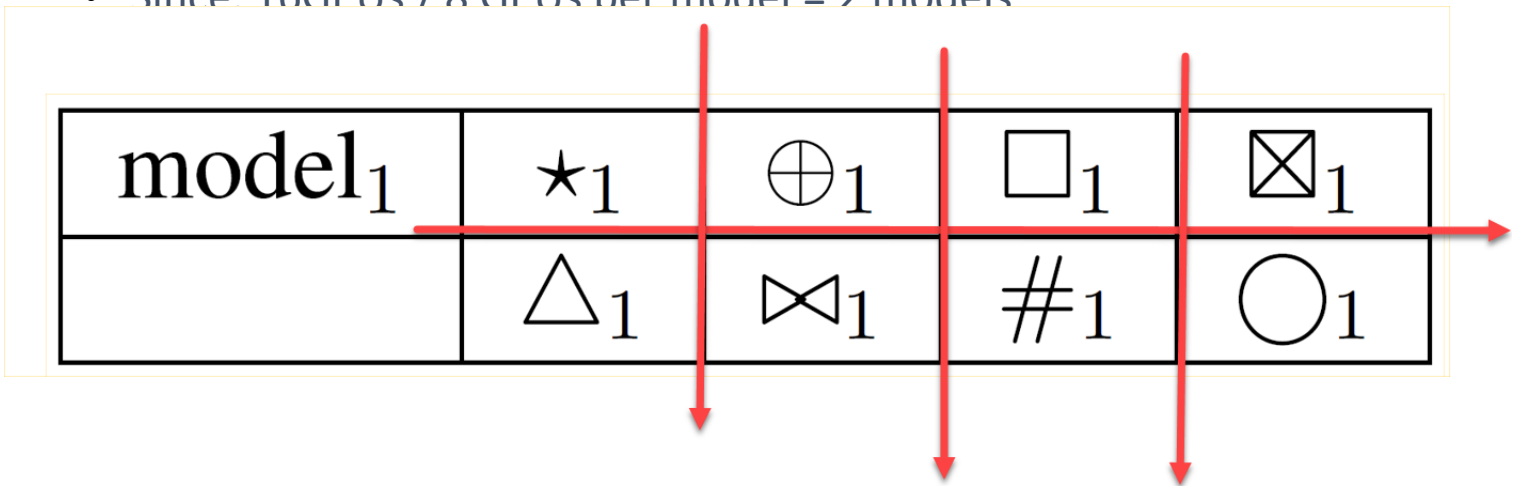
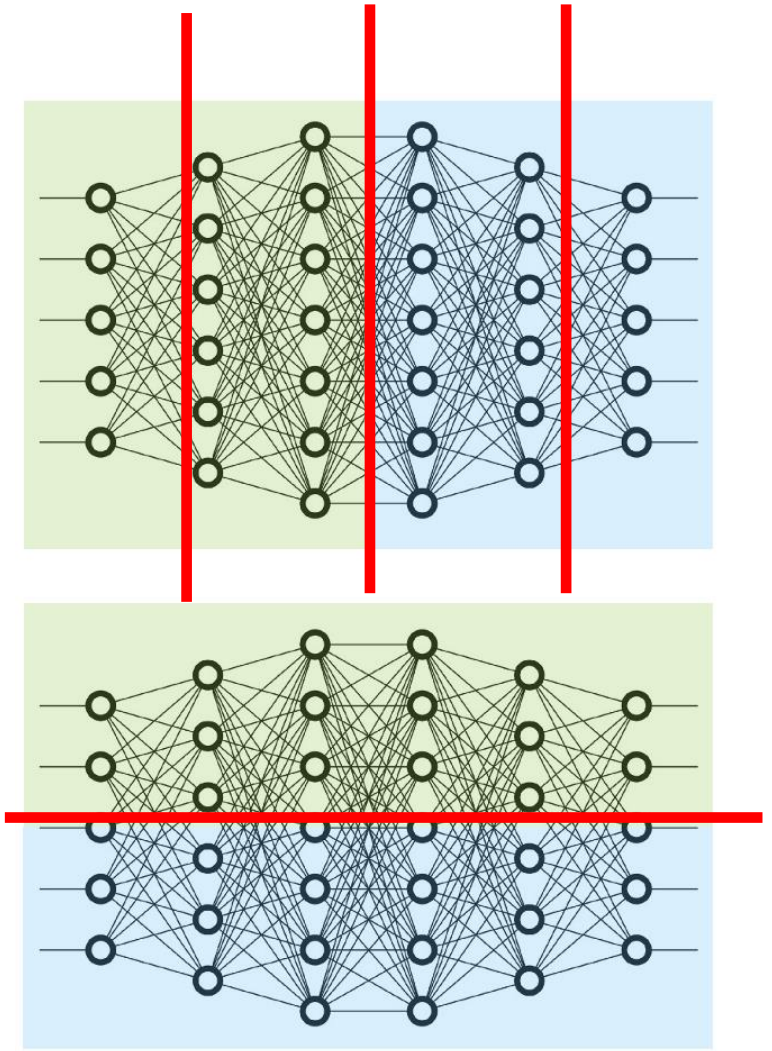
Model parallel in Megatron

