实验内容

1. attention算法的横向比较，FlashAttn、ChunkAttn(ns=0)、PagedAttn、RingAttn性能测试，Latency
2. ChunkAttn的np,ns拉大对比，同时和add\_seq时间比较

实验设置

批处理大小：32

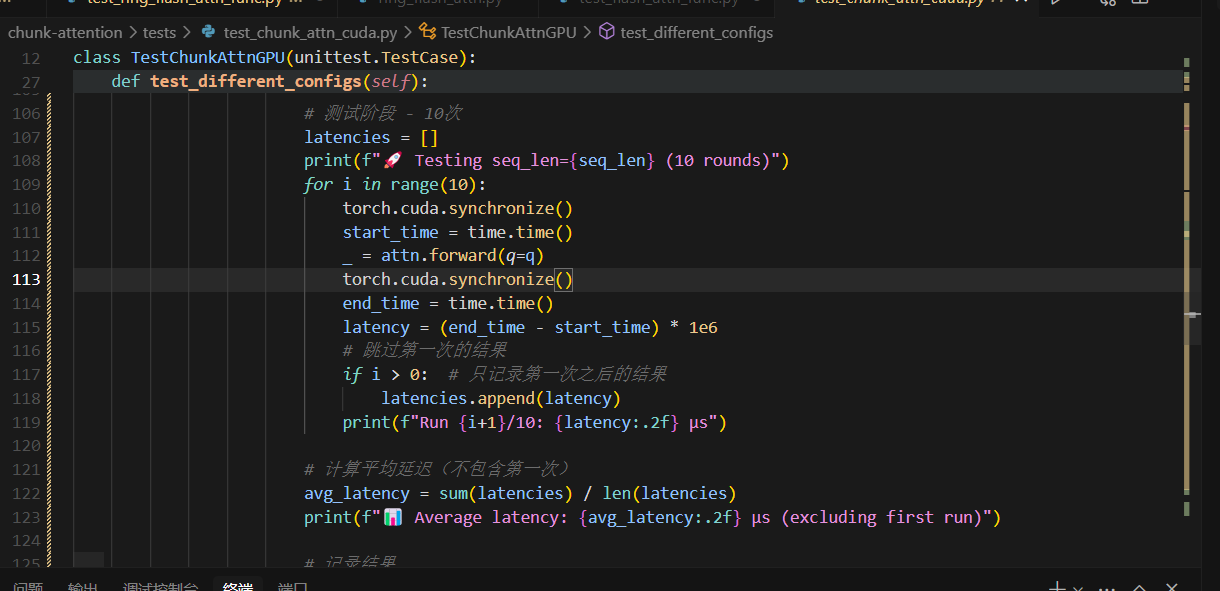
注意力头数：32

