

Project 4

Introduction

In this project, we are asked to implement a two layer NN in order to use OpenAcc & Triton to accelerate machine learning algorithms. Specifically, we use MINST to do handwriting recognition. We will implement forward pass and backward propagation in both sequential and parallel format, and then do optimization to it.

Compilation and Execution

To compile the program, please do the following steps:

```
cd project4
mkdir build && cd build
cmake ..
make
```

After compilation, in order to batch process the project in order to get the execution time, you can simply `sbatch` at the project root directory:

```
cd /path/to/project4
sbatch ./test.sh
```

The result is stored in `Project4-Results.txt`. You can use `vim` to open it.

In order to get profiling result, you can simply run the following commands using `nsys`:

```
#!/bin/bash
mkdir -p ./perf_results

# Path to the dataset
TRAIN_X=./MINST/train-images-idx3-ubyte
TRAIN_Y=./MINST/train-labels-idx1-ubyte
TEST_X=./MINST/t10k-images-idx3-ubyte
TEST_Y=./MINST/t10k-labels-idx1-ubyte

# Hyperparameters
HIDDEN_DIM=400
EPOCHS=10
LEARNING_RATE=0.001
BATCH=32

# Sequential
srun -n 1 --cpus-per-task 1 perf record -e cpu-cycles,cache-misses,page-faults -g -o ./perf_results/sequential.data ./build/sequential $TRAIN_X $TRAIN_Y $TEST_X $TEST_Y $HIDDEN_DIM $EPOCHS $LEARNING_RATE $BATCH

# Kernel
srun -n 1 --gpus 1 nsys profile -t cuda,nvtx,osrt,openacc -o ./perf_results/openacc_kernel.qdrep ./build/openacc_kernel $TRAIN_X $TRAIN_Y $TEST_X $TEST_Y $HIDDEN_DIM $EPOCHS $LEARNING_RATE $BATCH
```

```
# Fusion
srun -n 1 --gpus 1 nsys profile -t cuda,nvtx,osrt,openacc -o
./perf_results/openacc_fusion.qdrep ./build/openacc_fusion $TRAIN_X $TRAIN_Y
$TEST_X $TEST_Y $HIDDEN_DIM $EPOCHS $LEARNING_RATE $BATCH
```

For sequential version, we can use `perf report` to view the results under the `perf_results` folder. For OpenACC, we can use `nsys stats` to visualize the results as well.

Parallel Principle and Optimization

1. For sequential version, we implement the forward pass and back propagation using nested loops and chain rule. No speed up here.
2. For kernel version, we apply OpenACC pragma to each of the kernel function separately. For loops, we use `#pragma acc parallel loop` and `#pragma acc parallel loop collapse(2)` if two nested loop unfold is needed. Since each function only need to access its part of data, so a `copyin` and a `copyout` are needed for each function. Also, functions such as `update_bias` and `update_weights` need `#pragma acc loop reduction(+ : grad_sum)` to accelerate accumulation as well.
3. For fusion version, we aim to transfer all necessary data to the GPU at the start of the MLP training process and using a single `#pragma acc`. Therefore, we only copy all the data needed at the beginning of the training block.

Result

The result is shown below:

```
Sequential (Optimized with -O2)
Training two layer neural network 400 hidden units
| Epoch | Acc Rate | Training Time
| 1 | 92.040% | 45837 ms
| 2 | 93.670% | 45834 ms
| 3 | 94.560% | 45834 ms
| 4 | 95.390% | 45848 ms
| 5 | 95.880% | 45839 ms
| 6 | 96.270% | 45839 ms
| 7 | 96.710% | 45836 ms
| 8 | 96.860% | 45841 ms
| 9 | 97.080% | 45839 ms
| 10 | 97.250% | 45842 ms
Execution Time: 503778 milliseconds
```

```
OpenACC kernel
Training two layer neural network 400 hidden units
| Epoch | Acc Rate | Training Time
| 1 | 91.960% | 7706 ms
| 2 | 93.590% | 6461 ms
| 3 | 94.650% | 6459 ms
| 4 | 95.360% | 6478 ms
| 5 | 95.910% | 6481 ms
| 6 | 96.280% | 6511 ms
| 7 | 96.650% | 6511 ms
| 8 | 96.850% | 6550 ms
| 9 | 97.040% | 6562 ms
| 10 | 97.150% | 6566 ms
Execution Time: 68535 milliseconds
```

```

OpenACC fusion
Training two layer neural network 400 hidden units
| Epoch | Acc Rate | Training Time
| 1 | 92.040% | 4822 ms
| 2 | 93.690% | 4883 ms
| 3 | 94.560% | 4785 ms
| 4 | 95.390% | 4821 ms
| 5 | 95.880% | 4841 ms
| 6 | 96.280% | 4836 ms
| 7 | 96.680% | 4872 ms
| 8 | 96.830% | 4881 ms
| 9 | 97.060% | 4873 ms
| 10 | 97.270% | 4858 ms
Execution Time: 50678 milliseconds

