能计算，比你想象的更加有用

高性能计算应用概览

历军 等 编著

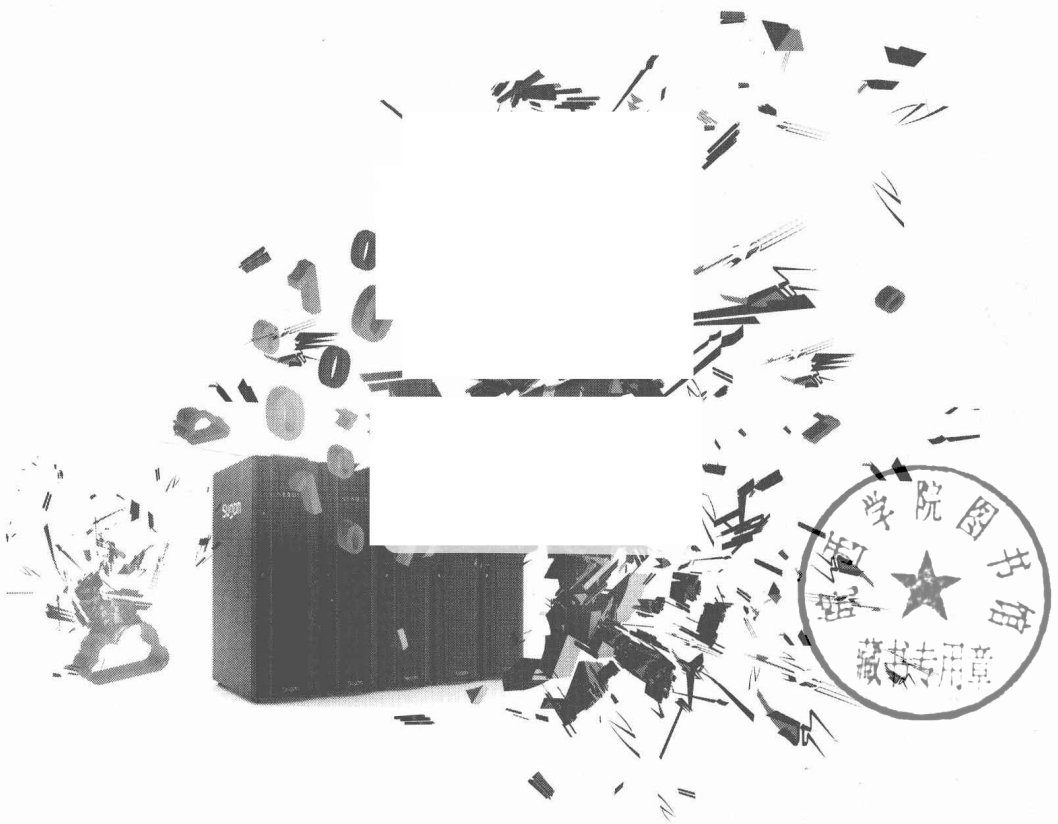
HPC: 大工具成就大事
HPC: Big Tool, Big Mission
高性能计算就在身边

HPC不仅仅是人工智能
高性能计算决定未来

清華大學出版社

高性能计算应用概览

历军 等 编著



清華大學出版社
北京

内 容 简 介

高性能计算技术，又称超级计算技术，近年来受关注度较高。从天河夺冠到谷歌的 AlphaGo，我们常常可以看到高性能计算的身影。对于中国来说，高铁和高性能计算并称为两大中国可以比肩甚至超过美国的技术。然而，高性能计算的应用却往往躲在屏幕之后，不为大众所熟知。此外，由于高性能计算的产业链长而复杂，很多 IT 研究人员对应用了解的并不多，而且不同应用方向的研究人员之间也是隔行如隔山。其次，中国的高性能计算机如星云、天河、太湖之光都在世界排名中名列前茅，不少国外的专家质疑我们只是用钱堆出个机器，并不是真正地把高性能计算机用起来。本书在一定程度上可以对上述这些问题有所解答。