注意力头的维度：128

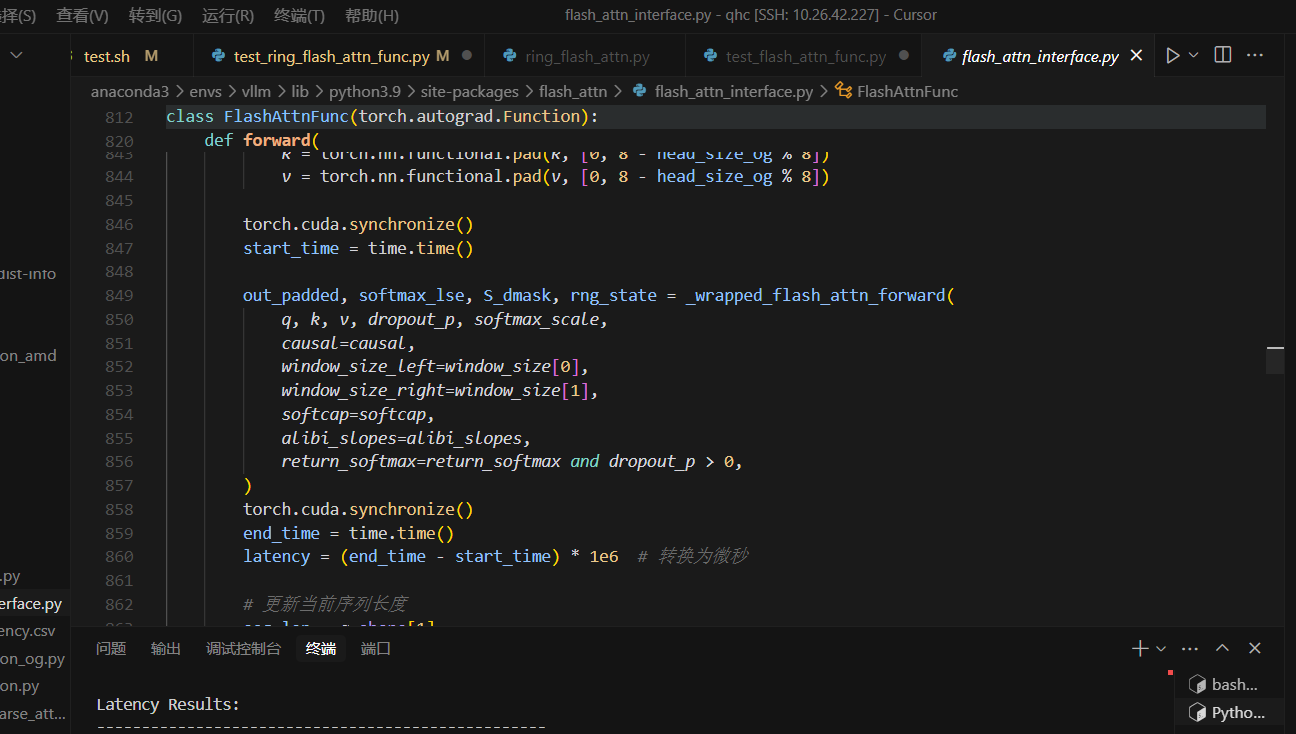
每次实验预热30次，后面再做十次测试取平均值作为数据

Chunk

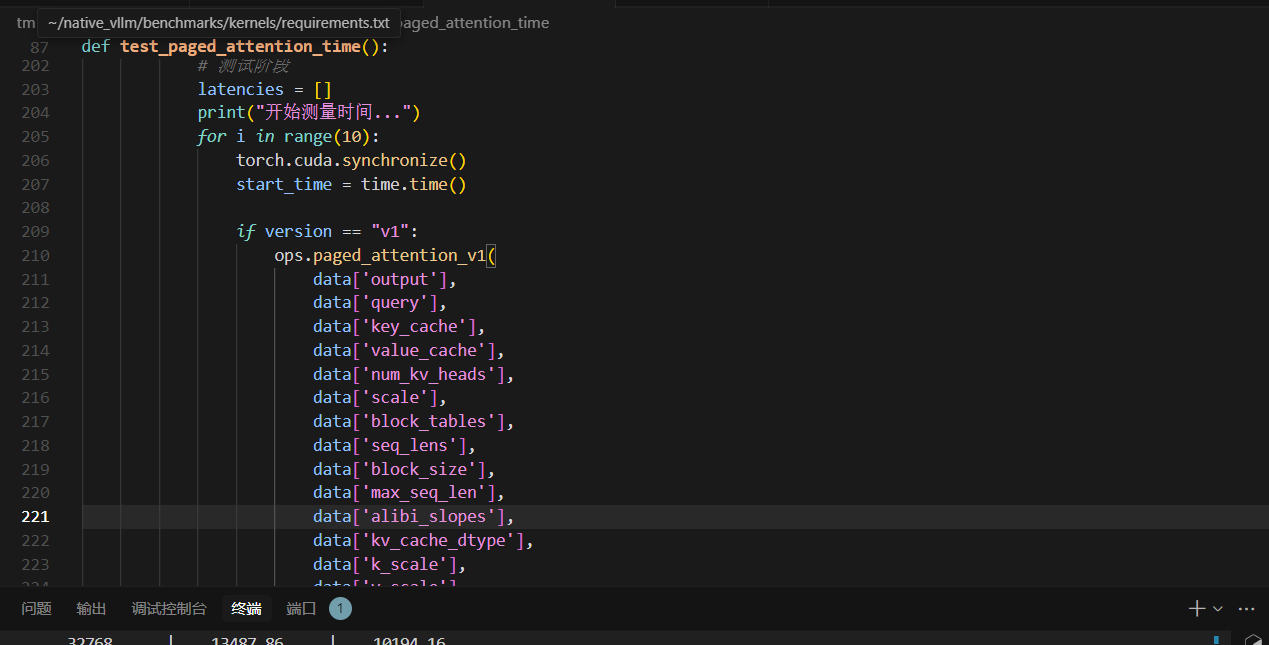
/home/qhc/chunk-attention/tests/test\_chunk\_attn\_cuda.py



