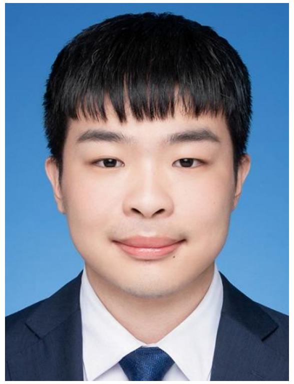
南京大學
NANJING UNIVERSITY

CVPR Nashville JUNE 11-15, 2025

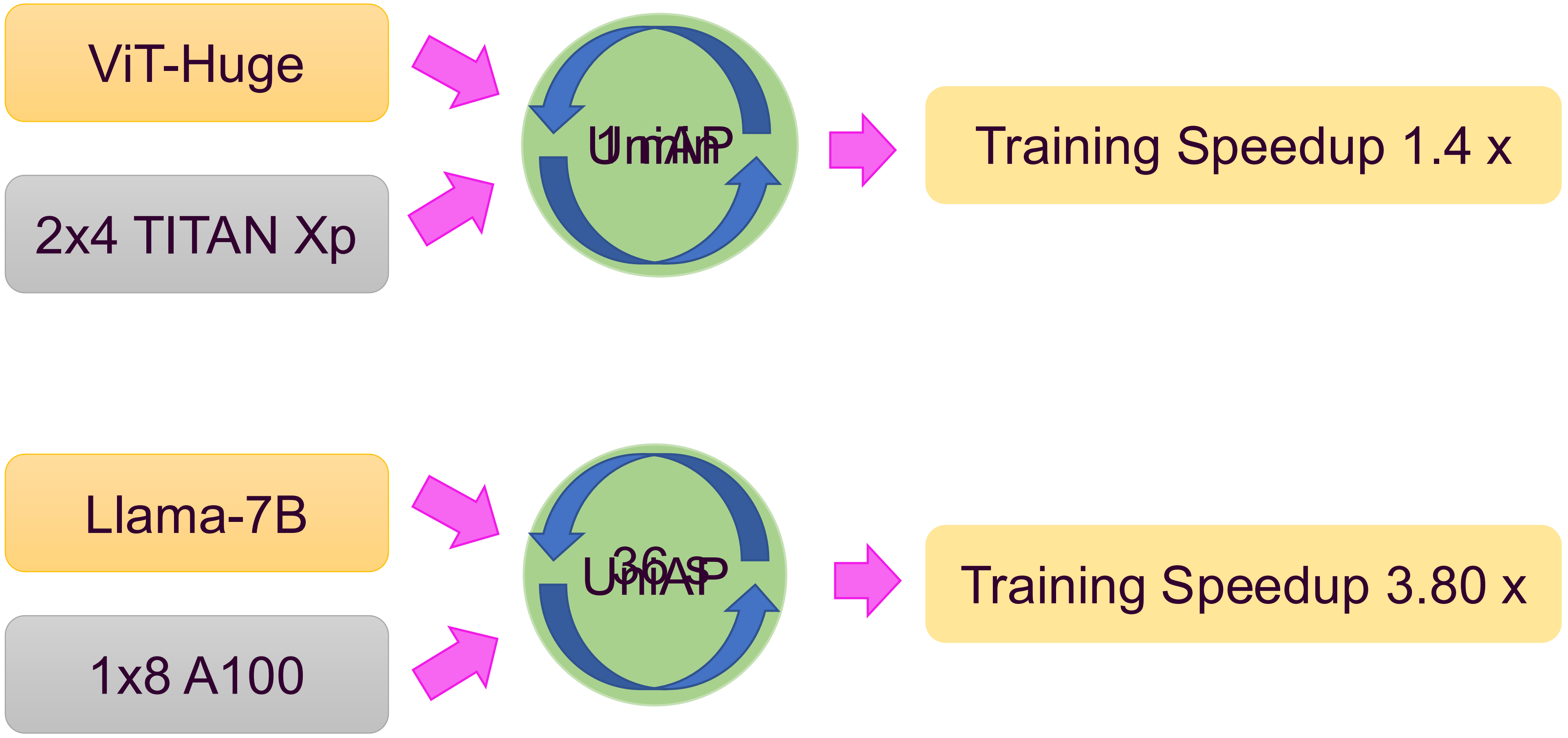
UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Quadratic Programming


Hao Lin[*] Ke Wu[*] Jie Li[*] Jun Li Wu-Jun Li[†]

CVPR 2025 Award Candidate

What is UniAP?

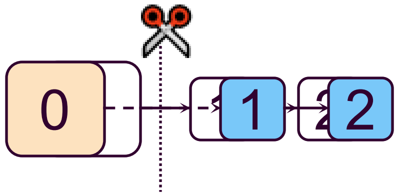


Contents

Contents

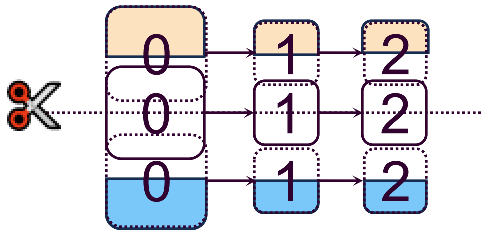
1.1 Parallel Strategies

- *Parallel strategies* have been proposed to accelerate distributed learning.



Inter-layer parallelism strategies

- Pipeline Parallelism（PP）
- ...

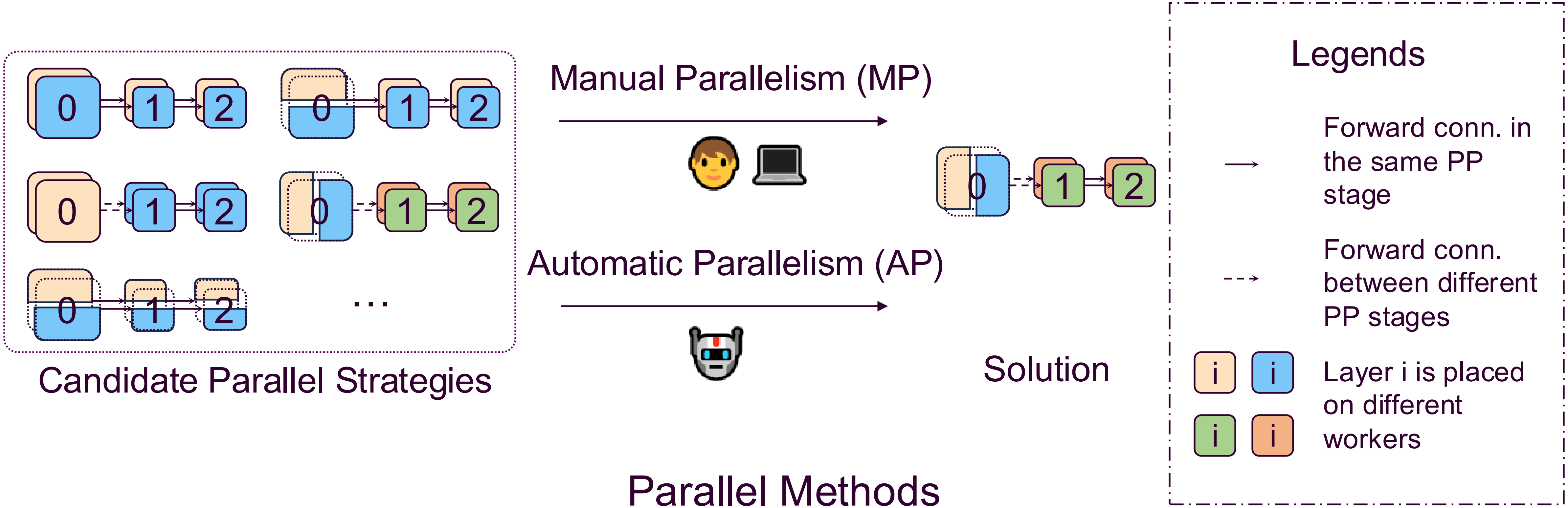


Intra-layer parallelism strategies

- Data Parallelism（DP）
- Tensor Parallelism（TP）
- Fully Sharded Data Parallelism（FSDP）
- ...

