计算机类专业
系统能力培养
系　列

HZ BOOKS
华章教育

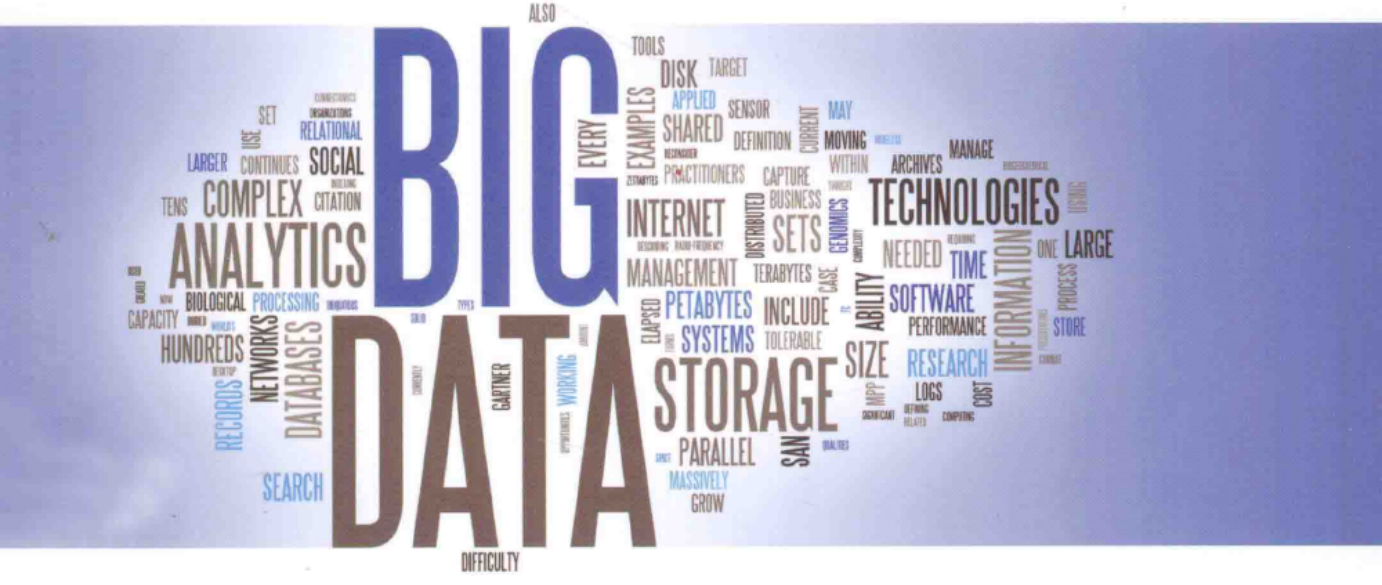
深入理解大数据

大数据处理与编程实践

UNDERSTANDING BIG DATA

BIG DATA PROCESSING AND PROGRAMMING

主编：黄宜华（南京大学）　副主编：苗凯翔（英特尔公司）



- 学术界与业界完美结合的结晶，从原理剖析到系统化算法设计与编程实践
- 多年来系统性教学实践和成果总结，一系列业界产品增强功能深度技术剖析
- 一系列大赛获奖算法、优秀课程设计以及来自科研课题及业界应用的实战案例

机械工业出版社
China Machine Press

深入理解大数据

大数据处理与编程实践

UNDERSTANDING BIG DATA

BIG DATA PROCESSING AND PROGRAMMING

主编：黄宜华（南京大学）　副主编：苗凯翔（英特尔公司）

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

深入理解大数据：大数据处理与编程实践 / 黄宜华主编 . —北京：机械工业出版社，2014.7
（计算机类专业系统能力培养系列教材）

ISBN 978-7-111-47325-1

I. ①深… II. ①黄… III. ①数据管理 – 高等学校 – 教材 IV. ① TP274

中国版本图书馆 CIP 数据核字（2014）第 145263 号

本书从 Hadoop MapReduce 并行计算技术与系统的基本原理剖析着手，在系统介绍基本工作原理、编程模型、编程框架和接口的基础上，着重系统化地介绍 MapReduce 并行算法设计与编程技术，较为全面地介绍了基本 MapReduce 算法设计、高级 MapReduce 编程技术以及一系列较为复杂的机器学习和数据挖掘并行化算法，并引入来自 Intel Hadoop 系统的一系列增强功能以及深度技术剖析；最后，为了提高读者的算法设计与编程实战能力，本书较为详细地介绍了一系列综合性和实战性大数据处理和算法设计问题，这些问题来自课程同学参加的全国性大数据大赛中的获奖算法、课程中的优秀课程设计以及来自本团队的科研课题及业界实际的大数据应用实战案例。书中第 8 章和第 10 章的所有算法均有完整实现代码可供下载学习。

