

FlexFlow

TASO/MetaFlow

FlexFlow

Distributed DNN Training on **Legion** 

Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

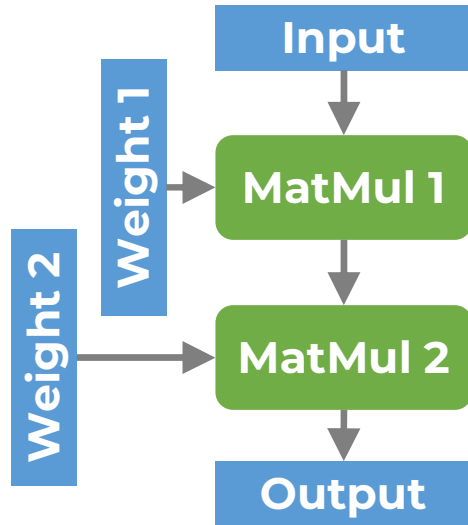
Hugging
Face

FlexFlow

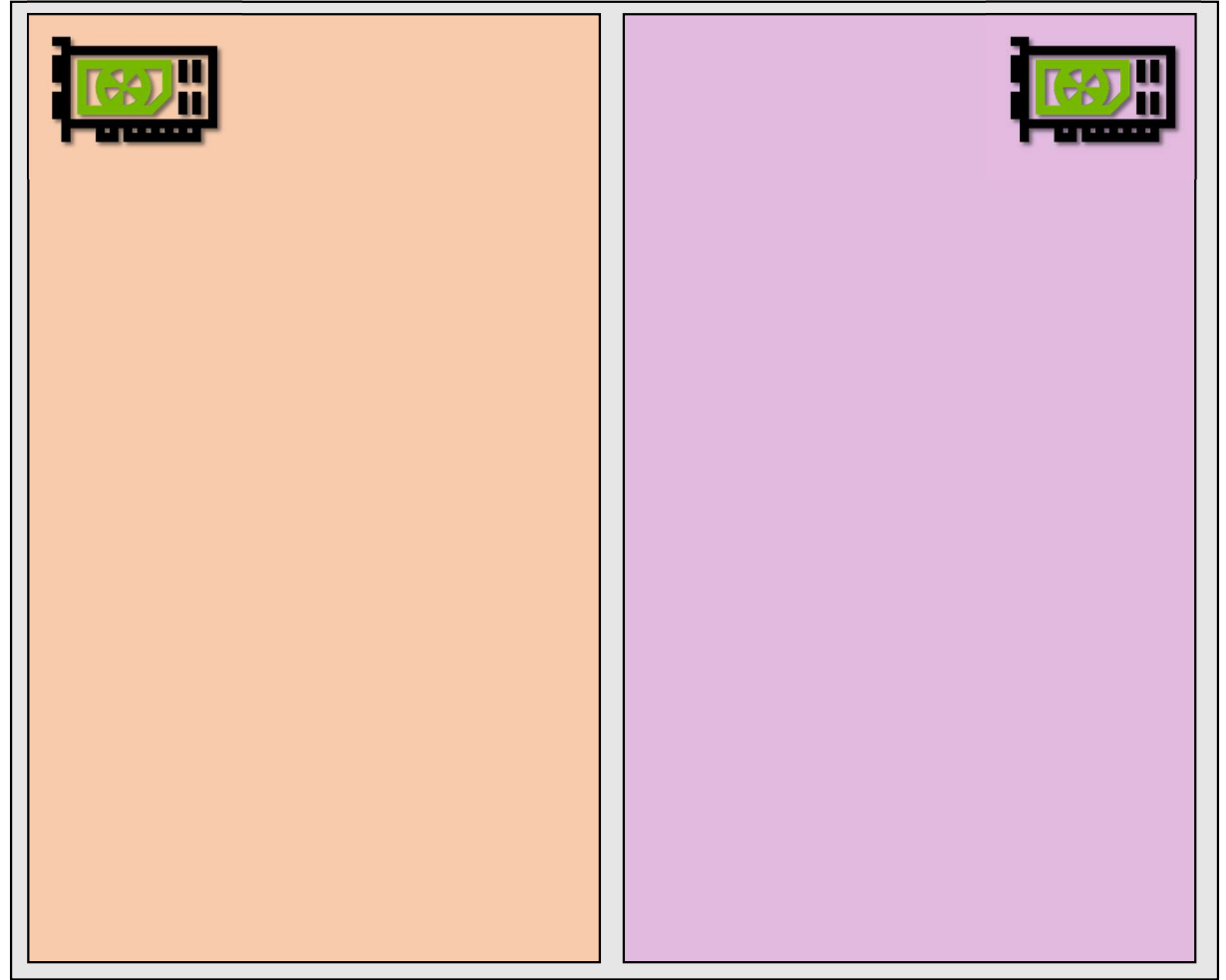
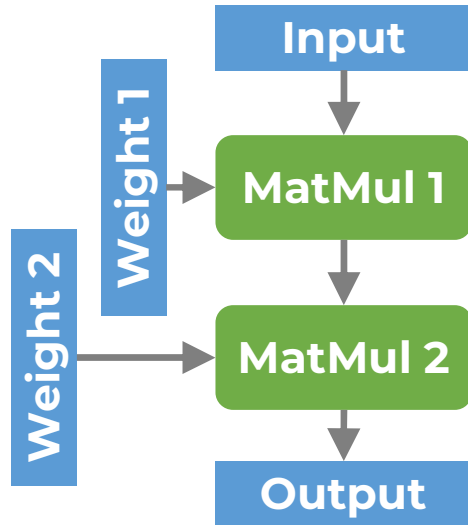
Distributed DNN Training

Data Parallel

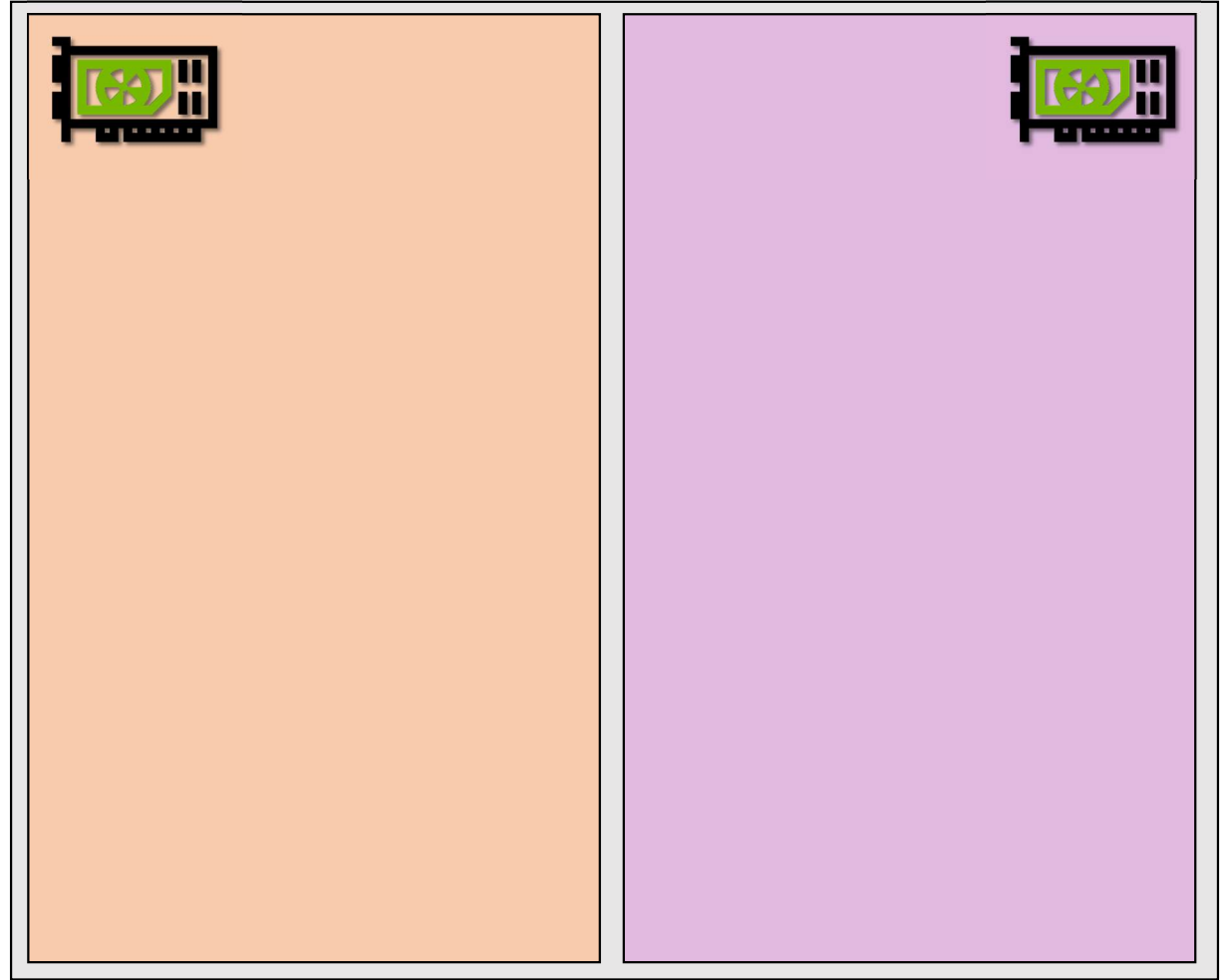
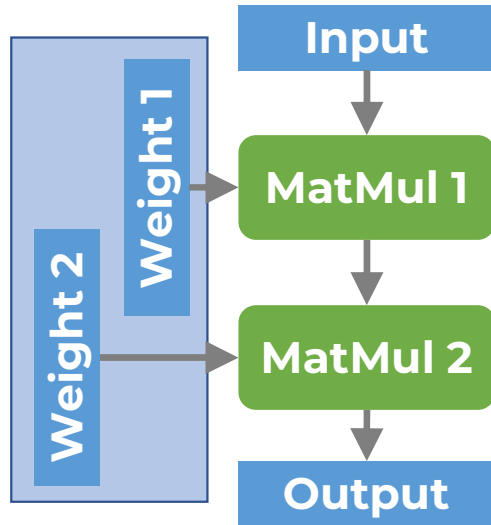
Data Parallel



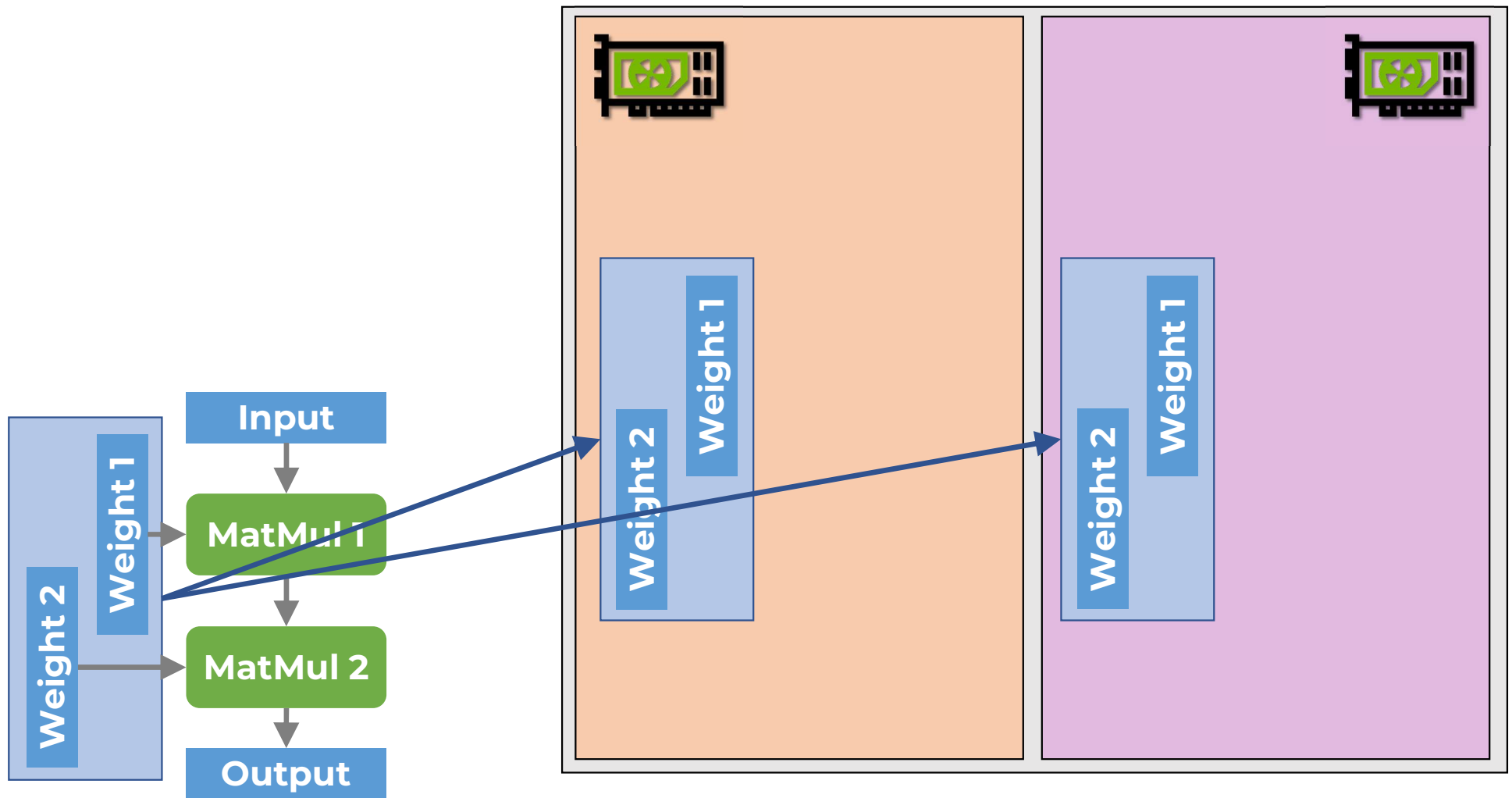
Data Parallel



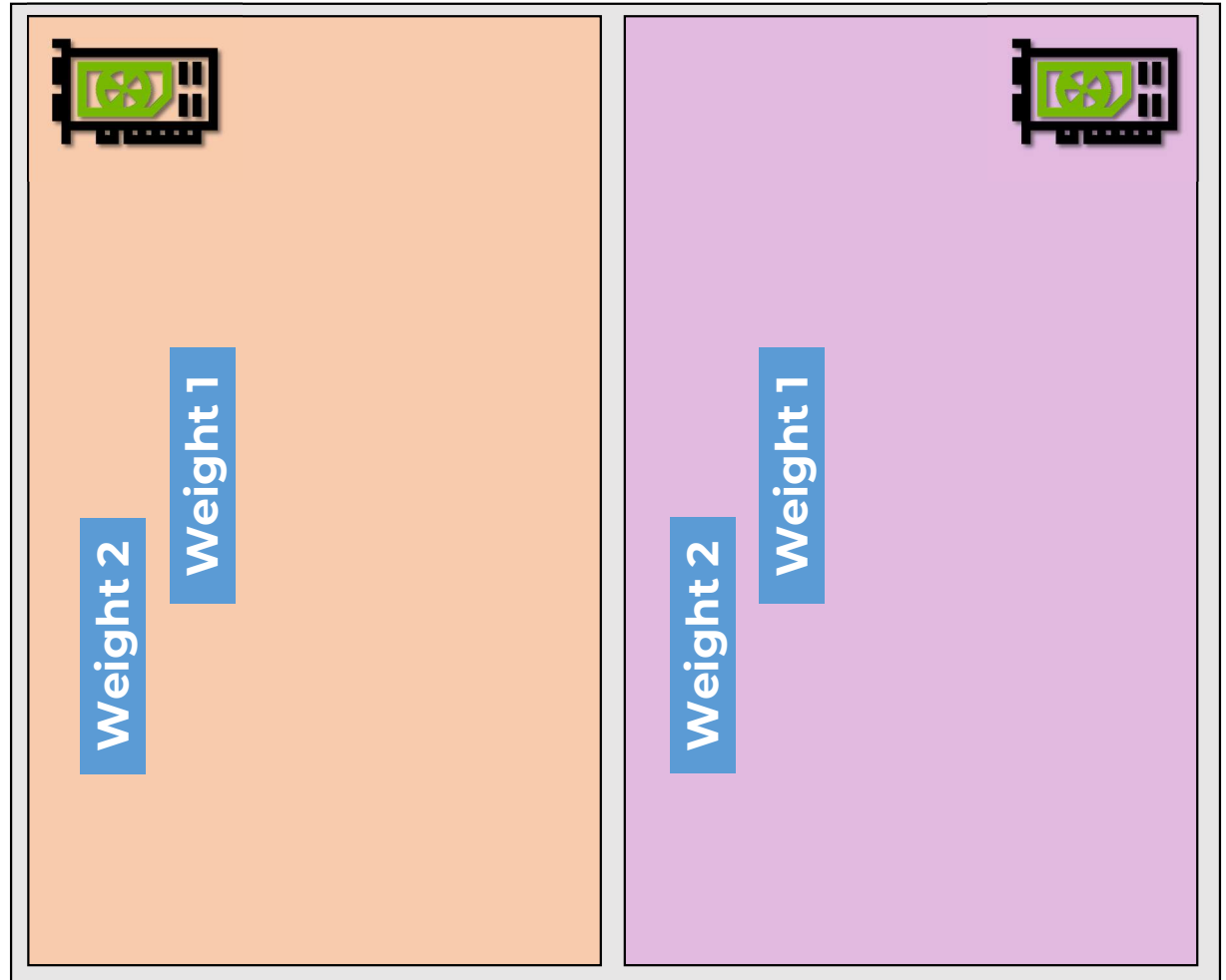
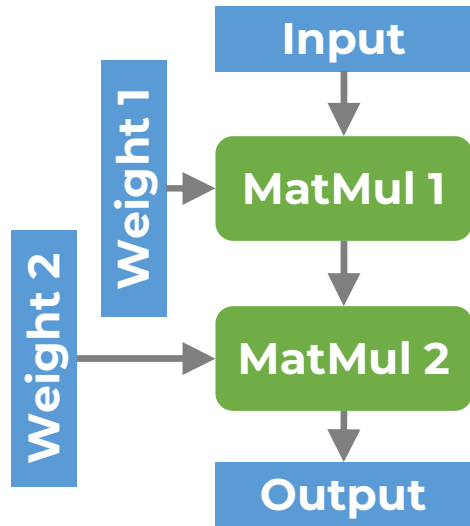
Data Parallel



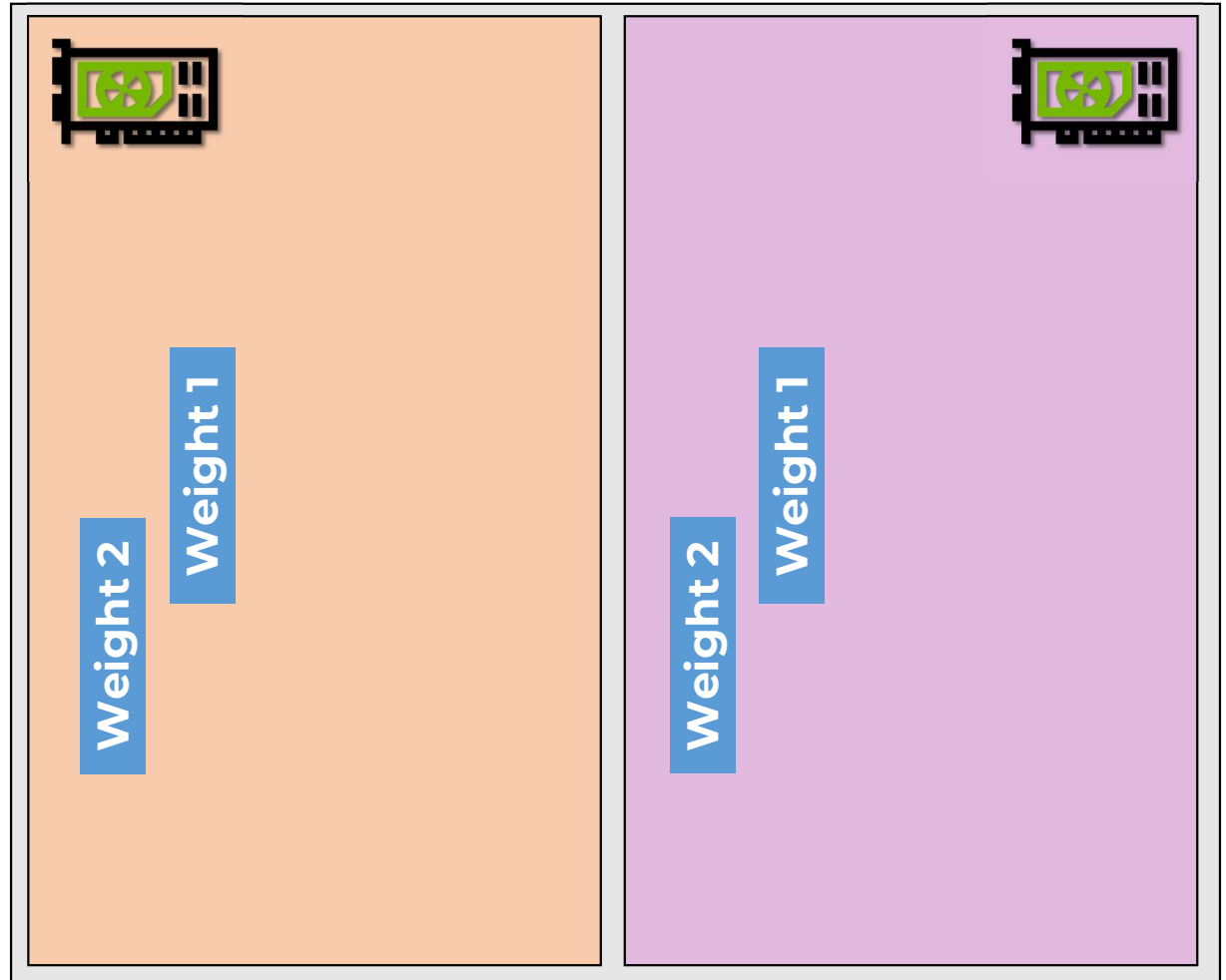
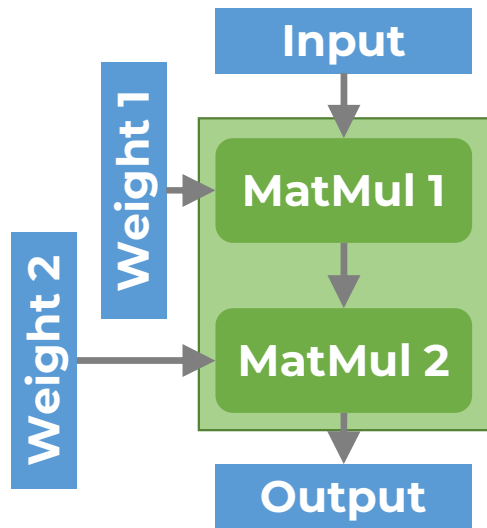
Data Parallel



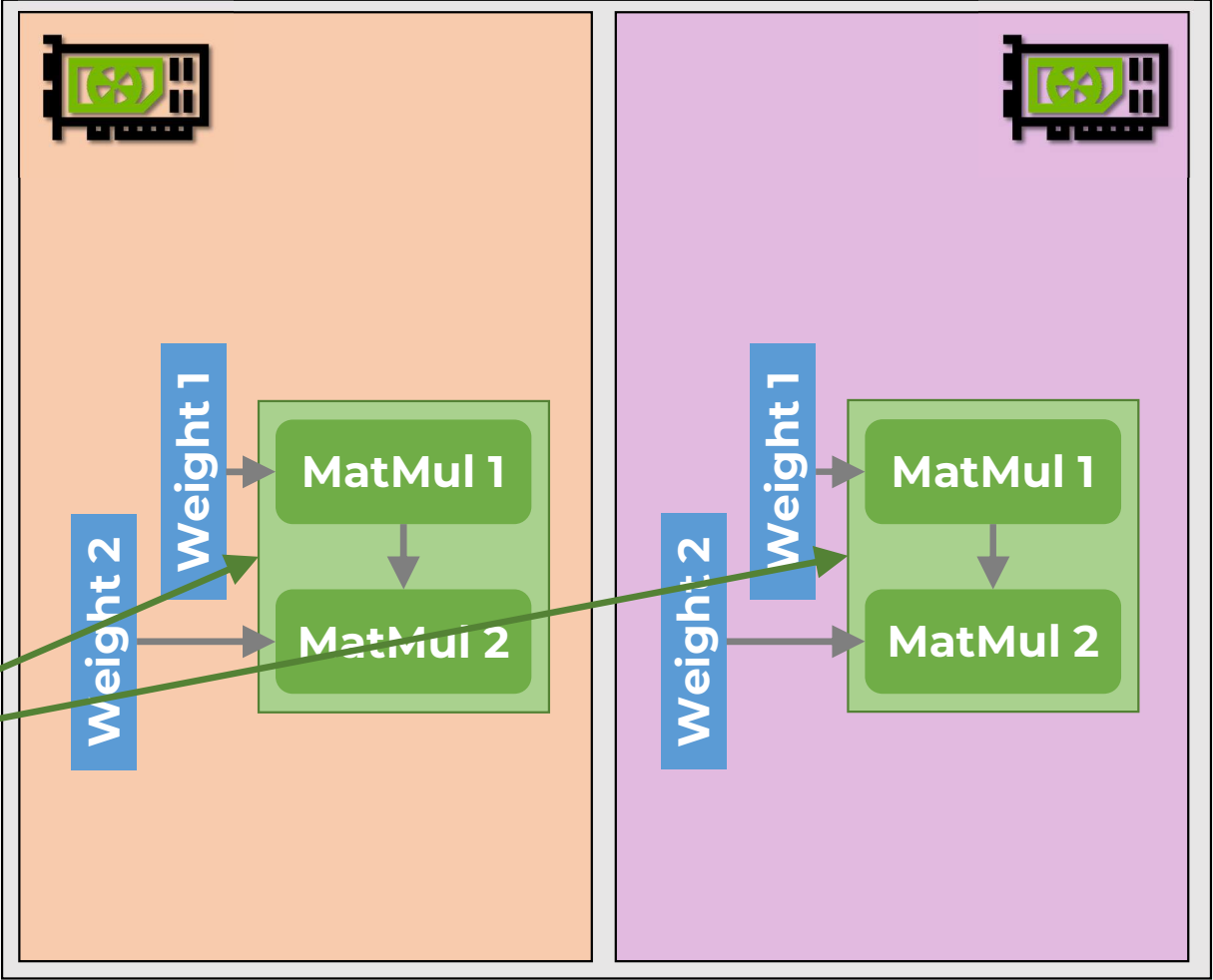
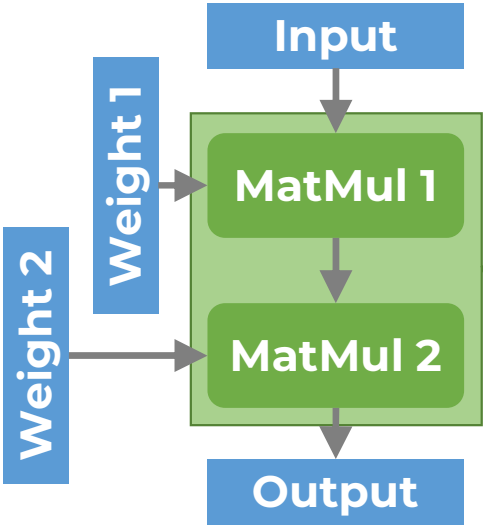
Data Parallel



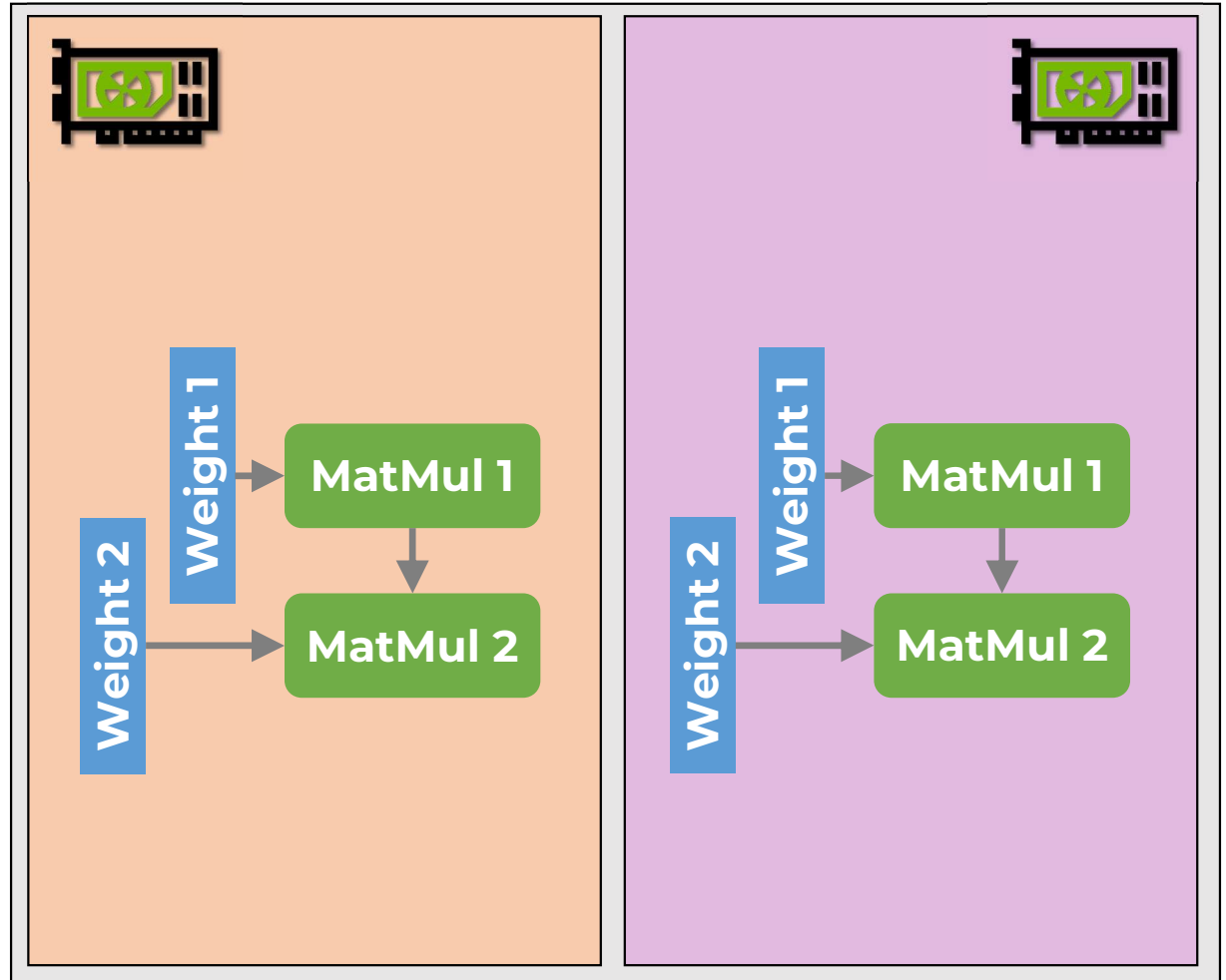
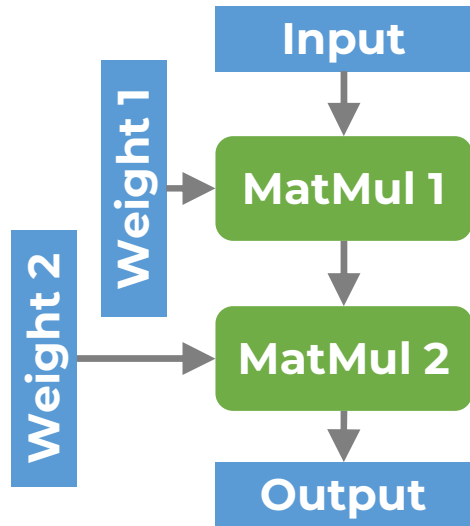
Data Parallel