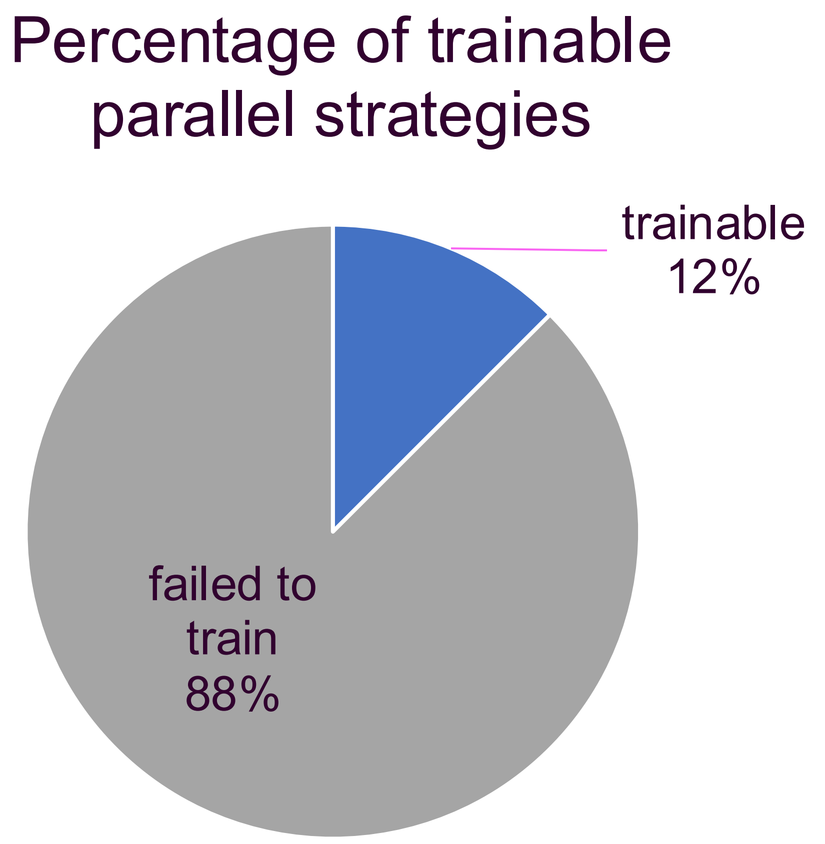
1.2 Parallel Methods

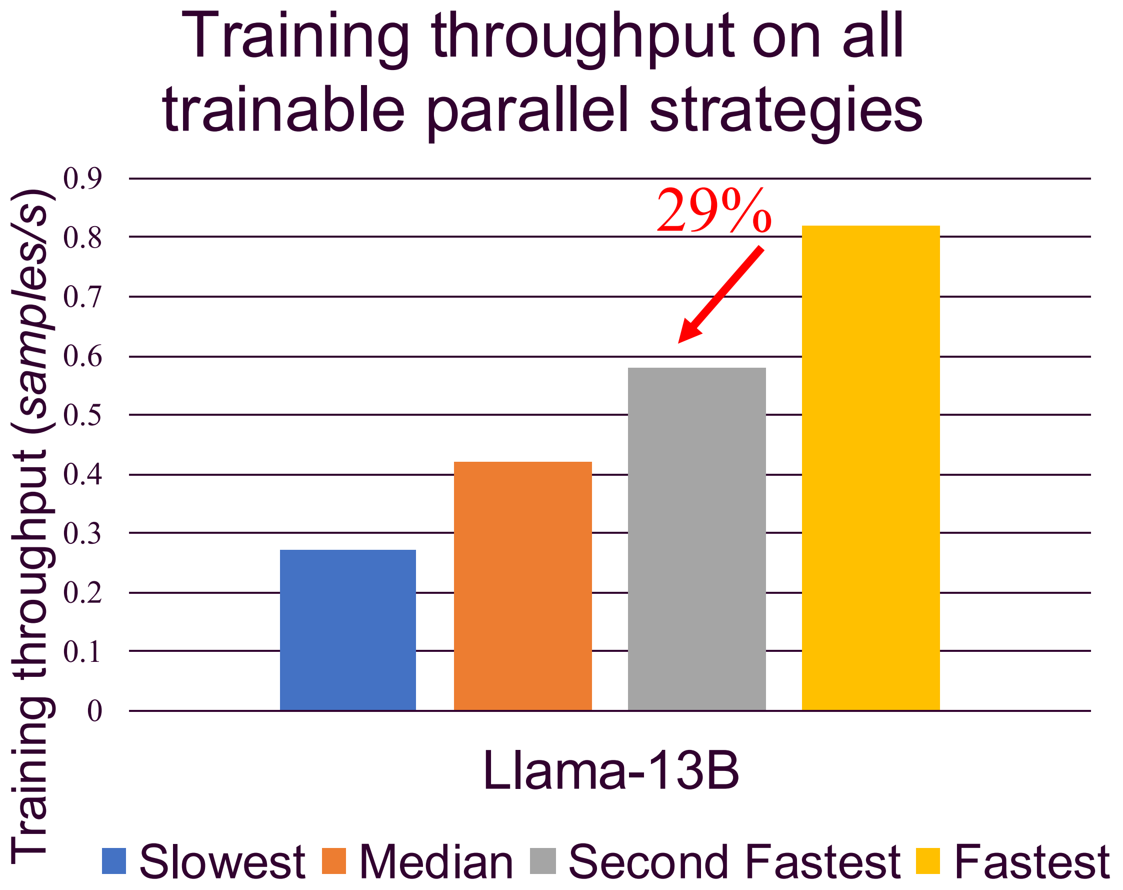
- The *parallel method* in distributed learning typically adopts one *parallel strategy* or a combination of several *parallel strategies.*



