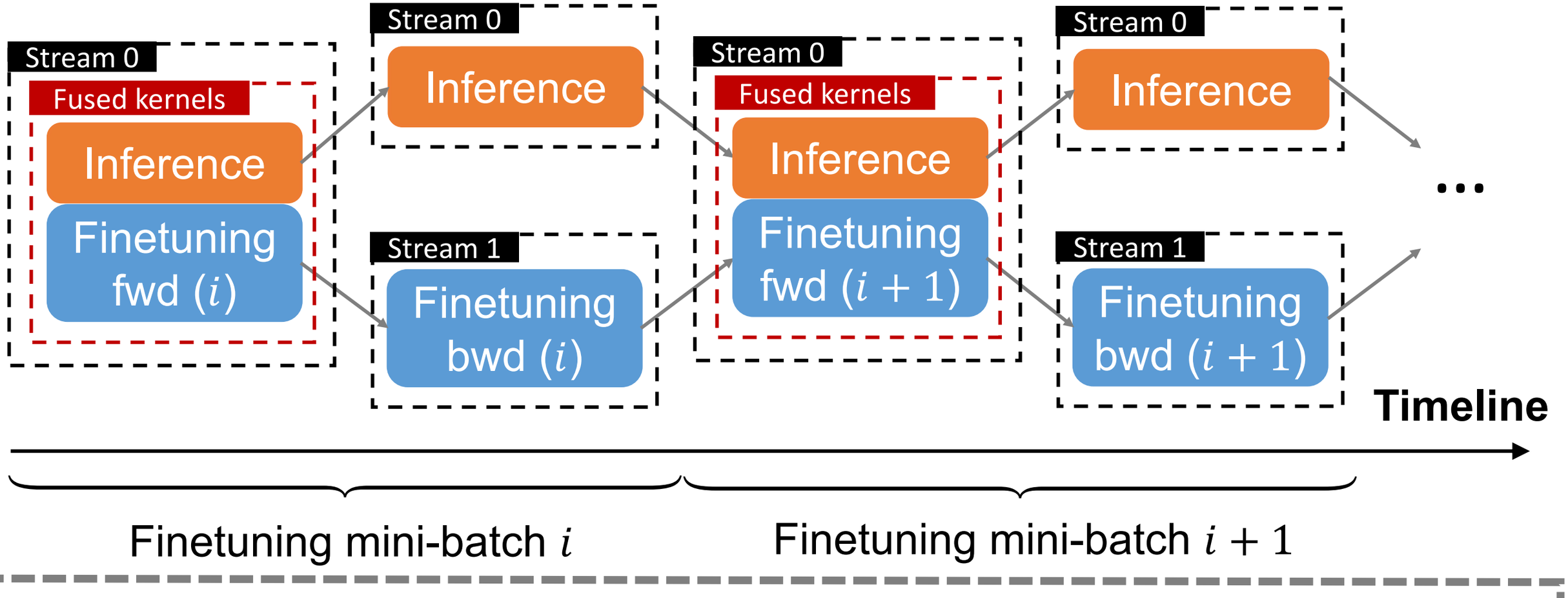
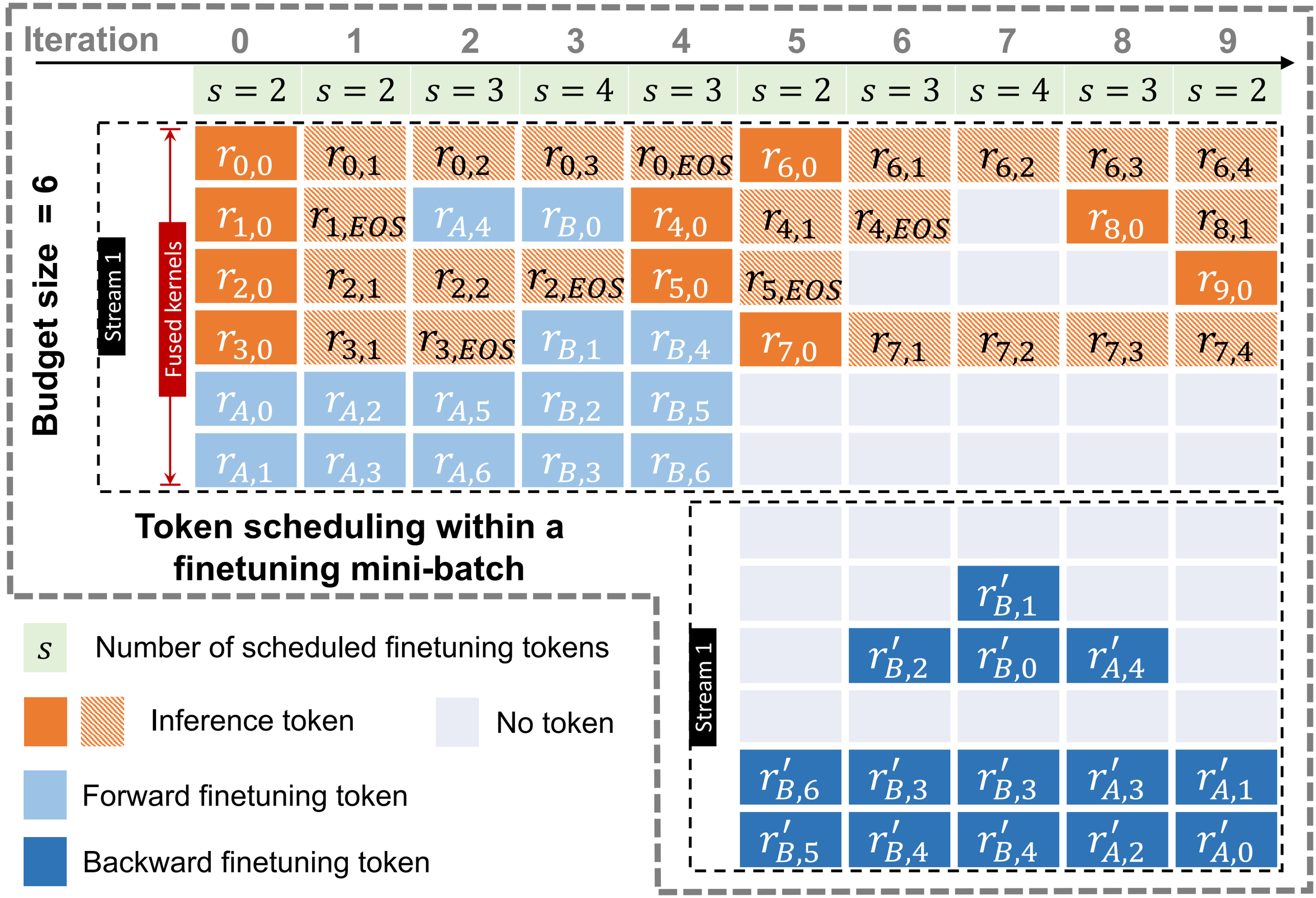

Stream 0
Fused kernels
Inference
Finetuning fwd (i)
Stream 0
Inference
Stream 1
Finetuning bwd (i)
Stream 0
Fused kernels
Inference
Finetuning fwd $(i+1)$
Stream 0
Inference
Stream 1
Finetuning bwd $(i+1)$
...
Timeline
Finetuning mini-batch i
Finetuning mini-batch $i+1$


Iteration 0 1 2 3 4 5 6 7 8 9
$s=2$ $s=2$ $s=3$ $s=4$ $s=3$ $s=2$ $s=3$ $s=4$ $s=3$ $s=2$
Budget size = 6
Stream 1
Fused kernels
$r_{0,0}$ $r_{0,1}$ $r_{0,2}$ $r_{0,3}$ $r_{0,EOS}$ $r_{6,0}$ $r_{6,1}$ $r_{6,2}$ $r_{6,3}$ $r_{6,4}$
$r_{1,0}$ $r_{1,EOS}$ $r_{A,4}$ $r_{B,0}$ $r_{4,0}$ $r_{4,1}$ $r_{4,EOS}$ $r_{8,0}$ $r_{8,1}$
$r_{2,0}$ $r_{2,1}$ $r_{2,2}$ $r_{2,EOS}$ $r_{5,0}$ $r_{5,EOS}$ $r_{9,0}$
$r_{3,0}$ $r_{3,1}$ $r_{3,EOS}$ $r_{B,1}$ $r_{B,4}$ $r_{7,0}$ $r_{7,1}$ $r_{7,2}$ $r_{7,3}$ $r_{7,4}$
$r_{A,0}$ $r_{A,2}$ $r_{A,5}$ $r_{B,2}$ $r_{B,5}$
$r_{A,1}$ $r_{A,3}$ $r_{A,6}$ $r_{B,3}$ $r_{B,6}$
Token scheduling within a finetuning mini-batch
s Number of scheduled finetuning tokens
Inference token
No token
Forward finetuning token
Backward finetuning token
Stream 1
$r'_{B,1}$
$r'_{B,2}$ $r'_{B,0}$ $r'_{A,4}$
$r'_{B,6}$ $r'_{B,3}$ $r'_{B,3}$ $r'_{A,3}$ $r'_{A,1}$
$r'_{B,5}$ $r'_{B,4}$ $r'_{B,4}$ $r'_{A,2}$ $r'_{A,0}$