Unity

Accelerating DNN Training Through Joint Optimization of Algebraic Transformations and Parallelization

Colin Unger *, [1]
Sina Lin [6]
Vinay Ramakrishnaiah [4]
Jamaludin Mohd-Yusof [4]
Jongsoo Park [3]

Zhihao Jia *, [2, 3]
Mandeep Baines [3]
Nirmal Prajapati [4]
Xi Luo [7]
Misha Smelyanskiy [3]

Wei Wu [4, 5]
Carlos Efrain Quintero Narvaez [3]
Pat McCormick [4]
Dheevatsa Mudigere [3]
Alex Aiken [1]


1

2


3


4

5


6

7

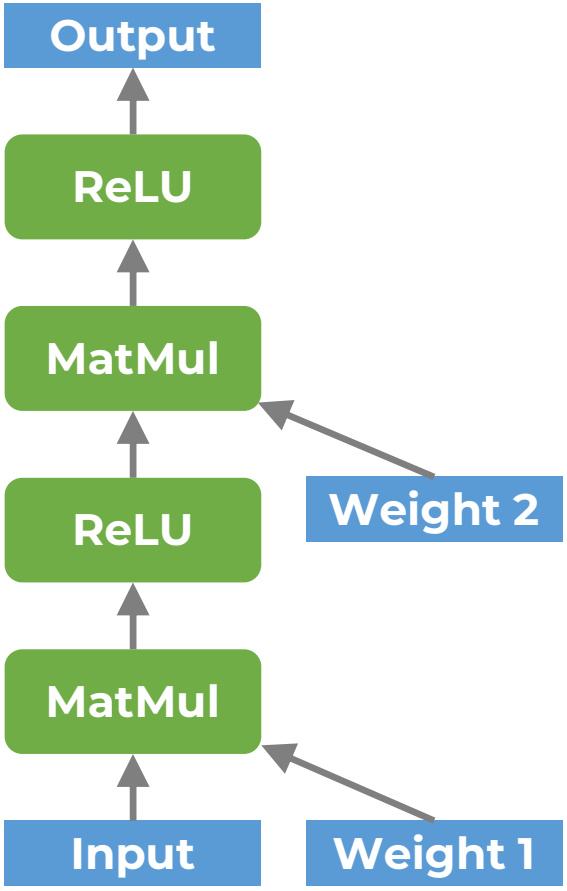
Unity

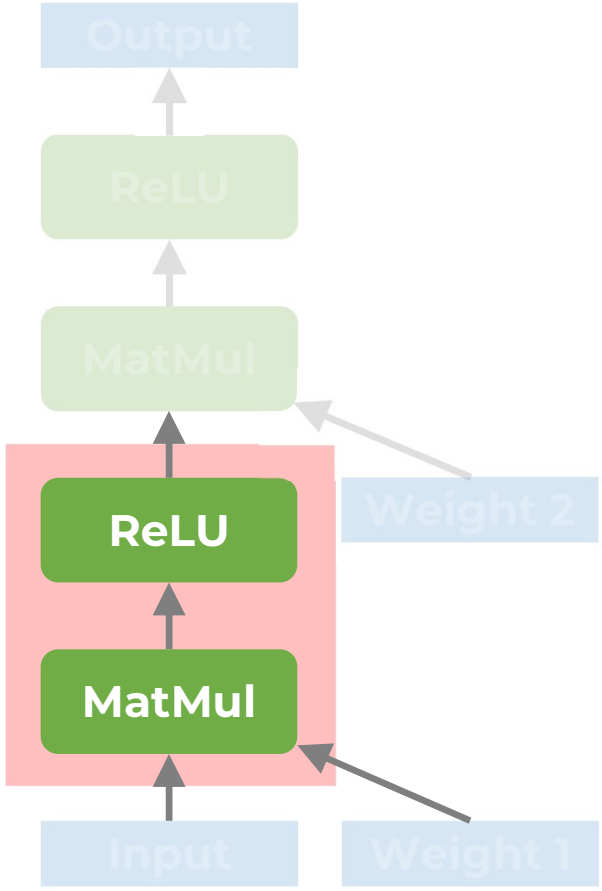
Accelerating DNN Training Through Joint Optimization
of Algebraic Transformations and Parallelization

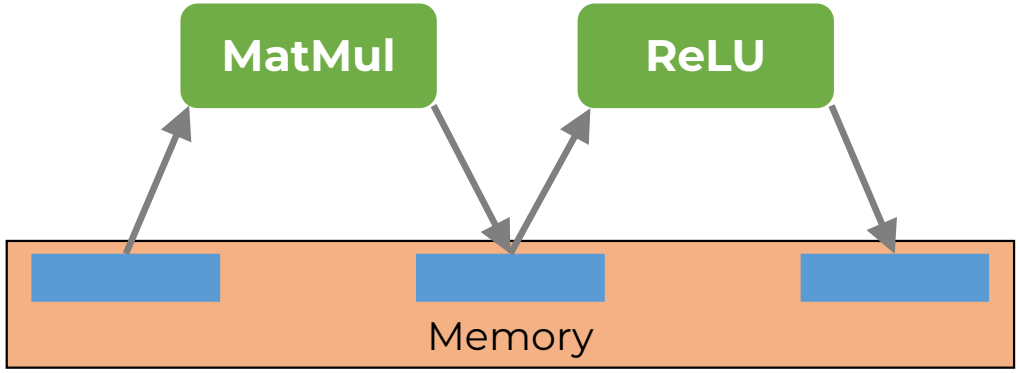
1. Algebraic Transformations

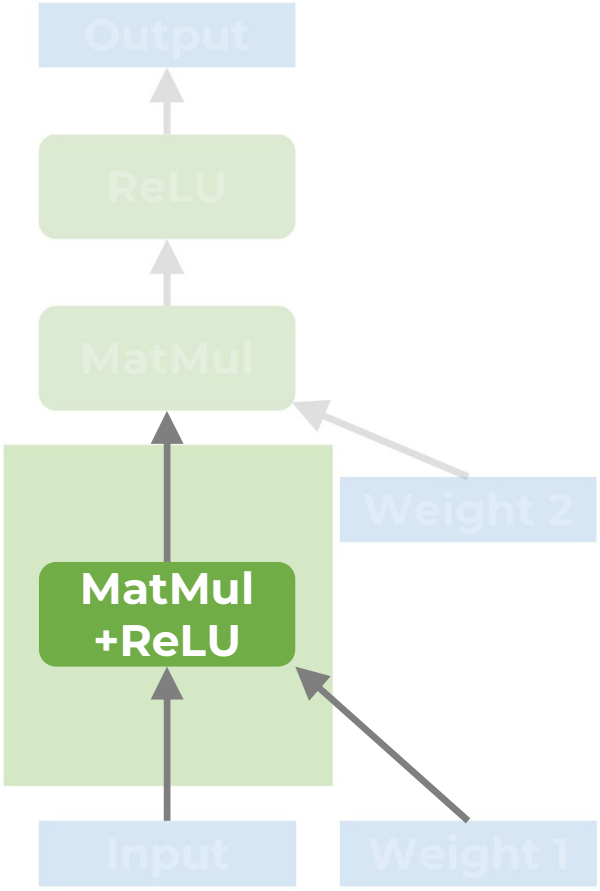
2. Parallelization

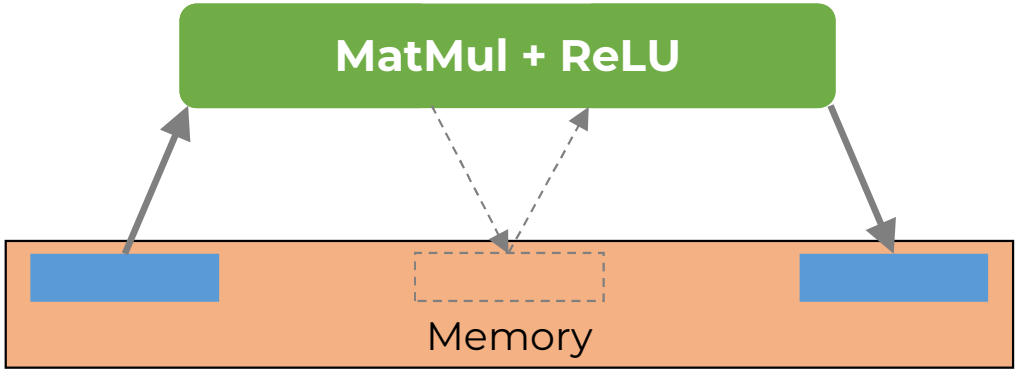
1. Algebraic Transformations

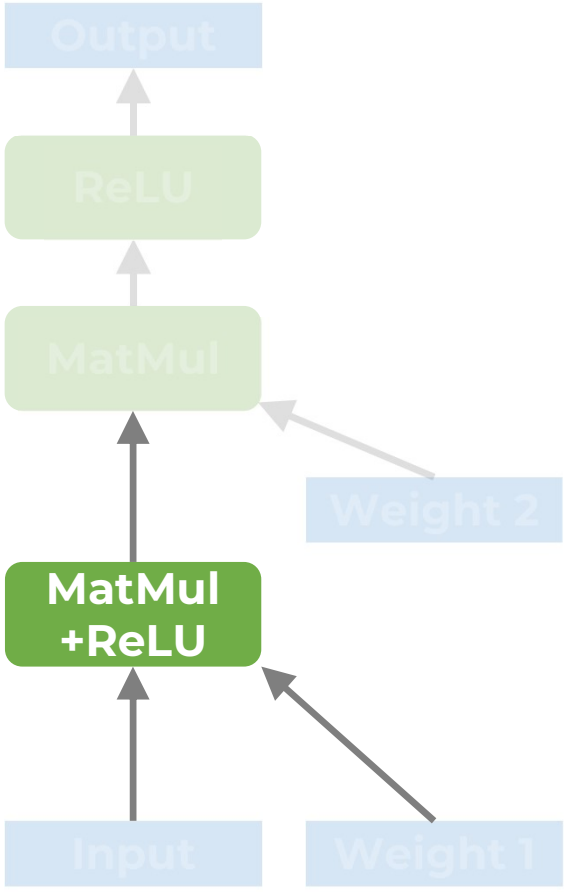

Output
ReLU
MatMul
Weight 2
ReLU
MatMul
Input
Weight 1


Output
ReLU
MatMul
Weight 2
ReLU
MatMul
Input
Weight 1

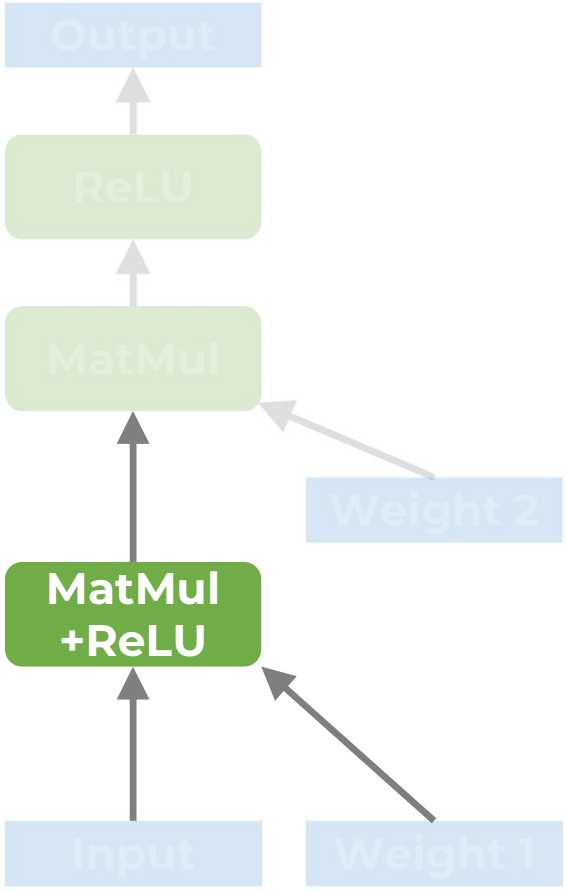

MatMul
ReLU
Memory


Output
ReLU
MatMul
Weight 2
MatMul
+ReLU
Input
Weight 1


MatMul + ReLU
Memory

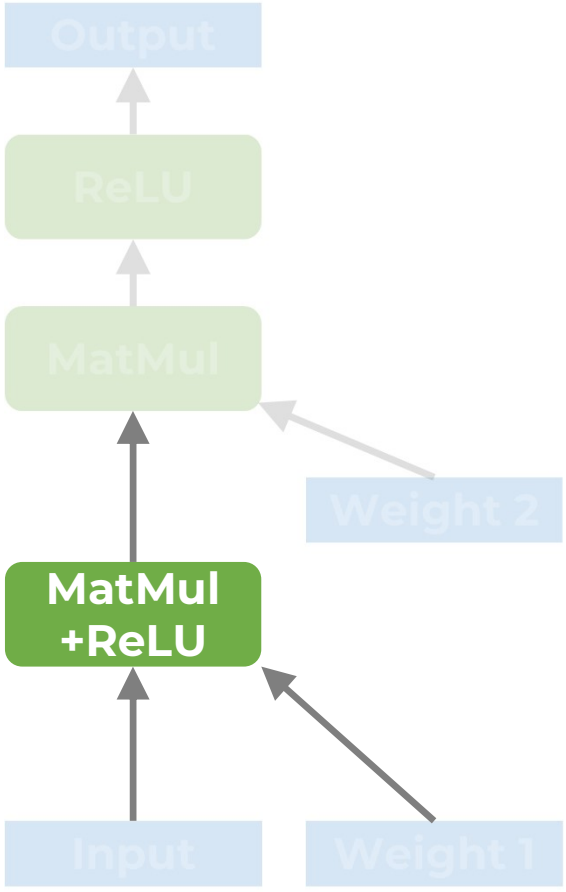


Operator Fusion



Operator Fusion

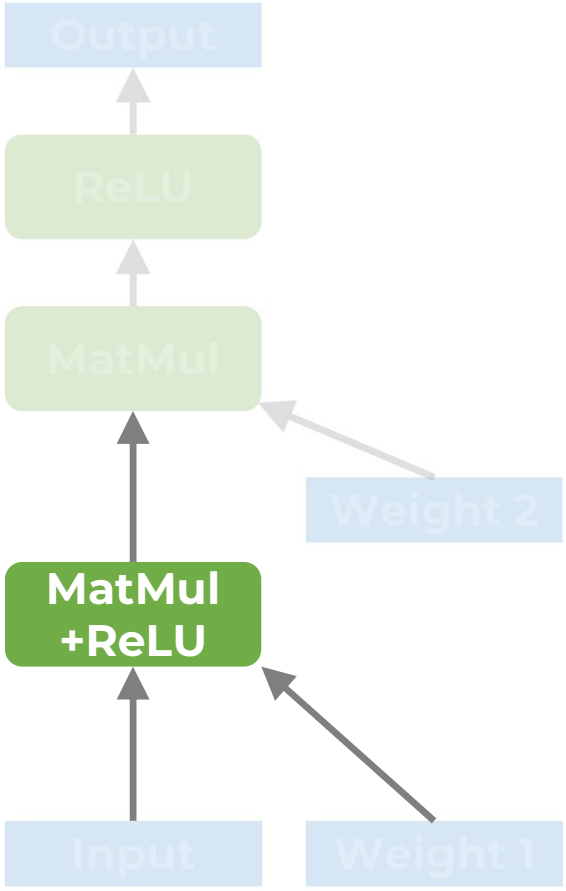
Operator Splitting



Operator Fusion

Operator Splitting

Operator Reordering



Operator Fusion

Operator Splitting

Operator Reordering

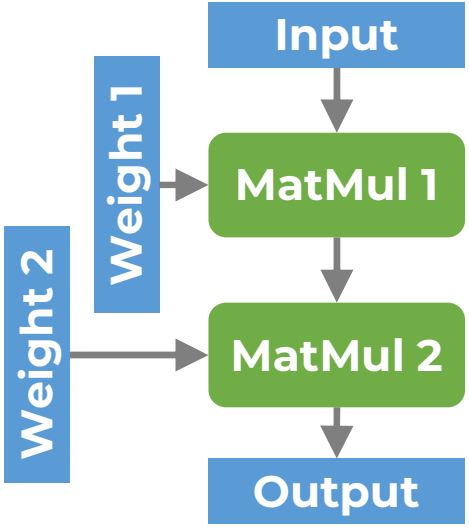
...

