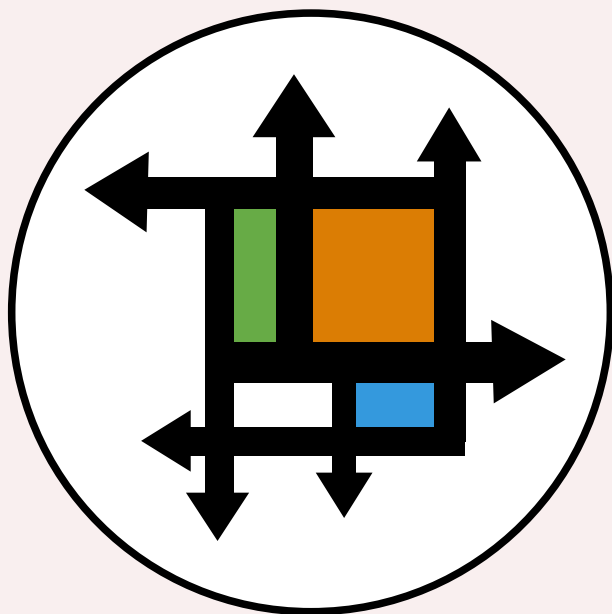


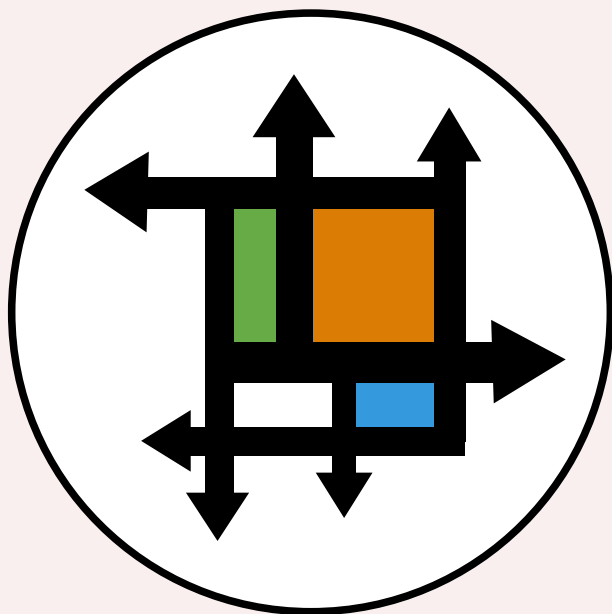
FlexFlow

Colin Unger



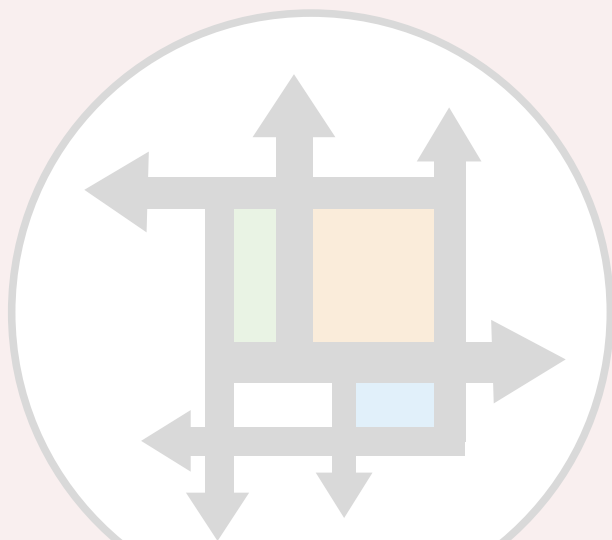
FlexFlow

distributed DNN



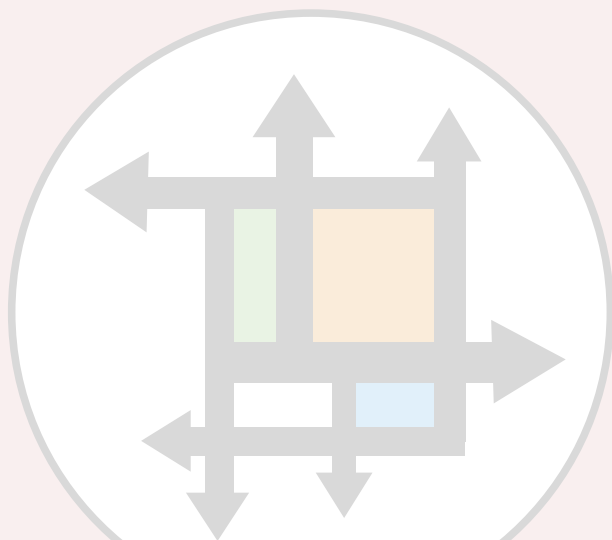
FlexFlow

distributed DNN training



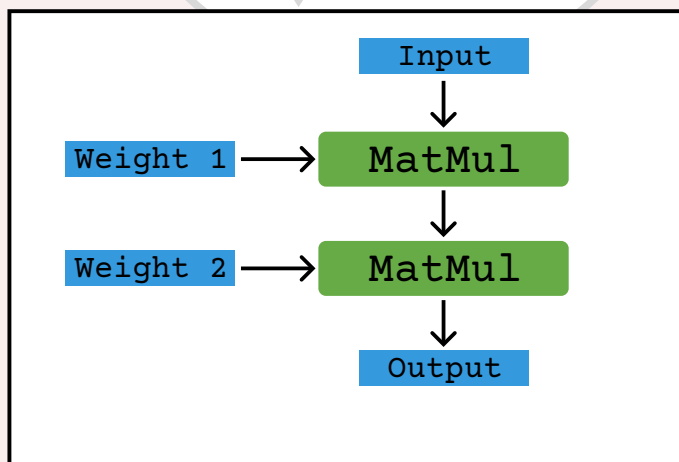
FlexFlow

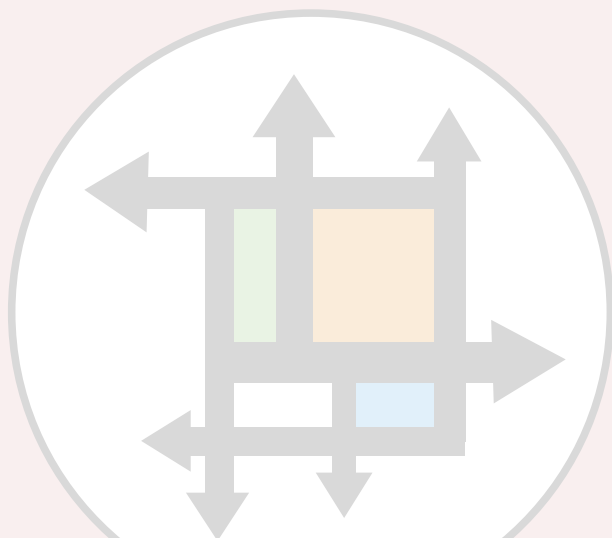
distributed DNN training



FlexFlow

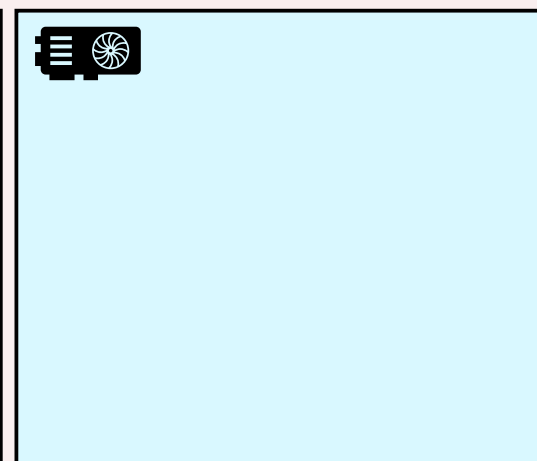
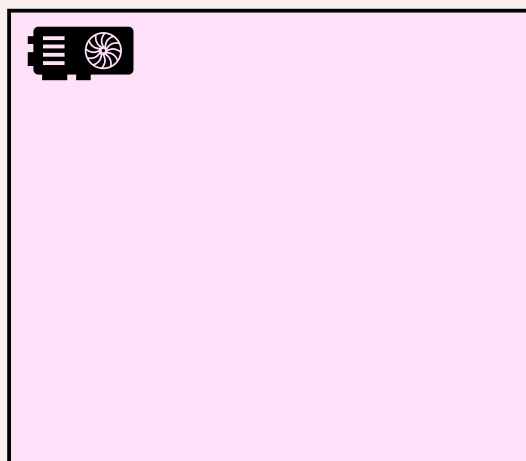
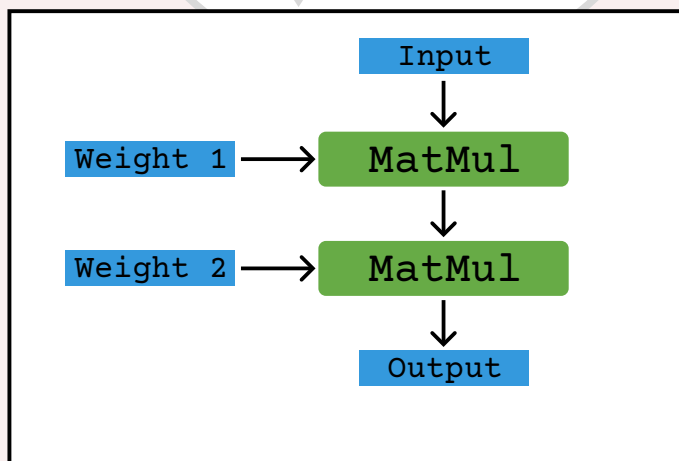
distributed DNN training

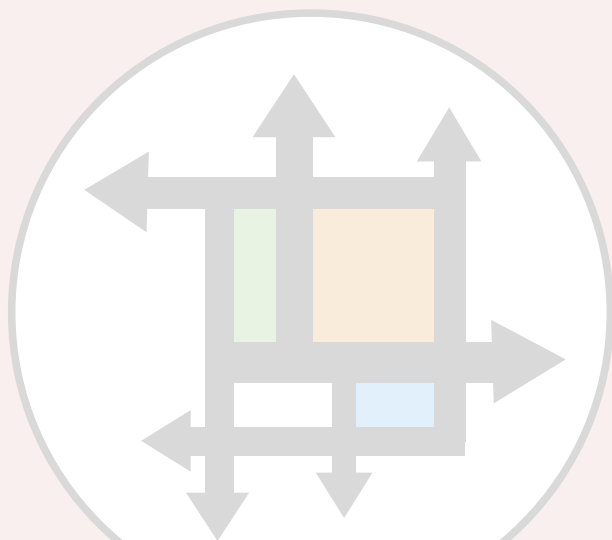




FlexFlow

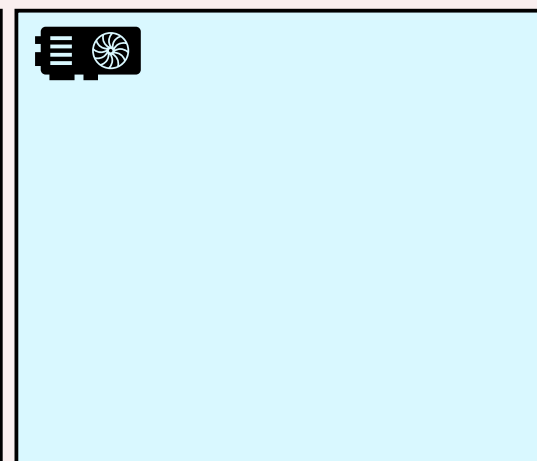
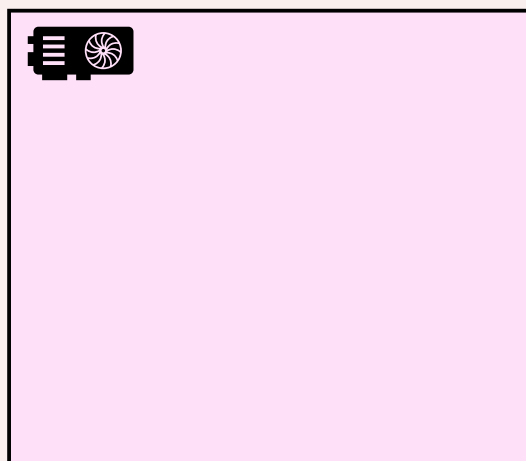
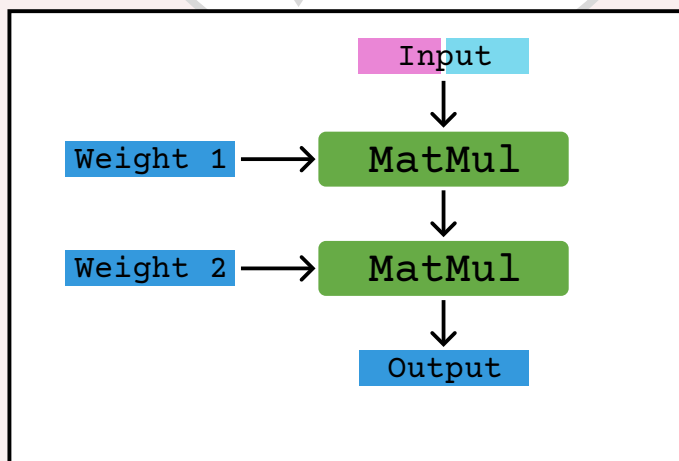
distributed DNN training

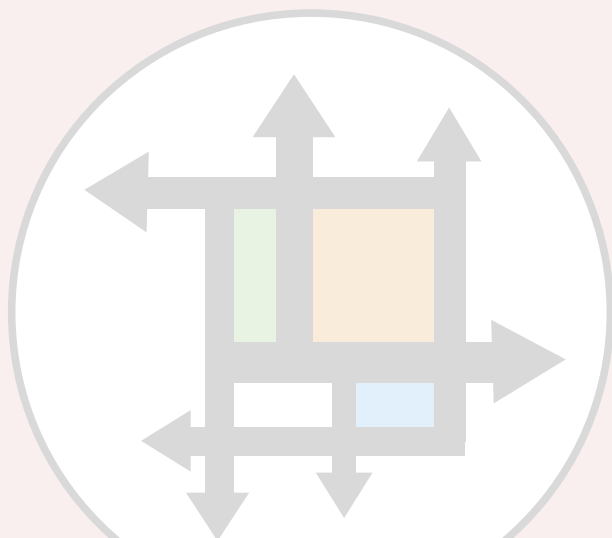




FlexFlow

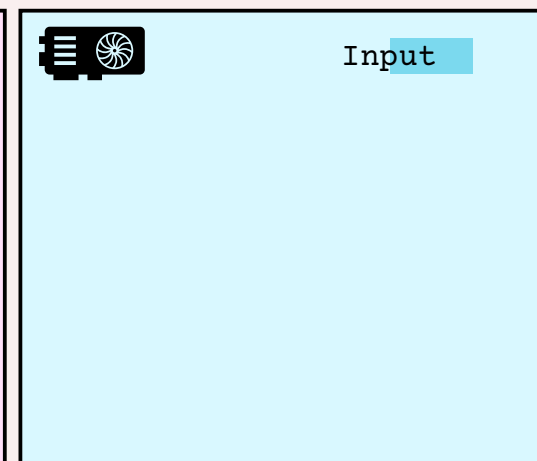
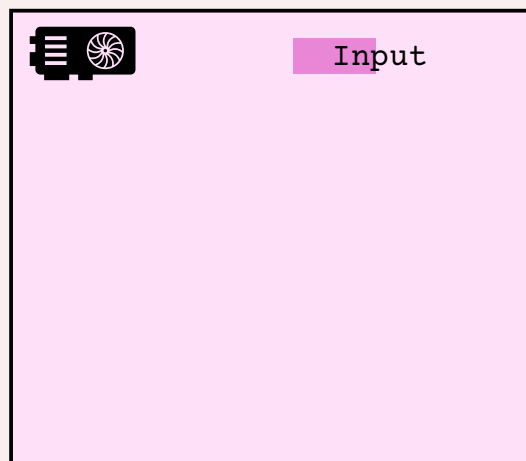
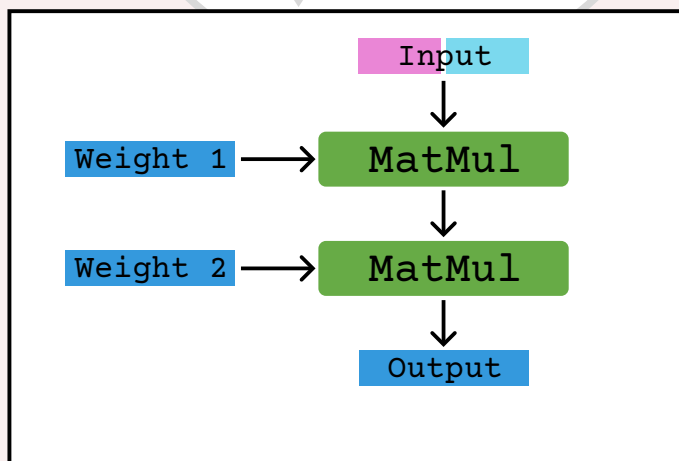
distributed DNN training

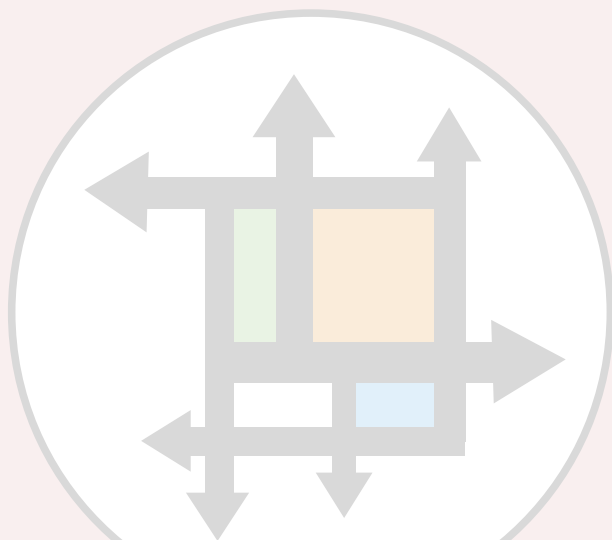




FlexFlow

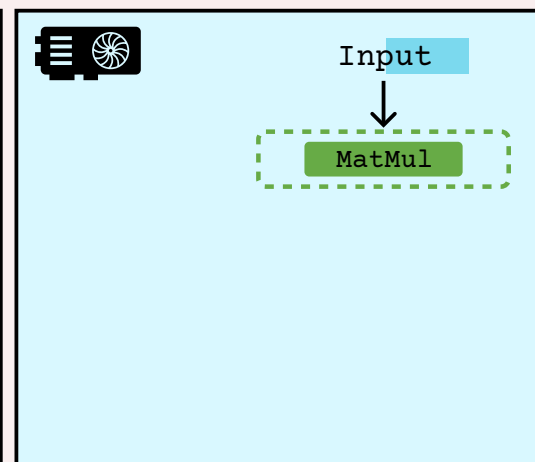
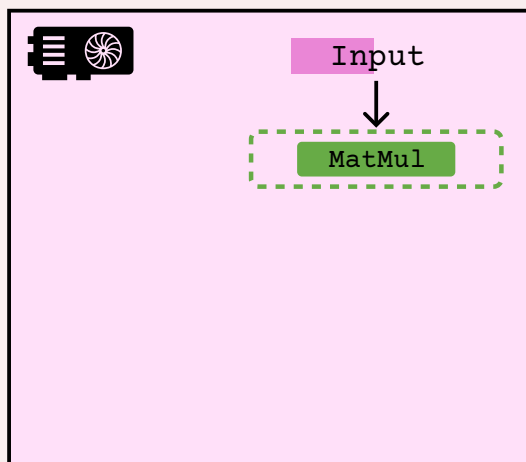
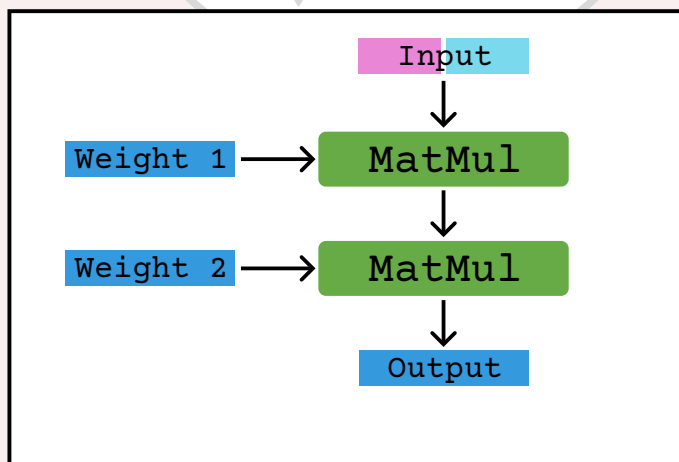
distributed DNN training

