hta_v2_test

June 27, 2025

[2]:

```python
from hta.trace_analysis import TraceAnalysis




# Load the trace
analyzer = TraceAnalysis(trace_dir="/home/tilak/Alinet_optim/
 ↪inference-optimization-blog-post/part-3/test_hta/")

# Get temporal breakdown
temporal_breakdown = analyzer.get_temporal_breakdown()
print("\nTemporal Breakdown:")
print(temporal_breakdown)

# Get kernel breakdown
kernel_breakdown = analyzer.get_gpu_kernel_breakdown()
print("\nKernel Breakdown:")
print(kernel_breakdown)

# Get idle time
idle_time = analyzer.get_idle_time_breakdown()
print("\nIdle Time:")
print(idle_time)

# For inference, you might want to look at specific iterations
# Since we used step_0, step_1, etc. in record_function
```

```
WARNING Task(Task-2) hta:trace_file.py:create_rank_to_trace_dict()- If the trace
file does not have the rank specified in it, then add the following snippet key
to the json files to use HTA; "distributedInfo": {"rank": 0}. If there are
multiple traces files, then each file should have a unique rank value.For now we
will default to rank = 0.
WARNING Task(Task-2) hta:trace_parser.py:parse_trace_dict()- Parsed
/home/tilak/Alinet_optim/inference-optimization-blog-
post/part-3/test_hta/test_profile_better.pt.trace.json time = 0.57 seconds
WARNING Task(Task-2) hta:trace_parser.py:round_down_time_stamps()- Rounding down
ns resolution events due to issue with events overlapping. ts dtype = float64,
dur dtype = float64.Please see https://github.com/pytorch/pytorch/pull/122425
```

```
WARNING Task(Task-2) hta:trace_parser.py:parse_trace_dataframe()- Parsed
/home/tilak/Alinet_optim/inference-optimization-blog-
post/part-3/test_hta/test_profile_better.pt.trace.json
backend=ParserBackend.JSON in 2.10 seconds; current PID:2899783
WARNING Task(Task-2) hta:trace.py:parse_trace_file()- Overall parsing of
/home/tilak/Alinet_optim/inference-optimization-blog-
post/part-3/test_hta/test_profile_better.pt.trace.json in 2.67 seconds; current
PID:2899783
WARNING Task(Task-2) hta:trace.py:parse_multiple_ranks()- leaving
parse_multiple_ranks duration=2.74 seconds
WARNING Task(Task-2) hta:trace.py:parse_traces()- leaving parse_traces
duration=2.74 seconds


Temporal Breakdown:
   rank  idle_time(us)  compute_time(us)  non_compute_time(us)  \
0     0       349521.0         1482212.0                6296.0

   kernel_time(us)  idle_time_pctg  compute_time_pctg  non_compute_time_pctg
0        1838029.0           19.02              80.64                   0.34

/home/tilak/envApril29/lib/python3.12/site-
packages/hta/analyzers/breakdown_analysis.py:517: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior, set
`pd.set_option('future.no_silent_downcasting', True)`

/home/tilak/envApril29/lib/python3.12/site-
packages/hta/analyzers/breakdown_analysis.py:517: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior, set
`pd.set_option('future.no_silent_downcasting', True)`



Kernel Breakdown:
(   kernel_type      sum  percentage
0  COMPUTATION  1482212        99.6
1       MEMORY     6296         0.4,
name  sum (us)  max (us)  \
0                                              others   80359.0   36088.0
1   sm80_xmma_fprop_implicit_gemm_tf32f32_tf32f32_…  433063.0  433063.0
2   void at::native::(anonymous namespace)::CatArr…   54894.0   54894.0
3   void at::native::(anonymous namespace)::upsamp…  142525.0  142525.0
4   void at::native::elementwise_kernel<128, 2, at…   57127.0   57127.0
5   void at::native::vectorized_elementwise_kernel…  132264.0  132264.0
6   void at::native::vectorized_elementwise_kernel…   76275.0   76275.0
7   void cudnn::bn_fw_inf_1C11_kernel_NCHW<float, …  156594.0  156594.0
8   void cudnn::engines_precompiled::nchwToNhwcKer…  124288.0  124288.0
9   void cutlass::Kernel<cutlass_80_tensorop_s1688…   99373.0   99373.0
10  void cutlass_cudnn_infer::Kernel<cutlass_tenso…  125450.0  125450.0
11                                  Memset (Device)    6296.0     647.0

    min (us)        stddev      mean (us)  kernel_type  rank
0       20.0  11670.219383    8928.777778  COMPUTATION     0
1   433063.0      0.000000  433063.000000  COMPUTATION     0
2    54894.0      0.000000   54894.000000  COMPUTATION     0
3   142525.0      0.000000  142525.000000  COMPUTATION     0
4    57127.0      0.000000   57127.000000  COMPUTATION     0
5   132264.0      0.000000  132264.000000  COMPUTATION     0
6    76275.0      0.000000   76275.000000  COMPUTATION     0
7   156594.0      0.000000  156594.000000  COMPUTATION     0
8   124288.0      0.000000  124288.000000  COMPUTATION     0
9    99373.0      0.000000   99373.000000  COMPUTATION     0
10  125450.0      0.000000  125450.000000  COMPUTATION     0
11       1.0    321.516243     314.800000       MEMORY     0  )


Idle Time:
(   rank stream idle_category  idle_time  idle_time_ratio
0     0      7     host_wait   349521.0              1.0
1     0      7         other        0.0              0.0, None)
```

