



大数据、云计算、人工智能、信息安全人才培养丛书 “互联网+”新形态一体化教材

大数据

大数据基础
数据可视化
数据清洗与治理
Hadoop应用与开发
数据挖掘基础
SEO搜索引擎优化
MySQL数据库
大数据采集与数据处理

● 大数据集群搭建维护与数据存储

云计算

云计算基础
虚拟化与容器
云安全运维
网络工程与组网技术
现代通信技术
路由交换技术
无线网络技术
现代网络SDN技术
数据网组建与维护
局域网组建与维护
云数据中心架构与SDN技术

人工智能

人工智能基础
物联网基础
深度学习
机器学习

信息安全

信息安全基础
Linux服务器安全高级运维
Web安全与防御
防火墙技术与应用
计算机病毒与防范
数据存储与恢复
密码学基础
计算机网络安全运维
网络设备配置与综合实战
无线网络安全技术
VPN虚拟专用网安全
终端数据存储与恢复
工控安全
渗透测试
恶意代码分析
网络空间安全态势感知
数据库安全技术
网络安全协议
企业级数据安全与灾备管理

信息安全法律法规
终端数据安全及防泄密

专业基础

Linux操作系统基础
计算机网络基础
Ubuntu服务器管理
Windows Server 2016配置与管理
数据结构
Python程序设计
Java程序设计
C语言程序设计
C#程序设计
Android程序设计
XML基础教程
JavaScript基础教程
Web前端开发
OpenStack应用与开发
Spark应用与开发
静态网页设计与制作
HTML5与JavaScript程序设计
数据库设计与应用
数据库应用基础
UML建模与设计模式
ERP原理与应用
综合布线



扫描二维码
关注上海交通大学出版社
官方微信



ISBN 978-7-313-25992-9

9 787313 259929 >

定价：49.80元

“互联网+”新形态一体化教材

大数据集群搭建维护与数据存储

主编◎童 杰 冉孟廷 陈 阳

上海交通大学出版社

大数据、云计算、人工智能、信息安全人才培养丛书
“互联网+”新形态一体化教材

大数据集群 搭建维护与数据存储

DASHUJU JIQUN DAJIAN WEIHU YU SHUJU CUNCHU

主编◎童 杰 冉孟廷 陈 阳



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

大数据、云计算、人工智能、信息安全人才培养丛书
“互联网+”新形态一体化教材

大数据集群 搭建维护与数据存储

DASHUJU JIQUN DAJIAN WEIHU YU SHUJU CUNCHU

主 编◎童 杰 冉孟廷 陈 阳

副主编◎梁 宾

扫一扫
学习资源库



◆素材源码

◆教学课件

◆电子教案



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

全书从项目实战的角度出发,以 CentOS 7 系统为平台,搭建基于 Hadoop 3.3.1 的高可用集群,讲解 Linux 系统环境配置、Hadoop 集群的高可用搭建、大数据集群生态平台的管理维护、数据存储、各组件间的版本兼容等实践操作。书中所有项目案例均源自笔者亲身编写和测试,项目切实可行,真实有效;并在已教授的学生中得到了检验。本书提供了大量翔实的源码和图文并茂的实现步骤描述,解决了学生在实践操作上经常会遇到的问题,方便学生阅读理解并掌握在大数据集群搭建维护与数据存储中的知识和技术。

图书在版编目(CIP)数据

大数据集群搭建维护与数据存储 / 童杰, 冉孟廷,
陈阳主编. — 上海: 上海交通大学出版社, 2022.3 (2024.7 重印)
ISBN 978-7-313-25992-9

I. ①大… II. ①童… ②冉… ③陈… III. ①数据管
理 ②数据存贮 IV. ① TP274

中国版本图书馆 CIP 数据核字 (2022) 第 004917 号

大数据集群搭建维护与数据存储
DASHUJU JIQUN DAJIAN WEIHU YU SHUJU CUNCHU

主 编: 童杰 冉孟廷 陈阳	地 址: 上海市番禺路 951 号
出版发行: 上海交通大学出版社	电 话: 6407 1208
邮政编码: 200030	
印 制: 北京荣玉印刷有限公司	经 销: 全国新华书店
开 本: 889 mm × 1194 mm 1/16	印 张: 10.5
字 数: 185 千字	
版 次: 2022 年 3 月第 1 版	印 次: 2024 年 7 月第 2 次印刷
书 号: ISBN 978-7-313-25992-9	
定 价: 49.80 元	

版权所有 侵权必究

告读者: 如发现本书有印装质量问题请与印刷厂质量科联系
联系电话: 010-6020 6144

前言

PREFACE

大数据专业建设，任重而道远。大数据专业源于教育部在 2016 年对电子信息类专业的增补，此后，大数据专业如雨后春笋般在各高校中开设，但该专业目前仍处于初创阶段，大家都在摸索中一边建设大数据专业，一边研究如何办好该专业。2021 年，人力资源社会保障部、工业和信息化部共同制定了大数据工程技术人员国家职业技能标准，指出大数据工程技术人员是从事大数据采集、清洗、分析、治理、挖掘等技术研究，并加以利用、管理、维护和服务的工程技术人员。这为大数据专业人才的培养目标指明了方向。

办好大数据专业，课程设置是重点。结合对大数据工程技术人员主要工作任务内容的解码，以及大数据高职段对专业知识面掌握程度的要求，笔者在重庆信息技术职业学院大数据专业建设中进行了探索与研究，根据大数据工程的数据采集、数据存储与管理、数据分析、数据应用等环节，进行大数据专业的课程设置。限于篇幅，将部分课程名称列举如下（见表 1）。

表 1 大数据部分专业课程设置

序号	课程名称	总学时	理论学时	实践学时
1	大数据导论	64	32	32
2	Python 编程基础	96	32	64
3	数据爬虫开发	64	32	32
4	关系型数据库管理与设计	64	32	32
5	Java 程序设计	96	32	64
6	Linux 系统管理与维护	80	32	48
7	NoSQL 数据库基础与应用	96	32	64
8	大数据集群搭建维护与数据存储	64	32	32
9	大数据采集与数据处理	96	32	64
10	大数据实时计算分析	96	32	64

理想很“丰满”，现实很“骨感”。课程的设置是对知识体系的解码，需要有满足对应课程需求的教材，但目前市场上的大数据类教材，很多是将大数据的整体知识体系融入一本书中，从基础到进阶面面俱到，但往往又难以面面俱到。因此，教材的选择变成了一件让人头疼的事情，这也让笔者萌生了自己动手写教材的想法。

本书的特色

全书从项目实战的角度出发,以 CentOS 7 系统搭建的 Hadoop 3.3.1 高可用大数据集群生态平台为开发环境,解决学习者在搭建 Hadoop 高可用集群、大数据集群生态平台的管理维护、数据存储(HDFS、HBase、Hive)、各组件间的版本兼容等一系列环节中的操作难题。本书以面向高职和应用型本科大数据专业学生为主,力求做到理论讲解通俗易懂、简洁,做到理论够用,案例充实。

项目中所采用的组件版本,均是截至 2021 年 8 月,大数据生态圈中的最新版或最新稳定版,分别是 hadoop-3.3.1、hbase-2.3.6、hive-3.1.2 等,书中所有项目案例均源自笔者亲身编写和测试,项目切实可行,真实有效,并在其已教授的学生中得到了检验。

学习者完全可以按照本书提供的大量翔实的源码和图文并茂的步骤描述,掌握在 Hadoop 高可用集群搭建与数据存储过程中的知识和技术。

本书不适合零基础的读者学习。在学习本教材之前,要求学习者应具备表 1 中的第 1、4、5、6 项知识。按知识体系和课程设置的划分,本书内容中只包含对 Hadoop 基础知识的讲解。

本书的内容安排

第 1 章:介绍 Linux 集群的搭建。

第 2 章:介绍 Hadoop 高可用集群的搭建与配置,以及 HDFS 的数据存储与操作。

第 3 章:介绍 Yarn 高可用集群的搭建配置。

第 4 章:介绍分布式数据存储系统 HBase 的安装与配置,以及数据的存储与操作。

第 5 章:介绍分布式数据仓库 Hive 的安装与配置,以及数据的存储与操作。

另外,在附录中介绍了实践中需要用到的在 Windows 环境下 Maven 工程开发环境的安装与配置内容。

本教材内容的编排,基于大数据工程技术知识体系,并兼顾考虑高职学生的学习基础与学习能力。目前大数据工程技术内容很多,比如数据采集、数据传递、数据迁移、数据的实时和离线分析处理等。考虑到高职学生的学习特点,将知识内容全部放在本书中讲解,容易给老师和学生造成课程压力,也难以在 64 ~ 96 个学时内完成。因此,对大数据工程技术内容应该有一个知识体系的划分,在学完基础内容,对 Hadoop 生态圈技术内容有了一定了解后,再去学习数据采集、数据传递、数据迁移、数据的实时和离线分析处理等进阶技术内容。我们已在本校的大数据专业教学中,按照这种思路进行了实践,效果比较良好。

为满足对 Hadoop 生态圈进阶知识学习的需要,笔者编写有对应表 1 中第 9 项课程的教材,书名为《大数据采集与数据处理》,书中主要讲解在 Hadoop 高可用大数据集群环境下,数据采集、数据传递、数据迁移、数据的实时和离线分析处理,以及最后的数据可视化等一系列环节中的实践操作。建议对 Hadoop 生态圈进阶知识感兴趣的读者,在学习完本教材中的内容后,可继续学习《大数据采集与数据处理》书中的内容,以提升自己的在大数据应用开发方面的能力。

本书采用的运行环境

本书搭建 Linux 系统的工具如下:

(1) VMware Workstation 15 Pro,可满足较新版本操作系统的安装与运行管理。

(2) Linux CentOS 7.x-x86_64-Minimal,该系统安装文件较小,安装后,能满足在同一台计算机运行多台 Linux 系统时,对内存的占用要求相对较小;CentOS 是 Red Hat 公司发布的基于商业版 Linux 服务器的免费版本,属于企业级 Linux 系统产品。

本书运行环境对硬件的需求如下：

(1) CPU，不低于 intel i3 系列。

(2) 内存，因为需要在同一台计算机上安装三台 Linux 虚拟机来组建 Hadoop 集群，因此建议计算机的内存容量不低于 16G。

(3) 固态硬盘，不低于 128G，建议使用，但属于非必须配置项，用于存放安装后的 Linux 虚拟机文件，可提高对虚拟机文件的读写速度。

本书中使用的应用软件和开发工具

(1) SecureCRT x64 8.5.4：一款 SSH 客户端软件，为 Linux 提供了终端模拟，便于同时发送命令到多台 Linux 虚拟机。

(2) Eclipse IDE for Java Developers - 2020-12：用于 Hadoop 项目的开发。

(3) JDK 版本：Java 64-Bit 1.8。

针对教师的服务

本书配有教学课件、教案、授课计划、软件安装包等教学资源包，可以节约教师的备课时间，做到事半功倍。有需要者可致电 13810412048 或发邮件至 2393867076@qq.com。

