

GPU Profiling pytorch programs

24 nov. 2018

https://docs.google.com/presentation/d/1DCe2gjAhfAHqpG47ZiUGlQrCHT_dDwjGpDgXcsjoLyQ/edit?usp=sharing

Occupancy ?

```
> watch -n 1 nvidia-smi
```

Every 1.0s: nvidia-smi

Wed Oct 24 07:00:40 2018

NVIDIA-SMI 396.26				Driver Version: 396.26			
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile	Uncorr. ECC	
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M.	
=====							
0	GeForce GTX 108...	Off	00000000:00:06:0	Off		N/A	
49%	84C	P2	197W / 250W	2379MiB / 11178MiB	97%	Default	

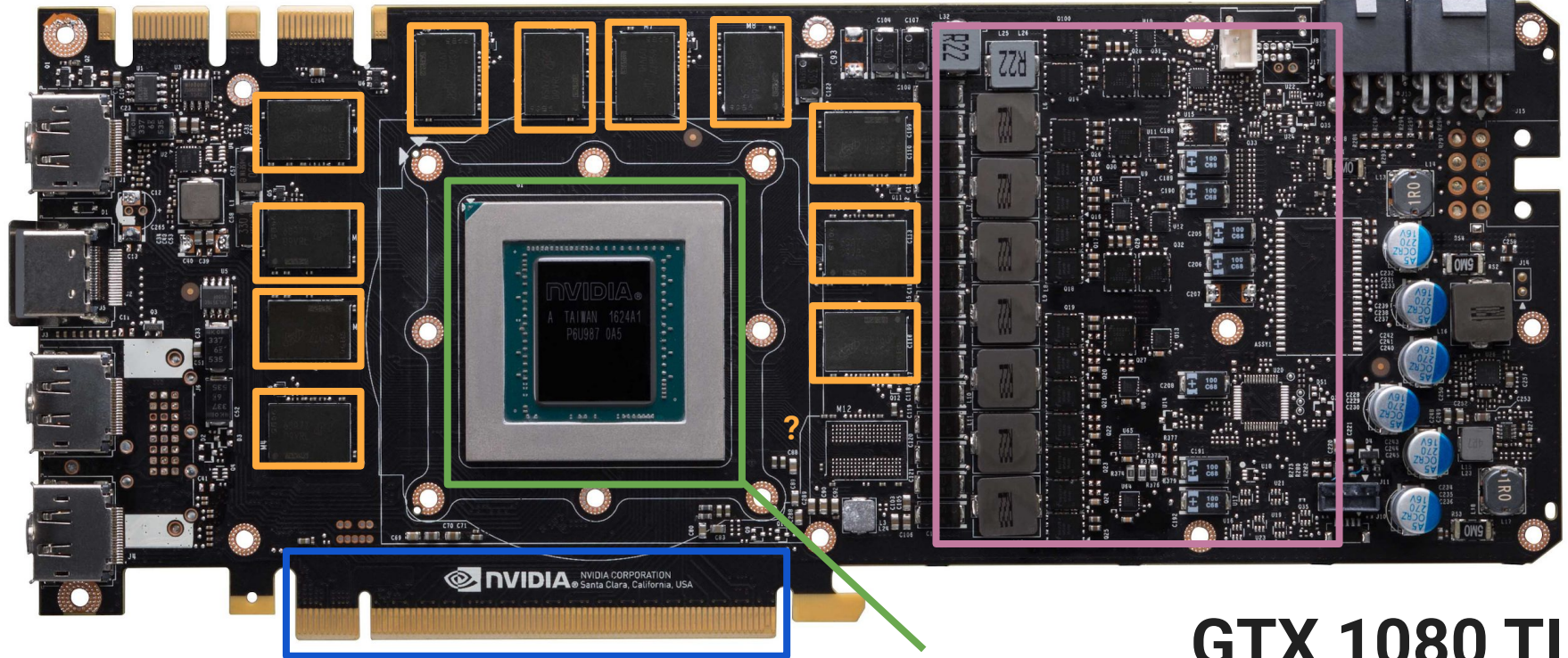
Processes:				GPU Memory
GPU	PID	Type	Process name	Usage
=====				
0	22841	C	python	2369MiB

???

GPU Usage != occupancy ?