Parallel Methods

1.3 Why AP, rather than MP?

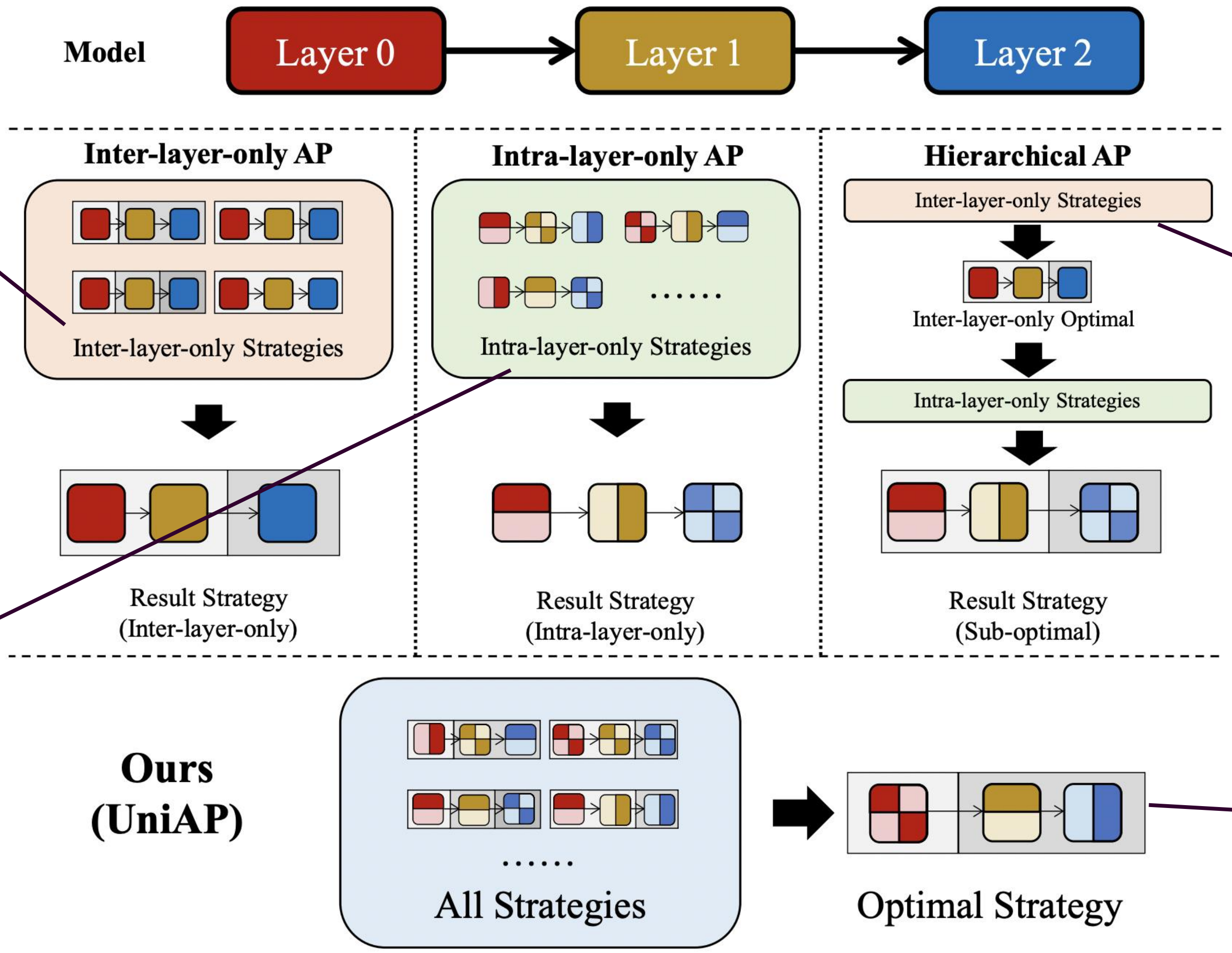




Statistics on the Candidate Parallel Strategies for Llama-13B on Megatron

1.4 Why UniAP, rather than other AP methods?

NANJING UNIVERSITY



PP only

DP, TP, FSDP, … only

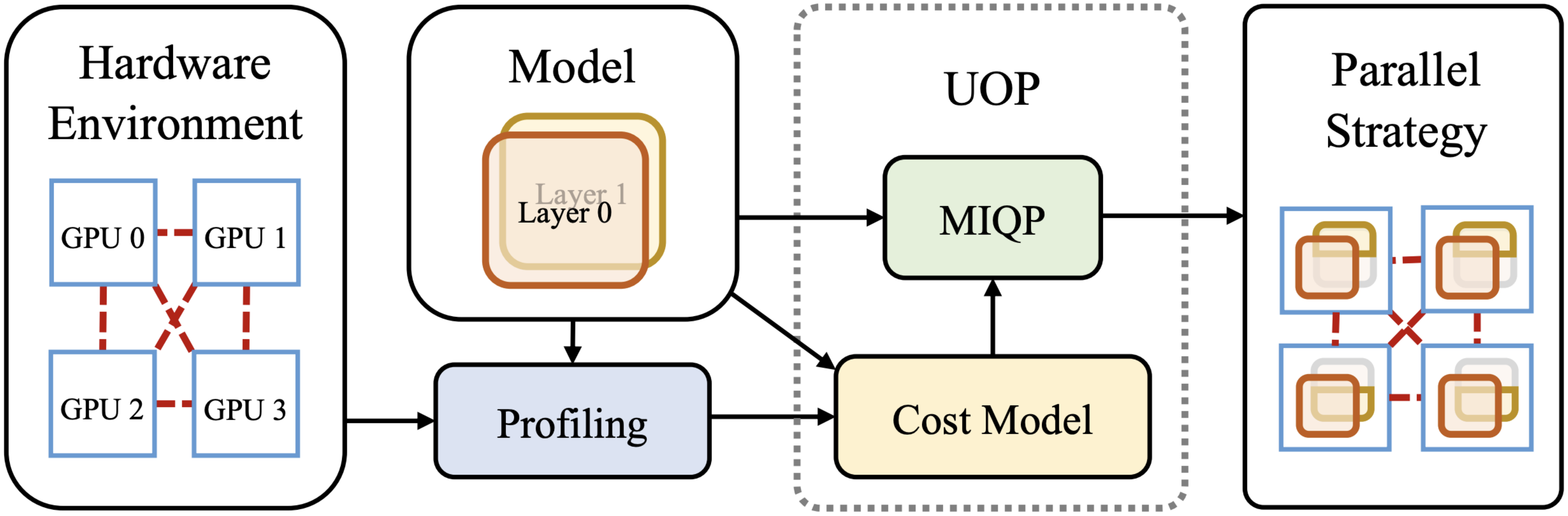
Optimize separately

Optimize JOINTLY 👍

Different AP methods

Contents

2.1 Overview

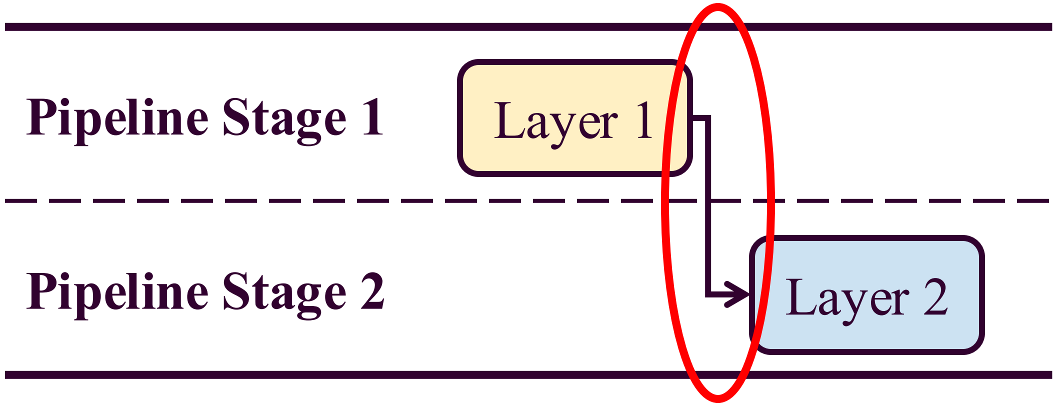


Flowchart of UniAP

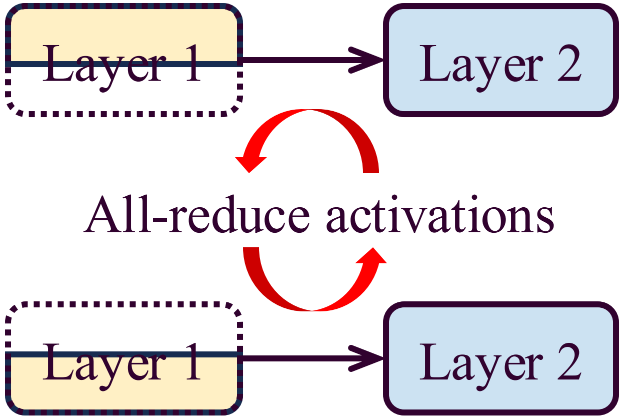
2.2 Profiling

- Environment performance information

