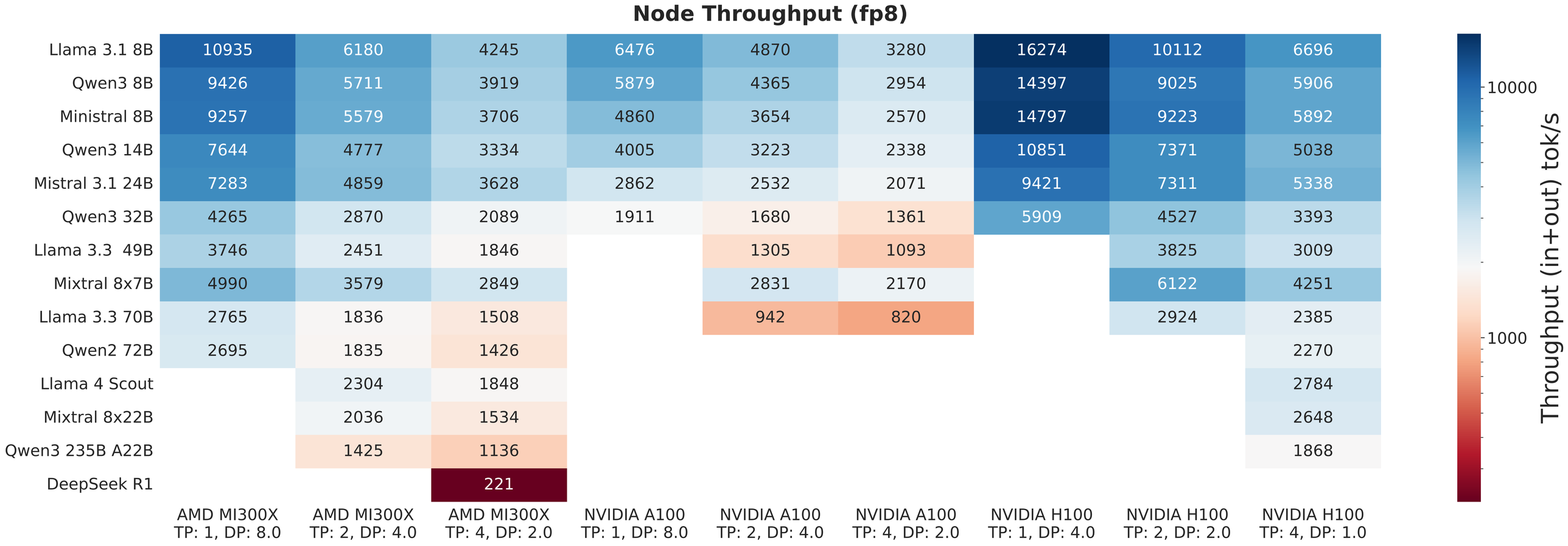
Node Throughput (fp8)
Llama 3.1 8B 10935 6180 4245 6476 4870 3280 16274 10112 6696
Qwen3 8B 9426 5711 3919 5879 4365 2954 14397 9025 5906
Ministral 8B 9257 5579 3706 4860 3654 2570 14797 9223 5892
Qwen3 14B 7644 4777 3334 4005 3223 2338 10851 7371 5038
Mistral 3.1 24B 7283 4859 3628 2862 2532 2071 9421 7311 5338
Qwen3 32B 4265 2870 2089 1911 1680 1361 5909 4527 3393
Llama 3.3 49B 3746 2451 1846 1305 1093 3825 3009
Mixtral 8x7B 4990 3579 2849 2831 2170 6122 4251
Llama 3.3 70B 2765 1836 1508 942 820 2924 2385
Qwen2 72B 2695 1835 1426 2270
Llama 4 Scout 2304 1848 2784
Mixtral 8x22B 2036 1534 2648
Qwen3 235B A22B 1425 1136 1868
DeepSeek R1 221
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