1. Algebraic Transformations

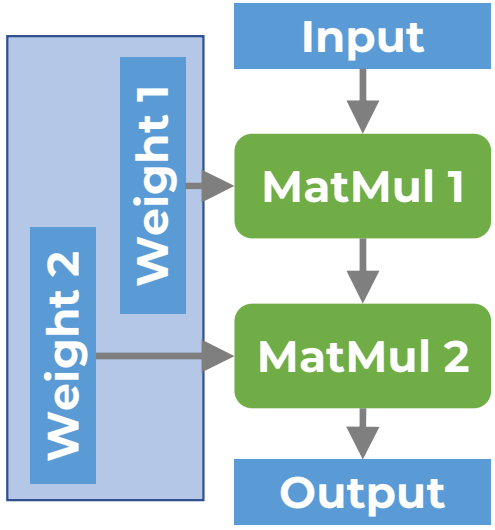
2. Parallelization

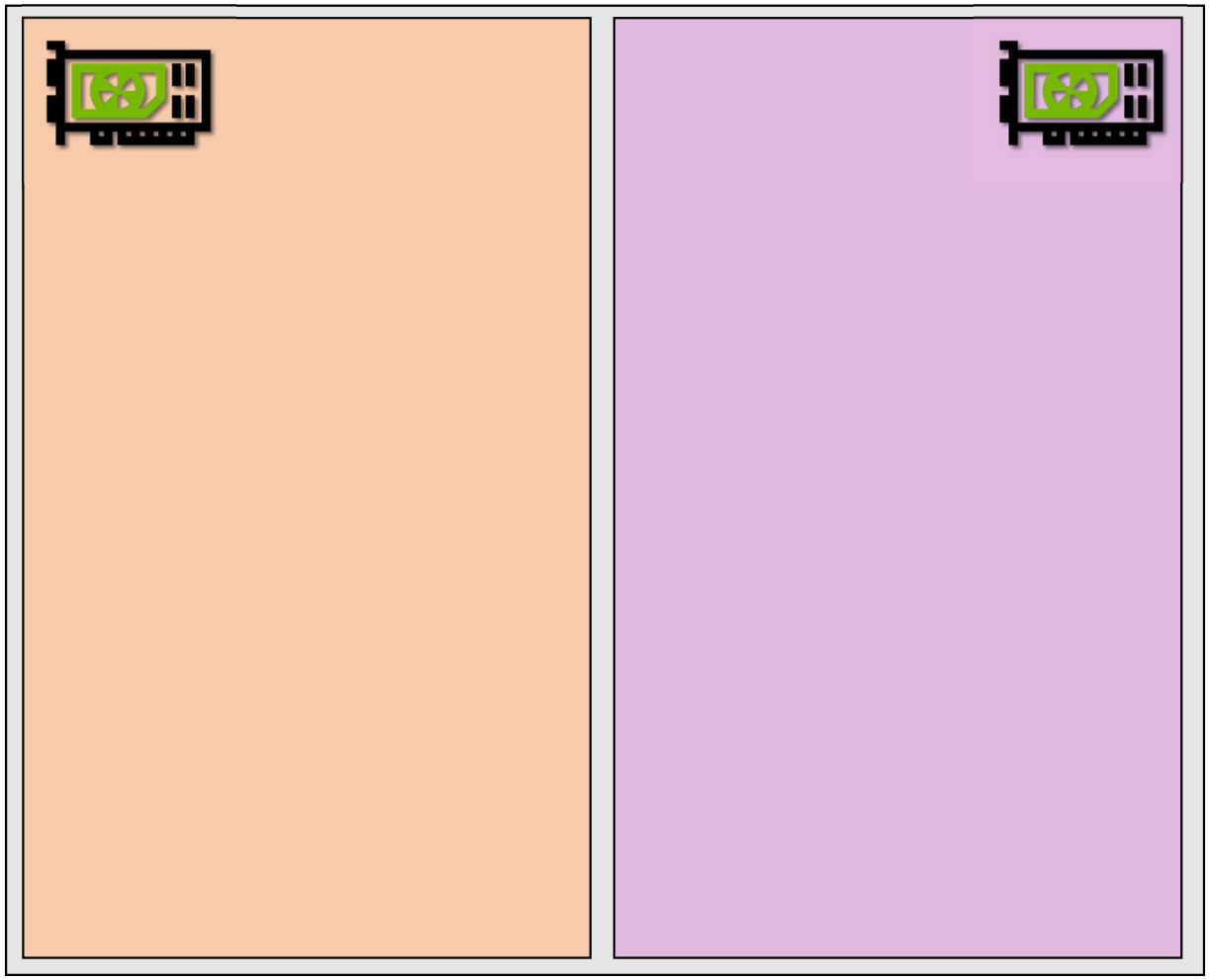
Data Parallel

Data Parallel

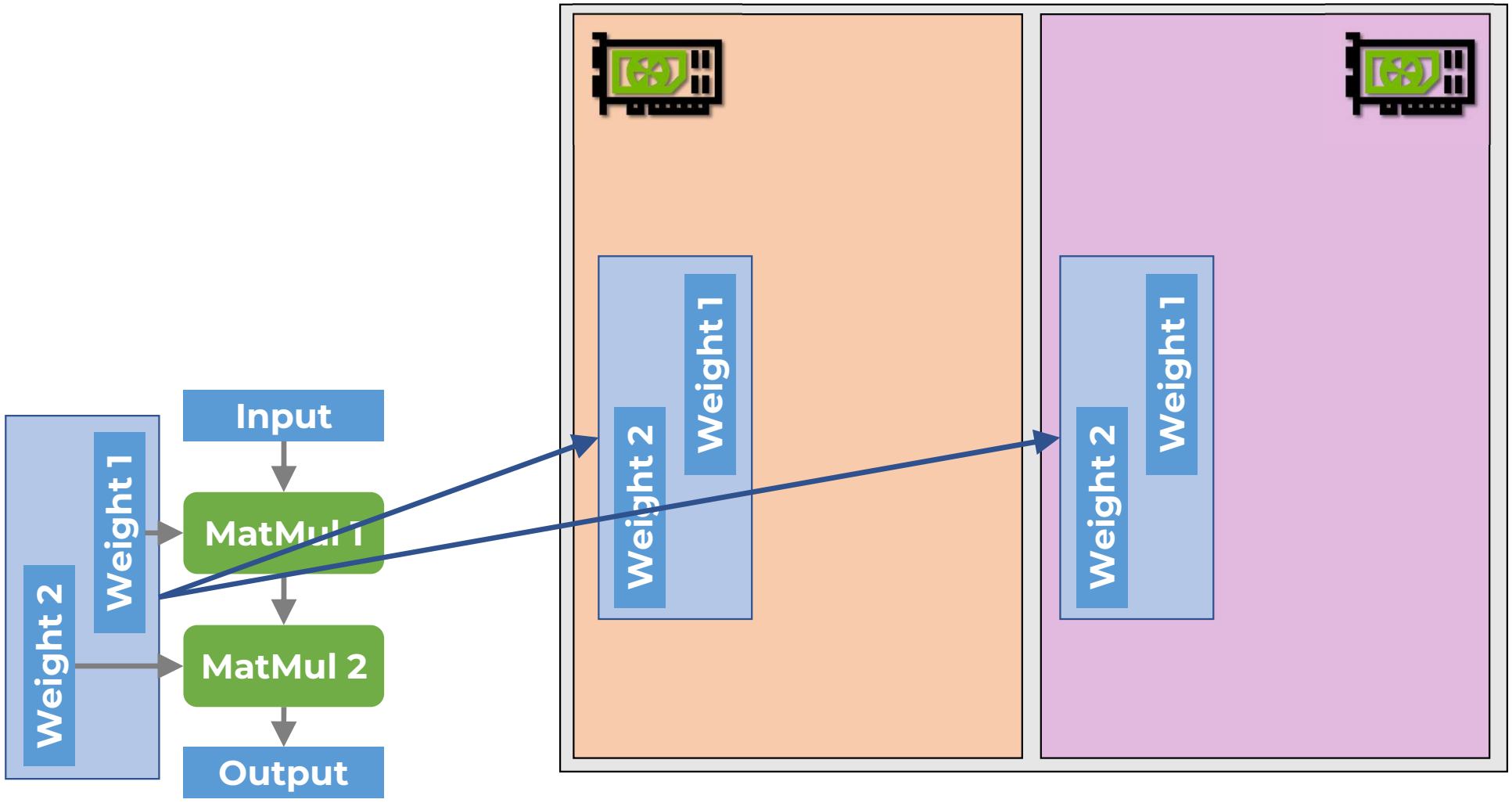


Data Parallel

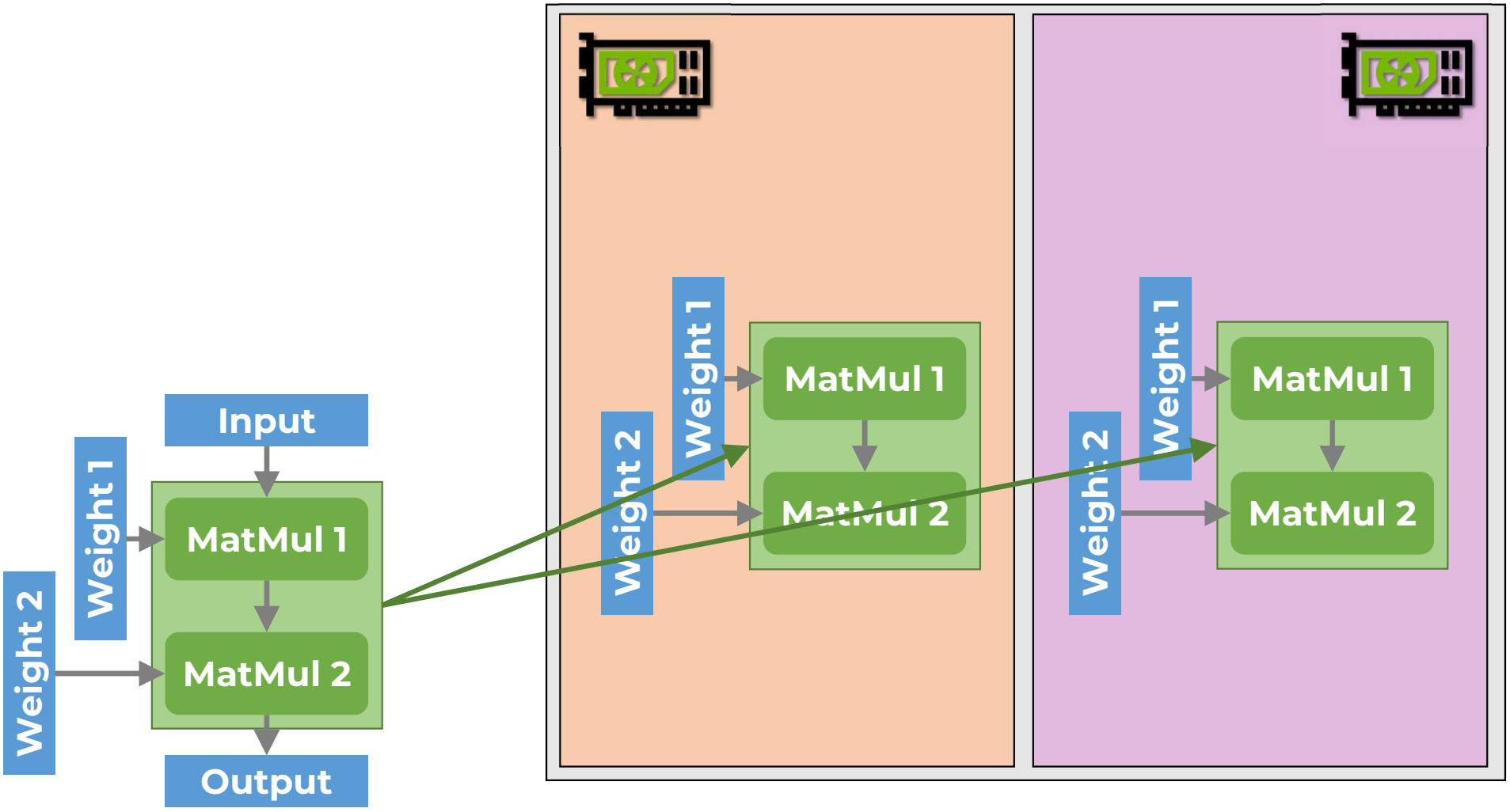




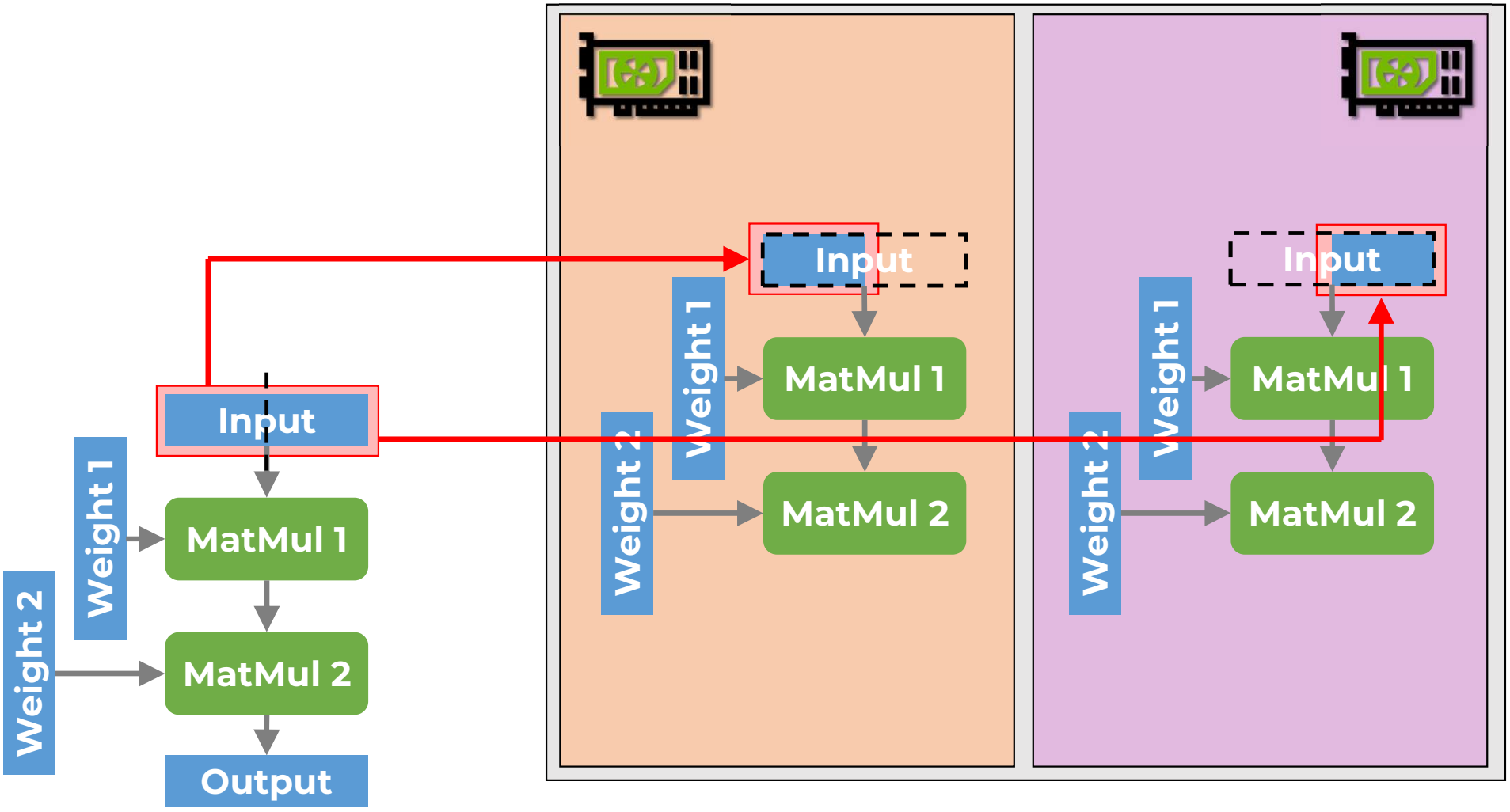
Data Parallel



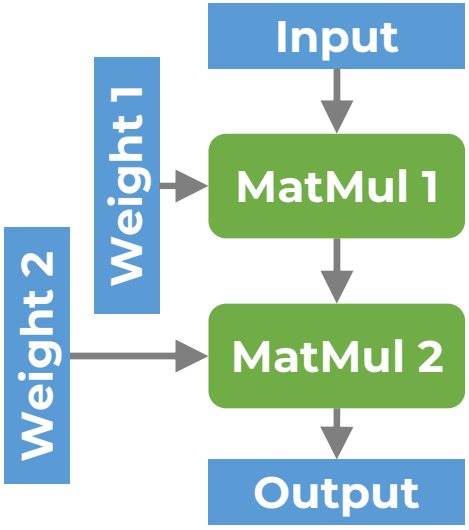
Data Parallel

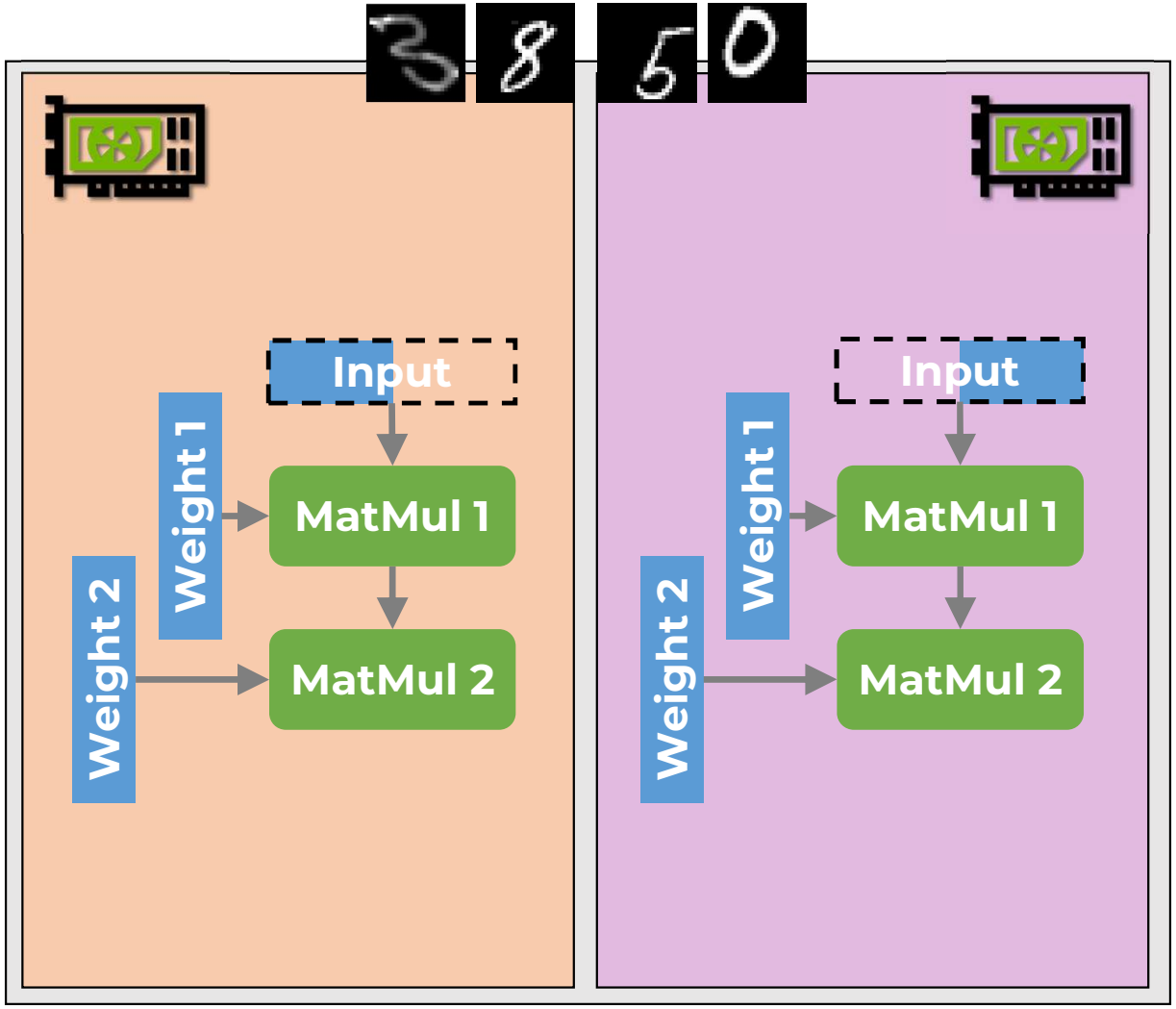


Data Parallel

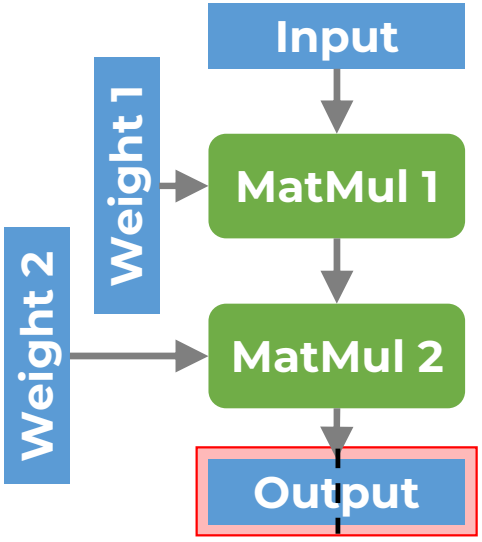


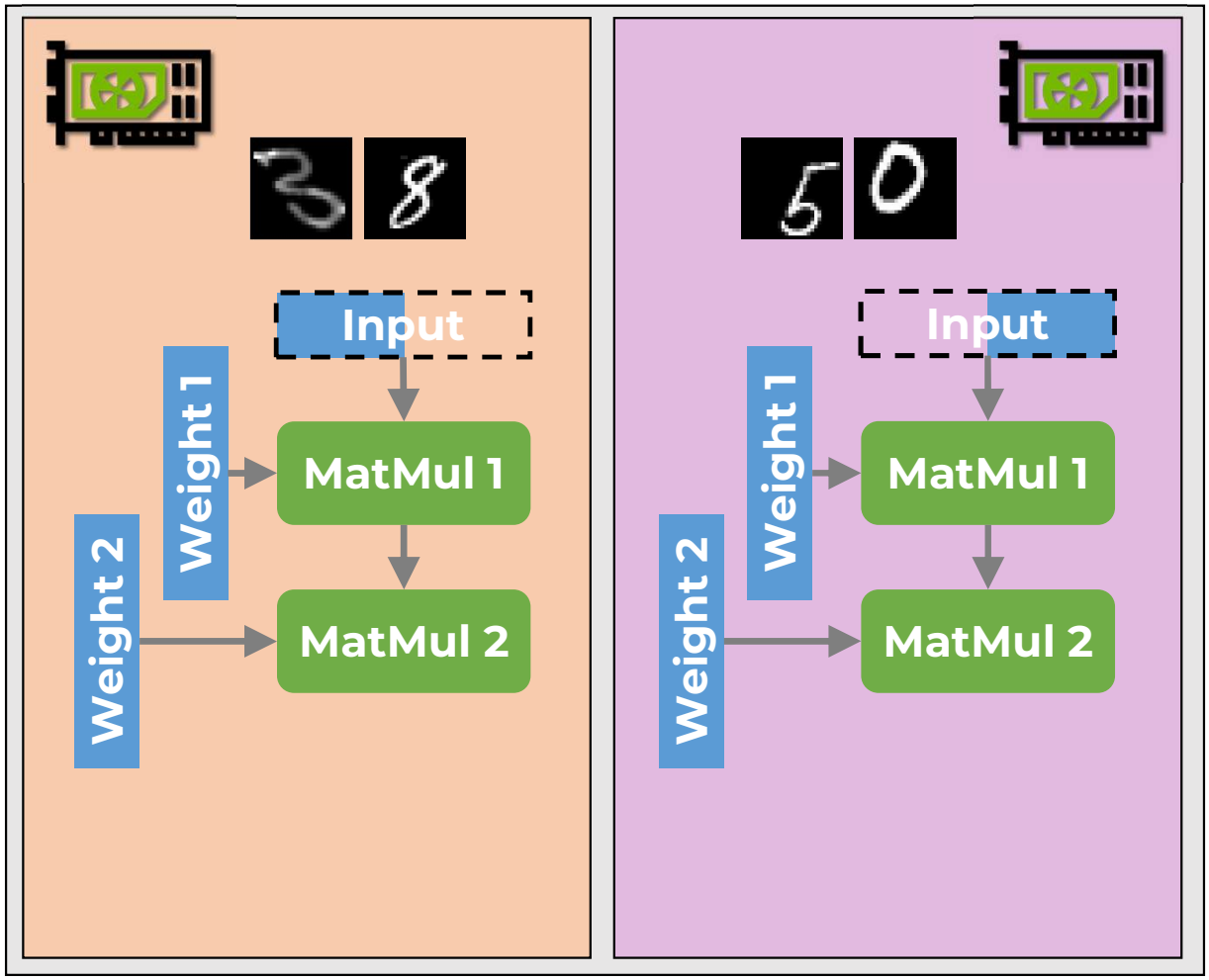
Data Parallel



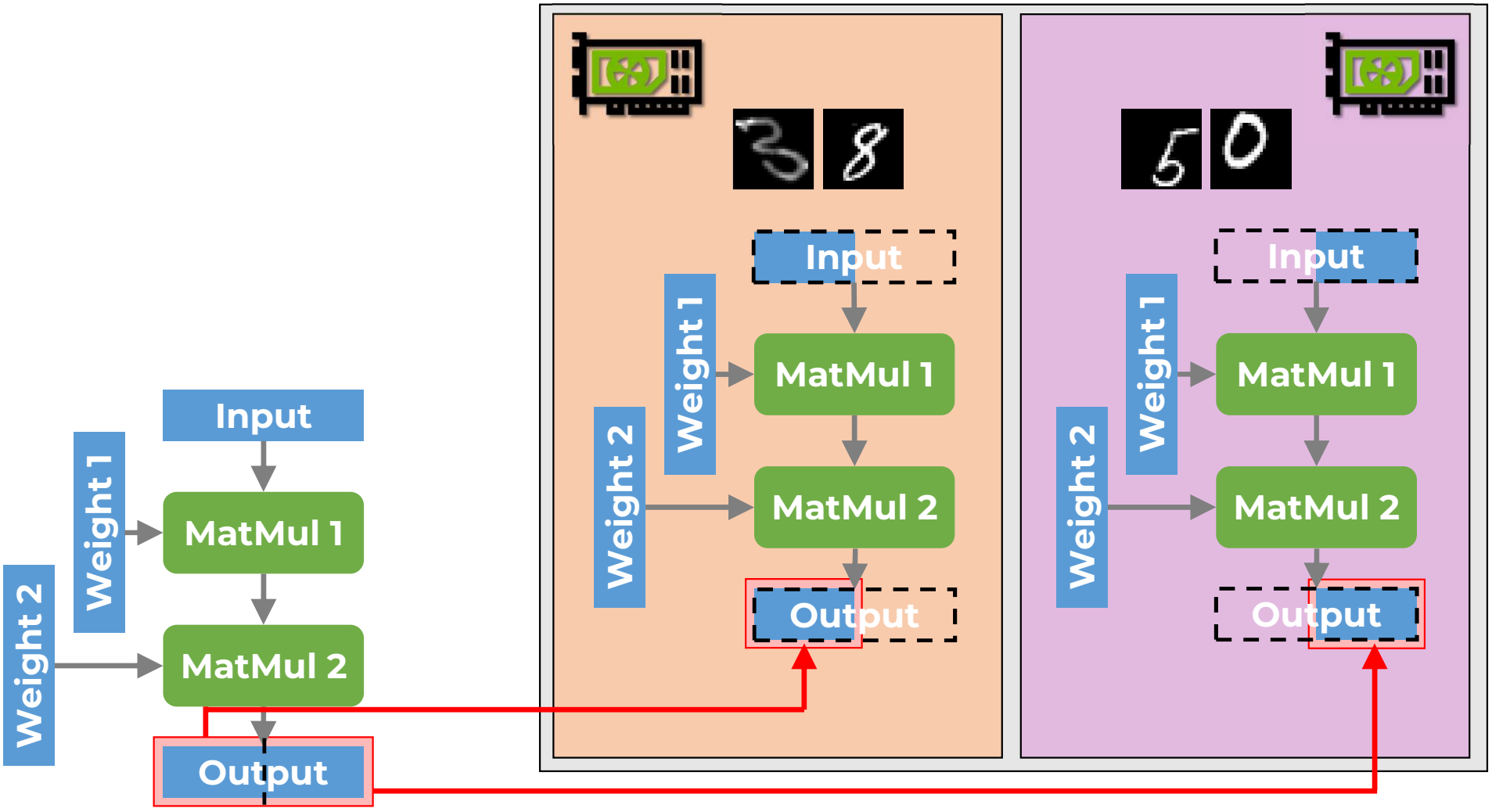


Data Parallel

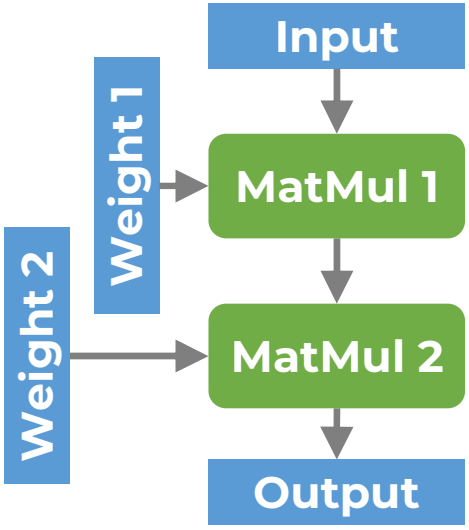


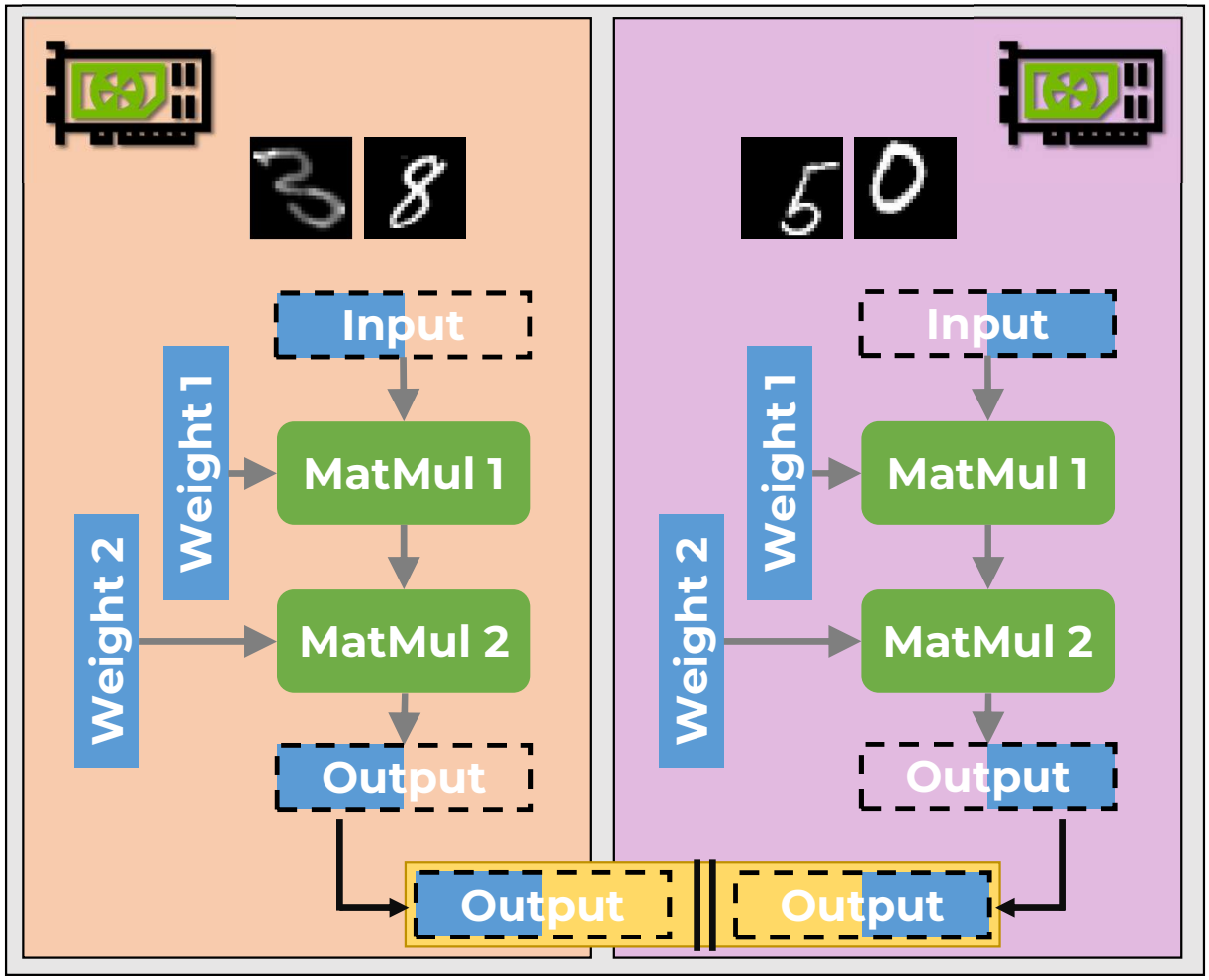


Data Parallel

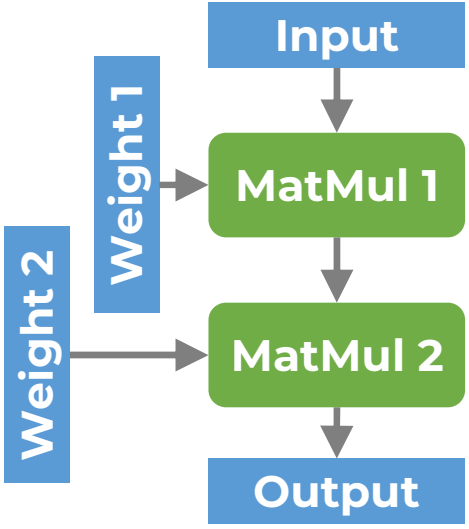


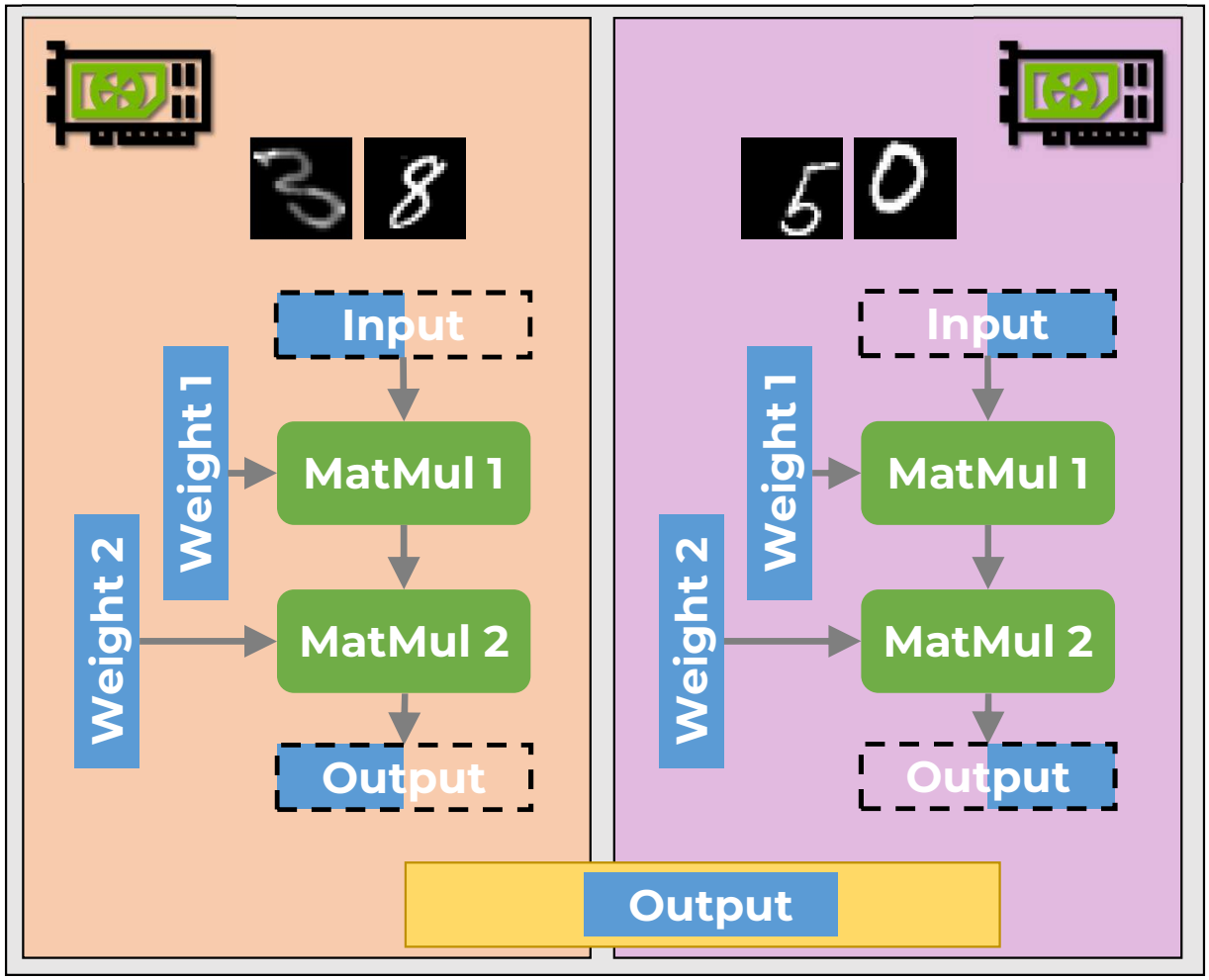
Data Parallel



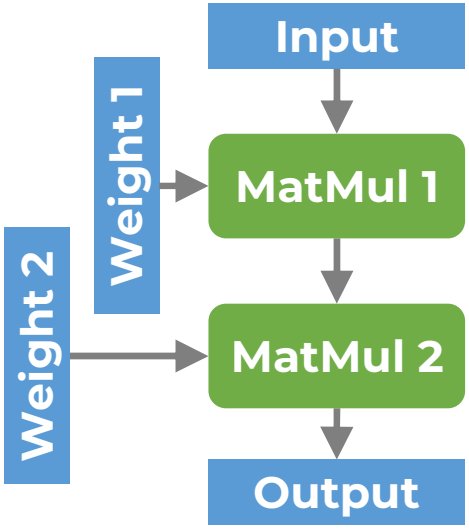


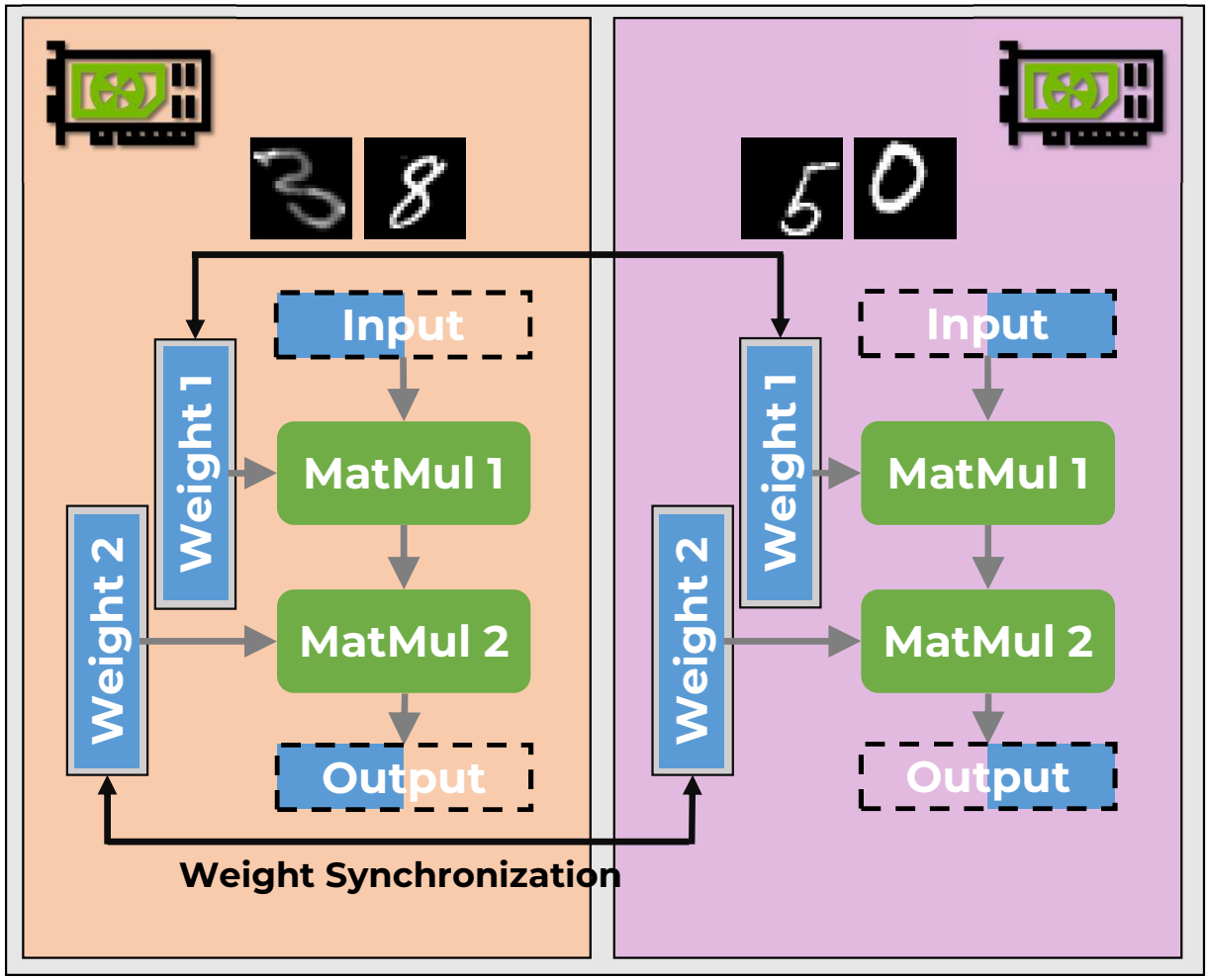
Data Parallel





Data Parallel





Data Parallel

Data Parallel

Model Parallel

Data Parallel

Model Parallel

Attribute Parallel

Operation	Parallelizable Dimensions		
	(S)ample	(A)ttribute	(P)arameter
1D pooling	sample	length, channel	
1D convolution	sample	length	channel
2D convolution	sample	height, width	channel
Matrix multiplication	sample		channel