GDDR5X memory - 11 GB
480 GB/s bandwidth

VRM -Voltage Regulator Module



Bus PCI-E v3 x16 = 16 x 985 MB/s
= 15.7 GB/s*
GPU ↔ CPU bus async I/O

Multiprocessors
1.2GHz - 2GHz

GTX 1080 TI

source : nvidia.com

Multiprocessors? (SM)

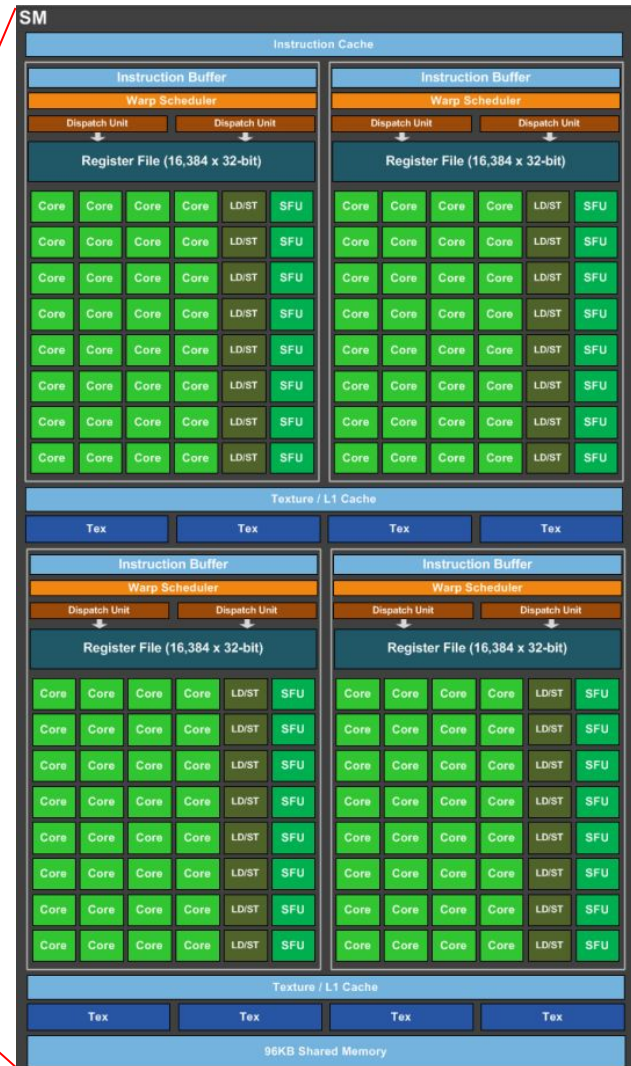
source : nvidia.com



Multiprocessors (SM) ?

28 multiprocessors in 1080TI

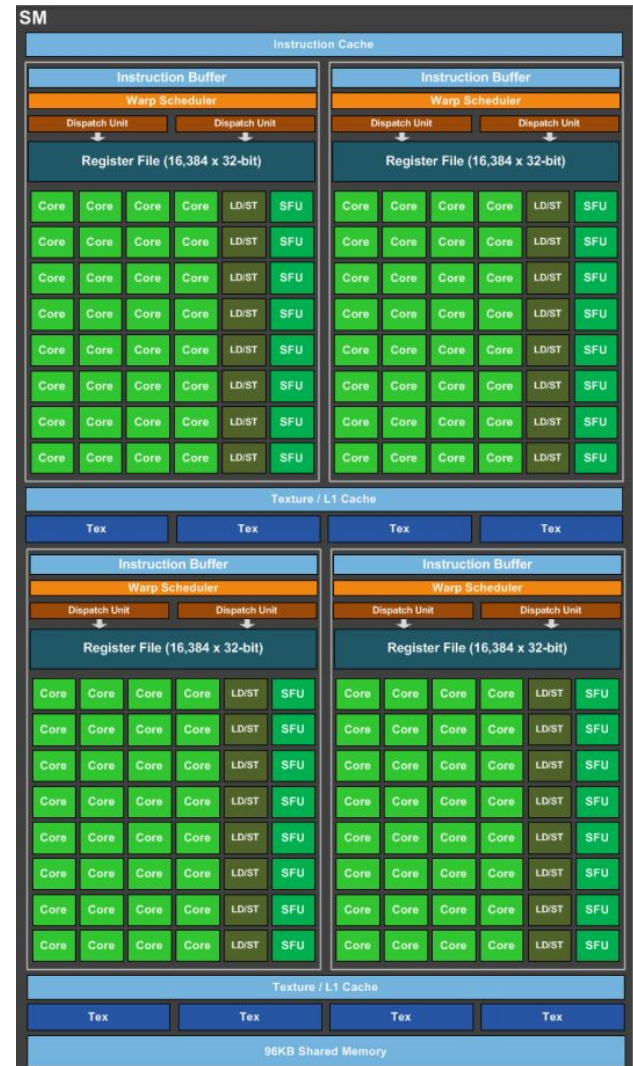
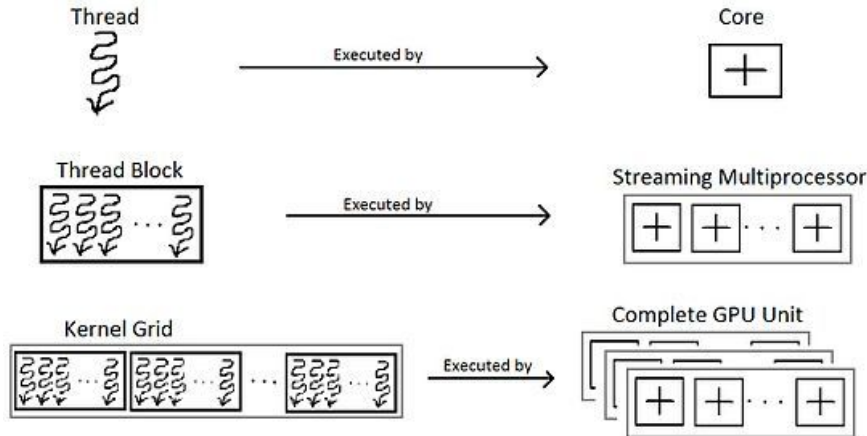
- Warp ?
- Thread ?
- Core ?
- Block ?
- Grid ?
- Kernel ?



