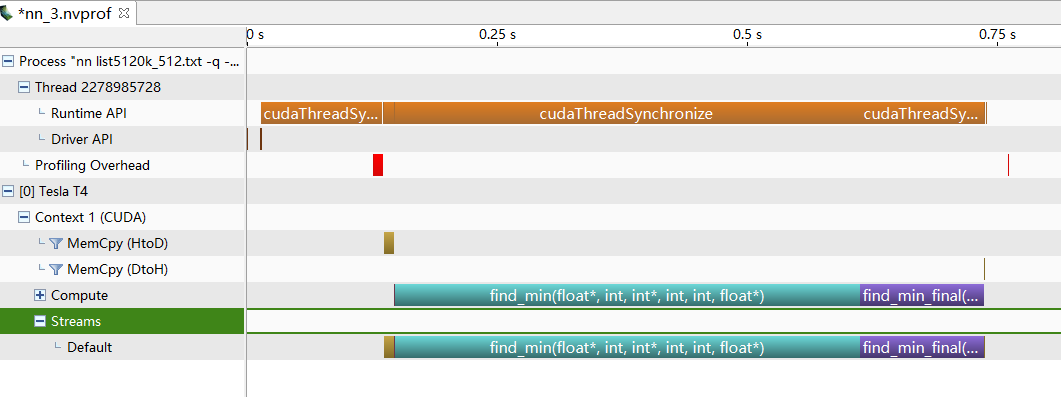
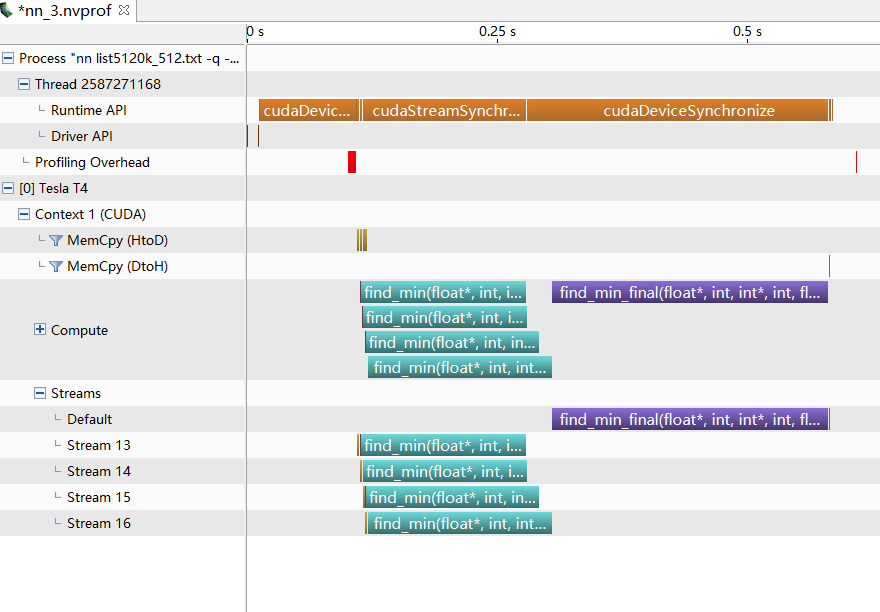
## Summery:

By using CUDA stream, computational tasks and data transport are segmented into multiple streams for concurrent processing on the GPU. the data will be divided into small blocks and transport to device simultaneously in each stream, kernel 'find\_min' will also be operated within these streams to find the the ResountCount of the smallest distance in each block . Once the 'find\_min' processes in all streams complete, the 'find\_min\_final' kernel will be executed in the default stream to find the ResountCount of the smallest distance for all data according to the result of 'find\_min'.



NVVP result without CUDA streams, datasize: 5120K\_512, ResultCount: 3000



NVVP result using 4 CUDA streams, datasize: 5120K\_512, ResultCount: 3000

## Stream Count Optimization:

Increasing the number of streams reduces the data size and execution time per stream for the 'find\_min' kernel. However, this increasing also multiplies the data size for the 'find\_min\_final' kernel significantly, making the execution time longer. To optimize efficiency, it’s important to adjust the number of streams.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Result Count** | **Origin K\_NN(s)** | **QuickSelect(ms)**  **(without Stream)** | **QuickSelect(ms)**  **(4 Streams)** | **QuickSelect(ms)**  **(10 Streams)** |
| 100 | 4.569 | 308.07 | 284.71 | 343.46 |
| 300 | 8.673 | 414.32 | 386.20 | 443.35 |
| 500 | 9.939 | 416.85 | 398.46 | 421.35 |
| 1000 | 17.097 | 555.68663 | 556.28 | 553.19 |
| 3000 | 45.063 | 683.45 | 538.55 | 812.35 |
| 5000 | 72.227 | 721.17 | 835.74 | 1008.23 |
| 10000 | 166.824 | 843.93 | 1155.86 | 1047.08 |
| 30000 | 417.156 | 1259.61 | 1474.87 | 1822.01 |
| 50000 | 697.713 | 1732.61 | 1891.54 | 2204.22 |
| 100000 | --- | 2587.19 | 2313.60 | 2869.76 |

Data Size: 5120K\_512

## Thread Per Block:

When data size is large enough, Thread Per Block has little effect on performance

|  |  |
| --- | --- |
| **Thread Per Block** | **Execution Time** |
| 64 | 572.76 |
| 128 | 570.63 |
| 256 | 556.06 |
| 512 | 563.30 |
| 1024 | 546.32 |

Using 4 CUDA streams, datasize: 5120K\_512, ResultCount: 3000