致谢

本书在编写过程中，得到了热心人士和编辑老师的指导和帮助，对教材提出了许多宝贵的意见，在此一并表示衷心的感谢！

作者从书籍和网络中学到很多知识，对相关知识的作者，在此也一并表示衷心的感谢！

最后，感谢我的家人，是他们的默默支持，让我能有更多的时间来从事该书的编写工作。

意见反馈

虽然笔者在编写时力求完美，但大数据技术更新很快，个人能力有限，书中存在的不足和疏漏之处，敬请读者批评指正，希望在下一版中，能给读者带来更好的体验。也欢迎读者给予宝贵意见，作者的联系方式：jacktone@foxmail.com。

童杰

2021 年 7 月于重庆万州

目录

CONTENTS

第 1 章 Linux 集群的安装与配置	1	2.3 安装配置 Hadoop 集群	55
1.1 集群规划	2	2.3.1 在主节点上操作	55
1.2 安装前的准备	2	2.3.2 将 Hadoop 复制到其他从节点	58
1.2.1 磁盘格式	2	2.3.3 格式化 Hadoop 的 NameNode	58
1.2.2 检测虚拟化技术	4	2.4 启动集群与测试	59
1.2.3 在 BIOS 中开启虚拟化设置	6	2.5 搭建 Hadoop 的 HA (高可用)	60
1.3 虚拟机的安装	8	2.5.1 安装与配置 ZooKeeper	60
1.4 CentOS 7 的安装	15	2.5.2 配置 HDFS	64
1.5 node2 节点虚拟机的配置	23	2.5.3 启动 JN 节点	68
1.6 克隆 node2 节点, 完成 node3 节点虚拟机		2.5.4 格式化 NameNode 与 ZKFC	68
的安装	25	2.5.5 验证 Hadoop 的 HA 集群	69
拓展练习	29	2.5.6 退出集群	72
第 2 章 Hadoop 集群搭建	30	2.6 操作 HDFS	73
2.1 Hadoop 概述	31	2.6.1 命令方式	73
2.1.1 什么是 Hadoop	31	2.6.2 编程方式	74
2.1.2 Hadoop 3 V.S. Hadoop 2	31	2.7 异常处理	90
2.2 搭建前的系统环境配置	32	2.7.1 NameNode 无法启动	90
2.2.1 虚拟网卡的检测及问题处理	32	2.7.2 DataNode 无法启动	90
2.2.2 配置 Linux 系统的网卡、修改主机名		2.7.3 NameNode 格式化失败	90
与主机映射	35	拓展练习	91
2.2.3 完成常用命令的安装	42	第 3 章 资源管理框架 YARN 集群	
2.2.4 安装 JDK	43	搭建	92
2.2.5 完成 yum 的阿里云源配置	45	3.1 什么是 YARN	93
2.2.6 升级 OpenSSL 安全套接层协议	45	3.2 YARN 的 HA 搭建	93
2.2.7 关闭 SELinux 与 THP 服务	50	3.2.1 编辑 “hadoop-env.sh” 文件	93
2.2.8 在所有节点上完成 SSH 免密码登录 ..	51	3.2.2 编辑 “mapred-site.xml” 文件	94
2.2.9 配置时间同步	52	3.2.3 编辑 “yarn-site.xml” 文件	94
		3.2.4 编辑 “yarn-env.sh” 文件	96

3.2.5 其他从节点上的配置	96	5.2.1 安装 MySQL 的开源 MariaDB 版	116
3.3 启动集群, 并验证 YARN 的 HA	96	5.2.2 进行字符集设置, 以兼容中文	117
3.3.1 启动集群	96	5.2.3 启动服务	117
3.3.2 验证 YARN 的 HA 是否成功	98	5.2.4 访问测试	117
3.4 启动历史查看服务器	98	5.2.5 远程访问 MySQL	118
3.5 退出集群	100	5.3 安装与配置 Hive	122
3.6 其他命令	100	5.3.1 下载安装	122
拓展练习	101	5.3.2 设置环境变量	124
第 4 章 分布式数据库 HBase	102	5.3.3 配置 Hive	124
4.1 什么是 HBase	103	5.3.4 启动 Hive	128
4.2 HBase 的 HA 搭建与操作	103	5.4 Hive 数据操作	132
4.2.1 HBase 的 HA 搭建	103	5.4.1 数据库操作	133
4.2.2 启动 HBase 集群	107	5.4.2 数据表操作	134
4.2.3 解决启动过程中可能出现的问题	110	5.4.3 表的分区与分桶	137
4.2.4 关闭 HBase 集群	111	5.4.4 内部表与外部表	139
4.2.5 操作 HBase	112	5.5 Hive 的其他操作	144
4.2.6 异常处理	114	5.6 异常处理	146
拓展练习	114	拓展练习	149
第 5 章 数据仓库工具 Hive	115	附录 Windows 环境下 Maven 工程开发环境 的安装与配置	150
5.1 什么是 Hive	116	参考文献	160
5.2 安装 MySQL	116		

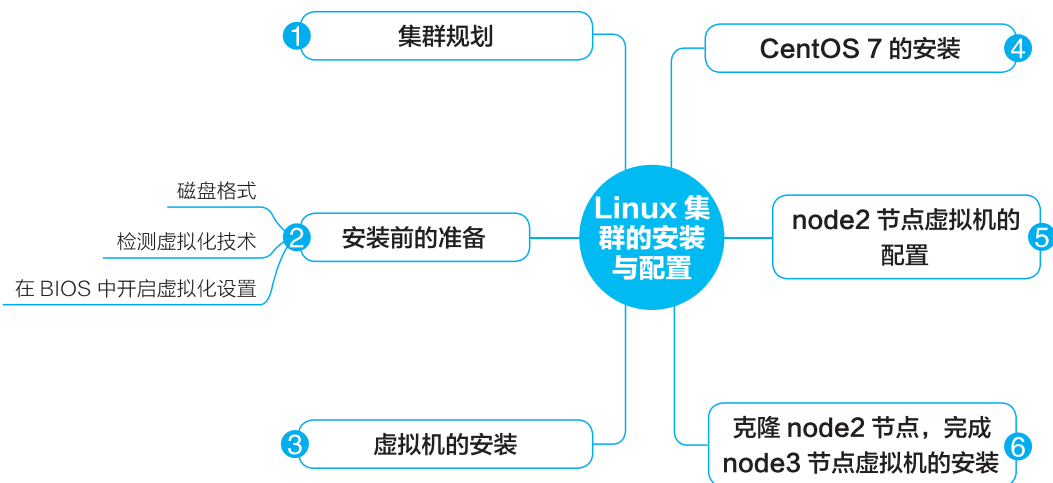
第 1 章

Linux 集群的安装与配置

学习目标

- 1 掌握 Linux 系统的安装。
- 2 掌握 Linux 集群节点环境的配置。
- 3 熟悉环境安装前的磁盘格式化。
- 4 熟悉虚拟化的开启。

知识导图





1.1 集群规划

搭建一个最小规模的 Hadoop HA（高可用）集群，至少需要三台 Linux 服务器（以下称为节点），对集群节点的设计规划如表 1.1 所示。

表 1.1 集群节点规划

序号	节点名称	分 工
1	node1	在集群中，同时担任管理节点、工作节点、主节点三种角色
2	node2	在集群中，同时担任备用管理节点、工作节点、主从节点三种角色
3	node3	在集群中，担任工作节点的角色

本章的目标，就是按设计规划的需求，先搭建三台 Linux 服务器，为后面 Hadoop 集群的搭建做准备。

采用的工具及软件如下：

- （1）VMware Workstation 15 Pro：可满足较新版本操作系统的安装与运行管理。
- （2）Linux CentOS 7.x-x86_64-Minimal：安装文件较小，安装后，能满足在同一台计算机运行多台 Linux 系统时，对内存的占用率相对较小；CentOS 是 Red Hat 公司发布的基于商业版 Linux 服务器的免费版本，属于企业级 Linux 服务器系统产品。

1.2 安装前的准备

1.2.1 磁盘格式

基于 VMware 的 Linux 安装，最好将存放 Linux 安装的磁盘格式设置为 NTFS 格式。NTFS (new technology file system) 是微软公司开发的一种新型磁盘格式，比起 FAT 格式，NTFS 支持更大容量文件的存储，便于 VMware 对大文件的保存。

在安装前，先检测用于安装 Linux 的磁盘格式是否是 NTFS 格式（见图 1.1），选择用于安装 Linux 系统的磁盘图标，单击鼠标右键，在快捷菜单中，单击“属性”。



图 1.1 磁盘“属性”菜单项

在弹出的磁盘“属性”窗口中（见图 1.2），查看当前磁盘的文件系统是否是 NTFS。



笔记



图 1.2 磁盘“属性”窗口

如果当前磁盘不是 NTFS 格式，则需要进行磁盘格式化（见图 1.3、图 1.4），选择用于安装 Linux 的磁盘图标，单击鼠标右键，在快捷菜单中，单击“格式化”。在弹出的格式化窗口中，选择 NTFS 文件系统，选择“快速格式化”选项，单击“开始”按钮，将磁盘格式化为 NTFS 格式。

