ASPLOS

Institute of Computing Technology, CAS

Research Center for Intelligent Computing Systems

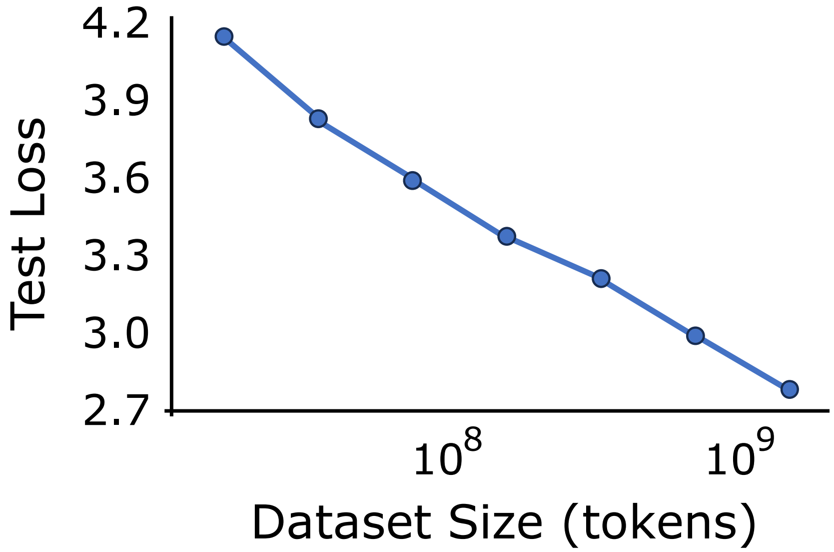
PrimePar: Efficient Spatial-temporal Tensor Partitioning for Large Transformer Model Training

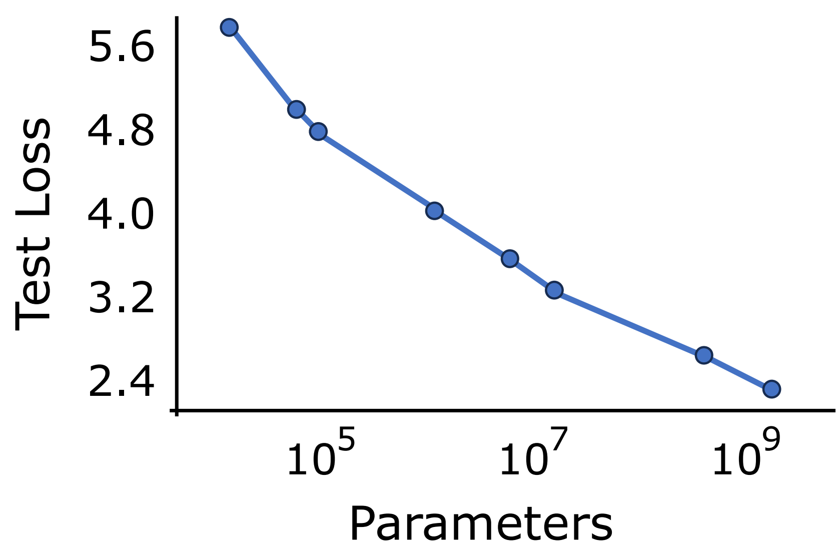
Haoran Wang, Lei Wang, Haobo Xu, Ying Wang, Yuming Li, Yinhe Han

Research Center for Intelligent Computing Systems
Institute of Computing Technology, Chinese Academy of Sciences

ASPLOS 2024

Training Large Language Models (LLM) is Challenging





Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv:2001.08361, 2020.

Model	Corpus size	Model Parameters
GPT	800M tokens	117M
GPT-3	300B tokens	175B
Llama 2	2T tokens	70B
Llama 3	15T tokens	70B

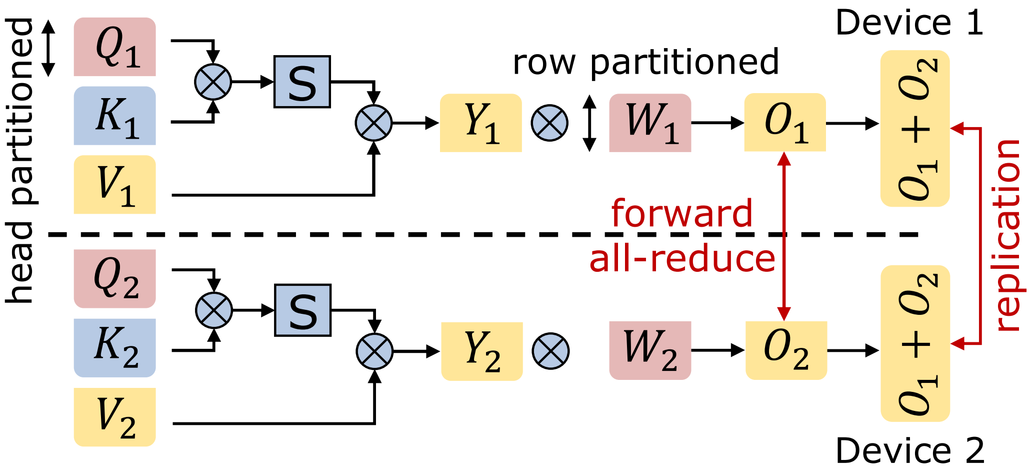
LLM training:

- Larger dataset size
- Larger model parameter size

Training LLM is Challenging

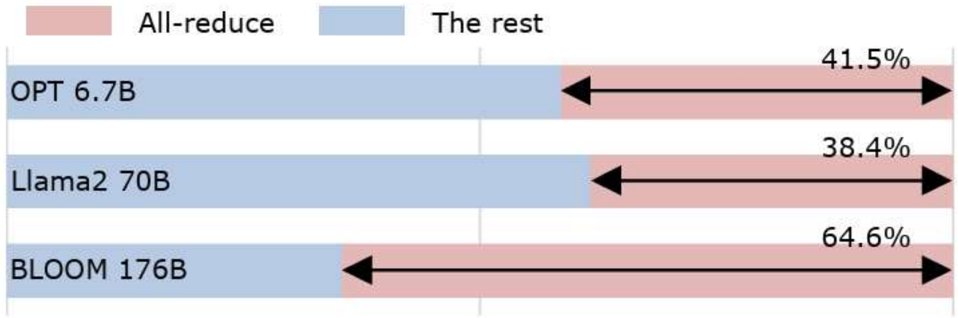
3D parallelism → Pipeline parallelism

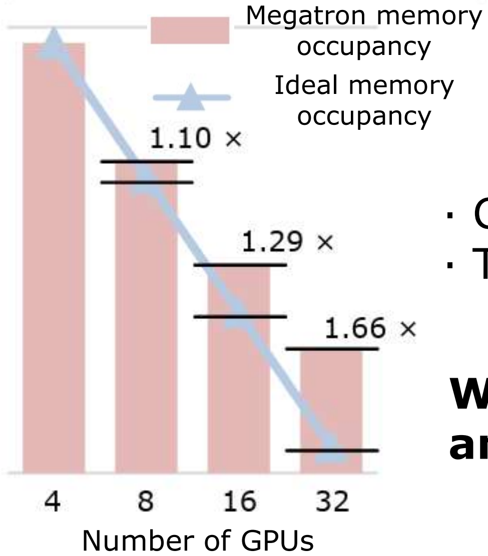
Tensor partition: data/model parallelism



SOTA tensor partition of attention layer

Shoeybi M, Patwary M, Puri R, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism[J]. arXiv preprint arXiv:1909.08053, 2019.





Inefficiencies:

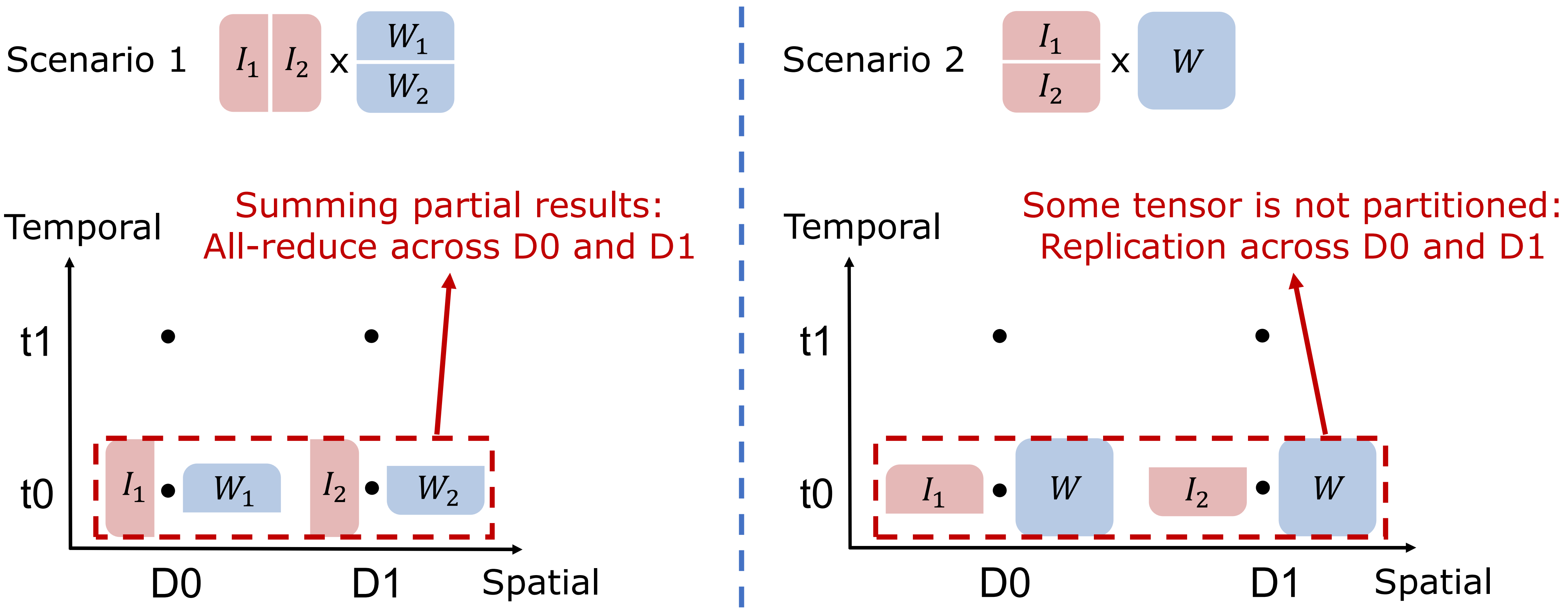
- Collective communication
- Tensor replication

Worse for larger model and greater parallelism

Focus of this work: better tensor partition with less collective communication and tensor replication

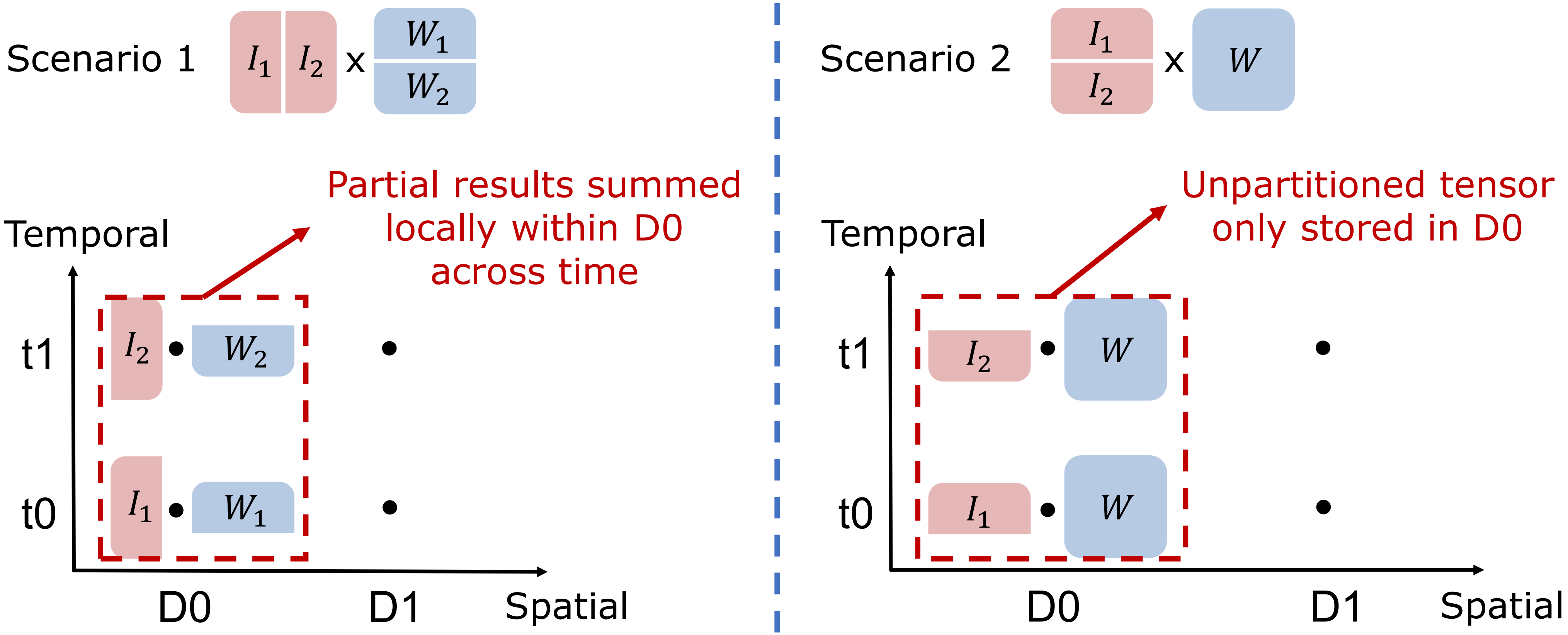
Motivational ideas

Distributing sub-operators along spatial dimension



Motivational ideas

Distributing sub-operators along temporal dimension provides extra opportunities

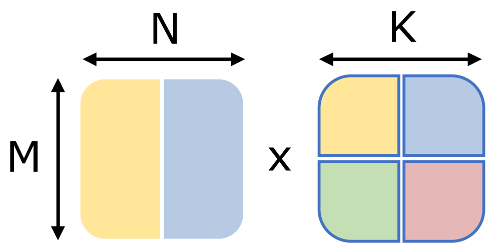


Optimizing both throughput and memory footprint

Tensor Partition Notations

Spatial index: device ID $\boldsymbol{D} = (d_1, d_2, \ldots, d_n),\ d_i = 0, 1$
Temporal index: $t = 0,1,2, \ldots$
Dimension slice index (DSI): $I_X(\boldsymbol{D}, t)$