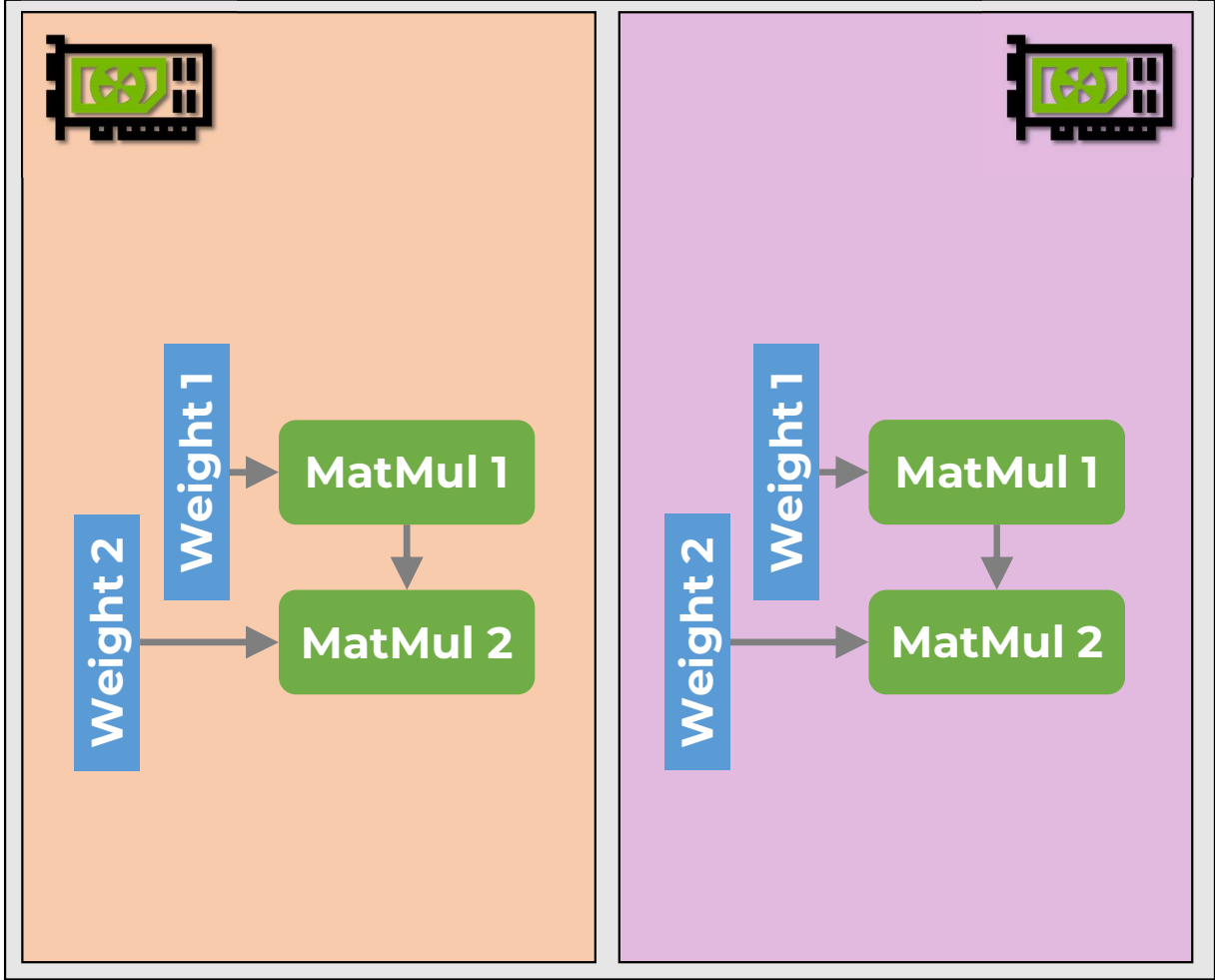
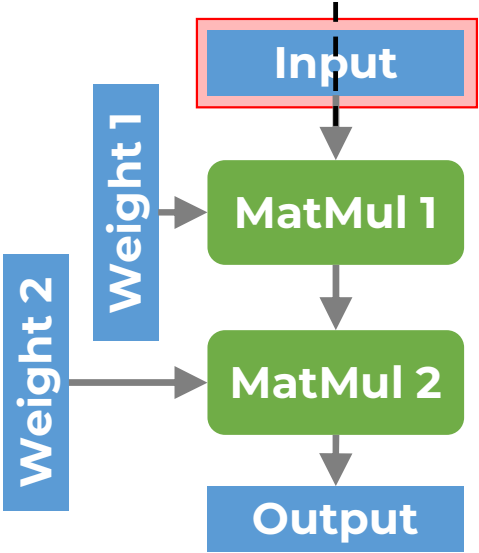
Data Parallel



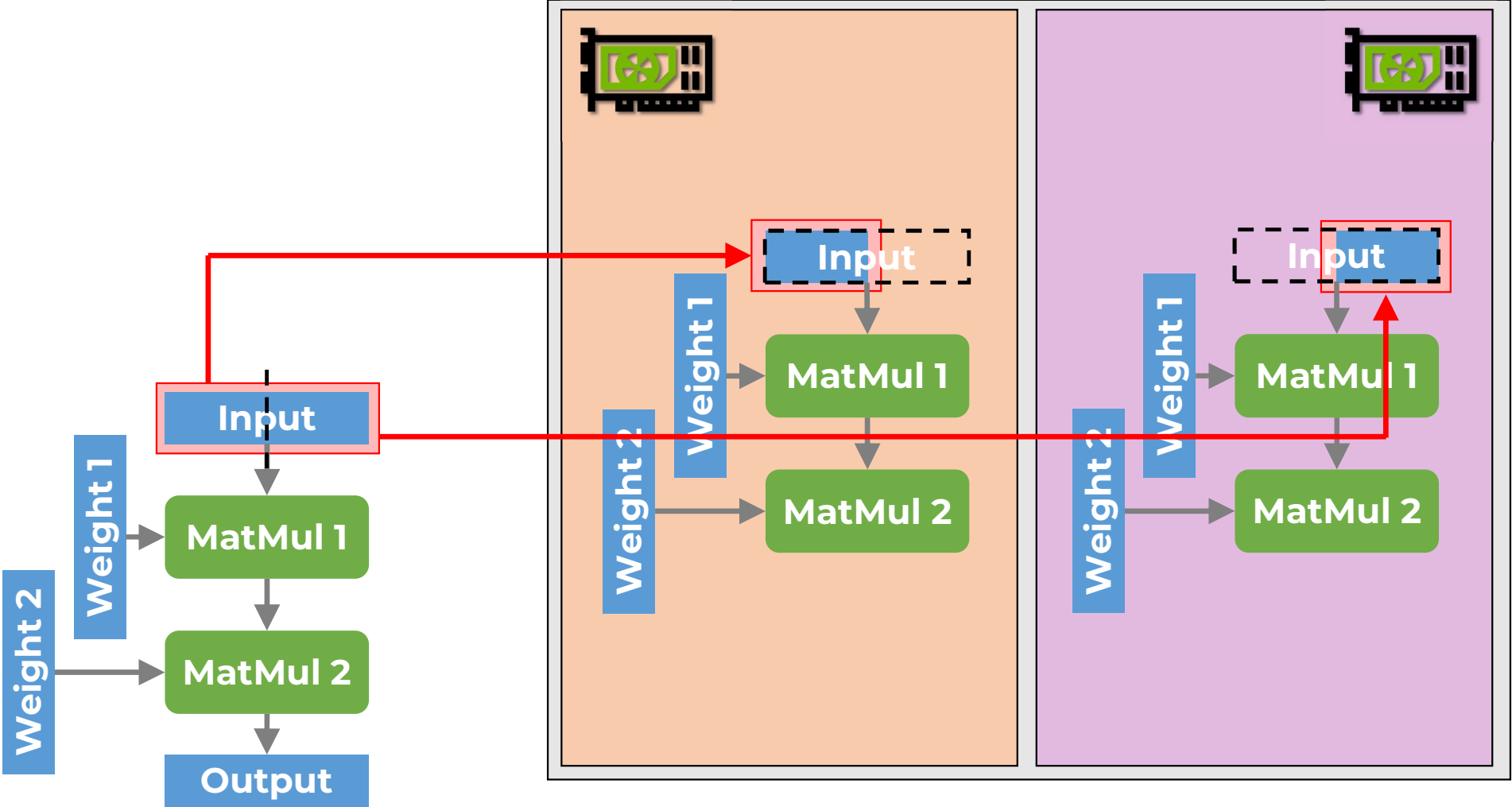
Data Parallel



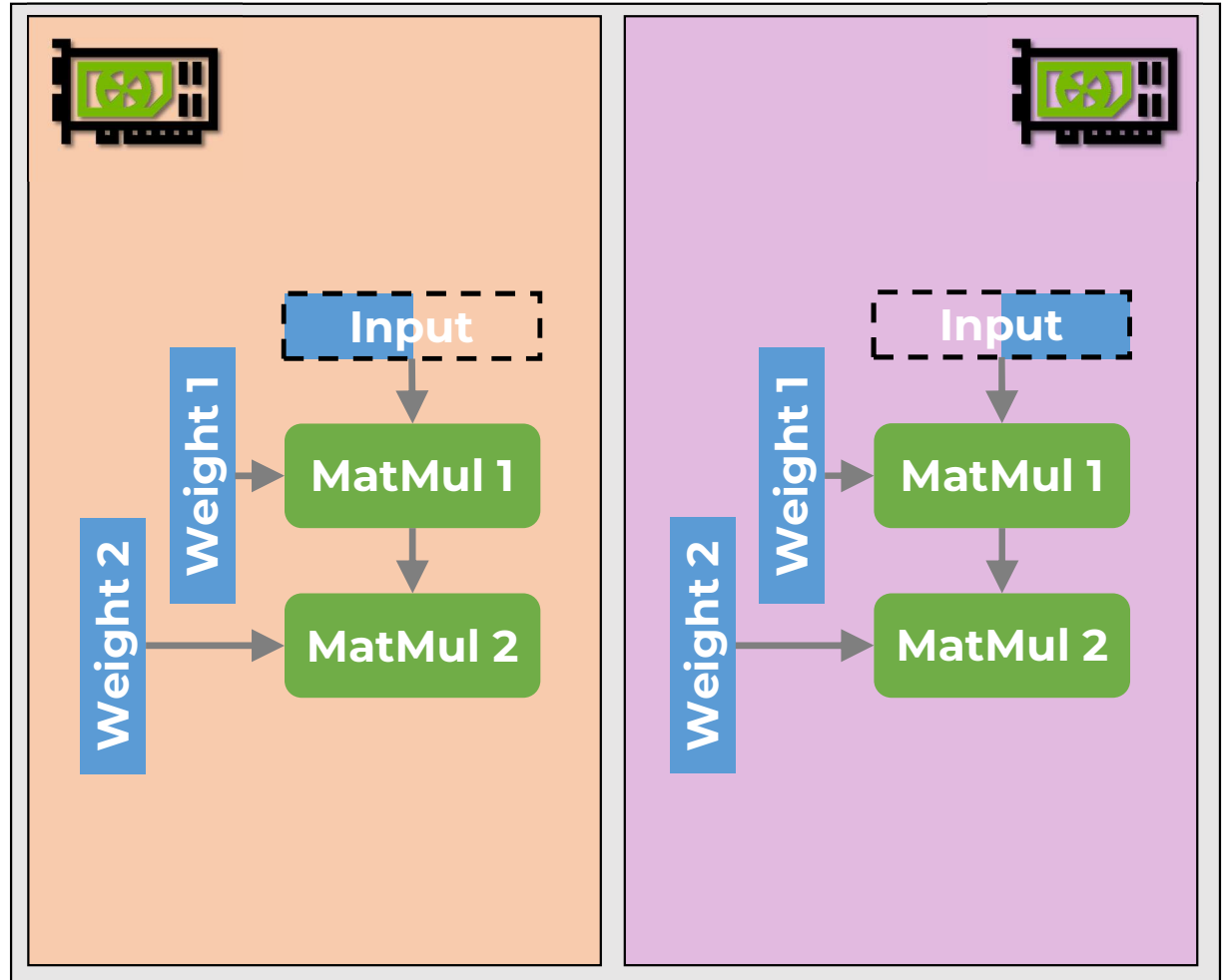
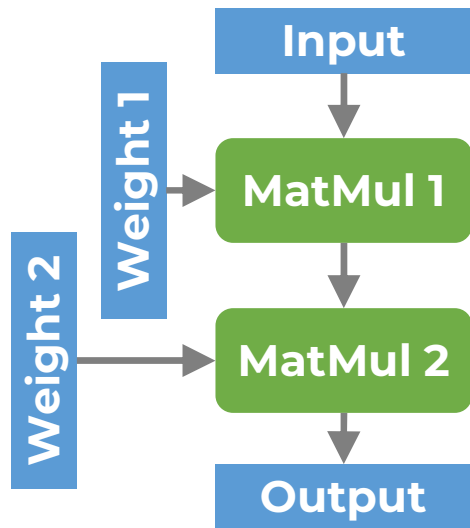
Data Parallel



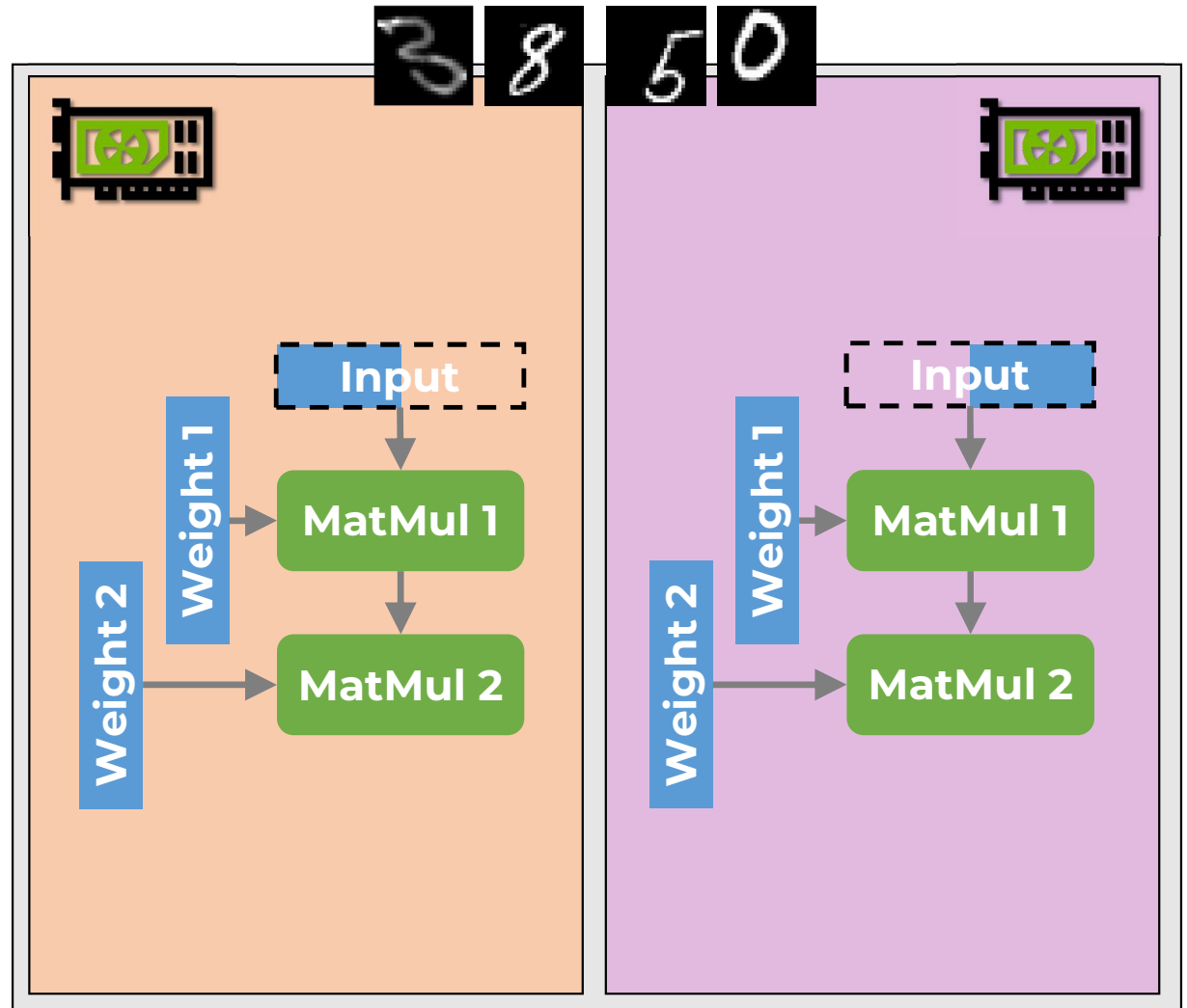
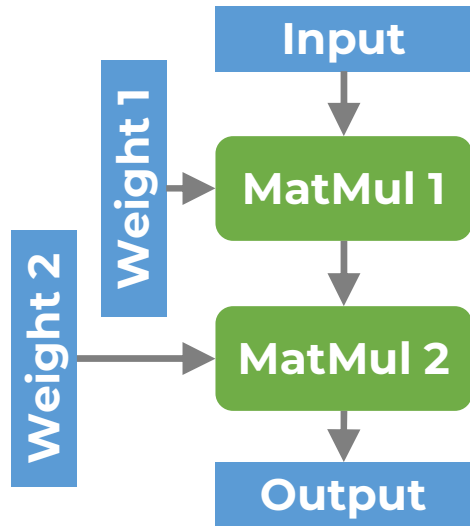
Data Parallel



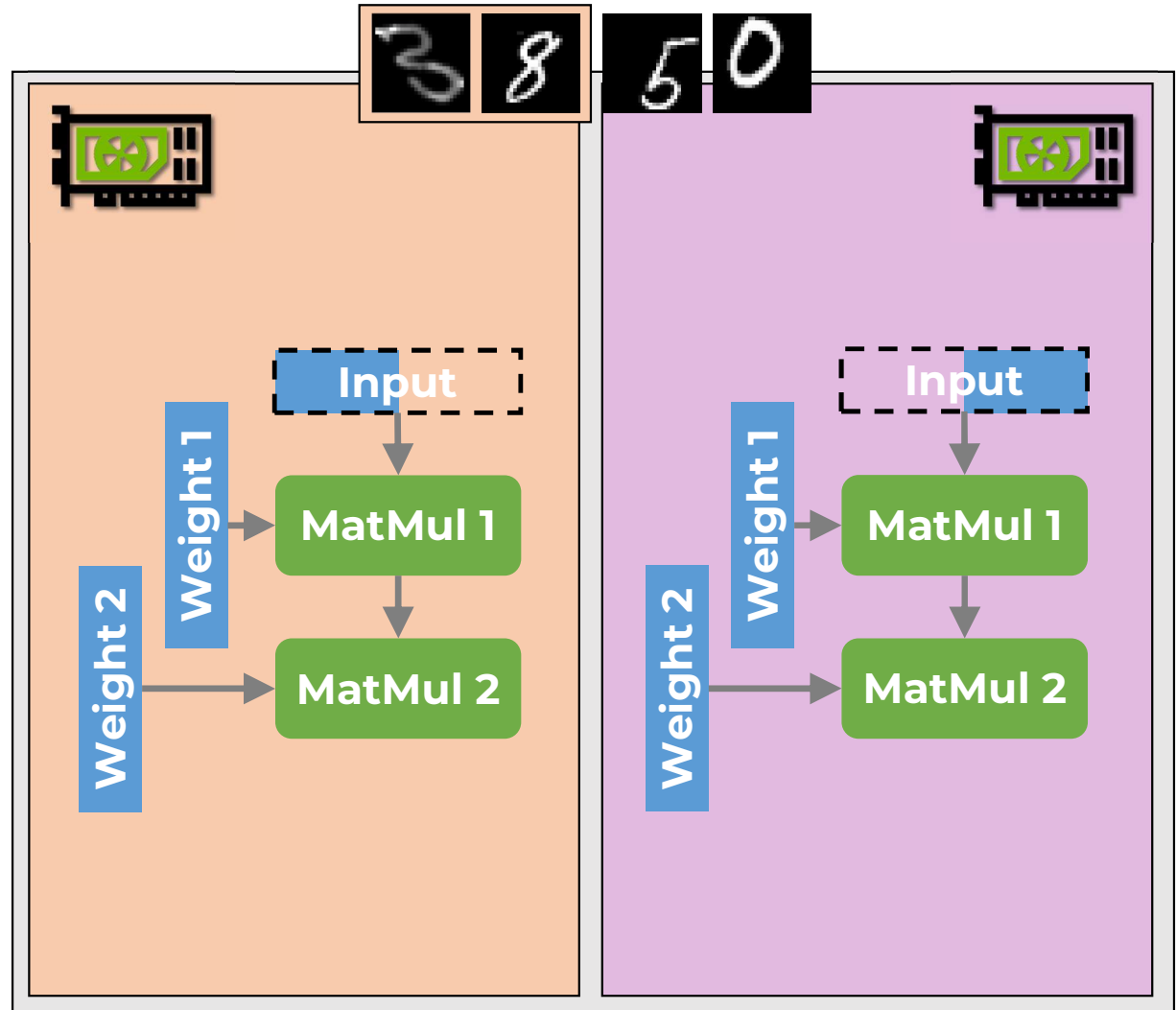
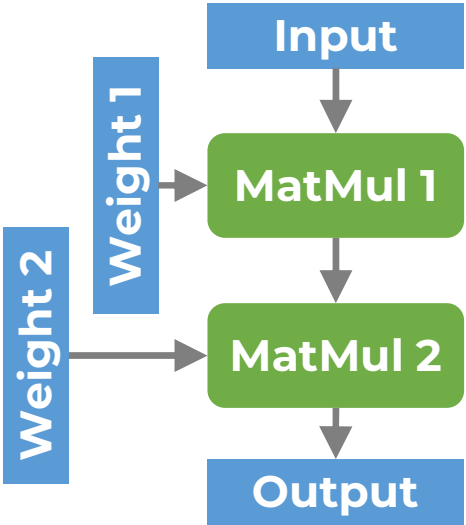
Data Parallel



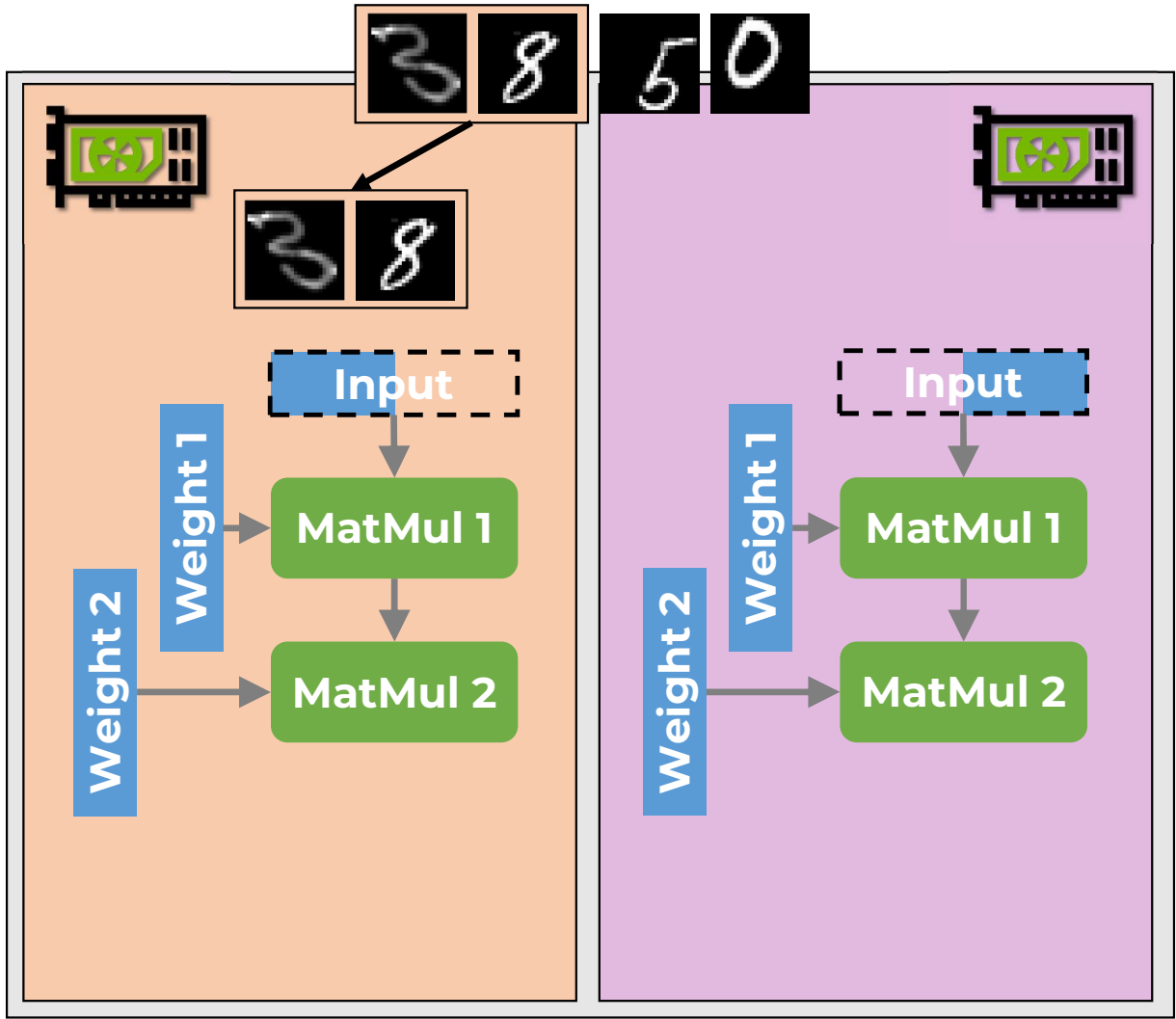
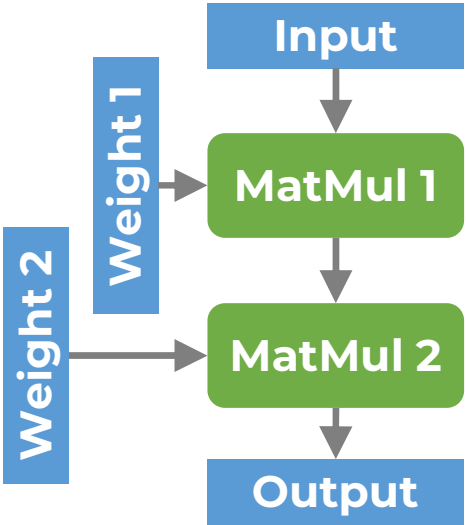
Data Parallel



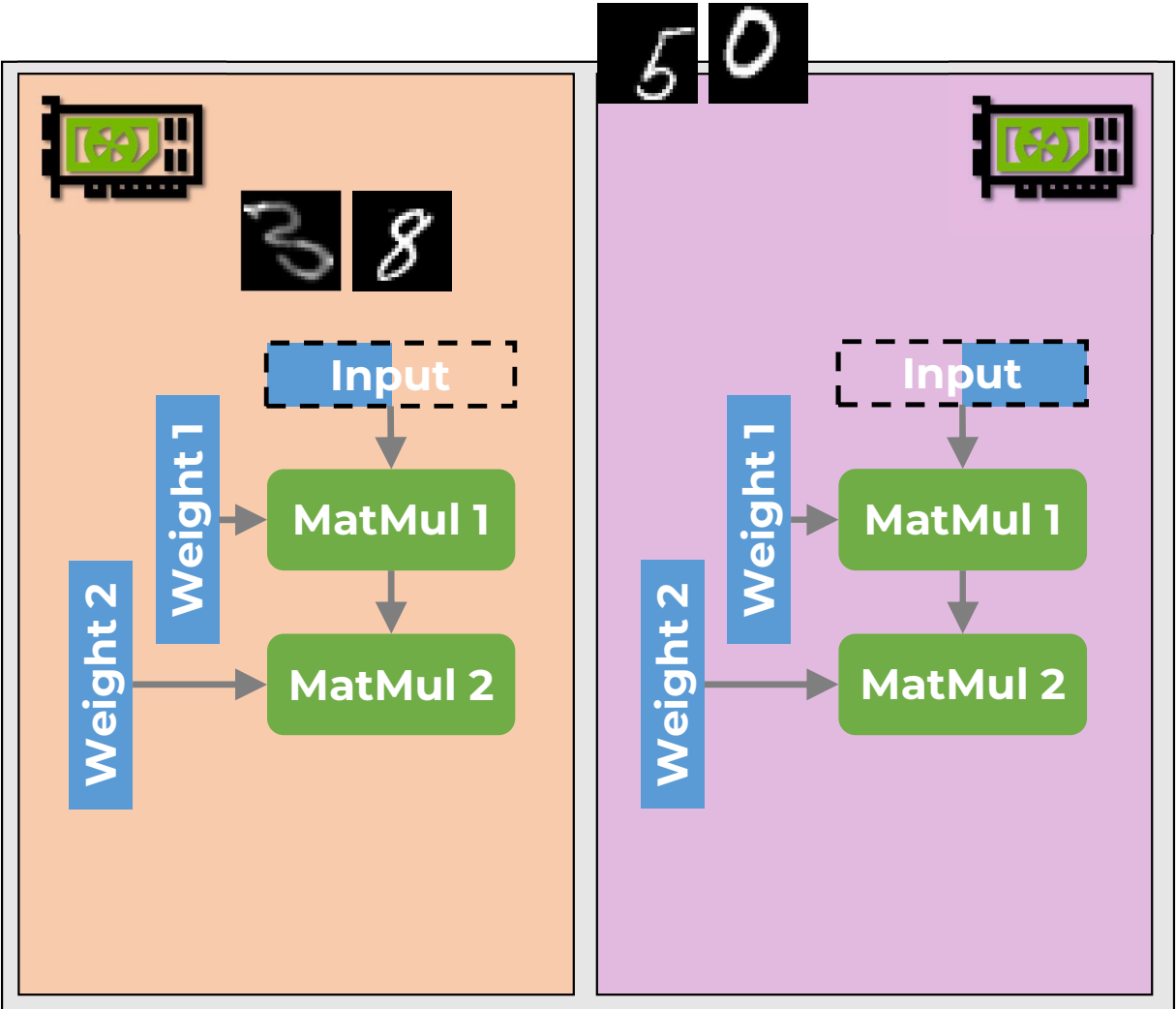
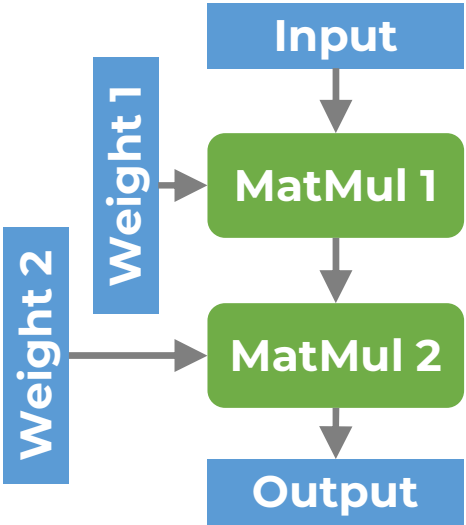
Data Parallel



