考试试题：

一、体系结构：

a) 假设一个事务是将 SQL 嵌入到 C 语言中实现的，在运行过程中，有大约 80%的时间是花在 SQL 语句执行上的，如果只对 SQL 语句实施并行，那么期望能够获得多大的加速比，请说明理由。

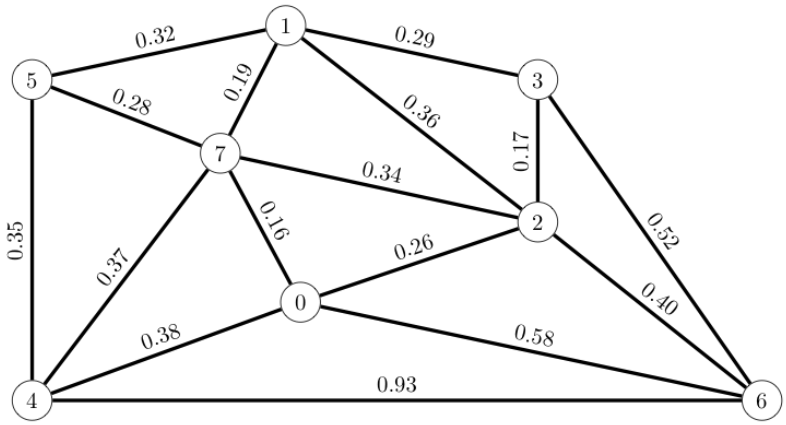
b) 在共享无内容(shared-nothing)系统中，远程数据访问可以通过远程过程调用（RPC）或发送消息来实现。但是远程直接内存访问（RDMA）提供了比这些方法更快速的数据访问机制。解释一下为什么 RDMA 会更快？

c) 假设有若干个共享资源 $r_1,r_2,\ldots,r_n$ 每个资源都由一个相应的标志位F_i来指示是否被锁定。
a. 使用测试并设置（test-and-set）指令描述实现 lock(r_i) 和 unlock(r_i) 的方法。
b. 使用比较并交换（compare-and-swap）指令描述实现 lock(r_i) 和 unlock(r_i) 的方法。
请分别描述如何使用这两种指令来实现对共享资源 r_i 的加锁（lock）和解锁（unlock）操作。

二、并行和分布式数据库

a) 请描述流水线并行的优点和缺点。

b) 对于下面这样一个图，要找一个从节点 1 到节点 6 的路径，使得路径上的权重和最小。请设计一个并行策略完成这个任务。



c) 请列举一个读一次、写所有可用（read one, write all available）方法导致错误状态的一个例子。

d) 题目：对于给定的两个关系 r 和 s ，执行连接操作 $r\bowtie_{r.A=s.A \wedge r.B<s.B} s$, 可以使用分区连接来优化此查询吗？请解释你的答案。

三、高级数据类型和新应用

a) 数据库中，时间有那些类型？

b) 假设你有一个空间数据库，支持使用圆形区域进行区域查询，但不支持最近邻查询。描述一种算法，通过利用多个区域查询来找到最近邻的邻居。

c) 假设你要设计一个用于存储城市地图数据的 R 树。每个城市表示为一个矩形区域（城市边界），你需要将这些城市的矩形区域存储在 R 树中以支持空间查询。
请回答以下问题：
i. 如何构建 R 树来存储城市地图数据？
ii. 如果要查询某个指定区域内包含的所有城市，你会如何使用 R 树进行查询？
iii. R 树在存储和查询城市地图数据时的优势是什么？

四、大数据

a) 假设你希望将一个大学的 schema 建模成一个图（graph），对于下面的每一个关系，他们是建模成节点还是边？ 这个模型能捕捉 sections 和 courses 之间的连接吗？

(1) student (2) instructor (3) course (4) section (5) takes (6) teaches

b) 假设一个数据流可以按照元组的时间戳顺序无序地发送元组。数据流应该提供哪些额外信息，以便流查询处理系统能够确定何时已经看到了窗口中的所有元组？

c) 使用 Map-Reduce 框架统计一组文档中每个单词的出现次数
例如，如果输入的文档 ID 和内容如下：
文档 1: "apple orange banana"
文档 2: "banana peach apple"
文档 3: "orange mango apple"
请给出相应的 Map 和 Reduce 函数的伪代码，以创建每个单词的出现次数统计。