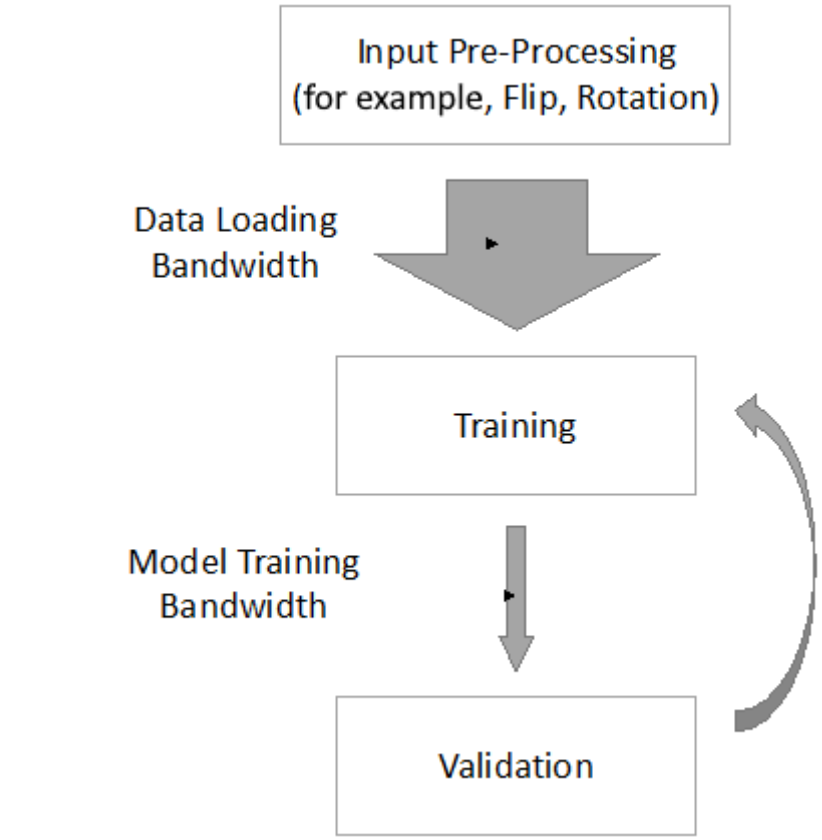
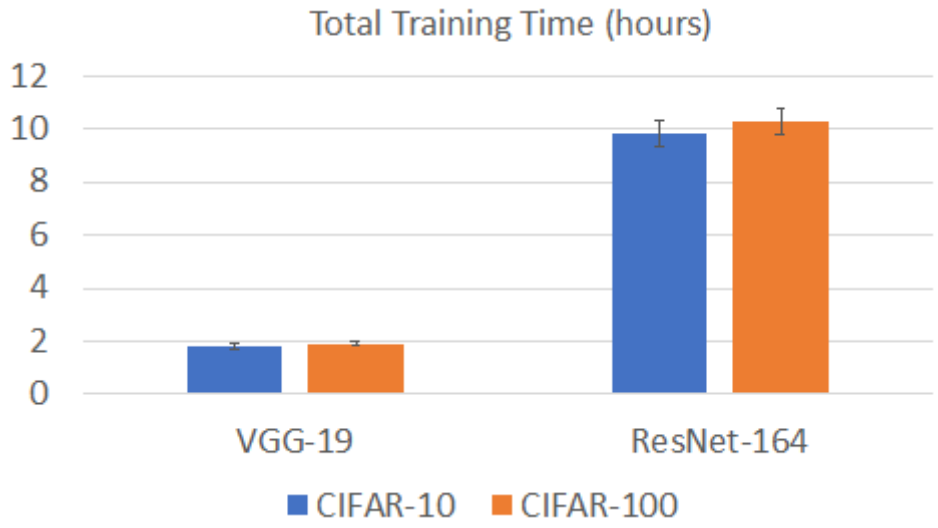
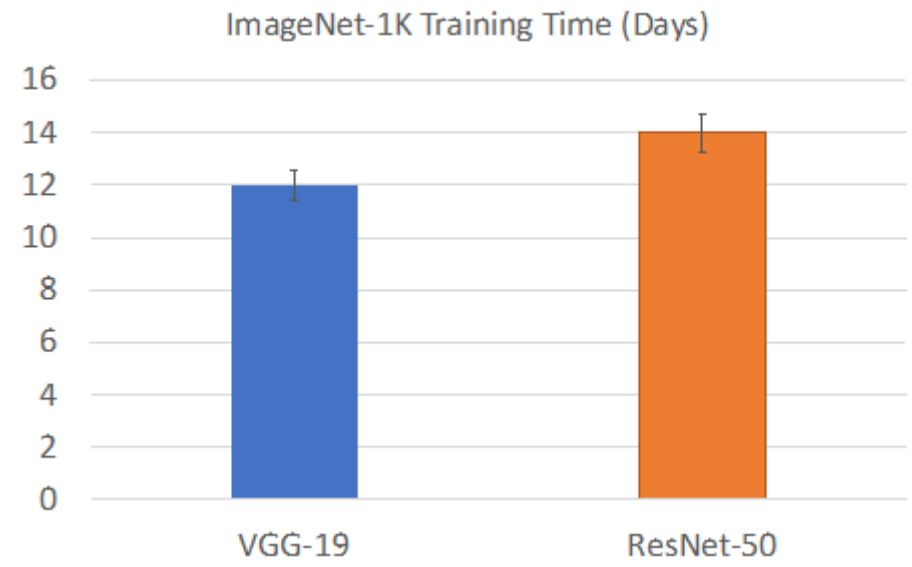
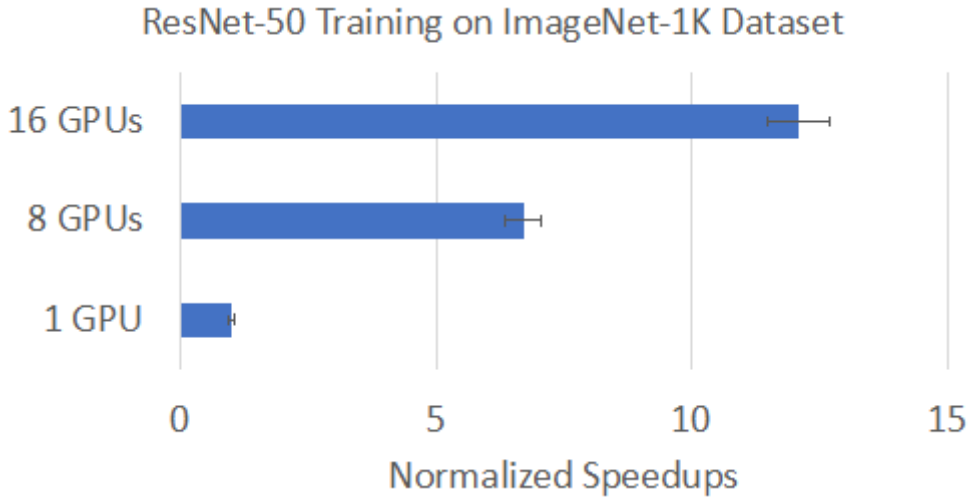
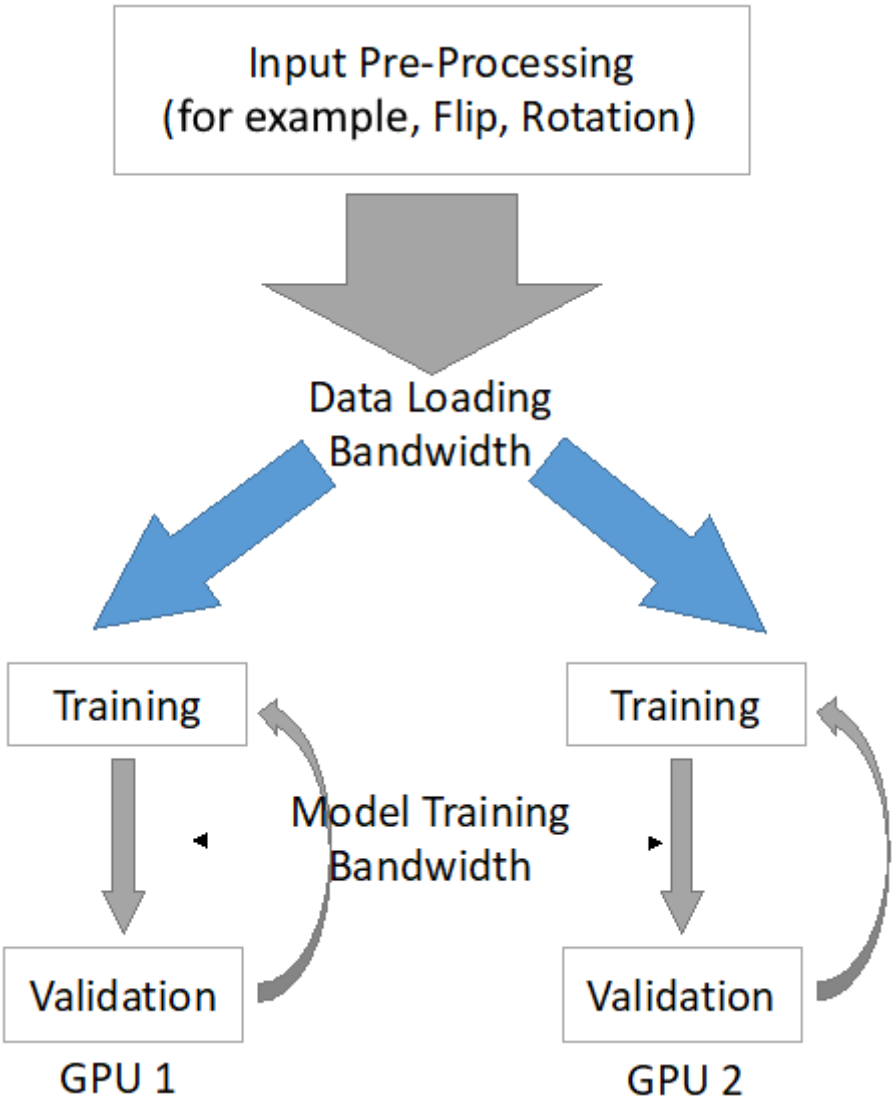
Chapter 1: Splitting Input Data

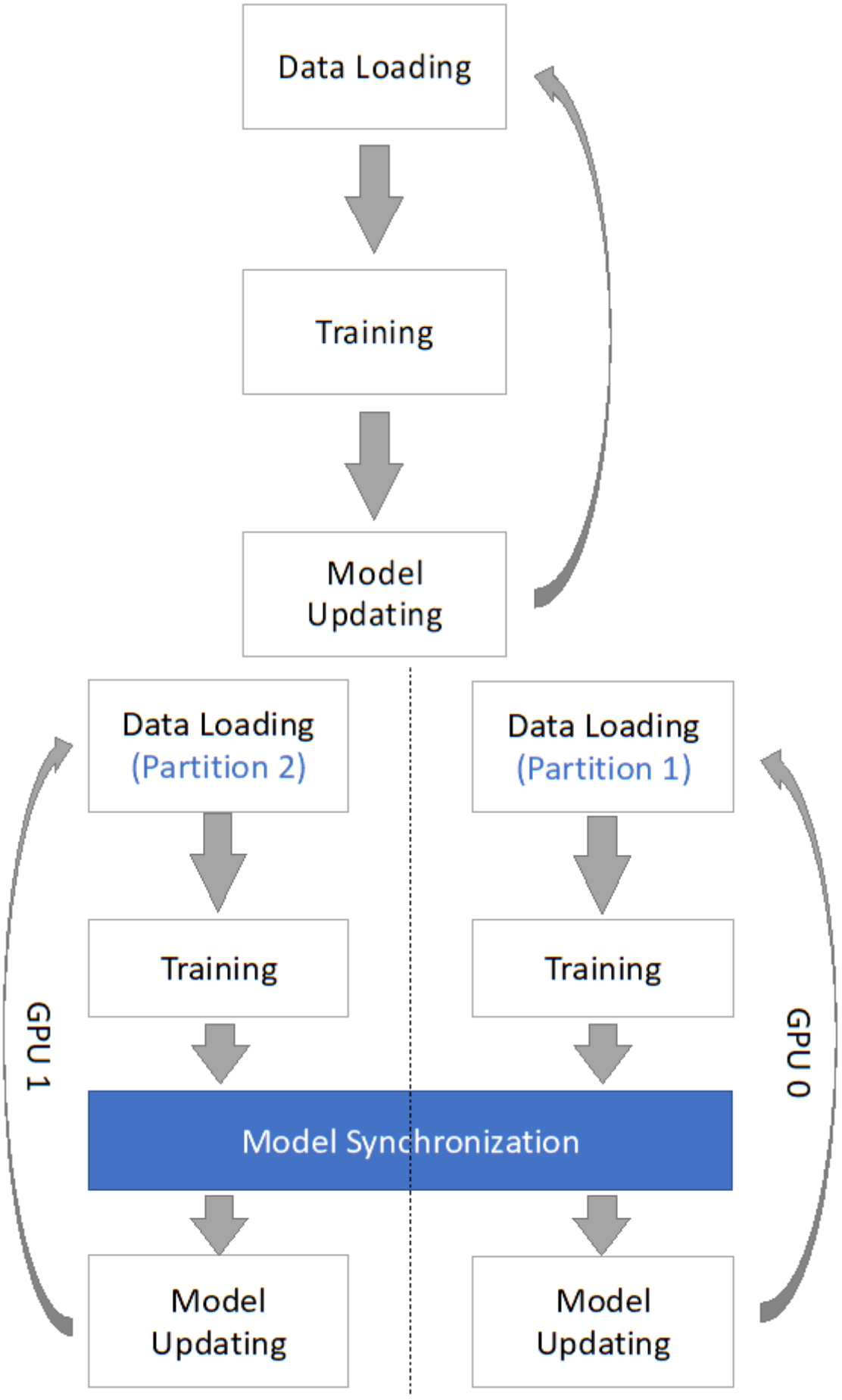


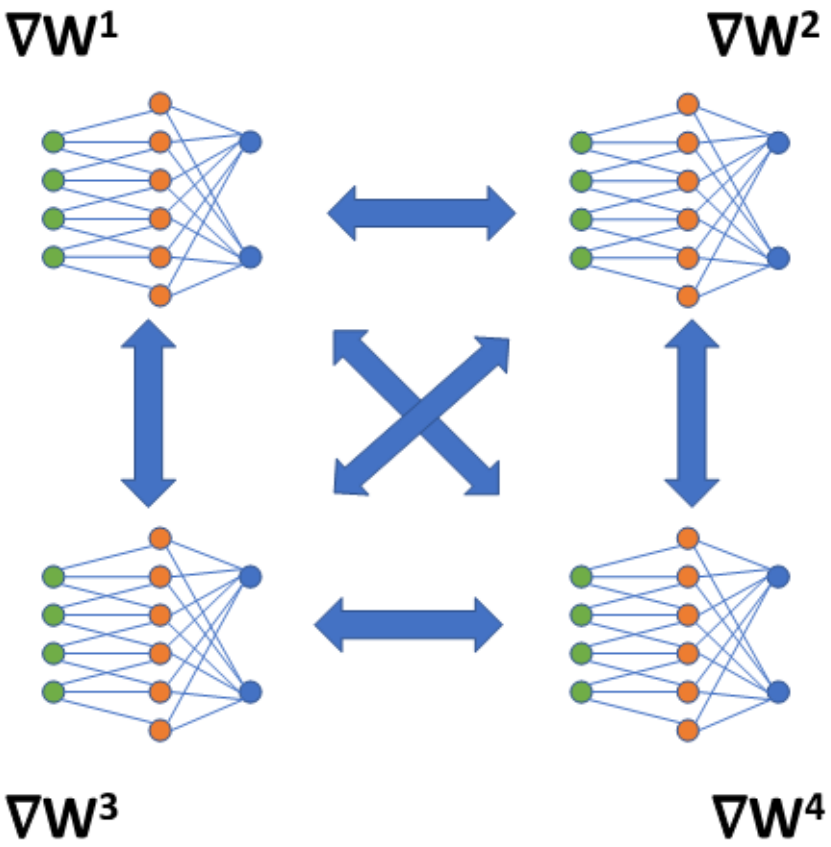



ImageNet-1K Training Time (Days)
16
14
12
10
8
6
4
2
0
VGG-19
ResNet-50


ResNet-50 Training on ImageNet-1K Dataset
16 GPUs
8 GPUs
1 GPU
0
5
10
15
Normalized Speedups


Input Pre-Processing
(for example, Flip, Rotation)
Data Loading
Bandwidth
Training
Training
Model Training
Bandwidth
Validation
Validation
GPU 1
GPU 2

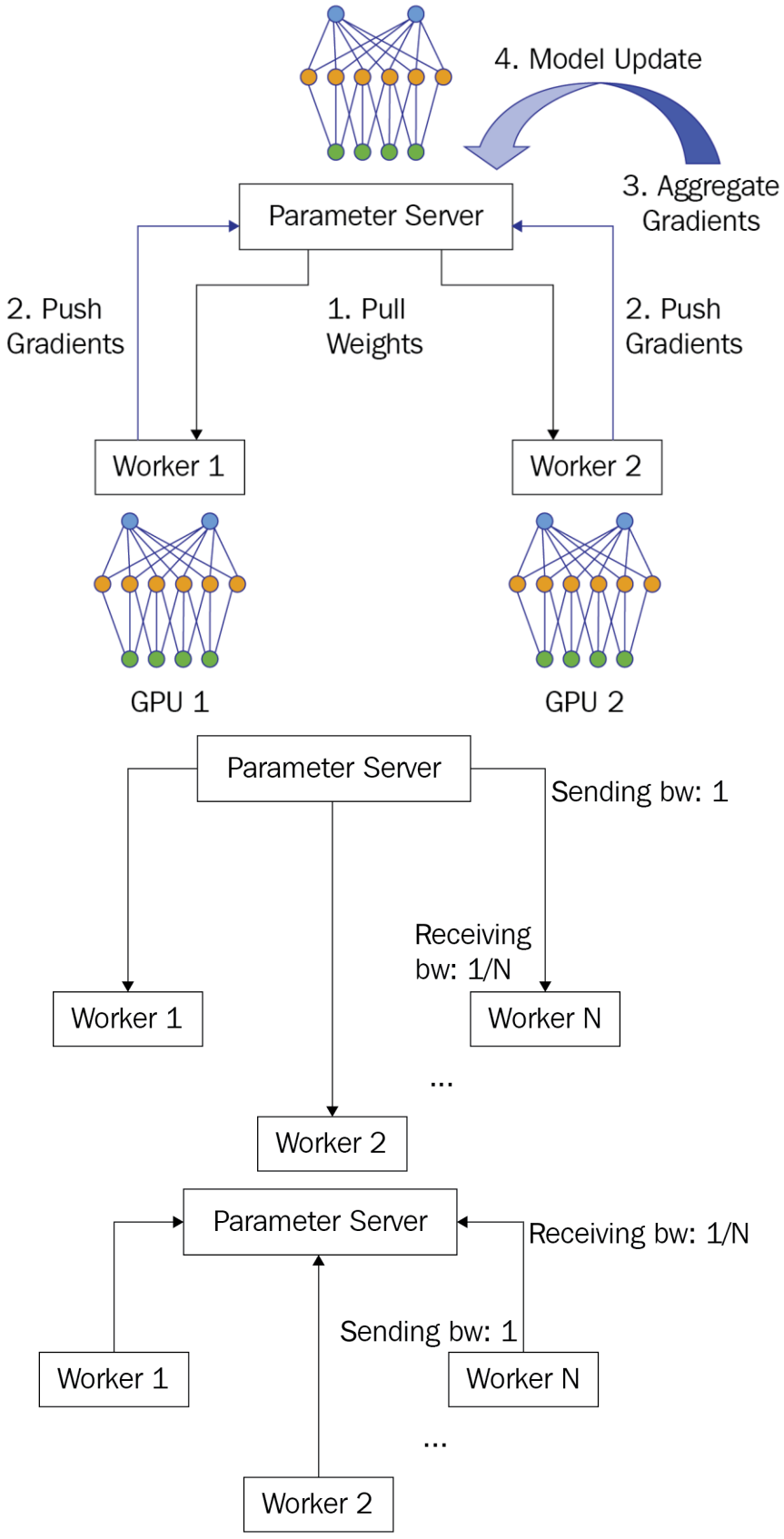

Data Loading
Training
Model Updating
Data Loading (Partition 2)
Data Loading (Partition 1)
Training
Training
GPU 1
GPU 0
Model Synchronization
Model Updating
Model Updating

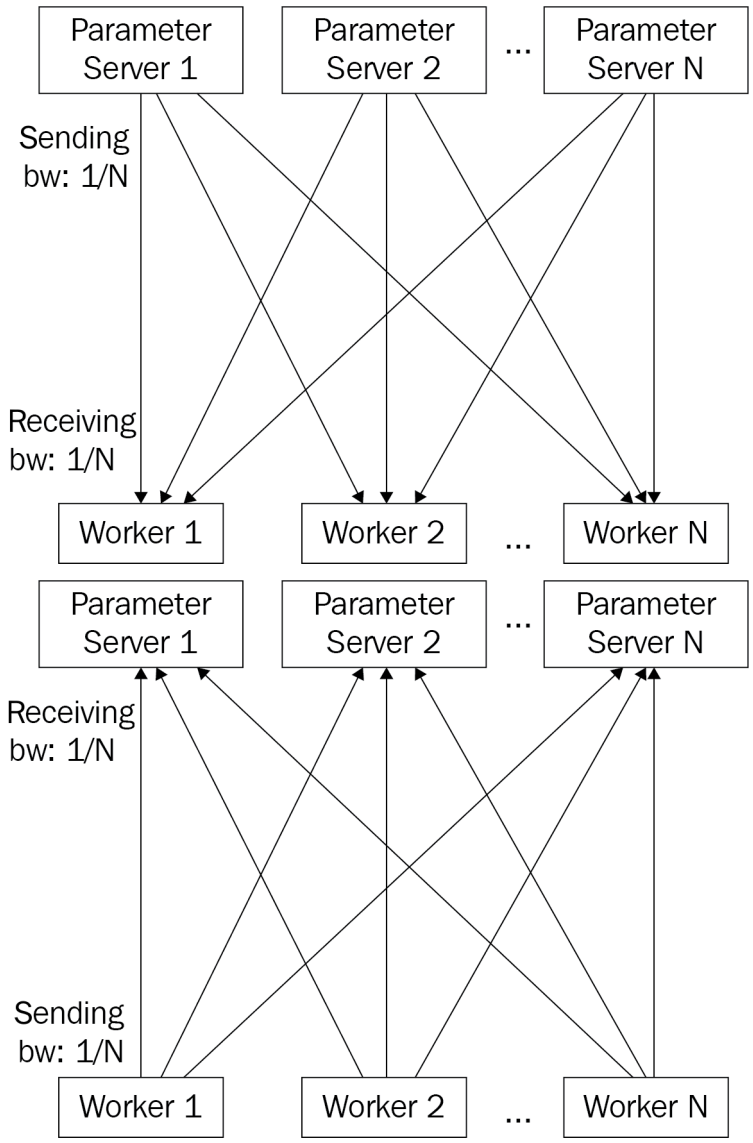


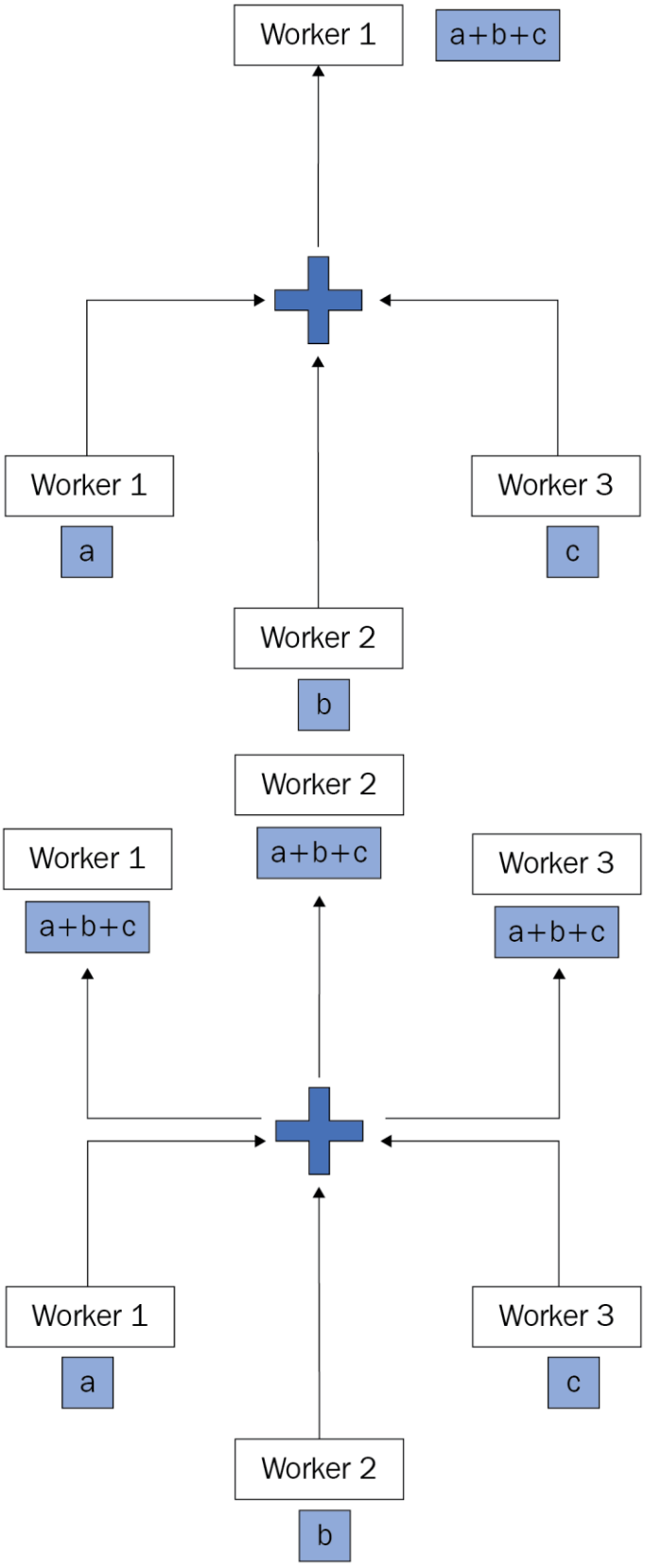
Synchronization

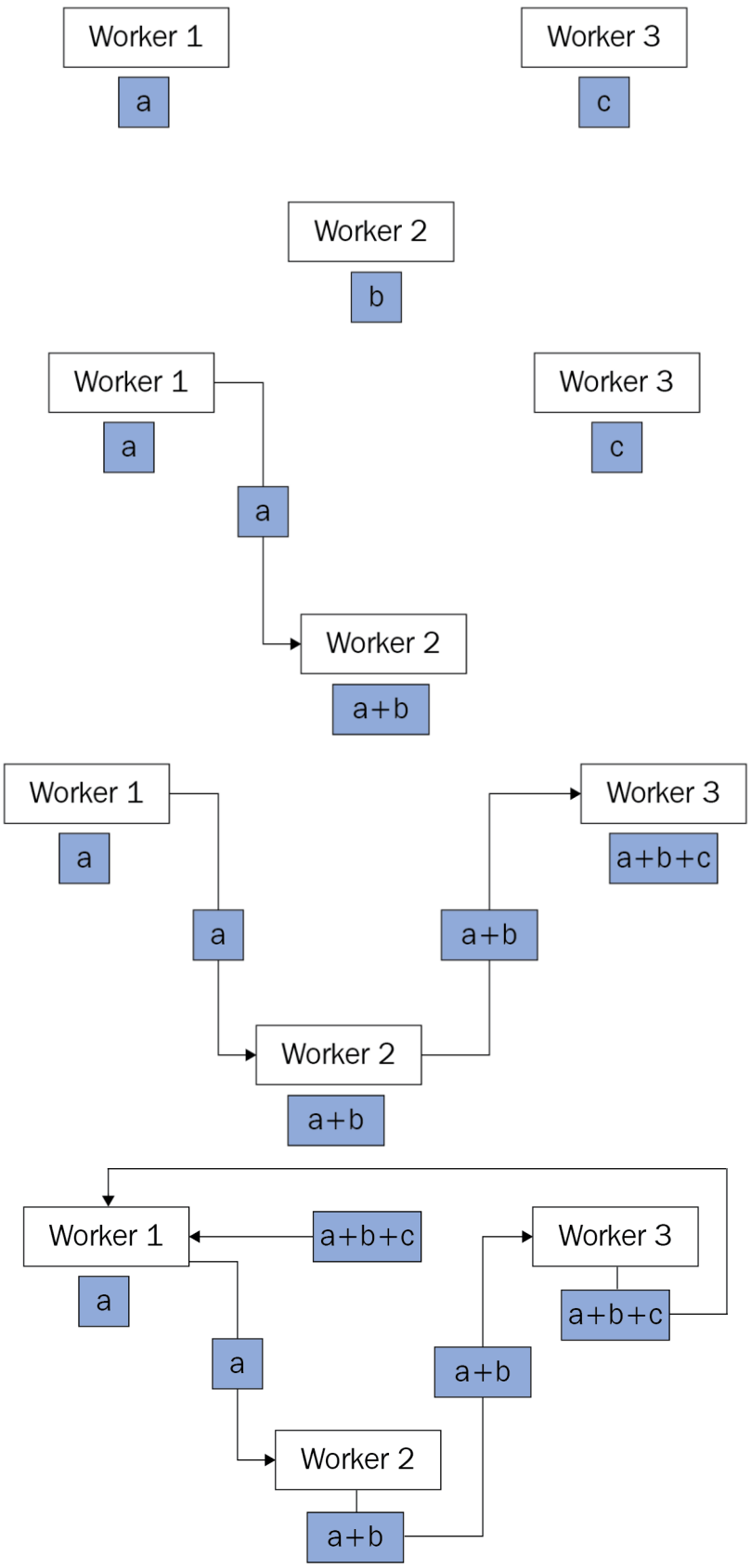
$$\nabla W = \nabla W^1 + \nabla W^2 + \cdots + \nabla W^N$$

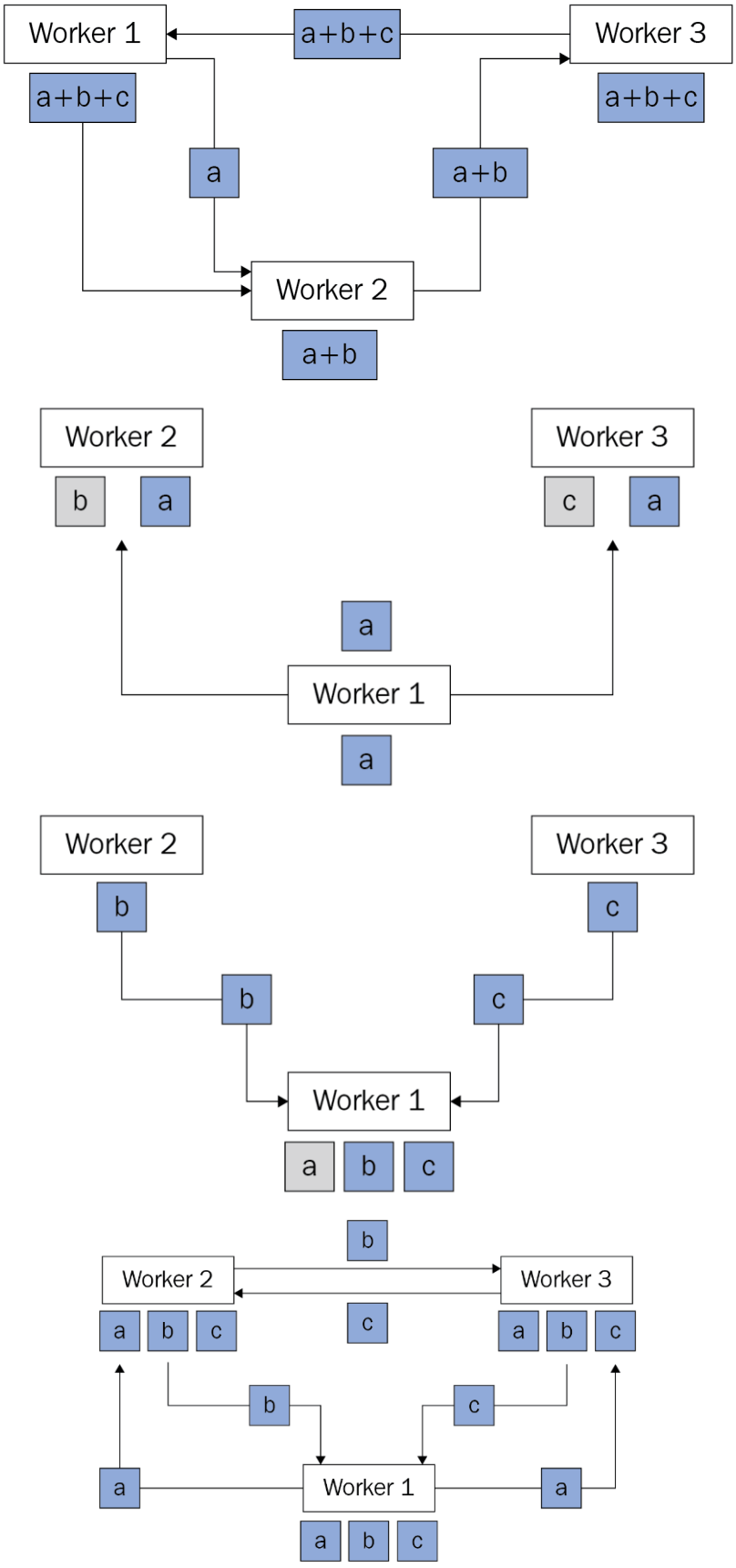
Chapter 2: Parameter Server and All-Reduce