1. P2P efficiency

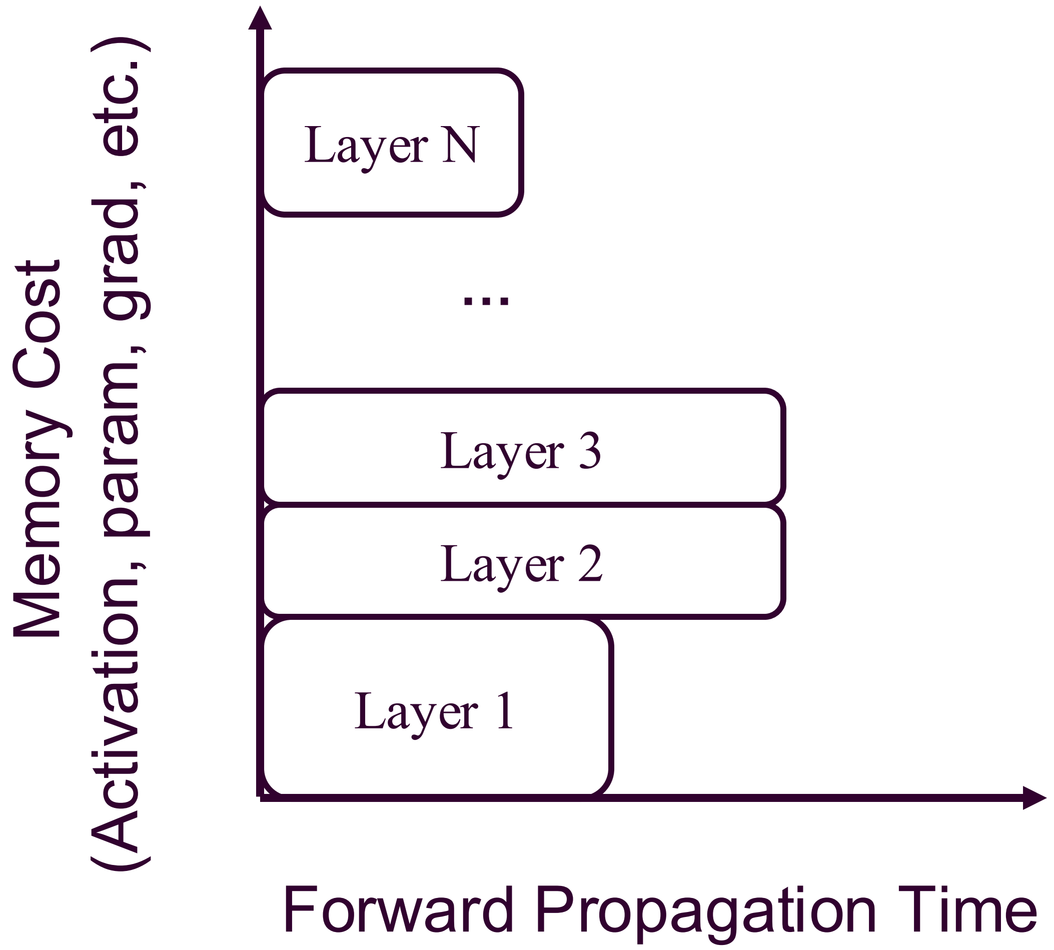


2. All-reduce efficiency



2.2 Profiling

- Model performance information



2.3 Cost Model

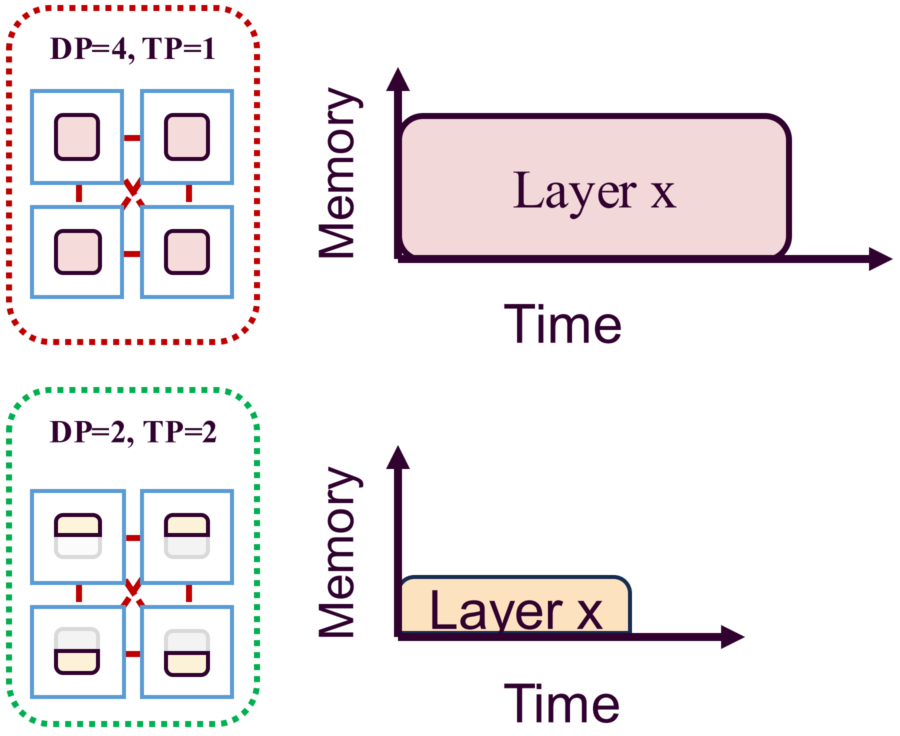
Efficiency coefficient

Forward Propagation Time

Memory Cost

Time Model

Memory Model



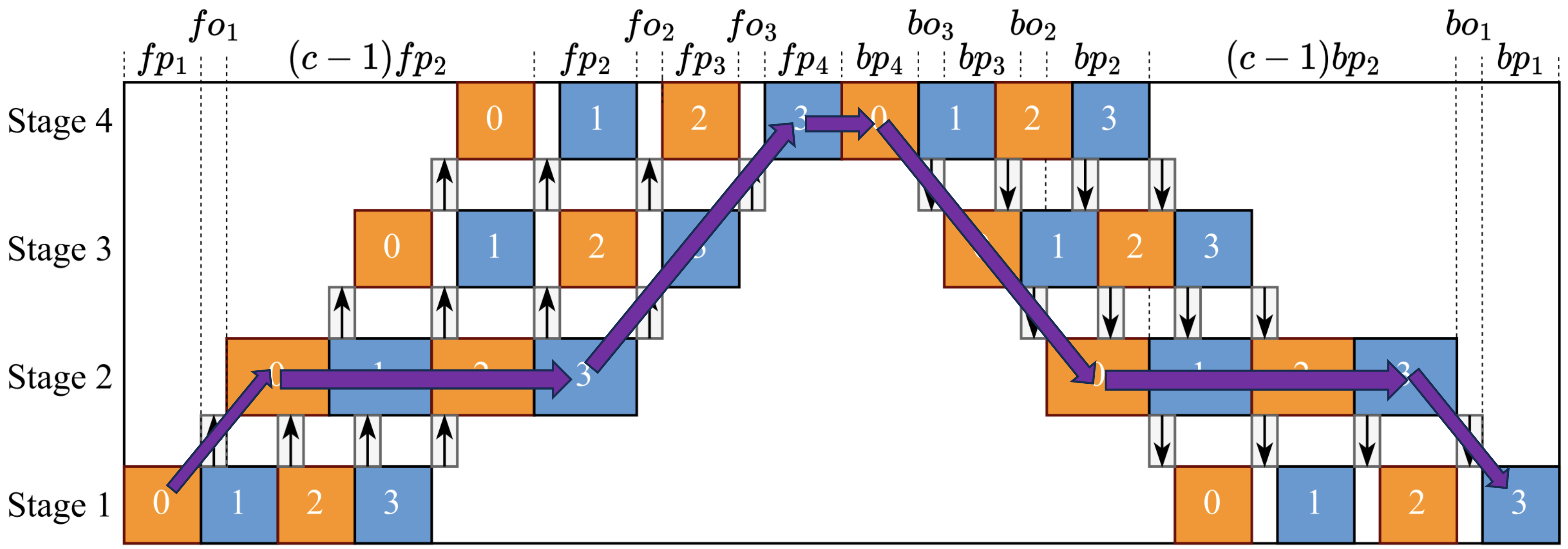
Profiling results

Cost modeling

Time and memory for each layer and each parallel strategy

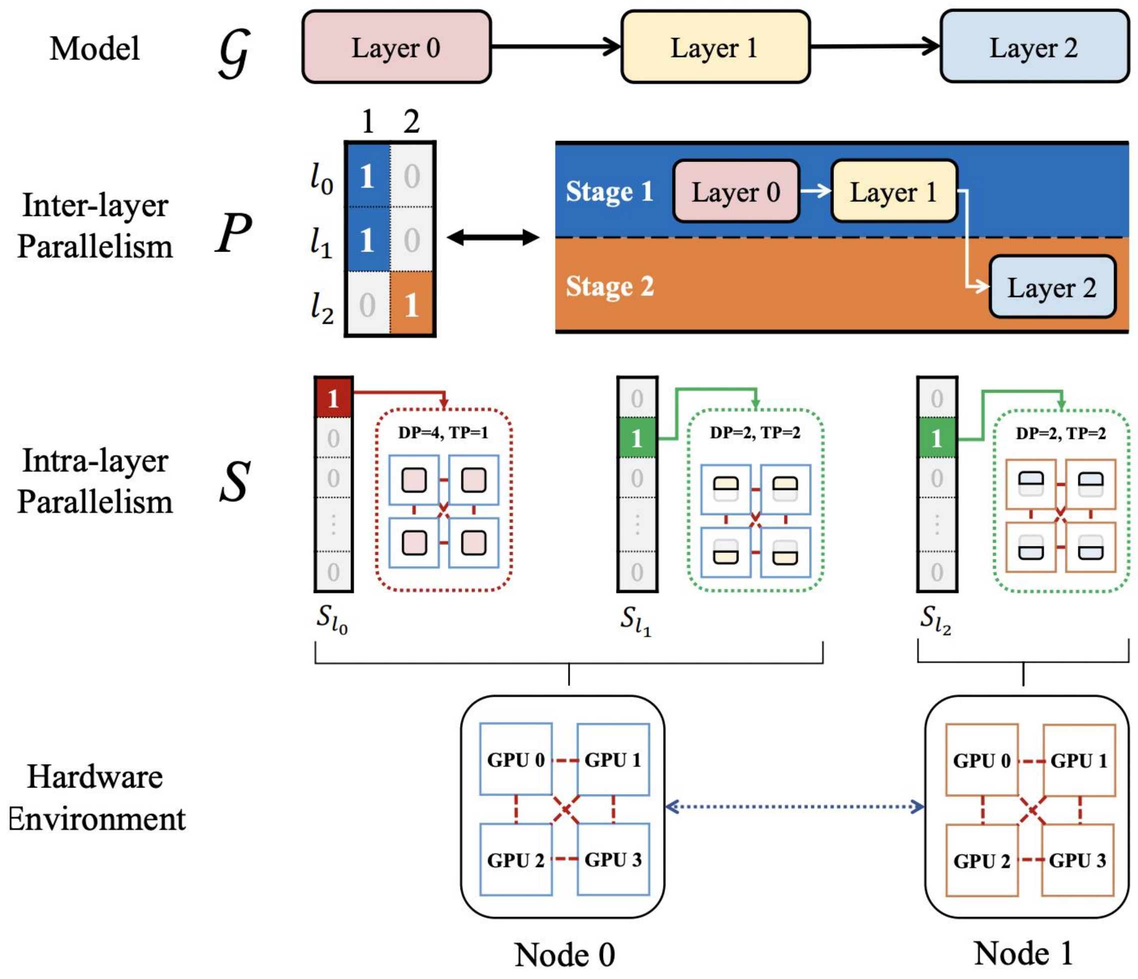
NANJING UNIVERSITY

- Objective function: to minimize the *Time-Per-Iteration* (TPI)



