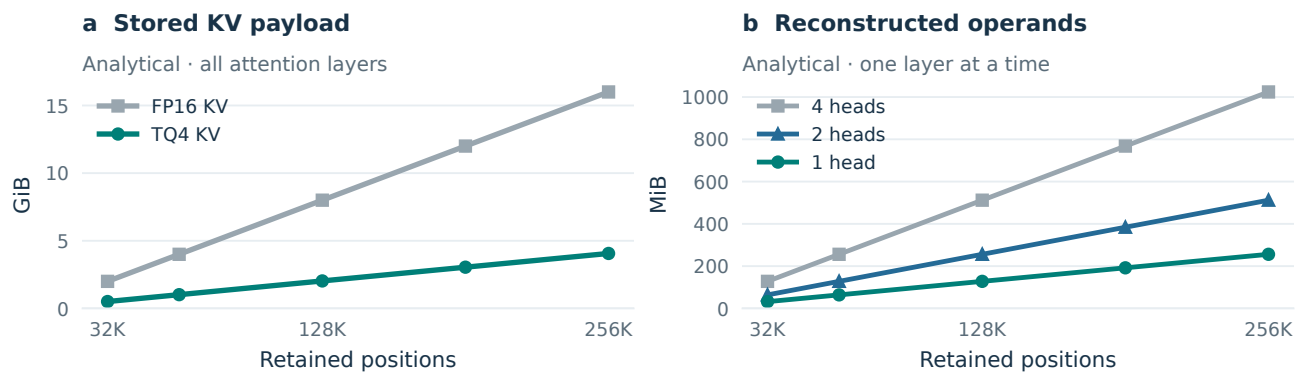

a Stored KV payload
Analytical · all attention layers
FP16 KV
TQ4 KV
GiB
15
10
5
0
32K
128K
256K
Retained positions
b Reconstructed operands
Analytical · one layer at a time
4 heads
2 heads
1 head
MiB
1000
800
600
400
200
0
32K
128K
256K
Retained positions