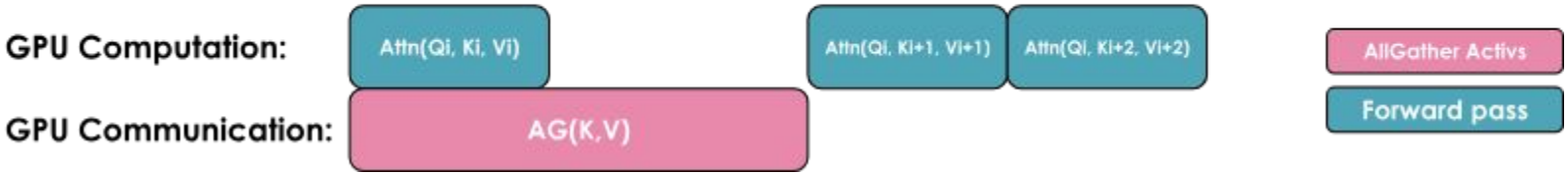
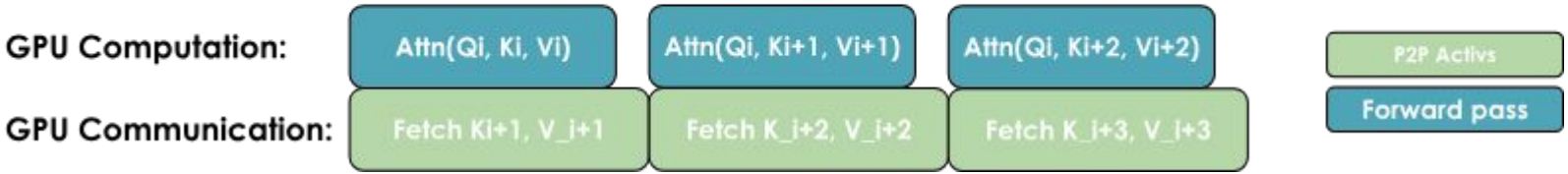
Pipeline Parallelism

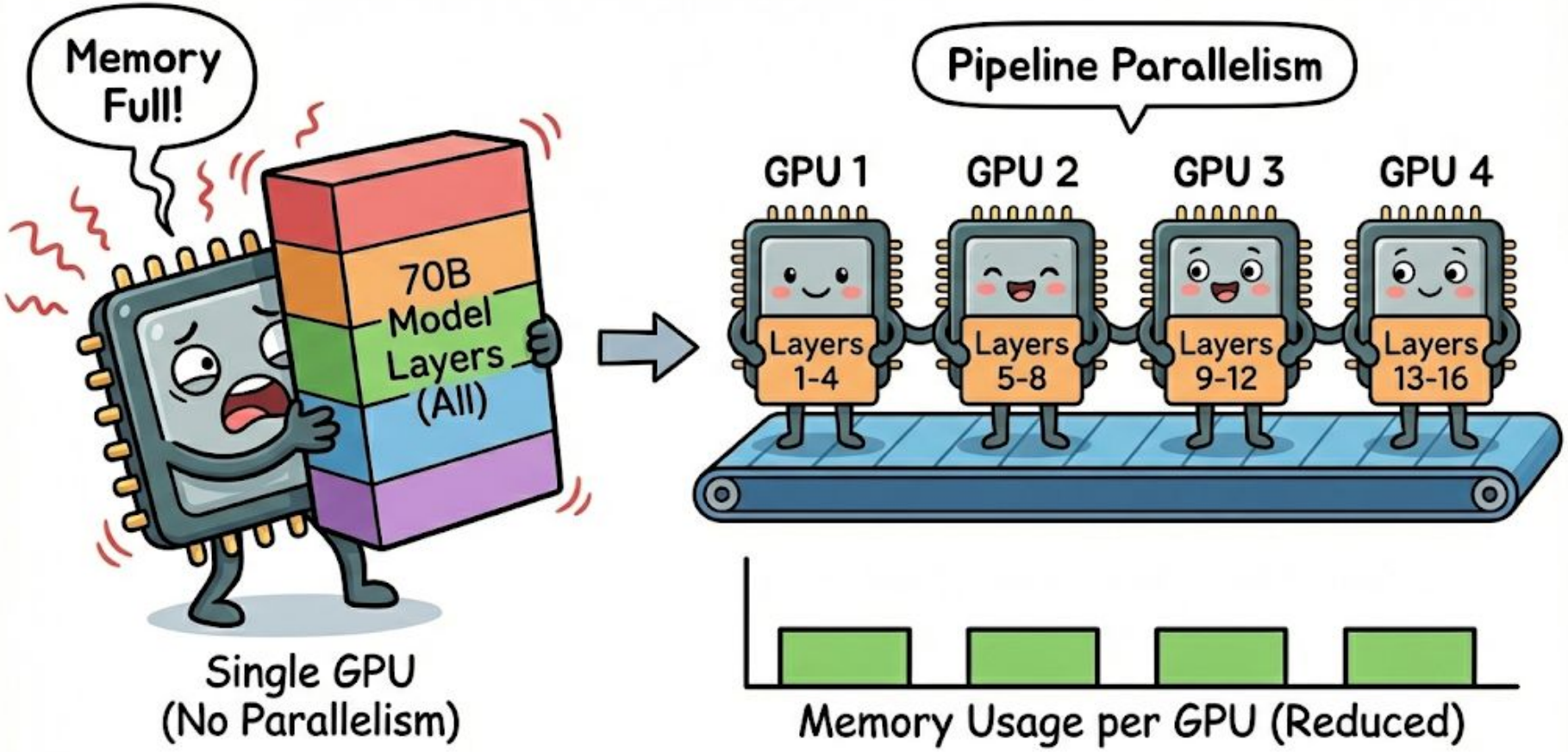
made with ❤️ for “Little ML book club”

