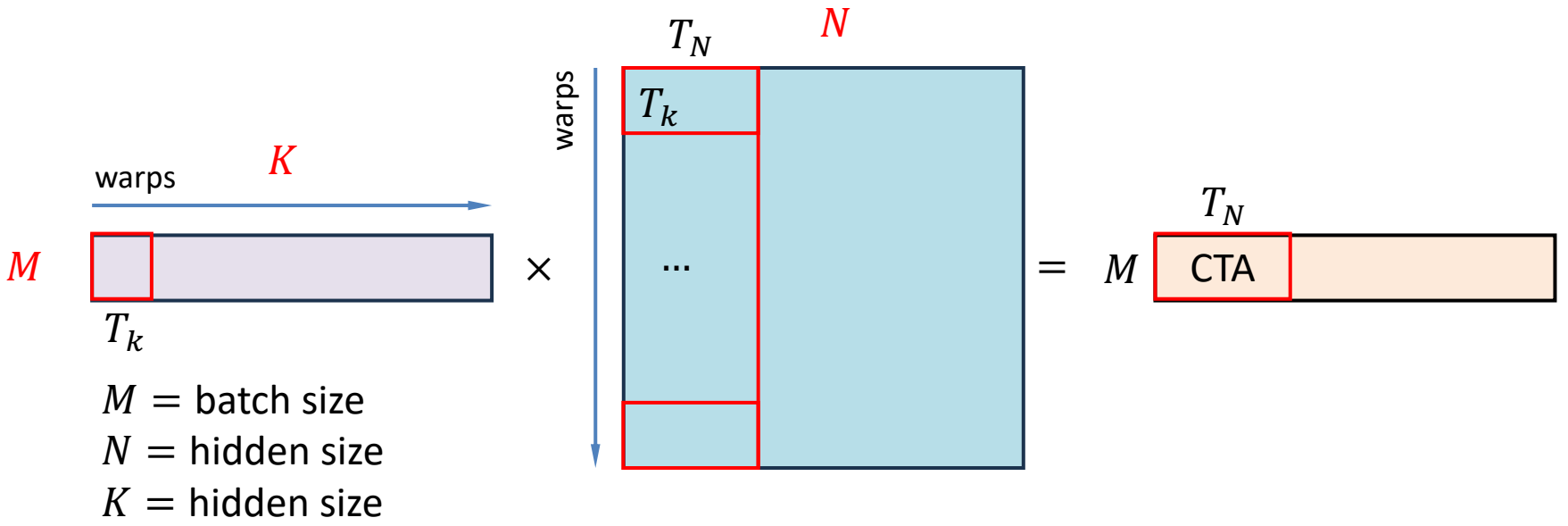
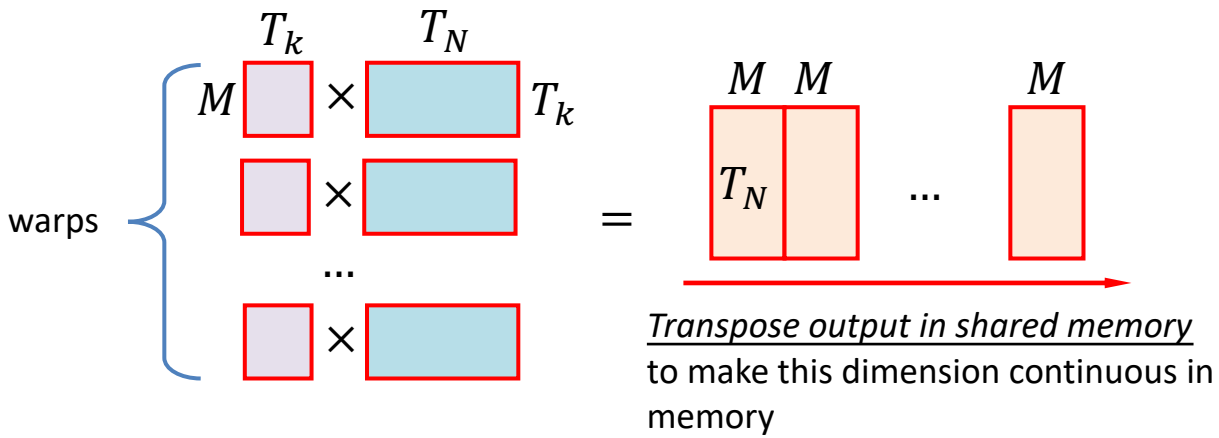