- Xianchao Wu
- xianchaow@nvidia.com

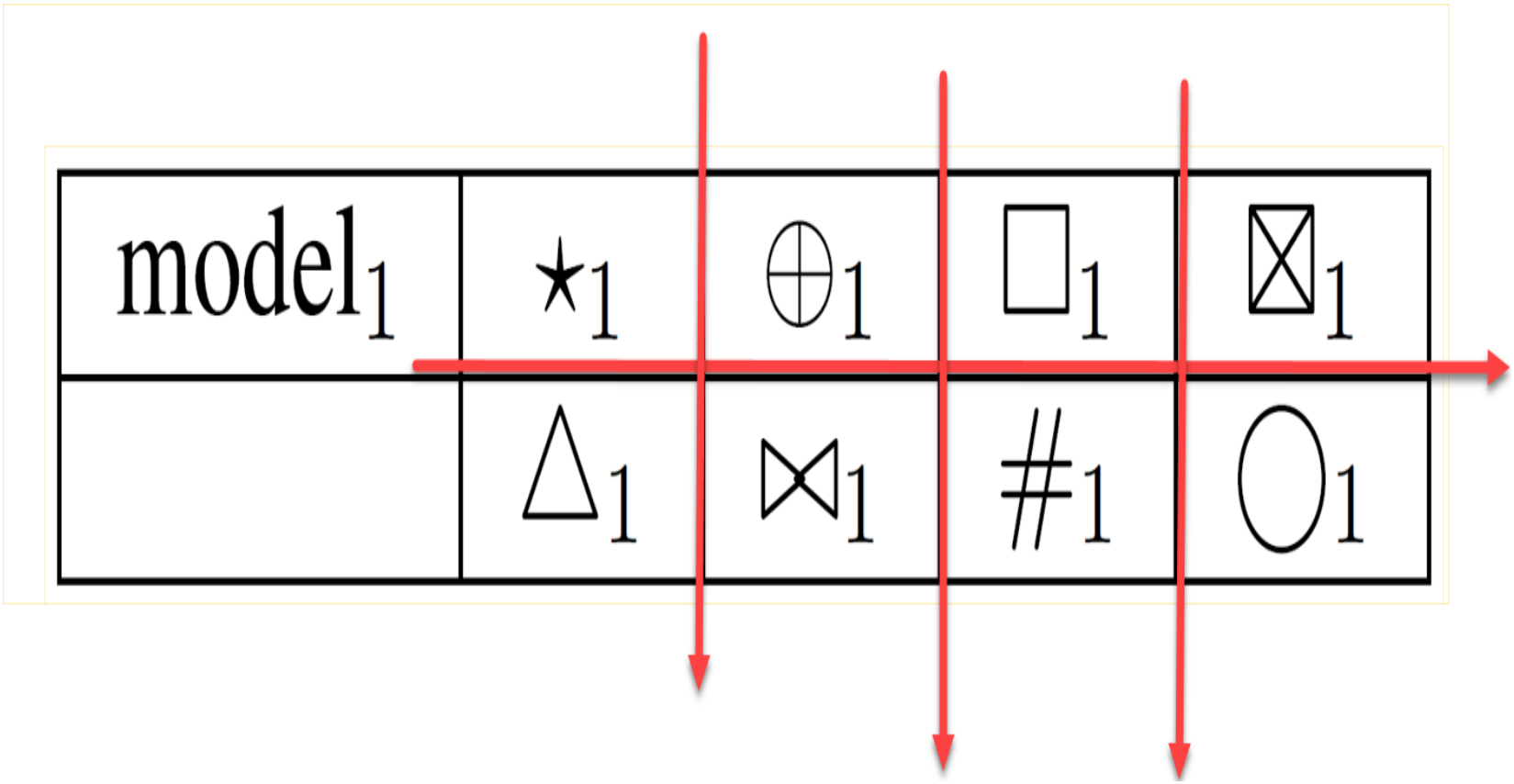
More complicated case

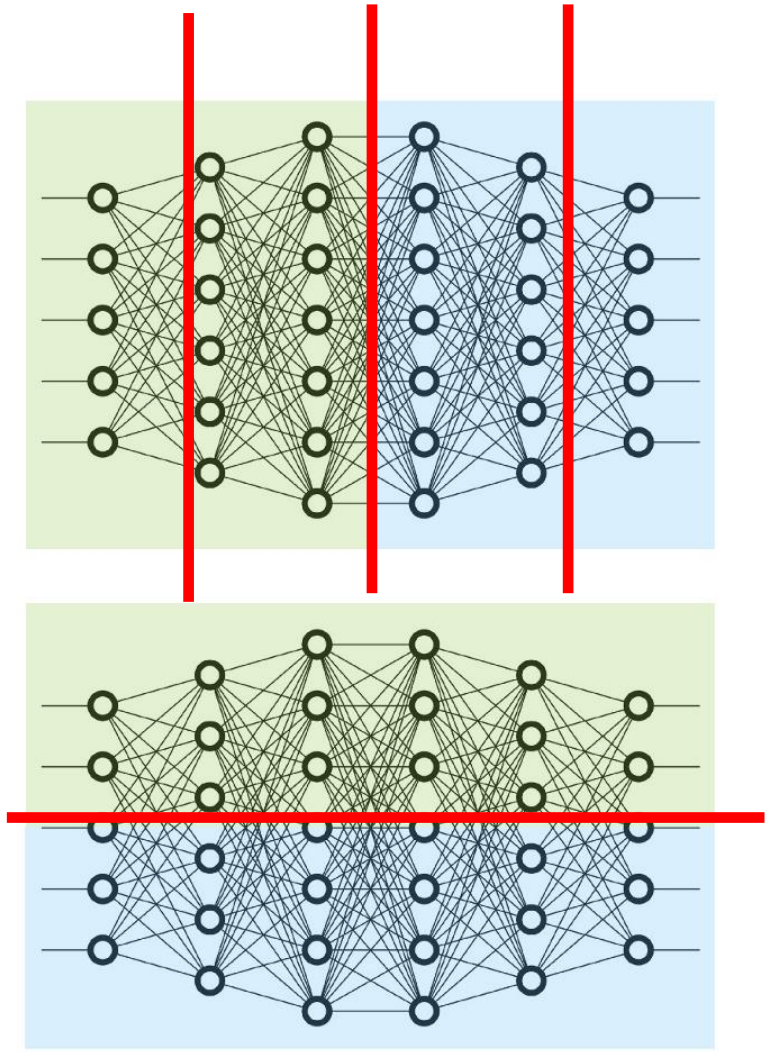
- 2 nodes, 8 GPUs per node -> 16 GPUs in total
- One model
  - Cut (row) 1 time -> tensor_parallel_size=2
  - cut (column) 3 times -> pipeline_parallel_size=4
  - So, 8 pieces for one model = one model takes 8 GPUs
- Thus, there are 2 complete model copies
  - Since, 16GPUs / 8 GPUs per model = 2 models