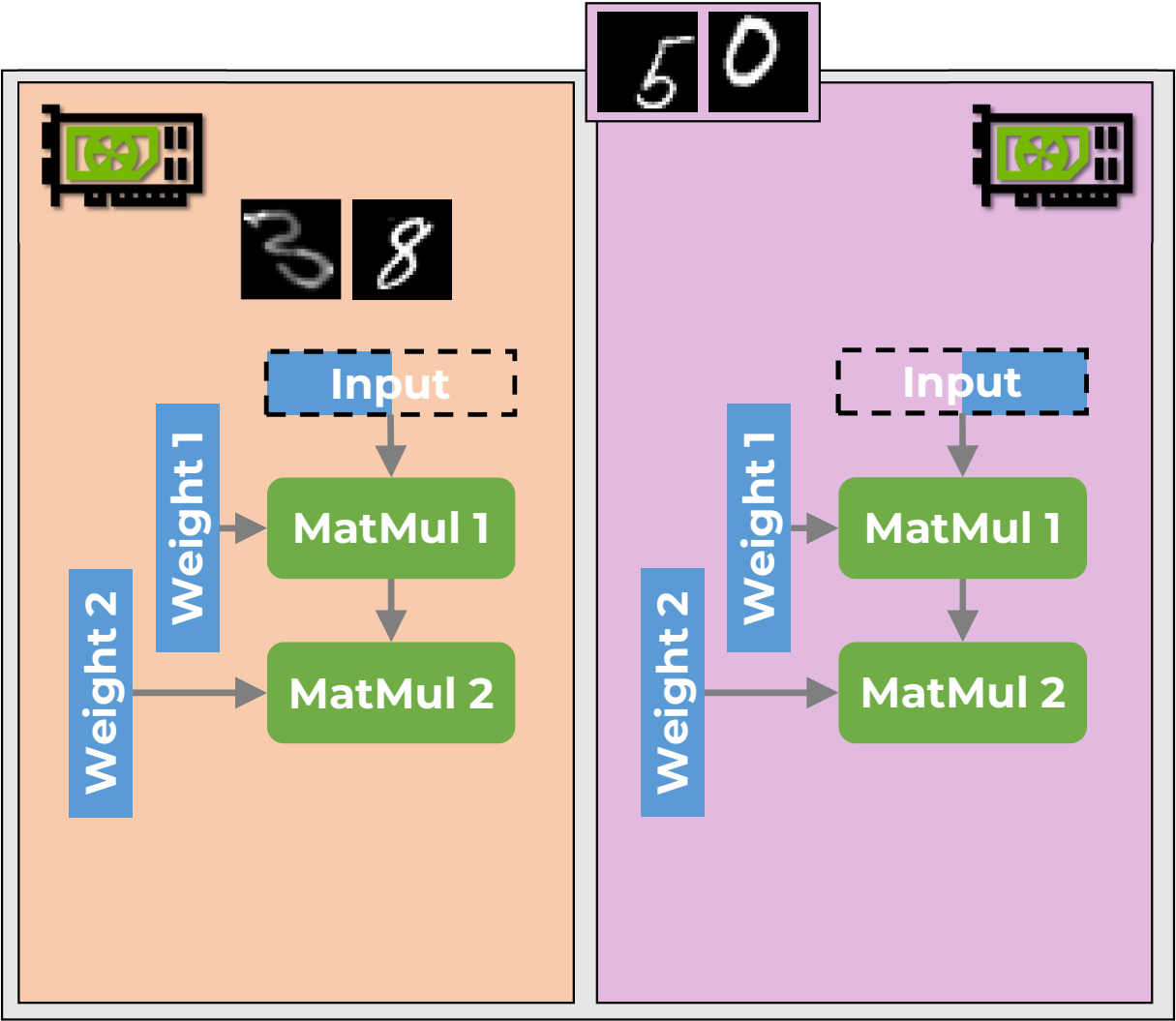
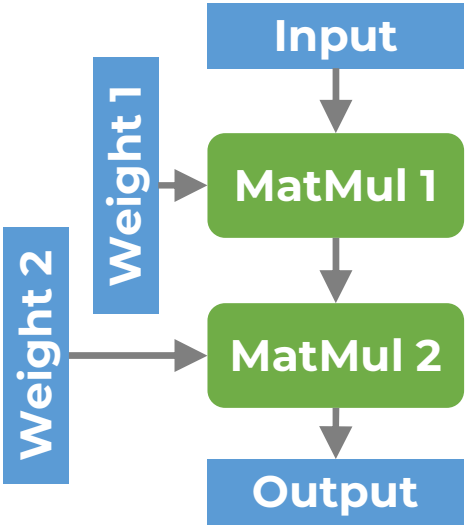
Data Parallel



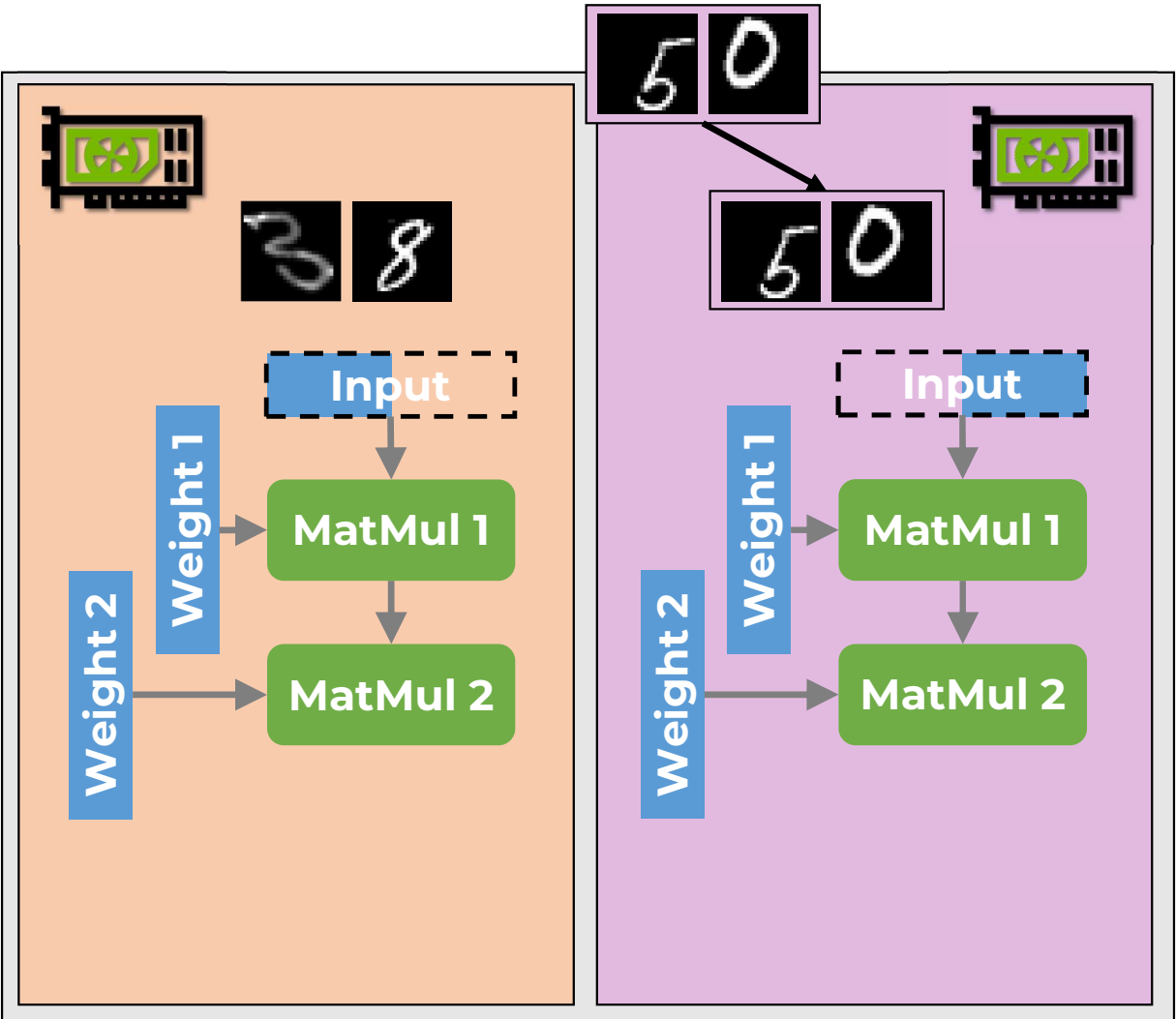
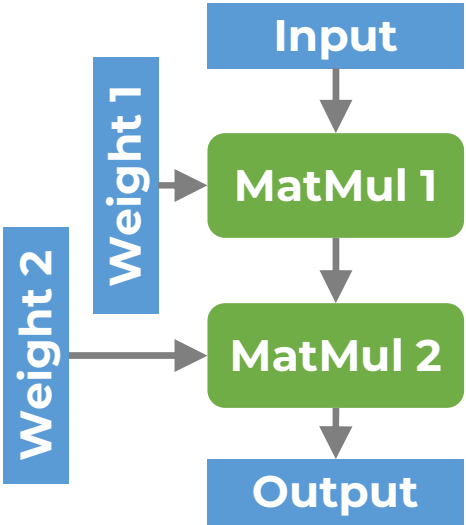
Data Parallel



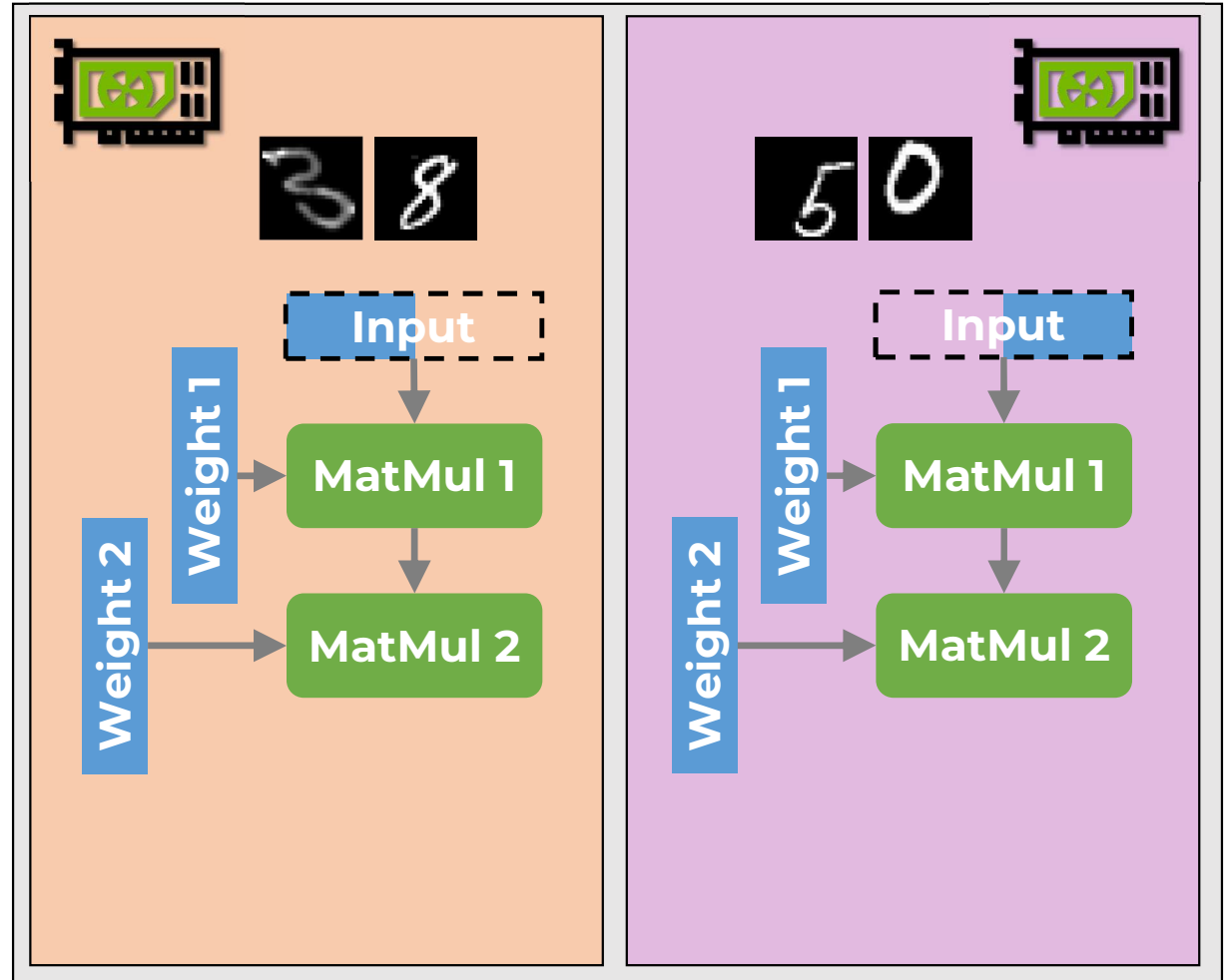
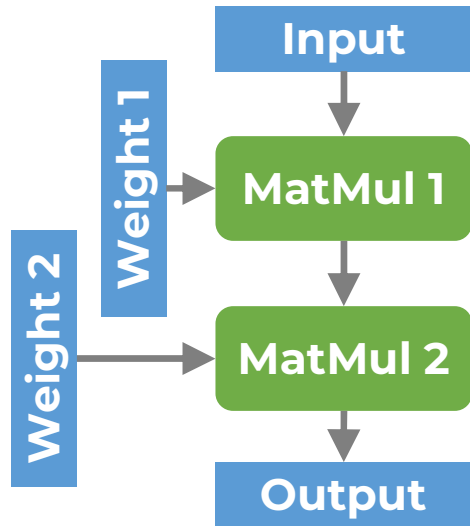
Data Parallel



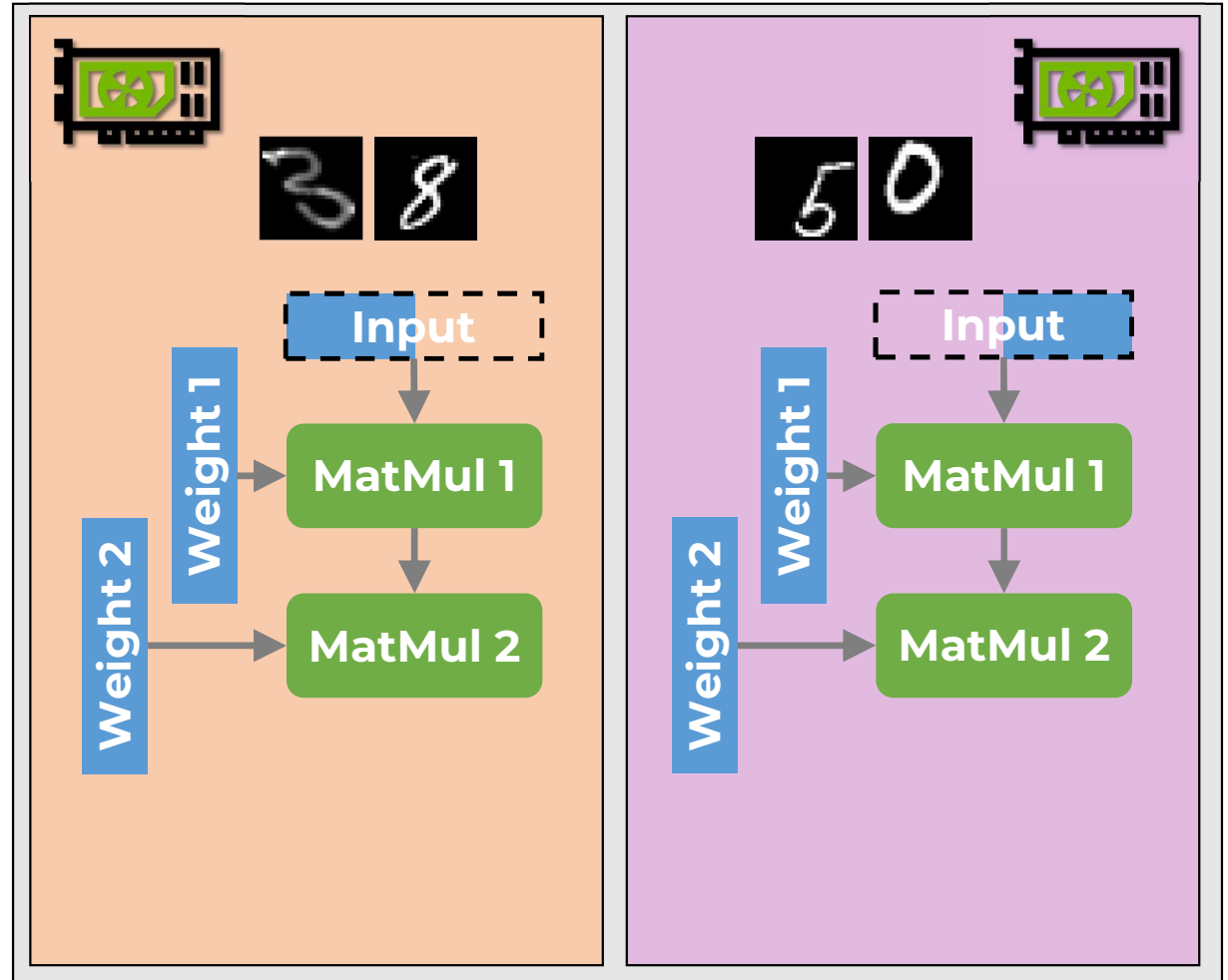
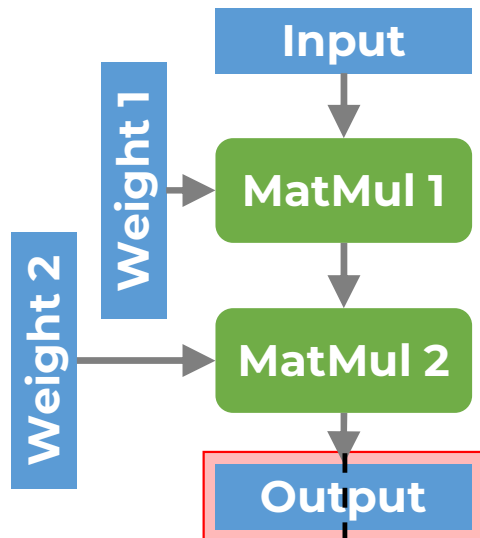
Data Parallel



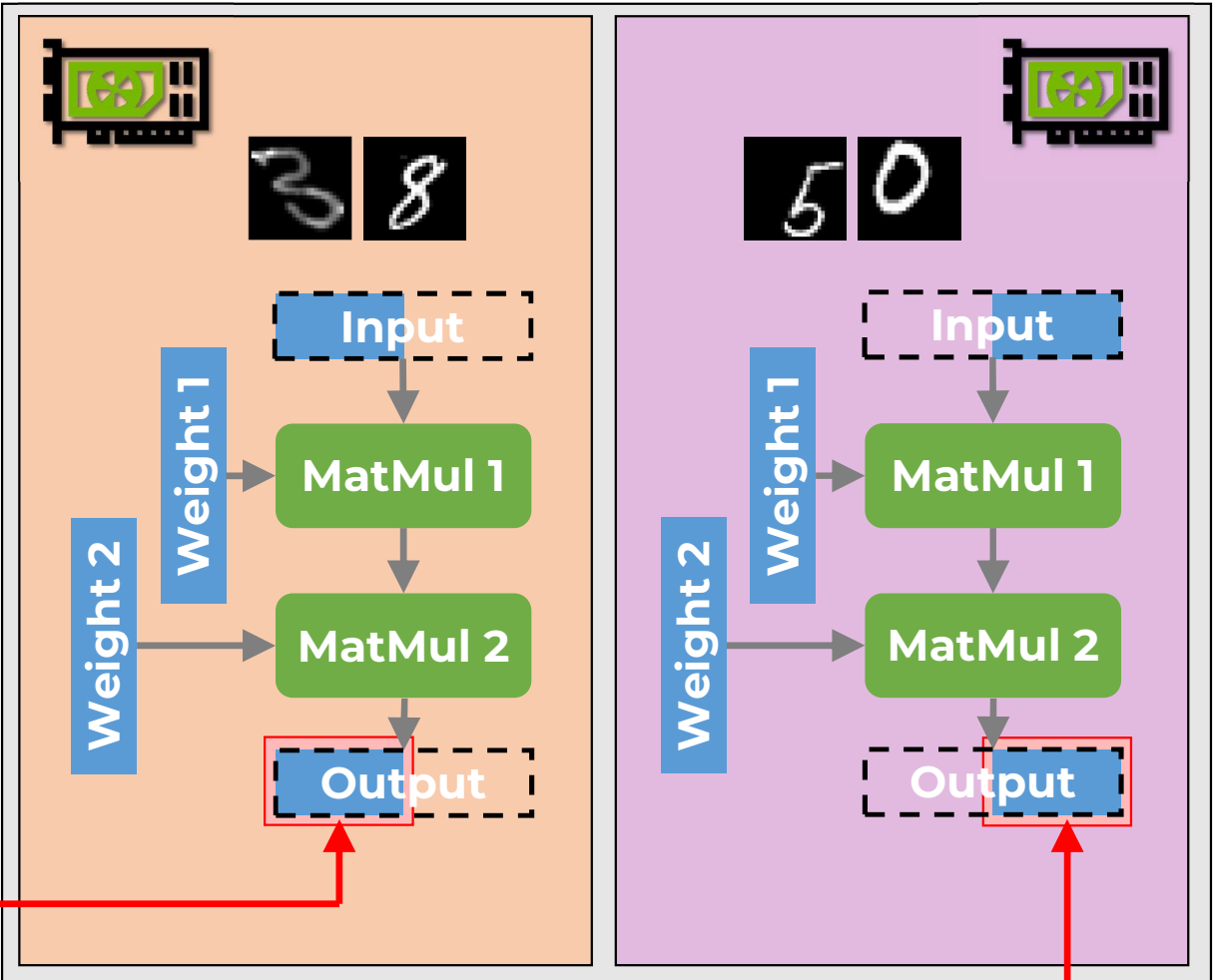
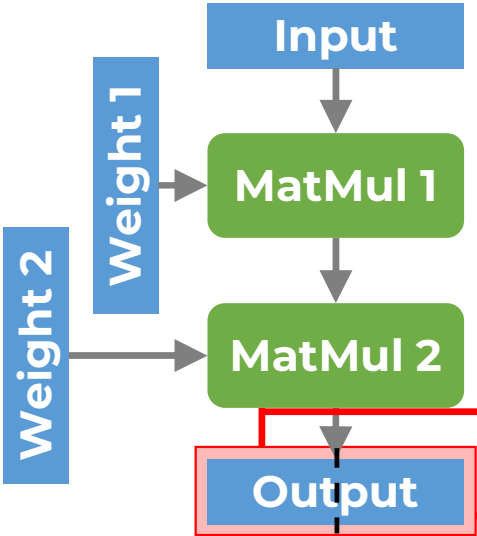
Data Parallel



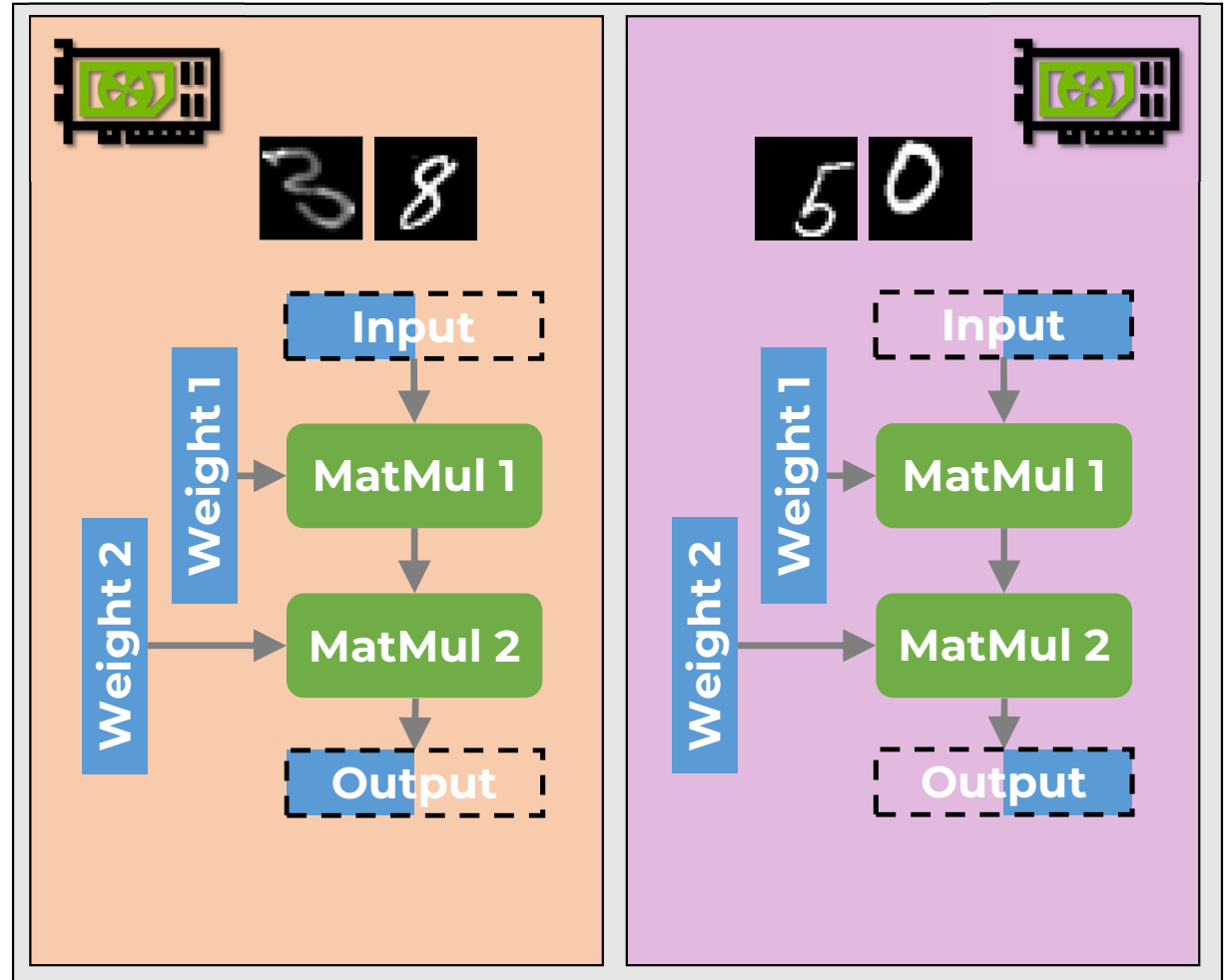
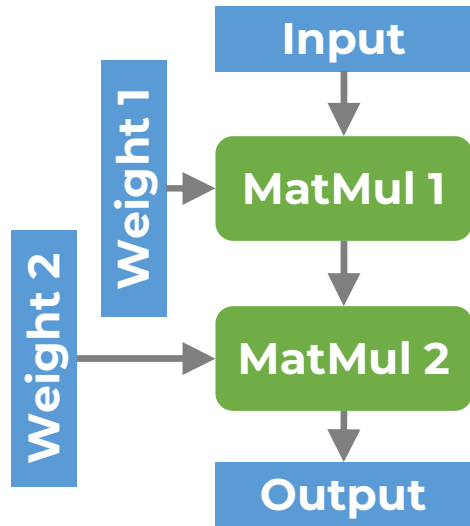
Data Parallel



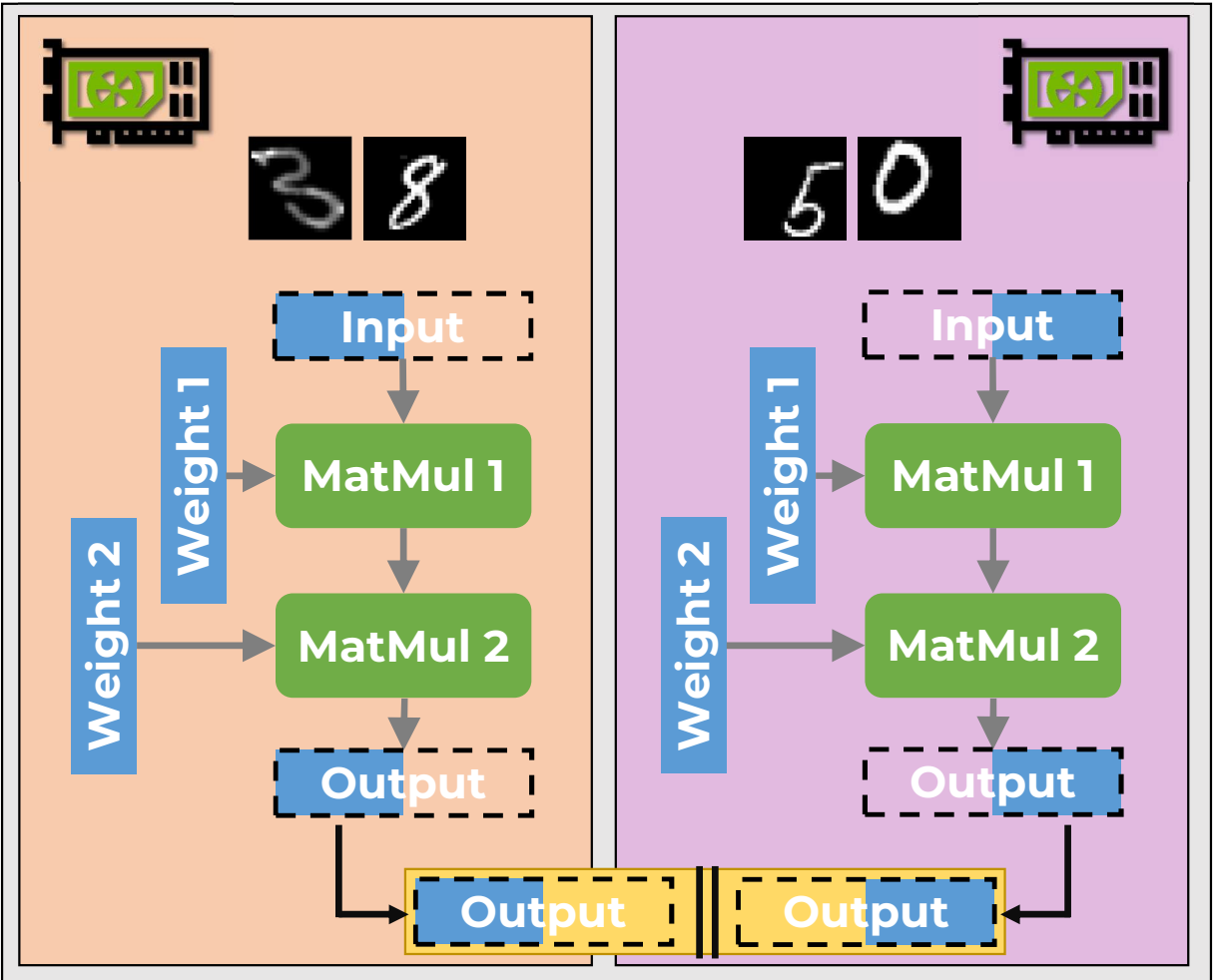
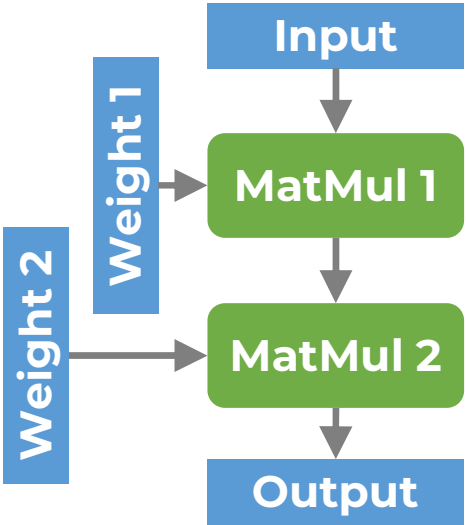
Data Parallel



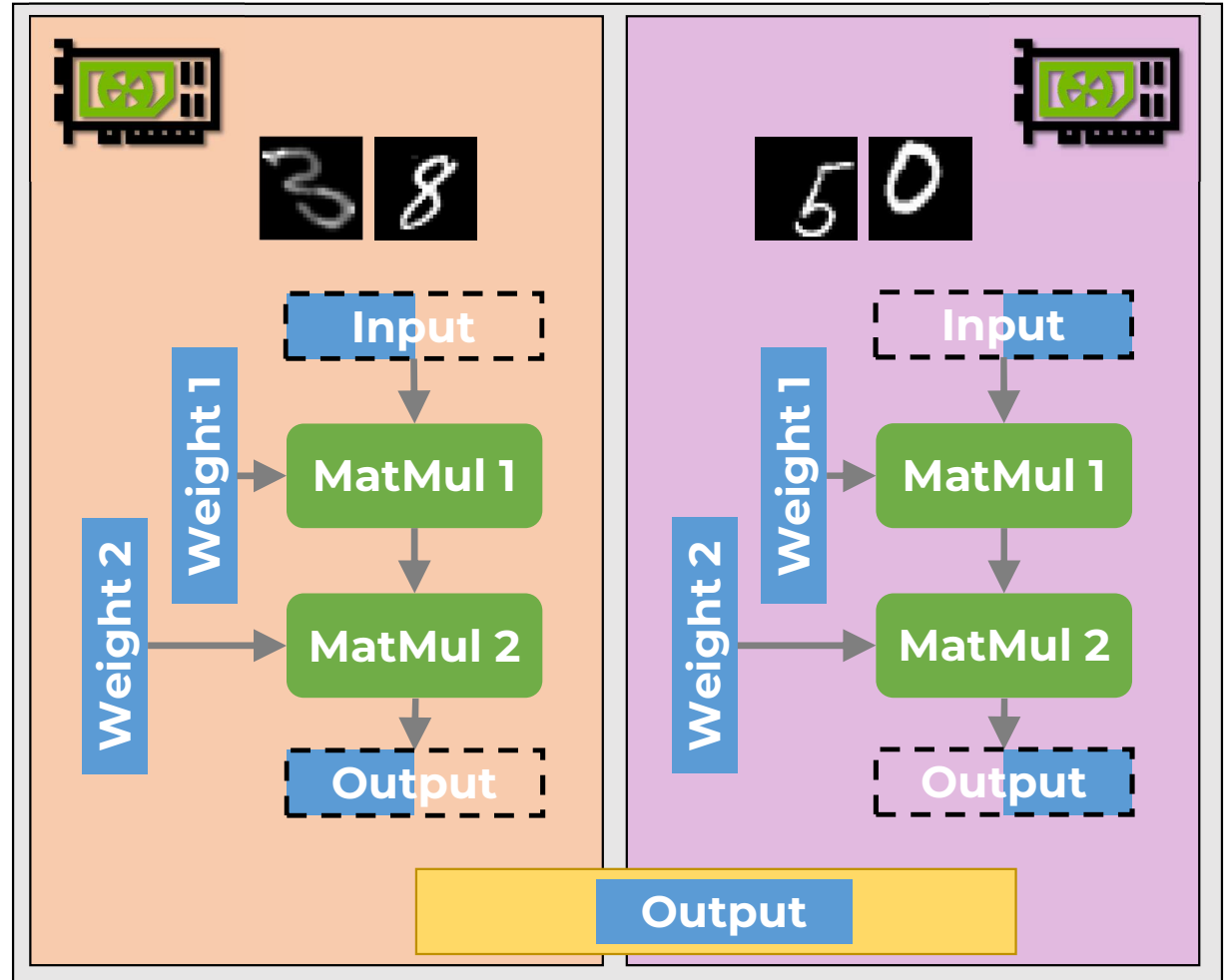
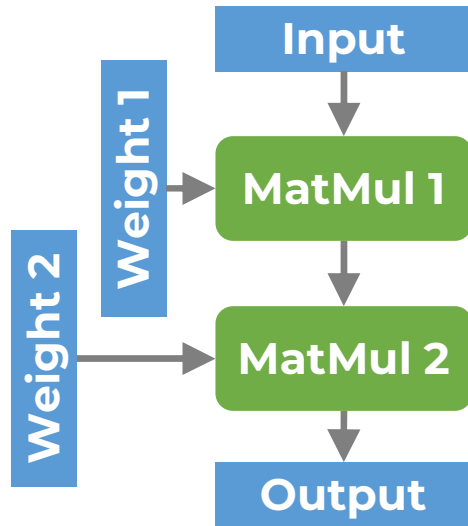
Data Parallel



