分布式训练系列

张量自动并行

ZOMI

BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

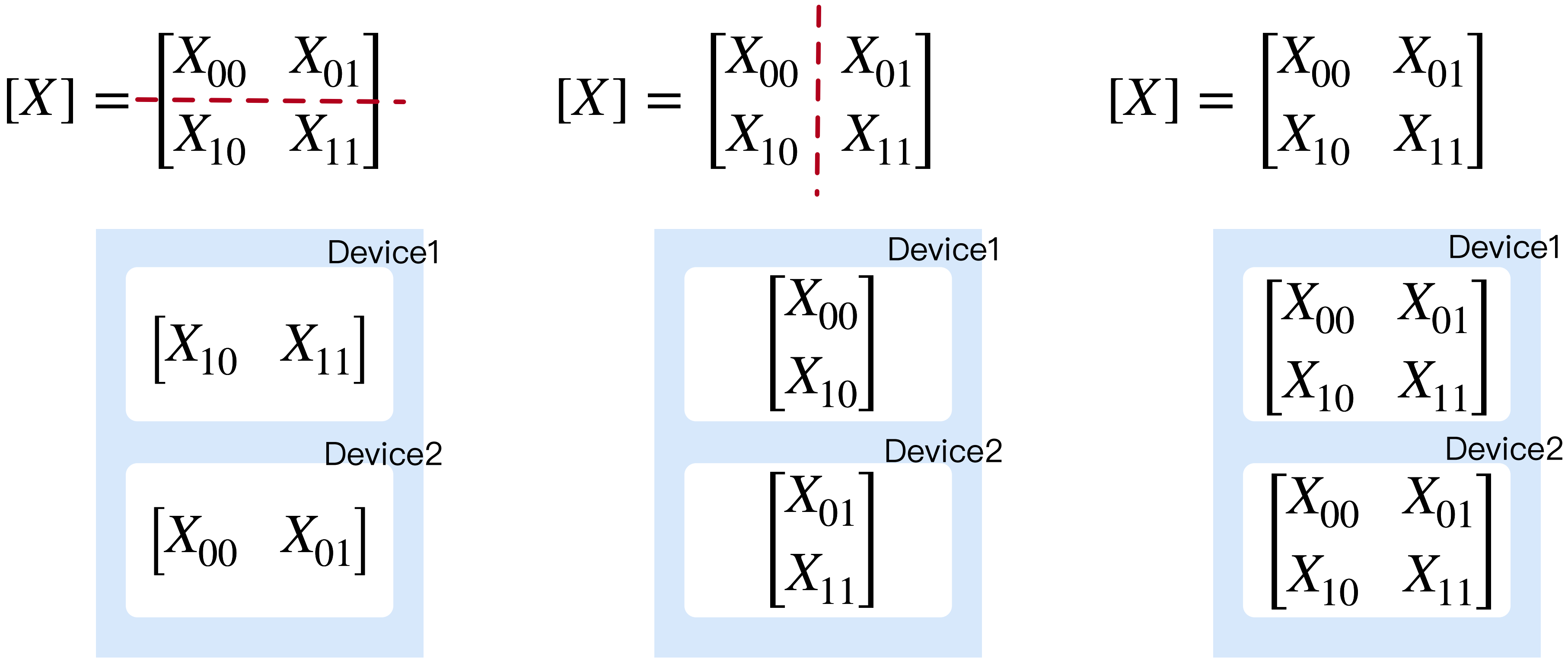
www.hiascend.com
www.mindspore.cn

Model Parallelism, MP 模型并行

- Tensor Parallelism 张量并行
  - Principles 并行原理
  - Matmul 算子并行
  - Loss 损失并行
  - Transformer 算子并行
  - Tensor Redistribution 张量重排（MindSpore）
  - Stochastic Control 随机控制
- Pipeline Parallelism 流水线并行