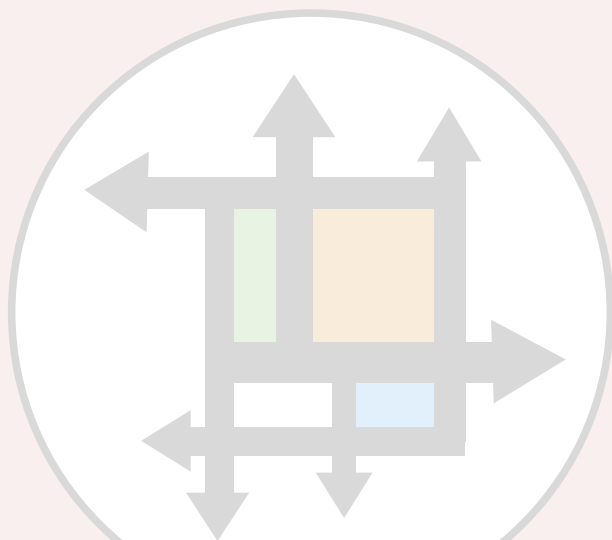




FlexFlow

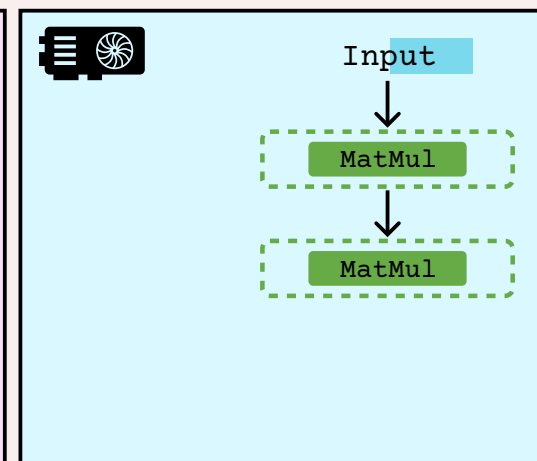
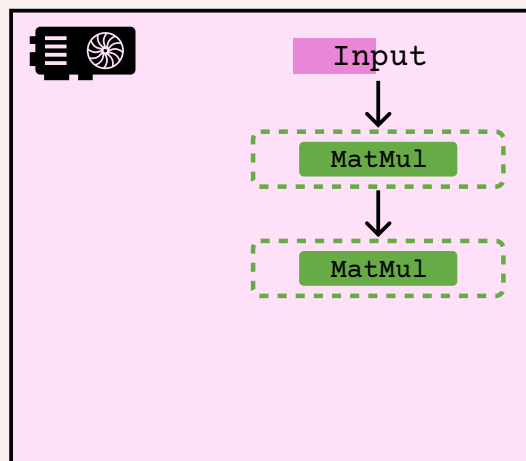
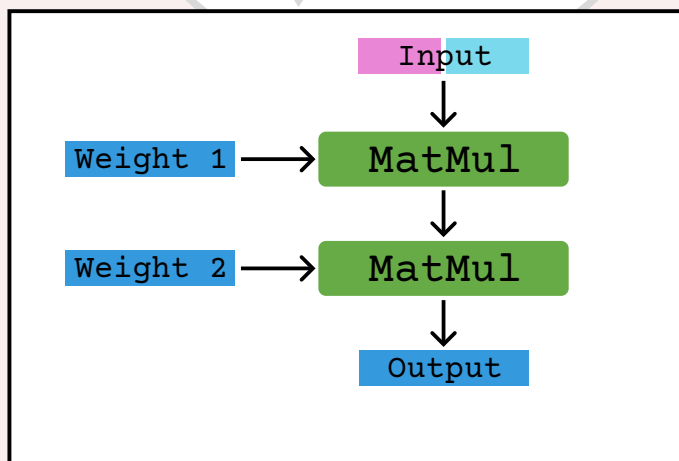
distributed DNN training

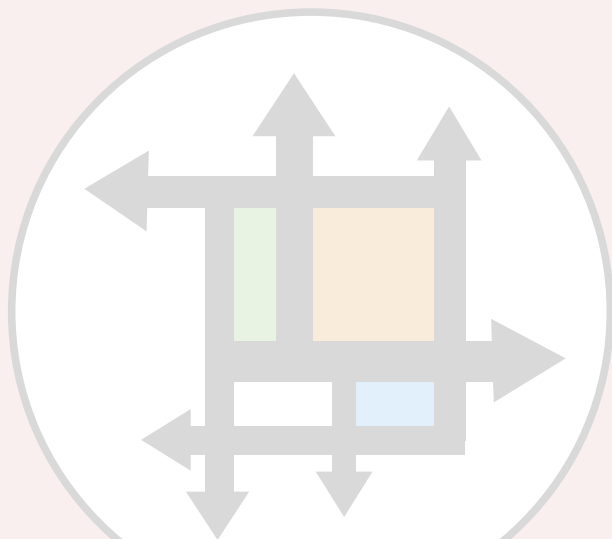




FlexFlow

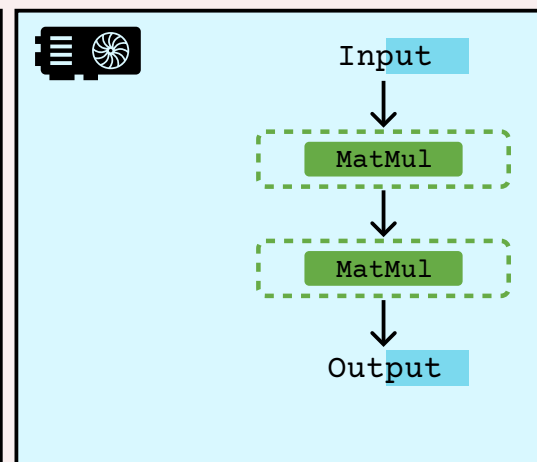
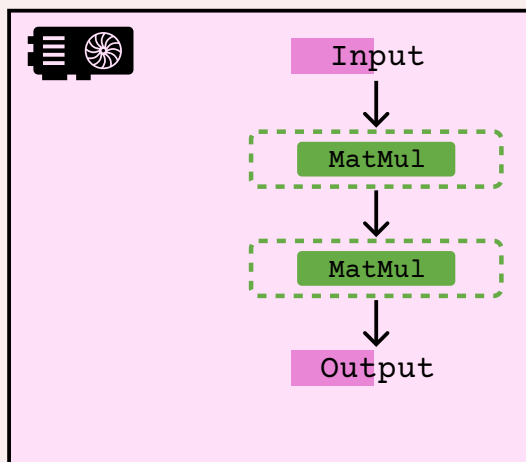
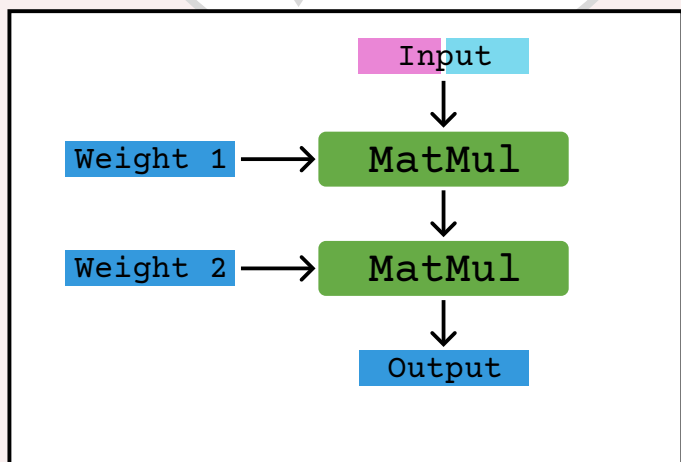
distributed DNN training

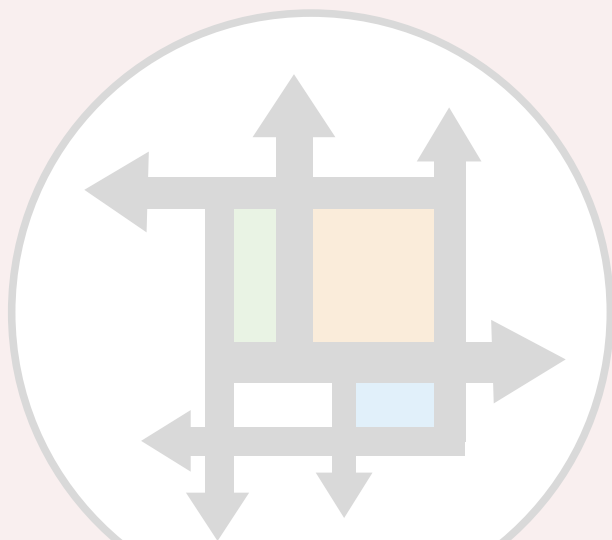




FlexFlow

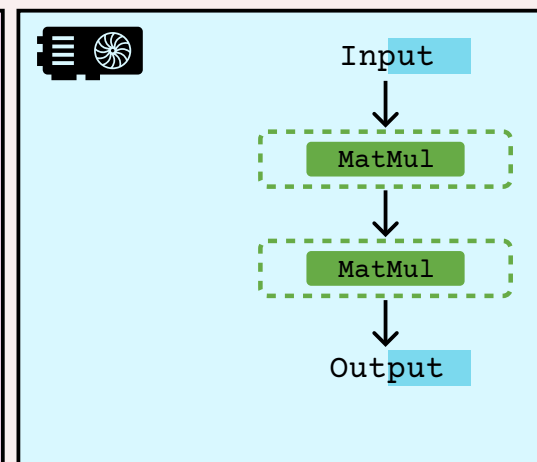
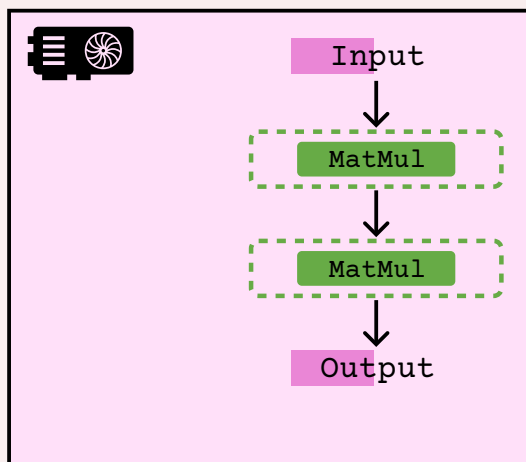
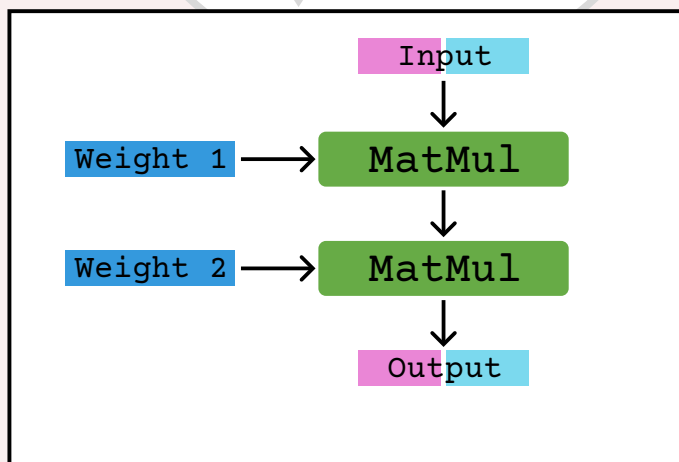
distributed DNN training

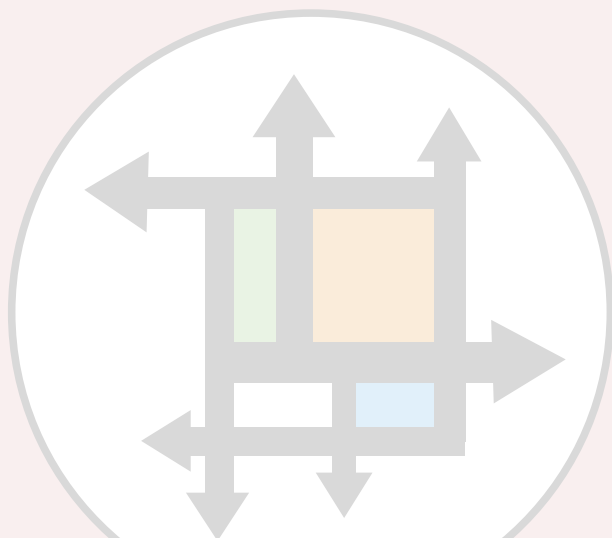




FlexFlow

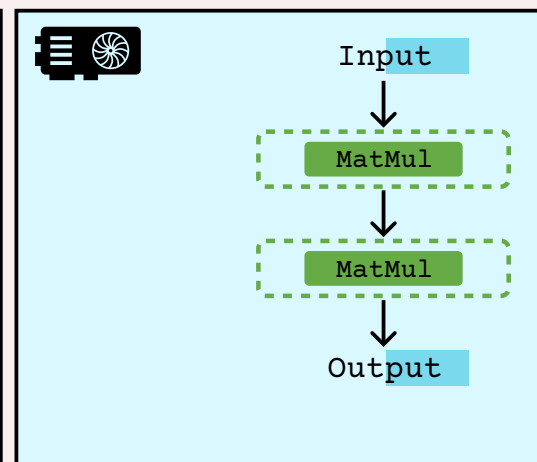
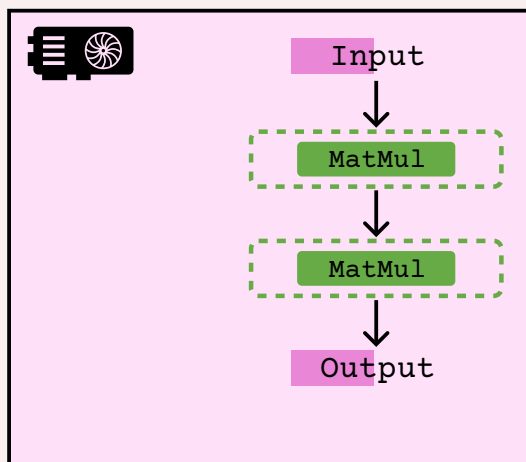
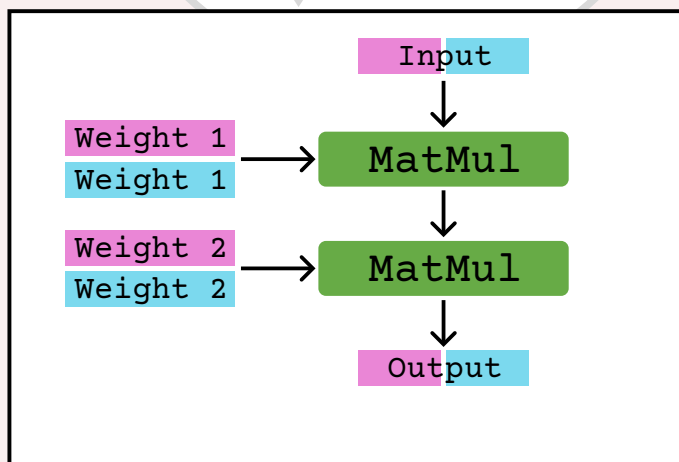
distributed DNN training

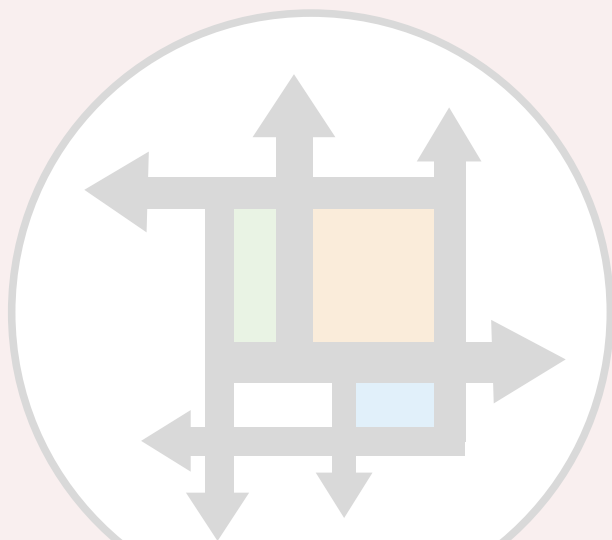




FlexFlow

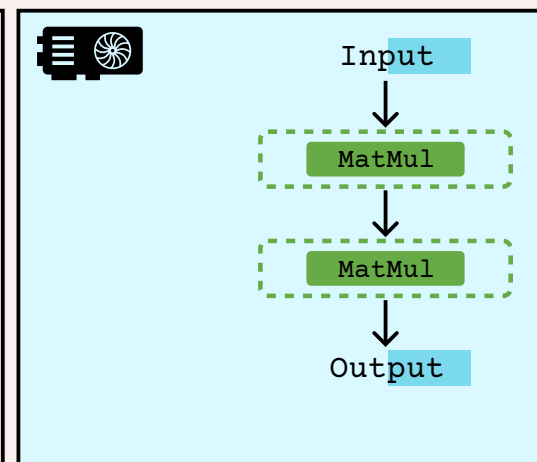
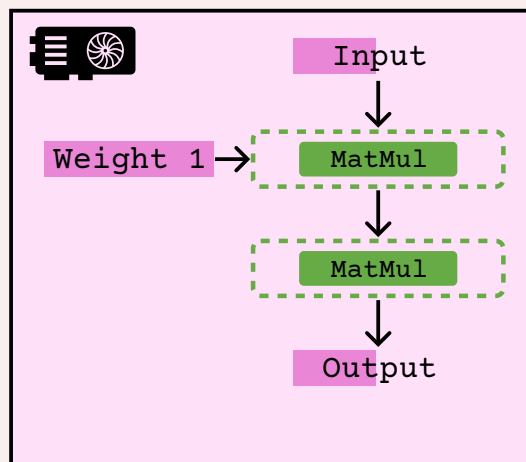
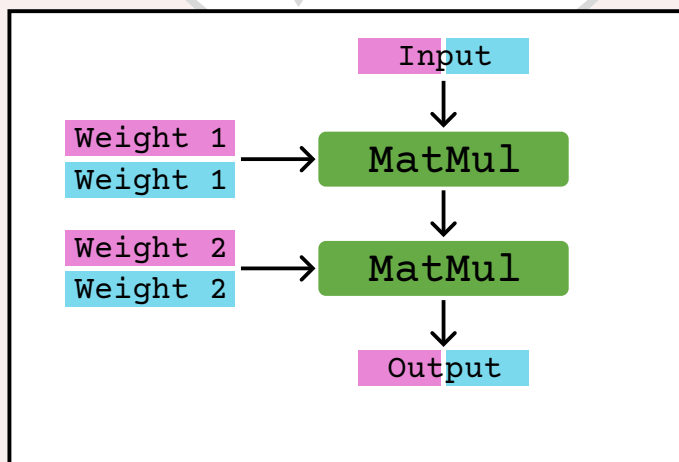
distributed DNN training

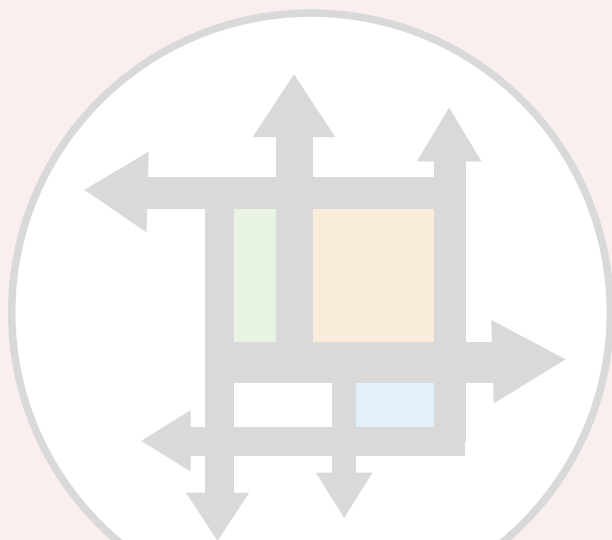




FlexFlow

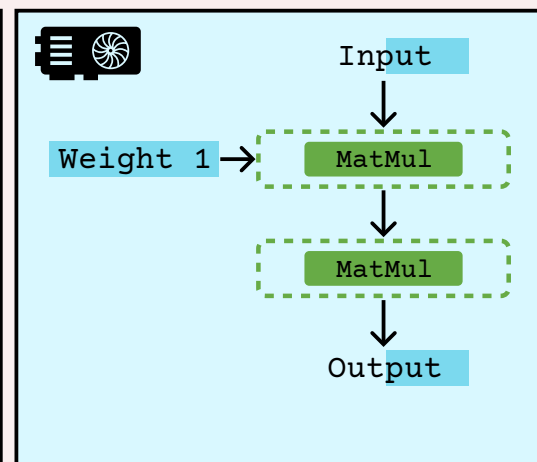
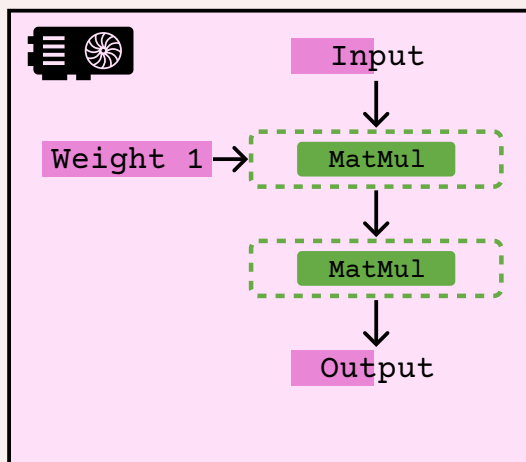
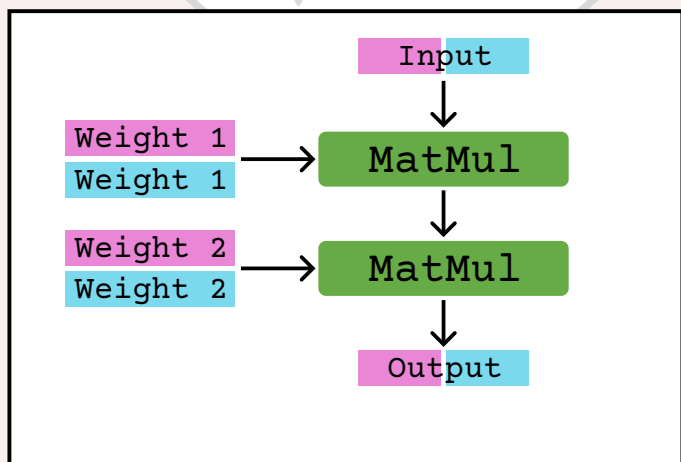
distributed DNN training

