**满血部署Deepseek V3/R1的服务器方案**

一、Deepseek V3/R1模型的综合研制成本估算

1、固定资产投入：按10000块H800计（公开信息是至少10000块A100和2000块H800），总投入至少19-20亿元。

2、年费用：资产折旧即使按10%计为2亿元左右，电费按20%利用率约1亿元，人力成本约1亿元，运维及其他成本约0.5亿元。

结论：我们拿到的免费Deepseek V3/R1模型，其研制综合成本至少为10-15亿元水平。

二、部署Deepseek V3/R1 FP8满血模型的方案

表1 各厂家方案比较情况

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 厂家 | GPU服务器 | 首字延时 | 推理速度 | 推理吞吐 | 训练可行性 | 造价水平 |
| 1 | 英伟达 | 1\*H200(8 GPU) | 工程优秀 | 工程优秀 | 工程优秀 | 优秀 | 250 |
| 2 | 2\*H20(8 GPU) | 工程可用 | 工程优秀 | 工程可用 | 可行 | 250 |
| 3 | 2\*H100(8 GPU) | 工程优秀 | 工程优秀 | 工程优秀 | 优秀 | 460 |
| 4 | AMD | 1\*MI300(8 GPU) | 工程优秀 | 工程优秀 | 工程优秀 | 生态不行 | 250 |
| 5 | 华为 | 4\*910B(8 GPU) | 工程可用 | 生态不行 | 生态不行 | 生态不行 | 600 |
| 6 | Apple | 8\*M4Ultra(80核) | 工程不可用 | demo可用 | 工程不可用 | 生态不行 | 50 |
| 7 | 其他 | 清华ktransformer | 工程不可用 | 个人可用 | 工程不可用 | 不可行 | 10 |

注：清华ktransformer方案包括1\*4090+1-2T内存。

附：GPU关键参数情况

附表1 GPU算力、卡内带宽、卡间带宽、节点间带宽对推理、训练的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 训练时长 | 推理出首字时间 | 推理速度 | 多用户推理吞吐 |
|  | (月) | (s) | (tokens/s) | (tokens/s) |
| GPU算力 | 完全反比 | 完全正比 | 无关 | 大致正比 |
| 卡内带宽 | 较大影响 | 无关 | 完全正比 | 较大影响 |
| 卡间、节点间带宽 | 极大影响 | 无关 | 较大影响 | 较大影响 |
| H200典型水平 | 1个月 | 3 | 80 | 3000 |
| H20典型水平 | 7个月 | 21 | 60 | 1000 |
| 910B典型水平 | 5-6个月 | 15 | 10 | 1000 |

附表2 GPU算力、卡内带宽、卡间带宽等参数情况

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 5090 | 4090 | 3090ti | 3090 | L40s | 6000ADA | A100SXM | H20SXM | H100SXM | H200SXM | 910B | m4max |
| 显存 | 32 | 24 | 24 | 24 | 48 | 48 | 80 | 96 | 80 | 141 | 64 | 128 |
| TF32 |  | 83 | 40 | 36 | 183 | 91 | 156 |  | 495 | 495 |  |  |
| FP16 Tensor | 419 | 330 | 160 | 142 | 362 | 340 | 312 | 148 | 990 | 990 | 320 | 150 |
| INT8 Tensor |  | 660 | 320 | 284 | 733 |  | 624 |  | 1979 | 1979 |  |  |
| INT4 Tensor |  | 1321 | 640 | 568 | 733 |  | 1248 |  |  |  |  |  |
| 卡内带宽 GB/s | 1792 | 1008 | 1008 | 936 | 864 | 960 | 2039 | 4000 | 3350 | 4800 | 2000 | 1000 |
| 卡间带宽 GB/s |  |  |  |  |  |  | 600 | 900 | 900 | 900 | 56 |  |
| 支持FP8 |  |  |  |  |  |  |  | 是 | 是 | 是 |  |  |

注：m4max为最新Apple的GPU产品