数据与计算知识点

第3周

计算机的功能：

营造一个和物理世界一致的数字化世界；

通过计算，精准揭示物理世界现状和规律；

通过模拟，预见物理世界的未来。

可计算函数的定义：能够在抽象计算机上编出程序计算其值的函数。

图片大小的计算：水平像素\*垂直像素\*1（黑白）或3（彩色）\*一种颜色深度bit数（黑白是1，彩色是8）/8/1024（单位为KB）。

对计算思维的理解：

计算思维是我们认知计算的过程中积累形成的思考“模式”；

计算思维教/学需要传递计算给我们带来的可能性以及实现这些可能的基本方法，具体表现为：想以前想不到之事，做以前做不到之事，做以前做不好之事；

算法是解读计算思维的最佳载体，是计算思维的核心概念，体现为抽象和自动化。

导致了我们的独特视角,驱动计算机技术和应用如此迅速地发展的因素：人类在认识计算和组织计算中遇到和解决独有问题时形成的思考模式。

作业：寻找假币问题

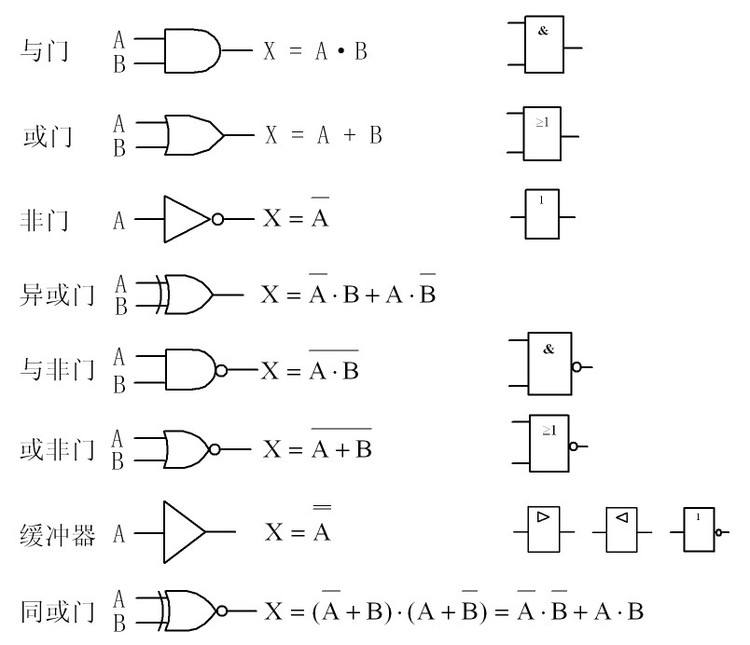
第4周

计算机内部的01的实现方式：

外存：磁介质存储：数据记录在磁性材料上，磁性材料磁化方向不同表征01；光介质存储：光盘表面单元面积上是否存在蚀刻出的“坑”表征01；

内存：半导体介质存储；存储单元的电气特征不同表征01。

各种门的符号：



字符编码的定义：它是一套法则，使用该法则能够对自然语言的字符的一个集合（如字母表或音节表），与其他东西的一个集合（如号码或电脉冲）进行配对。

余码和补码，浮点表示法

作业：二进制数的运算以及溢出问题，解释不同音乐格式APE、MP3在编码环节的区别

第6周

判断两个bit位是否相等：

equality test(x,y)

1.zero eq

2.test x flip eq

3.test y flip eq

计算两个1bit位的数之和：

add(x,y)

1. zero z0

2. zero z1

3. equality test(x,y)

4. test eq goto 7

5. flip z0

6. exit

7. zero t

8. flip t

9.equality test(x,t)

