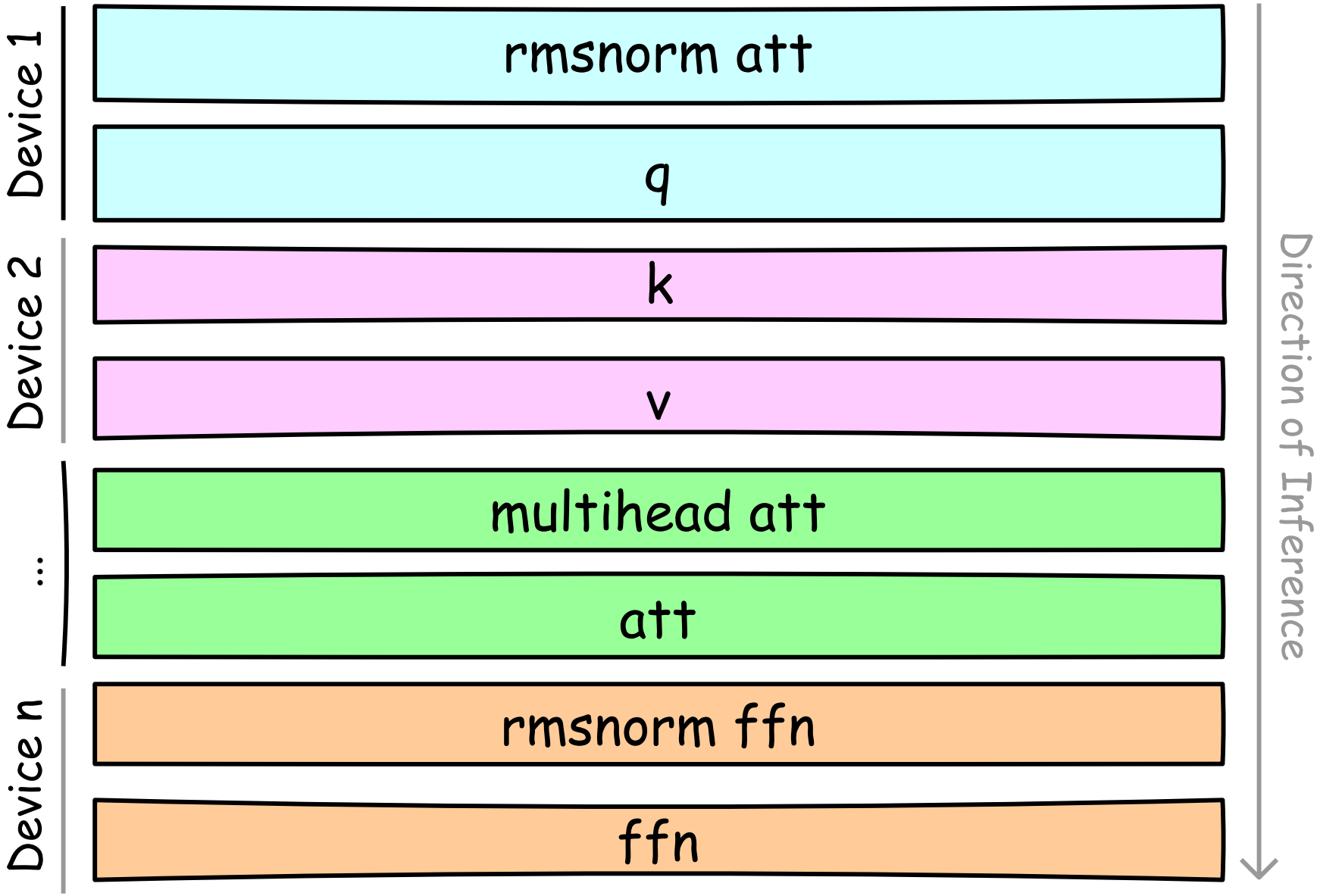
Device 1
rmsnorm att
q
Device 2
k
v
...
multihead att
att
Device n
rmsnorm ffn
ffn
Direction of Inference