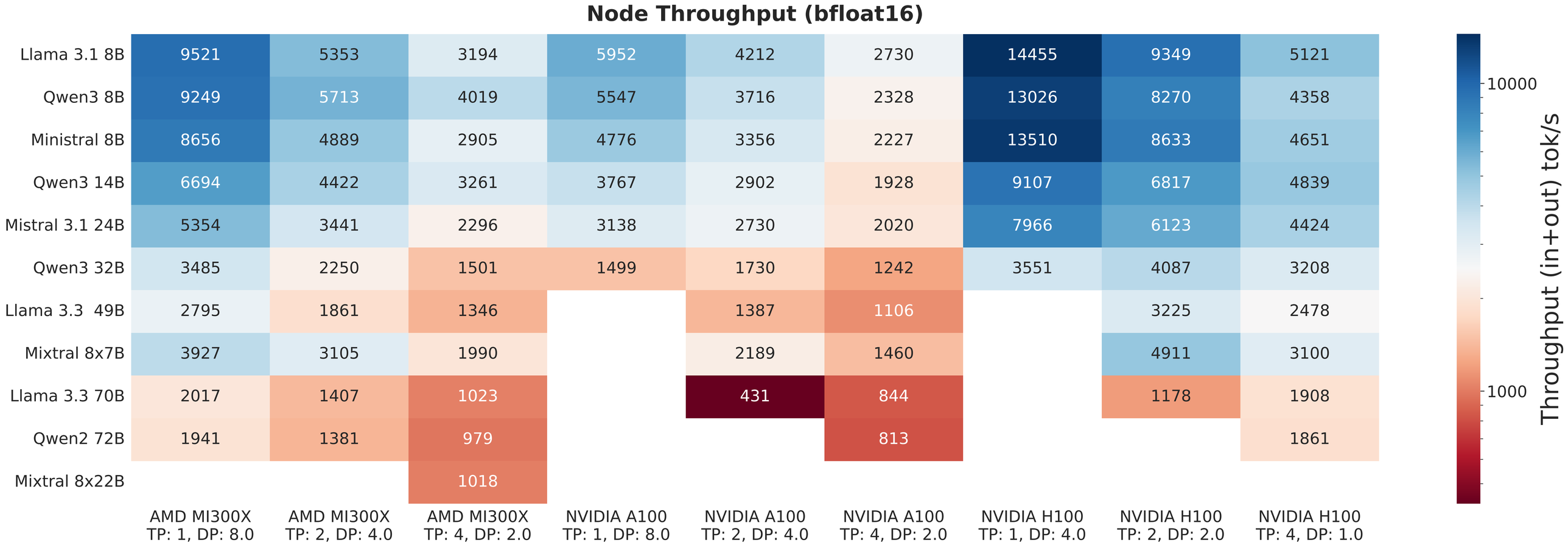
Node Throughput (bfloat16)
Llama 3.1 8B 9521 5353 3194 5952 4212 2730 14455 9349 5121
Qwen3 8B 9249 5713 4019 5547 3716 2328 13026 8270 4358
Ministral 8B 8656 4889 2905 4776 3356 2227 13510 8633 4651
Qwen3 14B 6694 4422 3261 3767 2902 1928 9107 6817 4839
Mistral 3.1 24B 5354 3441 2296 3138 2730 2020 7966 6123 4424
Qwen3 32B 3485 2250 1501 1499 1730 1242 3551 4087 3208
Llama 3.3 49B 2795 1861 1346 1387 1106 3225 2478
Mixtral 8x7B 3927 3105 1990 2189 1460 4911 3100
Llama 3.3 70B 2017 1407 1023 431 844 1178 1908
Qwen2 72B 1941 1381 979 813 1861
Mixtral 8x22B 1018
AMD MI300X TP: 1, DP: 8.0
AMD MI300X TP: 2, DP: 4.0
AMD MI300X TP: 4, DP: 2.0
NVIDIA A100 TP: 1, DP: 8.0
NVIDIA A100 TP: 2, DP: 4.0
NVIDIA A100 TP: 4, DP: 2.0
NVIDIA H100 TP: 1, DP: 4.0
NVIDIA H100 TP: 2, DP: 2.0
NVIDIA H100 TP: 4, DP: 1.0
10000
1000
Throughput (in+out) tok/s