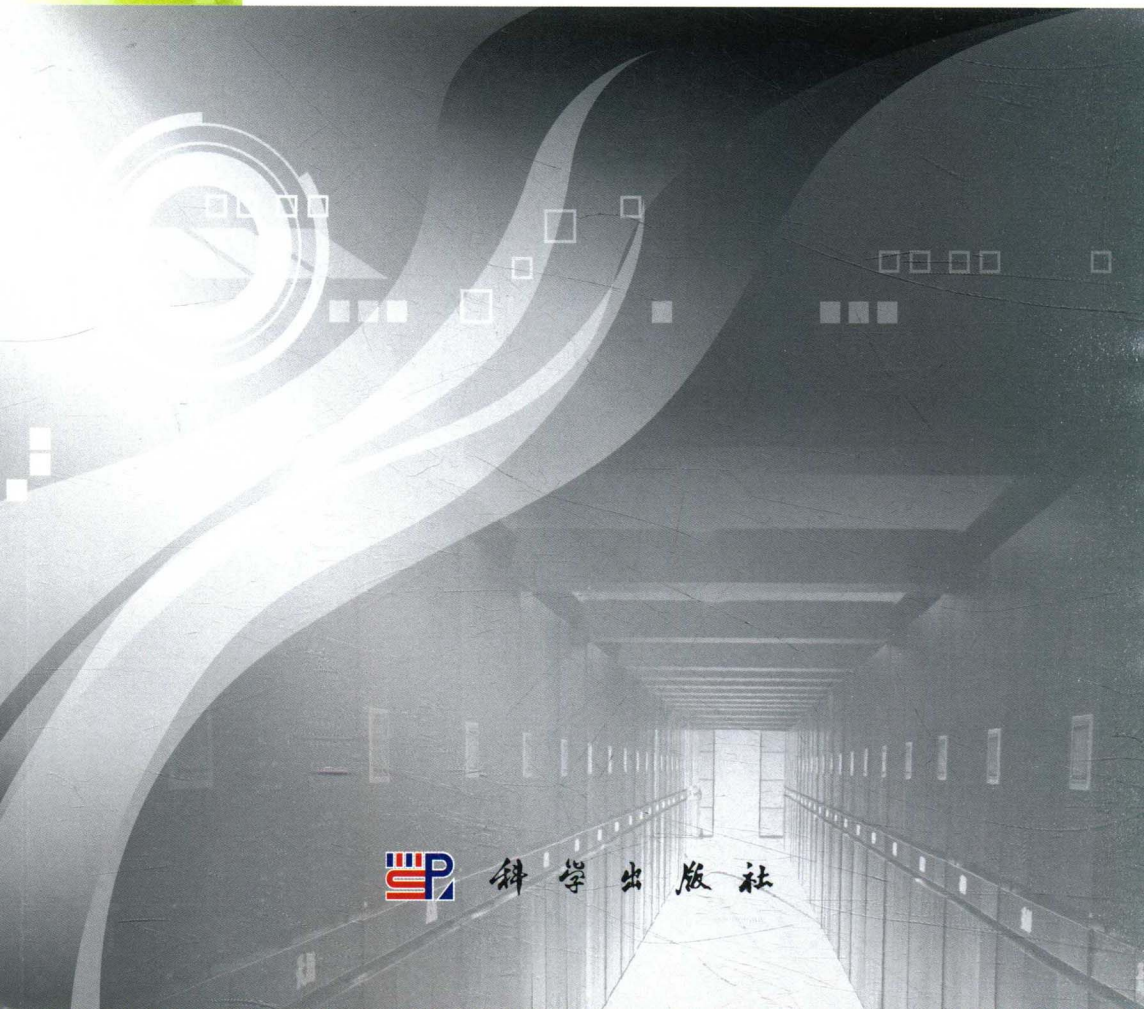
面向大规模应用的高性能计算编程与优化

文梅　柴俊　苏华友　董辛楠　张春元　著



科学出版社

面向大规模应用的高性能计算编程与优化

文 梅 柴 俊 苏华友 董辛楠 张春元 著

科学出版社

北 京

内 容 简 介

随着信息技术的不断发展，应用对计算的需求不断增加，需要借助高性能计算系统来解决相关领域的问题。如何高效地利用高性能计算资源解决工程和科学问题成为急需解决的问题。本书源自于作者基于天河系列超级计算机进行大规模应用开发的经验和研究成果，对高性能计算相关的基础知识和优化关键技术进行系统的介绍。

