Tutorial 3：调度与连续批处理观察

题目

1. [单选题] 调度器更接近下列哪一项职责？A. 直接提高 kernel 速度 B. 分配执行机会与运行顺序 C. 替代模型执行器 D. 只负责日志打印

---

2. [判断题] 只要系统吞吐更高，短请求体验也一定更好。

---

3. [多选题] 下列哪些现象通常说明连续批处理正在发生？A. waiting 与 running 集合持续动态变化 B. 新请求可以在旧请求未完成时进入运行集合 C. batch 大小永久固定不变 D. 请求会动态退出运行集合

---

4. [简答题] 为什么短请求常在混部场景中首先处于不利位置?

5. [计算题] 策略 A 在 20 秒内完成 100 个请求，其中 70 个满足 SLO；策略 B 在 20 秒内完成 90 个请求，其中 85 个满足 SLO。请分别计算两者的 throughput 与 goodput，并判断哪一项策略更适合在线服务。

---

6. [判断题] 抢占或回退主要属于算子优化问题，而不是调度问题。

---

7. [简答题] 为什么“设备利用率更高”有时会带来“用户等待更久”?

8. [伪代码题] 请写出一个简短伪代码：若存在“高交互优先”标记请求，则优先从 waiting 集合中调

度该请求；否则按常规顺序调度。

9. [多选题] 下列哪些现象较像调度器过度偏向长请求？A. 整体吞吐较高 B. 短请求 TTFT 与 P99 明显恶化 C. waiting 队列中短请求停留过久 D. 所有请求体验同时稳定改善

---

10. [简答题] 为什么调度解释必须与 workload 一起阅读?

参考答案

1. B。调度器主要负责执行机会分配和运行顺序管理。
2. 错。吞吐改善可能是以等待时间上升为代价换来的。
3. A、B、D。这些都是连续批处理的典型信号。
4. 因为长请求更容易持续占据运行窗口和状态资源。
5. 策略 A：throughput = 100 / 20 = 5 请求每秒，goodput = 70 / 20 = 3.5 请求每秒；策略 B：throughput = 90 / 20 = 4.5 请求每秒，goodput = 85 / 20 = 4.25 请求每秒。若目标是在线服务，策略 B 通常更合适。
6. 错。抢占或回退改变的是顺序、状态和运行安排，属于调度问题。
7. 因为更高利用率常伴随更长的排队窗口和更激进的资源占用。
8. 示例伪代码：

```
function pick_next(waiting_queue):
     for req in waiting_queue:
          if req.high_interactive == true:
               return req
     return waiting_queue[0]
```

9. A、B、C。这类现象常说明长请求被优先推进。
10. 因为不同 workload 对公平性、吞吐和时延的偏好并不相同。