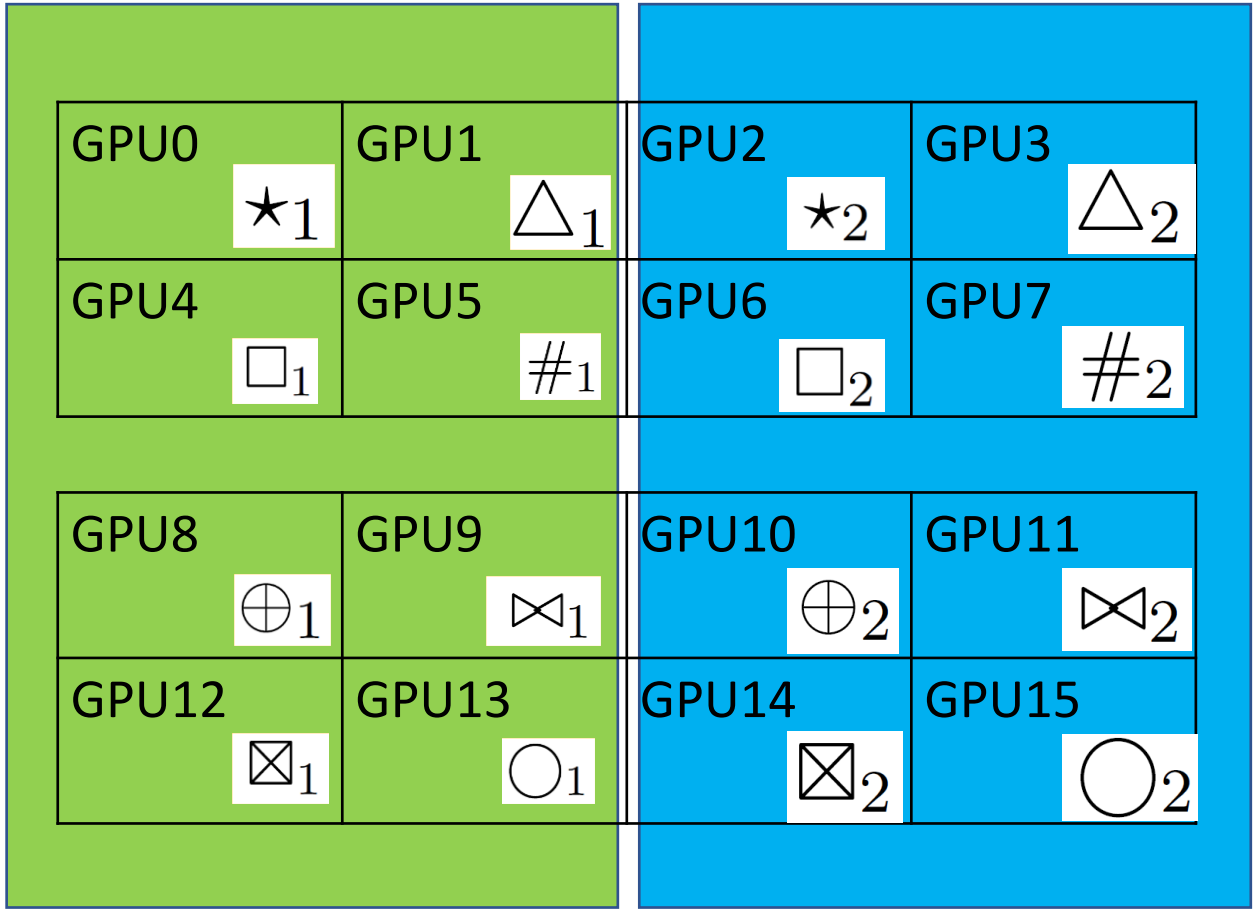
More complicated case



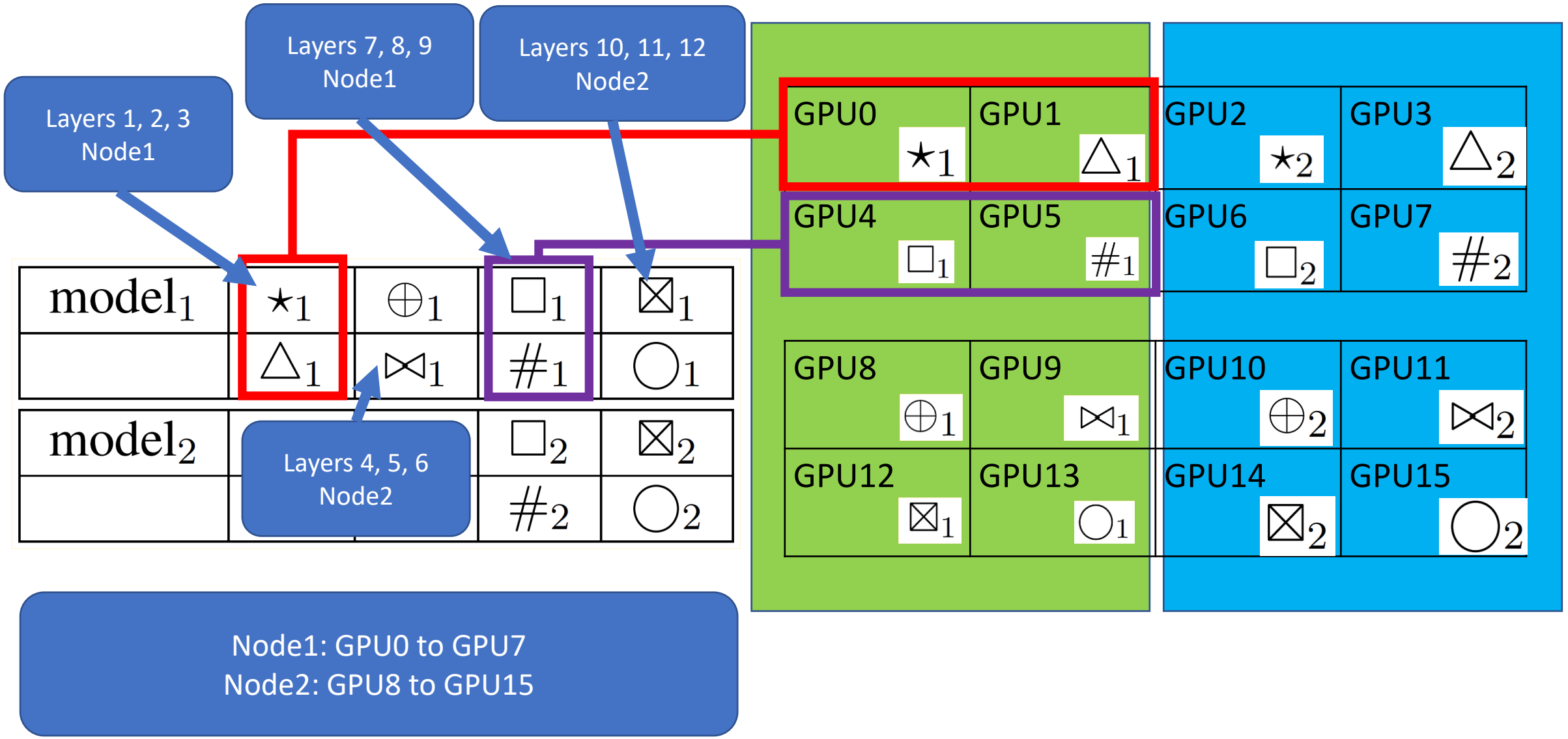


In the code (Megatron)

model_1	$\star_1$	$\oplus_1$	$\square_1$	$\boxtimes_1$
	$\triangle_1$	$\bowtie_1$	$\#_1$	$\bigcirc_1$
model_2	$\star_2$	$\oplus_2$	$\square_2$	$\boxtimes_2$
	$\triangle_2$	$\bowtie_2$	$\#_2$	$\bigcirc_2$



In the code (Megatron)