注意：

磁盘格式化会永久删除磁盘中原有的文件，因此，在磁盘格式化之前，提前做好磁盘中原有文件的备份。

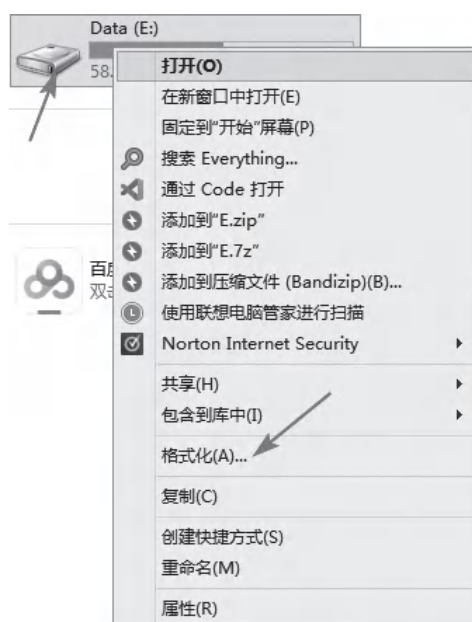


图 1.3 “格式化”菜单项

笔记

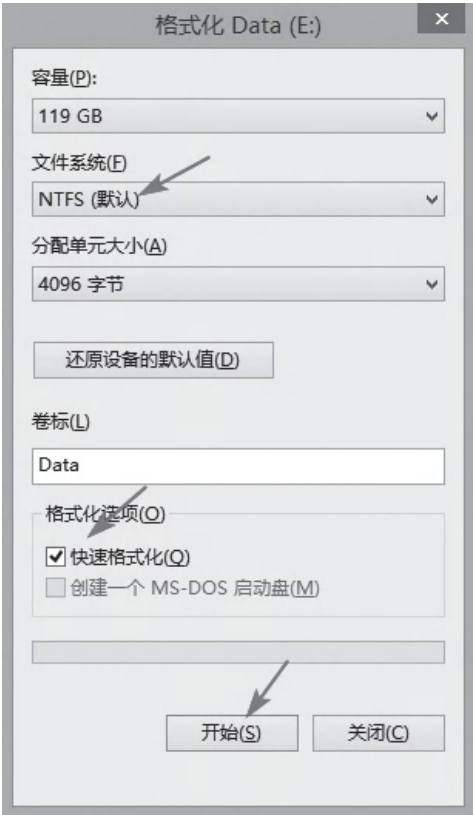


图 1.4 磁盘“格式化”窗口

1.2.2 检测虚拟化技术

基于 VMware 的 Linux 安装，需要 CPU 支持虚拟化技术，Intel CPU 的虚拟化技术称为 VT-x，AMD CPU 的虚拟化技术称为 AMD-V(SVM)。

如果一台计算机没有开启虚拟化技术，那么在 VMware 中安装 Linux 后，启动 Linux 时，计算机会给出警告窗口提示（见图 1.5），Linux 启动失败。

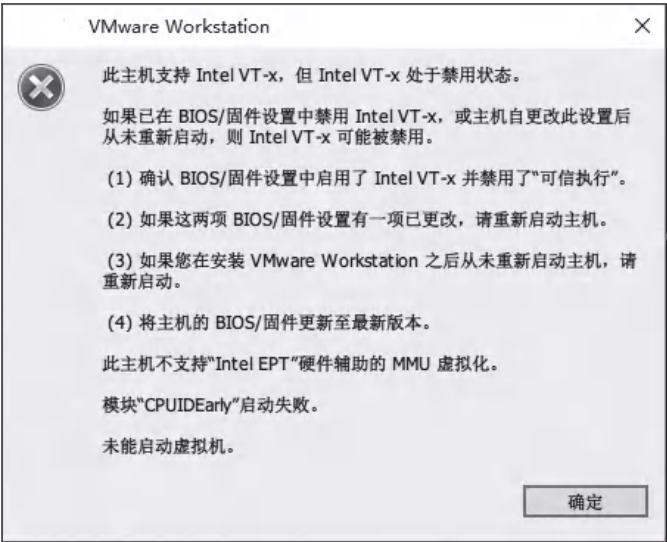


图 1.5 因虚拟化未开启而发出的警告

因此在安装 Linux 之前，应先查看计算机是否已开启了虚拟化技术，以 Windows 8 操作系统的计算机为例，打开计算机的“任务管理器”（见图 1.6），单击“任务管理器”的“性能”页，然后选择“CPU”选项，在窗口的右下方，如果“虚拟化”属性状态显示为“已启用”，则表明当前计算机的虚拟化技术已开启，可以直接跳过本书 1.2.3 节，从 1.3 节开始操作。

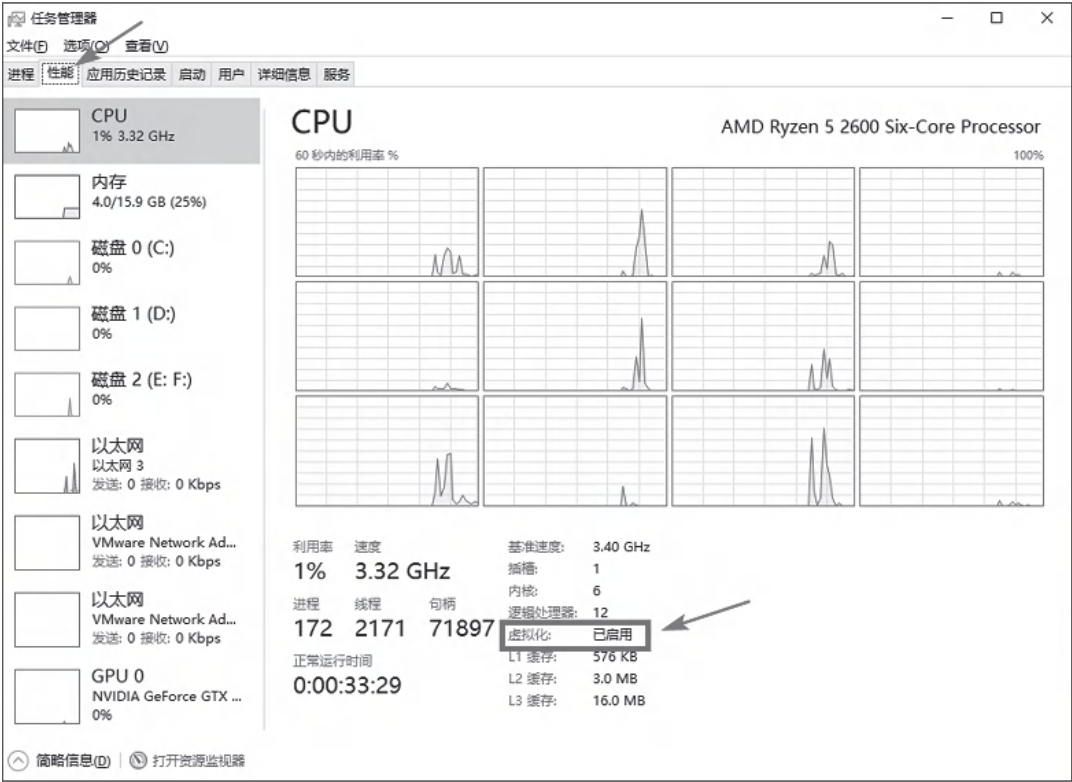


图 1.6 Windows “任务管理器”

如果“虚拟化”属性状态显示为“未启用”，可通过“CPU-z”工具（这是一款免费软件，可自行从网上下载）检测当前计算机的 CPU 是否支持虚拟化技术。打开“CPU-z”软件，以 Intel CPU 为例，在它的处理器页面的指令集处（见图 1.7），查看是否存在“VT-x”指令，如果不存在，则表明当前计算机不具备虚拟化技术，因此也就无法在 VMware 中正常启动 Linux，必须更换 CPU 或计算机。



图 1.7 “CPU-z” 软件窗口



如果“CPU-z”工具显示，存在“VT-x”指令或“AMD-V(SVM)”指令，则表明当前计算机支持虚拟化技术，接下来就需要进入计算机的 BIOS 中，开启虚拟化。

1.2.3 在 BIOS 中开启虚拟化设置

因厂商的原因，各品牌计算机的 BIOS 并不统一，因此不同品牌的计算机进入 BIOS 的快捷键方式各不相同，需要根据自己计算机品牌从网上查询进入 BIOS 的快捷键方式。可以使用百度搜索，搜索内容一般为“××× ××× 笔记本进 BIOS”，前一个“×××”指的是计算机品牌，第二个“×××”指的是计算机型号，查看搜索结果中关于计算机进入 BIOS 的方式。

以联想的 Thinkpad E 系列的笔记本电脑为例，在计算机开机或重启时，反复按下键盘上的“F1”，或“Fn+F1”键，或“F2”，或“Fn+F2”键，进入 BIOS Setup 设置界面。对于联想家用笔记本电脑，进入 BIOS Setup 设置界面（见图 1.8）后，依次选择“Configuration”→“Intel Virtual Technology”选项并按回车键。

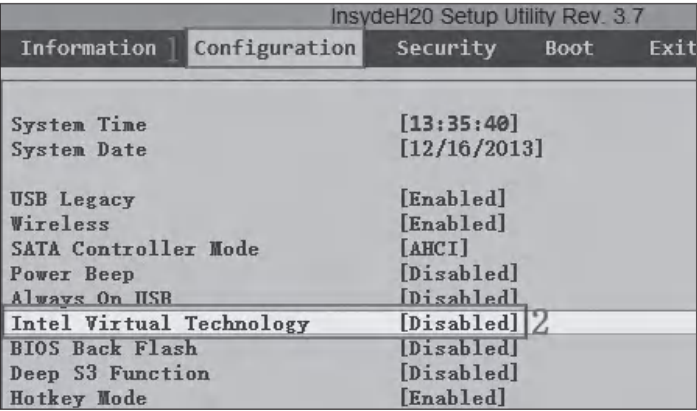


图 1.8 “Configuration” 菜单窗口

注意：
采用 Intel CPU 的机型使用的是“Intel Virtualization Technology”表示，采用 AMD CPU 的机型则使用的是“AMD SVM Support”或“AMD-V Technology”表示。

对于联想商用笔记本电脑，进入 BIOS Setup 设置界面（见图 1.9）后，依次选择“Security”→“Virtualization”选项并按回车键。

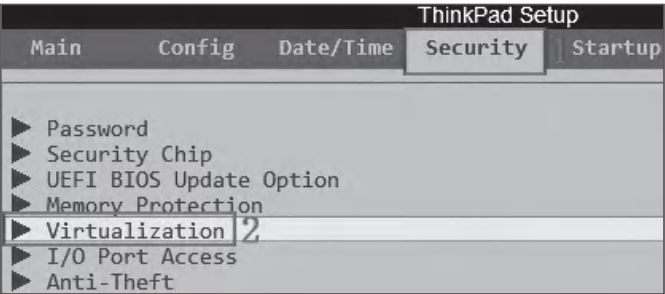


图 1.9 “Security” 菜单窗口

在“Virtualization”设置列表（见图 1.10）中，找到“Intel(R) Virtualization Technology”（或“AMD-V Technology”“AMD SVM Support”）并按回车键。