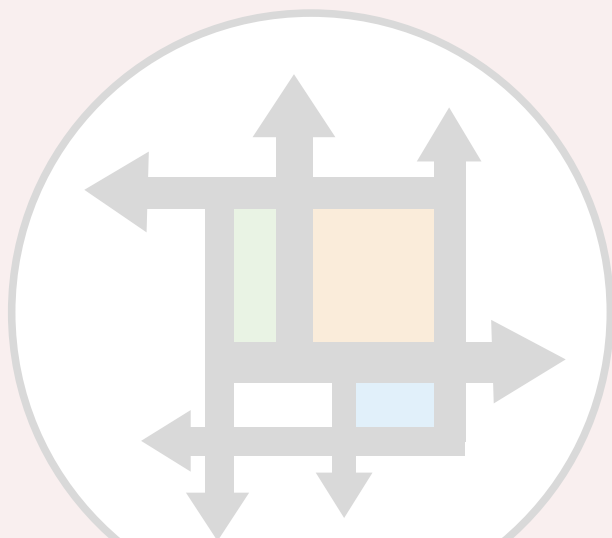




FlexFlow

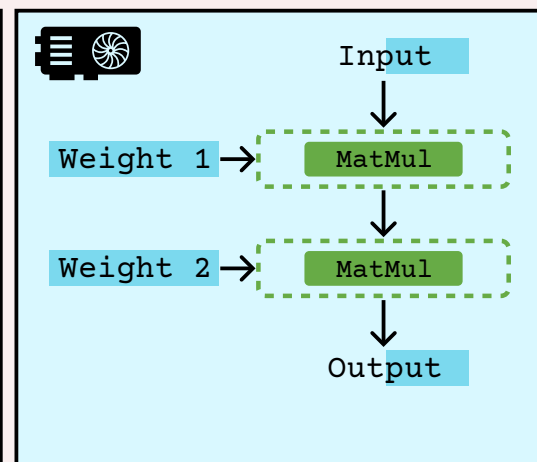
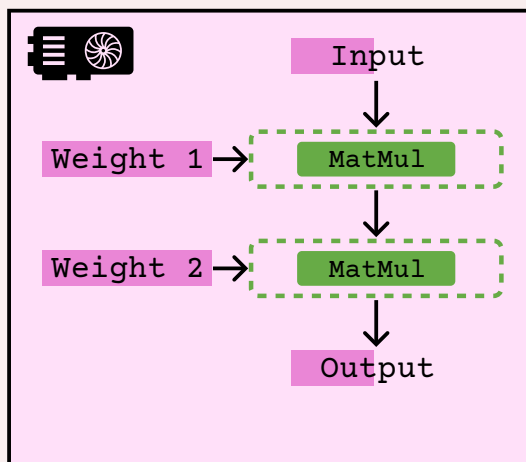
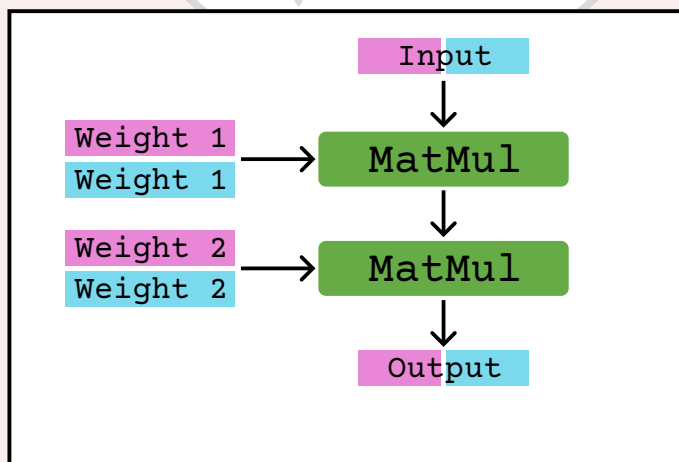
distributed DNN training

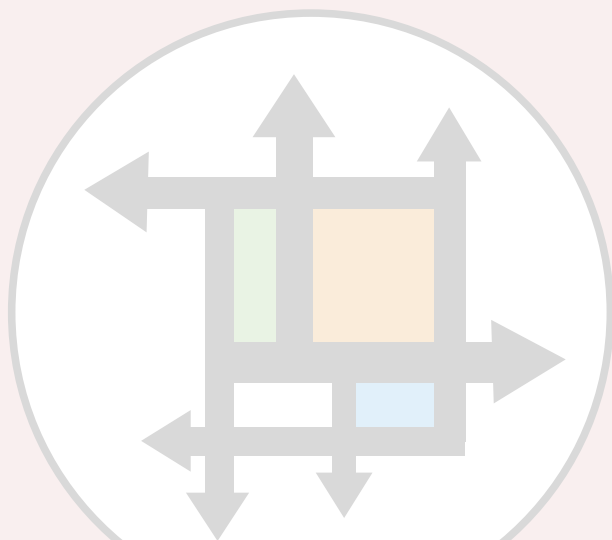




FlexFlow

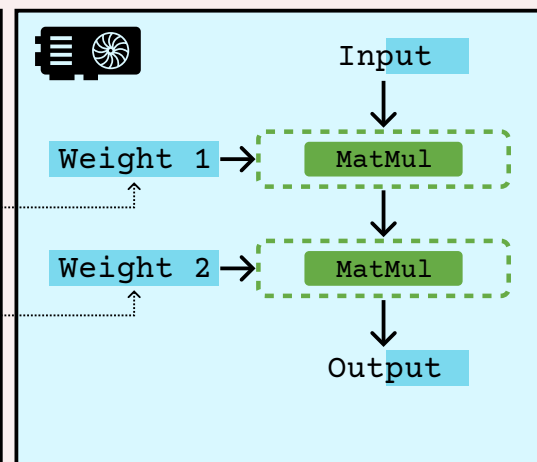
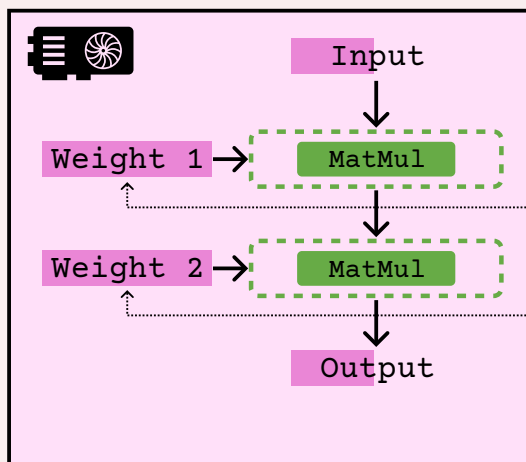
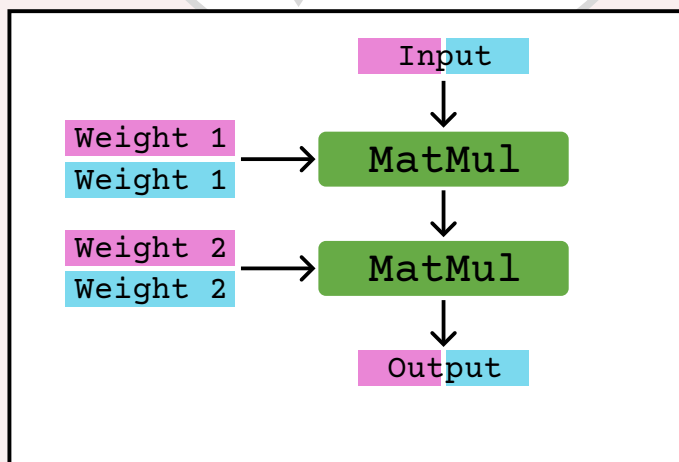
distributed DNN training

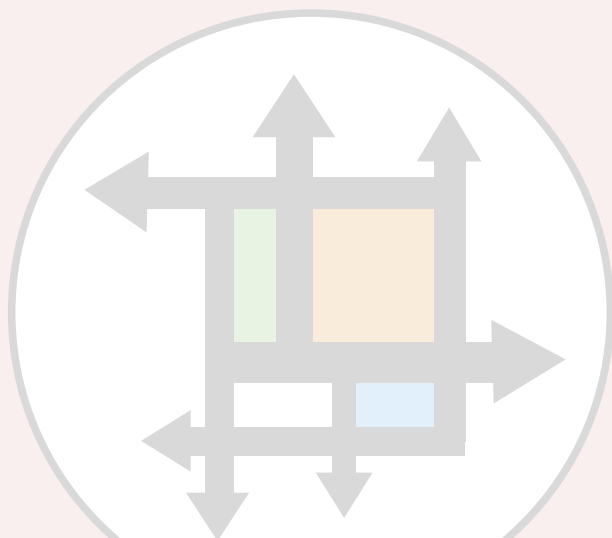




FlexFlow

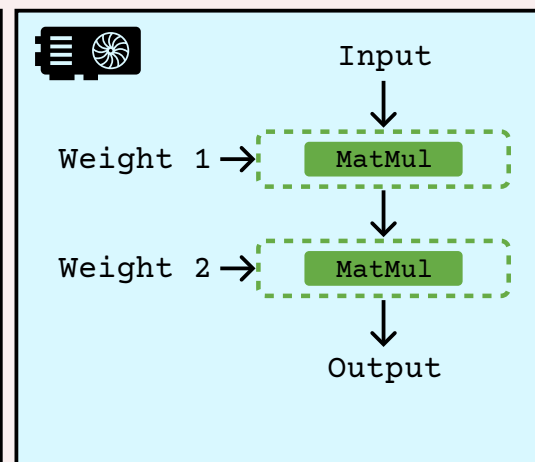
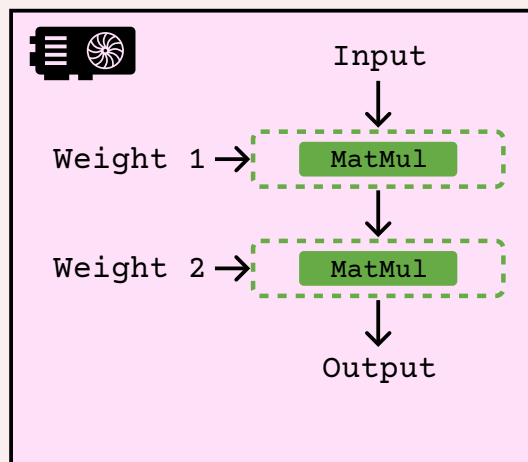
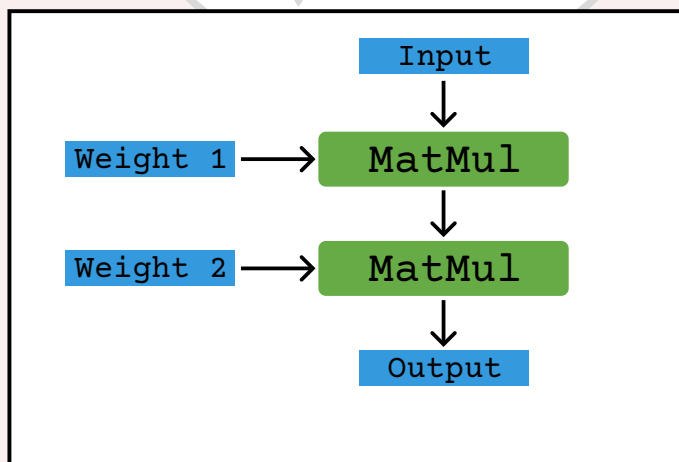
distributed DNN training

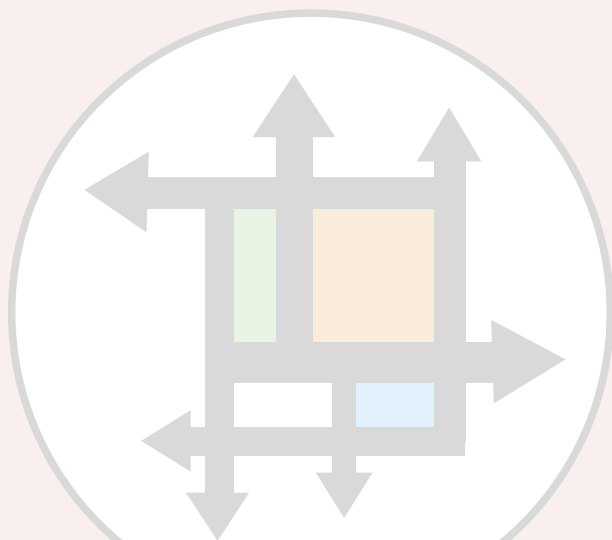




FlexFlow

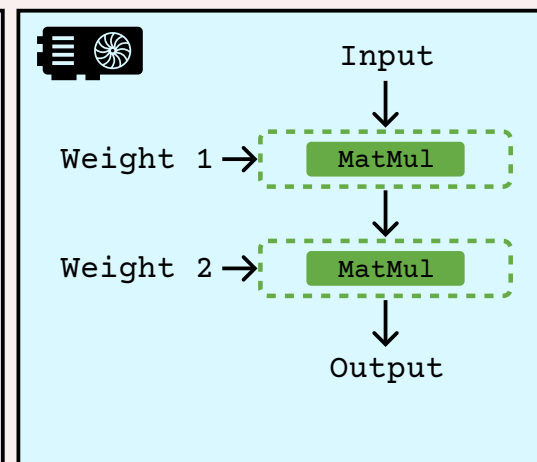
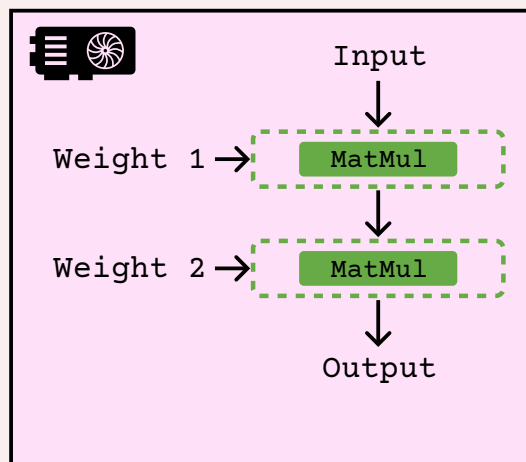
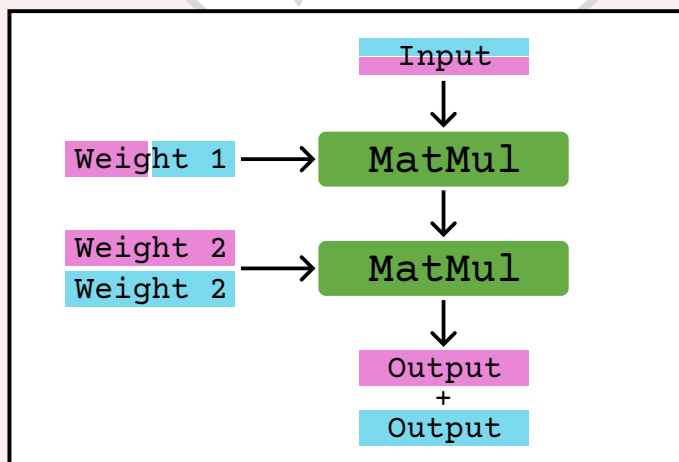
distributed DNN training

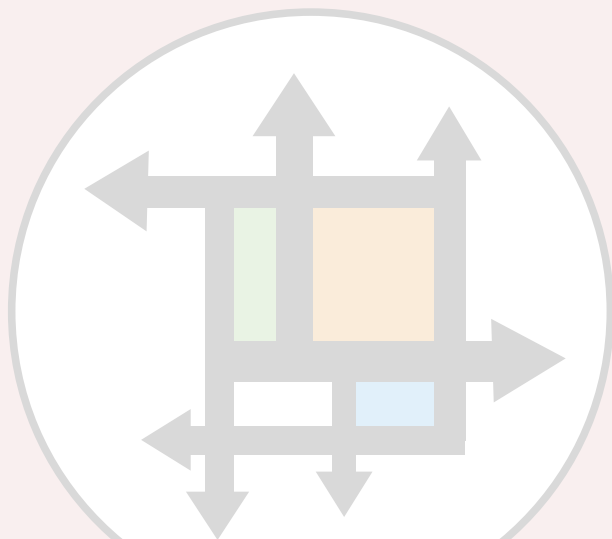




FlexFlow

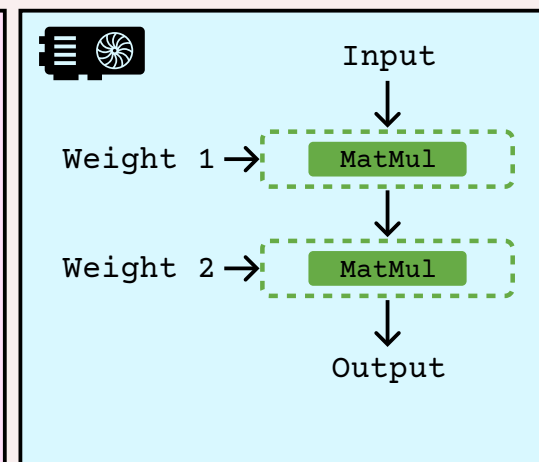
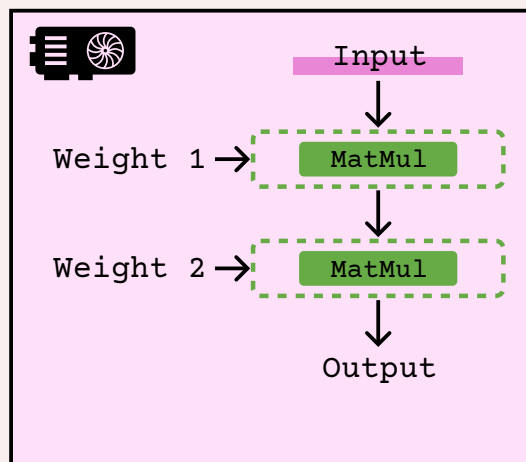
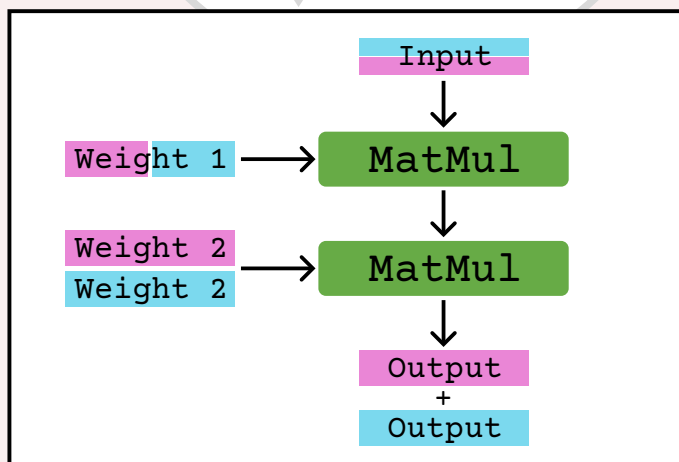
distributed DNN training

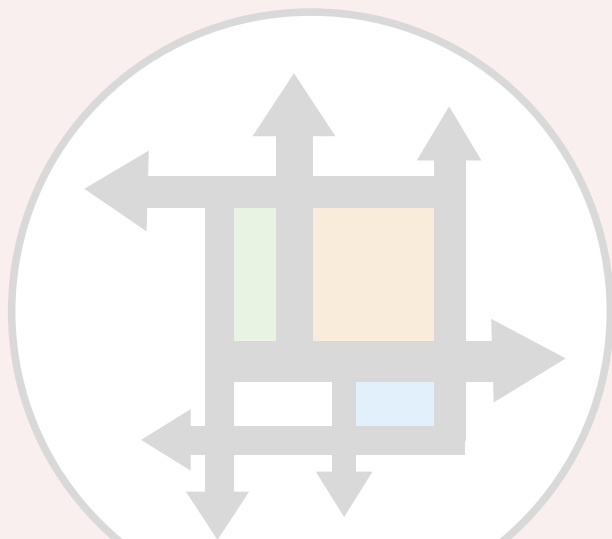




FlexFlow

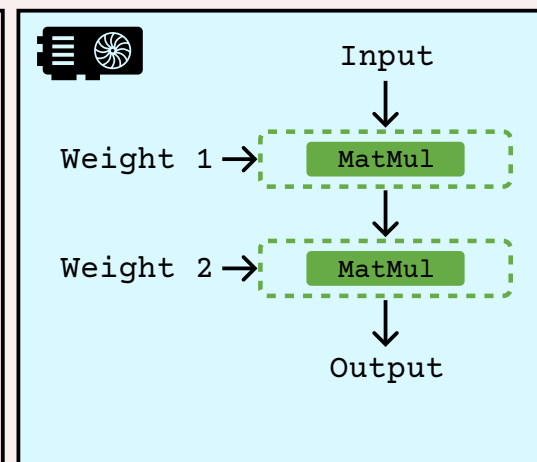
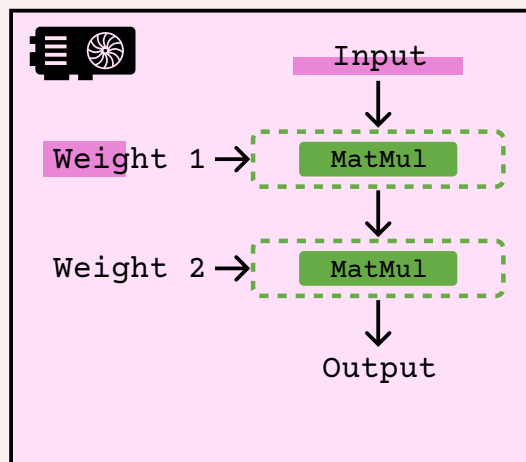
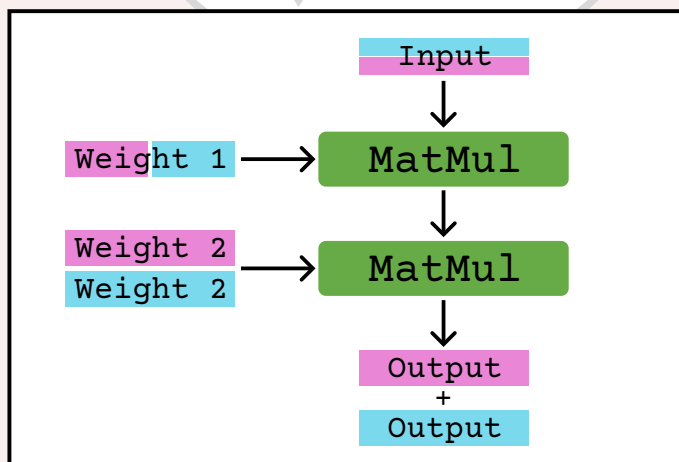
distributed DNN training

