

KV-cache vs Context Length — MHA vs GQA (multi-group)
(n_heads=24, emb_dim=2048, n_layers=48, batch=1, dtype=bf16)