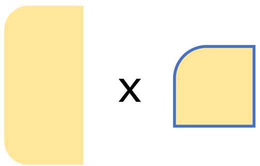
Example:



Given DSIs:

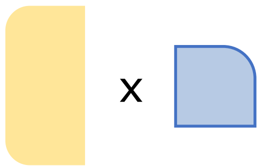
$I_M(\boldsymbol{D}, t) = 0$
$I_N(\boldsymbol{D}, t) = d_1$
$I_K(\boldsymbol{D}, t) = d_2$

Device $(d_1 = 0, d_2 = 0)$



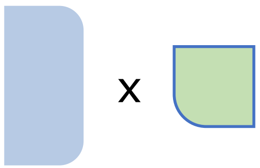
$I_M = 0$
$I_N = d_1 = 0$
$I_K = d_2 = 0$

Device $(d_1 = 0, d_2 = 1)$



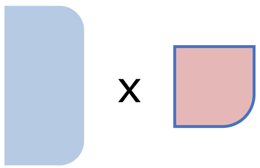
$I_M = 0$
$I_N = d_1 = 0$
$I_K = d_2 = 1$

Device $(d_1 = 1, d_2 = 0)$



$I_M = 0$
$I_N = d_1 = 1$
$I_K = d_2 = 0$

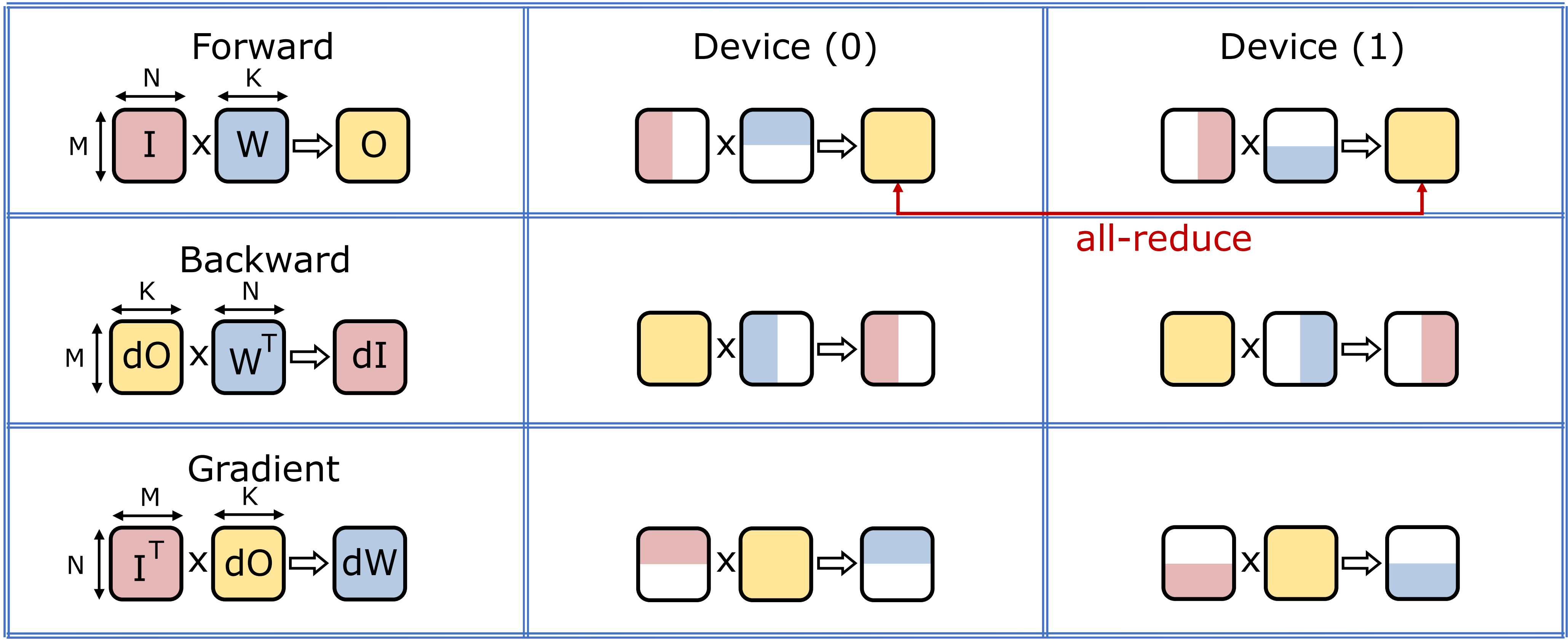
Device $(d_1 = 1, d_2 = 1)$



$I_M = 0$
$I_N = d_1 = 1$
$I_K = d_2 = 1$

Existing Spatial Tensor Partition

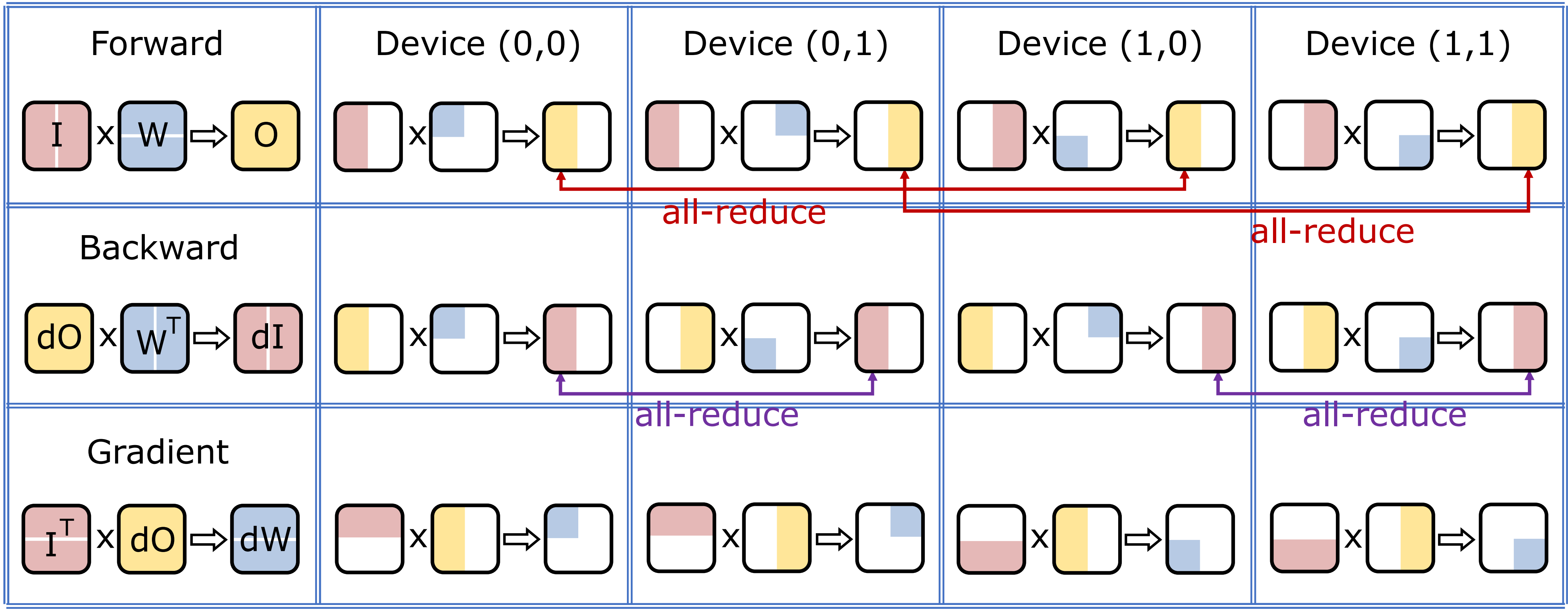
Partition dimension N $I_M^F = I_M^B = I_M^G = 0$ $I_N^F = I_N^B = I_N^G = d_1$ $I_K^F = I_K^B = I_K^G = 0$



Each time choose one dimension to partition and partition recursively

Existing Spatial Tensor Partition

Partition dimension K $\quad I_M^F = I_M^B = I_M^G = 0 \quad I_N^F = I_N^B = I_N^G = d_1 \quad I_K^F = I_K^B = I_K^G = d_2$



Each time choose one dimension to partition and partition recursively

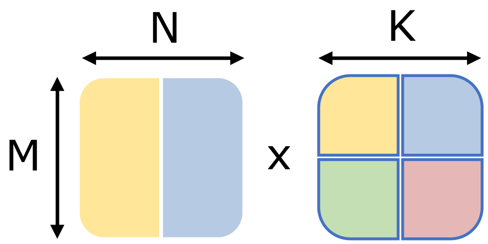
Essence of Inefficiencies: the DSI perspective

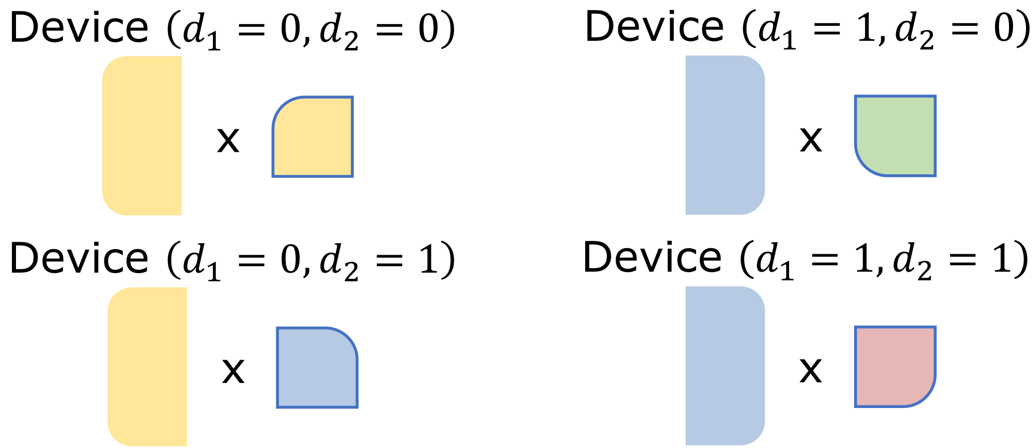
Spatial

$I_M = 0$

$I_N = d_1$

$I_K = d_2$





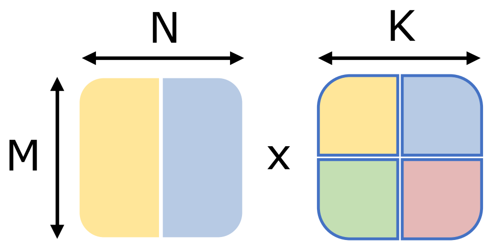
Essence of Inefficiencies: the DSI perspective