Mathematical Principles 数学原理

- 张量切分方式，双设备



$$[X] = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device1

$$\begin{bmatrix} X_{10} & X_{11} \end{bmatrix}$$

Device2

$$\begin{bmatrix} X_{00} & X_{01} \end{bmatrix}$$

行切分

$$[X] = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device1

$$\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$$

Device2

$$\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$$

列切分

$$[X] = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device1

$$\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

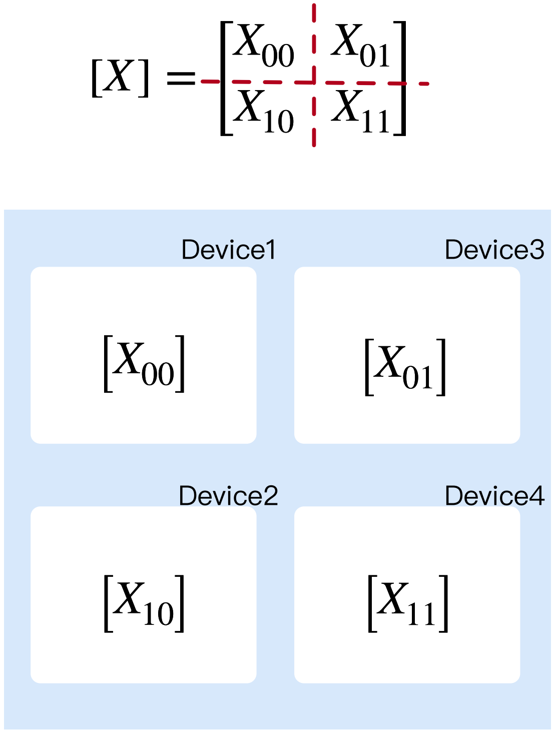
Device2

$$\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

复制

Mathematical Principles 数学原理

- 张量切分方式，四设备



$$[X] = \left[\begin{array}{c:c} X_{00} & X_{01} \\ \hdashline X_{10} & X_{11} \end{array}\right]$$

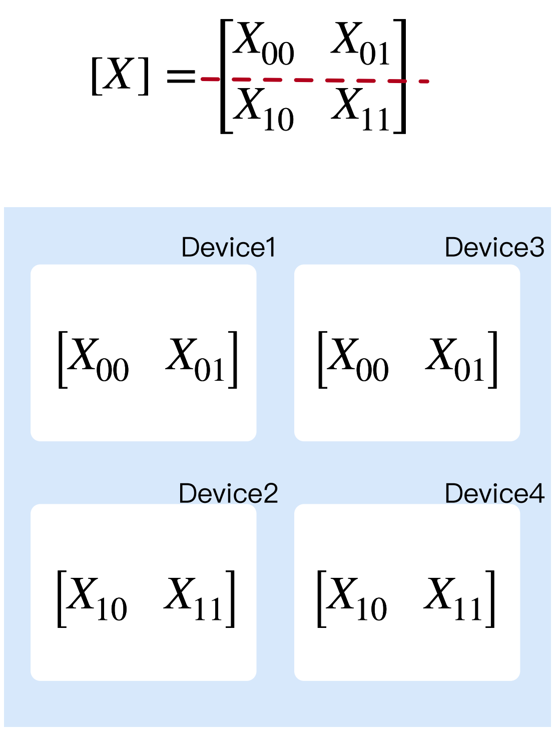
Device1 $[X_{00}]$

Device3 $[X_{01}]$

Device2 $[X_{10}]$

Device4 $[X_{11}]$

行列切分



$$[X] = \left[\begin{array}{cc} X_{00} & X_{01} \\ \hdashline X_{10} & X_{11} \end{array}\right]$$