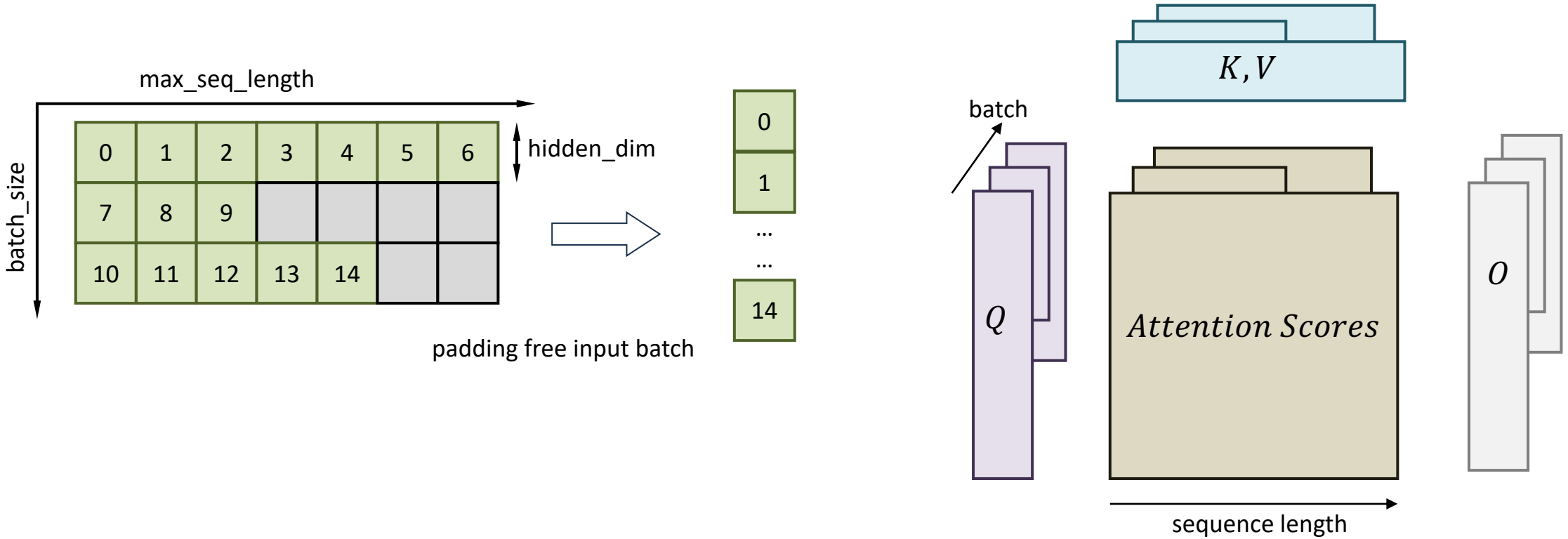
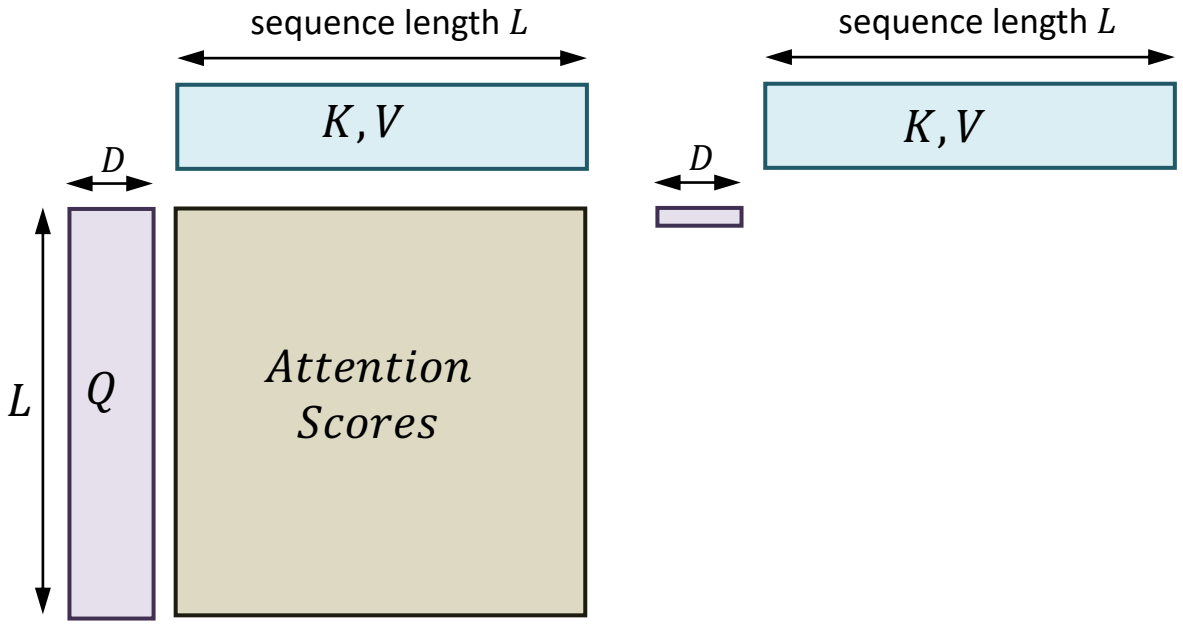