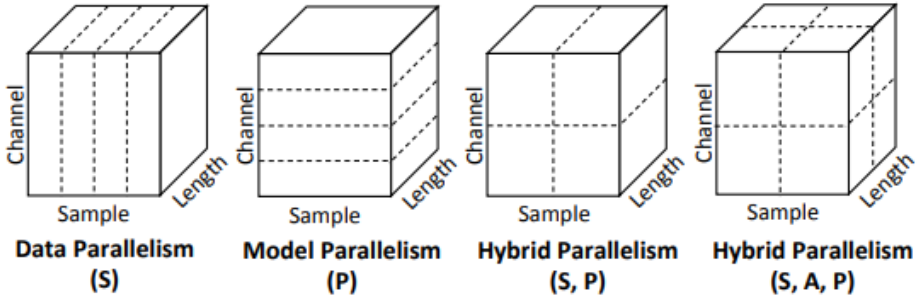


Figure 3: Example parallelization configurations for 1D convolution. Dashed lines show partitioning the tensor.

Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel

Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
Parameter Parallel

Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
Parameter Parallel
Pipeline Parallel

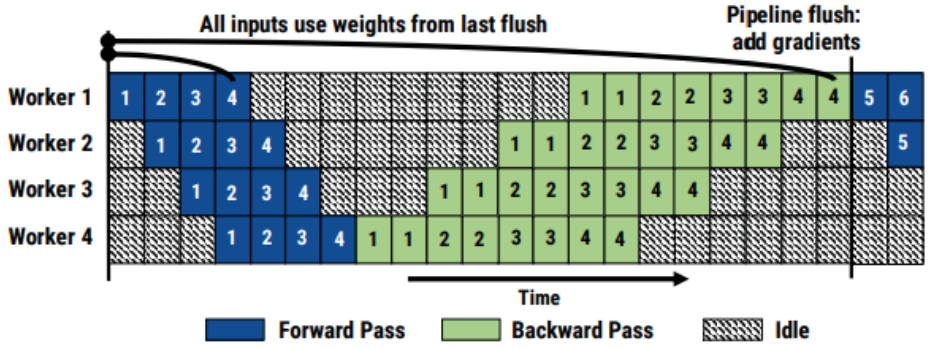


Figure 3: GPipe's inter-batch parallelism approach. Frequent pipeline flushes lead to increased idle time.

Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
Parameter Parallel
Pipeline Parallel
...

2 Parallelization

Algebraic Transformations

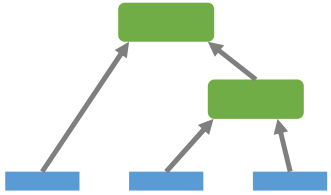
Operator Fusion
Operator Splitting
Operator Reordering
...

Auto-Parallelization

FlexFlow [MLSys 19]
Tofu [EuroSys 19]
PipeDream [SOSP 19]
automap [arXiv 19]
Whale [arXiv 21]
Alpa [OSDI 22]
...

Algebraic Optimizers

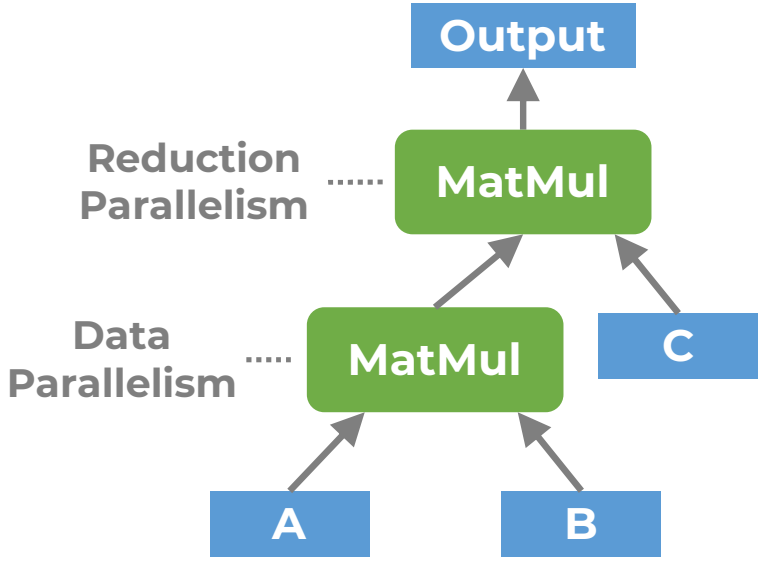
MetaFlow [MLSys 19]
TASO [SOSP 19]
PET [OSDI 21]
Tensat [MLSys 21]
...



Auto-Parallelization

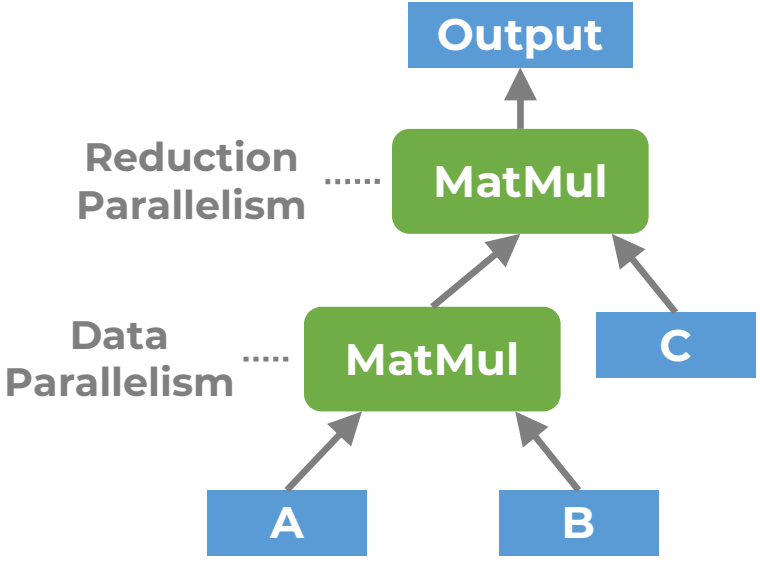
Algebraic Optimizer

?



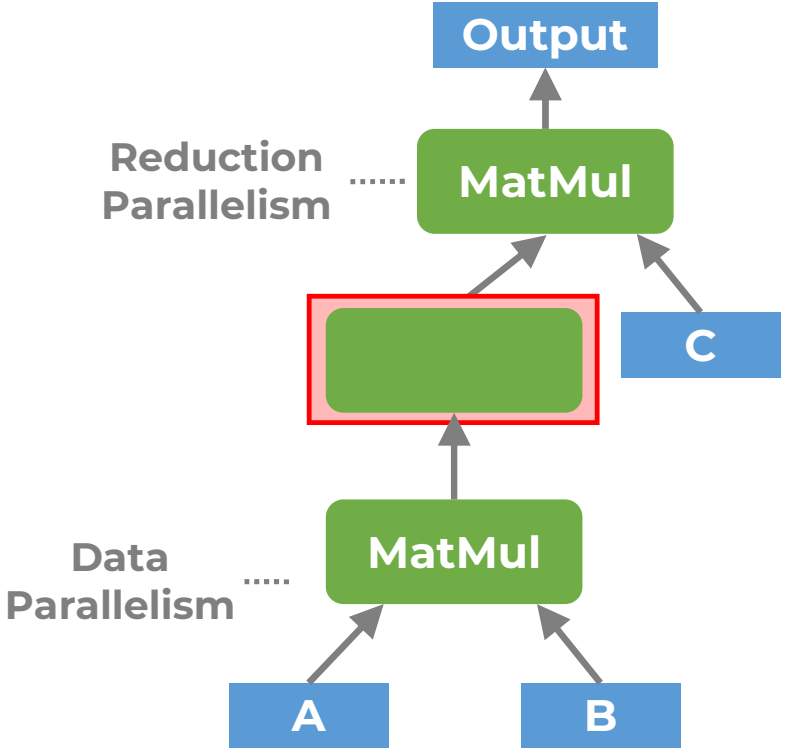
“annotated computation graph”


