GPU VRAM Blueprint

Estimating GPU Memory for Large Language Model Inference

How Much VRAM for Model Weights?

The first step is calculating the storage for model weights at different precisions. Lower precision drastically reduces the memory footprint.

7 Billion
Parameters

FP16 / BF16	~14 GB
INT8	~7 GB
INT4	~3.5 GB

13 Billion
Parameters

FP16 / BF16	~26 GB
INT8	~13 GB
INT4	~6.5 GB

30 Billion
Parameters

FP16 / BF16	~60 GB
INT8	~30 GB
INT4	~15 GB

65 Billion
Parameters

FP16 / BF16	~130 GB
INT8	~65 GB
INT4	~32.5 GB

Beyond Weights: The Real VRAM Consumers

Model weights are just the baseline. Real-world inference requires budgeting for several other critical components.

KV Cache

+100% or More

Memory to store attention keys/values. This is the largest overhead and grows with batch size and context length. **This can easily double your VRAM requirements.**

CUDA & System Overhead

+10-20%

The CUDA kernels, framework libraries (PyTorch), and various buffers all consume a baseline amount of VRAM just by being loaded.

Model Activations

+1-2%

Intermediate calculations stored during the forward pass. While smaller than other factors, it still contributes to the total memory load.

The Real-World VRAM Equation

Model Weights

+

KV Cache

+

Overhead

=

Total VRAM Needed

Key Takeaway: A 7B parameter model may only need 14GB for its weights (FP16), but with a large context window for the KV Cache, the **actual requirement can easily exceed 24GB**. Always profile, don't just calculate!

---

vLLM Course Content | Built for Delivery Engineers & Consultants