实验 4：KV 状态与缓存组织实验

1. 本节定位

本实验把 KV 从“attention 的中间量”提升为“推理系统中的长期状态对象”。重点是观察状态如何生成、驻留、扩展和释放，以及这种状态组织如何反过来约束并发和调度。

2. 核心问题

为什么很多看起来像“调度问题”或“延迟问题”的现象，根上其实是状态驻留和容量问题?

3. 学习目标

完成本节后，你应能够:

- 画出 KV 或请求状态对象的生命周期。
- 指出状态与请求、调度、并发之间的关系。
- 用一次状态容量观察说明“长上下文为何会压缩系统弹性”。
- 区分仅用于观测的探针信号和真实复用收益。

4. 开始前

4.1 先修要求

- 已完成实验三，能解释调度和等待。
- 已了解 KV 生命周期、块式管理或状态对象的基本概念。
- 教学代码能够支持状态规模的近似观测。

4.2 本节材料

- 理论配套: `build/tutorials/Tutorial_04_KV_Cache 与状态组织.pdf`
- 参考代码: `src/code/nano-vllm-hust/nanovllm/engine/block_manager.py`
- 相关状态对象: `src/code/nano-vllm-hust/nanovllm/engine/sequence.py`
- 学生练习:`src/code/nano-vllm-hust/student_exercises/nanovllm/engine/block_manager.py`
- 作业包:`src/experiments/nano-vLLM 实验课/experiment_4_KV 状态与缓存组织实验/assignment_spring-20`

5. 提交内容

本节提交物必须包含:

1. 一张 KV 生命周期图。
2. 一张状态容量观察表。

3. 一段 500-800 字分析，说明状态如何约束系统能力边界。
4. 如果做了探针，附一段明确的说明，交代哪些结论仍只是观测信号。

6. 时间安排建议

- 10 分钟：回顾 KV 是什么，不是什么。
- 20 分钟：定位状态对象。
- 25 分钟：完成状态容量观察。
- 20 分钟：整理生命周期图和因果链。
- 10 分钟：区分观测信号和真实收益。

7. 实验任务

任务 1：定位状态生命周期

请定位或近似定位以下事件：

1. prefill 产生 KV 或等价状态。
2. decode 继续读取和追加状态。
3. 请求完成后释放状态。
4. 多请求并发时状态如何绑定到 `request_id`。

任务 2：设计容量观测

至少准备三类输入：

- `short_ctx`：较短上下文
- `long_ctx`：较长上下文
- `mixed_ctx`：长短上下文混合

建议记录以下量：

- 当前存活请求数
- 每请求估计状态规模
- 总状态对象数或总块数
- TTFT
- queue wait

任务 3：写出一条明确因果链

至少写出一条明确因果链，例如：

上下文变长 -> 状态驻留增大 -> 可并发请求减少 -> 等待时间上升 -> TTFT 变差

任务 4：选择一项扩展观察

二选一：

1. 实现一个最小状态统计器。
2. 实现一个仅用于观测的前缀命中探针。

如果选择探针，报告中必须明确写出：

- 这是潜在重叠信号。
- 这不是已经兑现的真实复用收益。

任务 5：记录结果

请填写下表：

case	prompt_len	concurrency	live_requests	estimated_kv	TTFT	queue_wait	解释
short_ctx							
long_ctx							
mixed_ctx							

8. 证据要求

以下情况不能视为合格证据：

- 把 KV 只写成 attention 的中间结果。
- 只看显存或块数，不解释它如何影响并发和等待。
- 用仅用于观测的探针结果直接宣称有真实复用收益。
- 没有说明状态规模估计的近似边界。

9. 提交核对

- 已附 KV 生命周期图。
- 已附状态容量观察表。
- 已写出至少一条明确因果链。
- 如果做了探针，已附相应说明。

10. 评分关注点

- 是否建立了状态视角，而不是仍停留在单次计算视角。
- 因果链是否清楚。
- 是否诚实区分观测信号和真实收益。

11. 与后续实验的衔接

本节结果将直接进入实验五和课程项目：

- 实验五：把状态观测纳入规范实验记录。
- 课程项目：复用生命周期图、容量表和失败条件。

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 状态生命周期记录

生命周期事件	文件 / 函数 / 对象	判断依据
prefill 产生状态		
decode 读取 / 追加状态		
完成后释放状态		
状态绑定 request_id		

12.3 容量观察表

case	prompt_len	concurrency	live_requests	estimated_kv	TTFT	queue_wait	解释
short_ctx							
long_ctx							
mixed_ctx							

12.4 因果链

上下文变化 ->

12.5 扩展观察

选择的是: 状态统计器 / 仅用于观测的前缀命中探针
实现位置:
观测结论:

12.6 结果说明

请说明: 哪些结果只是观测信号，哪些结果可以被你当作真实收益来解释，哪些还不能。