Radix Select (Min)

Brief

Use radix select method to select top `k` minimal elements(result unordered)

Method

The method is similar to radix sort. But instead we sort the array, we only need to find the pivot that can split the array into two parts. One is all bigger than pivot, the other is all smaller than pivot and have size `k`

1. Read the 8 bits and collect histogram of it(256 bins). `func:colect_histogram`
2. Use prefix sum to determine which bin the pivot is in. `func:set_limit`
3. Copy the data that is smaller than pivot(and in the same bin with pivot) to new space to look next 8 bits.`func:relocation` Return to step 1.

Atomic operations is used to ensure no conflict in relocation.

Note: Atomic operation used as prefix sum method is as quick as global prefix scan as long as we do not care the order of the result.

Usage

```
template <typename T>
void radix_select(T *d_data, int n, T *result, int topk);
```

- TypeName: `T` unsigned or signed or pointer:
  - char, unsigned char, signed char
  - short , unsigned short, signed short
  - int, unsigned int, signed int
  - long long, unsigned long long, signed long long
  - *int, *float, *struct_name, *string, etc.
- T* `d_data` device pointer of source data
- int `n` size of source data;
- T* `result` host pointer where to store the result
- int `topk` top-k minimal value what to select

see example on `radix_select.cu` `func:origin`

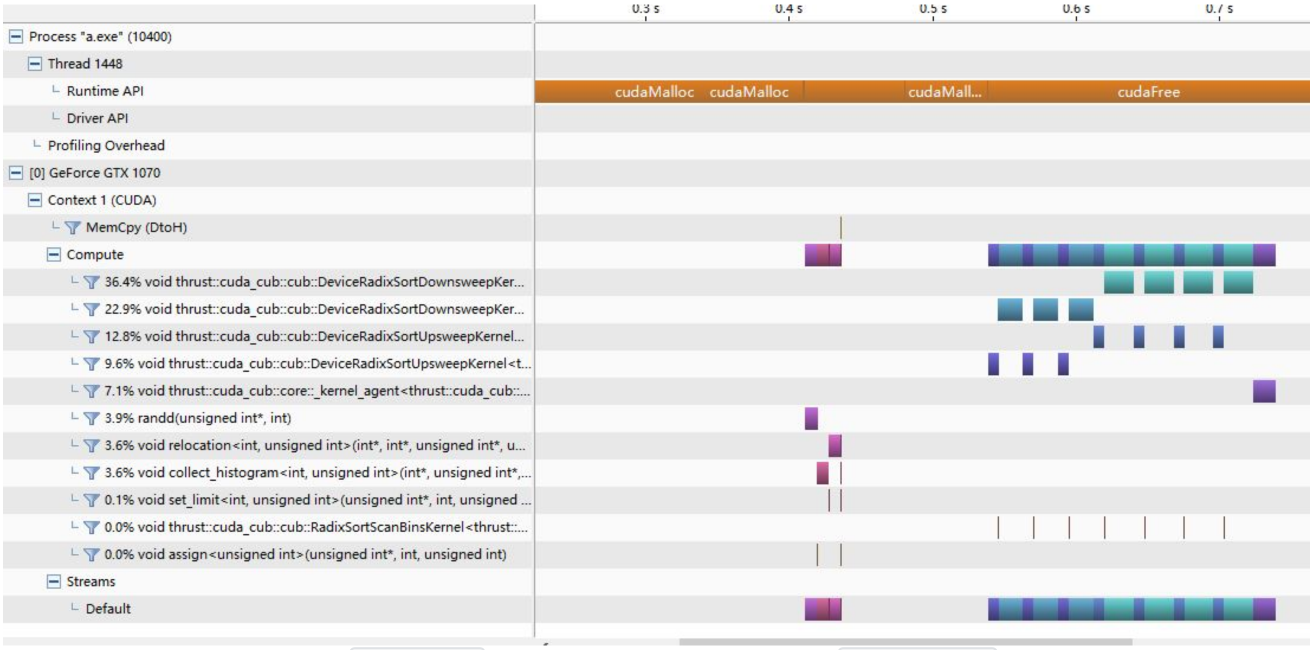
```
template <typename T, typename U>
void radix_select_detail(T *d_data, int n, T *d_data1, T *d_data2, int topk, unsigned *d_params,
U *d_limits);
```

- TypeName: `T` unsigned or signed or pointer:
    - char, unsigned char, signed char
    - short , unsigned short, signed short
    - int, unsigned int, signed int
    - long long, unsigned long long, signed long long
    - *int, *float, *struct_name, *string, etc.
- TypeName: U unsigned type as same length as T, data of type T convert to type U and which is actually doing the sorting:
    - unsigned char
    - unsigned short
    - unsigned int
    - unsigned long long
    - unsigned long long
- T* `d_data` device pointer of source data
- int `n` size of source data
- T* `d_data1` device pointer of temp storage data_1, showed be as same size as n
- T* `d_data2` device pointer of temp storage data_12, showed be as same size as n, if you do not want to keep the origin data, this can be same address as `d_data`.
- int `topk` top-k minimal value what to select
- unsigned* `d_params` temp storage of size (256 + 256 * 90 + sizeof(T) * 8 + 10):
    - histogram: 256
    - prefix: 90 * 256
    - d_total: sizeof(T) * 8+10
- U* `d_limits` temp storage of size 2;