本书是国内第一本基于多年课堂教学实践总结和撰写而成的大数据处理和并行编程技术书籍，因此，本书非常适合高等院校作为 MapReduce 大数据并行处理技术课程教材使用，同时也很适合于高等院校学生作为自学 MapReduce 并行处理技术的参考书。与此同时，由于本书包含了很多来自业界实际产品的深度技术内容、并包括了丰富的算法设计和编程实战案例，因此，本书也很适合作为 IT 和其他应用行业专业技术人员进行大数据处理应用开发和编程实现时的参考手册。

深入理解大数据：大数据处理与编程实践

主　编：黄宜华（南京大学）　副主编：苗凯翔（英特尔公司）

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：姚　蕾　　责任校对：董纪丽
印　　刷：中国电影出版社印刷厂　　版　　次：2014 年 8 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：32.5
书　　号：ISBN 978-7-111-47325-1　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

本书编写组

主　编：黄宜华　　南京大学教授

副主编：苗凯翔　　英特尔中国大数据首席技术官

编　委：**南京大学**

顾　荣　　赵　博　　金　磊
仇红剑　　赵　頔　　沈　仪
韦永壮　　唐　云　　笪　庆
陈　虎　　李相臣　　彭　岳
王　刚　　姬　浩　　张同宝

英特尔公司

姜伟华　　杜竟成　　陈建忠
陈新江　　王星宇　　王　毅
周　珊　　Manoj Shanmugasundaram

北京神州立诚科技有限公司

萧少聪　　韩小姣

南京大学　　英特尔公司

推荐序一

（中国工程院院士、中国计算机学会大数据专家委员会主任　李国杰[⊖]）

数据是与自然资源、人力资源一样重要的战略资源，掌控数据资源的能力是国家数字主权的体现。大数据研究和应用已成为产业升级与新产业崛起的重要推动力量，如果落后就意味着失守战略性新兴产业的制高点。近年来，大数据浪潮席卷全球，引起世界各国的高度关注，美国等发达国家出台了发展大数据的国家计划，全世界著名 IT 企业都在积极推动大数据技术的研发和应用，国内外很多高校和研究机构都在从事大数据技术和数据科学的研究。

学术界已总结了大数据的许多特点，包括体量巨大、速度极快、模态多样、潜在价值大等。对于处理大数据的技术人员，首先面对的困难是过去熟悉的处理系统和软件对付不了大数据，需要学会使用大数据处理和分析平台，进一步的需求是掌握大数据并行处理的算法和程序设计的方法。

Google 公司是大数据处理的先驱，其三大核心技术 MapReduce、GFS 和 BigTable 奠定了大数据分布式处理的基础。MapReduce 是一种分布式运算技术，也是简化的分布式编程模式。在 Google 公司三大核心技术基础上，Apache 社区开发的开源软件 Hadoop 是实现 MapReduce 计算模型的分布式并行编程框架。Hadoop 还提供一个分布式文件系统 (HDFS）及分布式数据库（HBase），将数据部署到各个计算节点上。Hadoop 的独特之处在于它的编程模型简单，用户可以很快地编写和测试分布式系统。2008 年以来，Hadoop 逐渐被互联网企业广泛接纳，这一开源的生态环境已成为大数据处理的主流和事实标准。

一般而言，大数据处理有三种模式：离线计算、在线处理和流计算。Hadoop 是目前使用较广泛的离线计算应用框架，在线处理与流计算尚未形成广泛使用的开源生态环境。大数据处理平台还在不断发展之中，2013 年出现的 Spark 在全面兼容 Hadoop 的基础上，通过更多的利

⊖ 李国杰院士，中国计算机学会大数据专家委员会主任，是我国计算机界的老一辈科学家，在并行处理、计算机体系结构、人工智能、组合优化等方面成果卓著，荣获过多项国家级奖励，领导中科院计算所和曙光公司为发展我国高性能计算机产业、研制龙芯高性能通用 CPU 芯片做出了重要贡献，对国内计算机科技、教育和产业的发展也提出过有影响的政策建议。

用内存处理大幅提高了系统性能。Spark 等新框架的出现并不是取代 Hadoop，而是扩大了大数据技术的生态环境，促使生态环境向良性化和完整化发展。

目前国内外大数据技术人才十分短缺。据麦肯锡公司预计，美国到 2018 年大数据分析技术人才缺口将达 19 万人。中国巨大的人口基数会带来更为巨量的数据，未来几年国内也将需要数十万以上的大数据技术人才。技术市场大规模的人才需求对高校大数据技术人才培养提出了很大的挑战。

作为国内最早从事大数据技术教学与研究的教师之一，南京大学黄宜华教授 2010 年就在 Google 公司资助下开设了“ MapReduce 大规模数据并行处理技术”研究生课程，并组织成立了南京大学 PASA 大数据技术实验室，开展了一系列大数据技术研究工作。在多年课程教学和科研工作基础上，以理论联系实际的方式，结合学术界的教学科研成果与来自业界的系统研发经验，他组织撰写了这本专业技术教材——《深入理解大数据》，着重介绍目前主流的 Hadoop MapReduce 大数据处理与编程技术。