Streaming Multiprocessors (SM)

Hardware perspective : SM ? block ? Warp ? core ?
... vs software perspective : Kernel ? Grid ? Thread ?

- SM = 4 warps (thread blocks)
- warp = group 32 cores (== threads)
- Kernel = mini-program / function



nvprof - collect human readable stats

```
> nvprof --log-file profile.log python train.py
```

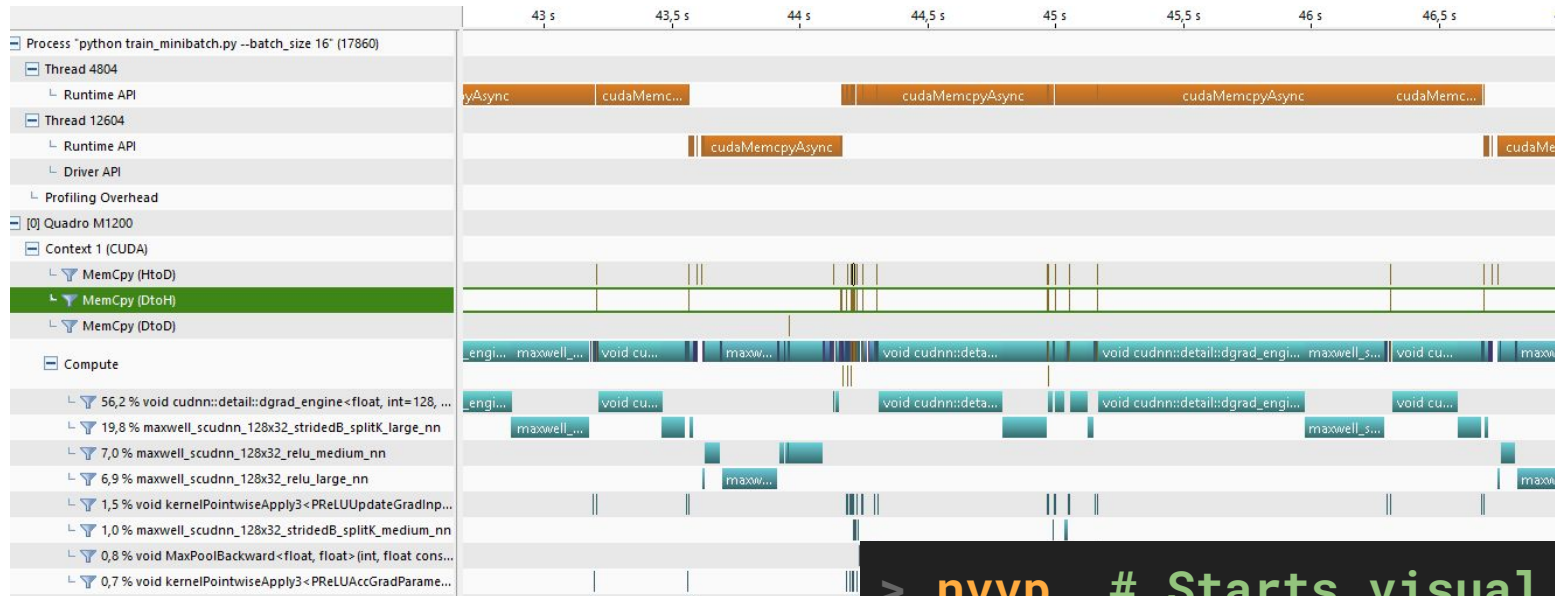
nvprof is usually available in /usr/local/cuda/bin

==20510== Profiling result:

Time(%)	Time	Calls	Avg	Min	Max	Name
48.97%	15.0900s	618	24.417ms	696.98us	80.159ms	void cudnn::detail::dgrad_engine<float, int=128, int=6, i
16.63%	5.12344s	618	8.2904ms	231.62us	27.441ms	maxwell_scudnn_128x32_stridedB_splitK_large_nn
8.88%	2.73659s	835	3.2773ms	191.94us	11.317ms	maxwell_scudnn_128x32_relu_medium_nn
8.59%	2.64821s	334	7.9288ms	338.22us	18.684ms	maxwell_scudnn_128x32_relu_large_nn
3.75%	1.15434s	1648	700.45us	12.001us	1.7575ms	void calc_bias_diff<int=2, float, float, int=128, int=0>(
2.65%	815.20ms	334	2.4407ms	553.81us	4.3517ms	void cudnn::detail::dgrad2d_alg1_1<float, int=0, int=6, i
1.25%	386.56ms	2671	144.73us	2.4640us	628.37us	void add_tensor_kernel_v3<int=2, float, float, int=128, i
1.03%	317.38ms	2838	111.83us	928ns	558.39us	void kernelPointwiseApply2<LeakyReLUUpdateOutput<float>,
0.97%	299.27ms	619	483.48us	90.371us	1.1689ms	maxwell_scudnn_128x32_stridedB_splitK_medium_nn
0.95%	291.26ms	1648	176.73us	1.0880us	834.97us	void kernelPointwiseApply3<LeakyReLUUpdateGradInput<float
0.70%	214.34ms	167	1.2835ms	628.09us	1.5355ms	maxwell_scudnn_128x64_relu_small_nn
0.56%	172.94ms	206	839.51us	69.731us	1.8050ms	maxwell_scudnn_128x32_stridedB_small_nn
0.56%	171.60ms	103	1.6661ms	814.68us	1.9186ms	maxwell_scudnn_128x64_stridedB_splitK_large_nn
0.51%	156.42ms	412	379.66us	115.52us	653.53us	void MaxPoolBackward<float, float>(int, float const *, lo
0.46%	140.24ms	668	209.94us	64.867us	314.57us	void MaxPoolForward<float, float>(int, float const *, int
0.44%	135.19ms	840	160.94us	672ns	3.9193ms	[CUDA memcpy HtoD]
0.40%	123.27ms	206	598.42us	32.833us	1.3077ms	maxwell_scudnn_128x32_stridedB_splitK_interior_nn

nvprof - visual profiler

```
> nvprof -o profile.prof python train.py
```



```
> nvvp # Starts visual profiler
```


nvprof - GPU / system stats

```
> nvprof --system-profiling -o profile.prof python train.py
```

[0] GeForce GTX 1080 Ti		
▼ Attributes		
Compute Capability	6.1	
▼ Maximums		
Threads per Block	1024	
Threads per Multiprocessor	2048	
Shared Memory per Block	48 KiB	
Shared Memory per Multiprocessor	96 KiB	
Registers per Block	65536	
Registers per Multiprocessor	65536	
Grid Dimensions	[2147483647, 65535, 65535]	
Block Dimensions	[1024, 1024, 64]	
Warps per Multiprocessor	64	
Blocks per Multiprocessor	32	
Half Precision FLOP/s	88,592 GigaFLOP/s	
Single Precision FLOP/s	11,34 TeraFLOP/s	
Double Precision FLOP/s	354,368 GigaFLOP/s	
▼ Multiprocessor		
Multiprocessors	28	
Clock Rate	1,582 GHz	
Concurrent Kernel	true	
Max IPC	6	
Threads per Warp	32	

nvprof - cross platform profiling

Important for cross platform analysis: use `--export-profile` instead of `-o` ?

```
> nvprof --export-profile profile.prof python train.py
```

nvprof - visual profiler

The screenshot displays the nvprof application window. The top menu bar includes 'Analysis', 'GPU Details', 'CPU Details', 'Console', and 'Settings'. Below the menu, there are icons for a list, a document, an upward arrow, and a PDF icon, followed by the text 'Export PDF Report'. The left sidebar contains two main sections: '1. CUDA Application Analysis' and '2. Check Overall GPU Usage'. The right pane, titled 'Results', contains two warning messages and one information message.

1. CUDA Application Analysis

2. Check Overall GPU Usage

The analysis results on the right indicate potential problems in how your application is taking advantage of the GPU's available compute and data movement capabilities. You should examine the information provided with each result to determine if you can make changes to your application to increase GPU utilization.

Results

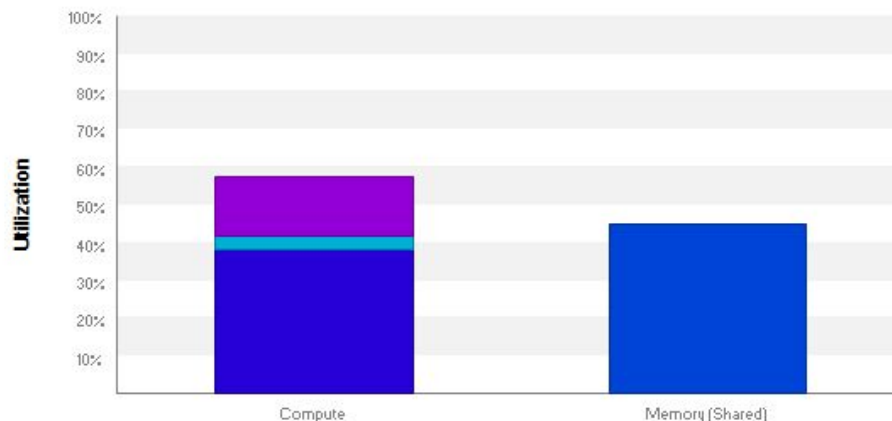
- ⚠ Low Kernel Concurrency** [0 ns / 57,199 ms = 0 %]
The percentage of time when two kernels are being executed in parallel is low.
- ⚠ Low Compute Utilization** [57,199 ms / 693,128 s = 0 %]
The multiprocessors of one or more GPUs are mostly idle.
- i Compute Utilization**
The device timeline shows an estimate of the amount of the total compute capacity being used by the kernels executing on the device.