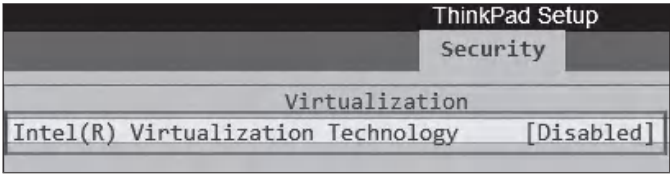


图 1.10 “Virtualization” 设置窗口

对于联想台式机计算机，进入 BIOS Setup 设置界面（见图 1.11）后，依次选择“Advanced”→“Intel(R) Virtualization Technology”（或“AMD-V Technology”，或“AMD SVM Support”）选项并按回车键。

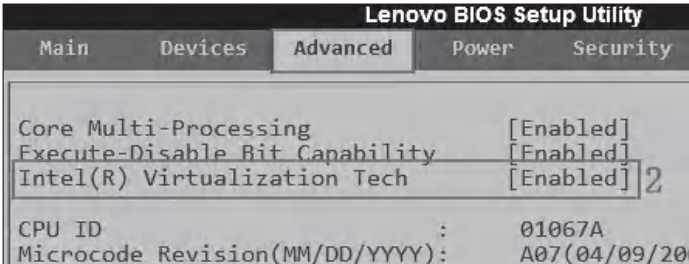


图 1.11 “Advanced” 菜单窗口



注意：

部分版本可能需要在“Advanced”菜单的“CPU Setup”选项下才能找到。

以上三种机型，在按下回车键后，在弹出的窗口（见图 1.12）中，使用上下方向键移动光标，选择“Enabled”选项后按回车键，“Enabled”表示启用，“Disabled”表示禁用。

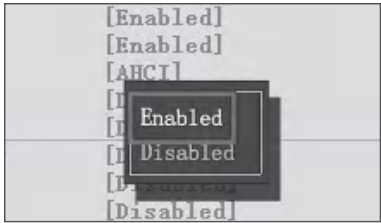


图 1.12 “Enabled” 选项窗口

最后按下键盘上的“F10”键，或部分机型可能需要按“Fn+F10”组合键，调出保存对话框（见图 1.13），使用方向键移动光标，选择“Yes”并按回车键，计算机会保存刚才的 BIOS 设置并退出 BIOS 设置，随后计算机会自动重启。

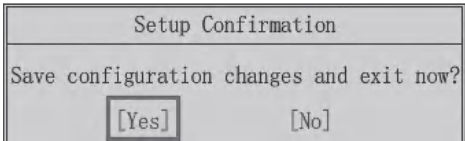


图 1.13 保存设置并退出的确认窗口



1.3 虚拟机的安装

在 VMWare 中安装三台 Linux CentOS 虚拟机，组成一个 Linux 集群。
首先开始的是集群中 node1 节点虚拟机的安装，打开安装好的 VMware 软件，在主窗口（见图 1.14）中，双击窗口中的“创建新的虚拟机”图标。

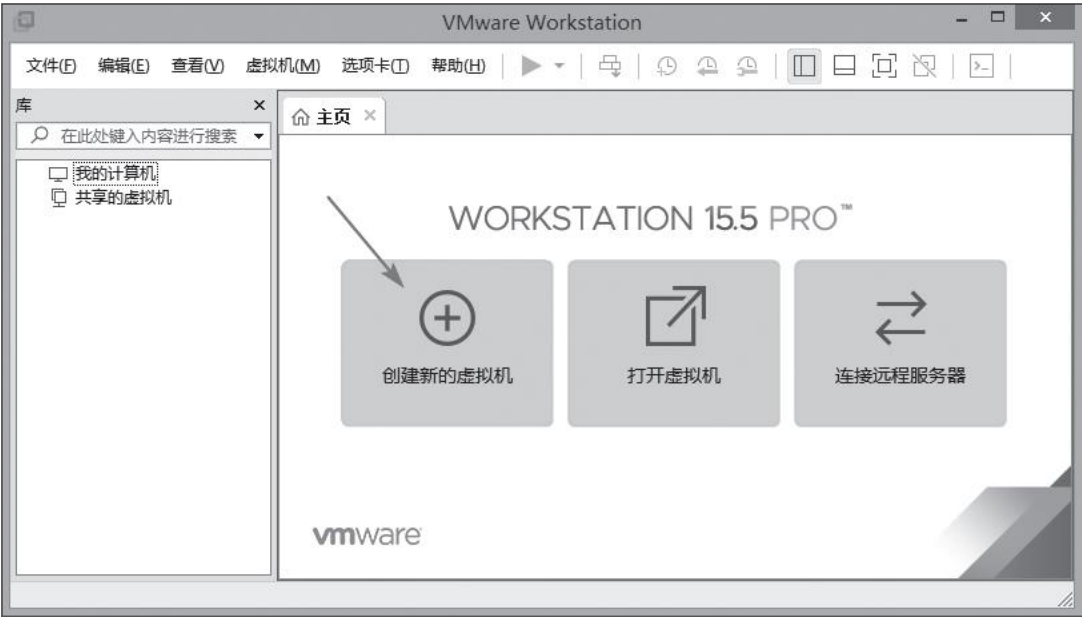


图 1.14 VMware 软件的主窗口

在弹出的“新建虚拟机向导”窗口（见图 1.15）中，默认选择“自定义（高级）”选项，再单击“下一步”按钮。

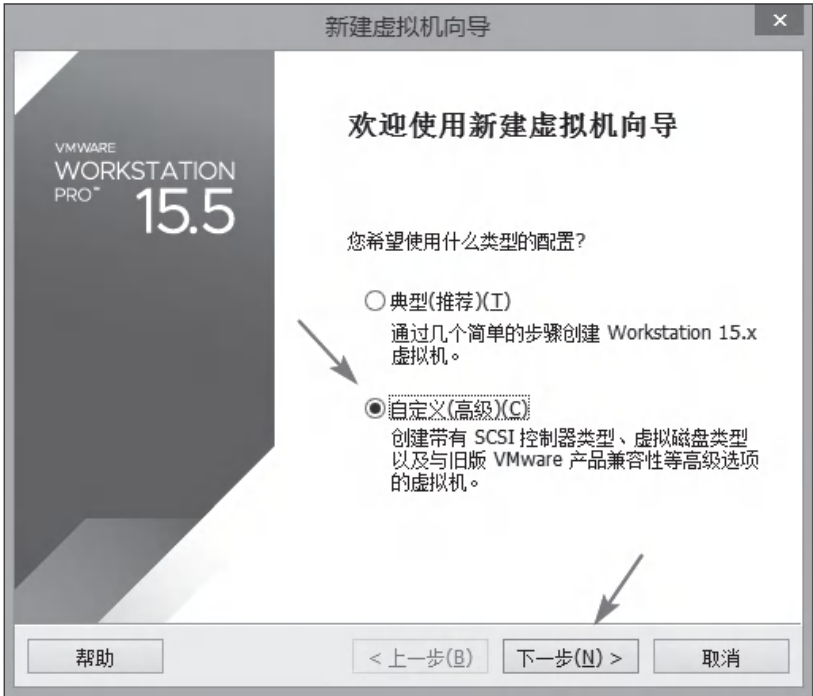


图 1.15 “新建虚拟机向导”窗口

在接下来弹出的“选择虚拟机硬件兼容性”窗口（见图 1.16）中，从下拉列表中选择“Workstation 15.x”选项，单击“下一步”按钮。



图 1.16 “选择虚拟机硬件兼容性”窗口

在接下来弹出的“安装客户机操作系统”窗口（见图 1.17）中，选择“安装程序光盘映像文件（iso）（M）”选项，然后单击“浏览”按钮，找到“CentOS-7-x86_64-Minimal-1810.iso”文件，将其加载到输入框中，然后单击“下一步”按钮。



图 1.17 “安装客户机操作系统”窗口

在接下来弹出的“命名虚拟机”窗口（见图 1.18）中，在“虚拟机名称”输入框中为虚拟机填写名称，然后单击“浏览”按钮，为新建虚拟机指定安装位置，建议安装路径上不要有中文目录，并将虚拟机的安装位置指定到固态硬盘，便于虚拟机运行时，可



提高读写操作的响应速度。然后单击“下一步”按钮。

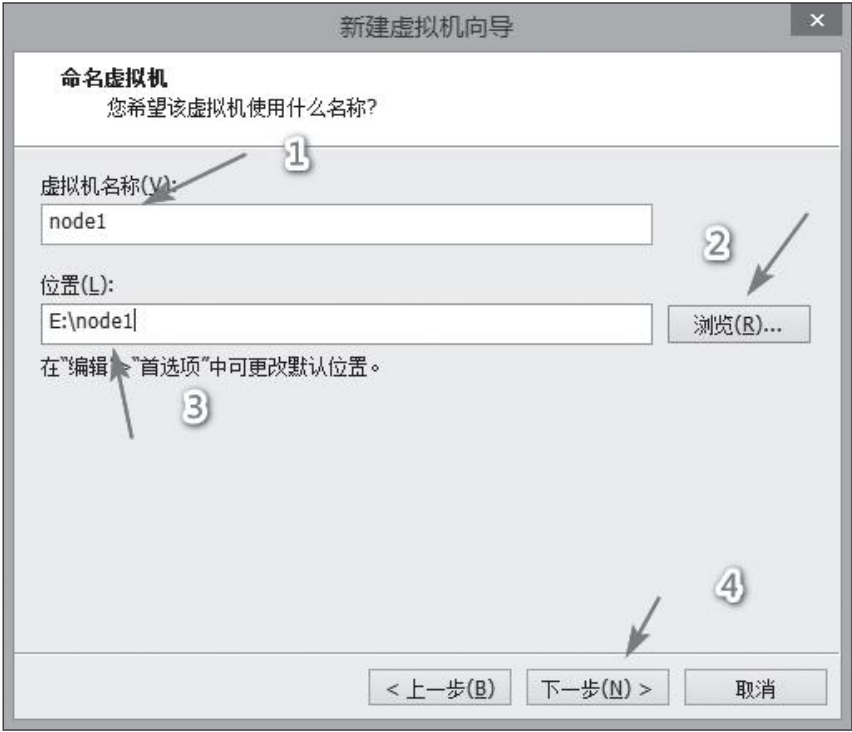


图 1.18 “命名虚拟机”窗口

在接下来弹出的“处理器配置”窗口（见图 1.19）中，从下拉列表中，分别为虚拟机的 CPU 指定个数以及为 CPU 内核指定个数，建议各设置为 1 个，如果设置为多个，则需要占用更多的内存资源。然后单击“下一步”按钮。