```

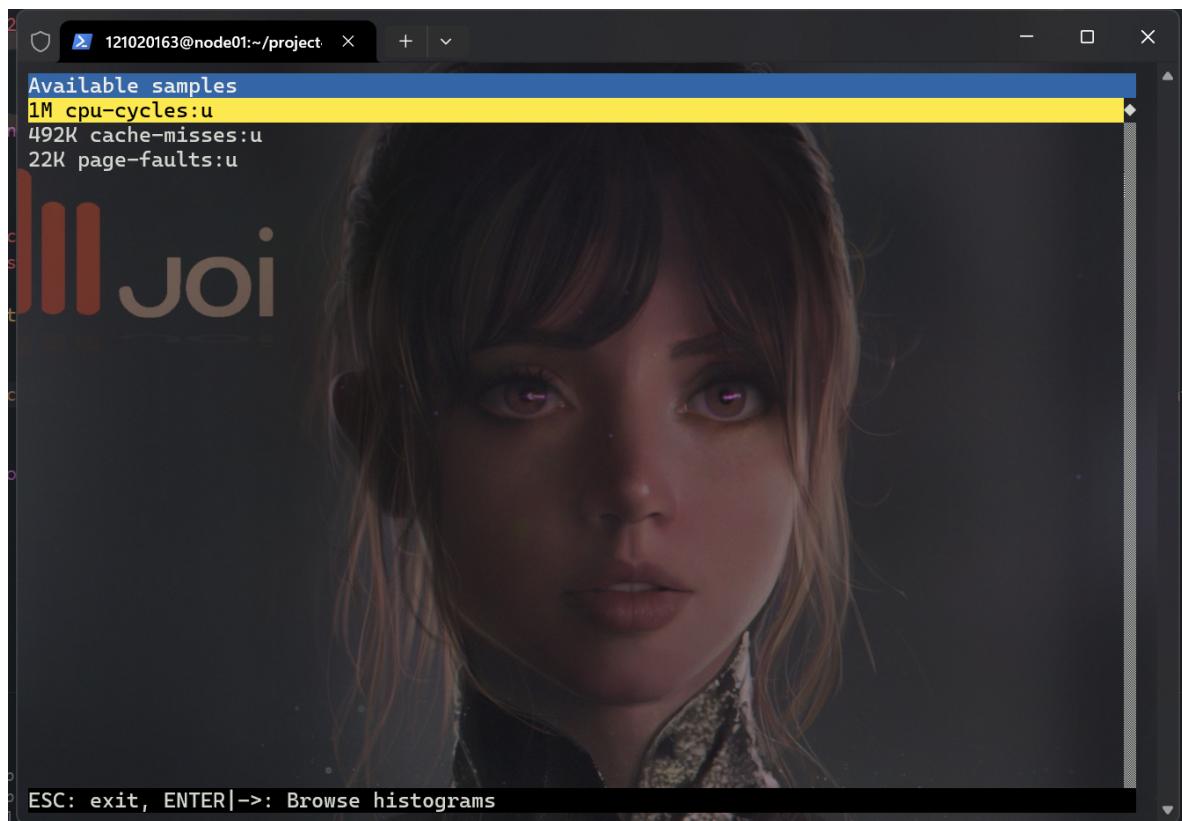
The first epoch speedup factors are shown as followed, using $S(p) = \frac{t_s}{t_p}$ and Sequential performance (45837 ms) as a baseline.

Methods	Speedup
Kernel	5.95
Fusion	9.51

We can see that the OpenACC Fusion method has the fastest speed, since it has already cache the memory inside GPU in one time, avoiding unnecessary overheads.

Profiling Results

Sequential:



Kernel:

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/cudaapisum.py openacc_kernel.sqlite]...