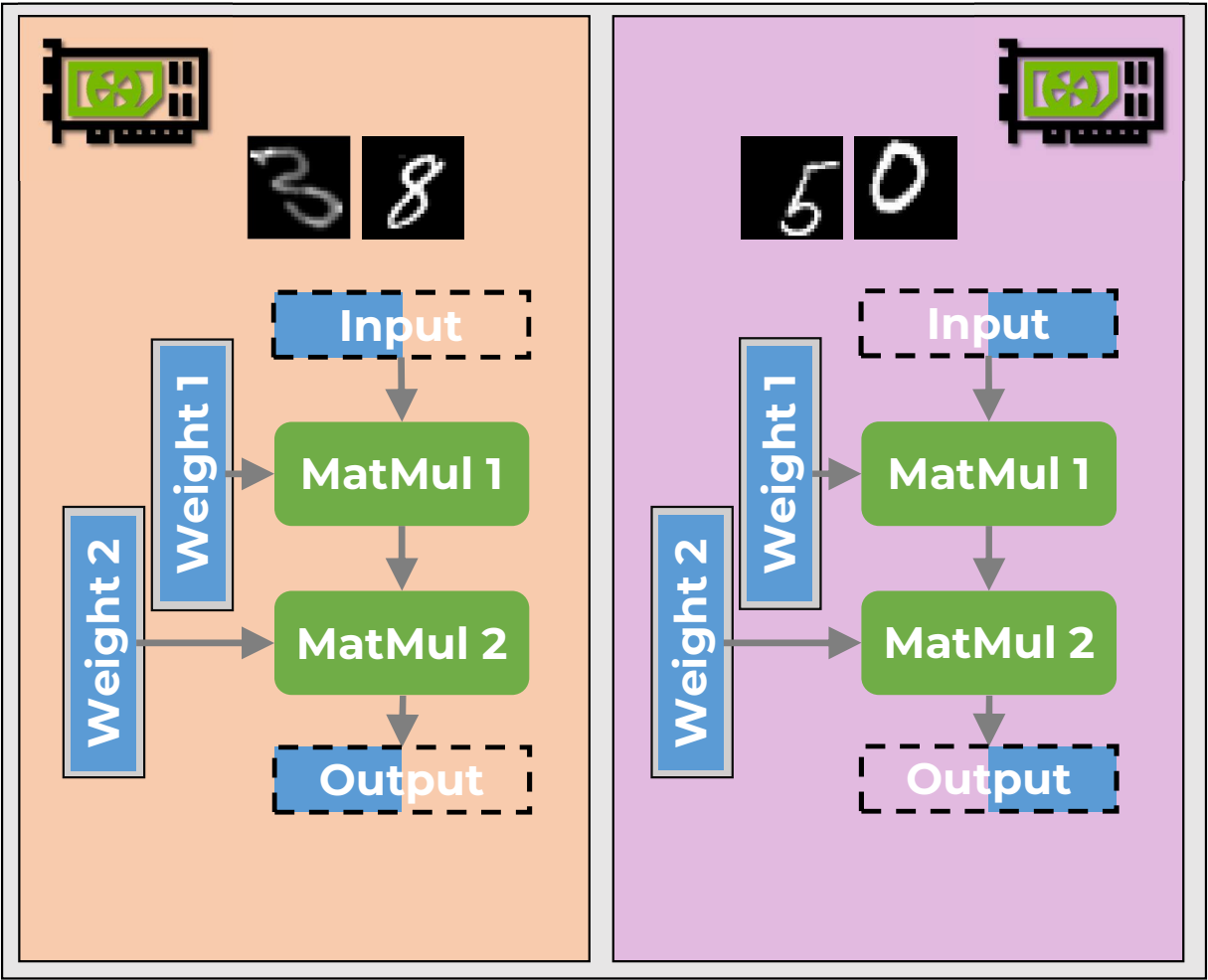
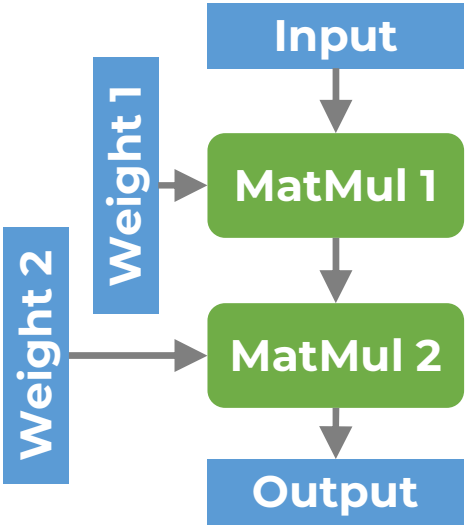
Data Parallel



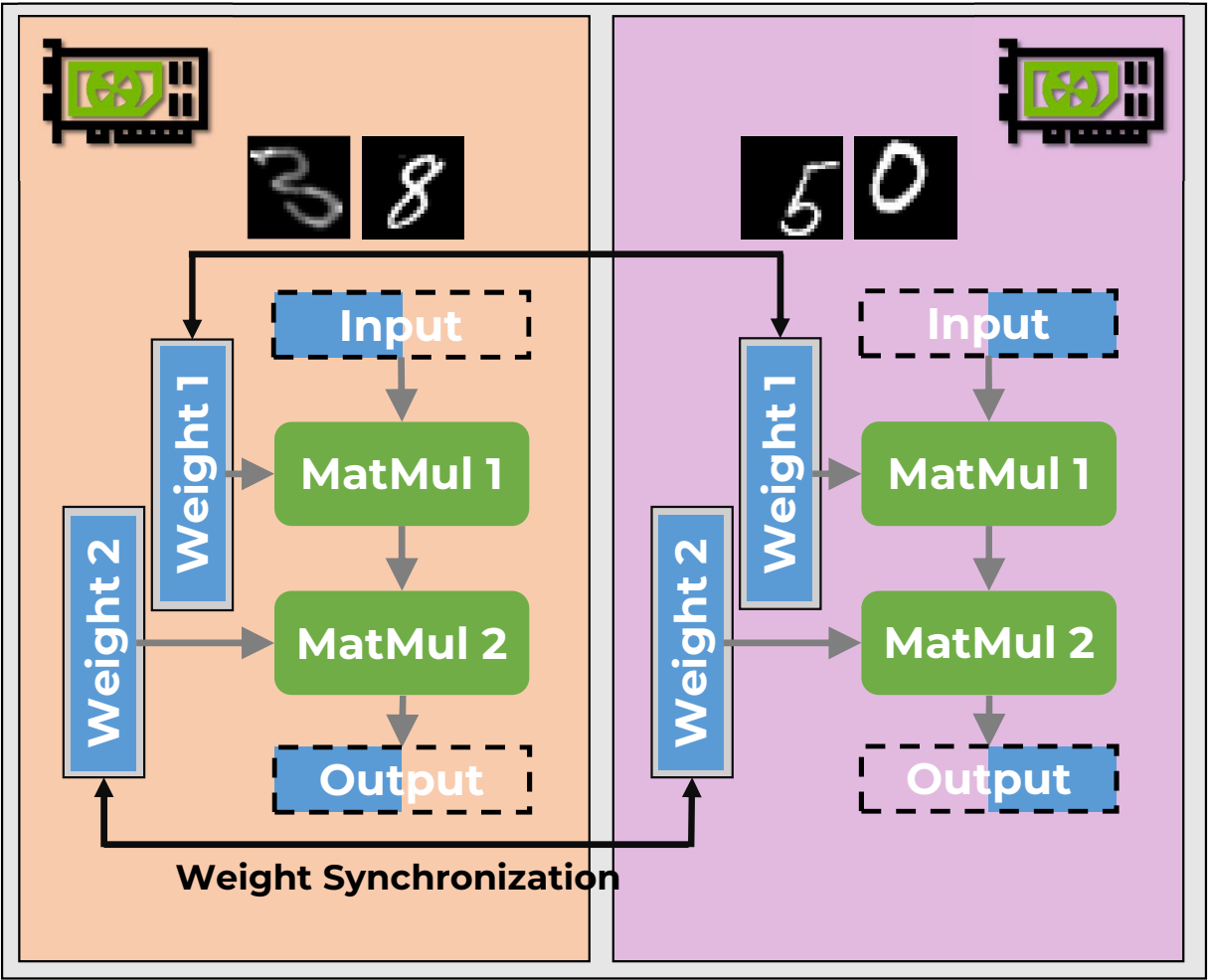
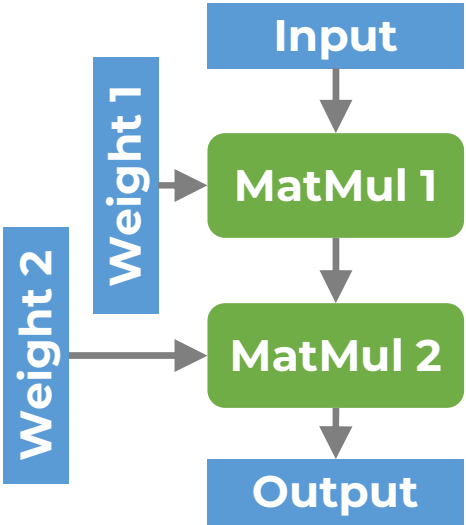
Data Parallel



Data Parallel



Data Parallel



Data Parallel

Data Parallel

Model Parallel

Data Parallel
Model Parallel
Attribute Parallel

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

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

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

Parallelization

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

2^{|Parallelization|}

FlexFlow

Distributed DNN Training

FlexFlow

Distributed DNN Training **Runtime**



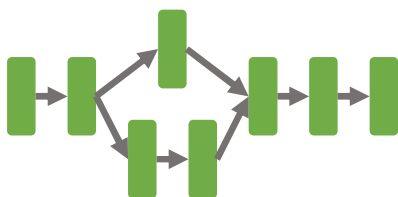
FlexFlow

Distributed DNN Training **Runtime**



FlexFlow

Distributed DNN Training **Runtime**





FlexFlow

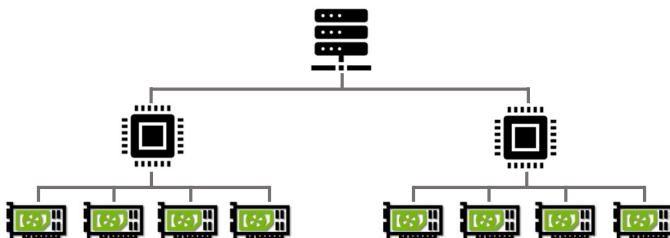
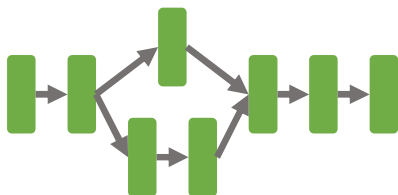
Distributed DNN Training **Runtime**





FlexFlow

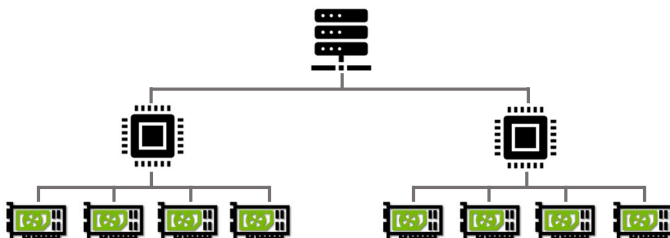
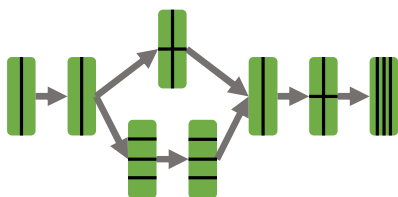
Distributed DNN Training Runtime





FlexFlow

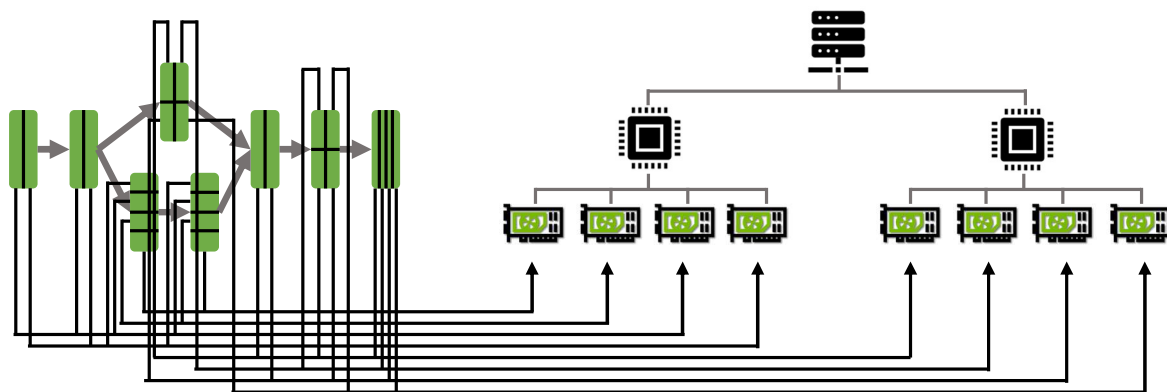
Distributed DNN Training Runtime





FlexFlow

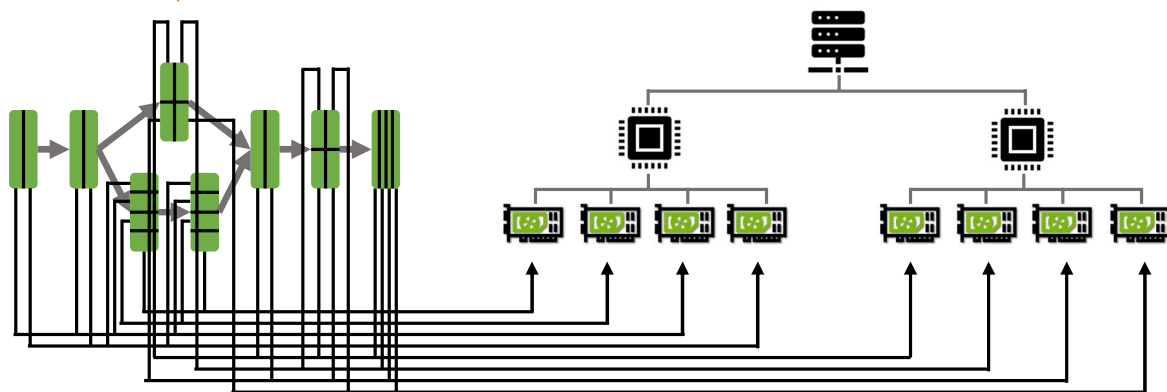
Distributed DNN Training **Runtime**





FlexFlow

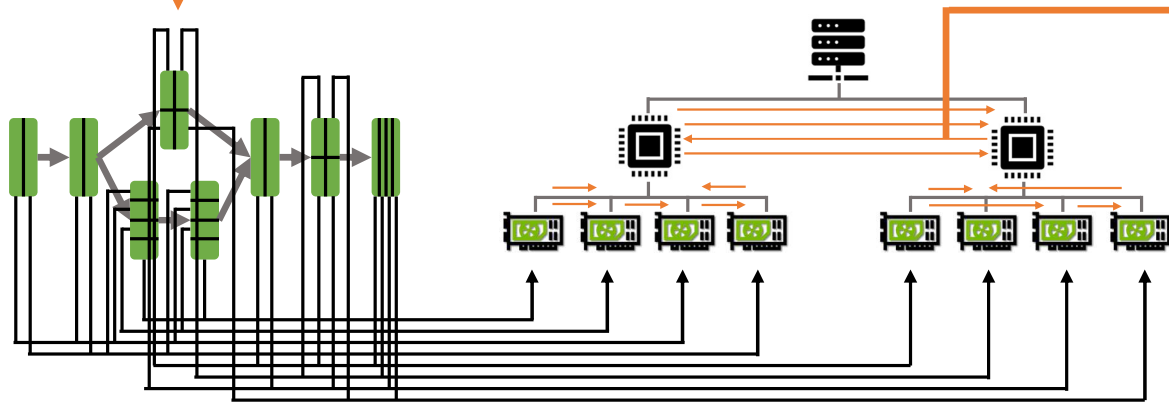
Distributed DNN Training **Runtime**





FlexFlow

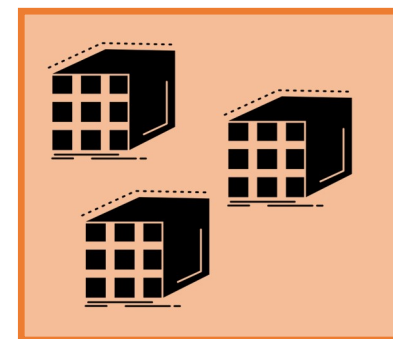
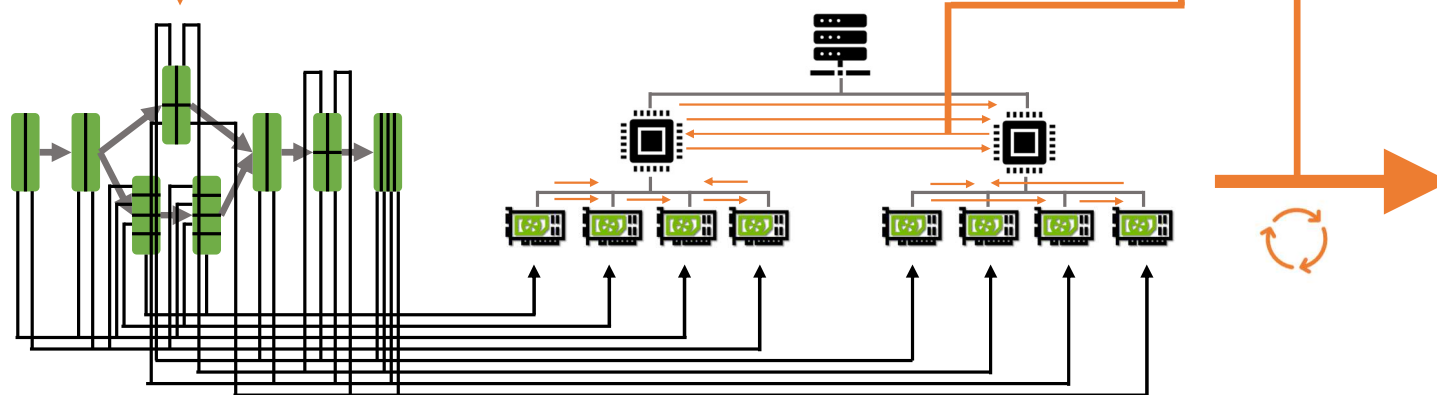
Distributed DNN Training Runtime





FlexFlow

Distributed DNN Training Runtime

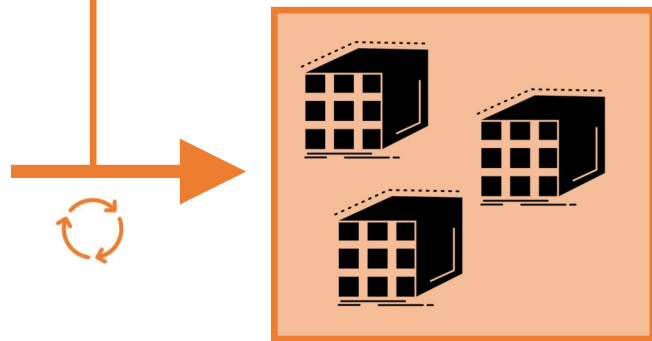
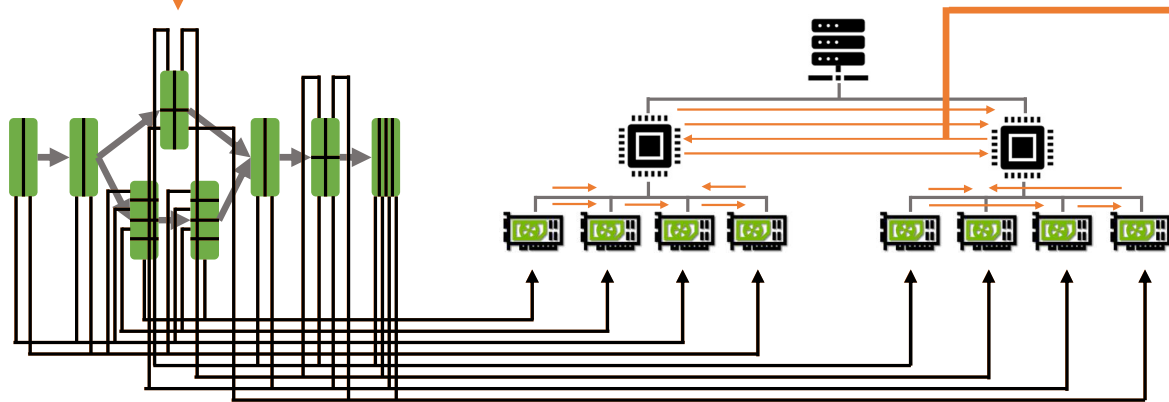




FlexFlow

Distributed DNN Training

Auto-Parallelizer

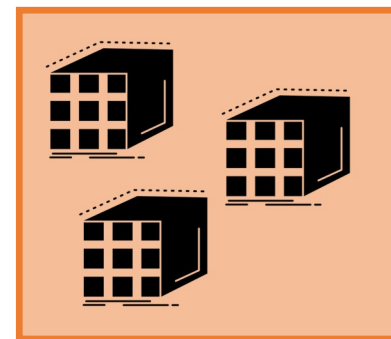
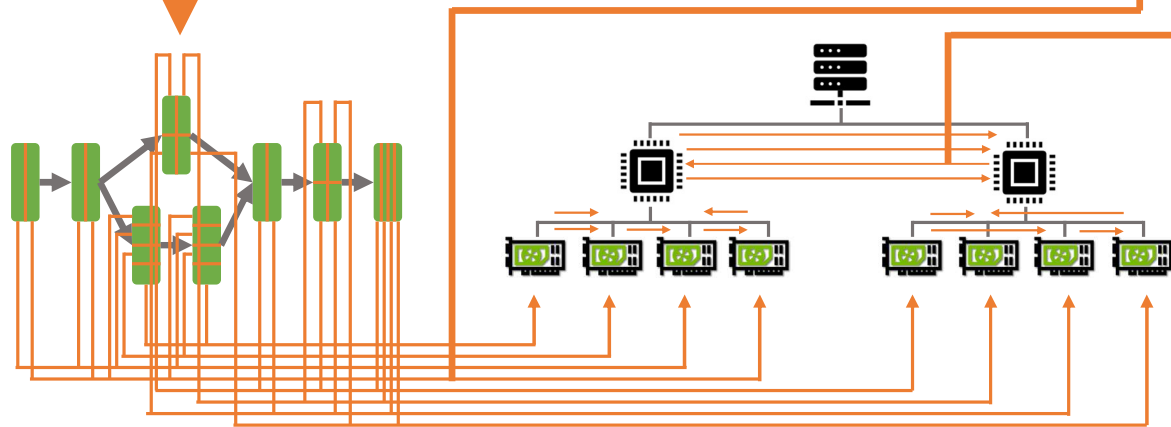




FlexFlow

Distributed DNN Training

Auto-Parallelizer

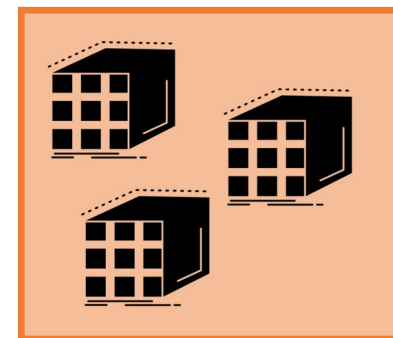
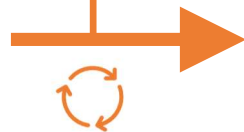
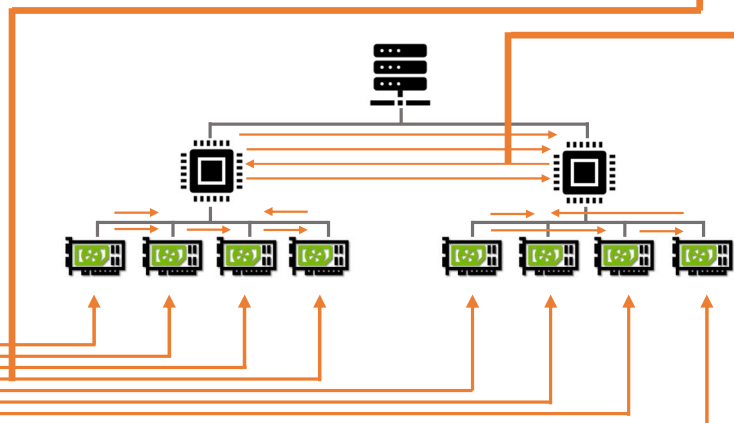
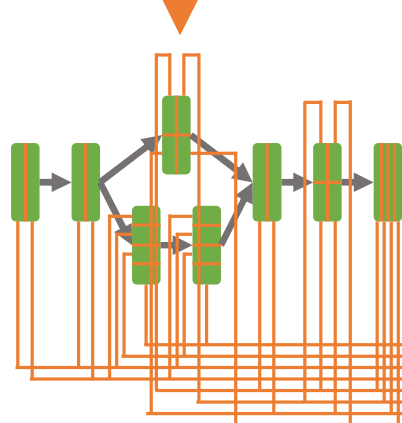




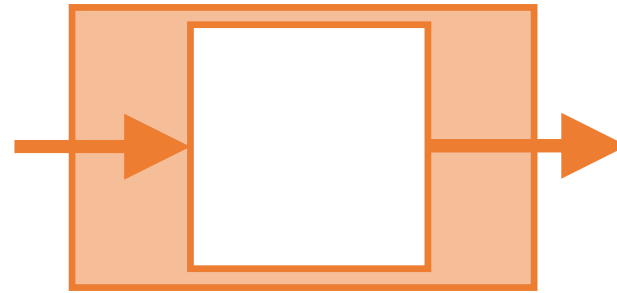
FlexFlow

Distributed DNN Training

Search-Based Auto-Parallelizer



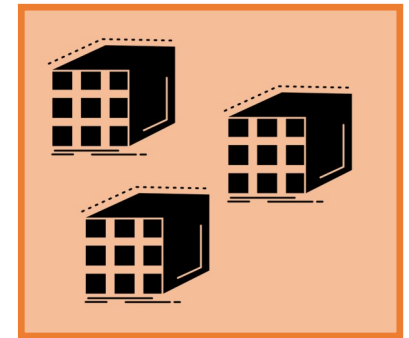
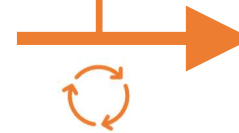
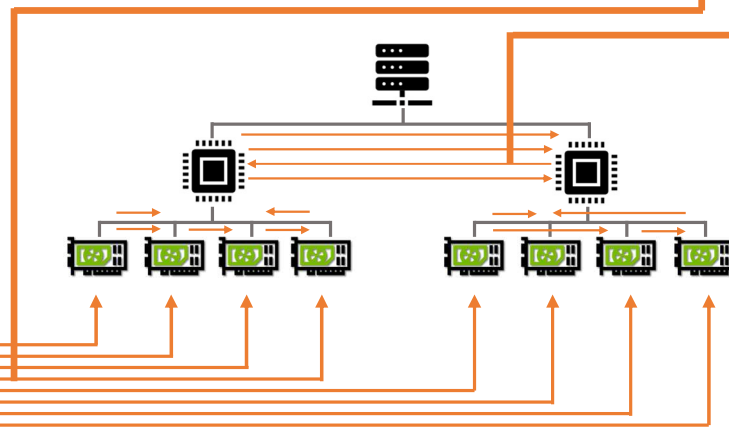
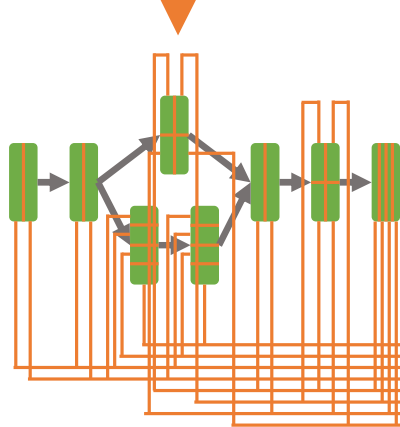
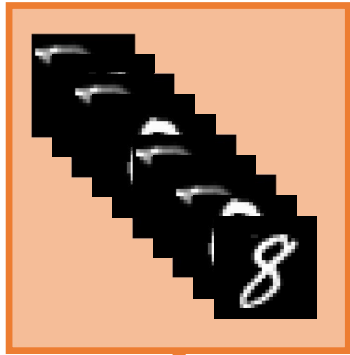
Optimizer



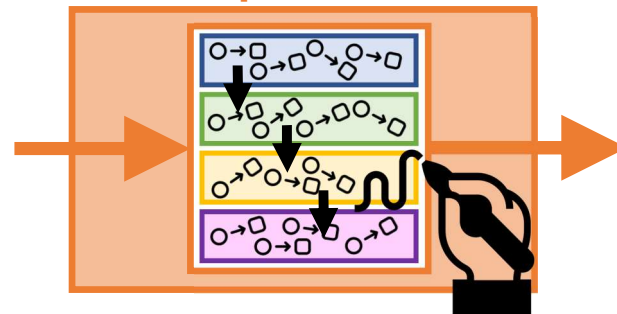
FlexFlow

Distributed DNN Training

Search-Based Auto-Parallelizer



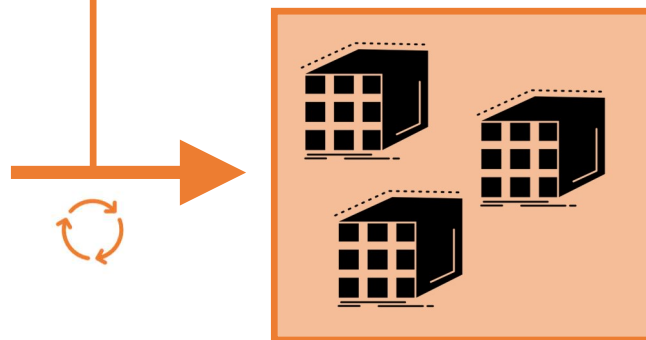
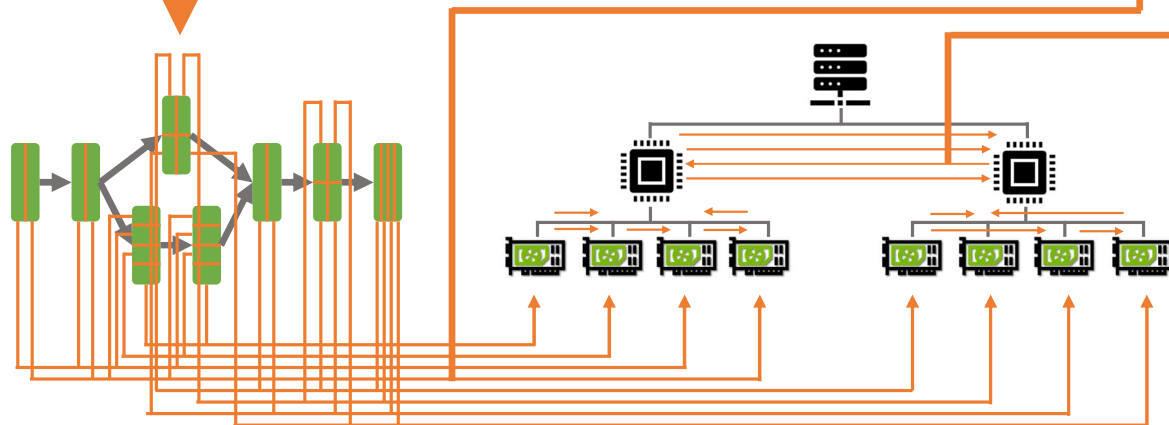
Optimizer

