PR #1764 完整报告

THUDM/slime

Add host memory metrics to available_memory function

合并时间：2026-04-03 11:52

原文链接：http://prhub.com.cn/THUDM/slime/pull/1764

执行摘要

本 PR 在 slime/utils/memory_utils.py 的 available_memory 函数中添加了主机内存指标（总内存、可用内存、已用内存、空闲内存），通过引入 psutil 库增强了系统资源监控的完整性。这是一个低风险的功能扩展，主要影响资源诊断和监控工具，建议关注依赖管理和字段命名一致性。

功能与动机

- 动机：从 PR 标题和代码变更推断，目的是扩展内存监控功能，使用户能够同时查看 GPU 和主机内存状态，便于全面了解系统资源使用情况，辅助调试和优化。
- 背景：PR body 和关联 Issue 为空，但结合仓库上下文（如近期 PR #1768 和 #1776 也涉及指标增强），可见团队正在持续提升系统的可观测性。

实现拆解

变更仅涉及一个文件，按模块拆解如下：

模块	文件	关键改动	说明
utils	slime/utils/memory_utils.py	1. 导入 psutil 库 2. 在 available_memory 函数中添加 psutil.virtual_memory() 调用 3. 返回字典新增四个字段：host_total_GB、host_available_GB、host_used_GB、host_free_GB	通过辅助函数 _byte_to_gb 将字节转换为 GB 单位，保持输出格式一致。

关键代码逻辑示例（基于 patch_excerpt）：

```
def available_memory():
    device = torch.cuda.current_device()
    free, total = torch.cuda.mem_get_info(device)
    vm = psutil.virtual_memory() # 新增：获取主机内存信息
    return {
        "gpu": str(device),
        "total_GB": _byte_to_gb(total),
        "free_GB": _byte_to_gb(free),
        "used_GB": _byte_to_gb(total - free),
```

```
        "allocated_GB": _byte_to_gb(torch.cuda.memory_allocated(device)),
        "reserved_GB": _byte_to_gb(torch.cuda.memory_reserved(device)),
        "host_total_GB": _byte_to_gb(vm.total), # 新增字段
        "host_available_GB": _byte_to_gb(vm.available), # 新增字段
        "host_used_GB": _byte_to_gb(vm.used), # 新增字段
        "host_free_GB": _byte_to_gb(vm.free), # 新增字段
    }
```

评论区精华

- 无 review 评论或讨论记录，变更直接合并，表明可能被视为简单、低争议的改进。
- 提交历史显示作者进行了三次 merge 操作（如提交 177c98d、f0a3d4c、c2242e5），可能为了同步主分支变更，但未引发额外讨论。

风险与影响

- 技术风险：
  - 新增 psutil 依赖：若未在项目依赖（如 requirements.txt 或 setup.py）中声明，可能导致导入错误。
  - 返回结构变更：添加了新字段，但未修改现有字段，因此向后兼容性较好；不过下游代码若严格依赖字典键值，可能需要适配。
  - 跨平台兼容性：psutil.virtual_memory() 在主流操作系统上行为一致，但极端环境下可能有差异。
- 影响评估：
  - 用户：获得更全面的内存数据，有助于资源监控和问题诊断，但需确保环境已安装 psutil。
  - 系统：轻微增加函数执行开销（调用 psutil），通常可忽略不计。
  - 团队：变更简单易维护，但建议更新相关文档（如工具使用说明）以反映新字段。

关联脉络

- 与历史 PR 的关联：
  - PR #1768（修复 wandb 指标上传）：同属 metrics 领域，都关注系统监控数据的收集与上报。
  - PR #1776（新增 trace 时间线查看器）：反映仓库在增强可观测性方面的持续趋势，本 PR 是这一方向的延续。
- 演进趋势：近期多个 PR（如 #1760、#1769、#1776）涉及功能扩展和监控增强，表明团队正积极完善系统的多模态支持、性能优化和诊断工具。本 PR 作为其中一环，强化了基础监控能力，为后续更复杂的资源管理特性奠定基础。