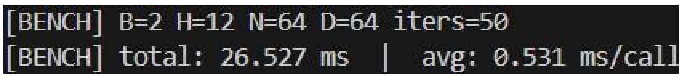
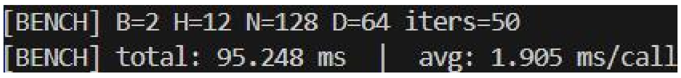
V0: naive

N=64



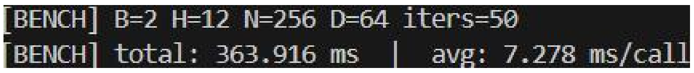
```
[BENCH] B=2 H=12 N=64 D=64 iters=50
[BENCH] total: 26.527 ms  |  avg: 0.531 ms/call
```

N=128



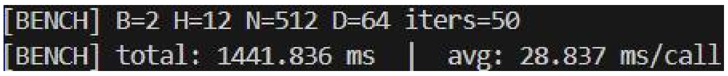
```
[BENCH] B=2 H=12 N=128 D=64 iters=50
[BENCH] total: 95.248 ms  |  avg: 1.905 ms/call
```

N=256



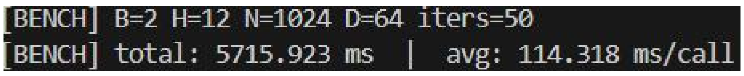
```
[BENCH] B=2 H=12 N=256 D=64 iters=50
[BENCH] total: 363.916 ms  |  avg: 7.278 ms/call
```

N=512

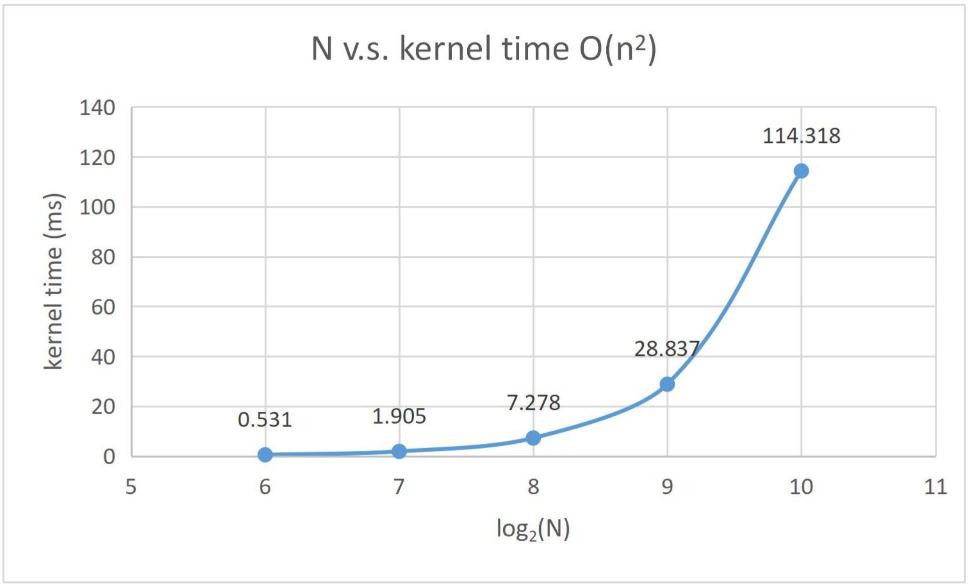


```
[BENCH] B=2 H=12 N=512 D=64 iters=50
[BENCH] total: 1441.836 ms  |  avg: 28.837 ms/call
```

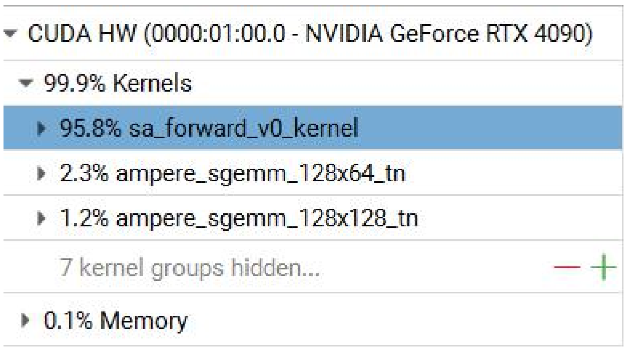
N=1024



```
[BENCH] B=2 H=12 N=1024 D=64 iters=50
[BENCH] total: 5715.923 ms  |  avg: 114.318 ms/call
```



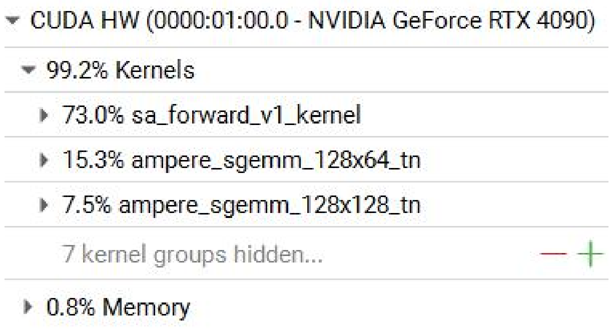
Bottleneck



CUDA HW (0000:01:00.0 - NVIDIA GeForce RTX 4090)
99.9% Kernels
95.8% sa_forward_v0_kernel
2.3% ampere_sgemm_128x64_tn
1.2% ampere_sgemm_128x128_tn
7 kernel groups hidden...
0.1% Memory