Spatial

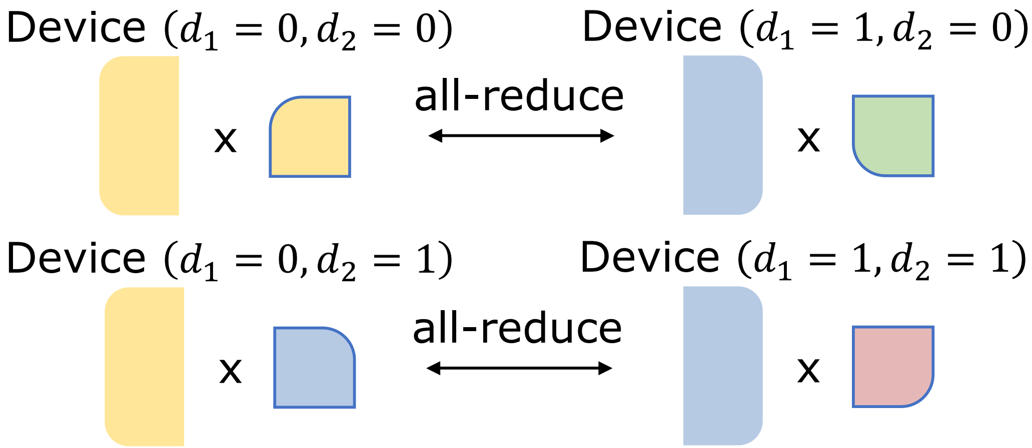
$I_M = 0$

❶ $\boxed{I_N = d_1}$

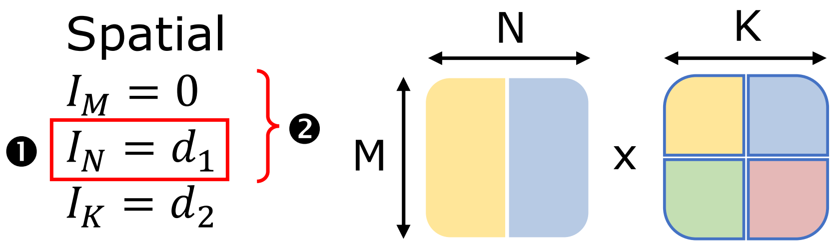
$I_K = d_2$



❶ Different d_1: all-reduce

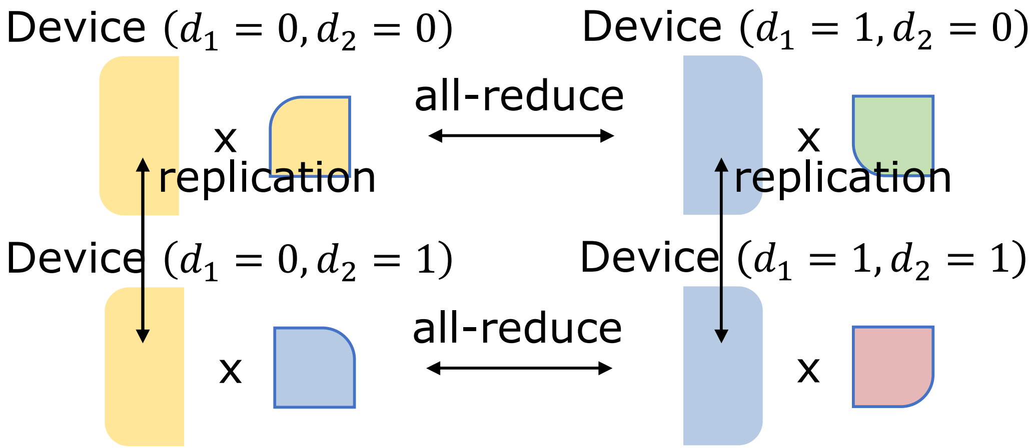


Essence of Inefficiencies: the DSI perspective



❶ Different d_1: all-reduce

❷ Same d_1 and different d_2: replication



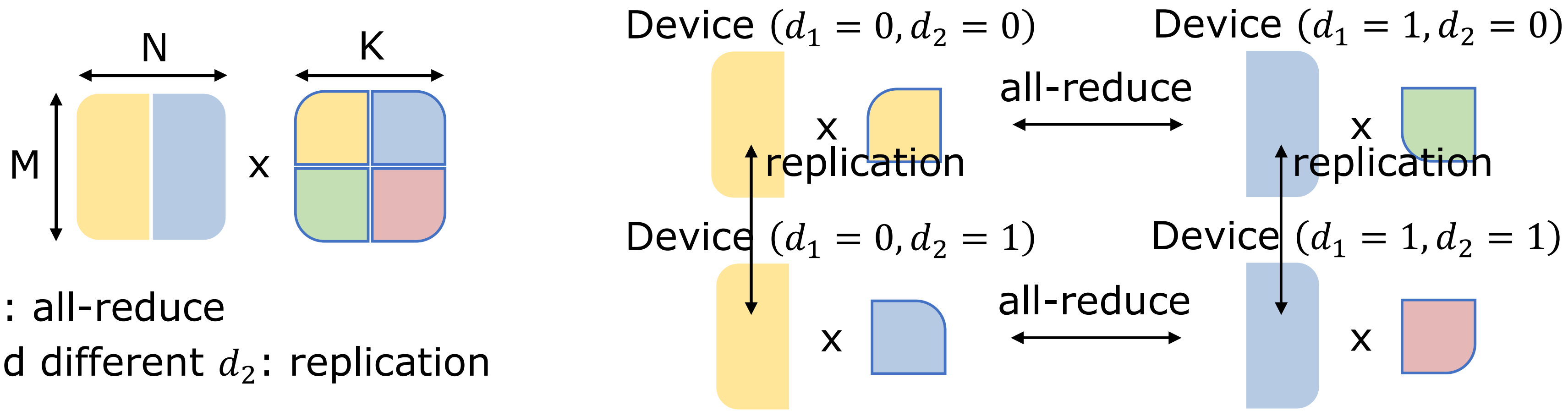
Essence of Inefficiencies: the DSI perspective

Spatial

$I_M = 0$

❶ $\boxed{I_N = d_1}$ ❷

$I_K = d_2$



❶ Different d_1: all-reduce

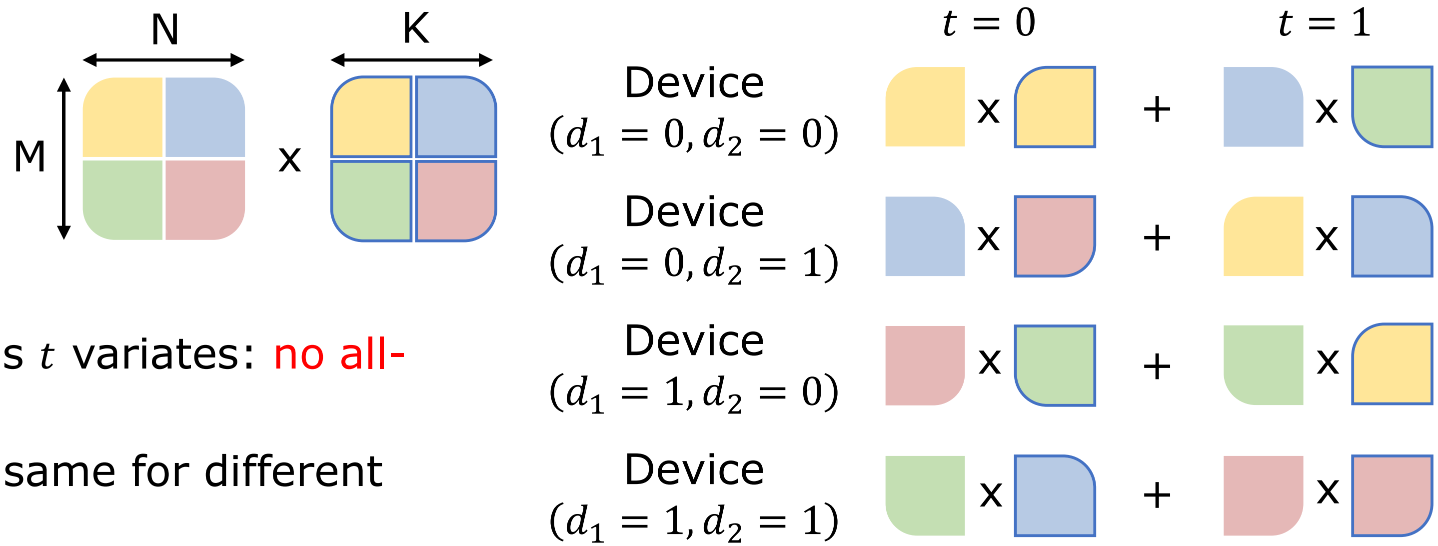
❷ Same d_1 and different d_2: replication

Spatial-temporal

$I_M = d_1 \ mod\ 2$

❶ $\boxed{I_N = (d_1 + d_2 + t) \ mod\ 2}$ ❷

$I_K = d_2 \ mod\ 2$



❶ I_N takes all possible values as t variates: no all-reduce

❷ Fixing t, (I_M, I_N) can't be the same for different devices: no replication

Novel Spatial-temporal Tensor Partition Primitive

Regard 2^{2k} devices as a square with row and column indices $0 \le r, c < 2^k$

Temporal index $0 \le t < 2^k$

Forward

$I_M = r \bmod 2^k$

$\boxed{I_N = (r + c + t) \bmod 2^k}$ ❶

$I_K = c \bmod 2^k$

❸

Backward

$I_M = r \bmod 2^k$

$\left.\begin{array}{l} I_N = (r + c - 1) \bmod 2^k \\ \boxed{I_K = (c + t) \bmod 2^k} \text{ ❶} \end{array}\right\}$ ❷

Gradient

$\boxed{I_M = (r + t) \bmod 2^k}$ ❶

$I_N = \left(r + c - 1 + \delta_{t,2^k-1}\right) \bmod 2^k$

$I_K = \left(c - 1 + \delta_{t,2^k-1}\right) \bmod 2^k$

❶ Collective communication free:
Summed-over dimensions take all possible values when t variates

❷ No tensor replication:

$$\begin{cases} (r + c - 1) \equiv (r' + c' - 1) \bmod 2^k \\ (c' + t) \equiv (c + t) \bmod 2^k \end{cases} \longrightarrow r = r', c = c'$$

❸ Continuity between training phases:

Forward last step

$I_N = \left(r + c + 2^k - 1\right) \bmod 2^k \quad I_K = c \bmod 2^k$

Backward first step

$I_N = (r + c - 1) \bmod 2^k \quad I_K = (c + 0) \bmod 2^k$

match

Example: k = 2   Forward step t = 0

Communicate tensor I

$(r, c+1, t)$: $\xrightarrow{\text{from right}}$ $(r, c, t+1)$:

$I_M = r \bmod 4$ $\quad$ $I_M = r \bmod 4$

$I_N = (r + c + 1 + t) \bmod 4$ $\xleftrightarrow{\text{match}}$ $I_N = (r + c + t + 1) \bmod 4$

Communicate tensor $\boldsymbol{W}$

$(r+1, c, t)$: $\xrightarrow{\text{from bottom}}$ $(r, c, t+1)$:

$I_N = (r + 1 + c + t) \bmod 4$ $\quad$ $I_N = (r + c + t + 1) \bmod 4$

$I_K = c \bmod 4$ $\xleftrightarrow{\text{match}}$ $I_K = c \bmod 4$

Forward DSIs

$I_M = r \bmod 4$

$I_N = (r + c + t) \bmod 4$

$I_K = c \bmod 4$

c →, r ↓

$O_{0,0} \mathrel{+}= I_{0,0} \times \boldsymbol{W}_{0,0}$	$O_{0,1} \mathrel{+}= I_{0,1} \times \boldsymbol{W}_{1,1}$	$O_{0,2} \mathrel{+}= I_{0,2} \times \boldsymbol{W}_{2,2}$	$O_{0,3} \mathrel{+}= I_{0,3} \times \boldsymbol{W}_{3,3}$
$O_{1,0} \mathrel{+}= I_{1,1} \times \boldsymbol{W}_{1,0}$	$O_{1,1} \mathrel{+}= I_{1,2} \times \boldsymbol{W}_{2,1}$	$O_{1,2} \mathrel{+}= I_{1,3} \times \boldsymbol{W}_{3,2}$	$O_{1,3} \mathrel{+}= I_{1,0} \times \boldsymbol{W}_{0,3}$
$O_{2,0} \mathrel{+}= I_{2,2} \times \boldsymbol{W}_{2,0}$	$O_{2,1} \mathrel{+}= I_{2,3} \times \boldsymbol{W}_{3,1}$	$O_{2,2} \mathrel{+}= I_{2,0} \times \boldsymbol{W}_{0,2}$	$O_{2,3} \mathrel{+}= I_{2,1} \times \boldsymbol{W}_{1,3}$
$O_{3,0} \mathrel{+}= I_{3,3} \times \boldsymbol{W}_{3,0}$	$O_{3,1} \mathrel{+}= I_{3,0} \times \boldsymbol{W}_{0,1}$	$O_{3,2} \mathrel{+}= I_{3,1} \times \boldsymbol{W}_{1,2}$	$O_{3,3} \mathrel{+}= I_{3,2} \times \boldsymbol{W}_{2,3}$