Attention

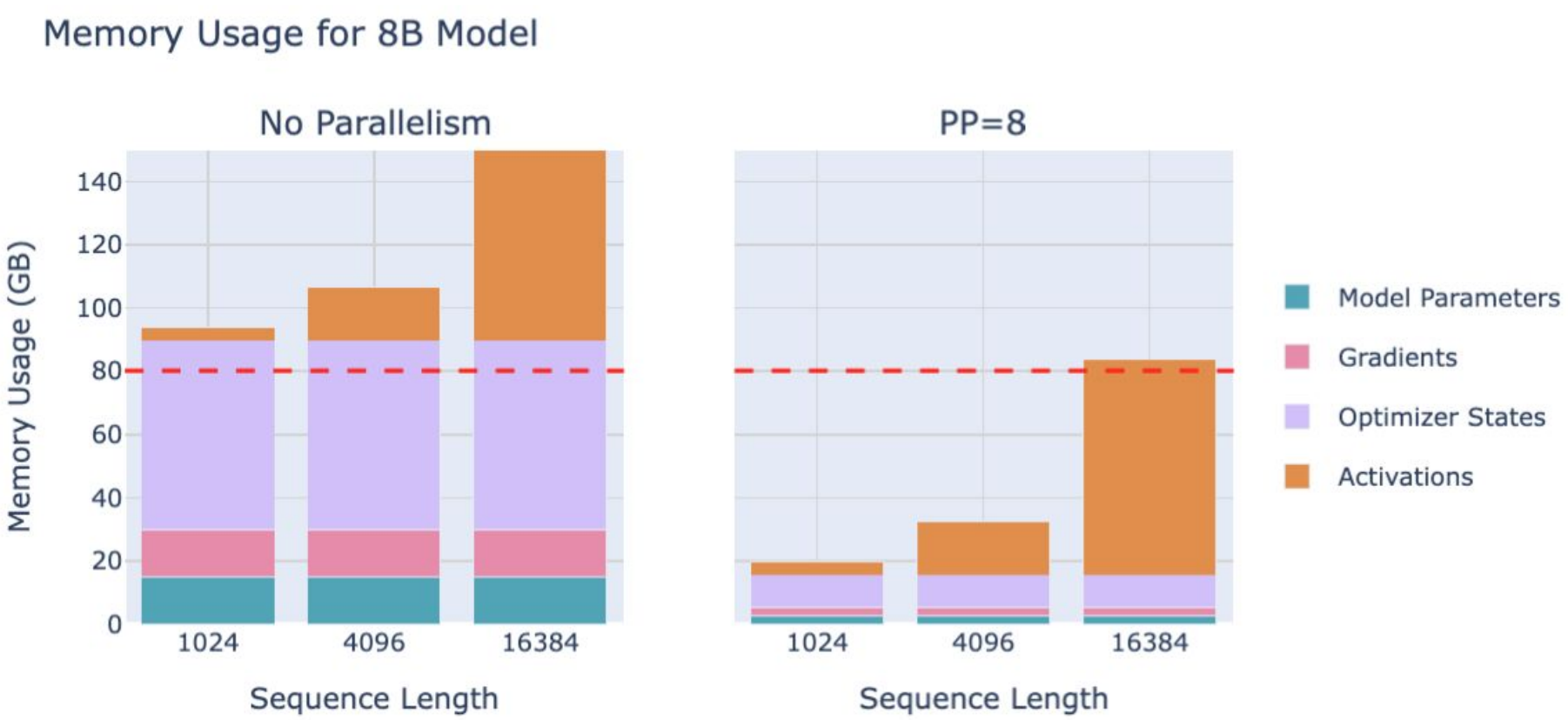


Attention

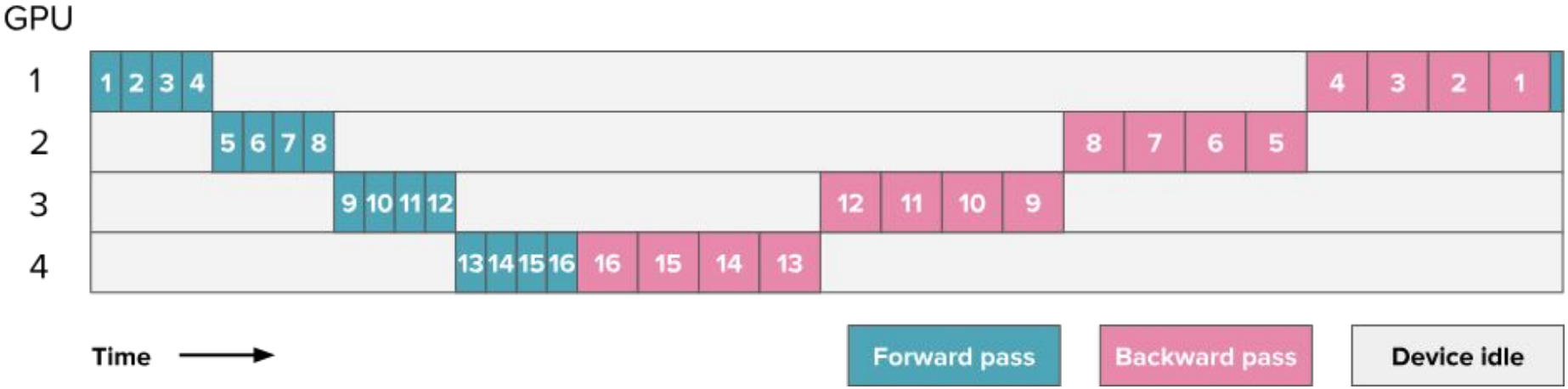


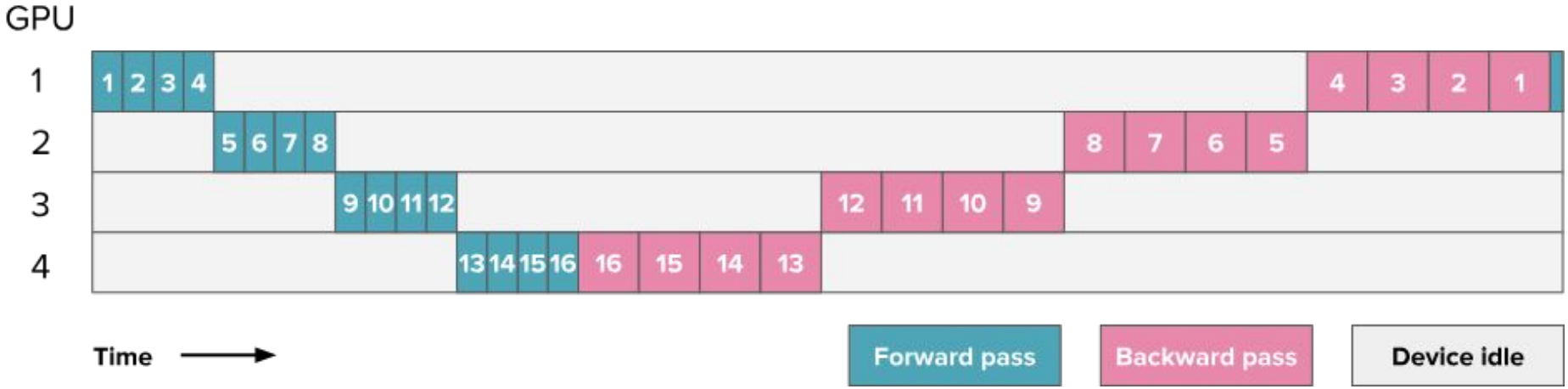


Efficiently splits model layers across multiple GPUs

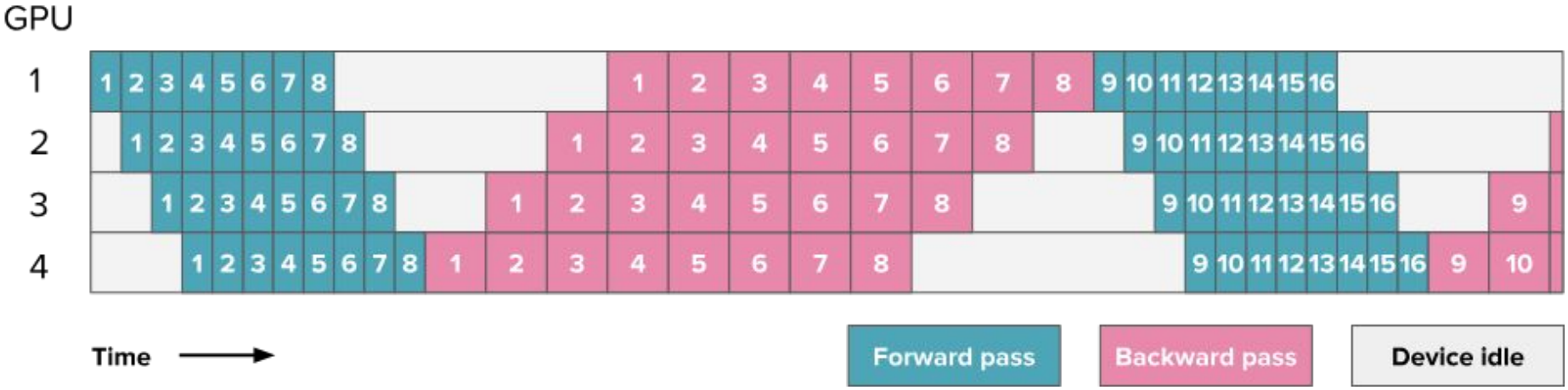

Memory Usage for 8B Model
No Parallelism
PP=8
Memory Usage (GB)
0
20
40
60
80
100
120
140
1024
4096
16384
Sequence Length
1024
4096
16384
Sequence Length
Model Parameters
Gradients
Optimizer States
Activations

