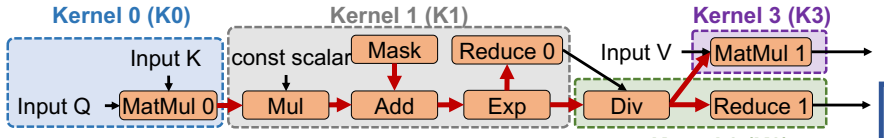
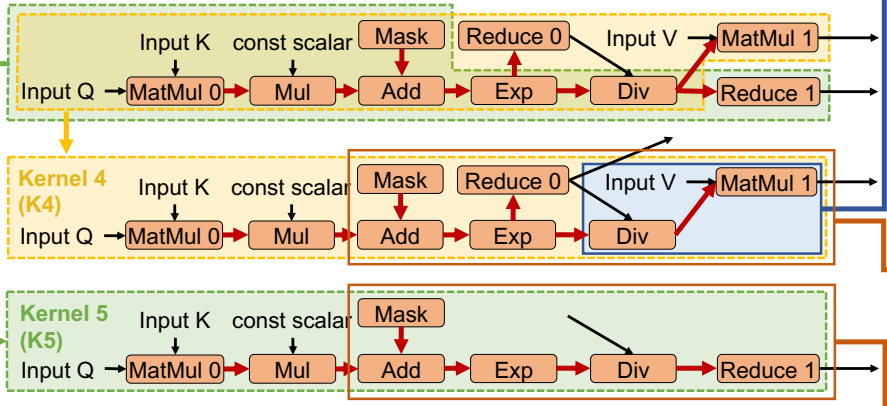
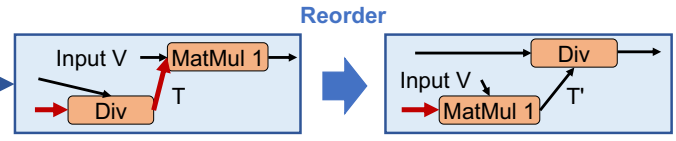



(a) TensorRT's Strategy for $H_2O$




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

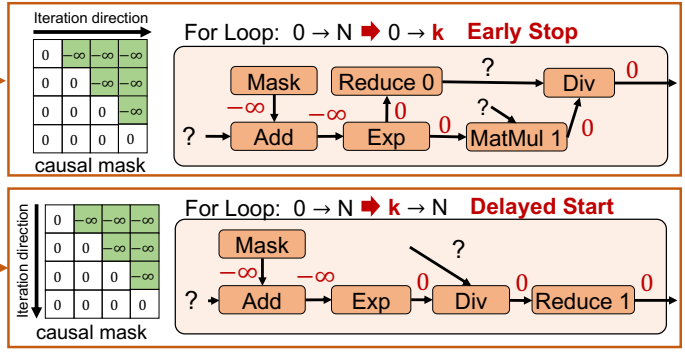
(b) FlashTensor's reduce dependency property-aware kernel mapping for $H_2O$. The table includes the identified reduce dependency for Kernel 1's inputs Q and K with common shape $(h, N, d)$.




| Before Reorder | h=32 | N=4096 | N'=4096 |
|---|---|---|---|
| **(T)Size:** | h×N×N'=536870912 | | |

| After Reorder | h=32 | N=4096 | d=128 |
|---|---|---|---|
| **(T')Size:** | h×N×d=16777216 | | |

(c) FlashTensor's broadcast and size property-aware transformation. The two tables represents 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