本书共 9 章，第 1 章绪论，主要介绍大规模应用对计算的需求，阐述编程方面的挑战；第 2～5 章介绍高性能计算的基础知识，重点介绍 GPU 和 MIC 编程及优化技术；第 6～8 章阐述作者基于天河-1A、天河 2 号超级计算机开发的三个典型应用案例，重点介绍大规模计算集群的优化技术；第 9 章介绍未来的高性能计算，E 级计算的挑战以及一些新兴应用，并讨论未来高性能计算可能的发展方向。

本书主要面向专门从事高性能计算的程序员和工程师以及使用大规模异构集群系统进行科学计算的科研人员，也可作为相关专业本科生和研究生的参考书。

图书在版编目 (CIP) 数据

面向大规模应用的高性能计算编程与优化 / 文梅等著. —北京：科学出版社，2015.11

ISBN 978-7-03-046259-6

Ⅰ. ①面… Ⅱ. ①文… Ⅲ. ①程序设计 Ⅳ. ①TP311

中国版本图书馆 CIP 数据核字 (2015) 第 265595 号

责任编辑：陈 静 / 责任校对：陈玉凤

责任印制：徐晓晨 / 封面设计：迷底书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京九州迅驰传媒文化有限公司印刷

科学出版社发行 各地新华书店经销

*

2015 年 11 月第 一 版 开本：720×1 000 B5

2015 年 12 月第二次印刷 印张：12 1/2 插页：2

字数：237 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

作 者 简 介

文梅，女，国防科学技术大学计算机学院研究员，硕士生导师。长期从事超高性能加速器体系结构、并行计算、媒体处理等研究。2011 年在挪威 Simula 实验室担任客座科学家。近年来，主持及参与国家重大项目 10 余项，其中包括世界上第一款 64 位流处理器 FT64 的研制、流处理系列国家自然科学基金项目（重点、面上、青年项目）、中挪合作项目等。目前研究兴趣包括深度学习加速器以及相关图像处理。在国际会议和期刊上以第一作者/通信作者身份发表论文 20 余篇，总计发表论文 100 余篇，其中 SCI 8 篇，EI 17 篇。完成学术专著 3 部。2008 年获湖南省优秀博士论文，2009 年获全国优秀博士学位论文提名。

柴俊，男，工程师，2014 年获得国防科学技术大学计算机学院博士学位，研究方向为并行编程、高性能科学计算、计算机系统结构。以第一作者发表论文 5 篇，其中 SCI 3 篇，EI 2 篇。

苏华友，男，助理研究员，2014 年获得国防科学技术大学计算机学院博士学位。研究方向为 GPGPU 并行计算、媒体处理等。以第一作者/通信作者身份发表论文 12 篇，其中 SCI 4 篇，EI 7 篇。

董辛楠，女，助理工程师，2014 年获得国防科学技术大学计算机学院硕士学位。研究方向为高性能计算，以第一作者发表 SCI 1 篇。

张春元，男，国防科学技术大学计算机学院教授，博士生导师，IEEE 会员，享受国务院政府特殊津贴。长期从事计算机体系结构、并行计算等领域研究和教学工作。研究领域主要涉及新型计算机系统结构技术、高性能计算、嵌入式系统及应用技术、并行与分布处理技术、Web 应用技术等。作为项目负责人和主要研究人员主持或参加的各类项目（包括国家自然科学基金、国家 863 高技术研究项目、国家 973 安全重大基础研究项目、国家重点型号项目和对外合作等）共计 20 多项。发表高水平科研论文 50 余篇，出版学术专著 3 部。

序

我所在的国防科学技术大学计算机学院研制的天河系列超级计算机多次位列世界高性能计算机排名 TOP500 的第一名，这是我们学院、中国计算机界乃至全中国人的骄傲。作为天河团队的一员，与有荣焉。2011 年我在挪威 Simula 国家实验室做为期 1 年的访问学者，因为天河，结识了 Oslo 大学的蔡行教授及其研究团队，他问起能不能在天河上合作做一些事情。作为资深的数值计算专家，又是海外华人，常年与高性能集群打交道，他对天河出现的兴奋在某种程度上尤甚于我。加上我所属研究团队的张春元教授致力于开启民智，主持我院学生超算事宜，我便一头扎入高性能计算的研究中，这一扎就是 4 年。