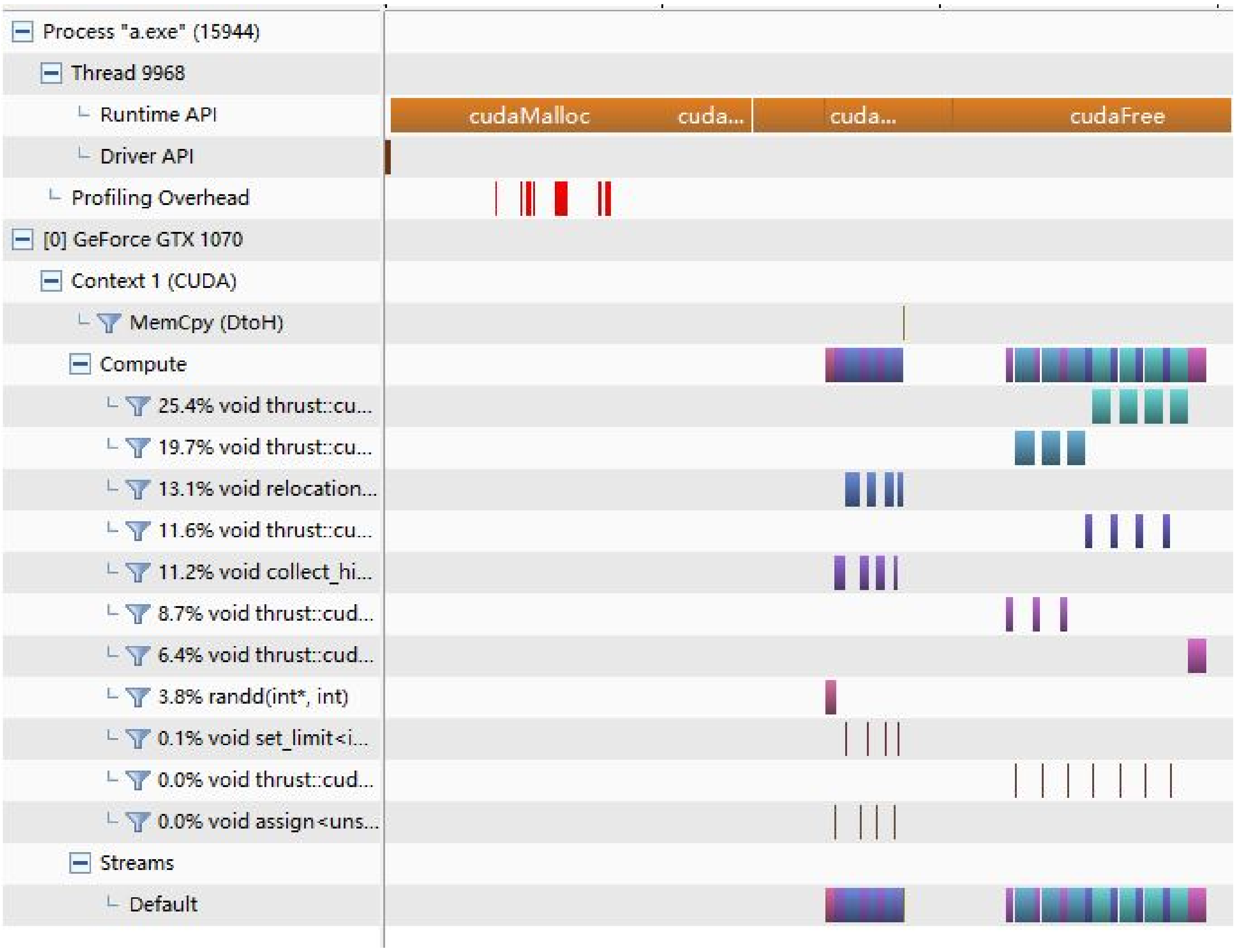
if you want to manage the memory yourself ,use `func:radix_select_detail`, see the example on `radix_selcet.inl` `func:radix_select`

Test

when data is sparse in space `sparse.exe`, radix select is 9X faster than `thrust:sort`!


0.3 s
0.4 s
0.5 s
0.6 s
0.7 s
Process "a.exe" (10400)
Thread 1448
Runtime API
Driver API
Profiling Overhead
[0] GeForce GTX 1070
Context 1 (CUDA)
MemCpy (DtoH)
Compute
36.4% void thrust::cuda_cub::cub::DeviceRadixSortDownsweepKer...
22.9% void thrust::cuda_cub::cub::DeviceRadixSortDownsweepKer...
12.8% void thrust::cuda_cub::cub::DeviceRadixSortUpsweepKernel...
9.6% void thrust::cuda_cub::cub::DeviceRadixSortUpsweepKernel<t...
7.1% void thrust::cuda_cub::core::_kernel_agent<thrust::cuda_cub::...
3.9% randd(unsigned int*, int)
3.6% void relocation<int, unsigned int>(int*, int*, unsigned int*, u...
3.6% void collect_histogram<int, unsigned int>(int*, unsigned int*,...
0.1% void set_limit<int, unsigned int>(unsigned int*, int, unsigned ...
0.0% void thrust::cuda_cub::cub::RadixSortScanBinsKernel<thrust::...
0.0% void assign<unsigned int>(unsigned int*, int, unsigned int)
Streams
Default
cudaMalloc
cudaMalloc
cudaMall...
cudaFree

when data is dense in space `dense.exe` radix select is 3X faster than `thrust:sort` !


Process "a.exe" (15944)
Thread 9968
Runtime API
Driver API
Profiling Overhead
[0] GeForce GTX 1070
Context 1 (CUDA)
MemCpy (DtoH)
Compute
25.4% void thrust::cu...
19.7% void thrust::cu...
13.1% void relocation...
11.6% void thrust::cu...
11.2% void collect_hi...
8.7% void thrust::cud...
6.4% void thrust::cud...
3.8% randd(int*, int)
0.1% void set_limit<i...
0.0% void thrust::cud...
0.0% void assign<uns...
Streams
Default
cudaMalloc
cuda...
cuda...
cudaFree