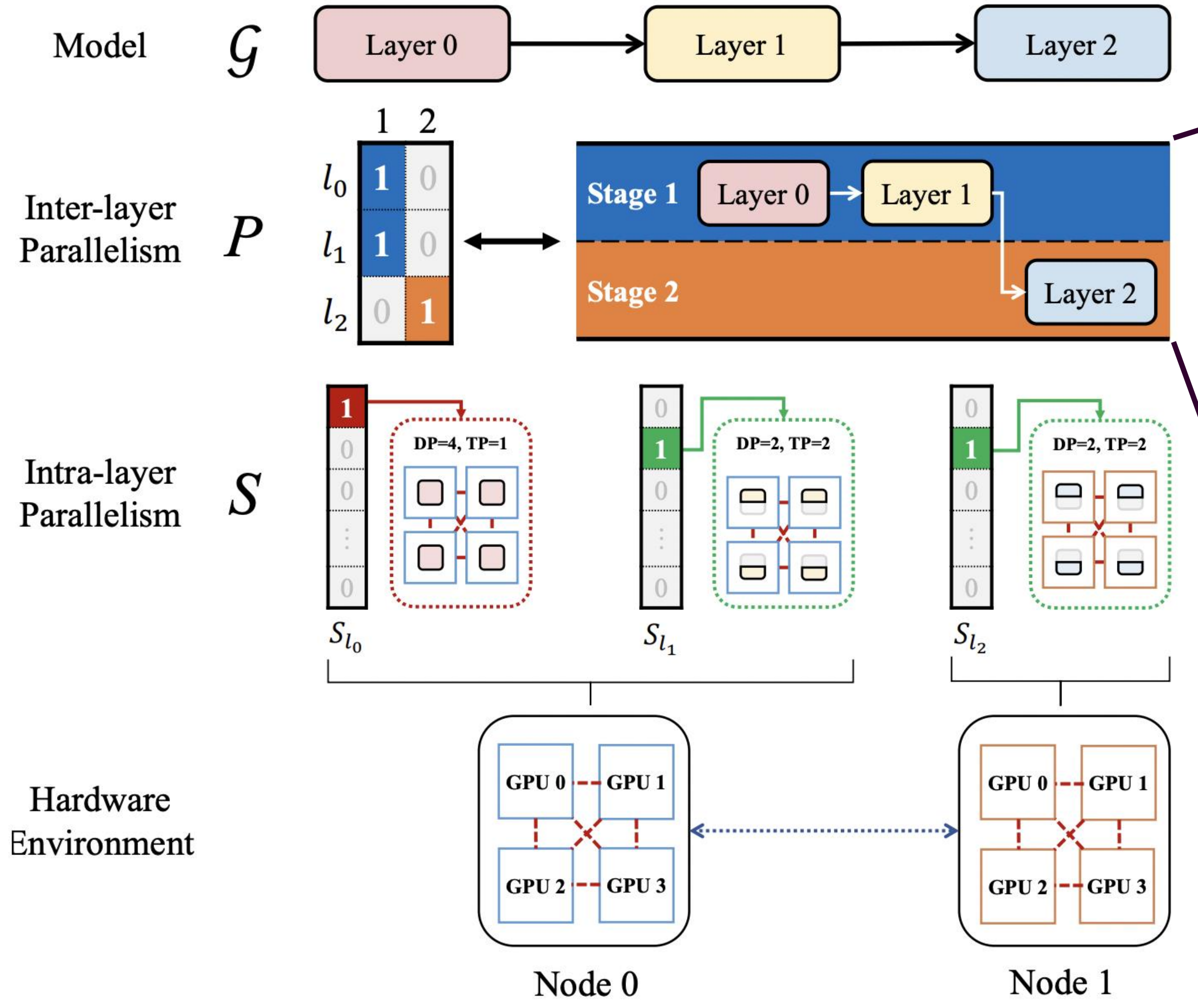
A GPipe-style pipeline

2.4 Mixed Integer Quadratic Programming



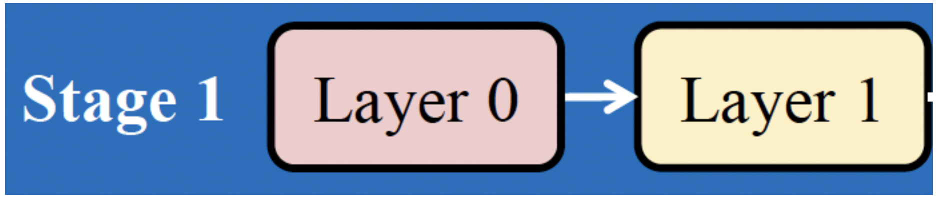
A candidate solution

2.4 Mixed Integer Quadratic Programming

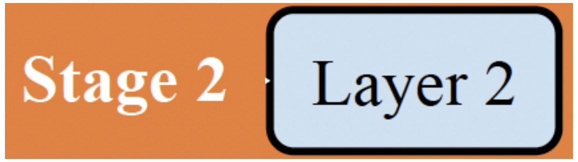


A candidate solution

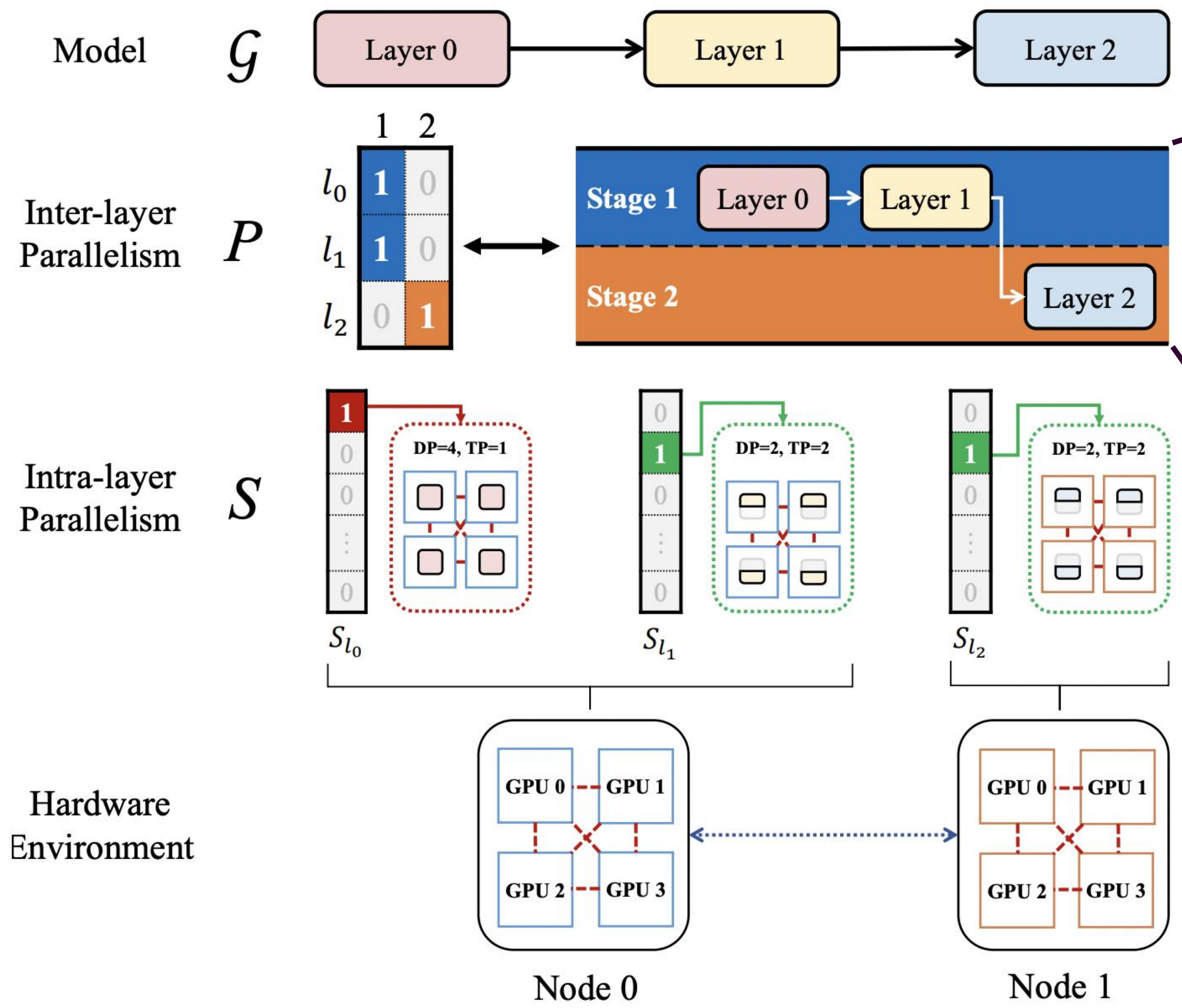
- Computation-stage constraint
  - Time cost of Stage 1 equals to:
    - Latency of Layer 0; plus
    - Latency of Layer 1; plus
    - Latency between Layer 0 and 1.



  - Time cost of Stage 2 equals to:
    - Latency of Layer 2.

