DeepSeek-R1-Distill-Llama-70B Performance Analysis
TP/PP Parallelism Configurations (Input: 2048 tokens, Output: 512 tokens)

Performance Metrics
Total Cost ($)
Total Time (s)
0
200
400
600
0
1
2
3
4
0
200
400
600
800
2x g5.12xlarge TP4, PP2: 144, 36, 180, $0.97, 7m6s
g5.48xlarge TP4, PP2: 117, 29, 146, $3.76, 13m47s
2x g5.12xlarge TP4, PP3: 537, 134, 671, $0.39, 1m54s
3x g5.12xlarge TP4, PP3: 537, 134, 671, $0.39, 1m54s
2x g5.12xlarge TP4, PP4: 574, 143, 717, $0.49, 1m47s
4x g5.12xlarge TP4, PP4: 574, 143, 717, $0.49, 1m47s
2x g5.12xlarge TP8, PP1: 123, 30, 153, $2.28, 8m22s
g5.48xlarge TP8, PP1: 130, 32, 162, $2.16, 7m54s
2x g5.12xlarge TP8, PP2: 270, 67, 337, $2.07, 3m48s
2x g5.48xlarge TP8, PP2: 269, 67, 336, $2.08, 3m49s
g5.48xlarge TP8, PP2: 268, 67, 335, $2.09, 3m49s
2x g5.12xlarge TP8, PP3: 347, 86, 433, $2.42, 2m57s
3x g5.48xlarge TP8, PP3: 357, 89, 446, $2.35, 2m52s
g5.48xlarge TP8, PP3: 367, 91, 458, $2.29, 2m48s
2x g5.12xlarge TP8, PP4: 399, 99, 499, $2.80, 2m34s
4x g5.48xlarge TP8, PP4: 381, 95, 476, $2.94, 2m42s
g5.48xlarge TP8, PP4: 364, 90, 454, $3.08, 2m49s
Performance Metrics, Cost & Time
Requests/sec
Input tokens/sec
Output tokens/sec
Total tokens/sec
Total Cost ($)
Total Time (s)
Performance Scaling Efficiency & Cost Efficiency Analysis
Performance Scaling Efficiency (%)
Cost Efficiency (%)
0
50
100
150
200
250
2x g5.12xlarge TP4, PP2: 100.0%, 100.0%
g5.48xlarge TP4, PP2: 81.2%, 40.6%
2x g5.12xlarge TP4, PP3: 248.4%, 248.4%
3x g5.12xlarge TP4, PP3: 248.4%, 248.4%
2x g5.12xlarge TP4, PP4: 199.1%, 199.1%
4x g5.12xlarge TP4, PP4: 199.1%, 199.1%
2x g5.12xlarge TP8, PP1: 85.0%, 42.5%
g5.48xlarge TP8, PP1: 90.1%, 45.0%
2x g5.12xlarge TP8, PP2: 93.6%, 46.8%
2x g5.48xlarge TP8, PP2: 93.3%, 46.6%
g5.48xlarge TP8, PP2: 92.9%, 46.5%
2x g5.12xlarge TP8, PP3: 80.1%, 40.1%
3x g5.48xlarge TP8, PP3: 82.5%, 41.2%
g5.48xlarge TP8, PP3: 84.8%, 42.4%
2x g5.12xlarge TP8, PP4: 69.2%, 34.6%
4x g5.48xlarge TP8, PP4: 66.1%, 33.1%
g5.48xlarge TP8, PP4: 63.0%, 31.5%
Efficiency Metrics (%)
Perfect Performance Scaling
Performance Scaling Efficiency (%)
Perfect Cost Efficiency
Cost Efficiency (%)
GPU Utilization Analysis
Avg SM Utilization (%)
Avg Memory BW Utilization (%)
0
20
40
60
80
100
2x g5.12xlarge TP4, PP2: 85.5%, 61.6%
g5.48xlarge TP4, PP2: 90.4%, 49.3%
2x g5.12xlarge TP4, PP3: 87.4%, 46.2%
3x g5.12xlarge TP4, PP3: 87.4%, 46.2%
2x g5.12xlarge TP4, PP4: 84.1%, 47.7%
4x g5.12xlarge TP4, PP4: 84.1%, 47.7%
2x g5.12xlarge TP8, PP1: 99.0%, 20.7%
g5.48xlarge TP8, PP1: 99.0%, 21.9%
2x g5.12xlarge TP8, PP2: 90.4%, 12.5%
2x g5.48xlarge TP8, PP2: 91.1%, 12.4%
g5.48xlarge TP8, PP2: 91.7%, 12.4%
2x g5.12xlarge TP8, PP3: 86.1%, 15.0%
3x g5.48xlarge TP8, PP3: 85.5%, 15.4%
g5.48xlarge TP8, PP3: 84.8%, 15.7%
2x g5.12xlarge TP8, PP4: 76.8%, 16.5%
4x g5.48xlarge TP8, PP4: 76.6%, 15.8%
g5.48xlarge TP8, PP4: 76.3%, 15.0%
GPU Utilization Metrics
100% Utilization
Avg SM Utilization (%)
Avg Memory BW Utilization (%)