与市场上现有的一些大数据处理和编程书籍相比，该书有较大的特色。该书较为全面地介绍了大数据处理相关的基本概念和原理，着重介绍了 Hadoop MapReduce 大数据处理系统的组成结构、工作原理和编程模型。在此基础上，该书由浅入深、循序渐进，重点介绍和分析了基于 MapReduce 的各种大数据并行处理算法和程序设计的思想方法，并辅以经过完整实现和验证的各种算法代码的分析介绍，内容涵盖了常用的基本算法以及较为复杂的机器学习和数据挖掘算法的设计与实现。该书还通过一系列来自全国性大赛获奖算法、部分优秀课程设计、部分科研成果、以及业界实际的大数据应用编程实战案例，较为深入地阐述了相关的大数据并行处理和编程技术。

作为国内第一本经过多年课堂教学实践总结而成的大数据并行处理和编程技术书籍，该书很适合高等院校作为 MapReduce 大数据并行处理技术课程的教材，同时也很适合于作为大数据处理应用开发和编程专业技术人员的参考手册。

此外，我很高兴地看到，该书已纳入了教育部计算机类专业教学指导委员会制定的计算机类专业系统能力培养计划，作为“计算机类专业系统能力培养系列教材”。从计算技术的角度看，大数据处理是一种涉及到几乎所有计算机技术层面的综合性计算技术，涉及到计算机软硬件技术的方方面面，因此，大数据处理是一门综合性、最能体现计算机系统能力培养的课程。为此，把大数据处理纳入计算机类专业系统能力培养课程体系中第三层次的核心课程，作为一门起到一定“收官”作用的综合性课程，这是在计算机系统能力培养方面的一个很好的尝试。

中国工程院院士

中国计算机学会大数据专家委员会主任

李国杰

2014 年 7 月，于北京

推荐序二

大数据处理技术——信息时代的金钥匙

可以毫不夸张地说，我们现在正处在信息爆炸的时代！随着移动互联网和物联网的迅猛发展，数据正在以前所未有的规模急剧增长。海量数据的收集、存储、处理、分析以及由此而产生的信息服务正在成为全球信息技术发展的主流。如果说大数据是信息时代的“石油”，那么大数据处理就是信息时代对这些数据“石油”的开采、运输、加工和提炼过程。可以预见，我们未来的生活将会像我们依赖石油化工产品一样依赖丰富多彩的大数据分析应用和信息服务。

在这一轮大数据的发展和变革中，中国以其过去三十几年信息化和近几年移动互联网和物联网发展所积累的基础，正面临着前所未有的良好发展机遇。与西方发达国家相比，中国在大数据技术发展方面有不少自身独特的优势：中国巨大的人口基数孕育着巨大的技术市场，同时也造就巨大的数据资源。因此，在这一轮新的变革中，一方面，我们面临着很多技术挑战，需要努力学习国外先进的技术和经验，需要去不断探索和发现很多新的数据分析应用商业模式；但另一方面，上述的一些独特优势，加上大数据领域目前还处于初期阶段，这些因素使得我们有很好的机会和一定的优势与其他发达市场一道，探索如何通过大数据技术来进一步提高数据信息分析应用的技术水平与服务质量，并以此改善我们的生活。因此，在大数据的研究与应用方面，中国和西方发达国家可以齐头并进、互相借鉴。

我很欣慰地看到，作为国内最早从事大数据技术研究和教学的团队之一，南京大学黄宜华教授和他的大数据实验室同仁们在大数据技术领域已经进行了多年系统深入的研究工作，取得了卓有成效的研究成果。我和黄教授相识于两年多前的中国大数据技术大会上，黄教授的学识和为人令人钦佩。此后我们在大数据研究方面展开了建设性的合作。

英特尔作为一家全球领先的计算技术公司，长期以来始终以计算技术的创新为己任。在大数据处理技术方面，我们也竭尽全力发挥出我们在软硬件平台的组合优势引领大数据技术的全

面发展和推广。让人欣喜的是，英特尔中国团队是英特尔 Hadoop 系统开发的主力军。这也为我们与黄教授的合作创造了得天独厚的条件。

这本《深入理解大数据》的力作正是我们双方在大数据领域共同努力的结晶，是以学术界和业界完美结合的方式，在融合了学术界系统化的研究教学工作和业界深度的系统和应用研发工作基础上，成功打造出的一本大数据技术佳作。

本书在总结多年的技术研发和教学内容的基础上，深入浅出地概括了大数据的基本概念和技术内容，然后重点介绍了主流大数据处理系统 Hadoop 的基本原理和架构；在此基础上逐章详细介绍了 Hadoop 平台下大数据分布存储、并行化计算和算法设计等一系列重要技术及其编程方法，尤其是详细介绍了大量实际的大数据处理算法设计和编程实现方法。相信这是一本适合软件技术人员和 IT 行业管理人员理解和掌握大数据技术的不可多得的技术书籍，也是一本适合于在校大学生和研究生学习和掌握大数据处理和编程技术的好教材。

