里选取3层神经网络的结构，如**图8**所示，每层的节点数依次为15、10、6，拟合得到体系内一个原子的势能，然后针对其他原子采用相同的方案进行拟合，最后将所有的原子的势能求和，即可得到总的原子体系的势能。将这一势能与数据库中给定的经典势能——EAM势能进行比较，以验证模型的精度。采用AIPM模型对1 000个粒子大小的Fe-Cu原子体系势能进行计算，验证了AIPM的可靠性。Fe-Cu原子体系神经网络计算结果见**表5**。

6 结束语

本文首次提出了材料数值计算大数据的概念，阐述了材料数值计算大数据的特点及研究意义，提出了一种适用于材料数

值计算的数值计算大数据存储体系，并基于该数据体系，在Frenkel缺陷对计算、MD中的缺陷团簇划分、类环状团簇发现以及势函数模型构建等多个方面取得了进展。尽管数值计算大数据很早就出现在研究工作中，但系统性的研究仍处于起步阶段。随着数值计算的规模越来越大以及技术上的瓶颈越来越多，数值计算大数据的研究在材料多尺度模拟研究中起到了越来越重要的作用，其价值有待进一步挖掘，尤其在改进物理模型和软件耦合方面，数值计算大数据将成为突破多尺度模拟难点和挑战的重要途径和手段。

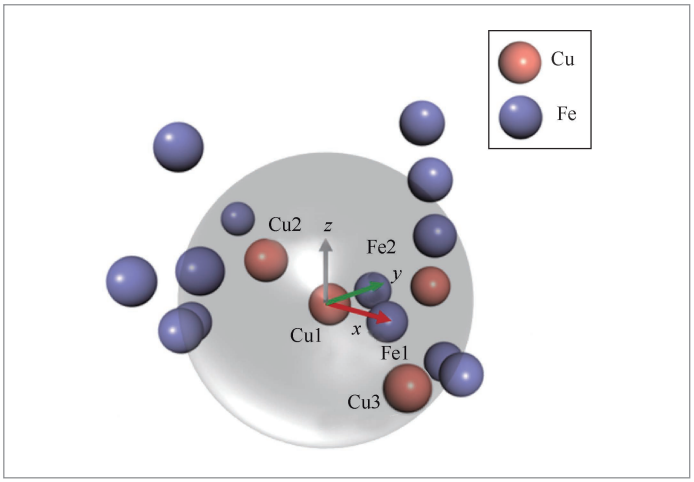

图 7　局域坐标系的建立方法

表 5　Fe-Cu 原子体系神经网络计算结果

模型	原子体系	总势能/J	计算时间/min
EAM	Fe-Cu	1.449 0	20.332
AIPM	Fe-Cu	1.459 2	18.769

参考文献:

[1] WANG J, LIU C, HUANG Y H. Auto tuning for new energy dispatch problem: a case study[J]. Future Generation Computer Systems, 2016, 54: 501-506.

[2] LI S G, ZHANG Y Q, XIANG C Y, et al. Fast convolution operations on many-core architectures[C]// Proceedings of 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th

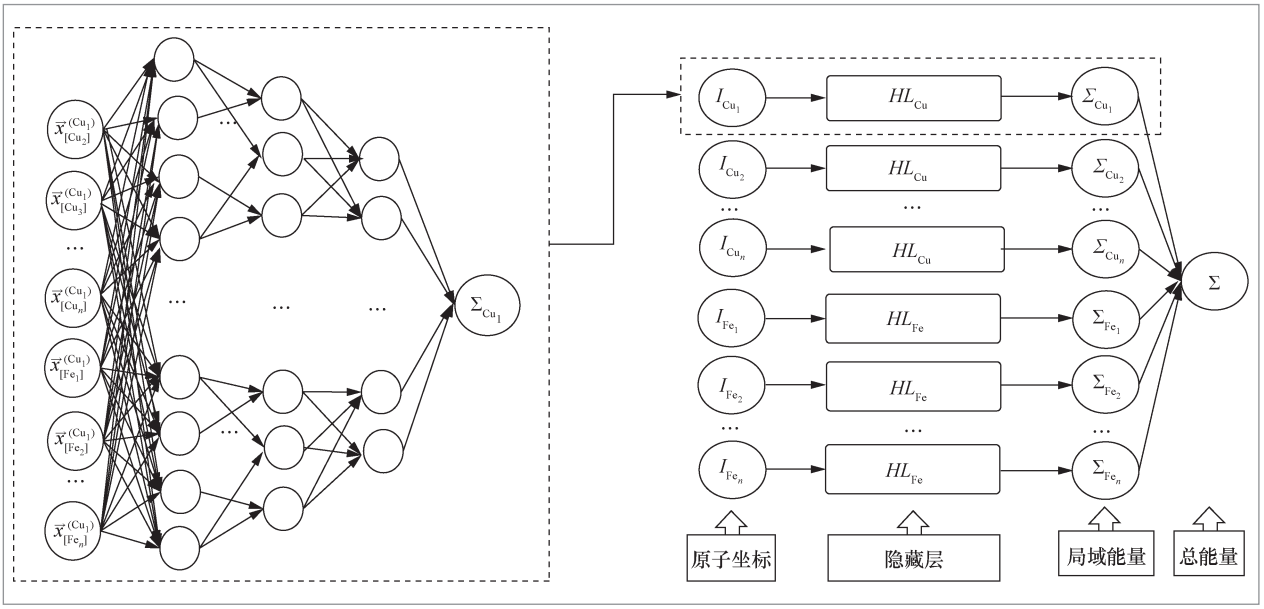

图 8　Fe-Cu 原子体系神经网络构建过程

International Conference on Embedded Software and Systems. Piscataway: IEEE Press, 2015: 316–323.