10. test eq flip z1

工资求和问题（数组求和），插入排序算法

作业：两个2bits位的数的判断是否相等以及求和，计算机解题四过程：理解问题，设计方案，执行计划，回顾；二分法查找学生信息

第7周

goto语句的缺陷，汉诺塔问题

变量的定义：变量是用于跟踪、操纵几乎所有数据的简单(通用)工具。

数据类型的作用：

定义了变量的变化范围；

定义了计算对变量的操作方式。

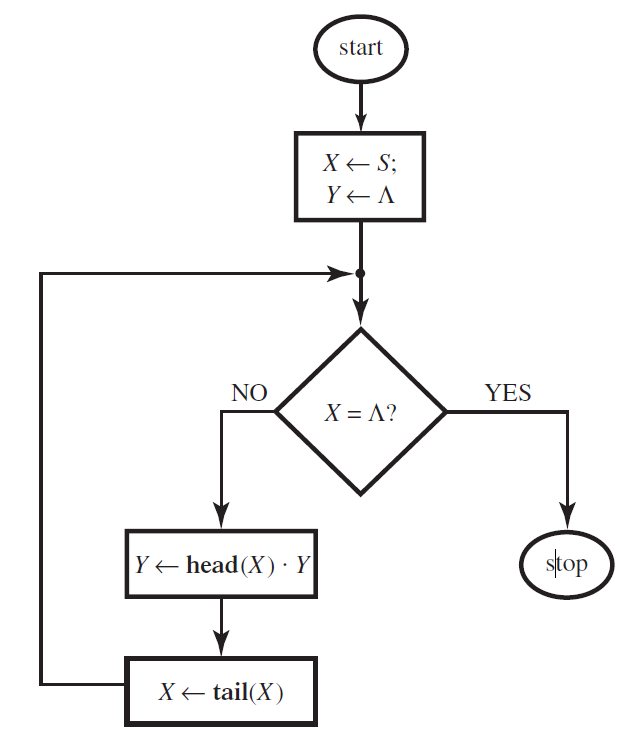
地址和指针

数组和队列的区别：前者具有使用时的随意性：任意存取，但代价更高；后者具有使用时的受限性：头存尾取，但代价更低。

作业：插入排序算法的程序框图，不同控制结构的转换，全排列问题

第8周（必考）

求一个字符串的逆序字符串：



部分正确性的定义：如果对于每个合法的输入，程序能运行到最后一步，则其能输出正确的结果。

完全正确性的定义：在部分正确性的基础上，程序确实运行到了最后一步，并输出了正确的结果。

checkpoint的定义：我们可以在算法的任意位置，臆想算法的暂停，检查算法的行为，这一位置称为checkpoint。

assertion的定义：判断数据状态是否符合某些规则或约定的条件。

assertion的作用：

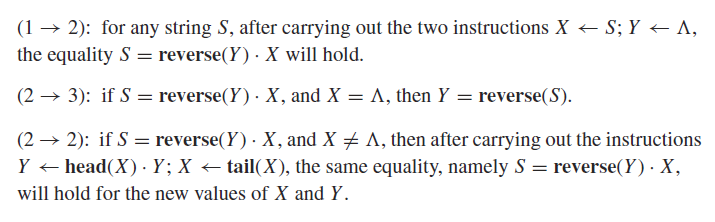
可以用来判断任意输入的合法性；

如果每个合法的输入，算法都能运行到最后一步并且该处的断言都为真，算法的部分正确性就能得到证明；

在循环内的断言（循环不变式）可以反映每次到达该点时处理逻辑必须遵循的约定，并判断一个动态执行的状态。

真断言链：(1)->(2)->(2)->……->(3)。其中(1)位于开头，判断输入的合法性；(2)位于循环内，起到循环不变式的作用；(3)位于结尾，判断输出的正确性。

证明算法的部分正确性（以字符串反转为例）：



循环不变式的定义：它是在循环过程中设计的断言，每进行一次循环，各个循环变量当前的值都满足循环不变式且向循环的目的更加靠近了一步。

作业：含两个循环的程序需要不变量（即断言）的数目，证明算法的部分/完全正确性

第9周

时间复杂度的定义：算法在给定一个输入的情况下，要执行多少条指令。

最坏时间复杂度：是指对于算法的任意一个输入，在最不利情况下的时间复杂度。

最好时间复杂度：是指对于算法的任意一个输入，在最好的情况下的时间复杂度。

平均时间复杂度：是指算法在平均情况（最一般的情况）下的时间复杂度。若某算法的最坏时间复杂度等于最好时间复杂度，则平均时间复杂度与两者相同。

时间复杂度的计算方法：

我们选择“代表性”语句，进行统计；

我们往往忽略代表性语句执行条数的“系数”而只保留其指数。

归并排序

性能函数的定义：它是一个接受数据规模，输出算法执行步骤数的函数。

时间复杂度o(n)可以理解为性能函数在n->正无穷时的主要影响部分，是性能函数的“主部”。常见的时间复杂度有：o(1),o(n),o(nlogn),o(n^2),o(n!)等等。