在未来十年里，我们将看到数以十亿计的移动设备、可穿戴设备和智能终端设备融入我们生活的方方面面，沉浸式计算体验（Immersive Computing Experience）将成为我们生活的常态。随之应运而生的海量数据，大数据的存储和处理将分布于数据生命周期的各个阶段，因此，我们需要更多、更便捷的大数据处理方法和能力。显然，这并非一个算法、一个软件或一个高性能处理器所能单独完成的事情。我们需要以开放和软硬件结合的综合性体系架构，来实现大规模的大数据分析处理系统的部署和使用。基于英特尔架构优化的 Hadoop 平台是一个很好的开端。

大数据技术带来的变革方兴未艾，我们正在打造开启信息技术新时代的金钥匙。我也衷心地希望这本书能成为读者打开大数据技术之门的钥匙！

英特尔亚太研发有限公司总经理

何京翔

推荐序三

据预测，到 2020 年，全球包含 PC、平板电脑、智能手机等联网设备将超过 300 亿台。实际上，随着物联网技术与可穿戴设备的飞速发展，终端设备会远远大于这个数量。随之而来各种应用也会爆炸式增长。大量应用会产生巨量的数据，数据内容的种类也会非常多样化，比如大量的普通数据、医疗影像数据以及越来越多城市摄像头所录下的视频数据等。

根据国际分析机构 IDC 的统计，全球不同设备产生的数据量，到 2020 年预计将会突破 40ZB。如此海量、持续、细粒度、多样化的数据，让各个行业都看到了数据的巨大潜在价值，这将大力推动大数据技术和应用的发展，为当今和未来的科技和经济发展以及社会的生产和生活带来重大影响。

目前，全球的大数据市场规模很大，并保持了 30% 以上的年增长率。在中国，据 2012 年的统计，中国占据全球数据总量的 13%；而据预计，随着中国的不断发展，作为全球第二大经济体，中国将拥有全球最高的终端设备出货量以及全球最高的物联网的用户数，并且我们的增长速度也将超过全球。到 2020 年，中国的整个数据量将超过 8ZB，也就是说，数据增长率将是 2012 年的 23 倍。虽然中国的大数据解决方案才刚刚起步，但是预计在未来五年内中国大数据市场将会保持 50% 的增长率。大数据市场在未来几年将拥有整体 IT 市场 4 倍的增长速度、服务器市场 5 倍的增长速度，并且将远远高于云计算市场的增长速度。大数据市场在中国的 IT 行业已经变得越来越重要。

近年来，大数据的各种应用层出不穷。在政府行业的“平安城市”项目中，很多摄像头在收集视频数据，通过这些视频数据的管理和分析，可有效预防犯罪、保障社会公共安全。此外，互联网舆情分析、地震预测、气象分析、人口信息综合分析等，也都是政府民生相关的大数据应用领域。

在金融行业，很多行业用户使用大数据解决方案，对海量结构化交易数据进行实时入库处理，并提供并发查询，进行金融欺诈分析监测以预防金融犯罪；还可以通过数据分析挖掘以进

行精准的金融营销，通过对金融分析发现更多的投资组合、避免投资风险等。

电信行业也开始使用大数据解决方案，提供对上亿电话通联详单数据的快速查询和分析，并可以通过对电信用户数据的分析，提供基于 LBS（基于位置的服务）的数据分析服务，进行电信产品和服务的精准营销、精准广告投放和促销等。

零售行业也开始使用大数据解决方案，通过对用户的交易数据进行关联性分析挖掘以决定商品在货架上的摆放位置，在方便用户购物的同时、提高用户的购买量。

交通领域也逐步采用智能交通管理方案，通过对道路的交通状况进行分析预测，实现智能化道路交通管理和分流，对违章进行自动检测和处理，完成套牌车辆检测、区域分析以及其他异常行为的检测、分析和预警。其他还有像铁路、航空等交通行业的客票和货运处理、物流管理等，都将成为典型的大数据应用领域。

在医疗行业，对于医疗影像（如 X 光片、CT 片等）、就诊、用药、手术、住院状况等医疗数据的信息化管理要求越来越高。同时目前医疗行业也开始关注如何通过对医疗大数据进行融合分析，为医生提供辅助诊断、医疗方案推荐、药物疗效分析、病因分析、专家治疗经验共享等基于大数据的智能化诊疗服务。

大数据广泛的应用前景代表了 IT 行业的未来。近年来，大数据的巨大应用需求推动了大数据处理技术取得了长足的进步和发展。但是，大数据的 4V 特性（大体量、多样性、时效性、以及精确性）决定了大数据处理仍然面临着巨大的技术困难和挑战，因此，我们还需要大力推动大数据技术的研发和应用。这就需要培养更多熟练掌握大数据处理技术的专业人才。而今天的人才市场上还极为缺乏这种熟练掌握大数据技术的专业人才。

