

服务器硬件&运行状态汇总报告

报告时间：2026-07-09 14:57:57 CST

系统版本：Ubuntu 22.04.5 LTS x86_64

一、CPU 核心配置

项目	参数详情
处理器型号	Intel Xeon Gold 6530 双路 CPU
物理插槽	2 颗
单颗规格	32 核心 64 线程
总逻辑 CPU	128 核 (0~127)
主频范围	最小 800MHz / 最大 4000MHz
缓存	L1:3MiB+2MiB; L2:128MiB; L3:320MiB
NUMA 架构	2 个 NUMA 节点 Node0: CPU 0-31、64-95 Node1: CPU 32-63、96-127
虚拟化支持	开启 VT-x
安全漏洞	所有高危漏洞均已修复 / 不受影响

二、内存配置与使用

1. 整体内存概况

- 物理总内存：**251GiB**
- 已使用：21GiB | 空闲：3.3GiB | 缓冲缓存：226GiB | 可用内存：**227GiB**
- Swap 交换分区：2GiB，已用 969MiB

2. 内存硬件插槽（DDR5 ECC 5600MT/s，降频运行 4800MT/s）

整机共**4 条 64GB DDR5 ECC 内存**，合计 256GB（系统识别 251GiB 为正常进制换算差异）

内存位置	容量	状态
P0_CHA_D0	64GB	已插
P0_CHE_D0	64GB	已插
P1_CHA_D0	64GB	已插
P1_CHE_D0	64GB	已插
其余全部插槽空置，支持后续扩容。		

三、GPU 显卡阵列（4 张 NVIDIA RTX 6000D）

基础参数

驱动版本：580.95.05 | CUDA 工具链：13.0
单卡标配显存：**85651MiB $\approx$ 83.6GB** | 功耗上限 450W

GPU 编号	PCIe 版本	当前链路	显存已用	温度	实时功耗	GPU 占用
GPU0	PCIe5.0	x16	500MiB	45°C	80.81W	0%
GPU1	PCIe5.0	x16	17MiB	35°C	30.71W	0%
GPU2	PCIe5.0	x16	17MiB	35°C	32.42W	0%
GPU3	PCIe5.0	x16	17MiB	34°C	29.90W	0%

GPU 拓扑绑定

4 张 GPU 全部隶属于 **NUMA Node1（CPU 32\63、96\127）**，跨 NUMA 访问会有性能损耗，推荐进程绑定对应 CPU 核。
无 NVLink 互联，仅 PCIe 通信。

正在占用 GPU 进程

- Xorg 桌面进程：四张卡均挂载，单卡占用极小显存
- gnome-remote-desktop-daemon（GPU0）：占用 468MiB 显存，远程桌面服务

四、硬盘存储

两块 Solidigm 3.5TB NVMe SSD

1. nvme1n1（系统盘）

- / 根目录：3.5T，已用 531G，使用率 16%
- /boot/efi 引导分区：512M

2. nvme0n1（数据盘）

- 挂载路径 `/mnt/nvme1n1`，几乎未使用

五、Docker 监控容器（全部正常运行）

容器名称	端口	用途	资源占用
prometheus	9090	指标采集	CPU1.39%，内存 172.8MiB
grafana	3000	可视化面板	CPU0.64%，内存 237.6MiB
loki	3100	日志存储	CPU0.36%，内存 86MiB
promtail	-	日志采集	CPU0.88%，内存 47.37MiB
node-exporter	9100	主机硬件监控	CPU7.81%，内存 25MiB
dcgm-exporter	9400	GPU 指标导出	CPU0.10%，内存 101.8MiB

六、本地大模型目录（/home/www/models）

1. DeepSeek-V4-Flash-DSpark

完整 48 分片权重文件，模型文件齐全

2. Qwen3.5-122B-A10B-FP8

完整 39 分片 FP8 量化权重

3. Unlimited-OCR

OCR 识别单文件模型权重

七、可补充安装工具

1. `numactl`：查看 / 管控 NUMA 绑定
2. `nvme-cli`：查看 NVMe 硬盘健康、温度、写入量
3. `ollama`：轻量化大模型运行工具

八、整体状态总结

1. **硬件配置拉满**：双路 128 线程 CPU、256GB ECC 内存、4×83GB 超大显存专业卡、双 3.5TB 高速 NVMe；
2. **负载极低**：CPU 无业务负载，GPU 纯待机，内存大部分为文件缓存，硬件资源富余；
3. **运维体系完善**：全套 Prometheus+Grafana+Loki+DCGM 监控栈已部署，可实时监控主机与 GPU；
4. **模型资产完备**：已预装 DeepSeek V4、通义千问 122B FP8、OCR 多套大模型权重，可直接部署推理服务。

（注：部分内容可能由 AI 生成）