MASTERING LLM PRESENTS

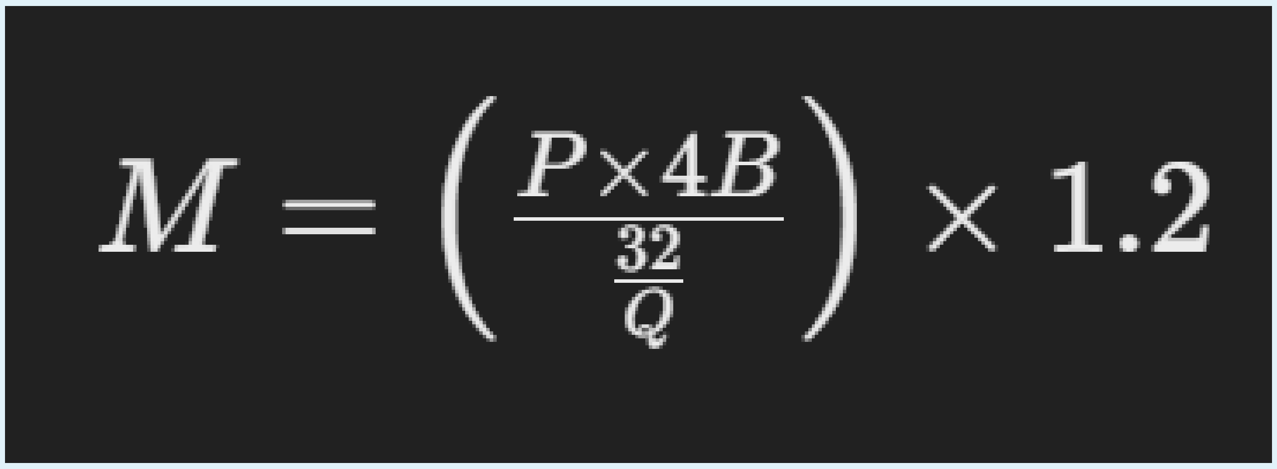
COFFEE BREAK CONCEPTS

How Much GPU Memory is Needed to Serve a Large Language Model (LLM)?

