

Table of contents in chinese  
(courtesy of Junlin Yao)

Book web page: <https://www.lix.polytechnique.fr/~nielsen/HPC4DS/>

# 目录

## I. 通过消息传递接口 (MPI) 的高性能计算 (HPC)

### 1. 高性能计算 (HPC)

1. 什么是高性能计算?
2. 为什么使用高性能计算?
3. 大数据: 大数据的四个 V
4. 并行计算的范式
5. 并行粒度: 细粒度与粗粒度
6. HPC 的架构: 存储与网络
7. 加速 (speed-up)
  1. 加速比、效率与可扩展性
  2. Amdahl 定律 (数据量恒定)
  3. Gustafson 定律: 数据量与处理器数成正比 (scale speed-up)
  4. 可扩展性与等效率
  5. 单机上的并行计算机模拟
  6. 大数据与并行文件系统
8. 关于 HPC 的 8 个误解
9. \*拓展: 备注、补充读物和讨论
10. 小结: 需要记住的要点
11. 习题

### 2. MPI 接口的介绍: 消息传递接口

1. 用于并行计算的 MPI 接口: 基于消息的通讯
2. 并行计算的模型, 线程与进程
3. 全局通讯
  1. 四个基本全局运算
  2. 阻塞与非阻塞点对点通讯
  3. 死锁 (deadlocks)
  4. 几个竞争假设: 本地计算与包封通讯
  5. 单向与双向通讯
  6. MPI 上的全局计算: 归纳计算 (reduce) 与并行前缀 (和) (scan)
  7. 通讯组: 通信机
4. 同步的阻碍: 进程的集合点
  1. 一个 MPI 下同步的测量执行时间的例子
  2. BSP 模型: 整体同步并行计算模型
5. MPI 的应用程序编程接口 (API)
  1. 基于 C++ 的 MPI 应用程序编程接口的“Hello World”
  2. 基于 C 的 MPI 编程
  3. \*基于 C++ 的使用 Boost 接口的 MPI
6. \*通过 OpenMP 使用 MPI
7. MPI 的主要通讯操作的句法

## Introduction to HPC with MPI for Data Science, ISBN 978-3-319-21902-8

1. MPI广播(broadcast)、散发(scatter)、收集(gather)、归约(reduce)和全局归约(Allreduce)
2. 其它通讯操作 / 并不常见的全局计算
8. MPI环的通讯
9. 任务调度 SLURM
10. 基于MPI的并行计算的若干例子及其加速
  1. 矩阵-向量乘法
  2. 通过蒙特卡洛模拟逼近
  3. 通过随机积分求得分子体积
11. \*拓展: 备注、补充读物和讨论
12. 小结: 需要记住的要点
13. 习题

### 3. 互连网络拓扑

14. 静态 / 动态网络与逻辑 / 物理网络
15. 互连网络: 基于图的建模
16. 拓扑的特征
  1. 图的度与直径
  2. 连通性与二等分
  3. 一个好的拓扑网络的标准
17. 常用拓扑: 简单静态网络
  1. 完全子图 (clique) : K完全图
  2. 星、环与弦
  3. 网格与环面
  4. 3D立方与立方体连接环 (Cube-connected cycles)
  5. 树与胖树
18. 超方形拓扑与格雷码 (Gray Code)
  1. 超方形的迭代创建
  2. 通过格雷码的结点编号
  3. C++上的格雷码生成
  4. 格雷码与二进制码的互相转换
  5. \*图的笛卡尔积 (算子)
19. 拓扑上的通讯算法
  1. 环上通讯
  2. 超方形的广播算法: 树状通讯
20. 将拓扑嵌入其它拓扑
21. \*复杂正则拓扑
22. 集成电路片的互连网络
23. \*拓展: 备注、补充读物和讨论
24. 小结: 需要记住的要点
25. 习题

### 4. 并行排序

26. 循序排序: 要点回顾
  1. 主要排序算法的简要回顾
  2. 排序算法的复杂度
27. 并行归并排序: MergeSort
28. 并行排列排序: RankSort
29. 并行快速排序: ParallelQuickSort
30. 改善的算法: HyperQuickSort

- 31. \* 并行正则采样排序 (PSRS)
- 32. \* 2D 网格排序: ShearSort
- 33. 比较排序网络
- 34. \* 通过比较器电路排序后的列合并
- 35. \* 迭代双调排序
- 36. \* 拓展: 备注、补充读物和讨论
- 37. 小结: 需要记住的要点
- 38. 习题

## 5. 并行线性代数运算

- 39. 分布式线性代数
  - 1. 用于大数据的线性代数
  - 2. 经典线性代数
  - 3. 矩阵-向量乘法:  $y = Ax$
  - 4. 矩阵数据并行化的动机
- 40. 环拓扑上的矩阵-向量乘法
- 41. 网格上的矩阵乘法 (外积算法)
- 42. 环面拓扑上的矩阵乘法
  - 1. Cannon 算法
  - 2. Fox 算法
  - 3. Snyder 算法
  - 4. 环面拓扑上三种矩阵乘法的比较
- 43. \* 拓展: 备注、补充读物和讨论
- 44. 小结: 需要记住的要点
- 45. 习题