本书由中科曙光牵头，参与单位近二十家，比较全面地反映了中国高性能计算应用的现状。主题内容涵盖高性能技术简介，高性能计算在材料研究、生物信息、大气海洋与气候研究、工业仿真、石油勘探与加工、渲染、遥感，以及深度学习等方面的应用。同时我们邀请了任职于中国超算创新联盟等的知名专家和企业家对未来的高性能计算技术和应用做了前瞻性展望，希望各位读者能够有所收获。

图书在版编目(CIP)数据

高性能计算应用概览/历军等编著. —北京：清华大学出版社，2018（2018.8重印）
ISBN 978-7-302-50472-6

Ⅰ. ①高…　Ⅱ. ①历…　Ⅲ. ①高性能计算机－研究　Ⅳ. ①TP38

中国版本图书馆 CIP 数据核字(2018)第 126839 号

责任编辑：贾　斌　薛　阳
封面设计：刘　键
责任校对：焦丽丽
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：20.25　　**字　　数**：490 千字
版　　次：2018 年 6 月第 1 版　　**印　　次**：2018 年 8 月第 2 次印刷
印　　数：2001～3000
定　　价：69.80 元

产品编号：077362-01

本书编委会

主　编　历　军

参　编　（姓氏笔画排序）

卜景德　马天星　王国江　王彦棡　王　鹏
方　娟　方跃文　帅　威　丛维涛　吉　青
许　涛　孙凝晖　杜夏威　李　峰　李　斌
杨　光　何铁宁　况吕林　迟学斌　张永民
张郭亮　陈　芳　林　茂　金　钟　赵江伟
赵琉涛　胡玉新　胡　辰　段纯刚　姜金良
顾蓓蓓　晏平仲　徐春明　高金森　席　萌
黄志新　梅林涛　塔依尔·伊布拉音
彭延国　韩　叙　蓝兴英　谭光明　翟　健
颜深根　潘震西　戴　荣

前言

高性能计算，又称超级计算，是计算机科学重要的前沿性分支，它不仅是一个国家综合科研实力的体现，更是对国家安全、经济和社会发展具有举足轻重的意义，是公认的国家科技发展水平和综合国力的重要标志，已成为各国竞相抢占的科技竞争战略制高点，全球仅美国、日本、中国拥有超级计算技术。高铁和超级计算是美国唯一公开承认中国能与之比肩甚至超过美国的技术和产业。

高性能计算是科技的基础产业，应用上可支撑：核试验模拟、石油勘探、气象预报、农业育种、医疗服务、新药研制、动漫渲染、材料设计、金融计算等，几乎涉及人类科学和生活的每一个领域。一般来说，凡是需要大规模数值模拟计算和大规模数据分析处理的情形都可以利用超级计算机进行加速，同时还可以协助探索超宏观（如宇宙）、超微观（如纳米级）、极端环境（如人造太阳）等实际工作环境难以实现的研究。据 IDC 报告，2015 年全球超算市场规模约 250 亿美元，其中，高性能计算机系统（包括服务器、存储和网络）约占 60%，软件和服务约占 35%；并预测 2015—2020 年超算市场规模将以 8.3%的复合增长率迅速扩大，在 2020 年将达到 440 亿美元。

另外，深度学习和人工智能被认为是 2020 年前最有希望颠覆人类生产和生活的技术，而它与超级计算密不可分。从深度学习的模型训练，到模型推理都依赖于超算技术。目前，深度学习已经渗透到文字、语音与影像的识别与处理、生物、医药与医疗、娱乐与媒体、精准营销、国防与安保、自动驾驶与无人飞行器等多个方面。到 2020 年，预计深度学习应用市场将达到 400 亿美元。超级计算同时也已经与大数据结合，成为大数据相关产业的技术基础，在此之上可以进行城市规划，实现相关惠民服务，包括政务、交通、社保、医疗、教育、就业、城市、帮扶、电商等惠民服务。此外，还有医疗大数据、空天大数据、气象大数据、环保大数据、金融大数据等一系列朝阳性应用。

