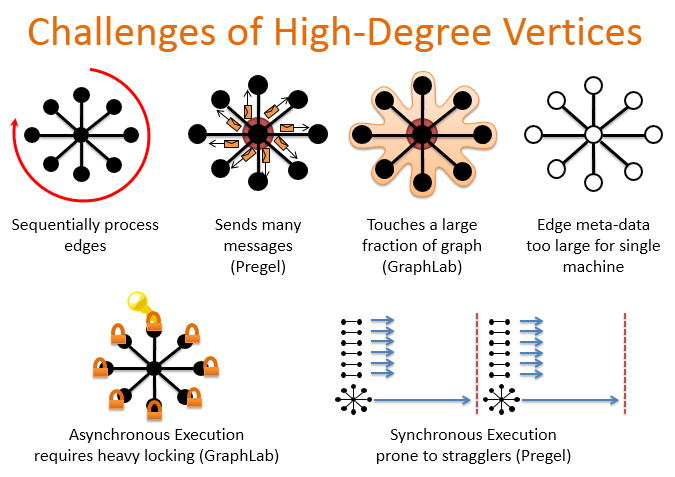
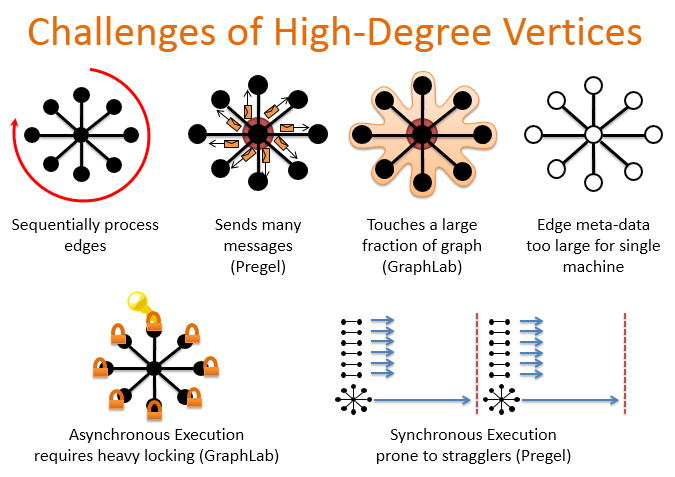
1. 请分析大数据相关特征，包括数据体量大(Volume)、数据流动性高(Velocity)以及数据种类繁多(Variety)分别对大数据处理带来了什么具体挑战，并简要阐述相应措施。
2. 简述复制技术为什么会带来数据一致性问题，并分析如何保证多副本的一致性。
3. 副本的放置需要遵循哪些原则？简述在集群环境下的多副本放置策略。
4. 简述分布式环境下CAP原理的内涵。
5. 向量钟和Merkle树分别是什么？简述他们的作用
6. NVM有什么特性？为什么要用DRAM和NVM一起构成异构内存系统？
7. 内存计算模式是什么？为什么内存系统对大数据处理很重要？
8. Dynamo如何对数据进行分区？
9. 什么是函数式编程？它有哪些特点？
10. 简述引用透明性、高阶函数、函数柯里化、惰性求值的内涵
11. \*简述Google MapReduce和Hadoop的异同点。
12. 假设有1百万个均匀分布的数，试给出抽取其中前100个最大数的MapReduce方案，并给出Map和Reduce的伪代码。
13. 请简述at most once、at least once和exactly once这三种语义保障。
14. \*简述Apache Storm和Spark Streaming 的特点和区别。
15. 简述几种主要的图计算编程模型，并分析它们的优点和缺点。
16. 采用Scatter-Gather编程模型，写出单源点最短连接算法SSSP的伪代码。
17. 对于幂律图中的高维度节点（例如Twitter中具有大量粉丝的美国总统），有以下几种处理方式：