为此，需要让更多的专业技术人员学习和掌握大数据技术，这正是我们编写这本 Hadoop 大数据处理技术书籍的主要动机和目的。本书希望通过对目前最为主流、最广为业界接受使用的 Hadoop 大数据处理和编程技术的深入介绍，对 IT 专业技术人员与学生学习和掌握大数据技术起到较大的帮助作用！

英特尔中国大数据首席技术官

苗凯翔博士

2014 年 3 月 20 日，于上海

丛书序言

——计算机专业学生系统能力培养和系统课程设置的研究

未来的5 ~ 10年是中国实现工业化与信息化融合，利用信息技术与装备提高资源利用率、改造传统产业、优化经济结构、提高技术创新能力与现代管理水平的关键时期，而实现这一目标，对于高效利用计算系统的其他传统专业的专业人员需要了解和掌握计算思维，对于负责研发多种计算系统的计算机专业的专业人员则需要具备系统级的设计、实现和应用能力。

1. 计算技术发展特点分析

进入本世纪以来，计算技术正在发生重要发展和变化，在上世纪个人机普及和Internet快速发展基础上，计算技术从初期的科学计算与信息处理进入了以移动互联、物物相联、云计算与大数据计算为主要特征的新型网络时代，在这一发展过程中，计算技术也呈现出以下新的系统形态和技术特征。

（1）四类新型计算系统

1）嵌入式计算系统 在移动互联网、物联网、智能家电、三网融合等行业技术与产业发展中，嵌入式计算系统有着举足重轻和广泛的作用。例如，移动互联网中的移动智能终端、物联网中的汇聚节点、“三网融合”后的电视机顶盒等是复杂而新型的嵌入式计算系统；除此之外，新一代武器装备，工业化与信息化融合战略实施所推动的工业智能装备，其核心也是嵌入式计算系统。因此，嵌入式计算将成为新型计算系统的主要形态之一。在当今网络时代，嵌入式计算系统也日益呈现网络化的开放特点。

2）移动计算系统 在移动互联网、物联网、智能家电以及新型装备中，均以移动通信网络为基础，在此基础上，移动计算成为关键技术。移动计算技术将使计算机或其他信息智能终端设备在无线环境下实现数据传输及资源共享，其核心技术涉及支持高性能、低功耗、无线连接和轻松移动的移动处理机及其软件技术。

3）并行计算系统 随着半导体工艺技术的飞速进步和体系结构的不断发展，多核 / 众核处理机硬件日趋普及，使得昔日高端的并行计算呈现出普适化的发展趋势；多核技术就是在处理器上拥有两个或更多一样功能的处理器核心，即将数个物理处理器核心整合在一个内核中，数个处理器核心在共享芯片组存储界面的同时，可以完全独立地完成各自操作，从而能在平衡功耗的基础上极大地提高 CPU 性能；其对计算系统微体系结构、系统软件与编程环境均有很大影响；同时，云计算也是建立在廉价服务器组成的大规模集群并行计算基础之上。因此，并行计算将成为各类计算系统的基础技术。

4）基于服务的计算系统 无论是云计算还是其他现代网络化应用软件系统，均以服务计算为核心技术。服务计算是指面向服务的体系结构（SOA）和面向服务的计算（SOC）技术，它是标识分布式系统和软件集成领域技术进步的一个里程碑。服务作为一种自治、开放以及与平台无关的网络化构件可使分布式应用具有更好的复用性、灵活性和可增长性。基于服务组织计算资源所具有的松耦合特征使得遵从 SOA 的企业 IT 架构不仅可以有效保护企业投资、促进遗留系统的复用，而且可以支持企业随需应变的敏捷性和先进的软件外包管理模式。Web 服务技术是当前 SOA 的主流实现方式，其已经形成了规范的服务定义、服务组合以及服务访问。

（2）“四化”主要特征

1）网络化 在当今网络时代，各类计算系统无不呈现出网络化发展趋势，除了云计算系统、企业服务计算系统、移动计算系统之外，嵌入式计算系统也在物联时代通过网络化成为开放式系统。即，当今的计算系统必然与网络相关，尽管各种有线网络、无线网络所具有的通信方式、通信能力与通信品质有较大区别，但均使得与其相联的计算系统能力得以充分延伸，更能满足应用需求。网络化对计算系统的开放适应能力、协同工作能力等也提出了更高的要求。

2）多媒体化 无论是传统 Internet 应用服务，还是新兴的移动互联网服务业务，多媒体化是其面向人类、实现服务的主要形态特征之一。多媒体技术是利用计算机对文本、图形、图像、声音、动画、视频等多种信息进行综合处理、建立逻辑关系和人机交互作用的新技术。多媒体技术使计算机可以处理人类生活中最直接、最普遍的信息，从而使得计算机应用领域及功能得到了极大的扩展，使计算机系统的人机交互界面和手段更加友好和方便。多媒体具有计算机综合处理多种媒体信息的集成性、实时性与交互性特点。