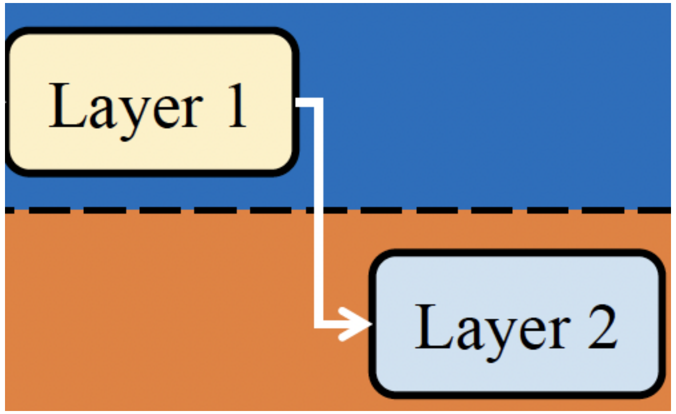


2.4 Mixed Integer Quadratic Programming

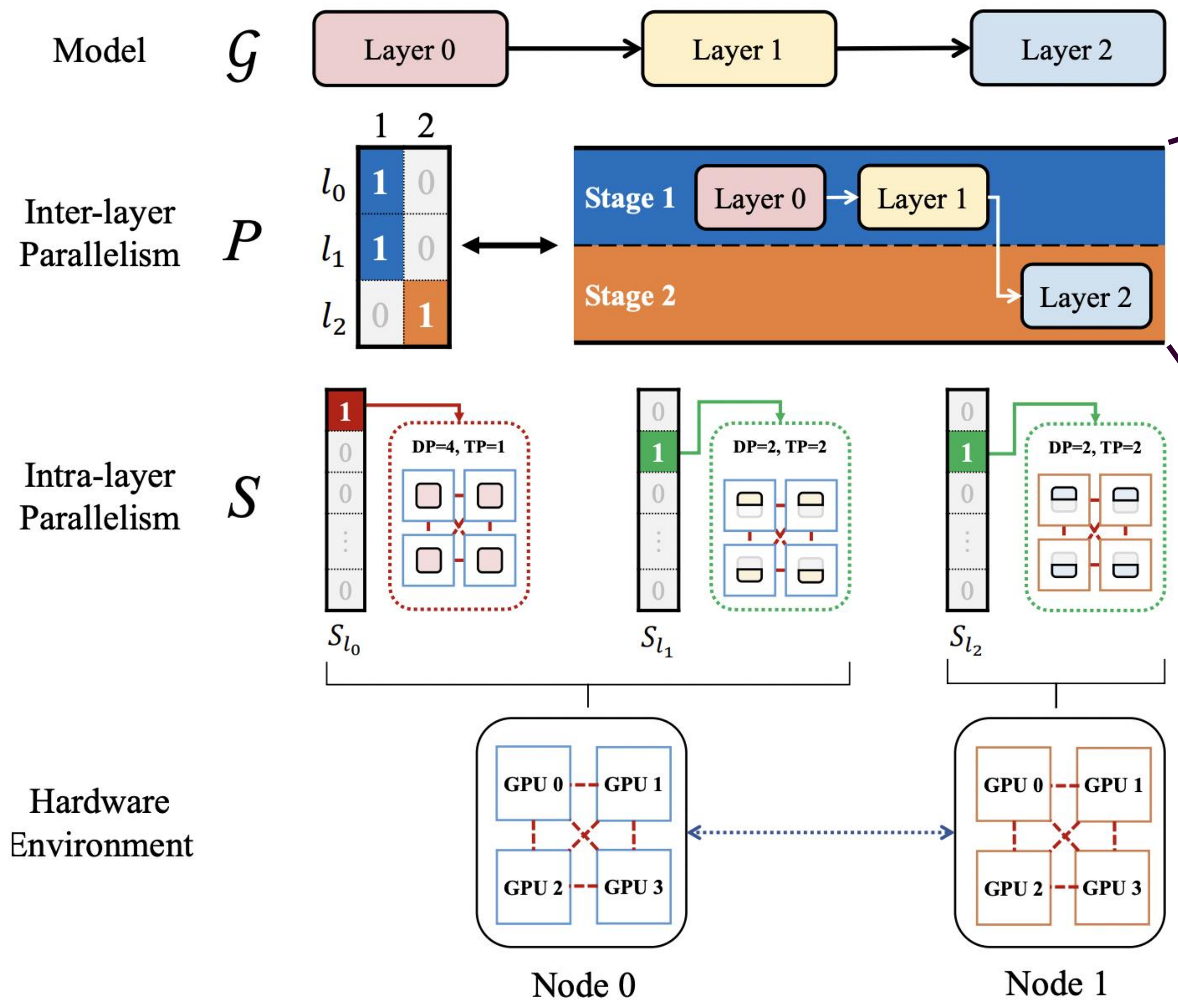


A candidate solution

- Communication-stage constraint
  - Time cost between Stage 1 and 2 equals to:
    - Latency between Layer 1 and 2.



2.4 Mixed Integer Quadratic Programming



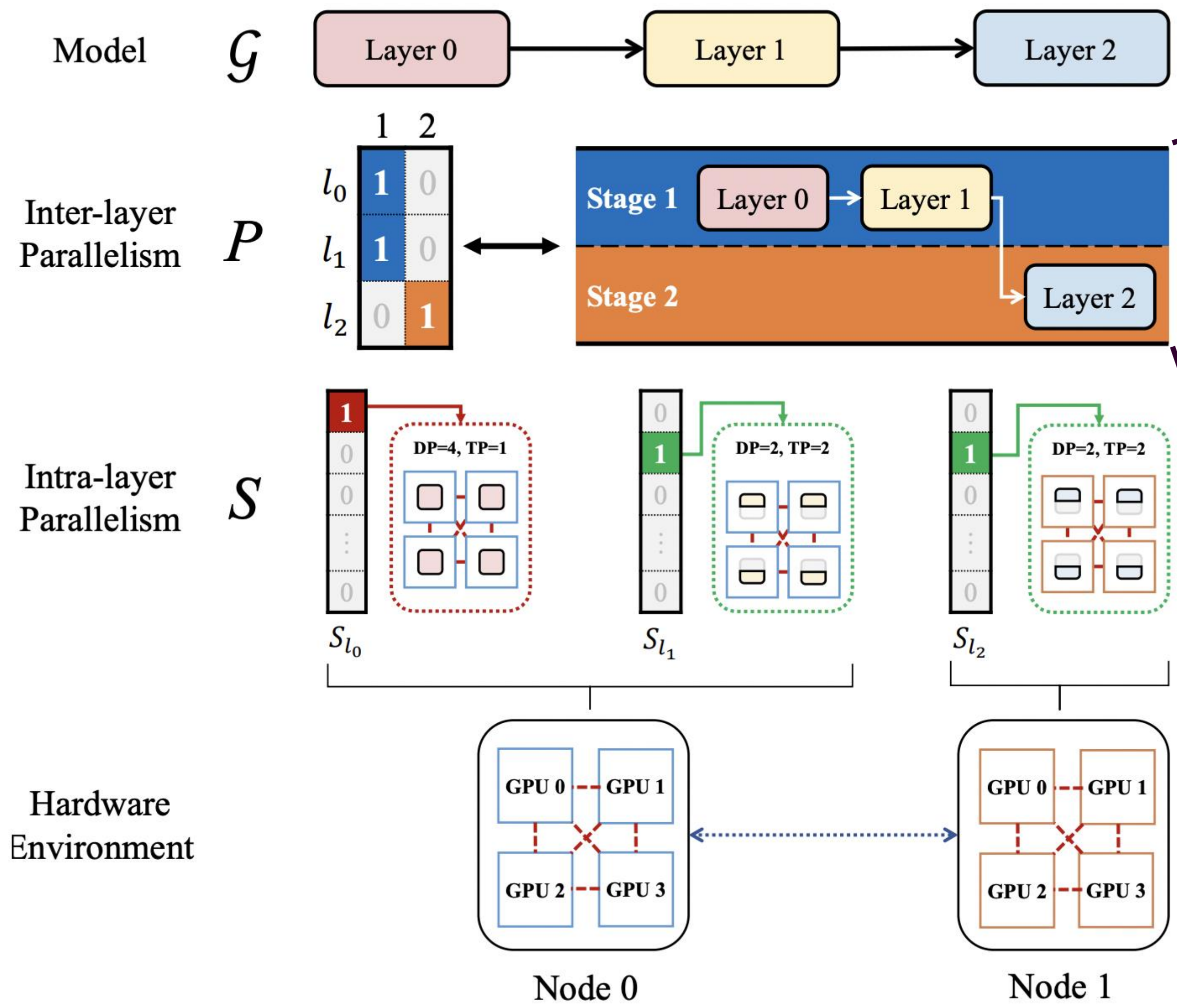
A candidate solution

- Layer-placement constraint
  - All layers should be placed on exactly one pipeline stage.
  - At least one layer should be placed on each pipeline stage.

$$P \quad \begin{array}{c|cc} & 1 & 2 \\ \hline l_0 & 1 & 0 \\ l_1 & 1 & 0 \\ l_2 & 0 & 1 \end{array}$$