顺序处理 发送消息 接触大部分图



共享状态：异步执行 共享状态：同步执行

请写出以上几种方法的缺陷。

观察多层划分算法的伪代码，尝试理解操作Coarsening和Uncoarsening的意义

Multilevel Graph Partition (Graph *G*)

① Take the weighted undirected graph as input

② While (|*V*|> coarsening threshold) perform

Coarsening (*Gi* to *Gi* +1 )

③ For all *v* in *V*

If vertex degree (*v*) is high perform

Balanced partitioning

Else

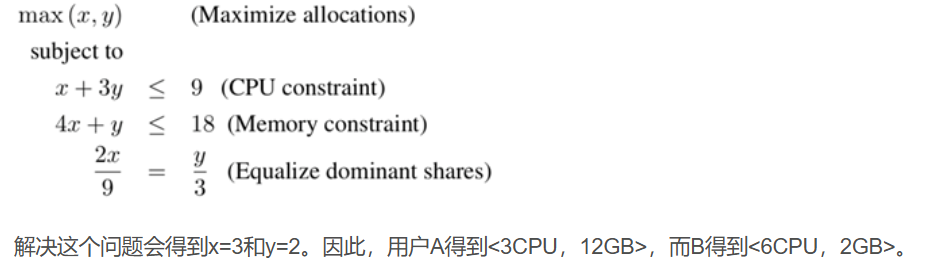
Perform Greedy partitioning

④ Simulated annealing(initial partition)

⑤ While (*Gi*!= *G*) perform

Uncoarsening (*Gi* to *Gi-1* )

1. 比较Raft协议和Paxos协议的区别与联系。
2. 计算机系统有9个CPU和18 GB内存，用户A的每个任务都请求（1CPU，4GB）资源；用户B的每个任务都请求（3CPU，1GB），根据主资源公平分配算法，写出资源的分配过程。



1. 集群中总的资源量为16 CPU，36 GB内存，3GB/s网络。现在有A、B、C三个MapReduce作业，A 有16个map任务和3个reduce任务，B和C都有8个 map任务和3个reduce任务。A的每个map任务的资源需求是<1 CPU, 2GB 内存>，B和C每个map任务的资源需求都是<2CPU, 1GB内存>。A、B、C的reduce任务都只需要1GB/s的网络带宽。请分别绘出用主资源公平调度算法和多资源打包调度算法的示意图。

主资源公平调度算法

| **核** | **16core** | **16core** | **16core** | **16core** | **0core** | **0core** | **0core** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **内存** | **16GB** | **16GB** | **16GB** | **16GB** | **0GB** | **0GB** | **0GB** |
| **网络** | **0GB/s** | **0GB/s** | **0GB/s** | **0GB/s** | **3GB/s** | **3GB/s** | **3GB/s** |
| **A** | **4任务阶段1** | **4任务阶段1** | **4任务阶段1** | **4任务阶段1** | **1任务阶段2** | **1任务阶段2** | **1任务阶段2** |
| **B** | **2任务阶段1** | **2任务阶段1** | **2任务阶段1** | **2任务阶段1** | **1任务阶段2** | **1任务阶段2** | **1任务阶段2** |
| **C** | **2任务阶段1** | **2任务阶段1** | **2任务阶段1** | **2任务阶段1** | **1任务阶段2** | **1任务阶段2** | **1任务阶段2** |

**多资源打包调度算法**

| **核** | **16core** | **16core** | **16core** | **0core** |
| --- | --- | --- | --- | --- |
| **内存** | **32GB** | **8GB** | **8GB** | **0GB** |
| **网络** | **0GB/s** | **3GB/s** | **3GB/s** | **3GB/s** |
| **A** | **16任务阶段1** | **3任务阶段2** |  |  |
| **B** | **0任务** | **8任务阶段1** | **3任务阶段2** |  |
| **C** | **0任务** | **0任务** | **8任务阶段1** | **3任务阶段2** |

1. 并行任务在分布式缓存系统中调度的基本原则是？存储优化中解决小文件问题的方法有哪些？