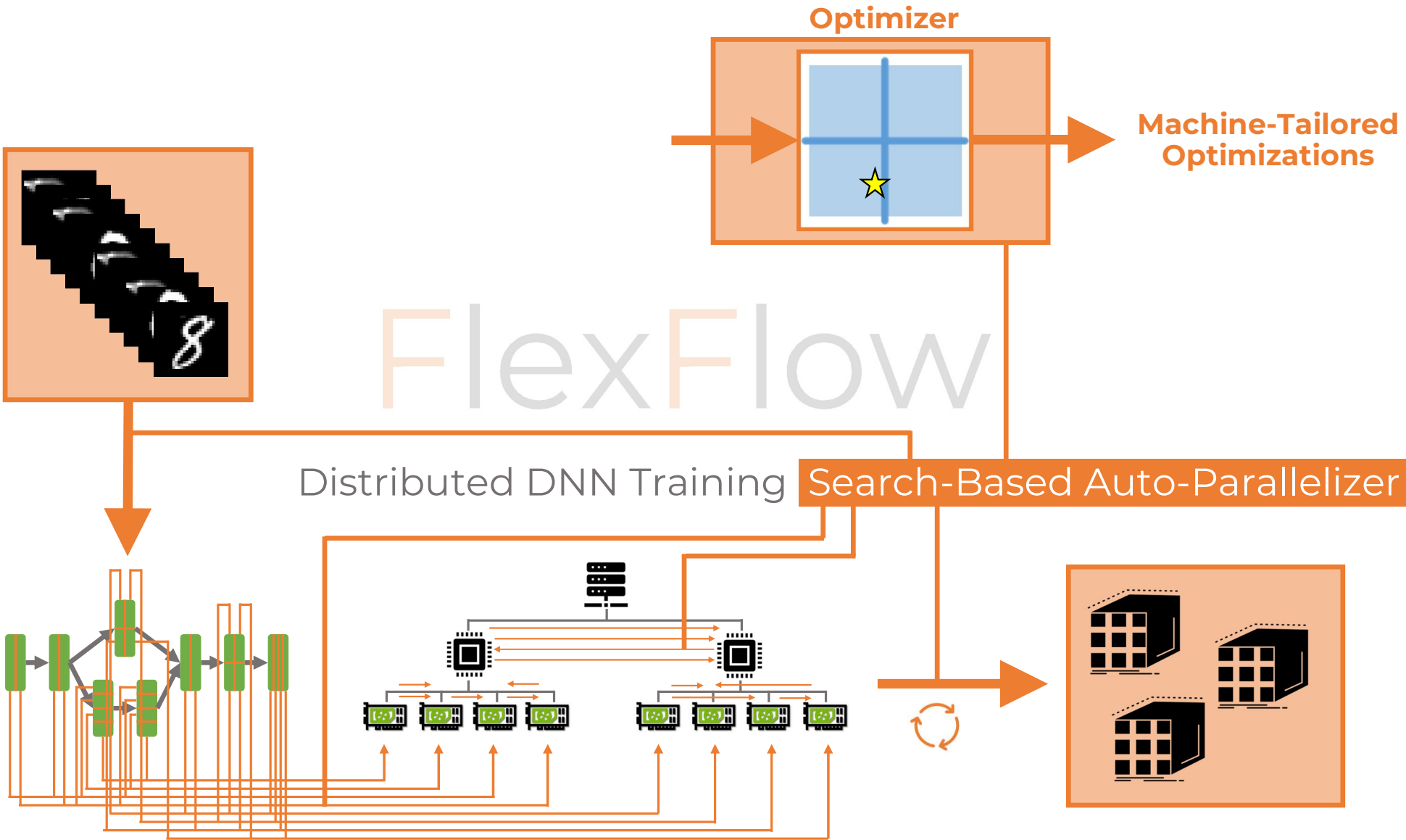


FlexFlow

Distributed DNN Training

Search-Based Auto-Parallelizer





FlexFlow

TASO/MetaFlow

Unity



FlexFlow

TASO/MetaFlow

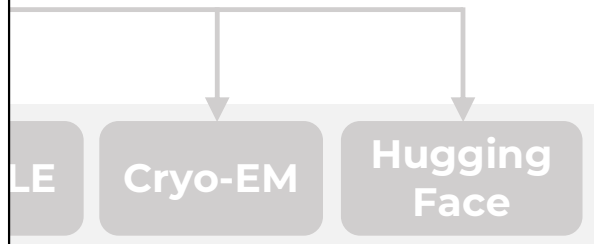
Unity

Unity

OSDI '22

Accelerating DNN Training Through **Joint Optimization**
of Algebraic Transformations and Parallelization

Colin Unger*, Zhihao Jia*, Wei Wu, Sina Lin, Mandeep Baines, Carlos Efrain Quintero Narvaez, Vinay Ramakrishnaiah, Nirmal Prajapati, Pat McCormick, Jamaludin Mohd-Yusof, Xi Luo, Dheevatsa Mudigere, Jongsoo Park, Misha Smelyanskiy, Alex Aiken



FlexFlow

TASO/MetaFlow

Unity



**Memory
Optimization**

**Generalized
Pipeline
Parallelism**

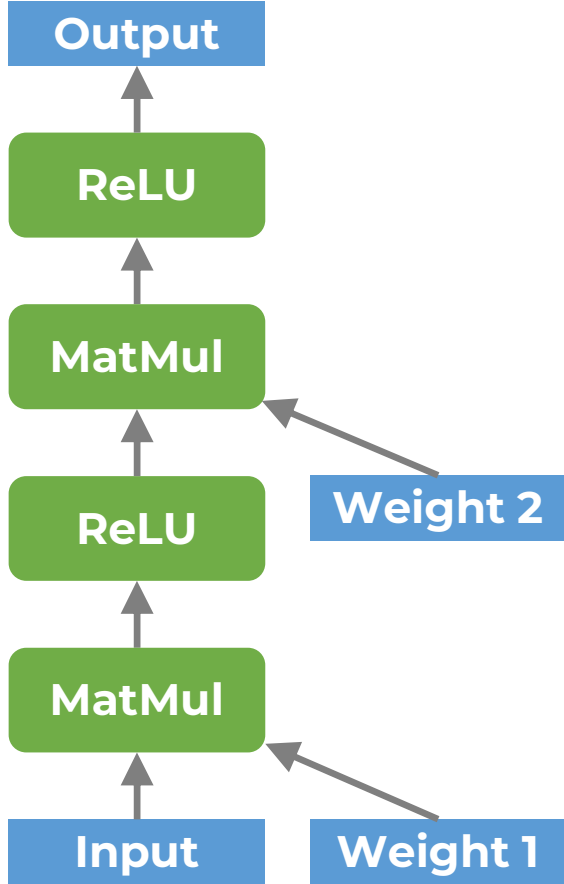
**MoE
Inference**

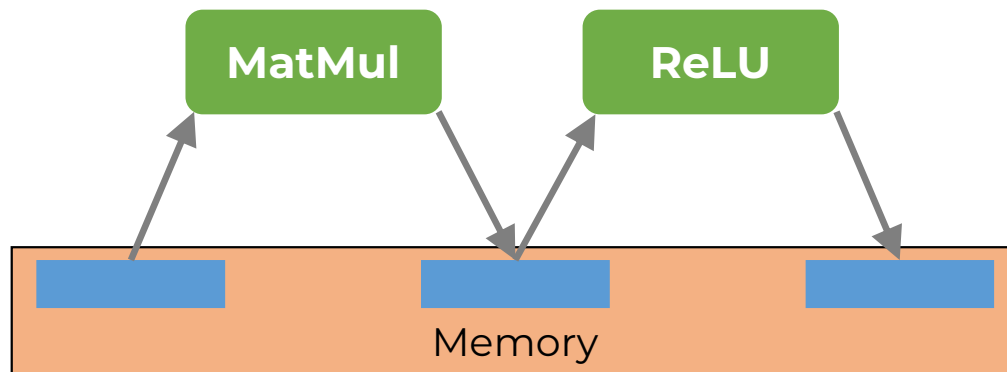
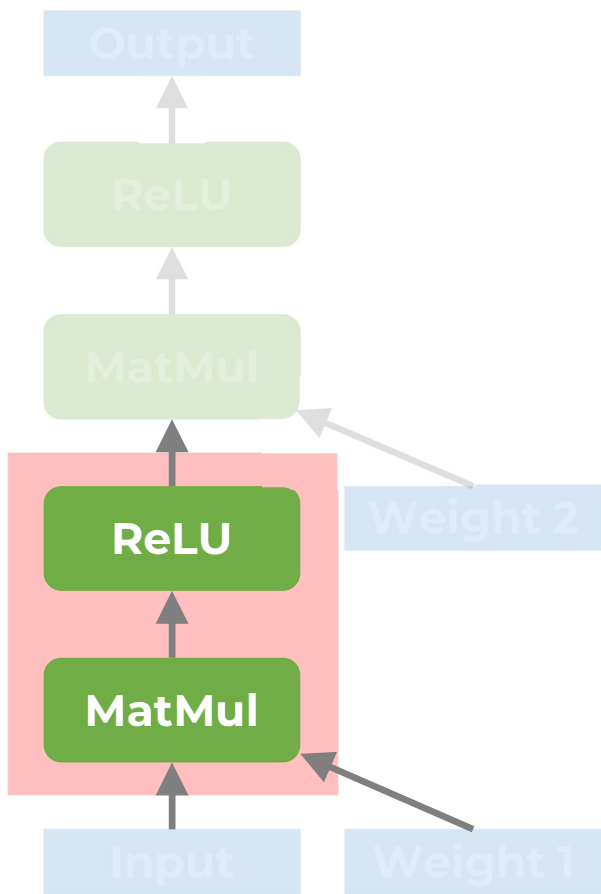
**JAX
Interface**

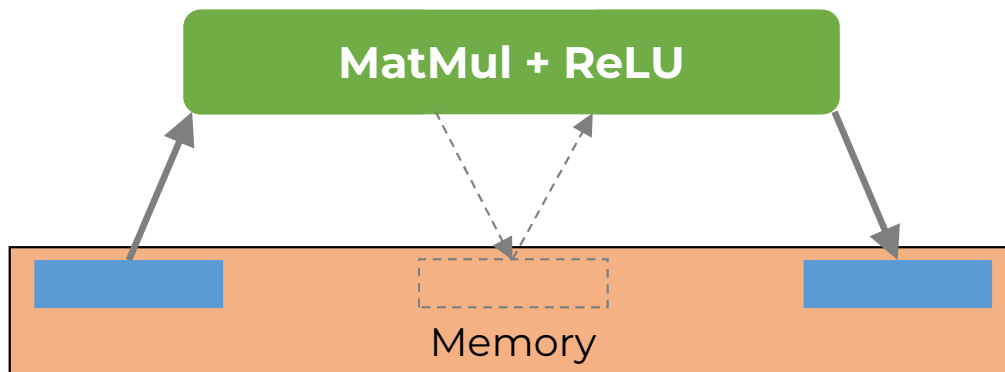
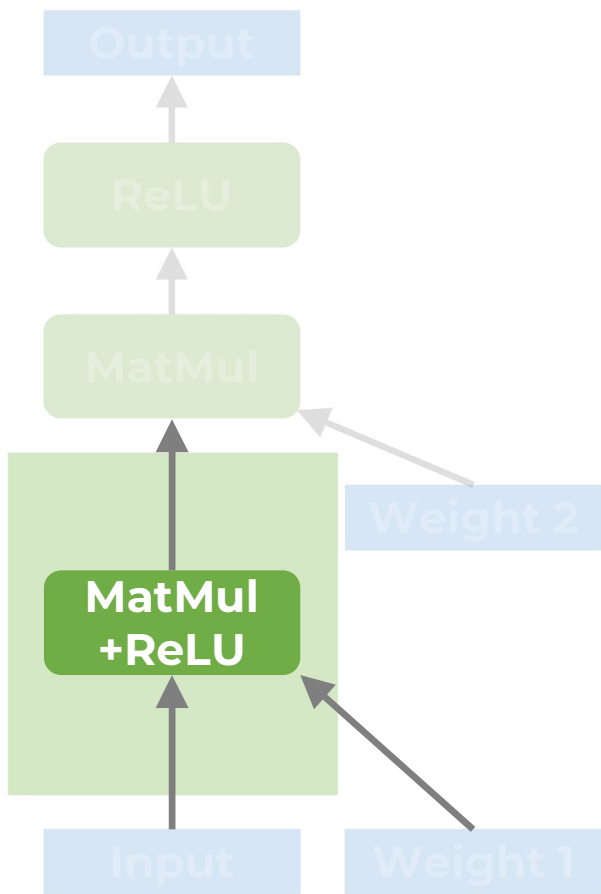
ECP-CANDLE

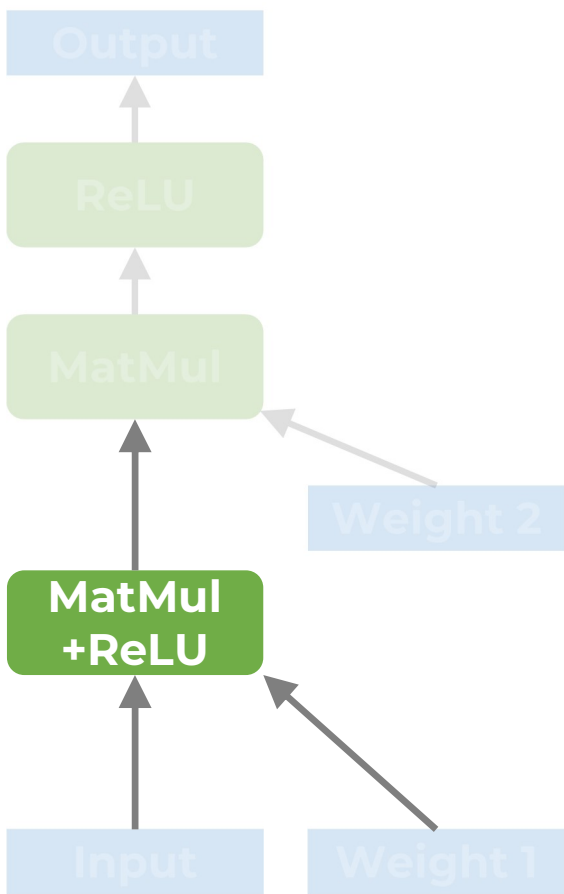
Cryo-EM

**Hugging
Face**

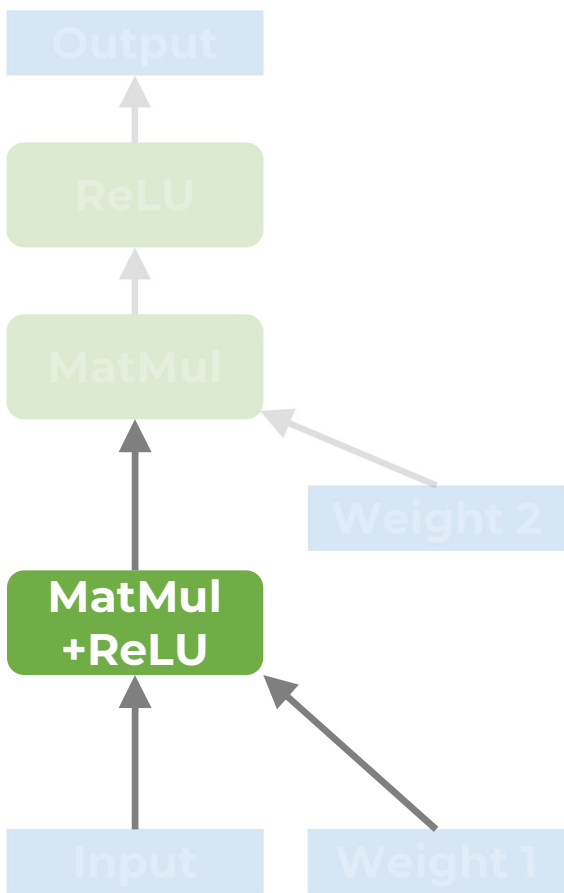




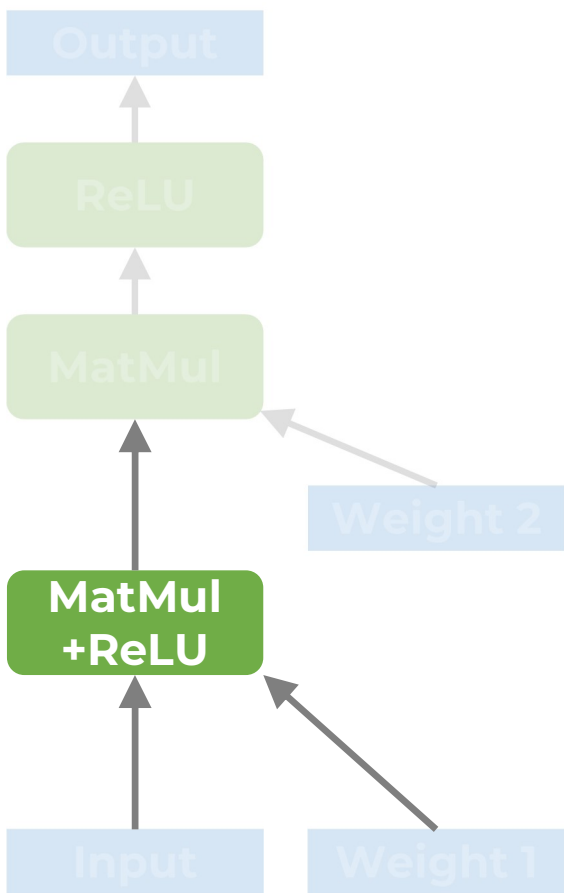




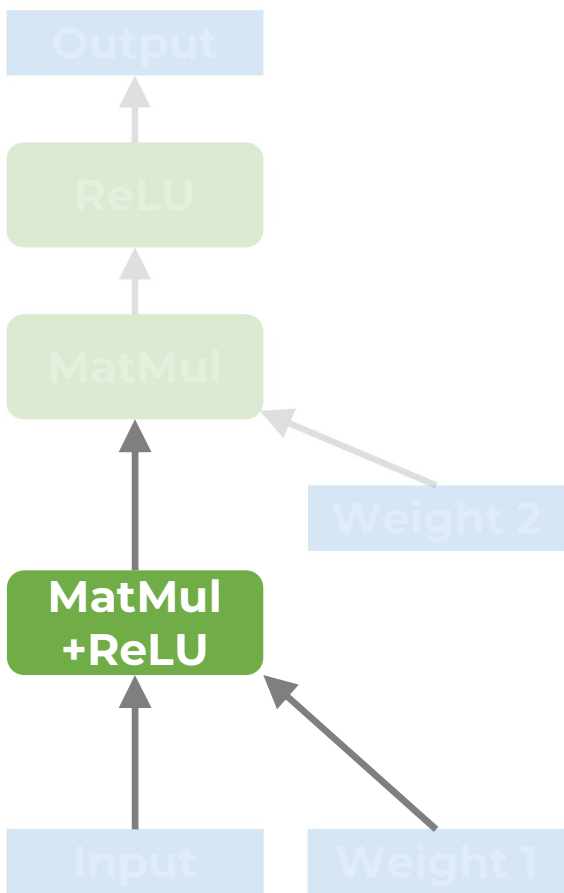
Operator Fusion



Operator Fusion
Operator Splitting



Operator Fusion
Operator Splitting
Operator Reordering



Operator Fusion
Operator Splitting
Operator Reordering
...

Operator Fusion
Operator Splitting
Operator Reordering
...

“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

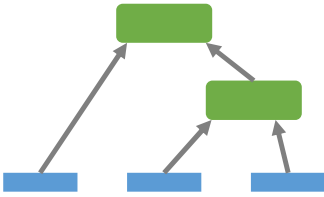
MetaFlow [MLSys 19]

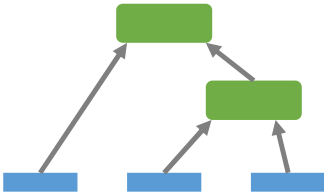
TASO [SOSP 19]

PET [OSDI 21]

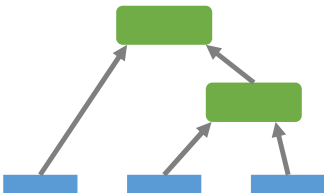
Tensat [MLSys 21]

...



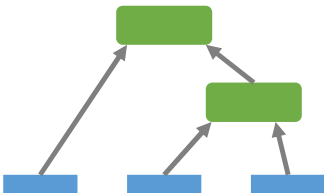


Auto-Parallelization



**Auto-
Parallelization**

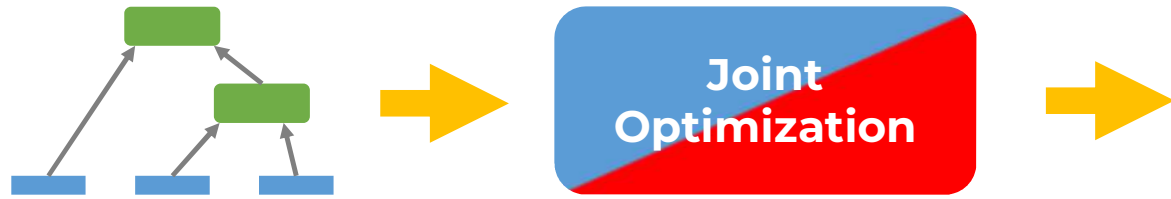
**Algebraic
Optimizer**

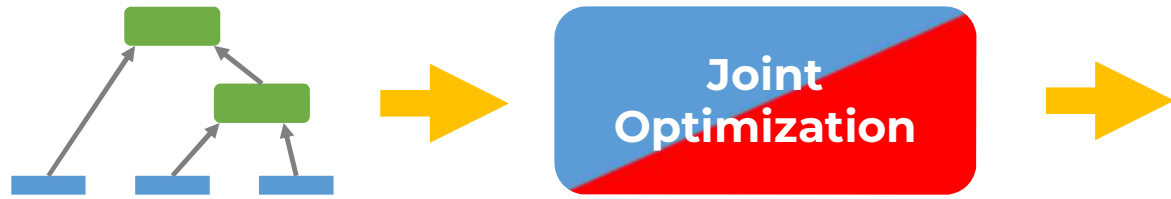


**Auto-
Parallelization**

**Algebraic
Optimizer**

?





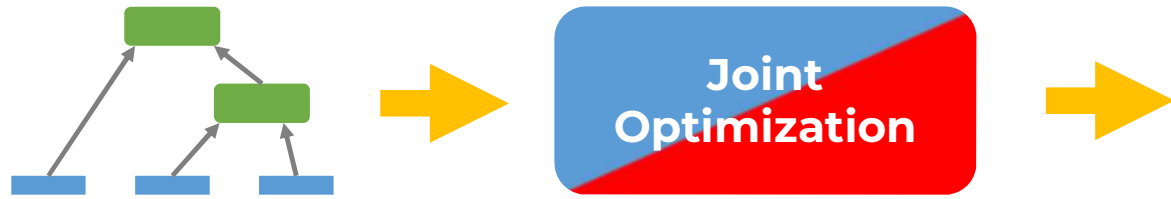
1.

2.



1. Representation

2.



1. Representation
2. Scalability

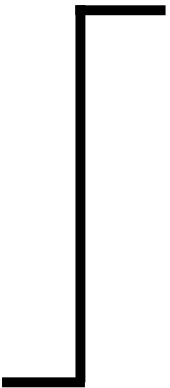
Unity

Unity — Representation

Representation

Parallel Computation
Graph (PCG)

Unity

A diagram consisting of a horizontal line extending from the right side of the word 'Unity', followed by a vertical line segment that turns right to point towards the text 'Parallel Computation Graph (PCG)'.

Representation

Parallel Computation
Graph (PCG)

Unity

Scalability

Representation

Parallel Computation
Graph (PCG)

Unity

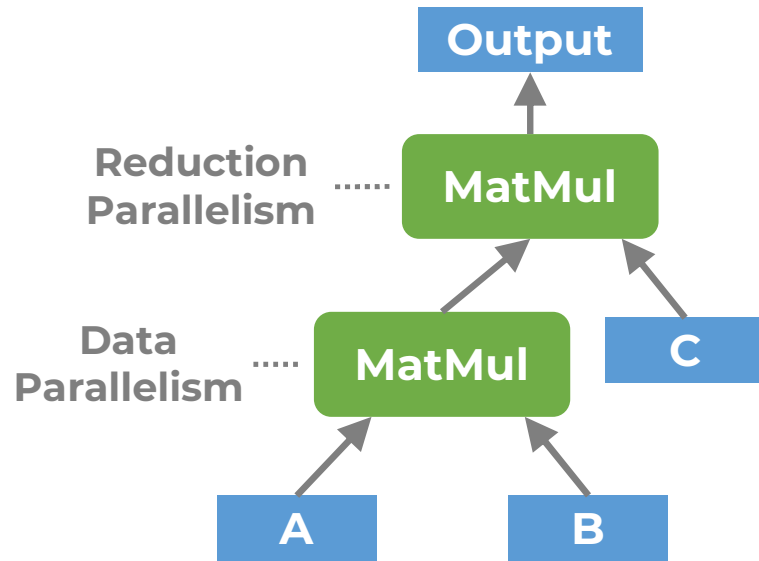
Scalability

Hierarchical Search
Algorithm

Representation

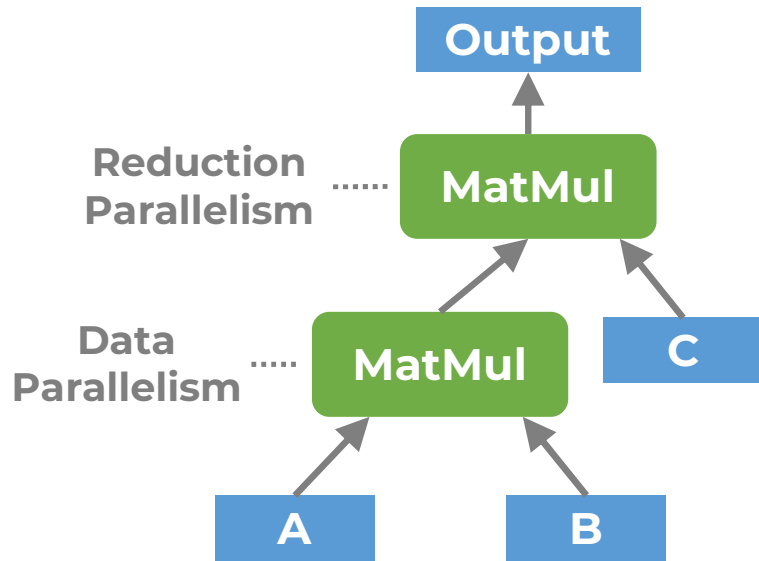
Parallel Computation Graph (PCG)

Parallel Computation Graph (PCG)

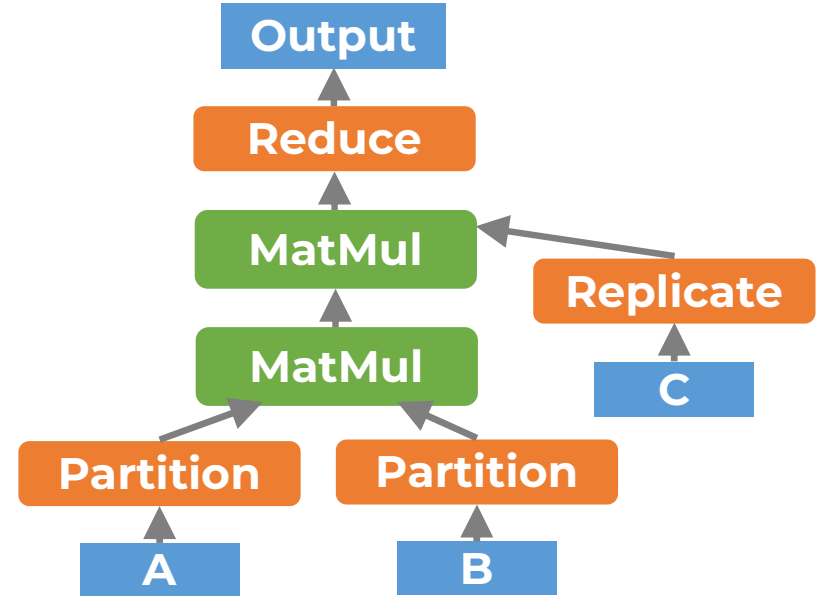


annotated computation graph

Parallel Computation Graph (PCG)

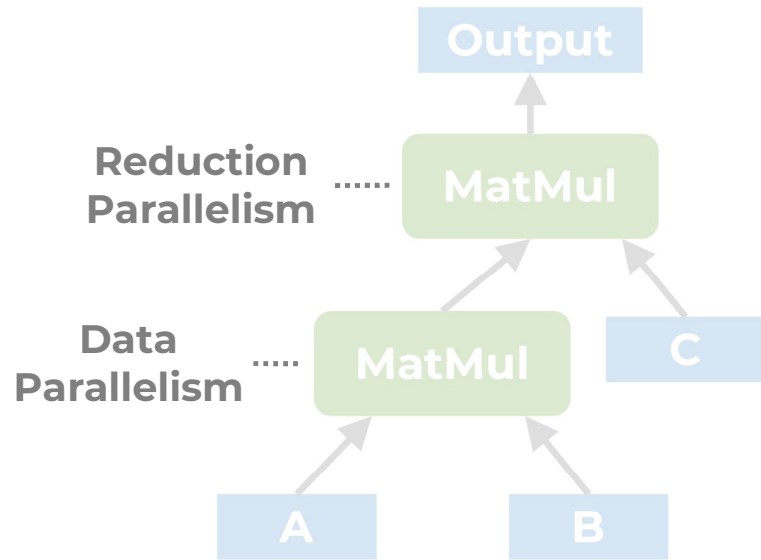


annotated computation graph

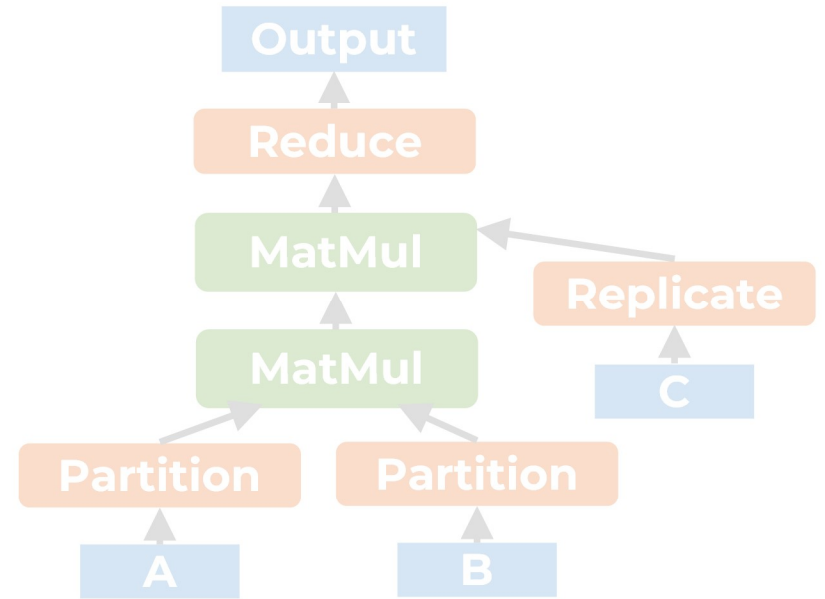
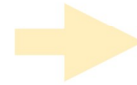


parallel computation graph (PCG)

Parallel Computation Graph (PCG)

