ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.19, No.1, January 2023

E-mail: kfyj@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

DOI:10.14004/j.cnki.ckt.2023.0011

省级重点实验室高性能计算平台的建设研究

朱莹[1,2],于泠[1],陈文通[1]

(1.江苏省大规模复杂系统数值模拟重点实验室,江苏 南京 210023;2.南京师范大学江苏省大型科学仪器开放实验室,江苏 南京 210023)

摘要:为满足各学科日益增长的高性能计算需求,2018年江苏省"大规模复杂系统数值模拟"重点实验室建成了高性能计算二期平台。该文详细阐述了平台建设的背景、软硬件配置情况以及系统性能测试结果,分析了计算平台的使用情况及科研产出,对实验室高性能计算平台的建设进行了深入的思考。

关键词:高性能计算;大规模复杂系统数值模拟;实验室HPC平台

中图分类号:G482 **文献标识码**:A

文章编号:1009-3044(2023)01-0121-03

开放科学(资源服务)标识码(OSID):

高性能计算(High Performance Computing,HPC)又称超级计算或并行计算,是应用超级计算机与并行处理技术解决大规模复杂计算问题的一种技术手段。目前高性能计算能力已经成为衡量一个国家综合国力的重要标志,是国家信息化建设的根本保证。

高性能计算平台为大规模复杂系统的研究提供了计算服务,是高校、科研院所进行科学研究、高层次人才培养、学科建设的支撑平台[1-2],因此,越来越多的高校都搭建了适合自身需求的高性能计算平台。随着大数据、云计算、人工智能的发展,高性能计算的新场景不断涌现,对平台也提出了更高的需求。

1 实验室HPC平台建设背景

大规模计算对HPC平台提出的需求不断增长,作为科学研究与人才培养重要基地的高等院校,建设HPC平台势在必行。

以江苏省部分高校为例,南京大学高性能计算中心目前拥有910台计算节点,理论CPU计算峰值为870TFlops,在2017年6月发布的全球超级计算机Top500排行榜中列第284位[3];东南大学大数据计算中心先后建设了两套集群,目前共有501台计算节点,理论CPU计算峰值为366.5 Tflops;南京信息工程大学高性能计算中心目前拥有172台计算节点,理论CPU计算峰值为180TFlops;南京航空航天大学高性能计算中心于2020年完成一期平台建设,共有64台计算节点,理论CPU计算峰值为204.8Tflops。

江苏省"大规模复杂系统数值模拟"重点实验室的前身是1995年成立的原南京师范大学科学与工程计算校重点实验室,2010年成为江苏省重点实验室。实验室不仅为大规模数值模拟提供软硬件服务,还为省经济建设、环境保护、疾病控制、交通管理等提供决策依据和解决方法。

实验室自成立以来,先后构建了两套HPC平台。一期平台于2012年建成,共108台计算节点,总浮点运算峰值超过14TFlops,存储容量达到198TB,是校级"共享型"计算平台,对我校化学、地理、生物、数学、物理等学科的科学研究和人才培养发挥了重要的作用[4]。

随着科学技术的不断发展,各学科对高性能计算的要求越来越高。

实验室各学科在科研项目年均使用机时数量级在十万到百万级别,潜在科研项目的需求更多,原有平台可用计算机时已经无法满足科研需求,实验室通过调研论证,在2018年建成第二套HPC平台,并于2018年4月正式投入使用。

2 实验室HPC平台的构建

构建一套完整的实验室HPC平台,需要进行一系列前期准备工作,对HPC平台的软硬件进行合理配置,并对计算系统性能进行测试。

2.1 前期准备

1)机房选址

为确保高性能计算集群正常运行,HPC平台除了各类计算服务器、存储器等组成的计算系统外,还包

收稿日期:2022-10-30
基金项目:江苏省大型科学仪器开放实验室项目,项目名称:高性能计算共享平台运行与服务模式的探索研究(项目编号:BM2020018)
作者简介:朱莹(1980—),女,江苏无锡人,实验师,硕士,研究方向为高性能计算平台建设与系统管理、并行计算方法。