Follow us on

The Formula to Estimate GPU Memory

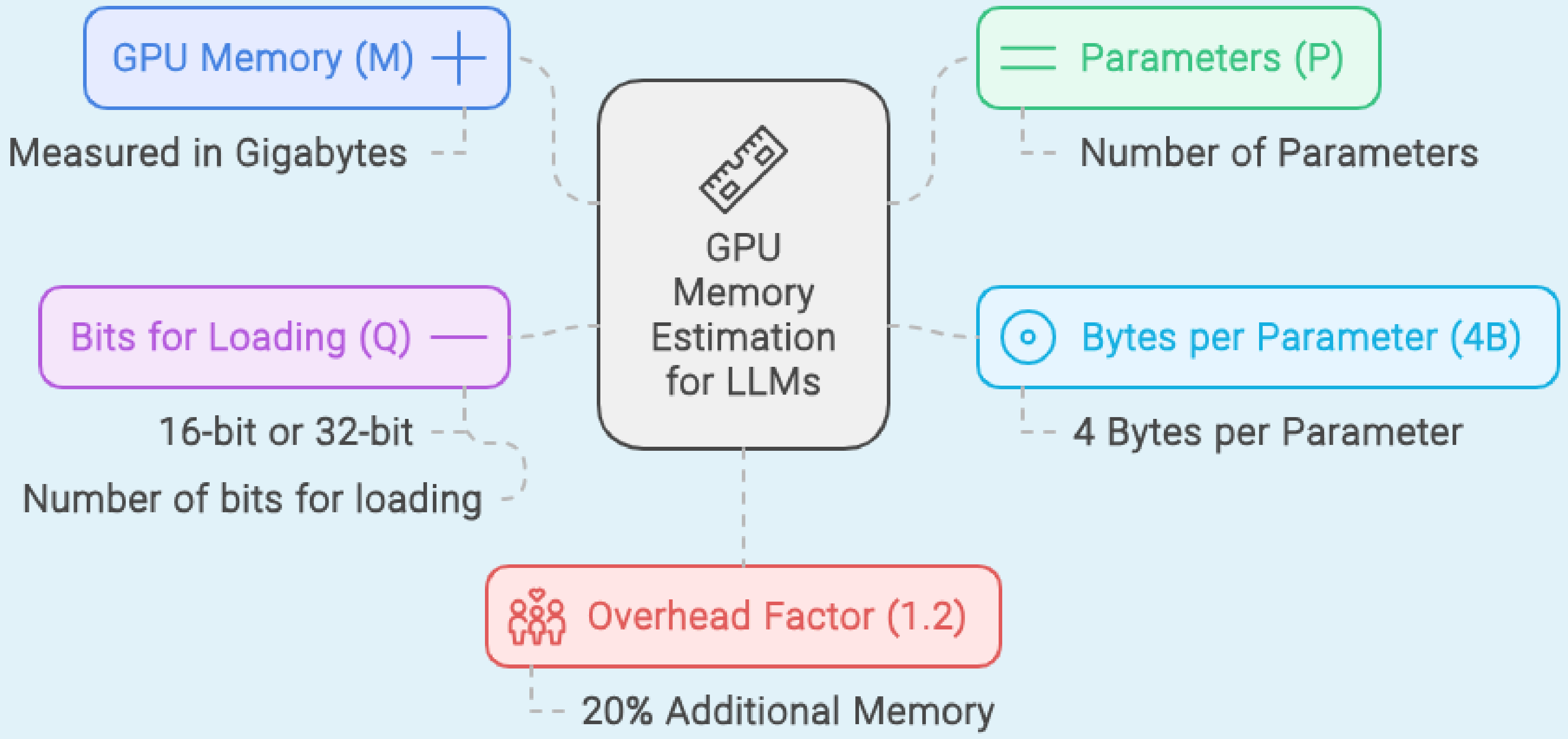
- To estimate the GPU memory required for serving a Large Language Model, you can use the following formula:



$$M = \left(\frac{P \times 4B}{\frac{32}{Q}} \right) \times 1.2$$

  - **M** is the GPU memory in Gigabytes.
  - **P** is the number of parameters in the model.
  - **4B** represents the 4 bytes used per parameter.
  - **Q** is the number of bits for loading the model (e.g., 16-bit or 32-bit).
  - **1.2** accounts for a 20% overhead.

