分析pp-3 fwd慢的原因

2k卡，pp8 3dtimeline：

https://megavision.byteintl.net/perfetto?perfettoUrlParams=%7B%22theme%22%3A%22light%22%7D&trace=%7B%22type%22%3A%22arnoldNdtimeline%22%2C%22params%22%3A%7B%22view%22%3A%22compute%22%2C%22run_id%22%3A11%2C%22role_id%22%3A0%2C%22rank_ids%22%3A%5B0%2C256%2C512%2C768%2C1024%2C1280%2C1536%2C1792%5D%2C%22trial_id%22%3A300752278%2C%22step_from%22%3A12%2C%22step_to%22%3A13%2C%22topology_type%22%3A%22megatron%22%2C%22topology_shape%22%3A%5B256%2C8%2C1%5D%7D%7D

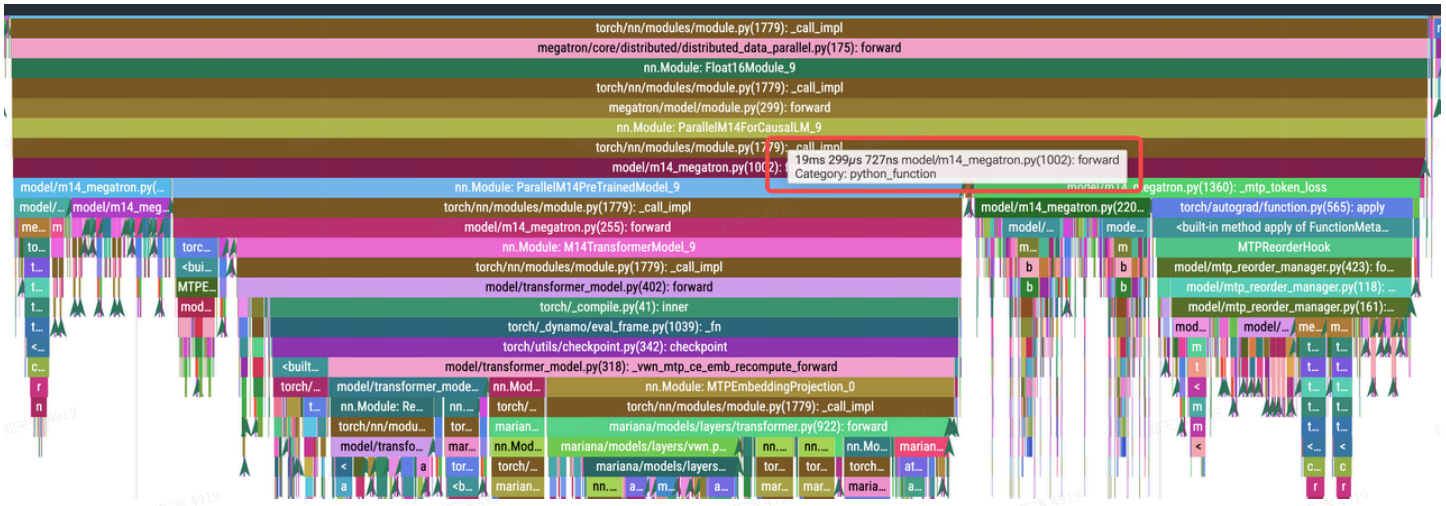
pp-3 profile：https://ui.perfetto.dev/#!/viewer?local_cache_key=1-json

pp-1 profile：https://ui.perfetto.dev/#!/viewer?local_cache_key=1-json

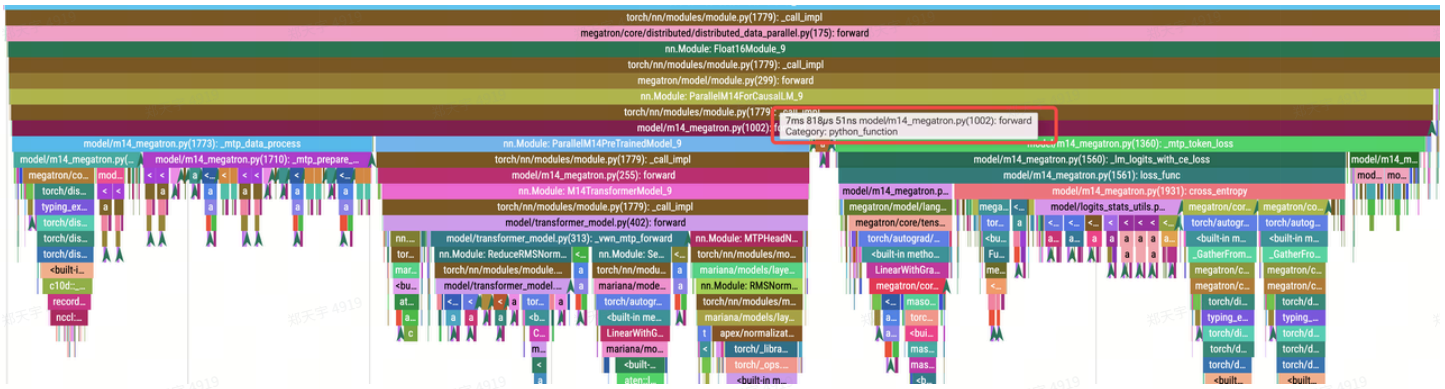
分析：

CPU time 对比同一个step，同一个micro_batch forward的时间：

pp-3:（19ms.299us）



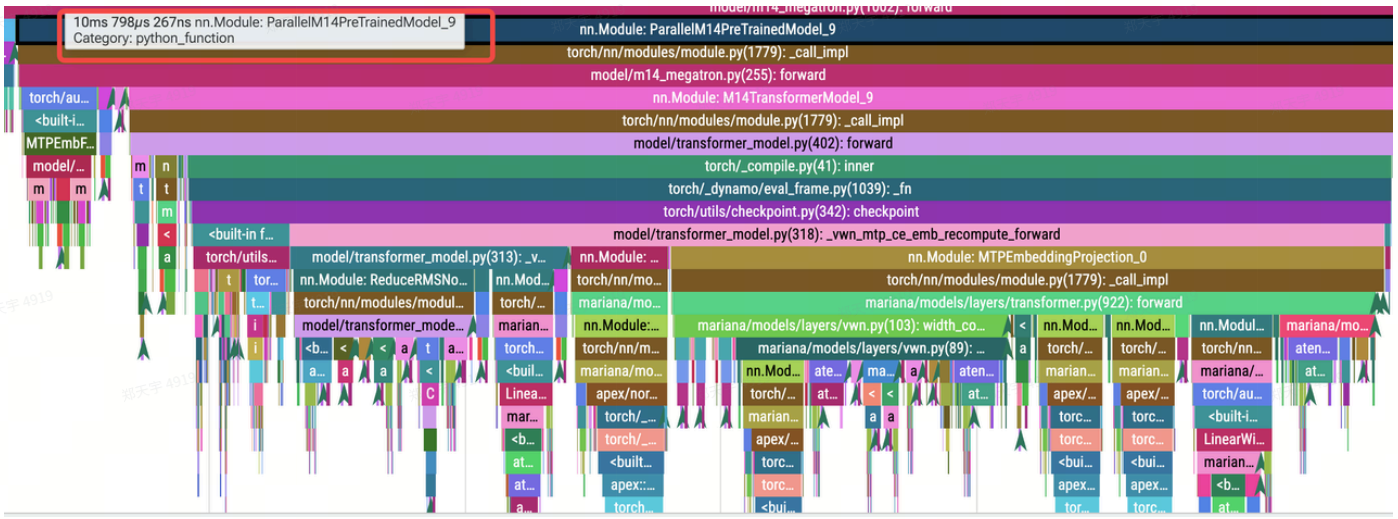
pp-1:(7ms818us)



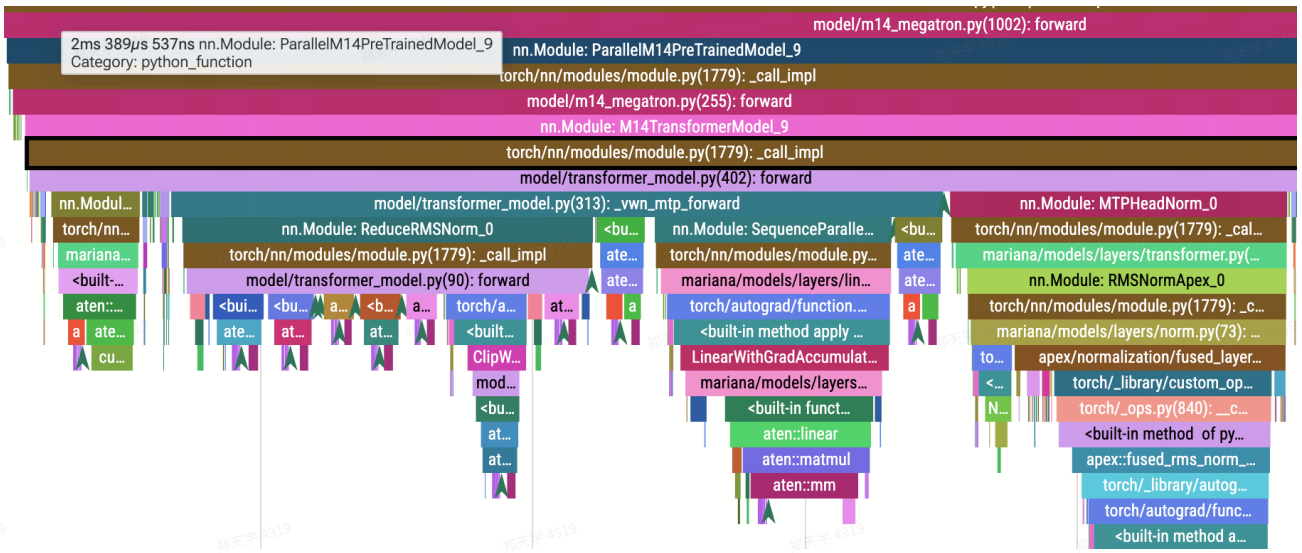
分析CPU Stack：

1.ParallelM14PreTrainedModel_9 forwrward(相差约8.5 ms)

pp-3:



pp-0:

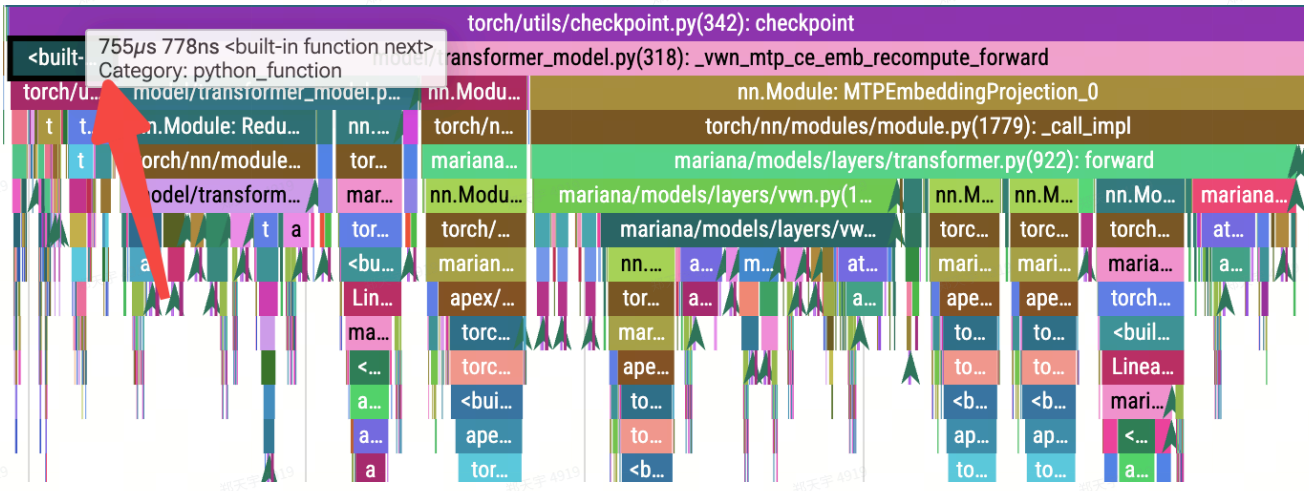


多出的时间如下：

1. MTPEmbFluxRecv

接受embedding_tokens + copy到本地数据 约1ms（pp-3独有）

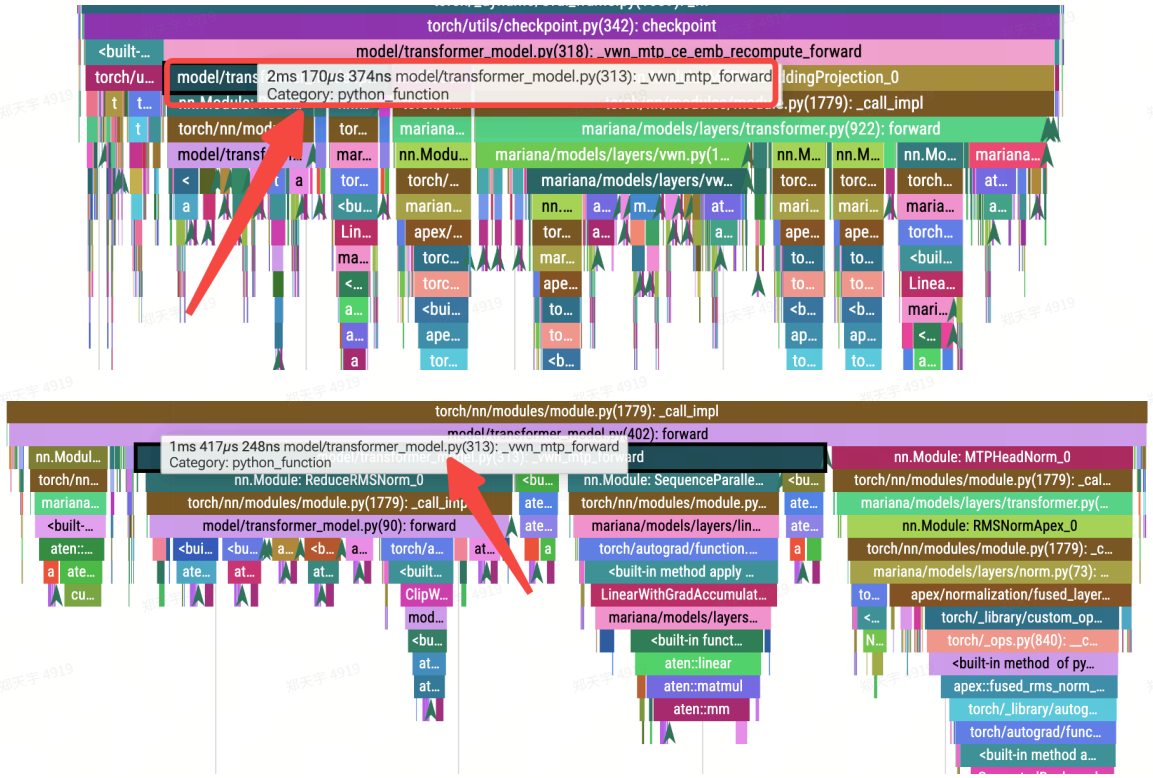
2. torch的checkpoint函数构建约1ms（pp-3独有）



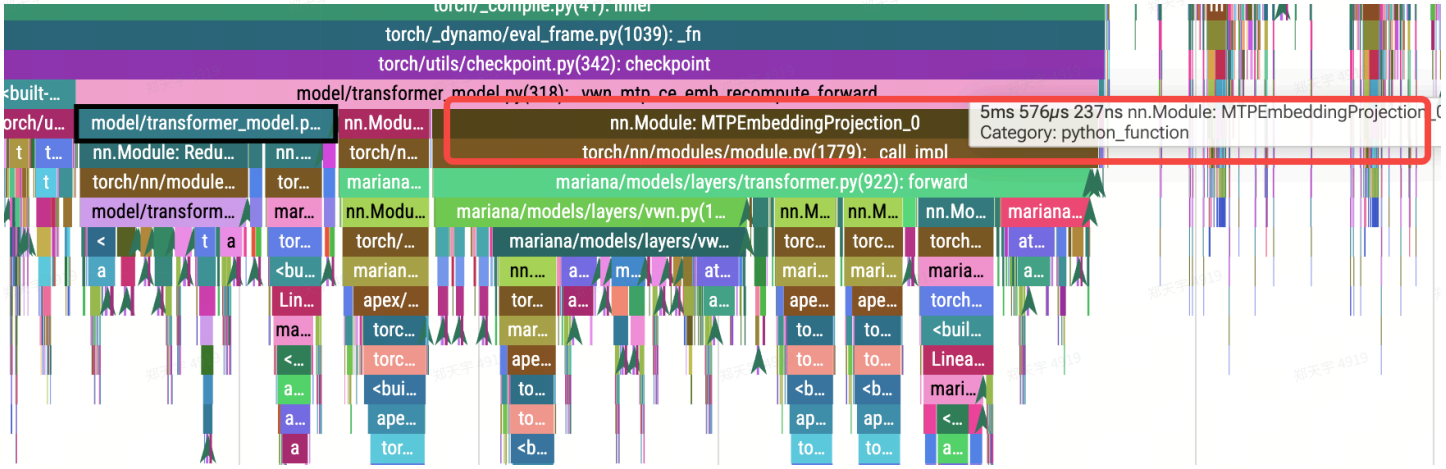
3. _vwn_mtp_forward(0.7ms)

注意 这里的_vwn_mtp_forward 并不是vwn算法的 width和depth操作，是为mtp特定做了两个layer，而实际的vwn算法的操作都用的fuse_kernel，因此那部分几乎没有增加时间。