含配电、制冷、消防和环境监控等辅助系统。因此，高性能计算机房的选址需考虑到面积、承重、层高、防潮、防盗、防干扰等问题。根据调研，国内大部分高性能计算机房一般都位于某大楼的一层或者负一层，而本实验室一期HPC平台位于办公楼五层，占地面积及楼层承重均无法满足二期平台的建设。从承重、安全及散热等角度出发，选择在非架空的大楼一楼建设二期平台。该选址达到了相应的环境需求。

2) 机房环境搭建

二期HPC平台机房有主机房、配电间和监控室3个房间，整体机房面积约152 m²，为HPC平台正常运行提供了环境支撑。

主机房内放置了一组冷通道封闭模块，包括IT机柜13架，用于放置计算系统的服务器、存储器、交换设备等；行间空调5台；电源头柜2架。其中5台总制冷量为190kW的行间精密空调设为4+1备份模式，全机组定期轮巡，最大限度提高了制冷系统的利用率和冗余度。此外，主机房内还配备了无管网七氟丙烷气体消防系统。

配电房内配备了一套配电系统及两套200kVA的UPS，后备电池时间能满足设备满载不低于30分钟。该套配电系统能确保计算系统采用UPS双路供电，行间空调采用市电供电。配电间也配备了无管网七氟丙烷气体消防系统。

监控室主要放置了用于对机房的所有动力设备以及环境参数进行实时检测并起预警的环境监控系统监视终端以及计算系统的管理、任务调度监视终端。

3) HPC平台计算系统设备选型

实验室HPC平台面向的学科种类繁多，应用各异，既有成熟商业软件、主流开源软件，又有众多自编程序。从一期使用情况来看，平台上有MPI并行程序、多线程并行程序、GPU程序和众多的串行程序。不同的应用对硬件资源的需求也千差万别，有计算密集、网络密集型、IO密集型和耦合密集型等不同需求。二期平台在设备选型时遵循高性能、低功耗、易管理、可扩展的总体设计原则，采用刀片式、胖节点、GPU节点、MIC节点服务器相结合的集群系统架构，配置高速的InfiniBand网络，以满足高带宽和低延迟的特性，配置容量为2PB的ESS存储系统，采用专业、可靠的国际主流商业并行文件系统，满足海量存储空间的需求特点。

2.2 HPC平台计算系统整体架构

二期平台系统拓扑图如图1所示，系统总浮点运算峰值超过116TFlops，存储总裸容量达到2PB，包含112台并行计算节点，4台登录管理节点，2台胖节点，2台GPU节点和1台MIC节点等主要硬件设备。具体配置如表1。