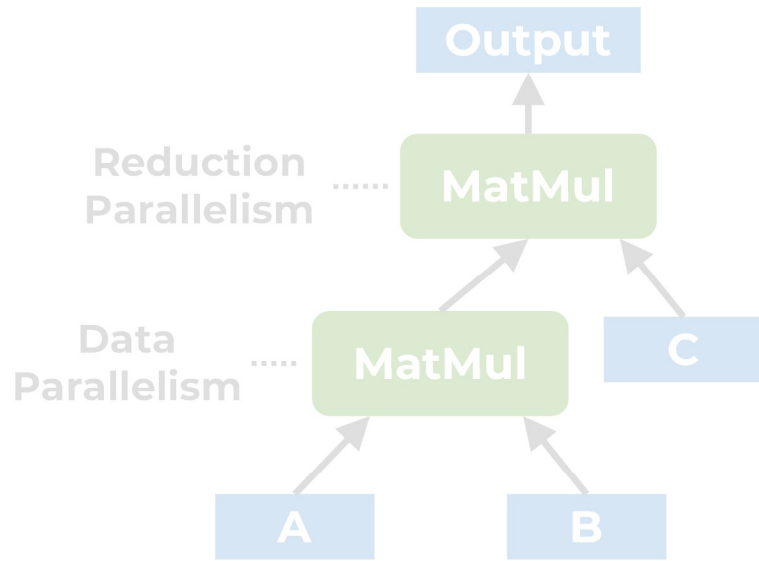


annotated computation graph

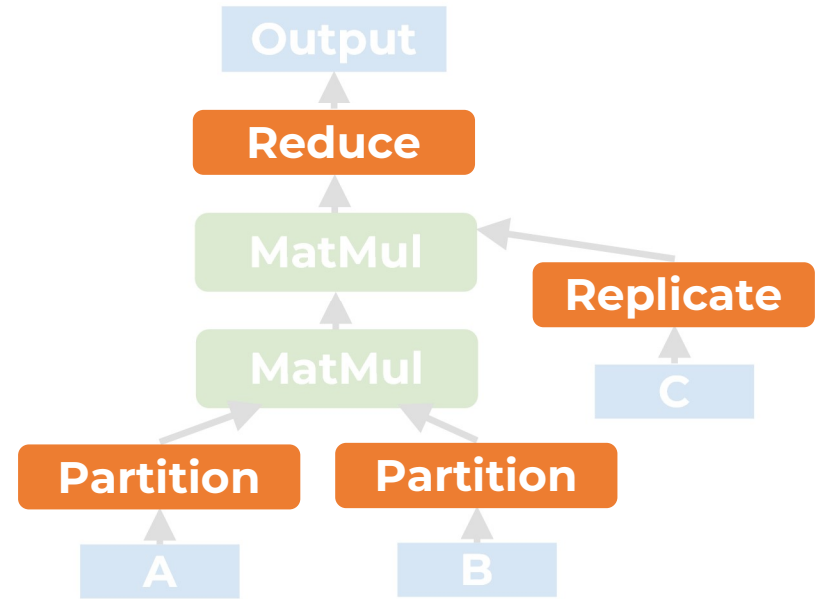
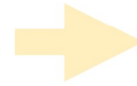


parallel computation graph (PCG)

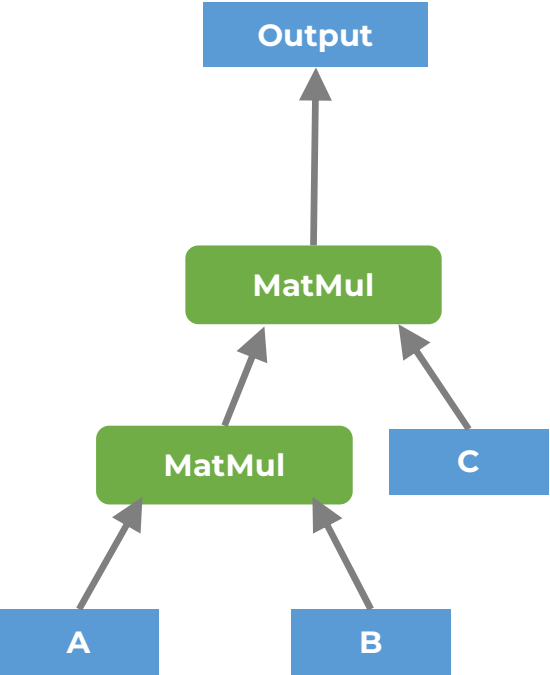
Parallel Computation Graph (PCG)

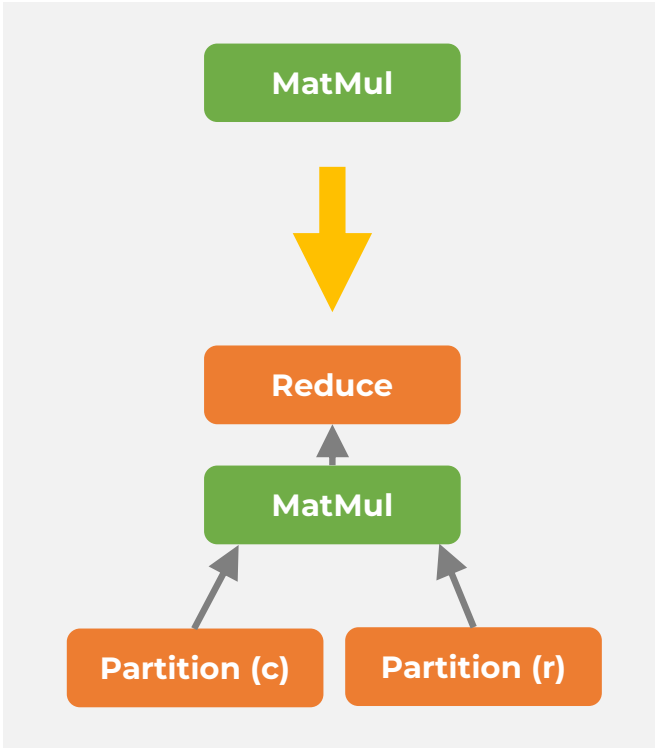
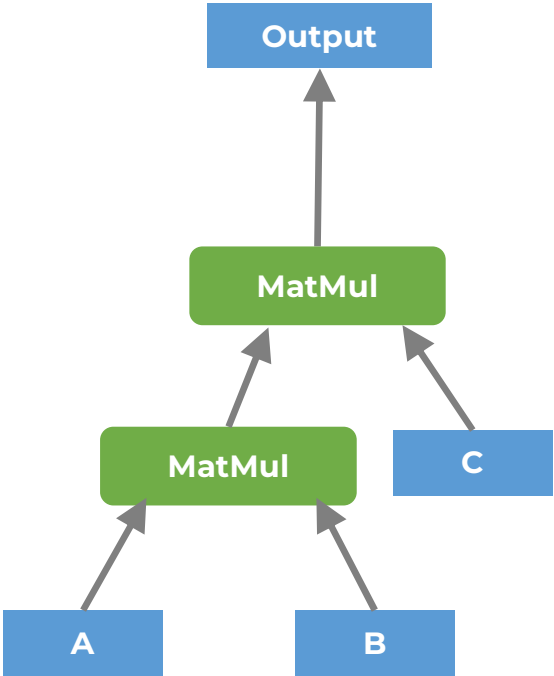


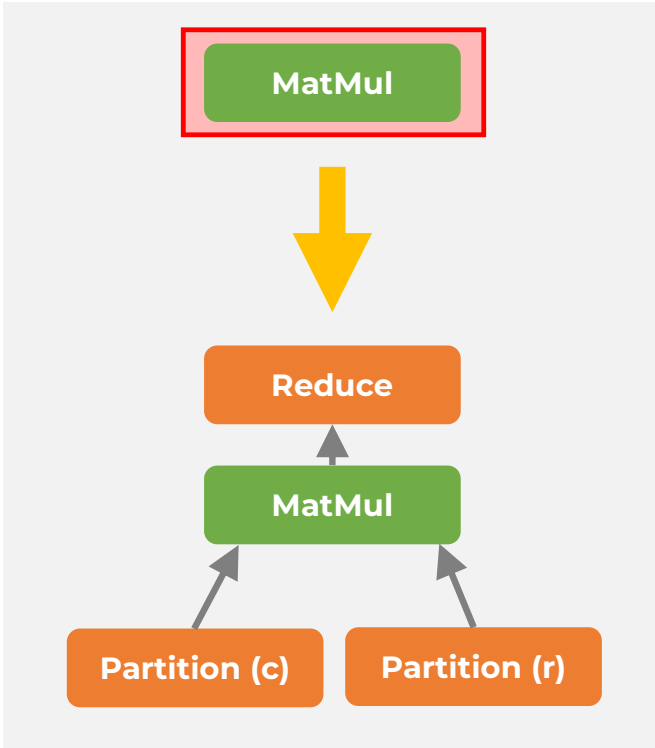
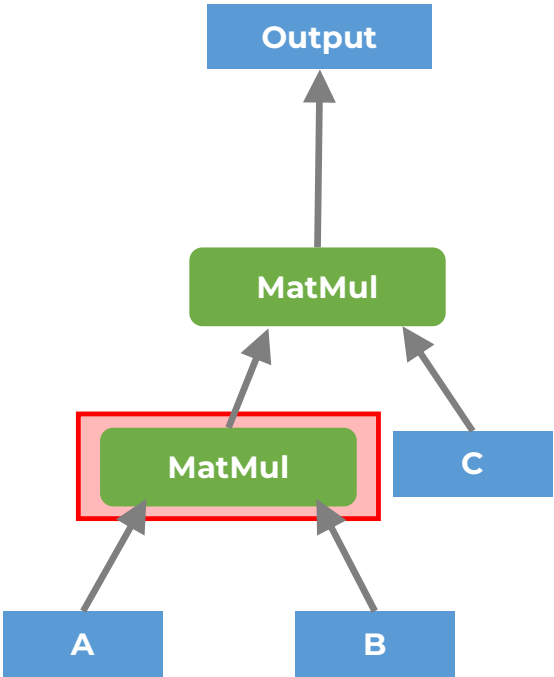
annotated computation graph

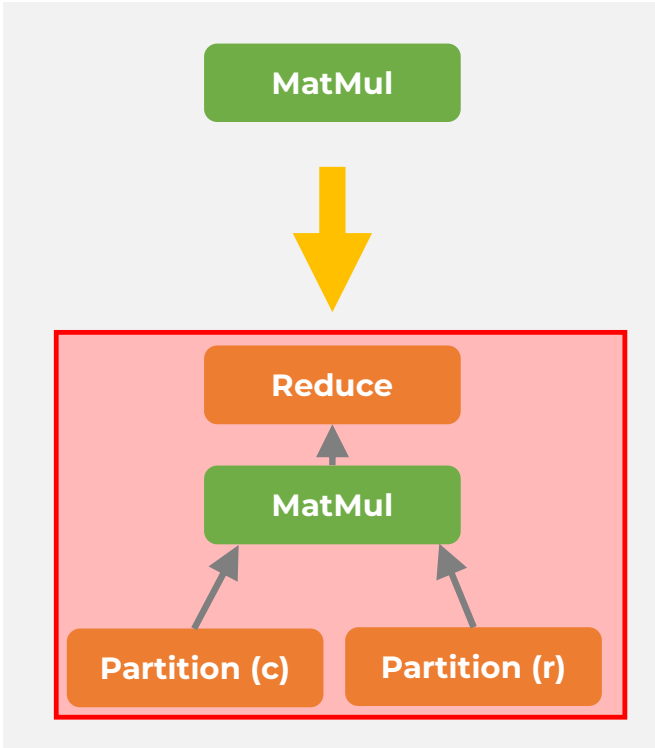
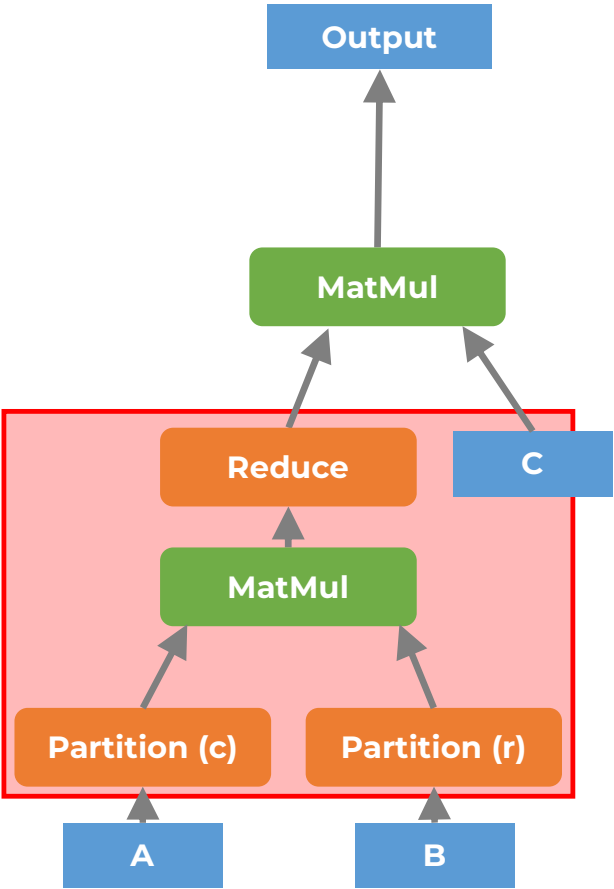


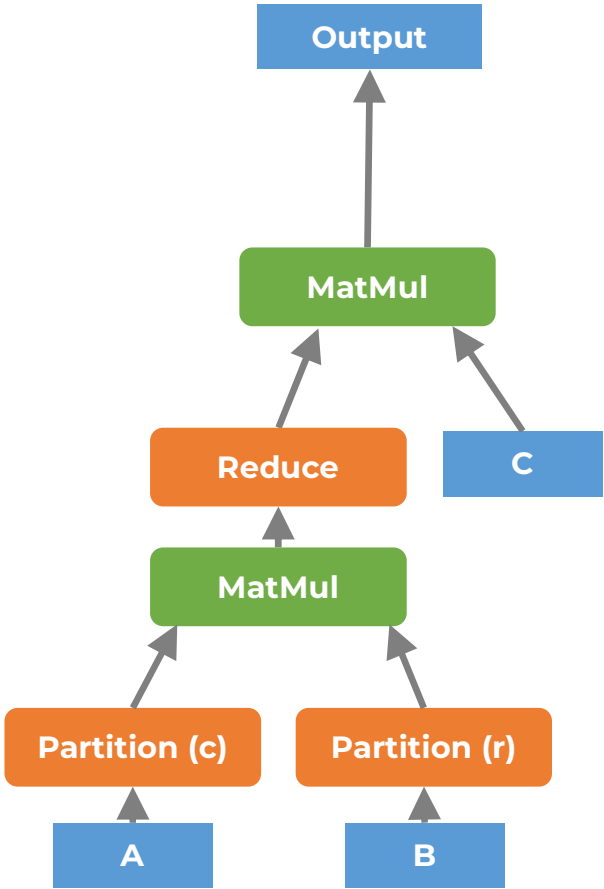
parallel computation graph (PCG)

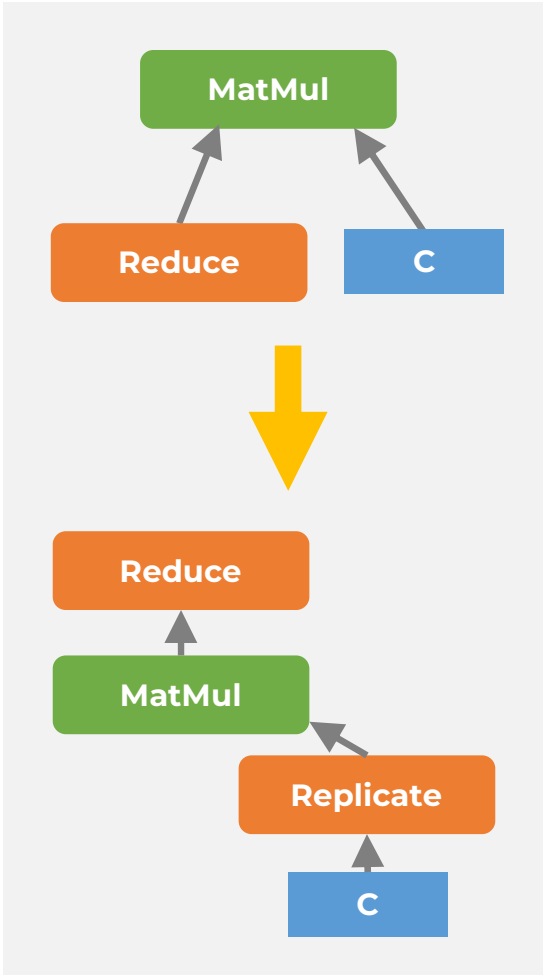
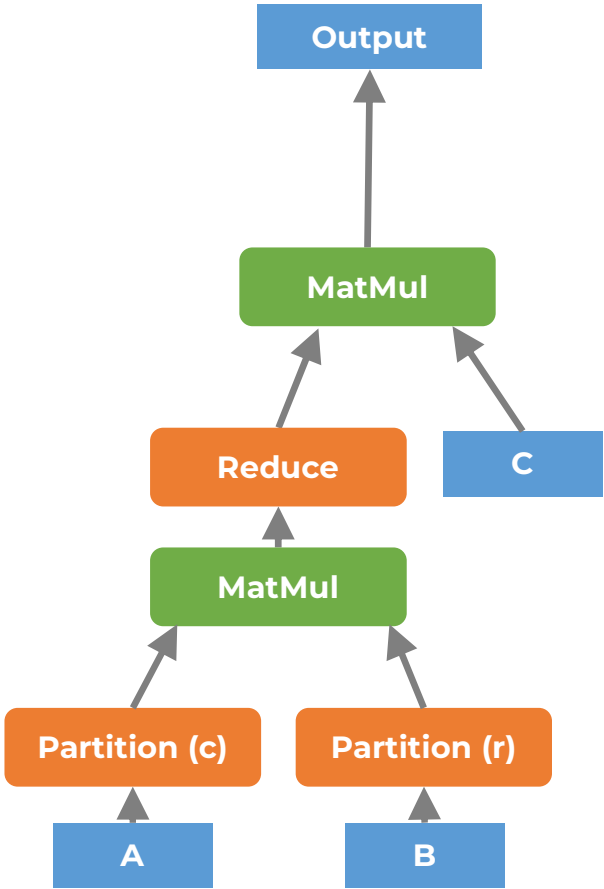


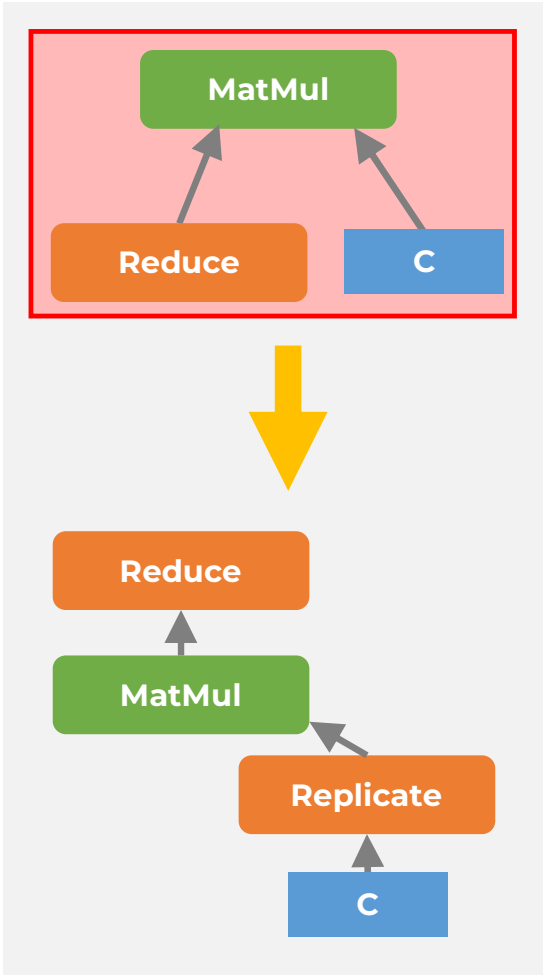
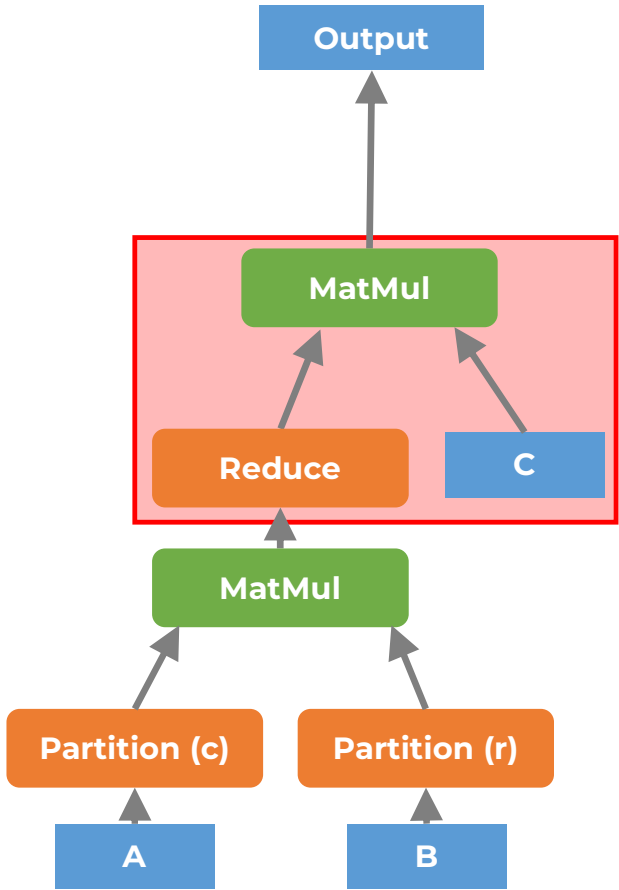


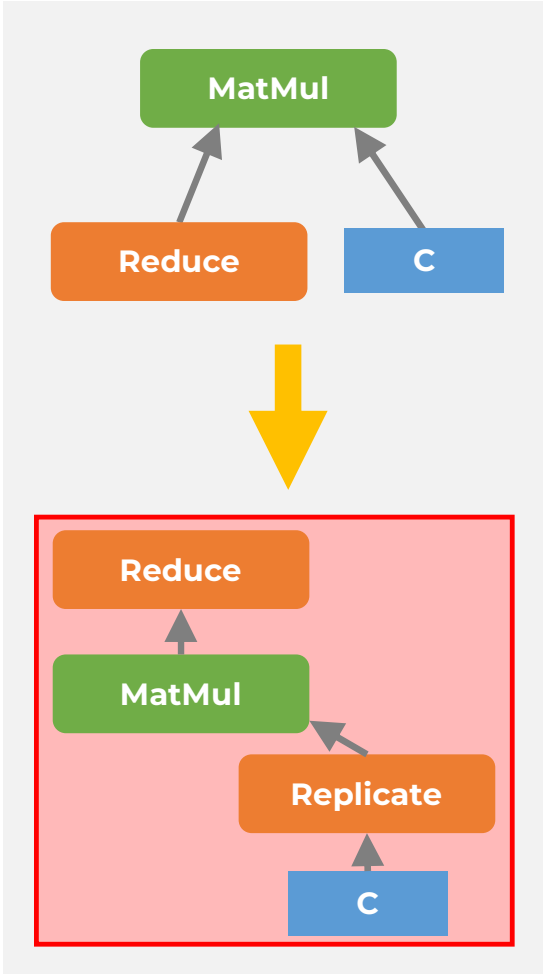
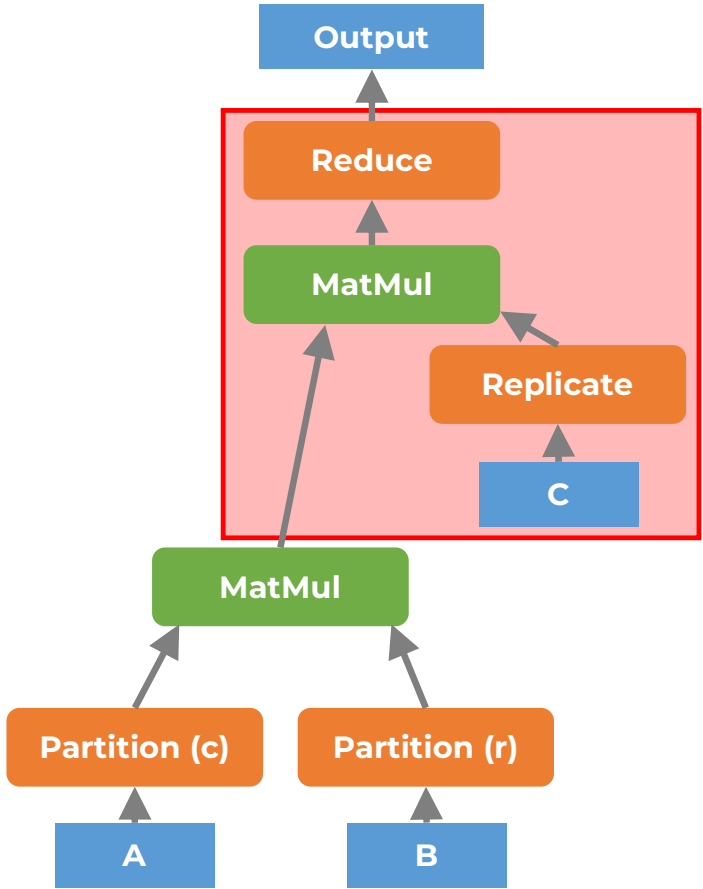


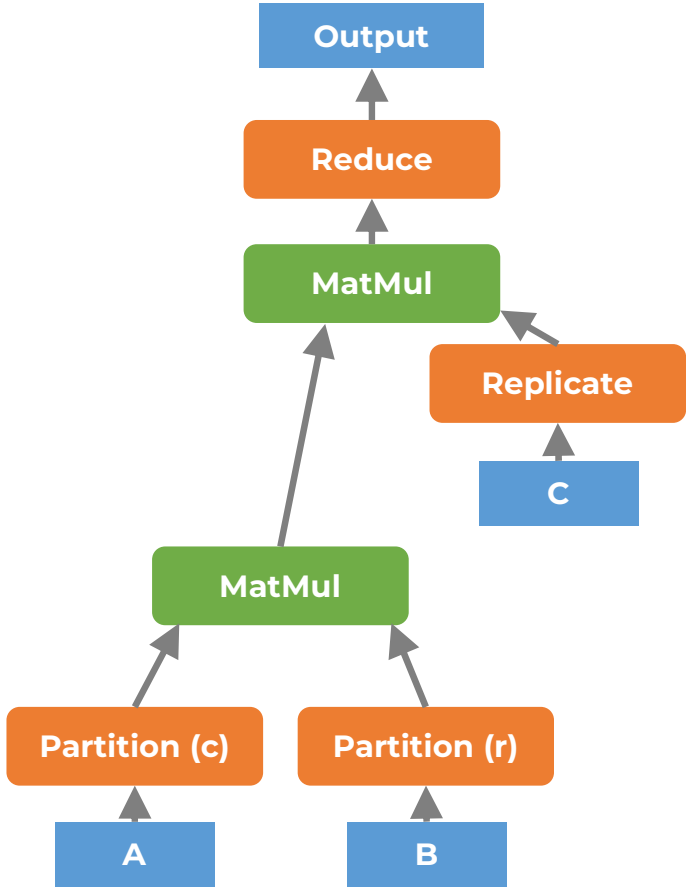












Hierarchical Search Algorithm

Hierarchical Search Algorithm

Algebraic Transformation

Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

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

Backtracking
Search

Parallelism Type



Parallelism Degree

Device Mapping

Hierarchical Search Algorithm

Algebraic Transformation

Backtracking
Search

Parallelism Type

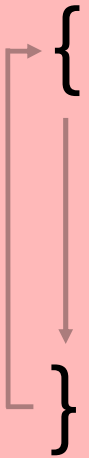


Parallelism Degree

Device Mapping



Dynamic
Programming



Models

BERT-Large

(Language Modeling)

Candle-UNO

(Precision Medicine)

MLP

(Regression)

DLRM

XDL

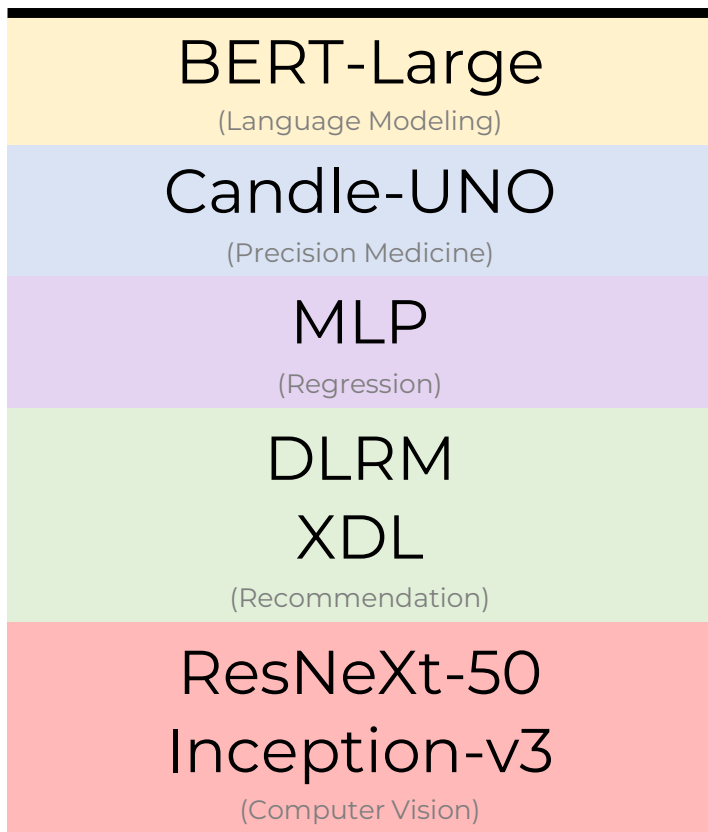
(Recommendation)

ResNeXt-50

Inception-v3

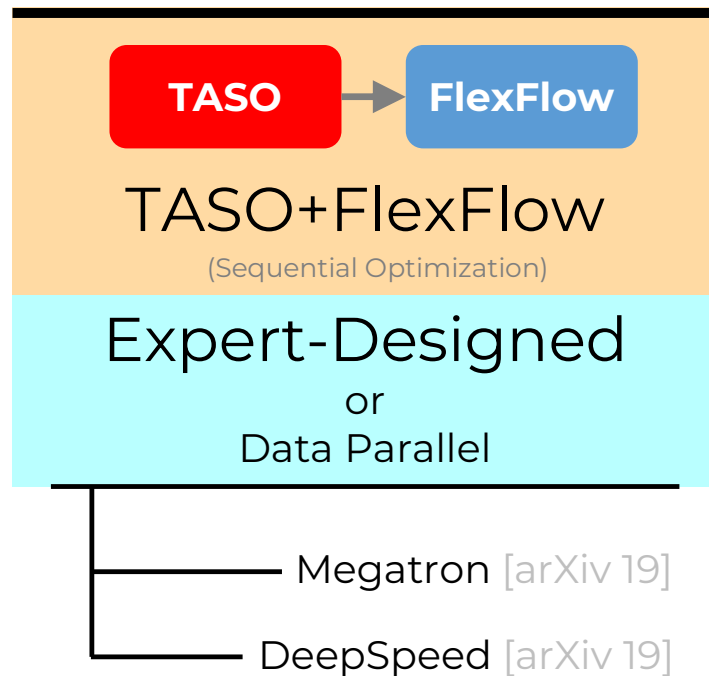
(Computer Vision)