Auto-
Parallelization
Algebraic
Optimizer


Output
Reduction
Parallelism
MatMul
Data
Parallelism
MatMul
C
A
B

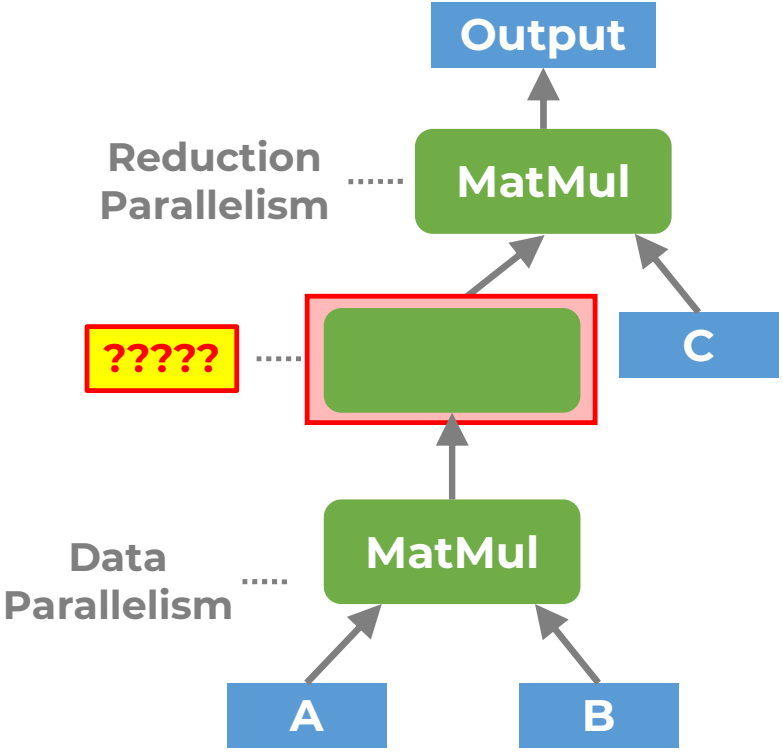
Auto-Parallelization

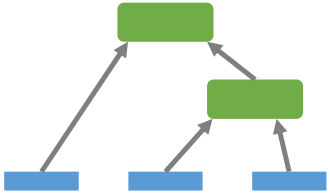
Algebraic Optimizer



Auto-Parallelization

Algebraic Optimizer



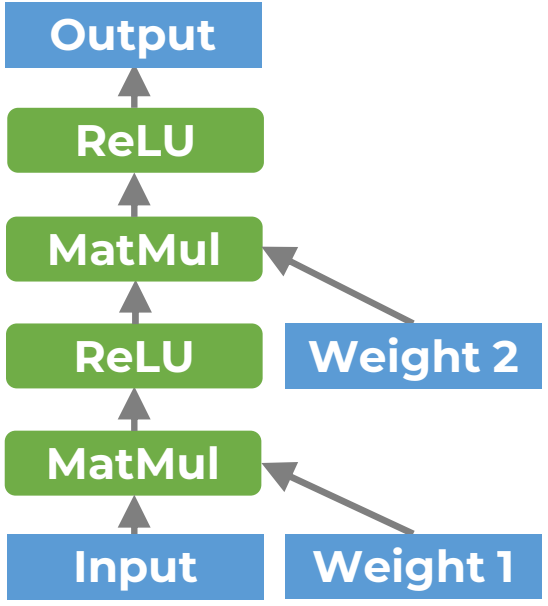


Algebraic Optimizer

Auto-Parallelization

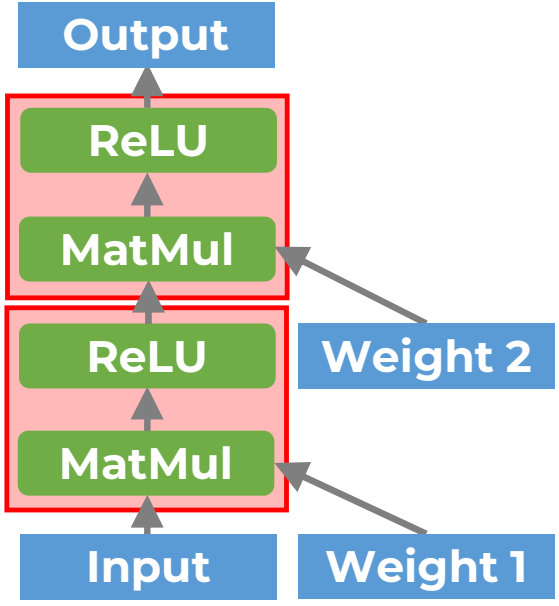
Algebraic Optimizer

Auto-Parallelization



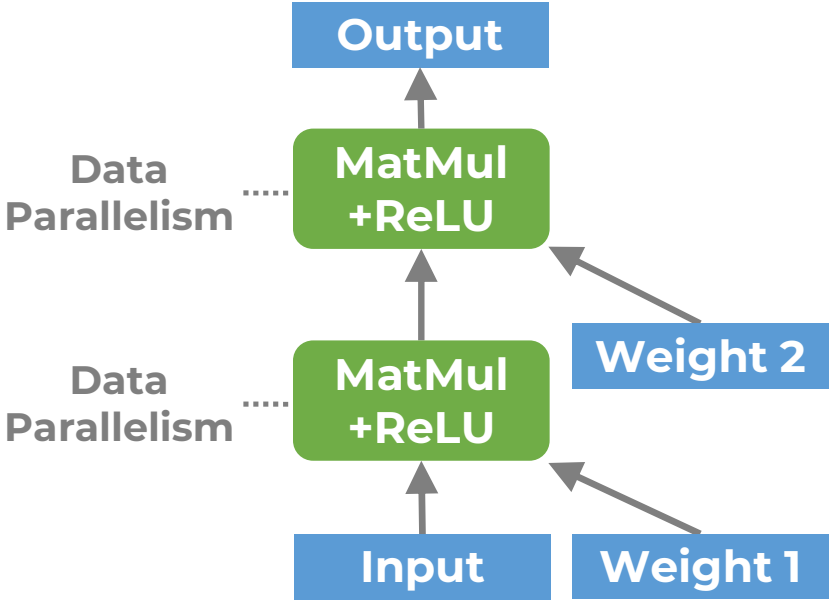
Algebraic Optimizer

Auto-Parallelization

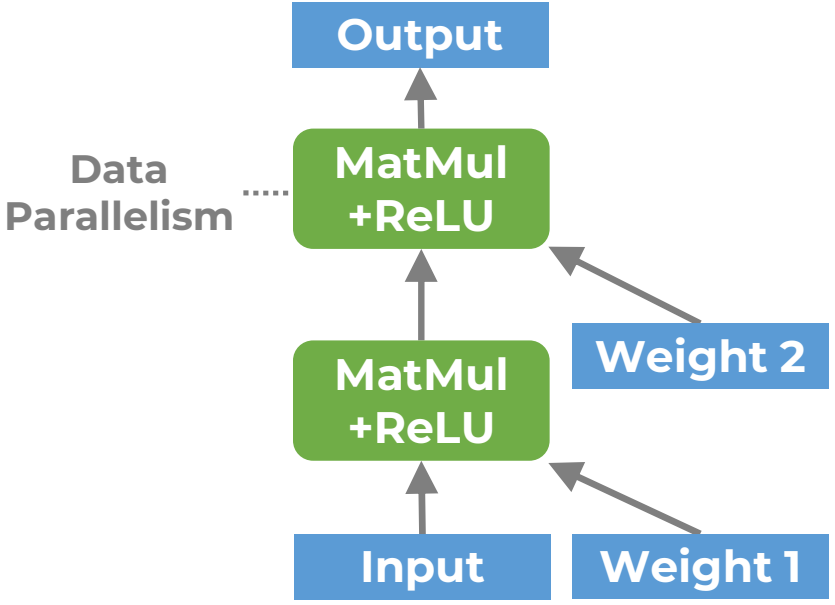


Algebraic Optimizer

Auto-Parallelization

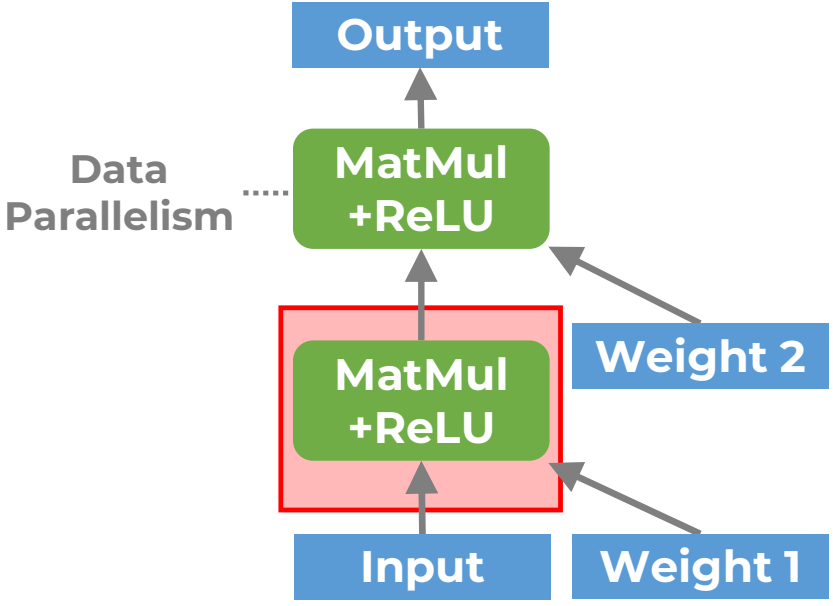


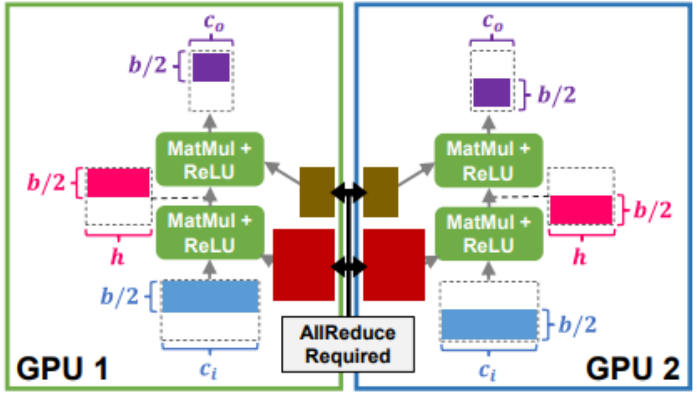

Algebraic Optimizer
Auto-Parallelization


Output
Data Parallelism
MatMul +ReLU
Weight 2
MatMul +ReLU
Input
Weight 1

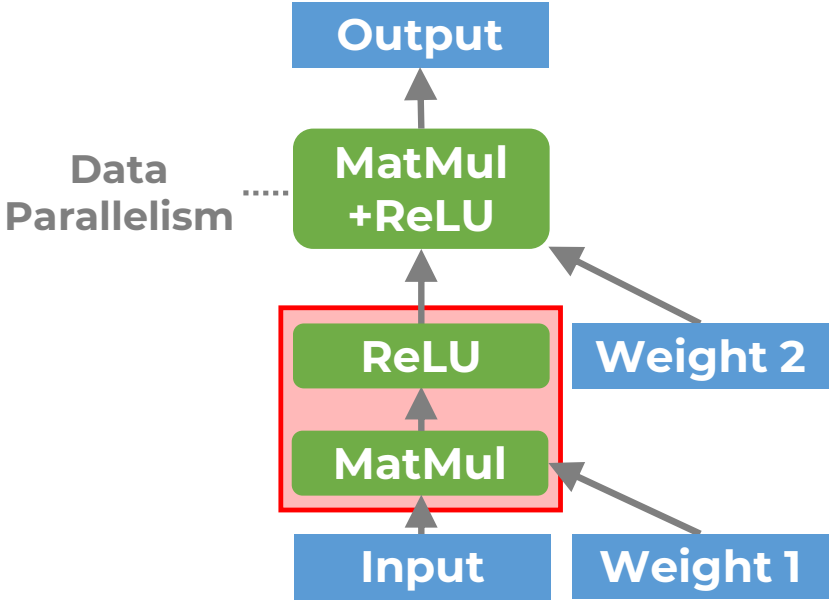

Algebraic
Optimizer

Auto-
Parallelization


Output
Data
Parallelism
MatMul
+ReLU
MatMul
+ReLU
Weight 2
Input
Weight 1

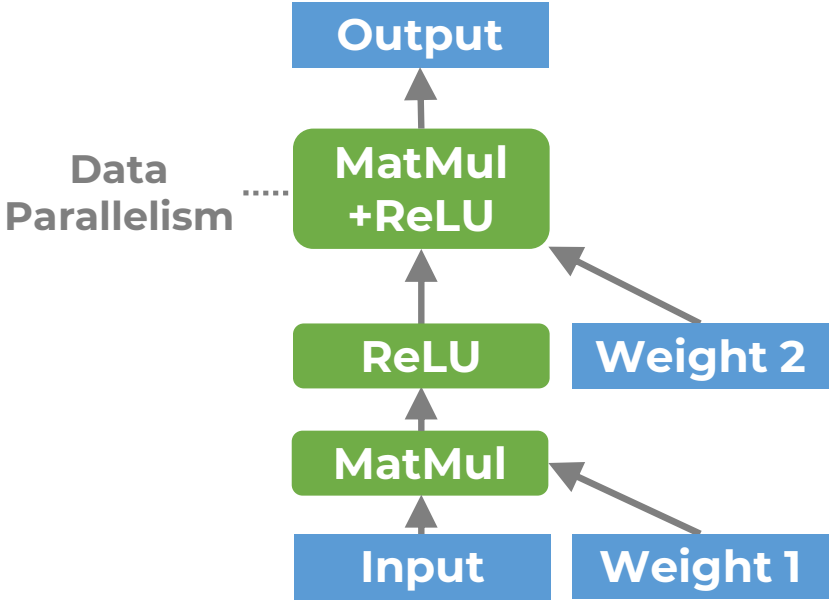

c_o
$b/2$
MatMul +
ReLU
h
c_i
AllReduce
Required
GPU 1
GPU 2


Algebraic Optimizer
Auto-Parallelization


Output
Data Parallelism
MatMul +ReLU
ReLU
Weight 2
MatMul
Input
Weight 1

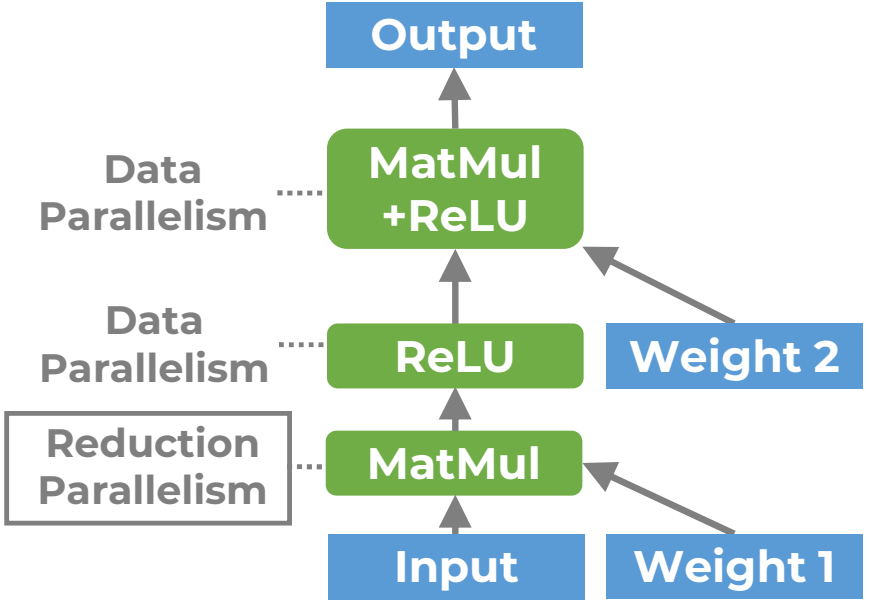
Algebraic Optimizer

Auto-Parallelization



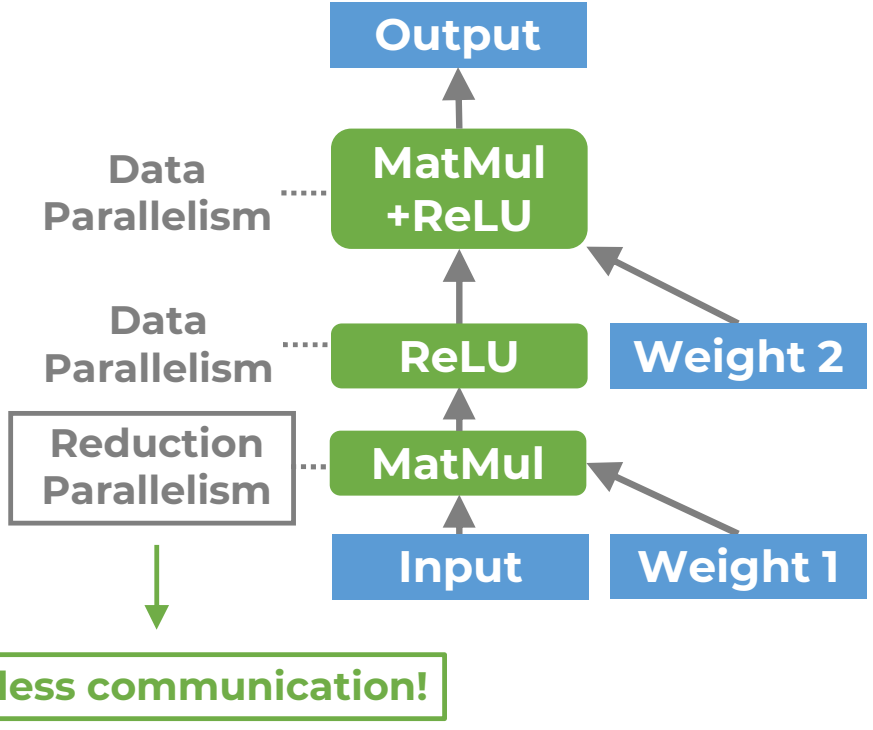
Algebraic Optimizer

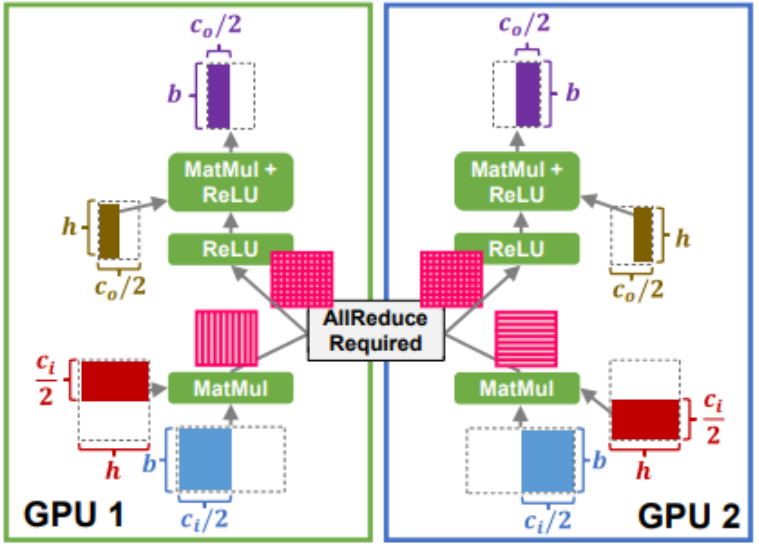
Auto-Parallelization



Algebraic Optimizer

Auto-Parallelization



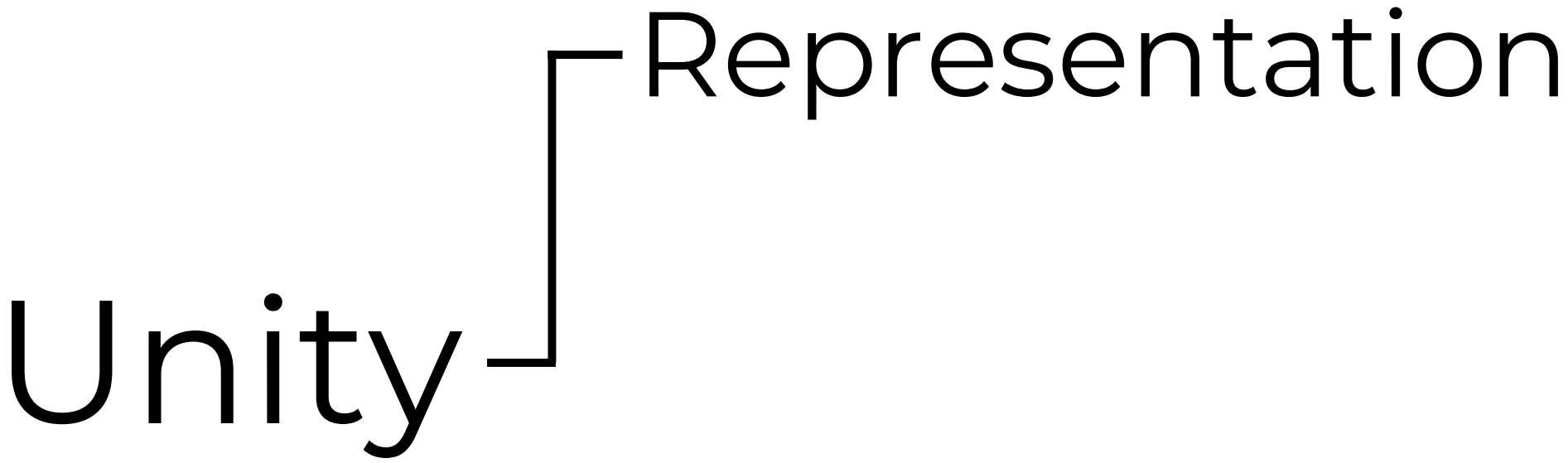


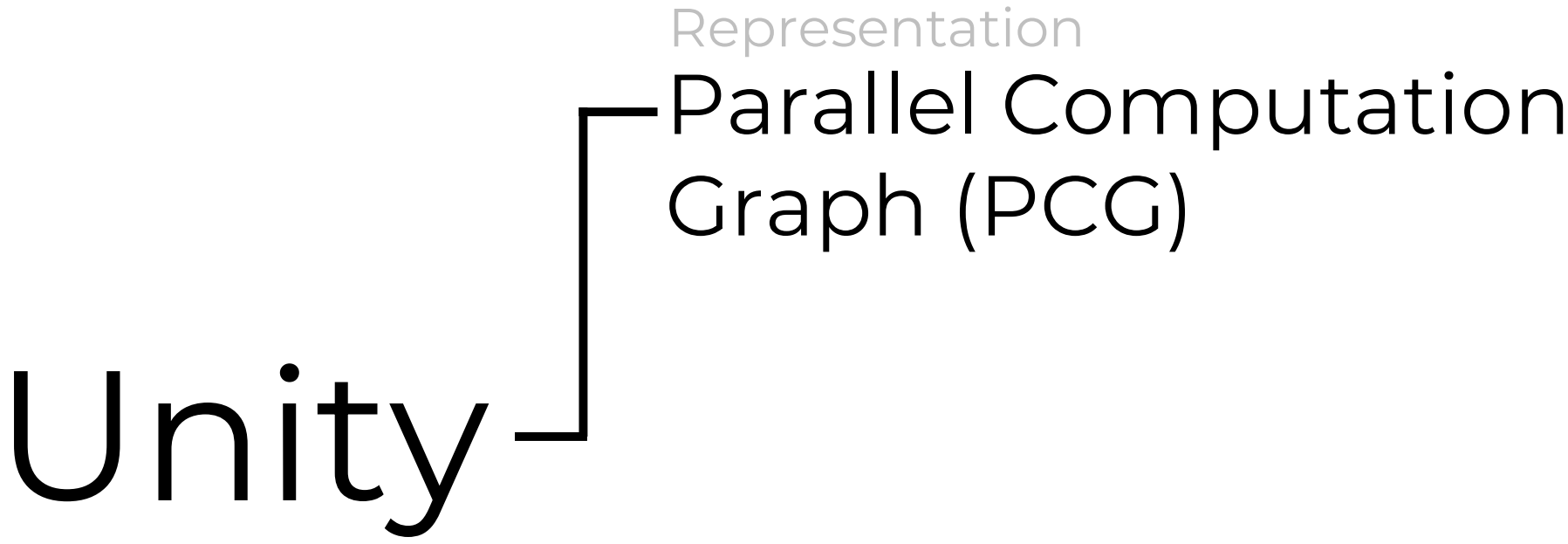
1. Representation

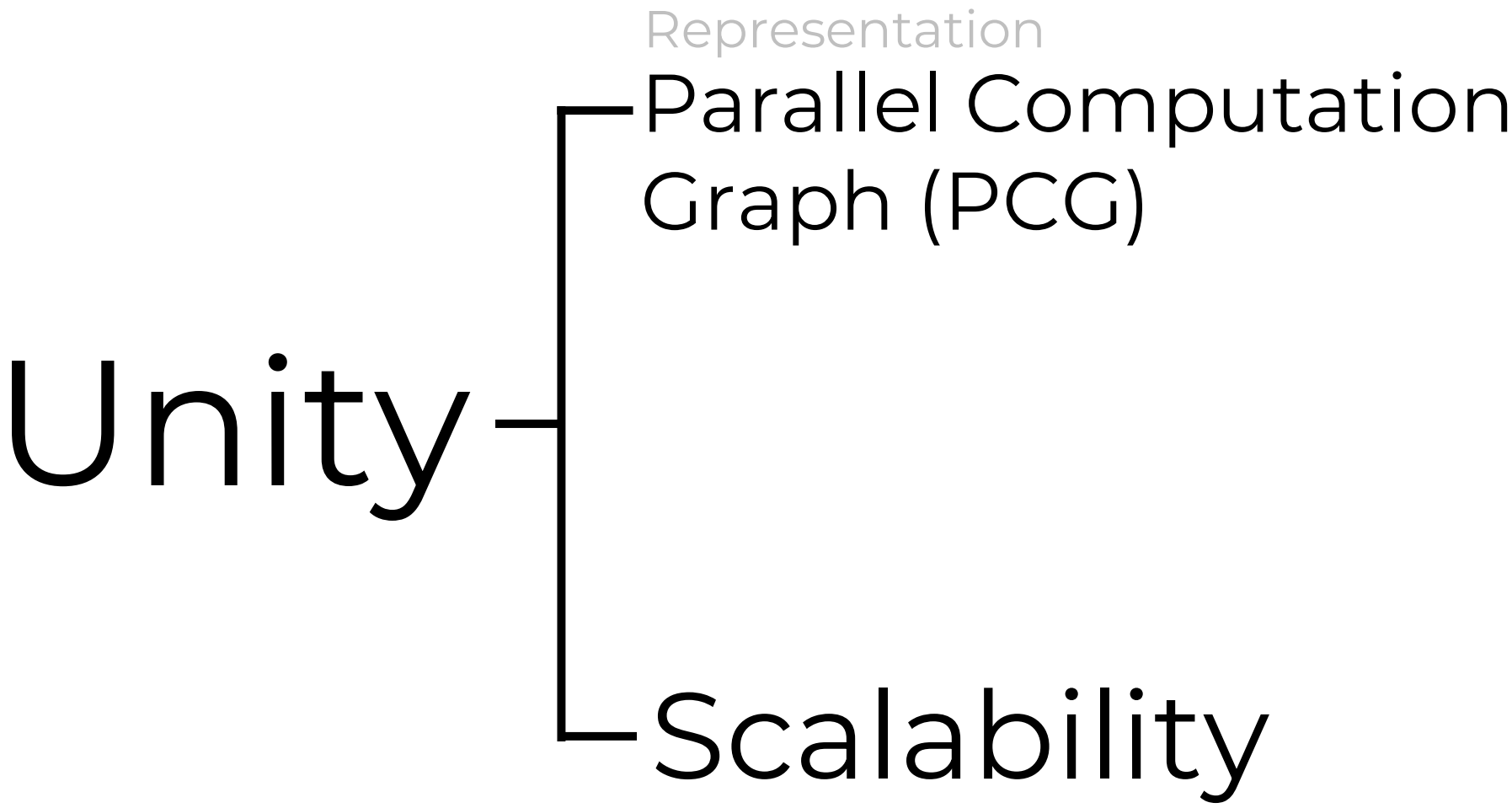
2.

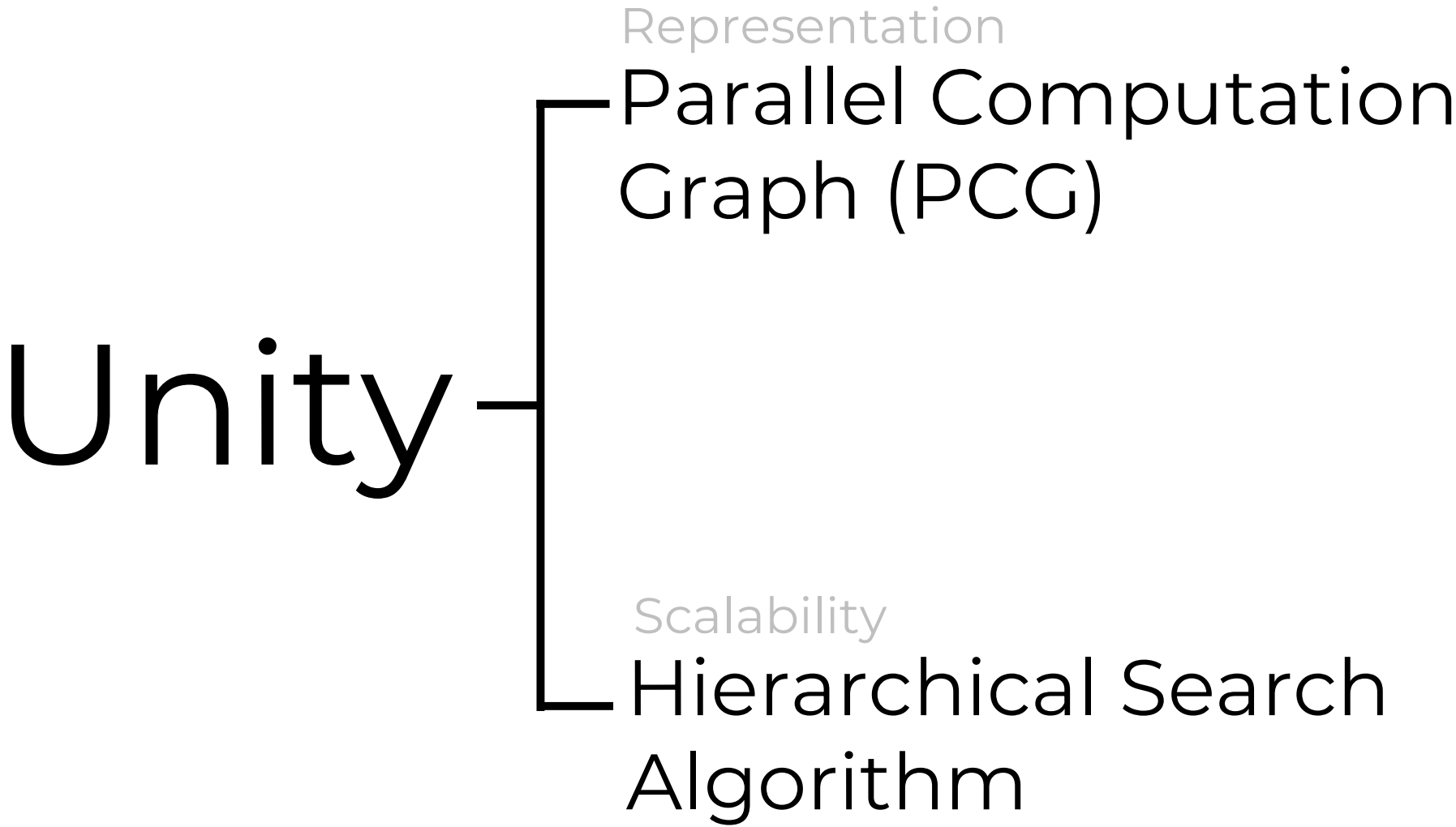
1. Representation
2. Scalability

Unity


Representation
Unity


Representation
Parallel Computation Graph (PCG)
Unity

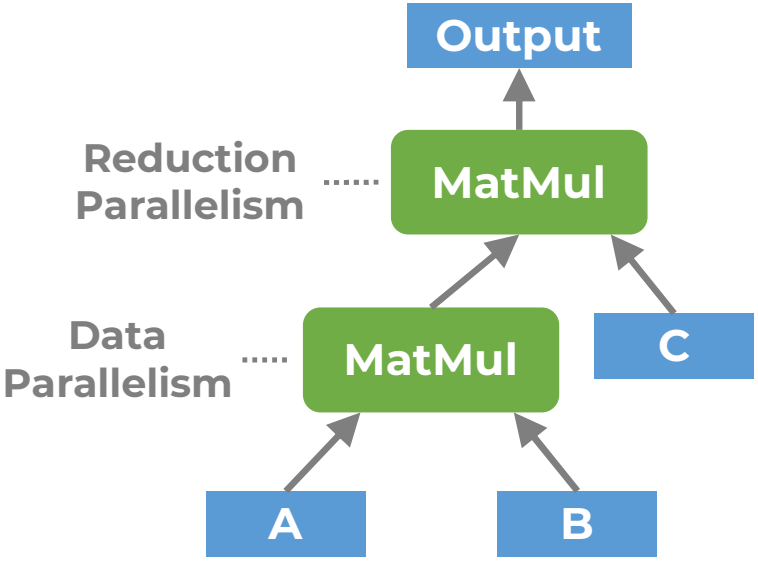

Representation
Parallel Computation Graph (PCG)
Unity
Scalability


Representation
Parallel Computation Graph (PCG)
Unity
Scalability
Hierarchical Search Algorithm

Representation

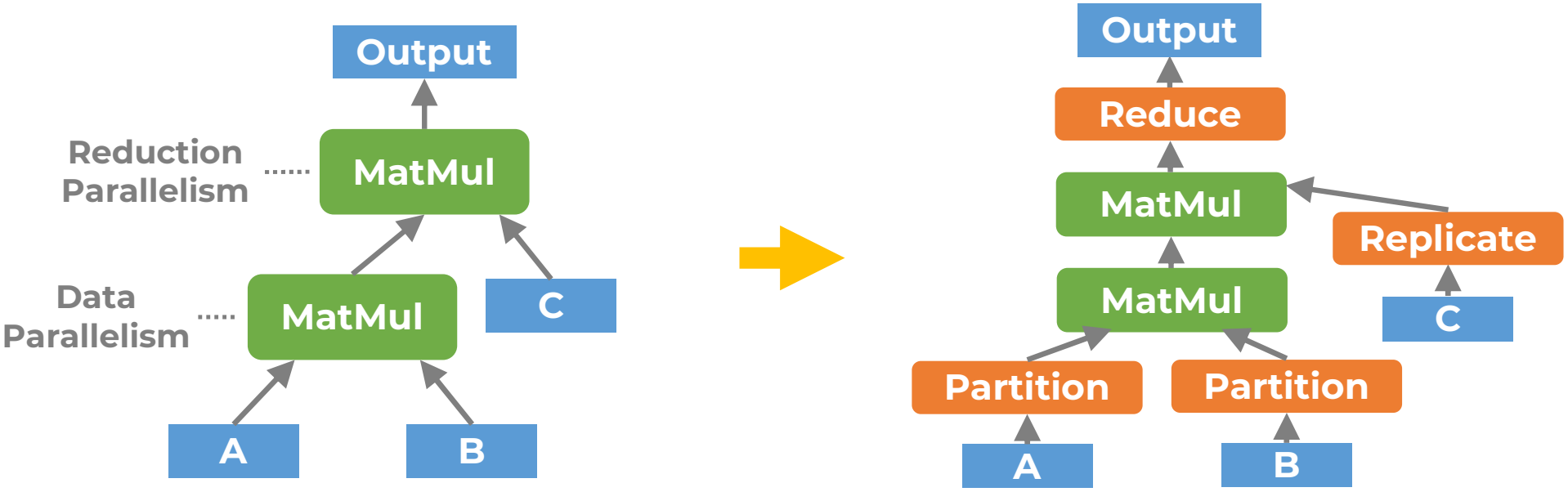
Parallel Computation Graph (PCG)

