比较排序资源组合并规划BAM处理服务器

用同一个BAM比较线程与每线程内存组合，分别记录排序耗时和结果验收，再测批量并发时的总资源。

比较 BAM 排序配置时，最有用的问题不是“哪组参数最快”，而是“在相同输入和资源边界下，哪组能稳定完成、临时盘可承受，并满足下游排序要求”。samtools sort 的 -m 是约每线程的内存参数，不是整项作业的硬上限；-@ 控制排序和压缩线程。配置比较必须把线程、每线程内存、临时目录和同时运行数一起记录。

先固定输入和比较条件

选一个具有代表性的 BAM，记录文件大小、记录数、samtools 版本和输入校验值。每组使用独立输出目录与临时前缀，输入文件、存储位置、压缩级别保持一致。若比较不同存储设备，应单独做一轮；不要同时改变线程、内存和磁盘，否则无法解释差异。重复运行时记下是否刚读过文件，系统缓存会影响耗时。

做小规模、可复核的参数对照

下面示例比较两组候选值，数值只是演示，必须按作业申请资源调整。每次只运行一组，避免测试本身争抢内存和 I/O。正式测试前确认输出目录不存在、临时盘有空间；任何一组失败都保留日志和退出状态，不把失败配置从比较表里删掉。

```
set -euo pipefail
: "${IN_BAM:?set representative BAM}"
: "${RUN_ROOT:?set new result directory}"
: "${TMP_ROOT:?set writable scratch directory}"
test -r "$IN_BAM" && test -d "$TMP_ROOT"
test ! -e "$RUN_ROOT"
mkdir -p "$RUN_ROOT"
samtools --version > "$RUN_ROOT/version.txt"
sha256sum "$IN_BAM" > "$RUN_ROOT/input.sha256"
samtools quickcheck -v "$IN_BAM"
for setting in "2 512M" "4 256M"; do
  read -r threads memory <<< "$setting"
  tag="t${threads}_m${memory}"
  out="$RUN_ROOT/$tag.bam"
  /usr/bin/time -v -o "$RUN_ROOT/$tag.time.txt" \
    samtools sort -@ "$threads" -m "$memory" \
      -T "$TMP_ROOT/$tag" -o "$out" "$IN_BAM" \
      2> "$RUN_ROOT/$tag.log"
  samtools quickcheck -v "$out"
  samtools view -c "$out" > "$RUN_ROOT/$tag.records.txt"
  samtools index "$out"
  du -h "$out" > "$RUN_ROOT/$tag.size.txt"
done
```

读结果时分开看速度、内存和临时盘

指标	怎么记录	怎么解释
耗时与峰值内存	time 输出、作业调度器记录	以作业实际资源统计为准
临时盘峰值	运行期间监控目标文件系统	结束后临时文件可能已删除，事后 df 看不到峰值
结果可用性	退出码、quickcheck、索引和下游抽查	记录数相同只是基本检查，不等于每条记录内容完全相同
稳定性	多次运行的成功率与波动	单次最快不代表日常最稳

把试跑结论转成作业配置

先排除失败和不满足下游要求的配置，再比较剩余选项。若提高线程让耗时下降但峰值内存超出作业限制，就不是可用配置；若降低 -m 后临时 I/O 明显增加，也要评估共享存储是否能承受。多样本并行时，把每个作业的实测峰值和并发数合并考虑，给系统及其他进程留出空间。

建议归档一张小表：输入校验值、软件版本、线程、-m、运行目录、耗时、峰值内存、临时盘峰值、退出码及输出检查。坐标排序结果才能按常规方式建立索引；name sort、tag sort 等特殊排序要按下游软件要求使用。试跑数据只能说明这份输入和当前存储环境，不能直接外推到不同样本和集群。

避免让基准测试误导配置选择

不同配置的结果要在同一资源配额和相近负载下比较。共享集群里若一组运行时节点空闲、另一组同时有多个任务，耗时差不能直接归因于参数；可以错开测试时间，或至少把节点负载和并发作业数记入结果表。临时文件创建后又被删除，单次结束后的磁盘用量无法代表峰值，需在运行时监测目标挂载点。还应检查输出头部的排序声明、索引能否建立，并用下游工具做小规模读取。校验值用于确认输入未变，记录数用于发现明显截断；两者都不能替代下游分析检查。最后选一组资源有余量、连续多次成功的参数，再将这组参数作为正式流程的起点，而不是把单次最快值直接推广到所有样本。

参考资料

samtools sort 官方手册（samtools 1.25） · 官方资料

samtools quickcheck 官方手册 · 官方资料

samtools 官方手册 · 官方资料