中国政府从“九五计划”开始就一直支持高性能计算的技术、产业、应用的发展。国家“863 计划”推出了一系列高效能计算机系统，2008 年的深腾 7000 的计算性能是每秒 150 万亿次，曙光 5000A 是每秒 230 万亿次；2010 年推出了曙光每秒 6000 万亿次和每秒 3000 万亿次，天河一号是每秒 4700 万亿次。2011 年，我们用国产的处理器推出了每秒千万亿次系统神威蓝光，这是一个里程碑式的成果，解决了国内用自主研发的处理器实现千万亿次系统的突破。2013 年 6 月，世界超级计算机 500 强中，天河二号名列第一，其峰值速度达到了每秒 5 亿亿次。“863 计划”也启动了第二台 10 亿亿次的计算机研究，由神威蓝光团队研制的基于自主芯片的太湖之光超级计算机，当前位列 Top500 排名第一。此外，截至 2016 年，中国科技部批准建立的国家超级计算中心共有 7 家，分别是国家超级计算合肥中心、国家超级

计算天津中心、国家超级计算广州中心、国家超级计算深圳中心、国家超级计算长沙中心、国家超级计算济南中心和国家超级计算无锡中心。可以看到，在系统研制和环境建设方面，我们已经走到了世界领先的地位。

近年来，E级计算成为高性能计算一个新的发展目标。2013年，以Prace为首的欧洲超算联盟又启动2020地平线计划及基于ARM架构的E级计算原型系统的“稻草人”计划。2015年，美国白宫提出了国家战略计算计划(National Strategic Computing Initiative)，用以最大化超级计算的研究、开发、部署能给美国社会所能带来的福利。

2016年以来，中国政府加大了对超算的支持的力度，发展E级高性能计算机及其相关技术。2016年10月9日，习近平总书记在中共中央政治局第三十六次集体学习时强调“要紧紧牵住核心技术自主创新这个‘牛鼻子’，……，推动高性能计算(超算)、移动通信、量子通信、核心芯片、操作系统等研发和应用取得重大突破”。2016年12月，国务院印发《“十三五”国家信息化规划》指出：“十三五”时期要大力发展先进计算技术，重点加强E级计算(超级计算全球最前沿技术，每秒运算性能达到百亿亿次)、云计算、量子计算、人本计算、异构计算、智能计算、机器学习等技术研发及应用。科技部已经按照“十三五”的规划要求，启动“高性能计算(超算)”重点专项2016年度项目和2017年度项目。其中，2016年专项围绕E级高性能计算机系统研制、高性能计算应用软件研发、高性能计算环境研发等三个创新链(技术方向)部署了20个重点研究任务；2017年专项则围绕E级计算机的编程模型、算法、示范应用及特定行业应用软件研制展开。

尽管如此，高性能计算的应用却往往躲在屏幕之后，不为大众所熟知。此外，由于高性能计算的产业链长而复杂，很多IT研究人员对应用了解的并不多，而且不同应用方向的研究人员之间也是隔行如隔山。其次，中国的超级计算机如星云、天河、太湖之光都在世界排名中名列前茅，不少国外的专家质疑我们只是用钱堆出个机器，并不是真正地把高性能计算机用起来。再次，E级计算机原型系统以及将来的E级计算机即将部署，我们需要对以往的高性能计算应用进行归纳，进而为运算速度更快的新一代超级计算机的高效运行与利用打好基础。正是因为这些考量，笔者邀请了业内相关专家学者一起编著了本书。

本书由中科曙光公司牵头，参与单位近二十家，较为全面地反映了中国高性能计算应用的现状。同时我们邀请了中国超算创新联盟对未来的高性能计算技术和应用做了展望，相信可以供高性能计算技术的研究人员、应用专家、相关政策的制订者，以及该技术的爱好者参考使用。

本书得到了国家重点研发计划高性能计算重点专项2016YFB0200300和2016YFB0200100的资助，特此表示感谢。

对于本书的编写工作，各位作者付出了极大的心血和努力，将自己多年积累的高性能计算相关知识和经验予以整理共同完成了此书。然而，编写时间仓促，精力有限，书中难免会有所疏漏，敬请读者批评指正。