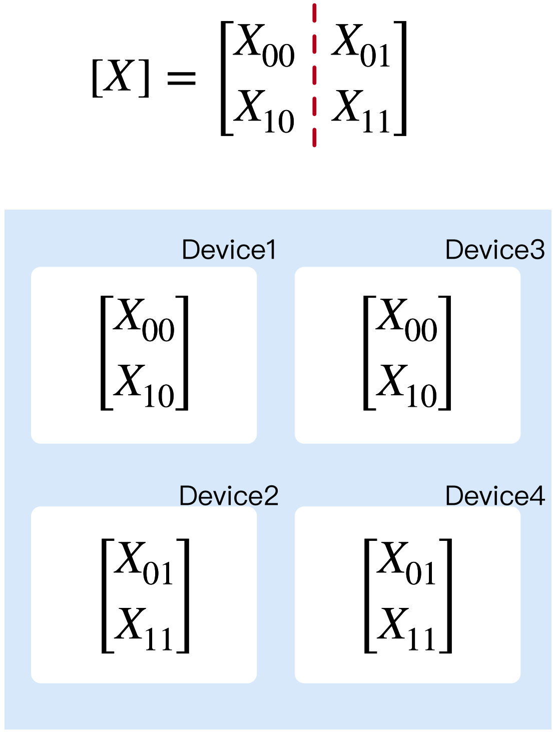
Device1 $[X_{00} \quad X_{01}]$

Device3 $[X_{00} \quad X_{01}]$

Device2 $[X_{10} \quad X_{11}]$

Device4 $[X_{10} \quad X_{11}]$

行切分+复制



$$[X] = \left[\begin{array}{c:c} X_{00} & X_{01} \\ X_{10} & X_{11} \end{array}\right]$$

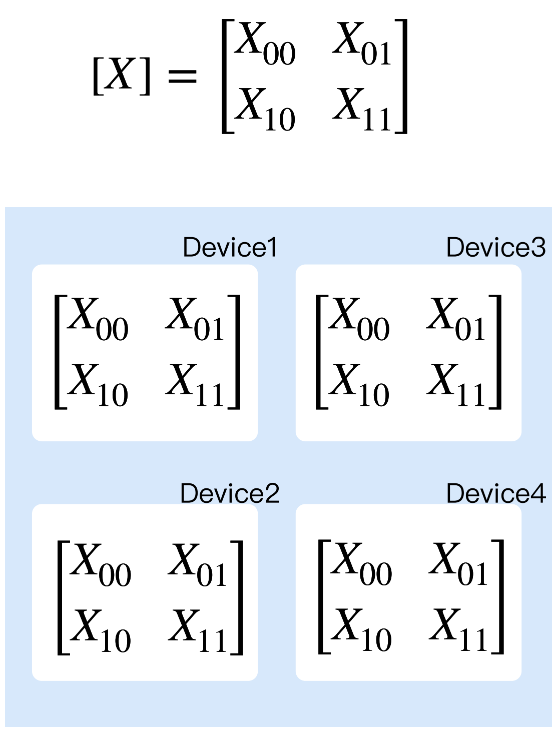
Device1 $\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$

Device3 $\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$

Device2 $\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$

Device4 $\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$

行列切分



$$[X] = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device1 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$