主要都是torch 的 checkpoint 插 hook导致的时间

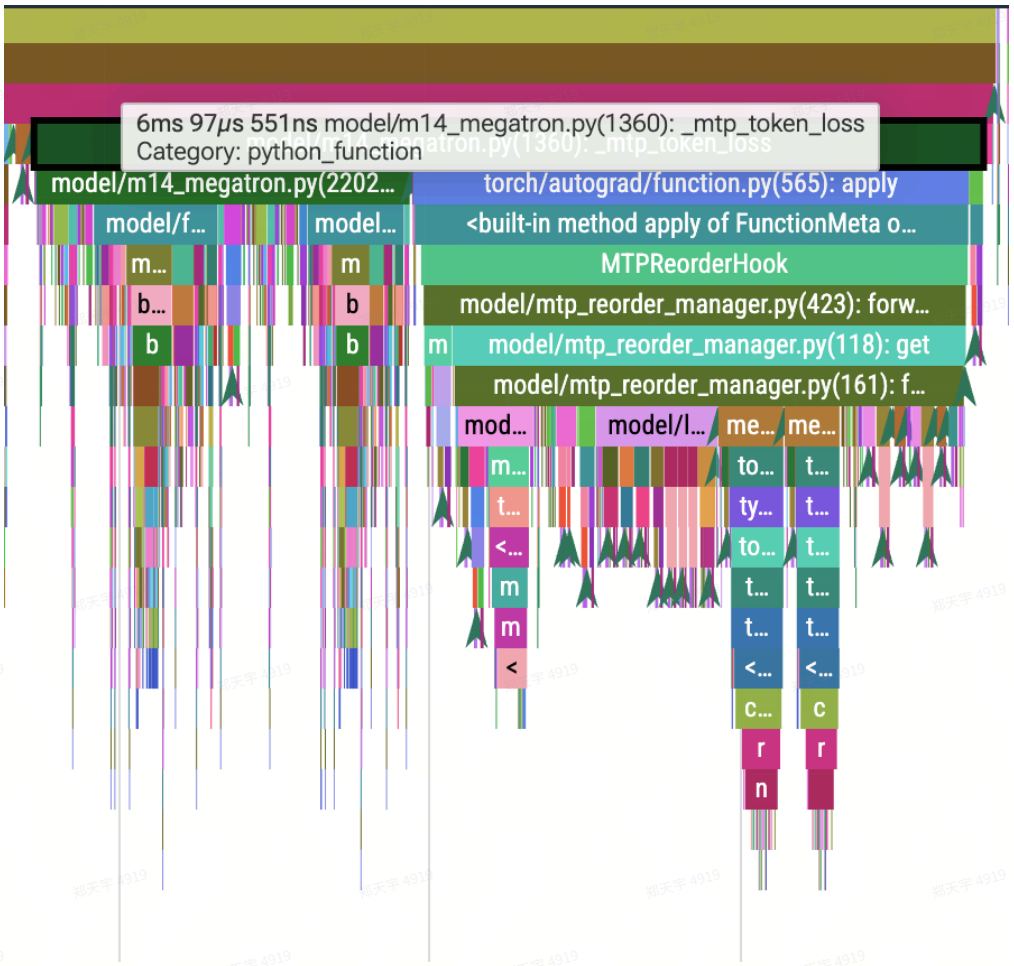


4. MTPEmbeddingProjection （约5.5ms）

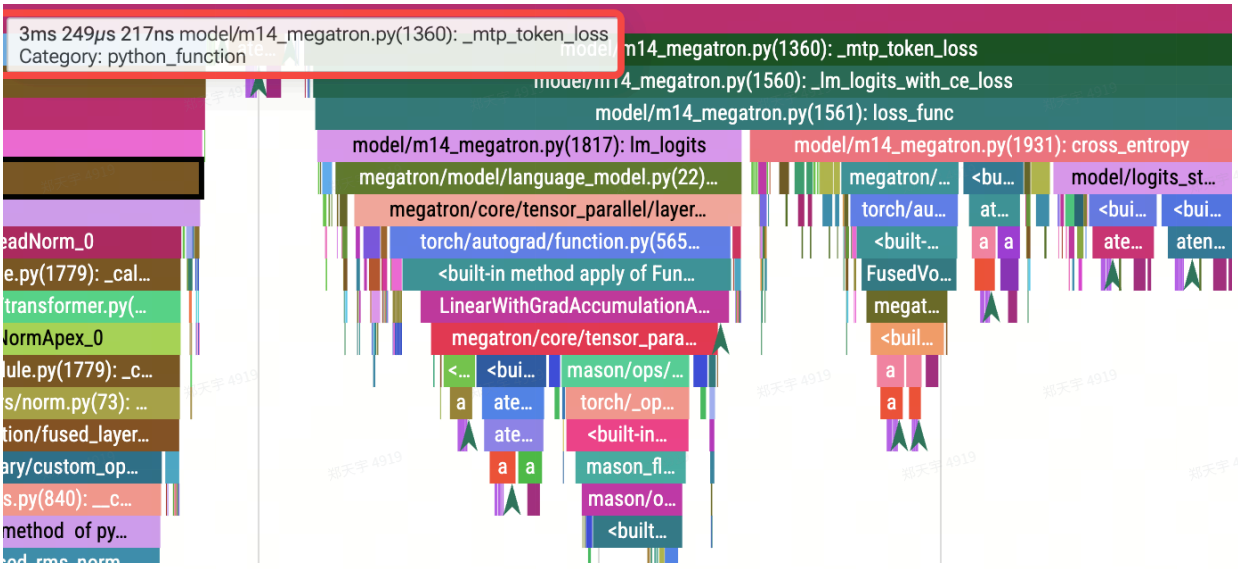


2.MTP_LOSS(约3ms)

- pp-3:



- pp-1:



分析GPU Stack

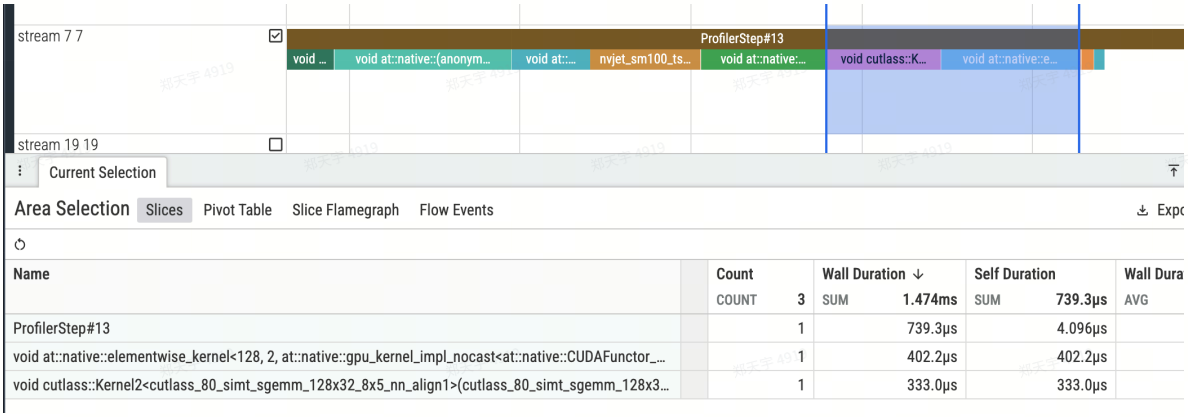
同一步同一个micro_batch的fwd：

大概查了一下，MTPEmbeddingProjection带的op有比较大的延迟

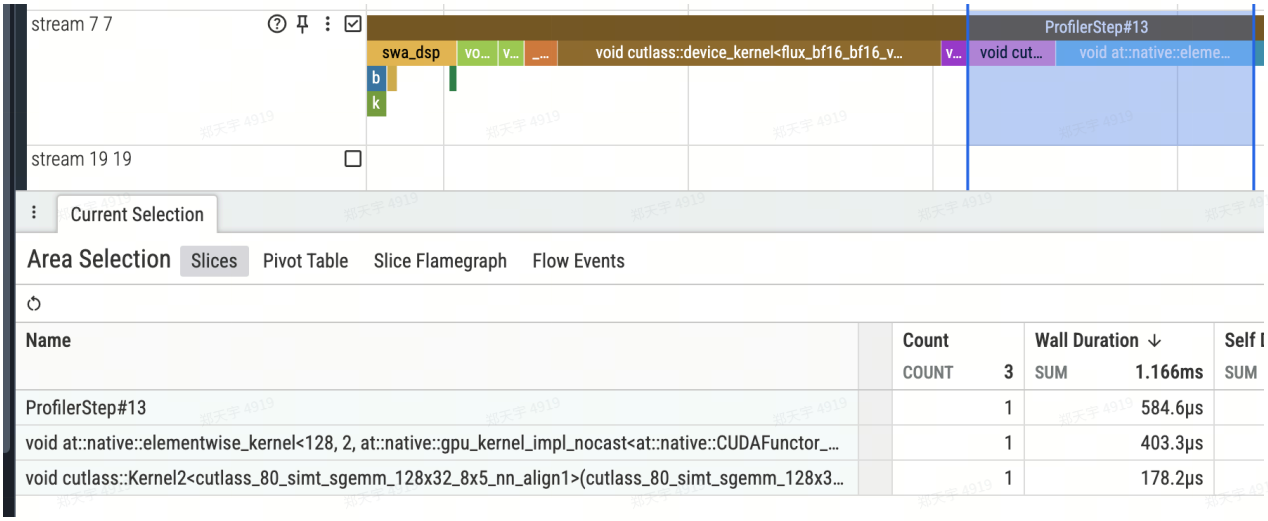
加入fuse vwn kernel后继续分析pp-3慢的原因：

- 首先确认一下当前的pp-3的depth_connection操作和bigop中的depth_connection的gpu执行相同（depth_connection都只用到这两个kernel，虽然两个调用的不同的接口函数，但都在python端实现的）：