Parallel Computation Graph (PCG)



annotated computation graph

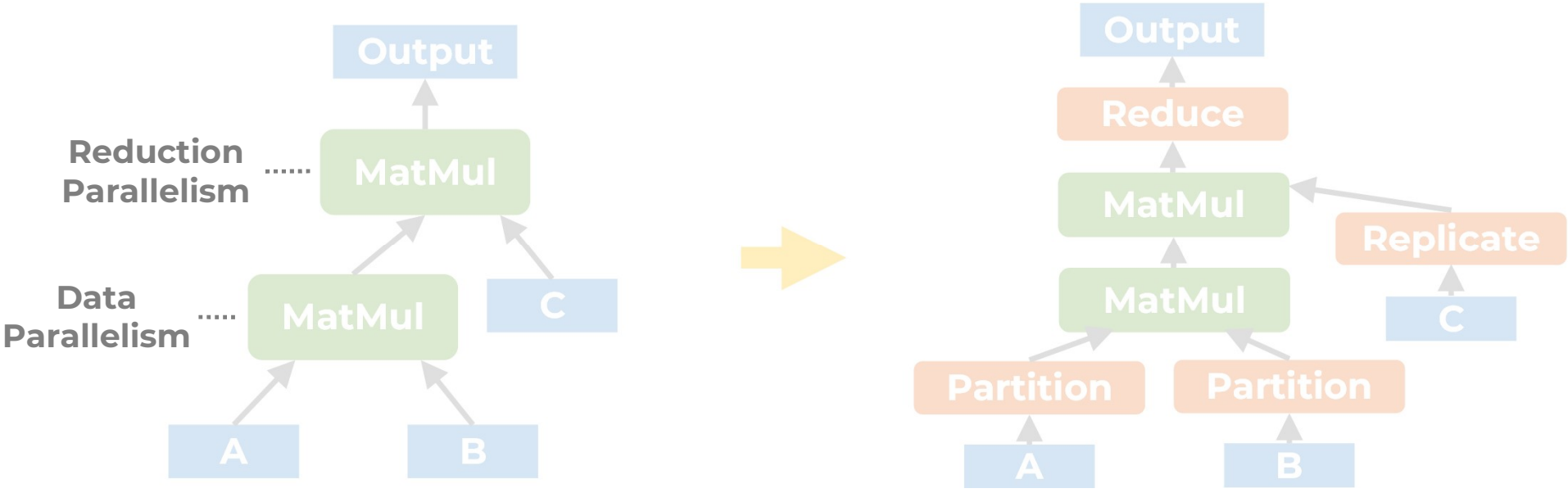
Parallel Computation Graph (PCG)



annotated computation graph

parallel computation graph (PCG)

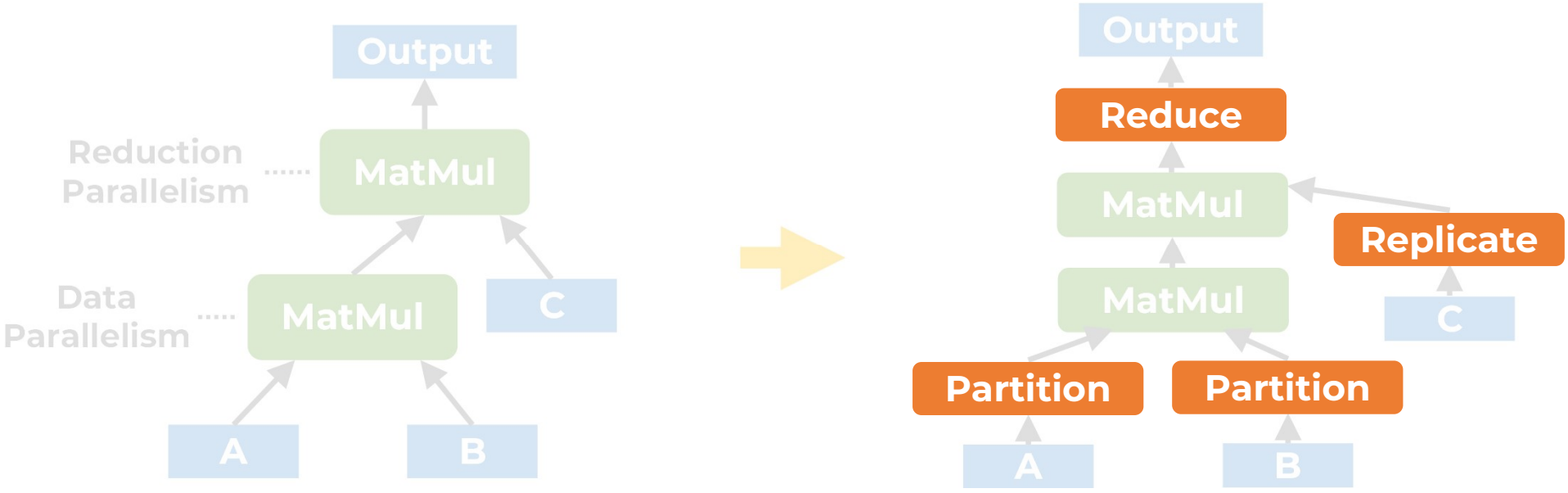
Parallel Computation Graph (PCG)



annotated computation graph

parallel computation graph (PCG)

Parallel Computation Graph (PCG)

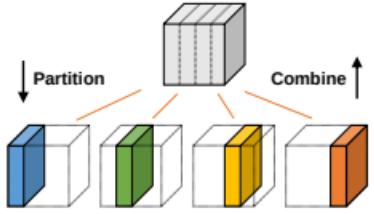


annotated computation graph

parallel computation graph (PCG)