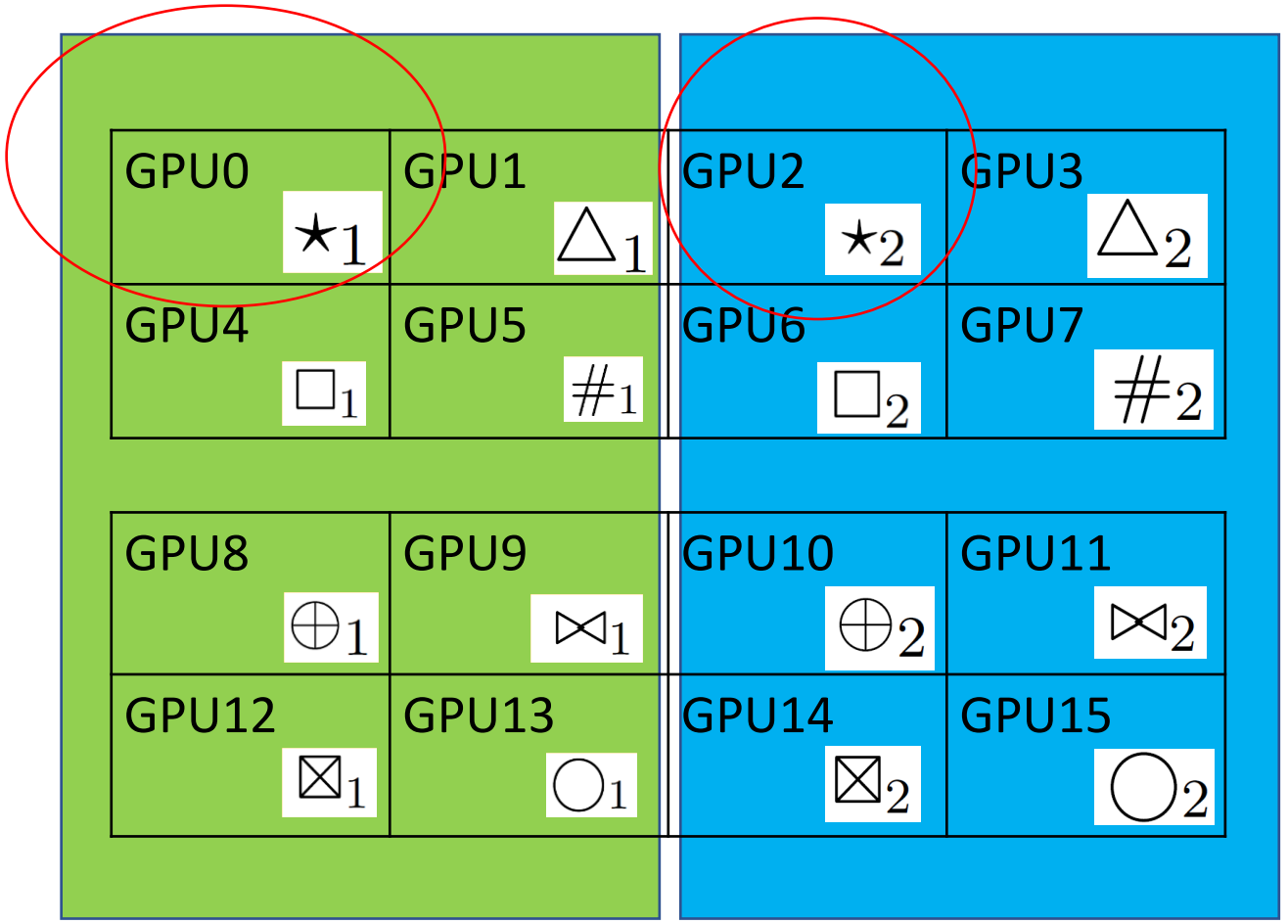


In the code (Megatron) – Data Parallel

model_1	$\star_1$	$\oplus_1$	$\square_1$	$\boxtimes_1$
	$\triangle_1$	$\bowtie_1$	$\#_1$	$\bigcirc_1$
model_2	$\star_2$	$\oplus_2$	$\square_2$	$\boxtimes_2$
	$\triangle_2$	$\bowtie_2$	$\#_2$	$\bigcirc_2$

- Data parallel groups:
- Model pieces with same parameters
- 0 2

Think: forget all others, just look at GPU0 and GPU2 -> then they are having the same model (piece) copy;
Then, we can send minibatches to them and update them

