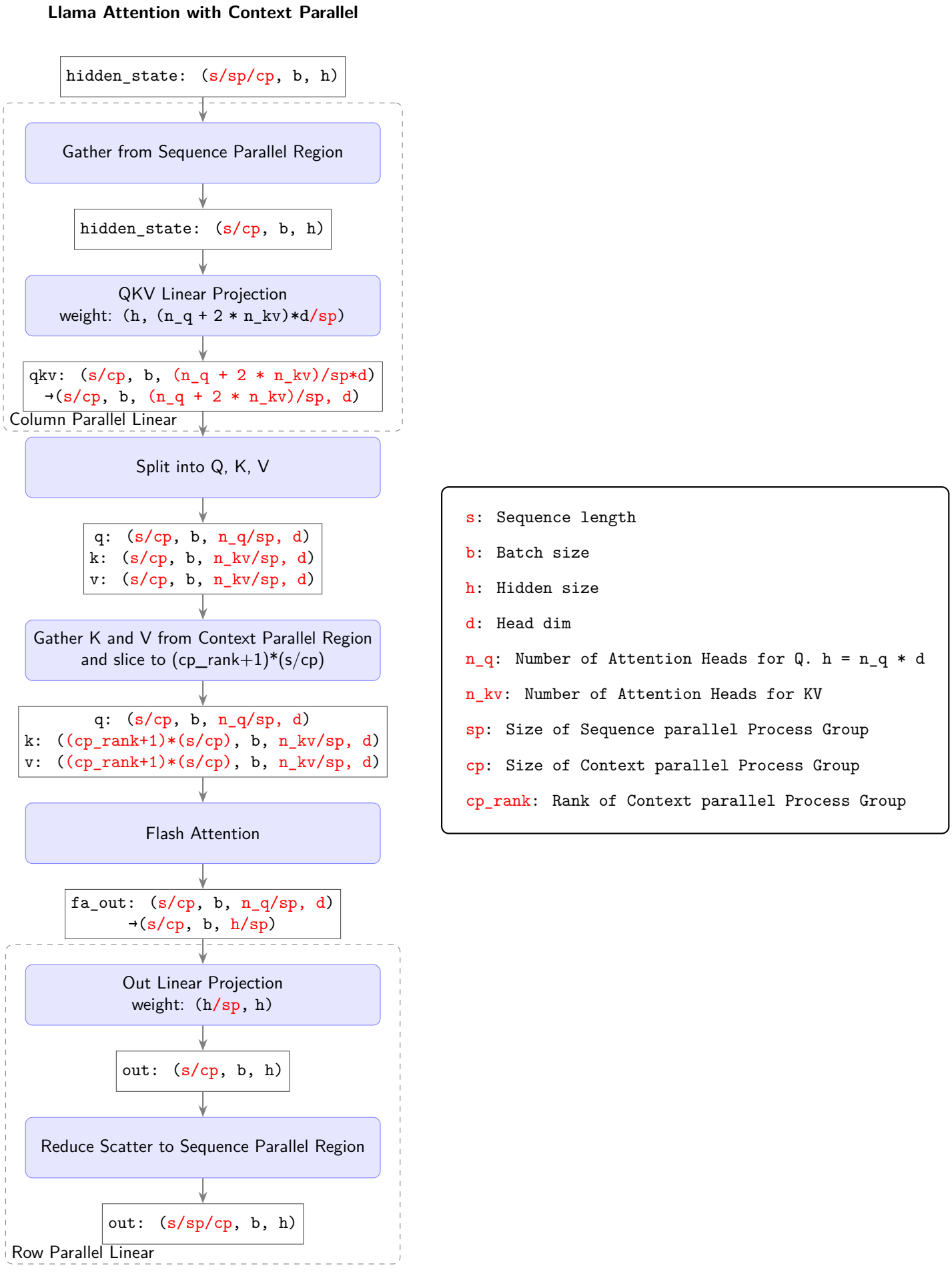
Llama Attention with Context Parallel
hidden_state: (s/sp/cp, b, h)
Gather from Sequence Parallel Region
hidden_state: (s/cp, b, h)
QKV Linear Projection
weight: (h, (n_q + 2 * n_kv)*d/sp)
qkv: (s/cp, b, (n_q + 2 * n_kv)/sp*d)
→(s/cp, b, (n_q + 2 * n_kv)/sp, d)
Column Parallel Linear
Split into Q, K, V
q: (s/cp, b, n_q/sp, d)
k: (s/cp, b, n_kv/sp, d)
v: (s/cp, b, n_kv/sp, d)
Gather K and V from Context Parallel Region
and slice to (cp_rank+1)*(s/cp)
q: (s/cp, b, n_q/sp, d)
k: ((cp_rank+1)*(s/cp), b, n_kv/sp, d)
v: ((cp_rank+1)*(s/cp), b, n_kv/sp, d)
Flash Attention
fa_out: (s/cp, b, n_q/sp, d)
→(s/cp, b, h/sp)
Out Linear Projection
weight: (h/sp, h)
out: (s/cp, b, h)
Reduce Scatter to Sequence Parallel Region
out: (s/sp/cp, b, h)
Row Parallel Linear
s: Sequence length
b: Batch size
h: Hidden size
d: Head dim
n_q: Number of Attention Heads for Q. h = n_q * d
n_kv: Number of Attention Heads for KV
sp: Size of Sequence parallel Process Group
cp: Size of Context parallel Process Group
cp_rank: Rank of Context parallel Process Group