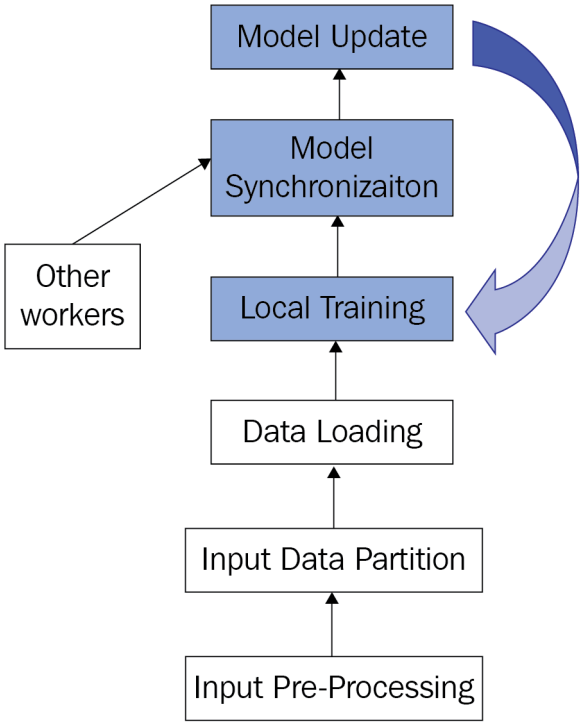

Parameter Server 1
Parameter Server 2
...
Parameter Server N
Sending bw: 1/N
Receiving bw: 1/N
Worker 1
Worker 2
...
Worker N
Parameter Server 1
Parameter Server 2
...
Parameter Server N
Receiving bw: 1/N
Sending bw: 1/N
Worker 1
Worker 2
...
Worker N


Worker 1
a+b+c
Worker 1
a
Worker 3
c
Worker 2
b
Worker 2
a+b+c
Worker 1
a+b+c
Worker 3
a+b+c
Worker 1
a
Worker 3
c
Worker 2
b

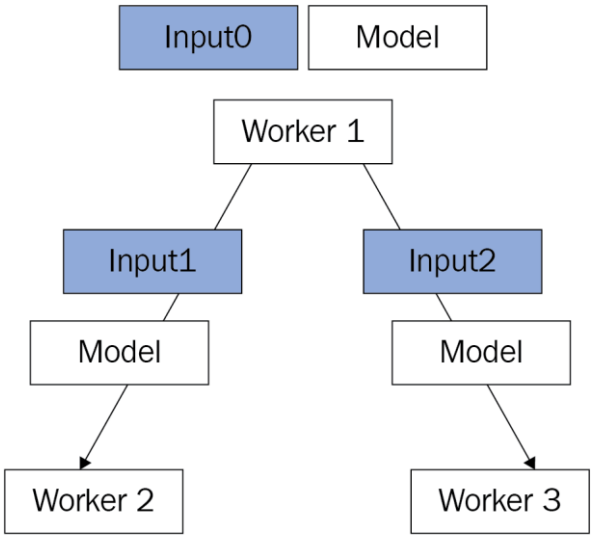

Worker 1
a
Worker 3
c
Worker 2
b
Worker 1
a
Worker 3
c
a
Worker 2
a+b
Worker 1
a
Worker 3
a+b+c
a
a+b
Worker 2
a+b
Worker 1
a+b+c
Worker 3
a
a+b+c
a
a+b
Worker 2
a+b

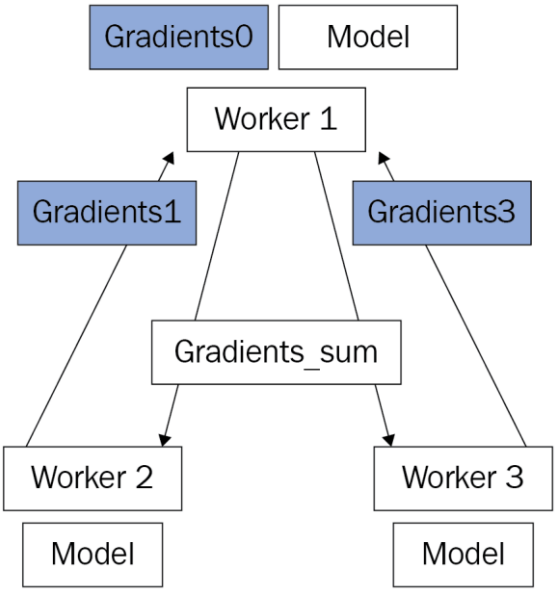

Worker 1
a+b+c
Worker 3
a+b+c
a+b+c
a
a+b
Worker 2
a+b
Worker 2
Worker 3
b
a
c
a
a
Worker 1
a
Worker 2
Worker 3
b
c
b
c
Worker 1
a
b
c
b
Worker 2
Worker 3
c
a
b
c
a
b
c
b
c
a
Worker 1
a
a
b
c

Chapter 3: Building a Data Parallel Training and Serving Pipeline

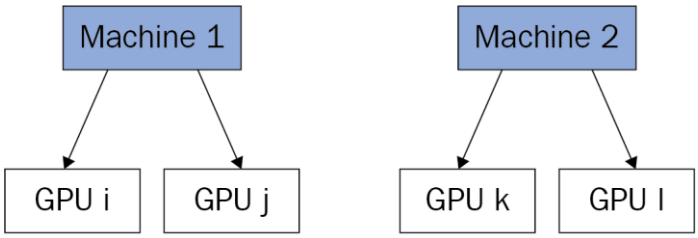



```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.119.03   Driver Version: 450.119.03   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1B.0 Off |                    0 |
| N/A   33C    P0    39W / 300W |      0MiB / 16160MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:00:1C.0 Off |                    0 |
| N/A   33C    P0    39W / 300W |      0MiB / 16160MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:00:1D.0 Off |                    0 |
| N/A   35C    P0    41W / 300W |      0MiB / 16160MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   35C    P0    41W / 300W |      0MiB / 16160MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```






```
+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A      7906      C   python                           1393MiB |
|    1   N/A  N/A      7906      C   python                           1393MiB |
|    2   N/A  N/A      7906      C   python                           1393MiB |
|    3   N/A  N/A      7906      C   python                           1035MiB |
+-----------------------------------------------------------------------------+
```



local rank:

0 1 0 1

rank:

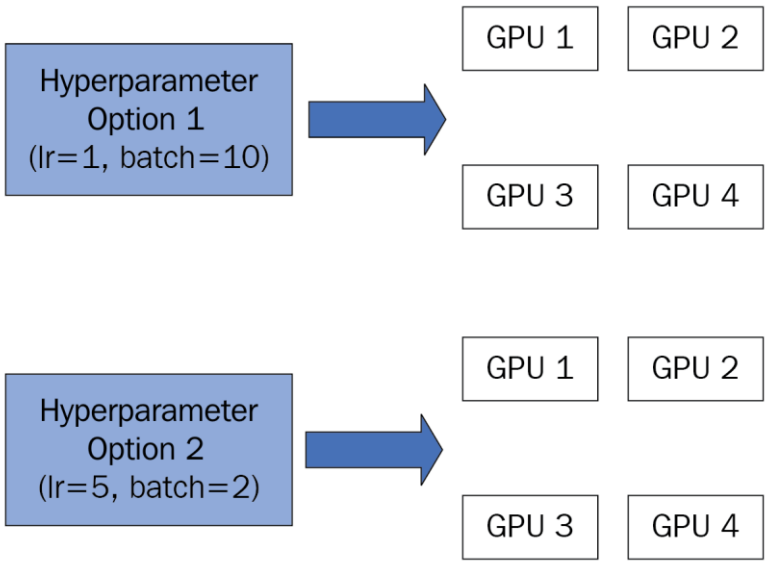
0 1 2 3

world_size:

4

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.119.03   Driver Version: 450.119.03   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1B.0 Off |                    0 |
| N/A   43C    P0    57W / 300W |   1768MiB / 16160MiB |     27%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:00:1C.0 Off |                    0 |
| N/A   42C    P0    60W / 300W |   1792MiB / 16160MiB |     29%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:00:1D.0 Off |                    0 |
| N/A   43C    P0    68W / 300W |   1792MiB / 16160MiB |     21%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   45C    P0    65W / 300W |   1768MiB / 16160MiB |     29%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     40074      C   ...rch_latest_p37/bin/python     1765MiB |
|    1   N/A  N/A     40075      C   ...rch_latest_p37/bin/python     1789MiB |
|    2   N/A  N/A     40076      C   ...rch_latest_p37/bin/python     1789MiB |
|    3   N/A  N/A     40077      C   ...rch_latest_p37/bin/python     1765MiB |
+-----------------------------------------------------------------------------+
```

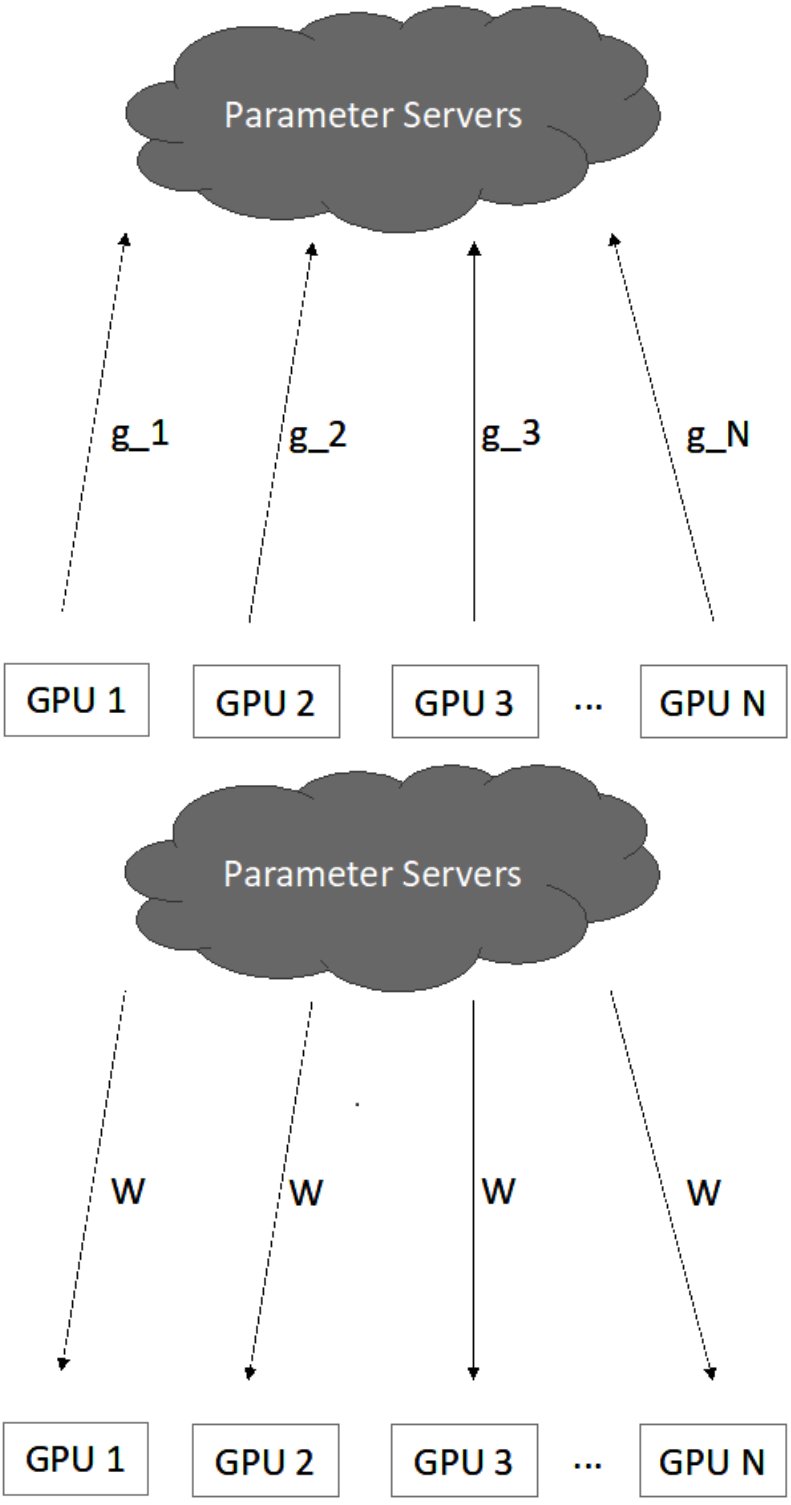
```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.142.00   Driver Version: 450.142.00   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1B.0 Off |                    0 |
| N/A   45C    P0    71W / 300W |   1768MiB / 16160MiB |     25%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:00:1C.0 Off |                    0 |
| N/A   46C    P0    67W / 300W |   1792MiB / 16160MiB |     24%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:00:1D.0 Off |                    0 |
| N/A   47C    P0    82W / 300W |   1792MiB / 16160MiB |     26%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   45C    P0    73W / 300W |   1768MiB / 16160MiB |     29%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     27914      C   ...rch_latest_p37/bin/python     1765MiB |
|    1   N/A  N/A     27915      C   ...rch_latest_p37/bin/python     1789MiB |
|    2   N/A  N/A     27916      C   ...rch_latest_p37/bin/python     1789MiB |
|    3   N/A  N/A     27917      C   ...rch_latest_p37/bin/python     1765MiB |
+-----------------------------------------------------------------------------+
```

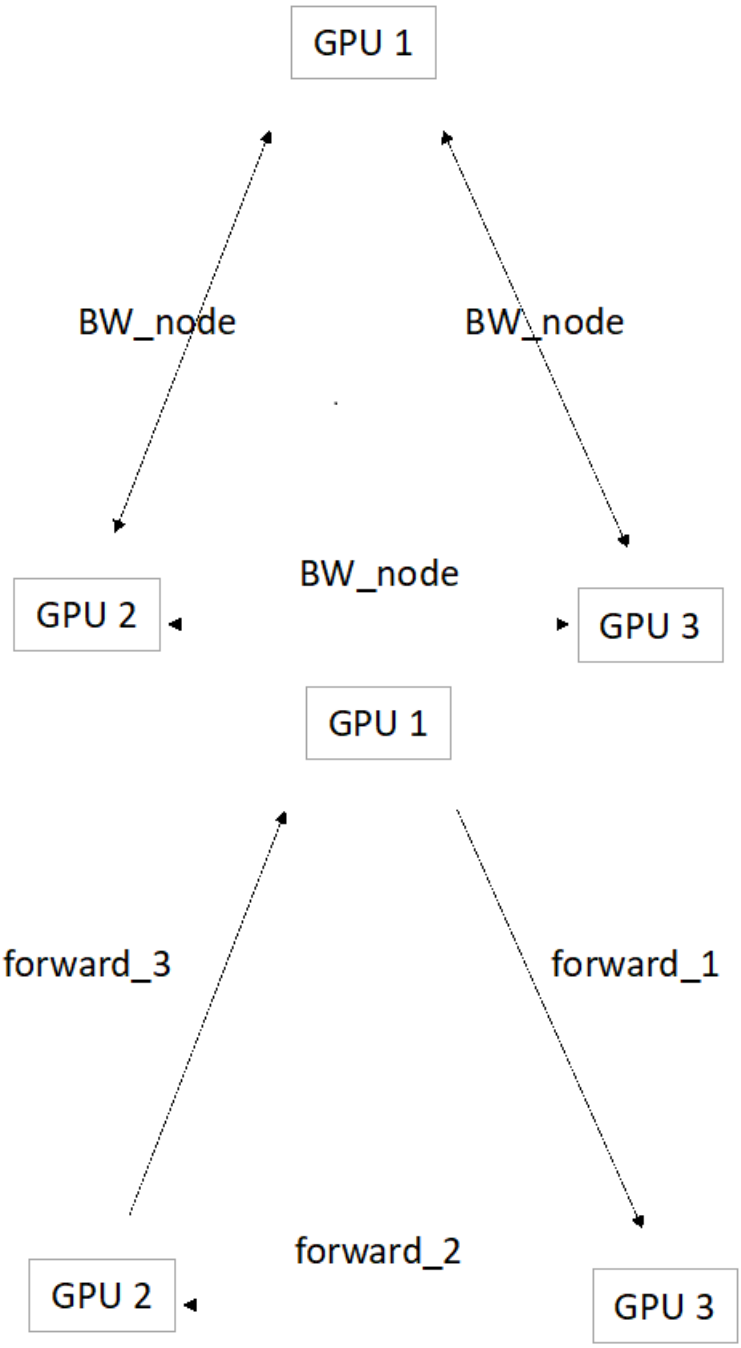


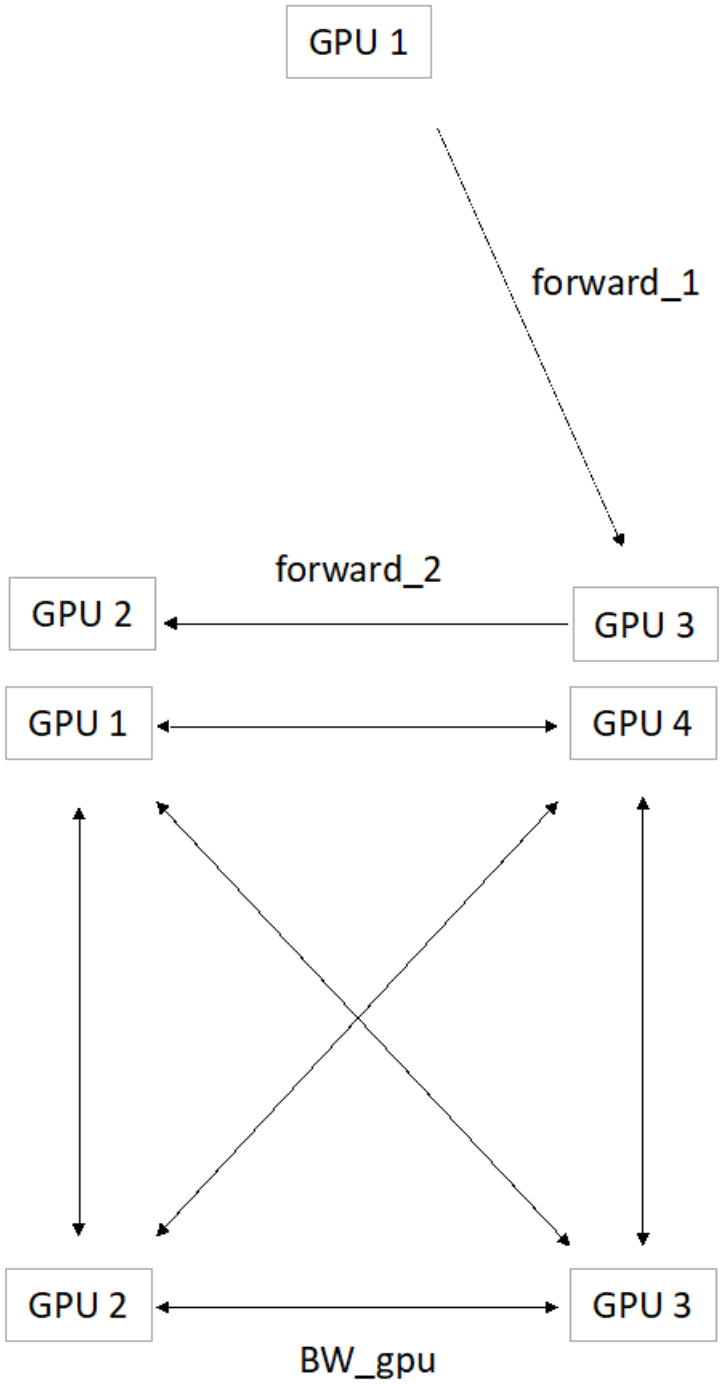
```
Load checkpoint 3
Load checkpoint 2
Load checkpoint 1
Load checkpoint 0
Checkpoint loading done!
Checkpoint loading done!
GPU 0, Test Accuracy 0.1198
Test Done!
GPU 1, Test Accuracy 0.1189
Test Done!
Checkpoint loading done!
Checkpoint loading done!
GPU 2, Test Accuracy 0.1192
Test Done!
GPU 3, Test Accuracy 0.1181
Test Done!
```

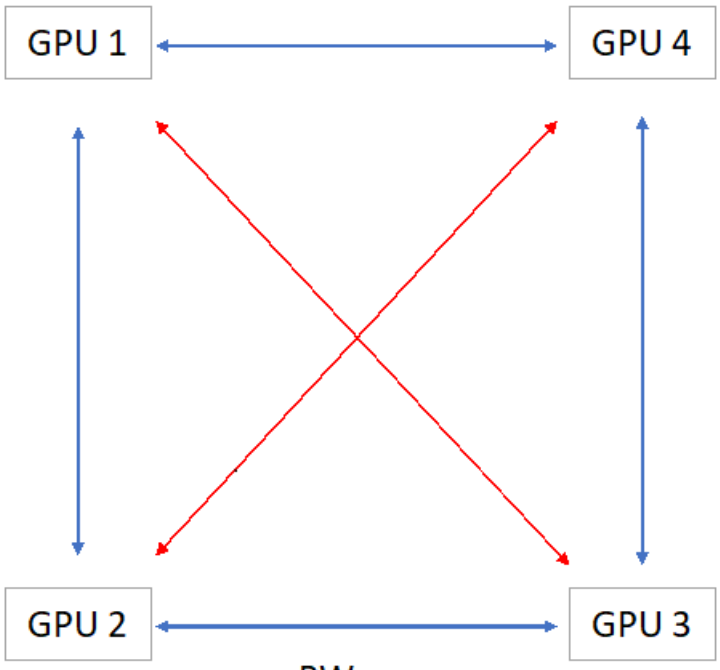
```
Load checkpoint 6
Load checkpoint 5
Load checkpoint 7
Load checkpoint 4
Checkpoint loading done!
Checkpoint loading done!
GPU 5, Test Accuracy 0.119
Test Done!
GPU 4, Test Accuracy 0.1202
Test Done!
Checkpoint loading done!
Checkpoint loading done!
GPU 7, Test Accuracy 0.1195
Test Done!
GPU 6, Test Accuracy 0.1193
Test Done!
```

Chapter 4: Bottlenecks and Solutions

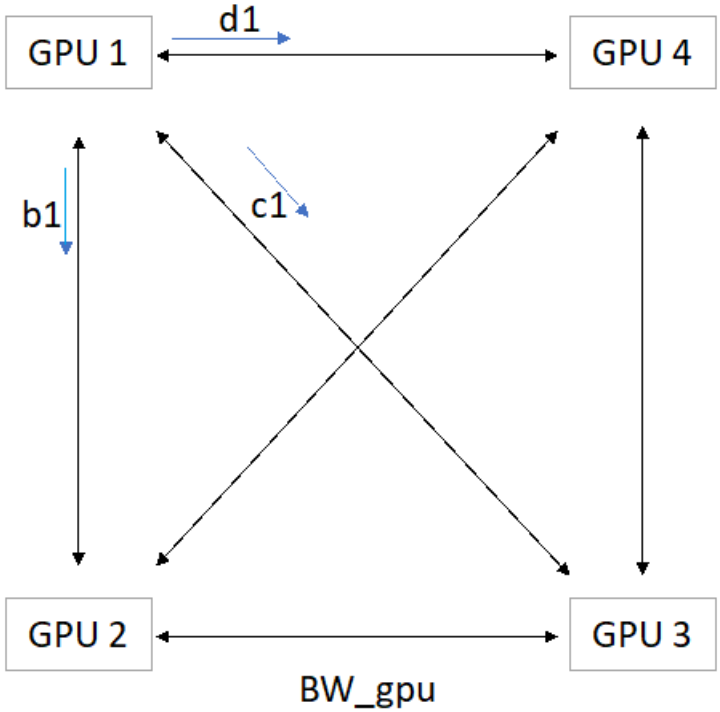



GPU 1
BW_node
BW_node
BW_node
GPU 2
GPU 3
GPU 1
forward_3
forward_1
forward_2
GPU 2
GPU 3


