实验 3：调度与连续批处理实验

1. 本节定位

本实验聚焦调度控制点。重点是从“系统正在执行 token”进一步进入“系统正在决定先执行谁”的层面。本节强调观察 batch 形成、队列等待和策略取舍，而不是写一个复杂调度器。

2. 核心问题

调度改变的到底是“算得更快”，还是“谁先被算、谁先完成、谁先感知到结果”？

3. 学习目标

完成本节后，你应能够：

- 说明当前调度器控制的对象是什么。
- 记录 batch 如何形成、扩展和回收。
- 比较不同 workload 与简单策略差异下的 TTFT、吞吐和等待时间。
- 说清调度改变的是顺序、公平性与资源利用，而不一定是算子本身更快。

4. 开始前

4.1 先修要求

- 已完成实验二，能区分 prefill 与 decode。
- 知道 waiting、running、finished 等基本状态概念。
- 能运行至少两类 workload。

4.2 本节材料

- 理论配套：build/tutorials/Tutorial_03_调度与连续批处理观察.pdf
- 参考代码：src/code/nano-vllm-hust/nanovllm/engine/scheduler.py
- 学生练习：src/code/nano-vllm-hust/student_exercises/nanovllm/engine/scheduler.py
- 作业包：src/experiments/nano-vLLM实验课/experiment_3_调度与连续批处理实验/assignment_spring-2027/

5. 提交内容

本节提交物必须包含：

1. 一份 batch 观测日志。
2. 一张至少两类 workload 的结果对比表。

3. 一页分析，解释策略取舍与用户体验的关系。
4. 如做了策略改动，附最小 diff 或关键代码说明。

6. 时间安排建议

- 10 分钟：回顾调度在生命周期中的位置。
- 20 分钟：建立 batch 观测日志。
- 25 分钟：记录两类 workload 的调度行为。
- 20 分钟：尝试一项最小策略改动或参数改动。
- 10 分钟：整理策略取舍说明。

7. 实验任务

任务 1：确定观测对象

在调度循环中至少记录以下对象：

- waiting 队列长度
- running 集合大小
- 当前 batch 中的 `request_id`
- 每个请求的 `prompt_len`
- 每个请求已经生成的 token 数

任务 2：设计工作负载

至少准备两类 workload：

1. 同质 workload：请求长度接近。
2. 异质 workload：长短请求混合。

如条件允许，可增加突发 workload。

任务 3：设计对照策略

可选方式：

- 修改最大 batch size。
- 修改每轮允许接纳的新请求数。
- 修改 decode 过程中是否插入新请求。
- 实现一个最小的短请求优先插入规则。

任何改动都必须写清：

- 改的是哪个控制点。
- 预期影响哪类指标。

- 哪个结果会推翻这个预期。

任务 4：记录结果

请填写下表：

workload	policy	batch_limit	TTFT	throughput	queue_wait	short_finish	long_finish	解释
homogeneous								
heterogeneous								
bursty								

记录要求：至少覆盖等待、公平性和吞吐中的两个维度，不能只比较 throughput。

任务 5：整理一段批处理日志

至少保留一段能看清下列现象的日志：

- 新请求何时进入 batch。
- 长请求是否持续占位。
- 短请求是否因为插入策略受益或受损。

8. 证据要求

以下情况不能视为合格证据：

- 只比吞吐，不比等待时间和短请求完成时间。
- 改了参数却没有说明控制点。
- 没有记录 batch 形成过程，却直接解释调度效果。
- 把策略收益写成绝对结论，没有给出失败条件。

9. 提交核对

- 已附 batch 观测日志。
- 已附两类 workload 以上的结果对比表。
- 已说明策略取舍与用户体验关系。
- 若做了改动，已给出最小 diff 或关键代码说明。

10. 评分关注点

- 是否真的观察到了调度行为，而不是只看最终数字。
- 指标选择是否覆盖等待、公平性和吞吐至少两个维度。
- 解释是否回到控制点，而不是停在“更快/更慢”。

11. 与后续实验的衔接

本节结果会直接进入：

- 实验四：解释状态容量为什么会影响调度弹性。
- 实验五：形成完整的对照实验记录。
- 课程项目：调度方向开题的基线证据。

12. 提交模板

12.1 基本信息

姓名 / 组员:
学号:
实验日期:
模型与环境:

12.2 工作负载设计

workload	说明	请求长度特征	到达方式
homogeneous			
heterogeneous			
bursty			

12.3 策略与控制点记录

改动的控制点:
改动方式:
预期影响指标:
哪个结果会推翻预期:

12.4 结果表

workload	policy	batch_limit	TTFT	throughput	queue_wait	short_finish	long_finish	解释
homogeneous								
heterogeneous								
bursty								

12.5 批处理日志摘录

```
[timestamp] batch_state ...
```

12.6 策略取舍说明

请用约 1 页说明：你的策略主要改变了谁先被算、谁先完成、谁先等待？它带来的收益和代价分别是什么？