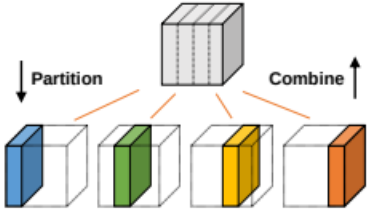
Partition

Combine



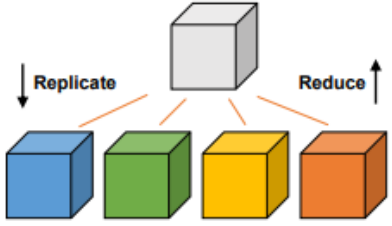
Partition

Combine



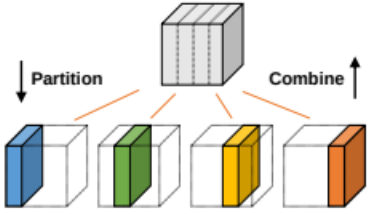
Replicate

Reduce



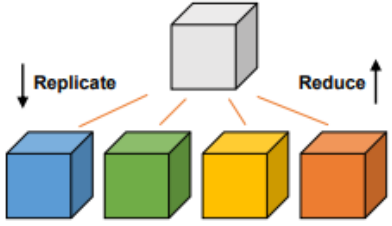
Partition

Combine



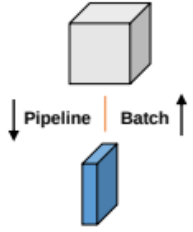
Replicate

Reduce



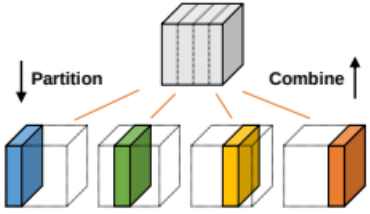
Pipeline

Batch



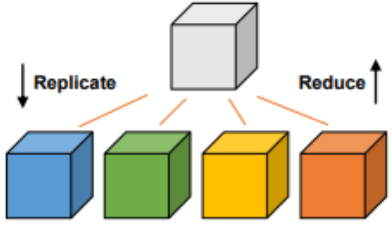
Partition

Combine



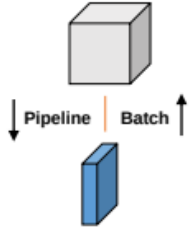
Replicate

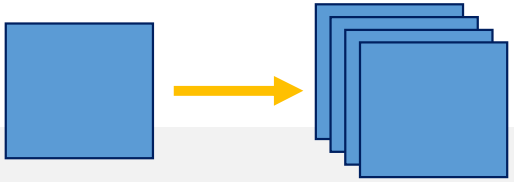
Reduce



Pipeline

Batch





Partition

Replicate

Pipeline

Combine

Reduce

Batch

Partition

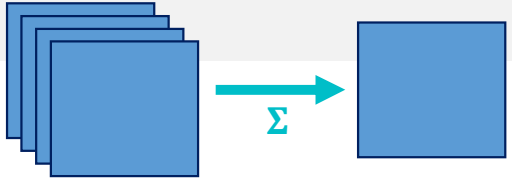
Replicate

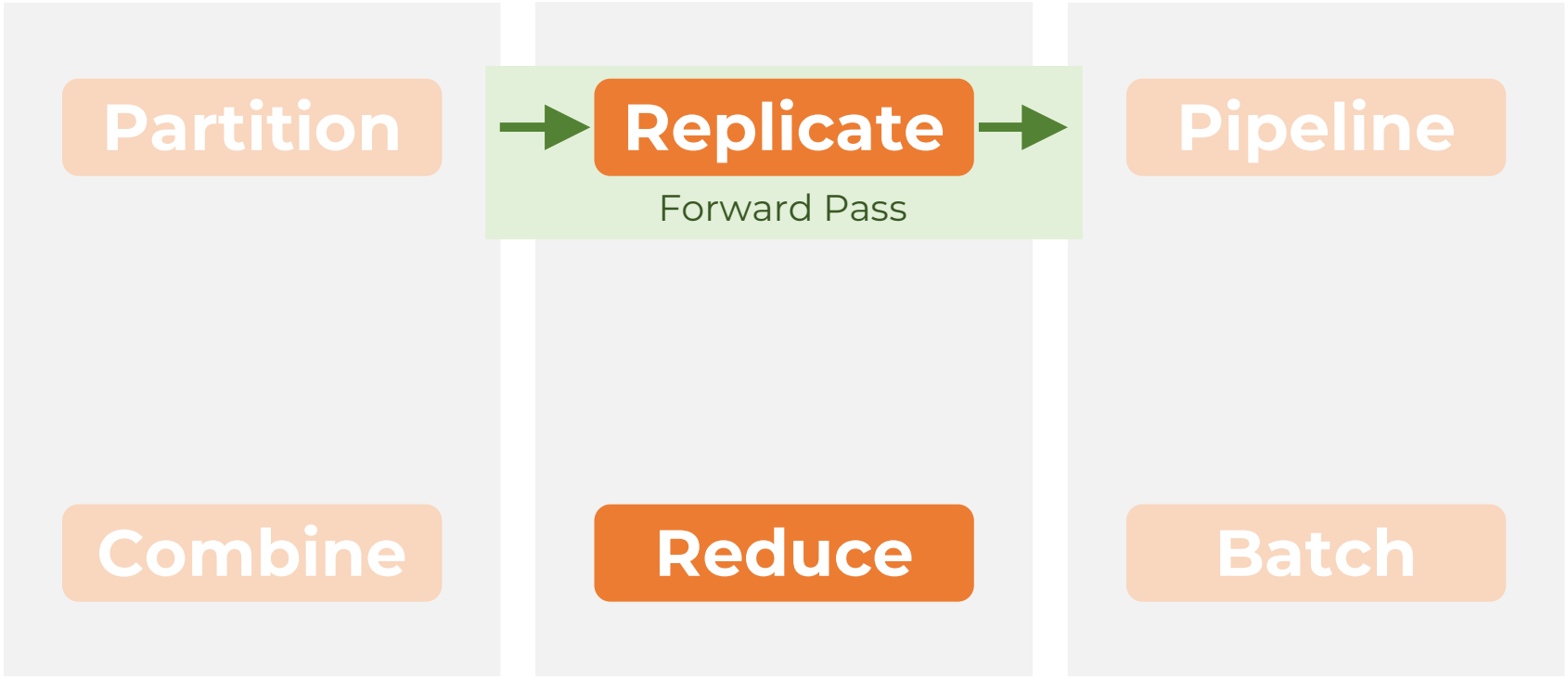
Pipeline

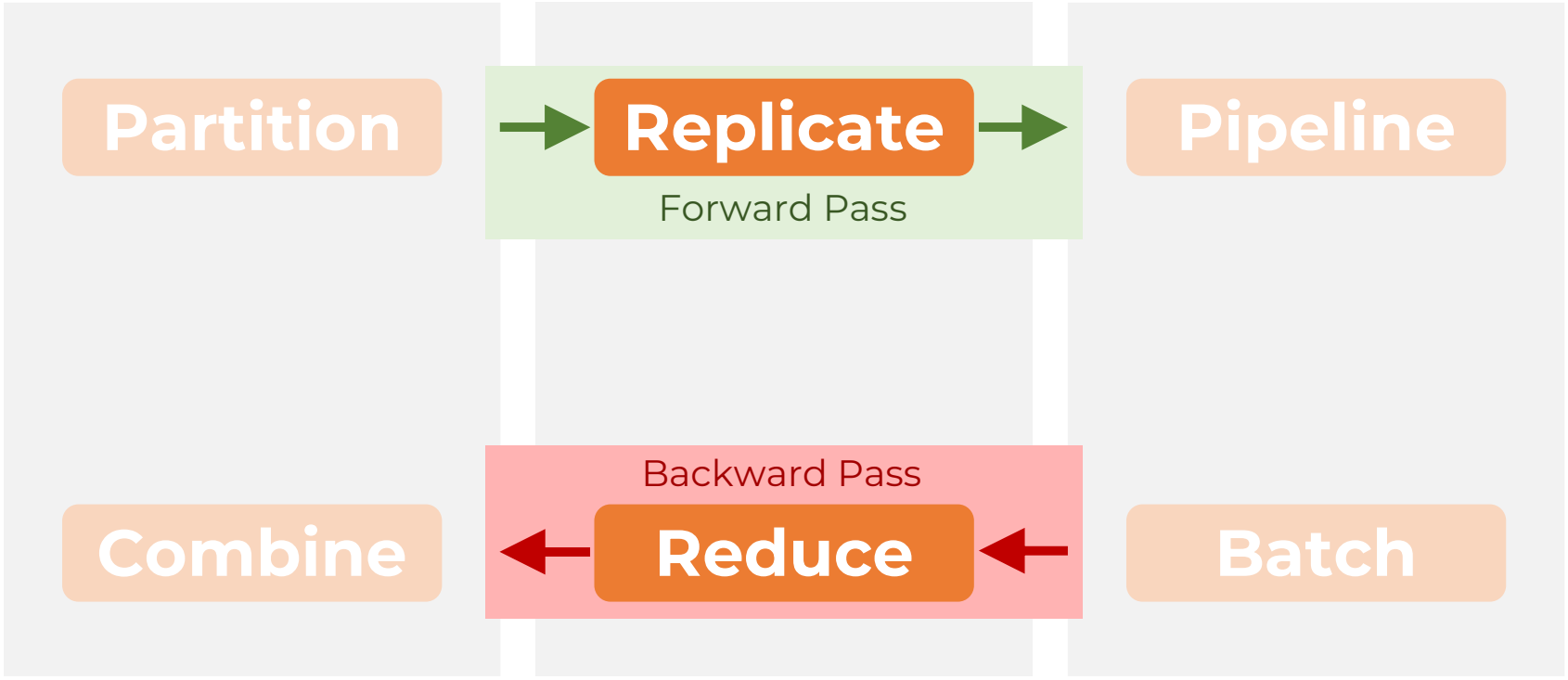
Combine

Reduce

Batch




Partition
Replicate
Pipeline
Forward Pass
Combine
Reduce
Batch


Partition
Replicate
Pipeline
Forward Pass
Backward Pass
Combine
Reduce
Batch

Partition

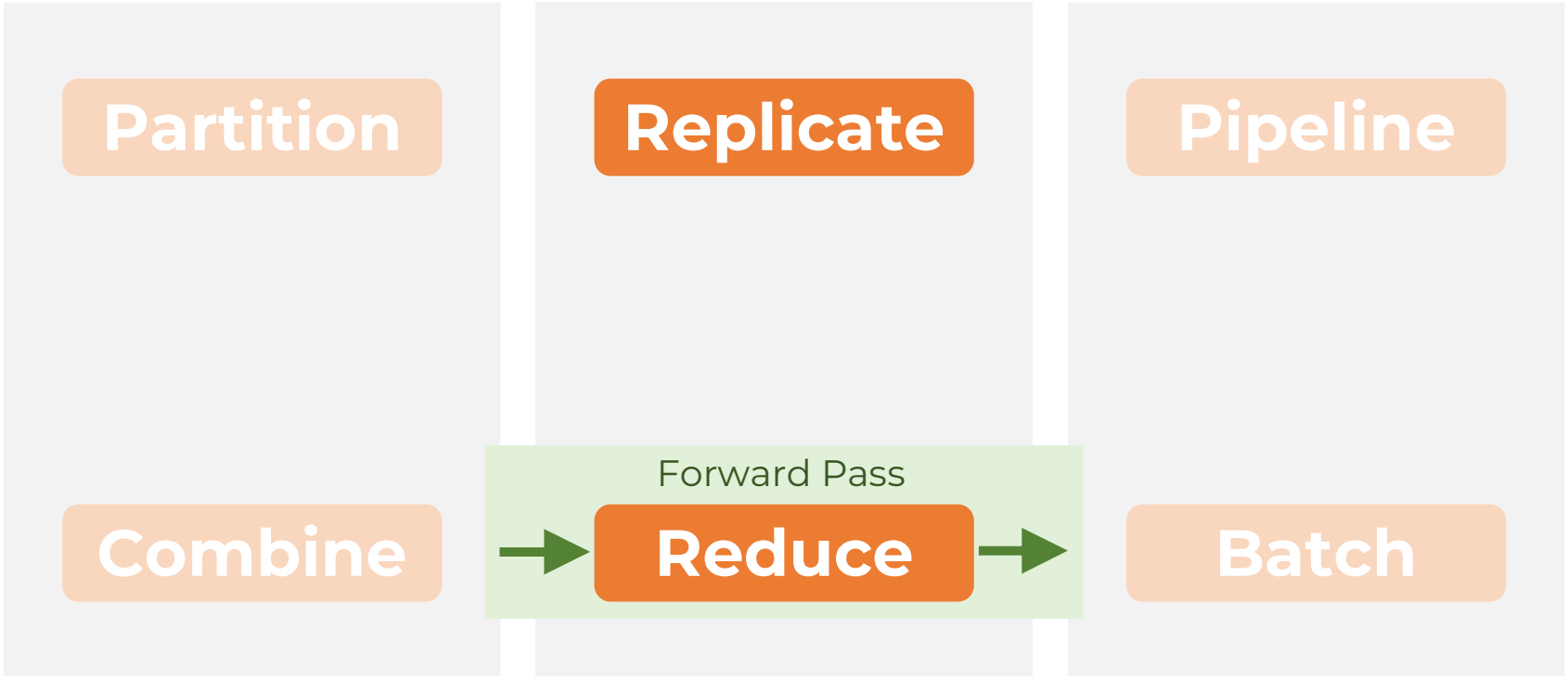
Replicate

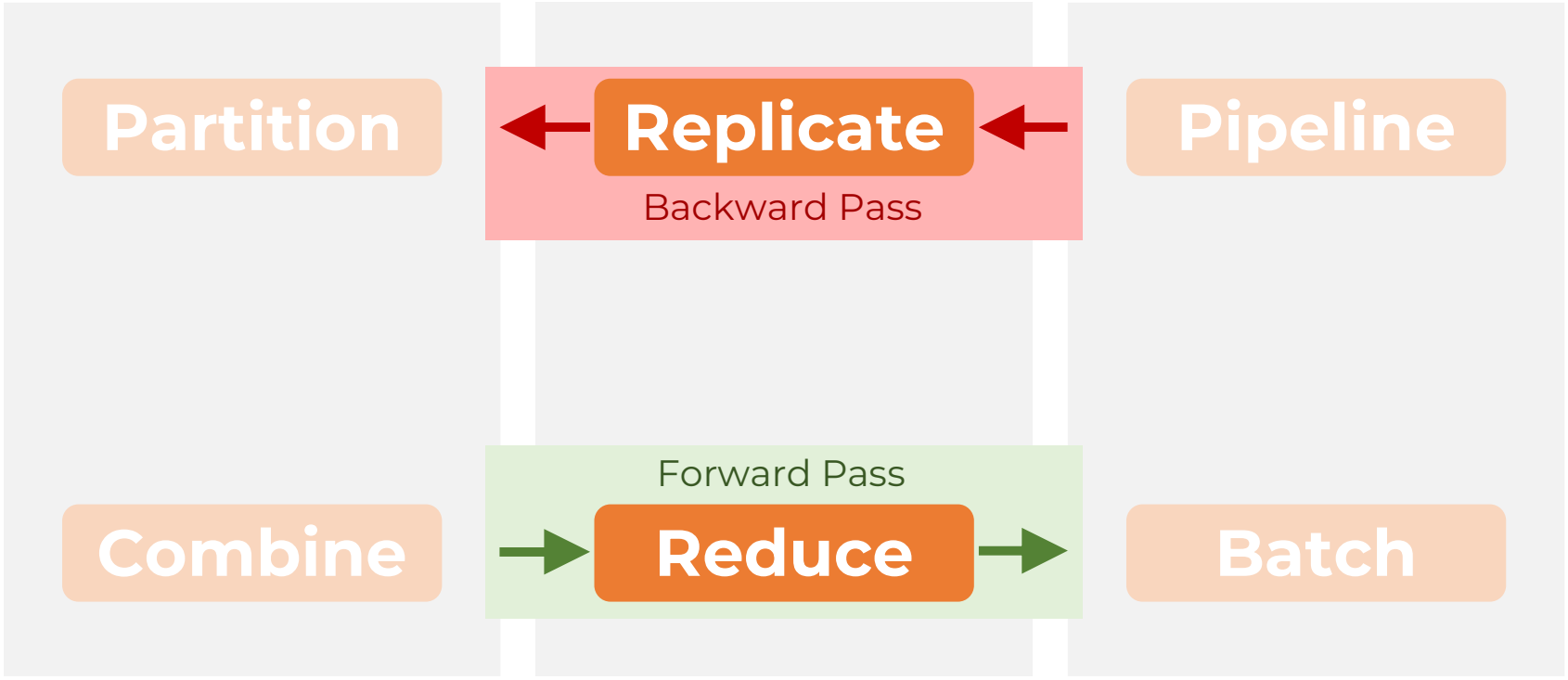
Pipeline

Combine

Reduce

Batch


Partition
Replicate
Pipeline
Forward Pass
Combine
Reduce
Batch


Partition
Replicate
Pipeline
Backward Pass
Forward Pass
Combine
Reduce
Batch

Partition

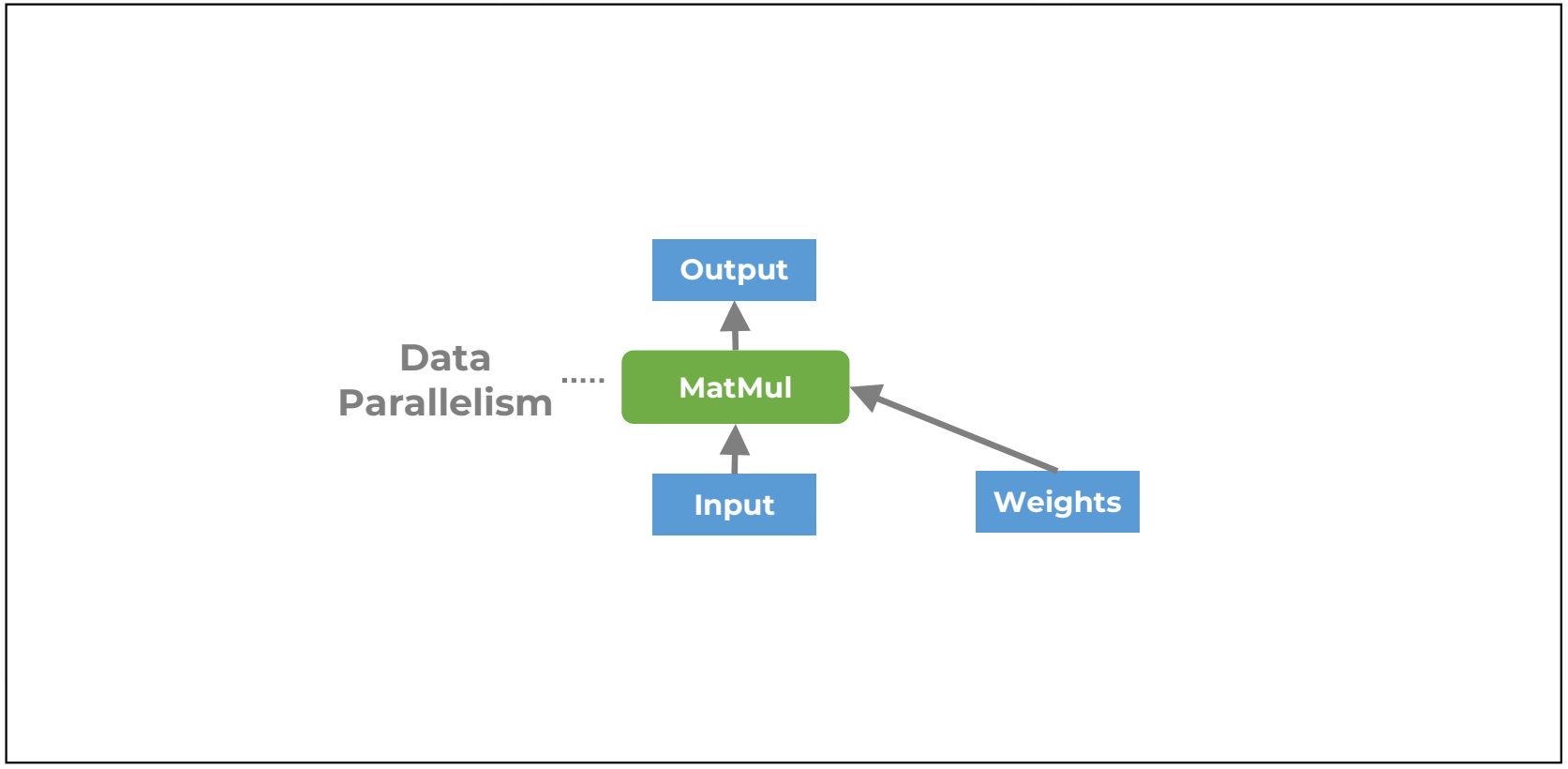
Replicate

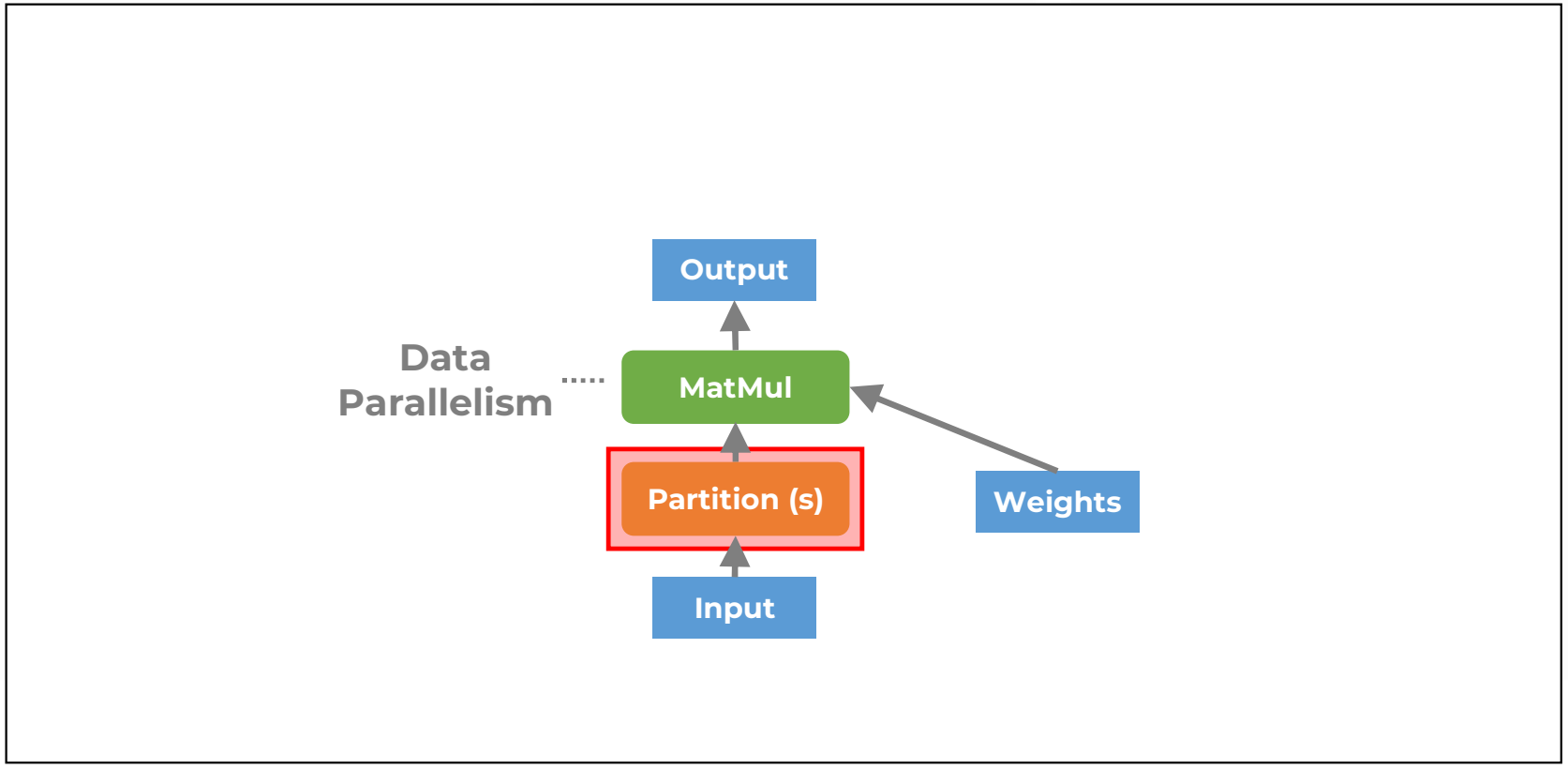
Pipeline

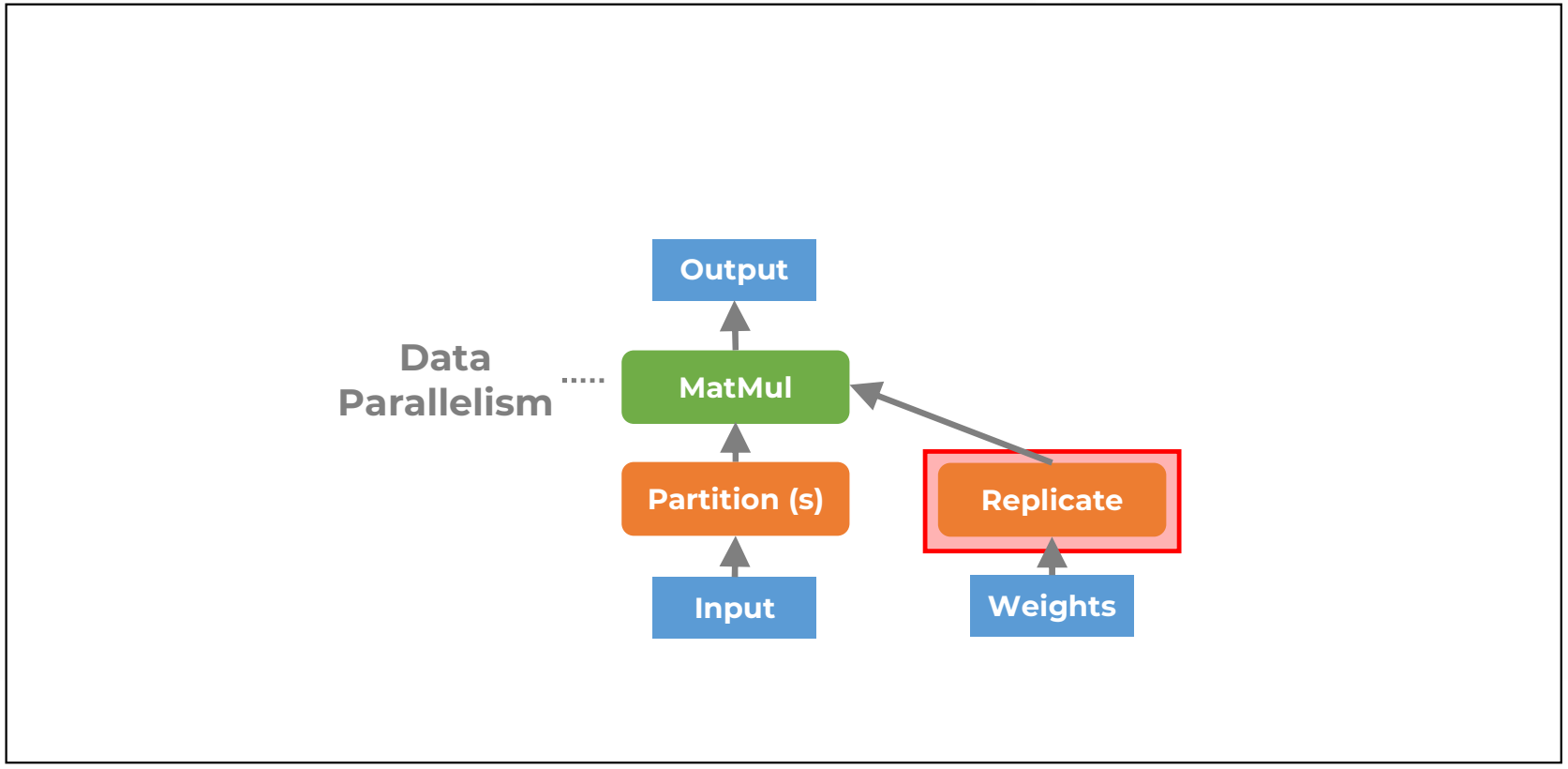
Combine

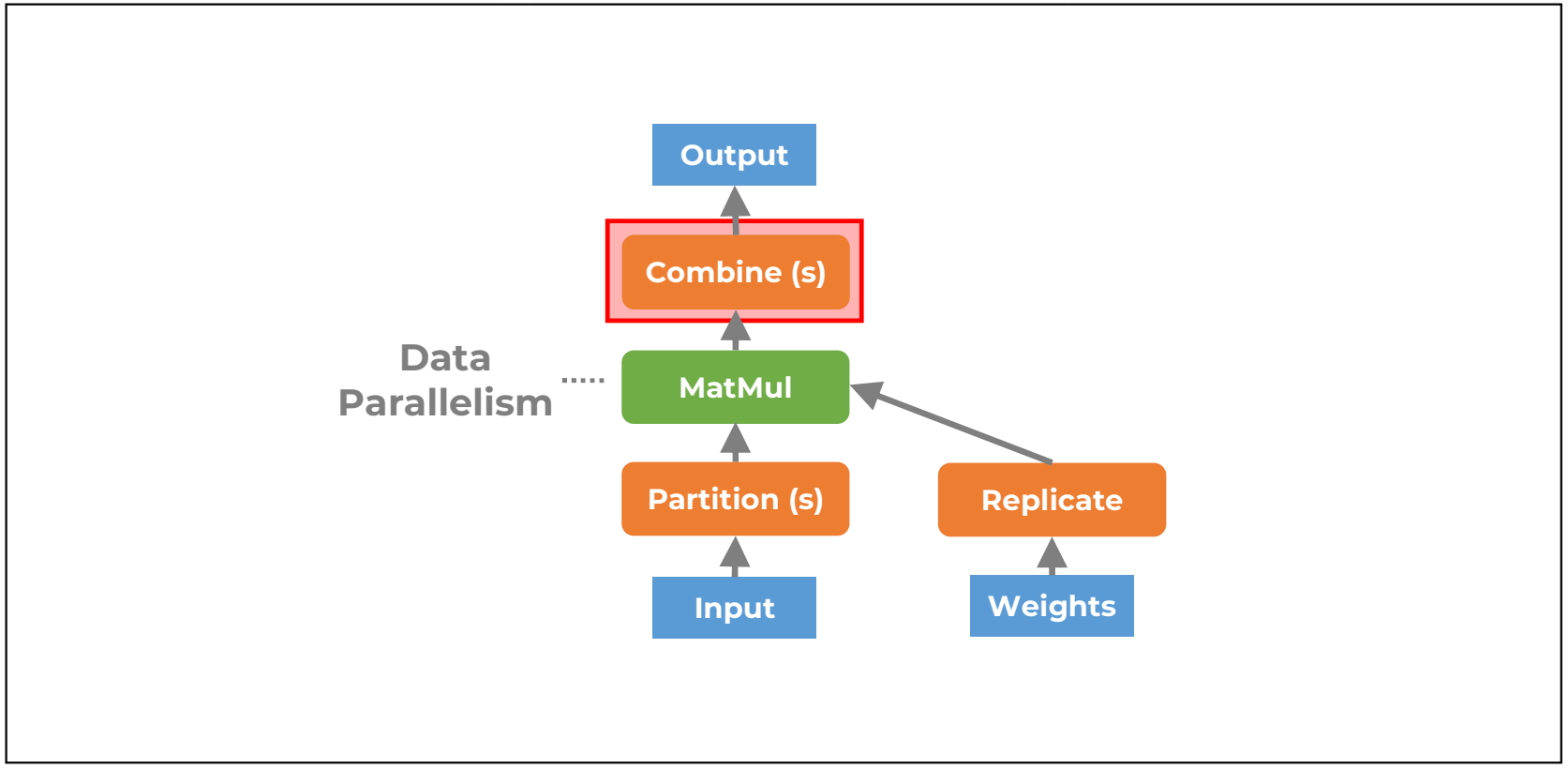
Reduce

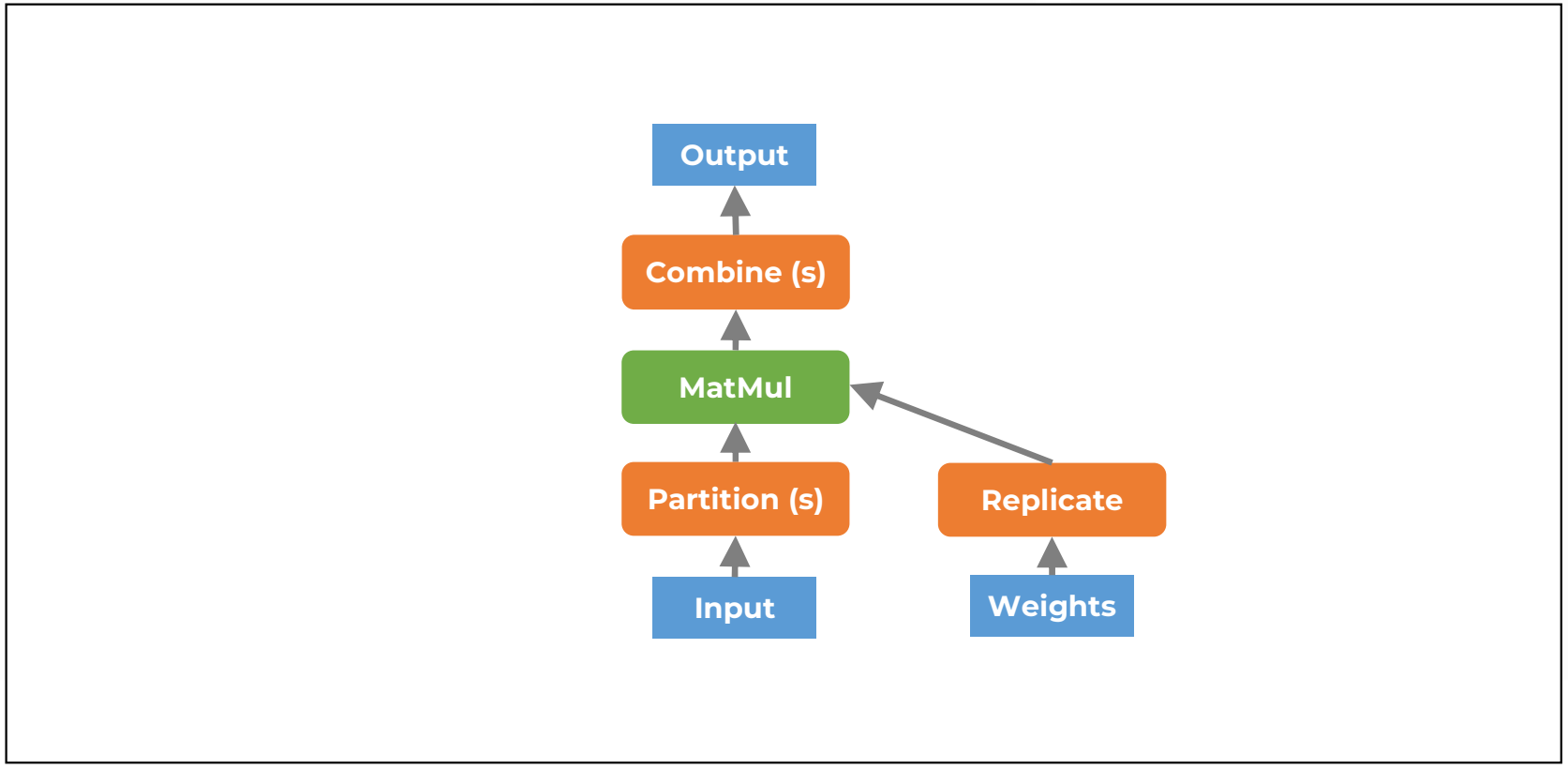
Batch

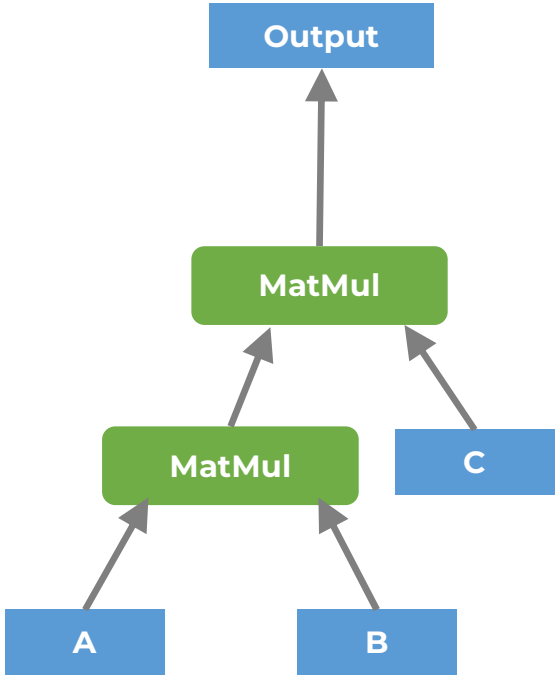

Data
Parallelism
Output
MatMul
Input
Weights


Data Parallelism
Output
MatMul
Partition (s)
Weights
Input

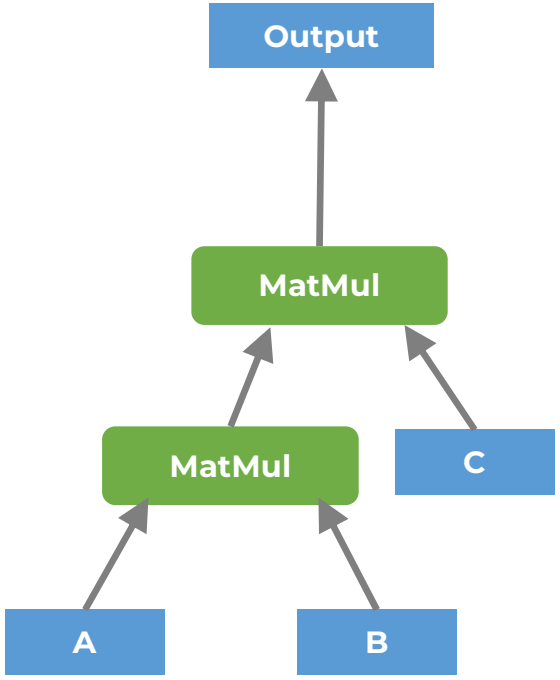

Data Parallelism
Output
MatMul
Partition (s)
Replicate
Input
Weights


Output
Combine (s)
Data Parallelism
MatMul
Partition (s)
Replicate
Input
Weights

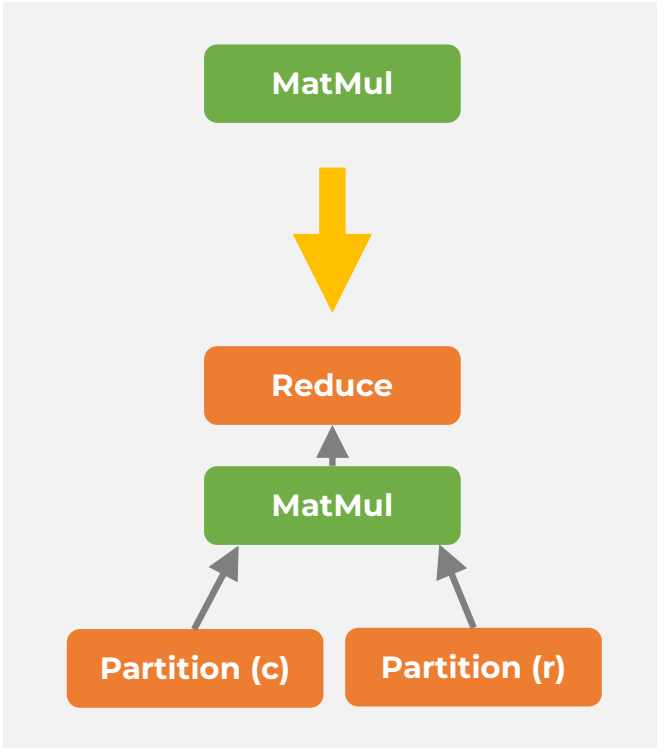

Output
Combine (s)
MatMul
Partition (s)
Replicate
Input
Weights


Output
MatMul
MatMul
C
A
B

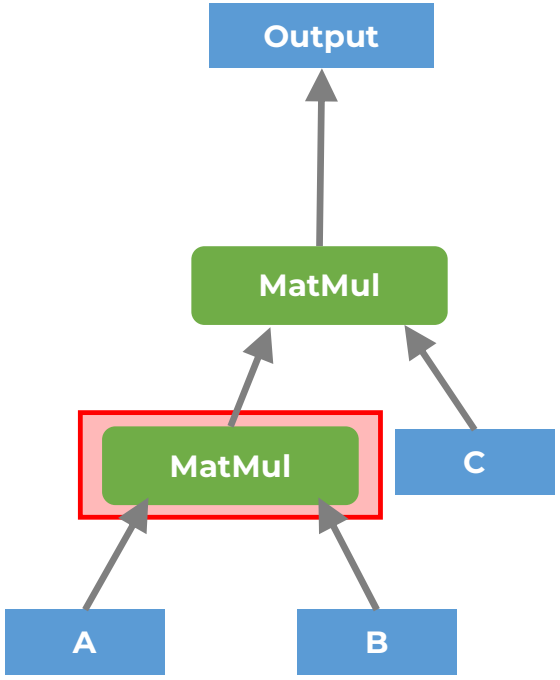
PCG



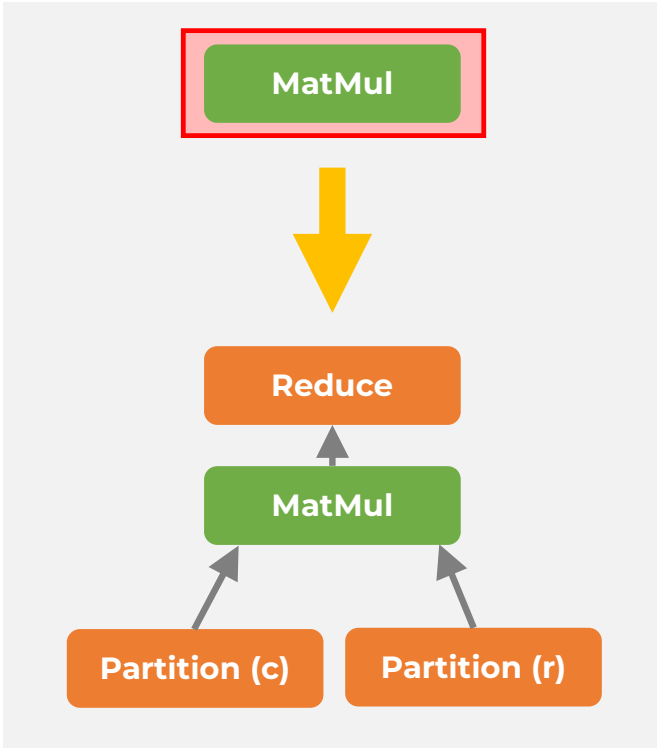
Substitution



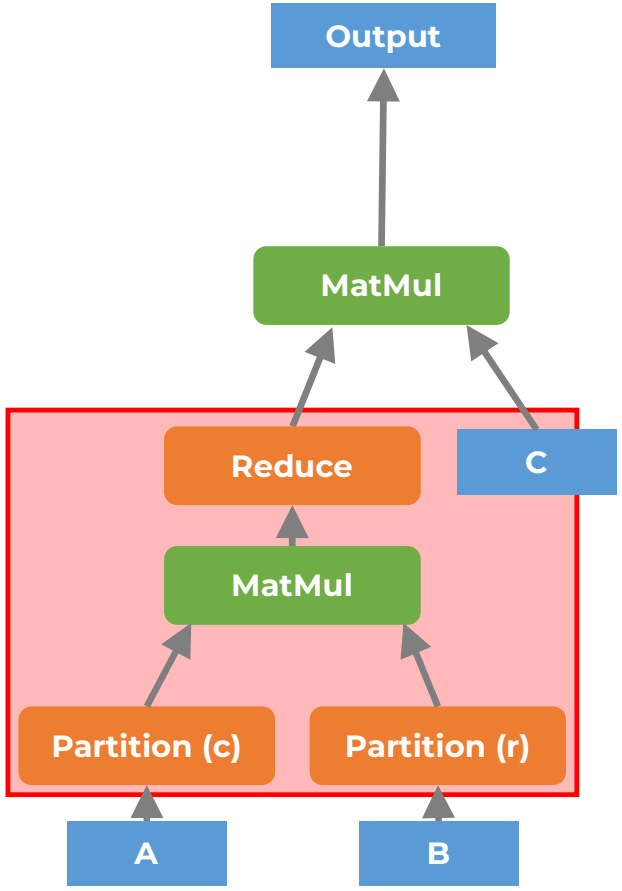
PCG



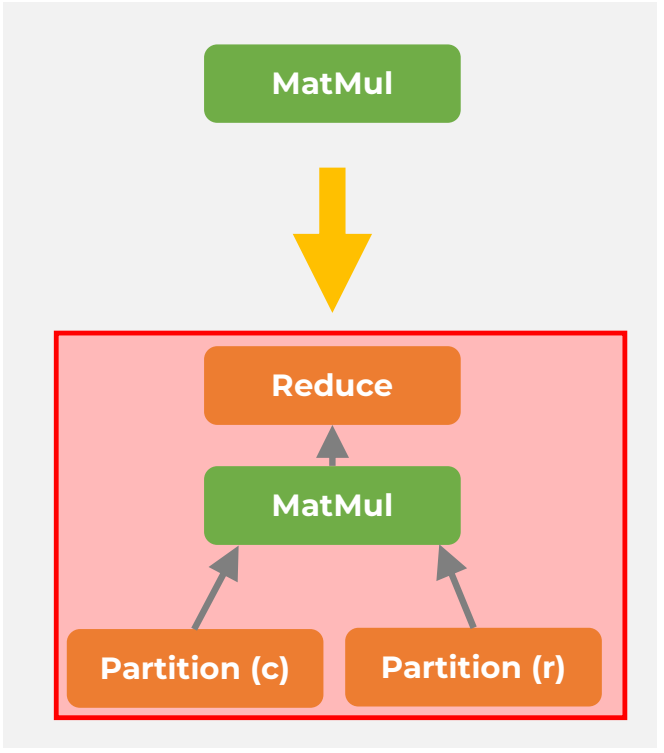
Substitution



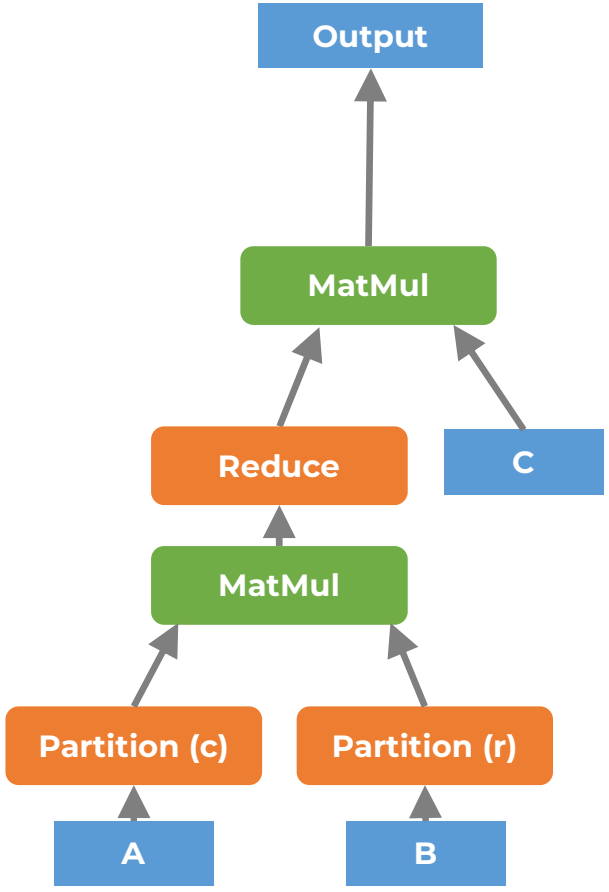
PCG



Substitution

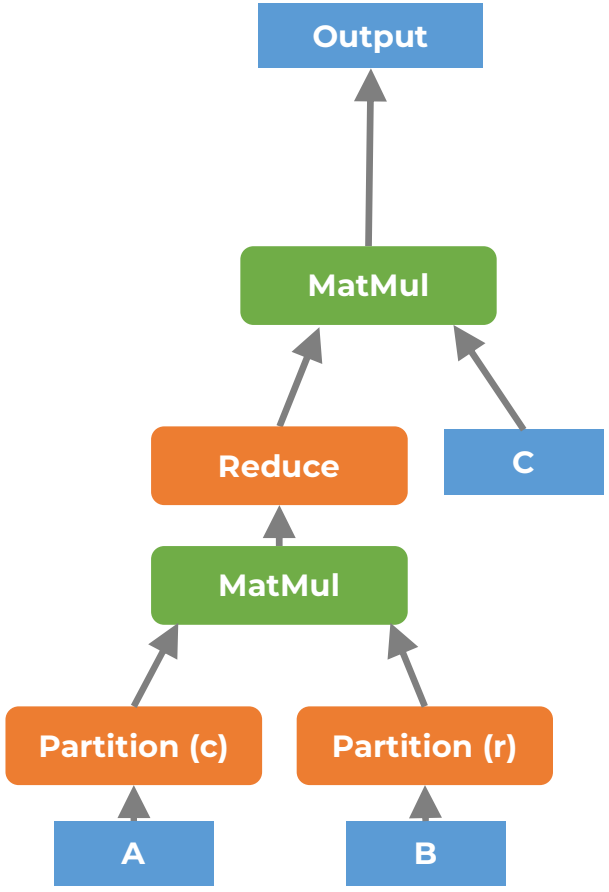


PCG

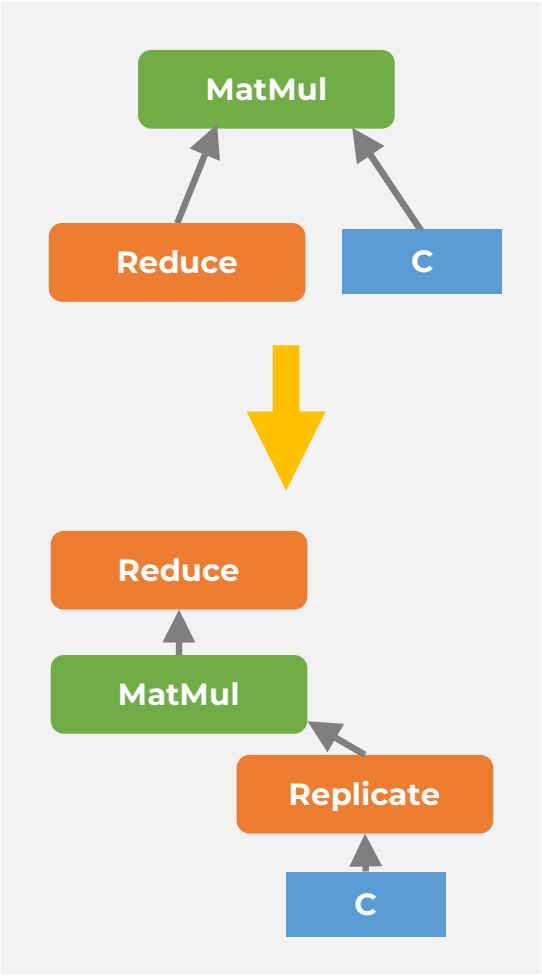