3）大数据化 随着物联网、移动互联网、社会化网络的快速发展，半结构化及非结构化的数据呈几何倍增长。数据来源的渠道也逐渐增多，不仅包括了本地的文档、音视频，还包括网络内容和社交媒体；不仅包括 Internet 数据，更包括感知物理世界的数据。从各种类型的数据中快速获得有价值信息的能力，称为大数据技术。大数据具有体量巨大、类型繁多、价值密度低、处理速度快等特点。大数据时代的来临，给各行各业的数据处理与业务发展带来重要变

革，也对计算系统的新型计算模型、大规模并行处理、分布式数据存储、高效的数据处理机制等提出了新的挑战。

4）智能化 无论是计算系统的结构动态重构，还是软件系统的能力动态演化；无论是传统 Internet 的搜索服务，还是新兴移动互联的位置服务；无论是智能交通应用，还是智能电网应用，无不显现出鲜明的智能化特征。智能化将影响计算系统的体系结构、软件形态、处理算法以及应用界面等。例如，相对于功能手机的智能手机是一种安装了开放式操作系统的手机，可以随意安装和卸载应用软件，具备无线接入互联网、多任务和复制粘贴以及良好用户体验等能力；相对于传统搜索引擎的智能搜索引擎是结合了人工智能技术的新一代搜索引擎，不仅具有传统的快速检索、相关度排序等功能，更具有用户角色登记、用户兴趣自动识别、内容的语义理解、智能信息化过滤和推送等功能，其追求的目标是根据用户的请求从可以获得的网络资源中检索出对用户最有价值的信息。

2. 系统能力的主要内涵及培养需求

（1）主要内涵

计算机专业学生的系统能力的核心是掌握计算系统内部各软件 / 硬件部分的关联关系与逻辑层次；了解计算系统呈现的外部特性以及与人和物理世界的交互模式；在掌握基本系统原理的基础上，进一步掌握设计、实现计算机硬件、系统软件以及应用系统的综合能力。

（2）培养需求

要适应“四类计算系统，四化主要特征”的计算技术发展特点，计算机专业人才培养必须“与时俱进”，体现计算技术与信息产业发展对学生系统能力培养的需求。在教育思想上要突现系统观教育理念，在教学内容中体现新型计算系统原理，在实践环节上展现计算系统平台技术。

要深刻理解系统化专业教育思想对计算机专业高等教育过程所带来的影响。系统化教育和系统能力培养要采取系统科学的方法，将计算对象看成一个整体，追求系统的整体优化；要夯实系统理论基础，使学生能够构建出准确描述真实系统的模型，进而能够用于预测系统行为；要强化系统实践，培养学生能够有效地构造正确系统的能力。

从系统观出发，计算机专业的教学应该注意教学生怎样从系统的层面上思考（设计过程、工具、用户和物理环境的交互），讲透原理（基本原则、架构、协议、编译以及仿真等），强化系统性的实践教学培养过程和内容，激发学生的辩证思考能力，帮助他们理解和掌控数字世界。

3. 计算机专业系统能力培养课程体系设置总体思路

为了更好地培养适应新技术发展的、具有系统设计和系统应用能力的计算机专门人才，我们需要建立新的计算机专业本科教学课程体系，特别是设立有关系统级综合性课程，并重新规划计算机系统核心课程的内容，使这些核心课程之间的内容联系更紧密、衔接更顺畅。

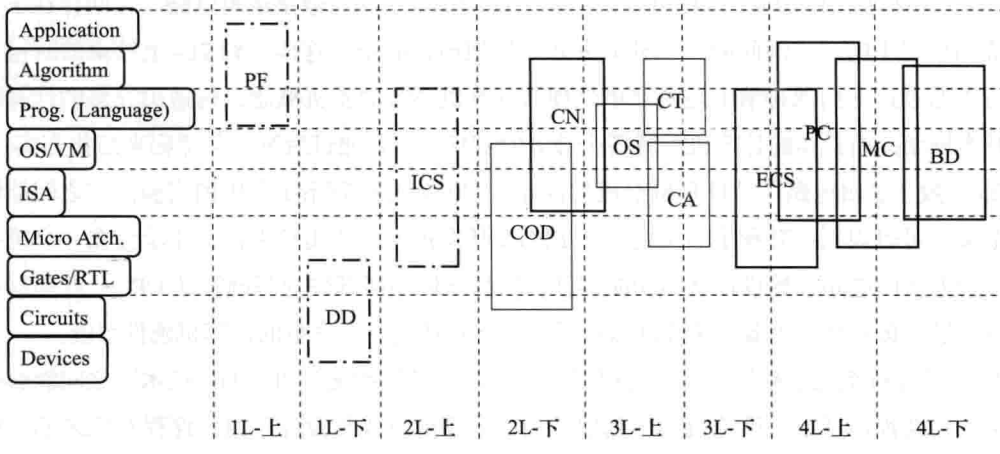
我们建议把课程分成三个层次：计算机系统基础课程、重组内容的核心课程、侧重不同计算系统的若干相关平台应用课程。