严格地说，我们以前的研究虽然可以列入高性能计算的范畴，但一直处于单芯片微处理器体系结构，再准确点是加速器体系结构，对大规模并行编程也是初次涉猎。只是刚好，天河系列超级计算机都是加速器增强型集群，而且天天在院里耳濡目染，因此有一个相对较好的启动基础。几年来的研究使我清楚地认识到，高性能计算是一个系统工程，要把成百上千个节点集合起来做一件事，尤其是真实应用，挑战是巨大的。分层次讲，任务涉及计算机硬件、系统软件，以及上层的应用。其中，硬件包括芯片、主板、网络；系统软件包括操作系统、作业系统、通信支持，以及各种系统软件库/应用开发库；应用方面包括领域专家提供真实的应用，数值计算专家对应用的数值方法实现，并行专家在庞大集群上的并行实行，而对于集群系统还可能涉及配套的外围环境，如冷却系统、机房保障等，每一个环节都必须紧密耦合，一旦配合不好就可能导致应用无法完成，或者扩展性不高，或者需要时间太长等。高性能计算一直都是国家综合实力的象征，因为无论从人力成本还是硬件成本来说，它的开销都非常巨大。同时它也是一个需要长期积累的，涉及多个科学领域。从另一个角度来说，军事、石油勘探、医学成像、气候模拟、自然科学研究等应用领域都离不开高性能计算。例如，石油工业广泛使用软件来模拟在油井、输油管线和油气处理设备中的油、气及水的运动状态。

在高性能计算领域，大家熟知的 TOP500 相当于“硬实力”的排名，已经得到国内外广泛的关注和认可。而另一个“软实力”的竞赛同样重要，即 Gordon Bell 奖，它是由美国计算机协会（Association for Computing Machinery，ACM）组织颁发的高性能计算应用领域的学术奖项，旨在奖励将高性能计算用于解决实际科学问题的杰出成就。纵观 Gordon Bell 奖从 1987 年设立以来的历史，美国的实力有目共睹，日本也拿过 7 次，而我国则是空白，这在一定程度上也体现了我们的真实应用水平与国际先进水平存在很大的差距。MASA 团队在该书中提到的真实算例使用的最大节点规模 1000 节点、

4000 节点，在当时分别都是天河-1A、天河 2 号应用的高水平，只大约相当于全部节点规模的 1/8、1/4，最高达到 1.2Pflops（10^{15} 次/s 双精度浮点运算）的性能，而 2013 年 Gordon Bell 奖的入围候选者（finalist）的门槛是 7.17Pflops。2014 年的 Gordon Bell 奖的入围候选者中最高的应用性能已经达到了 24.77Pflops（机器 Piz Daint：18600 节点，18600 GPU）。因此，尽管我们的高性能计算的硬实力有目共睹，但是软实力还需要科学家们继续努力。相信不久的将来，我国学者最终能够摘得这个大奖。

虽然高性能计算一直是“高大上”的学科，通常用于科学计算及高端工业计算（本质上这二者是一回事），但有一个神秘出口与我们每个人都相关，这就是“技术下移”。计算机芯片每 18 个月可多集成一倍的晶体管，以前的巨型机还比不上现在的台式机性能，因此，高性能计算积累的技术（如大规模、多节点并行计算），现在已经变为普通人接触到的片上多核计算技术。

随着互联网、大数据时代的到来，传统的高性能计算有机会焕发新生。例如，2014 年下半年 Google 开源的分布式集群管理系统 Kubernetes（用 Go 语言重写的 Borg），提出的思想就是“Manage a cluster of Linux containers as a single system to accelerate Dev and simplify Ops”。就是说，像一个单系统那样管理众多进程群（实际运行在可能异构的集群上）以便加速设备和简化操作。简而言之，用户可以像面对 1 台机器那样工作，这在传统高性能计算看来几乎是不可想象的。因为大数据任务天生并行，例如，搜索引擎，每个人的搜索任务是不相关的，所以实际上数据中心的集群计算相当于是利用多台机器完成多个任务，并行难度小，任务之间没有通信。数据中心的主要目标不是在最短的时间完成任务，而是如何处理更多的数据（包括具有极高的容错能力，面向巨量并且持续的商用需求），这也是 Google 可以采用廉价集群（单节点计算能力低下，低速互联网络）的原因。而传统高性能计算任务难以做到这样的任务并行度。