Flash

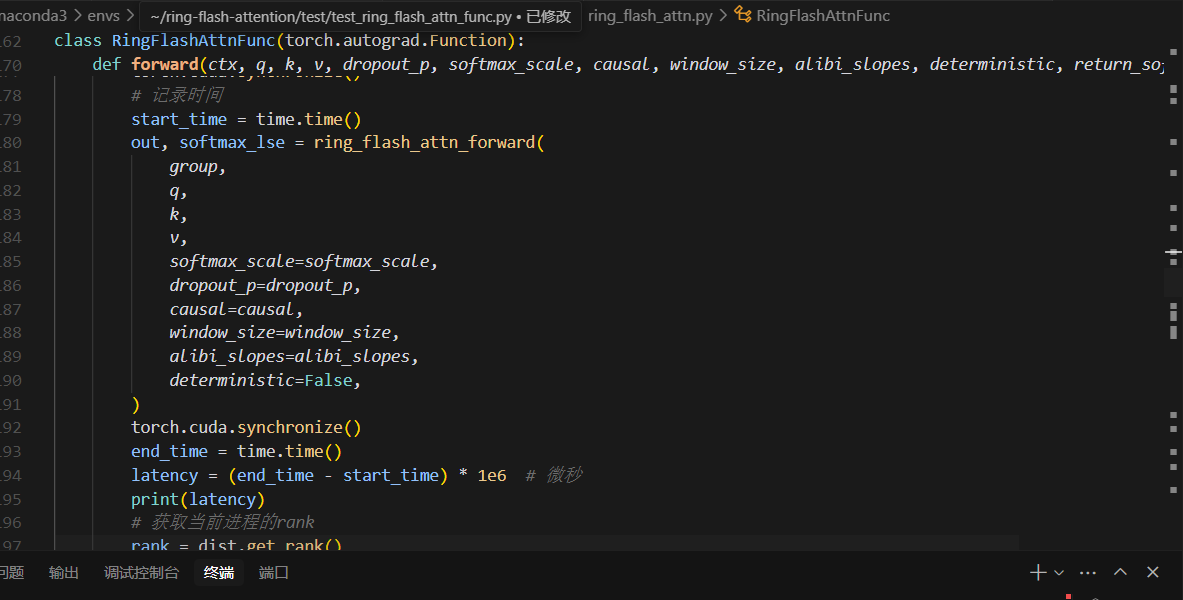
/home/qhc/test\_flash\_attn\_func.py

Paged

/home/qhc/tmp/tests/kernels/test\_paged.py

Ring

/home/qhc/anaconda3/envs/vllm/lib/python3.9/site-packages/ring\_flash\_attn/ring\_flash\_attn.py



