

Power and Throughput: FP16 vs 4-bit NF4