Example: k = 2 Forward step t = 1

Communicate tensor I

$$(r, c+1, t): \xrightarrow{\text{from right}} (r, c, t+1):$$

$I_M = r \; mod \; 4$ $\quad \longleftrightarrow \quad$ $I_M = r \; mod \; 4$

$I_N = (r + c + 1 + t) \; mod \; 4$ $\quad \xleftrightarrow[\text{match}]{} \quad$ $I_N = (r + c + t + 1) \; mod \; 4$

Communicate tensor $\boldsymbol{W}$

$$(r+1, c, t): \xrightarrow{\text{from bottom}} (r, c, t+1):$$

$I_N = (r + 1 + c + t) \; mod \; 4$ $\quad \longleftrightarrow \quad$ $I_N = (r + c + t + 1) \; mod \; 4$

$I_K = c \; mod \; 4$ $\quad \xleftrightarrow[\text{match}]{} \quad$ $I_K = c \; mod \; 4$

Forward DSIs

$I_M = r \; mod \; 4$

$I_N = (r + c + t) \; mod \; 4$

$I_K = c \; mod \; 4$

$O_{0,0} \mathrel{+}= I_{0,1} \times \boldsymbol{W}_{1,0}$	$O_{0,1} \mathrel{+}= I_{0,2} \times \boldsymbol{W}_{2,1}$	$O_{0,2} \mathrel{+}= I_{0,3} \times \boldsymbol{W}_{3,2}$	$O_{0,3} \mathrel{+}= I_{0,0} \times \boldsymbol{W}_{0,3}$
$O_{1,0} \mathrel{+}= I_{1,2} \times \boldsymbol{W}_{2,0}$	$O_{1,1} \mathrel{+}= I_{1,3} \times \boldsymbol{W}_{3,1}$	$O_{1,2} \mathrel{+}= I_{1,0} \times \boldsymbol{W}_{0,2}$	$O_{1,3} \mathrel{+}= I_{1,1} \times \boldsymbol{W}_{1,3}$
$O_{2,0} \mathrel{+}= I_{2,3} \times \boldsymbol{W}_{3,0}$	$O_{2,1} \mathrel{+}= I_{2,0} \times \boldsymbol{W}_{0,1}$	$O_{2,2} \mathrel{+}= I_{2,1} \times \boldsymbol{W}_{1,2}$	$O_{2,3} \mathrel{+}= I_{2,2} \times \boldsymbol{W}_{2,3}$
$O_{3,0} \mathrel{+}= I_{3,0} \times \boldsymbol{W}_{0,0}$	$O_{3,1} \mathrel{+}= I_{3,1} \times \boldsymbol{W}_{1,1}$	$O_{3,2} \mathrel{+}= I_{3,2} \times \boldsymbol{W}_{2,2}$	$O_{3,3} \mathrel{+}= I_{3,3} \times \boldsymbol{W}_{3,3}$

Example: k = 2   Forward step t = 2

Communicate tensor I

$(r, c+1, t)$: $\xrightarrow{\text{from right}}$ $(r, c, t+1)$:

$I_M = r \bmod 4$ $\quad$ $I_M = r \bmod 4$

$I_N = (r + c + 1 + t) \bmod 4$ $\xleftrightarrow[\text{match}]{}$ $I_N = (r + c + t + 1) \bmod 4$

Communicate tensor $\boldsymbol{W}$

$(r+1, c, t)$: $\xrightarrow{\text{from bottom}}$ $(r, c, t+1)$:

$I_N = (r + 1 + c + t) \bmod 4$ $\xleftrightarrow[\text{match}]{}$ $I_N = (r + c + t + 1) \bmod 4$

$I_K = c \bmod 4$ $\quad$ $I_K = c \bmod 4$

Forward DSIs

$I_M = r \bmod 4$

$I_N = (r + c + t) \bmod 4$

$I_K = c \bmod 4$

$O_{0,0} \mathrel{+}= I_{0,2} \times \boldsymbol{W}_{2,0}$	$O_{0,1} \mathrel{+}= I_{0,3} \times \boldsymbol{W}_{3,1}$	$O_{0,2} \mathrel{+}= I_{0,0} \times \boldsymbol{W}_{0,2}$	$O_{0,3} \mathrel{+}= I_{0,1} \times \boldsymbol{W}_{1,3}$
$O_{1,0} \mathrel{+}= I_{1,3} \times \boldsymbol{W}_{3,0}$	$O_{1,1} \mathrel{+}= I_{1,0} \times \boldsymbol{W}_{0,1}$	$O_{1,2} \mathrel{+}= I_{1,1} \times \boldsymbol{W}_{1,2}$	$O_{1,3} \mathrel{+}= I_{1,2} \times \boldsymbol{W}_{2,3}$
$O_{2,0} \mathrel{+}= I_{2,0} \times \boldsymbol{W}_{0,0}$	$O_{2,1} \mathrel{+}= I_{2,1} \times \boldsymbol{W}_{1,1}$	$O_{2,2} \mathrel{+}= I_{2,2} \times \boldsymbol{W}_{2,2}$	$O_{2,3} \mathrel{+}= I_{2,3} \times \boldsymbol{W}_{3,3}$
$O_{3,0} \mathrel{+}= I_{3,1} \times \boldsymbol{W}_{1,0}$	$O_{3,1} \mathrel{+}= I_{3,2} \times \boldsymbol{W}_{2,1}$	$O_{3,2} \mathrel{+}= I_{3,3} \times \boldsymbol{W}_{3,2}$	$O_{3,3} \mathrel{+}= I_{3,0} \times \boldsymbol{W}_{0,3}$

Example: k = 2   Forward step t = 3

Last step of Forward, no communication:

Forward $O = I \times W$
Backward $dI = dO \times W^T$

- $\boldsymbol{W}$ alignment

Forward $(r, c, t = 3)$:
$I_N = (r + c + 3)\ mod\ 4$
$I_K = c\ mod\ 4$

$\longleftrightarrow$ match

Backward $(r, c, t = 0)$:
$I_N = (r + c - 1)\ mod\ 4$
$I_K = (c + 0)\ mod\ 4$

$O_{0,0} \mathrel{+}= I_{0,3} \times \boldsymbol{W}_{3,0}$	$O_{0,1} \mathrel{+}= I_{0,0} \times \boldsymbol{W}_{0,1}$	$O_{0,2} \mathrel{+}= I_{0,1} \times \boldsymbol{W}_{1,2}$	$O_{0,3} \mathrel{+}= I_{0,2} \times \boldsymbol{W}_{2,3}$
$O_{1,0} \mathrel{+}= I_{1,0} \times \boldsymbol{W}_{0,0}$	$O_{1,1} \mathrel{+}= I_{1,1} \times \boldsymbol{W}_{1,1}$	$O_{1,2} \mathrel{+}= I_{1,2} \times \boldsymbol{W}_{2,2}$	$O_{1,3} \mathrel{+}= I_{1,3} \times \boldsymbol{W}_{3,3}$
$O_{2,0} \mathrel{+}= I_{2,1} \times \boldsymbol{W}_{1,0}$	$O_{2,1} \mathrel{+}= I_{2,2} \times \boldsymbol{W}_{2,1}$	$O_{2,2} \mathrel{+}= I_{2,3} \times \boldsymbol{W}_{3,2}$	$O_{2,3} \mathrel{+}= I_{2,0} \times \boldsymbol{W}_{0,3}$
$O_{3,0} \mathrel{+}= I_{3,2} \times \boldsymbol{W}_{2,0}$	$O_{3,1} \mathrel{+}= I_{3,3} \times \boldsymbol{W}_{3,1}$	$O_{3,2} \mathrel{+}= I_{3,0} \times \boldsymbol{W}_{0,2}$	$O_{3,3} \mathrel{+}= I_{3,1} \times \boldsymbol{W}_{1,3}$

Example: k = 2   Backward step t = 0

Communicate tensor dO

$(r, c+1, t)$: $\xrightarrow{\text{from right}}$ $(r, c, t+1)$:

$I_M = r \; mod \; 4$ | $I_M = r \; mod \; 4$

$I_K = (c + 1 + t) \; mod \; 4$ $\xleftrightarrow[\text{match}]{}$ $I_K = (c + t + 1) \; mod \; 4$

Communicate tensor $\boldsymbol{W}$

$(r-1, c+1, t)$: $\xrightarrow{\text{from right-top}}$ $(r, c, t+1)$:

$I_N = (r - 1 + c + 1 - 1) \; mod \; 4$ | $I_N = (r + c - 1) \; mod \; 4$

$I_K = (c + 1 + t) \; mod \; 4$ $\xleftrightarrow[\text{match}]{}$ $I_K = (c + t + 1) \; mod \; 4$

Backward DSIs

$I_M = r \; mod \; 4$

$I_N = (r + c - 1) \; mod \; 4$

$I_K = (c + t) \; mod \; 4$

$dI_{0,3} \mathrel{+}= dO_{0,0} \times \boldsymbol{W}^T_{0,3}$	$dI_{0,0} \mathrel{+}= dO_{0,1} \times \boldsymbol{W}^T_{1,0}$	$dI_{0,1} \mathrel{+}= dO_{0,2} \times \boldsymbol{W}^T_{2,1}$	$dI_{0,2} \mathrel{+}= dO_{0,3} \times \boldsymbol{W}^T_{3,2}$
$dI_{1,0} \mathrel{+}= dO_{1,0} \times \boldsymbol{W}^T_{0,0}$	$dI_{1,1} \mathrel{+}= dO_{1,1} \times \boldsymbol{W}^T_{1,1}$	$dI_{1,2} \mathrel{+}= dO_{1,2} \times \boldsymbol{W}^T_{2,2}$	$dI_{1,3} \mathrel{+}= dO_{1,3} \times \boldsymbol{W}^T_{3,3}$
$dI_{2,1} \mathrel{+}= dO_{2,0} \times \boldsymbol{W}^T_{0,1}$	$dI_{2,2} \mathrel{+}= dO_{2,1} \times \boldsymbol{W}^T_{1,2}$	$dI_{2,3} \mathrel{+}= dO_{2,2} \times \boldsymbol{W}^T_{2,3}$	$dI_{2,0} \mathrel{+}= dO_{2,3} \times \boldsymbol{W}^T_{3,0}$
$dI_{3,2} \mathrel{+}= dO_{3,0} \times \boldsymbol{W}^T_{0,2}$	$dI_{3,3} \mathrel{+}= dO_{3,1} \times \boldsymbol{W}^T_{1,3}$	$dI_{3,0} \mathrel{+}= dO_{3,2} \times \boldsymbol{W}^T_{2,0}$	$dI_{3,1} \mathrel{+}= dO_{3,3} \times \boldsymbol{W}^T_{3,1}$

Example: k = 2 Backward step t = 1

Communicate tensor dO

$$(r, c+1, t): \xrightarrow{\text{from right}} (r, c, t+1):$$

$$I_M = r \bmod 4 \qquad\qquad I_M = r \bmod 4$$

$$I_K = (c+1+t) \bmod 4 \xleftrightarrow[\text{match}]{} I_K = (c+t+1) \bmod 4$$

Communicate tensor $\boldsymbol{W}$

$$(r-1, c+1, t): \xrightarrow{\text{from right-top}} (r, c, t+1):$$

$$I_N = (r-1+c+1-1) \bmod 4 \qquad I_N = (r+c-1) \bmod 4$$

$$I_K = (c+1+t) \bmod 4 \xleftrightarrow[\text{match}]{} I_K = (c+t+1) \bmod 4$$

Backward DSIs

$$I_M = r \bmod 4$$
$$I_N = (r+c-1) \bmod 4$$
$$I_K = (c+t) \bmod 4$$

$dI_{0,3} \mathrel{+}= dO_{0,1} \times \boldsymbol{W}^T_{1,3}$	$dI_{0,0} \mathrel{+}= dO_{0,2} \times \boldsymbol{W}^T_{2,0}$	$dI_{0,1} \mathrel{+}= dO_{0,3} \times \boldsymbol{W}^T_{3,1}$	$dI_{0,2} \mathrel{+}= dO_{0,0} \times \boldsymbol{W}^T_{0,2}$
$dI_{1,0} \mathrel{+}= dO_{1,1} \times \boldsymbol{W}^T_{1,0}$	$dI_{1,1} \mathrel{+}= dO_{1,2} \times \boldsymbol{W}^T_{2,1}$	$dI_{1,2} \mathrel{+}= dO_{1,3} \times \boldsymbol{W}^T_{3,2}$	$dI_{1,3} \mathrel{+}= dO_{1,0} \times \boldsymbol{W}^T_{0,3}$
$dI_{2,1} \mathrel{+}= dO_{2,1} \times \boldsymbol{W}^T_{1,1}$	$dI_{2,2} \mathrel{+}= dO_{2,2} \times \boldsymbol{W}^T_{2,2}$	$dI_{2,3} \mathrel{+}= dO_{2,3} \times \boldsymbol{W}^T_{3,3}$	$dI_{2,0} \mathrel{+}= dO_{2,0} \times \boldsymbol{W}^T_{0,0}$
$dI_{3,2} \mathrel{+}= dO_{3,1} \times \boldsymbol{W}^T_{1,2}$	$dI_{3,3} \mathrel{+}= dO_{3,2} \times \boldsymbol{W}^T_{2,3}$	$dI_{3,0} \mathrel{+}= dO_{3,3} \times \boldsymbol{W}^T_{3,0}$	$dI_{3,1} \mathrel{+}= dO_{3,0} \times \boldsymbol{W}^T_{0,1}$