实验1 横向对比

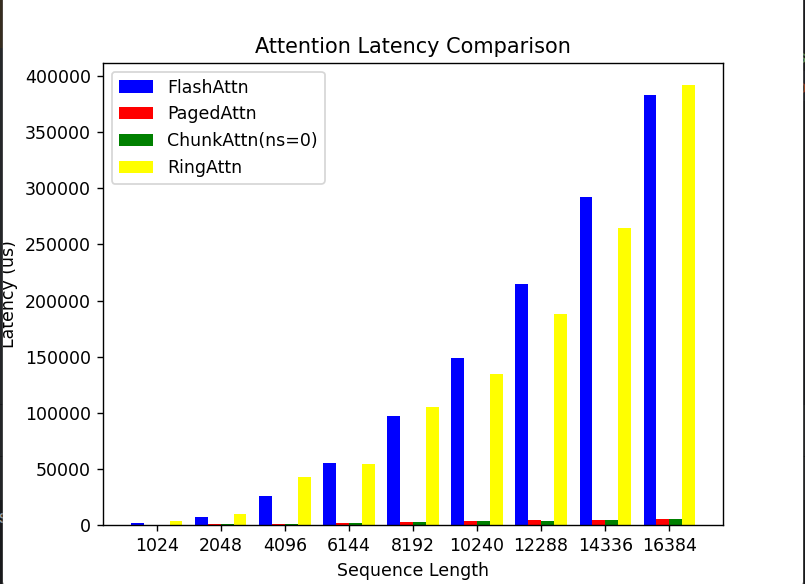
|  | **Attention Latency(us)** | | | |
| --- | --- | --- | --- | --- |
| **seq\_len** | **FlashAttn** | **PagedAttn** | **ChunkAttn(ns=0)** | **RingAttn** |
| 1024 | 2071.499824523926 | 333.74 | 342.92 | 3534.22 |
| 2048 | 6872.55859375 | 689.24 | 664.62 | 9541.79 |
| 4096 | 25605.58319091797 | 1271.61 | 1310.40 | 42766.21 |
| 6144 | 55296.08726501465 | 1903.08 | 1959.35 | 54399.49 |
| 8192 | 96689.39113616943 | 2624.70 | 2607.97 | 105382.01 |
| 10240 | 149172.54447937012 | 3324.34 | 3246.62 | 134161.99 |
| 12288 | 214458.5132598877 | 4051.66 | 3898.43 | 187990.78 |
| 14336 | 292490.69690704346 | 4528.86 | 4544.76 | 264774.87 |
| 16384 | 383563.37547302246 | 5181.19 | 4934.41 | 392073.46 |
| 32768 | 1569379.734992981 | 13487.86 | 9883.00 | - |