GPU 1
forward_1
forward_2
GPU 2
GPU 3
GPU 1
GPU 4
GPU 2
GPU 3
BW_gpu



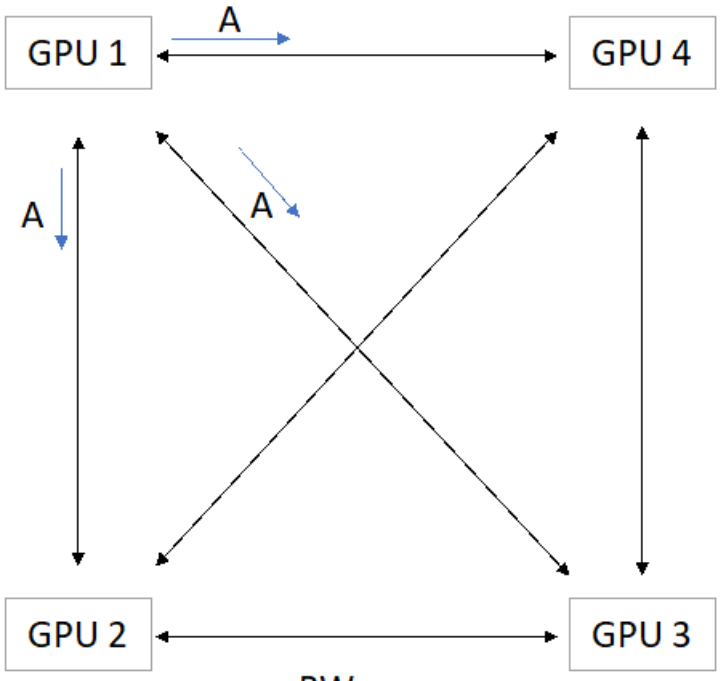
a1,b1,c1,d1 a4,b4,c4,d4



a2,b2,c2,d2 a3,b3,c3,d3

A,b1,c1,d1

a4,b4,c4,D



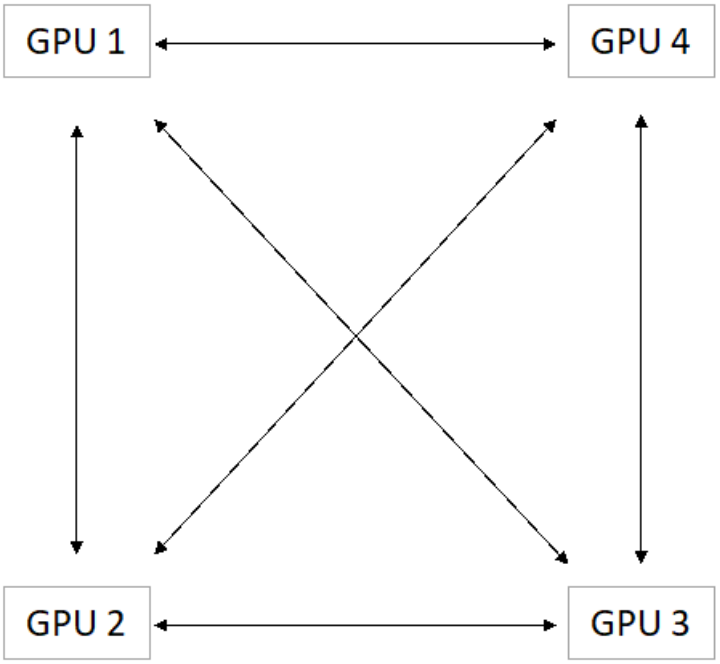
BW_gpu

a2,B,c2,d2

A,B,C,D

a3,b3,C,d3

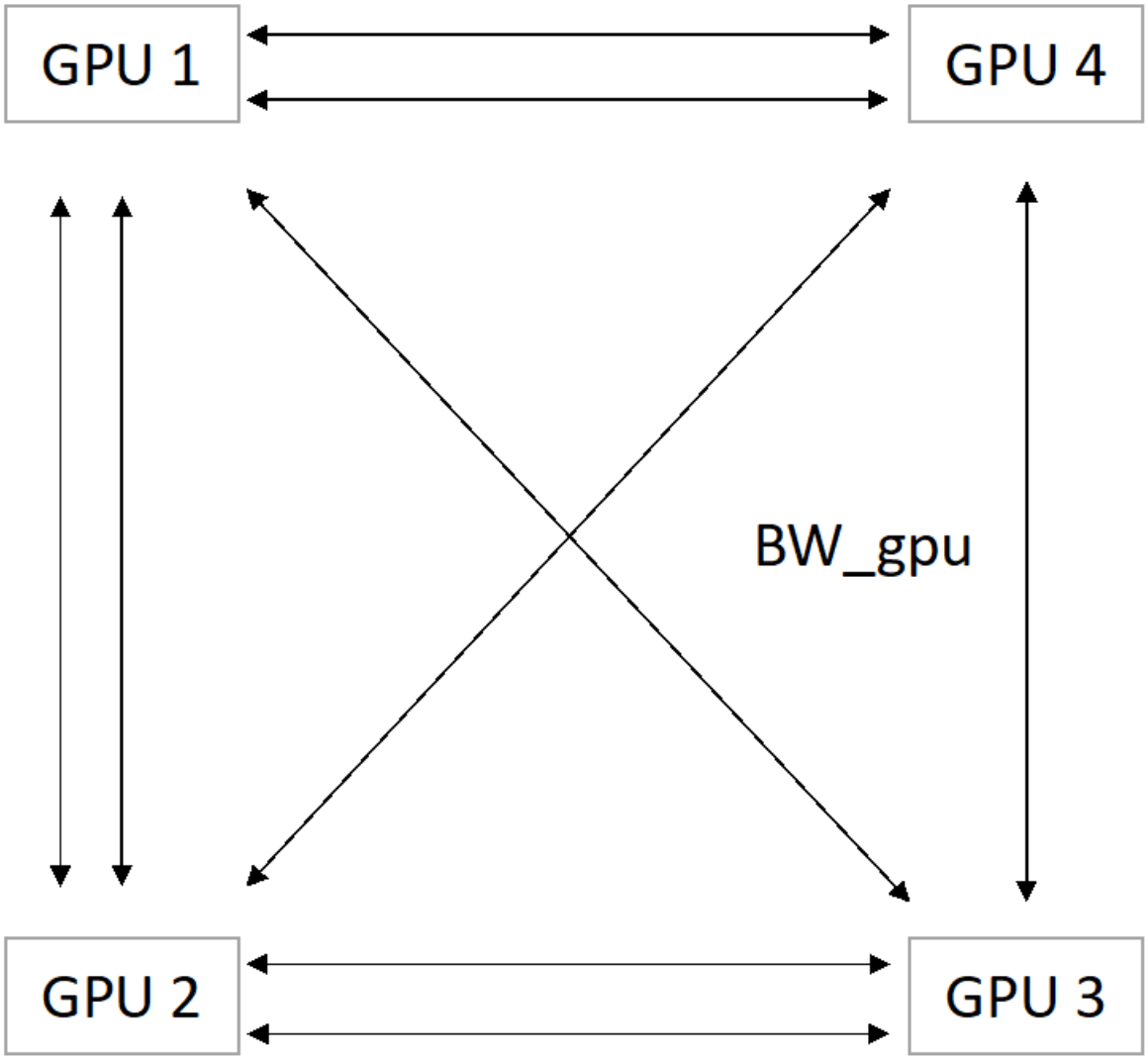
A,B,C,D

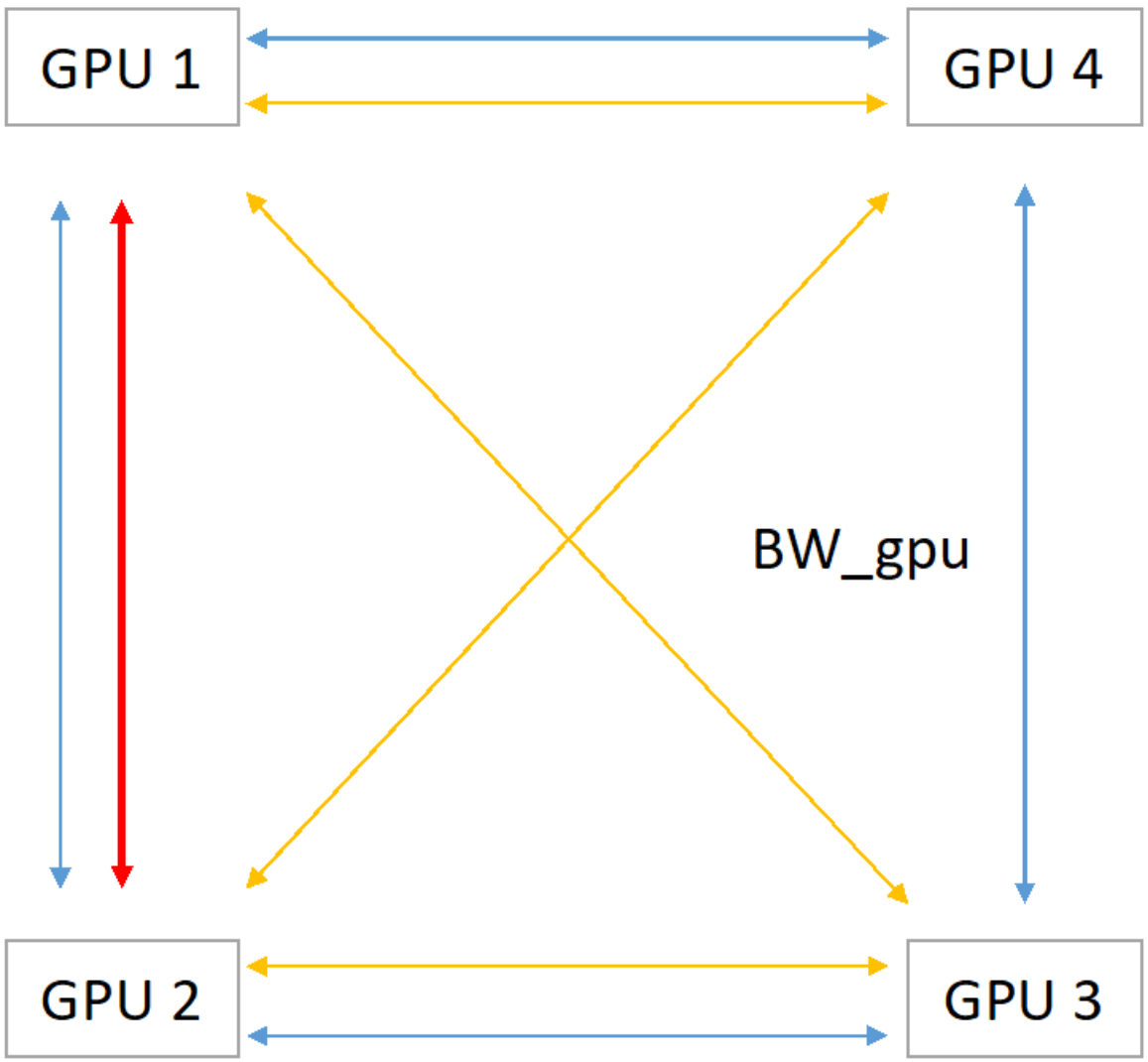


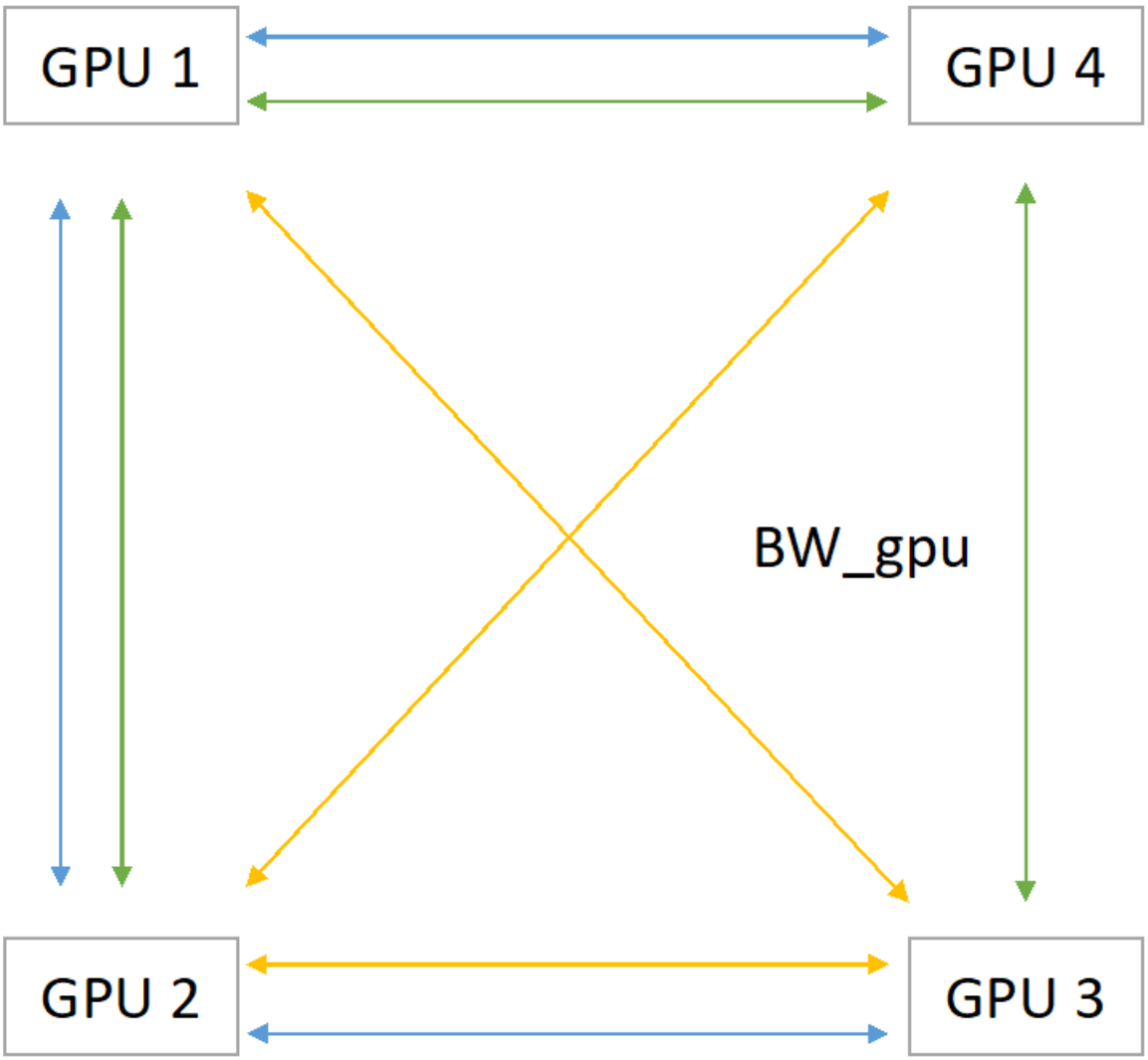
BW_gpu

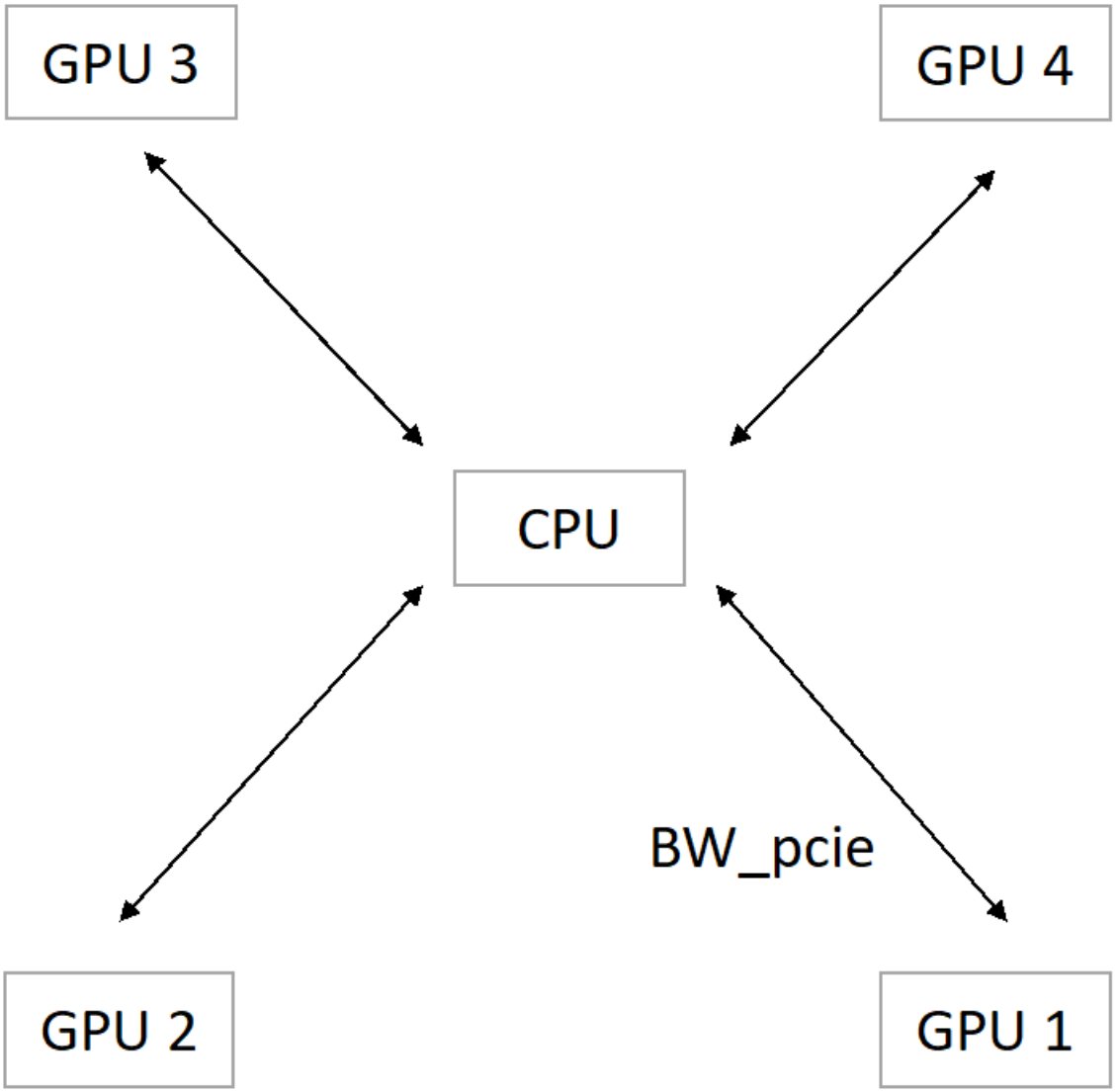
A,B,C,D

A,B,C,D


GPU 1
GPU 4
BW_gpu
GPU 2
GPU 3

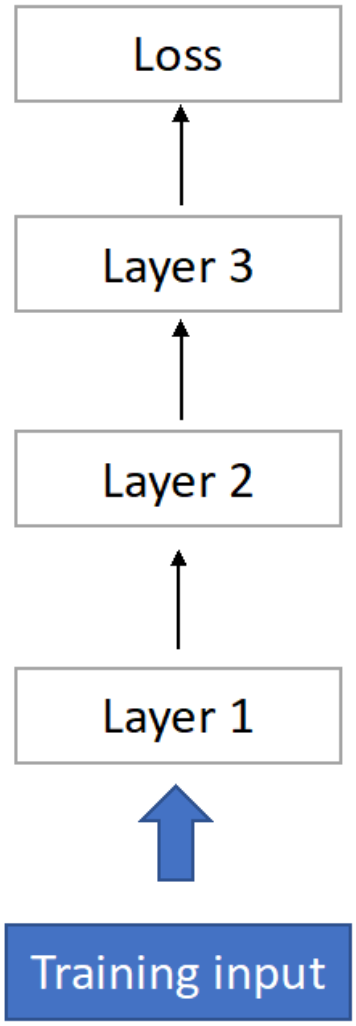

GPU 1
GPU 4
BW_gpu
GPU 2
GPU 3

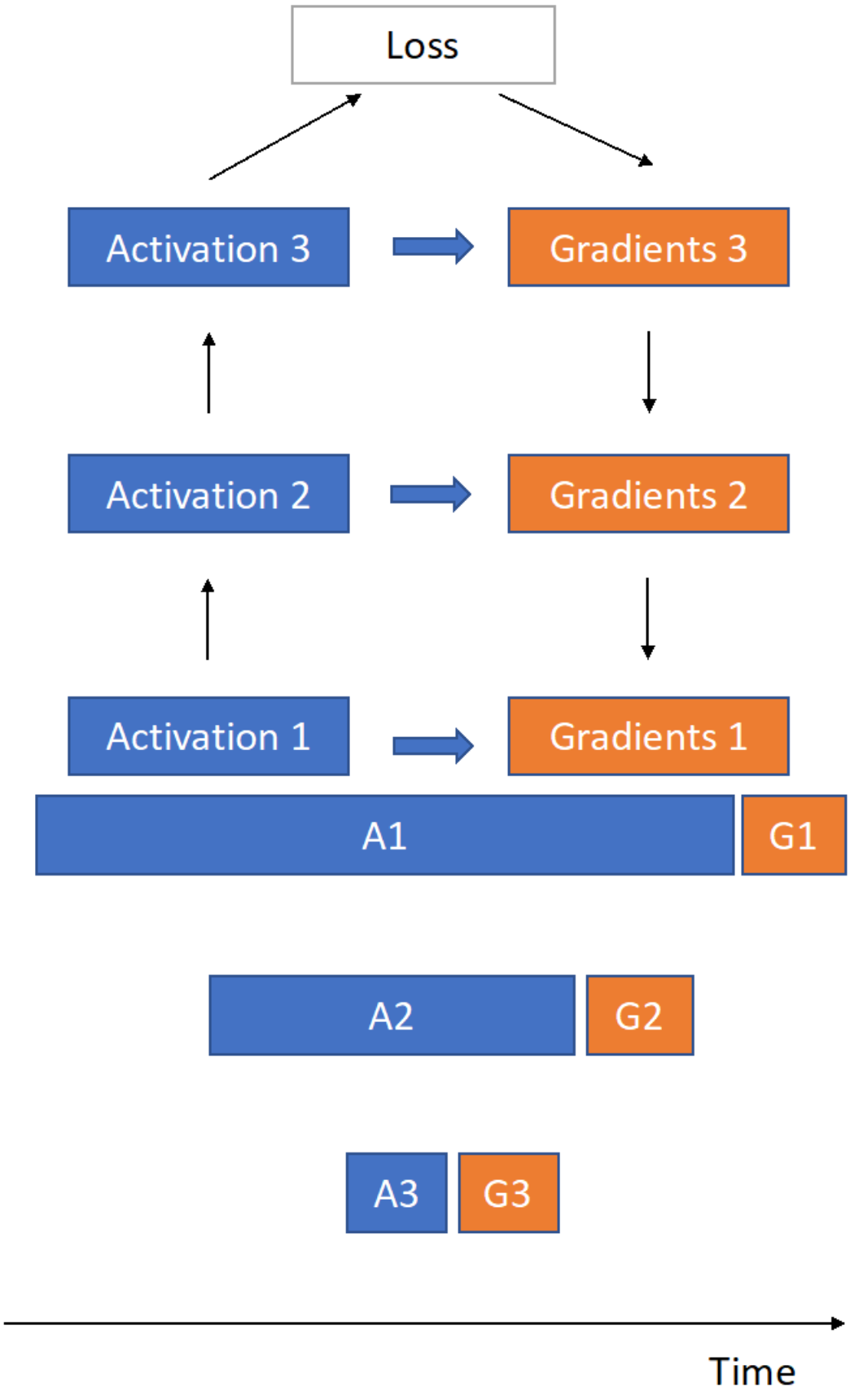

GPU 1
GPU 4
BW_gpu
GPU 2
GPU 3

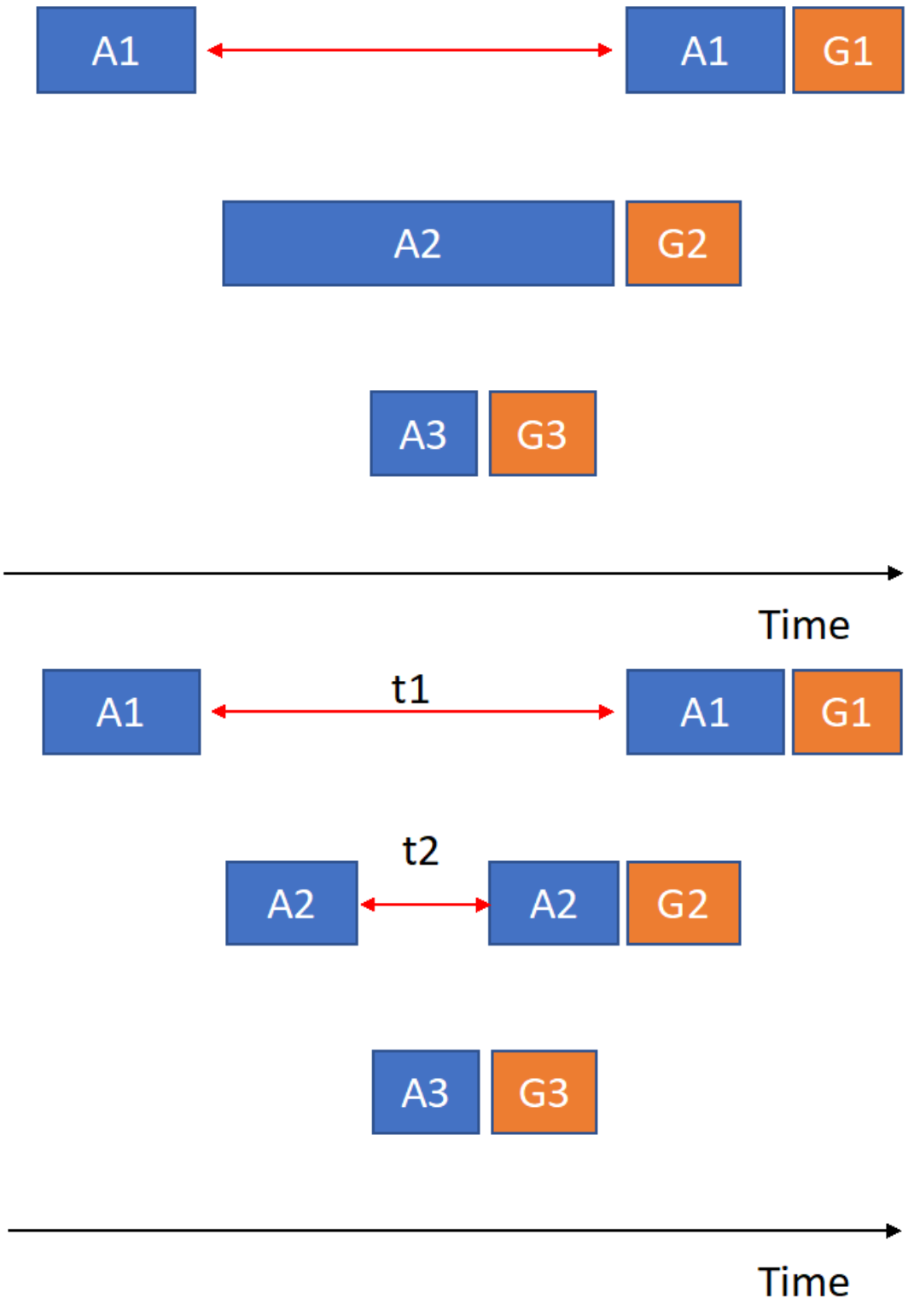


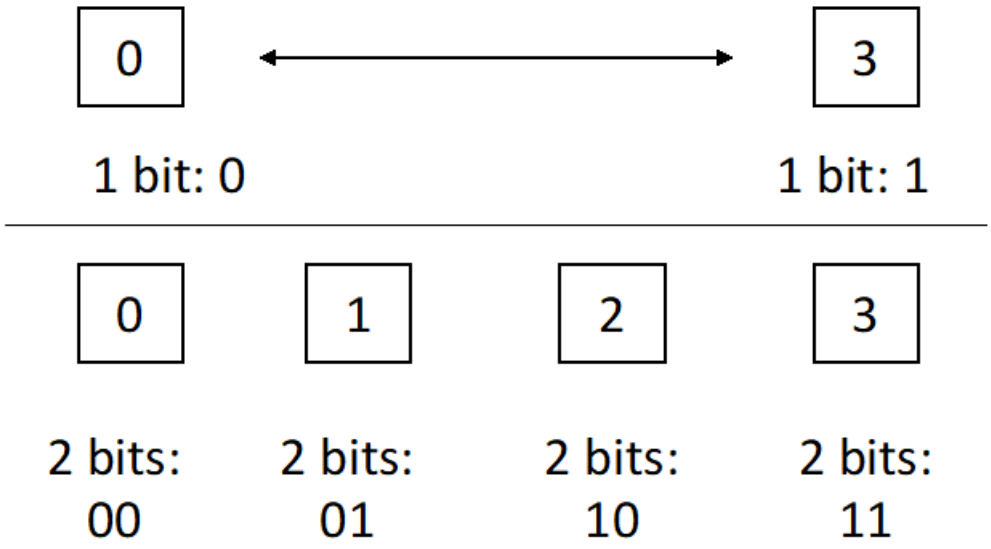
GPU type	On-device memory size
NVIDIA 1080	8 GB
NVIDIA RTX 2080	8 GB
NVIDIA K80	12 GB
NVIDIA V100	16 GB
NVIDIA A100	40 GB

	Recomputation	Quantization
Lossy/lossless	Lossless	Lossy
Reducing memory footprint	Yes	Yes
Computation overhead	Medium	Low




Loss
Activation 3
Gradients 3
Activation 2
Gradients 2
Activation 1
Gradients 1
A1
G1
A2
G2
A3
G3
Time


A1
A1
G1
A2
G2
A3
G3
Time
A1
t1
A1
G1
t2
A2
A2
G2
A3
G3
Time


0
3
1 bit: 0
1 bit: 1
0
1
2
3
2 bits:
00
2 bits:
01
2 bits:
10
2 bits:
11

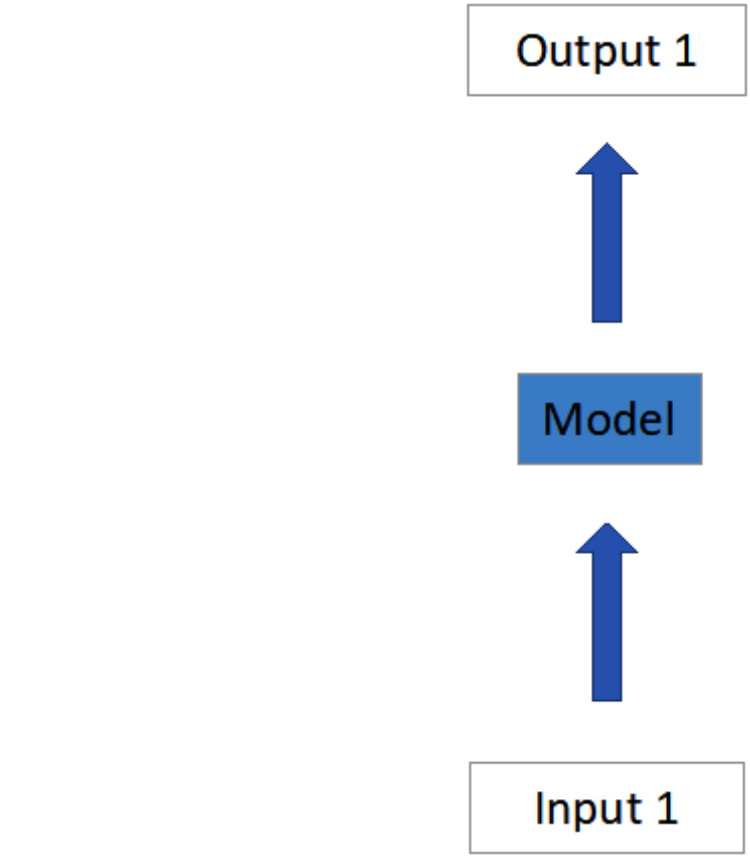
Chapter 5: Splitting the Model

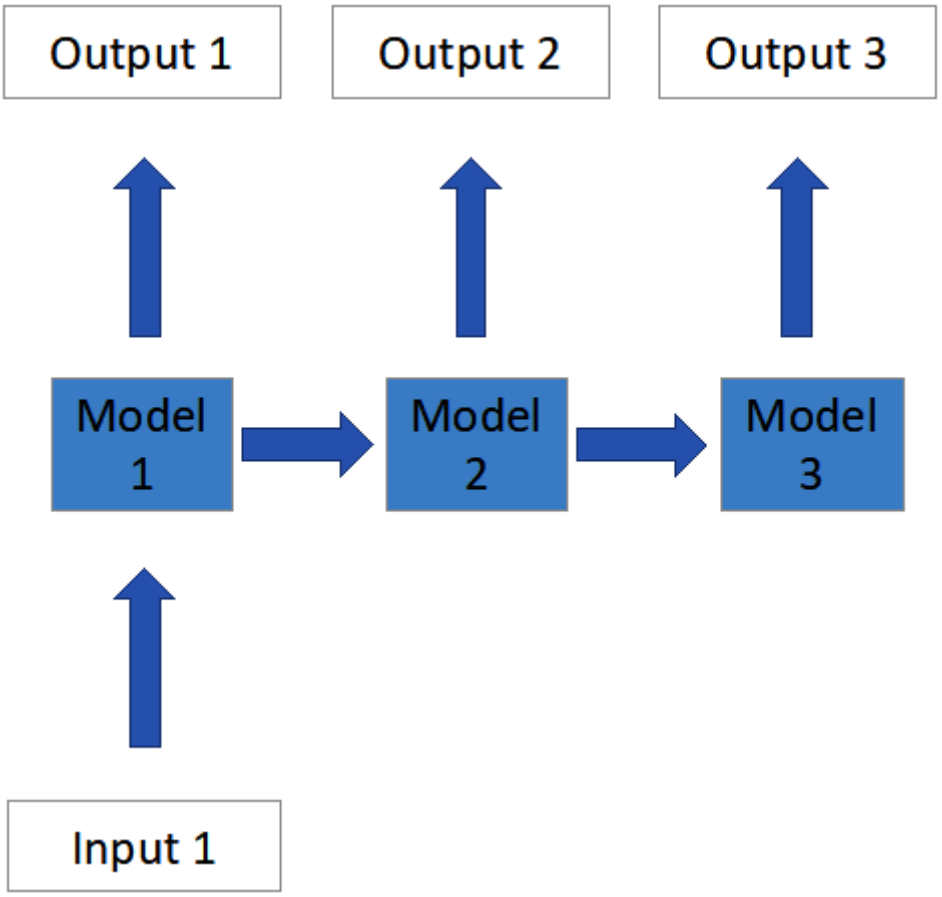
```
Python 3.7.10 | packaged by conda-forge | (default, Feb 19 2021, 16:07:37)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import transformers
>>> print(transformers.__version__)
4.10.3
>>>
```

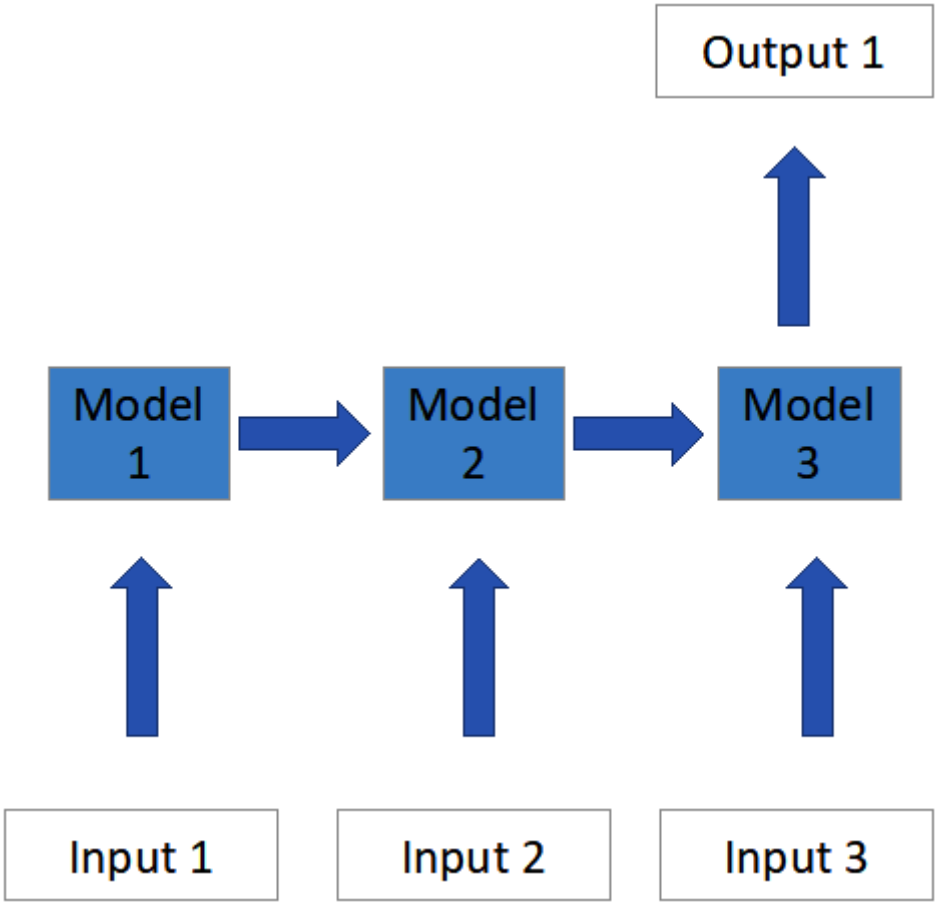
```
Training epoch  1
  0%|                                                                                          | 0/86136
[00:00<?, ?it/s]Traceback (most recent call last):
  File "bert.py", line 198, in <module>
    end_positions=end_token_idx, return_dict=False)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/modules/module.py",
line 889, in _call_impl
    result = self.forward(*input, **kwargs)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 1825, in forward
    return_dict=return_dict,
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/modules/module.py",
line 889, in _call_impl
    result = self.forward(*input, **kwargs)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 1000, in forward
    return_dict=return_dict,
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/modules/module.py",
line 889, in _call_impl
    result = self.forward(*input, **kwargs)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 589, in forward
    output_attentions,
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/modules/module.py",
line 889, in _call_impl
    result = self.forward(*input, **kwargs)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 511, in forward
    self.feed_forward_chunk, self.chunk_size_feed_forward, self.seq_len_dim, attention_output
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/modeling_utils.p
y", line 2196, in apply_chunking_to_forward
    return forward_fn(*input_tensors)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 522, in feed_forward_chunk
    intermediate_output = self.intermediate(attention_output)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/modules/module.py",
line 889, in _call_impl
    result = self.forward(*input, **kwargs)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/transformers/models/bert/mode
ling_bert.py", line 426, in forward
    hidden_states = self.intermediate_act_fn(hidden_states)
  File "/home/ubuntu/anaconda3/envs/pytorch_latest_p37/lib/python3.7/site-packages/torch/nn/functional.py", line
 1459, in gelu
    return torch._C._nn.gelu(input)
RuntimeError: CUDA out of memory. Tried to allocate 144.00 MiB (GPU 0; 15.78 GiB total capacity; 14.19 GiB alrea
dy allocated; 102.75 MiB free; 14.25 GiB reserved in total by PyTorch)
  0%|                                                                                          | 0/86136
[00:04<?, ?it/s]
```

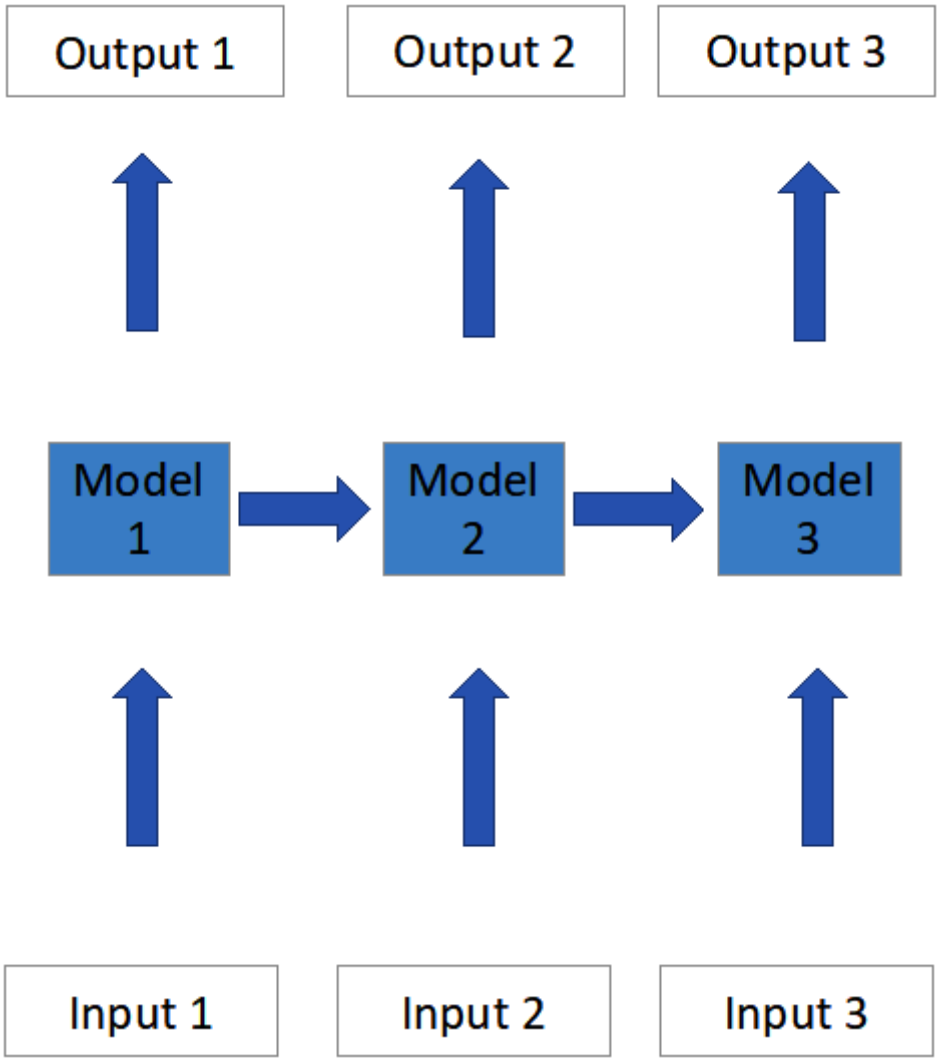
```
Mon Sep 27 19:40:34 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.119.03   Driver Version: 450.119.03   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   35C    P0    23W / 300W |      0MiB / 16160MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
Creating training points: 100%|████████████████| 442/442 [01:20<00:00,  5.47it/s]
Creating evaluation points: 100%|████████████████| 48/48 [00:09<00:00,  4.98it/s]
86136 training points created.
10331 evaluation points created.
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.119.03   Driver Version: 450.119.03   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   47C    P0   240W / 300W |   5361MiB / 16160MiB |     83%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     13322      C   python                           5359MiB |
+-----------------------------------------------------------------------------+
```