Example: k = 2   Backward step t = 2

Communicate tensor dO

$(r, c+1, t)$: $\xrightarrow{\text{from right}}$ $(r, c, t+1)$:

$I_M = r \bmod 4$ $\qquad$ $I_M = r \bmod 4$

$I_K = (c+1+t) \bmod 4$ $\xleftrightarrow[\text{match}]{}$ $I_K = (c+t+1) \bmod 4$

Communicate tensor $\boldsymbol{W}$

$(r-1, c+1, t)$: $\xrightarrow{\text{from right-top}}$ $(r, c, t+1)$:

$I_N = (r-1+c+1-1) \bmod 4$ $\qquad$ $I_N = (r+c-1) \bmod 4$

$I_K = (c+1+t) \bmod 4$ $\xleftrightarrow[\text{match}]{}$ $I_K = (c+t+1) \bmod 4$

Backward DSIs

$$I_M = r \bmod 4$$
$$I_N = (r+c-1) \bmod 4$$
$$I_K = (c+t) \bmod 4$$

$dI_{0,3} \mathrel{+}= dO_{0,2} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{2,3}}$	$dI_{0,0} \mathrel{+}= dO_{0,3} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{3,0}}$	$dI_{0,1} \mathrel{+}= dO_{0,0} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{0,1}}$	$dI_{0,2} \mathrel{+}= dO_{0,1} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{1,2}}$
$dI_{1,0} \mathrel{+}= dO_{1,2} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{2,0}}$	$dI_{1,1} \mathrel{+}= dO_{1,3} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{3,1}}$	$dI_{1,2} \mathrel{+}= dO_{1,0} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{0,2}}$	$dI_{1,3} \mathrel{+}= dO_{1,1} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{1,3}}$
$dI_{2,1} \mathrel{+}= dO_{2,2} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{2,1}}$	$dI_{2,2} \mathrel{+}= dO_{2,3} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{3,2}}$	$dI_{2,3} \mathrel{+}= dO_{2,0} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{0,3}}$	$dI_{2,0} \mathrel{+}= dO_{2,1} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{1,0}}$
$dI_{3,2} \mathrel{+}= dO_{3,2} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{2,2}}$	$dI_{3,3} \mathrel{+}= dO_{3,3} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{3,3}}$	$dI_{3,0} \mathrel{+}= dO_{3,0} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{0,0}}$	$dI_{3,1} \mathrel{+}= dO_{3,1} \times \boldsymbol{W}^{\boldsymbol{T}}_{\boldsymbol{1,1}}$

Example: k = 2   Backward step t = 3

- W alignment

Backward $(r, c+1, t=3)$: $\xrightarrow{\text{from right}}$ Forward $(r, c, t=0)$:

$I_N = (r + c + 1 - 1)\ mod\ 4$ $\quad$ $I_N = (r + c)\ mod\ 4$

$I_K = (c + 1 + 3)\ mod\ 4$ $\quad \xleftrightarrow{\text{match}} \quad$ $I_K = (c)\ mod\ 4$

Forward	$O = I \times W$
Backward	$dI = dO \times W^T$
Gradient	$dW = I^T \times dO$

- dO alignment

Backward $(r, c, t=3)$:

$I_M = r\ mod\ 4$

$I_K = (c + 3)\ mod\ 4$

$\xleftrightarrow{\text{match}}$

Gradient $(r, c, t=0)$:

$I_M = (r + 0)\ mod\ 4$

$I_K = (c - 1)\ mod\ 4$

- I alignment

Forward $(r, c, t=3)$:

$I_M = r\ mod\ 4$

$I_N = (r + c + 3)\ mod\ 4$

$\xleftrightarrow{\text{match}}$

Gradient $(r, c, t=0)$:

$I_M = (r + 0)\ mod\ 4$

$I_K = (r + c - 1)\ mod\ 4$

$dI_{0,3} \mathrel{+}= dO_{0,3} \times W^T_{3,3}$	$dI_{0,0} \mathrel{+}= dO_{0,0} \times W^T_{0,0}$	$dI_{0,1} \mathrel{+}= dO_{0,1} \times W^T_{1,1}$	$dI_{0,2} \mathrel{+}= dO_{0,2} \times W^T_{2,2}$
$dI_{1,0} \mathrel{+}= dO_{1,3} \times W^T_{3,0}$	$dI_{1,1} \mathrel{+}= dO_{1,0} \times W^T_{0,1}$	$dI_{1,2} \mathrel{+}= dO_{1,1} \times W^T_{1,2}$	$dI_{1,3} \mathrel{+}= dO_{1,2} \times W^T_{2,3}$
$dI_{2,1} \mathrel{+}= dO_{2,3} \times W^T_{3,1}$	$dI_{2,2} \mathrel{+}= dO_{2,0} \times W^T_{0,2}$	$dI_{2,3} \mathrel{+}= dO_{2,1} \times W^T_{1,3}$	$dI_{2,0} \mathrel{+}= dO_{2,2} \times W^T_{2,0}$
$dI_{3,2} \mathrel{+}= dO_{3,3} \times W^T_{3,2}$	$dI_{3,3} \mathrel{+}= dO_{3,0} \times W^T_{0,3}$	$dI_{3,0} \mathrel{+}= dO_{3,1} \times W^T_{1,0}$	$dI_{3,1} \mathrel{+}= dO_{3,2} \times W^T_{2,1}$

Example: k = 2 Gradient step t = 0

Communicate tensor $\boldsymbol{I}$

$$(r+1, c-1, t): \xrightarrow{\text{from bottom left}} (r, c, t+1):$$

$$I_M = (r+1+t)\ mod\ 4 \qquad I_M = (r+t+1)\ mod\ 4$$

$$I_N = \left(r+1+c-1-1+\delta_{0,3}\right)\ mod\ 4 \xleftrightarrow{\text{match}} I_N = \left(r+c-1+\delta_{1,3}\right)\ mod\ 4$$

Communicate tensor dO

$$(r+1, c, t): \xrightarrow{\text{from bottom}} (r, c, t+1):$$

$$I_M = (r+1+t)\ mod\ 4 \qquad I_M = (r+t+1)\ mod\ 4$$

$$I_K = \left(c-1+\delta_{0,3}\right)\ mod\ 4 \xleftrightarrow{\text{match}} I_K = \left(c-1+\delta_{1,3}\right)\ mod\ 4$$

Gradient DSIs

$$I_M = (r+t)\ mod\ 4$$
$$I_N = \left(r+c-1+\delta_{0,3}\right)\ mod\ 4$$
$$I_K = \left(c-1+\delta_{0,3}\right)\ mod\ 4$$

$dW_{3,3} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{3,0}} \times dO_{0,3}$	$dW_{0,0} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{0,0}} \times dO_{0,0}$	$dW_{1,1} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{1,0}} \times dO_{0,1}$	$dW_{2,2} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{2,0}} \times dO_{0,2}$
$dW_{0,3} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{0,1}} \times dO_{1,3}$	$dW_{1,0} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{1,1}} \times dO_{1,0}$	$dW_{2,1} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{2,1}} \times dO_{1,1}$	$dW_{3,2} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{3,1}} \times dO_{1,2}$
$dW_{1,3} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{1,2}} \times dO_{2,3}$	$dW_{2,0} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{2,2}} \times dO_{2,0}$	$dW_{3,1} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{3,2}} \times dO_{2,1}$	$dW_{0,2} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{0,2}} \times dO_{2,2}$
$dW_{2,3} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{2,3}} \times dO_{3,3}$	$dW_{3,0} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{3,3}} \times dO_{3,0}$	$dW_{0,1} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{0,3}} \times dO_{3,1}$	$dW_{1,2} \mathrel{+}= \boldsymbol{I}^{\boldsymbol{T}}_{\boldsymbol{1,3}} \times dO_{3,2}$

Example: k = 2   Gradient step t = 1

Communicate tensor $\boldsymbol{I}$

$$(r+1, c-1, t): \xrightarrow{\text{from bottom left}} (r, c, t+1):$$

$$I_M = (r+1+t) \bmod 4 \qquad I_M = (r+t+1) \bmod 4$$

$$I_N = \left(r+1+c-1-1+\delta_{1,3}\right) \bmod 4 \xleftrightarrow{\text{match}} I_N = \left(r+c-1+\delta_{2,3}\right) \bmod 4$$

Communicate tensor dO

$$(r+1, c, t): \xrightarrow{\text{from bottom}} (r, c, t+1):$$

$$I_M = (r+1+t) \bmod 4 \qquad I_M = (r+t+1) \bmod 4$$

$$I_K = \left(c-1+\delta_{1,3}\right) \bmod 4 \xleftrightarrow{\text{match}} I_K = \left(c-1+\delta_{2,3}\right) \bmod 4$$

Gradient DSIs

$$I_M = (r+t) \bmod 4$$
$$I_N = \left(r+c-1+\delta_{1,3}\right) \bmod 4$$
$$I_K = \left(c-1+\delta_{1,3}\right) \bmod 4$$

$dW_{3,3} \mathrel{+}= I^T_{3,1} \times dO_{1,3}$	$dW_{0,0} \mathrel{+}= I^T_{0,1} \times dO_{1,0}$	$dW_{1,1} \mathrel{+}= I^T_{1,1} \times dO_{1,1}$	$dW_{2,2} \mathrel{+}= I^T_{2,1} \times dO_{1,2}$
$dW_{0,3} \mathrel{+}= I^T_{0,2} \times dO_{2,3}$	$dW_{1,0} \mathrel{+}= I^T_{1,2} \times dO_{2,0}$	$dW_{2,1} \mathrel{+}= I^T_{2,2} \times dO_{2,1}$	$dW_{3,2} \mathrel{+}= I^T_{3,2} \times dO_{2,2}$
$dW_{1,3} \mathrel{+}= I^T_{1,3} \times dO_{3,3}$	$dW_{2,0} \mathrel{+}= I^T_{2,3} \times dO_{3,0}$	$dW_{3,1} \mathrel{+}= I^T_{3,3} \times dO_{3,1}$	$dW_{0,2} \mathrel{+}= I^T_{0,3} \times dO_{3,2}$
$dW_{2,3} \mathrel{+}= I^T_{2,0} \times dO_{0,3}$	$dW_{3,0} \mathrel{+}= I^T_{3,0} \times dO_{0,0}$	$dW_{0,1} \mathrel{+}= I^T_{0,1} \times dO_{0,1}$	$dW_{1,2} \mathrel{+}= I^T_{1,0} \times dO_{0,2}$

Example: k = 2   Gradient step t = 2

Communicate tensor I

$(r+1, c, t)$: $\xrightarrow{\text{from bottom}}$ $(r, c, t+1)$:

$I_M = (r+1+2)\ mod\ 4$ | $I_M = (r+3)\ mod\ 4$

$I_N = (r+1+c-1+\delta_{2,3})\ mod\ 4$ $\xleftrightarrow{\text{match}}$ $I_N = (r+c-1+\delta_{3,3})\ mod\ 4$

Communicate tensor $\boldsymbol{dO}$

$(r+1, c+1, t)$: $\xrightarrow{\text{from bottom right}}$ $(r, c, t+1)$:

$I_M = (r+1+2)\ mod\ 4$ | $I_M = (r+3)\ mod\ 4$

$I_K = (c+1-1+\delta_{2,3})\ mod\ 4$ $\xleftrightarrow{\text{match}}$ $I_K = (c-1+\delta_{3,3})\ mod\ 4$

Gradient DSIs

$$I_M = (r+t)\ mod\ 4$$
$$I_N = (r+c-1+\delta_{2,3})\ mod\ 4$$
$$I_K = (c-1+\delta_{2,3})\ mod\ 4$$