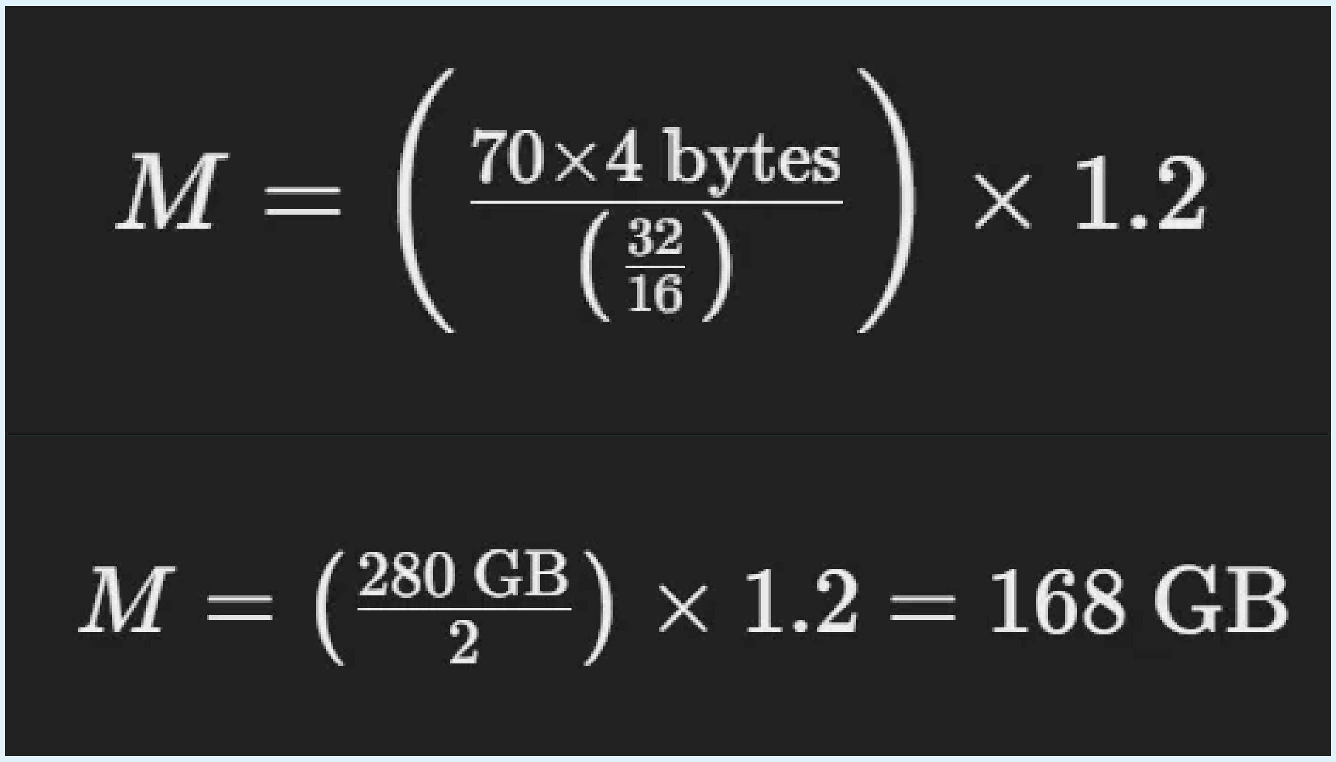
Breaking Down the Formula



- **Number of Parameters (P):** This represents the size of your model. LLaMA 70 billion --> 70 billion parameter
- **Bytes per Parameter (4B):** Each parameter typically requires 4 bytes of memory.
- **Bits per Parameter (Q):** Depending on whether you're loading the model in 16-bit or 32-bit precision, this value will change.
- **Overhead (1.2):** The 1.2 multiplier adds a 20% overhead to account for additional memory used during inference. This isn't just a safety buffer; it's crucial for covering the memory required for activations and other intermediate results during model execution.

Example Calculation

- Let’s consider you want to estimate the memory required to serve a LLaMA model with 70 billion parameters, loaded in 16-bit precision



