记录STAR峰值并设计样本并发与服务器配置

完整操作教程 · 参考固定 · 往返恢复验收

用固定输入顺序记录STAR线程组，再按CPU、RAM与工作盘限制样本并发，最后以任务组峰值复验。

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。

STAR单样本跑通后，批量服务器选型要同时考虑每份任务的线程收益与并发总内存。一个任务更快，不一定让整批更快；私有索引、排序和输出写盘同时出现高峰时，按核心数安排并发容易失败。本文建立可复用的对照协议和预算表。

先固定任务，再比较线程

选择已验收且规模接近正式上界的双端FASTQ，固定索引、STAR版本、过滤和输出。下面以私有索引和未排序BAM为例，4与8线程只是演示档位，需在账号许可额度内运行。两个测试顺序执行、各用新前缀，不让缓存或历史文件的差异隐身。

```
#!/usr/bin/env bash
set -euo pipefail
mkdir star-benchmark-001
for threads in 4 8; do
  out="star-benchmark-001/t$threads"
  mkdir "$out"
  /usr/bin/time -v -o "$out/time.txt" STAR     --genomeDir /refs/star_index --
genomeLoad NoSharedMemory     --runThreadN "$threads"     --readFilesIn /data/S01_
R1.fastq.gz /data/S01_R2.fastq.gz     --readFilesCommand zcat --outSAMtype BAM Unsorted
    --outFileNamePrefix "$out/"     > "$out/stdout.log" 2> "$out/stderr.log"
  samtools quickcheck "$out/Aligned.out.bam"
done
```

若正式目标需要排序，另建采用相同排序参数的测试组，并包含排序全过程。不能用未排序测试的峰值承诺排序任务也足够。需要重复时创建新的批次目录，保留全部有效运行，不能只挑最快一次。

候选方案	单任务线程	有效耗时	峰值RAM	计划并发	整批验收
方案A	____	____	____ GiB	____	____
方案B	____	____	____ GiB	____	____
记录条件	索引与输出相同	墙钟时间	任务与组统计分开	CPU/RAM/盘共同限制	日志、样本和计数齐全

把单任务记录转成初始并发

用实际账号CPU数、可用RAM和测得的单任务预算填下面演算。系统预留需要有依据，并覆盖其他必要程序；若一个样本的峰值明显高于其他样本，按分组或上界预算安排。示例数字纯属假设，结果不能直接当作硬件套餐。

```
from math import floor, ceil
# ■■■■■■■■■■■■■■■■■■■■■■■■
cpu, ram_gib, reserve_gib = 16, 128, 16
threads, one_job_gib, samples, hours = 4, 36, 24, 2
parallel = min(cpu // threads, floor((ram_gib-reserve_gib) / one_job_gib))
if parallel < 1: raise SystemExit("■■■■■■■■■■■■■■■")
batches = ceil(samples / parallel)
print("■■■■■■■", parallel, "■■■■■", batches)
print("■■■■■■■", batches * hours)
```

这组假设得到并发3、8批、理想16小时，未包含传输、样本差异与人工检查。GNU Time的单进程峰值不能替代实际重叠任务的总峰值；操作系统共享页和缓存也使简单相加只适合作为初始预算。

再测一组真实重叠任务

用两个代表样本在不同目录重叠运行，确认任务组内存、工作盘最高占用和整批结束时间。NoSharedMemory意味着每个作业自己的索引副本，不能假定索引自动只占一份。共享索引模式需要同索引、节点共享内存与清理安排，由环境维护者验证后再比较，不直接沿用私有模式数据。

同时观察读盘和写盘。更多并发可能让CPU等待存储，特别是BAM与干净FASTQ集中写到同一盘时。若增加并发后整批耗时没有改善，先检查存储与任务争用，不能把CPU忙碌程度当成收益。

资源与结果一起验收

确认每份任务退出成功，STAR日志中的输入量和比对分类可解释，BAM通过必要检查，计数和样本清单对应。失败或换了输出范围的记录单列。只有接近正式规模及计划并发通过复验，才将配置、租用窗口和盘容量写进运行方案。

若单任务已超过可用RAM，应先核对必要输出、参考和额度；确有容量缺口时再比较扩容。任务并发、软件或参考变化后保留旧记录并重新验证，不把一次测量变成对所有RNA-seq项目的保证。

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

STAR私有索引与输出参数 · 来源资料

GNU Time语义 · 来源资料

samtools quickcheck · 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。