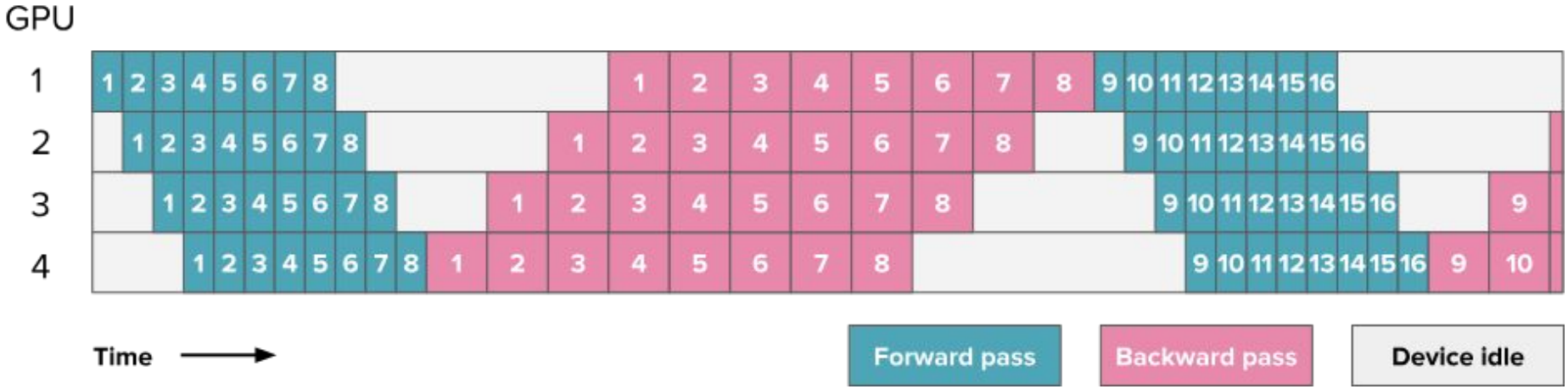
> 📝 **Note**
>
> This is because each GPU needs to perform PP forward passes before starting the first backward pass. Since each GPU handles 1/PP of the layers but needs to process PP micro-batches before the first backward, it ends up storing $PP \times (activs/PP) \approx activs$, which means the activation memory requirement remains roughly the same as without pipeline parallelism.


GPU
1
2
3
4
1 2 3 4
4 3 2 1
5 6 7 8
8 7 6 5
9 10 11 12
12 11 10 9
13 14 15 16
16 15 14 13
Time
Forward pass
Backward pass
Device idle

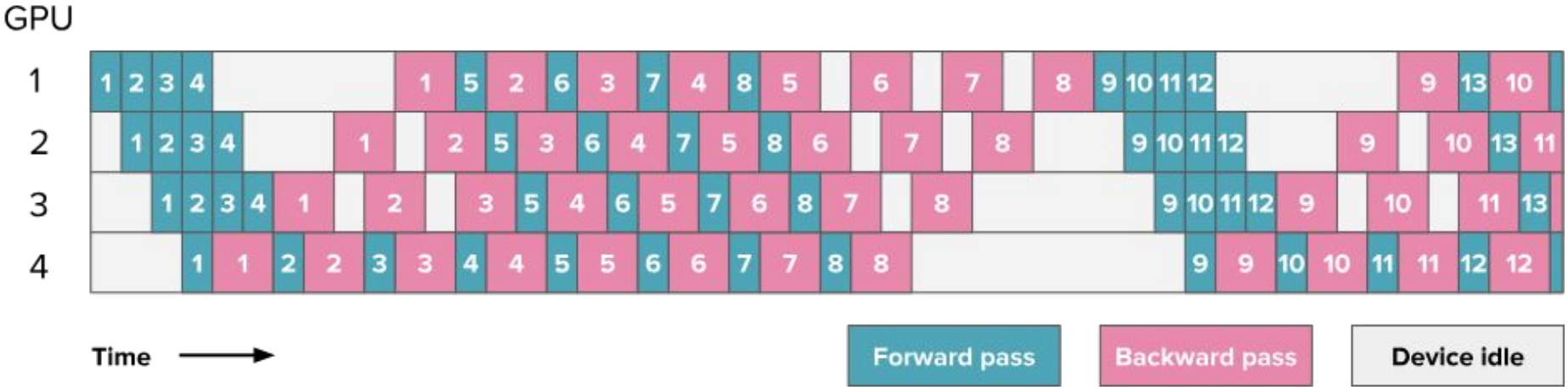


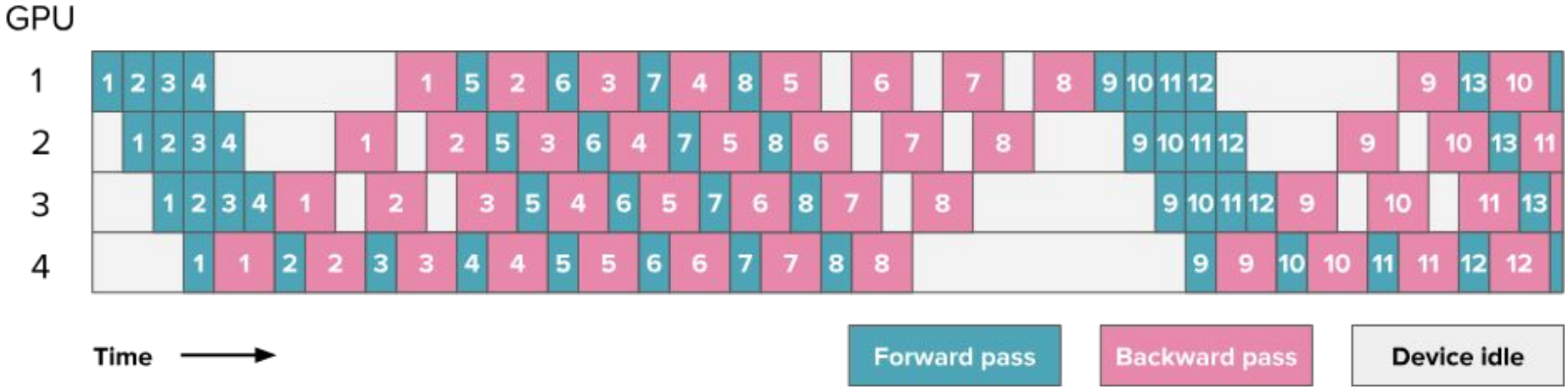
GPU activation memory = 1/4 of total activation memory