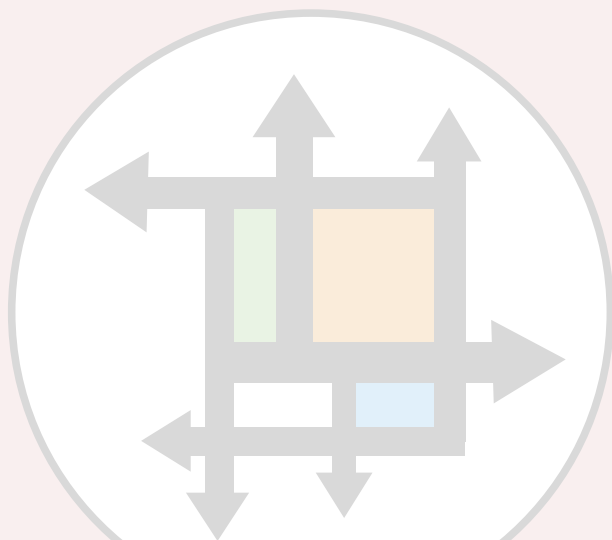




FlexFlow

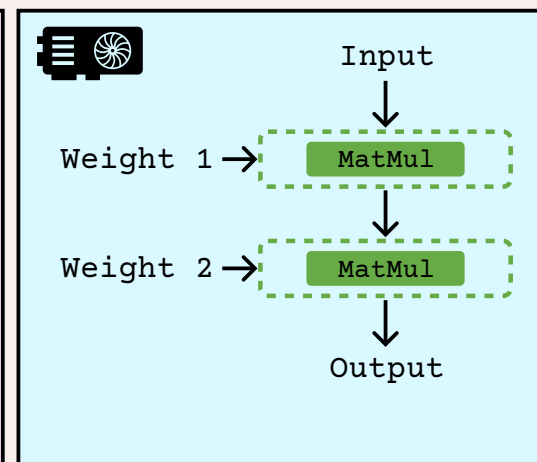
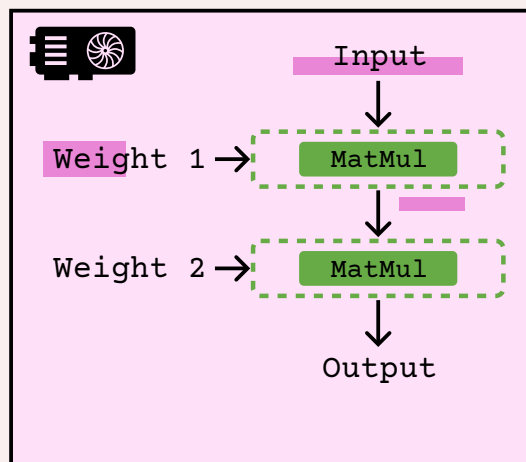
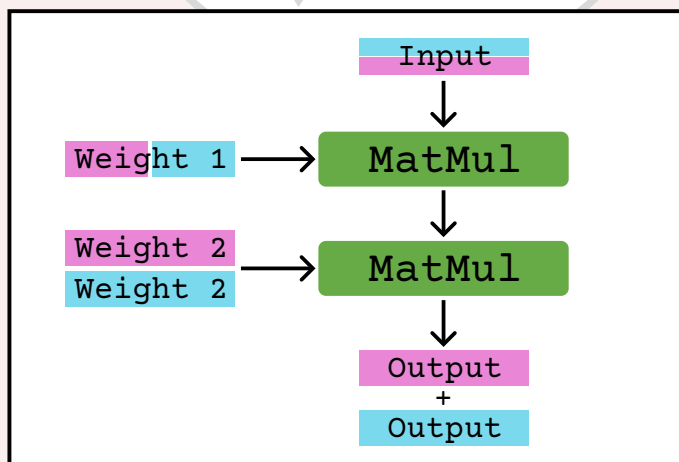
distributed DNN training

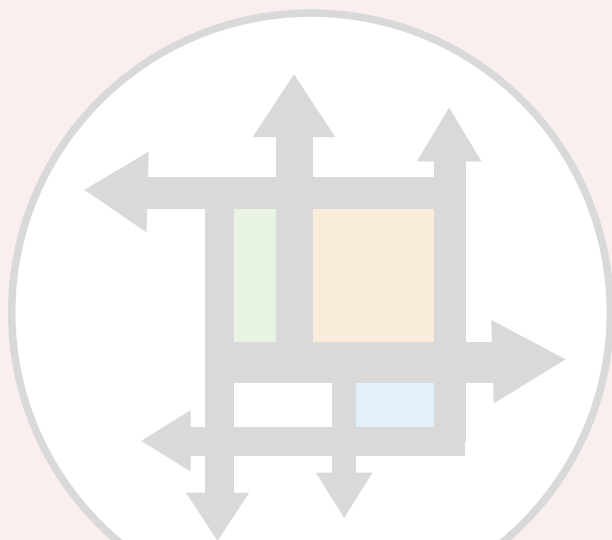




FlexFlow

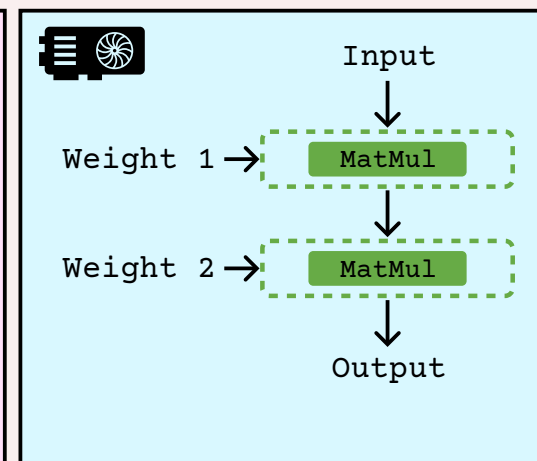
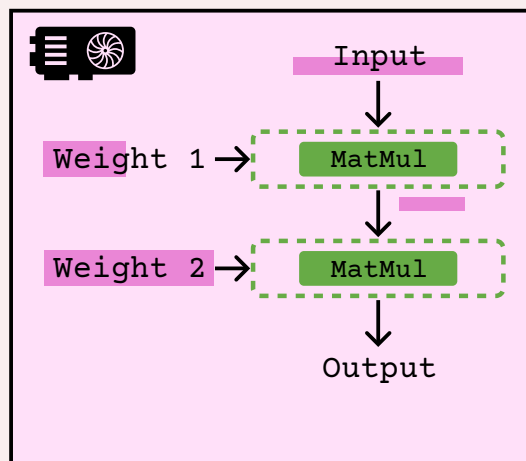
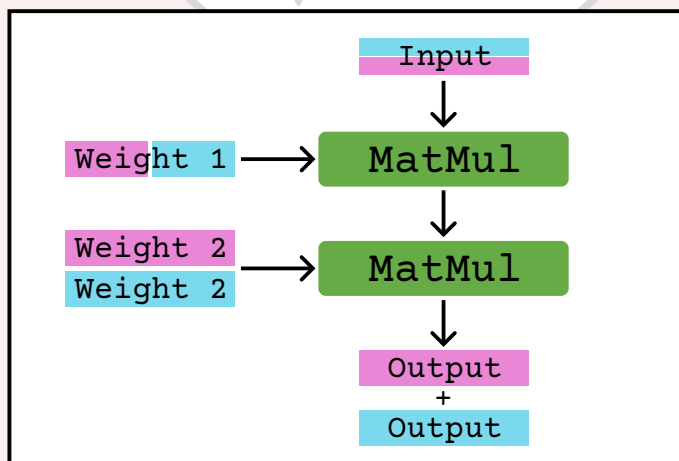
distributed DNN training

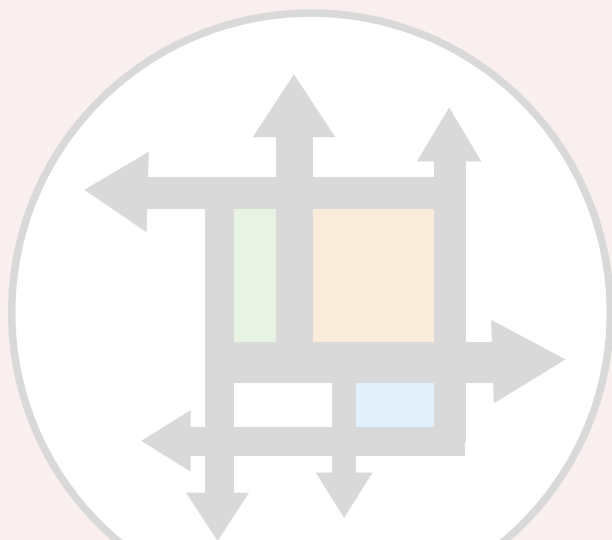




FlexFlow

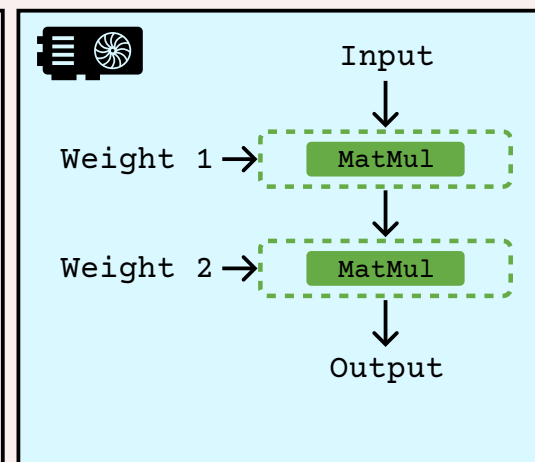
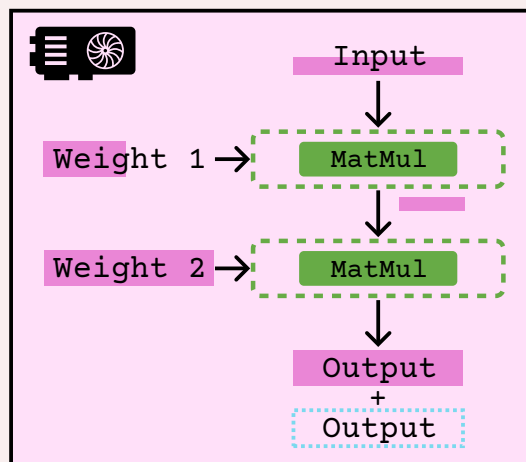
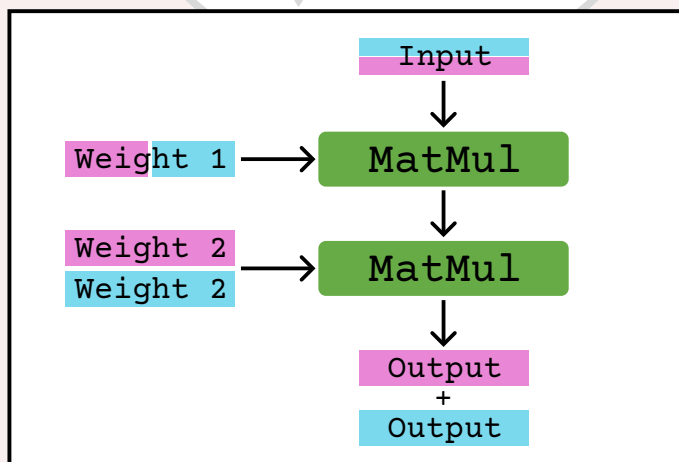
distributed DNN training

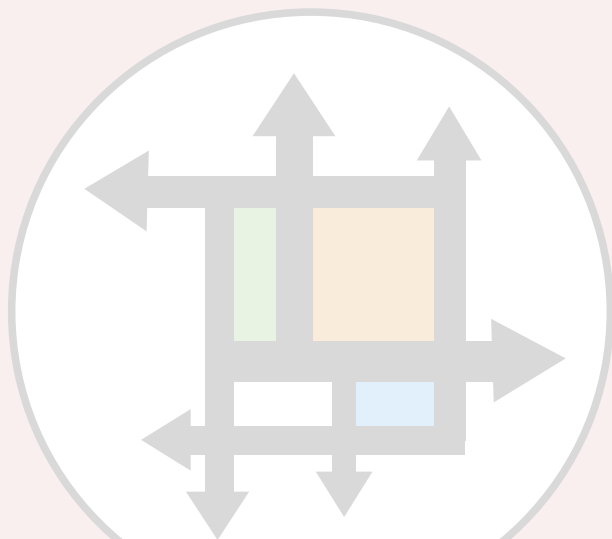




FlexFlow

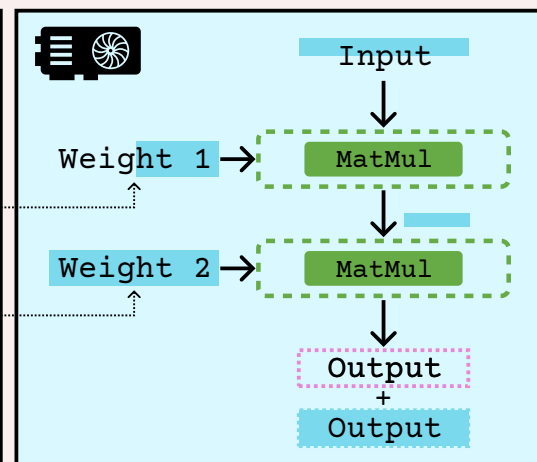
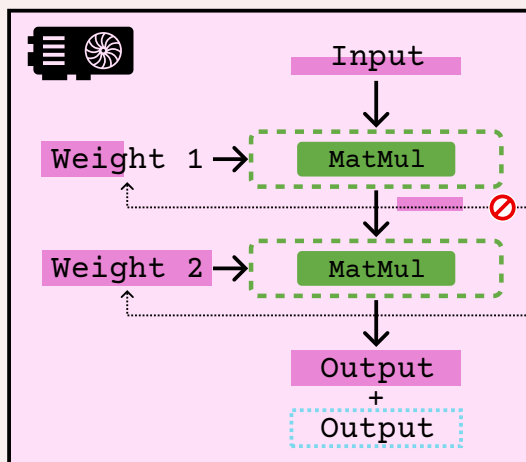
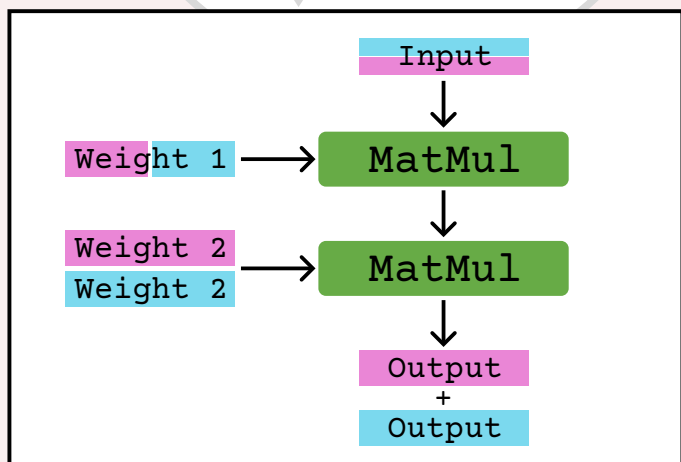
distributed DNN training

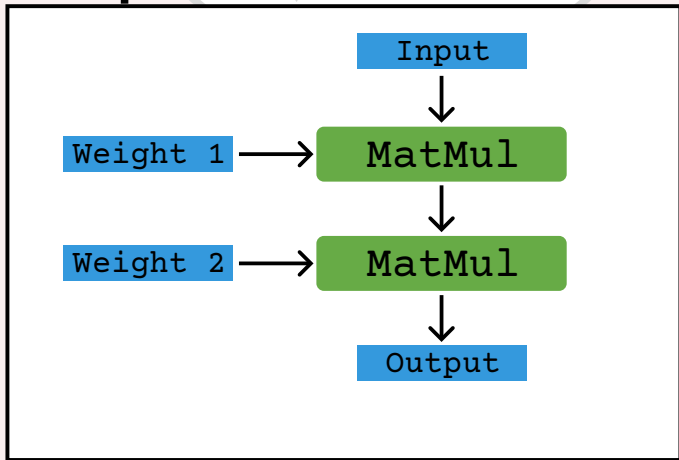
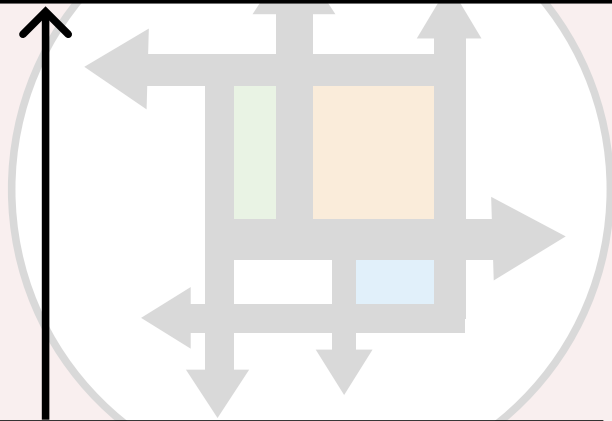
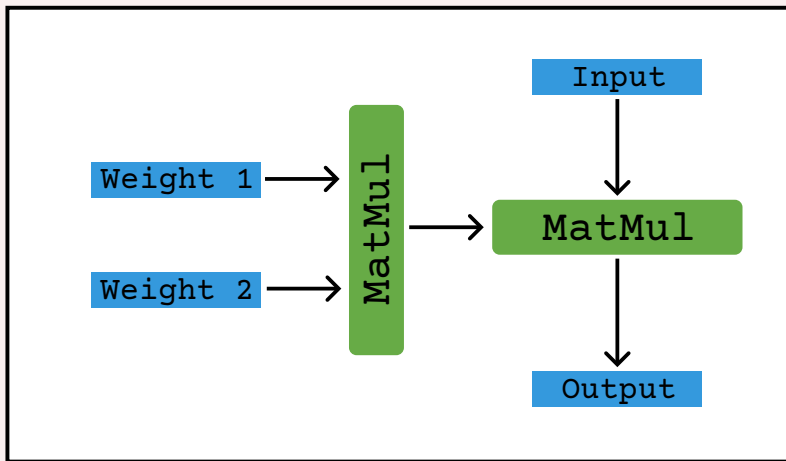




FlexFlow

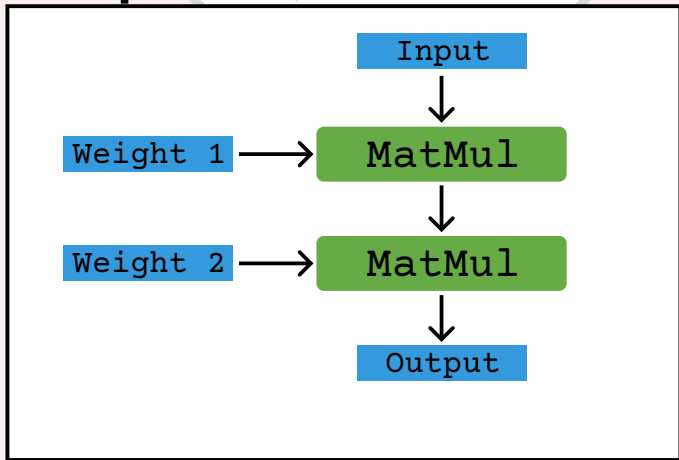
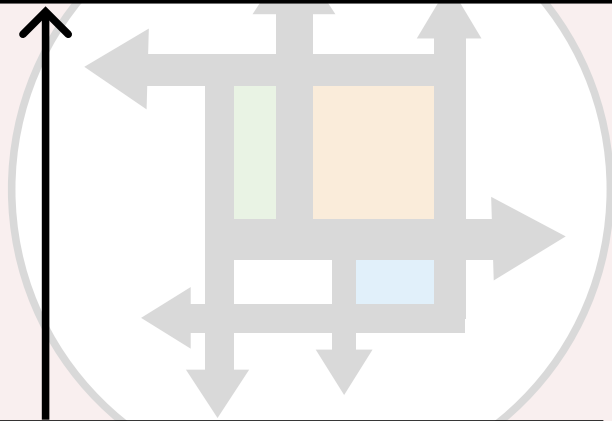
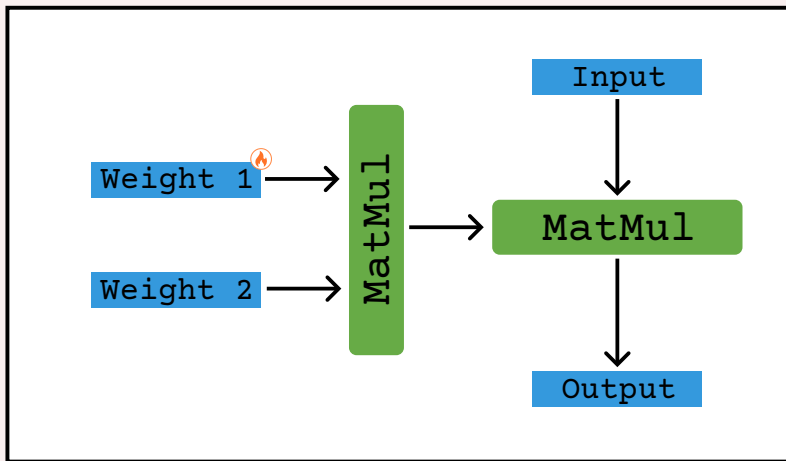
distributed DNN training