图 1.19 “处理器配置”窗口

在接下来弹出的“此虚拟机的内存”窗口（见图 1.20）中，拖动窗口左边的游标，为虚拟机指定内存大小，建议设置为 2GB 大小，然后单击“下一步”按钮。



笔记

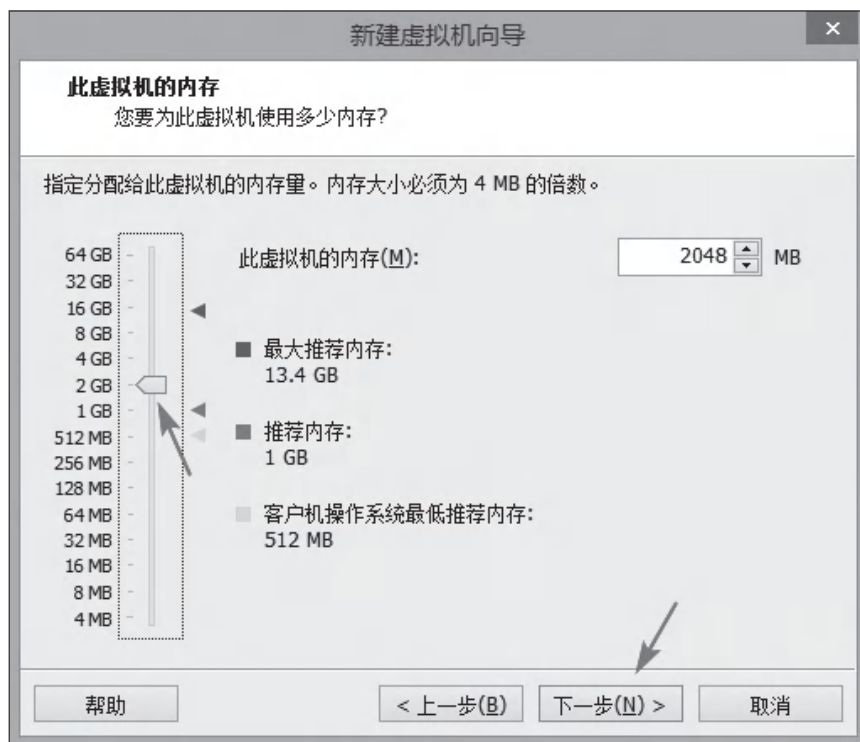


图 1.20 “此虚拟机的内存”窗口

在接下来弹出的“网络类型”窗口（见图 1.21）中，默认选择“使用网络地址转换 (NAT)(E)”选项，然后单击“下一步”按钮。

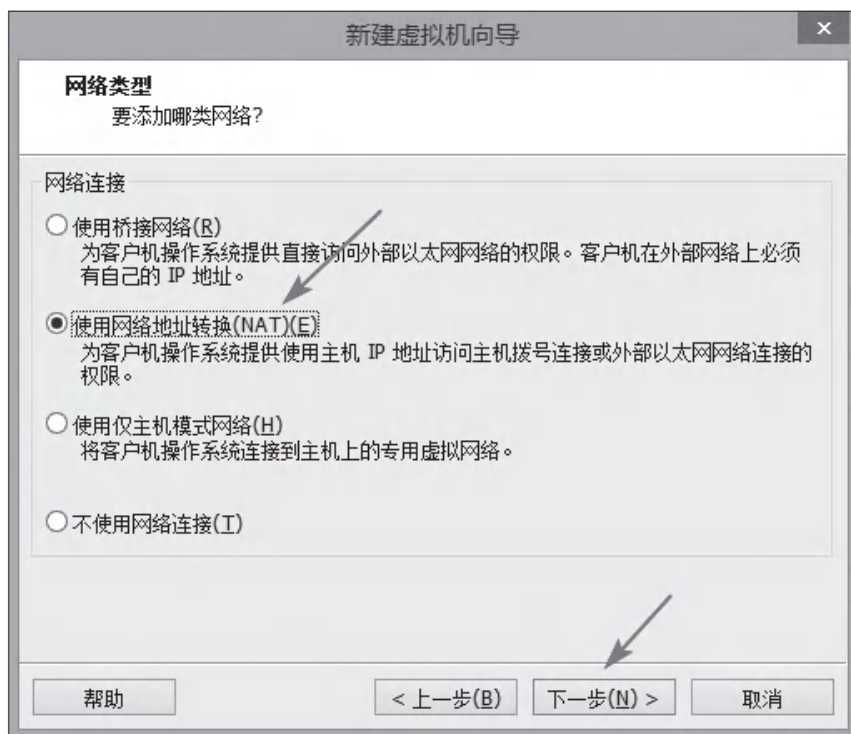


图 1.21 “网络类型”窗口



笔记

在接下来弹出的“选择 I/O 控制器类型”窗口（见图 1.22）中，默认选择“LSI Logic(L)”选项，然后单击“下一步”按钮。

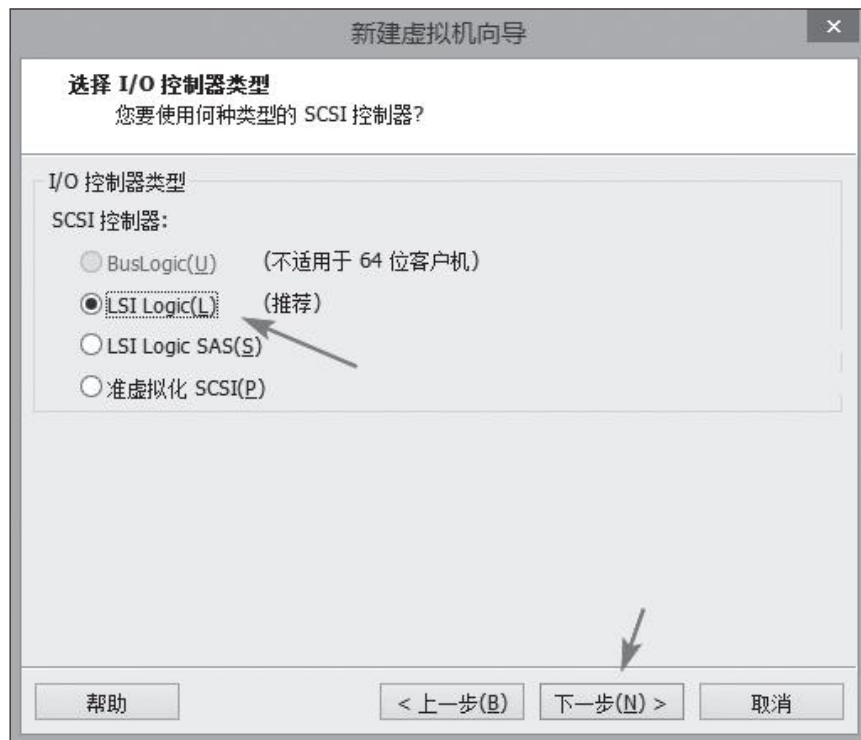


图 1.22 “选择 I/O 控制器类型”窗口

在接下来弹出的“选择磁盘类型”窗口（见图 1.23）中，默认选择“SCSI(S)”选项，然后单击“下一步”按钮。



图 1.23 “选择磁盘类型”窗口

在接下来弹出的“选择磁盘”窗口（见图 1.24）中，默认选择“创建新虚拟磁盘(V)”选项，然后单击“下一步”按钮。



笔记

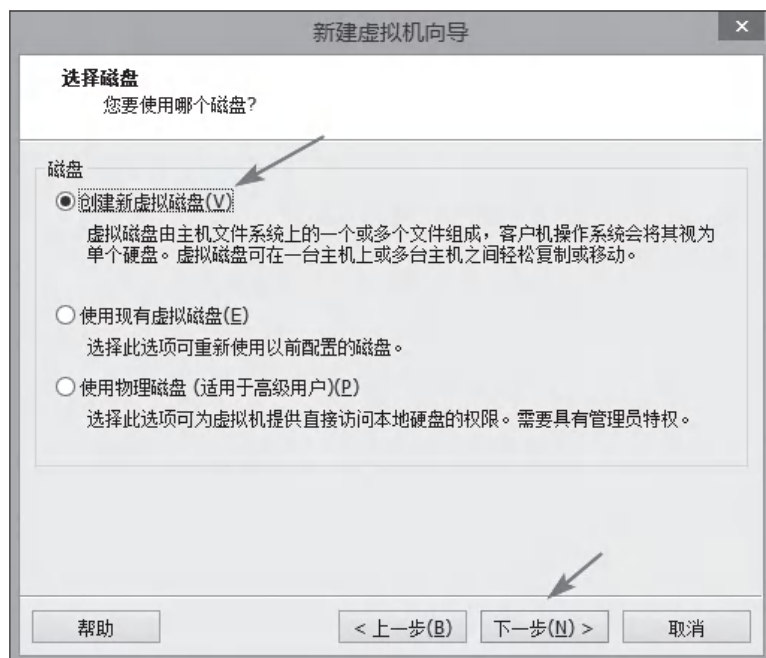


图 1.24 “选择磁盘”窗口

在接下来弹出的“指定磁盘容量”窗口（见图 1.25）中，在输入框中填写 30.0，将 node1 节点虚拟机的磁盘空间设置为 30G（因为 node1 节点将作为 Hadoop 集群的主节点，下载和安装的文件相对于集群中的其他节点较多，所以磁盘容量相对其他节点要大，才能满足工作要求），然后默认选择“将虚拟磁盘拆分成多个文件”选项，单击“下一步”按钮。

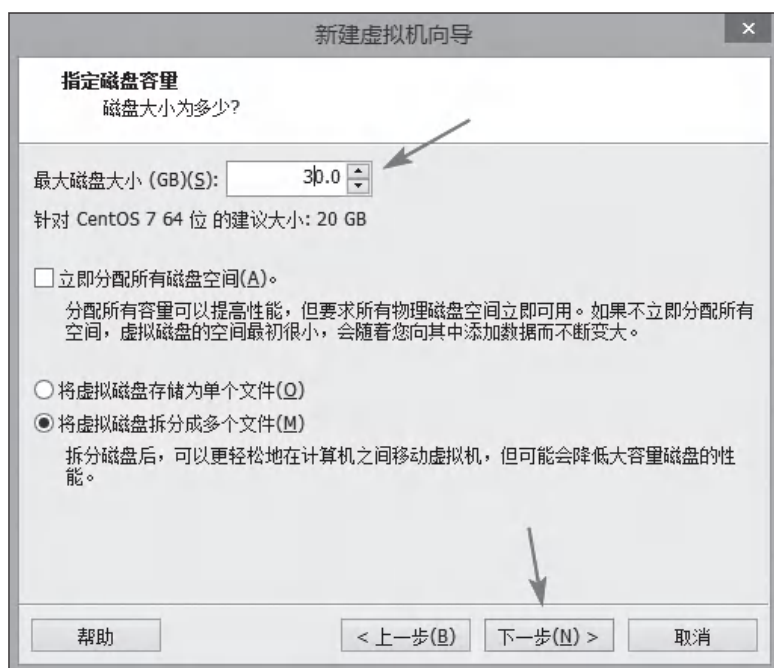


图 1.25 “指定磁盘容量”窗口



笔记

在接下来弹出的“指定磁盘文件”窗口（见图 1.26）中，确认输入框内给出的默认名称，不用修改，直接单击“下一步”按钮。



图 1.26 “指定磁盘文件”窗口

在接下来弹出的“已准备好创建虚拟机”窗口（见图 1.27）中，直接单击“完成”按钮，至此，虚拟机的安装已完成。接下来，VMware 软件的主窗口会自动进入 Linux CentOS 系统的安装界面。