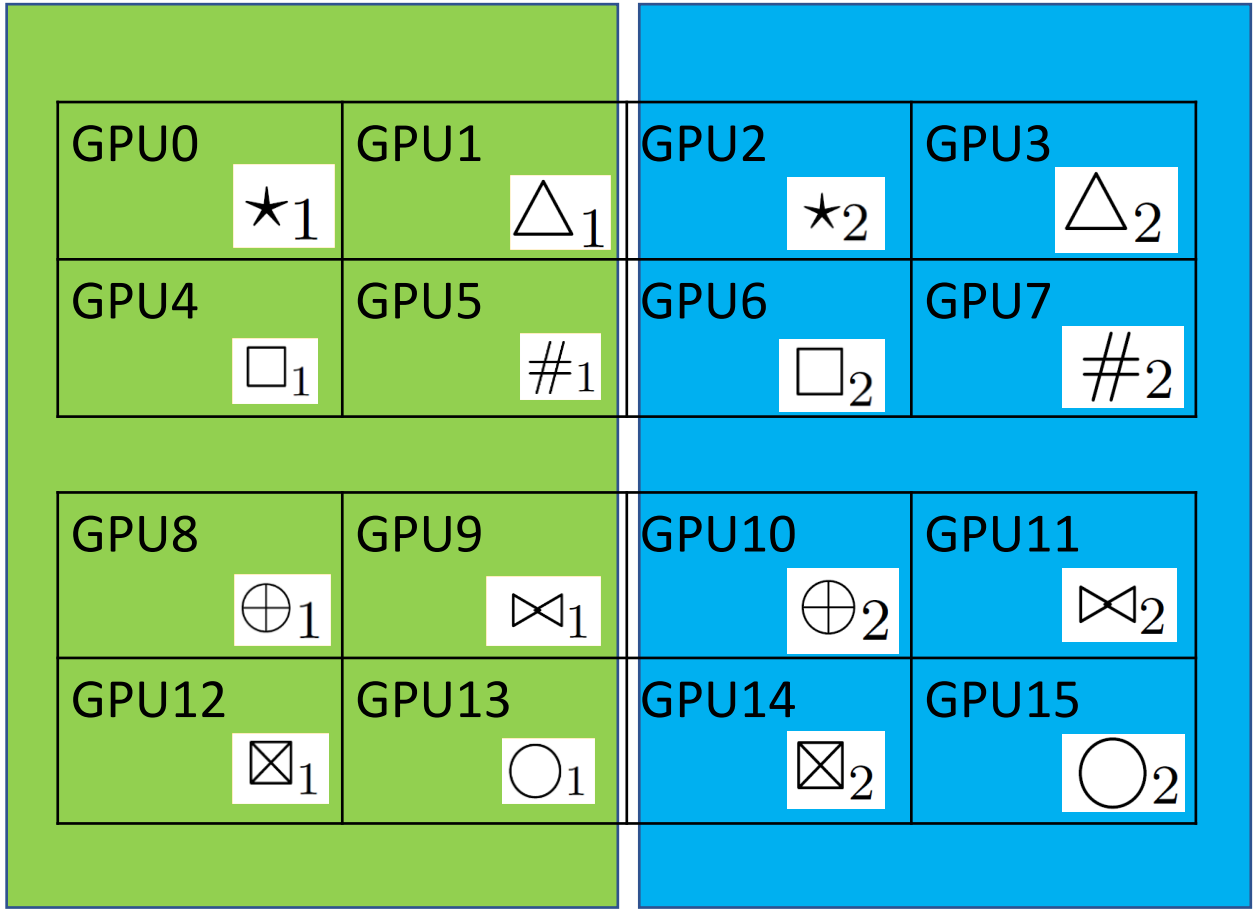


In the code (Megatron) – Data Parallel

model_1	$\star_1$	$\oplus_1$	$\square_1$	$\boxtimes_1$
	$\triangle_1$	$\bowtie_1$	$\#_1$	$\bigcirc_1$
model_2	$\star_2$	$\oplus_2$	$\square_2$	$\boxtimes_2$
	$\triangle_2$	$\bowtie_2$	$\#_2$	$\bigcirc_2$

- Data parallel groups:
- Model pieces with same parameters
- 0 2
- 1 3
- 4 6
- 5 7
- 8 10
- 9 11
- 12 14
- 13 15