pp-3(MTPEmbeddingProjection)：

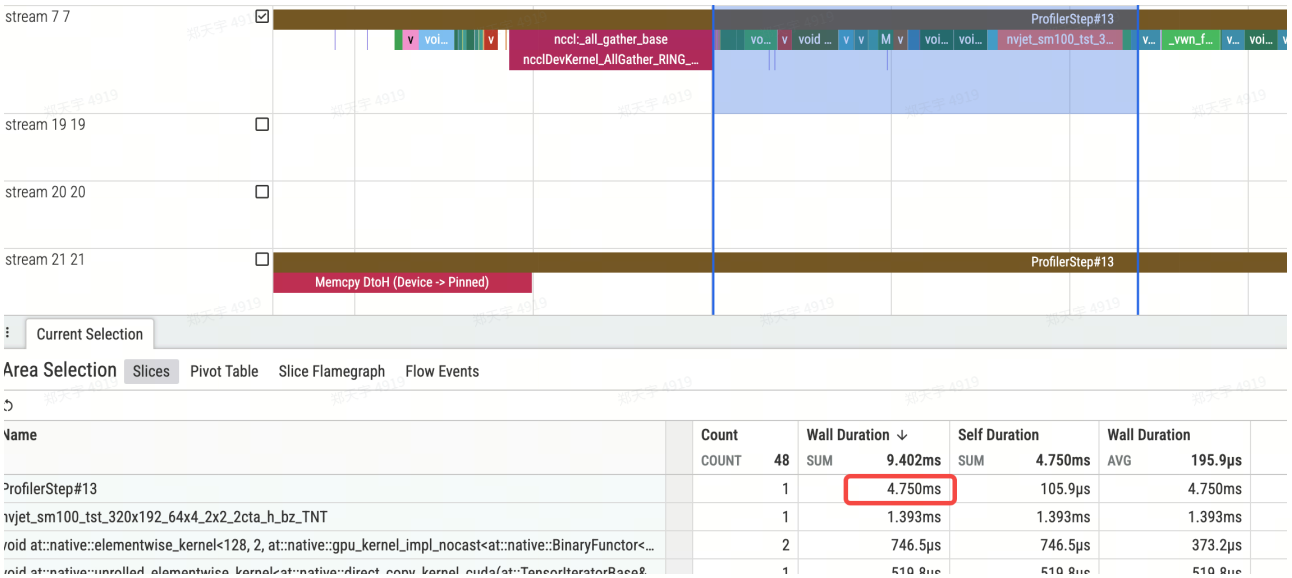


pp-1(TransformerBlock)：

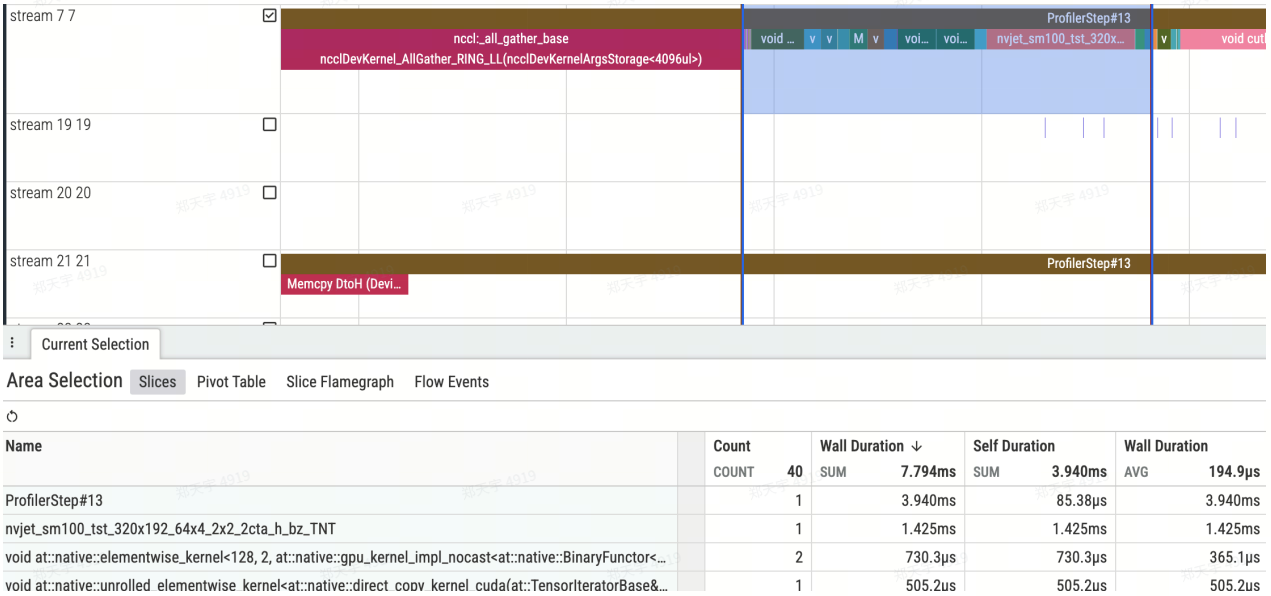


- 对比MTP层的forward
  - mtp_data_process -> MTPNormHead (pp-1和pp-3共有的前半部分):

  pp-3:(4.75ms)



pp-1:(3.94ms)

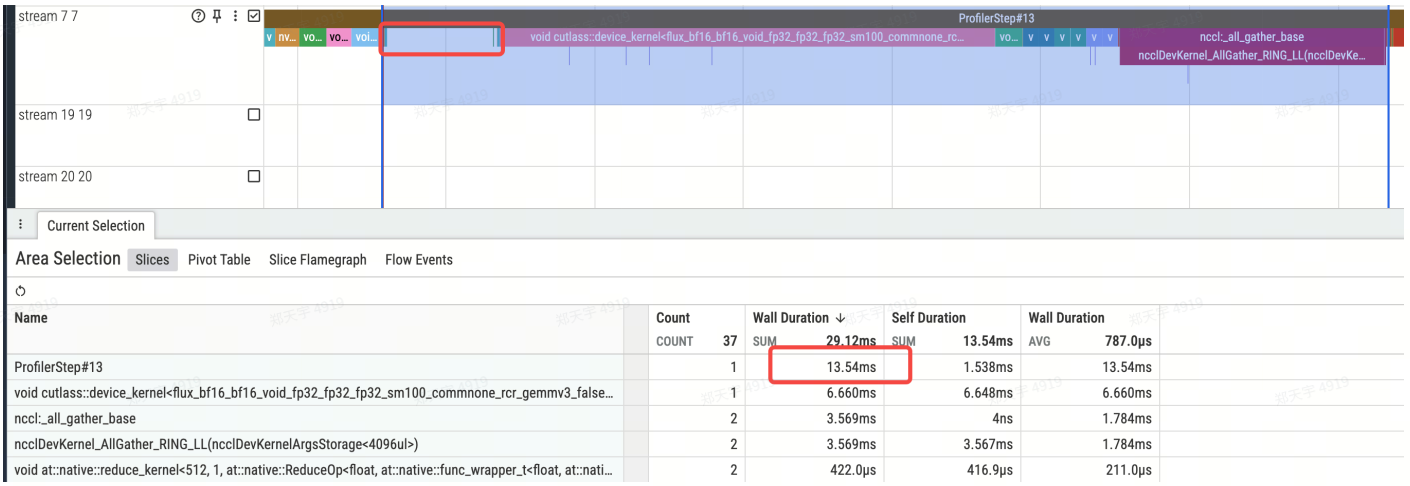


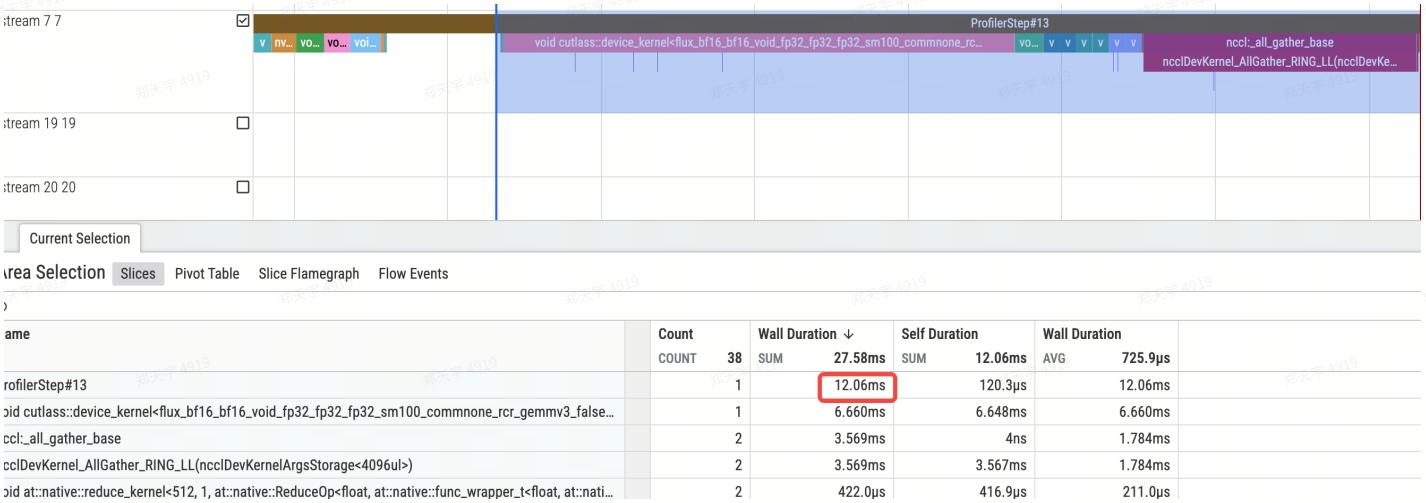
- mtp_token_loss计算(pp-1和pp-3共有的后半部分):

loss这部分已经做了fuse后面再分析还慢不慢

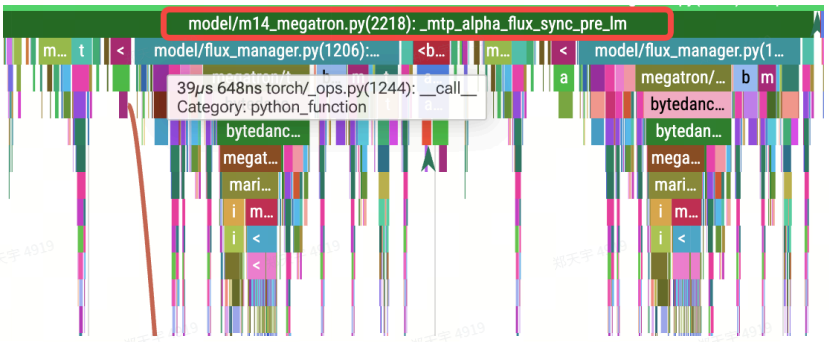
pp-3(13.54ms):