第一层次核心课程包括："程序设计基础（PF）"、"数字逻辑电路（DD）"和"计算机系统基础（ICS）"。

第二层次核心课程包括："计算机组成与设计（COD）"、"操作系统（OS）"、"编译技术（CT）"和"计算机系统结构（CA）"。

第三层次核心课程包括："嵌入式计算系统（ECS）"、"计算机网络（CN）"、"移动计算（MC）"、"并行计算（PC）"和"大数据并行处理技术（BD）"。

基于这三个层次的课程体系中相关课程设置方案如下图所示。



图中左边部分是计算机系统的各个抽象层，右边的矩形表示课程，其上下两条边的位置标示了课程内容在系统抽象层中的涵盖范围，矩形的左右两条边的位置标示了课程大约在哪个年级开设。点划线、细实线和粗实线分别表示第一、第二和第三层次核心课程。

从图中可以看出，该课程体系的基本思路是：先讲顶层比较抽象的编程方面的内容；再讲底层有关系统的具体实现基础内容；然后再从两头到中间，把顶层程序设计的内容和底层电路的内容按照程序员视角全部串起来；在此基础上，再按序分别介绍计算机系统硬件、操作系统和编译器的实现细节。至此的所有课程内容主要介绍单处理器系统的相关内容，而计算机体系结构主要介绍各不同并行粒度的体系结构及其相关的操作系统实现技术和编译器实现技术。第三层次的课程没有先后顺序，而且都可以是选修课，课程内容应体现第一和第二层次课程内容的螺旋式上升趋势，也即第三层次课程内容涉及的系统抽象层与第一和第二层次课程涉及的系统抽象层是重叠的，但内容并不是简单重复，应该讲授在特定计算系统中的相应教学内容。例如，对于"嵌入式计算系统（ECS）"课程，虽然它所涉及的系统抽象层与"计算机系统基础

（ICS）”课程涉及的系统抽象层完全一样，但是，这两门课程的教学内容基本上不重叠。前者着重介绍与嵌入式计算系统相关的指令集体系结构设计、操作系统实现和底层硬件设计等内容，而后者着重介绍如何从程序员的角度来理解计算机系统设计与实现中涉及的基础内容。

与传统课程体系设置相比，最大的不同在于新的课程体系中有一门涉及计算机系统各个抽象层面的能够贯穿整个计算机系统设计和实现的基础课程：“计算机系统基础（ICS）”。该课程讲解如何从程序员角度来理解计算机系统，可以使程序员进一步明确程序设计语言中的语句、数据和程序是如何在计算机系统中实现和运行的，让程序员了解不同的程序设计方法为什么会有不同的性能等。

此外，新的课程体系中，强调课程之间的衔接和连贯，主要体现在以下几个方面。

1）“计算机系统基础”课程可以把“程序设计基础”和“数字逻辑电路”之间存在于计算机系统抽象层中的“中间间隔”填补上去并很好地衔接起来，这样，到 2L- 上结束的时候，学生就可以通过这三门课程清晰地建立单处理器计算机系统的整机概念，构造出完整的计算机系统的基本框架，而具体的计算机系统各个部分的实现细节再通过后续相关课程来细化充实。

2）“数字逻辑电路”、“计算机组成与设计”、“嵌入式计算系统”中的实验内容之间能够很好地衔接，可以规划一套承上启下的基于 FPGA 开发板的综合实验平台，让学生在一个统一的实验平台上从门电路开始设计基本功能部件，然后再以功能部件为基础设计 CPU、存储器和外围接口，最终将 CPU、存储器和 I/O 接口通过总线互连为一个完整的计算机硬件系统。

3）“计算机系统基础”、“计算机组成与设计”、“操作系统”和“编译技术”之间能够很好地衔接。新课程体系中“计算机系统基础”和“计算机组成与设计”两门课程对原来的“计算机系统概论”和“计算机组成原理”的内容进行了重新调整和统筹规划，这两门课程的内容是相互密切关联的。对于“计算机系统基础”与“操作系统”、“编译技术”的关系，因为“计算机系统基础”以 Intel x86 为模型机进行讲解，所以它为“操作系统”（特别是 Linux 内核分析）提供了很好的体系结构基础。同时，在“计算机系统基础”课程中为了清楚地解释程序中的文件访问和设备访问等问题，会从程序员角度简单引入一些操作系统中的相关基础知识。此外，在“计算机系统基础”课程中，会讲解高级语言程序如何进行转换、链接以生成可执行代码的问题；“计算机组成与设计”中的流水线处理等也与编译优化相关，而且“计算机组成与设计”以 MIPS 为模型机进行讲解，而 MIPS 模拟器可以为“编译技术”的实验提供可验证实验环境，因而“计算机系统基础”和“计算机组成与设计”两门课程都与“编译技术”有密切的关联。“计算机系统基础”、“计算机组成与设计”、“操作系统”和“编译技术”这四门课程构成了一组计算机系统能力培养最基本的核心课程。