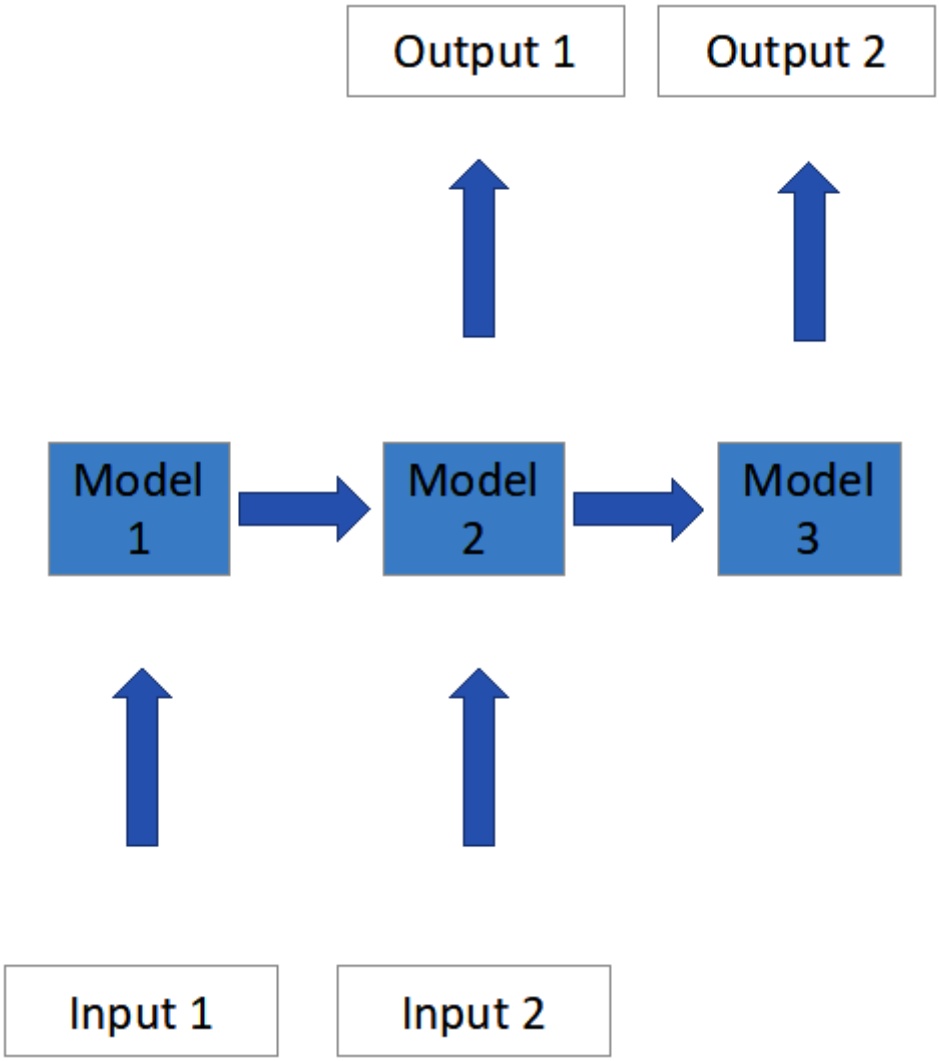
```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.119.03   Driver Version: 450.119.03   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:1E.0 Off |                    0 |
| N/A   57C    P0   118W / 300W |   3727MiB / 16160MiB |     65%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     14498      C   python                           3725MiB |
+-----------------------------------------------------------------------------+
```

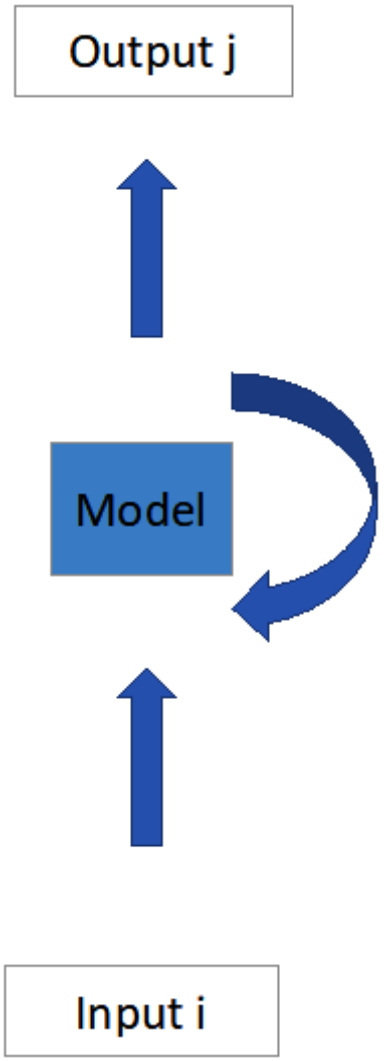


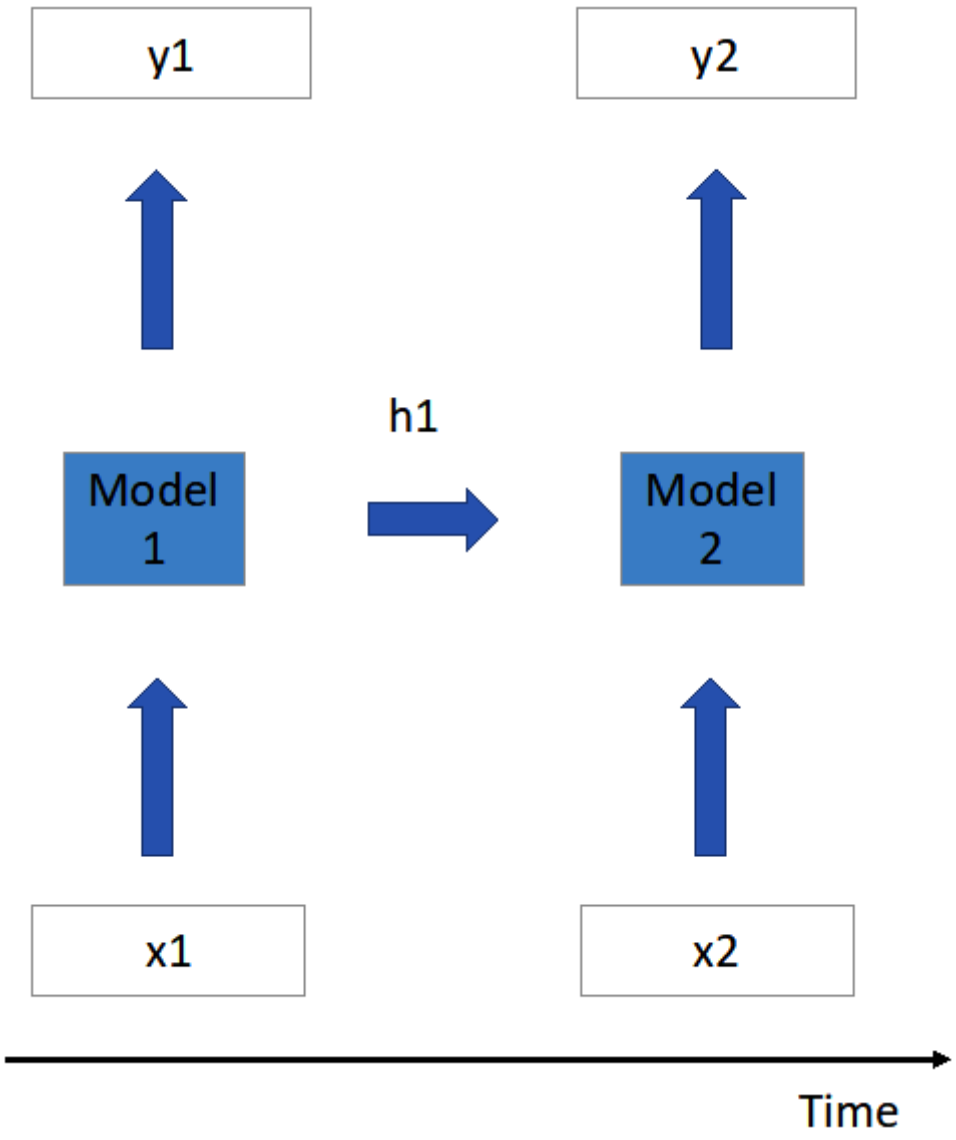

Output 1
Output 2
Output 3
Model 1
Model 2
Model 3
Input 1

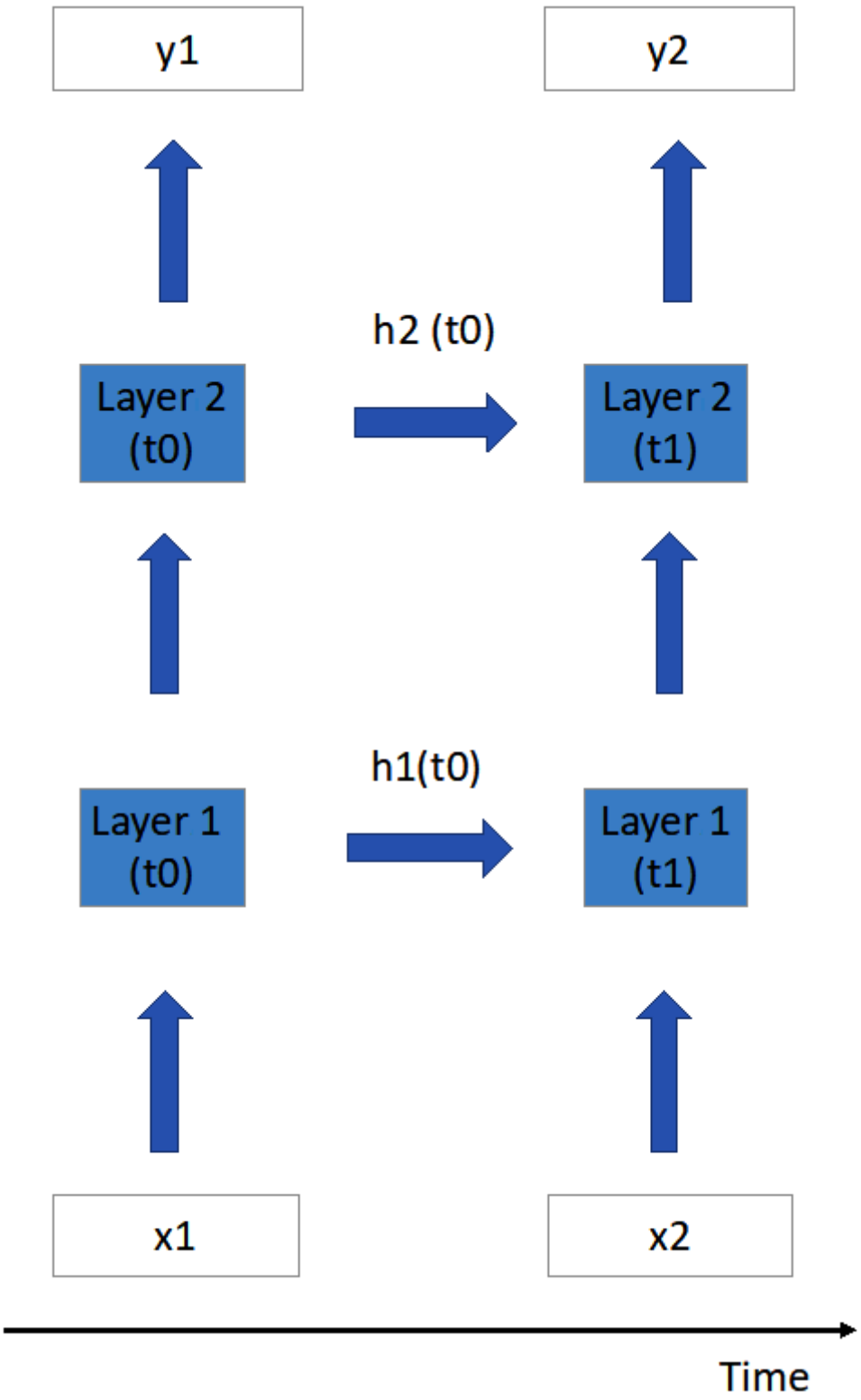

Output 1
Model 1
Model 2
Model 3
Input 1
Input 2
Input 3

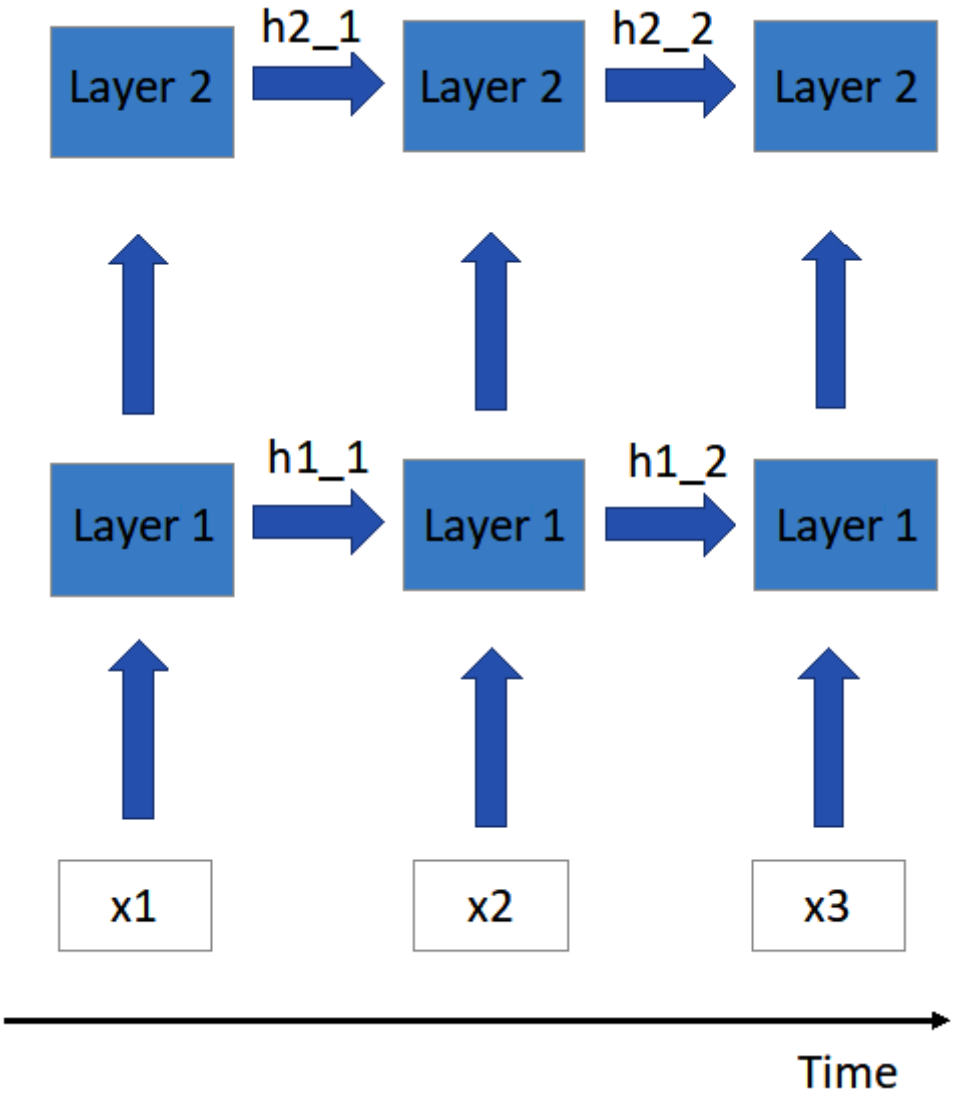

Output 1
Output 2
Output 3
Model 1
Model 2
Model 3
Input 1
Input 2
Input 3

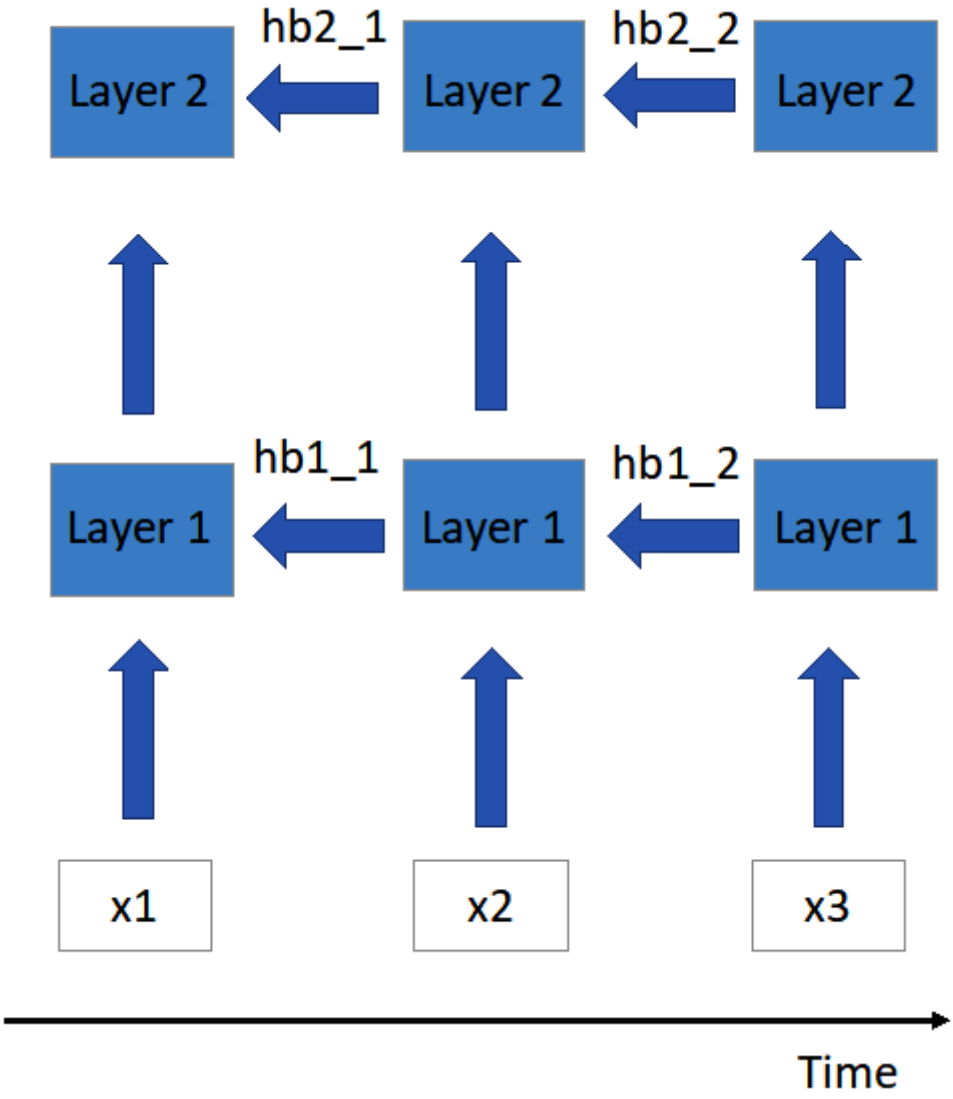

Output 1
Output 2
Model 1
Model 2
Model 3
Input 1
Input 2


Output j
Model
Input i

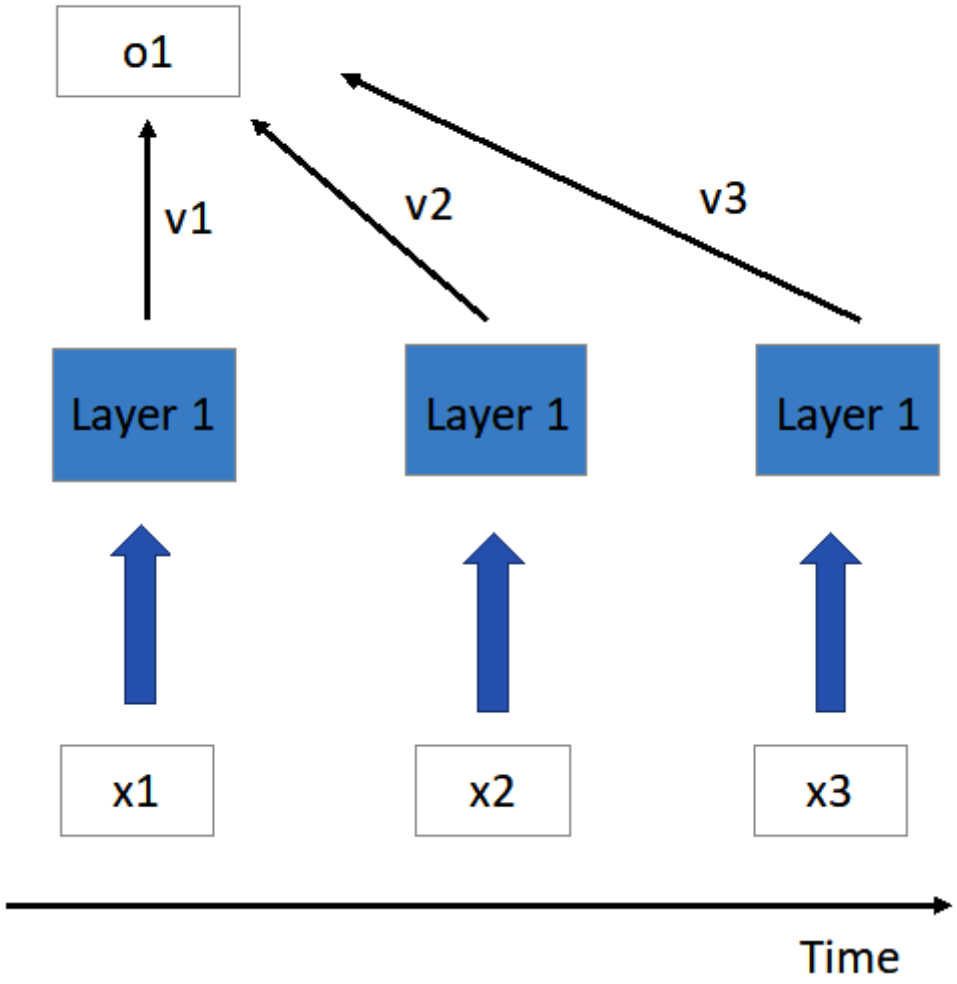

y1
y2
h1
Model 1
Model 2
x1
x2
Time

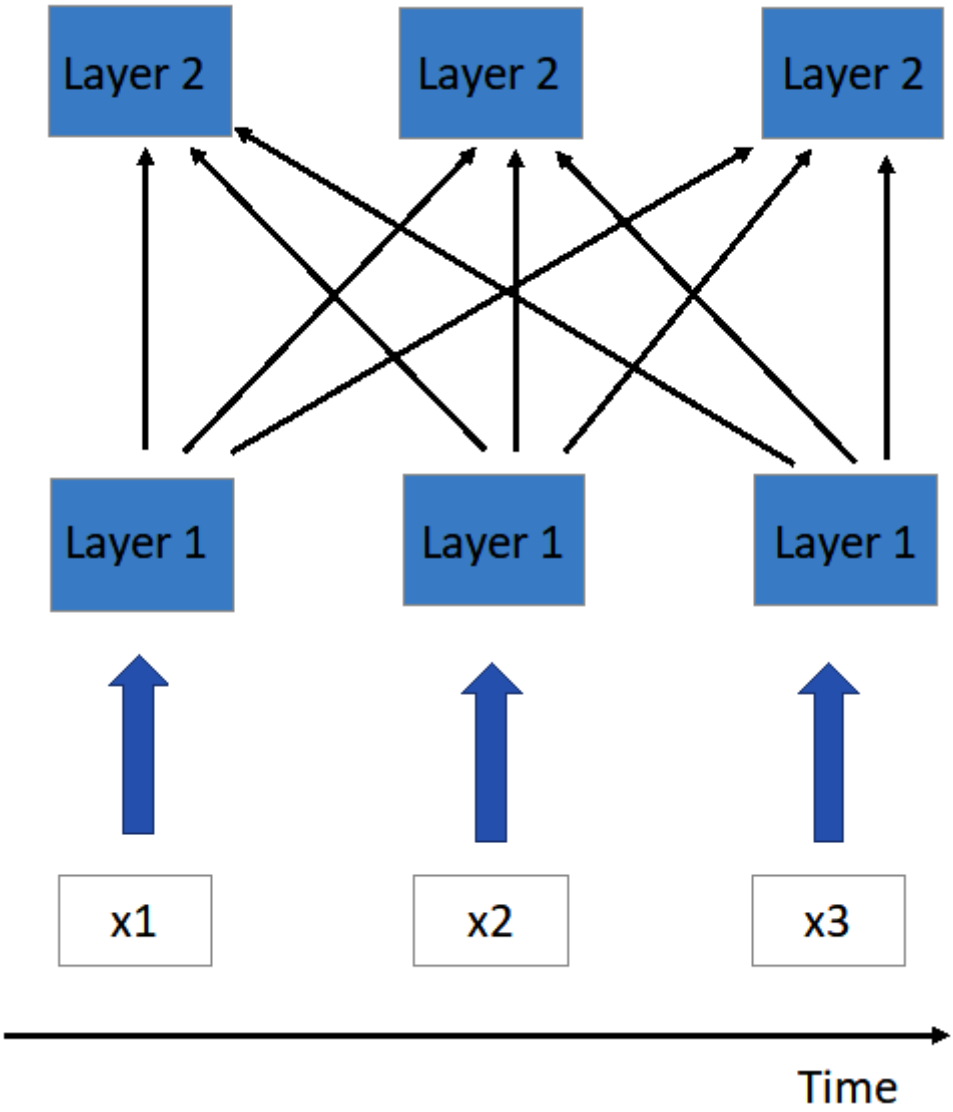

y1
y2
h2 (t0)
Layer 2 (t0)
Layer 2 (t1)
h1(t0)
Layer 1 (t0)
Layer 1 (t1)
x1
x2
Time

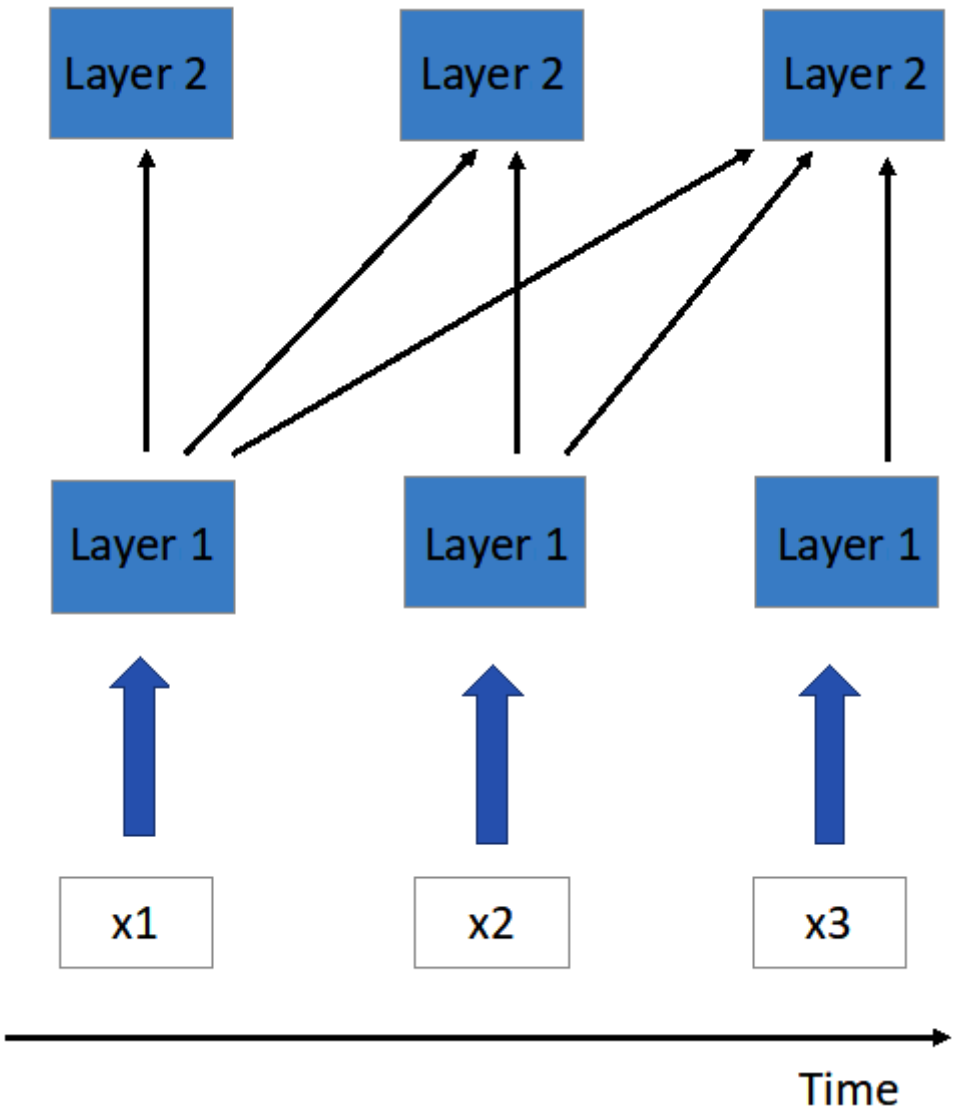

Layer 2
h2_1
Layer 2
h2_2
Layer 2
Layer 1
h1_1
Layer 1
h1_2
Layer 1
x1
x2
x3
Time

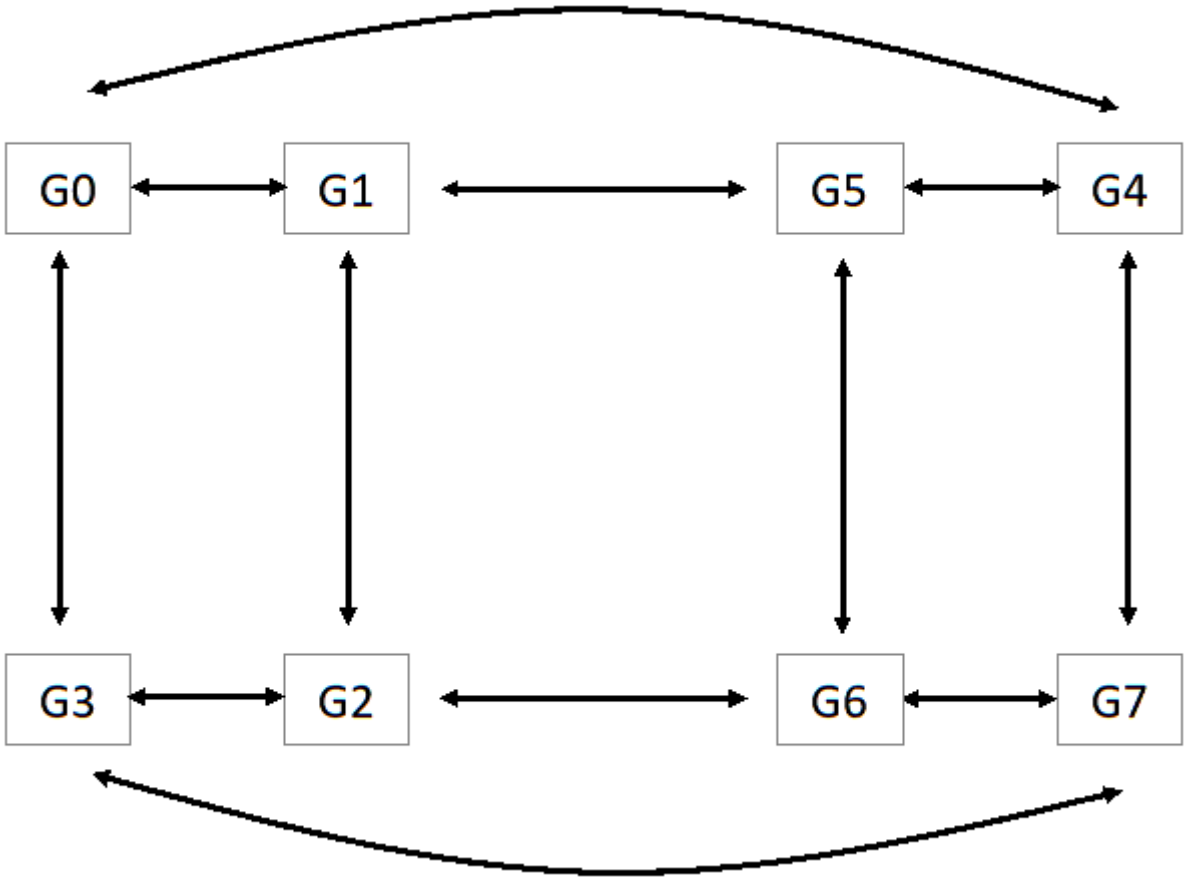

hb2_1
hb2_2
Layer 2
Layer 2
Layer 2
hb1_1
hb1_2
Layer 1
Layer 1
Layer 1
x1
x2
x3
Time