FlashAttn ，RingAttn延迟随序列长度增长急剧上升；

PagedAttn 性能略低于Chunk，随序列长度基本线性增长

Chunkattn (ns=0) 性能最优

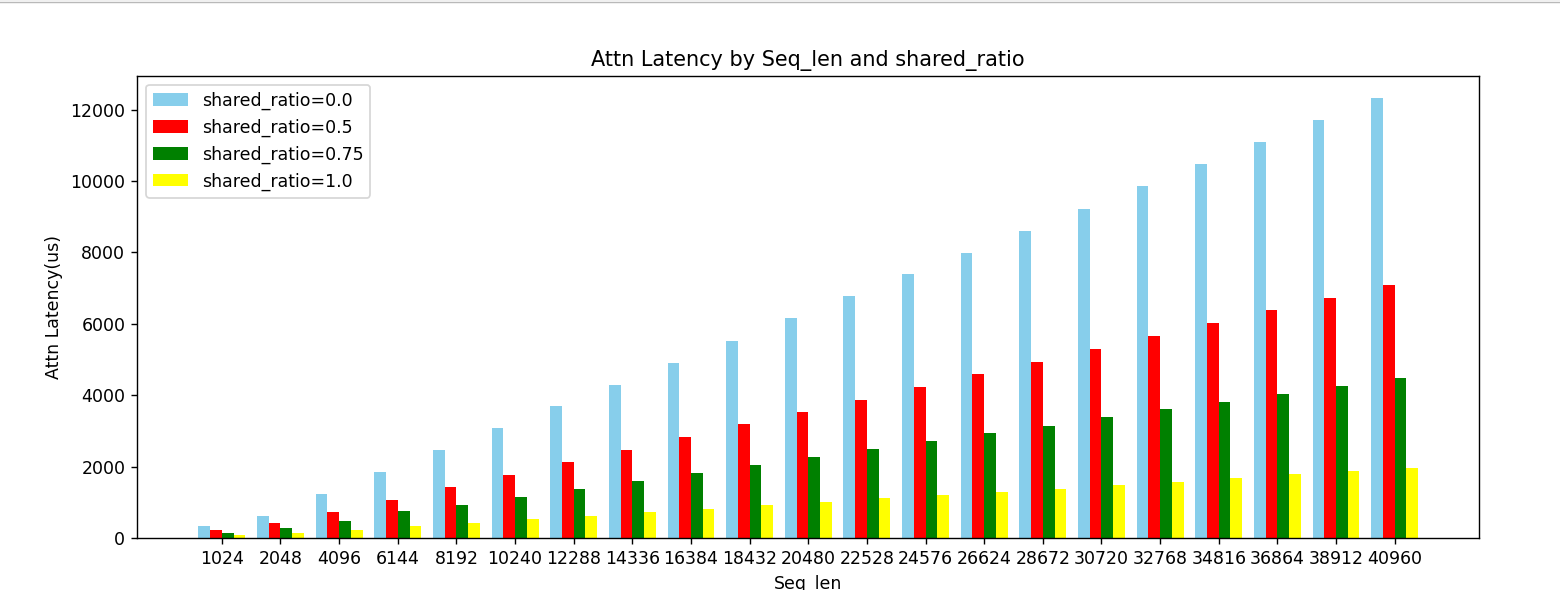
生成柱状图



实验2: ChunkAttn纵向对比

2-1: 1-52k的chunk对比

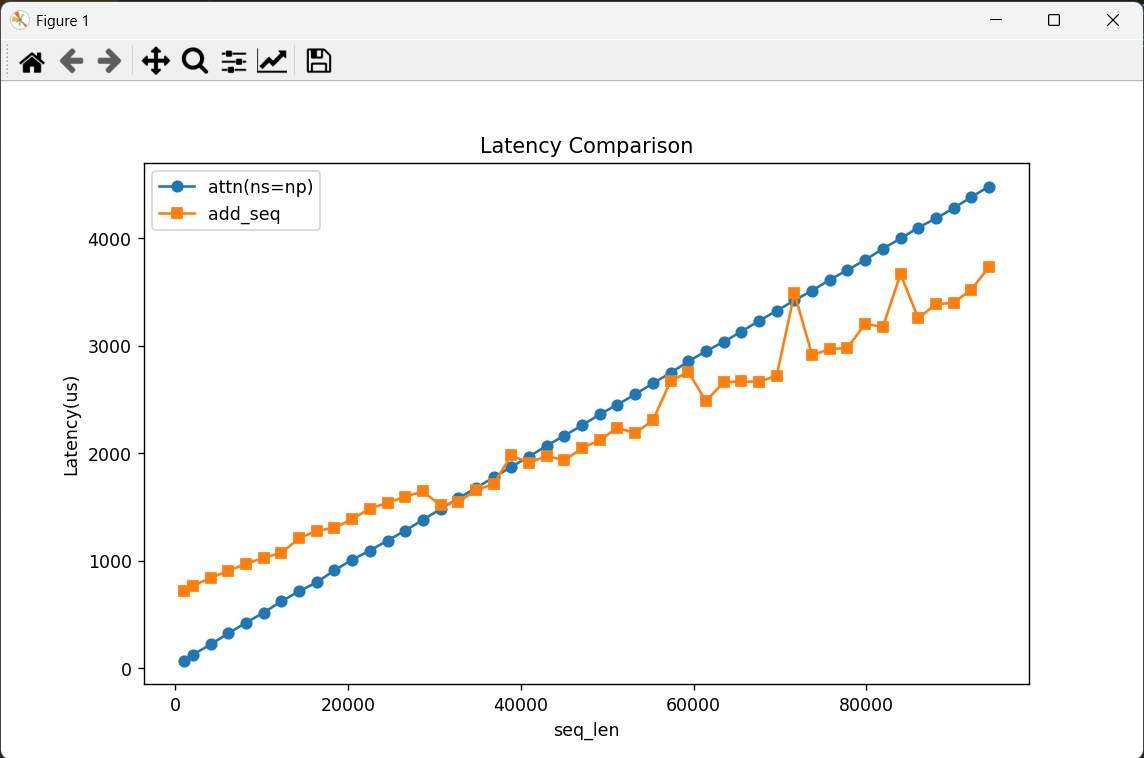
| **Seq\_len** | ****Attn Latency(us)**** | | | |
| --- | --- | --- | --- | --- |
| ****Ratio(ns/np)**** | | | |
| 0.0 | 0.5 | 0.75 | 1.0 |
| 1024 | 323.43 | 216.35 | 149.83 | 72.56 |
| 2048 | 628.13 | 408.33 | 267.82 | 129.12 |
| 4096 | 1238.40 | 730.01 | 478.82 | 221.52 |
| 6144 | 1847.03 | 1076.78 | 744.74 | 328.06 |
| 8192 | 2460.74 | 1423.86 | 926.60 | 418.80 |
| 10240 | 3068.21 | 1771.79 | 1144.78 | 519.78 |
| 12288 | 3684.41 | 2131.57 | 1367.17 | 623.54 |
| 14336 | 4294.58 | 2478.44 | 1591.76 | 717.67 |
| 16384 | 4911.24 | 2825.79 | 1815.61 | 815.79 |
| 18432 | 5529.01 | 3179.58 | 2034.32 | 914.71 |
| 20480 | 6148.42 | 3525.12 | 2257.74 | 1011.42 |