8 data parallel groups
Each group is with a size of 2

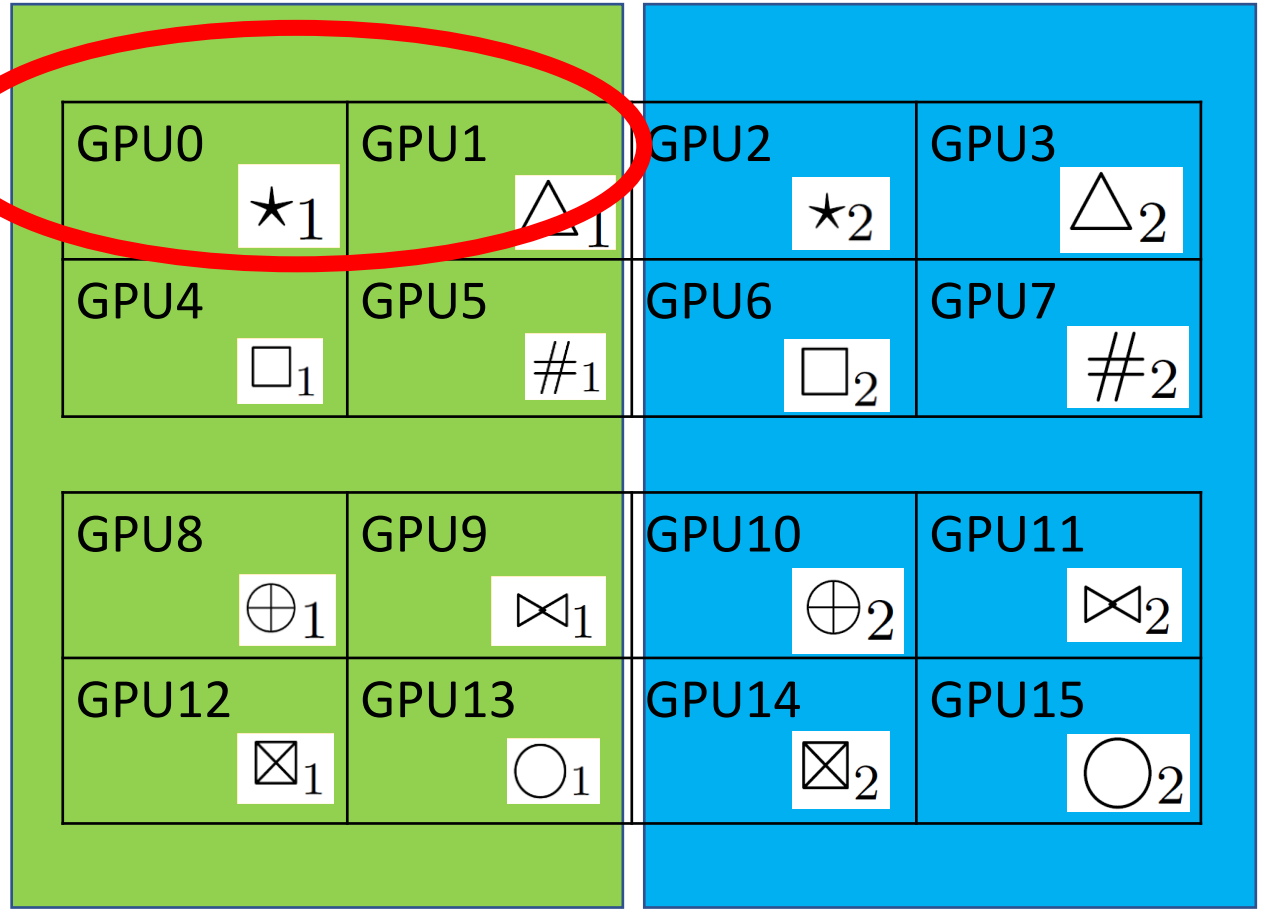


In the code (Megatron) – Tensor Model-Parallel

model_1	$\star_1$	$\oplus_1$	$\square_1$	$\boxtimes_1$
	$\triangle_1$	$\bowtie_1$	$\#_1$	$\bigcirc_1$
model_2	$\star_2$	$\oplus_2$	$\square_2$	$\boxtimes_2$
	$\triangle_2$	$\bowtie_2$	$\#_2$	$\bigcirc_2$

- **Tensor model-parallel** groups:
- 0 1
- 2 3
- 4 5
- 6 7
- 8 9
- 10 11
- 12 13
- 14 15

8 tensor model parallel groups
Each group is with a size of 2



In the code (Megatron) – Pipeline Model-Parallel

model_1	$\star_1$	$\oplus_1$	$\square_1$	$\boxtimes_1$
	$\triangle_1$	$\bowtie_1$	$\#_1$	$\bigcirc_1$
model_2	$\star_2$	$\oplus_2$	$\square_2$	$\boxtimes_2$
	$\triangle_2$	$\bowtie_2$	$\#_2$	$\bigcirc_2$

- **Pipeline model-parallel** groups:
  - 0 4 8 12
  - 1 5 9 13
  - 2 6 10 14
  - 3 7 11 15

4 pipeline model parallel groups
Each group is with a size of 4