[3] SAND A E, DUDAREV S L, NORDLUND K. High–energy collision cascades in tungsten: dislocation loops structure and clustering scaling laws[J]. EPL (Europhysics Letters), 2013, 103(4): 46003.

[4] WANG J L, ENOMOTO M, SHANG C J. First–principles study on the interfacial segregation at coherent Cu precipitate/Fe matrix interface[J]. Scripta Materialia, 2020, 185: 42–46.

[5] HU C J, BAI H, HE X F, et al. Crystal MD: the massively parallel molecular dynamics software for metal with BCC structure[J]. Computer Physics Communications, 2017, 211: 73–78.

[6] KAWAMURA T, NODA T, IDOMURA Y. In–situ visual exploration of multivariate volume data based on particle based volume rendering[C]// Proceedings of 2016 2nd Workshop on In Situ Infrastructures for Enabling Extreme–Scale Analysis and Visualization. Piscataway: IEEE Press, 2016: 18–22.

[7] BHARDWAJ U, SAND A E, WARRIER M. Pattern matching and classification of clusters in collision cascades[J]. arXiv preprint, 2018, arXiv:1811.10923.

[8] BHARDWAJ U, HEMANI H, MAJALEE A, et al. Analysis and visualization of collision cascades data[J]. Computational Materials Science, 2020, 172: 109364.

[9] BHARDWAJ U, SAND A E, WARRIER M. Graph theory based approach to characterize self interstitial defect morphology[J]. Computational Materials Science, 2021, 195: 110474.

[10] PODRYABINKIN E V, TIKHONOV E V, SHAPEEV A V, et al. Accelerating crystal structure prediction by machine–learning interatomic potentials with active learning[J]. Physical Review B, 2019, 99: 064114.

[11] PILANIA G, WANG C C, JIANG X, et al. Accelerating materials property predictions using machine learning[J]. Scientific Reports, 2013, 3: 2810.

[12] JIA W L, WANG H, CHEN M H, et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning[C]// Proceedings of the 2020: International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE Press, 2020: 1–14.

[13] 汪岸, 任帅, 苗雪, 等. 数值核反应堆大数据及其应用[J]. 大数据, 2021, 7(5): 40–56.
WANG A, REN S, MIAO X, et al. Big data of numerical nuclear reactor and its application[J]. Big Data Research, 2021, 7(5): 40–56.

[14] LI S G, WU B D, ZHANG Y Q, et al. Massively scaling the metal microscopic damage simulation on Sunway TaihuLight supercomputer[C]// Proceedings of the 47th International Conference on Parallel Processing. New York: ACM Press, 2018: 1–11.

[15] TCHIPEV N, SECKLER S, HEINEN M, et al. TweTriS: twenty trillion–atom simulation[J]. The International Journal of High Performance Computing Applications, 2019, 33(5): 838–854.

[16] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785–794.

[17] AUGER P, PAREIGE P, AKAMATSU M, et al. APFIM investigation of clustering in neutron–irradiated Fe–Cu alloys and pressure vessel steels[J]. Journal of Nuclear Materials, 1995, 225: 225–230.

[18] ETIENNE A, HERNÁNDEZ–MAYORAL

M, GENEVOIS C, et al. Dislocation loop evolution under ion irradiation in austenitic stainless steels[J]. Journal of Nuclear Materials, 2010, 400(1): 56-63.

作者简介

任帅(1992-),男,北京科技大学计算机与通信工程学院博士生,主要研究方向为机器学习、大数据及数据挖掘等。

陈丹丹(1995-),女,北京科技大学计算机与通信工程学院博士生,主要研究方向为软件工程、数值计算及数据挖掘等。

储根深(1994-),男,北京科技大学计算机与通信工程学院博士生,主要研究方向为并行算法、数值计算及材料多尺度算法等。

白鹤(1992-),男,北京科技大学计算机与通信工程学院博士生,主要研究方向为并行算法、数值计算等。

李慧昭(1999-),男,北京科技大学计算机与通信工程学院硕士生,主要研究方向为计算机科学与技术、机器学习等。

何远杰（1999– ），男，北京科技大学计算机与通信工程学院硕士生，主要研究方向为计算机科学与技术、大数据等。

胡长军（1963– ），男，博士，北京科技大学计算机与通信工程学院教授、博士生导师。智能超算融合应用技术教育部工程研究中心主任，北京科技大学学术委员会委员，北京科技大学计算机科学技术学科负责人，北京市重点学科计算机系统结构负责人，计算机学会高性能计算专业委员会委员，*Journal Citation Reports*等国际期刊客座编辑。曾任清华大学信息技术研究院特聘研究员、中国原子能科学研究院特聘专家，主要研究方向为大数据工程及计算智能、超级计算机体系结构及系统软件、大规模并行应用软件系统。

收稿日期： 2021-04-29
通信作者： 陈丹丹，chendandan@xs.ustb.edu.cn
基金项目： 国家自然科学基金资助项目（No.U1867217）
Foundation Item: The National Natural Science Foundation of China (No.U1867217)