Time(%) StdDev	Total Time (ns) Name	Num Calls	Average	Minimum	Maximum	
-----	-----	-----	-----	-----	-----	-----
88.0	50,469,912,090	834,520	60,477.8	710	3,932,584	
309,590.8	cuStreamSynchronize					
3.9	2,230,766,410	459,450	4,855.3	2,884	730,235	
3,243.7	cuMemcpyHtoDAsync_v2					
3.7	2,118,680,060	278,180	7,616.2	4,620	851,179	
3,421.0	cuLaunchKernel					
2.6	1,519,261,251	278,170	5,461.6	3,363	833,328	
2,937.8	cuMemcpyDtoHAsync_v2					
1.3	731,231,878	282,452	2,588.9	1,862	191,732	
1,131.2	cuEventRecord					
0.4	241,215,466	282,450	854.0	457	123,749	
945.4	cuEventSynchronize					
0.0	26,952,151	1	26,952,151.0	26,952,151	26,952,151	
0.0	cuMemHostAlloc					
0.0	1,220,469	1	1,220,469.0	1,220,469	1,220,469	
0.0	cuMemAllocHost_v2					
0.0	415,124	14	29,651.7	4,270	145,985	
48,886.2	cuMemAlloc_v2					
0.0	387,827	1	387,827.0	387,827	387,827	
0.0	cuModuleLoadDataEx					
0.0	84,346	10	8,434.6	6,386	12,343	
2,016.0	cuMemsetD32Async					
0.0	22,877	1	22,877.0	22,877	22,877	
0.0	cuStreamCreate					
0.0	17,721	4	4,430.3	647	9,821	
4,007.7	cuEventCreate					

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpukernsum.py openacc_kernel.sqlite]...

Time(%)	Total Time (ns)	Instances	Average Name	Minimum	Maximum	StdDev
-----	-----	-----	-----	-----	-----	-----
88.7	39,754,166,488	37,500	1,060,111.1	27,456	3,930,745	
1,033,670.1	update_weight_133_gpu(float*, float const*, float const*, unsigned long, float, unsigned long, unsi...					
8.4	3,777,858,373	43,760	86,331.3	54,208	206,880	
28,356.2	gemm_6_gpu(float const*, float const*, float*, unsigned long, unsigned long, unsigned long)					
1.6	726,849,551	43,760	16,609.9	2,175	57,696	
14,335.1	add_bias_24_gpu(float*, float*, float const*, unsigned long, unsigned long)					
0.5	206,461,576	18,750	11,011.3	10,496	20,448	
460.2	Softmax_49_gpu(float*, float*, unsigned long, unsigned long)					

0.3	139,474,430	37,500	3,719.3	2,527	8,704
1,108.3	update_bias_99_gpu(float*, float const*, unsigned long, float, unsigned long)				
0.1	60,283,057	18,750	3,215.1	3,072	5,664
123.3	input_grad_115_gpu(float const*, float const*, float*, float*, unsigned long, unsigned long, unsigned long)				
0.1	42,395,897	18,750	2,261.1	2,175	4,096
89.7	cross_entropy_loss_grad_87_gpu(float const*, float const*, float*, unsigned long, unsigned long)				
0.1	40,364,083	21,880	1,844.8	1,408	3,329
72.6	Relu_38_gpu(float*, float*, unsigned long)				
0.1	37,462,913	18,750	1,998.0	1,919	3,552
83.8	relu_grad_154_gpu(float const*, float*, unsigned long, unsigned long)				
0.1	35,833,267	18,750	1,911.1	1,823	8,064
93.1	vector_to_one_hot_matrix_68_gpu(unsigned char const*, float*, unsigned long, unsigned long)				
0.0	25,505	10	2,550.5	2,496	2,592
30.4	argmax_182_gpu(float const*, unsigned char*, unsigned long, unsigned long)				
0.0	21,856	10	2,185.6	2,144	2,272
37.1	mean_acc_168_gpu(unsigned char const*, unsigned char const*, unsigned long, unsigned long)				
0.0	21,504	10	2,150.4	2,112	2,240
36.3	mean_acc_168_gpu__red(unsigned char const*, unsigned char const*, unsigned long, unsigned long)				

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpumemtimesum.py openacc_kernel.sqlite]...

