用小规模数据测峰值，制定单细胞服务器配置表

完整操作教程 · 参考固定 · 往返恢复验收

选单细胞服务器时，一个有记录的小规模试跑，比只看“推荐128G”的经验句更有用。试跑的目的并非跑一个最容易成功的样例，而是找出正式任务最可能出现的峰值，并把它转成配置、并发和验收条件。本文给出一份可填写的试跑协议，适合已有分析脚本、准备确定租赁规格的课题组。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

选单细胞服务器时，一个有记录的小规模试跑，比只看“推荐128G”的经验句更有用。试跑的目的并非跑一个最容易成功的样例，而是找出正式任务最可能出现的峰值，并把它转成配置、并发和验收条件。本文给出一份可填写的试跑协议，适合已有分析脚本、准备确定租赁规格的课题组。

先固定试跑协议，避免比较失真

写下软件与包版本、参考数据、输入文件、随机种子和要执行的步骤。代表性输入应覆盖正式任务的数据结构：多个样本要保留样本差异，空间数据要保留必要位置和特征，整合任务不能只用一个样本来测试。抽样可能改变稀有群和算法行为，因此这里只把子集作为资源探测，不把子集结果当正式生物学结论。

记录项	试跑前填写	为什么需要
输入	文件校验值、细胞与样本数	确保每次比较处理同一批数据
分析范围	读入、标准化、整合或差异步骤	区分峰值来自哪个环节
运行设置	线程、worker数、随机种子	防止并行差异混入配置比较
环境	R、Seurat、系统与包锁定记录	避免把版本变化当性能变化
验收	目标输出、必要维度与日志要求	运行结束不等于分析有效

在独立进程中测一次完整步骤

把已有的、已核对输入输出的分析脚本保存为pilot.R；这里不替你选择整合方法。以下使用Linux上的GNU time，绝对路径用于区分shell自带的time关键字。每次使用新的日志前缀，避免覆盖上次记录。

```
mkdir -p pilot-logs
/usr/bin/time -v -o pilot-logs/run01.resources.txt \
  Rscript --vanilla pilot.R \
  > pilot-logs/run01.stdout.txt \
  2> pilot-logs/run01.stderr.txt
status=$?
printf 'exit_status=%s\n' "$status" > pilot-logs/run01.exit.txt
grep -E 'Elapsed|Maximum resident|Exit status' \
  pilot-logs/run01.resources.txt
```

Maximum resident set size通常以KiB报告，转GiB时除以1024的平方；同时保留原始单位。对于多进程或容器任务，单个time统计不一定等于整个任务所有并发进程的总峰值，应再查看调度器或cgroup层的统计，并询问资源提供方如何定义记账口径。日志里还应记录输出文件大小和当前工作盘剩余量，内存足够却临时盘满，同样不能验收。

至少比较两个规模，检查非线性增长

下表数值都是演算示例，不是任何软件或硬件的实测承诺。若2万细胞峰值为12 GiB、4万细胞为26 GiB，不能据此保证8万只需52 GiB。邻域图、整合锚点、特征范围和中间对象可能导致非线性变化。应增加一个接近正式规模、但仍可安全运行的验证点；若趋势突然上升，先定位算法步骤和对象表示。

试跑	细胞量（示例）	峰值（示例）	下一步
小规模	20,000	12 GiB	确认脚本与输出有效
中规模	40,000	26 GiB	判断是否出现新的高占用阶段
接近目标	按项目填写	待测	确认正式规格而非只做线性外推

把峰值转换成运行约定

先确定一项任务需要的实际可用内存，再确定同时允许多少项任务。设单任务经验证峰值为M，允许并发k，主机还需保留系统和其他服务的空间R，则总预算至少要覆盖k×M＋R，并为尚未覆盖的输入变化保留明确余量。余量不是固定百分比的万能规则，应结合失败代价、数据增长和已有测试覆盖度填写。CPU和磁盘分别规划，不要因为内存翻倍就机械地把核心数也翻倍。

最终配置表应包含正式输入范围、并发、账号额度、工作盘、环境版本、试跑日志和超限处理方式。在候选环境运行同一输入复验，确认结果有效、资源边界清楚，且没有被容器配额或多人使用改变条件。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

参考资料

GNU time资源统计手册（Debian发布） · 来源资料

Seurat大规模数据与sketch教程 · 来源资料

Linux内核cgroup v2文档 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。