Roofline Model: A100 Performance Limits

Achieved Performance = min(Peak Compute, Bandwidth × Arithmetic Intensity)

Peak: 312 TFLOPS

Bandwidth: 2 TB/s

Performance (TFLOPS)

Large Matmul
AI=256
280 TF

Attn n=512
AI=200
220 TF

Attn n=2048
AI=50
78 TF

Small Matmul
AI=32
50 TF

Softmax

350
300
200
100
50
0

1   10   50   100   200

Operation Types

- Compute-Bound (90% peak)
- Memory-Bound (<25% peak)
- Borderline (70% peak)

Key Insight

Long-context attention (n=2048) is memory-bound: O(n²) memory grows faster than O(n²d) compute.
More powerful GPUs won't help—need efficient attention variants or more bandwidth.