Device3 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$

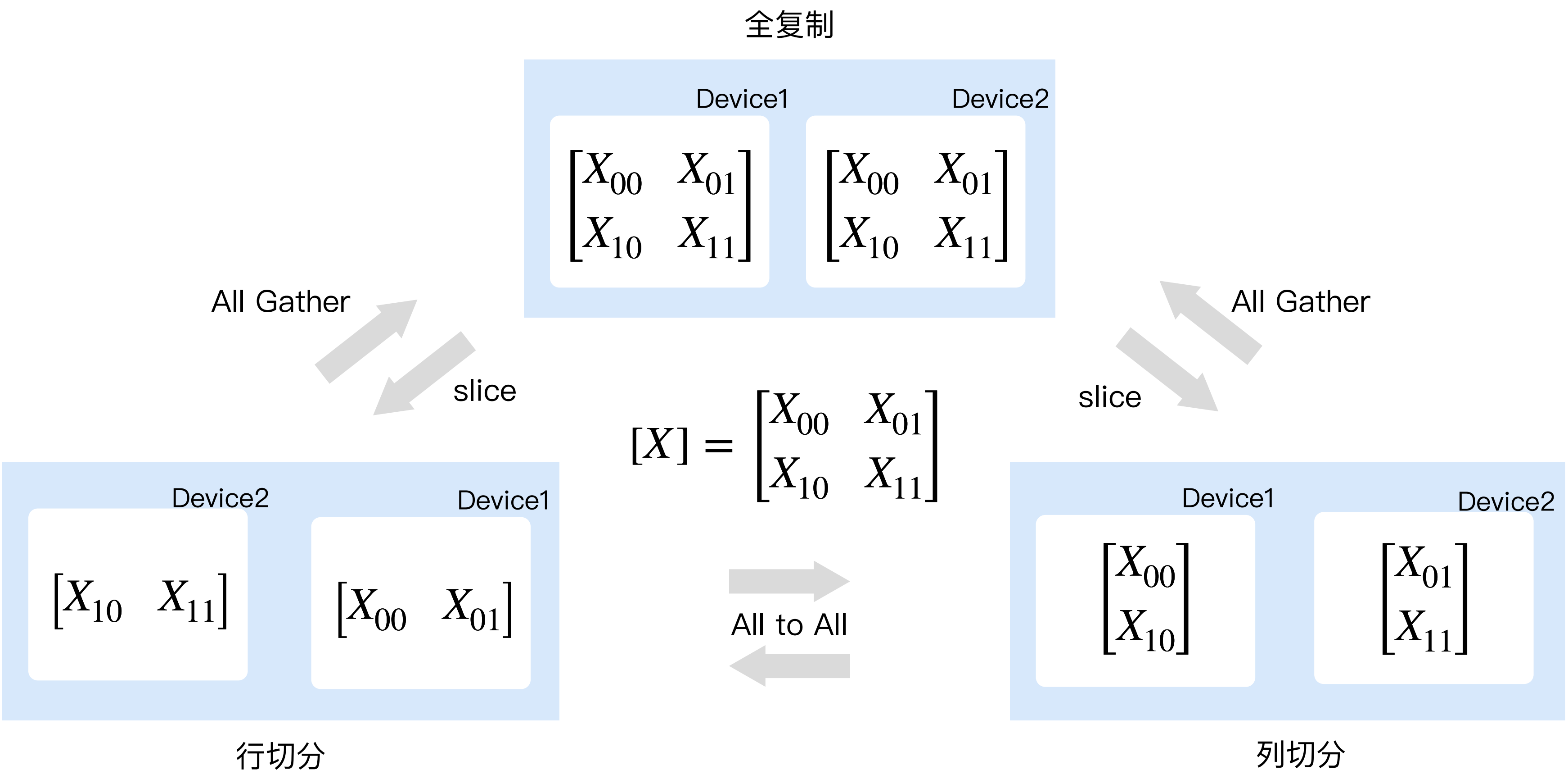
Device2 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$