$dW_{3,3} \mathrel{+}= I_{3,2}^T \times \boldsymbol{dO}_{2,3}$	$dW_{0,0} \mathrel{+}= I_{0,2}^T \times \boldsymbol{dO}_{2,0}$	$dW_{1,1} \mathrel{+}= I_{1,2}^T \times \boldsymbol{dO}_{2,1}$	$dW_{2,2} \mathrel{+}= I_{2,2}^T \times \boldsymbol{dO}_{2,2}$
$dW_{0,3} \mathrel{+}= I_{0,3}^T \times \boldsymbol{dO}_{3,3}$	$dW_{1,0} \mathrel{+}= I_{1,3}^T \times \boldsymbol{dO}_{3,0}$	$dW_{2,1} \mathrel{+}= I_{2,3}^T \times \boldsymbol{dO}_{3,1}$	$dW_{3,2} \mathrel{+}= I_{3,3}^T \times \boldsymbol{dO}_{3,2}$
$dW_{1,3} \mathrel{+}= I_{1,0}^T \times \boldsymbol{dO}_{0,3}$	$dW_{2,0} \mathrel{+}= I_{2,0}^T \times \boldsymbol{dO}_{0,0}$	$dW_{3,1} \mathrel{+}= I_{3,0}^T \times \boldsymbol{dO}_{0,1}$	$dW_{0,2} \mathrel{+}= I_{0,0}^T \times \boldsymbol{dO}_{0,2}$
$dW_{2,3} \mathrel{+}= I_{2,1}^T \times \boldsymbol{dO}_{1,3}$	$dW_{3,0} \mathrel{+}= I_{3,1}^T \times \boldsymbol{dO}_{1,0}$	$dW_{0,1} \mathrel{+}= I_{0,1}^T \times \boldsymbol{dO}_{1,1}$	$dW_{1,2} \mathrel{+}= I_{1,1}^T \times \boldsymbol{dO}_{1,2}$

Example: k = 2   Gradient step t = 3

- dW alignment

Gradient $(r, c+1, t<3)$: $\xrightarrow{\text{from right}}$ Gradient $(r, c, t=3)$:

$I_N = \left(r + c + 1 - 1 + \delta_{t,3}\right)\ mod\ 4$ $I_N = \left(r + c - 1 + \delta_{3,3}\right)\ mod\ 4$

$I_K = \left(c + 1 - 1 + \delta_{t,3}\right)\ mod\ 4$ $\xleftrightarrow{\text{match}}$ $I_K = \left(c - 1 + \delta_{3,3}\right)\ mod\ 4$

match

Forward $(r, c, t=0)$:

$I_N = (r + c + 0)\ mod\ 4$

$I_K = c\ mod\ 4$

Accumulated dW when t < 3

$dW_{3,3}$	$dW_{0,0}$	$dW_{1,1}$	$dW_{2,2}$
$dW_{0,3}$	$dW_{1,0}$	$dW_{2,1}$	$dW_{3,2}$
$dW_{1,3}$	$dW_{2,0}$	$dW_{3,1}$	$dW_{0,2}$
$dW_{2,3}$	$dW_{3,0}$	$dW_{0,1}$	$dW_{1,2}$

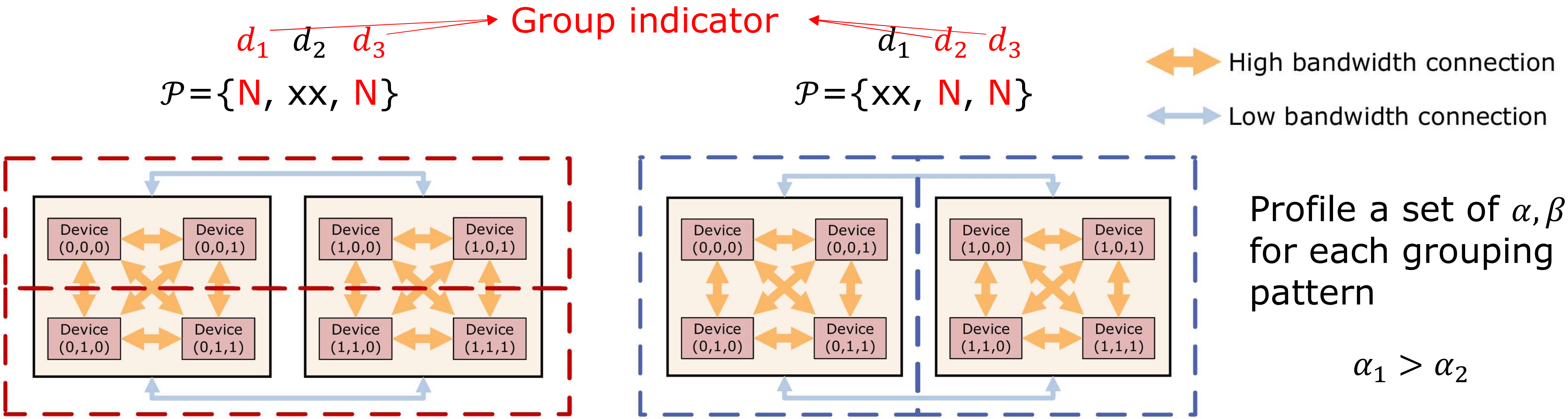
Add dW computed during step t = 3 with shifted accumulated dW

$dW_{0,0} \mathrel{+}= I_{0,3}^T \times dO_{3,0}$	$dW_{1,1} \mathrel{+}= I_{1,3}^T \times dO_{3,1}$	$dW_{2,2} \mathrel{+}= I_{2,3}^T \times dO_{3,2}$	$dW_{3,3} \mathrel{+}= I_{3,3}^T \times dO_{3,3}$
$dW_{1,0} \mathrel{+}= I_{1,0}^T \times dO_{0,0}$	$dW_{2,1} \mathrel{+}= I_{2,0}^T \times dO_{0,1}$	$dW_{3,2} \mathrel{+}= I_{3,0}^T \times dO_{0,2}$	$dW_{0,3} \mathrel{+}= I_{0,0}^T \times dO_{0,3}$
$dW_{2,0} \mathrel{+}= I_{2,1}^T \times dO_{1,0}$	$dW_{3,1} \mathrel{+}= I_{3,1}^T \times dO_{1,1}$	$dW_{0,2} \mathrel{+}= I_{0,1}^T \times dO_{1,2}$	$dW_{1,3} \mathrel{+}= I_{1,1}^T \times dO_{1,3}$
$dW_{3,0} \mathrel{+}= I_{3,2}^T \times dO_{2,0}$	$dW_{0,1} \mathrel{+}= I_{0,2}^T \times dO_{2,1}$	$dW_{1,2} \mathrel{+}= I_{1,2}^T \times dO_{2,2}$	$dW_{2,3} \mathrel{+}= I_{2,2}^T \times dO_{2,3}$

Cost Model

Intra-operator communication: all-reduce, ring

Example:
all-reduce of forward linear operator output tensor O – induced by partition N



Profile a set of α, β for each grouping pattern

$$\alpha_1 > \alpha_2$$

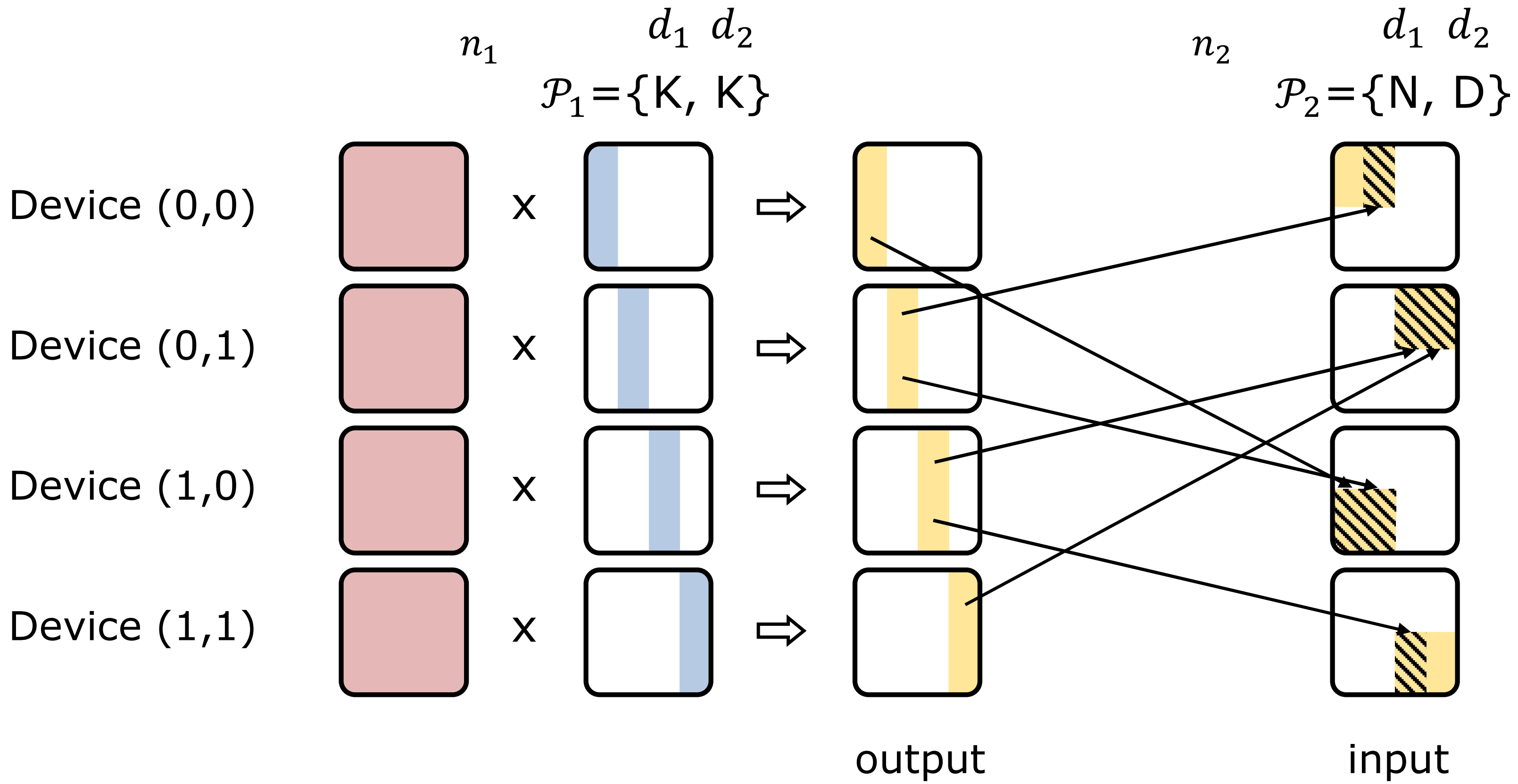
$$latency = \alpha_1 \cdot sizeof(O) + \beta_1$$

$$latency = \alpha_2 \cdot sizeof(O) + \beta_2$$

Cost Model

Inter-operator communication: redistribution between operators

Example:
redistribution during forward between linear (n_1) and relu (n_2)



Shadow: where the input and output tensor do not intersect, need communication

Cost Model

Overall cost

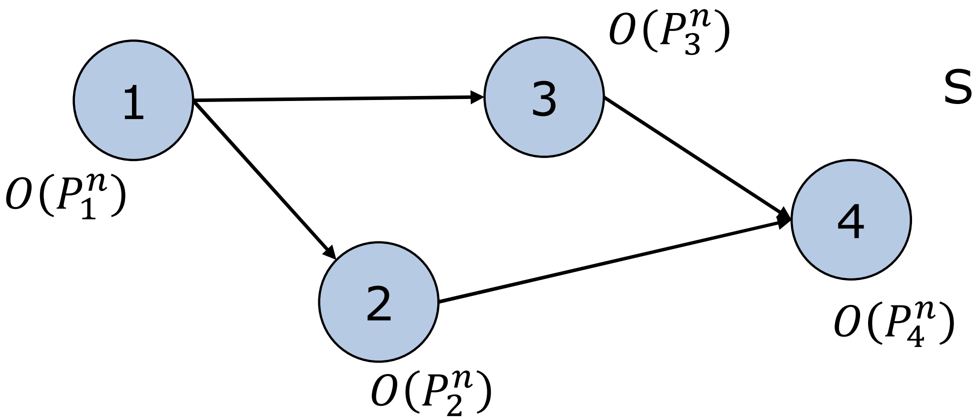
Counting all intra- and inter- operator cost

Computation graph $G =< N, E >$, suppose operator n_i is partitioned with strategy $\mathcal{P}_i$

$$Cost = \sum_{n_i \in N} intraCost(n_i, \mathcal{P}_i) + \sum_{(n_i, n_j) \in E} interCost(n_i, n_j, \mathcal{P}_i, \mathcal{P}_j)$$

To 2^n devices

- Number of partition primitives of operator n_i: P_i
- Tensor partition space size of n_i: $O(P_i^n)$