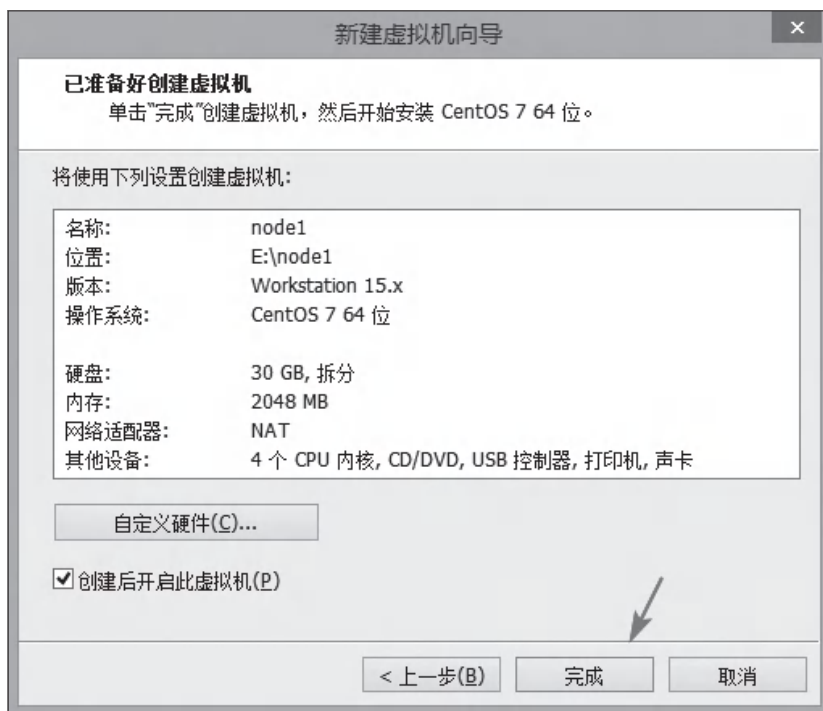


图 1.27 “已准备好创建虚拟机”窗口

1.4 CentOS 7 的安装



接下来，为 node1 节点虚拟机安装 CentOS 7 系统。如图 1.28 所示，此时系统默认选择安装界面中的第一个选项 “Install CentOS 7”，可以将鼠标指向安装界面，然后单击鼠标左键，让光标切入安装界面中（如想让光标由虚拟机切回到 Windows 系统，可以按下 “Ctrl+Alt” 组合键），通过键盘上的上下方向键，可以在安装界面的选项之间进行切换，选择第一个选项后，按回车键。

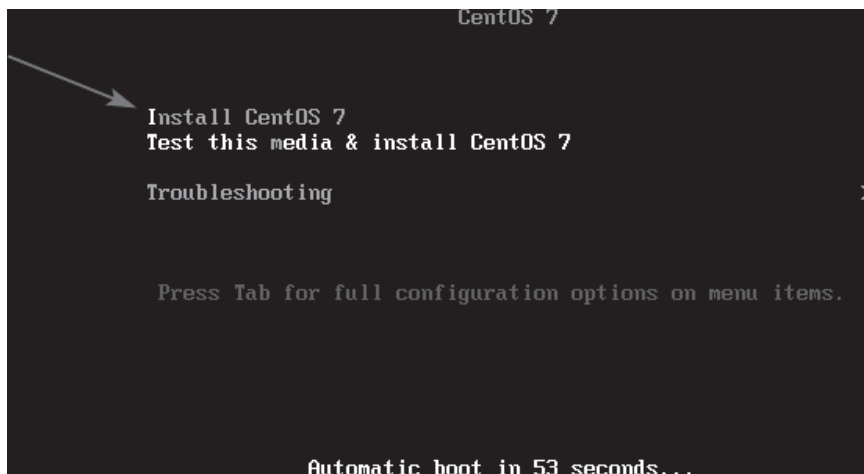


图 1.28 CentOS 系统的安装界面

接下来，CentOS 系统会进入安装检测阶段（见图 1.29），项目前的绿色 “OK” 字母，表示检测通过，每一项应该都被检测通过，否则会给后续的系统安装带来麻烦。

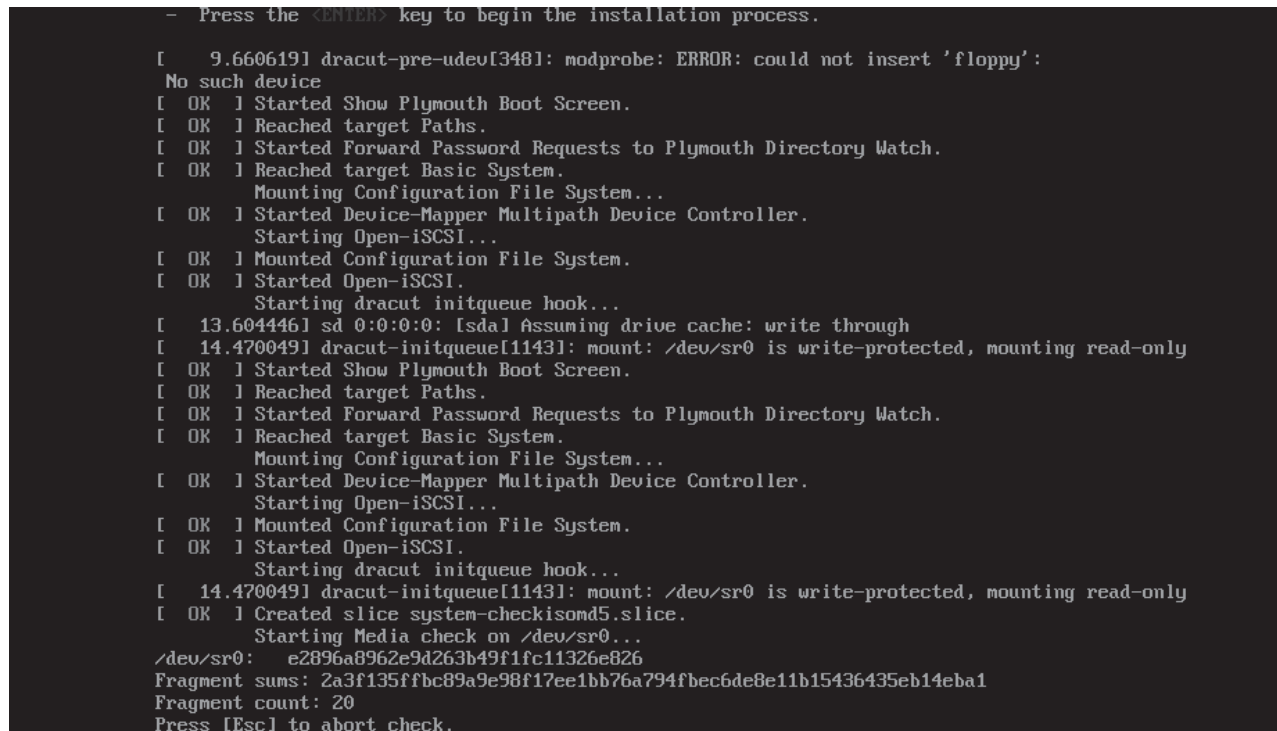


图 1.29 CentOS 系统的安装检测阶段

笔记

安装检测完成后，系统进入语言环境配置阶段（见图 1.30），在输入框中输入“ch”，然后选择左边文本框中的“中文”选项（注意：这个只会应用在当前系统安装阶段的中文显示，当系统安装好后，系统的运行，仍然是英文命令行状态），右边的文本框默认选择“简体中文（中国）”，如果没有，需要手动选择，确认无误后，单击右下角的“继续”按钮，进行下一步配置。



图 1.30 语言环境配置

接下来，在弹出的系统配置界面（见图 1.31）中，单击“安装位置”图标。



图 1.31 系统配置界面

接下来，在弹出的“设备选择”界面（见图 1.32）中，首先选择下方的“我要配置分区(I)”选项，然后单击左上角的“完成”按钮。



图 1.32 设备选择界面

单击图 1.32 中的“完成”按钮后，安装程序会弹出“手动分区”配置窗口（见图 1.33），在该窗口的左下角，单击“+”号按钮，进行手动分区配置。



图 1.33 “手动分区”配置窗口 1

接下来，在弹出的“添加新挂载点”窗口（见图 1.34）中，首先从下拉列表中，选择“/”选项，此选项是 Linux 系统的根目录。

笔记

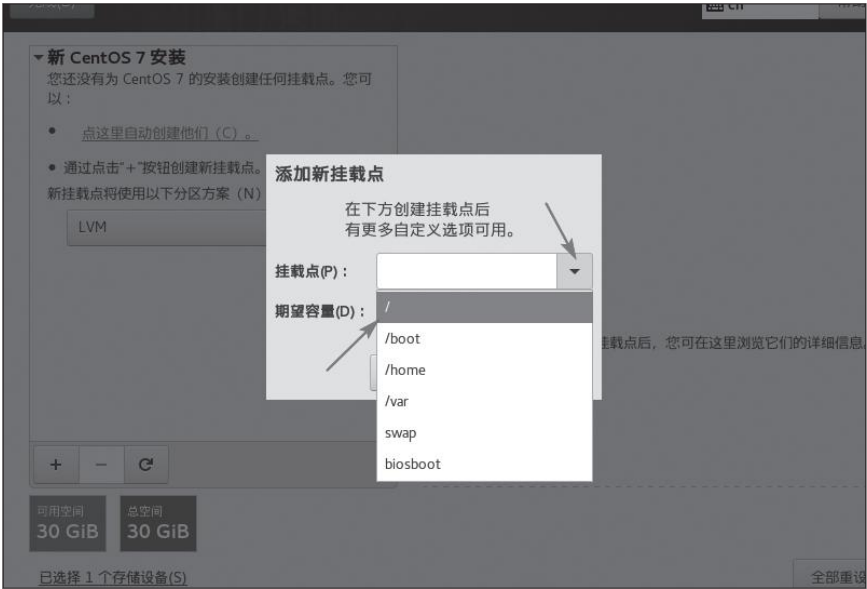


图 1.34 “添加新挂载点”窗口

选择“/”选项后，在“期望容量”的输入框（见图 1.35）中，填写 20GB，注意字母大小写的区分，填写后，单击“添加挂载点（A）”按钮。

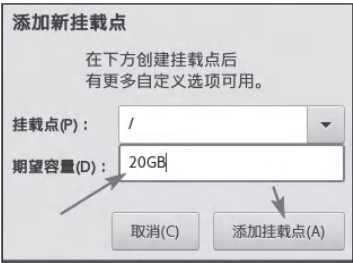


图 1.35 填写“期望容量”

