

分布计算系统

徐 辰

cxu@dase.ecnu.edu.cn

華東師範大學



课程名称

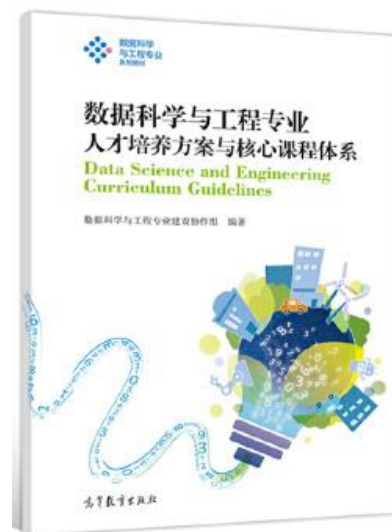
2

□ 研究生

- 2018、2019年：大数据处理系统
- 2020年：大规模数据处理系统
- 2021年：分布式计算系统

□ 本科生

- 2018、2019年：分布式模型与编程
- 2021、2022年：分布式编程模型与系统
- 2023年：分布式计算系统



课程背景

3

□ 大数据处理系统 → 分布式计算系统

- ✚ Hadoop、Spark、Flink等

- ✚ “大数据”的涵义过于宽泛

□ 其它类似课程/教材

- ✚ 英文论文的翻译：不同论文的体系可能不一致

- ✚ 针对某一系统的工具手册：时效性

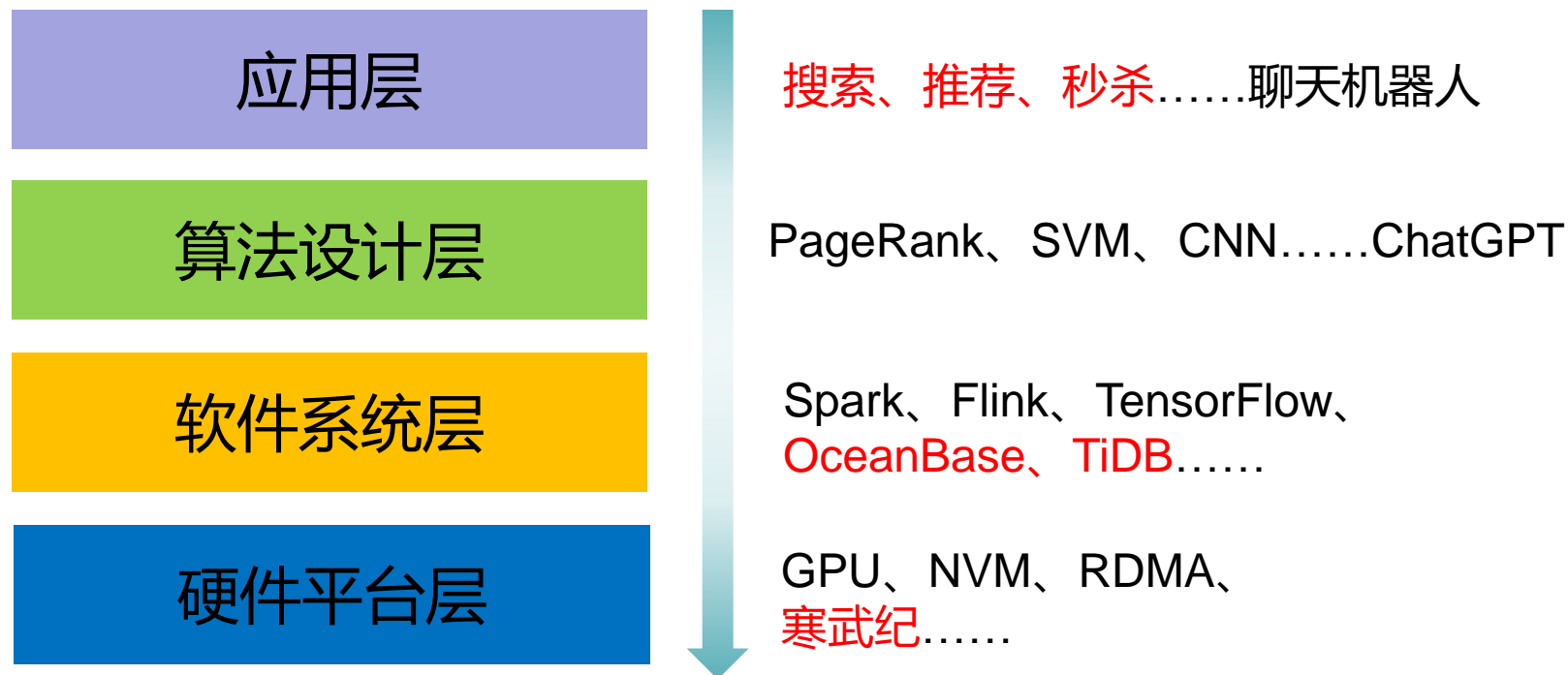
□ 本课程/教材

- ✚ 强调系统设计、原理、编程的结合

课程目的

4

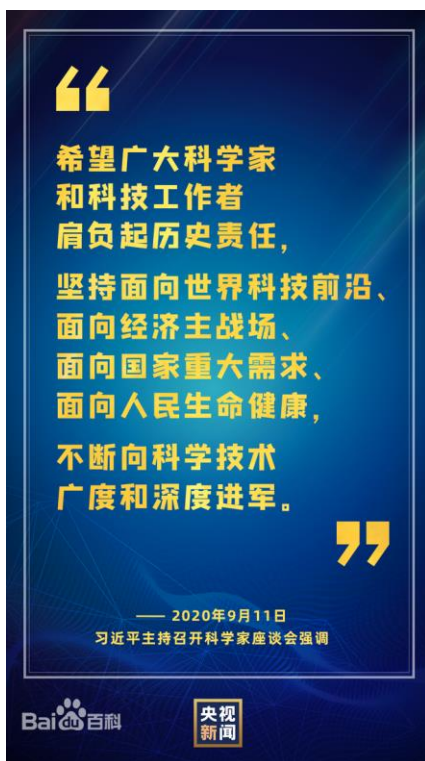
□ 培养“系统思维”



课程目的

5

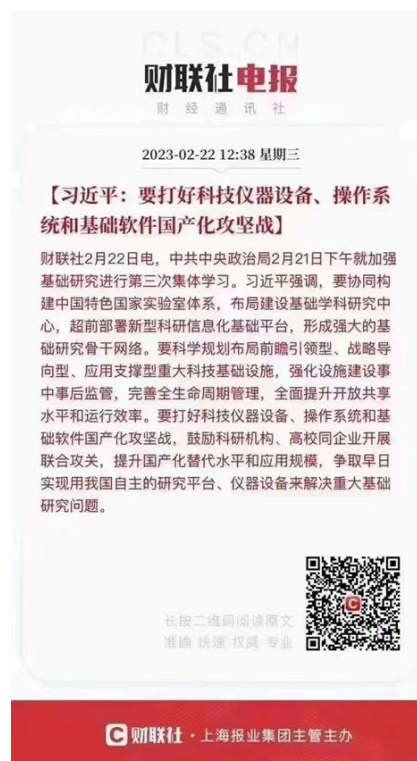
□ 基础软件系统是国家重大战略需求，支撑国民经济和社会发展