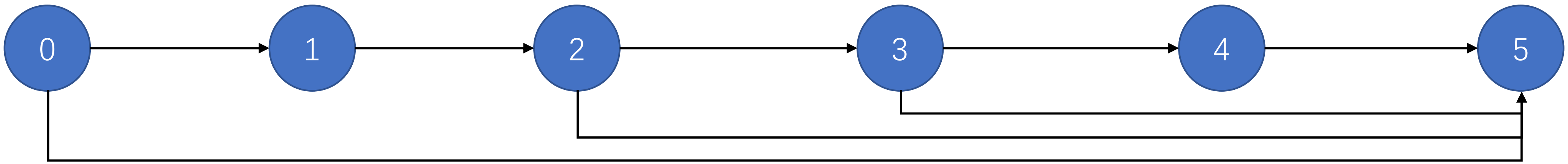


Search space size

$$\prod_i P_i^n$$

Optimization Algorithm: naïve dynamic programming

Complicated optimal substructure



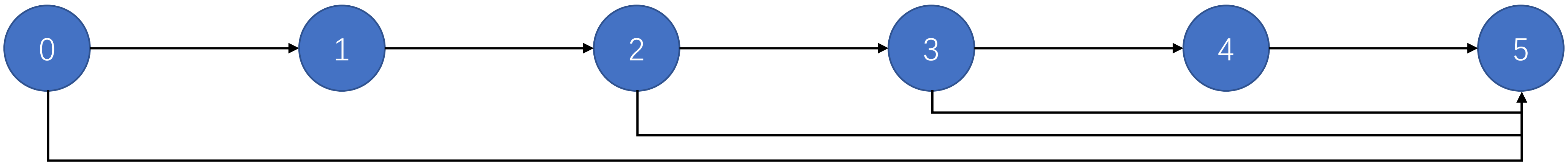
$$\mathcal{C}_0(\mathcal{P}_0) \rightarrow \mathcal{C}_{0,1}(\mathcal{P}_0, \mathcal{P}_1) \rightarrow \mathcal{C}_{0,2}(\mathcal{P}_0, \mathcal{P}_2) \rightarrow \mathcal{C}_{0,3}(\mathcal{P}_0, \mathcal{P}_2, \mathcal{P}_3) \rightarrow \mathcal{C}_{0,4}(\mathcal{P}_0, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4) \rightarrow \mathcal{C}_{0,5}(\mathcal{P}_0, \mathcal{P}_5)$$

$$O(P_0^n P_1^n) \quad O(P_0^n P_1^n P_2^n) \quad O(P_0^n P_2^n P_3^n) \quad O(P_0^n P_2^n P_3^n P_4^n) \quad O(P_0^n P_2^n P_3^n P_4^n P_5^n)$$

☹

Overall complexity $O(P_0^n P_2^n P_3^n P_4^n P_5^n)$

Optimization Algorithm: segmented dynamic programming



$C_0(\mathcal{P}_0) \rightarrow C_{0,1}(\mathcal{P}_0, \mathcal{P}_1) \rightarrow C_{0,2}(\mathcal{P}_0, \mathcal{P}_2)$

$O(P_0^n P_1^n)$ $\quad$ $O(P_0^n P_1^n P_2^n)$

Merge $C_{0,5}(\mathcal{P}_0, \mathcal{P}_5)$ $\quad$ $O(P_0^n P_2^n P_5^n)$

Merge $C_{2,5}(\mathcal{P}_2, \mathcal{P}_5)$ $\quad$ $O(P_2^n P_3^n P_5^n)$

$C_2(\mathcal{P}_2) \rightarrow C_{2,3}(\mathcal{P}_2, \mathcal{P}_3)$

$O(P_2^n P_3^n)$

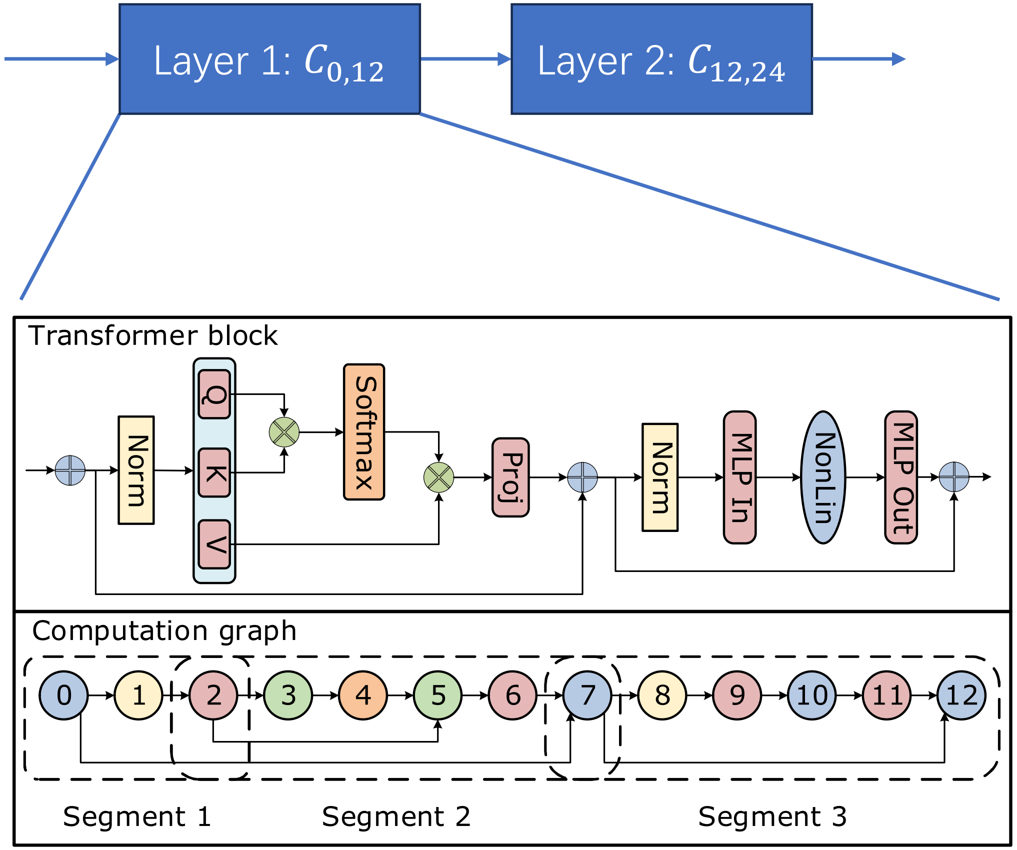
$C_3(\mathcal{P}_3) \rightarrow C_{3,4}(\mathcal{P}_3, \mathcal{P}_4) \rightarrow C_{3,5}(\mathcal{P}_3, \mathcal{P}_5)$

$O(P_3^n P_4^n)$ $\quad$ $O(P_3^n P_4^n P_5^n)$

☺

Overall complexity $O(P_i^n P_j^n P_k^n)$

Segmentation of Transformer Models



- Dynamic programming within each segment: Optimal substructures $C_{0,2}, C_{2,7}, C_{7,12}$

- Merge segments:

$$C_{0,7}(\mathcal{P}_0, \mathcal{P}_7) = \min_{\mathcal{P}_2}\{C_{0,2}(\mathcal{P}_0, \mathcal{P}_2) + C_{2,7}(\mathcal{P}_2, \mathcal{P}_7) - n_2(\mathcal{P}_2) + e_{0,7}(\mathcal{P}_0, \mathcal{P}_7)\}$$

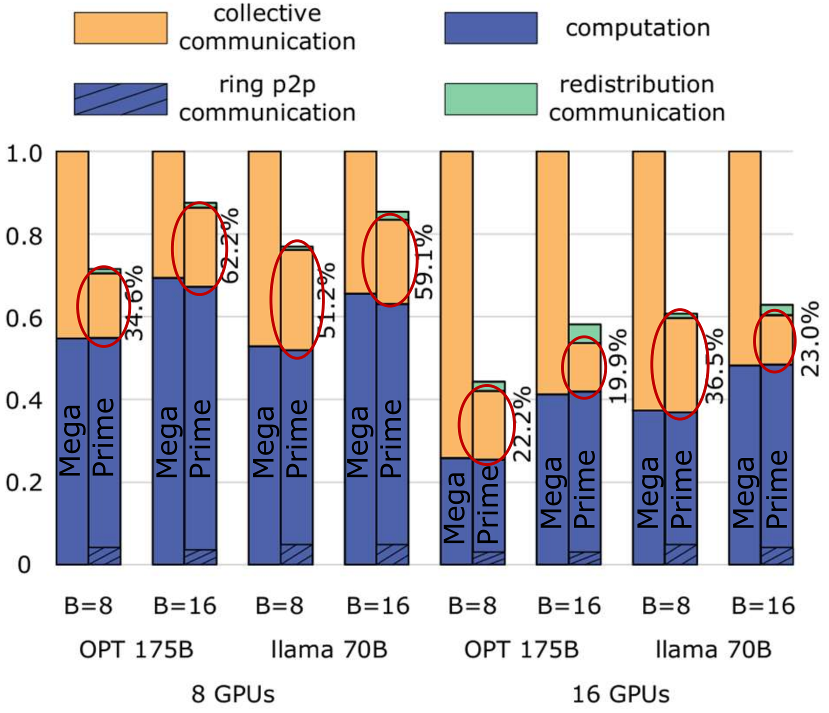
$$C_{0,12}(\mathcal{P}_0, \mathcal{P}_{12}) = \min_{\mathcal{P}_7}\{C_{0,7}(\mathcal{P}_0, \mathcal{P}_7) + C_{7,12}(\mathcal{P}_7, \mathcal{P}_{12}) - n_7(\mathcal{P}_7)\}$$

- Merge layers:

$$C_{0,24}(\mathcal{P}_0, \mathcal{P}_{24}) = \min_{\mathcal{P}_{12}}\{C_{0,12}(\mathcal{P}_0, \mathcal{P}_{12}) + C_{12,24}(\mathcal{P}_{12}, \mathcal{P}_{24}) - n_{12}(\mathcal{P}_{12})\}$$

Evaluation: Breakdown and Ablation

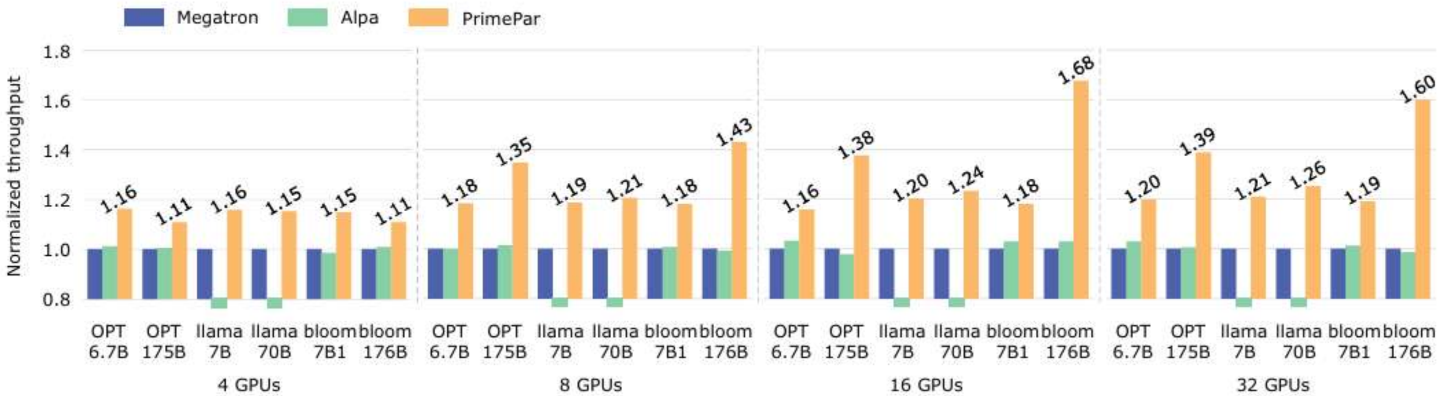
MLP blocks latency breakdown comparison



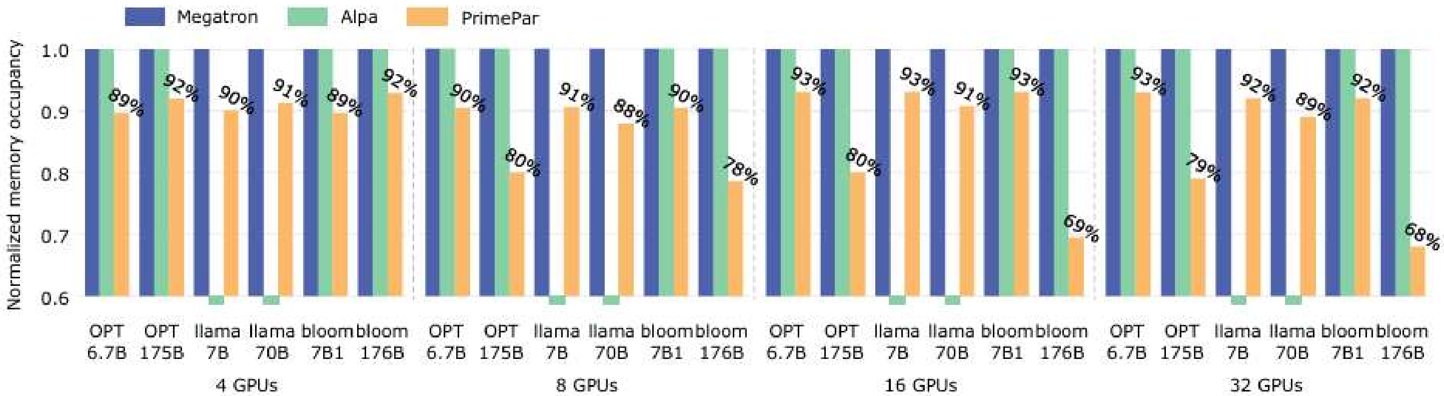
- The latency of collective communications are reduced to 19.9–62.2%