Substitution

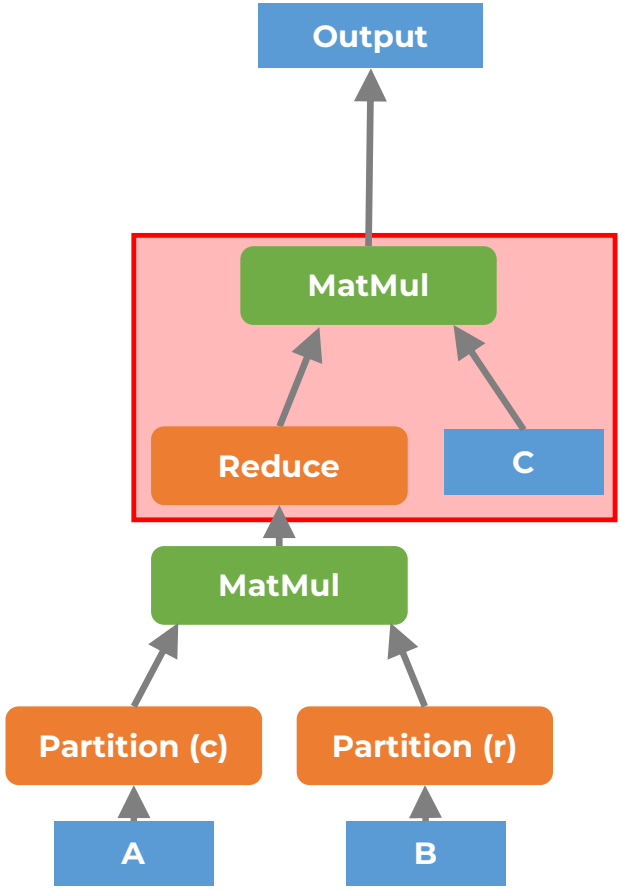
PCG



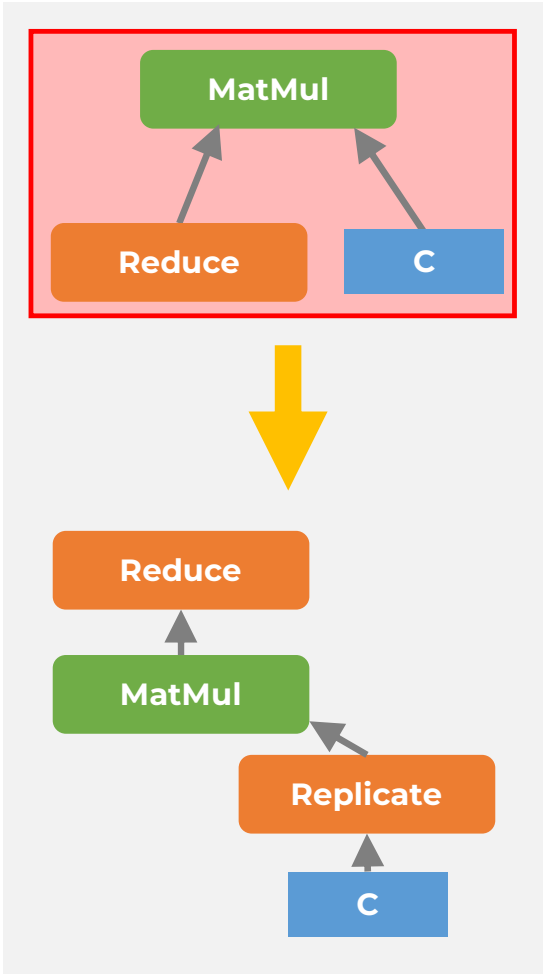
Substitution



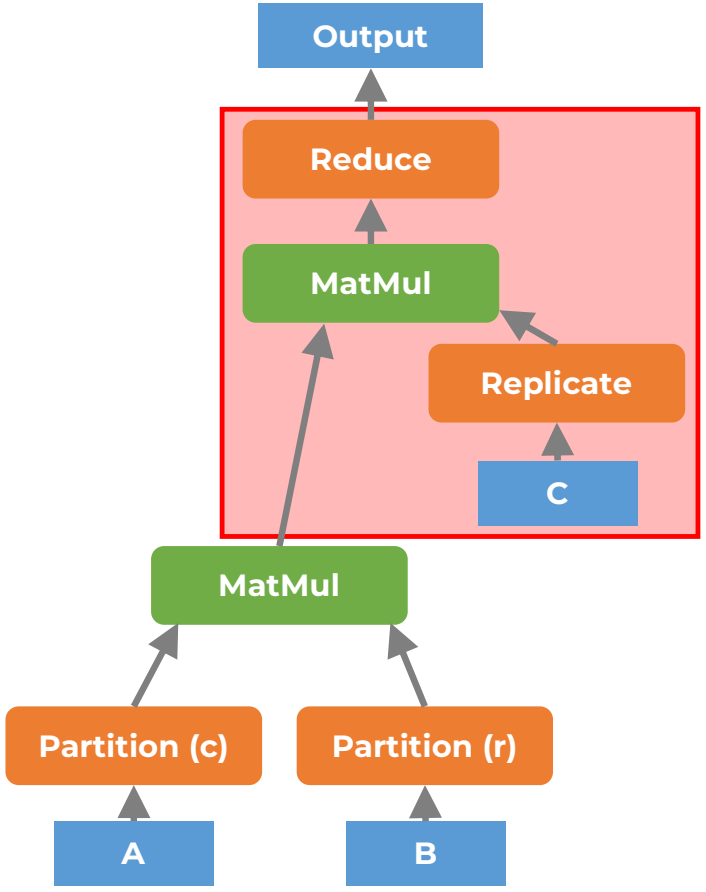
PCG



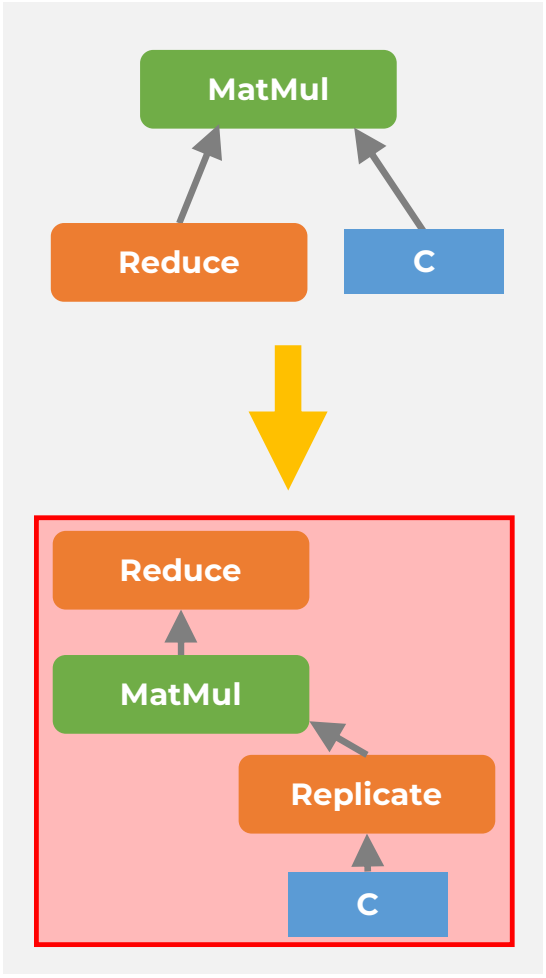
Substitution

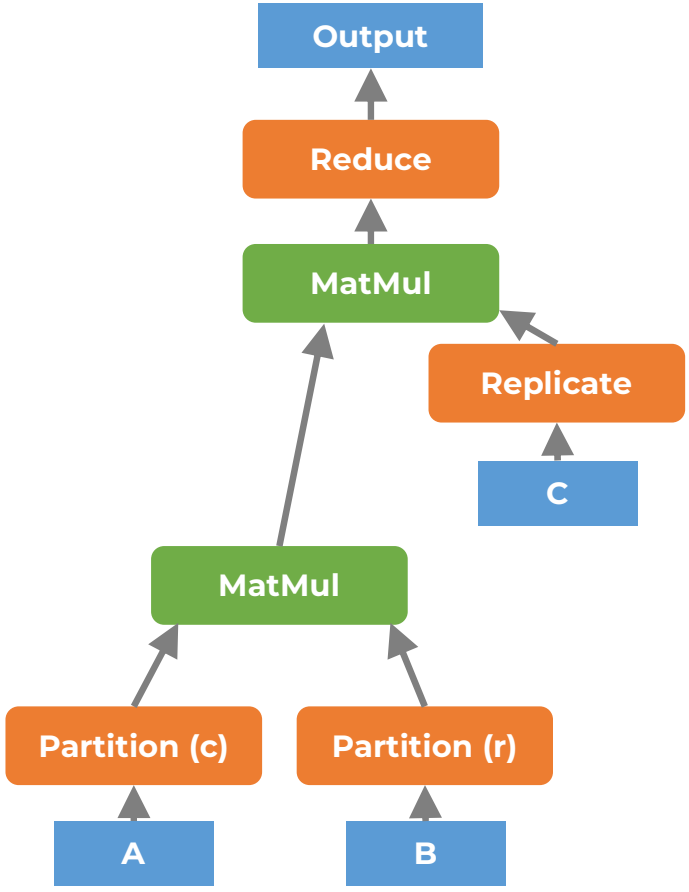


PCG



Substitution




Output
Reduce
MatMul
Replicate
C
MatMul
Partition (c)
Partition (r)
A
B

Separation of concerns

Separation of concerns

Automatically generate substitutions

Separation of concerns

Automatically generate substitutions

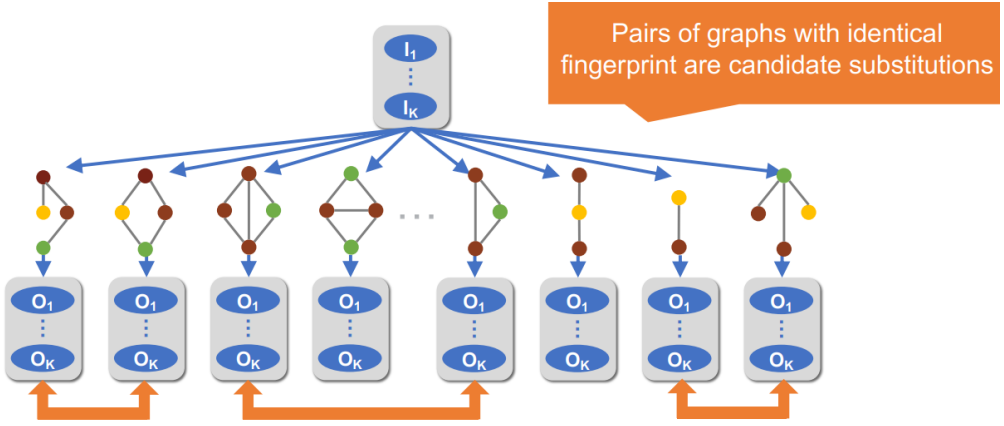
New operators

Separation of concerns

Automatically generate substitutions

New operators

New forms of parallelism



Separation of concerns

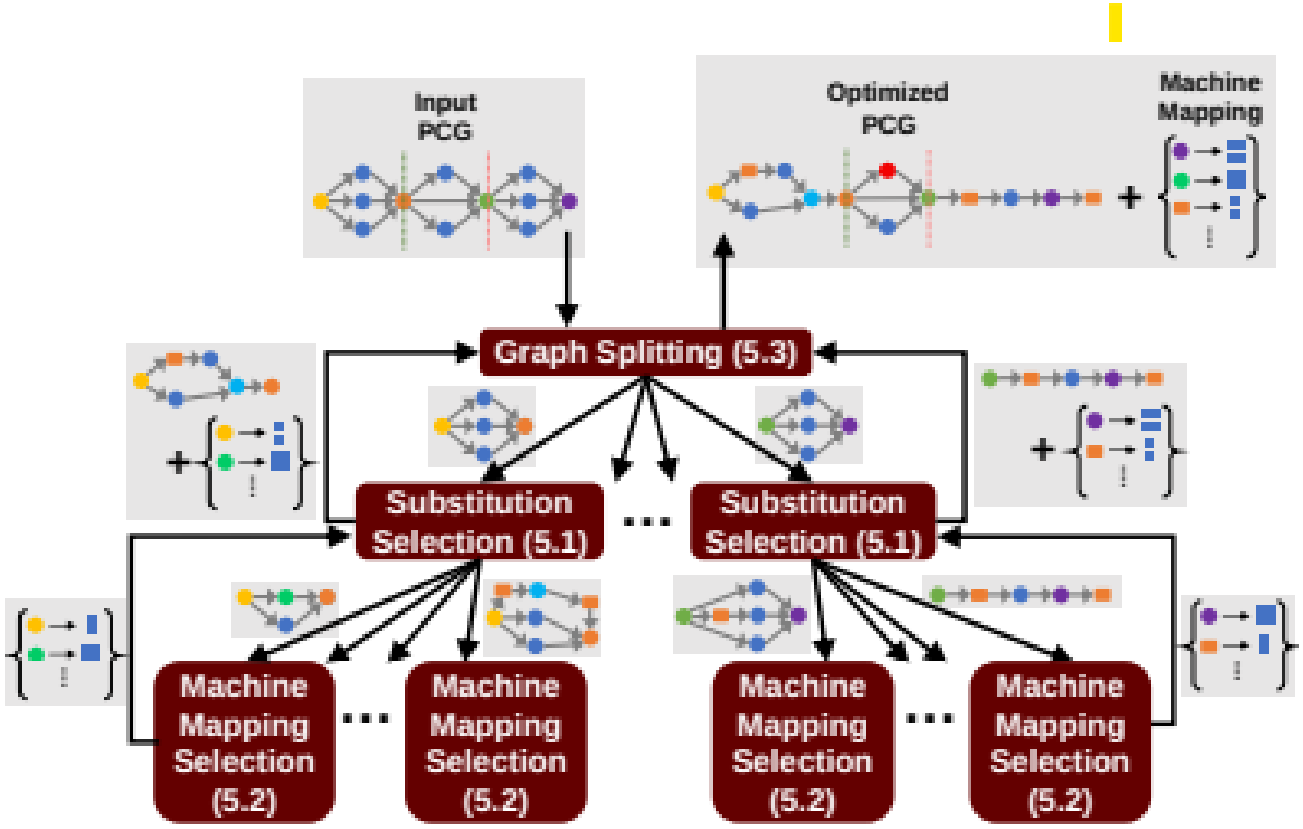
Explicitly represents communication

Separation of concerns

Explicitly represents communication

Concise

Hierarchical Search Algorithm



Algebraic Transformation

Parallelism Type

Parallelism Degree

Device Mapping

Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Device Mapping

Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Backtracking Search

Device Mapping

Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Backtracking Search

Device Mapping

Dynamic Programming

Evaluation

Models

BERT-Large
(Language Modeling)

Candle-UNO
(Precision Medicine)

MLP
(Regression)

DLRM
XDL
(Recommendation)

ResNeXt-50
Inception-v3
(Computer Vision)

Models

BERT-Large
(Language Modeling)

Candle-UNO
(Precision Medicine)

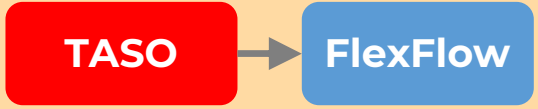
MLP
(Regression)

DLRM
XDL
(Recommendation)

ResNeXt-50
Inception-v3
(Computer Vision)

×

Baselines



TASO+FlexFlow
(Sequential Optimization)

Models

BERT-Large
(Language Modeling)

Candle-UNO
(Precision Medicine)

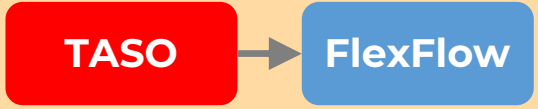
MLP
(Regression)

DLRM
XDL
(Recommendation)

ResNeXt-50
Inception-v3
(Computer Vision)

×

Baselines



TASO+FlexFlow
(Sequential Optimization)

Expert-Designed
or
Data Parallel

Models

BERT-Large
(Language Modeling)

Candle-UNO
(Precision Medicine)

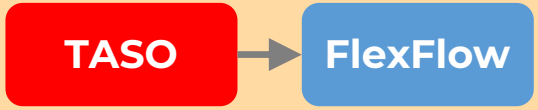
MLP
(Regression)

DLRM
XDL
(Recommendation)

ResNeXt-50
Inception-v3
(Computer Vision)

×

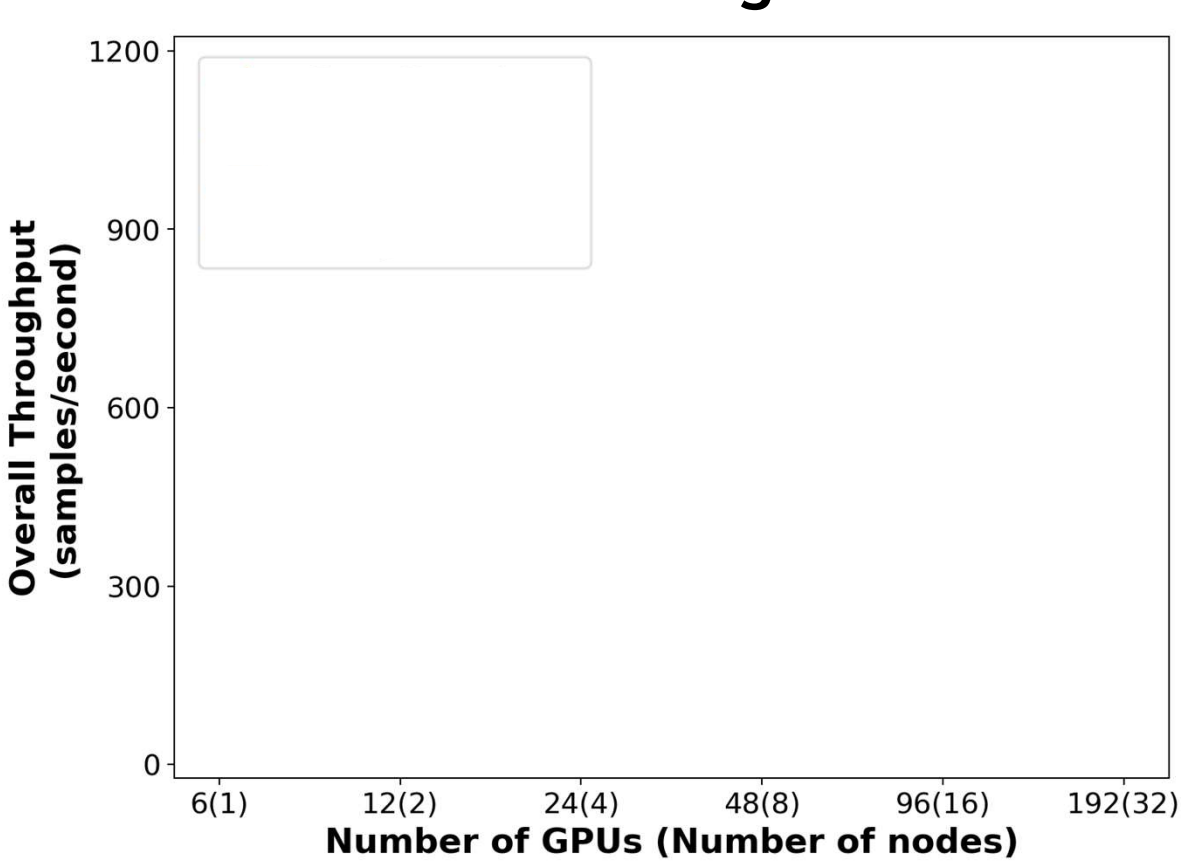
Baselines



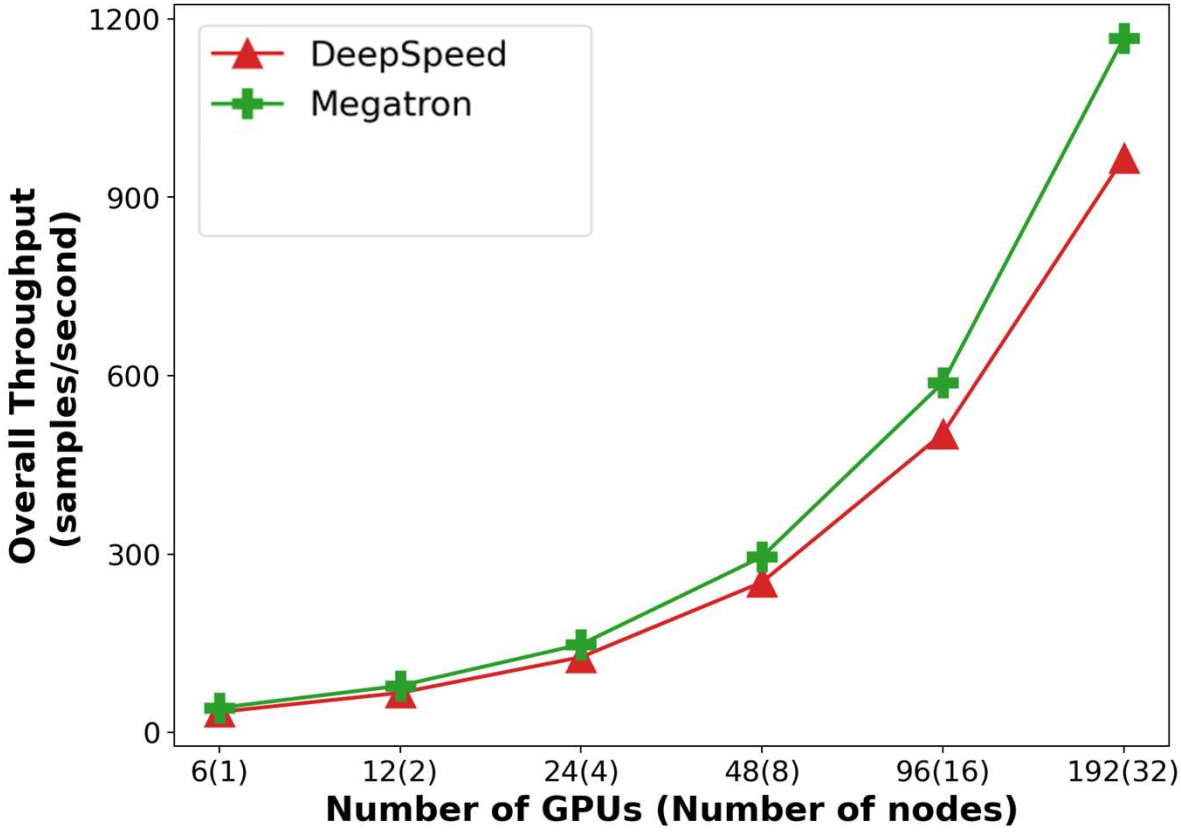
TASO+FlexFlow
(Sequential Optimization)

