

VWN接入MTPEmbeddingProjection

效果：

1. loss: 不能bitwise对齐，但是是deterministic的，符合预期

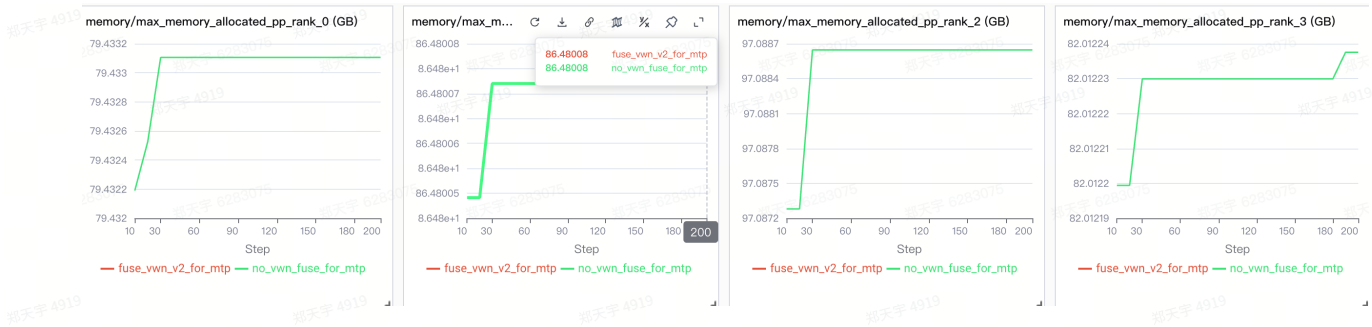


2. Mfu: 普遍提高



3. memory

峰值显存不变

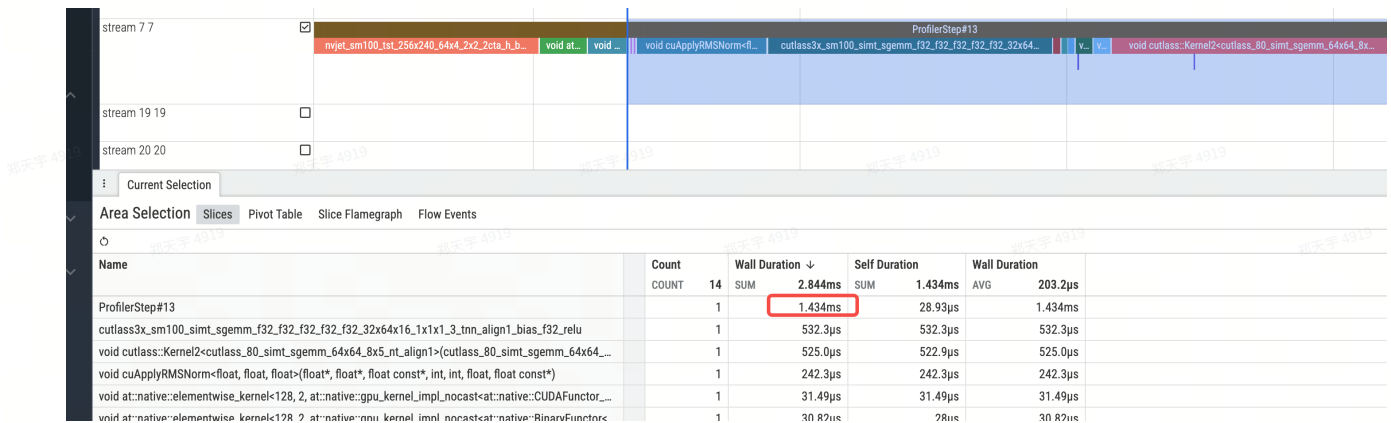


4. kernel减少时间:

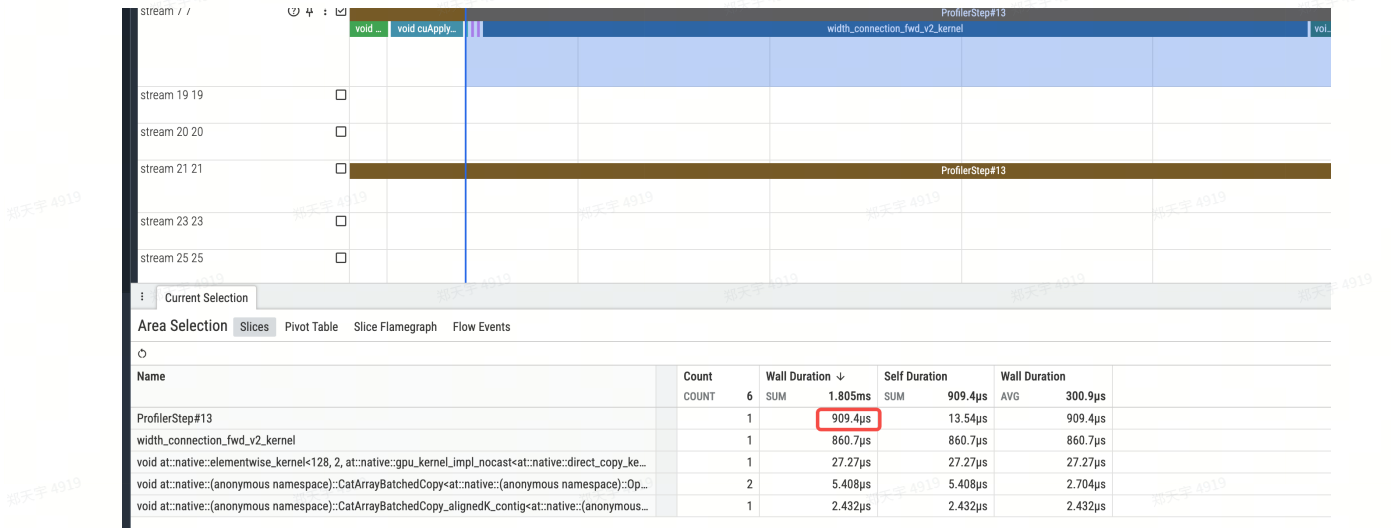
- width_connection

1.434 ms->0.909 ms

Before:



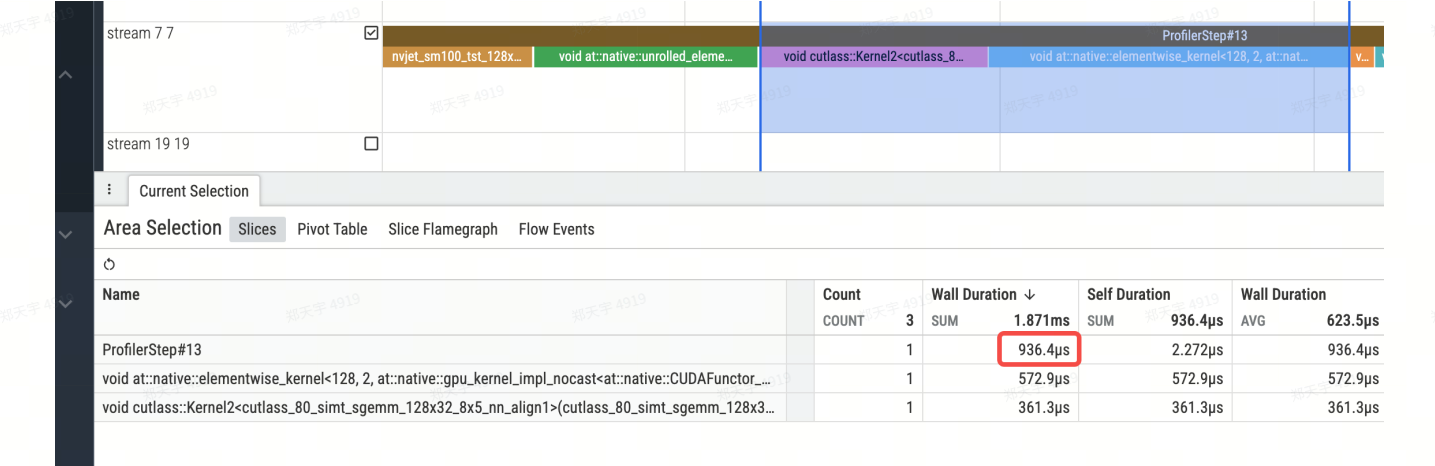
After:



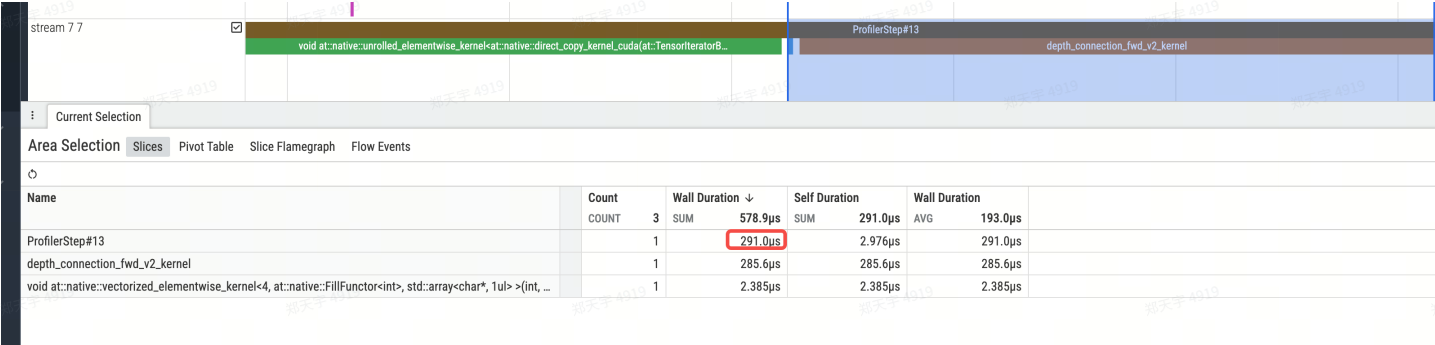
- depth_connection

0.936 ms->0.291 ms

Before:



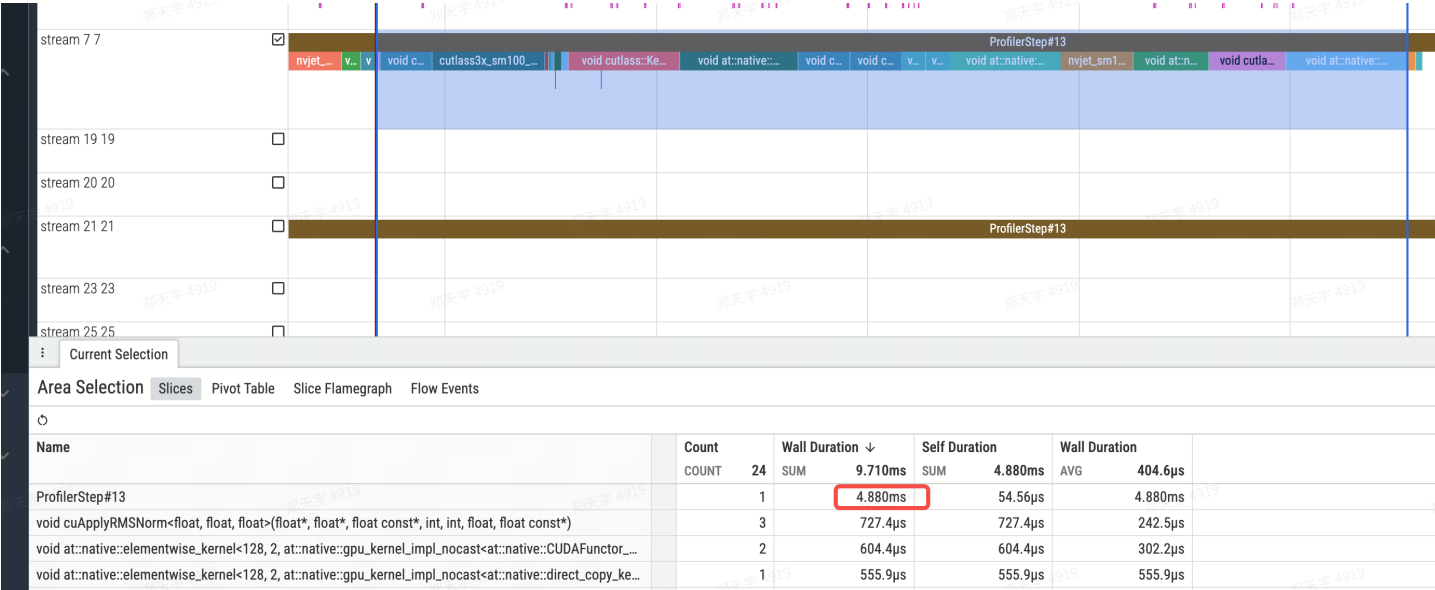
After:



- MTPEmbeddingProject整个时间

4.880->3.116

Before:



After:

stream 7 7	☒	v void at::n... v... nvjet_sm100_tst_256x240_64x4_2x2_2cta_h... v v	ProfilerStep#13 width_connection_fwd_v2_ker... void c... void c... v... v... void at::nativ... nvjet_sm1... void at::n... depth_c...
stream 19 19	☐		
stream 20 20	☐		
stream 21 21	☐		ProfilerStep#13
stream 23 23	☐		

⋮

Current Selection

Area Selection

Slices

Pivot Table

Slice Flamegraph

Flow Events

Name							
	Count		Wall Duration ↓		Self Duration		Wall Duration
	COUNT	12	SUM	6.207ms	SUM	3.116ms	AVG 517.3µs
ProfilerStep#13		1		3.116ms		23.94µs	3.116ms
width_connection_fwd_v2_kernel		1		860.7µs		860.7µs	860.7µs
void at::native::(anonymous namespace)::CatArrayBatchedCopy_alignedK_contig<at::native::(anonymous...		1		503.7µs		503.7µs	503.7µs
void cuAnnluDMSNorm(float float float/(float* float* float cnnct* int int float float cnnct*)		2		490.0µs		490.0µs	245.0µs