Ulysses Attention Shape Transformation

hidden_state: (s/sp/cp, b, h)

Column Parallel Linear

Gather from Sequence Parallel Region

(s/cp, b, h)

QKV Linear Projection
(reduce h to h/sp)

(s/cp, b, h/sp)

All2all

(s, b, h/sp/cp)

Flash Attention

(s, b, h/sp/cp)

All2all

(s/cp, b, h/sp)

Row Parallel Linear

Out Linear Projection
(aggregate h/sp to h)

(s/cp, b, h)

Reduce Scatter to Sequence Parallel Region
on sequence dimension

(s/sp/cp, b, h)

---
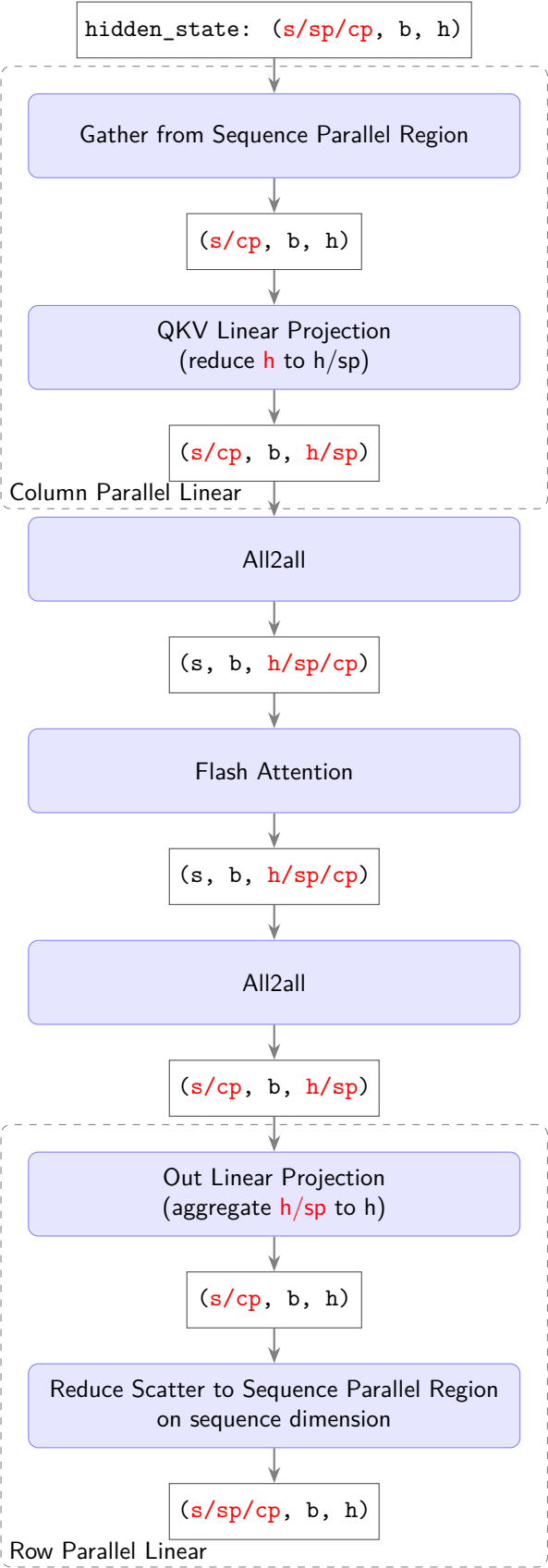
s: Sequence length

b: Batch size

h: Hidden dimension

sp: Sequence parallel size

cp: Context parallel size