“
希望广大科学家
和科技工作者
肩负起历史责任，
坚持面向世界科技前沿、
面向经济主战场、
面向国家重大需求、
面向人民生命健康，
不断向科学技术
广度和深度进军。”

—— 2020年9月11日
习近平主持召开科学家座谈会强调

Baidu 百科 央视新闻



CFLP NEWS
财联社电报
财经通讯社

2023-02-22 12:38 星期三

【习近平：要打好科技仪器设备、操作系统和基础软件国产化攻坚战】

财联社2月22日电，中共中央政治局2月21日下午就加强基础研究进行第三次集体学习。习近平强调，要协同构建中国特色国家实验室体系，布局建设基础学科研究中心，超前部署新型科研信息化基础平台，形成强大的基础研究骨干网络。要科学规划布局前瞻引领型、战略导向型、应用支撑型重大科技基础设施，强化设施建设事中事后监管，完善全生命周期管理，全面提升开放共享水平和运行效率。要打好科技仪器设备、操作系统和基础软件国产化攻坚战，鼓励科研机构、高校同企业开展联合攻关，提升国产化替代水平和应用规模，争取早日实现用我国自主的研究平台、仪器设备来解决重大基础研究问题。

长按二维码阅读原文
准确 快速 权威 专业

财联社 · 上海报业集团主管主办

课程内容

6

□ When: 背景

□ Why: 设计

□ What: 架构

□ How:

✚ 原理: 系统层面

✚ 编程: 用户层面

课程安排

7

□ 理论课程：每周3学时，紧跟节奏思考

- ✚ 设计思想：为什么？
- ✚ 系统架构：是什么？不同系统的联系与区别
- ✚ 编程思路：不是教API

□ 实践课程：每周2学时

- ✚ 开源系统部署：保持耐心、坑很多，不要奢望照着实验说明就能一步到位
- ✚ 基本编程开发、代码调试：
 - 使用Java开发，动手能力强的自学Scala
 - 熟练使用IntelliJ IDE、maven



课程安排(tentative)

8

周	周一	周四
1	绪论	实验一：准备工作
2	HDFS	实验二：Hadoop 1部署
3	MapReduce设计思想与架构	实验三：Hadoop 2部署
4	MapReduce工作原理	实验三：Hadoop 2部署
5	MapReduce编程	实验四：Hadoop 2编程
6	Spark设计思想与架构	实验四：Hadoop 2编程
7	Spark工作原理	实验五：Spark部署
8	Spark编程	实验五：Spark部署
9	MapReduce习题课	实验六：Spark编程



课程安排(tentative)

9

周	周一	周四
10	五一放假	实验六: Spark编程
11	Yarn	实验七: Spark+Yarn
12	Flink设计思想与架构	实验八: Flink部署
13	Flink工作原理	实验八: Flink部署
14	Flink编程	实验九: Flink编程
15	Spark习题课	实验九: Flink编程
16	Flink习题课	实验十: Flink+Yarn
17	答疑	端午节
18	考试	



课后训练

10

□ 理论复习

- ✚ 多思考，不要死记硬背

□ 动手编程是最关键的

- ✚ 上机作业：在线提交

- ✚ 增强调试代码的能力

 - 设断点调试

 - 遇到错误，多用Google搜索



课程成绩评定

11

□ 平时成绩：50%

- ✚ 考勤：10%

- ✚ 课堂讨论：15%

- ✚ 编程作业（2+1次）：30%

- ✚ 上机实验（10个实验共3组）：45%

□ 期末成绩：50%

- ✚ 期末考试：100%

课程要求

12

- 考勤：无故缺勤扣10分，扣完为止
- 课堂讨论：请积极参与课堂讨论
- 编程作业：3-4周时间内完成
- 上机实验：
 - ✚ 单机实验(90%)：独立完成，助教检查
 - ✚ 分布式实验 (bonus, 10%)：非必须内容，1-4人合作完成，完成后请主动找助教登记
 - ✚ 三组实验报告：实验1-4、实验5-7、实验7-10



实验报告

13

- 报告内容：需要记下来以后参考的内容
 - ✚ 不是抄实验说明，仅需记录你踩过什么坑，即：遇到了什么问题？是如何解决的？
 - ✚ 每个实验的报告不少于1页但不超过2页，每组实验报告的篇幅控制在6-8页
- 命名格式：学号-姓名-实验ABCD-版本X.pdf
 - ✚ X取值范围 $[1, +\infty)$
 - ✚ 例如：10061-张三-实验1234-版本1.pdf
- 注意：网上提交，规定的deadline一般是实验课结束一周左右，不交0分，迟交扣分



编程作业(tentative)

14

□ 每次5道题目，每题10分

难度系数	权重	备注
无	10%	赠送
易	20%	
中	30%	
中	30%	
难	10%	可选

教材及参考书目

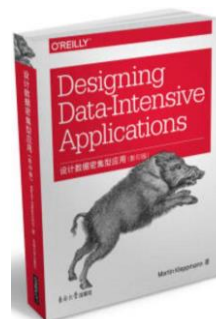
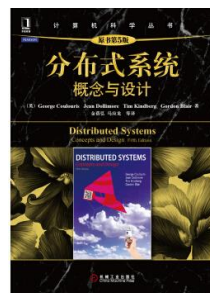
15

□ 教材

- ✚ 徐辰，分布式计算系统，高等教育出版社 2022

□ 参考书

- ✚ 分布式系统概念与设计，George Coulouris等著，金蓓弘等译
- ✚ 设计数据密集型应用，Martin Kleppmann



□ 教学团队

✚ 主讲教师：徐辰

✚ 助教：

➤ 孙玉书：2022级硕士

➤ 许珈赫：2022级硕士

➤ 刘明熹：2019级本科

□ 课程信息

✚ 钉钉群：课程通知、实验注意事项等

✚ 主页：<https://dasebigdata.github.io/>

谢谢! Q&A