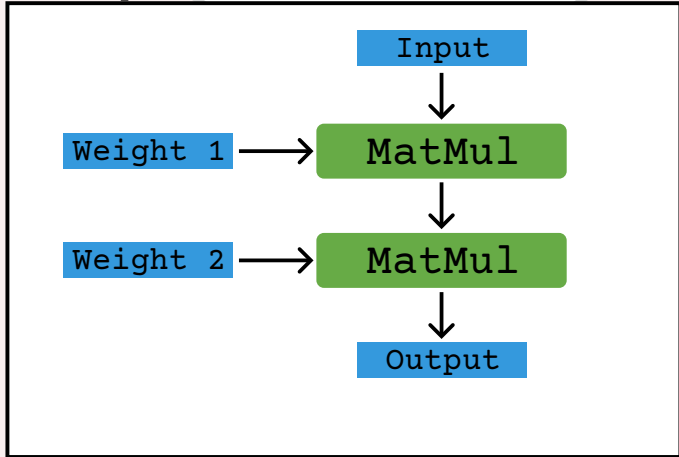
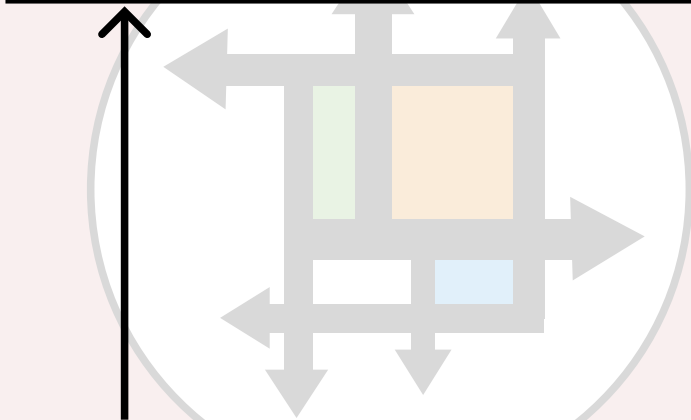
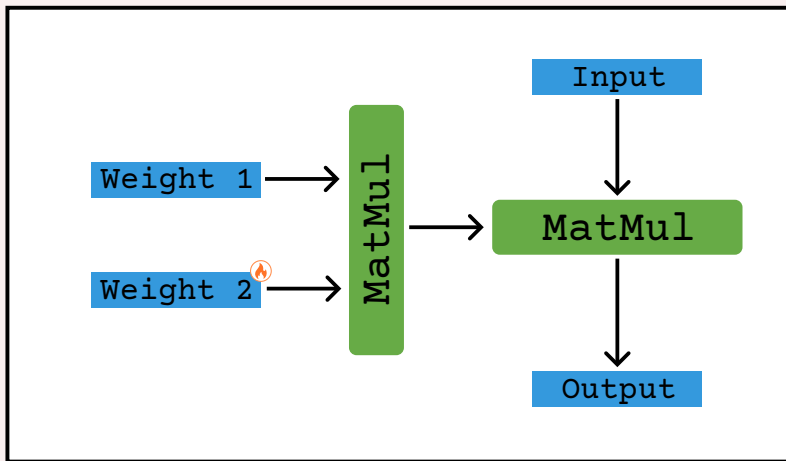
FlexFlow

distributed DNN training



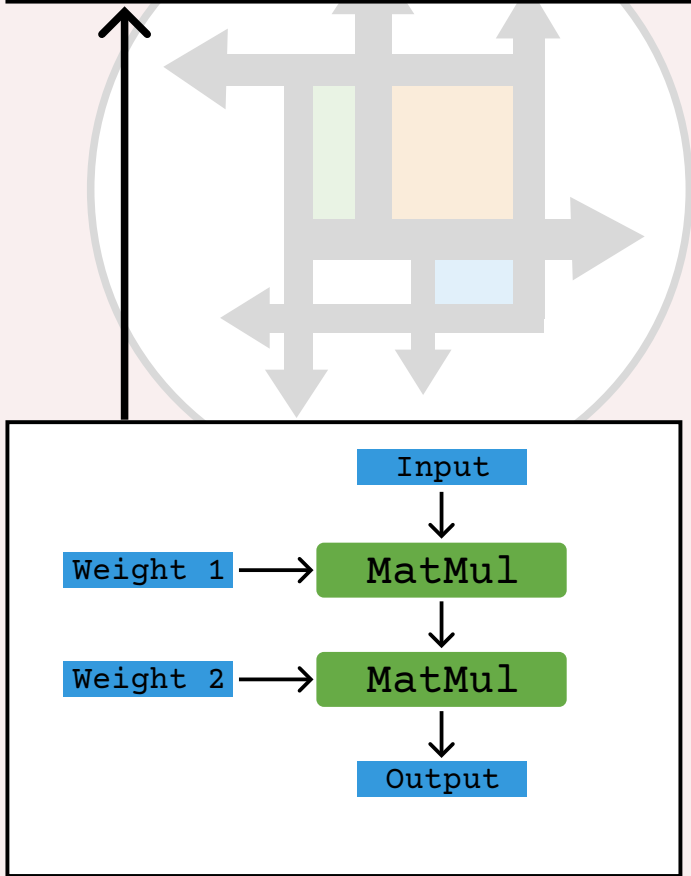
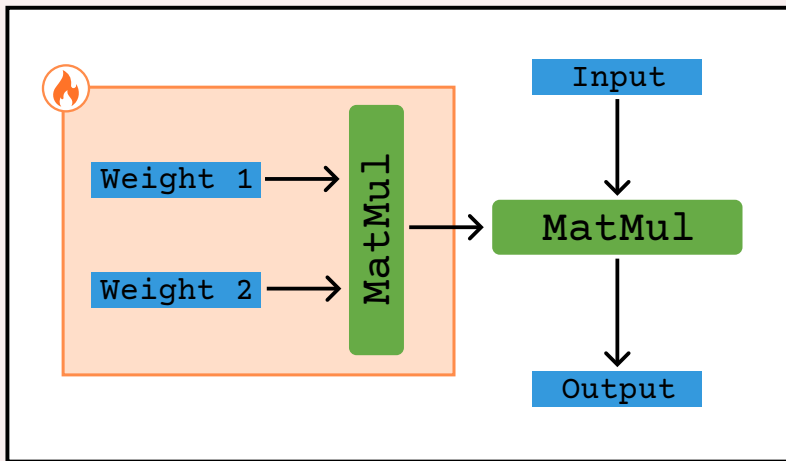
FlexFlow

distributed DNN training



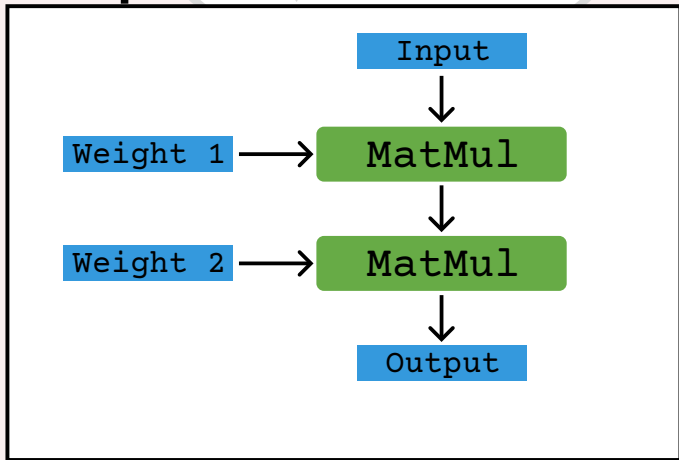
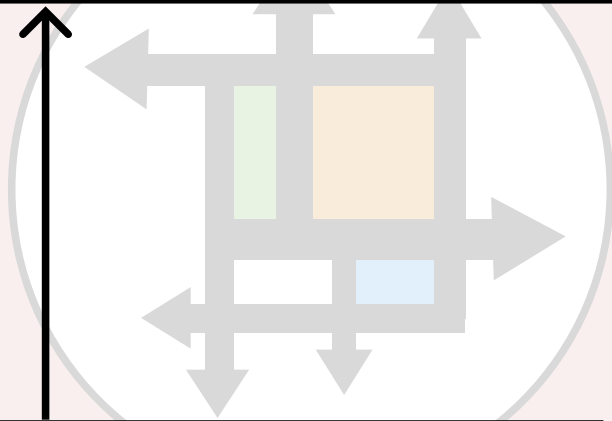
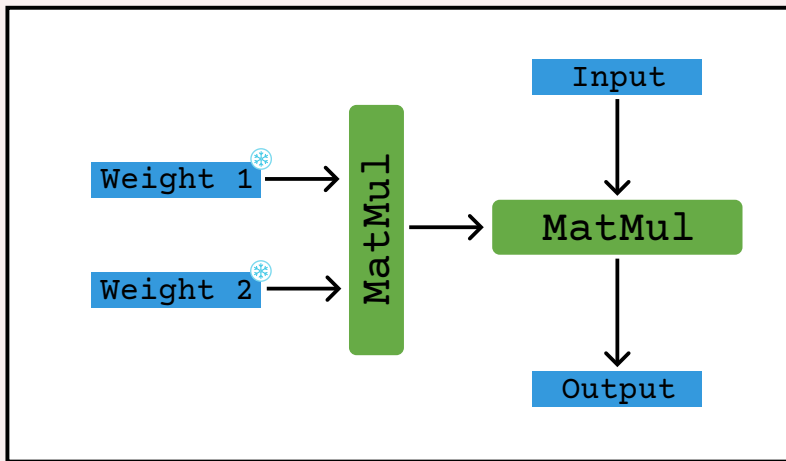
FlexFlow

distributed DNN training



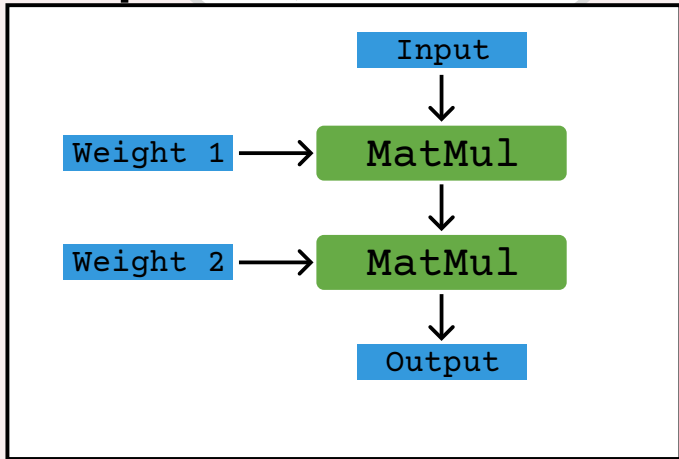
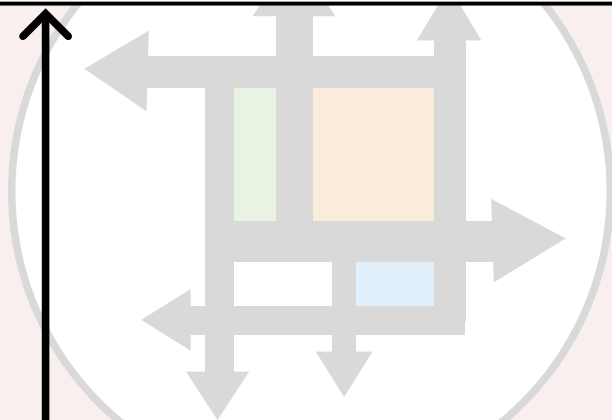
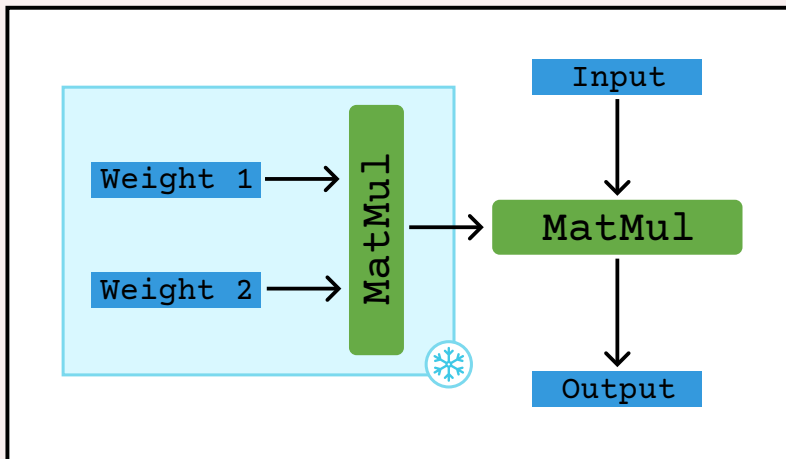
FlexFlow

distributed DNN training



FlexFlow

distributed DNN training



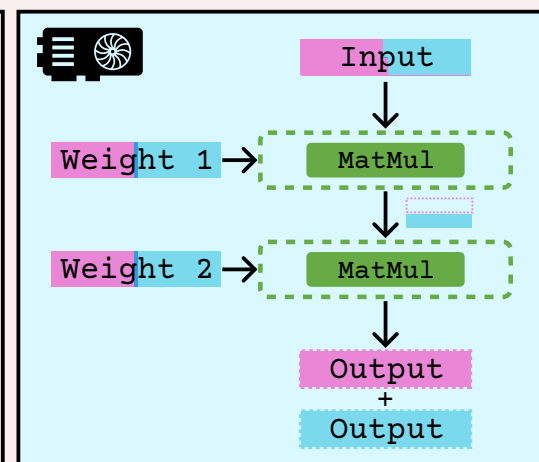
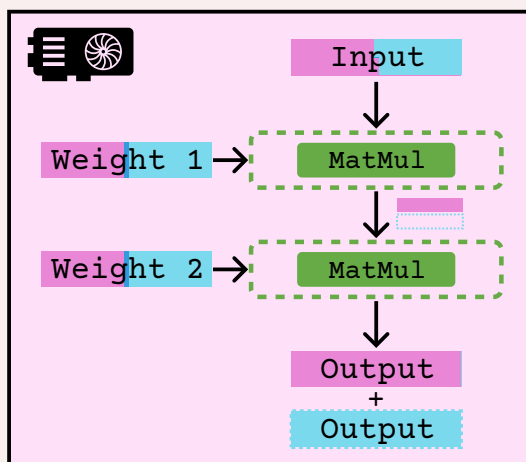
FlexFlow

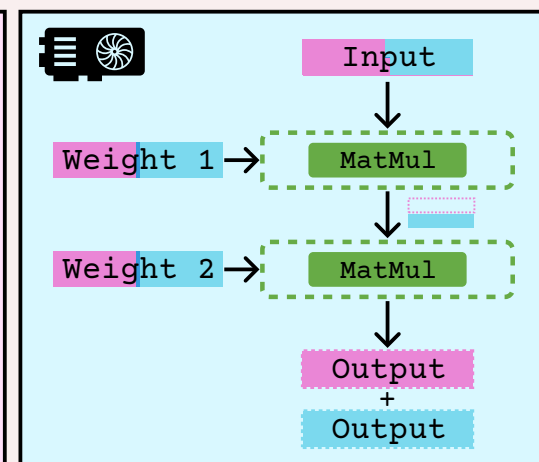
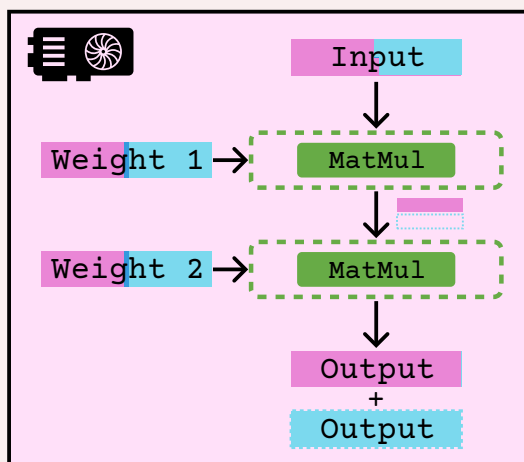
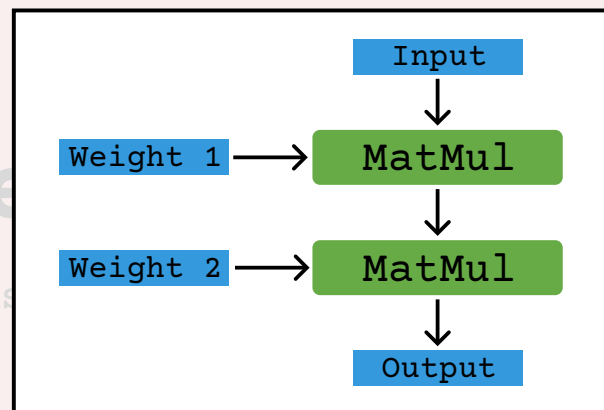
distributed DNN training

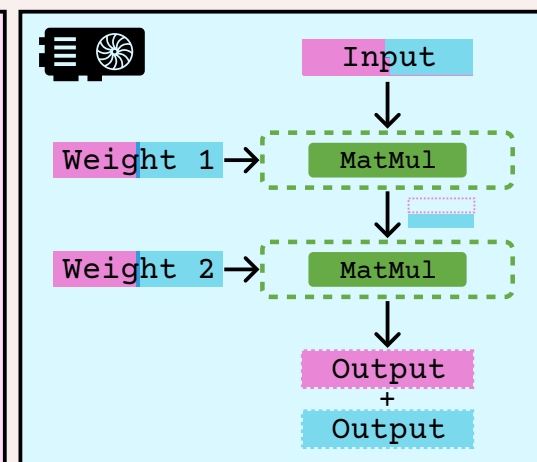
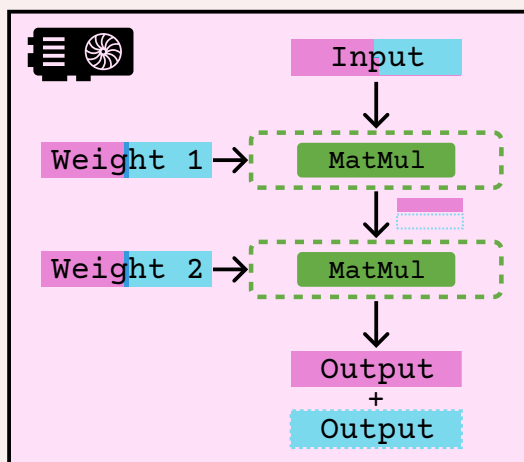
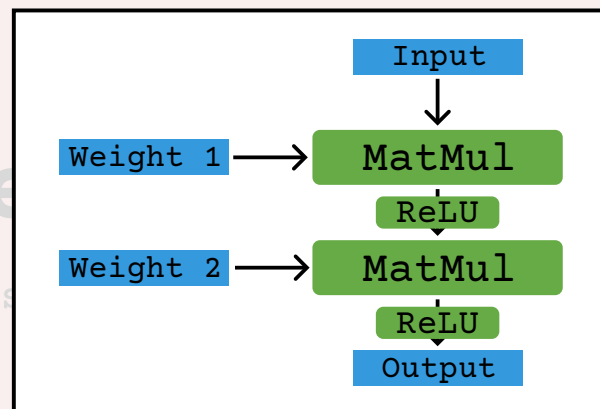


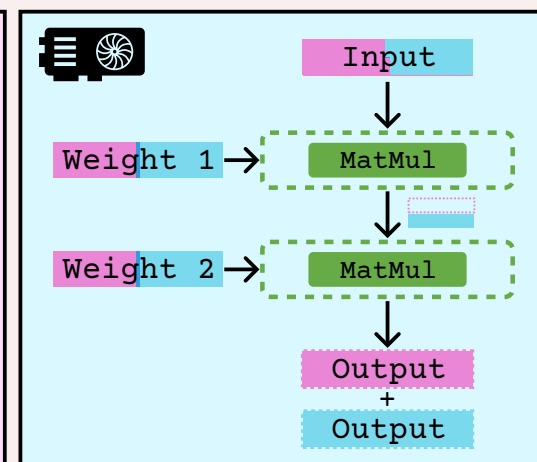
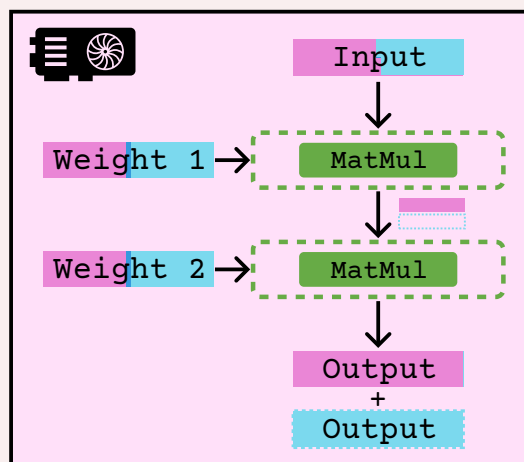
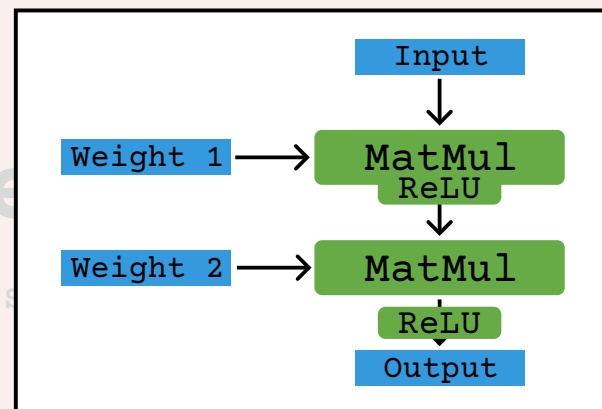
FlexFlow

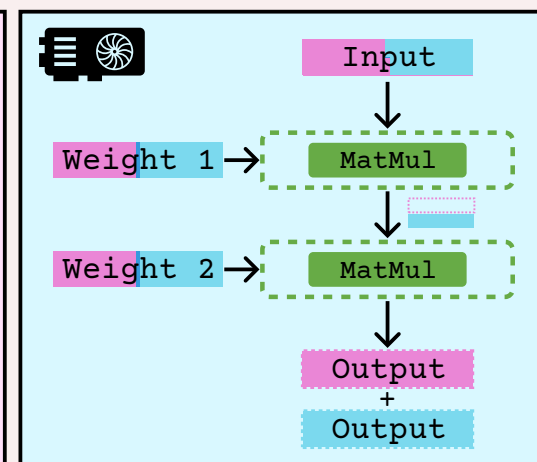
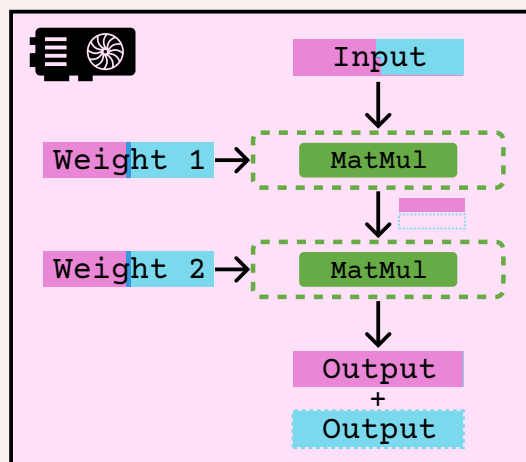
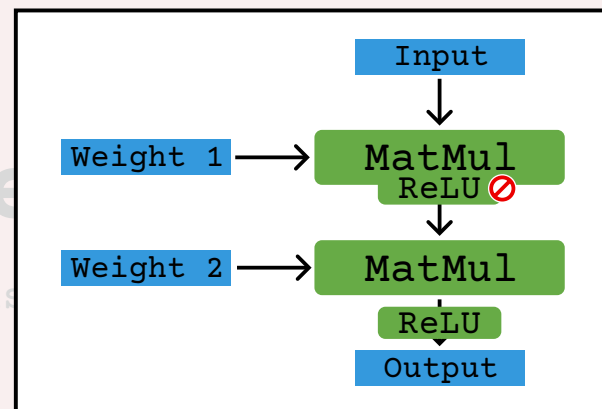
distributed DNN training