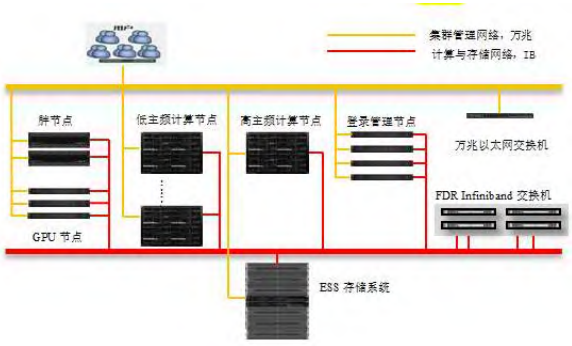

图1　HPC平台计算系统拓扑图

表1　HPC平台硬件配置表

硬件名称	配　置	数量	功　能
管理登录节点	X3650M5机架式服务器 CPU：2*Intel Xeon E5-2640 v4 10C 2.5GHz 内存：8*8GB TruDDR4； 硬盘：2*600GB 10K 12Gbps 2.5寸SAS	4	用户登录HPC平台以及HPC作业的调度管理
低主频计算节点	Flex System x240 M5刀片节点 CPU：2*Intel Xeon Processor E5-2680 v4 14C 2.4GHz 内存：8*16GB TruDDR4 硬盘：1*240GB 10K 12Gbps 2.5寸SSD	98	用于普通HPC任务
高主频计算节点	Flex System x240 M5刀片节点 CPU：2* Intel Xeon Processor E5-2667v4 8C 3.2 GHz 内存：8*16GB TruDDR4 硬盘：1*240GB 10K 12Gbps 2.5寸SSD	14	用于需高主频处理器的HPC任务。
胖节点	X3850M5机架式服务器 CPU：4* Intel Xeon Processor E7-8890 v4 24C 2.2 GHz 内存：64* 16TruDDR4 硬盘：4*600GB 10K 12Gbps 2.5寸SAS	2	用于需海量内存的大型HPC任务
GPU节点	X3650M5机架式服务器 CPU：2*Intel Xeon Processor E5-2680 v4 14C 2.4GHz 内存：8*16GB TruDDR4 硬盘：2*600GB 10K 12Gbps 2.5寸SAS GPU卡：1块NVIDIA Tesla K40	2	用于需GPU计算的任务
MIC节点	X3650M5机架式服务器 CPU：2*Intel Xeon Processor E5-2680 v4 14C 2.4GHz 内存：8*16GB TruDDR4 硬盘：2*600GB 10K 12Gbps 2.5寸SAS 协处理器：1块Intel Xeon Phi 7120A	1	用于高度并行化应用的任务
磁盘阵列	IBM Elastic Storage Server GL6高性能存储阵列 含2台IO节点，可用容量达到1.5PB	1	用于数据存储
计算网络	Mellanox SX6025，36口IB 56Gb FDR交换机	5	用于HPC计算

2.3 HPC平台计算系统软件配置

该平台采用Linux操作系统，为校内外用户提供统一的登录界面，为用户提供了一个稳定、安全、高效的高性能计算共享平台。具体软件配置如下：

1) 操作系统：采用CentOS6.8 Linux 64位操作系统。

2) 编译器：Intel ParallelStudioXE2018.1.163(C++/Fortran)。

3) 数学库：Intel MKL。

4) MPI实现：openmpi3.0.0/openmpi2.1.1。

5) HPC集群管理软件：IBM XCat2.11。

6) HPC集群作业调度软件：Easycluster v1.6。

7) 并行文件系统：IBM Spectrum Scale。

8) 其他应用软件：平台还安装了VASP、Python、Gromacs、Lammps、Gaussian、WRF等众多学科领域用户常用的科学计算软件。

2.4 计算系统性能测试结果

笔者对计算系统的linpack值和存储IO性能分别进行了测试，测试结果都超过预期。

针对每种架构，使用相同架构的所有节点，采用Intel linpack测试工具，进行Linpack性能测试，得到每种架构节点服务器的Linpack测试结果。

1) 14台高主频计算节点Linpack测试结果：实测值为10901Gflops，效率为 10901/(14*2*8*3.2*16)=95.05%。

2) 98台低主频计算节点Linpack测试结果：实测值为87144Gflops，效率为87114/(98*2*14*2.4*16)=82.70%。

因此，平台HPC系统整体计算能力理论值为116838.4 Gflops，实测值为98046.1Gflops，Linpack实测效率为98046.1/116838.4=83.91%。

采用IOzone工具测试文件系统的读写性能[5]。

聚合带宽实际测试结果如表2所示。

表2　聚合带宽实际测试结果表

操 作	测试文件大小(GB)	节点内存大小(GB)	-r 1024k	-r 4096k	-r 8192k
Write	128*32	128	23.2	23.3	23.2
Read	128*32	128	30.3	30.3	30.4

单节点带宽测试命令如下：

iozone -i 0 - i 1 -+n -r 1024k -s 128g -t 16 -e -+m /xxx/host.list -Rb /cclba/output.xls