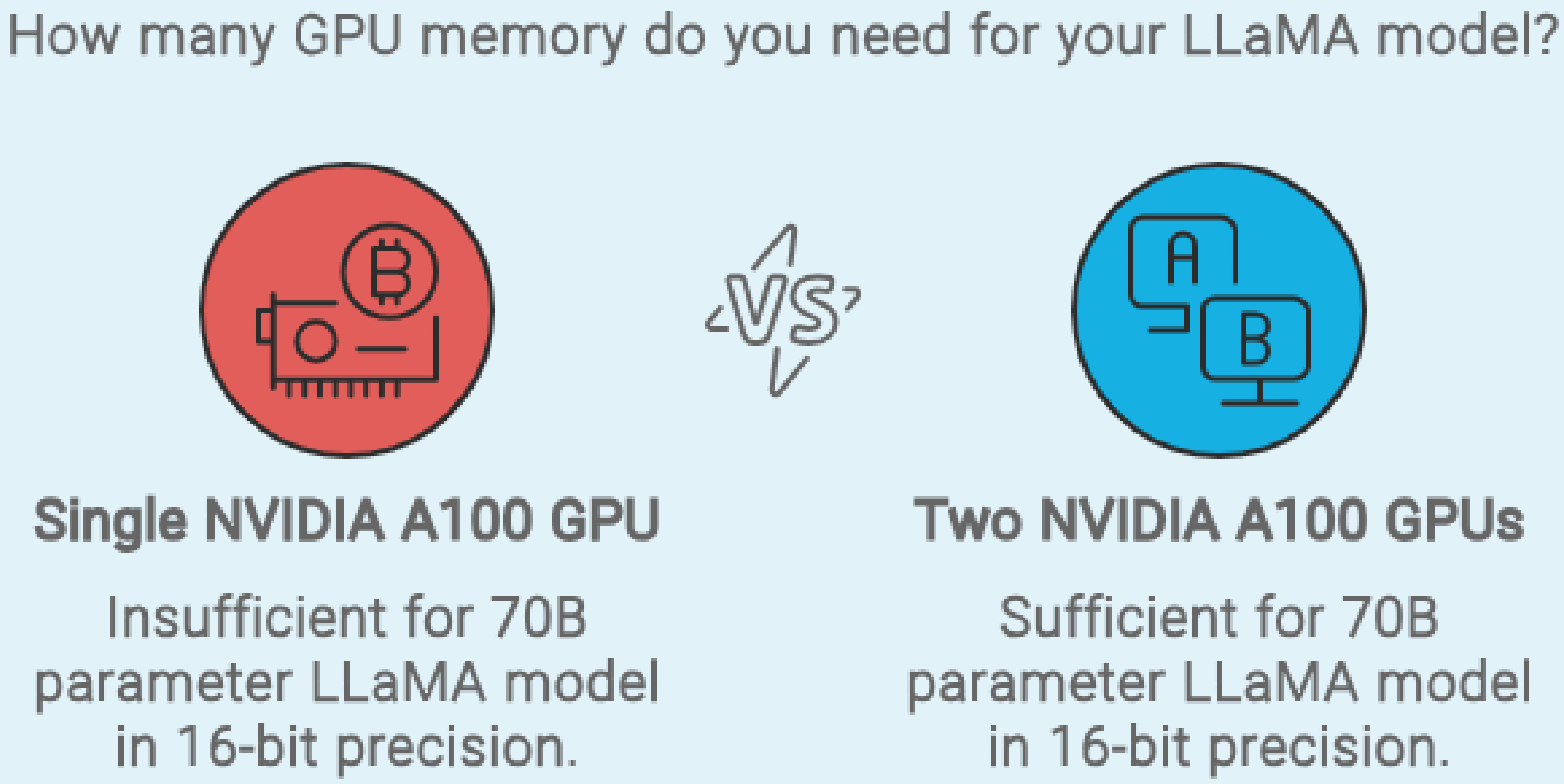
$$M = \left(\frac{70 \times 4 \text{ bytes}}{\left(\frac{32}{16}\right)} \right) \times 1.2$$

$$M = \left(\frac{280 \text{ GB}}{2}\right) \times 1.2 = 168 \text{ GB}$$

This calculation tells you you would need approximately **168 GB of GPU** memory to serve the **LLaMA model with 70 billion parameters in 16-bit mode**.

Practical Implications

- The calculation helps identify sufficient GPU to serve a model.
- This also helps handle the memory load of GPU efficiently.



How many GPU memory do you need for your LLaMA model?

Single NVIDIA A100 GPU

Insufficient for 70B parameter LLaMA model in 16-bit precision.