存在1ms的空转，空转的两个kernel之间的看了一下cpu上的操作是flux_write



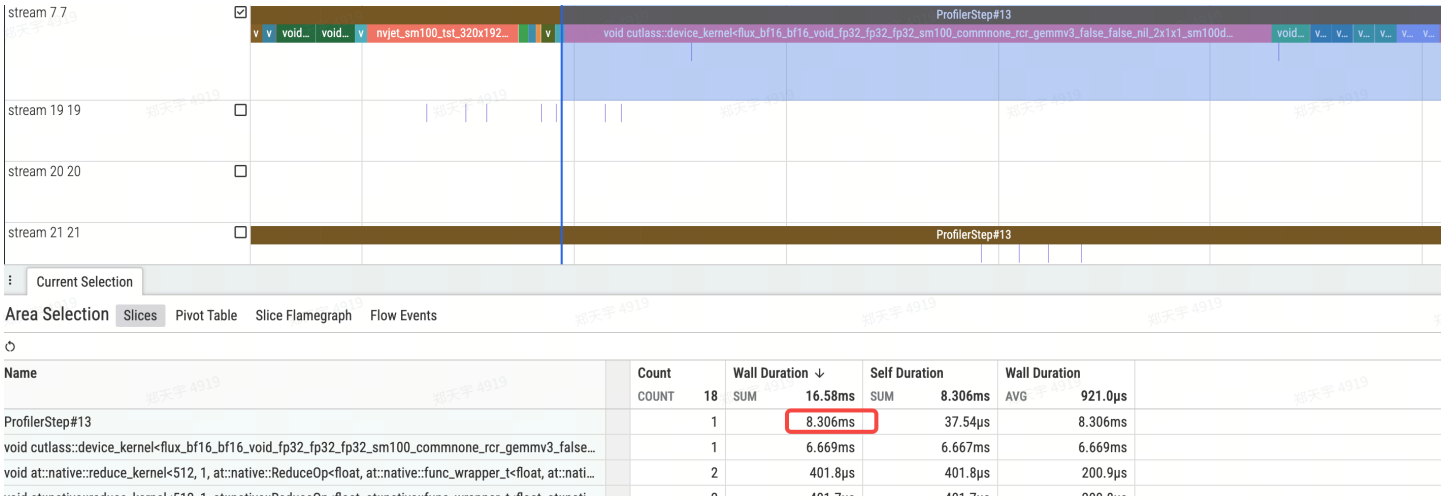


flux_write:

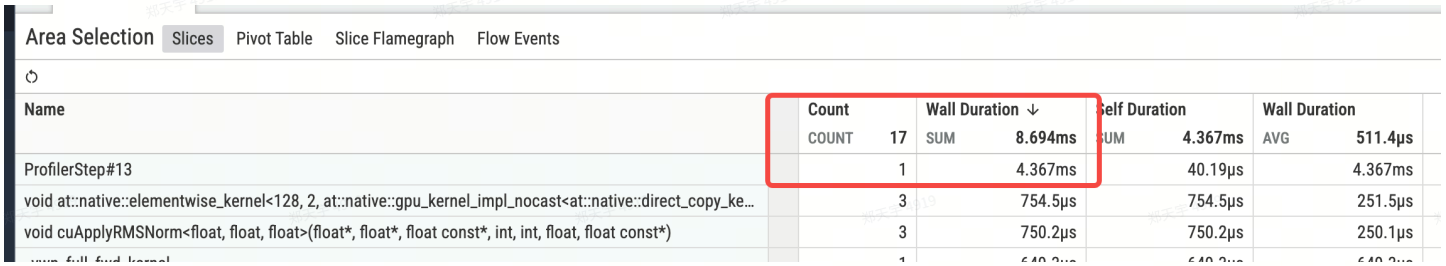


后续看看能不能异步?

pp-1(8.306ms):



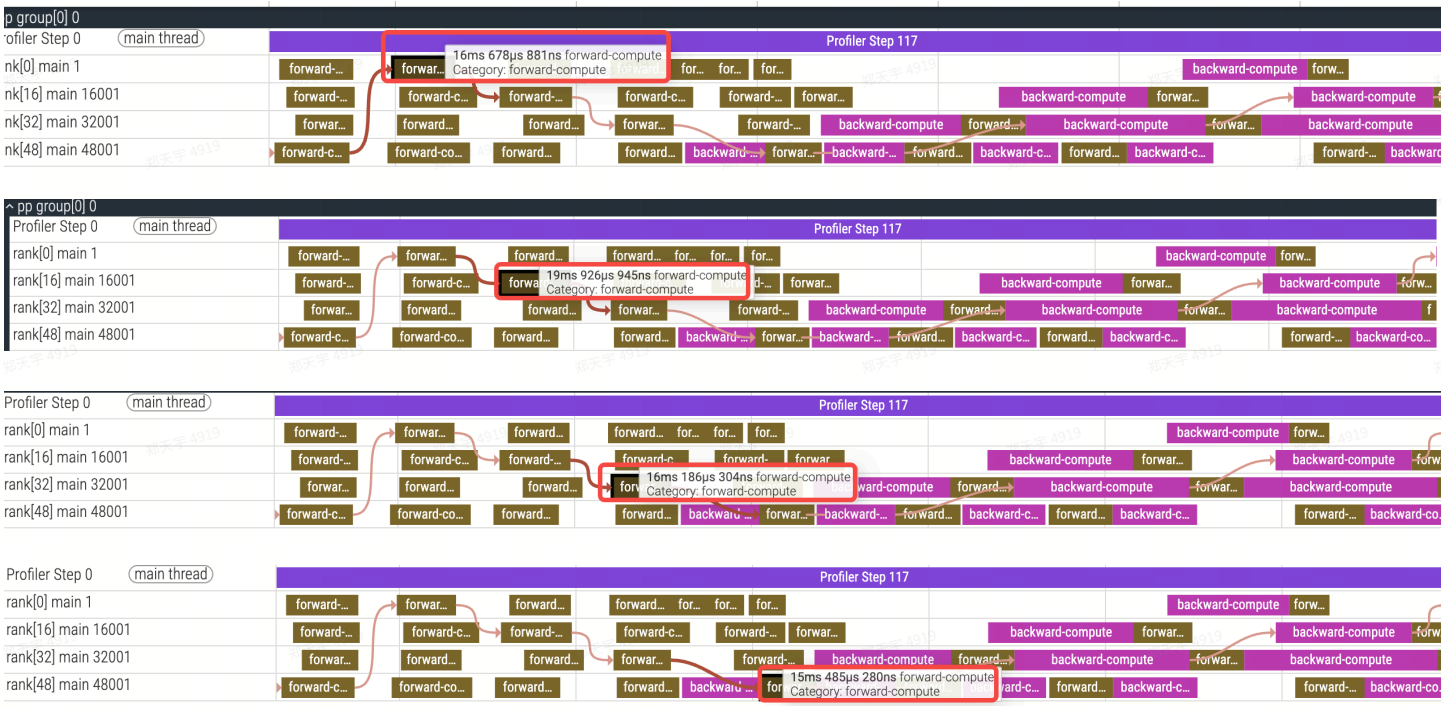
- pp-3独有的embedding(4.367ms)