从“计算机系统基础”课程的内容和教学目标以及开设时间来看，位于较高抽象层的先行

课（如程序设计基础和数据结构等课程）可以按照原来的内容和方式开设和教学，而作为新的“计算机系统基础”和“计算机组成与设计”先导课的“数字逻辑电路”，则需要对传统的教学内容，特别是实验内容和实验手段方面进行修改和完善。

有了“计算机系统基础”和“计算机组成与设计”课程的基础，作为后续课程的操作系统、编译原理等将更容易被学生从计算机系统整体的角度理解，课程内容方面不需要大的改动，但是操作系统和编译器的实验要以先行课程实现的计算机硬件系统为基础，这样才能形成一致的、完整的计算机系统整体概念。

本研究还对 12 门课程的规划思路、主要教学内容及实验内容进行了研究和阐述，具体内容详见公开发表的研究报告。

4. 关于本研究项目及本系列教材

机械工业出版社华章公司在较早的时间就引进出版了 MIT、UC-Berkeley、CMU 等国际知名院校有关计算机系统课程的多种教材，并推动和组织了计算机系统能力培养相关的研究，对国内计算机系统能力培养起到了积极的促进作用。

本项研究是教育部 2013 ~ 2017 年计算机类专业教学指导委员会“计算机类专业系统能力培养研究”项目之一，研究组成员由国防科技大学王志英、北京航空航天大学马殿富、西北工业大学周兴社、南开大学吴功宜、武汉大学何炎祥、南京大学袁春风、北京大学陈向群、中国科技大学安虹、天津大学张刚、机械工业出版社华章公司温莉芳等组成，研究报告分别发表于中国计算机学会《中国计算机科学技术发展报告》及《计算机教育》杂志。

本系列教材编委会在上述研究的基础上对本套教材的出版工作经过了精心策划，选择了对系统观教育和系统能力培养有研究和实践的教师作为作者，以系统观为核心编写了本系列教材。我们相信本系列教材的出版和使用，将对提高国内高校计算机类专业学生的系统能力和整体水平起到积极的促进作用。

“计算机类专业系统能力培养系列教材”编委会组成如下：

主　任　王志英

副主任　马殿富

委　员　周兴社　吴功宜　何炎祥　袁春风　陈向群　安　虹　温莉芳

秘　书　姚　蕾

此外，本系列教材的出版得到赛灵思电子科技有限公司和英特尔有限公司的支持。

“计算机类专业系统能力培养系列教材”编委会

2014 年 5 月

前　言

2012 年以来，大数据（Big Data）技术在全世界范围内迅猛发展，在全球学术界、工业界和各国政府得到了高度关注和重视，掀起了一场可与 20 世纪 90 年代的信息高速公路相提并论的发展热潮。

大数据技术如此重要，已经被我国政府提升到国家重大发展战略的高度。2014 年我国政府工作报告中指出："设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展"。由此可见，大数据已经被我国政府列为推动国家科技创新和引领经济结构优化升级、赶超国际先进水平、引领国家未来产业发展的战略性计划。两会期间，CCTV 中央电视台的新闻报道开创性地引入了大数据新闻报道手段，以大数据说话，高频率使用大数据报道两会重大新闻，引起了全国民众的普遍关注和兴趣。

大数据也同样成为各发达国家政府高度关注的战略性高科技技术和产业。2012 年 3 月，美国总统奥巴马签署并发布了一个"大数据研究发展创新计划"（Big Data R&D Initiative），投资 2 亿美元启动大数据技术和工具研发，这是继 1993 年美国宣布"信息高速公路"计划后的又一次重大科技发展部署。美国政府认为大数据是"未来的新石油"，将大数据研究上升为国家意志，认为大数据将对未来的科技与经济发展带来重大影响，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家间和企业间新的争夺焦点。在随后的近两年里，英国、法国、德国、日本等发达国家政府都纷纷推出了相应的大数据发展战略计划。

《大数据时代》一书的作者、英国牛津大学教授、被誉为"大数据时代预言家"的维克托·迈尔－舍恩伯格认为："大数据开启了一次重大的时代转型"，认为大数据将带来巨大的变革，改变我们的生活、工作和思维方式，改变我们的商业模式，影响我们的经济、政治、科技和社会等各个层面。他认为，大数据将成为企业的核心竞争力，成为一种商业资本，成为企业的重要资产。

大数据技术最大的推动力来自于行业应用需求。过去几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据量呈爆炸性增长。动辄达到 PB 级规模的行业 / 企业大数据已经远远超出了现有传统的计算技术和信息系统的处