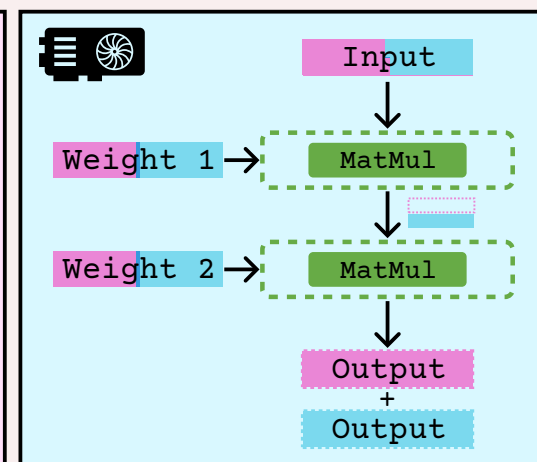
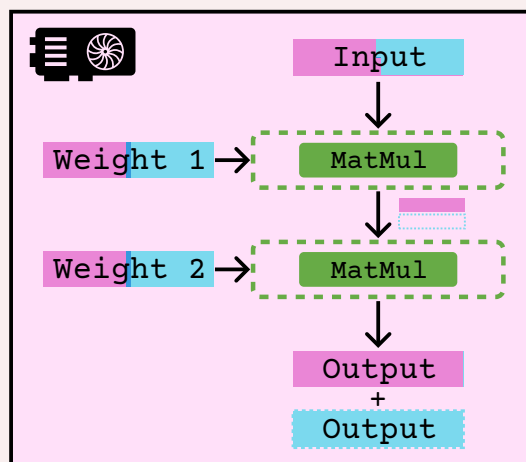
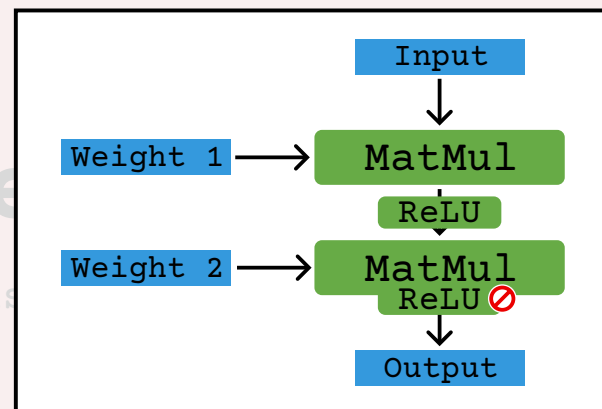















FlexFlow

distributed DNN training



**optimal
training
strategy**

:

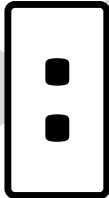
algebraic
transformations

FlexFlow

distributed DNN training



**optimal
training
strategy**

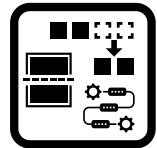


algebraic
transformations

parallelization
— pipeline parallel optimizer
— data parallel buffers
— tensor model parallel activation
— reduction parallel weights
— tensor placement/sharding

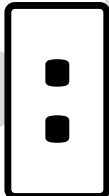


×



FlexFlow
distributed DNN training

**optimal
training
strategy**



FlexFlow

distributed DNN training

algebraic
transformations

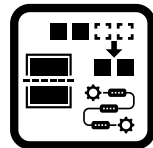
parallelization

— pipeline parallel	optimizer
— data parallel	buffers
— tensor model parallel	activation
— reduction parallel	weights
— tensor placement/sharding	

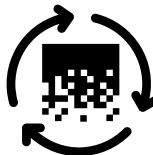
activation
rematerialization



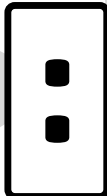
×



×



**optimal
training
strategy**



FlexFlow
distributed DNN training

algebraic
transformations

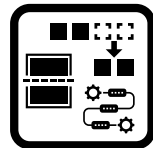
parallelization
— pipeline parallel optimizer
— data parallel buffers
— tensor model parallel activation
— reduction parallel weights
— tensor placement/sharding

activation
rematerialization

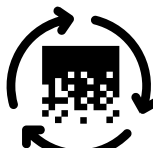
tensor offloading



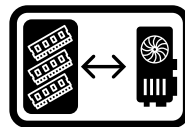
×



×

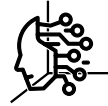


×

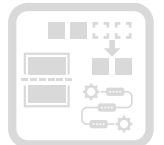


model shape

optimal
training
strategy



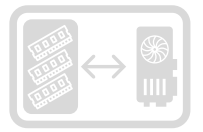
x



x

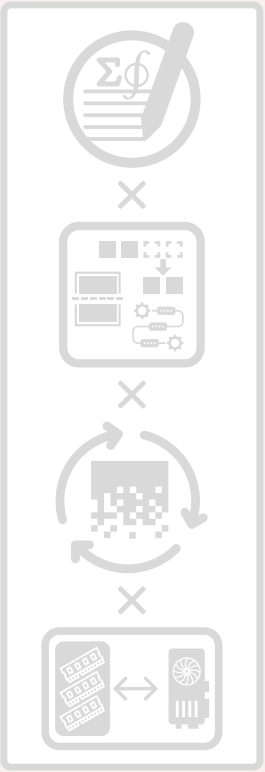
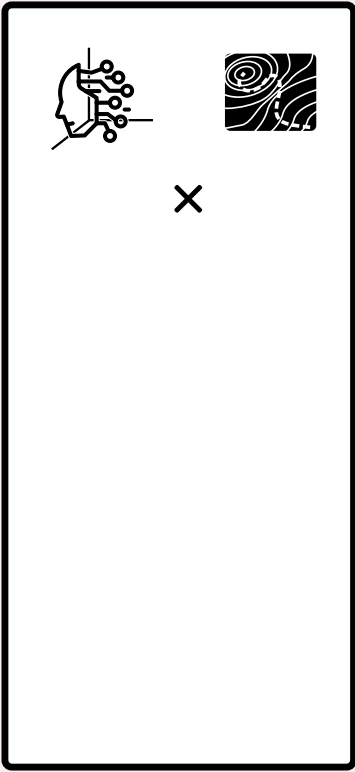
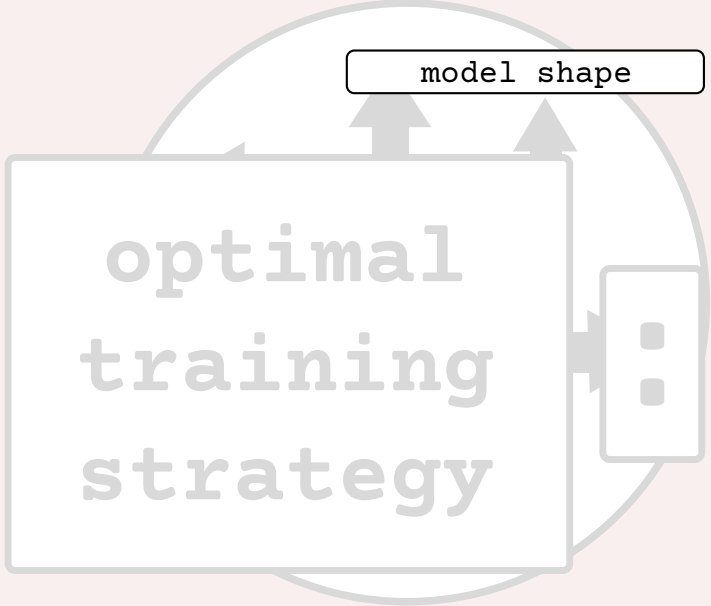


x



model shape

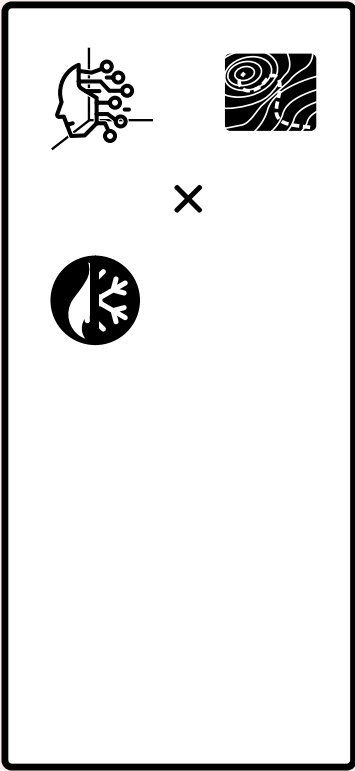
optimizer



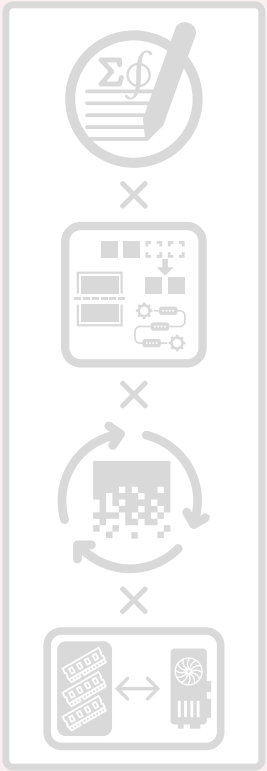
model shape

weight freezing

optimal
training
strategy

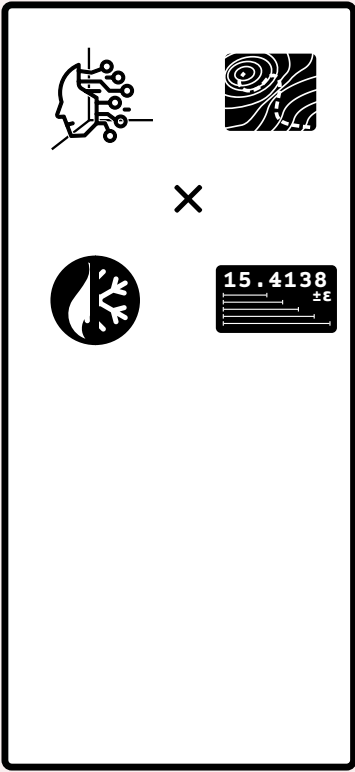


optimizer



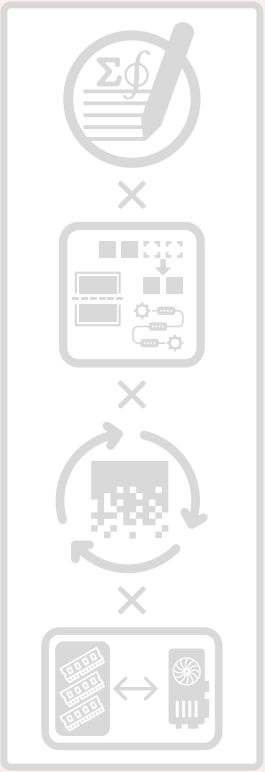
model shape

weight freezing



optimizer

quantization



optimal
training
strategy

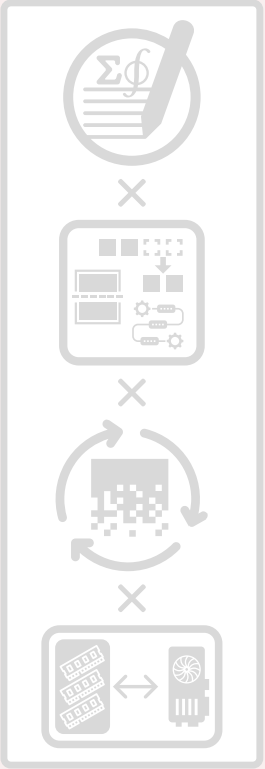
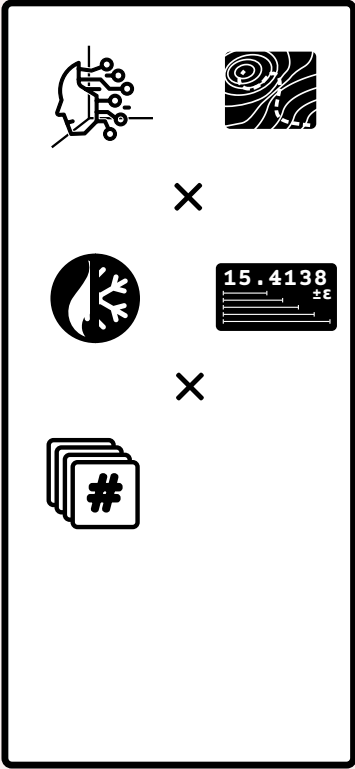
model shape

weight freezing

batch size

optimizer

quantization

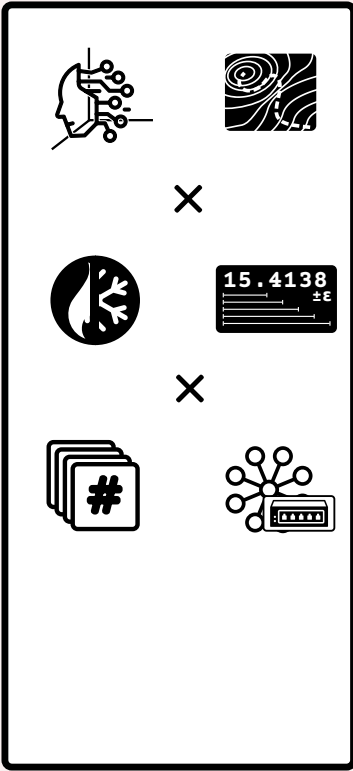


optimal
training
strategy

model shape

weight freezing

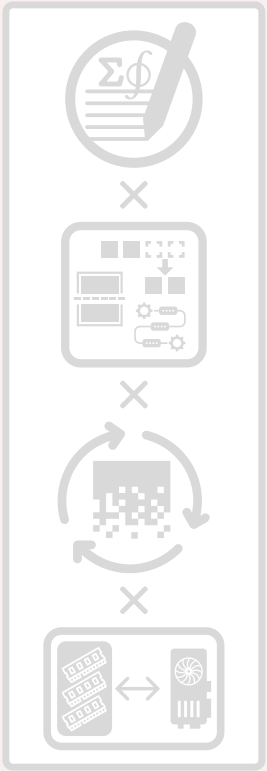
batch size



optimizer

quantization

network topology



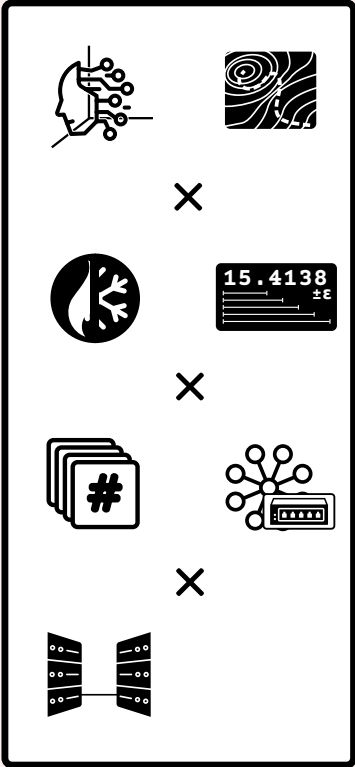
optimal
training
strategy

model shape

weight freezing

batch size

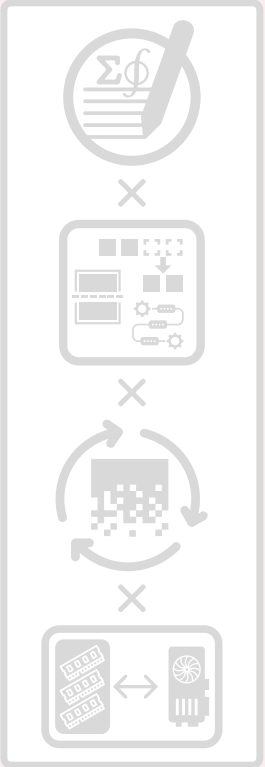
cluster/job size



optimizer

quantization

network topology



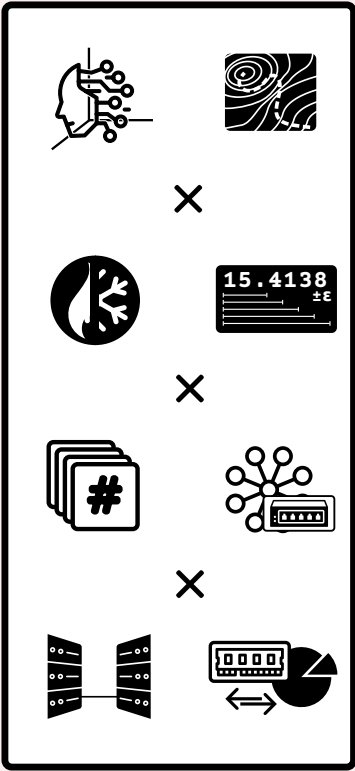
optimal
training
strategy

model shape

weight freezing

batch size

cluster/job size

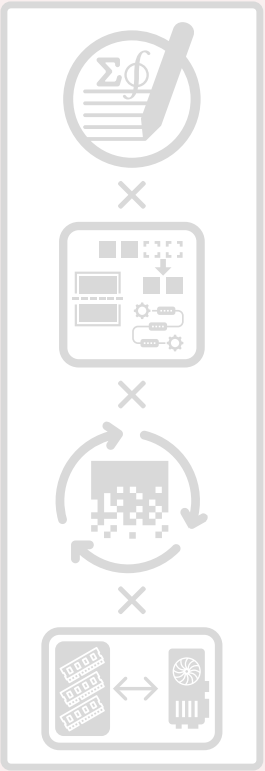


optimizer

quantization

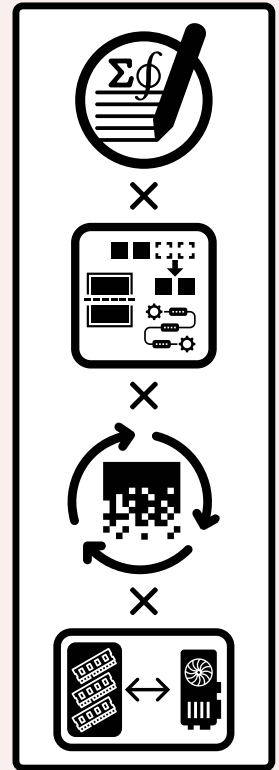
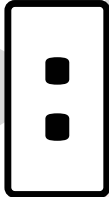
network topology

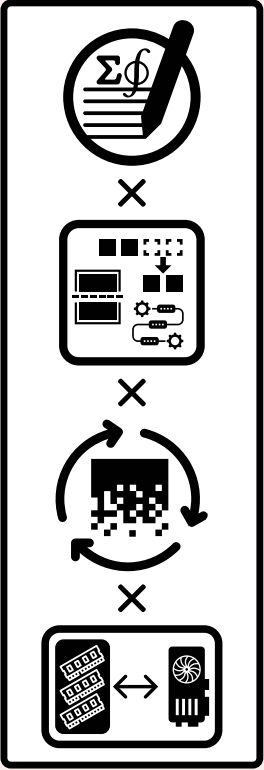
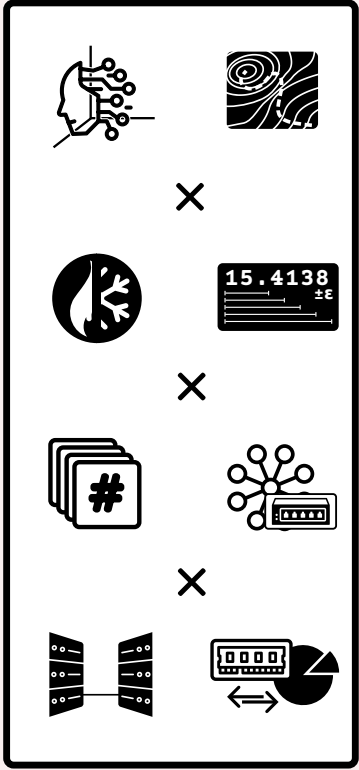
memory size & bandwidth

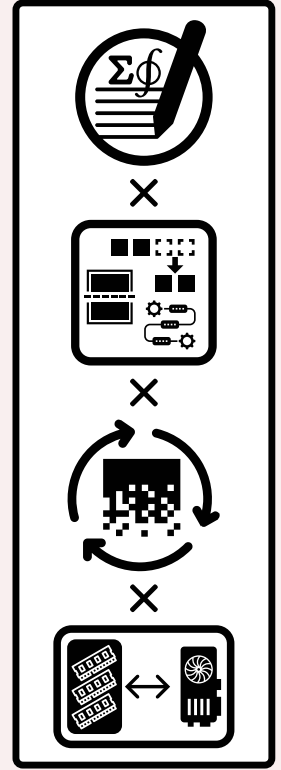
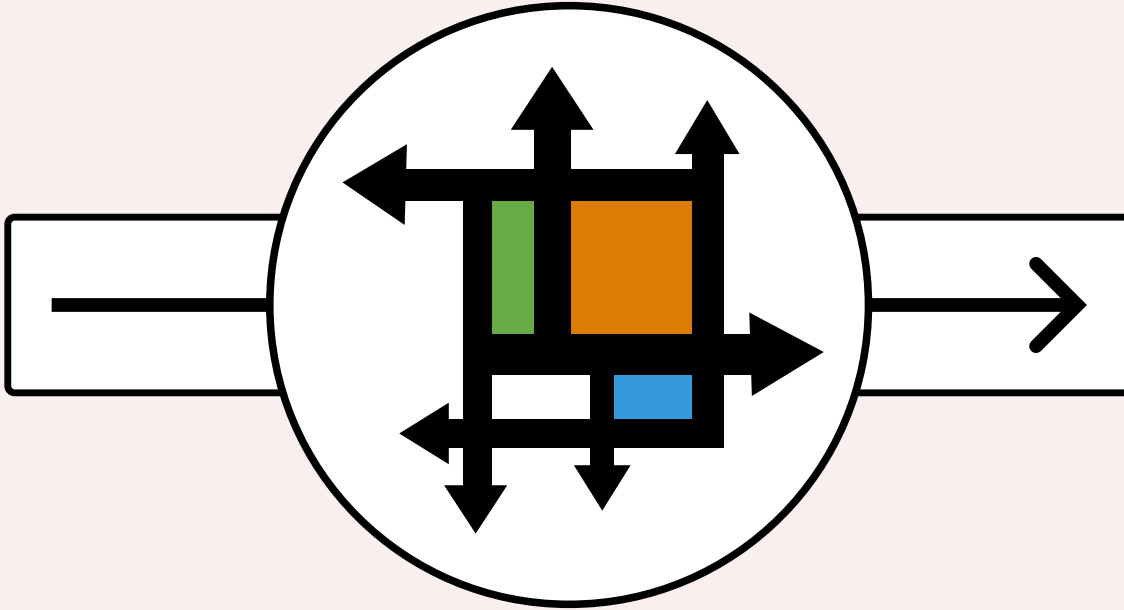
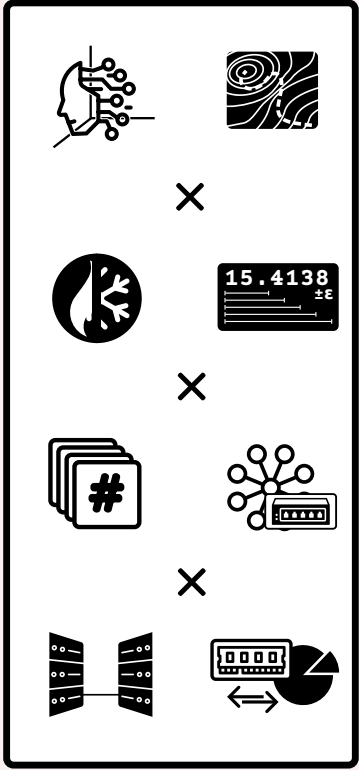


optimal
training
strategy

**optimal
training
strategy**







FlexFlow/SOAP

TASO/Metaflow



2018-2021

FlexFlow/SOAP

TASO/Metaflow

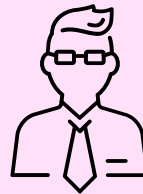


2018-2021

Unity

GraphPipe

Inference



2021-2024

FlexFlow/SOAP

TASO/Metaflow



2018-2021

Unity

GraphPipe

Inference

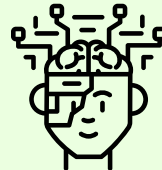


2021-2024

Memory Opt.