| 22528 | 6769.26 | 3879.04 | 2484.67 | 1106.61 |
| 24576 | 7381.57 | 4236.70 | 2706.37 | 1204.62 |
| 26624 | 7995.26 | 4587.41 | 2928.84 | 1286.75 |
| 28672 | 8610.09 | 4941.46 | 3149.03 | 1382.67 |
| 30720 | 9225.74 | 5305.71 | 3378.34 | 1488.71 |
| 32768 | 9851.85 | 5652.22 | 3599.33 | 1582.23 |
| 34816 | 10467.87 | 6015.62 | 3820.26 | 1680.43 |
| 36864 | 11088.56 | 6373.56 | 4045.14 | 1780.83 |
| 38912 | 11704.15 | 6722.90 | 4271.03 | 1881.84 |
| 40960 | 12323.96 | 7079.55 | 4495.12 | 1971.93 |
| 43008 | 12933.07 | 7438.05 | 4713.85 | 2071.57 |
| 45056 | 13544.64 | 7786.43 | 4937.09 | 2170.09 |
| 47104 | 14177.69 | 8143.64 | 5164.99 | 2266.27 |
| 49152 | 14789.24 | 8502.46 | 5388.02 | 2367.13 |
| 51200 | 15405.52 | 8865.04 | 5617.33 | 2459.92 |
| 53248 | 16030.42 | 9228.55 | 5846.50 | 2554.34 |



共享前缀带来显著优化

2-2：1-92k 的 add\_seq时间和attn时间开销比较（ns=np)