$$o1 = w1*v1 + w2*v2 + w3*v3$$

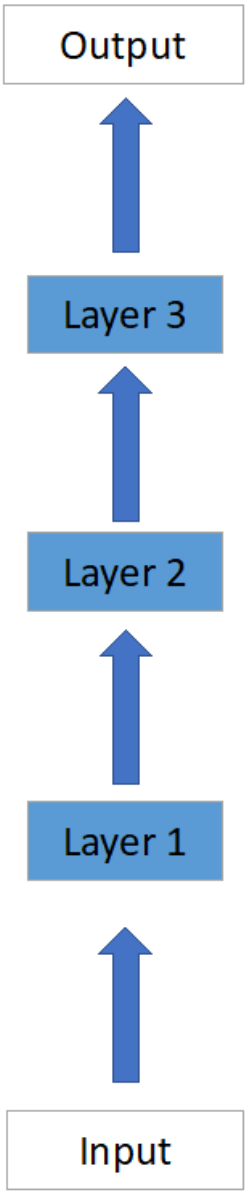


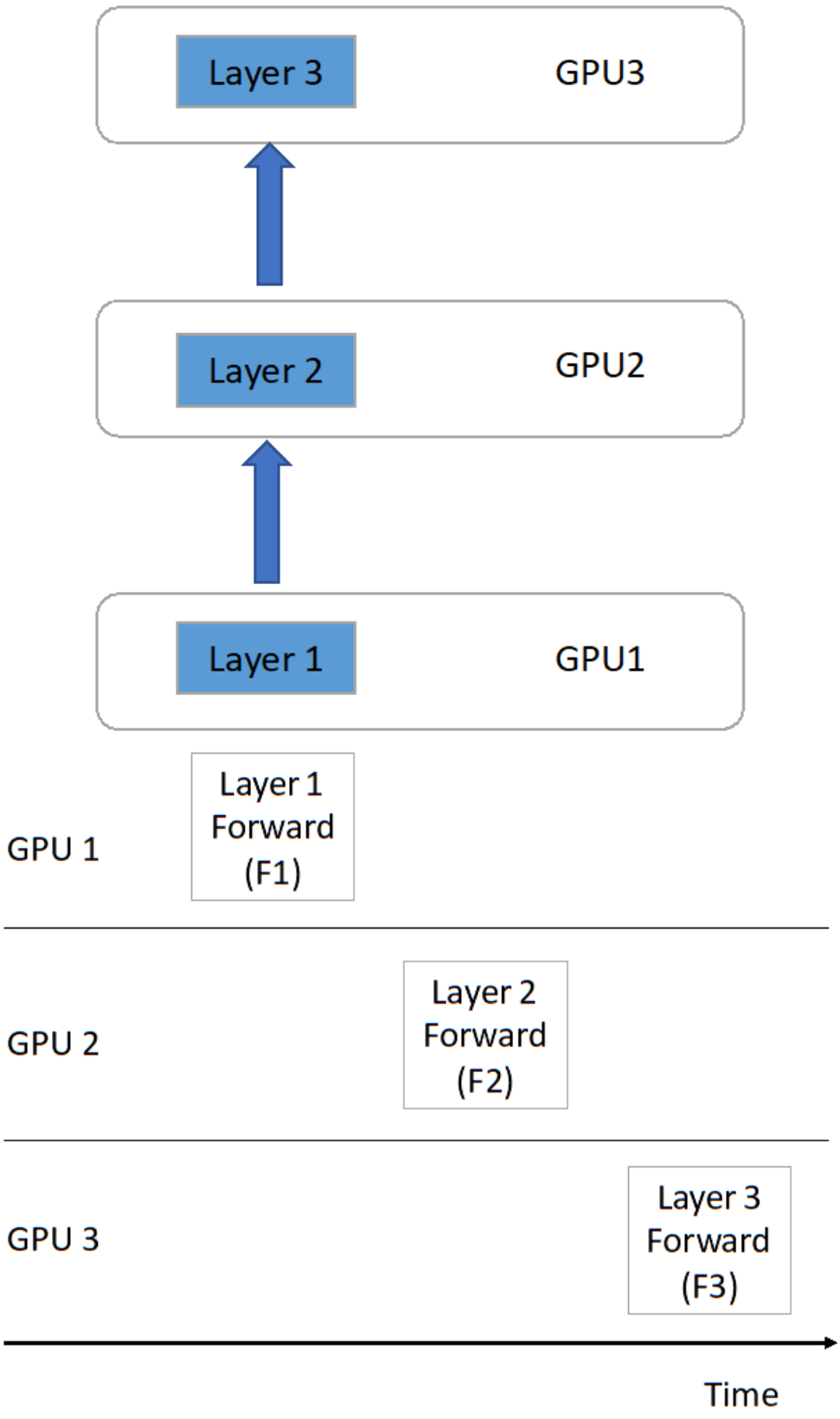

Layer 2
Layer 2
Layer 2
Layer 1
Layer 1
Layer 1
x1
x2
x3
Time

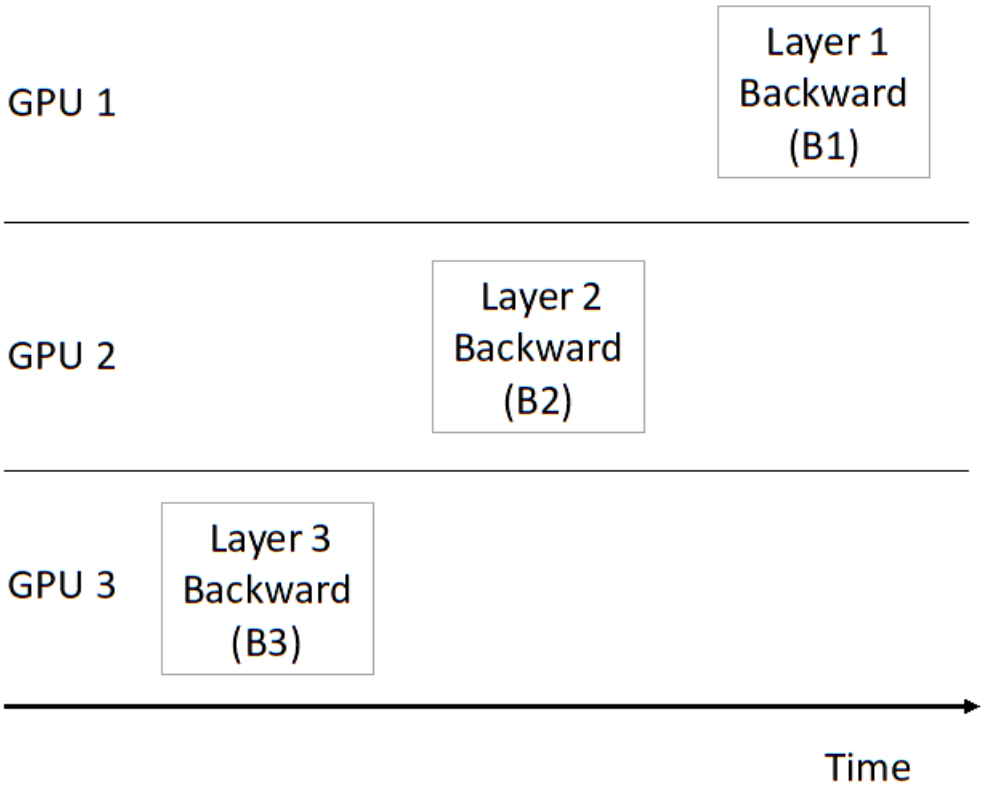

Layer 2
Layer 2
Layer 2
Layer 1
Layer 1
Layer 1
x1
x2
x3
Time

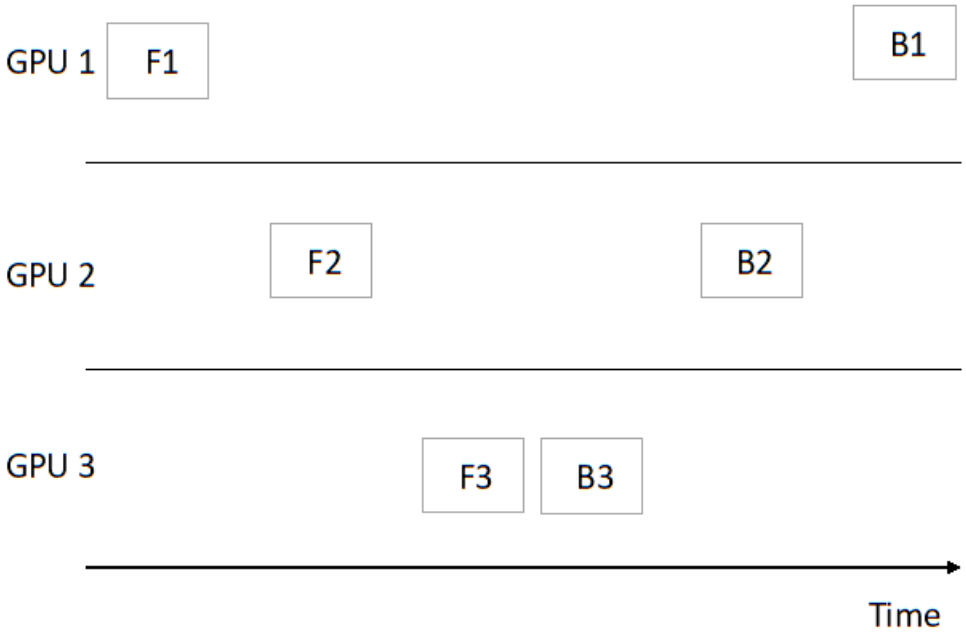

G0
G1
G5
G4
G3
G2
G6
G7

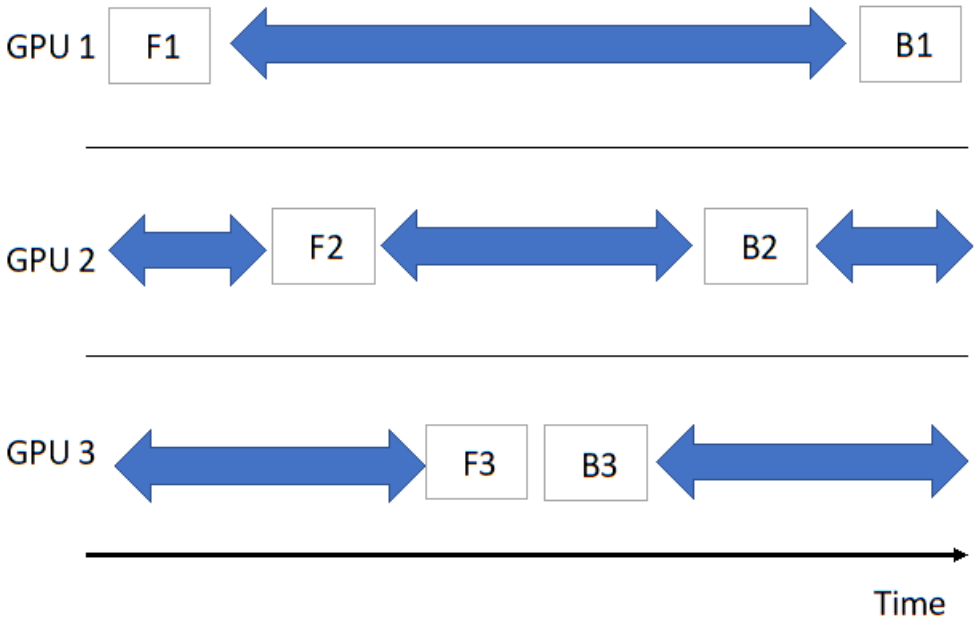
Chapter 6: Pipeline Input and Layer Split

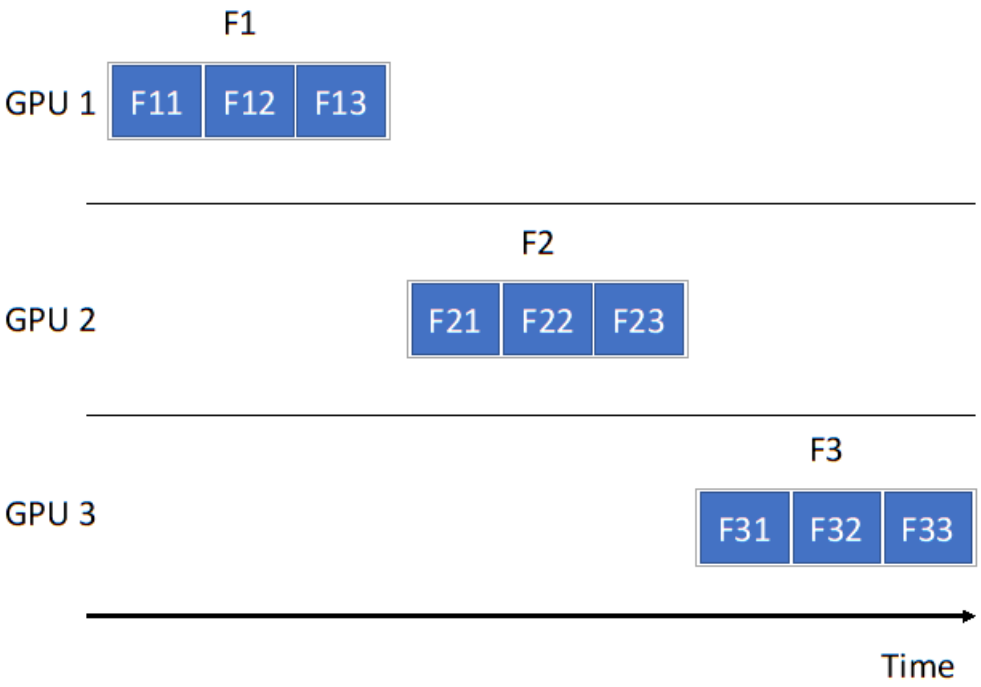


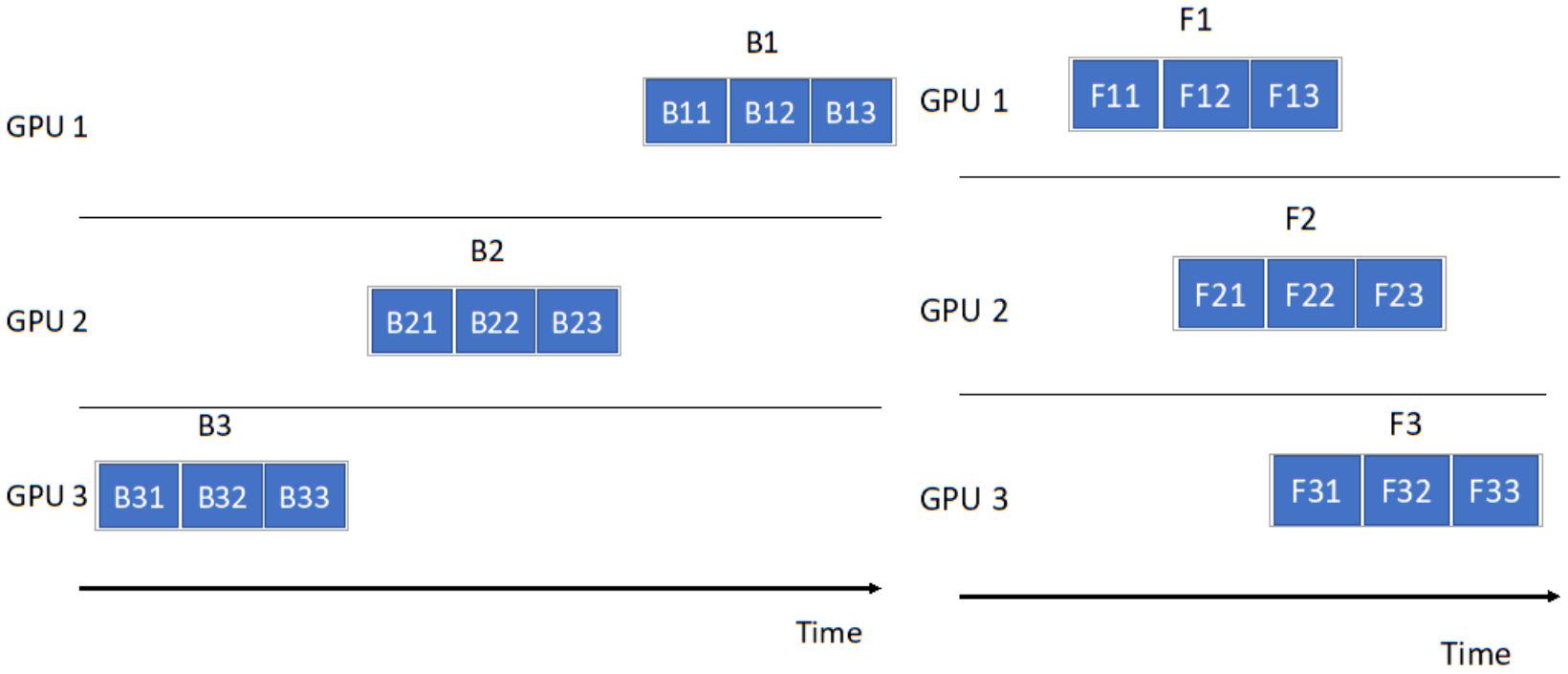

Layer 3
GPU3
Layer 2
GPU2
Layer 1
GPU1
GPU 1
Layer 1 Forward (F1)
GPU 2
Layer 2 Forward (F2)
GPU 3
Layer 3 Forward (F3)
Time

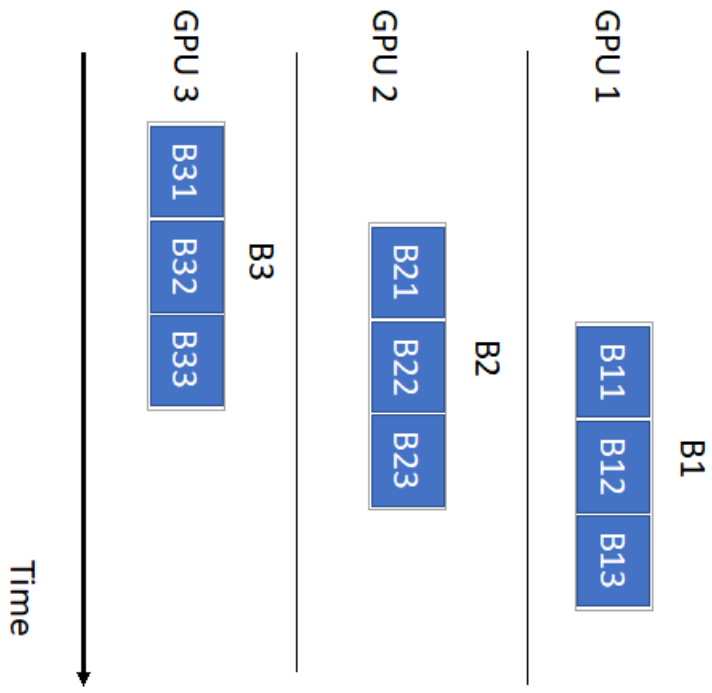

GPU 1
Layer 1
Backward
(B1)
GPU 2
Layer 2
Backward
(B2)
GPU 3
Layer 3
Backward
(B3)
Time


GPU 1
F1
B1
GPU 2
F2
B2
GPU 3
F3
B3
Time


GPU 1
F1
B1
GPU 2
F2
B2
GPU 3
F3
B3
Time


F1
GPU 1
F11
F12
F13
F2
GPU 2
F21
F22
F23
F3
GPU 3
F31
F32
F33
Time


B1
GPU 1
B11
B12
B13
B2
GPU 2
B21
B22
B23
B3
GPU 3
B31
B32
B33
Time
F1
GPU 1
F11
F12
F13
F2
GPU 2
F21
F22
F23
F3
GPU 3
F31
F32
F33
Time



Layer 1 Matrix

w(0,0)	w(1,0)	w(2,0)	w(3,0)
w(0,1)	w(1,1)	w(2,1)	w(3,1)
w(0,2)	w(1,2)	w(2,2)	w(3,2)
w(0,3)	w(1,3)	w(2,3)	w(3,3)

Input Matrix (batch size = 4)

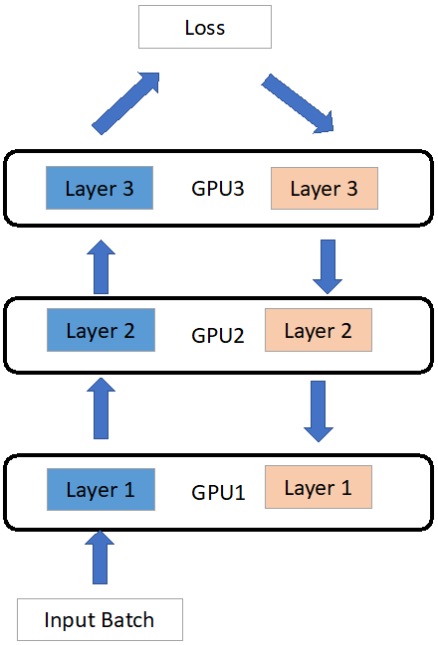
x(0,0)	x(0,1)	w(0,2)	w(0,3)
x(1,0)	x(1,1)	x(1,2)	x(1,3)
x(2,0)	x(2,1)	x(2,2)	x(2,3)
x(3,0)	x(3,1)	x(3,2)	x(3,3)

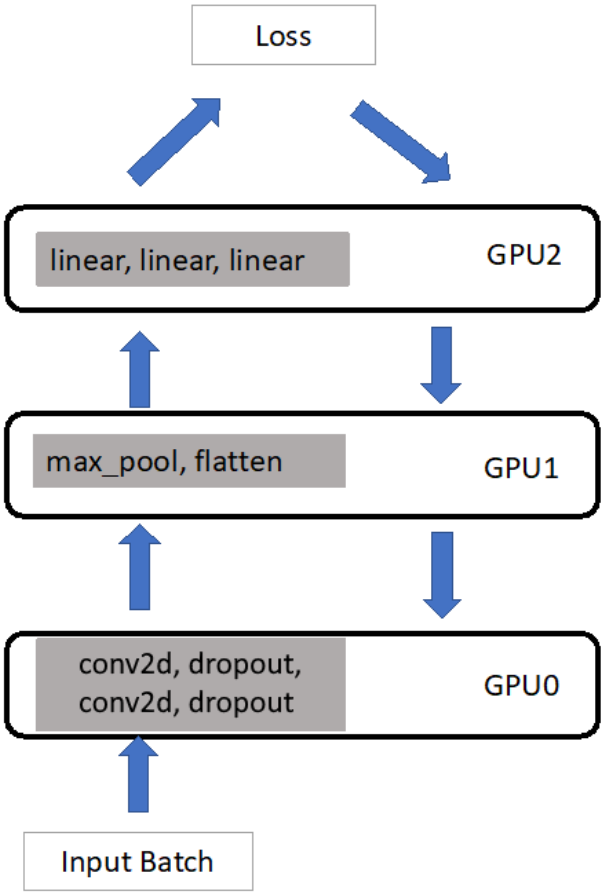
Layer 1 Matrix Splits
(Column-wise)

w(0,0)	w(1,0)	w(2,0)	w(3,0)
w(0,1)	w(1,1)	w(2,1)	w(3,1)
w(0,2)	w(1,2)	w(2,2)	w(3,2)
w(0,3)	w(1,3)	w(2,3)	w(3,3)

A_01 A_23

Chapter 7: Implementing Model Parallel Training and Serving Workflows

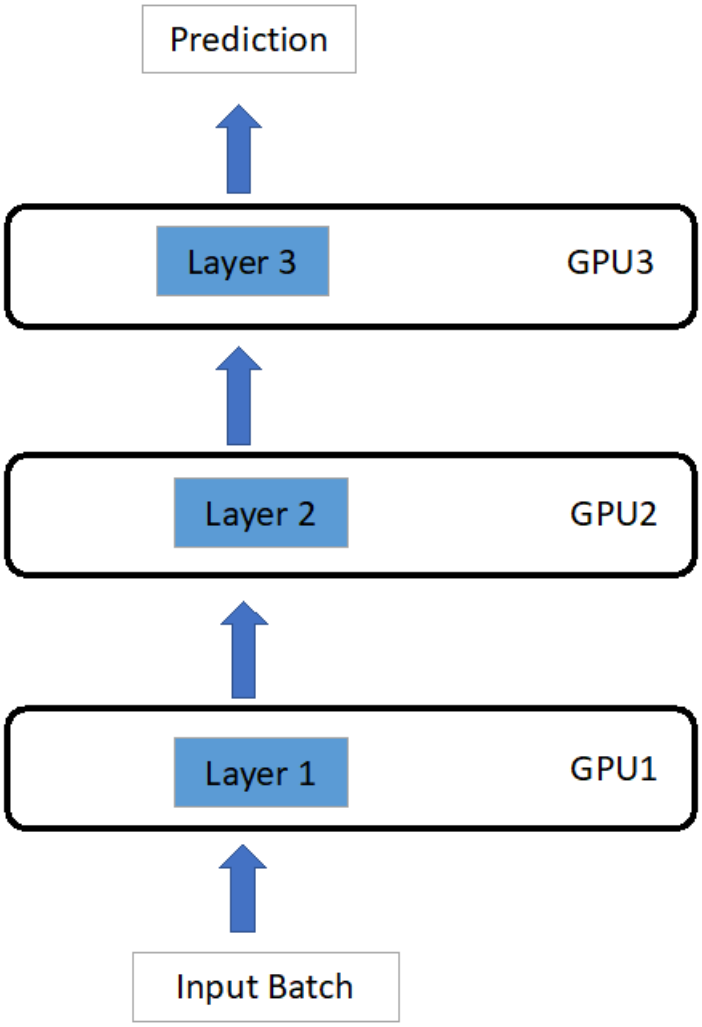





```python
import torch
import torch.nn as nn
import torch.nn.functional as F

class MyNet(nn.Module):
    def __init__(self):
        super(MyNet, self).__init__()
        self.seq1 = nn.Sequential(
                        nn.Conv2d(1,32,3,1),
                        nn.Dropout2d(0.5),
                        nn.Conv2d(32,64,3,1),
                        nn.Dropout2d(0.75)).to('cuda:0')
        self.seq2 = nn.Sequential(
                        nn.Linear(9216, 128),
                        nn.Linear(128,20),
                        nn.Linear(20,10)).to('cuda:2')

    def forward(self, x):
        x = self.seq1(x.to('cuda:0'))
        x = F.max_pool2d(x,2).to('cuda:1')
        x = torch.flatten(x,1).to('cuda:1')
        x = self.seq2(x.to('cuda:2'))
        output = F.log_softmax(x, dim = 1)
        return output
```

```
Epoch 0
batch 0 training :: loss 2.367696523666382
batch 1 training :: loss 2.358067274093628
batch 2 training :: loss 2.3166911602020264
batch 3 training :: loss 2.3472657203674316
batch 4 training :: loss 2.3291213512420654
batch 5 training :: loss 2.341862201690674
batch 6 training :: loss 2.3476767539978027
batch 7 training :: loss 2.3589253425598145
batch 8 training :: loss 2.3385939598083496
batch 9 training :: loss 2.314199209213257
batch 10 training :: loss 2.357100486755371
batch 11 training :: loss 2.341332197189331
batch 12 training :: loss 2.3510727882385254
batch 13 training :: loss 2.305490732192993
batch 14 training :: loss 2.2896692752838135
batch 15 training :: loss 2.2965853214263916
batch 16 training :: loss 2.289027452468872
batch 17 training :: loss 2.318589687347412
batch 18 training :: loss 2.314786911010742
batch 19 training :: loss 2.292377471923828
batch 20 training :: loss 2.311783790588379
batch 21 training :: loss 2.3006303310394287
batch 22 training :: loss 2.2897908687591553
batch 23 training :: loss 2.309767246246338
batch 24 training :: loss 2.326434373855591
batch 25 training :: loss 2.3054540157318115
batch 26 training :: loss 2.3287947177886963
batch 27 training :: loss 2.309558391571045
batch 28 training :: loss 2.289318084716797
batch 29 training :: loss 2.3383259773254395
batch 30 training :: loss 2.2959561347961426
batch 31 training :: loss 2.2574143409729004
batch 32 training :: loss 2.293168783187866
batch 33 training :: loss 2.2815194129943848
batch 34 training :: loss 2.2899670600891113
batch 35 training :: loss 2.2440366744995117
batch 36 training :: loss 2.2733407020568848
batch 37 training :: loss 2.2578611373901367
batch 38 training :: loss 2.2523033618927
batch 39 training :: loss 2.2961771488189697
batch 40 training :: loss 2.269951820373535
batch 450 training :: loss 0.4500477612018585
batch 451 training :: loss 0.41694992780685425
batch 452 training :: loss 0.5335432291030884
batch 453 training :: loss 0.3785797357559204
batch 454 training :: loss 0.5250097513198853
batch 455 training :: loss 0.48590853810310364
batch 456 training :: loss 0.4359087646007538
batch 457 training :: loss 0.5516181588172913
batch 458 training :: loss 0.4193853735923767
batch 459 training :: loss 0.24893827736377716
batch 460 training :: loss 0.4412848949432373
batch 461 training :: loss 0.6855418086051941
batch 462 training :: loss 0.60863196849823
batch 463 training :: loss 0.6327939629554749
batch 464 training :: loss 0.4109138548374176
batch 465 training :: loss 0.3921489715576172
batch 466 training :: loss 0.3610058128833771
batch 467 training :: loss 0.4983370304107666
batch 468 training :: loss 0.497358113527298
Training Done!
```

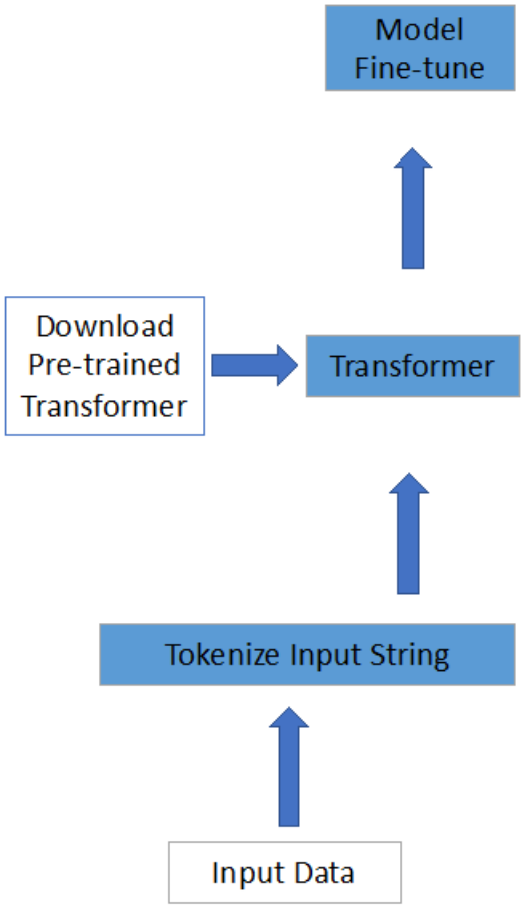
```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.142.00   Driver Version: 450.142.00   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla M60           On   | 00000000:00:1B.0 Off |                    0 |
| N/A   32C    P0    82W / 150W |   1008MiB /  7618MiB |     69%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M60           On   | 00000000:00:1C.0 Off |                    0 |
| N/A   25C    P0    38W / 150W |    708MiB /  7618MiB |     18%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla M60           On   | 00000000:00:1D.0 Off |                    0 |
| N/A   29C    P0    39W / 150W |    772MiB /  7618MiB |     21%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla M60           On   | 00000000:00:1E.0 Off |                    0 |
| N/A   23C    P8    14W / 150W |      3MiB /  7618MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A       722      C   python                           1005MiB |
|    1   N/A  N/A       722      C   python                            705MiB |
|    2   N/A  N/A       722      C   python                            769MiB |
+-----------------------------------------------------------------------------+
```




```
Test Accuracy 0.0019666666666666665
Test Accuracy 0.0037333333333333333
Test Accuracy 0.005683333333333334
Test Accuracy 0.0076
Test Accuracy 0.009483333333333333
Test Accuracy 0.0114
Test Accuracy 0.013216666666666666
Test Accuracy 0.015066666666666667
Test Accuracy 0.0168
Test Accuracy 0.01875
Test Accuracy 0.020566666666666667
Test Accuracy 0.0225
Test Accuracy 0.0243
Test Accuracy 0.026333333333333334
```

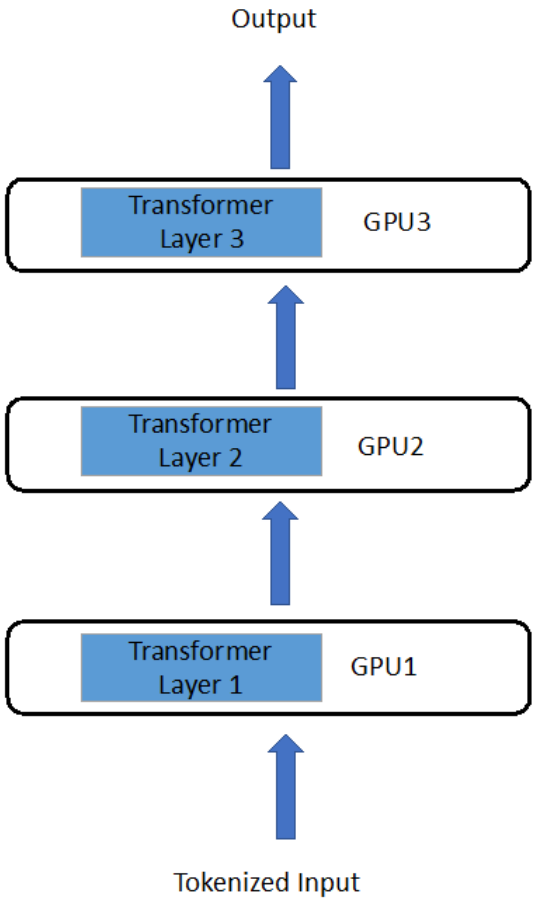
```
Test Accuracy 0.8768166666666667
Test Accuracy 0.8786333333333334
Test Accuracy 0.88055
Test Accuracy 0.8826333333333334
Test Accuracy 0.8844666666666666
Test Accuracy 0.8865666666666666
Test Accuracy 0.88795
Test Done!
```

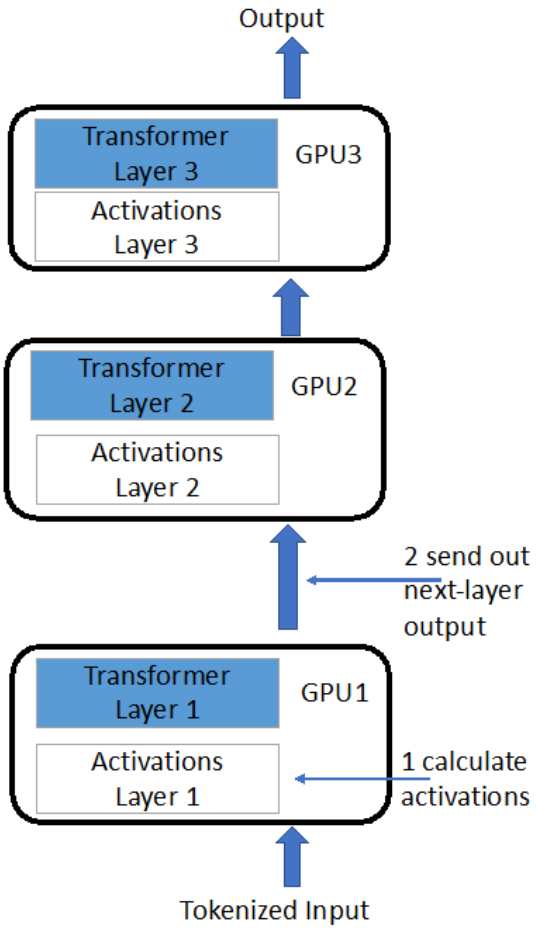
```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.142.00   Driver Version: 450.142.00   CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla M60           On   | 00000000:00:1B.0 Off |                    0 |
| N/A   38C    P0    78W / 150W |   1008MiB /  7618MiB |     42%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M60           On   | 00000000:00:1C.0 Off |                    0 |
| N/A   27C    P0    38W / 150W |    708MiB /  7618MiB |     20%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla M60           On   | 00000000:00:1D.0 Off |                    0 |
| N/A   30C    P0    39W / 150W |    772MiB /  7618MiB |     21%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla M60           On   | 00000000:00:1E.0 Off |                    0 |
| N/A   23C    P8    14W / 150W |      3MiB /  7618MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     10485      C   python                           1005MiB |
|    1   N/A  N/A     10485      C   python                            705MiB |
|    2   N/A  N/A     10485      C   python                            769MiB |
+-----------------------------------------------------------------------------+
```