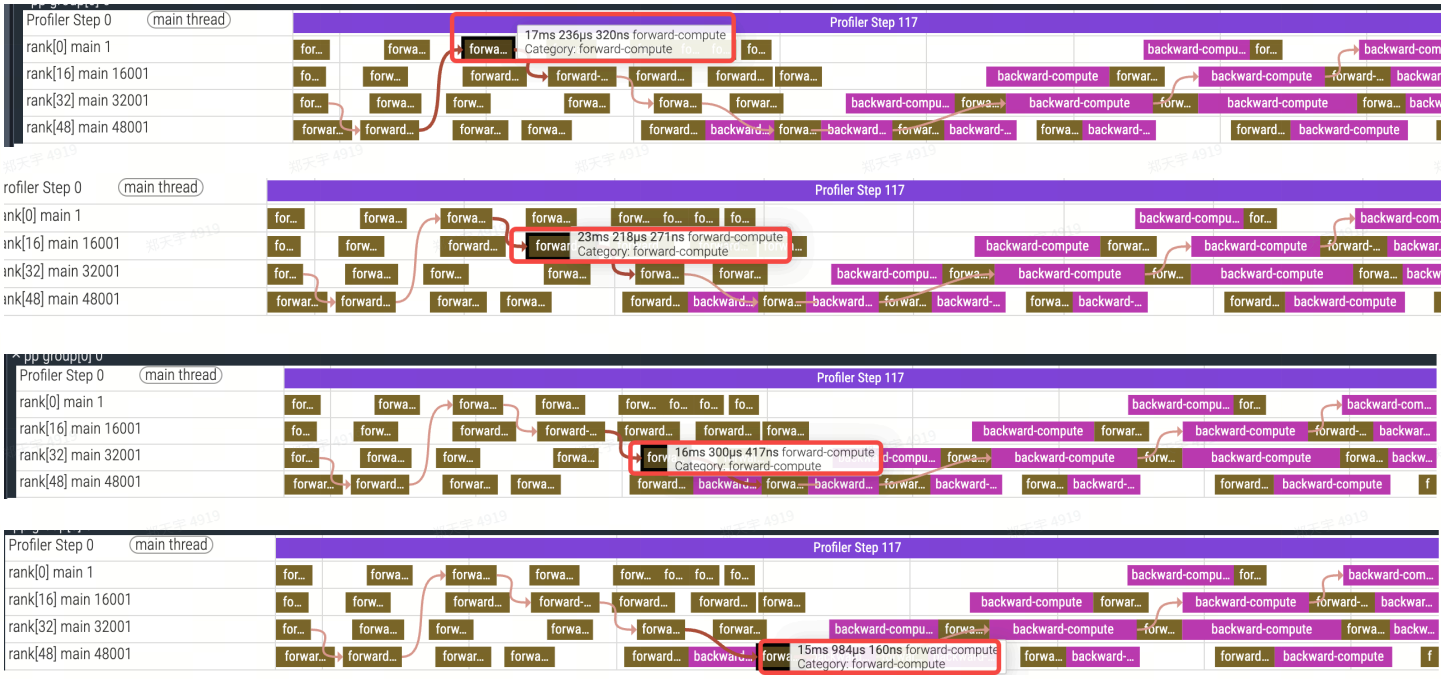
最后补一下loss fuse在pp-3上的效果：

看起来fuse后pp-3的相对时间确实缩短了

fuse：

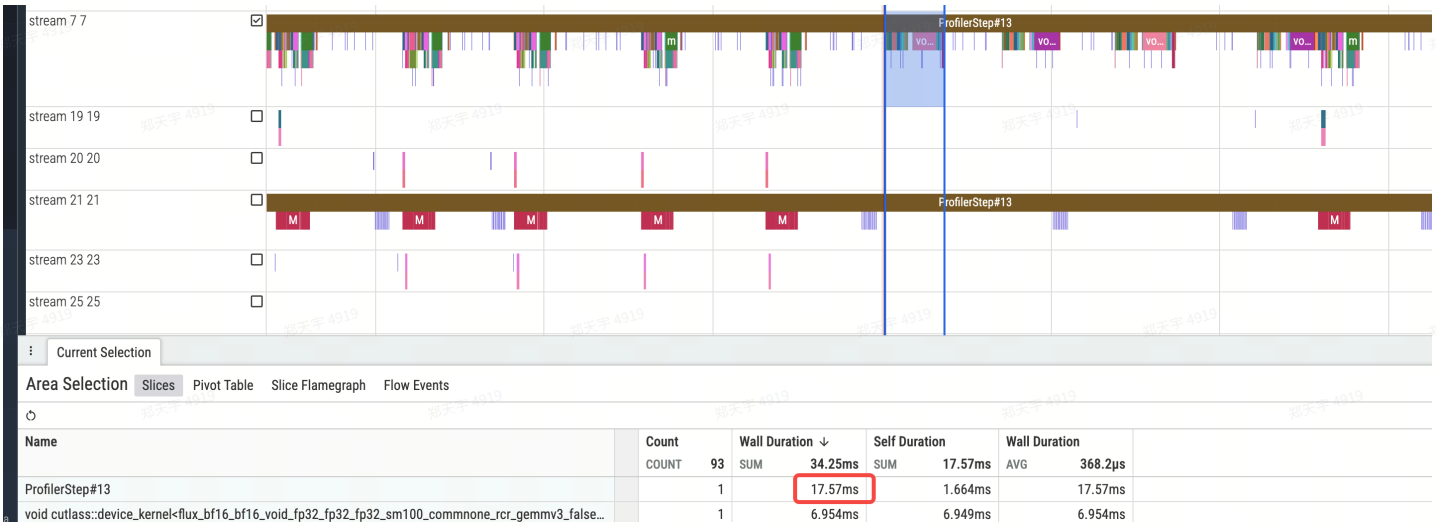


非fuse：

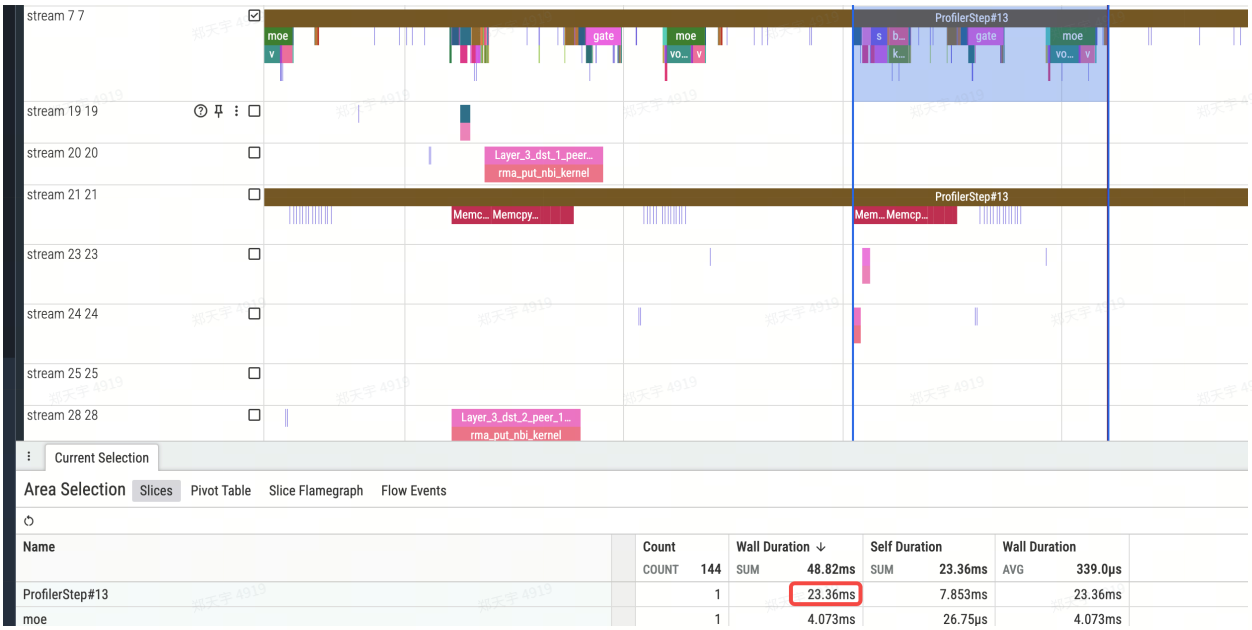


所有fuse，查看profile：

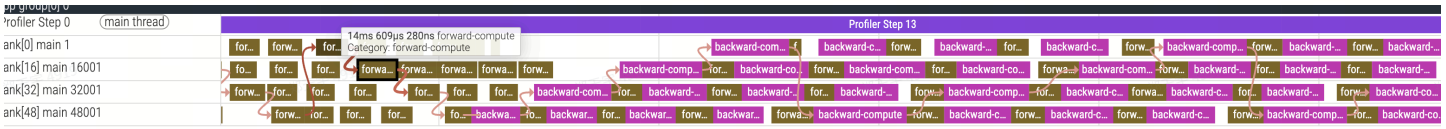
- 同一个step的pp-3的forward：

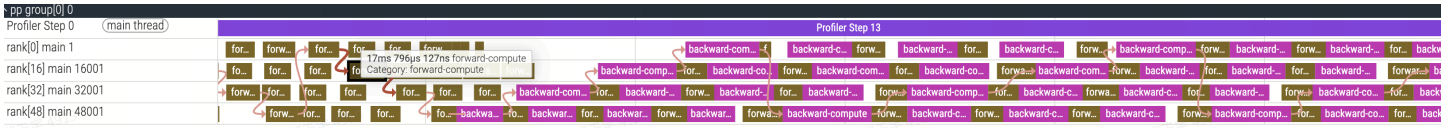


- pp0的forward：



但是ndtimeline的pp上显示pp_0比pp-3快?





以下为个人表格记录，不用看

name	count	wall duration	self duration
`ProfilerStep#13`	1	12.11ms	167.6µs
`void cutlass::device_kernel<flux_bf16_bf16_void_fp32_fp32_fp32_sm100_commnone_rcr_gemmv3_false_false_nil_2x1x1_sm100dense2sm_blockscalemperrow_blockscalenperblock_nil_256x128x64_gemmdefault_0_rasterheuristic>(flux_bf16_bf16_void_fp32_fp32_fp32_sm100_commnone_rcr_gemmv3_false_false_nil_2x1x1_sm100dense2sm_blockscalemperrow_blockscalenperblock_nil_256x128x64_gemmdefault_0_rasterheuristic::Params)`	1	7.151ms	7.147ms