T_N
N
warps
T_k
warps
K
M
T_k
...
$\times$
$=$
M
T_N
CTA
M = batch size
N = hidden size
K = hidden size

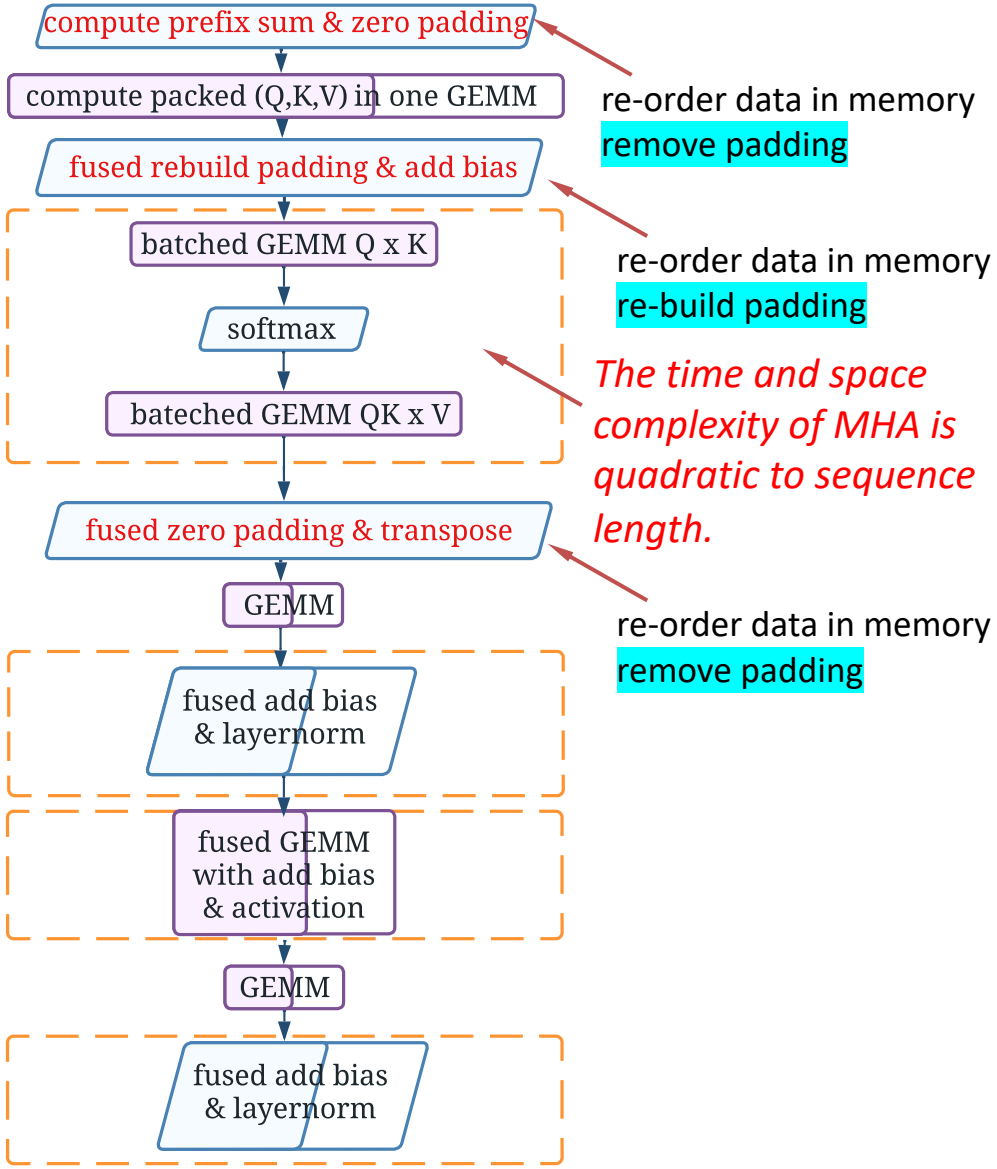

T_k
T_N
warps
M
$\times$
T_k
$\times$
...
$\times$
$=$
M M
M
T_N
...
Transpose output in shared memory
to make this dimension continuous in memory