Device4 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$

全复制

Mathematical Principles 数学原理

- 切分到两个节点的 Tensor 重排



全复制

Device1

$$\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device2

$$\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

All Gather

slice

All Gather

slice

$$[X] = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Device2

$$\begin{bmatrix} X_{10} & X_{11} \end{bmatrix}$$

Device1

$$\begin{bmatrix} X_{00} & X_{01} \end{bmatrix}$$

行切分

All to All

Device1

$$\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$$

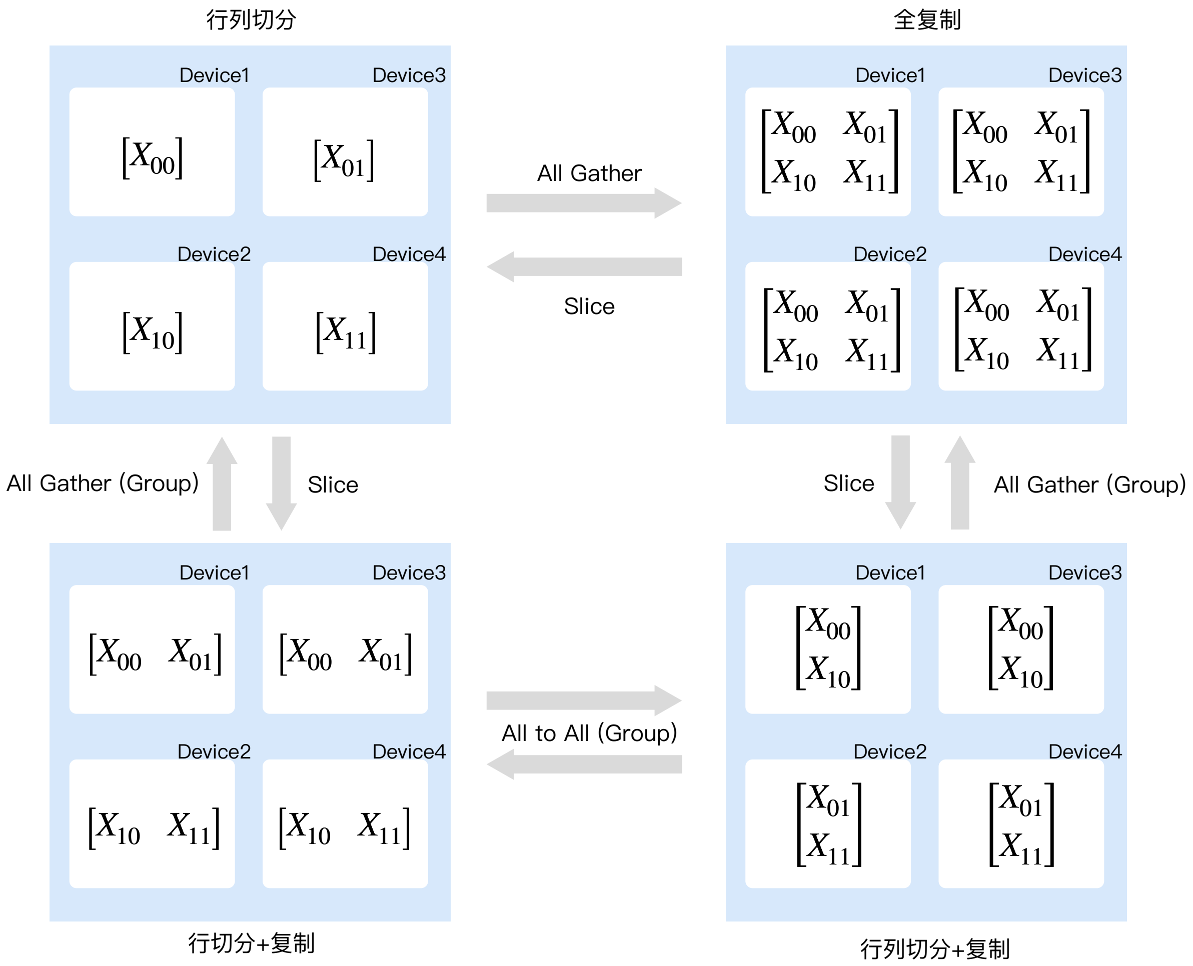
Device2

$$\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$$

列切分

Mathematical Principles 数学原理

- 切分到四个节点的 Tensor 重排


行列切分
Device1 $[X_{00}]$
Device3 $[X_{01}]$
Device2 $[X_{10}]$
Device4 $[X_{11}]$
All Gather
Slice
全复制
Device1 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$
Device3 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$
Device2 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$
Device4 $\begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$
All Gather (Group)
Slice
Slice
All Gather (Group)
Device1 $[X_{00} \quad X_{01}]$
Device3 $[X_{00} \quad X_{01}]$
Device2 $[X_{10} \quad X_{11}]$
Device4 $[X_{10} \quad X_{11}]$
行切分+复制
All to All (Group)
Device1 $\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$
Device3 $\begin{bmatrix} X_{00} \\ X_{10} \end{bmatrix}$
Device2 $\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$
Device4 $\begin{bmatrix} X_{01} \\ X_{11} \end{bmatrix}$
行列切分+复制

Tensor Redistribution

1 $$Y = XA = \begin{bmatrix}X_1\\X_2\\X_3\\X_4\end{bmatrix} \times A = \begin{bmatrix}X_1A\\X_2A\\X_3A\\X_4A\end{bmatrix} = \begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\end{bmatrix}$$

2 $$Z = YB = Y \times \begin{bmatrix}B_1\\B_2\\B_3\\B_4\end{bmatrix}^T = \begin{bmatrix}YB_1\\YB_2\\YB_3\\YB_4\end{bmatrix} = \begin{bmatrix}Z_1\\Z_2\\Z_3\\Z_4\end{bmatrix}^T$$