Time(%)	Total Time (ns)	Operations	Average	Minimum	Maximum	StdDev	Operation
61.9	3,985,089,015	459,450	8,673.6	703	131,743	21,801.6	[CUDA memcpy HtoD]
38.1	2,450,970,939	278,170	8,811.1	831	105,984	23,852.4	[CUDA memcpy DtoH]
0.0	7,777	10	777.7	736	992	76.9	[CUDA memset]

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpumemsizesum.py openacc_kernel.sqlite]...

Total	Operations	Average	Minimum	Maximum	StdDev	Operation
39,595,542.188	459,450	86.180	0.031	1,225.000	256.472	[CUDA memcpy HtoD]

28,573,095.742 278,170 102.718 0.004 1,225.000 302.560 [CUDA memcpy DtoH]

0.039 10 0.004 0.004 0.004 0.000 [CUDA memset]

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/osrtsum.py openacc_kernel.sqlite]...

Time(%)	Total Time (ns)	Num Calls	Average	Minimum	Maximum
StdDev	Name				
-----	-----	-----	-----	-----	-----
49.9	76,954,346,193	780	98,659,418.2	9,075	100,297,321
11,949,652.0	poll				
49.7	76,522,594,428	153	500,147,676.0	500,076,238	500,588,801
51,992.9	pthread_cond_timedwait				
0.3	472,164,285	23	20,528,882.0	1,880	401,889,856
84,315,999.9	read				
0.1	111,231,786	720	154,488.6	1,038	27,337,755
1,419,408.5	ioctl				
0.0	3,821,251	68	56,194.9	4,148	1,675,569
200,199.7	mmap64				
0.0	2,841,419	50	56,828.4	1,434	466,161
62,038.5	pthread_mutex_lock				
0.0	1,959,214	10	195,921.4	105,205	935,069
259,860.4	sem_timedwait				
0.0	1,371,819	28	48,993.5	4,228	281,332
74,711.7	fclose				
0.0	1,120,734	4	280,183.5	171,967	560,934
187,863.9	fopen64				
0.0	650,551	88	7,392.6	3,020	21,474
2,980.5	open64				
0.0	496,566	24	20,690.3	2,395	147,316
33,872.2	mmap				
0.0	429,120	5	85,824.0	75,784	95,851
9,375.8	pthread_create				
0.0	311,543	30	10,384.8	1,730	56,409
9,882.7	fopen				
0.0	123,286	12	10,273.8	6,139	11,600
1,468.5	fflush				
0.0	106,680	2	53,340.0	10,554	96,126
60,508.5	socket				
0.0	98,125	9	10,902.8	1,415	49,468
14,943.7	fgetc				
0.0	50,755	6	8,459.2	5,115	12,762
2,791.9	open				
0.0	46,096	11	4,190.5	2,765	6,440
1,161.0	write				
0.0	42,972	1	42,972.0	42,972	42,972
0.0	fgets				
0.0	41,565	11	3,778.6	1,218	18,679
5,081.5	fwrite				
0.0	23,706	7	3,386.6	2,513	4,609
717.7	munmap				
0.0	13,699	5	2,739.8	2,427	3,120
274.6	mprotect				
0.0	9,732	1	9,732.0	9,732	9,732
0.0	fread				

0.0	8,282	6	1,380.3	1,025	2,131
457.2	fcntl				
0.0	7,284	1	7,284.0	7,284	7,284
0.0	connect				
0.0	6,206	1	6,206.0	6,206	6,206
0.0	pipe2				
0.0	1,987	1	1,987.0	1,987	1,987
0.0	bind				
0.0	1,275	1	1,275.0	1,275	1,275
0.0	listen				

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/nvtxsum.py openacc_kernel.sqlite]... SKIPPED:
openacc_kernel.sqlite does not contain NV Tools Extension (NVTX) data

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/openmpevtsum.py openacc_kernel.sqlite]... SKIPPED:
openacc_kernel.sqlite does not contain OpenMP event data.

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/vulkanmarkerssum.py openacc_kernel.sqlite]... SKIPPED:
openacc_kernel.sqlite does not contain Vulkan Debug Extension (Vulkan Debug util) data

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/pixsum.py openacc_kernel.sqlite]... SKIPPED:
openacc_kernel.sqlite does not contain DX11/DX12 CPU debug markers

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/khrdebugsum.py openacc_kernel.sqlite]... SKIPPED:
openacc_kernel.sqlite does not contain KHR Extension (KHR_DEBUG) data

Fusion:

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/cudaapisum.py openacc_fusion.sqlite]...