GPU
1
2
3
4
Time
Forward pass
Backward pass
Device idle

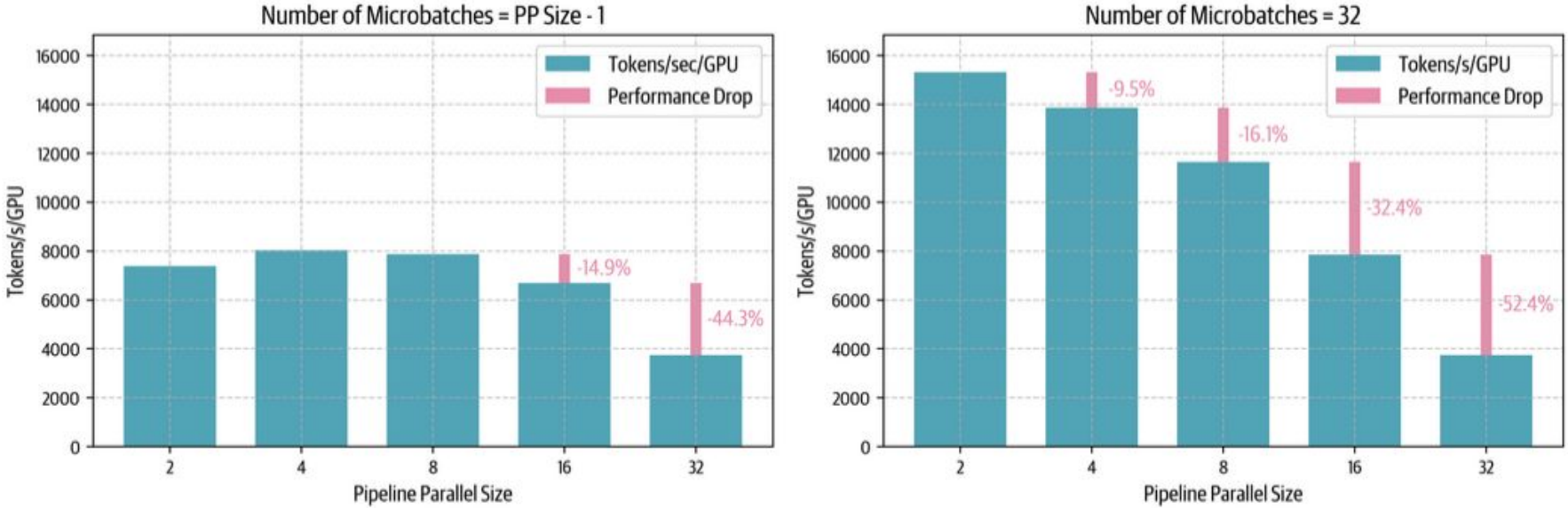


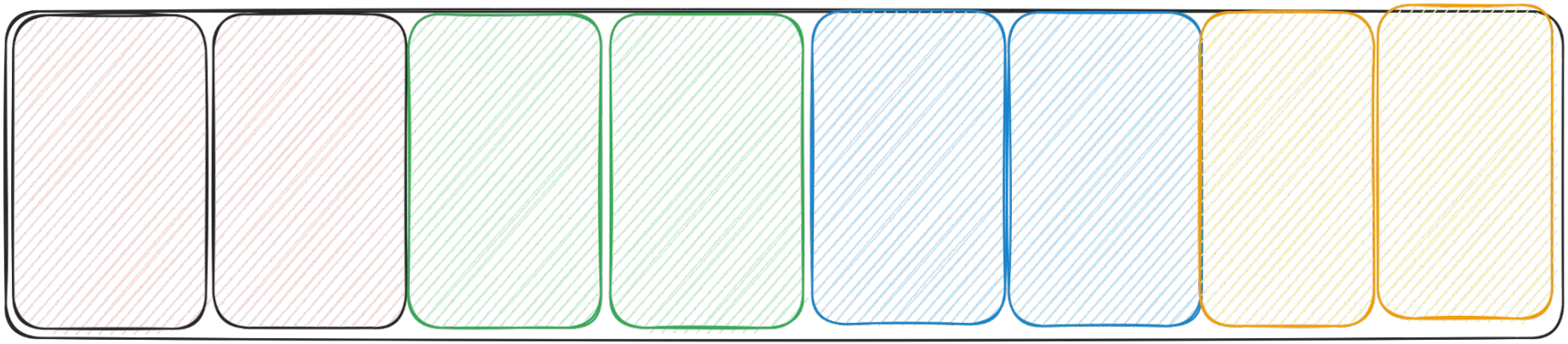
GPU activation memory = 8/4 of total activation memory