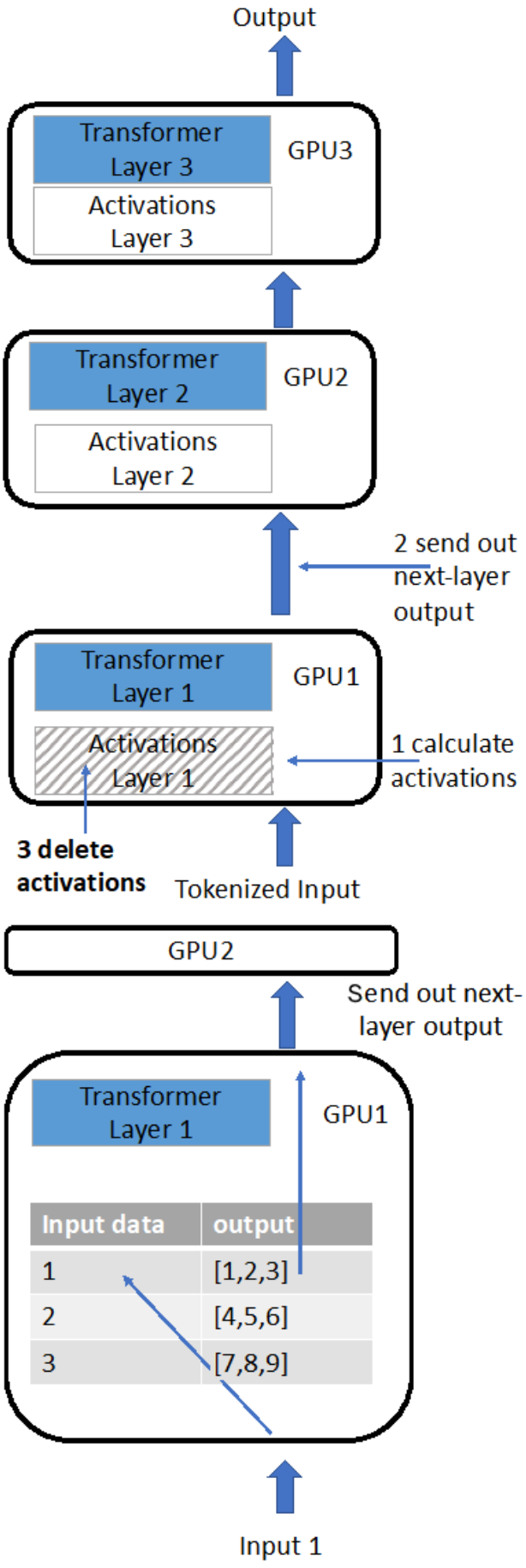


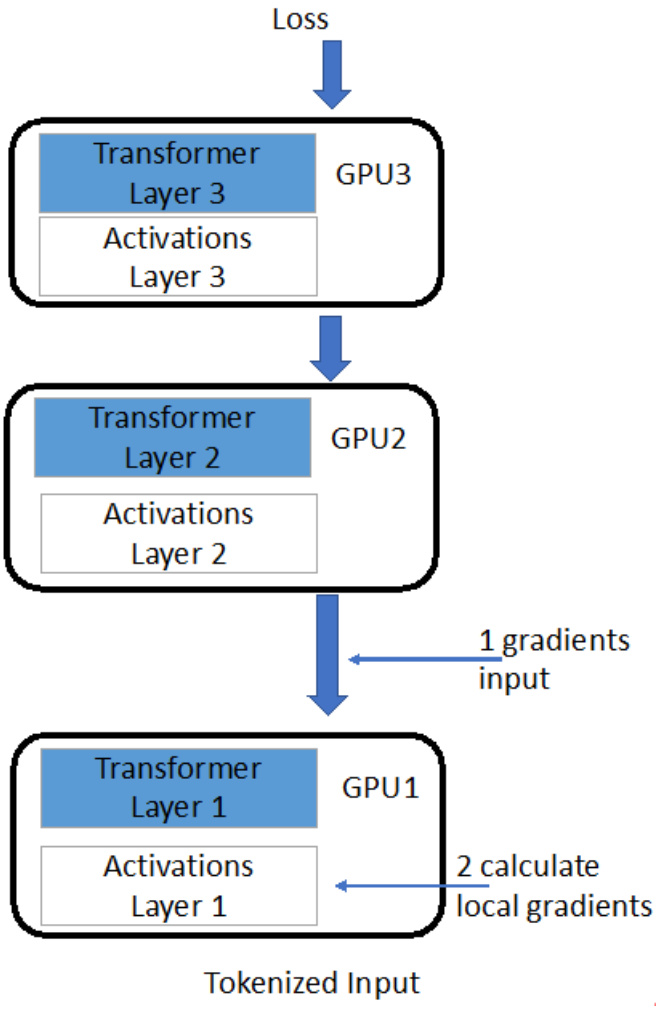

```
+-----------------------------------------+------------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla M60           On   | 00000000:00:1B.0 Off |                    0 |
| N/A   38C    P0    78W / 150W |   1008MiB /  7618MiB |     42%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M60           On   | 00000000:00:1C.0 Off |                    0 |
| N/A   27C    P0    38W / 150W |    708MiB /  7618MiB |     20%      Default |
|                               |                      |                  N/A |
```

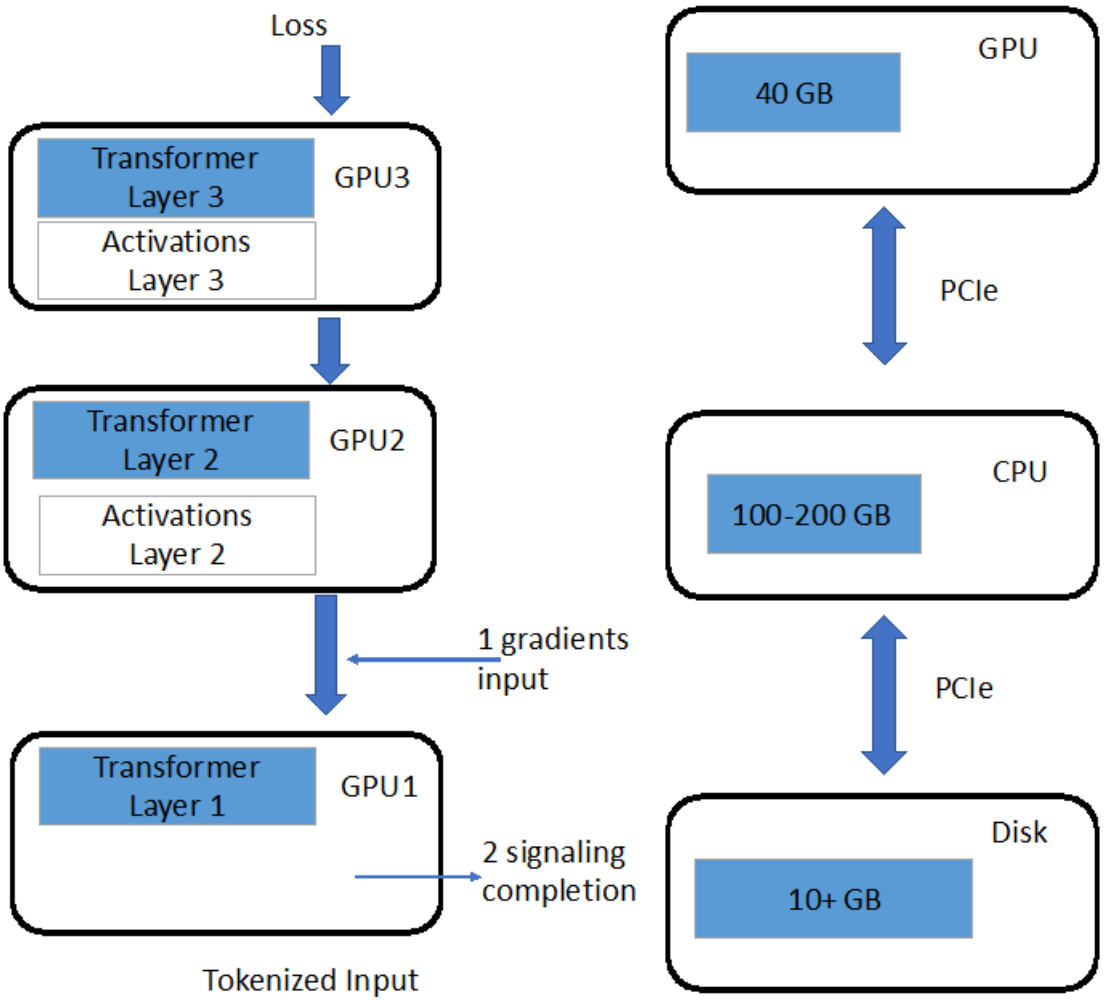
Chapter 8: Achieving Higher Throughput and Lower Latency

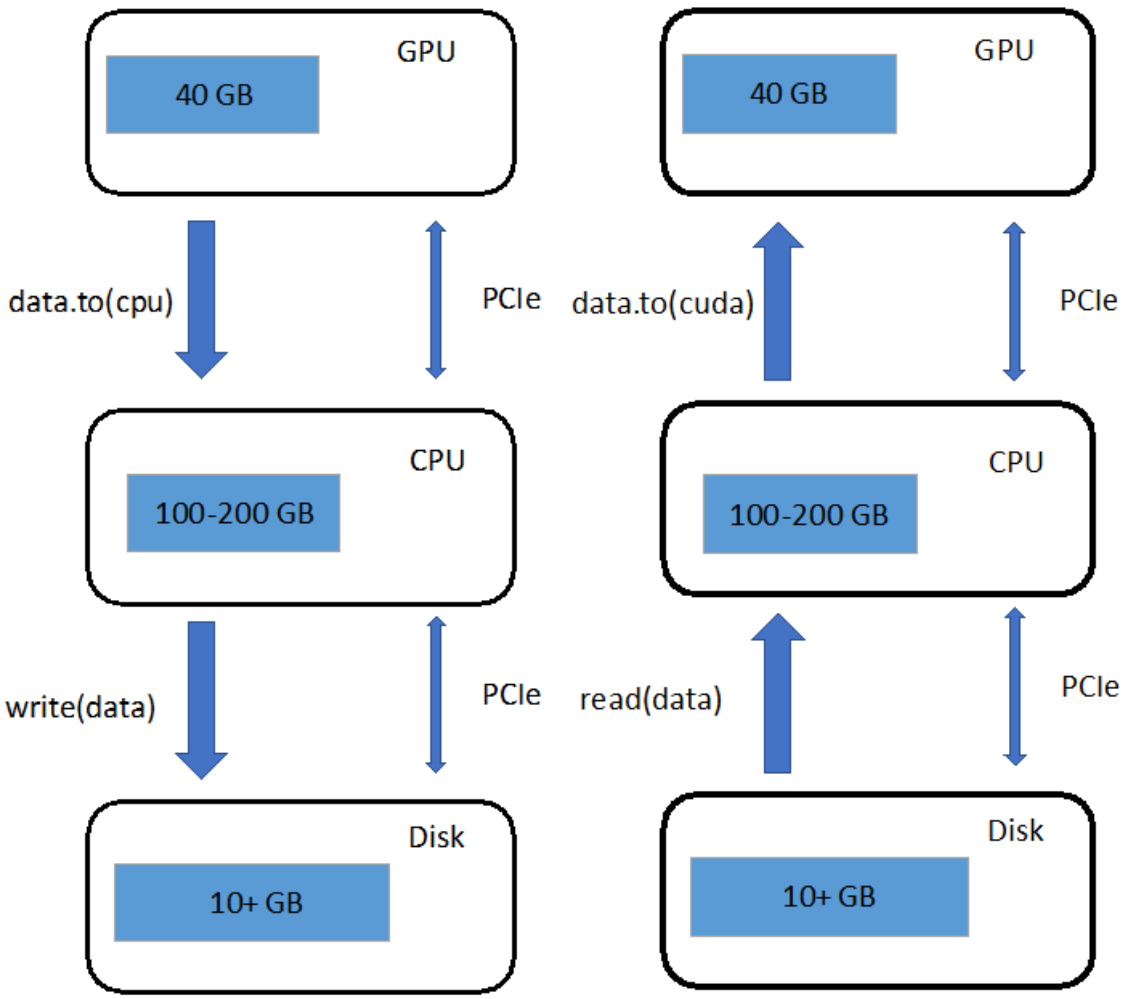


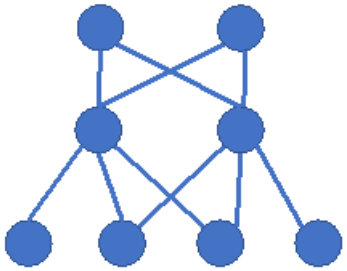

Output
Transformer Layer 3
GPU3
Activations Layer 3
Transformer Layer 2
GPU2
Activations Layer 2
2 send out next-layer output
Transformer Layer 1
GPU1
Activations Layer 1
1 calculate activations
Tokenized Input


Output
Transformer Layer 3
GPU3
Activations Layer 3
Transformer Layer 2
GPU2
Activations Layer 2
2 send out next-layer output
Transformer Layer 1
GPU1
Activations Layer 1
1 calculate activations
3 delete activations
Tokenized Input
GPU2
Send out next-layer output
Transformer Layer 1
GPU1
Input data
output
1
[1,2,3]
2
[4,5,6]
3
[7,8,9]
Input 1

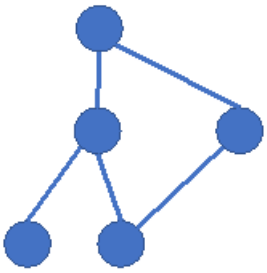

Loss
Transformer Layer 3
GPU3
Activations Layer 3
Transformer Layer 2
GPU2
Activations Layer 2
1 gradients input
Transformer Layer 1
GPU1
Activations Layer 1
2 calculate local gradients
Tokenized Input


Loss
Transformer Layer 3
GPU3
Activations Layer 3
Transformer Layer 2
GPU2
Activations Layer 2
1 gradients input
Transformer Layer 1
GPU1
2 signaling completion
Tokenized Input
GPU
40 GB
PCIe
CPU
100-200 GB
PCIe
Disk
10+ GB

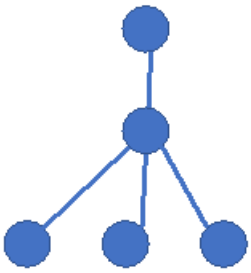

GPU
40 GB
data.to(cpu)
PCIe
CPU
100-200 GB
write(data)
PCIe
Disk
10+ GB
GPU
40 GB
data.to(cuda)
PCIe
CPU
100-200 GB
read(data)
PCIe
Disk
10+ GB



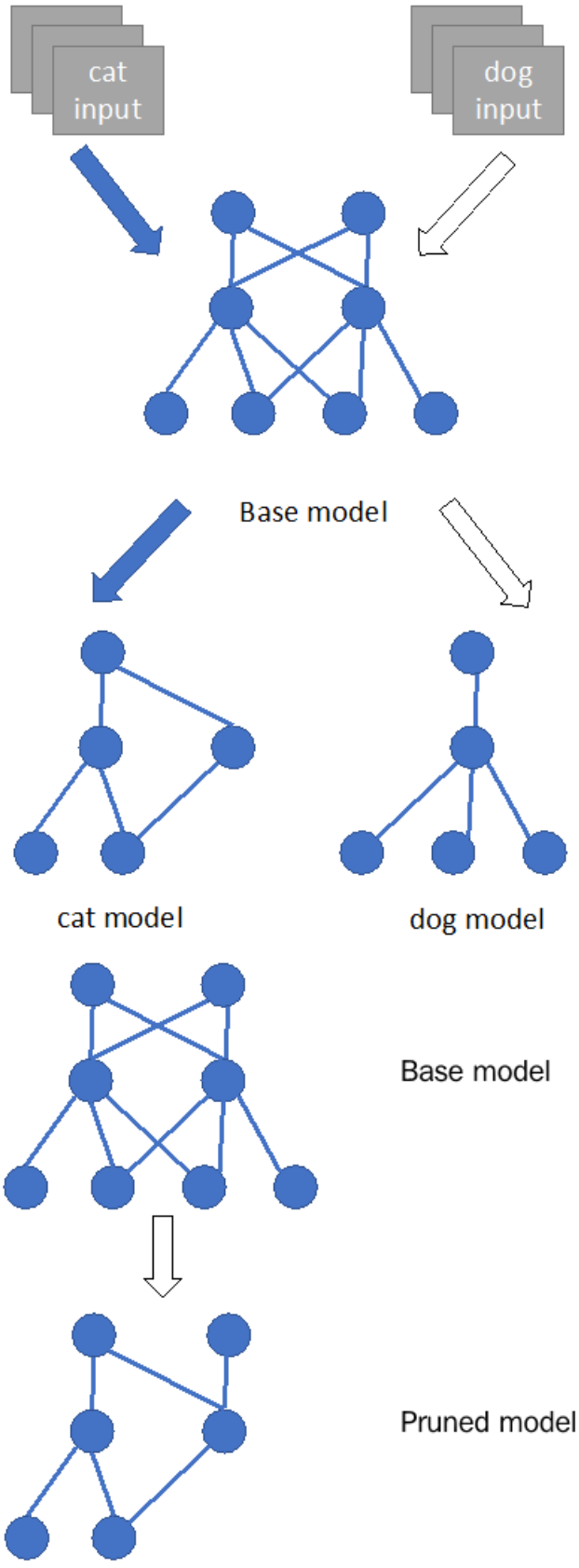
Base model

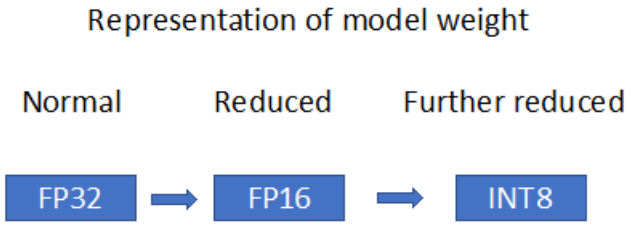


cat model



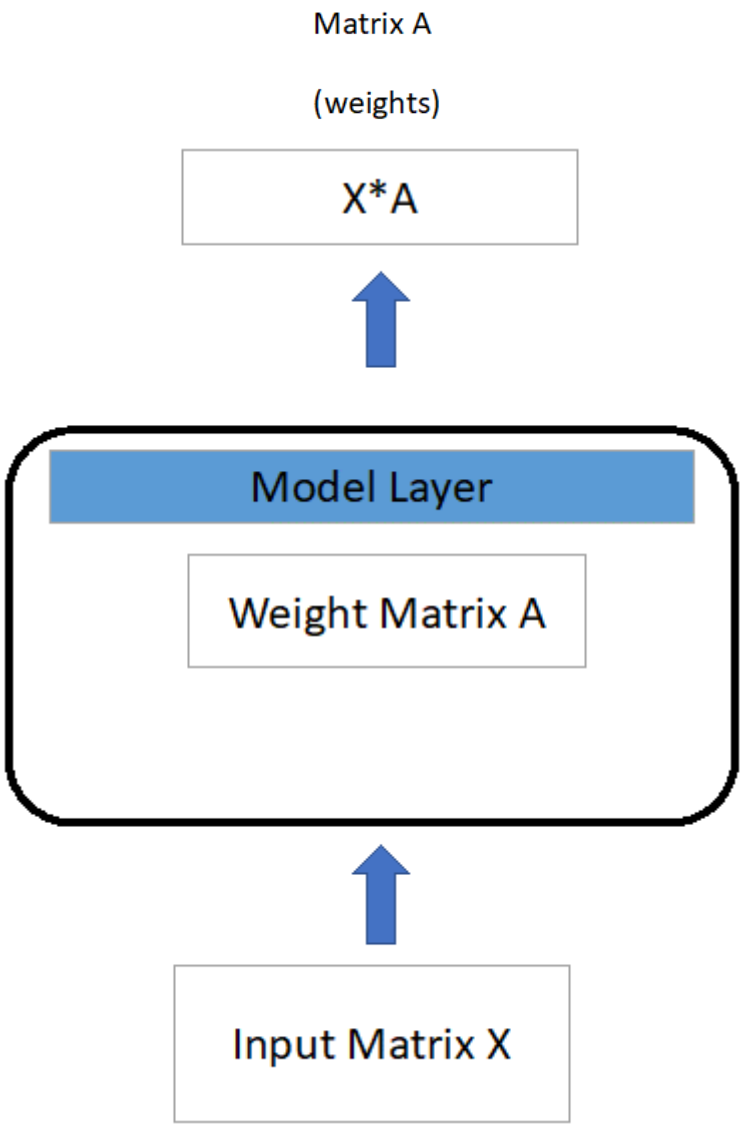
dog model


cat
input
dog
input
Base model
cat model
dog model
Base model
Pruned model


Representation of model weight
Normal
Reduced
Further reduced
FP32
FP16
INT8

Chapter 9: A Hybrid of Data and Model Parallelism

A(0,0)	A(0,1)	A(0,2)	A(0,3)
A(1,0)	A(1,1)	A(1,2)	A(1,3)
A(2,0)	A(2,1)	A(2,2)	A(2,3)
A(3,0)	A(3,1)	A(3,2)	A(3,3)



X(0,0)	X(0,1)	X(0,2)	X(0,3)
X(1,0)	X(1,1)	X(1,2)	X(1,3)
X(2,0)	X(2,1)	X(2,2)	X(2,3)
X(3,0)	X(3,1)	X(3,2)	X(3,3)

Matrix X

(Input)

A(0,0)	A(0,1)	A(0,2)	A(0,3)
A(1,0)	A(1,1)	A(1,2)	A(1,3)

A[0]

A(2,0)	A(2,1)	A(2,2)	A(2,3)
A(3,0)	A(3,1)	A(3,2)	A(3,3)

A[1]

Matrix A
(weights)

X(0,0)	X(0,1)
X(1,0)	X(1,1)
X(2,0)	X(2,1)
X(3,0)	X(3,1)

X[0]

X(0,2)	X(0,3)
X(1,2)	X(1,3)
X(2,2)	X(2,3)
X(3,2)	X(3,3)

X[1]

Matrix X

(Input)

A(0,0)	A(0,1)	A(0,2)	A(0,3)
A(1,0)	A(1,1)	A(1,2)	A(1,3)

A[0]

X(0,0)	X(0,1)
X(1,0)	X(1,1)
X(2,0)	X(2,1)
X(3,0)	X(3,1)

X[0]

X[0]*A[0]

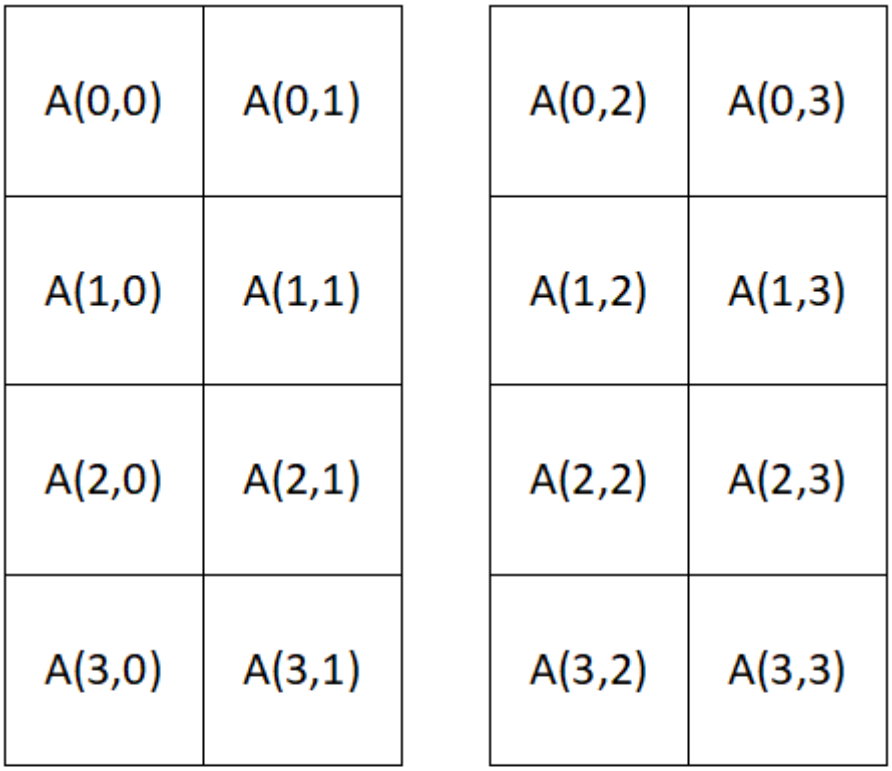
A(2,0)	A(2,1)	A(2,2)	A(2,3)
A(3,0)	A(3,1)	A(3,2)	A(3,3)

A[1]

X(0,2)	X(0,3)
X(1,2)	X(1,3)
X(2,2)	X(2,3)
X(3,2)	X(3,3)

X[1]*A[1]

X[1]



Matrix A

(weights)

X(0,0)	X(0,1)	X(0,2)	X(0,3)
X(1,0)	X(1,1)	X(1,2)	X(1,3)
X(2,0)	X(2,1)	X(2,2)	X(2,3)
X(3,0)	X(3,1)	X(3,2)	X(3,3)

X

A(0,0)	A(0,1)
A(1,0)	A(1,1)
A(2,0)	A(2,1)
A(3,0)	A(3,1)

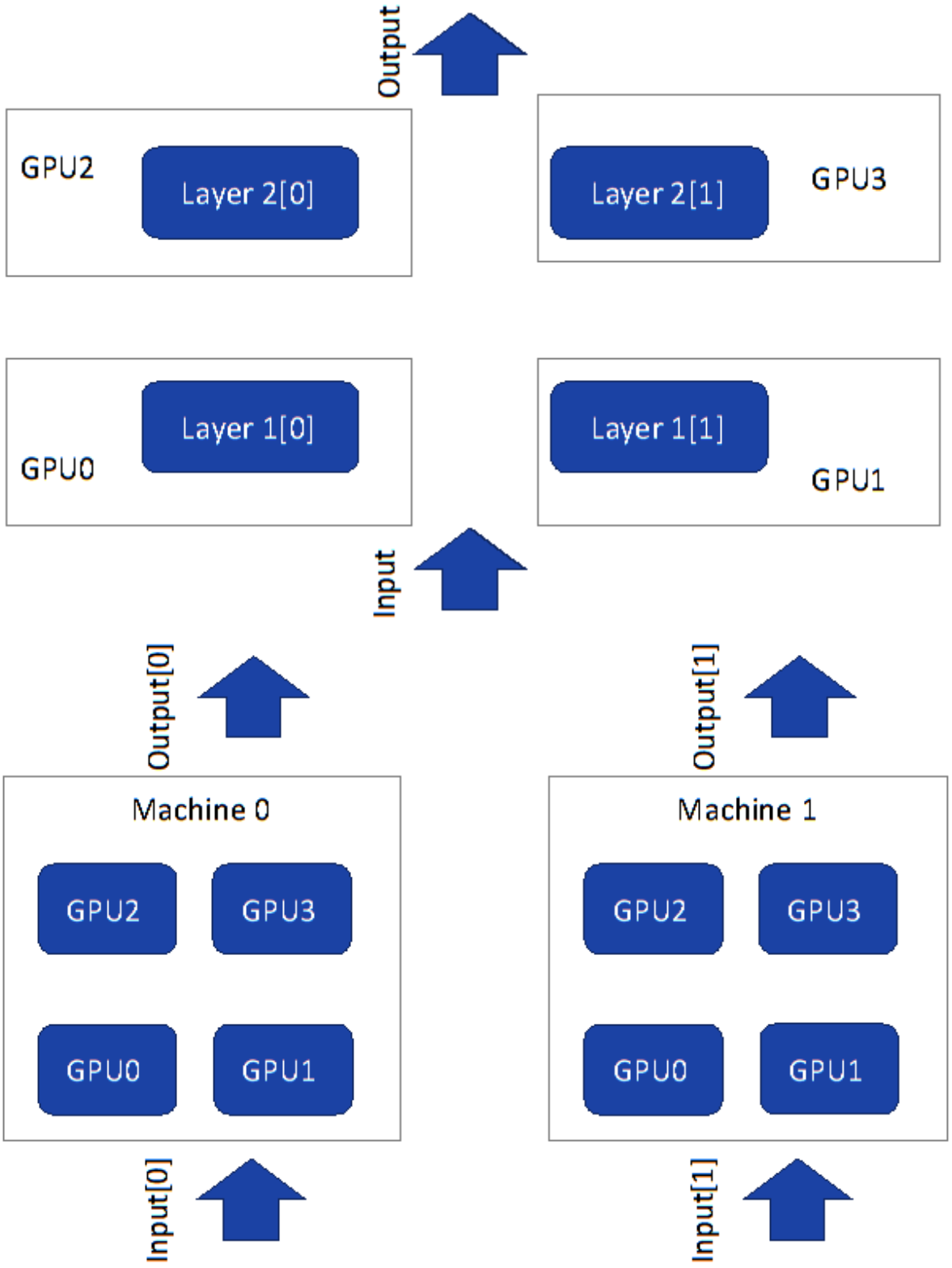
A’[0]

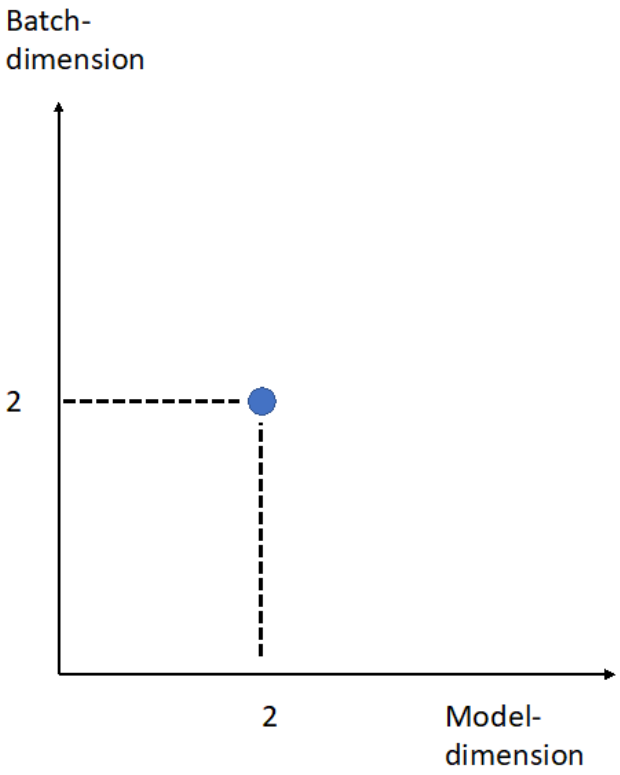
X(0,0)	X(0,1)	X(0,2)	X(0,3)
X(1,0)	X(1,1)	X(1,2)	X(1,3)
X(2,0)	X(2,1)	X(2,2)	X(2,3)
X(3,0)	X(3,1)	X(3,2)	X(3,3)

X

A(0,2)	A(0,3)
A(1,2)	A(1,3)
A(2,2)	A(2,3)
A(3,2)	A(3,3)

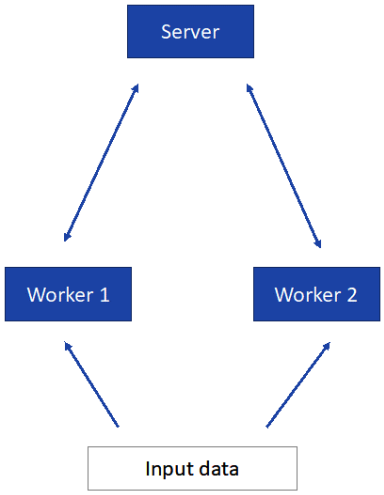
A’[1]


Output
GPU2
Layer 2[0]
Layer 2[1]
GPU3
GPU0
Layer 1[0]
Layer 1[1]
GPU1
Input
Output[0]
Output[1]
Machine 0
Machine 1
GPU2
GPU3
GPU0
GPU1
GPU2
GPU3
GPU0
GPU1
Input[0]
Input[1]

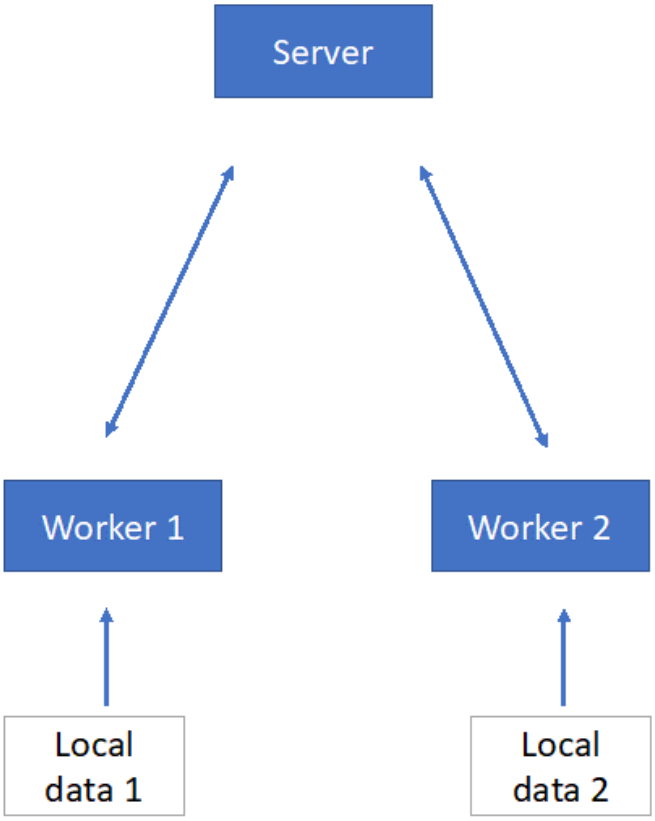


Mesh-2D

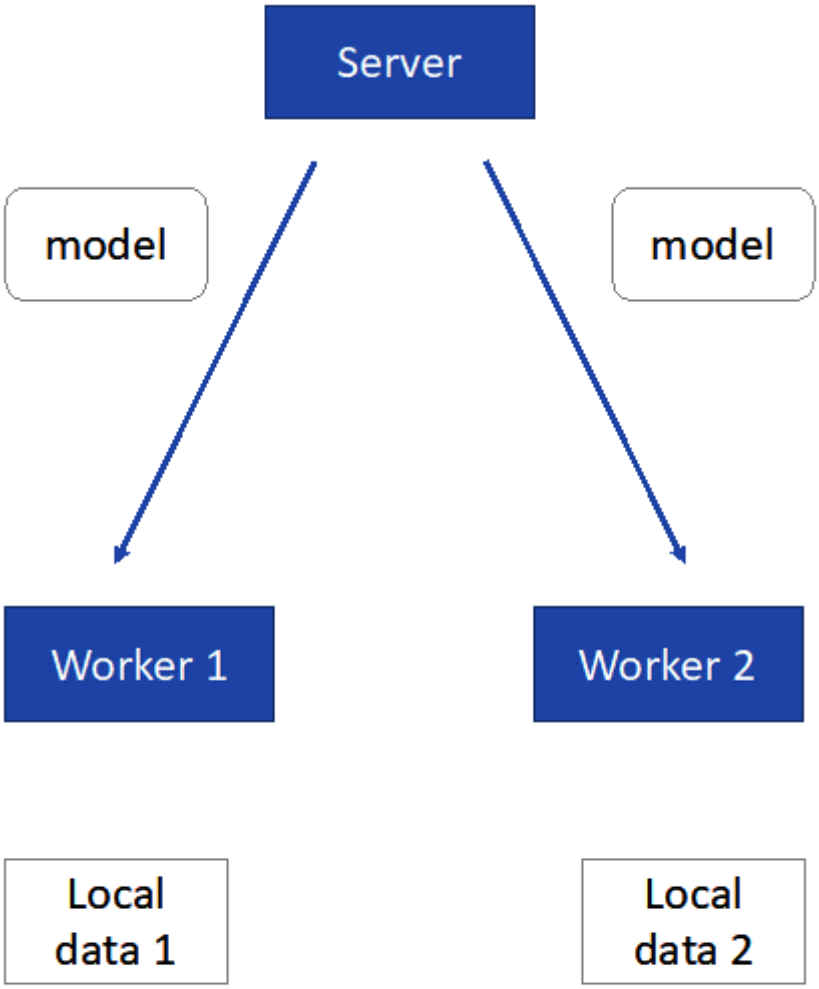
Chapter 10: Federated Learning and Edge Devices