历　军

2018年3月

中关村软件园

目录

第1章

总 述

孙凝晖[1]，谭光明[1]，吉 青[2]
1. 中国科学院计算技术研究所 2. 中科曙光

1.1 高性能计算概述

1.1.1 系统架构

计算机的起源可以追溯到欧洲文艺复兴时期。16—17世纪的思想解放和社会大变革，大大促进了自然科学技术的发展，其中，制造一台能帮助人类进行计算的机器，就是最耀眼的思想火花之一。1614年，苏格兰人John Napier发表了关于可以计算四则运算和方根运算的精巧装置的论文。1642年，法国数学家Pascal发明能进行8位计算的计算尺。1848年，英国数学家George Boole创立了二进制代数学。1880年，美国普查人口用了7年的时间进行统计，而1890年，Herman Hollerith用穿孔卡片存储数据，并设计了机器，仅用了6周时间就得出了准确的数据(62 622 250人)。1896年，Herman Hollerith创办了IBM公司的前身。这些“计算机”，都是基于机械运行方式，还没有计算机的灵魂——逻辑运算。而在这之后，随着电子技术的飞速发展，计算机开始了质的转变。

1943—1959年的计算机通常被称作第一代计算机，使用真空电子管，所有程序都是用机器码编写，使用穿孔卡片。1946年，John W. Mauchly和J. Presper Eckert负责研制的ENIAC (Electronic Numerical Integrator and Computer)是第一台真正意义上的数字电子计算机。重30吨，18 000个电子管，功率25kW，主要用于弹道计算和氢弹研制。1947年，Bell实验室的William B. Shockley、John Bardeen和Walter H. Brattain发明了晶体管，电子计算机才找到了腾飞的起点，开辟了电子时代新纪元。1949年，《科学》杂志就大胆预测“未来的计算机不会超过1.5吨”。真空管时代的计算机尽管已经步入了现代计算机的范畴，但其体积之大、能耗之高、故障之多、价格之贵大大制约了它的普及。

1959—1964年的计算机一般被称为第二代计算机，大量采用了晶体管和印刷电路板，计算机体积不断缩小，功能不断增强，可以运行FORTRAN和COBOL，接收英文字符命

令，并出现大量应用软件。尽管晶体管的采用大大缩小了计算机的体积，降低了其价格，减少了故障，但离人们的要求仍很远，而且各行业对计算机也产生了较大的需求，生产更强、更轻便、更便宜的机器成了当务之急，而集成电路的发明，不仅使计算机体积得以减小，更使其速度加快，故障减少。1958 年，在 Robert Noyce(Intel 公司的创始人)的领导下，继集成电路后，又推出了微处理器。

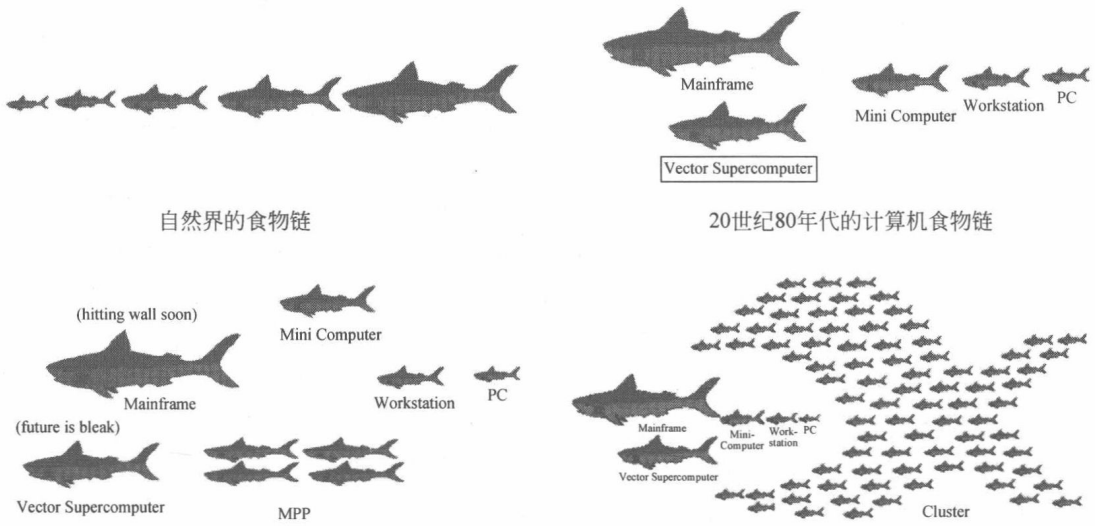
1964—1972 年的计算机一般被称为第三代计算机，大量使用集成电路，典型的机型是 IBM360 系列。