如图 1.36 所示，继续在“手动分区”配置窗口的左下角单击“+”号按钮。



图 1.36 “手动分区”配置窗口 2

在接下来的“添加新挂载点”窗口（见图 1.37）中，从下拉列表中，选择“/boot”选项，此选项是 Linux 系统的系统启动目录，并在“期望容量 (D)”输入框中填写 300mb，注意字母大小写的区分，填写后，单击“添加挂载点”按钮。

 笔记



图 1.37 配置系统启动目录新挂载点

如图 1.38 所示，继续在“手动分区”配置窗口的左下角单击“+”号按钮。



图 1.38 “手动分区”配置窗口 3

在接下来的“添加新挂载点”窗口（见图 1.39）中，从下拉列表中，选择“swap”选项，此选项是 Linux 系统的虚拟内存交换区，并在输入框填写 3.5GB，注意字母大小写的填写，填写后，单击“添加挂载点”按钮。

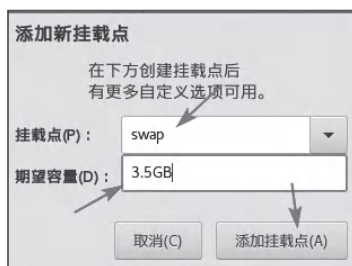


图 1.39 配置系统虚拟内存交换区新挂载点



添加完三个系统挂载点后，单击“手动分区”配置窗口上方的“完成”按钮。如图 1.40 所示。



图 1.40 “手动分区”配置窗口 4

在接下来的“更改摘要”窗口（见图 1.41）中，单击窗口右下角的“接受更改”按钮，返回安装主界面窗口。

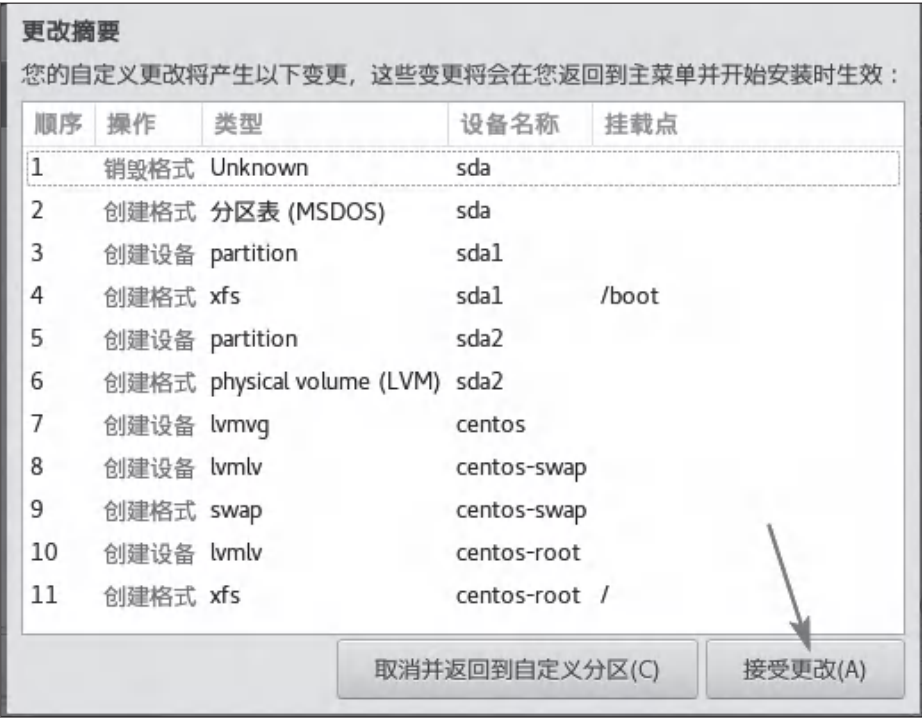


图 1.41 “更改摘要”窗口

在接下来的安装主界面窗口（见图 1.42），单击窗口右下角的“开始安装”按钮，开始 CentOS 系统的安装。



图 1.42 安装主界面窗口

在接下来的“用户设置”窗口（见图 1.43）中，单击界面中的“ROOT 密码”图标，进入管理员密码设置窗口。



图 1.43 单击“ROOT 密码”设置图标



在接下来的密码设置输入框（见图 1.44）中，输入密码，两次均是小写的 root（输入时，注意键盘的大小写切换，以避免无意中输入的是大写字母），然后单击左上角的“完成”按钮，因为密码长度不够，所以“完成”按钮需要单击两次。



图 1.44 管理员密码设置窗口

CentOS 系统安装完成后，窗口（见图 1.45）中会显示“完成！”字样，此时，单击窗口右下角的“重启”按钮。



图 1.45 系统完成安装界面

等待系统重启好后，系统进入命令行的登录界面，如图 1.46 所示。在“localhost login:”处，输入管理员账号“root”，并按回车键；在“Password:”处，输入在前面密码设置阶段中配置的管理员密码“root”，此时，出于 Linux 系统的安全设计，不会显示出你输入的密码，等输入完密码后，按回车键。



笔记

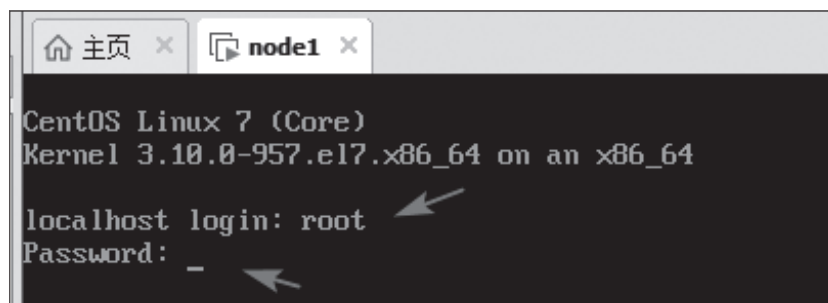


图 1.46 登录界面

系统成功登录后，窗口（见图 1.47）中会显示“[root@localhost ~]#”，表示已进入 CentOS 系统，等待输入新的指令。

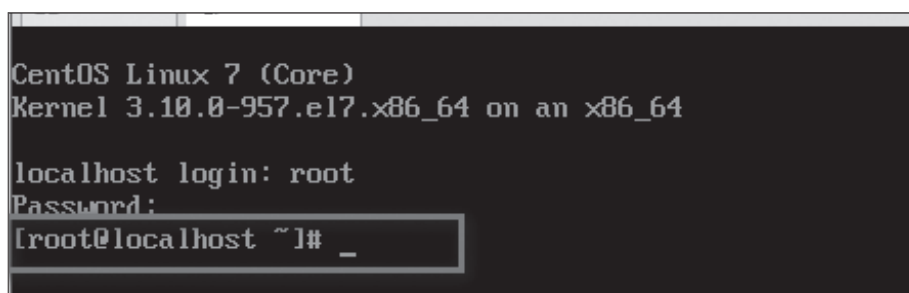


图 1.47 成功登录后的系统状态

知识拓展

对“[root@localhost ~]#”的解读

root：指当前登录的账号，当前登录用的是系统管理员账号，因此显示的是 root；
localhost：指当前的主机名称，系统安装时，没有配置主机名，因此显示的是默认主机名；
~：指当前登录账号的主目录，因为登录用的是系统管理员账号，~是管理员主目录符号；
#：命令标识符，因为登录用的是系统管理员账号，#是管理员默认命令标识符，普通用户的命令标识符是\$。

1.5 node2 节点虚拟机的配置

node2 节点虚拟机的配置，与前面 node1 节点虚拟机的配置有以下三点不同：



(1) 虚拟机名称不同，如图 1.48 所示。但注意，图 1.48 中所列位置信息只是格式说明，不能作为实际路径。为便于管理，三台虚拟机的位置最好在同一父目录下。



图 1.48 node2 节点虚拟机的命名

(2) 磁盘空间可以设为 20GB，因为 node2 节点在将来的 Hadoop 集群中是备用主节点角色，因此磁盘空间相比 node1 节点空间要小。如图 1.49 所示。