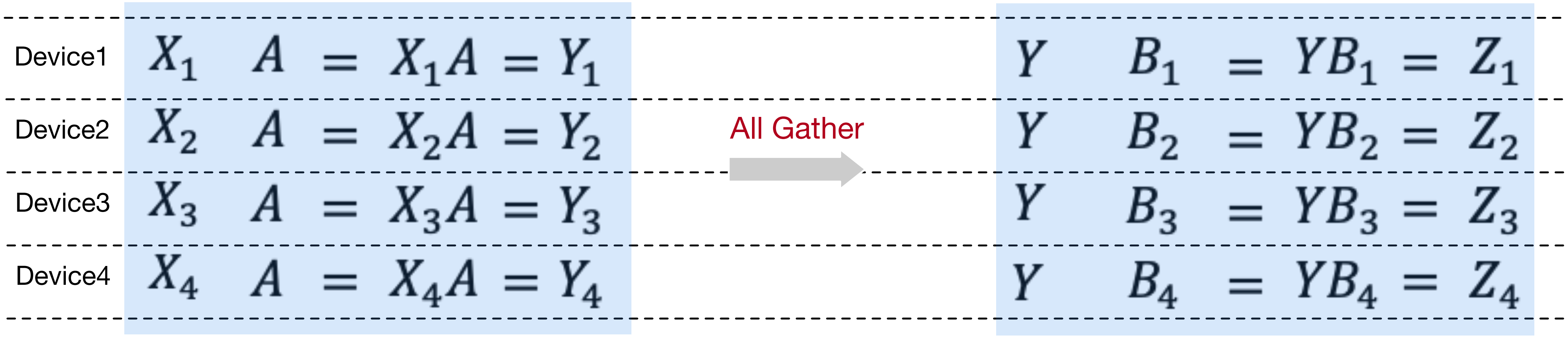
$Y = XA$

X 行切分

Tensor Redistribution
张量重排

$Z = YB$

B 列切分



Device1	X_1	A	$=$	X_1A	$= Y_1$		Y	B_1	$=$	YB_1	$=$	Z_1
Device2	X_2	A	$=$	X_2A	$= Y_2$	All Gather	Y	B_2	$=$	YB_2	$=$	Z_2
Device3	X_3	A	$=$	X_3A	$= Y_3$		Y	B_3	$=$	YB_3	$=$	Z_3
Device4	X_4	A	$=$	X_4A	$= Y_4$		Y	B_4	$=$	YB_4	$=$	Z_4

Tensor Redistribution

1 $Y = XA = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \times A = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}^T = \begin{bmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{bmatrix}$

2 $Z = YB = \begin{bmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$

$Y = XA$

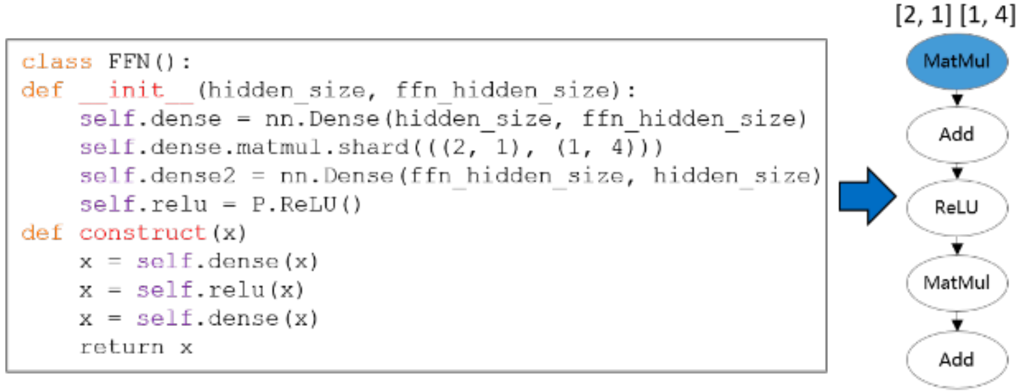
X 行切分、A 列切分

$Z = YB$

Y 行列切分、B 行切分

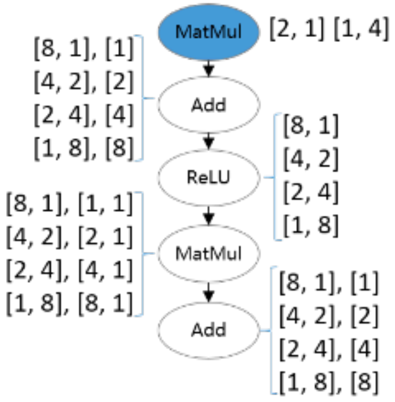
Device1	$X_1 \quad A_1 = X_1 A_1 = Y_{00}$	$=$	$Y_{00} \quad B_1 = Y_{00} B_1$	All Reduce	Z_1
Device2	$X_1 \quad A_2 = X_1 A_2 = Y_{01}$	$=$	$Y_{01} \quad B_2 = Y_{01} B_2$		Z_1
Device3	$X_2 \quad A_1 = X_2 A_1 = Y_{10}$	$=$	$Y_{10} \quad B_1 = Y_{10} B_1$	All Reduce	Z_2
Device4	$X_2 \quad A_2 = X_2 A_2 = Y_{11}$	$=$	$Y_{11} \quad B_2 = Y_{11} B_2$		Z_2