|  | **Latency(us)** | |
| --- | --- | --- |
| **seq\_len** | **attn** | **add\_seq** |
| 1024 | 70.52 | 720.82 |
| 2048 | 126.89 | 767.14 |
| 4096 | 222.05 | 841.83 |
| 6144 | 325.92 | 907.54 |
| 8192 | 424.46 | 971.97 |
| 10240 | 518.61 | 1028.22 |
| 12288 | 623.12 | 1074.25 |
| 14336 | 717.51 | 1208.81 |
| 16384 | 800.21 | 1280.15 |
| 18432 | 911.74 | 1307.13 |
| 20480 | 1010.47 | 1387.76 |
| 22528 | 1096.14 | 1486.81 |
| 24576 | 1188.83 | 1539.81 |
| 26624 | 1281.31 | 1597.34 |
| 28672 | 1381.64 | 1646.60 |
| 30720 | 1482.38 | 1515.92 |
| 32768 | 1582.12 | 1551.27 |
| 34816 | 1678.28 | 1655.23 |
| 36864 | 1775.58 | 1717.01 |
| 38912 | 1875.16 | 1980.74 |
| 40960 | 1969.92 | 1913.34 |
| 43008 | 2071.57 | 1976.37 |
| 45056 | 2165.85 | 1934.19 |
| 47104 | 2260.42 | 2049.97 |
| 49152 | 2363.71 | 2124.22 |
| 51200 | 2453.57 | 2237.74 |
| 53248 | 2549.73 | 2188.77 |
| 55296 | 2651.35 | 2306.70 |
| 57344 | 2749.05 | 2671.42 |
| 59392 | 2854.80 | 2758.51 |
| 61440 | 2951.54 | 2487.34 |
| 63488 | 3036.95 | 2661.06 |
| 65536 | 3133.38 | 2669.89 |
| 67584 | 3232.93 | 2665.05 |
| 69632 | 3327.24 | 2724.19 |
| 71680 | 3423.29 | 3488.79 |
| 73728 | 3514.50 | 2916.38 |
| 75776 | 3614.69 | 2967.60 |
| 77824 | 3706.32 | 2982.26 |
| 79872 | 3799.33 | 3204.51 |
| 81920 | 3904.29 | 3176.39 |
| 83968 | 3999.34 | 3668.51 |
| 86016 | 4099.95 | 3257.74 |
| 88064 | 4186.23 | 3387.10 |
| 90112 | 4281.12 | 3401.64 |
| 92160 | 4384.89 | 3523.66 |
| 94208 | 4482.83 | 3733.80 |



在ns=np时Add\_seq开销与Attn计算开销基本持平

2-3: 78-100k的一些综合对比

| **Seq\_len** | **Attn Latency(us)** | | | | **Add\_seq Latency(us)** | |
| --- | --- | --- | --- | --- | --- | --- |
| **ns/np** | | | | **ns/np** | |
| 0.0 | 0.5 | 0.75 | 1.0 | 0.5 | 1.0 |
| 79872 | 24074.53 | 13883.46 | 8815.18 | 3806.30 | 383893.19 | 3010.57 |
| 81920 | - | 14240.56 | 9001.15 | 3904.42 | 397366.42 | 3062.29 |
| 83968 | - | 14603.32 | 9264.71 | 3992.00 | 403763.67 | 3098.30 |
| 86016 | - | 14961.77 | 9493.75 | 4097.81 | 394025.02 | 3113.39 |
| 88064 | - | 15345.23 | 9691.93 | 4189.12 | 413613.47 | 3198.31 |
| 90112 | - | 15678.49 | 9964.33 | 4282.29 | 407649.49 | 3238.15 |
| 92160 | - | 16045.89 | 10216.10 | 4386.45 | 416251.52 | 3321.29 |
| 94208 | - | 16402.40 | 10422.79 | 4472.79 | 424239.69 | 3353.92 |
| 96256 | - | 16780.91 | 10645.26 | 4569.21 | 434907.74 | 3504.69 |
| 98304 | - | 17093.58 | 10846.11 | 4662.86 | 448864.98 | 3575.22 |
| 100352 | - | 17450.73 | 11078.70 | 4775.84 | 458419.60 | 3631.45 |
| 102400 | - | 17812.76 | 11308.86 | 4868.69 | 467814.45 | 3605.44 |

注：这些数据比较大，不同的ratio我申请了不同的memory\_mb才能跑，ratio相同时的add\_seq时间和attn时间比较才有意义！

Add\_seq Latency随序列变化小幅增长

而共享前缀占比较小时，add\_seq时间开销巨大！