Septal

aSP



2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021

Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



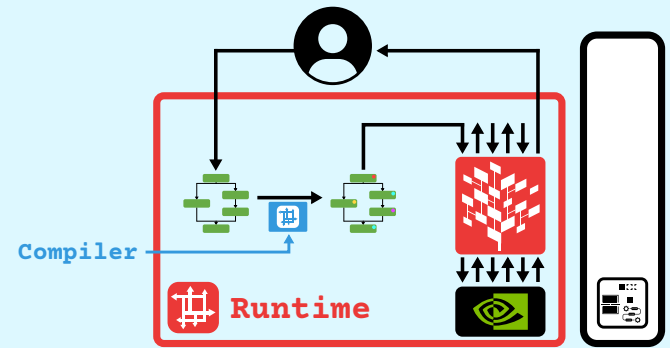
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



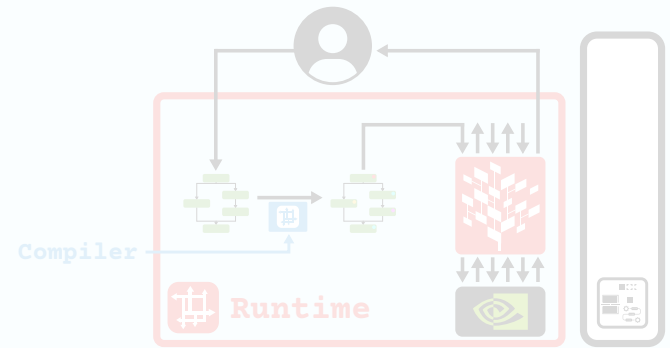
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



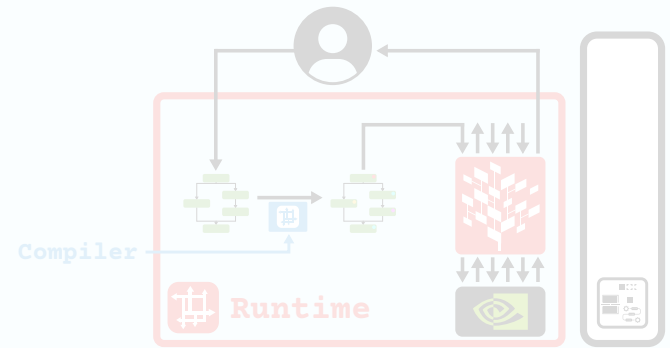
2024-????

FlexFlow/SOAP

TASO/Metaflow



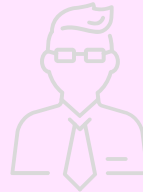
2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

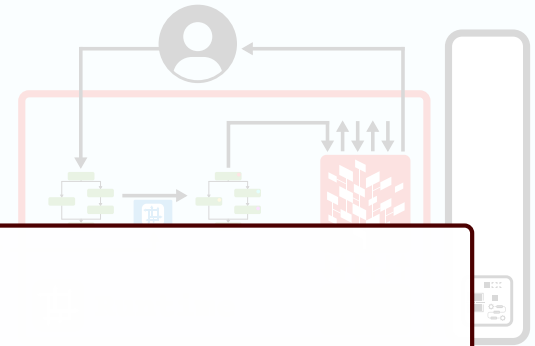
GraphPipe

Inference

Memory Opt.

Septal

aSP



base case: ○



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

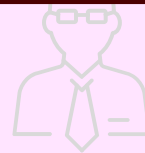
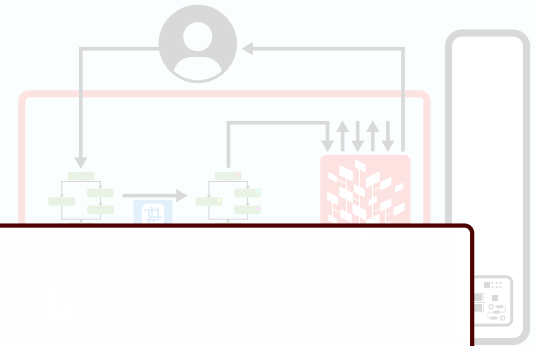
Memory Opt.

Septal

aSP

base case: ○

$$P \left(\begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \rightarrow \circ, \begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \rightarrow \begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \right) =$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

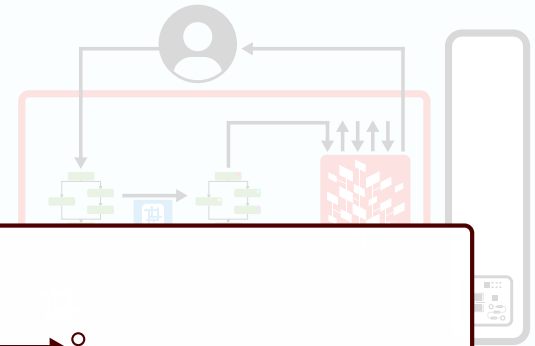
GraphPipe

Inference

Memory Opt.

Septal

aSP

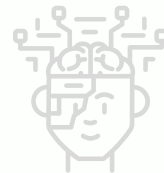


base case: \circ

$$P \left(\begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \rightarrow \circ, \begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \rightarrow \begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \right) = \begin{matrix} \circ \\ \circ \\ \circ \end{matrix} \rightarrow \circ$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

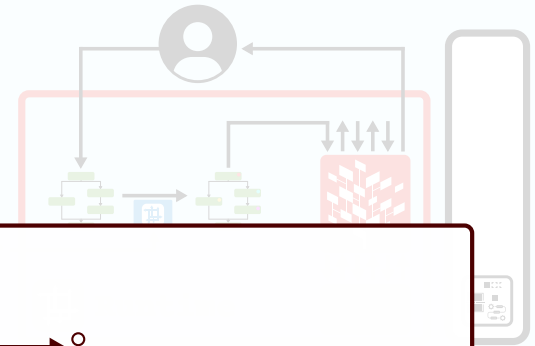
Memory Opt.

Septal

aSP

base case: ○

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

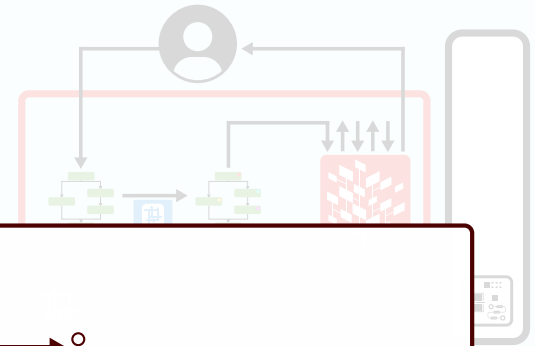
GraphPipe

Inference

Memory Opt.

Septal

aSP



base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) =$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

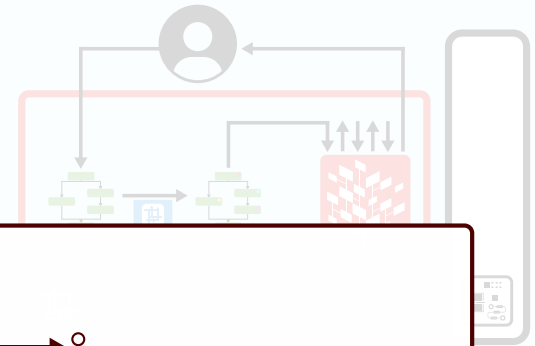
GraphPipe

Inference

Memory Opt.

Septal

aSP



base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \end{array}$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

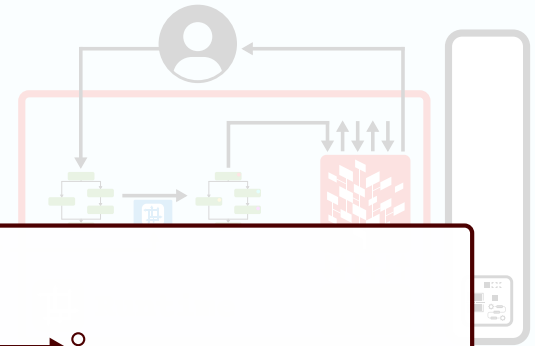
GraphPipe

Inference

Memory Opt.

Septal

aSP



base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array} \quad \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array}$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

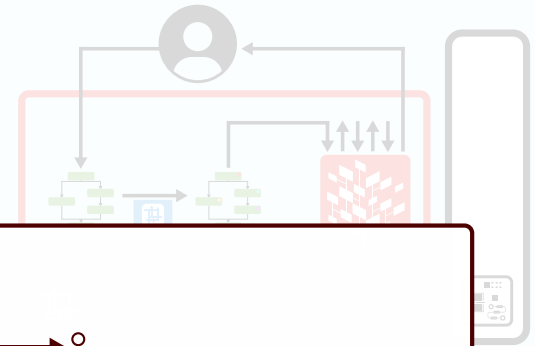
GraphPipe

Inference

Memory Opt.

Septal

aSP



base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



2021-2024



2024-????

FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

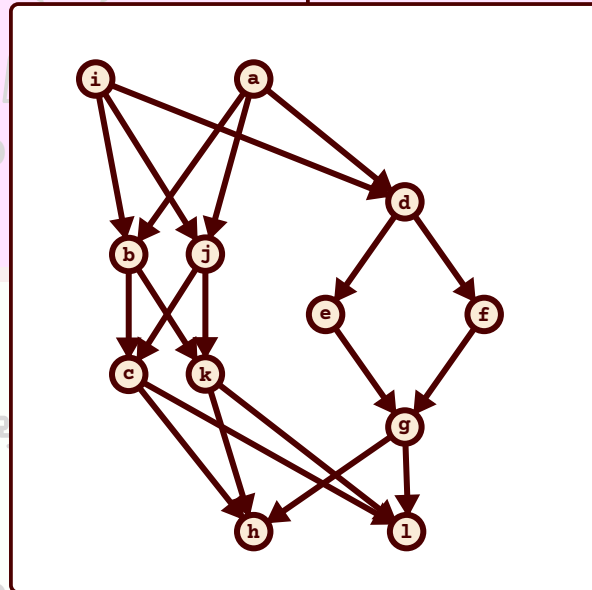
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

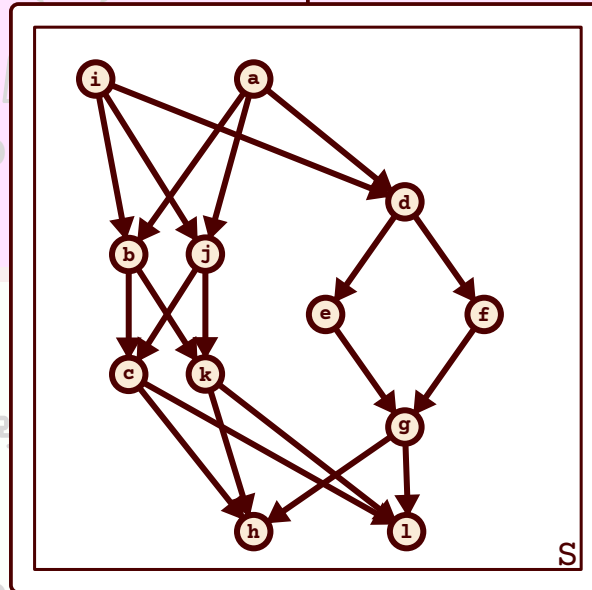
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

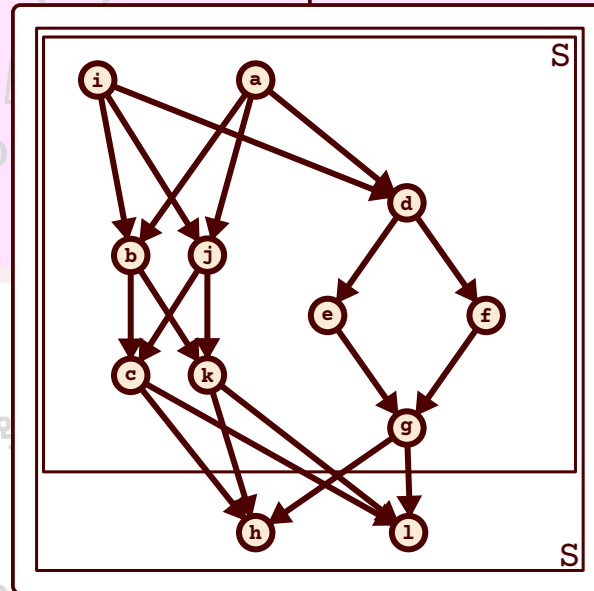
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

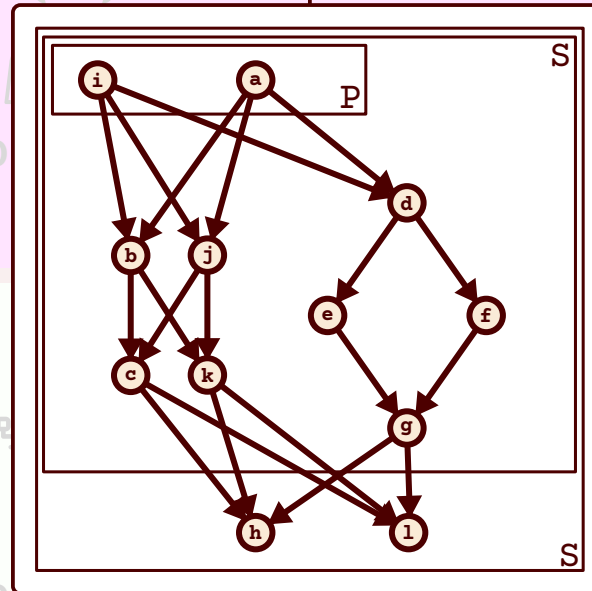
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

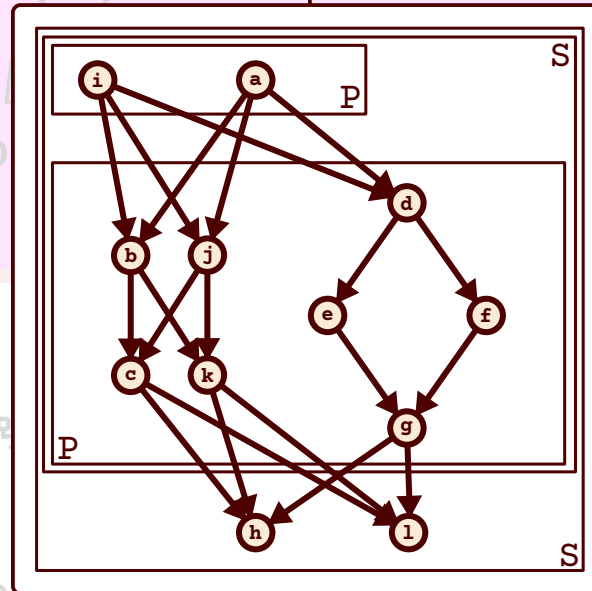
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

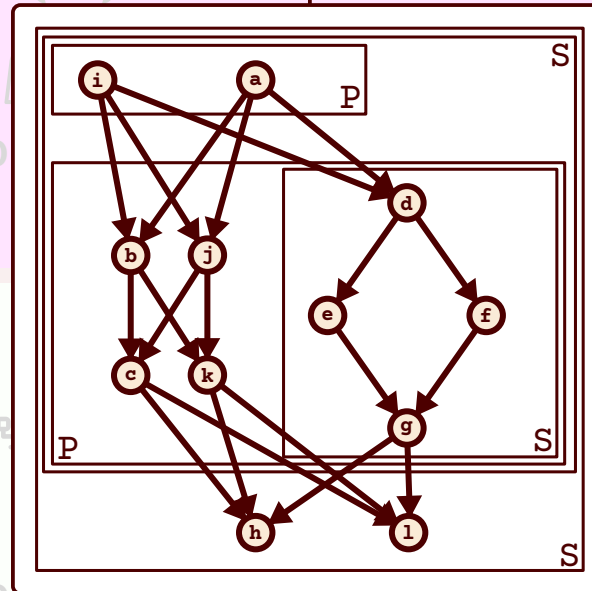
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

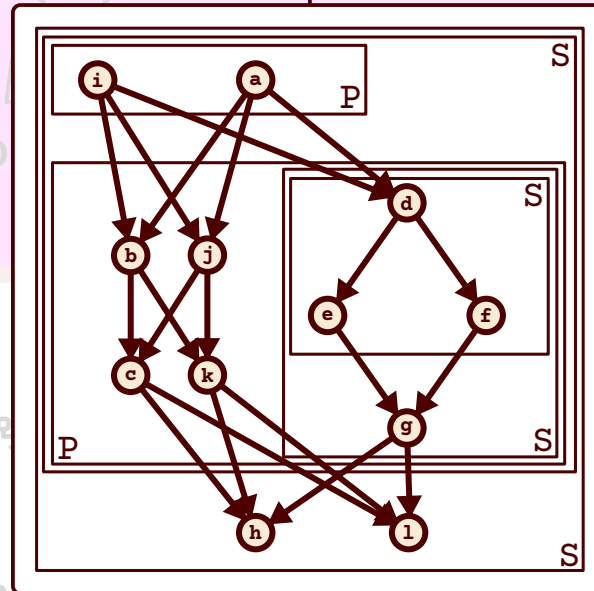
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

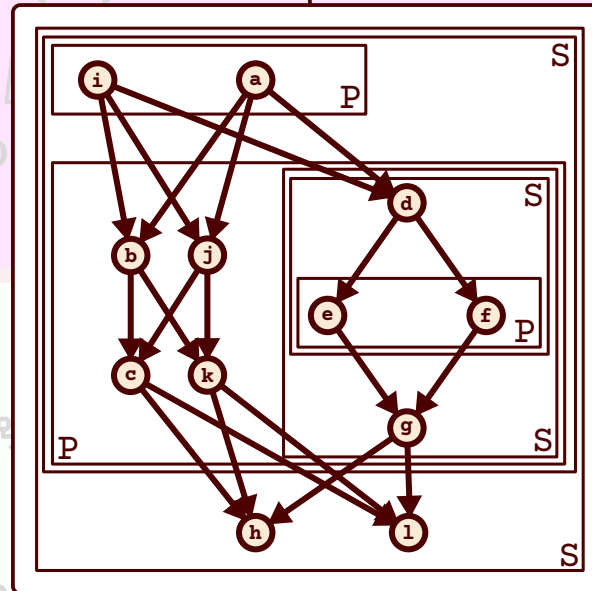
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \circ \rightarrow \circ \circ \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

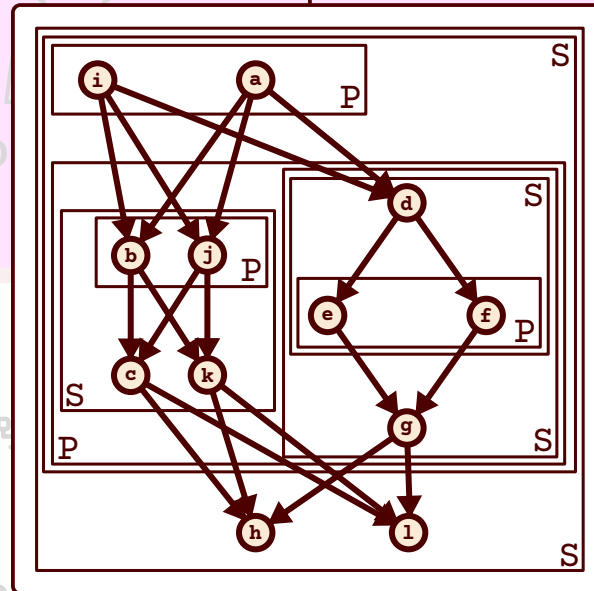
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

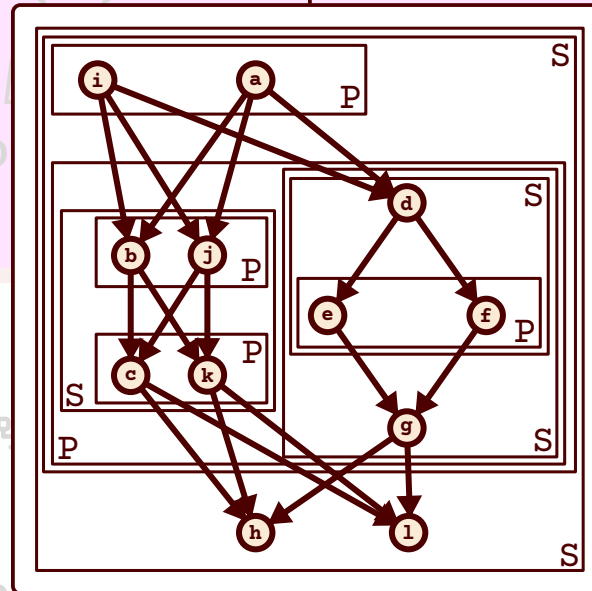
Septal

aSP

base case: \circ

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \circ \\ \circ \rightarrow \circ \\ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$$



FlexFlow/SOAP

TASO/Metaflow

Unity

GraphPipe

Inference

Memory Opt.

