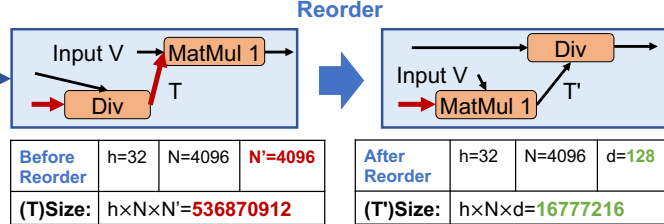
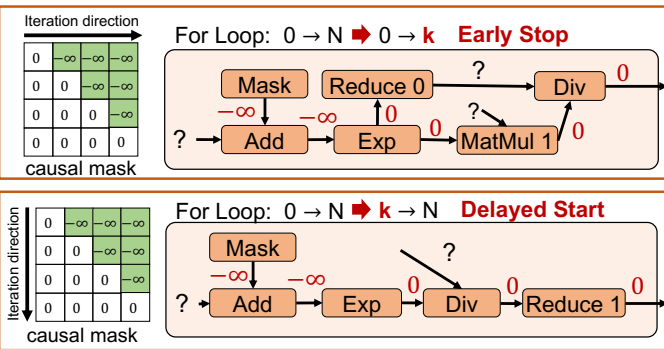


Kernel 5	h=32	N=4096	d=128
Q Dependency:	Batch	NonPara	NonPara
K Dependency:	Batch	Reuse	NonPara

(b) FlashTensor's reduce dependency property-aware kernel mapping for H₂O. The table includes the identified reduce dependency for Kernel 1's inputs Q and K with common shape (h, N, d) .



(c) FlashTensor's broadcast and size property-aware transformation. The two tables represent the size for the intermediate tensor T and T' before and after reorder.



(d) FlashTensor's value property-aware optimization by early stop and delayed start of For-loop by causal mask