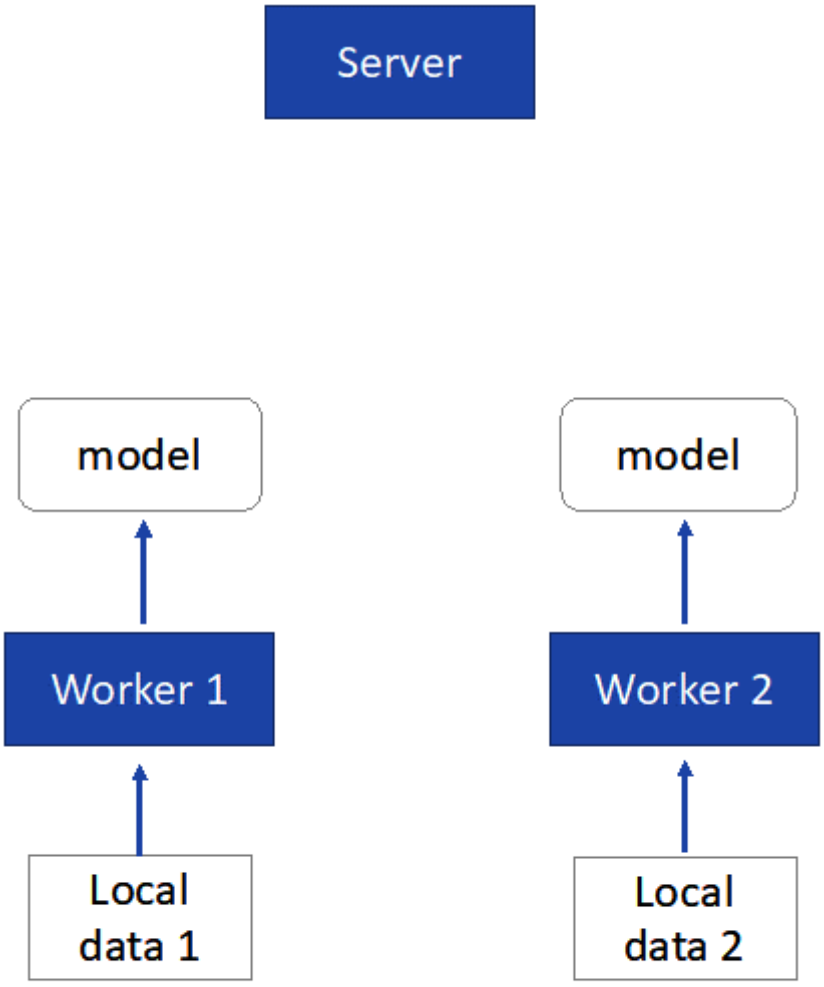
Data parallel training
(2 workers and 1 server)



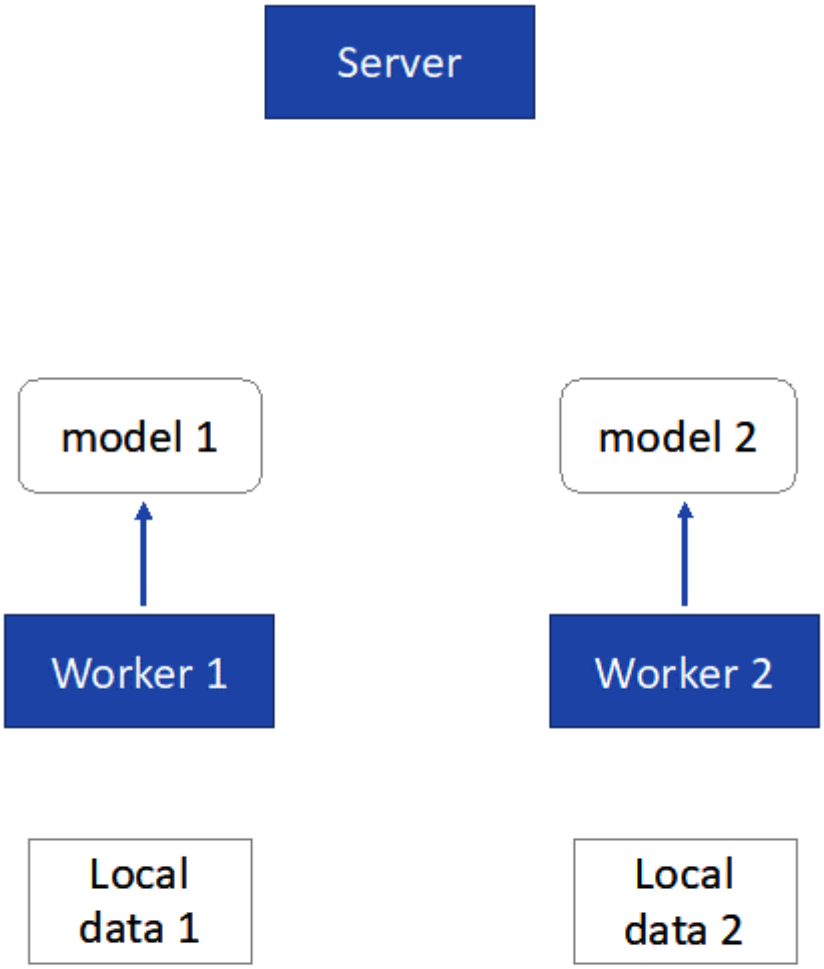
Federated Learning
(2 workers and 1 server)



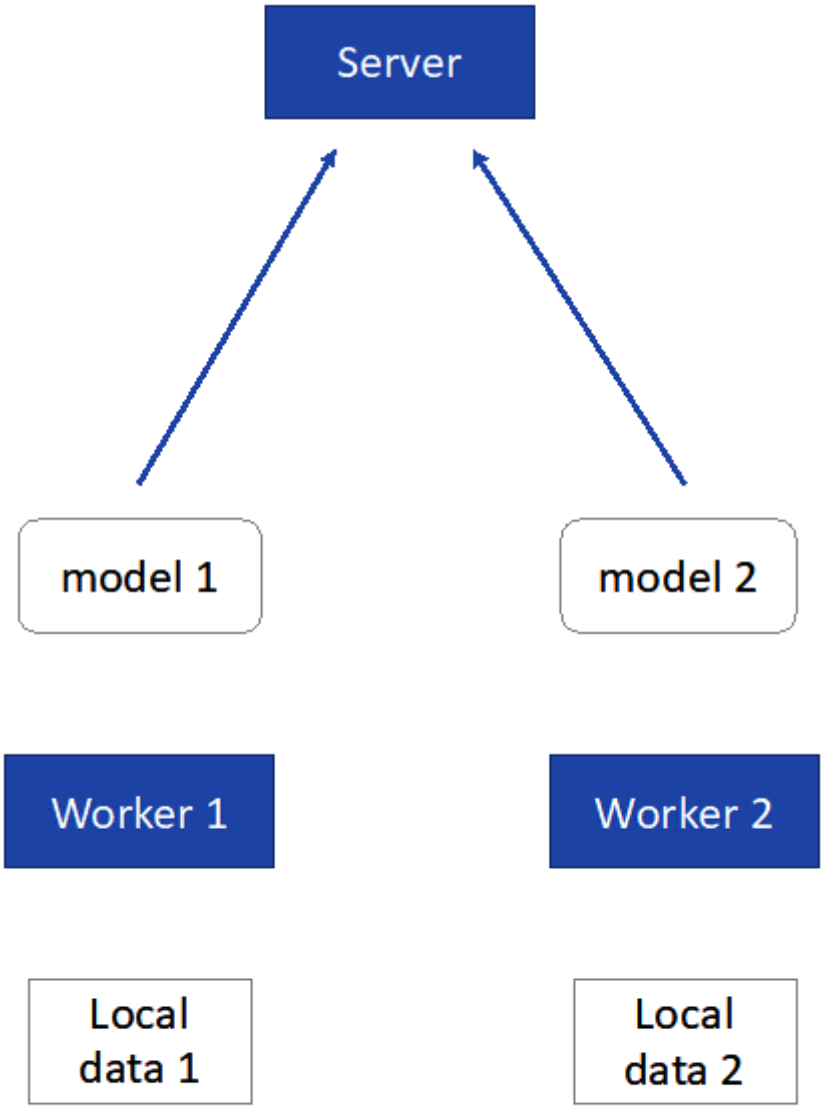
Federated Learning
(2 workers and 1 server)



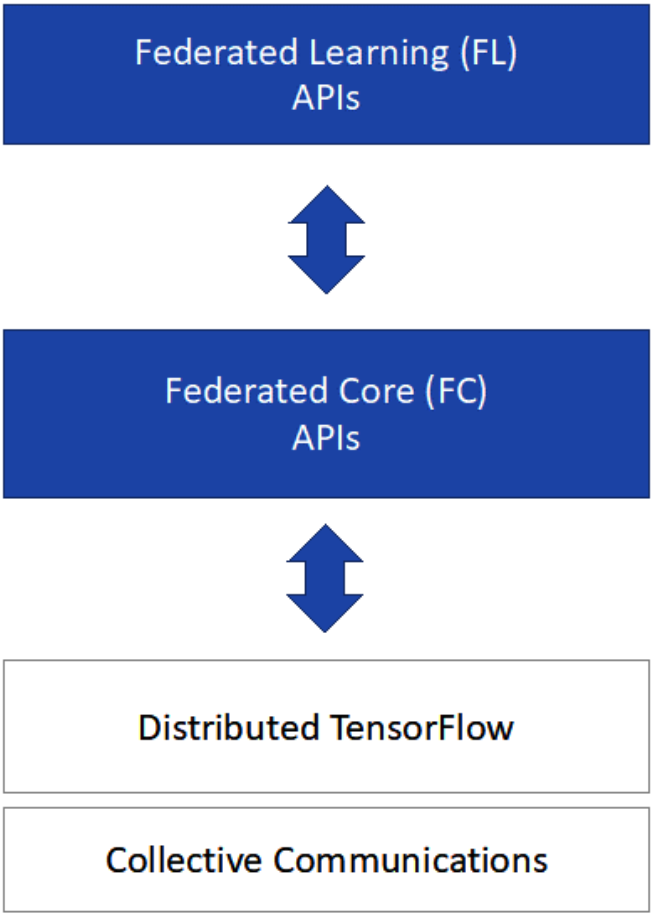
Federated Learning
(2 workers and 1 server)



Federated Learning
(2 workers and 1 server)

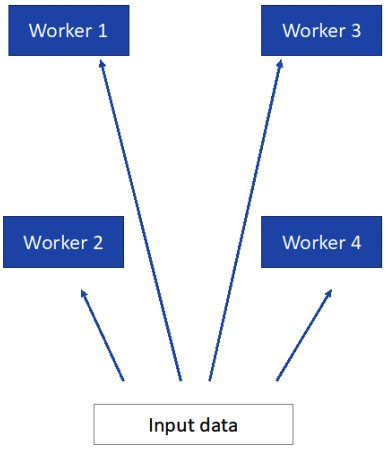


Federated Learning
(2 workers and 1 server)

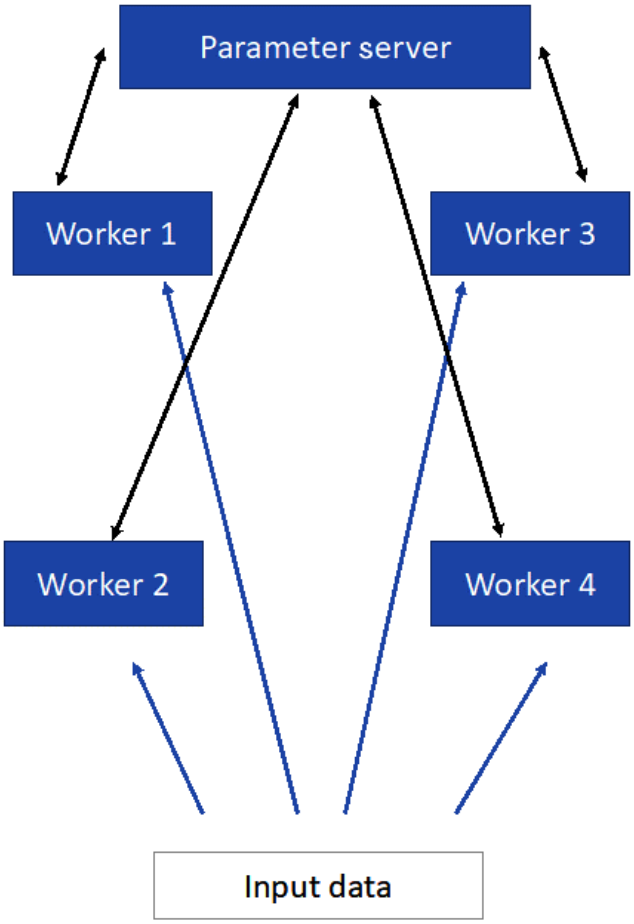


TensorFlow Federated
(two top layers)

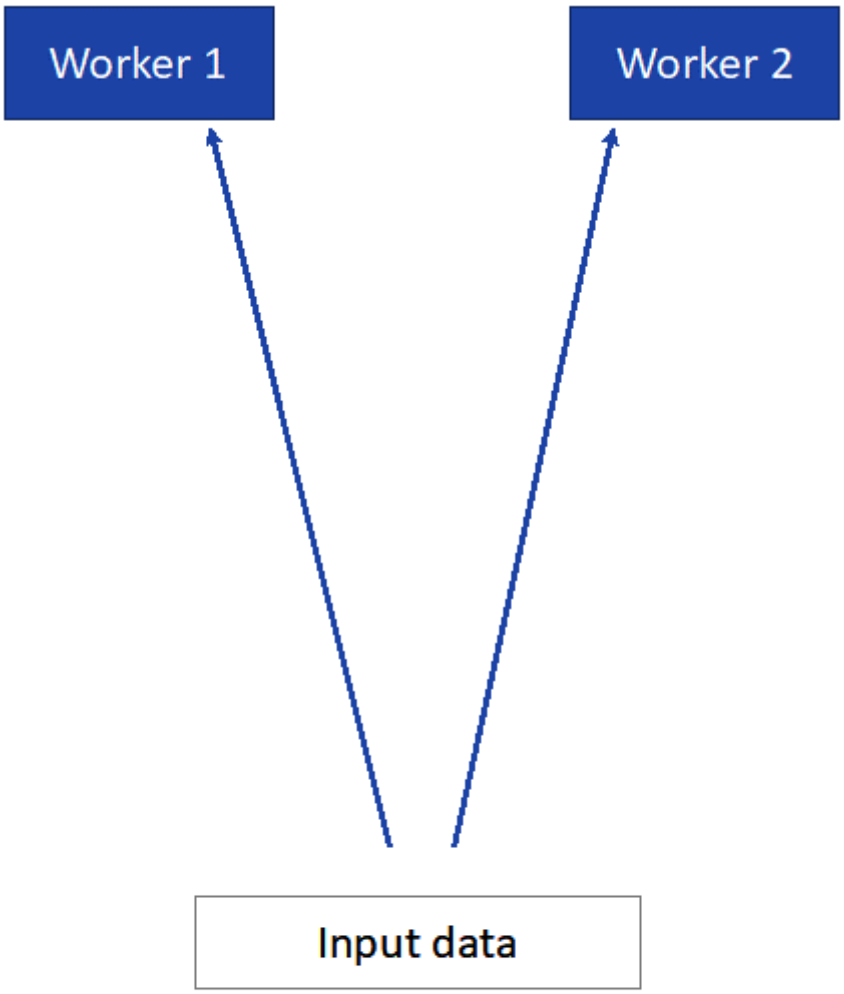
Chapter 11: Elastic Model Training and Serving



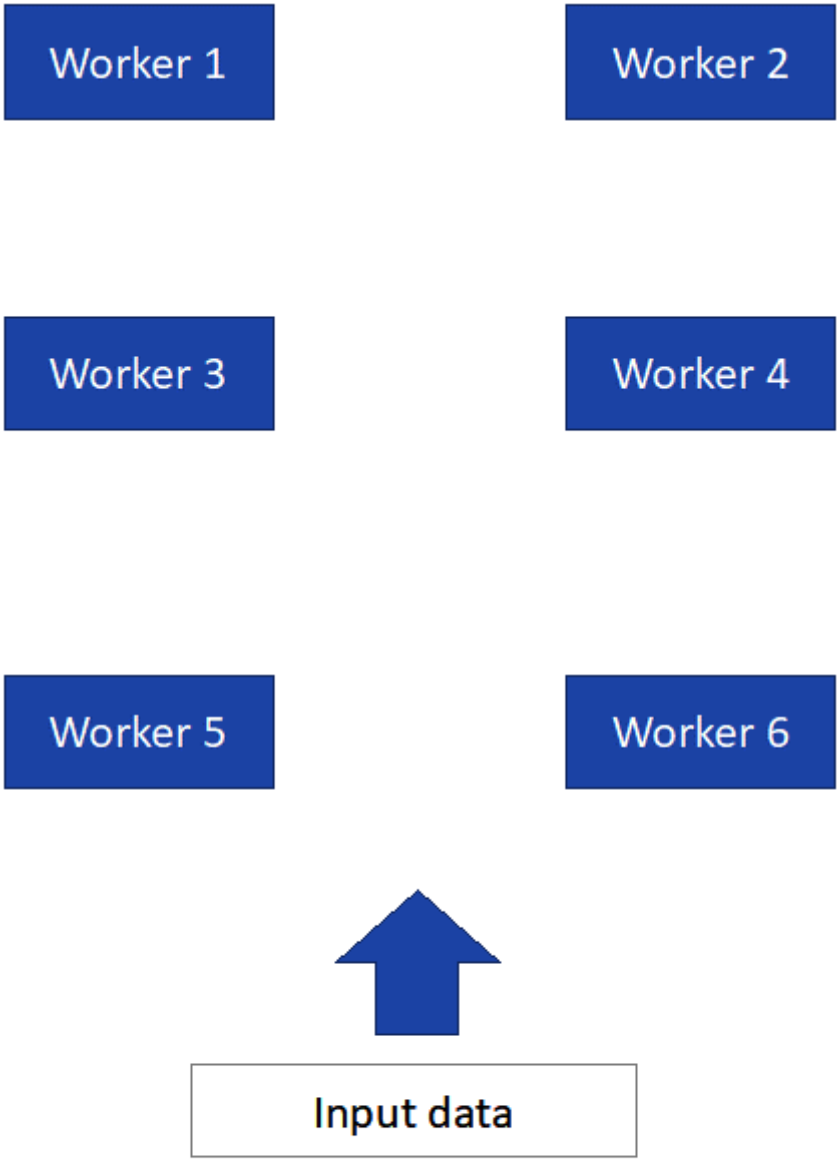
Data-parallel training
(AllReduce-based)



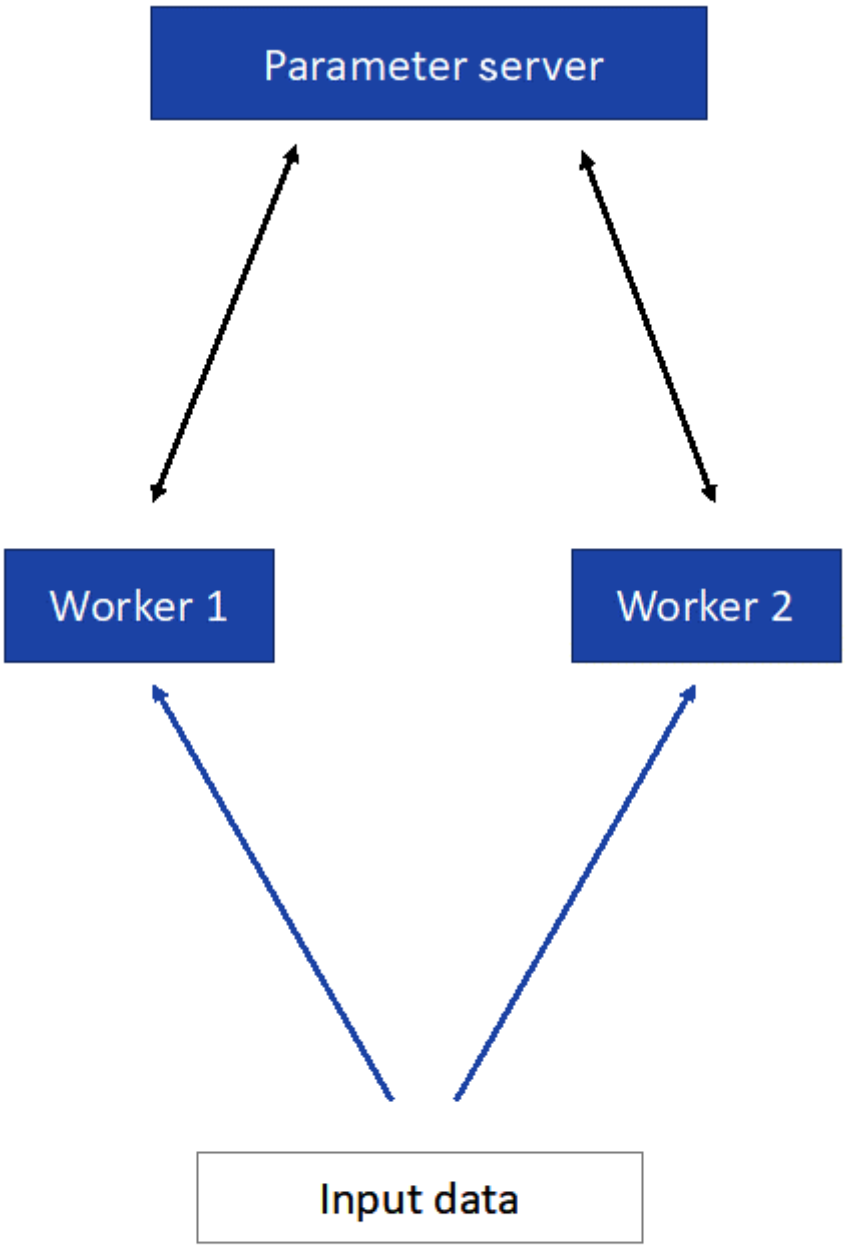
Data-parallel training
(Parameter server-based)



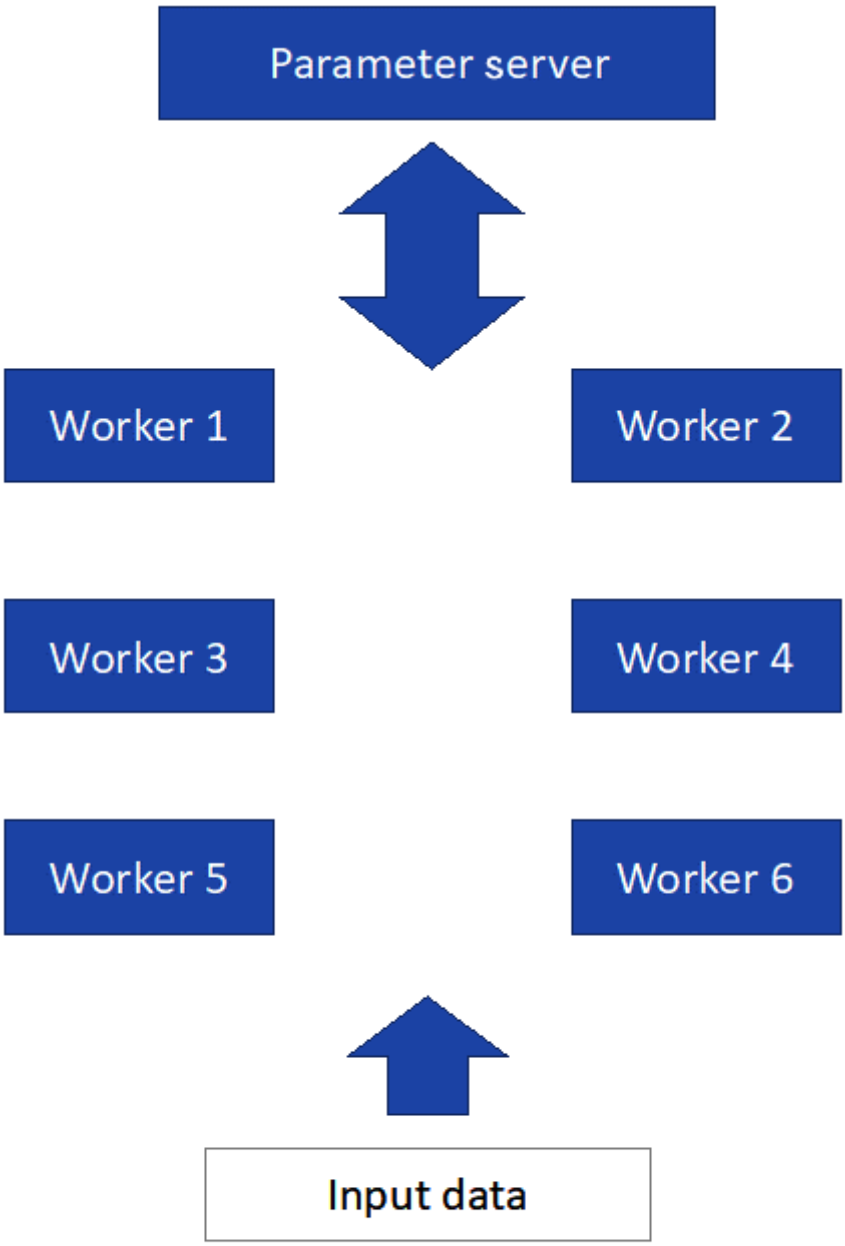
Adaptive data-parallel training
(AllReduce-based, early stage)



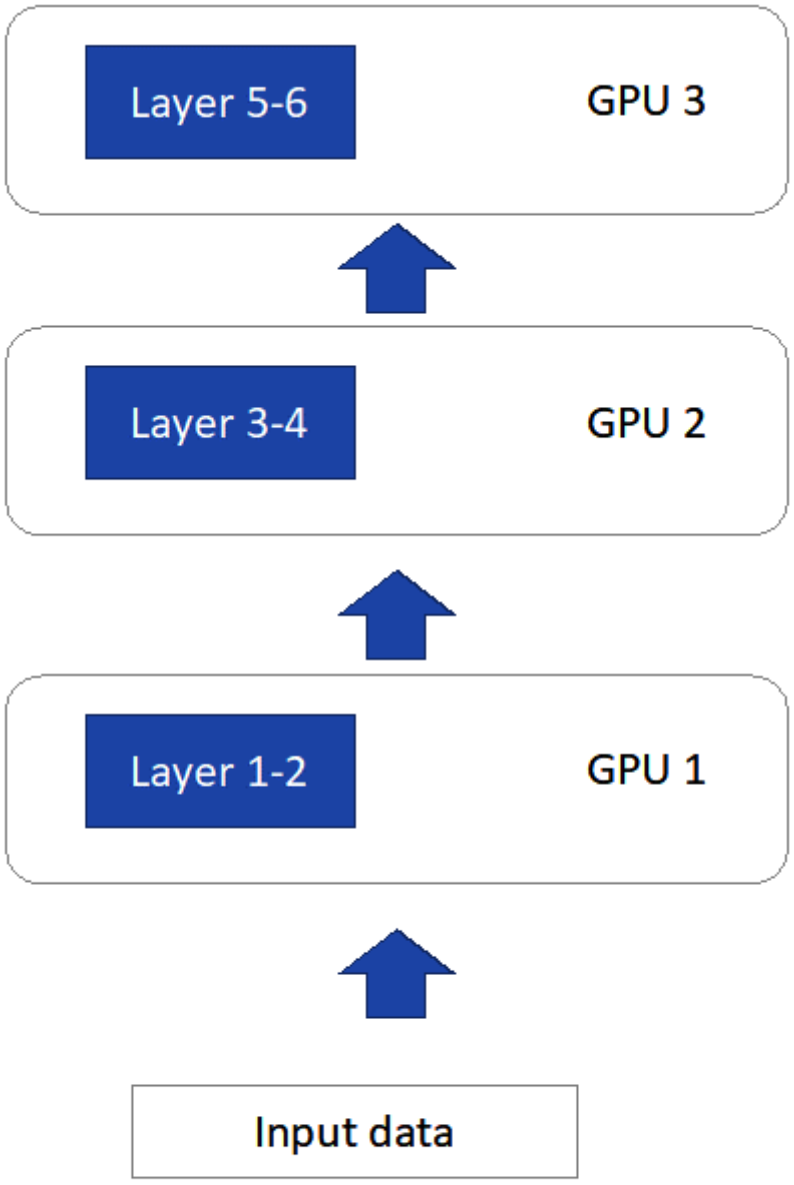
Adaptive data-parallel training
(AllReduce-based, late stage)



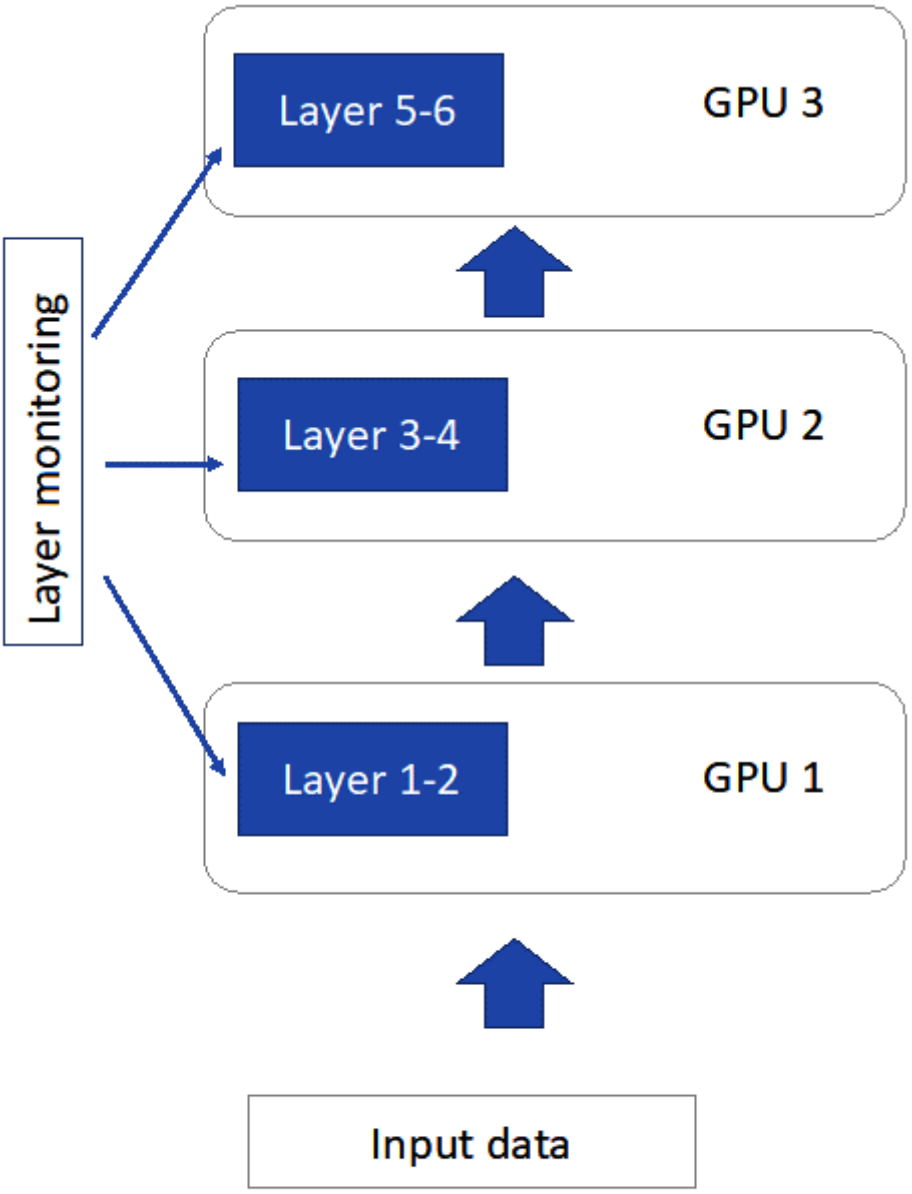
Data-parallel training
(Parameter server-based, early stage)