`nvjet_sm100_tst_256x240_64x4_2x2_2cta_h_bz_TNT`	1	1.218ms	1.218ms
`void at::native::elementwise_kernel<128, 2, at::native::gpu_kernel_impl_nocast<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > > (at::TensorIteratorBase&, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > const&)::{lambda(int)#1}>(int, at::native::gpu_kernel_impl_nocast<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > > (at::TensorIteratorBase&, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > const&)::{lambda(int)#1})`	2	694.4µs	694.4µs
`void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::MeanOps<float, float, float, float>, unsigned int, float, 4, 4> > (at::native::ReduceOp<float, at::native::MeanOps<float, float, float, float>, unsigned int, float, 4, 4>)`	3	515.5µs	515.5µs
`nccl:_all_gather_base`	3	445.6µs	6ns
`ncclDevKernel_AllGather_RING_LL(ncclDevKernelArgsStorage<4096ul>)`	3	445.6µs	441.3µs
`void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MinNanFunctor<float> >, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MinNanFunctor<float> >, unsigned int, float, 4, 4>)`	2	403.3µs	403.3µs
`void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MaxNanFunctor<float> >, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float,`	2	402.5µs	402.5µs

at::native::MaxNanFunctor<float> >, unsigned int, float, 4, 4>)			
void cross_entropy_fwd_kernel<8, float>(float*, float*, float const*, long const*, long)	1	352.5µs	352.5µs
Memcpy DtoD (Device -> Device)	8	193.6µs	193.6µs
void at::native::vectorized_elementwise_kernel<4, at::native::(anonymous namespace)::pow_tensor_scalar_kernel_impl<float, float>(at::TensorIteratorBase&, float)::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::(anonymous namespace)::pow_tensor_scalar_kernel_impl<float, float>(at::TensorIteratorBase&, float)::{lambda(float)#1}, std::array<char*, 2ul>)	1	149.2µs	149.2µs
void at::native::vectorized_elementwise_kernel<8, at::native::bfloat16_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::bfloat16_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda(float)#1}, std::array<char*, 2ul>)	2	139.4µs	139.4µs
void at::native::unrolled_elementwise_kernel<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>, 4, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1> >(int, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1>)	1	83.36µs	83.36µs
void cuApplyRMSNorm<float, float, float>(float*, float*, float const*, int, int, float, float	1	70.62µs	70.62µs

const*)			
void at::native::vectorized_elementwise_kernel<4, at::native::BUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul> >(int, at::native::BUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul>)	4	45.30µs	45.30µs
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::sum_functor<float, float, float>::operator()(at::TensorIterator&)::{lambda(float, float)#1}>, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::sum_functor<float, float, float>::operator()(at::TensorIterator&)::{lambda(float, float)#1}>, unsigned int, float, 4, 4>)	4	15.90µs	15.90µs
void at::native::roll_cuda_kernel<long>(long const*, long*, long, long, long, long, long, long)	6	12.06µs	12.06µs
void at::native::index_elementwise_kernel<128, 4, at::native::index_fill_kernel_impl<at::native::OpaqueType<8> >(at::TensorIterator&, long, long, long, at::native::OpaqueType<8>)::{lambda(int)#1}>(long, at::native::index_fill_kernel_impl<at::native::OpaqueType<8> >(at::TensorIterator&, long, long, long, at::native::OpaqueType<8>)::{lambda(int)#1})	6	10.62µs	10.62µs
Stream Wait Event	4	8.256µs	8.256µs
void at::native::vectorized_elementwise_kernel<4, at::native::FillFunctor<float>, std::array<char*, 1ul> >(int, at::native::FillFunctor<float>, std::array<char*, 1ul>)	3	6.656µs	6.656µs
void at::native::unrolled_elementwise_kernel<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#4}::operator()() const::{lambda(long)#1}, std::array<char*, 2ul>, 4,	3	5.472µs	5.472µs

TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1> >(int, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#4}::operator()() const::{lambda(long)#1}, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1>)			
void at::native::unrolled_elementwise_kernel<at::native::CUDAFunctorOnSelf_add<int>, std::array<char*, 2ul>, 4, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithoutCast, at::native::memory::StoreWithoutCast>(int, at::native::CUDAFunctorOnSelf_add<int>, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithoutCast, at::native::memory::StoreWithoutCast)	3	4.800µs	4.800µs
void at::native::vectorized_elementwise_kernel<2, at::native::BinaryFunctor<long, long, long, at::native::binary_internal::MulFunctor<long> >, std::array<char*, 3ul> >(int, at::native::BinaryFunctor<long, long, long, at::native::binary_internal::MulFunctor<long> >, std::array<char*, 3ul>)	2	4.064µs	4.064µs
void at::native::unrolled_elementwise_kernel<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 3ul>, 4, TrivialOffsetCalculator<2, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<2>, at::native::memory::StoreWithCast<1> >(int, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 3ul>, TrivialOffsetCalculator<2,	1	3.808µs	3.808µs

unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<2>, at::native::memory::StoreWithCast<1>)			
void at::native::vectorized_elementwise_kernel<4, at::native::CUDAFunctorOnSelf_add<float>, std::array<char*, 2ul> >(int, at::native::CUDAFunctorOnSelf_add<float>, std::array<char*, 2ul>)	2	3.776µs	3.776µs
void at::native::(anonymous namespace)::CatArrayBatchedCopy_alignedK_contig<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 1, 128, 1, 16>(at::native::(anonymous namespace)::OpaqueType<4u>*, at::native::(anonymous namespace)::CatArrInputTensorMetadata<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 128, 1>, at::native::(anonymous namespace)::TensorSizeStride<unsigned int, 4u>, int, unsigned int)	1	2.208µs	2.208µs
void at::native::vectorized_elementwise_kernel<4, at::native::(anonymous namespace)::launch_clamp_scalar(at::TensorIteratorBase&, c10::Scalar, c10::Scalar, at::native::detail::ClampLimits)::{lambda()#1}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::(anonymous namespace)::launch_clamp_scalar(at::TensorIteratorBase&, c10::Scalar, c10::Scalar, at::native::detail::ClampLimits)::{lambda()#1}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>)	1	2.112µs	2.112µs
void at::native::vectorized_elementwise_kernel<4, at::native::rsqrt_kernel_cuda(at::TensorIteratorBase&)::{lambda()#2}::operator()() const::{lambda()#2}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::rsqrt_kernel_cuda(at::TensorIteratorBase&)::{lambda()#2}::operator()() const::	1	1.792µs	1.792µs

`{lambda()#2}::operator()() const:: {lambda(float)#1}, std::array<char*, 2ul>)`			
`void at::native::vectorized_elementwise_kernel<4, at::native::AUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul> >(int, at::native::AUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul>)`	1	1.504µs	1.504µs

以下是按照您给出的数据逐行整理的表格，包含 `Name`、`Count`、`总 Wall Duration`、`总 Self Duration`、`平均 Wall Duration`。

所有数据顺序与原文本一致，未做任何省略或修改。

Name	Count	总 Wall Duration	总 Self Duration	平均 Wall Duration
COUNT	115	37.25ms	19.34ms	323.9µs
ProfilerStep#13	1	19.34ms	1.722ms	19.34ms
void cutlass::device_kernel<flux_bf16_bf16_void_fp32_fp32_fp32_sm100_commnone_rcr_gemmv3_false_false_nil_2x1x1_sm100dense2sm_blockscalemperrow_blockscalenperblock_nil_256x128x64_gemmdefault_0_rasterheuristic> (flux_bf16_bf16_void_fp32_fp32_fp32_sm100_commnone_rcr_gemmv3_false_false_nil_2x1x1_sm100dense2sm_blockscalemperrow_blockscalenperblock_nil_256x128x64_gemmdefault_0_rasterheuristic::Params)	1	7.192ms	7.181ms	7.192ms
nvjet_sm100_tst_256x240_64x4_2x2_2cta_h_bz_TNT	1	1.388ms	1.386ms	1.388ms
void cuApplyRMSNorm<float, float, float>(float*, float*, float const*, int, int, float, float const*)	4	790.9µs	790.9µs	197.7µs
void at::native::unrolled_elementwise_kernel<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&):: {lambda()#3}::operator()() const:: {lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>, 4, TrivialOffsetCalculator<1,	4	781.8µs	781.8µs	195.4µs

unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1> >(int, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1>)				
void at::native::elementwise_kernel<128, 2, at::native::gpu_kernel_impl_nocast<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > > (at::TensorIteratorBase&, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > const&)::{lambda(int)#1}>(int, at::native::gpu_kernel_impl_nocast<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > > (at::TensorIteratorBase&, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> > const&)::{lambda(int)#1})	3	742.5µs	740.4µs	247.5µs
void at::native::elementwise_kernel<128, 2, at::native::gpu_kernel_impl_nocast<at::native::CUDAFunctor_add<float> >(at::TensorIteratorBase&, at::native::CUDAFunctor_add<float> const&)::{lambda(int)#1}>(int, at::native::gpu_kernel_impl_nocast<at::native::CUDAFunctor_add<float> >(at::TensorIteratorBase&, at::native::CUDAFunctor_add<float> const&)::{lambda(int)#1})	2	594.1µs	594.1µs	297.1µs
void at::native::elementwise_kernel<128, 2, at::native::gpu_kernel_impl_nocast<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}> (at::TensorIteratorBase&, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}	1	547.6µs	547.6µs	547.6µs

const&)::{lambda(int)#1}>(int, at::native::gpu_kernel_impl_nocast<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1}>(at::TensorIteratorBase&, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::{lambda()#7}::operator()() const::{lambda(float)#1} const&)::{lambda(int)#1})				
cutlass3x_sm100_simt_sgemm_f32_f32_f32_f32_f32_32x64x16_1x1x1_3_tnn_align1_bias_f32_relu	1	532.0µs	532.0µs	532.0µs
void cutlass::Kernel2<cutlass_80_simt_sgemm_64x64_8x5_nt_align1>(cutlass_80_simt_sgemm_64x64_8x5_nt_align1::Params)	1	527.4µs	527.4µs	527.4µs
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::MeanOps<float, float, float, float>, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::MeanOps<float, float, float, float>, unsigned int, float, 4, 4>)	3	517.0µs	517.0µs	172.3µs
void at::native::(anonymous namespace)::CatArrayBatchedCopy_alignedK_contig<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 4, 128, 1, 16>(at::native::(anonymous namespace)::OpaqueType<4u>*, at::native::(anonymous namespace)::CatArrInputTensorMetadata<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 128, 1>, at::native::(anonymous namespace)::TensorSizeStride<unsigned int, 4u>, int, unsigned int)	1	501.0µs	501.0µs	501.0µs
Memcpy DtoD (Device -> Device)	10	430.9µs	424.9µs	43.09µs
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MinNanFunctor<float> >, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float,	2	401.4µs	401.4µs	200.7µs