Time(%)	Total Time (ns)	Num Calls	Average	Minimum	Maximum
StdDev	Name				
96.0	47,290,488,091	306,411	154,336.8	734	3,929,891
521,457.7	cuStreamSynchronize				
3.5	1,740,531,561	278,180	6,256.9	4,082	1,012,970
2,636.8	cuLaunchKernel				
0.2	85,315,489	15,770	5,410.0	3,046	1,322,382
10,636.2	cuMemcpyDtoHAsync_v2				
0.1	60,725,024	12,686	4,786.8	3,314	120,110
2,613.1	cuMemcpyHtoDAsync_v2				
0.1	40,979,282	15,950	2,569.2	1,610	168,254
2,223.1	cuEventRecord				
0.1	26,269,826	1	26,269,826.0	26,269,826	26,269,826
0.0	cuMemHostAlloc				
0.0	13,250,564	15,948	830.9	411	90,951
1,178.7	cuEventsSynchronize				
0.0	1,234,759	1	1,234,759.0	1,234,759	1,234,759
0.0	cuMemAllocHost_v2				

0.0	1,052,234	24	43,843.1	2,028	370,261
91,252.0	cuMemAlloc_v2				
0.0	419,185	1	419,185.0	419,185	419,185
0.0	cuModuleLoadDataEx				
0.0	87,633	10	8,763.3	6,854	11,673
1,819.1	cuMemsetD32Async				
0.0	30,654	13	2,358.0	474	12,268
3,588.8	cuEventCreate				
0.0	24,463	1	24,463.0	24,463	24,463
0.0	cuStreamCreate				

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpukernsum.py openacc_fusion.sqlite]...

Time(%)	Total Time (ns)	Instances	Average Name	Minimum	Maximum	StdDev
88.5	41,636,003,242	37,500	1,110,293.4	28,960	3,928,452	
1,081,382.9	update_weight_133_gpu(float*, float const*, float const*, unsigned long, float, unsigned long, unsi...					
8.6	4,045,601,678	43,760	92,449.8	59,264	307,712	
31,525.9	gemm_6_gpu(float const*, float const*, float*, unsigned long, unsigned long, unsigned long)					
1.6	769,403,166	43,760	17,582.3	2,335	61,856	
15,115.1	add_bias_24_gpu(float*, float*, float const*, unsigned long, unsigned long)					
0.5	218,941,444	18,750	11,676.9	11,359	20,736	
362.1	Softmax_49_gpu(float*, float*, unsigned long, unsigned long)					
0.3	147,383,205	37,500	3,930.2	2,687	8,672	
1,139.3	update_bias_99_gpu(float*, float const*, unsigned long, float, unsigned long)					
0.1	63,802,626	18,750	3,402.8	3,263	5,856	
103.1	input_grad_115_gpu(float const*, float const*, float*, float*, unsigned long, unsigned long, unsign...					
0.1	45,703,993	18,750	2,437.5	2,335	3,968	
75.5	cross_entropy_loss_grad_87_gpu(float const*, float const*, float*, unsigned long, unsigned long)					
0.1	42,375,368	21,880	1,936.7	1,824	3,232	
60.0	Relu_38_gpu(float*, float*, unsigned long)					
0.1	39,508,462	18,750	2,107.1	1,951	3,968	
94.7	vector_to_one_hot_matrix_68_gpu(unsigned char const*, float*, unsigned long, unsigned long)					
0.1	39,484,954	18,750	2,105.9	2,015	3,455	
68.6	relu_grad_154_gpu(float const*, float*, unsigned long, unsigned long)					
0.0	27,263	10	2,726.3	2,688	2,753	
25.3	argmax_182_gpu(float const*, unsigned char*, unsigned long, unsigned long)					


```

0.0          25,857          10          2,585.7    2,560    2,656
39.3 mean_acc_168_gpu__red(unsigned char const*, unsigned char const*, unsigned
long, unsigned long)

```

```

0.0          23,967          10          2,396.7    2,368    2,432
27.9 mean_acc_168_gpu(unsigned char const*, unsigned char const*, unsigned
long, unsigned long)

```

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpumemtimesum.py openacc_fusion.sqlite]...