nvprof - fine-grained kernel analysis

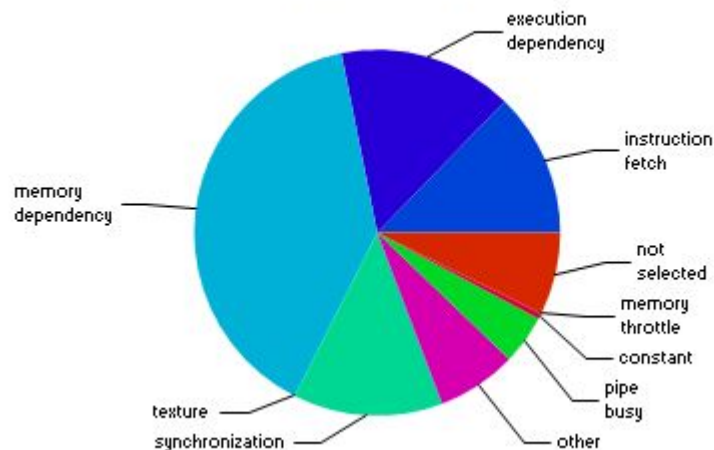
```
> nvprof --analysis-metrics -o profile.prof python train.py
```

i Kernel Performance Is Bound By Instruction And Memory Latency

This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance levels indicate that the performance of the kernel is most likely limited by the latency of arithmetic or memory and/or memory bandwidth below 60 % of peak typically indicates latency issues.