单计算节点带宽测试结果见表3所示。

表3　单计算节点带宽测试结果表

操 作	测试文件大小(GB)	节点内存大小(GB)	-r 1024k	-r 4096k	-r 8192k
Write	128*16	128	5.73	5.75	5.74
Read	128*16	128	6.36	6.36	6.36

综上，系统IO实测总聚合读带宽大于30GB/s，实测总聚合写带宽大于23GB/s，单计算节点实测读带宽6.36 GB/s，单计算节点实测写带宽5.35 GB/s，文件系统读写性能均超过了预期，为整个平台系统的高可用性提供了保障。

3 结束语

经过3年的运行，江苏省“大规模复杂系统数值模拟”重点实验室HPC平台为数学、计算机、物理、化学、地理、生物、能源与机械、食品与制药等众多学科领域的用户提供了高性能计算和数值模拟的软硬件服务环境，满足了实验室的科研计算需求，推动了创新人才培养，促进了学科间的交叉和融合，成效显著。

开放共享是大型平台的必然趋势，作为省级实验室的平台，我们将进一步探索开放共享模式，统筹平台管理，优化软件，完善平台激励机制，根据平台特点推进开放共享，开发针对实验室高性能计算平台的服务管理系统。

参考文献：

[1] 解通，孙勇，魏泽发，等．校级高性能计算平台建设的探索与实践[J]．软件，2021，42(6)：74-76，80.

[2] 黄建强，孟永伟，曹腾飞，等．青海大学三江源数据分析中心高性能计算集群的构建与设备管理[J]．实验技术与管理，2014，31(12)：237-240.

[3] 盛乐标，周庆林，游伟倩，等．高校大型高性能计算平台建设模式探讨[J]．实验室科学，2019，22(6)：158-161.

[4] 于泠，王雨顺，朱莹，等．科学计算中心建设与服务的探索与实践[J]．实验技术与管理，2015，32(2)：159-162.

[5] 游伟倩，盛乐标，周庆林，等．高性能计算集群存储系统搭建方式的对比研究[J]．电脑知识与技术，2018，14(9)：22-23.

【通联编辑：唐一东】

(上接第113页)

[4] 宋轩，高云君，李勇，等．空间数据智能：概念、技术与挑战[J]．计算机研究与发展，2022，59(2)：255-263.

[5] 郭名静，景琳．空间数据挖掘驱动城市疫情监测常态化的作用研究[J]．商业经济，2022(2)：11-13，16.

[6] 苏瑶．基于GIS+BIM的空间数据可视化研究[J]．自动化与仪器仪表，2021(12)：28-31，35.

[7] 周群．基于GIS的综合交通空间数据管理系统研究[J]．地理空间信息，2021，19(11)：75-78，8.

[8] 杜雪，王景弟，白彦锋，等．基于克里金插值法的湖南省慈利县森林碳储量专题图研究[J]．西北林学院学报，2022，37(1)：198-204.

[9] 陈新．论空间数据挖掘和知识发现的理论与方法[J]．电脑知识与技术，2021，17(33)：20-21，31.

[10] 崔梦真，贺晗，王虎，等．便携式采集装置及铁路基础设施三维空间数据管理系统设计[J]．铁路计算机应用，2021，30(10)：24-29.

[11] 石伟伟，刘皓宇，程丽丽，等．超大规模空间数据管理及计算框架关键技术研究[J]．国土资源信息化，2021(5)：15-21.

[12] 胡平昌，孙毅中，朱杰，等．顾及共享交换的行业地理空间数据关联模式[J]．测绘地理信息，2021，46(5)：127-130.

[13] 刘善磊，张大骞．多源矢量空间数据关联分析及应用[J]．测绘与空间地理信息，2021，44(8)：68-70，74.

[14] 栾绍鹏，徐东升，刘洪春，等．考虑卡尔曼滤波的地理空间数据局部特征分布[J]．测绘科学技术学报，2021，38(4)：410-415.

【通联编辑：梁书】