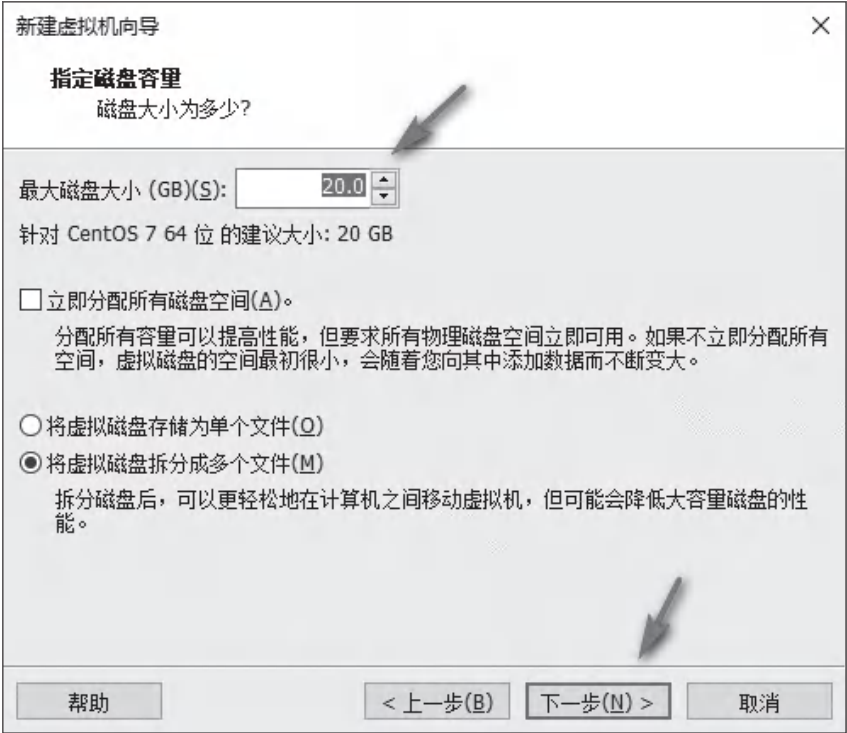


图 1.49 node2 节点虚拟机的磁盘空间配置

(3) 因为磁盘空间变小, 所以 “/” 根目录的空间大小可以设为 10GB, 如图 1.50 所示。



图 1.50 node2 节点虚拟机的 “/” 根目录配置

除以上三点区别之外, node2 节点的安装步骤与前面 node1 节点的安装步骤完全一致, 因此不再详细列举 node2 节点虚拟机的安装过程。

1.6 克隆 node2 节点, 完成 node3 节点虚拟机的安装

在开始克隆前, 需将 node2 节点虚拟机关闭, 首先在 VMWare 的左侧窗口中选择 “node2” 节点虚拟机的名称选项, 如图 1.51 所示。

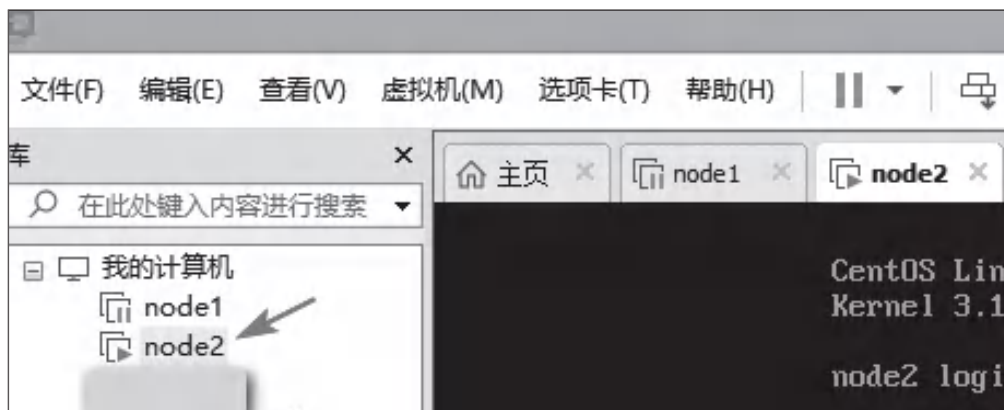


图 1.51 选择 “node2” 虚拟机



然后在 VMWare 的菜单栏（见图 1.52）单击菜单列表中的“关闭客户机（D）”菜单项。

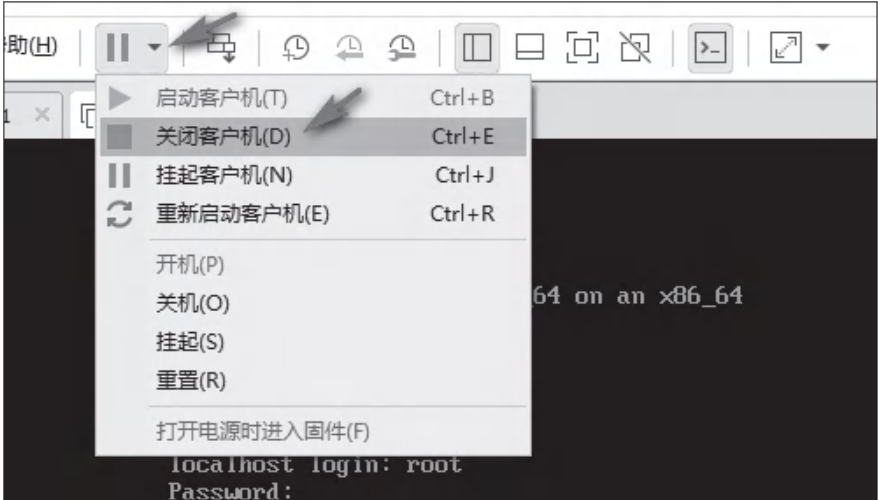


图 1.52 关闭“node2”虚拟机

接下来，继续在 VMWare 的左侧窗口（见图 1.53）中选择 node2 节点虚拟机的名称选项，单击鼠标右键，在快捷菜单中，依次单击“管理（M）”→“克隆”菜单项。

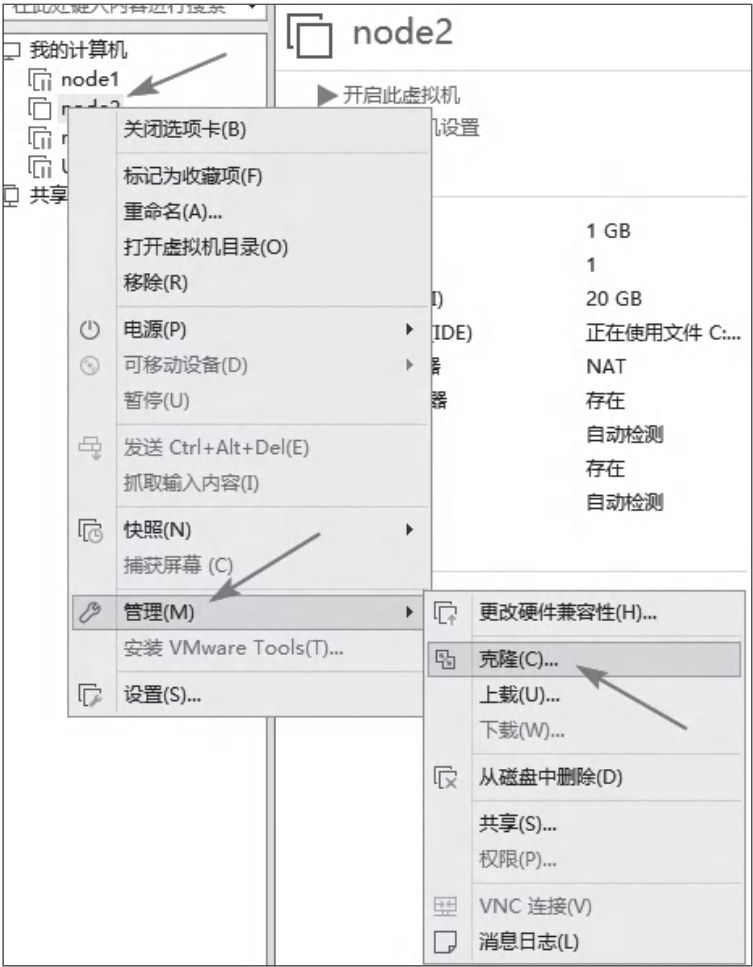


图 1.53 单击“克隆”菜单项

接下来，在弹出的“克隆虚拟机向导”窗口（见图 1.54）中，单击“下一步”按钮。

笔记



图 1.54 “克隆虚拟机向导”窗口

接下来，在弹出的“克隆源”窗口（见图 1.55）中，选择“虚拟机中的当前状态 (C)”选项，单击“下一步”按钮。



图 1.55 选择“克隆源”



笔记

接下来，在弹出的“克隆类型”窗口（见图 1.56）中，选择“创建完整克隆（F）”选项，单击“下一步”按钮。

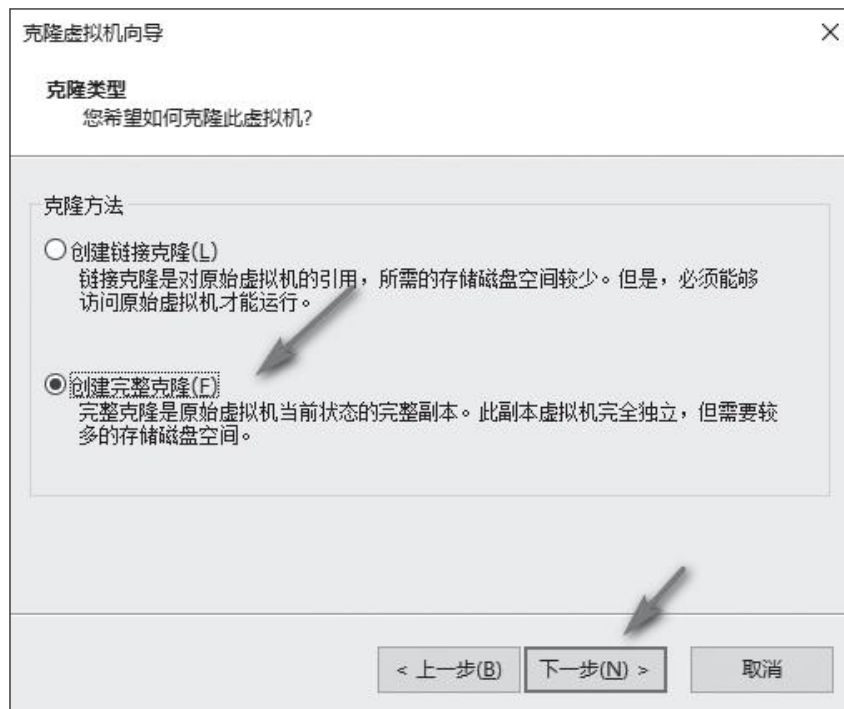


图 1.56 选择“克隆类型”

接下来，在弹出的“新虚拟机名称”窗口（见图 1.57）中，在“虚拟机名称”的输入框中为 node3 节点虚拟机命名，并单击“浏览”按钮，将 node3 节点虚拟机的位置指定到 node1、node2 节点虚拟机所在目录，最后单击“完成”按钮。



图 1.57 为新虚拟机命名

接下来，VMWare 会立刻进行克隆任务，并弹出如图 1.58 所示界面，提示正在克隆。

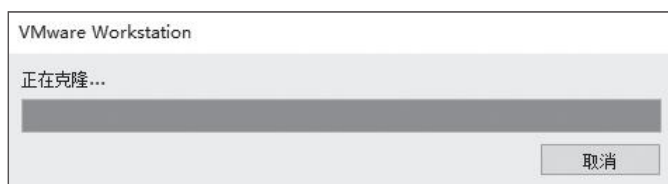


图 1.58 克隆正在进行

克隆任务完成后，会弹出如图 1.59 所示界面，单击“关闭”按钮，退出。



图 1.59 克隆完成

至此，完成 node3 节点虚拟机的安装。

拓展练习

参照本章讲解的内容，搭建一个至少包含三个节点的 Linux 集群，要求节点名称以自己姓名的汉语拼音首字母缩写 + 编号命名，例如，学生姓名“张三”，则三个节点的名称应分别是“zs01”“zs02”“zs03”。