利用DeepSpeed的ZeRO-2模式全参数微调8B模型，使用bfloat16精度时，显存占用取决于GPU数量（DP，数据并行度），因为ZeRO-2会将梯度和优化器状态分区到各GPU上，而模型参数在每个GPU上完整复制。以下假设典型配置中使用8个GPU（常见于8B模型的全参数微调，如A100 80GB GPU集群），并聚焦参数相关内存（不包括激活内存，激活内存取决于batch size和sequence length，通常为几GB到数十GB）。

计算依据

- 模型参数数 P = 8 × 10^9。
- bfloat16精度：每个参数2字节。
- 模型内存 = P × 2字节 = 16 GB（每个GPU完整复制，不分区）。
- 梯度内存 = P × 2字节 = 16 GB（全量），在ZeRO-2中分区为16 / DP GB每个GPU。另外，在混合精度训练中，通常有临时fp32梯度缓冲区（4字节/参数），全量32 GB，分区32 / DP GB。因此，梯度总占用约(16 + 32) / DP = 48 / DP GB。
- 优化器内存（Adam）：在混合精度下，包括fp32 master参数拷贝（4字节/参数）、动量m（4字节/参数）和方差v（4字节/参数），全量12字节/参数 = 96 GB，在ZeRO-2中分区为96 / DP GB。
- 总显存（参数相关） ≈ 16 + (48 + 96) / DP = 16 + 144 / DP GB（实际中可能有少量开销，如缓冲区，使其略高）。

对于DP=8的计算结果

- **总显存占用**：约34 GB每个GPU（16 GB模型 + 6 GB梯度 + 12 GB优化器 + 少量开销；实测值常为34-36 GB在低batch/seq配置下）。
- **模型占用**：16 GB。
- **梯度占用**：6 GB（包括bf16梯度和临时fp32缓冲区）。
- **优化器占用**：12 GB。

如果不包括临时fp32梯度缓冲区，则梯度占用2 GB，总显存约30 GB。实际总显存还会加上激活

等（例如sequence length=2048、batch=1时，可增加20-50 GB，需使用gradient checkpointing优化）。如果使用cpu_offload_optimizer，优化器可オフload到CPU，进一步降低GPU内存至约20 GB（但速度变慢）。如果DP不同（如DP=4），总显存≈16 + 36 =52 GB每个GPU。