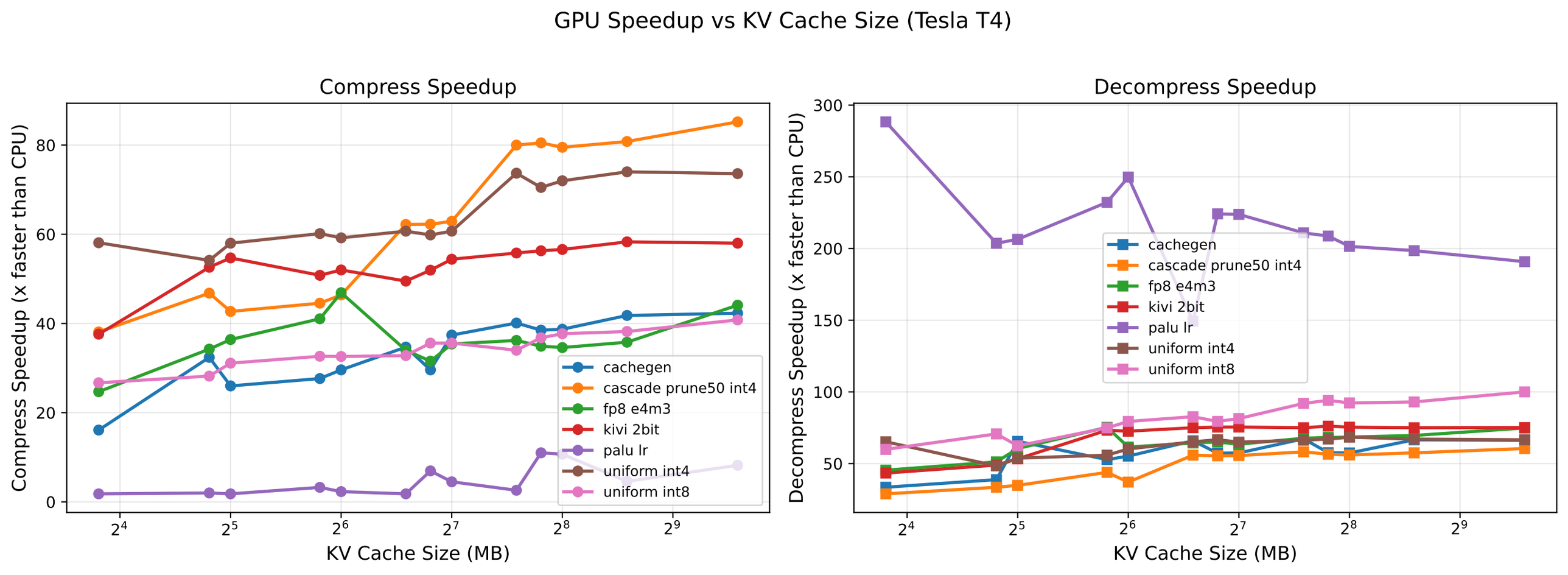
GPU Speedup vs KV Cache Size (Tesla T4)
Compress Speedup
Compress Speedup (x faster than CPU)
KV Cache Size (MB)
Decompress Speedup
Decompress Speedup (x faster than CPU)
KV Cache Size (MB)
cachegen
cascade prune50 int4
fp8 e4m3
kivi 2bit
palu lr
uniform int4
uniform int8