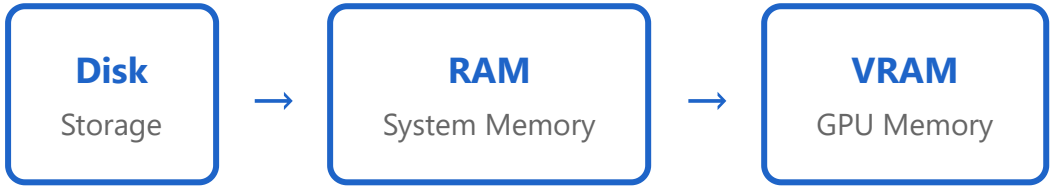
ML Models Loading and Memory Management



Challenge: **LLaMA-70B = 140GB (FP16)** exceeds single GPU VRAM

Solutions for Large Models

1

Model Sharding

Distribute across multiple GPUs (tensor parallelism)

2

Offloading

Keep parameters in RAM, load when needed

3

Quantization

Reduce memory footprint (8-bit, 4-bit)

Key Insights

- Memory bandwidth limits loading speed
- Disk I/O is not the bottleneck
- Use memory-mapped files (mmap)

Popular Tools

- Hugging Face Accelerate
- DeepSpeed ZeRO