Inference loop: 8.423 sec for 10 batch

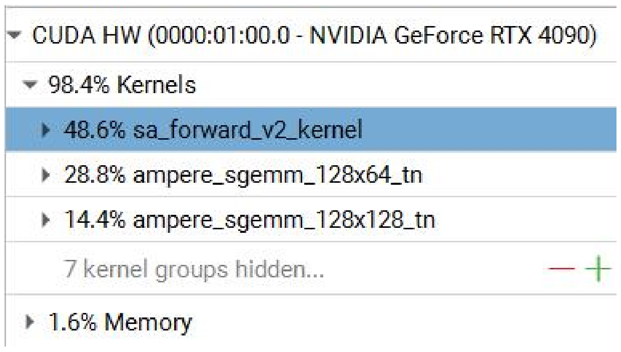
V1: online softmax , KV shared memory tile, single-pass output accumulation



```
▾ CUDA HW (0000:01:00.0 - NVIDIA GeForce RTX 4090)
  ▾ 99.2% Kernels
    ▸ 73.0% sa_forward_v1_kernel
    ▸ 15.3% ampere_sgemm_128x64_tn
    ▸ 7.5% ampere_sgemm_128x128_tn
      7 kernel groups hidden...
  ▸ 0.8% Memory
```

Inference loop: 1.276 sec for 10 batch

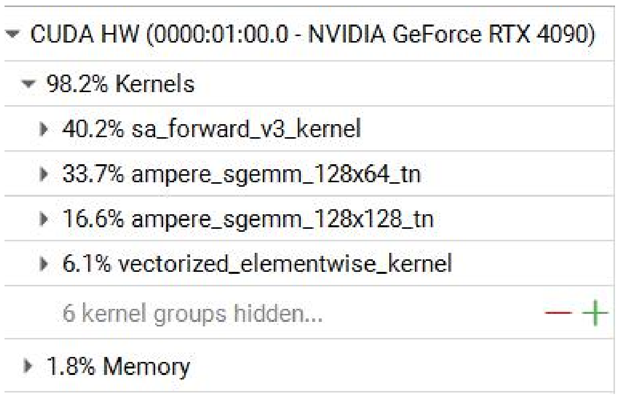
V2: warp-level design(one warp per query row), reuse K/V tile for multiple(8) Q rows, vectorized half2 loads and compute, warp shuffle reduction for dot-product



```
▾ CUDA HW (0000:01:00.0 - NVIDIA GeForce RTX 4090)
  ▾ 98.4% Kernels
    ▸ 48.6% sa_forward_v2_kernel
    ▸ 28.8% ampere_sgemm_128x64_tn
    ▸ 14.4% ampere_sgemm_128x128_tn
      7 kernel groups hidden...
  ▸ 1.6% Memory
```

Inference loop: 686.089 ms for 10 batch

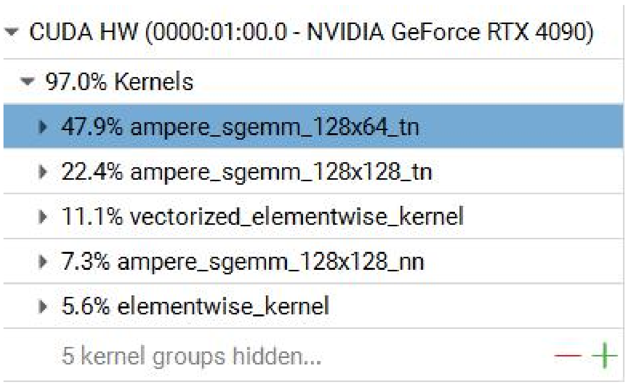
V3: Tensor Core calculate Q×K dot products



```
▾ CUDA HW (0000:01:00.0 - NVIDIA GeForce RTX 4090)
  ▾ 98.2% Kernels
    ▸ 40.2% sa_forward_v3_kernel
    ▸ 33.7% ampere_sgemm_128x64_tn
    ▸ 16.6% ampere_sgemm_128x128_tn
    ▸ 6.1% vectorized_elementwise_kernel
      6 kernel groups hidden...
  ▸ 1.8% Memory
```

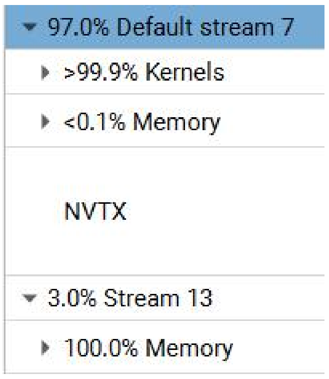
Inference loop: 578.692 ms for 10 batch

Pytorch: utilize Tensor Cores, highly optimized GEMM


CUDA HW (0000:01:00.0 - NVIDIA GeForce RTX 4090)
97.0% Kernels
47.9% ampere_sgemm_128x64_tn
22.4% ampere_sgemm_128x128_tn
11.1% vectorized_elementwise_kernel
7.3% ampere_sgemm_128x128_nn
5.6% elementwise_kernel
5 kernel groups hidden...

Inference loop: 386.165 ms for 10 batch

Pytorch: multi-stream hiding memory latency


97.0% Default stream 7
>99.9% Kernels
<0.1% Memory
NVTX
3.0% Stream 13
100.0% Memory

Inference loop: 365.383 ms for 10 batch