性能函数阶的比较，就是比较当n->正无穷时两者相除的结果，类似于无穷小量阶的比较。

数据的组织和结构涉及到的内容：

数据自身的逻辑模型；

数据在计算机中的存储模型。

作业：优化程序，求算法时间复杂度，二叉树排序，归并排序和桶排序的最坏，最好，平均时间复杂度的分析

第10周

问题求解的基本方法：

确定合理的解空间，并表示为某种“结构”；

利用已知的限制条件（知识）尽可能快的压缩可能的解空间；

当解空间已经足够小，我们就可以“直接”解题。

难问题的定义：如果很难确定解空间的范围，或者解空间的范围无限，或者很难有效地缩小解空间，这个问题就属于难问题。

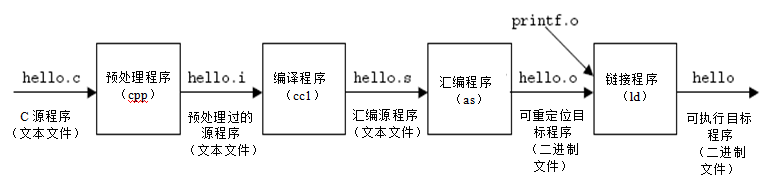
贪心算法的定义：在对问题求解时，总是做出在当前看来是最好的选择。也就是说，不从整体最优上加以考虑，他所做出的是在某种意义上的局部最优解。

动态规划的定义：按照问题规模的某种序，求解小规模问题，将小规模问题的解存起来，供解稍大规模问题使用的方法。

作业：算法问题的解，可行解和解空间，简单问题和难问题的例子，两种找中位数的方法（实质是排序算法）的比较

第11周（必考）

高级语言程序转换为机器代码的过程：



Hello程序被启动后，计算机的动作过程如下：

读取字符串“./hello”中各字符到寄存器，然后存放到主存；

“Enter”键输入后，操作系统内核根据主存中的字符串“hello”到磁盘上找到特定的hello目标文件，将其包含的指令代码和数据从磁盘读到主存，并将控制权转交给hello程序，即将hello程序的第一条指令的地址送到PC中；

处理器从hello主程序的指令代码开始执行；

Hello程序将“hello, world\n”串中的字节从主存读到寄存器，再从寄存器输出到显示器上。

指令码的定义：处理器制造厂商在芯片内部预先定义的一串二进制代码，它规定了处理器的功能和任务，也称为机器语言。

可执行程序以机器语言形式存在。

机器级操纵指令的符号化（汇编语言）的意义：能够帮助程序员记忆指令功能。

汇编语言的定义：汇编语言是建立在指令集体系结构基础上的，基本语言成分是指令集中的指令的，不关心指令集如何被具体实现、物理执行的一种编程语言。

汇编语言的特点：

用助记符表示操作码；

用标号表示位置；

用助记符表示寄存器。

汇编编程的优点：

不需记忆指令码，编写方便；

可读性比机器语言强；

与绝对位置无关，扩展性好；

更多的计算机硬件细节被开放给程序员；

可以编写时空更高效的程序。

runtime（运行时）的程序员可见状态：

寄存器组：CPU内部最高速存储空间，共8个，32位，用于存储地址或者数据，记录临时数据或者程序运行状态；

PC(程序计数器)：指示下一条指令在存储器中的地址（通常是相对位移）；

条件码：保存最近执行的指令结果状态信息；供后期用于管理数据流或者控制流；

状态寄存器：有若干标志位，每位(组合位)表示一个程序运行中的状态。它可以暂存数据，记录堆栈实时地址，代码指针，部分寄存器还能记录当前的测试状态。

AT&T 指令格式的基本操作：

在寄存器或存储器的数据上执行算术操作；

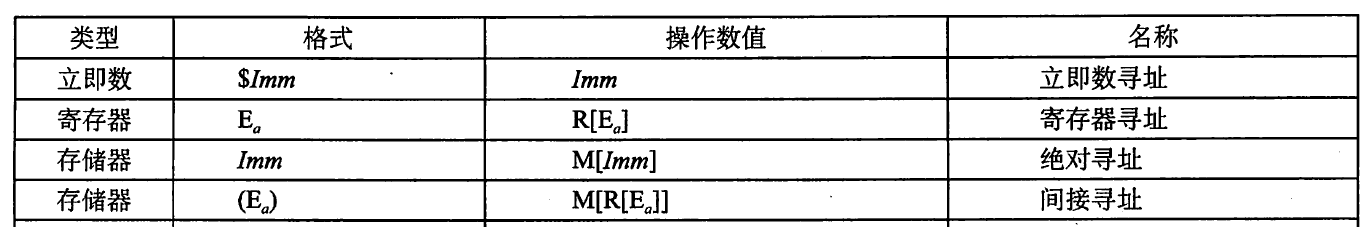
在存储器和寄存器间传送数据；

转移控制：无条件跳转，函数调用/返回，条件分支。

典型的指令格式：

操作符opcode 源操作数src, 目标操作数dst

三种表达（寻址）方式：立即数，寄存器，存储器。



算术和逻辑操作符：



leal的含义

条件码：记录当前操作运行状态；

比较指令：进行一次比较，通过设置条件码表征“比较”状态；

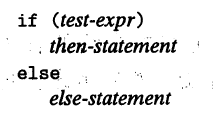
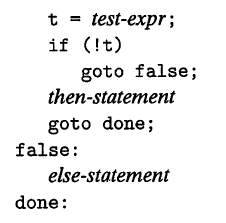
跳转指令：根据条件码，设置PC，确定下一条待运行指令；