at::native::MinNanFunctor<float> >, unsigned int, float, 4, 4>)				
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MaxNanFunctor<float> >, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::MaxNanFunctor<float> >, unsigned int, float, 4, 4>)	2	400.2µs	400.2µs	200.1µs
void at::native::vectorized_elementwise_kernel<8, at::native::bfloat16_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::bfloat16_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda(float)#1}, std::array<char*, 2ul>)	4	387.6µs	387.6µs	96.89µs
void cutlass::Kernel2<cutlass_80_simt_sgemm_128x32_8x5_nn_align1> (cutlass_80_simt_sgemm_128x32_8x5_nn_align1::Params)	1	365.0µs	365.0µs	365.0µs
void cross_entropy_fwd_kernel<8, float>(float*, float*, float const*, long const*, long)	1	354.2µs	354.2µs	354.2µs
nvjet_sm100_tst_128x256_64x6_2x1_2cta_v_bz_TNT	1	330.5µs	330.5µs	330.5µs
nccl:_all_gather_base	3	267.2µs	6ns	89.06µs
ncclDevKernel_AllGather_RING_LL(ncclDevKernelArgsStorage<4096ul>)	3	267.2µs	267.2µs	89.06µs
void at::native::vectorized_elementwise_kernel<4, at::native::AUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul> >(int, at::native::AUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul>)	4	258.0µs	258.0µs	64.50µs
void at::native::vectorized_elementwise_kernel<4, at::native::(anonymous namespace)::pow_tensor_scalar_kernel_impl<float, float>(at::TensorIteratorBase&, float):: {lambda(float)#1}, std::array<char*, 2ul> >(int,	1	148.7µs	148.7µs	148.7µs

at::native::(anonymous namespace)::pow_tensor_scalar_kernel_impl<float, float>(at::TensorIteratorBase&, float):: {lambda(float)#1}, std::array<char*, 2ul>)				
void at::native::vectorized_elementwise_kernel<4, at::native::BUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul> >(int, at::native::BUnaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 2ul>)	5	40.90µs	40.90µs	8.179µs
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::sum_functor<float, float, float>::operator() (at::TensorIterator&)::{lambda(float, float)#1}>, unsigned int, float, 4, 4> >(at::native::ReduceOp<float, at::native::func_wrapper_t<float, at::native::sum_functor<float, float, float>::operator() (at::TensorIterator&)::{lambda(float, float)#1}>, unsigned int, float, 4, 4>)	6	24.93µs	24.93µs	4.155µs
Stream Wait Event	9	21.02µs	21.02µs	2.336µs
void at::native::vectorized_elementwise_kernel<4, at::native::(anonymous namespace)::launch_clamp_scalar(at::TensorIteratorBase&, c10::Scalar, c10::Scalar, at::native::detail::ClampLimits):: {lambda()#1}::operator()() const:: {lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::(anonymous namespace)::launch_clamp_scalar(at::TensorIteratorBase&, c10::Scalar, c10::Scalar, at::native::detail::ClampLimits):: {lambda()#1}::operator()() const:: {lambda()#7}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>)	2	13.22µs	13.22µs	6.608µs
void at::native::roll_cuda_kernel<long>(long const*, long*, long, long, long, long, long, long)	6	12.13µs	12.13µs	2.021µs
void at::native::vectorized_elementwise_kernel<4, at::native::tanh_kernel_cuda(at::TensorIteratorBase&):: {lambda()#2}::operator()() const::	1	11.49µs	11.49µs	11.49µs

{lambda()#2}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::tanh_kernel_cuda(at::TensorIteratorBase&):: {lambda()#2}::operator()() const:: {lambda()#2}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>)				
void at::native::unrolled_elementwise_kernel<at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 3ul>, 4, TrivialOffsetCalculator<2, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<2>, at::native::memory::StoreWithCast<1> >(int, at::native::BinaryFunctor<float, float, float, at::native::binary_internal::MulFunctor<float> >, std::array<char*, 3ul>, TrivialOffsetCalculator<2, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<2>, at::native::memory::StoreWithCast<1>)	3	11.42µs	11.42µs	3.808µs
void at::native::index_elementwise_kernel<128, 4, at::native::index_fill_kernel_impl<at::native::OpaqueType<8> >(at::TensorIterator&, long, long, long, at::native::OpaqueType<8>)::{lambda(int)#1}>(long, at::native::index_fill_kernel_impl<at::native::OpaqueType<8> >(at::TensorIterator&, long, long, long, at::native::OpaqueType<8>)::{lambda(int)#1})	6	9.504µs	9.504µs	1.584µs
void at::native::vectorized_elementwise_kernel<4, at::native::FillFunctor<float>, std::array<char*, 1ul> > (int, at::native::FillFunctor<float>, std::array<char*, 1ul>)	4	5.952µs	5.952µs	1.488µs
void at::native::unrolled_elementwise_kernel<at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&):: {lambda()#3}::operator()() const:: {lambda()#4}::operator()() const::{lambda(long)#1}, std::array<char*, 2ul>, 4, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1> >(int, at::native::direct_copy_kernel_cuda(at::TensorIteratorBase&)::{lambda()#3}::operator()() const::	3	4.832µs	4.832µs	1.611µs

{lambda()#4}::operator()() const::{lambda(long)#1}, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithCast<1>, at::native::memory::StoreWithCast<1>)				
void at::native::(anonymous namespace)::CatArrayBatchedCopy<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 2, 64, 64>(at::native::(anonymous namespace)::OpaqueType<4u>*, at::native::(anonymous namespace)::CatArrInputTensorMetadata<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 64, 64>, at::native::(anonymous namespace)::TensorSizeStride<unsigned int, 4u>, int, unsigned int)	2	4.672µs	4.672µs	2.336µs
void at::native::vectorized_elementwise_kernel<2, at::native::BinaryFunctor<long, long, long, at::native::binary_internal::MulFunctor<long> >, std::array<char*, 3ul> >(int, at::native::BinaryFunctor<long, long, long, at::native::binary_internal::MulFunctor<long> >, std::array<char*, 3ul>)	2	4.288µs	4.288µs	2.144µs
void at::native::vectorized_elementwise_kernel<4, at::native::CUDAFunctorOnSelf_add<float>, std::array<char*, 2ul> >(int, at::native::CUDAFunctorOnSelf_add<float>, std::array<char*, 2ul>)	2	3.872µs	3.872µs	1.936µs
void at::native::unrolled_elementwise_kernel<at::native::CUDAFunctorOnSelf_add<int>, std::array<char*, 2ul>, 4, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithoutCast, at::native::memory::StoreWithoutCast>(int, at::native::CUDAFunctorOnSelf_add<int>, std::array<char*, 2ul>, TrivialOffsetCalculator<1, unsigned int>, TrivialOffsetCalculator<1, unsigned int>, at::native::memory::LoadWithoutCast, at::native::memory::StoreWithoutCast)	3	3.840µs	3.840µs	1.280µs
	1	2.496µs	2.496µs	2.496µs

void at::native::(anonymous namespace)::CatArrayBatchedCopy_alignedK_contig<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 2, 128, 1, 16>(at::native::(anonymous namespace)::OpaqueType<4u>*, at::native::(anonymous namespace)::CatArrInputTensorMetadata<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 128, 1>, at::native::(anonymous namespace)::TensorSizeStride<unsigned int, 4u>, int, unsigned int)				
void at::native::vectorized_elementwise_kernel<4, at::native::CUDAFunctor_add<float>, std::array<char*, 3ul> >(int, at::native::CUDAFunctor_add<float>, std::array<char*, 3ul>)	1	1.824µs	1.824µs	1.824µs
void at::native::vectorized_elementwise_kernel<4, at::native::rsqrt_kernel_cuda(at::TensorIteratorBase&)::{lambda()#2}::operator()() const::{lambda()#2}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul> >(int, at::native::rsqrt_kernel_cuda(at::TensorIteratorBase&)::{lambda()#2}::operator()() const::{lambda()#2}::operator()() const::{lambda(float)#1}, std::array<char*, 2ul>)	1	1.760µs	1.760µs	1.760µs
void at::native::(anonymous namespace)::CatArrayBatchedCopy_alignedK_contig<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 1, 128, 1, 16>(at::native::(anonymous namespace)::OpaqueType<4u>*, at::native::(anonymous namespace)::CatArrInputTensorMetadata<at::native::(anonymous namespace)::OpaqueType<4u>, unsigned int, 128, 1>, at::native::(anonymous namespace)::TensorSizeStride<unsigned int, 4u>, int, unsigned int)	1	1.632µs	1.632µs	1.632µs
void at::native::vectorized_elementwise_kernel<2, at::native::FillFunctor<float>, std::array<char*, 1ul> >(int, at::native::FillFunctor<float>, std::array<char*, 1ul>)	1	1.504µs	1.504µs	1.504µs
void (anonymous namespace)::elementwise_kernel_with_index<int,	1	1.344µs	1.344µs	1.344µs

at::native::arange_cuda_out(c10::Scalar const&, c10::Scalar const&, c10::Scalar const&, at::Tensor&):: {lambda()#1}::operator()() const:: {lambda()#4}::operator()() const::{lambda(long)#1}> (int, at::native::arange_cuda_out(c10::Scalar const&, c10::Scalar const&, c10::Scalar const&, at::Tensor&):: {lambda()#1}::operator()() const:: {lambda()#4}::operator()() const::{lambda(long)#1}, function_traits<at::native::arange_cuda_out(c10::Scalar const&, c10::Scalar const&, c10::Scalar const&, at::Tensor&)::{lambda()#1}::operator()() const:: {lambda()#4}::operator()() const:: {lambda(long)#1}>::result_type*)				