Time(%)	Total Time (ns)	Operations	Average	Minimum	Maximum	StdDev
81.9	219,697,345	12,686	17,318.1	768	1,390,145	128,199.7
[CUDA memcpy HtoD]						
18.1	48,487,260	15,770	3,074.7	863	4,960	1,752.0
[CUDA memcpy DtoH]						
0.0	8,000	10	800.0	800	800	0.0
[CUDA memset]						

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/gpumemsizesum.py openacc_fusion.sqlite]...

Total	Operations	Average	Minimum	Maximum	StdDev
478,723.945	15,770	30.357	0.004	50.000	23.886
[CUDA memcpy DtoH]					
2,346,710.977	12,686	184.984	0.039	16,384.000	1,530.555
[CUDA memcpy HtoD]					
0.039	10	0.004	0.004	0.004	0.000
[CUDA memset]					

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/osrtsum.py openacc_fusion.sqlite]...

Time(%)	Total Time (ns)	Num Calls	Average	Minimum	Maximum
50.0	54,100,418,130	552	98,008,003.9	9,973	100,334,556
14,159,387.6	poll				
49.9	54,013,996,604	108	500,129,598.2	499,806,132	500,160,134
32,724.5	pthread_cond_timedwait				
0.1	110,193,591	726	151,781.8	1,039	26,580,582
1,377,504.0	ioctl				
0.0	30,928,905	23	1,344,735.0	1,811	26,751,729
5,603,083.7	read				
0.0	3,573,675	68	52,554.0	4,313	1,533,507
183,119.2	mmap64				
0.0	1,960,416	10	196,041.6	105,534	947,731
264,186.5	sem_timedwait				
0.0	1,509,501	4	377,375.3	307,442	546,729
114,171.9	fopen64				

0.0	1,394,078	28	49,788.5	4,307	265,443
69,945.5	fclose				
0.0	673,555	88	7,654.0	3,051	27,531
3,762.6	open64				
0.0	586,709	26	22,565.7	2,278	200,474
43,109.7	mmap				
0.0	400,240	5	80,048.0	65,102	89,015
10,076.2	pthread_create				
0.0	388,427	9	43,158.6	35,897	57,411
8,002.8	pthread_mutex_lock				
0.0	310,072	30	10,335.7	1,337	51,147
9,037.4	fopen				
0.0	191,954	2	95,977.0	9,920	182,034
121,703.0	socket				
0.0	117,460	12	9,788.3	5,166	11,513
1,672.0	fflush				
0.0	89,280	10	8,928.0	1,003	44,601
12,951.4	fgetc				
0.0	57,882	6	9,647.0	5,275	17,440
4,493.8	open				
0.0	48,709	12	4,059.1	1,216	6,851
1,562.3	write				
0.0	37,695	11	3,426.8	1,422	15,554
4,095.8	fwrite				
0.0	34,359	1	34,359.0	34,359	34,359
0.0	fgets				
0.0	31,000	8	3,875.0	2,682	6,544
1,289.3	munmap				
0.0	12,674	5	2,534.8	2,259	2,940
285.2	mprotect				
0.0	10,159	1	10,159.0	10,159	10,159
0.0	fread				
0.0	8,881	6	1,480.2	1,084	2,184
427.7	fcntl				
0.0	7,573	1	7,573.0	7,573	7,573
0.0	connect				
0.0	5,393	1	5,393.0	5,393	5,393
0.0	pipe2				
0.0	3,055	1	3,055.0	3,055	3,055
0.0	bind				

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/nvtxsum.py openacc_fusion.sqlite]... SKIPPED:
openacc_fusion.sqlite does not contain NV Tools Extension (NVTX) data

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/openmpevtsum.py openacc_fusion.sqlite]... SKIPPED:
openacc_fusion.sqlite does not contain OpenMP event data.

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/vulkanmarkerssum.py openacc_fusion.sqlite]... SKIPPED:
openacc_fusion.sqlite does not contain Vulkan Debug Extension (Vulkan Debug Util) data

Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-linux-x64/reports/pixsum.py openacc_fusion.sqlite]... SKIPPED:
openacc_fusion.sqlite does not contain DX11/DX12 CPU debug markers

```
Running [/opt/nvidia/hpc_sdk/Linux_x86_64/21.7/profilers/Nsight_Systems/target-  
linux-x64/reports/khrdebugsum.py openacc_fusion.sqlite]... SKIPPED:  
openacc_fusion.sqlite does not contain KHR Extension (KHR_DEBUG) data
```