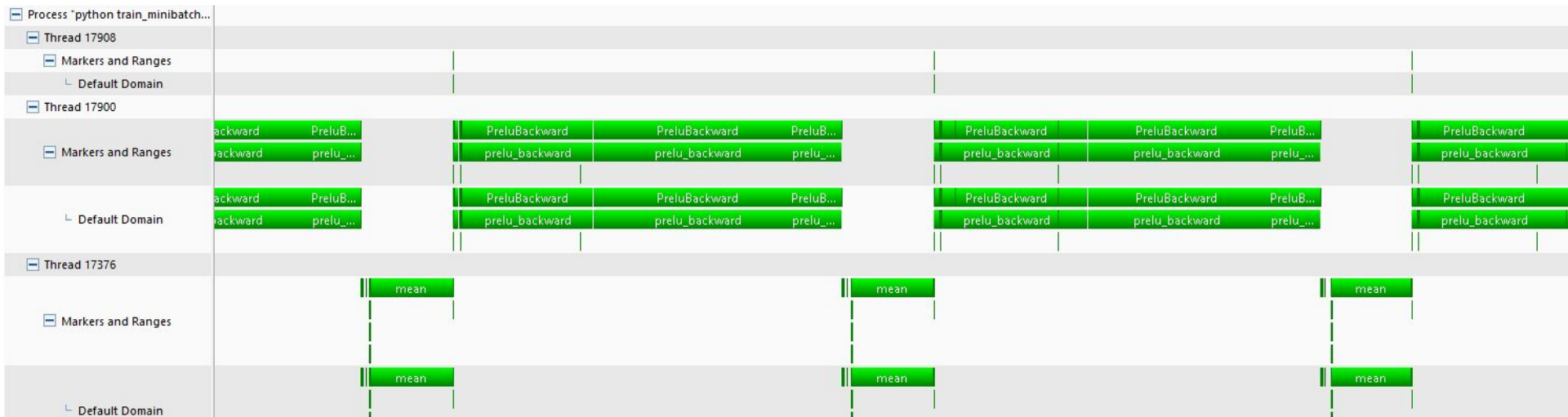


Stall Reasons

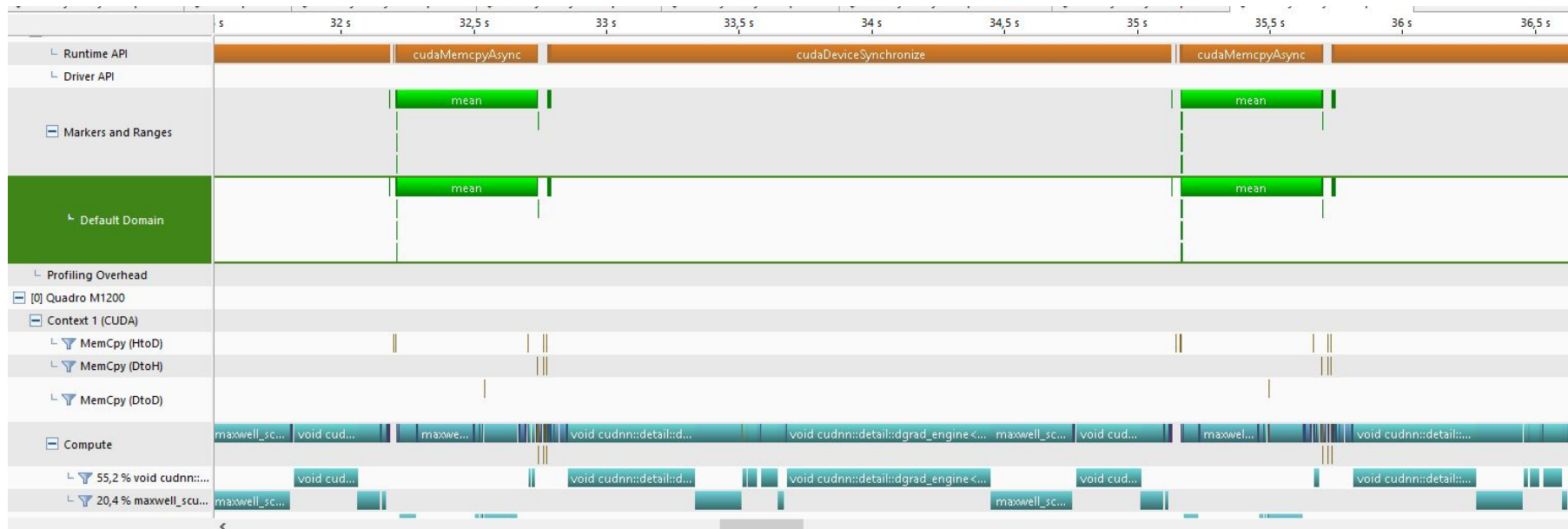


nvprof - Add pytorch annotations

```
with torch.autograd.profiler.emit_nvtx(enabled=True):  
    density_map = net(im_data, gt_data, gt_class_label, class_wts, nb_pixels)  
    loss = net.loss  
    optimizer.zero_grad()  
    loss.backward()  
    optimizer.step()
```



nvprof - Add pytorch annotations



nvprof - Add pytorch annotations



kernprof

```
> pip install line_profiler
> kernprof -l -o train.py.lprof train.py
> python -m line_profiler train.py.lprof
```

Line #	Hits	Time	Per Hit	% Time	Line Contents
--------	------	------	---------	--------	---------------

=====					
-------	--	--	--	--	--

50					@profile
51					def build_loss(self, density_map, density
52					
53	1000	679970.0	680.0	0.8	loss_mse = self.loss_mse_fn(density_m
54					
55					# manipulate loss_mse to get sum of s
56	1000	1220858.0	1220.9	1.4	loss_mse = loss_mse.sum(dim=3).sum(di
57					
58					# divide total squared error by numbe
59	1000	430174.0	430.2	0.5	loss_mse = loss_mse / nb_pixels
60					
61					# return mean of error as loss for ba
62	1000	78014267.0	78014.3	87.2	loss_mse = torch.sum(loss_mse)
63					
64	1000	9120701.0	9120.7	10.2	cross_entropy = self.loss_bce_fn(density_cls_score, gt_cls_label)
65					
66	1000	17252.0	17.3	0.0	return loss_mse, cross_entropy

```
@profile
```

```
def forward(self, im_data, gt_data=None, g
density_map, density_cls_score = self.
density_cls_prob = F.softmax(density_c
if self.training:
    self.loss_mse, self.cross_entropy
    density_map, density_cls_prob,
return density_map
```


The screenshot displays the Windows Performance Analyzer (PPW-SOMA) interface, which is used for analyzing system performance. The interface is divided into several panes:

- Top Pane:** Shows a timeline of events across multiple processors (P0, P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21, P22, P23, P24, P25, P26, P27, P28, P29, P30, P31, P32, P33, P34, P35, P36, P37, P38, P39, P40, P41, P42, P43, P44, P45, P46, P47, P48, P49, P50, P51, P52, P53, P54, P55, P56, P57, P58, P59, P60, P61, P62, P63, P64, P65, P66, P67, P68, P69, P70, P71, P72, P73, P74, P75, P76, P77, P78, P79, P80, P81, P82, P83, P84, P85, P86, P87, P88, P89, P90, P91, P92, P93, P94, P95, P96, P97, P98, P99, P100, P101, P102, P103, P104, P105, P106, P107, P108, P109, P110, P111, P112, P113, P114, P115, P116, P117, P118, P119, P120, P121, P122, P123, P124, P125, P126, P127, P128, P129, P130, P131, P132, P133, P134, P135, P136, P137, P138, P139, P140, P141, P142, P143, P144, P145, P146, P147, P148, P149, P150, P151, P152, P153, P154, P155, P156, P157, P158, P159, P160, P161, P162, P163, P164, P165, P166, P167, P168, P169, P170, P171, P172, P173, P174, P175, P176, P177, P178, P179, P180, P181, P182, P183, P184, P185, P186, P187, P188, P189, P190, P191, P192, P193, P194, P195, P196, P197, P198, P199, P200, P201, P202, P203, P204, P205, P206, P207, P208, P209, P210, P211, P212, P213, P214, P215, P216, P217, P218, P219, P220, P221, P222, P223, P224, P225, P226, P227, P228, P229, P230, P231, P232, P233, P234, P235, P236, P237, P238, P239, P240, P241, P242, P243, P244, P245, P246, P247, P248, P249, P250, P251, P252, P253, P254, P255, P256, P257, P258, P259, P260, P261, P262, P263, P264, P265, P266, P267, P268, P269, P270, P271, P272, P273, P274, P275, P276, P277, P278, P279, P280, P281, P282, P283, P284, P285, P286, P287, P288, P289, P290, P291, P292, P293, P294, P295, P296, P297, P298, P299, P300, P301, P302, P303, P304, P305, P306, P307, P308, P309, P310, P311, P312, P313, P314, P315, P316, P317, P318, P319, P320, P321, P322, P323, P324, P325, P326, P327, P328, P329, P330, P331, P332, P333, P334, P335, P336, P337, P338, P339, P340, P341, P342, P343, P344, P345, P346, P347, P348, P349, P350, P351, P352, P353, P354, P355, P356, P357, P358, P359, P360, P361, P362, P363, P364, P365, P366, P367, P368, P369, P370, P371, P372, P373, P374, P375, P376, P377, P378, P379, P380, P381, P382, P383, P384, P385, P386, P387, P388, P389, P390, P391, P392, P393, P394, P395, P396, P397, P398, P399, P400, P401, P402, P403, P404, P405, P406, P407, P408, P409, P410, P411, P412, P413, P414, P415, P416, P417, P418, P419, P420, P421, P422, P423, P424, P425, P426, P427, P428, P429, P430, P431, P432, P433, P434, P435, P436, P437, P438, P439, P440, P441, P442, P443, P444, P445, P446, P447, P448, P449, P450, P451, P452, P453, P454, P455, P456, P457, P458, P459, P460, P461, P462, P463, P464, P465, P466, P467, P468, P469, P470, P471, P472, P473, P474, P475, P476, P477, P478, P479, P480, P481, P482, P483, P484, P485, P486, P487, P488, P489, P490, P491, P492, P493, P494, P495, P496, P497, P498, P499, P500, P501, P502, P503, P504, P505, P506, P507, P508, P509, P510, P511, P512, P513, P514, P515, P516, P517, P518, P519, P520, P521, P522, P523, P524, P525, P526, P527, P528, P529, P530, P531, P532, P533, P534, P535, P536, P537, P538, P539, P540, P541, P542, P543, P544, P545, P546, P547, P548, P549, P550, P551, P552, P553, P554, P555, P556, P557, P558, P559, P560, P561, P562, P563, P564, P565, P566, P567, P568, P569, P570, P571, P572, P573, P574, P575, P576, P577, P578, P579, P580, P581, P582, P583, P584, P585, P586, P587, P588, P589, P590, P591, P592, P593, P594, P595, P596, P597, P598, P599, P600, P601, P602, P603, P604, P605, P606, P607, P608, P609, P610, P611, P612, P613, P614, P615, P616, P617, P618, P619, P620, P621, P622, P623, P624, P625, P626, P627, P628, P629, P630, P631, P632, P633, P634, P635, P636, P637, P638, P639, P640, P641, P642, P643, P644, P645, P646, P647, P648, P649, P650, P651, P652, P653, P654, P655, P656, P657, P658, P659, P660, P661, P662, P663, P664, P665, P666, P667, P668, P669, P670, P671, P672, P673, P674, P675, P676, P677, P678, P679, P680, P681, P682, P683, P684, P685, P686, P687, P688, P689, P690, P691, P692, P693, P694, P695, P696, P697, P698, P699, P700, P701, P702, P703, P704, P705, P706, P707, P708, P709, P710, P711, P712, P713, P714, P715, P716, P717, P718, P719, P720, P721, P722, P723, P724, P725, P726, P727, P728, P729, P730, P731, P732, P733, P734, P735, P736, P737, P738, P739, P740, P741, P742, P743, P744, P745, P746, P747, P748, P749, P750, P751, P752, P753, P754, P755, P756, P757, P758, P759, P760, P761, P762, P763, P764, P765, P766, P767, P768, P769, P770, P771, P772, P773, P774, P775, P776, P777, P778, P779, P780, P781, P782, P783, P784, P785, P786, P787, P788, P789, P790, P791, P792, P793, P794, P795, P796, P797, P798, P799, P800, P801, P802, P803, P804, P805, P806, P807, P808, P809, P810, P811, P812, P813, P814, P815, P816, P817, P818, P819, P820, P821, P822, P823