Models

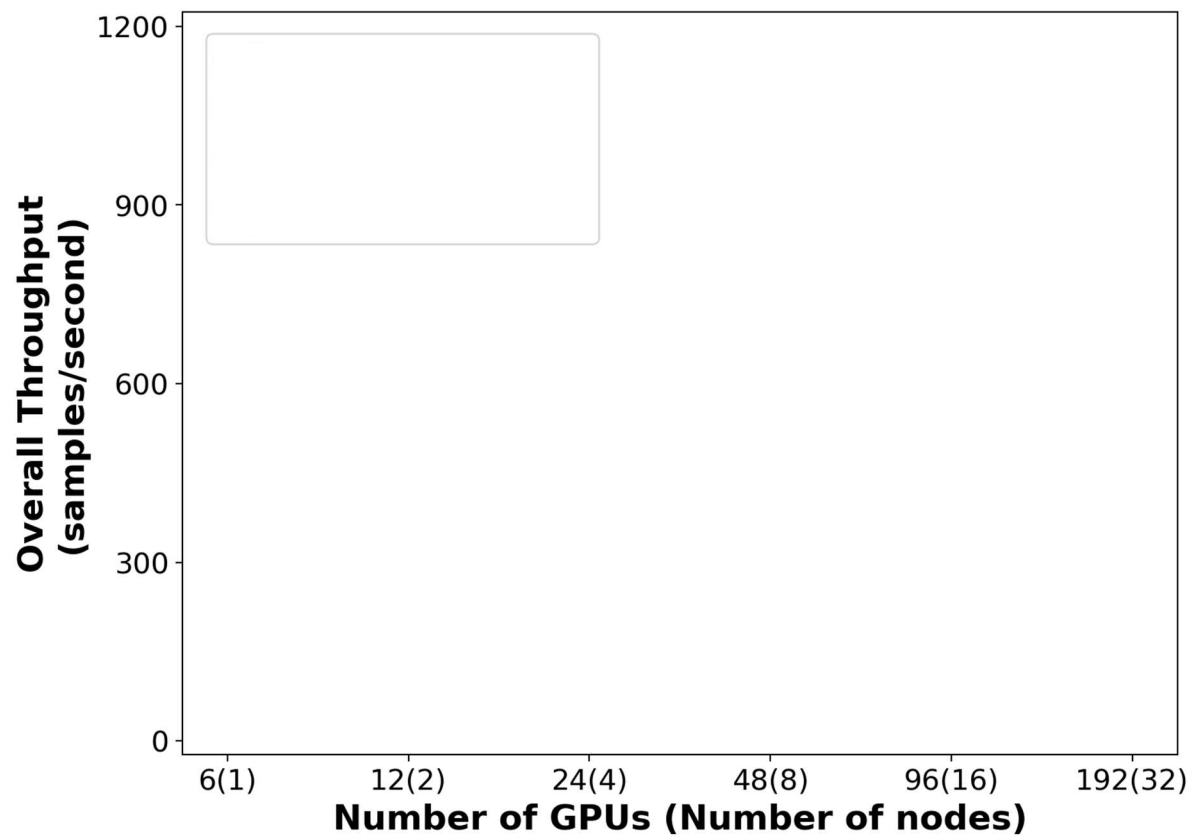


×

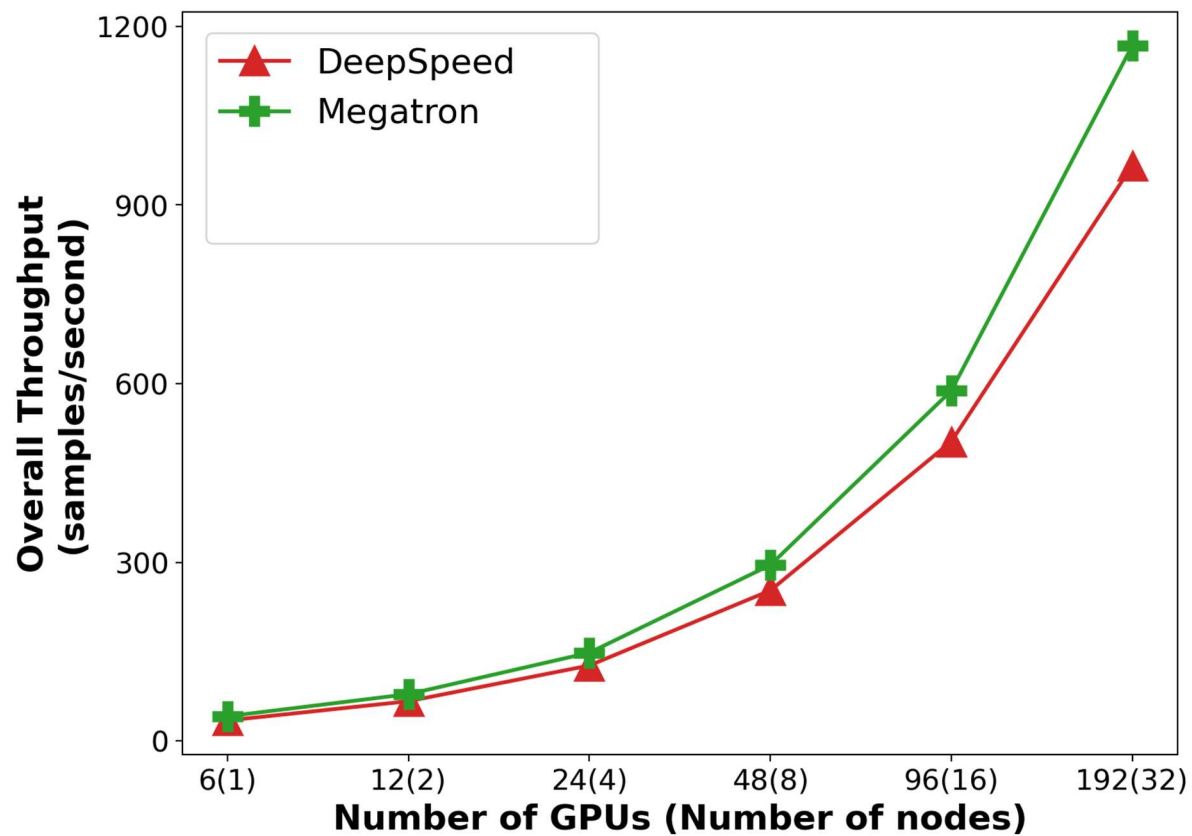
Baselines



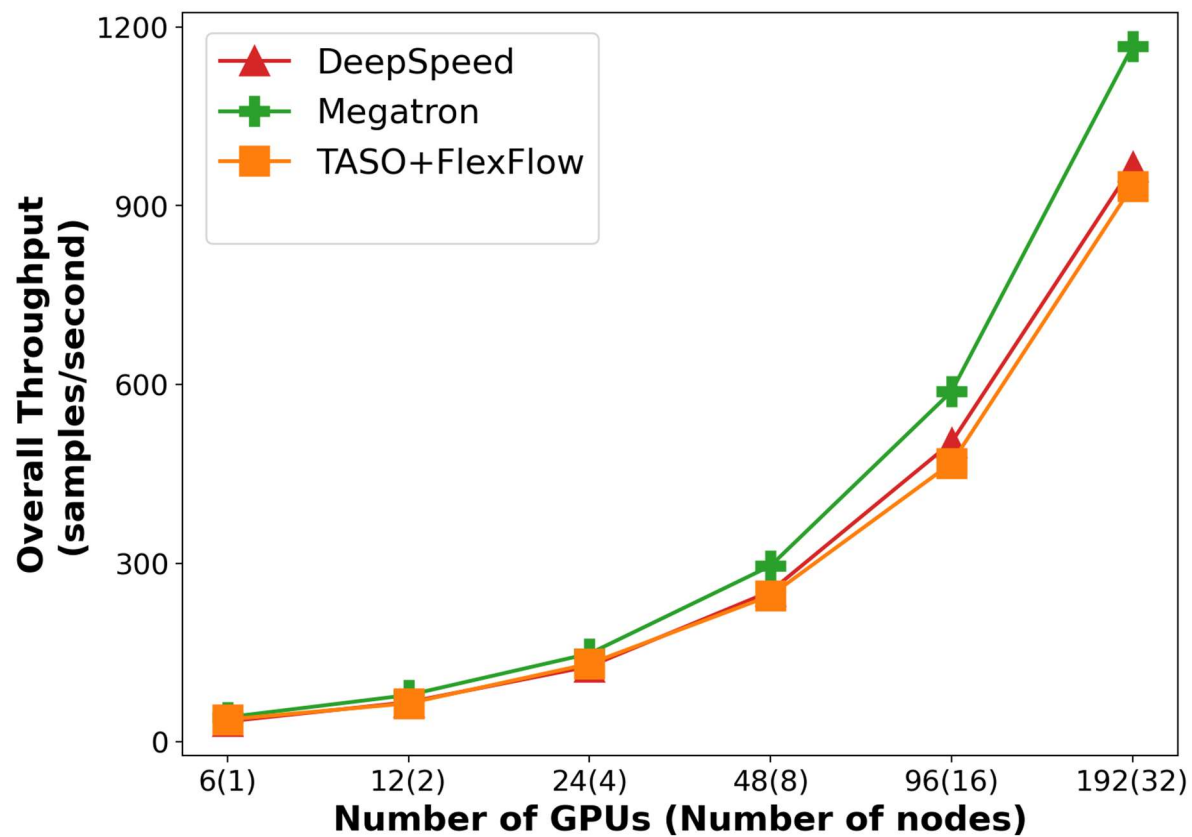
BERT-Large



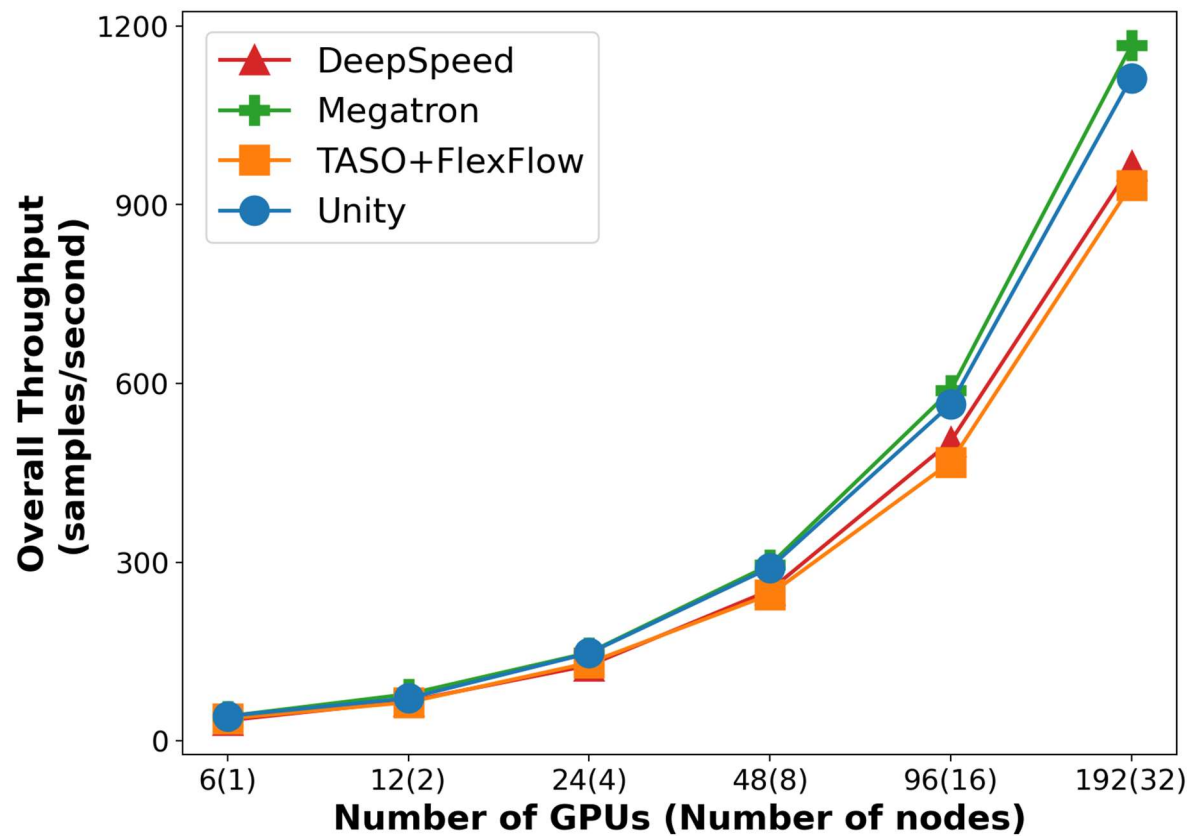
BERT-Large



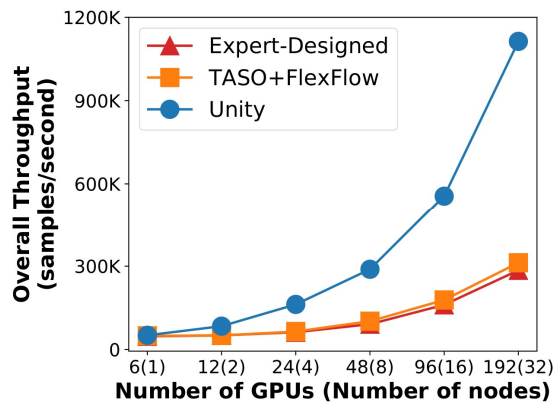
BERT-Large



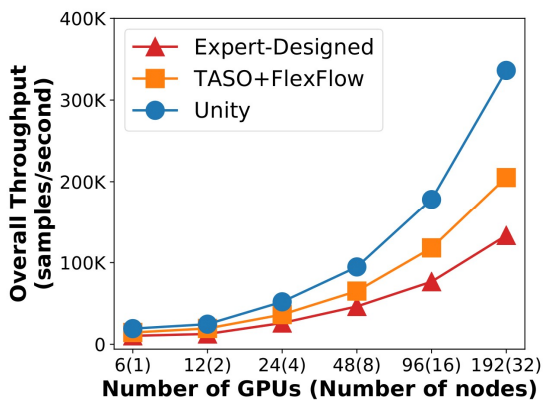
BERT-Large



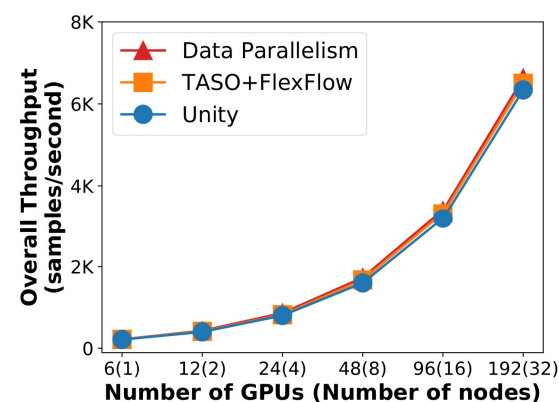
DLRM



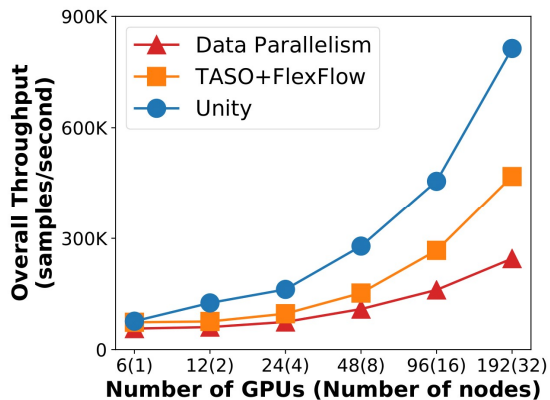
CANDLE-Uno



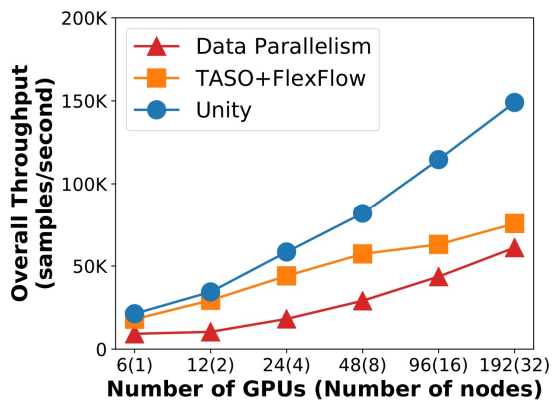
ResNeXt-50



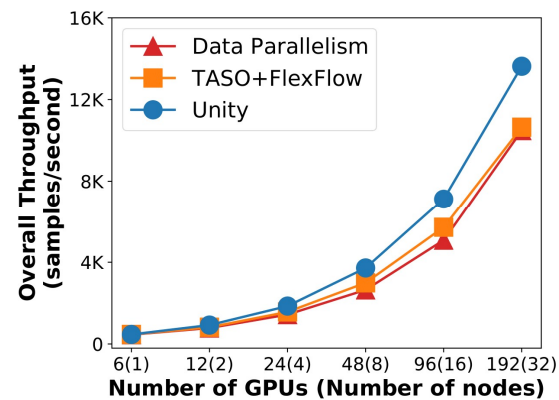
XDL

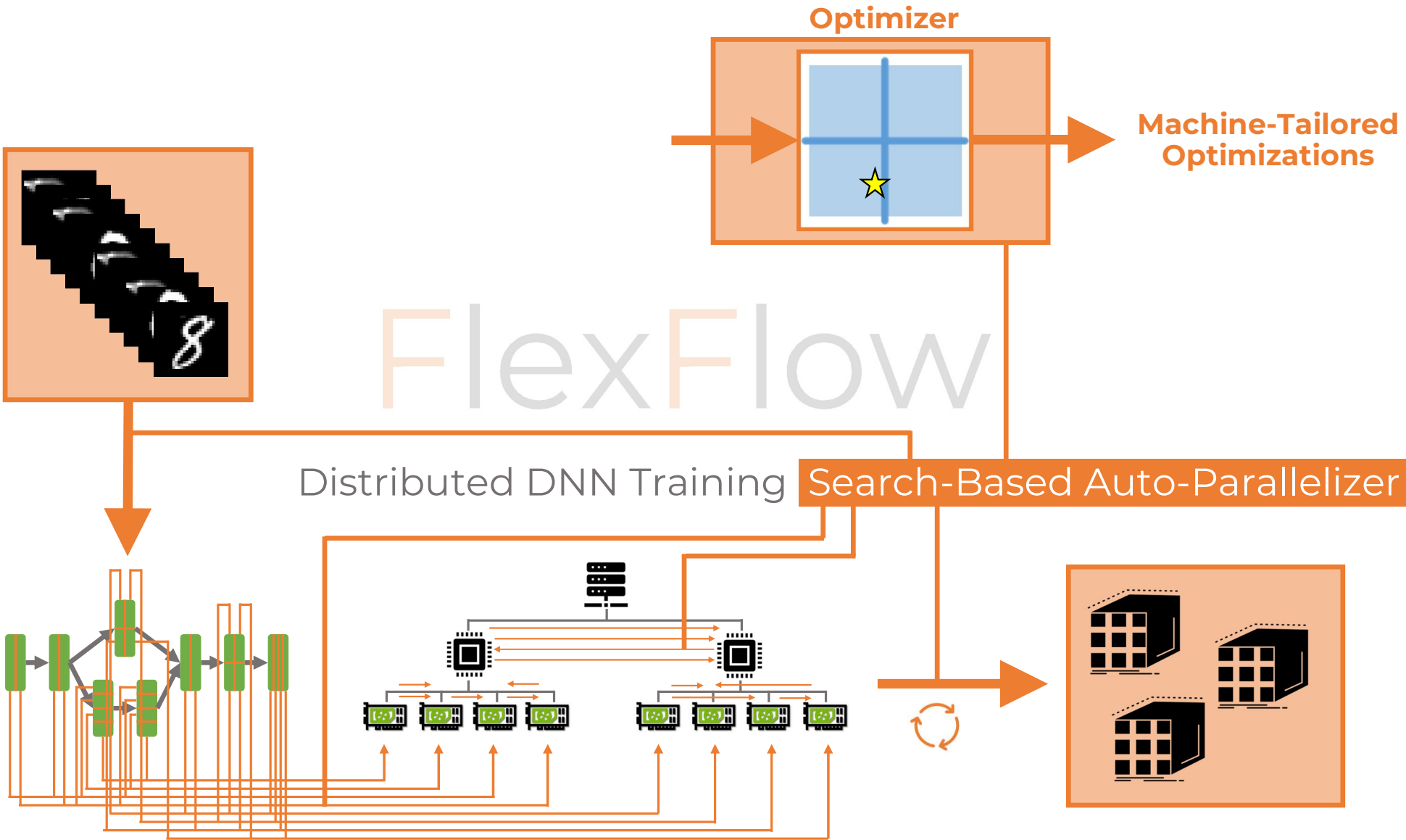


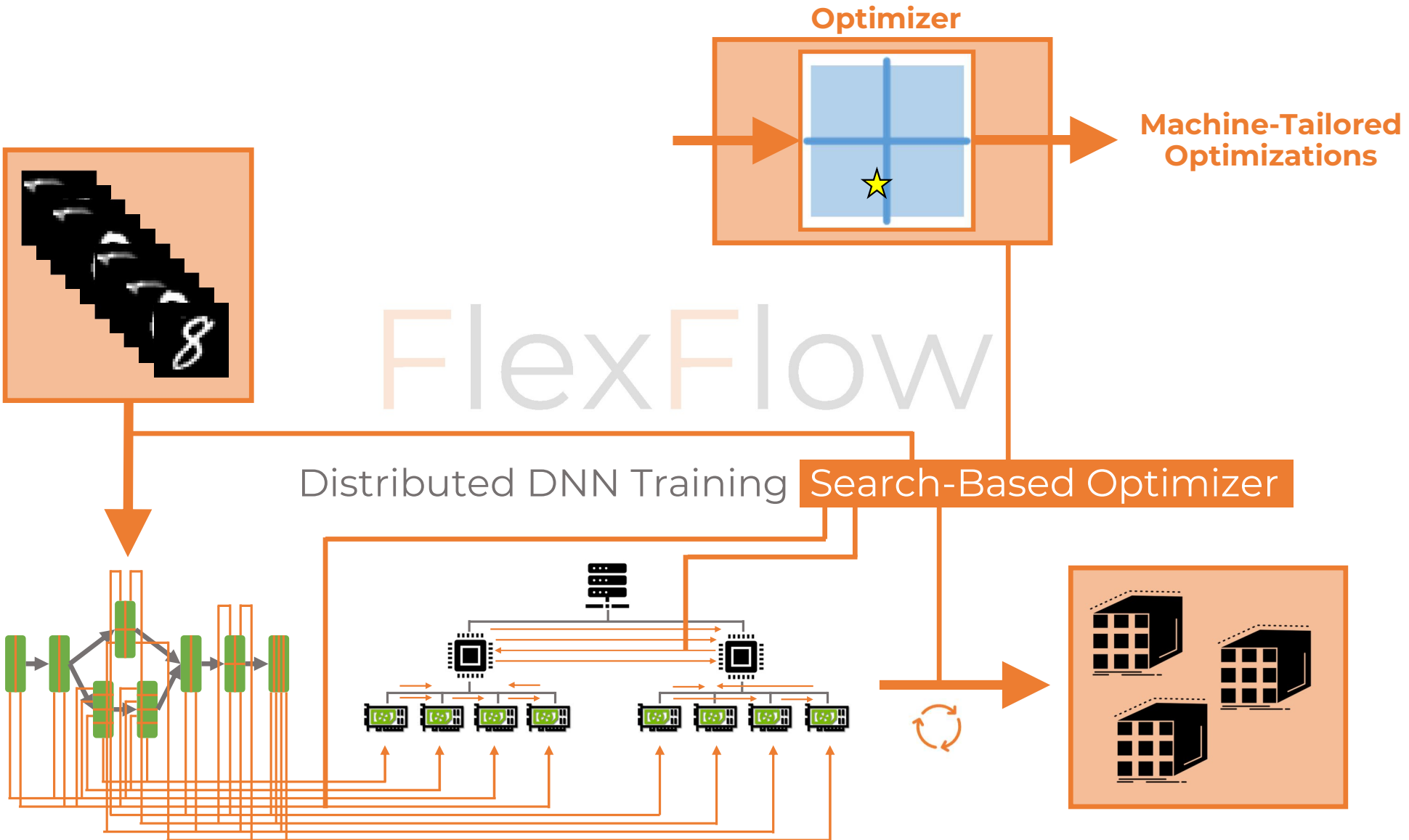
MLP



Inception-v3







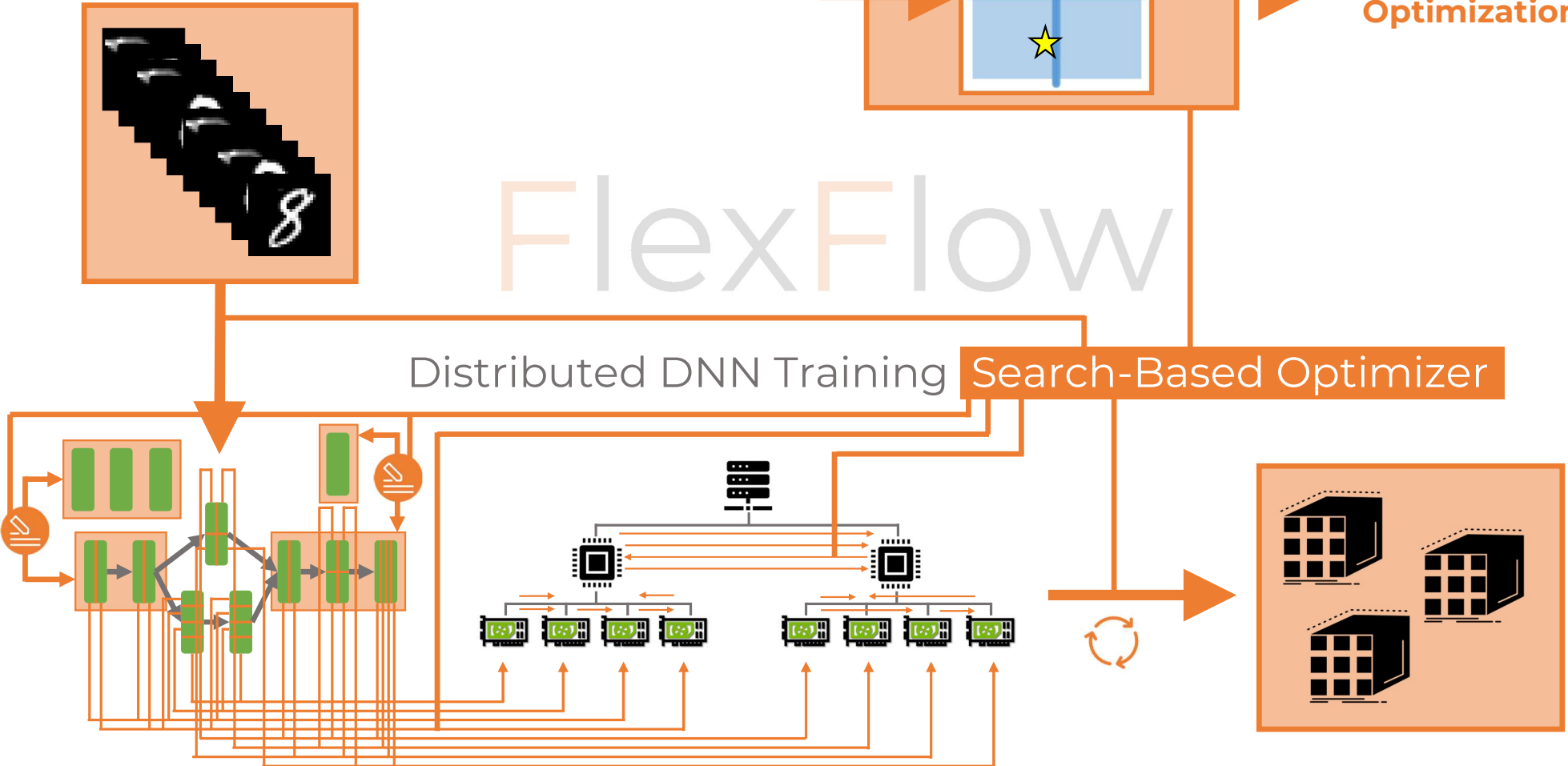
Optimizer

Machine-Tailored
Optimizations

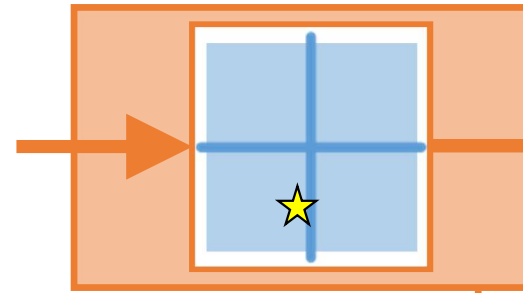
FlexFlow

Distributed DNN Training

Search-Based Optimizer



Optimizer

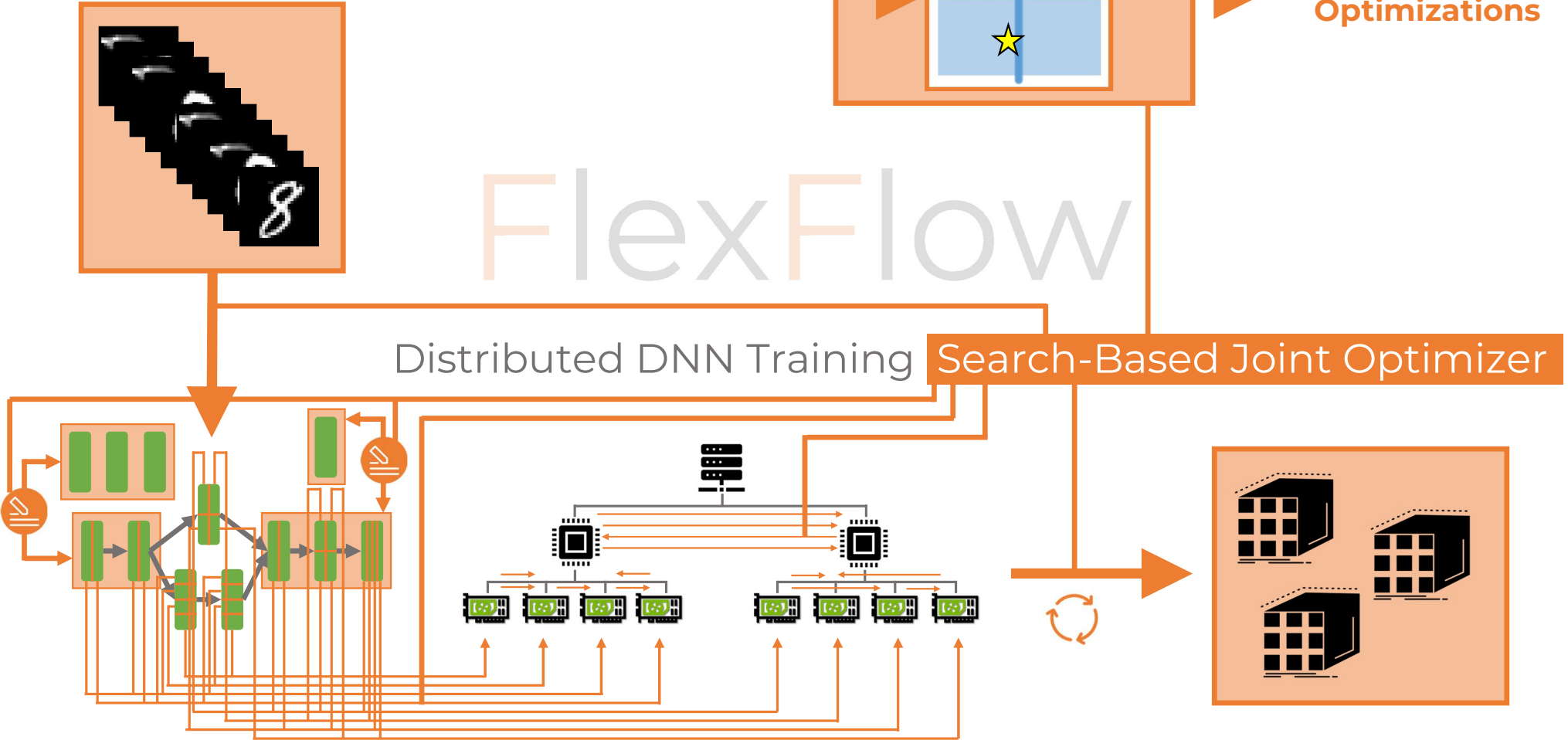


Machine-Tailored Optimizations

FlexFlow

Distributed DNN Training

Search-Based Joint Optimizer



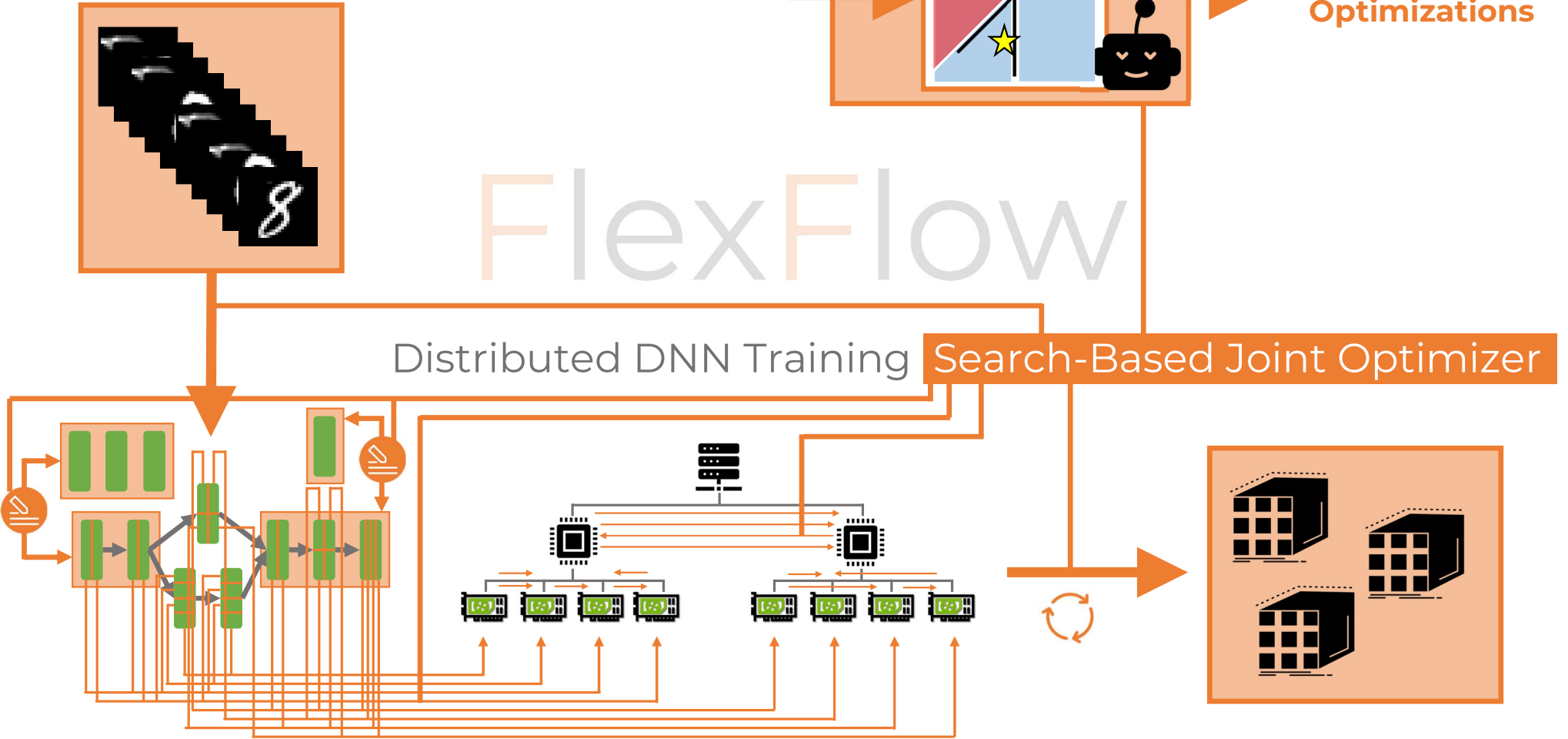
Optimizer

Machine-Tailored
Optimizations

FlexFlow

Distributed DNN Training

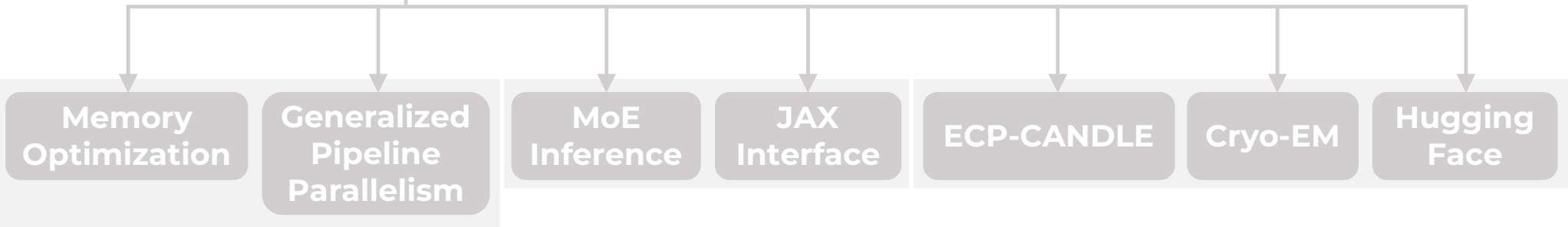
Search-Based Joint Optimizer



FlexFlow

TASO/MetaFlow

Unity



FlexFlow

TASO/MetaFlow

Unity



**Memory
Optimization**

**Generalized
Pipeline
Parallelism**

**MoE
Inference**

**JAX
Interface**

ECP-CANDLE

Cryo-EM

**Hugging
Face**

FlexFlow

TASO/MetaFlow

Unity



FlexFlow

TASO/MetaFlow

Unity

Memory Optimization

Generalized Pipeline Parallelism

MoE Inference

JAX Interface

ECP-CANDLE

Cryo-EM

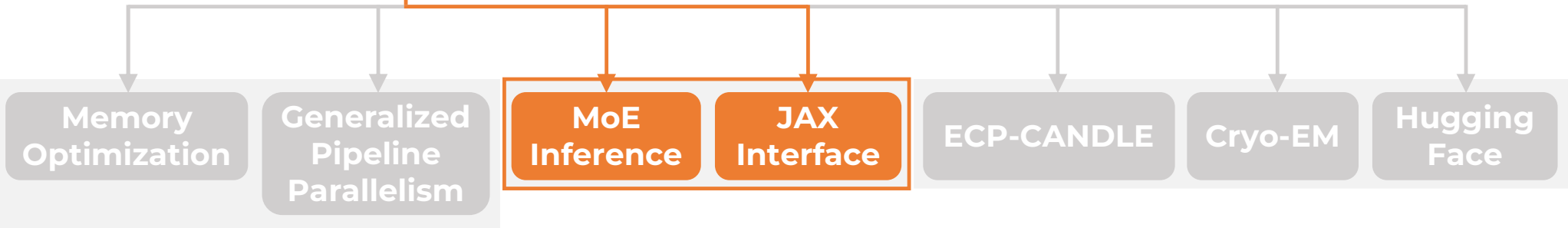
Hugging Face

Extending Unity

FlexFlow

TASO/MetaFlow

Unity

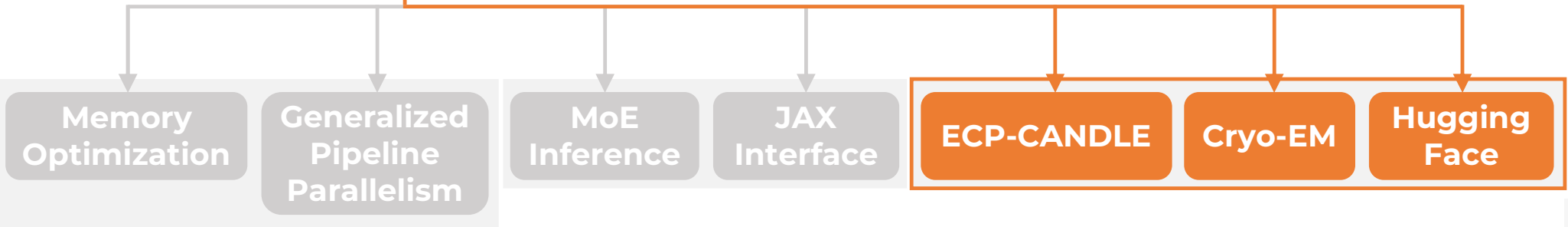


New Features

FlexFlow

TASO/MetaFlow

Unity

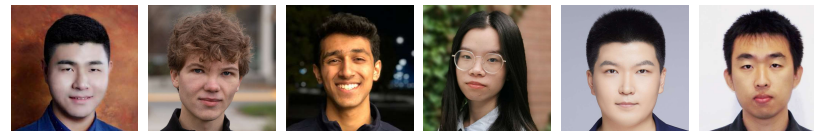


Applications

**Algebraic
Transformations**

Parallelization

Memory Constraints



**Memory
Optimization**

**Generalized
Pipeline
Parallelism**

**MoE
Inference**

**JAX
Interface**

ECP-CANDLE

Cryo-EM

**Hugging
Face**

Extending Unity

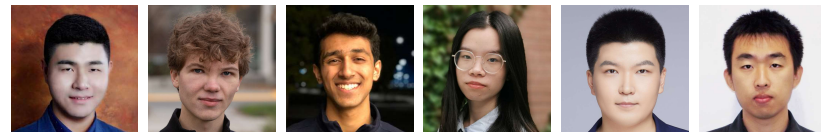
Algebraic
Transformations

Parallelization

Memory Constraints



Maximizes throughput *but can exceed memory capacity*



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

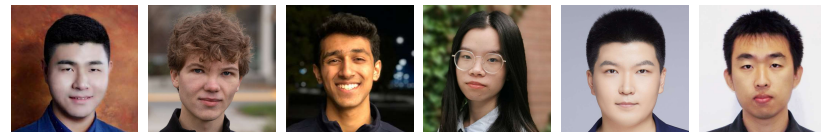
Extending Unity

**Algebraic
Transformations**

Parallelization

**Memory
Optimizations**

Maximizes throughput *but can exceed memory capacity*



**Memory
Optimization**

**Generalized
Pipeline
Parallelism**

**MoE
Inference**

**JAX
Interface**

ECP-CANDLE

Cryo-EM

**Hugging
Face**

Extending Unity

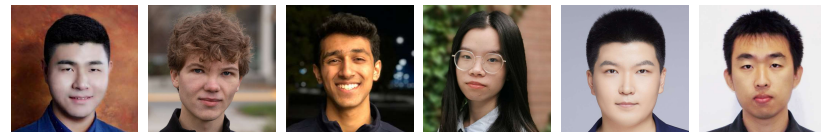
**Algebraic
Transformations**

Parallelization

**Memory
Optimizations**

Recomputation
Offloading

Maximizes throughput *but can exceed memory capacity*



**Memory
Optimization**

**Generalized
Pipeline
Parallelism**

**MoE
Inference**

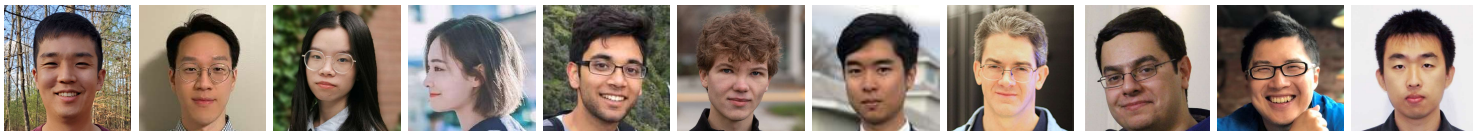
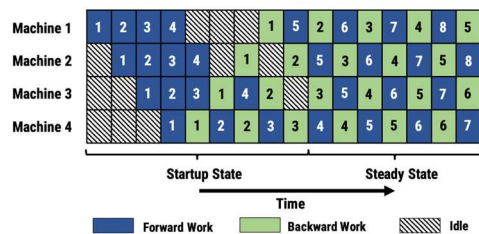
**JAX
Interface**

ECP-CANDLE

Cryo-EM

**Hugging
Face**

Extending Unity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

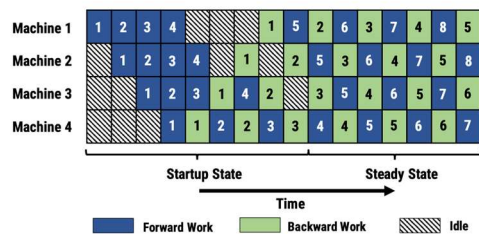
JAX
Interface

ECP-CANDLE

Cryo-EM

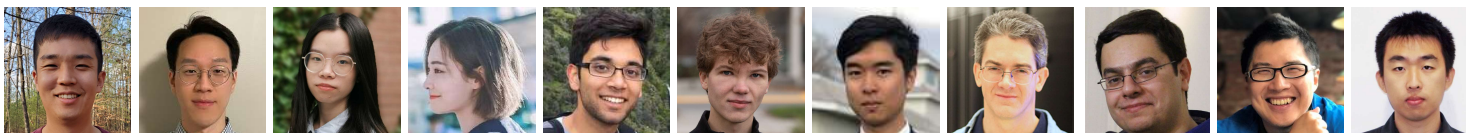
Hugging
Face

Extending Unity



Pipeline

Batch



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

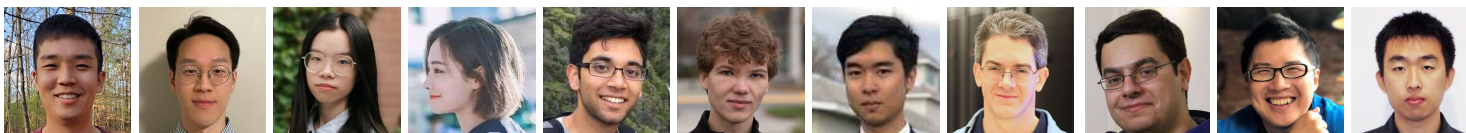
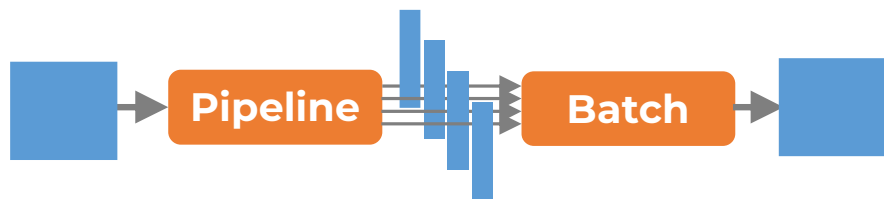
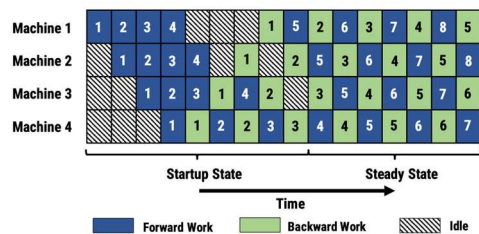
JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Extending Unity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

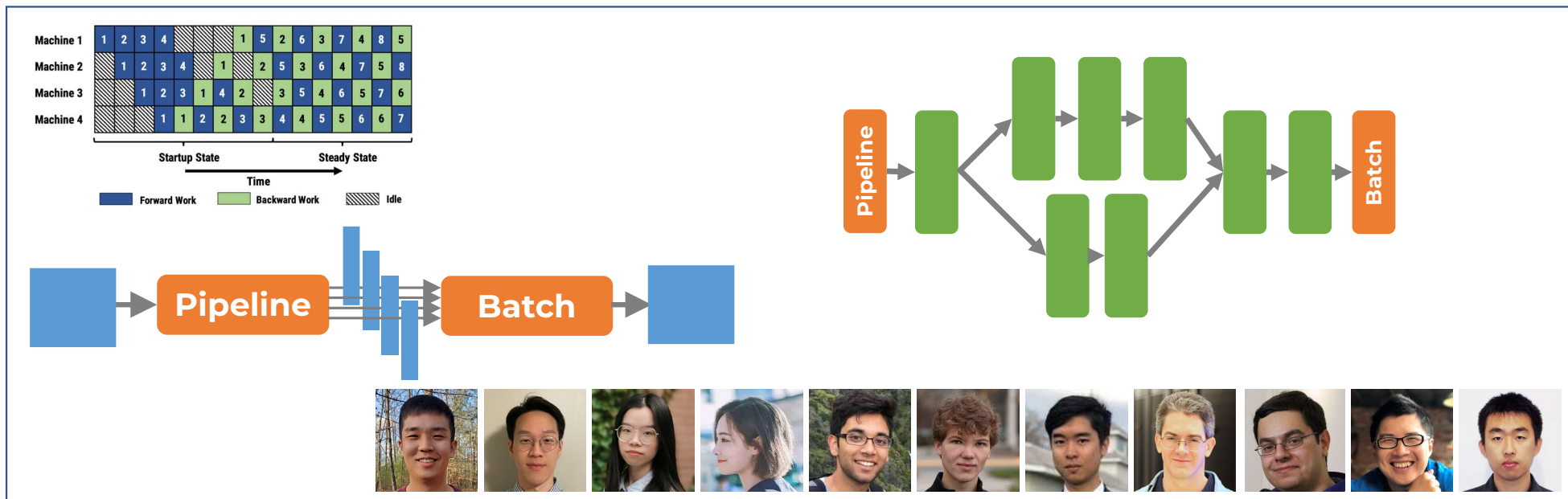
JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Extending Unity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

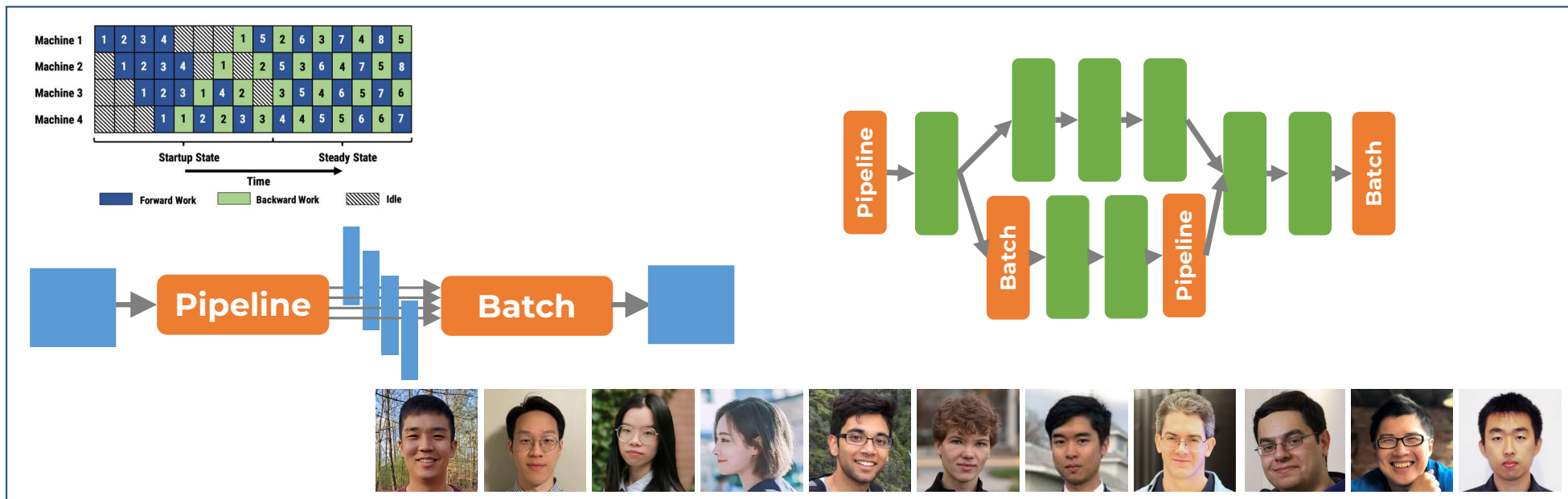
JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Extending Unity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

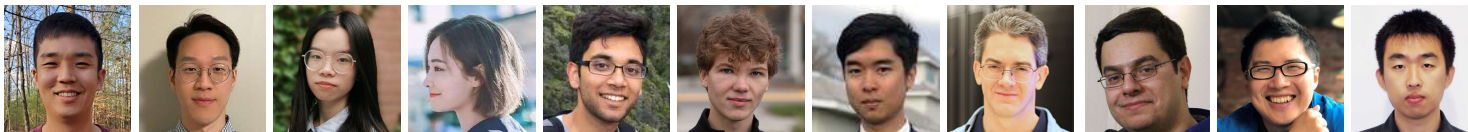
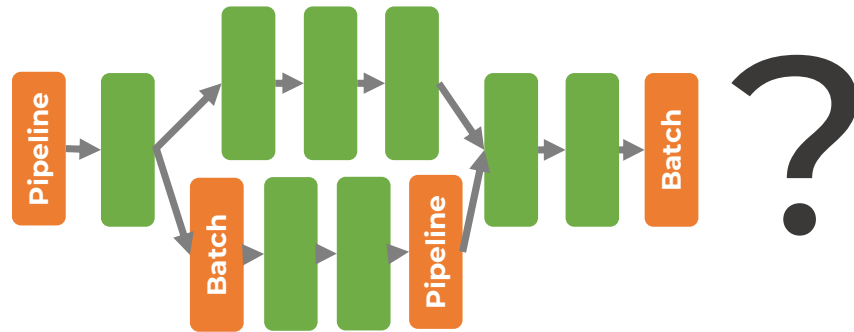
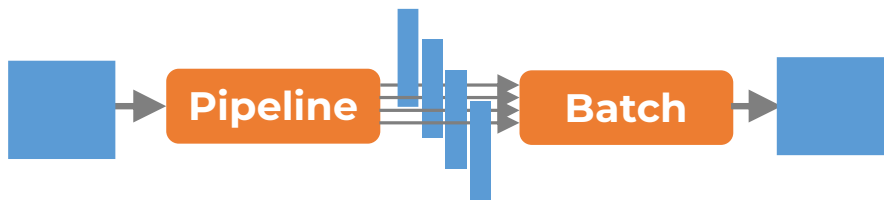
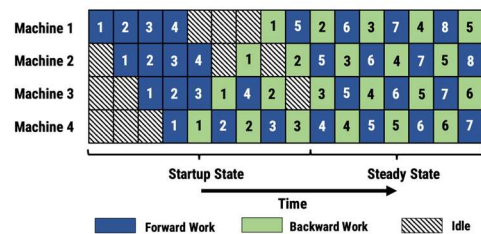
JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Extending Unity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Extending Unity



“sparsely-gated MoE” (2017)

GShard (2020)

GLaM (2021)

Switch Transformer (2021)

ST-MoE (2022) V-MoE (2022)

Expert Choice (2022)

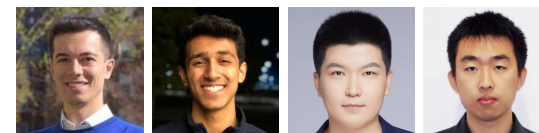
Tencent 腾讯

SpeechMoE (2021)

SpeechMoE2 (2021)



DeepSpeed-MoE (2022)



Memory
Optimization

Generalized
Pipeline
Parallelism

**MoE
Inference**

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

New Features



"sparsely-gated"
GShard (2020)
GLaM (2021)
Switch Trans
ST-MoE (2022)
Expert Choice (2022)



Mixture of
Experts

**data-dependent
control flow
+
distributed execution**



ed-MoE (2022)