1972 年以后的计算机习惯上被称为第四代计算机，基于大规模集成电路，及后来的超大规模集成电路，计算机功能更强，体积更小。在这之前，计算机技术主要集中在大型计算机和小型计算机领域发展，但随着超大规模集成电路和微处理器技术的进步，计算机进入寻常百姓家的技术障碍已被突破。特别是从 Intel 发布其面向个人计算机的微处理器 8080 开始，互联网技术、多媒体技术也得到了空前的发展，计算机真正开始改变了人们的生活。

每一个计算时代都从体系结构的发展开始，接着是系统软件(特别是编译器与操作系统)、应用软件，最后随着问题求解环境的发展而达到顶峰。1976 年，Cray-1——第一台商用高性能计算机(又称超级计算机)问世，集成了 20 万个晶体管，每秒进行 1.5 亿次浮点运算，从此掀开了高性能计算机的篇章。

高性能计算中最为重要的核心技术是并行计算(Parallel Computing)，它也是相对串行计算而言的。并行计算是指同时使用多种计算资源解决计算问题的过程，是提高计算机系统计算速度和处理能力的一种有效手段。它的基本思想是用多个处理器来协同求解同一问题，也可理解为将被求解的问题分解成若干个部分，各部分均由一个独立的处理机来并行计算。并行计算系统既可以是专门设计的、含有多个处理器的超级计算机，也可以是以某种方式互连的若干台独立计算机构成的集群。通过并行计算集群完成数据的处理，再将处理的结果返回给用户。创建和使用并行计算机的主要原因是并行计算机是解决单处理器速度瓶颈的最好方法之一。

如图 1-1 所示为不同时期的计算机食物链。



自然界的食物链　　20世纪80年代的计算机食物链

20世纪90年代的计算机食物链　　当前的计算机食物链

图 1-1　计算机的食物链

开展并行计算必须具备并行机、应用问题具有并行度、并行编程三个基本条件。实现并行计算的计算机系统结构主要分为以下几个。

1. 集群系统

计算机集群将一组松散集成的计算机软件或硬件连接起来高度紧密地协作完成计算工作。在某种意义上,它们可以被看作是一台计算机。集群系统中的单个计算机通常称为结点,通常通过局域网连接,但也有其他的可能连接方式。集群计算机通常用来改进单个计算机的计算速度和/或可靠性。一般情况下,集群计算机比单个计算机比如工作站或超级计算机性价比要高得多。

2. 对称多处理器和分布式共享内存系统

在均匀存储器访问(UMA)系统中,一个共享存储器可以为所有处理器通过一个互联网络进行访问,就如同一个单处理器访问它的存储器一样。所有处理器对任何存储单元有相同的访问时间。用于 UMA 中的互联网络可以是单总线、多总线或者是交叉开关。因为对共享存储器的访问是平衡的,故这类系统称为 SMP(Symmetrical Multi-Processing,对称多处理器)系统。每个处理器有相等的机会读/写存储器,也有相同的访问速度。

分布式共享内存(Distributed Shared Memory,DSM)可以通过硬件或软件来实现。分布式共享内存主要使用在集群计算机中,集群计算机中的每一个网络结点都有非共享的内存空间与共享的内存空间。该共享内存的位置空间在所有结点是一致的。简单地说,同一时间下在结点 A 读取 0x00001234 会和结点 B 读取 0x00001234 得到一样的值。

