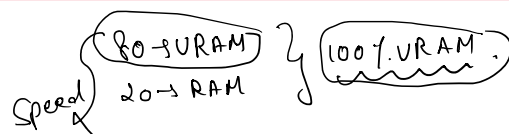
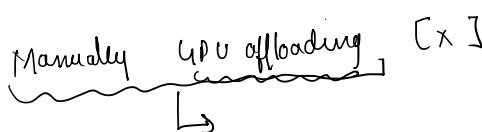
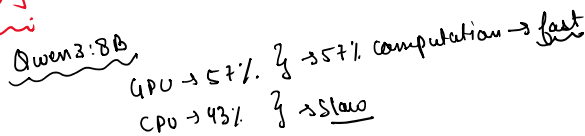
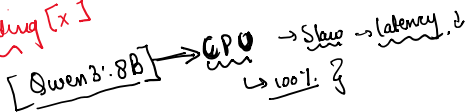
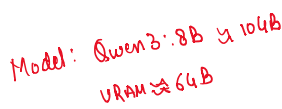
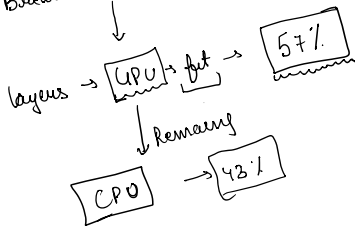
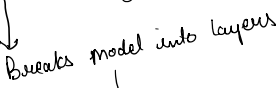
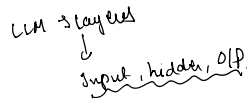
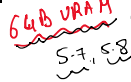
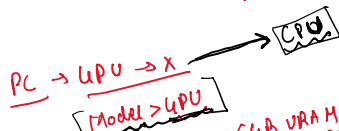
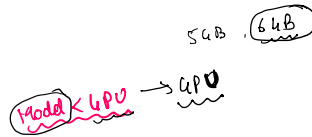
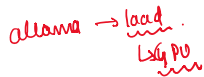
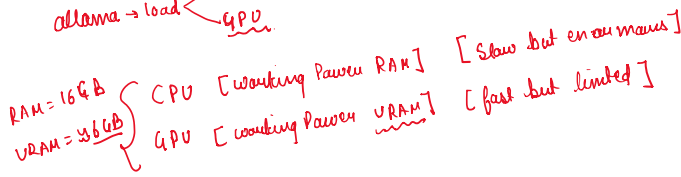


GPU OFFLOADING



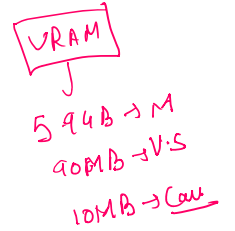
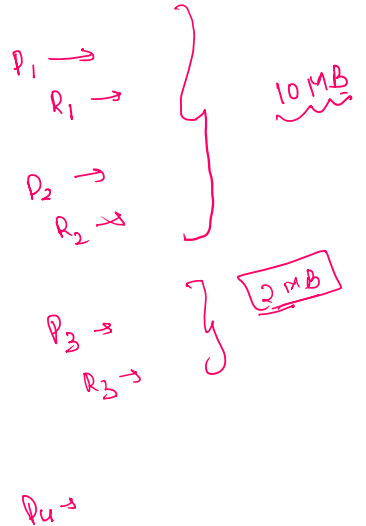
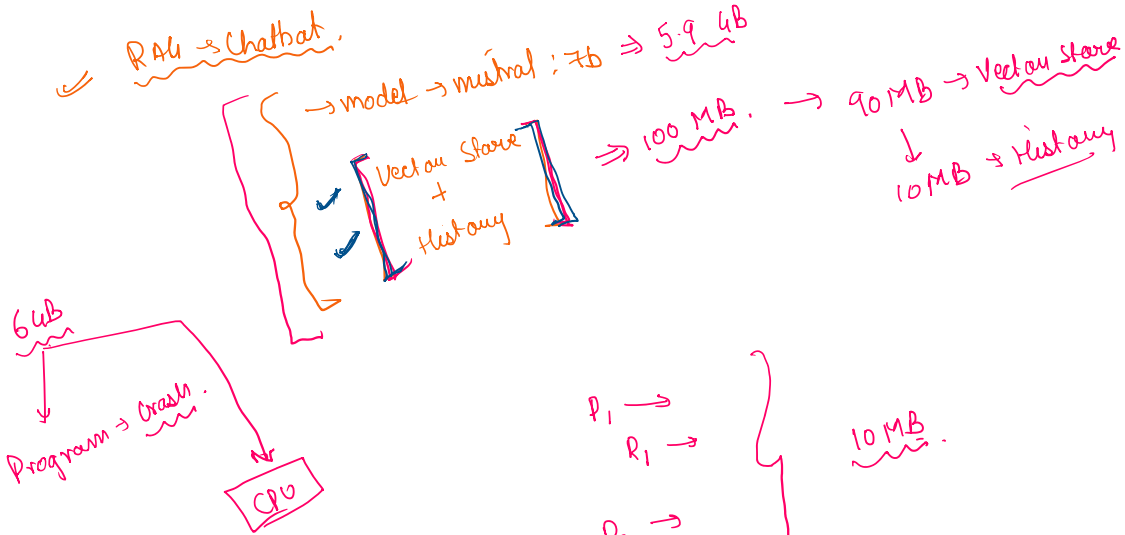
... output

LLM → VRAM

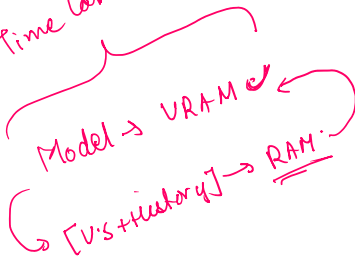
- model [weight & bias]
- KV Cache → Information about current conv;
- History, U.D.

RAM → Chatbot.

64B VRAM.



Time Latency [latency].



VRAM → 5.94B.

[History + V.S.] → RAM

manually 4PU offloading.

VRAM → 14B → storage free