+ improved inference support!

Memory
Optimization

Generalized
Pipeline
Parallelism

**MoE
Inference**

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

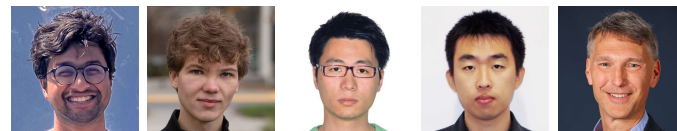
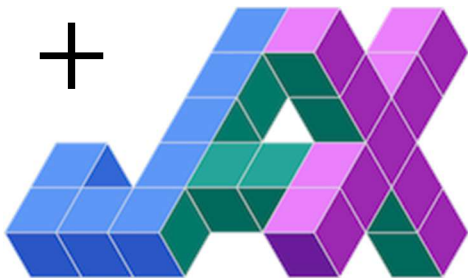
New Features

 Keras

 PyTorch

 ONNX

+



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

New Features

 Keras

 PyTorch

 ONNX

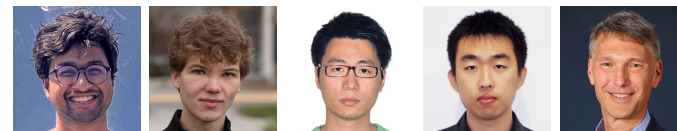
+



A functional deep learning framework



immutability, compositionality, modularity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

New Features

 Keras

 PyTorch

 ONNX

+

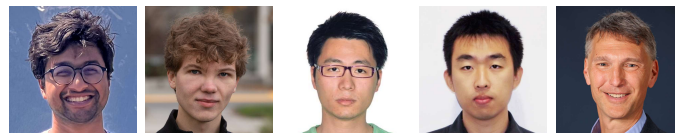


`jaxpr`

A functional deep learning framework



immutability, compositionality, modularity



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

New Features

 Keras

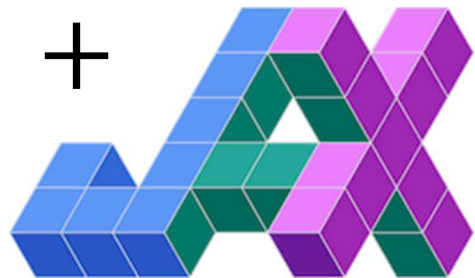
 PyTorch

 ONNX

A functional deep learning framework

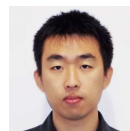
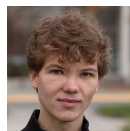


immutability, compositionality, modularity



+

`jaxpr`



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

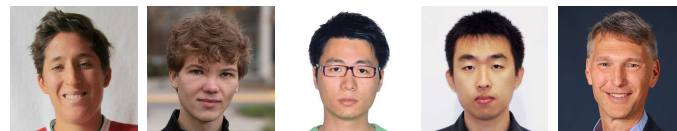
Cryo-EM

Hugging
Face

New Features



Set of Deep Neural Networks for
Precision Medicine



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

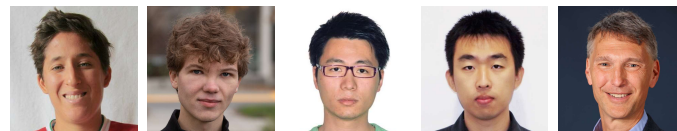
Hugging
Face

Applications



Set of Deep Neural Networks for Precision Medicine

Difficult to scale w/
existing frameworks



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

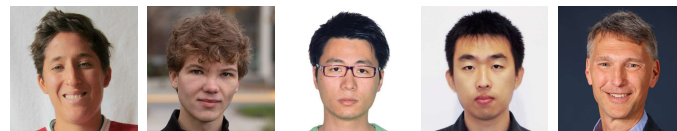
Hugging
Face

Applications



Set of Deep Neural Networks for
Precision Medicine

Difficult to scale w/
existing frameworks → *Hybrid Parallelism*



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

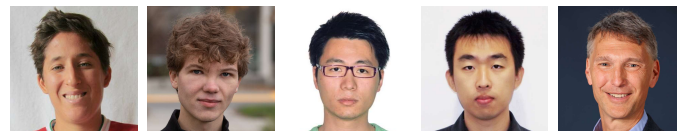
Hugging
Face

Applications



Set of Deep Neural Networks for
Precision Medicine

Difficult to scale w/
existing frameworks → **FlexFlow**
Hybrid Parallelism



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

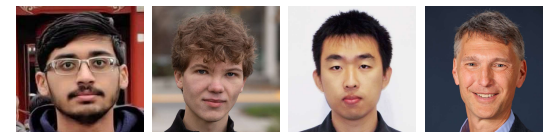
Cryo-EM

Hugging
Face

Applications

Cryo-EM

↓
electron microscopy



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

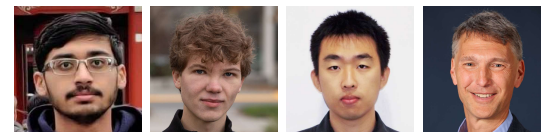
Hugging
Face

Applications

Cryo-EM

↓
electron microscopy

Raw EM Data
(slices)



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

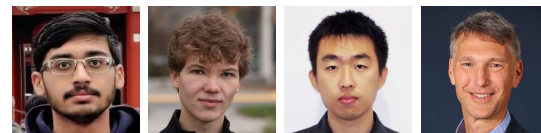
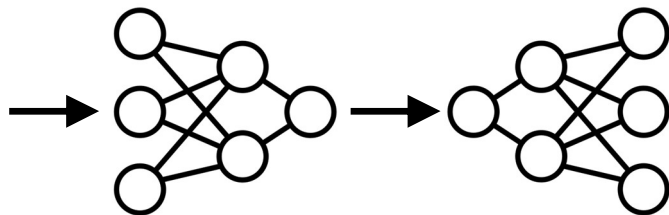
Hugging
Face

Applications

Cryo-EM

↓
electron microscopy

Raw EM Data
(slices)



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

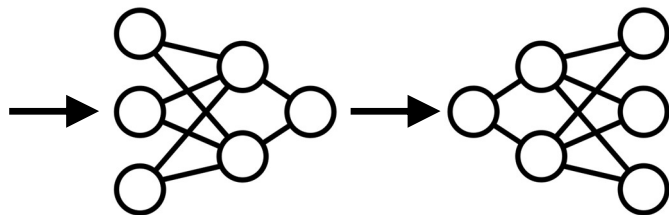
Hugging
Face

Applications

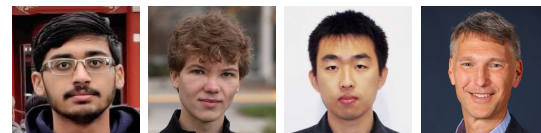
Cryo-EM

electron microscopy

Raw EM Data
(slices)



3D Protein Structure



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

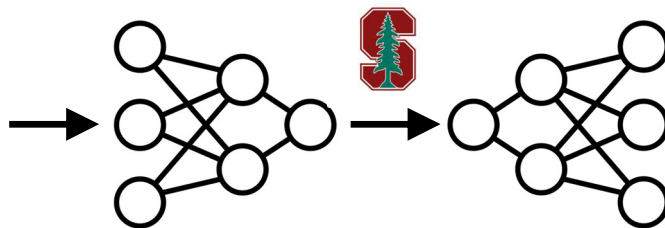
Hugging
Face

Applications

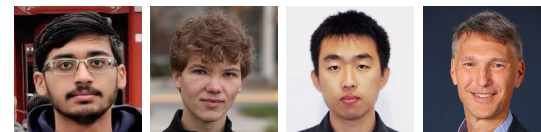
Cryo-EM

↓
electron microscopy

Raw EM Data
(slices)



→ 3D Protein Structure



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

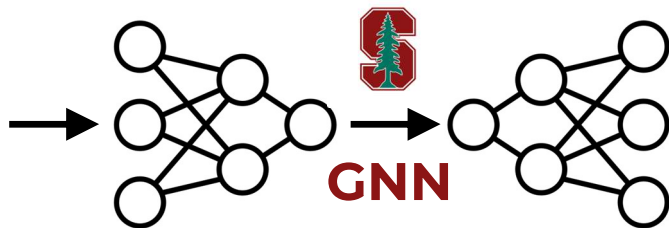
Hugging
Face

Applications

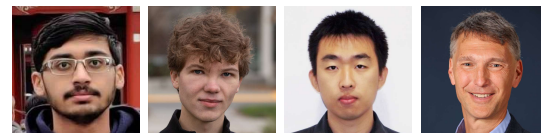
Cryo-EM

electron microscopy

Raw EM Data
(slices)



3D Protein Structure



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

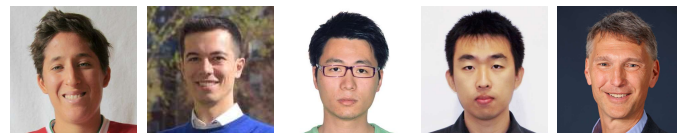
Applications



Library for
Transformer-based
Models

Ease-of-use

Performance



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

ECP-CANDLE

Cryo-EM

Hugging
Face

Applications

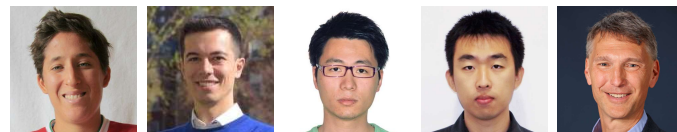


Library for
Transformer-based
Models

Ease-of-use

Performance

FlexFlow



Memory
Optimization

Generalized
Pipeline
Parallelism

MoE
Inference

JAX
Interface

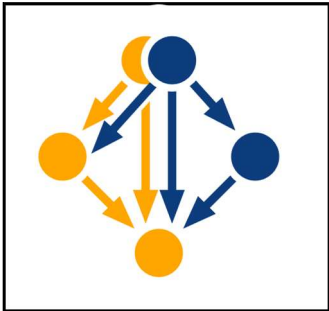
ECP-CANDLE

Cryo-EM

Hugging
Face

Applications

<https://github.com/flexflow/FlexFlow>



kadinzhang Added tests for Linear operator in align/linear (#264)		4884234 · 19 days ago · 1,339 commits
·circled	[CircleCI] update script	12 months ago
align	Added tests for Linear operator in align/linear (#264)	19 days ago
bootcamp_demo	[python] make num_samples a property (#121)	14 months ago
cmake	[CMake] retrieve legion installation path for spack	9 months ago
conda	[Conda]: fix conda	17 months ago
config	[Alignment] Add embedding alignment (#255)	3 months ago
deps	[Legion] version update	8 months ago
docker	Adding Dockerfile	2 years ago
docs	[Docs]: done with docs except tutorials	16 months ago

A distributed deep learning framework that supports flexible parallelization strategies.

- Readme
- Apache-2.0 license
- Code of conduct
- 458 stars
- 21 watching
- 100 forks

Releases

Release 21.09 (September 30th ...)

on Oct 6, 2021



 Keras  PyTorch  ONNX

Questions?