尽管从源头上，大数据与高性能计算有差别，但是随着大数据任务的发展，也逐渐出现了一些交叉。例如，近年来非常热的深度学习网络的训练就需要同时考虑性能需求，因为大量的样本图片和视频训练的时间是非常可观的，一次训练通常以周甚至月计算，而为了获得一个好的训练网络，需要不断调整，反复尝试。因此 Hinton 等纷纷采用高性能计算节点来进行训练。但同时该任务又有传统高性能计算所不具备的大量数据的特征。从这个角度上来说，我们认为未来在并行计算、分布式计算（通常包括云计算、大数据计算）等领域，都将看到高性能计算产生的持续影响。

本书的目的在于将我们近几年的研究成果总结出来，抛砖引玉，希望给正在或将要投身到高性能计算编程方面的同行提供一点经验和参考。由于作者的水平有限，书中不足之处在所难免，恳请读者批评指正。本想写得更为科普，可以适应更多的读者，受限于时间和水平，难免有些学究古板，希望读者也可多多包涵。

文　梅

2015 年 6 月

前　言

本书基于高性能计算集群，特别是 GPU/MIC 众核加速器的异构系统，结合我们近年来的课题研究，选取了 3 个具有一定代表性的真实大规模科学与工程的应用，介绍了大规模并行应用软件设计和实现技术，以及性能优化策略，为开发设计高效的异构混合计算应用提供参考，并为进一步拓展大规模 GPU/MIC 加速器异构系统的应用领域奠定基础。

全书共分 9 章，内容安排如下。

第 1 章绪论，主要介绍大规模应用的计算需求，高性能计算硬件基础知识，重点介绍加速器增强型异构计算系统，阐述编程方面的挑战及相关研究现状。

第 2 章高性能计算并行基础，介绍相关基础知识，包括并行的分类、并行计算的度量指标以及一些典型的基准测试集。

第 3 章并行程序设计，介绍以通用处理器 CPU 为主的单节点及多节点编程模型，包括 OpenMP、MPI 及混合编程模型 MPI+OpenMP 等，并给出实例，同时阐述大规模集群系统通用的关键优化——节点间通信优化问题。

第 4 章 GPU 并行计算，介绍 GPU 体系结构、CUDA 编程模型、单芯片性能优化方法并给出实例。同时介绍单节点多 GPU 以及大规模 CPU-GPU 异构计算的混合编程模型。

第 5 章 MIC 并行计算，介绍 MIC 体系结构、单芯片编程模型、性能优化策略，并给出实例。同时介绍单节点多 MIC、大规模 CPU-MIC 异构计算的混合编程模型。

第 6～8 章分别基于天河-1A、天河 2 号超级计算机，介绍以下三个应用的大规模并行编程实例。

（1）贝叶斯分析（Bayesian analysis）构建物种进化树。贝叶斯分析根据各物种的基因序列构建出物种的进化树，是生物信息学领域的典型方法之一。进化树被广泛地用于医药和生物学研究，对科学和社会具有重大的价值和意义。贝叶斯分析是一种根据排列好的分子序列数据和形态数据矩阵来推测进化树的标准方法。然而，使用贝叶斯分析方法来推测大型进化树时，会导致对计算能力的巨大需求。例如，在个人桌面计算机上对于数百物种，数千字符长的序列来构建一个可靠的进化树，可能需要数天甚至数月的时间。

（2）盆地演化模拟。计算机模拟地层学（stratigraphy）的一个有趣的主题是调研超过若干地质年代的海底盆地的地质沉降（sediment deposition）。与许多计算科学分支类似，该数学模型为一个非线性偏微分方程系统组。因为要分辨出足够多的物理和数值上的细节，这些耦合的方程通常要被离散化，并且在并行计算机上进行数值求解。

（3）纳米精度的亚细胞级（subcellular-level）。心脏钙离子动力学模拟通过对心肌细胞中的 Ca^{2+} 释放/钙波生成扩散的数学建模，使得心电模拟能够在亚细胞级水平上直接利用计算机技术对一些复杂的心脏活动假说进行验证与预测。与第二个应用在细胞级模拟类似，也同样用于指导实验研究，模拟各种心脏疾病，对研究心脏电生理学、心律不齐、药物作用等，以及心脏疾病的预防和治疗具有非常重要的作用。下降到 1nm 精度的亚细胞级钙动力数值模拟可以作为一个重要的工具，用于探索很多心脏疾病的生理原因。模拟一个肌纤维活动 1ms 的时间，可能需要总的浮点操作数在 10^{19} 的量级。对巨大计算能力的需求，使得亚细胞钙动力学在纳米精度的模拟极具挑战性。