[2]:

```
kernel_breakdown = analyzer.get_gpu_kernel_breakdown()
```

```
/home/tilak/envApril29/lib/python3.12/site-
packages/hta/analyzers/breakdown_analysis.py:517: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior, set
`pd.set_option('future.no_silent_downcasting', True)`

/home/tilak/envApril29/lib/python3.12/site-
packages/hta/analyzers/breakdown_analysis.py:517: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior, set
`pd.set_option('future.no_silent_downcasting', True)`
```

[3]:

```
#kernel_breakdown_df = analyzer.get_gpu_kernel_breakdown()
idle_time = analyzer.get_idle_time_breakdown()

# Print results
print(temporal_breakdown)
print(kernel_breakdown)
```

```
   rank  idle_time(us)  compute_time(us)  non_compute_time(us)  \
0     0       349521.0         1482212.0                6296.0

   kernel_time(us)  idle_time_pctg  compute_time_pctg  non_compute_time_pctg
0        1838029.0           19.02              80.64                   0.34
(   kernel_type      sum  percentage
0  COMPUTATION  1482212        99.6
1       MEMORY     6296         0.4,
name  sum (us)  max (us)  \
0                                              others   80359.0   36088.0
1   sm80_xmma_fprop_implicit_gemm_tf32f32_tf32f32_…  433063.0  433063.0
2   void at::native::(anonymous namespace)::CatArr…   54894.0   54894.0
3   void at::native::(anonymous namespace)::upsamp…  142525.0  142525.0
4   void at::native::elementwise_kernel<128, 2, at…   57127.0   57127.0
5   void at::native::vectorized_elementwise_kernel…  132264.0  132264.0
6   void at::native::vectorized_elementwise_kernel…   76275.0   76275.0
7   void cudnn::bn_fw_inf_1C11_kernel_NCHW<float, …  156594.0  156594.0
8   void cudnn::engines_precompiled::nchwToNhwcKer…  124288.0  124288.0
9   void cutlass::Kernel<cutlass_80_tensorop_s1688…   99373.0   99373.0
10  void cutlass_cudnn_infer::Kernel<cutlass_tenso…  125450.0  125450.0
11                                     Memset (Device)    6296.0     647.0

    min (us)        stddev      mean (us)  kernel_type  rank
0       20.0  11670.219383    8928.777778  COMPUTATION     0
1   433063.0      0.000000  433063.000000  COMPUTATION     0
2    54894.0      0.000000   54894.000000  COMPUTATION     0
3   142525.0      0.000000  142525.000000  COMPUTATION     0
4    57127.0      0.000000   57127.000000  COMPUTATION     0
5   132264.0      0.000000  132264.000000  COMPUTATION     0
6    76275.0      0.000000   76275.000000  COMPUTATION     0
7   156594.0      0.000000  156594.000000  COMPUTATION     0
8   124288.0      0.000000  124288.000000  COMPUTATION     0
9    99373.0      0.000000   99373.000000  COMPUTATION     0
10  125450.0      0.000000  125450.000000  COMPUTATION     0
11       1.0    321.516243     314.800000       MEMORY     0  )
```