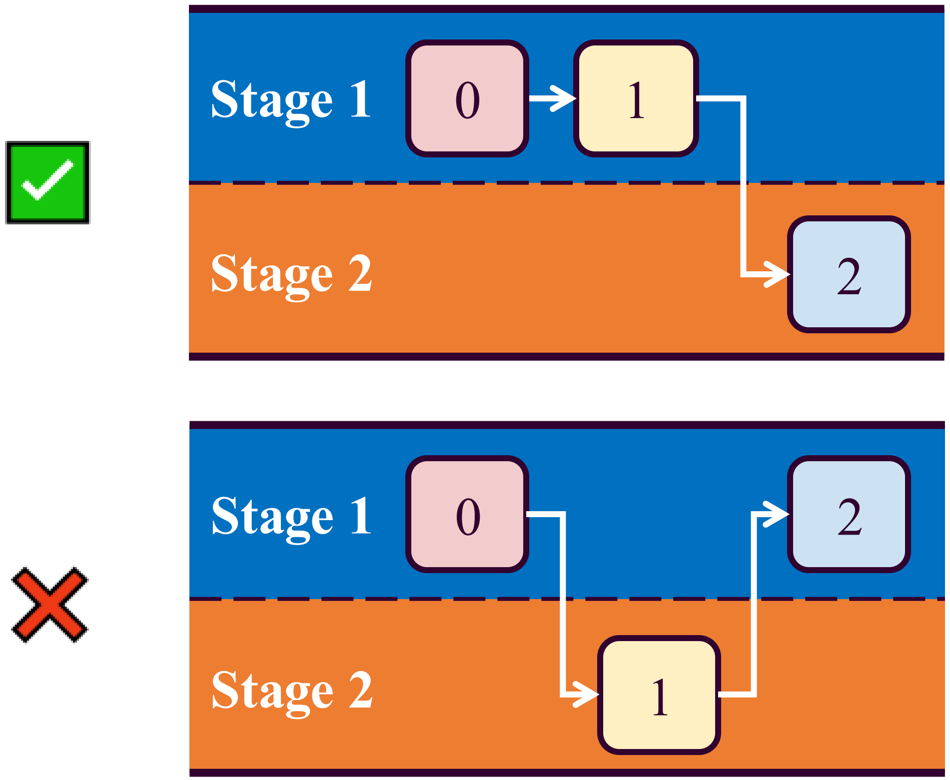
2.4 Mixed Integer Quadratic Programming

南京大學
NANJING UNIVERSITY

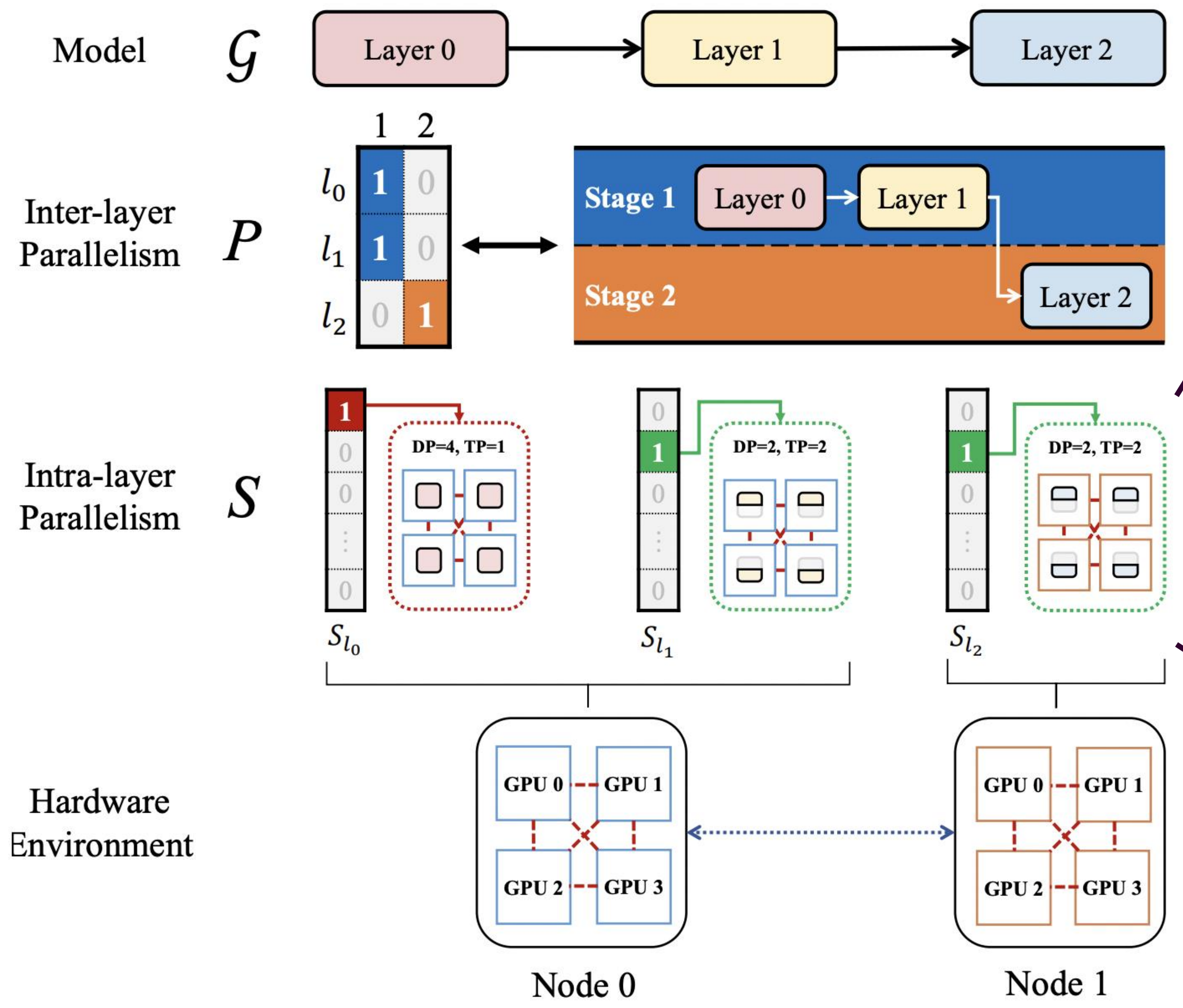


A candidate solution

- Order-preserving constraint
  - Layer 0, 1, and 2 are placed on Stage 1 and 2 *consecutively*.

