**Phase 3 Inference Throughput
(batch=100k, bits=64)**

Throughput (MOps/s)

JAX/GPU (RTX 3080): 10,959

NumPy/CPU (INT8): 29

380x