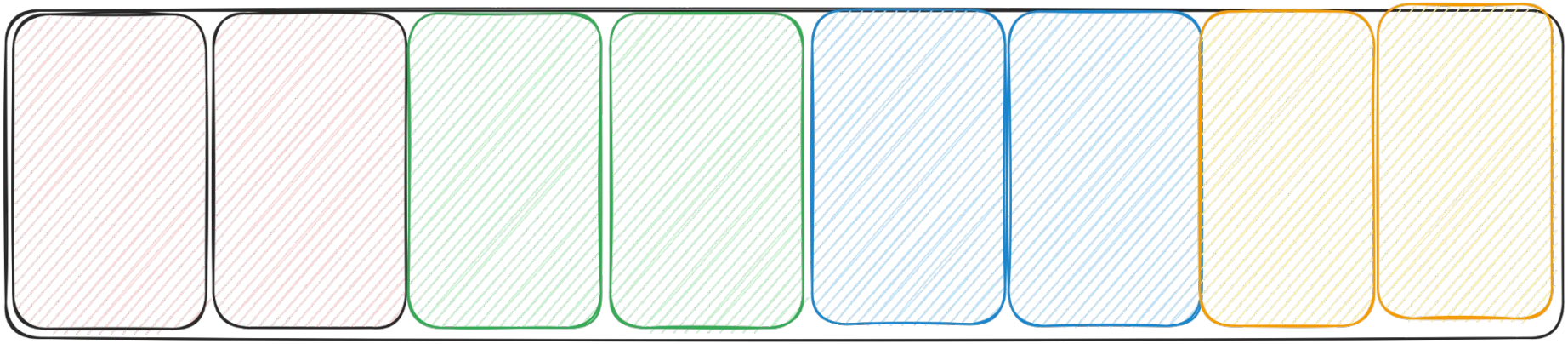

GPU
1
2
3
4
Time
Forward pass
Backward pass
Device idle

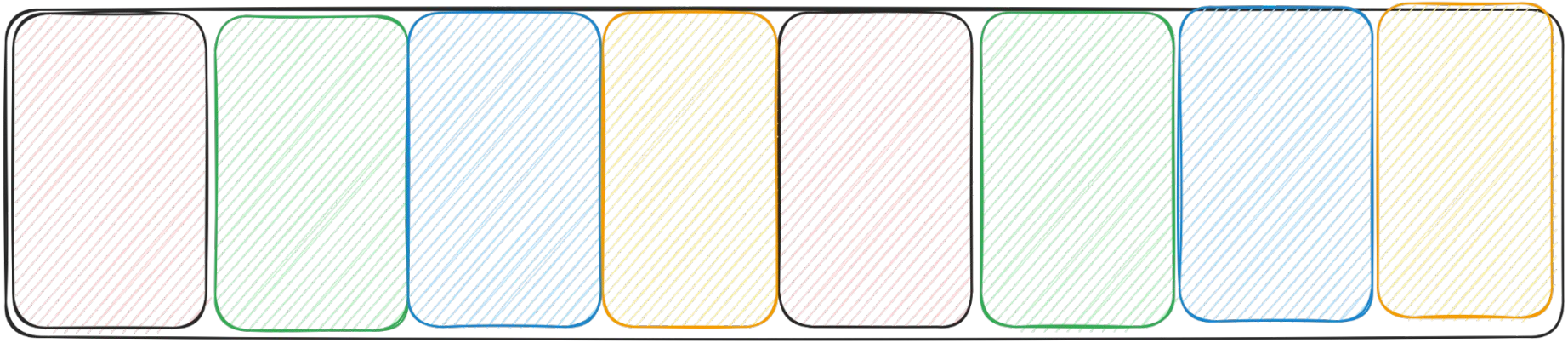


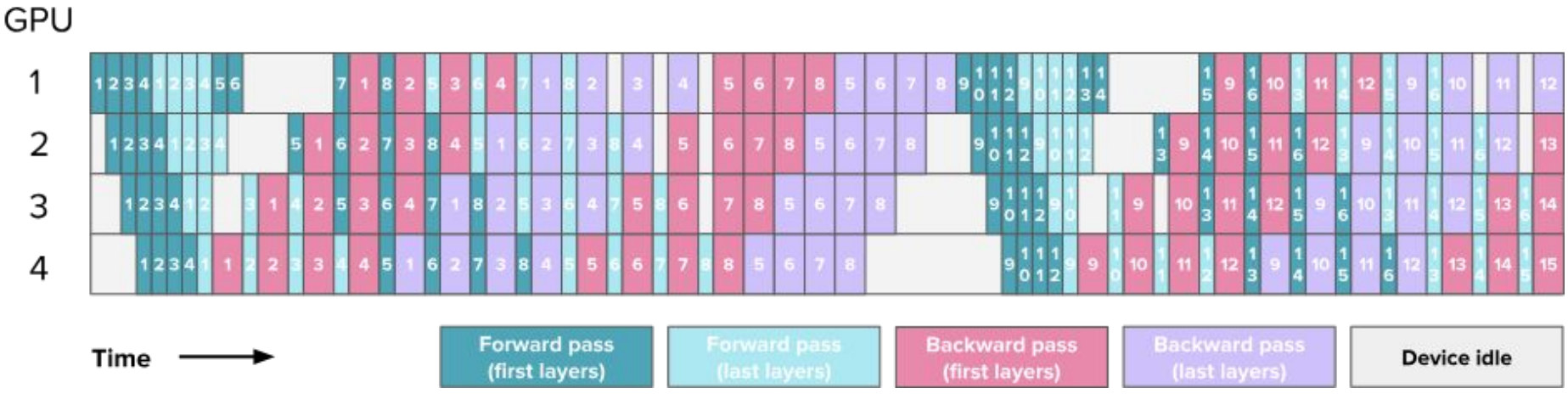
GPU activation memory = 4/4 of total activation memory