3. 大规模并行处理系统

向量计算机是相对于标量计算机系统而言的,各种 CISC 或 RISC 类型的计算机均是标量计算机。向量计算机的发展历史比较长,与其他架构的系统不同,Vector CPU 硬件和指令集针对科学计算进行过专门的优化设计,比如矩阵运算,标量计算机一条指令通常仅能够完成一个矩阵元素的运算,而向量系统则能够实现一条向量指令完成一批元素的运算。因此,在目前的所有架构中,Vector 系统的科学计算效率是最高的。但是,随着计算机技术的发展,特别是 Cluster 系统的流行,现在 Vector 系统的发展空间变得非常狭窄。

大规模并行处理机(Massively Paraller Processing,MPP)系统的研究工作于 20 世纪 60 年代就已经开始,但近十年才成为工业产品。MPP 系统主要的应用领域是气象、流体动力学、人类学和生物学、核物理、环境科学、半导体和超导体研究、视觉科学、认识学、物理探测等极大运算量的领域。RISC 处理器和处理器间高效互连技术的发展使得 MPP 系统在很多领域取得了比传统的向量巨型计算机好得多的性能价格比,并比向量计算机有高得多的发展潜力,开始成为巨型计算机的主要品种。1995 年,MPP 系统已经出现峰值达 355GFLOPS 的品种。到 1996 年年底已经出现每秒运算 1 万亿次(TFLOPS)的品种,2000 年则达到了 10～30TFLOPS 的高水平,当前已经出现了超过 100PFLOPS 的顶级高性能计算机。

4. 星群系统

星群系统(Constellation)包括刀片服务器集群、InfiniBand 交换机、StorageTek 存储硬件和多核处理器。操作系统可以是开源 Solaris 和 Linux。2007 年,Sun 公司曾经成功地把星群系统应用在得克萨斯大学超级计算中心,并主打互联网市场。

5. 混合集群系统

采用 nVIDIA GPU、Intel MIC/FPGA 和 AMD/ATI 作为加速卡，来加速高性能计算，近十年来已经成为 HPC 业界的热点。无论是曙光当年冲击全球高性能计算机排行榜 TOP2 的千万亿次超级计算机“星云 CPU＋GPU”，还是广州超算中心的“天河 2 CPU＋MIC”(6 次 Top 500 第一)，都使用了混合集群系统方案。这种方案的突破来源于 2010 年天津超算中心的“天河 1 CPU＋GPU”方案一举拿下当年 Top500 榜首。

1.1.2 硬件基础

高性能计算机是一种超大型的计算机。人们首先看到的是它的硬件部分。以下通过对中国科学院超算中心新升级的高性能计算机“元”(如图 1-2 所示)，来介绍高性能计算机的硬件基本组成。

图 1-2 中国科学院超算中心最新升级的高性能计算机“元”

“元”超级计算系统安装在中国科学院计算机网络信息中心怀柔分中心，是中国科学院超级计算环境总中心的新一代超级计算系统，用于替换已运行近六年的深腾 7000 超级计算系统。“元”超级计算系统采用混合架构，支持异构计算，总计算能力约 2.3PFLOPS，其中 CPU 通用计算能力 700TFLOPS，采用 Intel MIC 及 nVIDIA GPU 的计算能力共 1.6PFLOPS。系统内存总量约 140TB，存储总裸容量超过 6PB。“元”超级计算系统由曙光公司研制，分两期建设，一期已于 2014 年建设完成并投入使用。第一期建设总体计算能力 303.4TFLOPS，其中，CPU 计算能力 152.32TFLOPS，MIC 和 GPU 共 151.08TFLOPS，具体配置如下。

(1) 刀片计算结点：共有 270 台曙光 CB60-G16，双路刀片，CPU 整体性能达到 120.96TFLOPS。每台刀片计算结点配置两颗 Intel E5-2680 V2(Ivy Bridge|10C|2.8GHz) 处理器，64GB DDR3 ECC 1866MHz 内存。其中，CPU 13.44TFLOPS，GPU 70.2TFLOPS。每台 GPGPU 计算结点配置两块 nVIDIA Tesla K20 GPGPU 卡，两颗 Intel E5-2680 V2 (Ivy Bridge | 10C | 2.8GHz) 处理器，64GB DDR3 ECC 1866MHz 内存。支持 CUDA、OpenACC、OpenCL，支持 GPU Direct。