Processeur
25% 2,37 GHz



Mémoire
21,0/31,9 Go (66%)



Disque 0 (C:)
1%



Ethernet
Non connecté



Ethernet
E : 0 R : 0 Kbits/s



Ethernet
Non connecté



Wi-Fi
E : 0 R : 0 Kbits/s



Ethernet
E : 0 R : 0 Kbits/s



Processeur graphique
Intel(R) HD Graphics 630
6%



Processeur graphique
NVIDIA Quadro M1200
77%

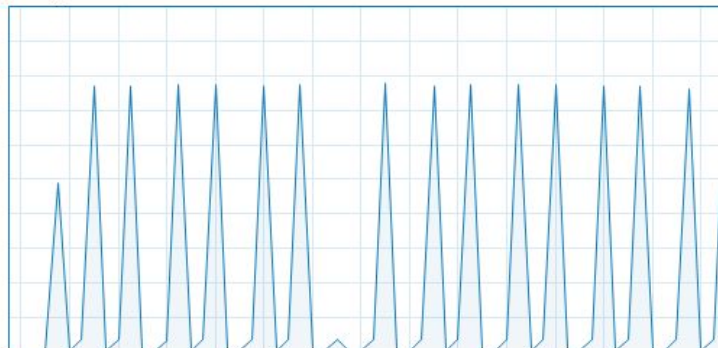
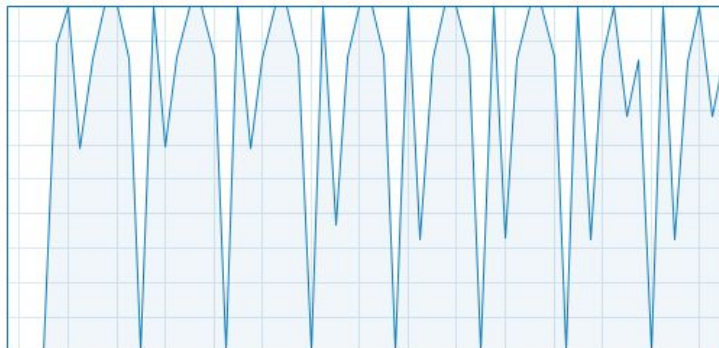
Processeur graphique

NVIDIA Quadro M1200

▼ Compute_0

85% ▼ Copy

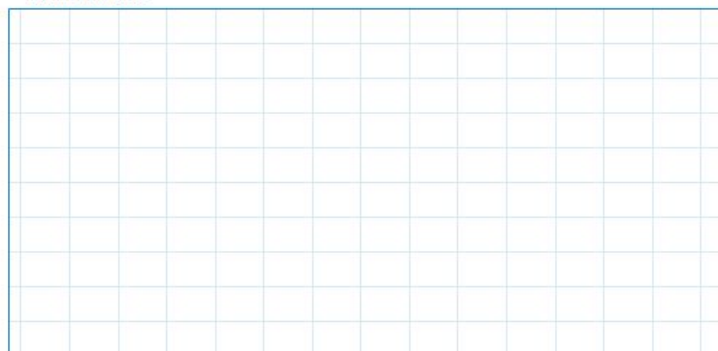
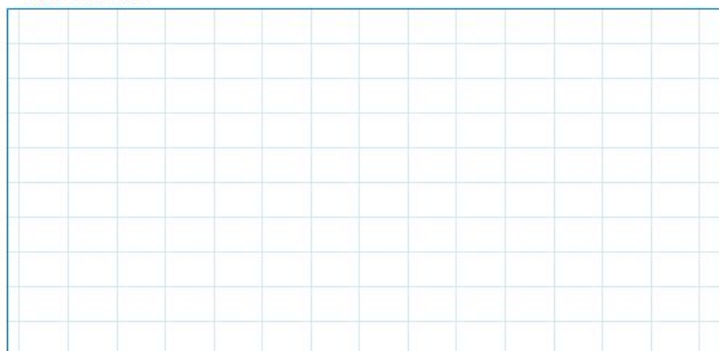
77%



▼ Video Encode

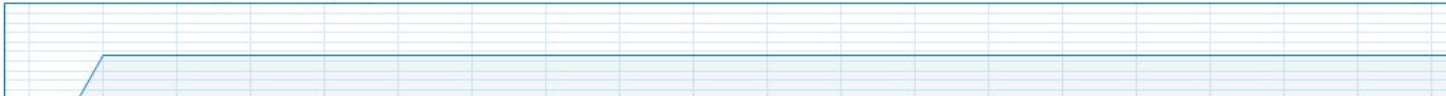
0% ▼ Video Decode

0%



Utilisation de la mémoire du processeur graphique dédié

4,0 Go



Utilisation de la mémoire du processeur graphique partagé

15,9 Go