从CPU占用、内存峰值和I/O判断瓶颈

完整操作教程 · 参考固定 · 往返恢复验收

用同时间段的进程、内存和磁盘记录区分计算、内存与I/O瓶颈，再决定是否增加核心或内存。

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。

分析很慢时，“加内存”与“加核心”都可能是错误动作。程序可能在等待磁盘、只在一个阶段使用多线程，或者受到账号配额限制。先收集同一段运行时间内的进程、内存和I/O记录，才能把资源选择对应到真正的瓶颈。

下面使用只读命令观察可正常启动的任务。先处理输入或程序报错，再分析正常运行阶段的性能。

先记录任务所处阶段和实际额度

记下软件版本、当前命令、输入规模、线程选项和日志中的阶段。检查账号或作业被分到的CPU与内存，而不只看主机总量。容器或调度环境可能使用cgroup限制资源；内核文档明确区分系统资源与受控组的上限，主机内存很多并不保证你的作业能使用。

把任务进程号替换到下面命令中。pidstat、iostat通常来自sysstat工具集；若未安装，先向环境维护者确认，不需要为了监控随意改动整套软件环境。不同版本输出字段可能略有差异，可先查看本机帮助。

```
pid=12345  # ■■■■■■■■■■■■
ps -p "$pid" -o pid,ppid,etime,%cpu,rss,comm
grep -E 'VmRSS|VmHWM|VmSwap|Threads' "/proc/$pid/status"
grep -E 'MemTotal|MemAvailable|SwapTotal|SwapFree' /proc/meminfo
pidstat -u -r -d -p "$pid" 1 10
vmstat 1 10
iostat -xz 1 10
```

持续采样比只截一张图更有用。vmstat等工具的首份报告可能是自开机以来的平均值，判断当前任务时应看后续间隔报告。对派生多个子进程的流程，仅监控父进程可能漏掉主要工作，需结合进程树、工作流报告或作业级统计。

把内存压力与缓存占用区分开

Linux使用内存缓存文件是常见现象，空闲量小本身不能证明内存不足。结合MemAvailable、任务常驻内存、近期换页和任务日志判断。VmHWM可以辅助观察进程内存高水位，但不是所有并发进程内存的同时合计。

若持续出现内存压力、频繁换页、分配失败或确认的OOM，再检查是否存在不必要的对象复制、过高并发和稠密化操作。先修复合理范围内的浪费；必要步骤的工作集仍放不下时，增加可用内存才有明确依据。不要只因为曾经使用过swap，就认为当前慢一定来自换页。

同一阶段观察到的现象	先验证什么	可能的下一步
CPU忙，增加受支持线程能缩短耗时	是否真有并行收益、是否受配额限制	比较更多核心或更多独立任务并发
内存压力与换页持续，任务变慢或分配失败	对象表示、复制、并发与作业额度	减少冗余，确认后增加可用内存
CPU较空闲而读写等待增加	输入、临时盘、共享存储争用	优化I/O或工作盘，不直接加核心
只有部分阶段忙，其他阶段单线程	算法和软件本身的并行范围	按阶段配置，评估吞吐量而非只看占用
主机资源空闲但任务受到限制	容器、调度配额和进程限制	核对约定额度或调整运行方式

磁盘忙不能只看一个百分比

sysstat的iostat手册说明，await包含请求排队与服务时间；需要结合队列、吞吐量、设备特征和任务阶段看变化。对于能够并行处理请求的现代SSD或RAID，%util不能单独代表性能极限。看到某个百分比很高，就断言必须换盘或加CPU，证据都不充分。

可以先保持同一输入和并发，在可用且获允许的不同工作盘上比较；也可以暂时错开其他备份或压缩任务。若这样显著改变该步骤耗时，说明下一轮应重点核查I/O条件。不同存储的缓存、网络路径和容量都要记录，不能只写“SSD更快”。

将监控记录转成一次可验证的调整

每次只改一个因素，例如线程数、独立任务并发，或者工作盘。保留前后日志与结果验收。若增加线程后CPU更忙但耗时没有改善，可能出现争用或并行开销；这不能作为继续增加核心的证据。

用记录支持一句具体判断，例如“必要阶段仍有内存容量缺口”。若改变一个因素后现象没有改善，应回到日志重新检查假设。

参考与版本核对

用进程与系统两种视角定位瓶颈

资源排查要同时观察进程级和节点级数据。进程级可看CPU百分比、RSS、线程数和运行时间；节点级看可用内存、swap、磁盘吞吐与等待、CPU iowait及调度器配额。单次ps快照不能还原峰值，建议用固定间隔采样或读取调度器/cgroup作业统计。记录采样间隔和任务PID，避免把其他用户作业的资源占用算到本任务。

在多进程流程中，主进程RSS未必覆盖所有子进程，GNU time的统计口径也要看具体调用方式。容器或作业被cgroup限制时，主机free显示的总量不能代表任务能使用的上限。若进程被OOM killer终止，核对内核或调度器事件；若磁盘满，核对对应挂载点而不是只看另一个盘的剩余空间。