MindSpore Tensor Sharded Strategy

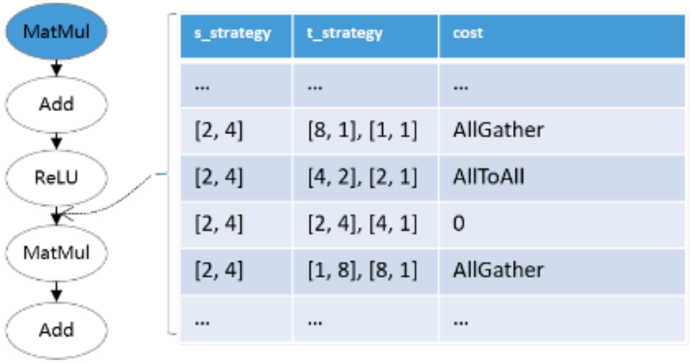


```
class FFN():
def __init__(hidden_size, ffn_hidden_size):
    self.dense = nn.Dense(hidden_size, ffn_hidden_size)
    self.dense.matmul.shard(((2, 1), (1, 4)))
    self.dense2 = nn.Dense(ffn_hidden_size, hidden_size)
    self.relu = P.ReLU()
def construct(x):
    x = self.dense(x)
    x = self.relu(x)
    x = self.dense(x)
    return x
```

(a) 由模型定义脚本转换成带有切分策略的计算图

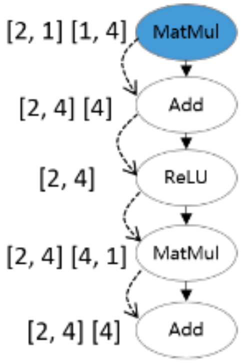


(b) 为每个未配置切分策略的算子枚举可行的策略



s_strategy	t_strategy	cost
...	...	...
[2, 4]	[8, 1], [1, 1]	AllGather
[2, 4]	[4, 2], [2, 1]	AllToAll
[2, 4]	[2, 4], [4, 1]	0
[2, 4]	[1, 8], [8, 1]	AllGather
...	...	...

(c) 枚举每条边的重排布策略和相应的代价，这里只列了ReLU->MatMul这条边的部分策略



(d) 由已配置策略的算法出发，传播到整张计算图

Summary 总结

1. 模型并行分为张量并行和流水线并行，张量并行主要层内并行、流水线主要层间并行，一般来说机内使用张量并行，机间使用数据并行；
2. 张量并行主要是对数据进行切分，切分方式有行（Row）切分和列（Col）切分，而通过复制组合可以形成多种通信形式；
3. 张量并行最常见的是 MatMul 算子并行，通过 MatMul 可以拓展到 Embedding、MLP、 Transformer等算子并行；
4. 张量并行的时候值得注意的是随机性问题，需要注意带有随机性算子的随机种子设置；

Inference

1. https://zhuanlan.zhihu.com/p/450854172 全网最全-超大模型+分布式训练架构和经典论文
2. https://developer.nvidia.com/blog/training-a-recommender-system-on-dgx-a100-with-100b-parameters-in-tensorflow-2/
3. https://developer.nvidia.com/blog/fast-terabyte-scale-recommender-training-made-easy-with-nvidia-merlin-distributed-embeddings/
4. https://www.mindspore.cn/docs/zh-CN/r1.7/design/operator_parallel.html
5. https://www.mindspore.cn/docs/zh-CN/r1.7/design/distributed_training_design.html
6. https://colossalai.org/zh-Hans/docs/features/2D_tensor_parallel/
7. https://zhuanlan.zhihu.com/p/507877303
8. https://zhuanlan.zhihu.com/p/450689346
9. https://zhuanlan.zhihu.com/p/497672789

BUILDING A BETTER CONNECTED WORLD

THANK YOU

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.