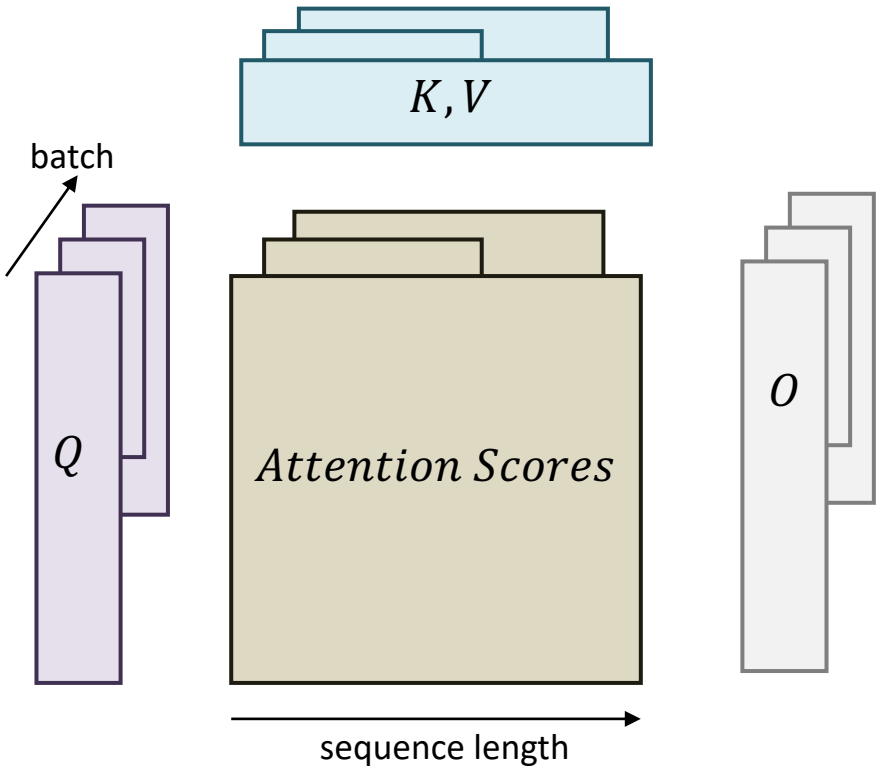

max_seq_length
batch_size
hidden_dim
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
padding free input batch
0
1
...
...
14
batch
K, V
Q
$Attention\ Scores$
O
sequence length