观察信号	可能原因	下一步验证
CPU高、磁盘等待低	计算阶段受CPU限制	固定输入测试线程边际收益
CPU低、iowait高	输入/输出盘吞吐受限	查看iostat和文件所在挂载点
内存接近额度且OOM	工作集超过作业限制	核对cgroup峰值、并发和日志阶段
内存尚有但进程退出	容器配额、临时盘或程序错误	查作业事件、退出码和stderr
swap持续增长	内存压力导致换页	降低并发或检查不必要的大对象

把采样和日志时间对齐

基准开始前记录任务ID、PID、命令行、运行目录和挂载点；运行期间每隔固定时间采集进程及I/O状态，结束时保存退出码和最后日志。若工具分阶段运行，把时间戳与阶段边界对应起来。这样可以判断峰值出现在索引加载、矩阵构建、排序还是写出阶段。没有阶段日志时，至少用标准错误输出或调度器事件定位失败前的状态。

若使用pidstat、iostat等工具，确认系统已有对应包并记录版本；不要为采样而无授权安装系统软件。没有工具时可用/proc和任务平台指标，但说明采样频率和范围。短时间峰值可能被低频采样漏掉，因此OOM事件比间歇观察free更有判定价值。

判断该加什么资源

CPU利用率持续饱和且工作量可并行时，固定其他条件逐步增加线程，记录总耗时是否继续下降。若增加线程后速度几乎不变，检查串行阶段、I/O和内存带宽。内存问题要看实际峰值和失败证据；如果算法对象本身大，减少并发可能比增加核心更有效。I/O问题先检查数据位置、文件系统和临时盘是否拥塞，必要时用相同数据在本地高速盘试跑。

资源调整之后复跑同一输入，记录从启动到完成的时间及输出检查。只缩小数据规模后成功不能证明正式任务已解决；小样本可作为诊断，但仍要在真实代表性规模下验证。若改变了参数或过滤步骤，结果可能不再可比，应将科学设置变化与资源变化分开记载。

生成可复用的瓶颈报告

简短报告包括软件版本、输入规模、命令、资源额度、峰值数据及来源、瓶颈证据、已尝试调整和结果是否通过。使用“观察到CPU持续满载”这类证据表达，不用“服务器太慢”等无法核实的判断。报告同时说明下一步建议和不确定性，使课题组可以决定增加CPU、内存、I/O能力还是改变排期。

一个可复用的排查顺序

任务变慢时先保存当前命令和最后一段日志，确认进程仍在运行且没有等待输入；随后观察CPU、内存、I/O和调度器事件。CPU持续高占用且磁盘等待低时，测试线程数；CPU低而磁盘等待高时，将同一输入复制到合规的本地高速盘复测；内存峰值触及配额或有OOM事件时，先降低并发或拆分流程。每次只改变一项，保留前后日志和输出统计。

如果监控显示CPU、内存和磁盘都不高，检查程序是否在等待网络、许可证、锁文件或人工交互。反复增加核心通常无法解决等待问题。对于容器任务，核对CPU quota和memory.max；对于调度作业，查看实际申请和退出原因。主机总体监控只能作背景，不可替代作业级数据。

避免采集结果被误读

多个任务并行时，分别标记进程和项目目录，并把同一时刻的节点总量与单作业用量分开保存。RSS不能简单相加为精确物理内存，因为共享页可能重复计数；调度器或cgroup报告通常更适合判断作业额度。I/O等待高也可能来自其他用户共享文件系统，复测时记录节点负载和数据位置。

排查报告给出观察窗口、采样频率、软件版本、输入规模和统计工具。若工具只能看到瞬时值，注明可能漏掉短时峰值。失败任务的stderr、退出码和OOM事件比“看起来内存够”更有证明力。

例如某任务在读取阶段CPU不高、磁盘等待升高，而将同一输入放到本地工作盘后耗时下降，证据支持I/O限制。若只看CPU平均利用率，可能错误地增加核心。把这种前后对照、输入校验和与存储路径一并保留，之后才能判断该结论是否适用于其他样本。

建立一张每次排查都能复用的记录表：任务ID、输入校验和、程序版本、节点、作业额度、采样时间、日志路径、失败阶段、证据、调整项和复测结果。记录使同一问题可以跨用户交接，也能避免同一个项目重复尝试已经失败的配置。

在条件允许时，对重要结论做一次复测，并保留复测时间和节点负载。若资源指标波动较大，应报告波动范围并延长观察，而不是挑选最好的一次作为标准。正式数据分析的容量申请以代表性运行和已确认额度为依据。

问题关闭前安排复核人确认调整后的任务正常完成、输出可读且资源峰值符合额度。把未解决的风险和适用范围写进交接，避免之后把一次性诊断结果当作全流程结论。

将资源结论写成有条件的建议：说明已测试的数据规模、软件版本、任务并发和结果验收范围。对于未覆盖的输入或环境变化，标注需要复测。不要把一次短时间成功等同于稳定运行，也不要仅凭设备规格推断正式任务耗时。这样其他同学可以按相同条件复核，并判断何时需要更新资源计划。

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

Linux内核：/proc文件系统与内存统计 · 来源资料

Linux内核：cgroup v2资源控制 · 来源资料

sysstat官方：iostat手册 · 来源资料

sysstat官方项目与工具清单 · 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。