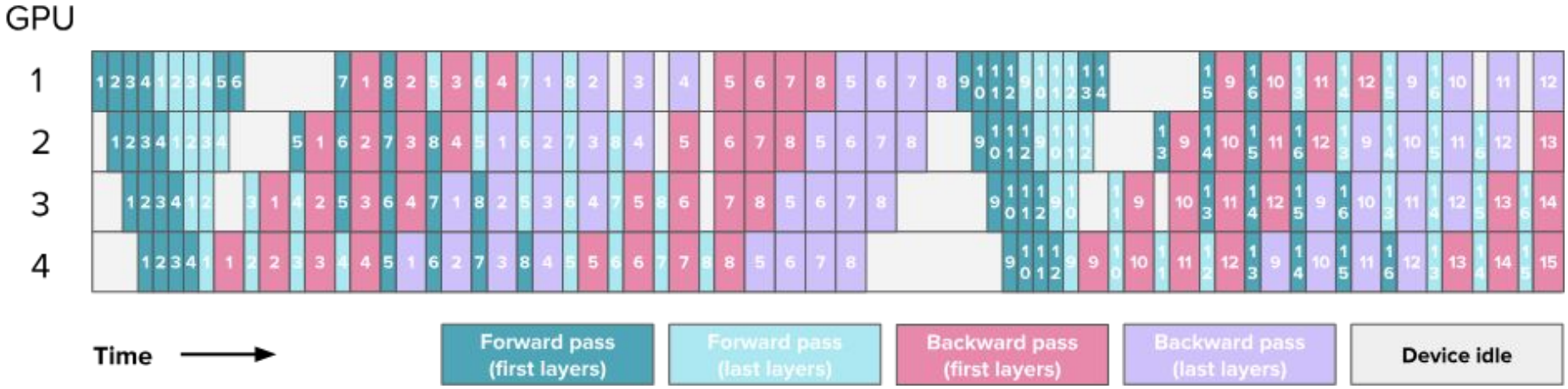

Throughput Scaling with Pipeline Parallelism (1F1B schedule)
Number of Microbatches = PP Size - 1
Tokens/sec/GPU
Performance Drop
Tokens/s/GPU
-14.9%
-44.3%
Pipeline Parallel Size
Number of Microbatches = 32
Tokens/s/GPU
Performance Drop
-9.5%
-16.1%
-32.4%
-52.4%
Pipeline Parallel Size








GPU
1
2
3
4
Time
Forward pass (first layers)
Forward pass (last layers)
Backward pass (first layers)
Backward pass (last layers)
Device idle



GPU activation memory = 4/4 of total activation memory

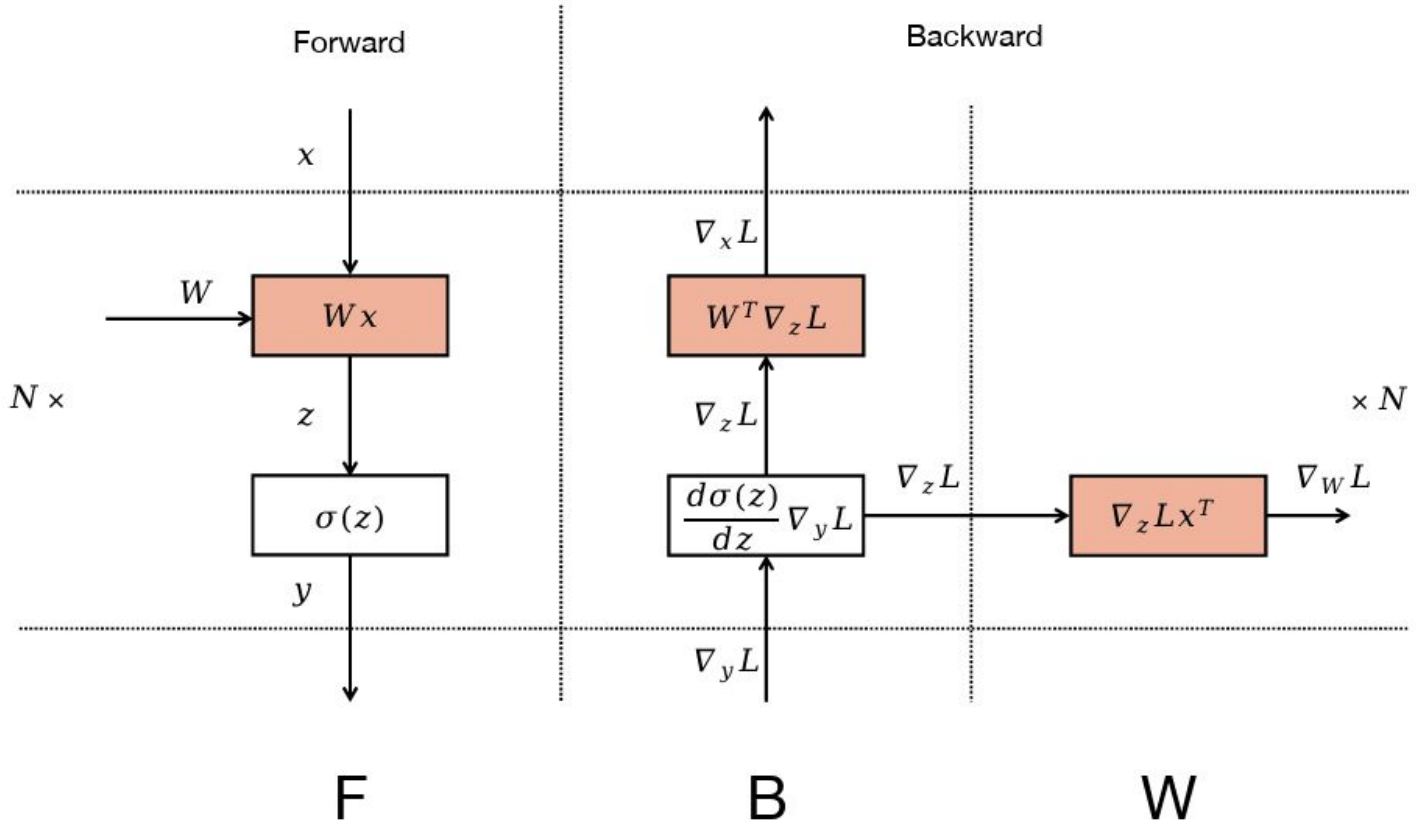


Figure 1: Computation Graph for MLP.