Two NVIDIA A100 GPUs

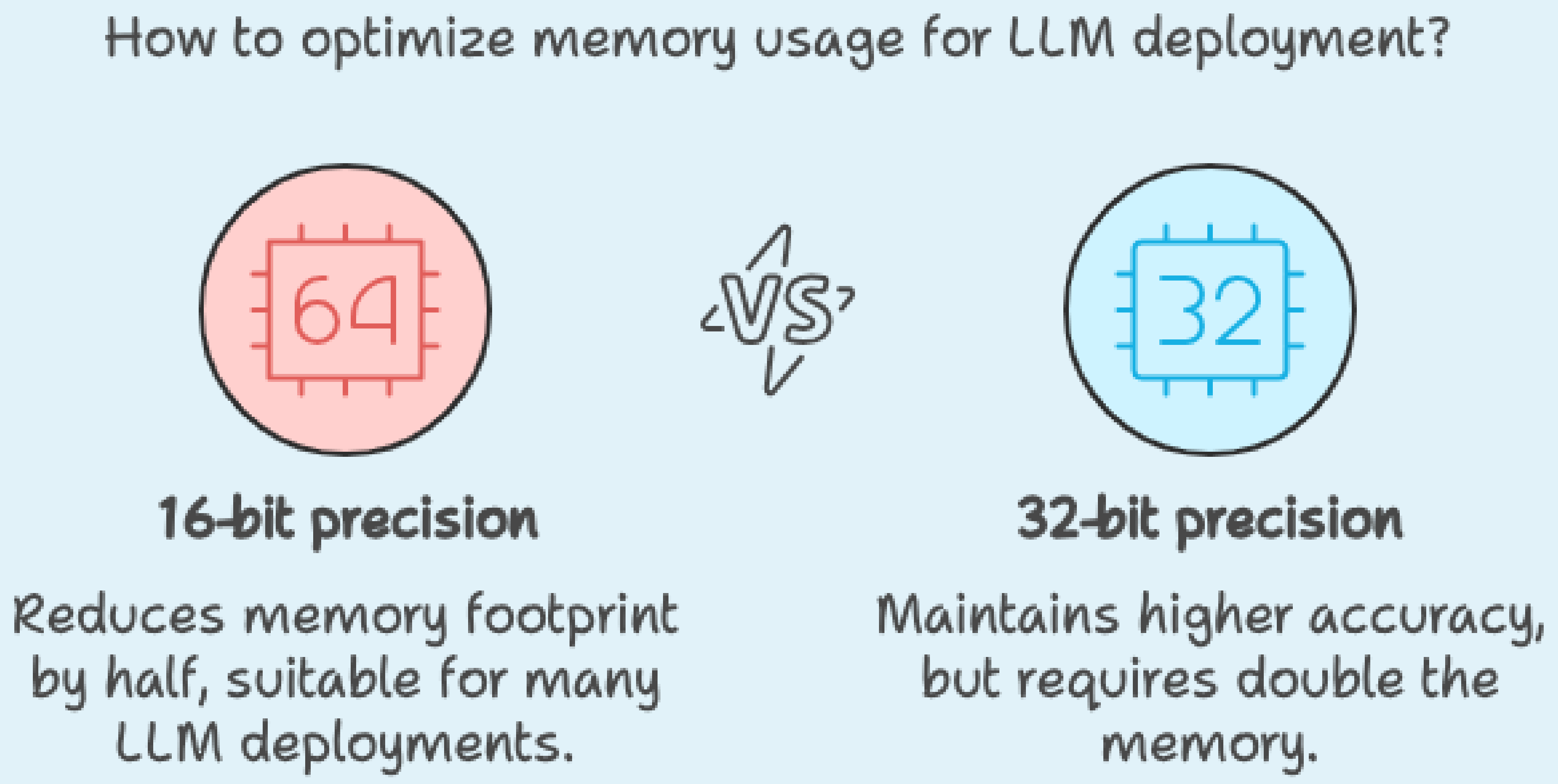
Sufficient for 70B parameter LLaMA model in 16-bit precision.

Summary

☑ **Use the formula to estimate approximate GPU memory require to infer a model.**

☑ **Larger models with more parameters require significantly more GPU memory, making accurate estimation essential for efficient deployment.**

☑ **Loading models in 16-bit precision (half-precision) reduces memory usage compared to 32-bit precision**



How to optimize memory usage for LLM deployment?

16-bit precision

Reduces memory footprint by half, suitable for many LLM deployments.

32-bit precision

Maintains higher accuracy, but requires double the memory.

www.masteringllm.com

AgenticRAG with LlamaIndex

Want to learn why AgenticRAG is future of RAG?

- ✓ Master **RAG fundamentals** through practical case studies
- ✓ Understand how to overcome **limitations of RA**G
- ✓ Introduction to **AgenticRAG** & techniques like **Routing Agents, Query planning agents, Structure planning agents, and React agents with human in loop**.
- ✓ **5** real-time **case studies with code walkthroughs**