Evaluation: Performance and Memory Occupation

Normalized training throughput



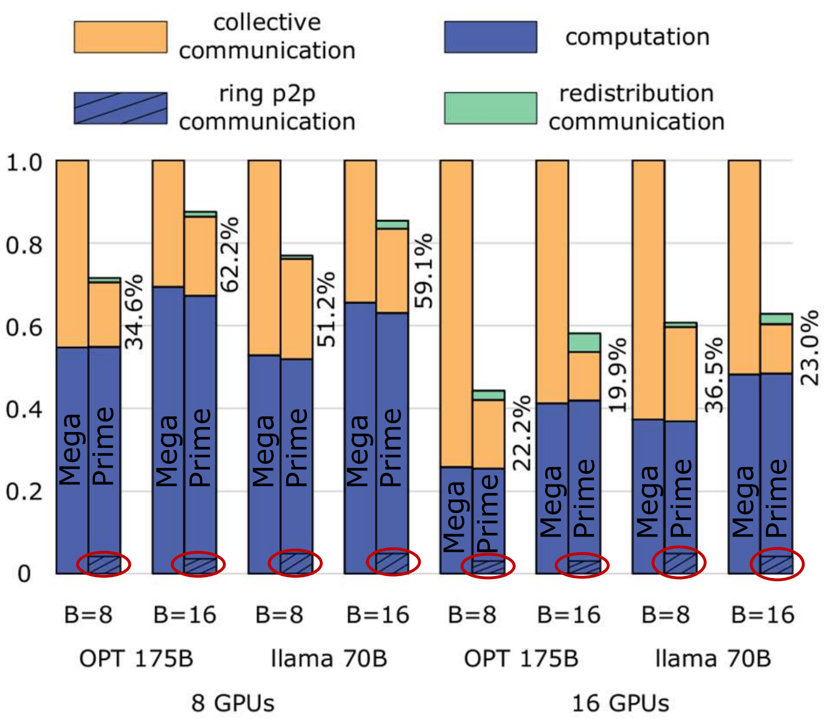
Peak memory occupation



- 1.11–1.68x training speedup and 68–93% peak memory
- Optimized tensor partitions improve training speed and save memory simultaneously
- Benefits are more significant when scaling larger models to more GPUs

Evaluation: Breakdown and Ablation

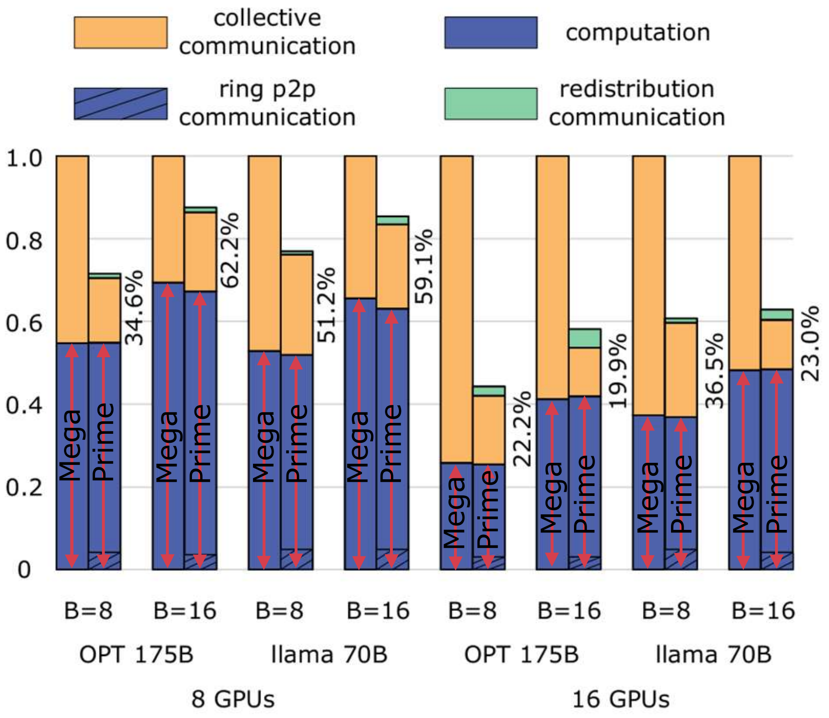
MLP blocks latency breakdown comparison



- The latency of collective communications are reduced to 19.9–62.2%
- Induced ring point-to-point communications are cheaper and fully overlapped with computation latency

Evaluation: Breakdown and Ablation

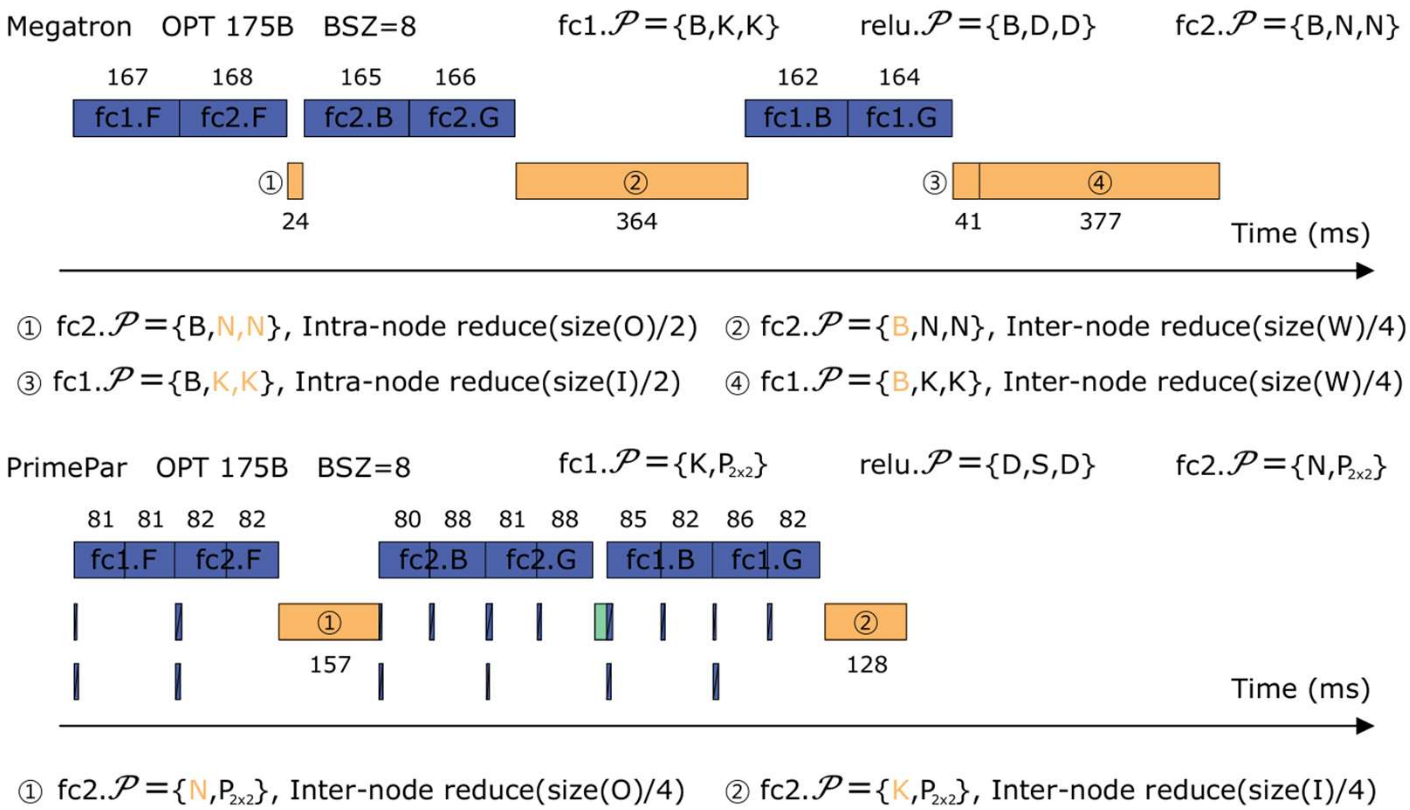
MLP blocks latency breakdown comparison



- The latency of collective communications are reduced to 19.9–62.2%
- Induced ring point-to-point communications are cheaper and fully overlapped with computation latency
- Computation latency remains the same: does not compromise computation efficiency

Evaluation: Breakdown and Ablation

Kernel execution timelines of the MLP block



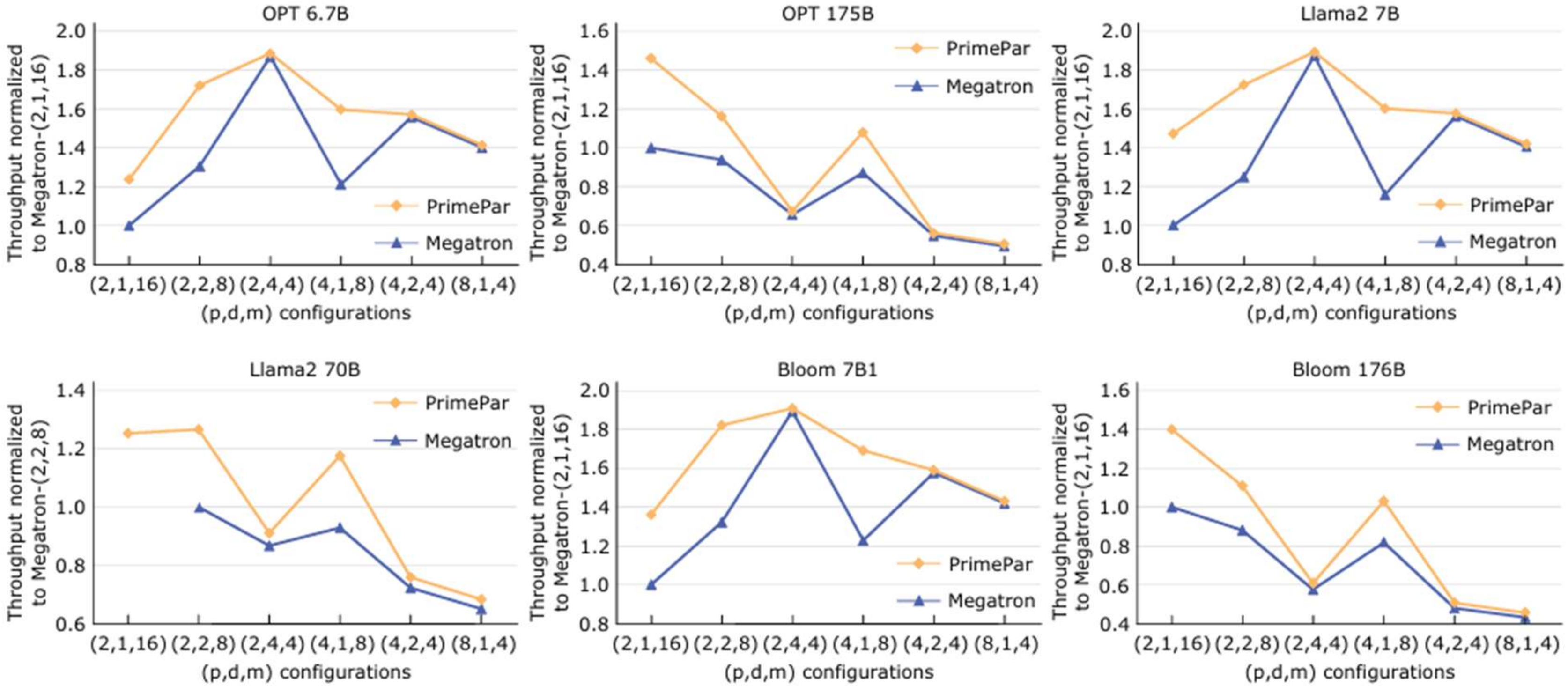
- Baseline:
  Intra-node collective:
  size(O)/2 + size(I)/2

  Inter-node collective:
  size(W)/2

- PrimePar:
  Intra-node collective:
  0

  Inter-node collective:
  size(O)/4 + size(I)/4
  < size(W)/2

Evaluation: Impact on 3D Parallelism



- 1.46, 1.27, 1.40x speedups for OPT 175B, Llama2 70B, Bloom 176B
- Larger models prefer higher degree of model parallelism, where PrimePar yields greater performance improvements

Conclusion

- Spatial-temporal tensor partition: more efficient communication and better utilization of hardware resources
- Formulize spatial-temporal sub-operator distribution: help design efficient tensor partition primitive and analyze communication patterns
- Further exploration into spatial-temporal tensor partition space is worthwhile

Thank you!

Please contact us at the email address below if you have any questions:

wanghaoran20g@ict.ac.cn