条件码寄存器：最常用有4个单独位，表示计算中最近的操作结果的属性。

jX 指令：根据条件码跳转到程序不同的部分



分支结构的转换模板：

do-while的翻译：

***loop*:**

***Body***

**if (*Test*)**

**goto *loop***

while的翻译：

**if (!*Test*)**

**goto *done*;**

***loop*:**

***Body***

**if (*Test*)**

**goto *loop*;**

***done*:**

for的翻译：

**for (*Init*; *Test*; *Update* )**

***Body***

***Init*;**

**if (!*Test*)**

**goto *done*;**

**loop:**

***Body***

***Update* ;**

**if (*Test*)**

**goto *loop*;**

**done:**

栈的定义：栈是一种采用“先进后出”方式进行访问的一块存储区，用于嵌套过程调用。从高地址向低地址增长。

一个C过程的大致结构如下：

准备阶段：

形成帧底：push指令 和 mov指令；

生成栈帧（如果需要的话）：sub指令 或 and指令；

保存现场（如果有被调用者保存寄存器）：mov指令；

过程（函数）体：

分配局部变量空间，并赋值；

具体处理逻辑，如果遇到函数调用时；

准备参数：将实参送栈帧入口参数处；

CALL指令：保存返回地址并转被调用函数；

在EAX中准备返回参数；

结束阶段：

退栈：leave指令 或 pop指令；

取返回地址返回：ret指令；

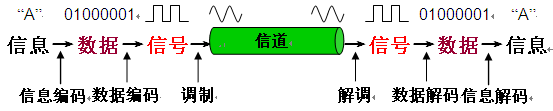
计算机系统的不同抽象层次：上层是下层的抽象，下层是上层的实现，下层为上层提供支撑环境。

作业：汇编作业

第12周

做到（实现）“跨机器”计算的关键：必须准确地实现两个计算“分量”之间的数据传递。

两台计算机通信的基本原理：



IP地址的作用：用来表示该主机在网络中的相对位置,并用于该主机的所有通信。

IP地址的特点：具有唯一性。

IP地址的结构：IP地址=网络号+主机号，其中网络号标志主机所连接到的网络，主机号标志该主机（路由器）。

作业：图灵机模型，跨网络计算，ISO七层网络协议，云电视

第14周

大数据产生的场合：社交网络，淘宝商城，金融银行，移动通信，物联网等等。

大数据的基本特征：大量化，多样化，快速化，有价值。

大数据的应对需求：

有效应对“大”的机理和技术；

有效应对“非结构化”的机理和技术；

有效应对“流”的机理和技术。

数据挖掘的定义：

它是针对大规模观测数据的分析方法，数据规模和复杂性比较高，针对现实数据而非实验数据；

它的目的是获得相应规律，发现数据中隐藏着的未知关系，确定该关系的正确性和规律性；

它的最终目的是帮助数据拥有者进行决策。

大数据分析的经典方法：分类法，聚类法，关联关系分析法。

两个对象之间相异度的衡量标准：距离。

不变性的两种表现形式：逻辑不变性和统计不变性。

适合并行计算的问题的特点：如果有大量结构一致的数据要处理，且数据可以分解成相同大小的部分，那我们就可以设法并行进行这些处理。

作业：大数据的“大”的体现，分类法，聚类法和频繁项集法的基本思想，Map和Reduce的含义

第17周

计算机病毒的定义：编制者在计算机程序中插入的破坏计算机功能或者破坏数据、影响计算机使用并且能够自我复制传播的一组计算机指令或者程序代码。

计算机病毒的特点：

它是人为制造的程序；

它会潜伏，适机运行；

它会自我复制并传播。

几种典型的计算机病毒：蠕虫病毒，逻辑炸弹，特洛伊木马，后门。

木马的攻击手段：观察、窃取，修改，操控。

木马的显式存在和隐式存在：

显式存在：

木马的运行，在计算机系统中以”进程”或”服务”独立存在，被计算机系统管理；

隐蔽存在：

将木马的代码织入到其他应用程序或者管理程序中；

将木马的运行线程注入到其他应用或者管理进程的地址空间中去。

芯片后门：是一段可以被有条件触发的恶意电路。

人肉搜索的定义：是指从片段的信息出发，利用计算机网络及社会网络，不断发现、挖掘和某个人的信息相关联的其它信息，进而不断丰富和完善直至得到满意的结果。