第 9 章未来的高性能计算，介绍 E 级计算的挑战以及一些新兴应用，并讨论未来高性能计算可能的发展方向。

本书由张春元教授和文梅研究员策划和统筹，由 MASA 课题组部分成员合作完成，第 1、2、9 章由文梅研究员撰写，第 3 章由苏华友博士和董辛楠硕士撰写，第 4 章和第 7 章由苏华友博士撰写，第 5 章由柴俊博士和董辛楠硕士撰写，第 6 章和第 8 章由柴俊博士撰写。伍楠博士及博士（硕士）研究生：蓝强、杨静、乔寓然、陈照云、沈俊忠、时洋、方皓、王彦鹏为本书提供了丰富的素材，并参与资料收集与整理工作。本书在写作过程中，参阅了国内外许多论文和著作。

本书中的三大算例除贝叶斯分析构建物种进化树外，均来源于挪威 Oslo 大学 Simula 国家实验室高性能计算系主任蔡行教授及研究团队卫文劼、Johanness Langguth、Johan Hake、Glenn Lines 的研究，在此表示感谢！

感谢国家自然科学基金项目（61033008、61272145），国家教育部博士点基金项目（20104307110002），挪威 SIU（Norwegian Center for International Cooperation in Education）合作项目（NFR-214113）的资助！

感谢伍楠博士为本书研究做出的贡献，他现在离开了天河团队，祝愿他可以开创更好的事业！

感谢国防科学技术大学计算机学院、天河团队，天津超算中心，广州超算中心对我们的研究工作给予的大力支持！

MASA 小组　文梅

2015 年 6 月于长沙

目　录

第1章 绪 论

高性能计算（High Performance Computing，HPC）是指使用高端处理器的高端服务器（处理器可以是多个、多种类型，如 CPU、GPU 等），或者是由多个这样的服务器构成的集群（单个这样的服务器称为一个节点），节点之间以高速互联网络，如天河网络、InfiniBand 或 Myrinet 连接的计算系统来进行计算的统称。传统的高性能计算通常指科学计算，广泛应用于军事、石油勘探、医学成像、气候模拟、自然科学研究等领域。作为解决国家挑战性问题的重要手段，以及解决制约国家经济发展瓶颈问题的重要工具，高性能计算有非常重要的战略意义，高性能计算的水平是国家综合国力的体现。本章从应用需求、硬件平台、编程挑战等几个方面介绍高性能计算的背景，并简单介绍本书中重点关注的加速器增强型异构集群。

1.1 大规模应用对高性能计算的迫切需求

大规模科学与工程计算应用领域对计算能力的需求是推动并行计算机发展的源动力[1]。21 世纪人类面临的一系列挑战性的重要科技问题，如卫星成像数据处理、全球天气预报、核爆炸模拟、石油勘探、地震数据处理、飞行器数值模拟和大型事务处理、基因工程、生物医学模拟等，数据规模高达 TB（10^{12}B）或者 PB（10^{15}B）量级①，每秒需要执行万亿次、百万亿次乃至千万亿次浮点运算，高性能计算已经成为当前科学研究不可或缺的重要手段。

下面以生物计算里的心电计算模拟为例来说明。近年来，心脏病已经成为人类三大疾病之一，我国心脏性猝死发生率上升很快，这一现象导致我国心脏性猝死研究工作任务艰巨，需求迫切。绝大多数猝死事件发生在医院外，一旦发生，存活比例甚低，据西方国家报道，院外猝死抢救存活率仅为 2%～15%。心脏性猝死作为人类疾病的主要死亡方式之一，迄今仍是威胁人类的重大健康问题。估计全球每年有 350 万例心脏性猝死发生，美国为 40 万～45 万例，德国为 8 万～10 万例，文献[2]调查首次得出我国心脏性猝死发生率为 41.84 例/10 万人。若以 13 亿人口推算，则我国心脏性猝死总人数高达 54.4 万例/年，位居全球各国之首。在医学领域对心电的研究主要依靠对心电现象的直接观察、对心电规律的总结和动物实验。心脏心电活动只有在有生命的个体上才能真实地表现出来。无论在伦理、观测深度还是实验便捷性上，直接观察都无