## 6. 映射归纳 (MapReduce) 计算的模型

- 46. 迅速处理大数据的挑战
- 47. MapReduce 的基本原理
  - 1. 映射过程 mappers 与归纳过程 reducers
  - 2. \* 函数式编程中的 map 和 reduce 函数
- 48. MapReduce 的配型与元算法
- 49. C++ 中 MapReduce 程序的完整例子
- 50. MapReduce 的执行模型与架构
- 51. \* 拓展: 备注、补充读物和讨论
- 52. 小结: 需要记住的要点
- 53. 习题

## I. 基于 MPI 的数据科学

### 7. K-平均算法: 划分聚类

- 54. 通过聚类的初步研究
  - 1. 划分聚类
  - 2. 聚类的代价与基于模型的聚类
- 55. K-平均的代价函数
  - 1. 代价函数的另一种形式: 聚类或分离数据
  - 2. 可计算度: K-平均的计算复杂度
- 56. K-平均的 Lloyd 局部启发法
- 57. K-平均的初始化
  - 1. 随机初始化 (Forgy)
  - 2. 全局 K-平均初始化

3. K-平均++的有概率保证的初始化
58. 向量的量化与K-平均
  1. 量化
  2. \*Lloyd 局部极小值生成 Voronoï 划分
59. K-平均的物理意义：惯性的分解
60. 分类数目 k 的选择：模型选择
  1. 肘部法则
  2. 能被 k 解释的方差比例
61. 用于大数据的计算机群上的 K-平均
62. 聚类方法划分结果的评估与比较
  1. Rand 指数
  2. 标准互信息 (NMI)
63. \*拓展：备注、补充读物和讨论
64. 小结：需要记住的要点
65. 习题

## 7. 层次聚类

66. 升序与降序层次聚类：树状图
67. 定义一个适合的链接距离的策略
  1. 凝聚层次聚类算法
  2. 基础距离的选择
68. Ward 合并准则与重心
69. 基于树状图的划分
70. 超度量距离与演化树
71. \*拓展：备注、补充读物和讨论
72. 小结：需要记住的要点
73. 习题

## 8. 通过 K-近邻算法的监督聚类

74. 监督学习
75. 近邻规则 (PPV)
  1. 欧式距离下的近邻计算优化
  2. PPV 规则与沃罗诺伊图 (Voronoi Diagram)
  3. K-近邻算法的规则
76. 分类器性能的评估
  1. 分类错误率 (misclassification)
  2. 混淆矩阵和假阳性 / 阴性
  3. 精确率、召回率和 F-值
77. 统计学习与贝叶斯最小误差
  1. 概率密度的非参估计
  2. 最小误差：误差的概率与贝叶斯误差
  3. PPV 的概率误差
78. 基于分布式存储的并行架构上的 PPV
79. \*拓展：备注、补充读物和讨论
80. 小结：需要记住的要点
81. 习题

## 9. 通过核心集 (coresets) 与降维的优化

82. 庞大数据下的优化
  1. 一个需要处理高维数据的例子
  2. 高维空间关于距离的现象
  3. 从大数据到小数据

83. 核心集的定义 (coresets)
84. 最小封闭球下的核心集
85. 快速逼近最小封闭球的启发法
  1. 收敛证明
  2. 近似封闭球与近似线性分离器
86. \* K-平均的核心集
87. 数据的快速降维
  1. 维度的诅咒
  2. 两个处理高维数据的详细例子
  3. 线性降维
  4. Johnson-Lindenstrauss 定理
  5. 随机映射矩阵
88. \* 拓展: 备注、补充读物和讨论
89. 小结: 需要记住的要点
90. 习题

## 10. 图的并行算法

91. (大) 图中的稠密图检测
  1. 问题定义
  2. 问题的复杂度与贪婪启发法
  3. 易于并行化的启发法
92. 测试 (小) 图的同构
  1. 枚举算法的一般原理
  2. 测试同构的 Ullmann 算法
  3. 并行枚举算法
93. \* 拓展: 备注、补充读物和讨论
94. 小结: 需要记住的要点
95. 习题

## II. C / C++ / Shell

### A. 在 java 的基础上学习 C

1. 头文件 .h, 源文件 .c, 宏与预处理器
96. 内存分配与表的销毁
97. 使用 struct 构造结构体
98. 函数申明
99. C 的其它特点
100. 通过冒泡排序法阐释输入输出

### B. 在 C 的基础上学习 C++

1. 关于函数的回顾: 值传递
101. 类与对象
  1. 类的定义
  2. 继承与类的层次
102. 方法的关键词 const
103. 表的创建与销毁
104. 运算符的重载
105. C++ 的派生
106. STL 标准库
107. C++ 的输入输出
108. 用于矩阵 (ublas) 的 Boost 库
109. C++ 的其它特点

**C. 使用 shell 指令控制进程与输入 / 输出**

- 1. 初始配置文件 .bashrc
- 110. Unix 管道 (pipeline) 命令与输入 / 输出重导向
- 111. 任务处理

**D. 机房电脑表 (Liste des ordinateurs en salle machines)**