按参考库和线程试跑匹配大内存分类环境

固定GTDB-Tk参考和基因组集合，以独立新进程比较pplacer线程，检查是否真的进入树放置阶段，再形成容量记录。

GTDB-Tk classify_wf 的内存峰值主要与参考树和 pplacer 阶段有关，增加 --cpus 不能保证缩短总耗时，调大 pplacer 并行度还会改变资源表现。GTDB-Tk 2.8.0 官方安装说明给出的参考资源约为：古菌 100 GB，细菌 140 GB；使用 --full_tree 的细菌分类需要 950 GB 以上。官方 FAQ 还提醒，pplacer 的虚拟内存报告可能大于实际物理占用，因为多个进程通过 copy-on-write 共享内存。

先确认任务模式和输入规模

记录细菌与古菌基因组数量、reference release、是否启用 --full_tree 或 --place_species、CPU 和作业内存。classify_wf 的 ANI screen 可直接给部分基因组分配物种分类，默认不再把这部分放入参考树；开启 place_species 会增加树放置工作。不要因为系统监控里 VIRT 很高，就直接认定物理 RAM 已同等耗尽。

现象	看哪类指标	可尝试的方向
pplacer 被内存限制终止	RES、cgroup/调度器峰值、pplacer_cpus	先减少 pplacer 并行度或批次规模
VIRT 很高但物理内存余量充足	RES 与节点实际内存，不只看进程 VIRT 求和	按 FAQ 的 copy-on-write 说明核实
内存仍不足	full_tree、scratch、并发任务	评估 --scratch_dir 或合适的内存节点
分类耗时很长	步骤日志、CPU、I/O、pplacer 阶段	先定位瓶颈，避免盲目加总线程

对照试跑并测量物理峰值

选一批具有代表性的 MAG，固定软件、reference release 和模式，先用较低 pplacer 并行度运行，记录阶段耗时、进程 RES、调度器峰值和临时盘。不要用虚拟内存总和简单相加估算物理需求；集群 cgroup 或调度器统计更适合用于作业申请。不同基因组数、域和 tree 模式会改变运行量，单次试跑不能替代后续资源监控。

```
set -euo pipefail
: "${GENOMES:?set FASTA genome directory}"
: "${OUT:?set new output directory}"
: "${THREADS:?set allocated CPU count}"
: "${PPLACER_THREADS:?set pplacer CPU count}"
test -d "$GENOMES"
test ! -e "$OUT"
gtdbtk classify_wf --genome_dir "$GENOMES" \
  --out_dir "$OUT" --cpus "$THREADS" \
  --pplacer_cpus "$PPLACER_THREADS" \
  --scratch_dir "$OUT.scratch"
```

scratch_dir 可通过写盘降低 pplacer 内存需求，但会变慢并增加临时存储压力；先确认目录所在文件系统容量和 I/O 能力。示例没有开启 full_tree 或 place_species。正式分析要依据研究目标决定是否需要这些模式，而非为了输出更多列默认开启。

根据证据确定资源而非照抄推荐数字

把参考数据库安装空间和运行时 RAM、scratch 空间分开规划；官方安装页同时给出约 100 GB reference data 存储以及不同分类模式的内存需求。若作业需多批运行，使用相同版本与 release，记录每批输入 ID 和输出，避免 ID 冲突。最终通过 summary 文件和日志核对所有预期基因组均完成分类，再结合基因组质量评估解释结果。

安排批次时也要核算峰值

把完整 classify_wf 拆成批次可能降低单次处理的输入规模，但每批使用相同 reference tree，不能据此把 reference 相关的基础内存压到零。建议从几十个高质量代表基因组开始，确认 identify、align、classify 各阶段均正常，再扩大批次。样本数量增加后记录总峰值和耗时，并考虑同一节点是否还有 pplacer 或其他大内存进程。

若分类只需要 species-level 结果，可先确认默认 ANI screen 是否已足够；需要把 ANI 命中的基因组也放入系统树时再考虑 --place_species，因为这会增加 pplacer 工作和内存。若因内存选择 scratch_dir，比较实际磁盘峰值与速度，不要只看 RAM 使用下降就认为配置更优。归档 summary、运行日志、reference release、输入 ID 清单和资源数据后，再将同一策略应用到下一批。

参考资料

GTDB-Tk 2.8.0 官方安装与硬件要求 · 官方资料

GTDB-Tk classify_wf 官方文档 · 官方资料

GTDB-Tk 官方 FAQ：pplacer 内存问题 · 官方资料

GTDB-Tk check_install 官方文档 · 官方资料