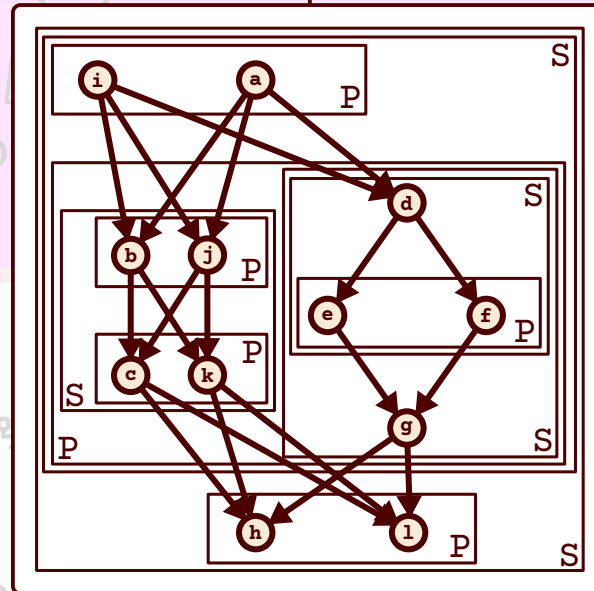
Septal

aSP

base case: ○

$$P \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \\ \circ \circ \rightarrow \circ \end{array}$$

$$S \left(\begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \circ, \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \rightarrow \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \right) = \begin{array}{c} \circ \circ \rightarrow \circ \circ \rightarrow \circ \end{array}$$

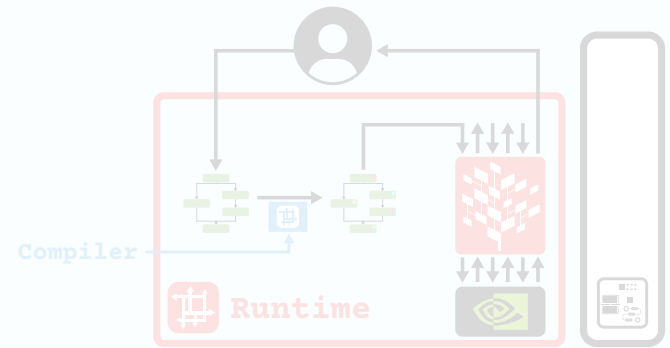


FlexFlow/SOAP

TASO/Metaflow



2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



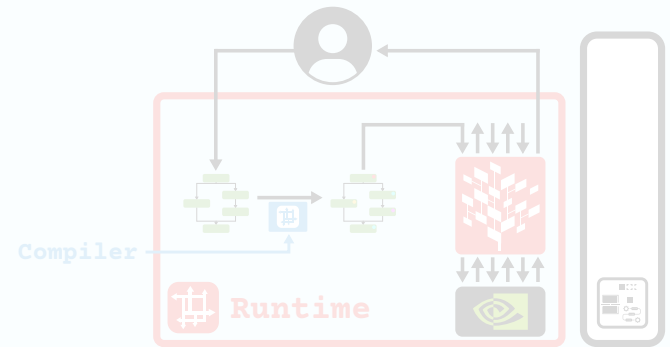
2024-????

FlexFlow/SOAP

TASO/Metaflow



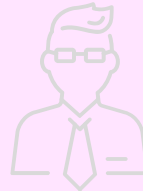
2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



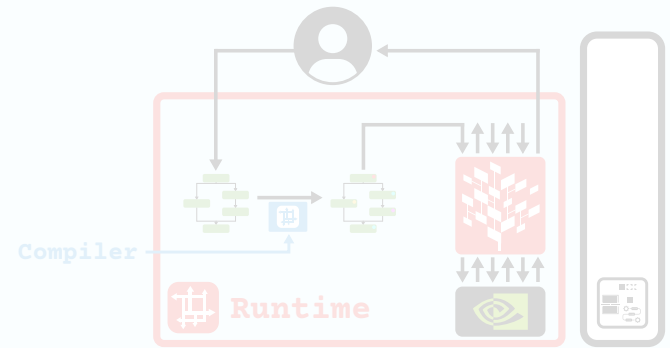
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



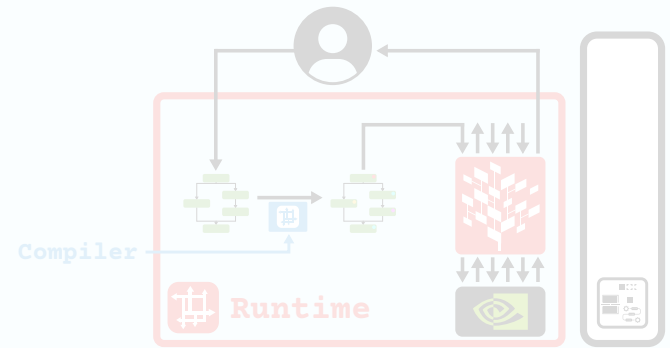
2024-????

FlexFlow/SOAP

TASO/Metaflow



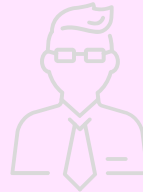
2018-2021



Unity

GraphPipe

Inference



2021-2024

Memory Opt.

Septal

aSP



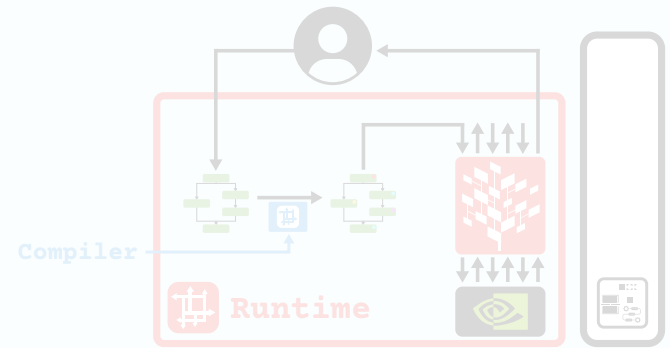
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



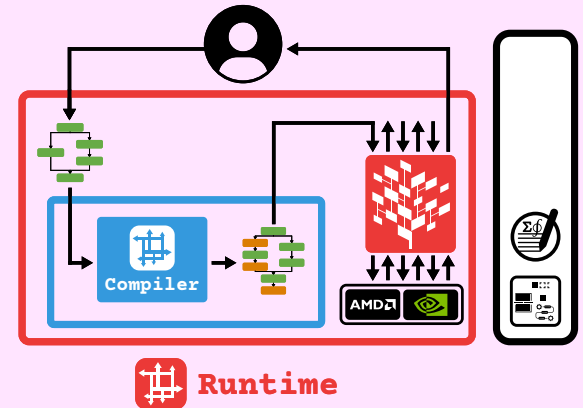
Unity

GraphPipe

Inference



2021-2024



Memory Opt.

Septal

aSP



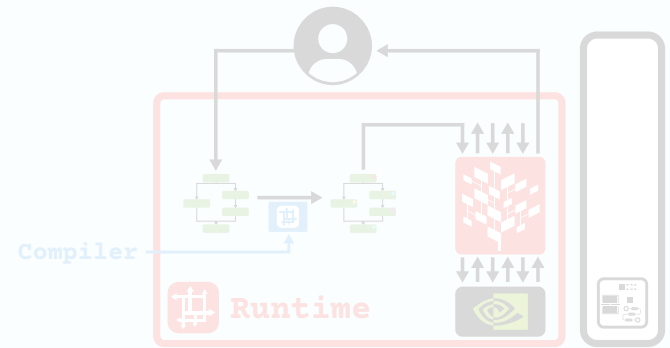
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



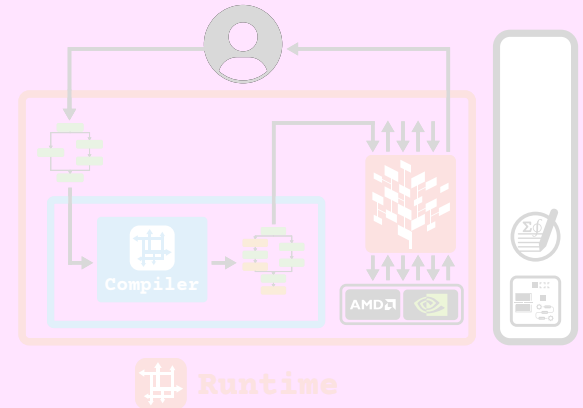
Unity

GraphPipe

Inference



2021-2024

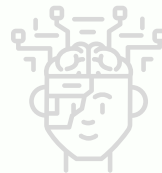


Compiler Refactor

Memory Opt.

Septal

aSP



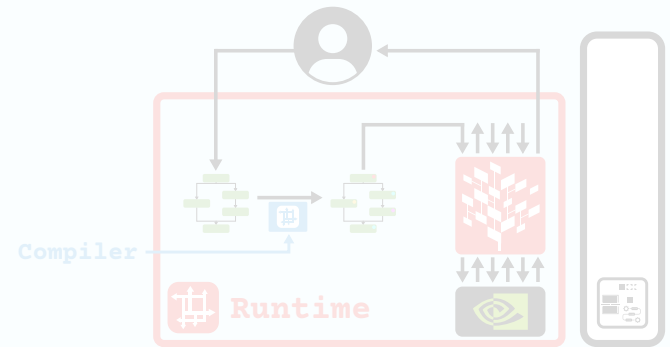
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



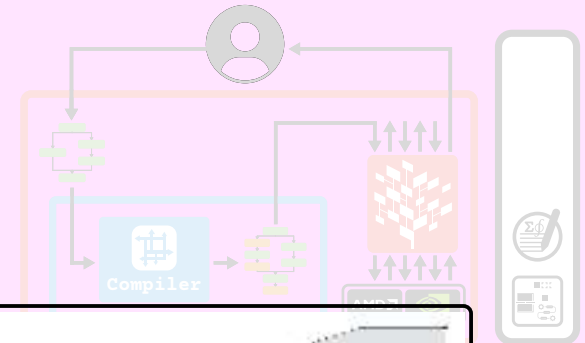
Unity

GraphPipe

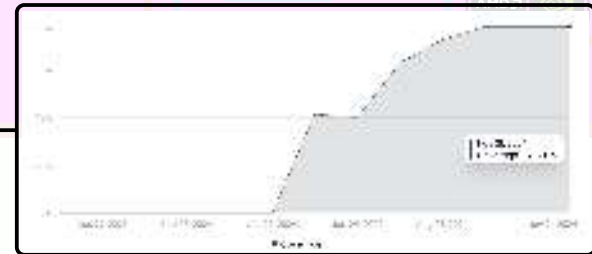
Inference



2021-2024



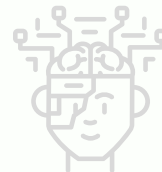
Compiler Refactor



Memory Opt.

Septal

aSP



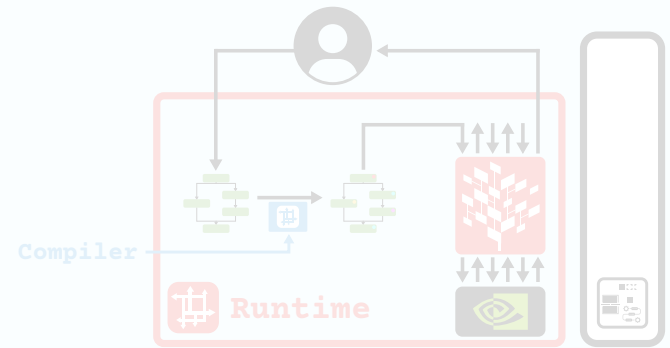
2024-????

FlexFlow/SOAP

TASO/Metaflow



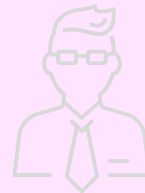
2018-2021



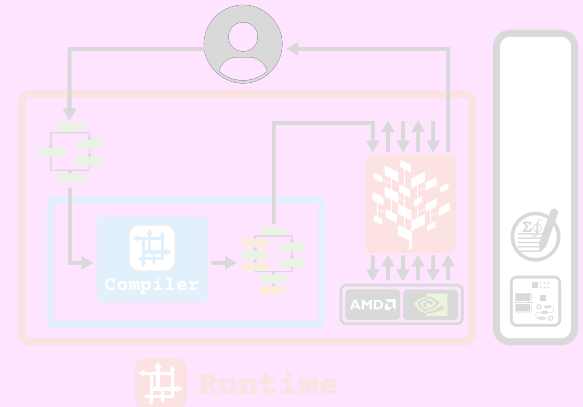
Unity

GraphPipe

Inference



2021-2024

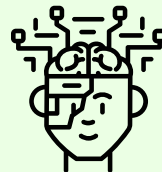


Compiler Refactor

Memory Opt.

Septal

aSP



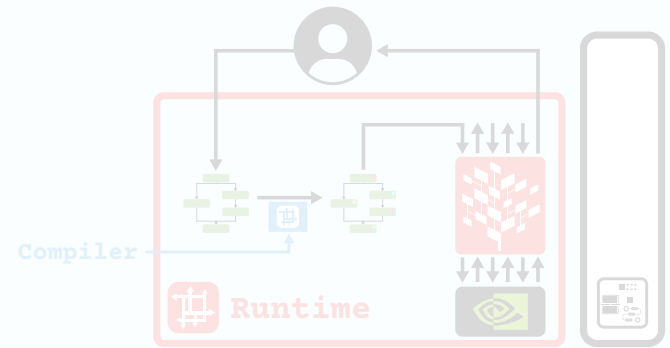
2024-????

FlexFlow/SOAP

TASO/Metaflow



2018-2021



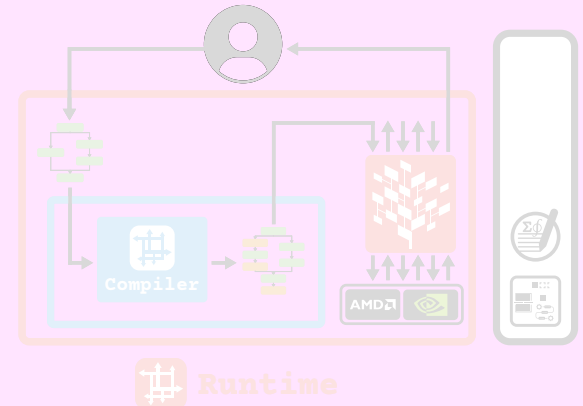
Unity

GraphPipe

Inference



2021-2024



Compiler Refactor

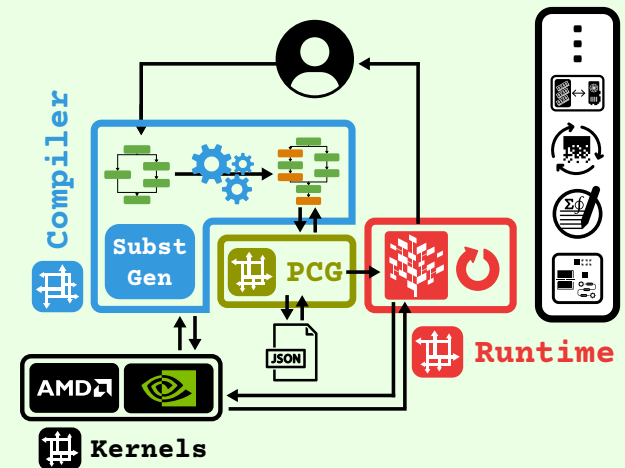
Memory Opt.

Septal

aSP



2024-????

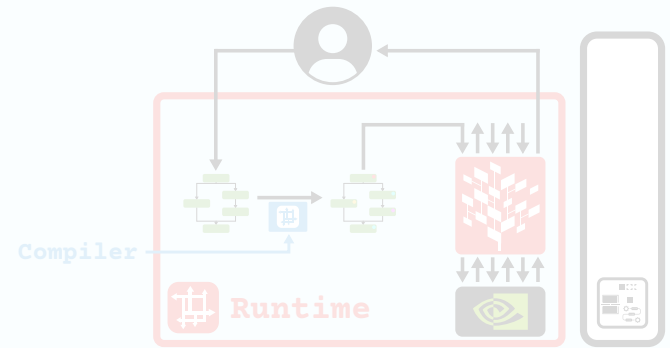


FlexFlow/SOAP

TASO/Metaflow



2018-2021



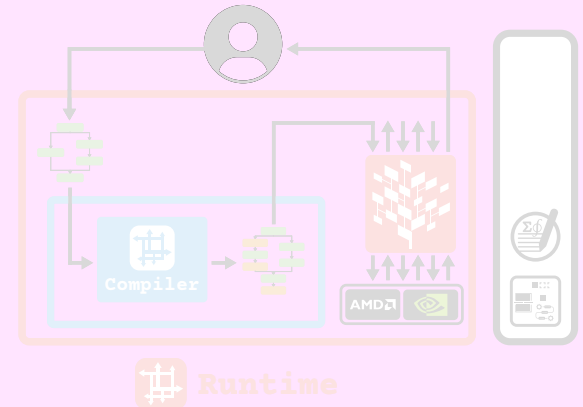
Unity

GraphPipe

Inference



2021-2024



Compiler Refactor

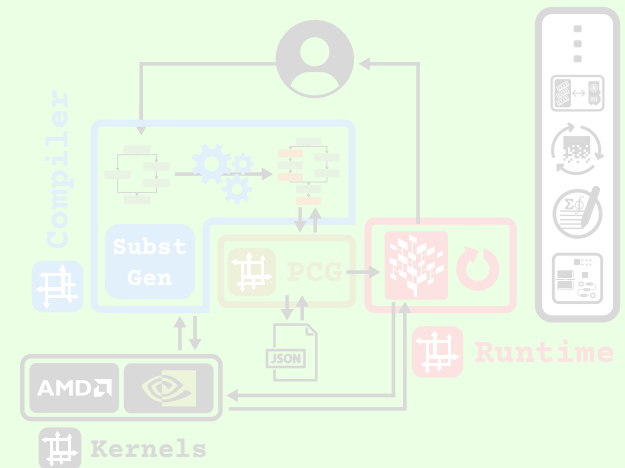
Memory Opt.

Septal

aSP



2024-????

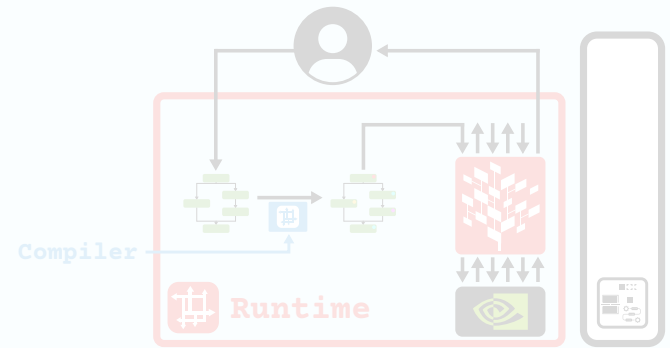


FlexFlow/SOAP

TASO/Metaflow



2018-2021



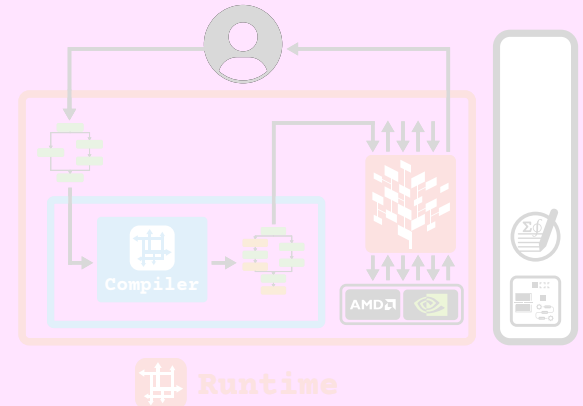
Unity

GraphPipe

Inference



2021-2024



Compiler Refactor

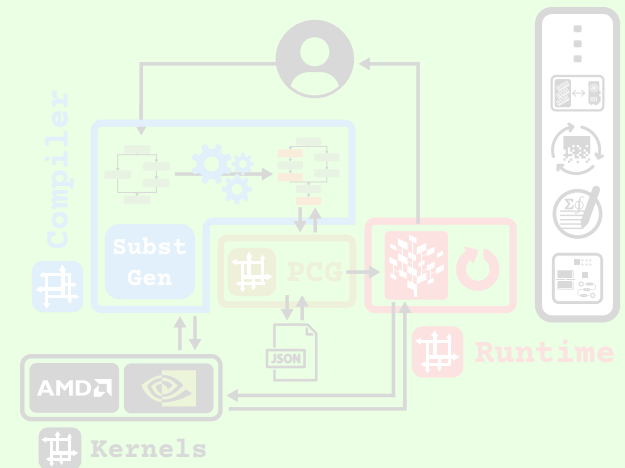
Memory Opt.

Septal

aSP



2024-????

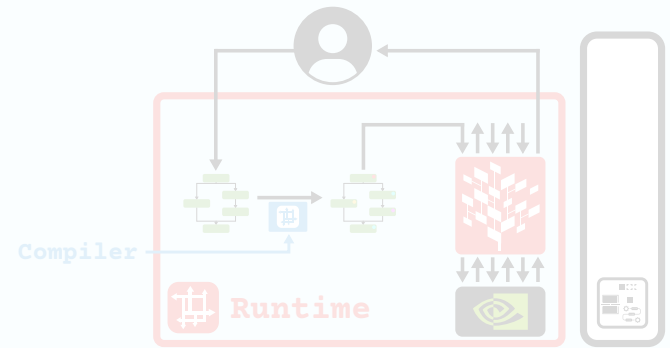


FlexFlow/SOAP

TASO/Metaflow



2018-2021



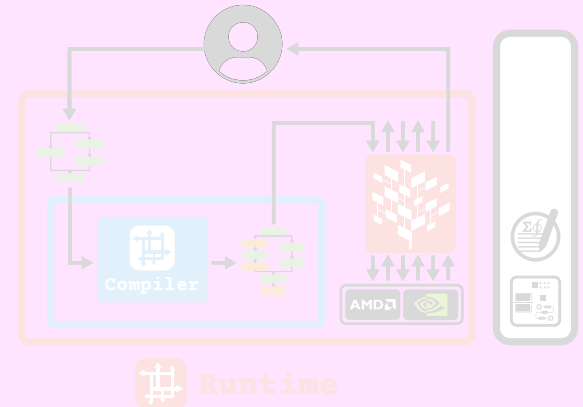
Unity

GraphPipe

Inference



2021-2024



Compiler Refactor

Memory Opt.

Septal

aSP



2024-????