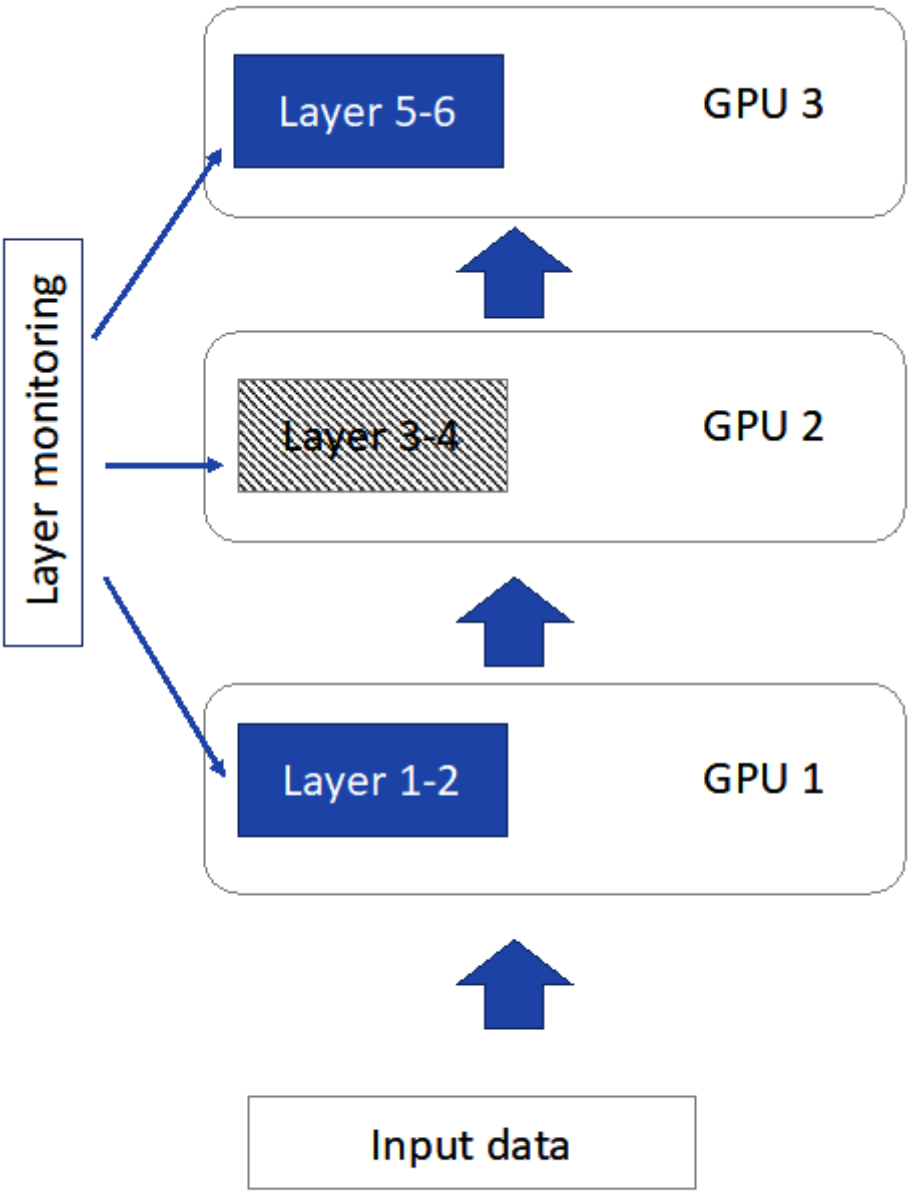
Data-parallel training
(Parameter server-based, late stage)



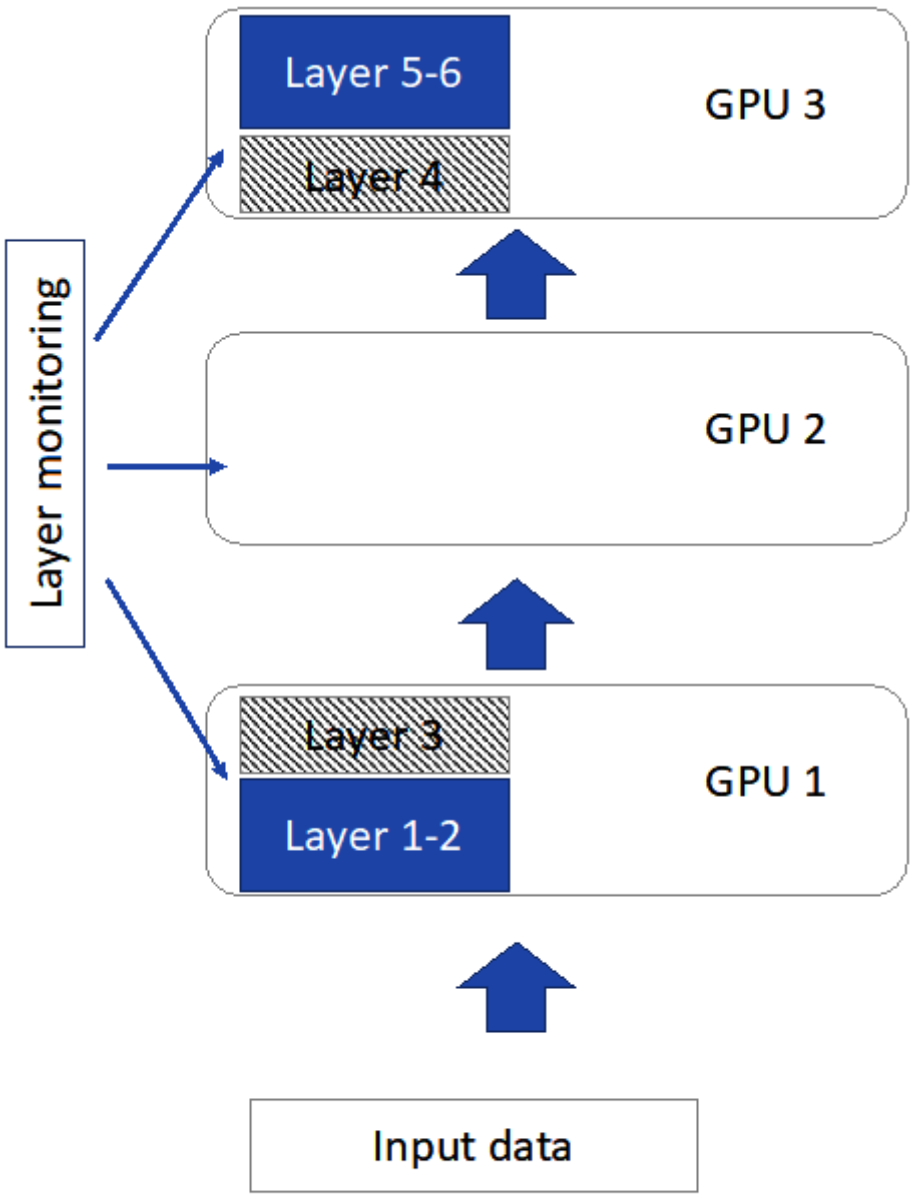
Model-parallel Training



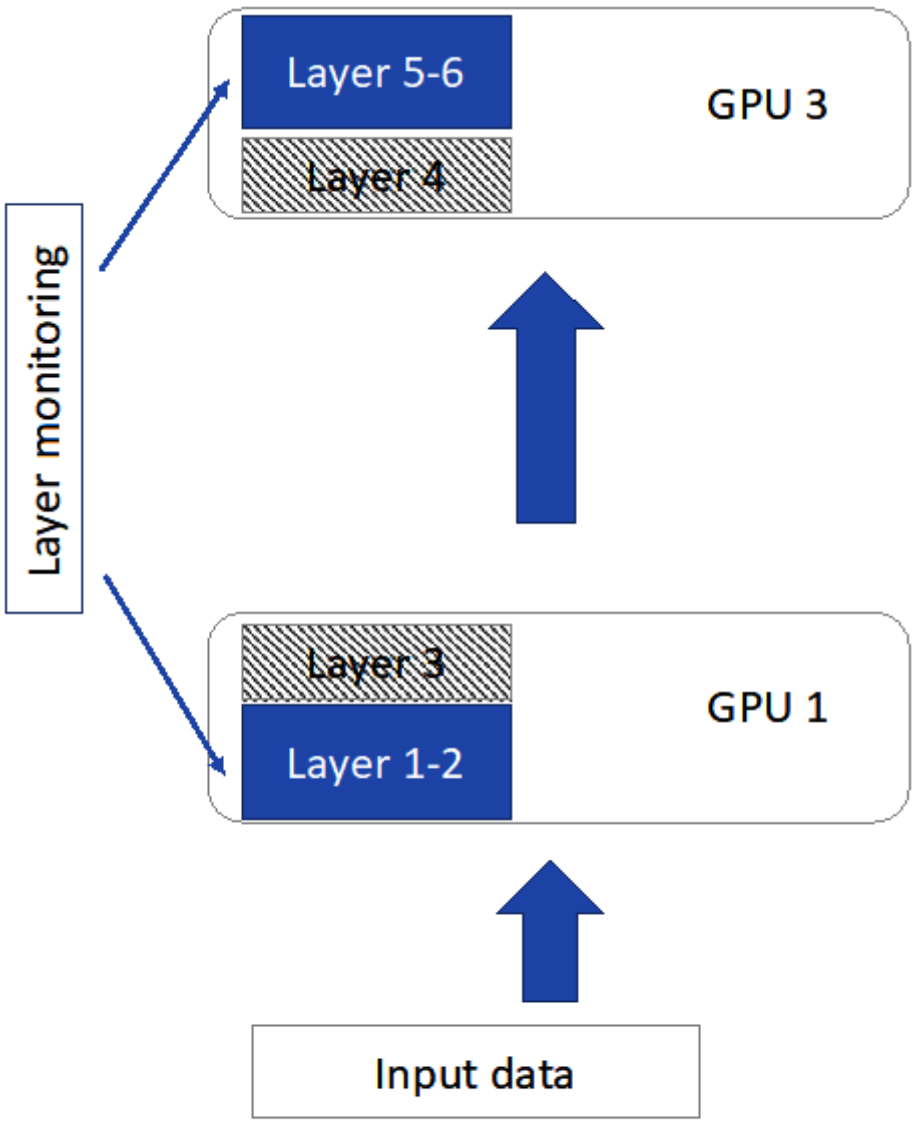
Adaptive model-parallel training
(early stage)



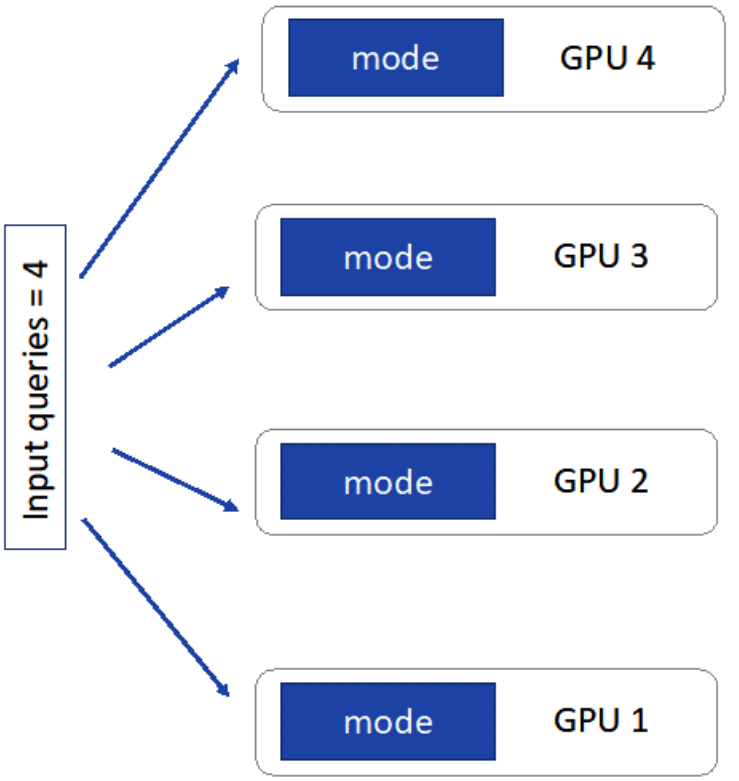
Adaptive model-parallel training
(after early stage)



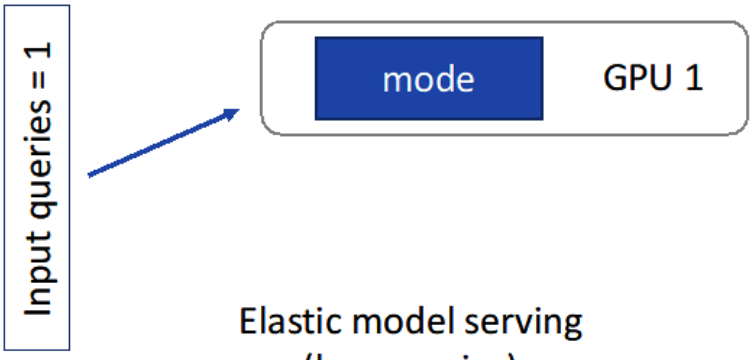
Adaptive model-parallel training
(before late stage)



Adaptive model-parallel training
(late stage)

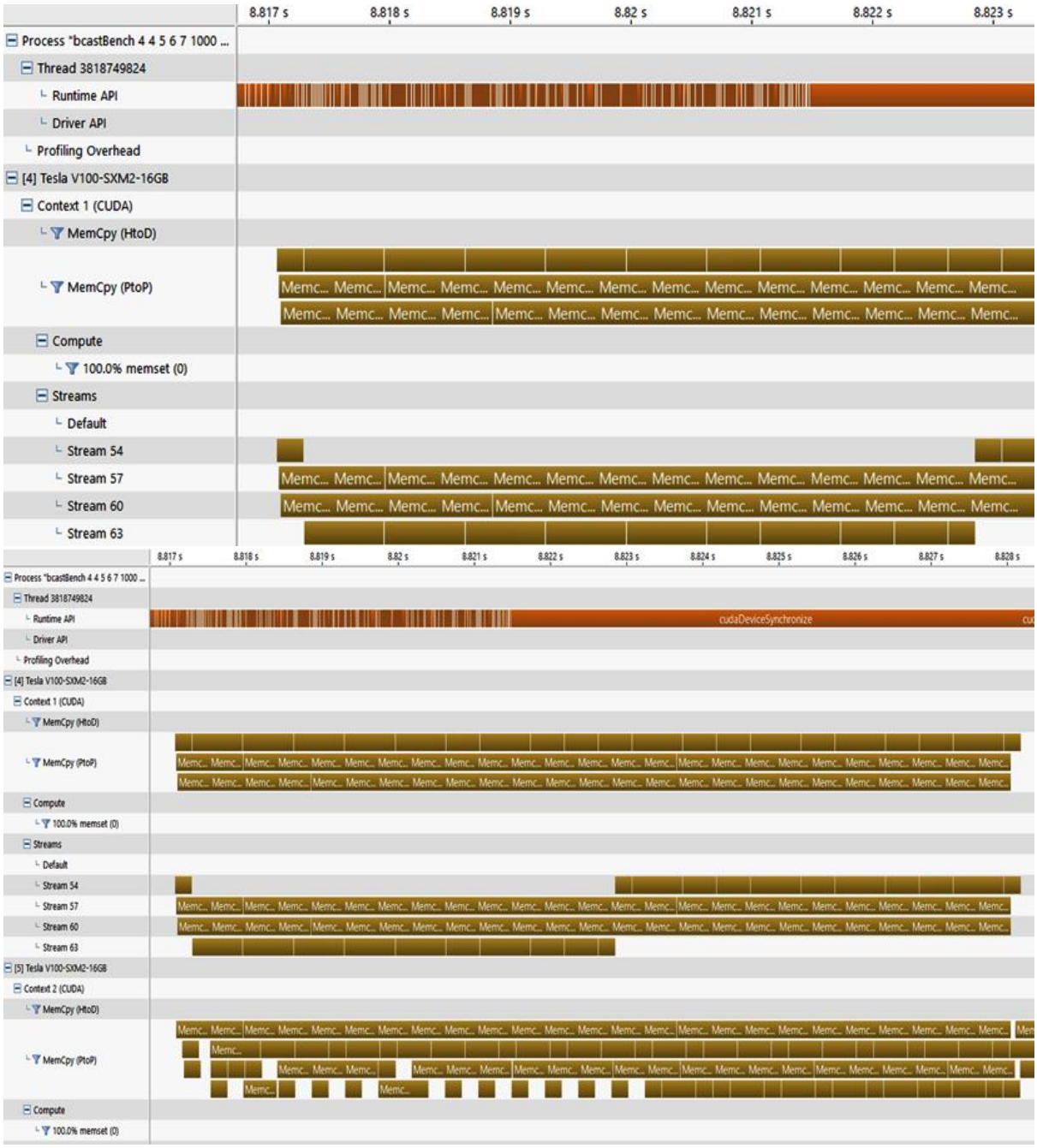


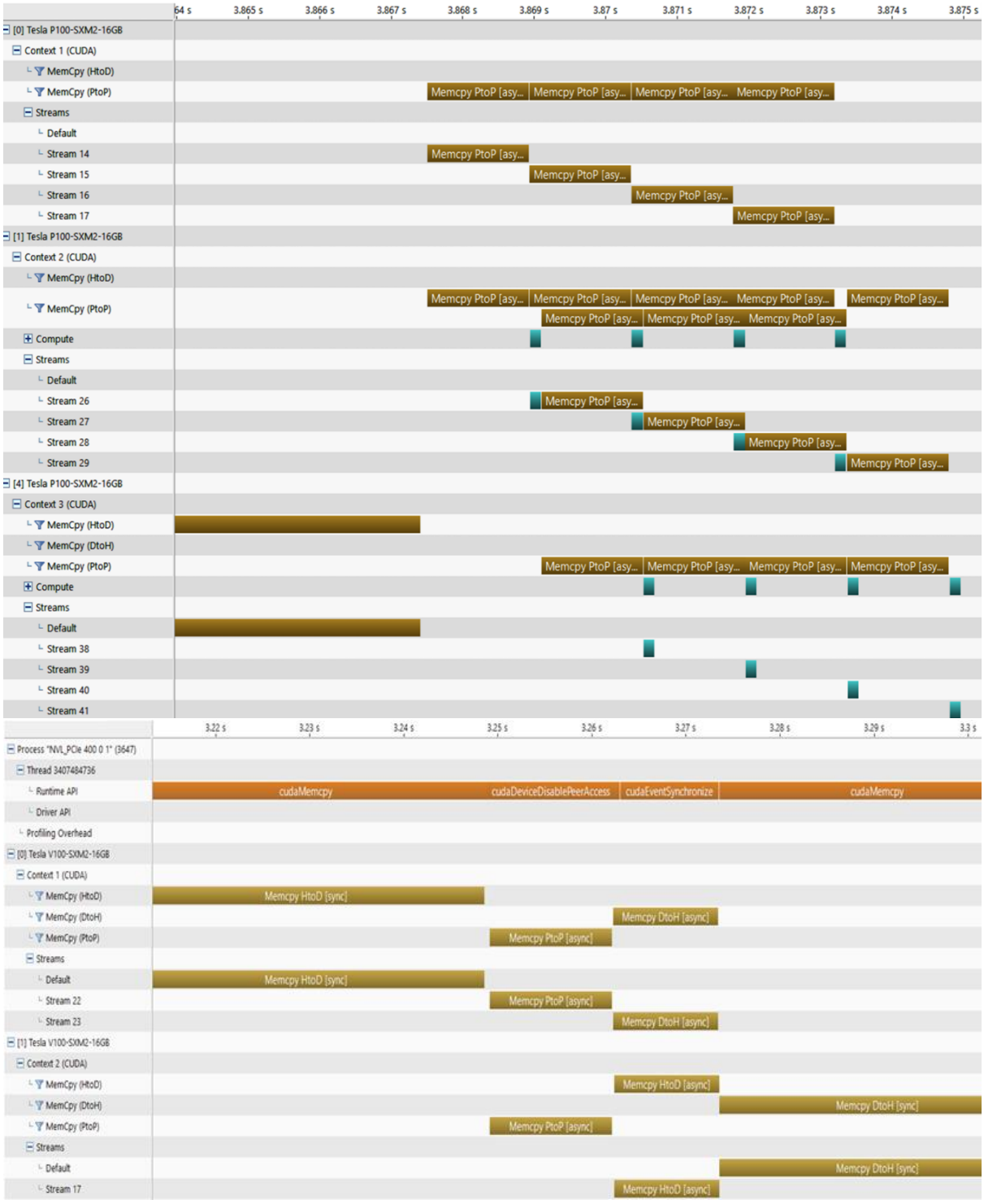
Elastic model serving
(more queries)

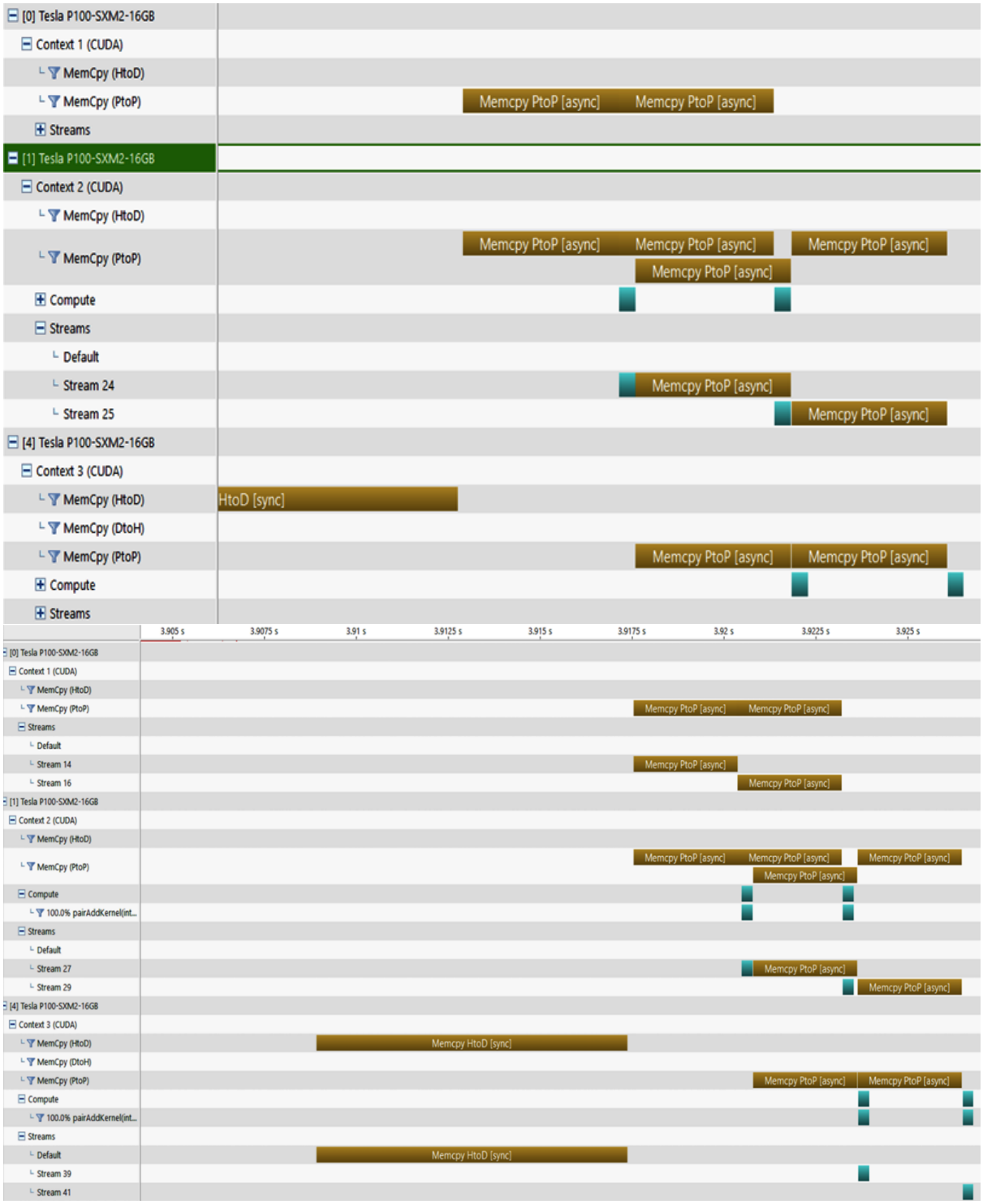


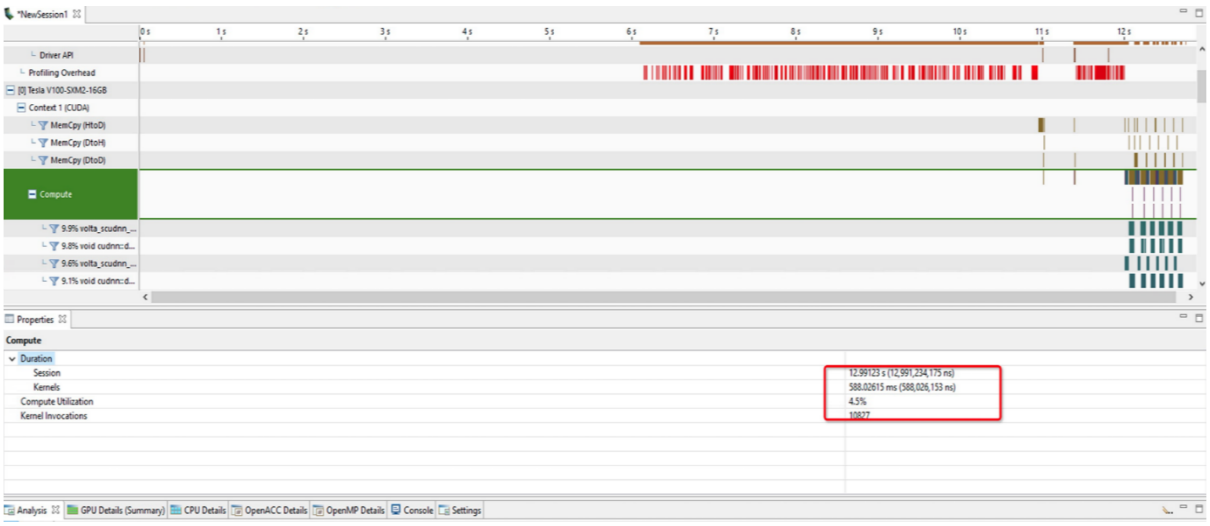
Elastic model serving
(less queries)

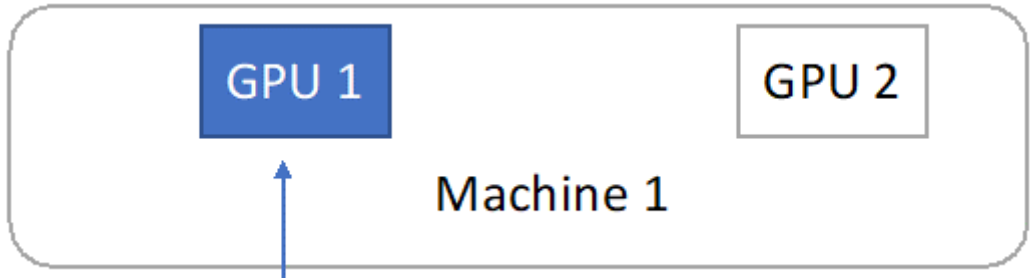
Chapter 12: Advanced Techniques for Further Speed-Ups

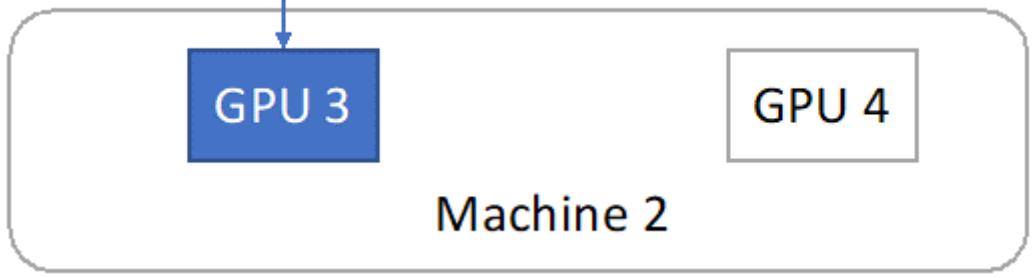


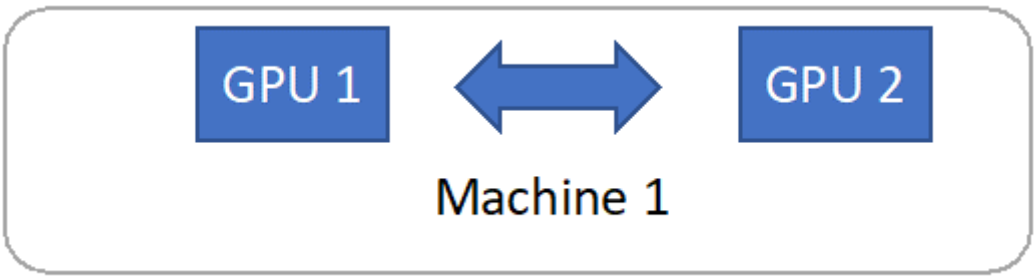

3.865 s
3.866 s
3.867 s
3.868 s
3.869 s
3.87 s
3.871 s
3.872 s
3.873 s
3.874 s
3.875 s
[0] Tesla P100-SXM2-16GB
Context 1 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Streams
Default
Stream 14
Stream 15
Stream 16
Stream 17
[1] Tesla P100-SXM2-16GB
Context 2 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Compute
Streams
Default
Stream 26
Stream 27
Stream 28
Stream 29
[4] Tesla P100-SXM2-16GB
Context 3 (CUDA)
MemCpy (HtoD)
MemCpy (DtoH)
MemCpy (PtoP)
Compute
Streams
Default
Stream 38
Stream 39
Stream 40
Stream 41
Memcpy PtoP [asy...
3.22 s
3.23 s
3.24 s
3.25 s
3.26 s
3.27 s
3.28 s
3.29 s
3.3 s
Process "NVL_PCIe 400 0 1" (3647)
Thread 3407484736
Runtime API
Driver API
Profiling Overhead
[0] Tesla V100-SXM2-16GB
Context 1 (CUDA)
MemCpy (HtoD)
MemCpy (DtoH)
MemCpy (PtoP)
Streams
Default
Stream 22
Stream 23
[1] Tesla V100-SXM2-16GB
Context 2 (CUDA)
MemCpy (HtoD)
MemCpy (DtoH)
MemCpy (PtoP)
Streams
Default
Stream 17
cudaMemcpy
cudaDeviceDisablePeerAccess
cudaEventSynchronize
cudaMemcpy
Memcpy HtoD [sync]
Memcpy DtoH [async]
Memcpy PtoP [async]
Memcpy HtoD [sync]
Memcpy PtoP [async]
Memcpy DtoH [async]
Memcpy HtoD [async]
Memcpy DtoH [sync]
Memcpy PtoP [async]
Memcpy DtoH [sync]
Memcpy HtoD [async]

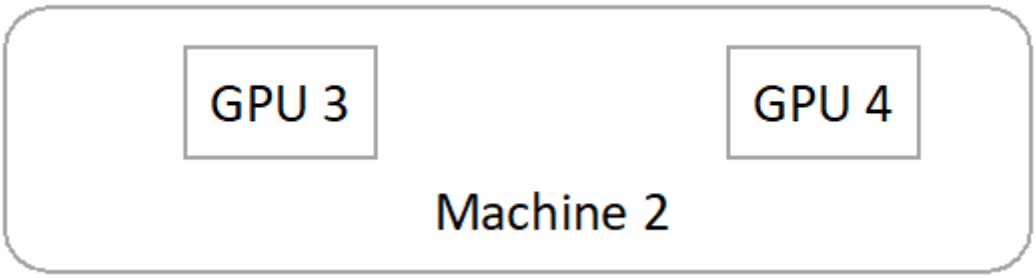

[0] Tesla P100-SXM2-16GB
Context 1 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Streams
[1] Tesla P100-SXM2-16GB
Context 2 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Memcpy PtoP [async]
Memcpy PtoP [async]
Compute
Streams
Default
Stream 24
Memcpy PtoP [async]
Stream 25
Memcpy PtoP [async]
[4] Tesla P100-SXM2-16GB
Context 3 (CUDA)
MemCpy (HtoD)
HtoD [sync]
MemCpy (DtoH)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Compute
Streams
3.905 s
3.9075 s
3.91 s
3.9125 s
3.915 s
3.9175 s
3.92 s
3.9225 s
3.925 s
[0] Tesla P100-SXM2-16GB
Context 1 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Streams
Default
Stream 14
Memcpy PtoP [async]
Stream 16
Memcpy PtoP [async]
[1] Tesla P100-SXM2-16GB
Context 2 (CUDA)
MemCpy (HtoD)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Memcpy PtoP [async]
Memcpy PtoP [async]
Compute
100.0% pairAddKernel(int...
Streams
Default
Stream 27
Memcpy PtoP [async]
Stream 29
Memcpy PtoP [async]
[4] Tesla P100-SXM2-16GB
Context 3 (CUDA)
MemCpy (HtoD)
Memcpy HtoD [sync]
MemCpy (DtoH)
MemCpy (PtoP)
Memcpy PtoP [async]
Memcpy PtoP [async]
Compute
100.0% pairAddKernel(int...
Streams
Default
Memcpy HtoD [sync]
Stream 39
Stream 41


*NewSession1
Driver API
Profiling Overhead
Tesla V100-SXM2-16GB
Context 1 (CUDA)
MemCpy (HtoD)
MemCpy (DtoH)
MemCpy (DtoD)
Compute
Properties
Compute
Duration
Session
Kernels
Compute Utilization
Kernel Invocations
12.99123 s (12,991,234,175 ns)
588.02615 ms (588,026,153 ns)
4.5%
10827
Analysis
GPU Details (Summary)
CPU Details
OpenACC Details
OpenMP Details
Console
Settings


GPU 1
GPU 2
Machine 1


GPU 3
GPU 4
Machine 2


GPU 1
GPU 2
Machine 1


GPU 3
GPU 4
Machine 2