① 新兴互联网大数据应用的数据规模更庞大，可以说是无限的。此处主要指传统高性能计算，即科学计算的数据量。

法满足人类心脏心电特性研究需要。因此，在心脏分子和细胞学研究领域，通过对心肌细胞中的离子动作电位进行包含心肌细胞电生理特征的数学建模，可以精确表述真实的心肌细胞收缩，能够在细胞水平上直接利用计算机技术对一些复杂的心脏活动假说进行验证与预测。

然而，心脏电活动的空间和时间的跨度很大，空间跨度可以从细胞膜蛋白分子直径的 1nm 到整个机体 10cm 的尺度，相差达 8 个数量级；时间跨度从布朗运动的 1μs 到人们心脏跳动持续的周期数十分钟，相差 9 个数量级。建立一个多层次的心脏细胞模型，用于精确描述健康或者生理病理的心脏细胞活动，仍是难以捉摸的。用单个细胞的模型构建整个心脏的模型（从左至右是从微观到宏观，亚细胞级、细胞级、组织级到器官级，如图 1.1 所示），这种自下而上的方法不仅建立了将细胞/亚细胞级别的生物物理现象与心脏器官活动模式关联的可能性，还对心电模拟的空间分辨率、时间分辨率的提升和优化打下基础，成为心电建模与仿真发展的一个主要方向。整个心脏由约 10^{10} 个细胞组成，这么多细胞级别甚至亚细胞级别模型模拟的计算量是极其巨大的。更为严重的是，由于心脏电子脉冲传播的时空特征，心脏状态的变化极快，要求极高的时间分辨率，而且心电波阵面是急剧升降的，所以又需要非常精细的空间分辨率。综合来说，这两个因素使得仅模拟一次心脏跳动，就要上万甚至数十万次求解一个庞大的方程系统，同时，因为需要设置大量不同的参数和场景组合，这些模拟通常必须反复运行[3]，所以需要庞大的计算能力来保证系统数学模型的精确性和正确性[2]。

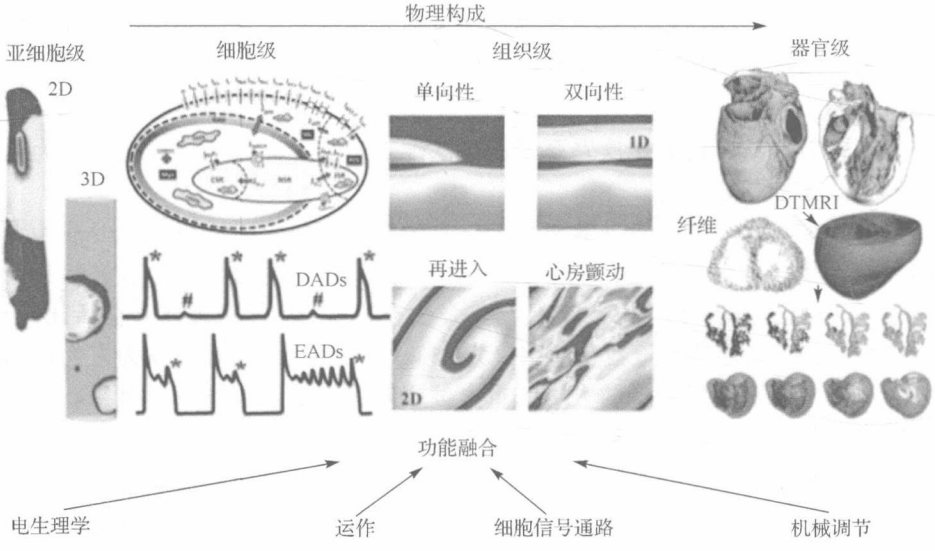


图 1.1　心电模拟自下而上的四个层次（从左至右是从微观到宏观）

当前超级计算机已经进入千万亿次（Peta-flops，10^{15}）浮点计算能力的时代，但诸如高能核物理、材料化学、生命科学等一系列应用表现出对计算能力的超高需求，

计算和模拟的时间空间分辨率，规模，以及实时性的需求永无止境，也就是说应用对计算能力的需求同样永无止境。近年来，美国、欧洲联盟（简称欧盟）、日本、俄罗斯和印度纷纷制定百亿亿次（Exascale，10^{18}）级计算的计划[4]，简称 E 级计算。例如，美国计算机科学中心为 Jaguar/Titan 超级计算机选定了 6 个面向百亿亿级的核心程序：S3D、CAM-SE、DENOVO、LAMMPS、PFLOTRAN 和 WL-LSMS[5]。这些代码的特点是在不同的尺度进行高分辨率模拟或直接数值模拟，但有限元/有限差分和分子动力学模拟这两大类基本数值方法仍将是未来百亿亿级科学与工程计算的核心应用算法[6]。