Expert-Designed
or
Data Parallel

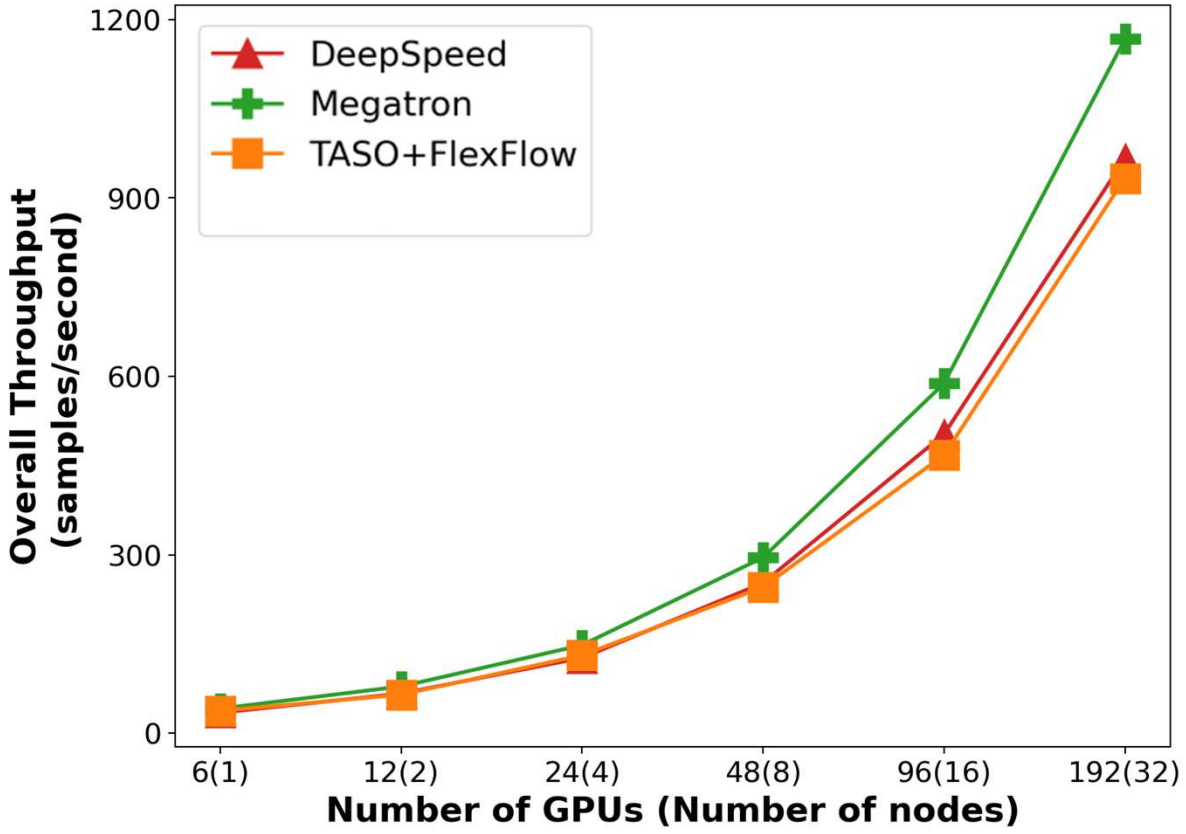
- Megatron [arXiv 19]
- DeepSpeed [arXiv 19]


BERT-Large
Overall Throughput (samples/second)
1200
900
600
300
0
6(1)
12(2)
24(4)
48(8)
96(16)
192(32)
Number of GPUs (Number of nodes)

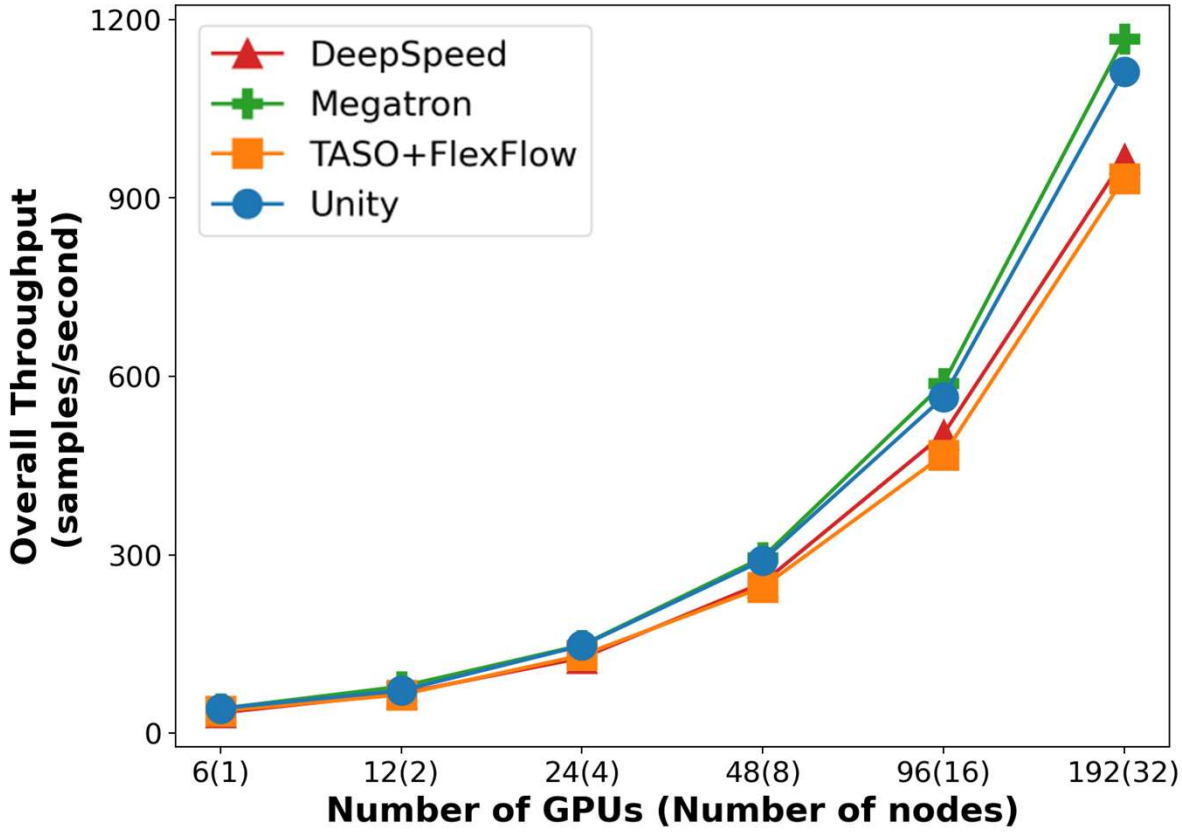
BERT-Large


DeepSpeed
Megatron
Overall Throughput (samples/second)
0
300
600
900
1200
6(1)
12(2)
24(4)
48(8)
96(16)
192(32)
Number of GPUs (Number of nodes)

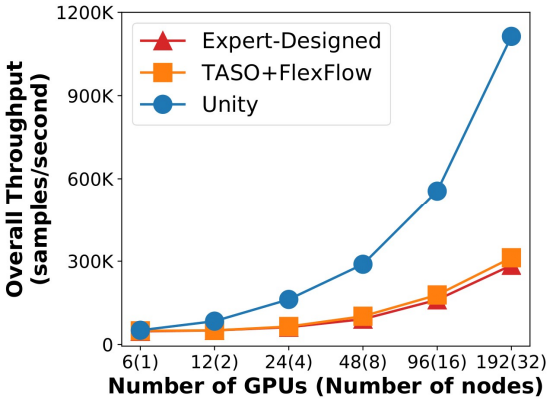
BERT-Large


DeepSpeed
Megatron
TASO+FlexFlow
Overall Throughput (samples/second)
0
300
600
900
1200
6(1)
12(2)
24(4)
48(8)
96(16)
192(32)
Number of GPUs (Number of nodes)

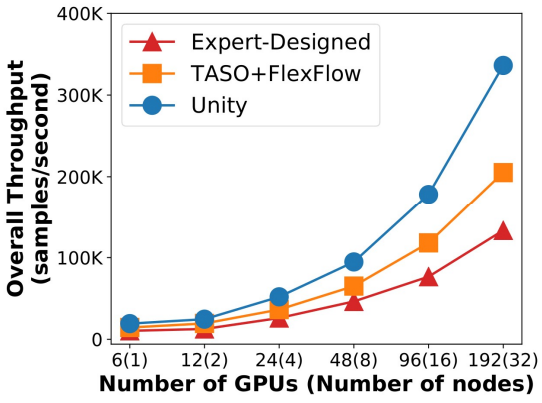
BERT-Large


DeepSpeed
Megatron
TASO+FlexFlow
Unity
Overall Throughput (samples/second)
0
300
600
900
1200
6(1)
12(2)
24(4)
48(8)
96(16)
192(32)
Number of GPUs (Number of nodes)

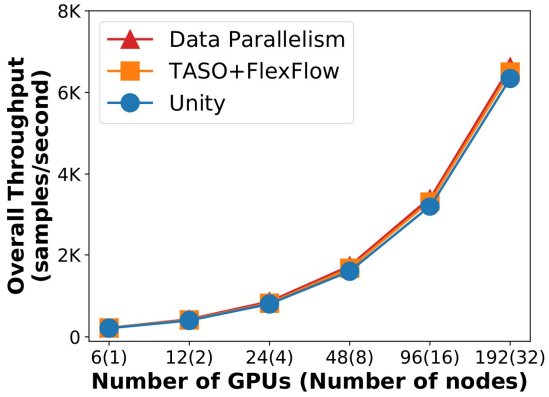
DLRM



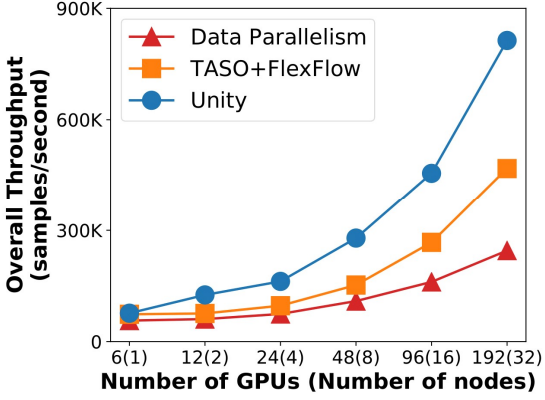
CANDLE-Uno



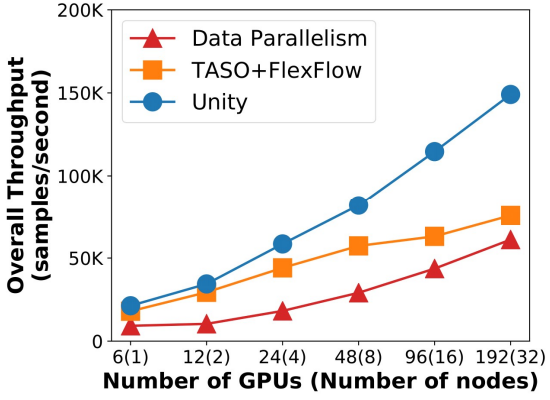
ResNeXt-50



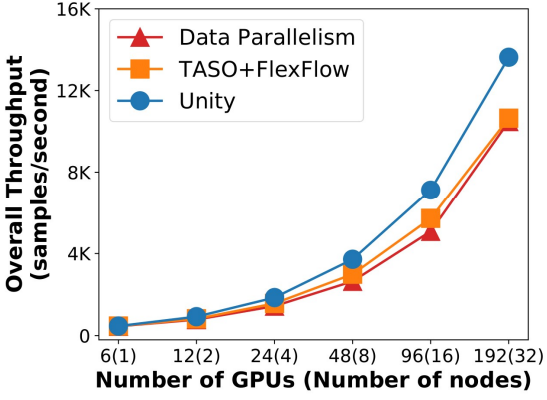
XDL

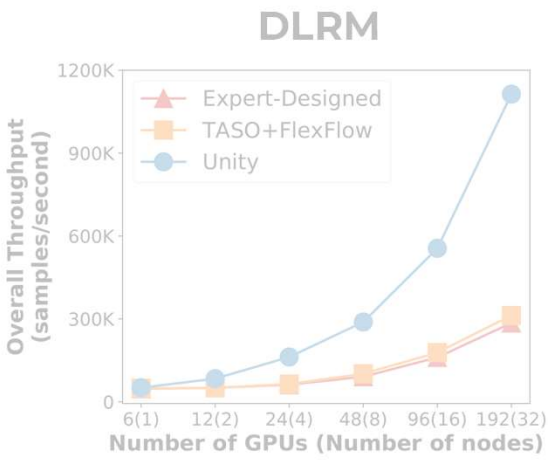


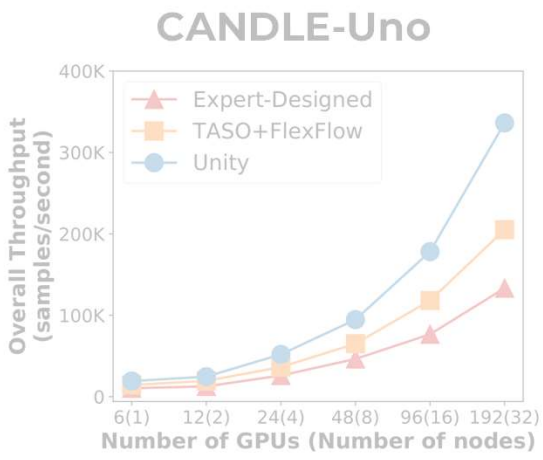
MLP

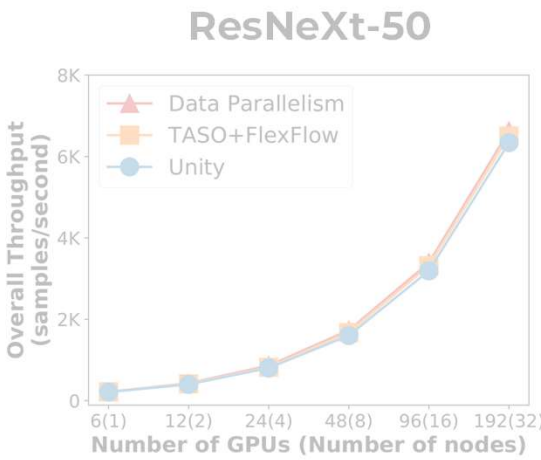


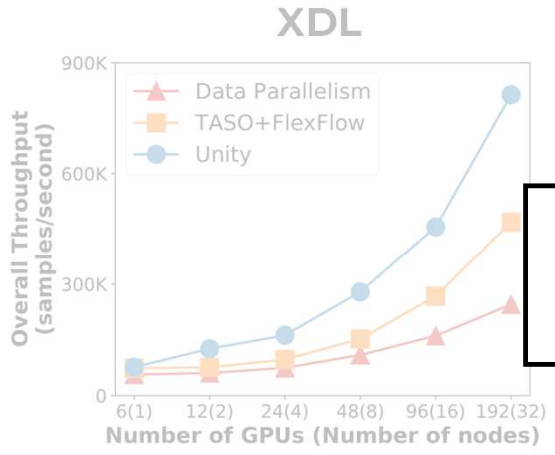
Inception-v3

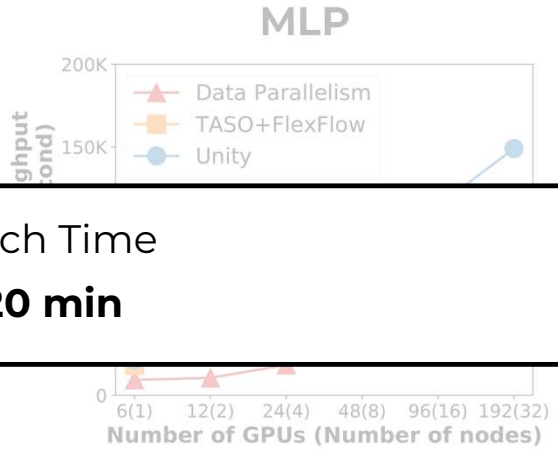


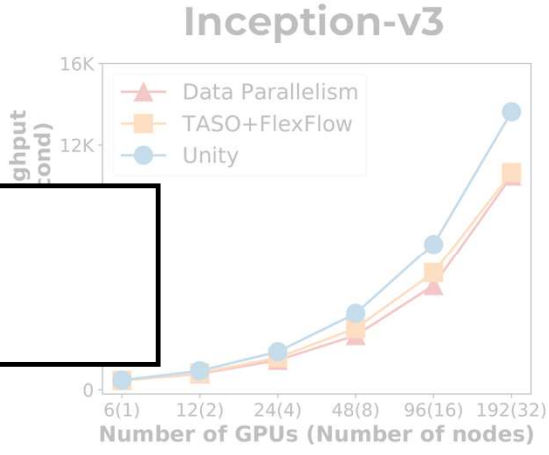






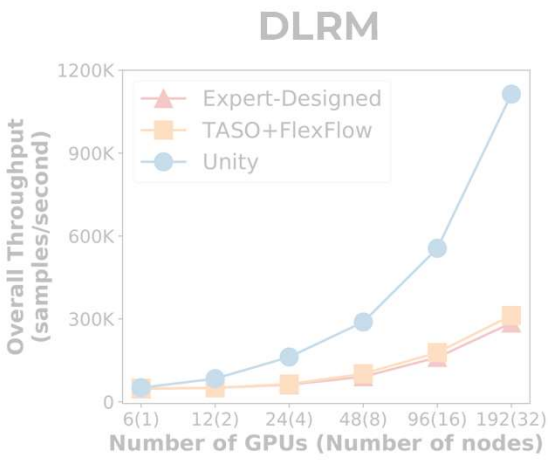


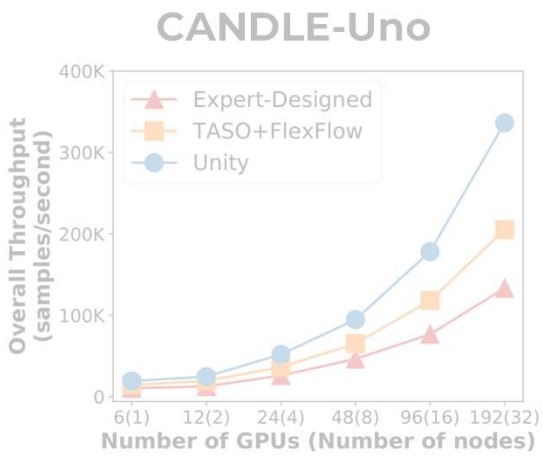


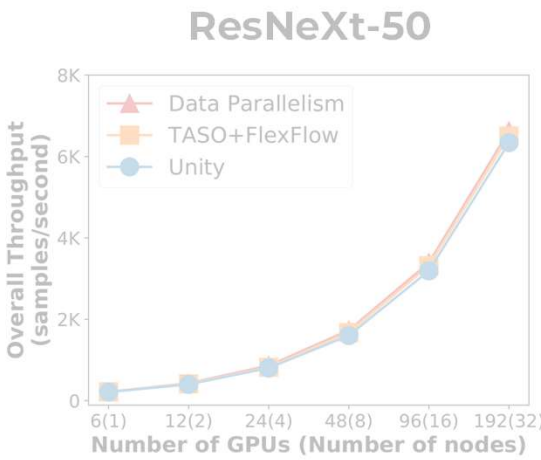


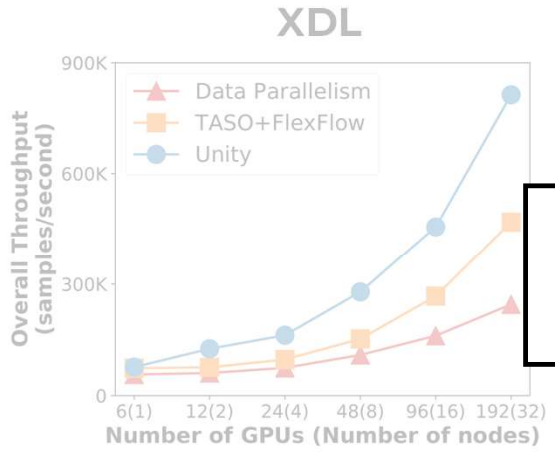
Search Time

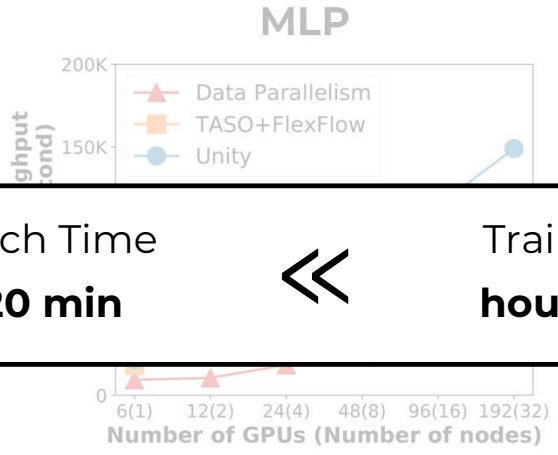
< 20 min

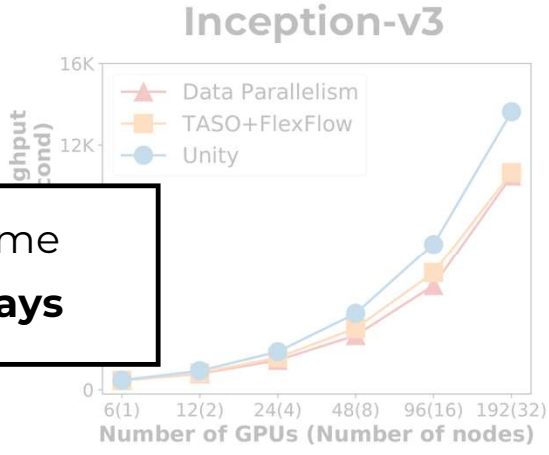








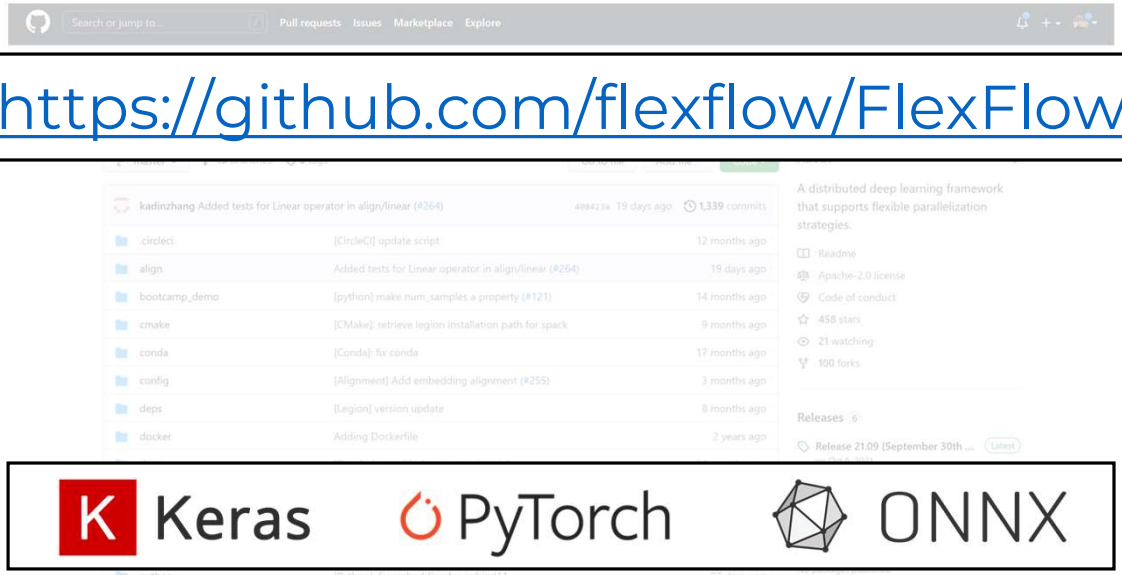




Search Time
< 20 min

≪

Training Time
hours or days


https://github.com/flexflow/FlexFlow
Keras
PyTorch
ONNX