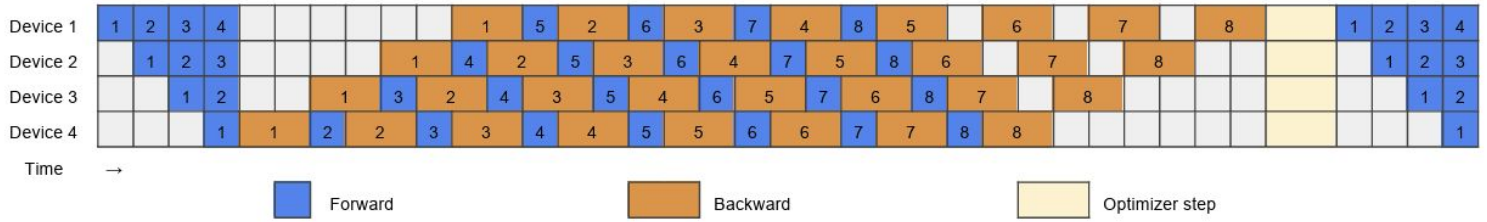


Figure 2: 1F1B pipeline schedule.

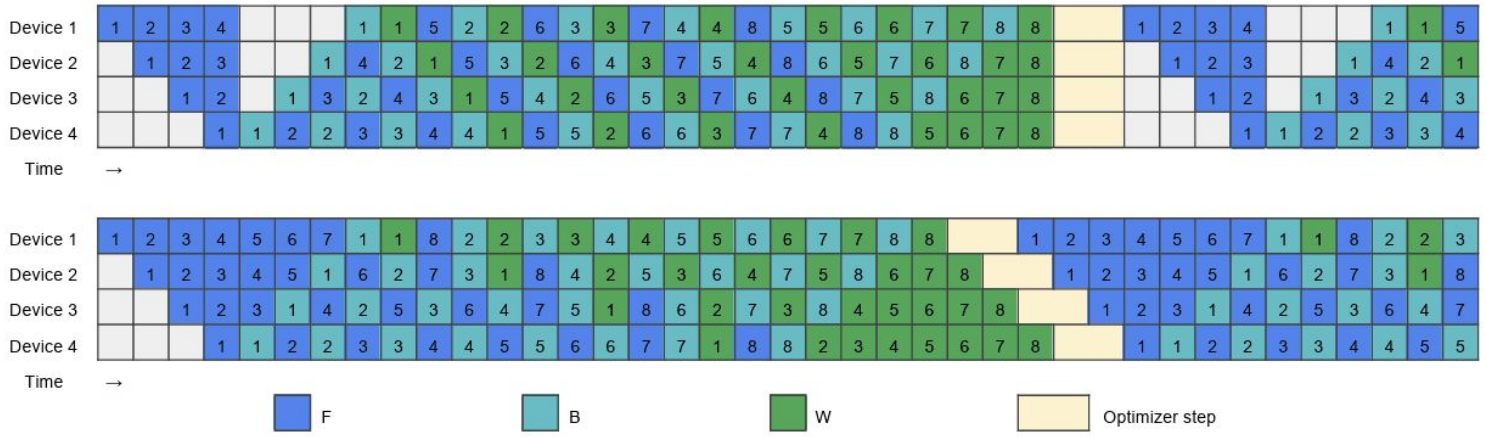
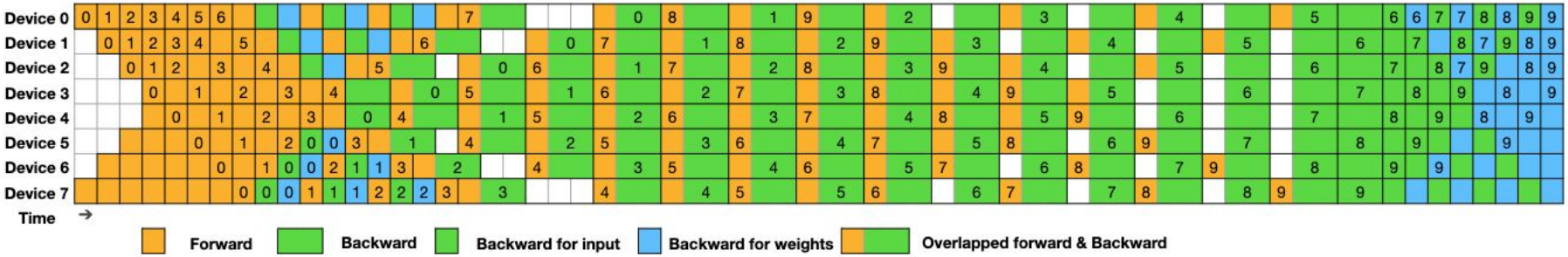


Figure 3: Handcrafted pipeline schedules, top: ZB-H1; bottom: ZB-H2



Method	Bubble	Parameter	Activation
1F1B	$(PP-1)(F+B)$	1×	PP
ZB1P	$(PP-1)(F+B-2W)$	1×	PP
DualPipe (Ours)	$(\frac{PP}{2}-1)(F\&B+B-3W)$	2×	$PP+1$

see you next time