综上所述，应用领域对计算能力存在着巨大的需求，而当前基于加速器增强型异构高性能计算系统的发展正好为这类计算问题的解决提供了良好的机遇。

1.2 高性能计算硬件基础

高性能计算系统从硬件角度来看，是由集中存放的高性能服务器节点通过高速网络互连组成的。高性能服务器的核心就是高端处理器（高端指性能高）。下面简要介绍主流的两种类型的处理器，通用处理器（CPU）和加速器，同时介绍高速网络和典型的加速器增强型异构计算系统。

1.2.1 多核通用处理器

传统通用处理器的特点是面向广泛的应用领域，如桌面应用、Web 服务、SPEC 等应用设计，多数工作负载以事务处理、不规则标量、非计算密集等特点为主[①]，因此可以独立作为处理器或者作为加速器的主处理器存在。由于通用处理器通常采用全局寄存器文件、Cache、深度流水线等，能效比相对专用处理器、加速器等较低。

三十多年来，大规模和超大规模集成电路和体系结构技术的发展，使得处理器的性能一直保持着指数规律飞速增长。根据摩尔定律（Moore's law）的预测，集成电路的集成度的速度增长为每 18 个月翻一倍，目前单芯片的晶体管数已接近十亿个，工作频率高达数 GHz。VLSI 工艺的发展提供了丰富片上资源，为处理器体系结构的发展提供动力，同时也带来挑战。提高片上单处理器性能一直是传统处理器体系结构的设计方向。然而传统的依靠开发指令级并行（Instruction Level Parallelism，ILP）和提高处理器频率来提高处理器性能的方式导致流水线越来越深，处理器频率越来越高。随着集成电路工艺向深亚微米发展，物理上功耗和散热的增加限制了性能的提高，导致可靠性下降，频率也很难像以前那样快速提高，4GHz 的频率成为处理器厂商难以逾越的关口。工艺、材料和功耗的限制使得摩尔定律中描述的性能翻倍时间加长，性能提升遭遇瓶颈。

多核革命（multi-core revolution）趋势的出现克服了单处理器性能提升的物理限

① 稠密矩阵乘是计算密集型的应用，也是多种类别处理器的基准测试程序，本书中如无特指，矩阵乘指的是稠密矩阵乘。

制，体系结构设计方向转向发展在单芯片上集成大量并行执行的计算单元，即多核/众核处理器（multi-core/many-core processor）[7]。多核/众核处理器在单片上集成了多个处理器核一起并行工作，不仅能开发单核传统的指令级并行，更能开发核间的数据级和任务级并行，可以在更低的频率下提供相比单核处理器高得多的性能。充分利用大量的片上资源，在提高性能的同时又满足了功耗和散热的限制，而且相对简单的处理器核降低了设计难度，提高了设计效率。因此，几乎所有微处理器厂商都转而研发多核/众核处理器。自从 2005 年 Intel 和 AMD 正式推出双核 CPU，此后各大厂商都陆续推出 4 核、8 核、12 核等的通用多核 CPU。以 Intel 芯片为例，目前 Sandy Bridge 和 IvyBridge 都是其高端处理器体系结构。

1.2.2　众核加速器

为了改善通用处理器的低能效问题，新型加速器不断出现。加速器的特点是仅面向特定的某一类或者某几类应用，如媒体处理、图形图像处理等①。这些工作负载的特点突出，如流式应用、计算密集等[8]，因此体系结构可以不使用或者少部分使用 Cache，而采用软件管理的存储，增加向量单元提供定点/浮点计算能力，简化控制逻辑等方法来提高处理器能效[9]。

典型的加速器有流处理器、GPU、Tile64、MIC（其芯片称为 Xeon Phi）等，其中 GPU 和 MIC 是主流的商用加速器体系结构。

2007 年 NVIDIA 公司推出了全新统一计算设备架构（Compute Unified Device Architecture，CUDA）的面向通用计算的 GPU 众核处理器，在单芯片上集成了数百个计算核，以 CPU 协处理器的方式工作，相比通用多核 CPU，其更加强大的浮点计算能力使之成为天然的加速器。目前 NVIDIA 新款的 GPU 已经在单片上集成了 2880 个计算核，双精度浮点性能为 1.43Tflops②。