(2) MIC 计算结点：共有 40 台曙光 I620-G15，总性能达到 98.8 TFLOPS，其中，CPU

17.92TFLOPS,MIC 80.88TFLOPS。每台 MIC 计算结点配置两块 Intel Xeon Phi 5110P (8GB 内存)卡,两颗 Intel E5-2680 V2(Ivy Bridge|10C|2.8GHz)处理器,64GB DDR3 ECC 1866MHz 内存。支持对 Xeon Phi 的 Offload 卸载、Symmetric、Native 原生模式调用。

(3) 大内存结点:一台 NUMA 结构的 SGI UV2000 结点,为有大内存需求的用户提供计算服务。系统共有 32 颗 Intel Xeon E5-4620 八核处理器,主频 2.60GHz,系统共享内存 4TB,采用 NUMALink 6 (NL6; 6.7GB/s bidirectional)连接,单一系统映像,系统峰值约 5TFLOPS。

(4) 存储系统:采用两套文件系统,使用 Stornext 并行文件系统做用户 HOME 和公共软件区存储,可靠性高,用户可用容量 165TB; 采用曙光 ParaStor200 做高性能工作区存储系统,I/O 带宽高,用户可用容量为 1.3PB。

① 用户 HOME 目录:/home 采用 SNFS 文件系统。

② 公共软件安装目录:/soft 采用 SNFS 文件系统。

③ 用户工作目录:/work1 采用曙光 ParaStor200 并行存储系统。

(5) 系统配置 1 套采用 56Gb/s FDR InfiniBand,全线速互连。

(6) 系统配置 4 台登录结点,通过均衡负载实现单 IP 登录。

由这个例子可以看出,典型的 HPC 集群系统主要由 5 类计算设备和 3 类网络组成。5 类设备主要指管理结点、计算结点(包括普通计算结点如刀片,及加速计算结点如 GPU 和 MIC)、存储结点、交换设备和 I/O 结点。3 类网络是指:管理网络、计算网络和存储网络。其中,管理网络用于管理结点,并管理各计算结点与 IO 结点的互连。管理网络所连接的机器就是集群内部的本地机器,所以高带宽和低延迟的要求并不高,同时可以容忍一定的过预订率。计算网络用于各计算结点间的互连,是并行任务执行时的进程间通信的专用网络,并行计算的核心就是它同集群内的其他结点交换信息的能力,通常称之为 IPC(Inter-Process Communications,进程间通信),它需要高性能的网络来进行快速交换,因此要求延迟小,带宽大。系统的互连结构和带宽非常重要,它决定了系统架构、性能以及可适应应用等。存储网络需要向 HPC 集群的结点提供数据访问服务。在 HPC 领域,根据数据存储和访问方式的不同,有几种不同的存储形式。在最低级别,可以有两种方法访问数据,一是数据由外部文件系统提供文件级别的访问,包括 NAS; 二是数据块级别的访问,包括 DAS 或 SAN, SAN 可以分别使用基于 SCSI 或 SCSI RDMA (SRP)协议的光纤通道或 IB 存储。此外,整个系统所需要的机房、配电、制冷等也是不可缺少的内容。

1.1.3 并行算法

近二十年来,以并行计算技术、并行算法和并行计算机结果为核心的并行化技术受到了国际国内计算数学界、计算机科学界乃至整个工程技术与科学界的广泛重视。早在 1989 年 3 月美国国防部提出的一份旨在保持其国际技术领先地位的报告中明确地将“并行处理”列为 22 项重大项目的第 3 项。日本政府则将并行技术与软件工程和人工智能并列为重点发展的三大技术。

并行算法是适合在并行计算机上实现的算法。一个“好的”并行算法应该能够充分发挥并行计算机多处理机的计算能力。Kung 于 1980 年在《并行算法结构》一文中将并行算法定义为“多个并发进程的集合,这些进程同时并相互协作地进行运行处理,从而达到对给定