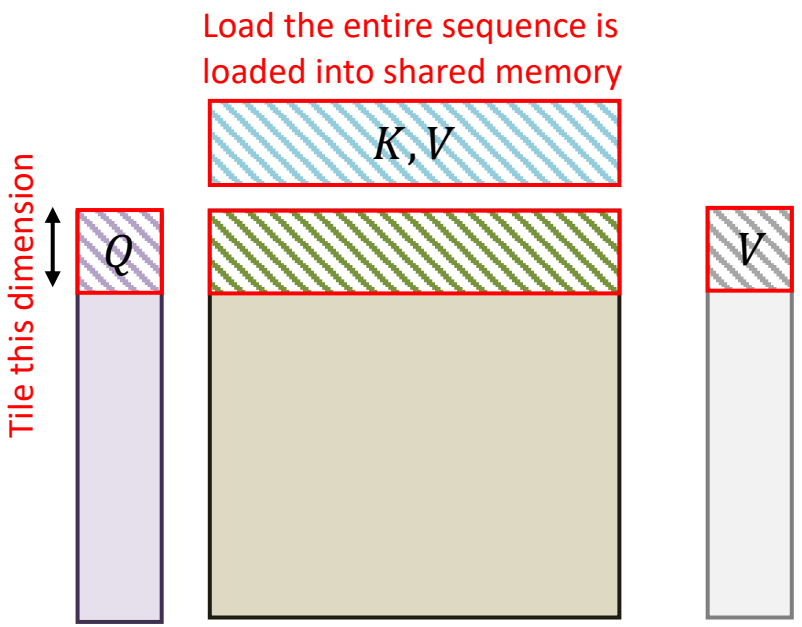


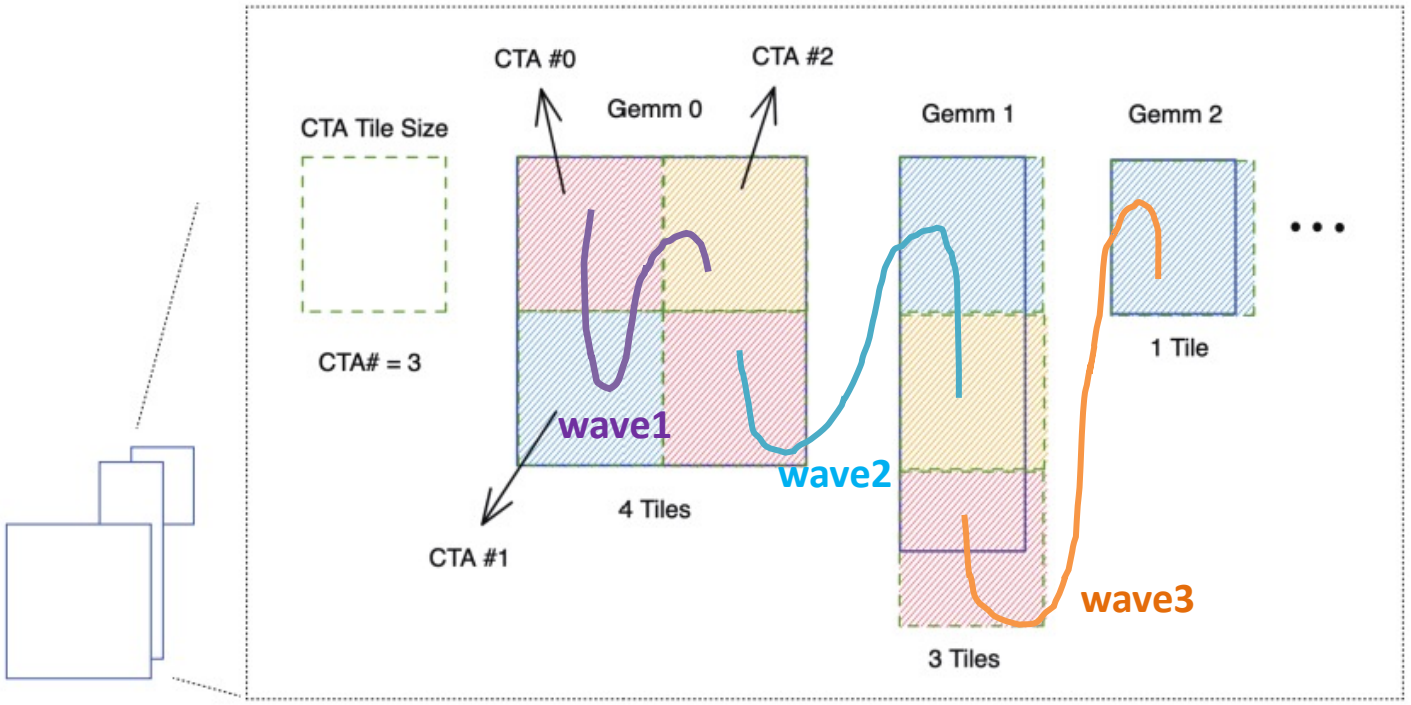
(1). Prefill stage

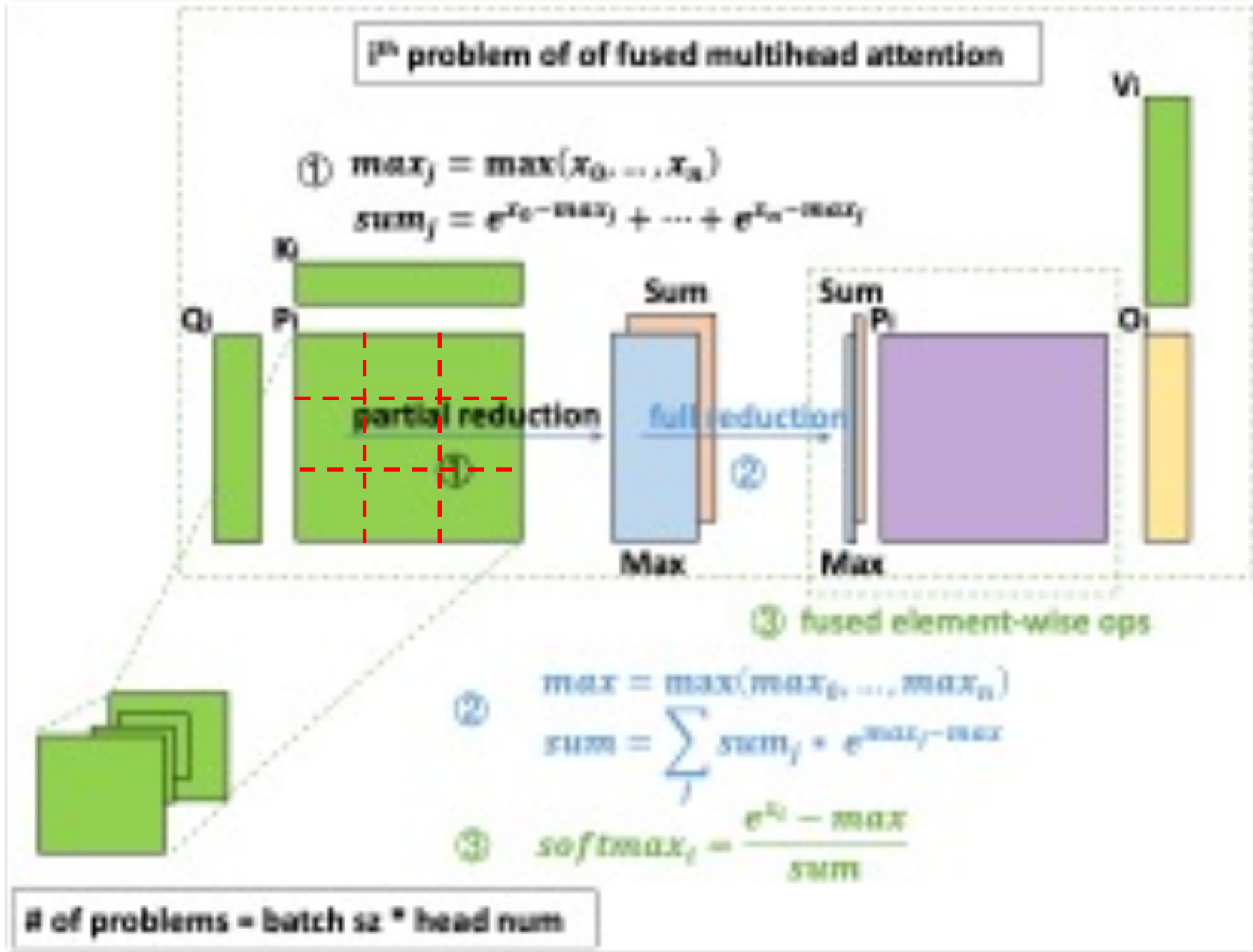
(2). Generate stage


compute prefix sum & zero padding
compute packed (Q,K,V) in one GEMM
fused rebuild padding & add bias
batched GEMM Q x K
softmax
bateched GEMM QK x V
fused zero padding & transpose
GEMM
fused add bias & layernorm
fused GEMM with add bias & activation
GEMM
fused add bias & layernorm
re-order data in memory
remove padding
re-order data in memory
re-build padding
The time and space complexity of MHA is quadratic to sequence length.
re-order data in memory
remove padding


K, V
batch
Q
Attention Scores
O
sequence length


Load the entire sequence is
loaded into shared memory
K, V
Tile this dimension
Q
V


CTA #0
CTA #2
Gemm 0
Gemm 1
Gemm 2
CTA Tile Size
CTA# = 3
wave1
wave2
wave3
4 Tiles
CTA #1
3 Tiles
1 Tile


i[th] problem of of fused multihead attention
① $max_j = \max(x_0, \ldots, x_n)$
$sum_j = e^{x_0 - max_j} + \cdots + e^{x_n - max_j}$
Qi
Ki
Pi
Vi
Oi
Sum
Max
Sum
Max
partial reduction
full reduction
②
③ fused element-wise ops
② $max = \max(max_0, \ldots, max_n)$
$sum = \sum_j sum_j * e^{max_j - max}$
③ $softmax_i = \frac{e^{x_i} - max}{sum}$
# of problems = batch sz * head num