2012 年 Intel 公司推出了新一代 MIC 架构的众核协处理器 Xeon Phi，单片 50+的计算核，支持 200+的硬件线程，双精度浮点峰值性能超过 1Tflops。

众核加速器在性能功耗比方面比通用处理器更有优势，从根据实测 Linpack 的性能功耗比进行排名的 Green500 全球超级计算机最近几期榜单就可以看出：2012 年 11 月的榜单头名是美国田纳西大学国家计算科学研究院的 Beacon 阵列，48 节点，每节点配有 4 个 Intel Xeon Phi 5110P MIC 加速器，全系统性能功耗比为 2.5Gflops/W；2013 年 6 月的榜单[10]中，前 3 名都是基于 GPU 或者 MIC 的系统；而 2013 年 11 月的榜单上[11]，前 10 名全是基于 GPU 的系统，第 1 名是日本东京工业大学（TITech）的 TSUBAME-KFC 阵列，40 个节点，单节点配有 4 个 NVIDIA K20x GPU 作为加速器，

① 随着 GPU 的应用领域不断扩展，GPGPU（General Purpose GPU）计算逐渐普及，科学计算、人工智能等领域也使用 GPU 来进行加速计算，它们的特征都是计算密集的。

② 本章数据截至 2014 年 5 月。

全系统性能功耗比达到4.5Gflops/W，同时TSUBAME-KFC也获得了衡量大数据计算功耗效率的Green Graph500同期榜单的第1名[12]。

目前，从巨型机的高性能计算到普通PC的桌面计算，多核/众核处理器已经被广泛使用于各个领域，计算技术发展已经全面进入多核/众核时代。特别是在高性能计算领域，GPU与MIC作为众核加速器的典型代表，以较高的性能功耗比，有力地推动了高性能计算的发展。基于通用多核处理器+众核加速器（称为加速器增强型）的异构并行计算，已经成为高性能计算的重要发展方向。

1.2.3 加速器增强型异构系统

图1.2所示结构为一个异构节点，节点间通过InfiniBand[13]、定制网络（TH Express-2）[14]或其他高速网络互连，形成节点内异构，节点间同构的基于加速器的大规模异构计算机阵列系统，称为加速器增强型异构系统，也是全书关注的计算平台。

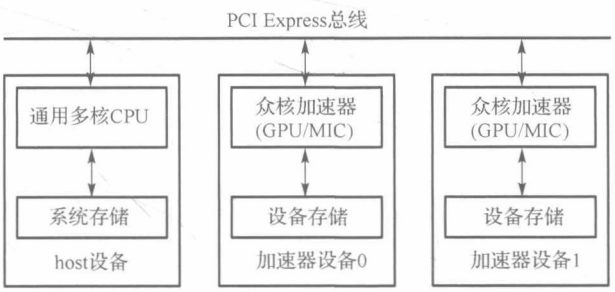


图1.2 单节点异构系统示意图

TOP500组织[15]每半年发布一次全球高性能计算机Linpack[16]计算性能前500排行榜，代表了全世界高性能计算机研制的最高水平，反映了各个国家的高性能计算实力以及高性能计算机的发展趋势。

图1.3是根据最近6年12次的TOP500榜单统计得到的异构系统数量变化趋势。从图中可以发现，基于加速器的异构系统总数量总体上呈逐年增长的趋势。其中基于NVIDIA GPU的系统占有了几乎统治性的比例，特别是在2012年6月达到最高峰的53台。基于Cell的系统起步最早，但已经消失。而基于ATI和AMD公司的GPU系统数量一直较少。从2012年6月出现基于MIC的异构系统后，其呈现逐年快速增长的趋势。因此，可以预见未来的异构系统还是主要以NVIDIA GPU和Intel MIC为主要加速设备，并且后者具有新的活力。从2013年11月的最新榜单看，虽然总计53台异构系统只占到500台机器的10%左右，然而在排名前10的机器中，有4台是异构系统（2台基于NVIDIA GPU、2台基于Intel MIC），并且排名1、2名的都是异构系统。这也说明了基于GPU和MIC的异构系统以其高性能、低功耗的优势已经并将继续引领高性能计算机的发展趋势。正如NVIDIA前首席架构设计师Scott所说：“同时