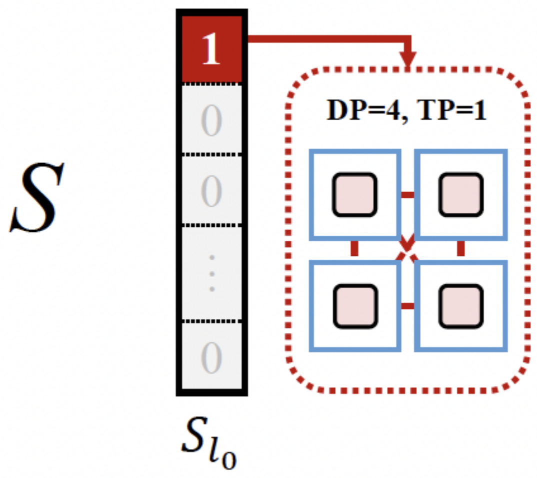


2.4 Mixed Integer Quadratic Programming

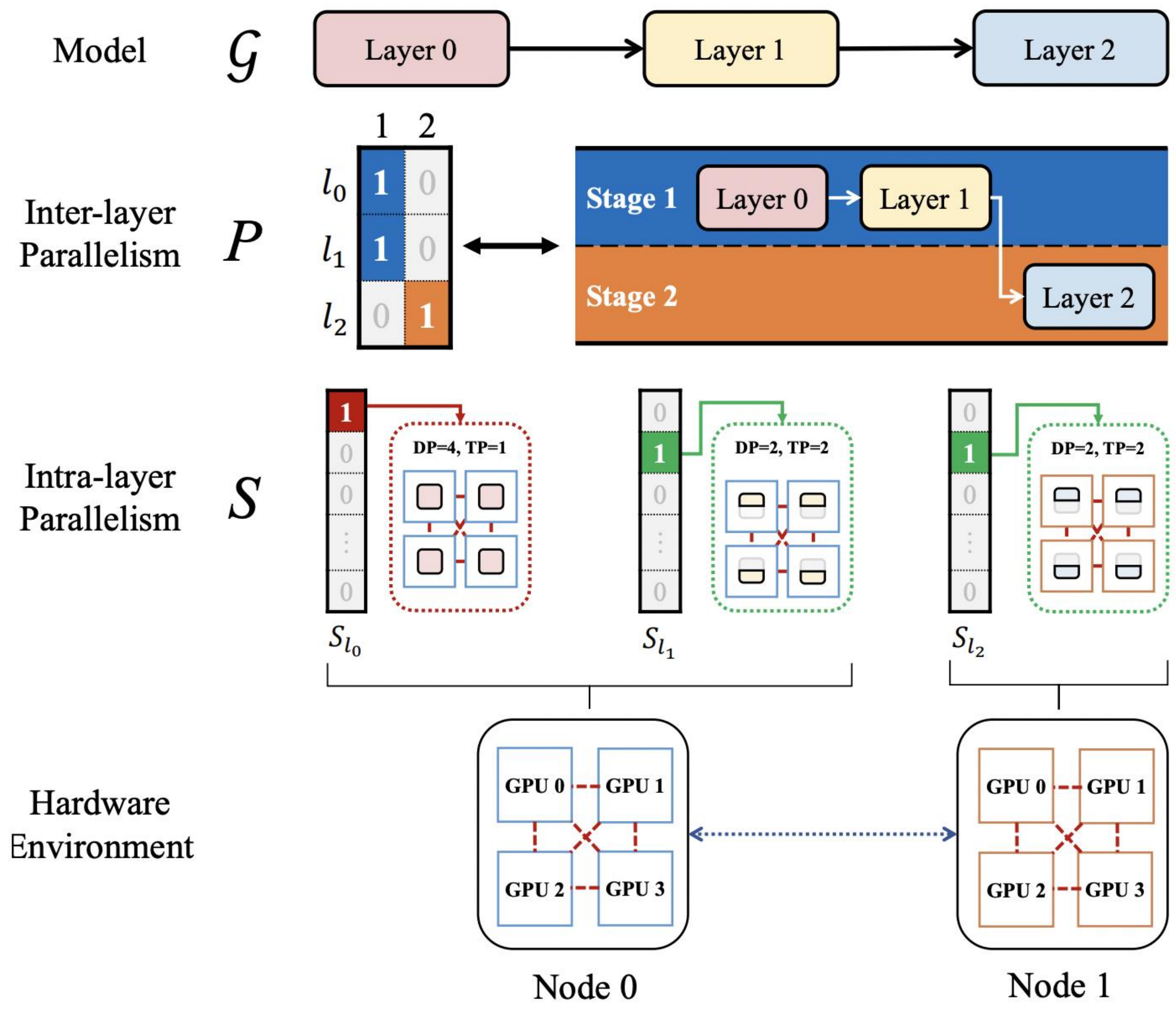


A candidate solution

- Strategy-selection constraint
  - Each layer must choose a strategy.
  - Each layer can only choose one strategy.

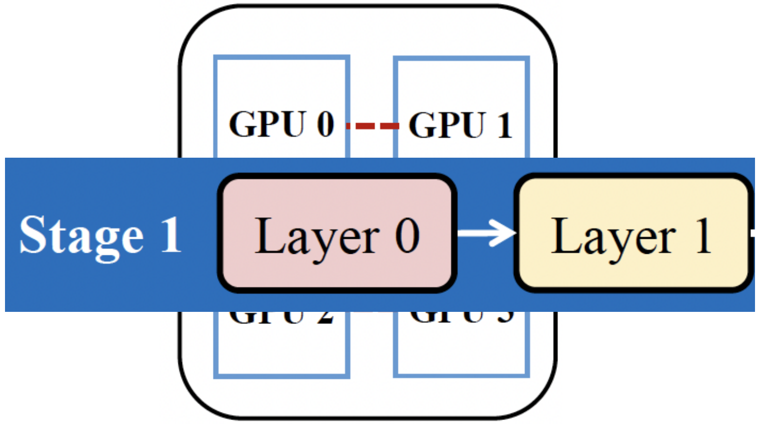


2.4 Mixed Integer Quadratic Programming



A candidate solution

- Memory constraint
  - Stage 1 with Layer 0 and Layer 1 can be placed on Node 0.

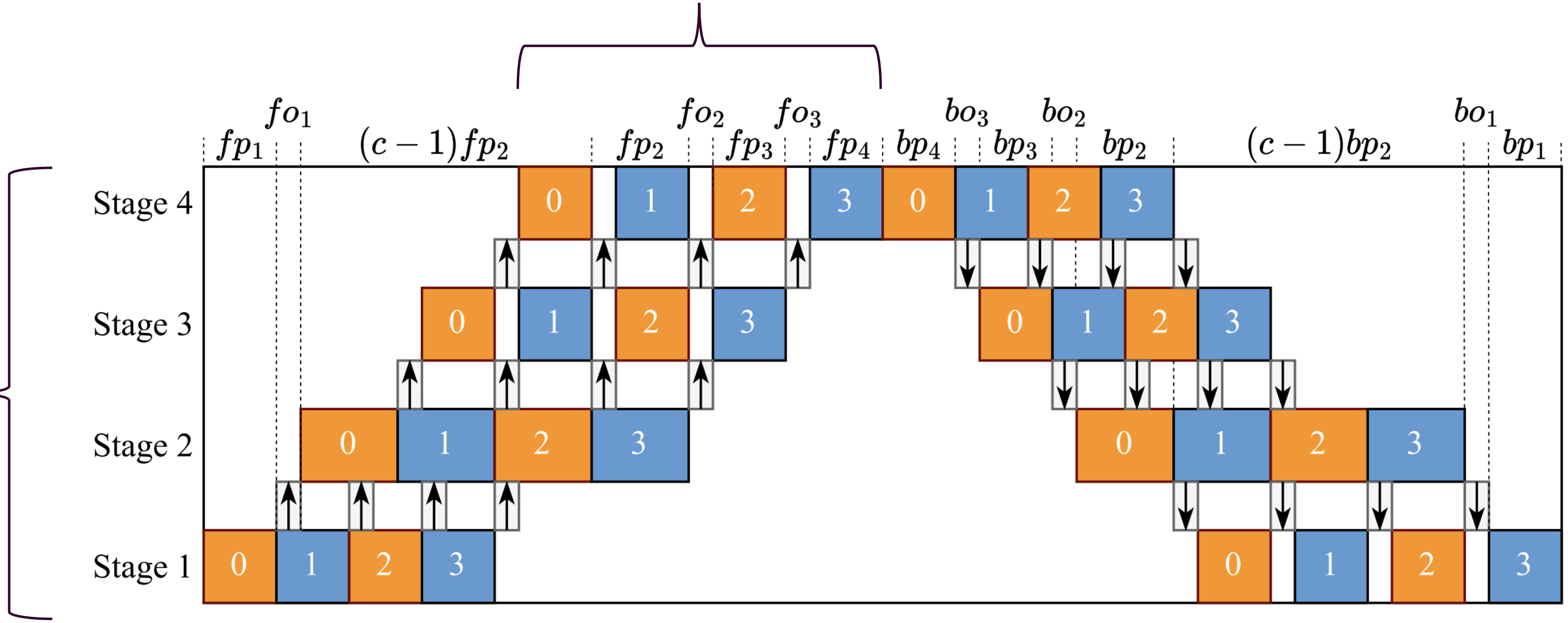


  - Layers on Stage 2 can be placed on Node 1.

2.5 Unified Optimization Process

micro-batches

Pipeline Degree



Two hyper-parameters to enumerate

Contents

3.1 Results on NVIDIA GPU

南京大學 NANJING UNIVERSITY

Env.	Model	Training throughput (samples/s)			Minimum speedup	Maximum speedup
		Galvatron	Alpa	UniAP		
EnvA	BERT-Huge	**33.46 ± 0.28**	31.56 ± 0.04	**33.46 ± 0.28**	1.00	1.06
	T5-Large	**23.29 ± 0.04**	MEM× [2)]	**23.29 ± 0.04**	1.00	1.00
	ViT-Huge	**109.51 ± 0.07**	97.66 ± 1.42	**109.51 ± 0.07**	1.00	1.12
	Swin-Huge	CUDA× [3)]	N/A [4)]	**67.96 ± 0.12**	N/A [4)]	N/A [4)]
EnvB-8	BERT-Huge	6.27 ± 0.17	8.95 ± 0.06	**10.77 ± 0.13**	1.20	1.71
	T5-Large [1)]	**8.06 ± 0.06**	MEM× [2)]	7.98 ± 0.05	0.99	0.99
	ViT-Huge	32.20 ± 0.17	38.74 ± 0.20	**45.58 ± 0.54**	1.18	1.41
	Swin-Huge	13.90 ± 0.17	N/A [4)]	**19.08 ± 0.10**	1.37	1.37
EnvC	Llama-7B	1.22 ± 0.01	N/A [4)]	**4.63 ± 0.007**	3.80	3.80

Throughput
max 3.8x ⬆

Env.	Model	Strategy optimization time (min.)			Minimum speedup	Maximum speedup
		Galvatron	Alpa	UniAP		
EnvA	BERT-Huge	6.44 ± 0.588	> 40	**0.37 ± 0.002**	17.29	> 107.41
	T5-Large	12.41 ± 0.122	MEM× [2)]	**0.89 ± 0.007**	13.98	13.98
	ViT-Huge	6.29 ± 0.464	> 40	**0.57 ± 0.009**	10.95	> 69.60
	Swin-Huge	11.88 ± 0.666	N/A [4)]	**2.16 ± 0.004**	5.49	5.49
EnvB-8	BERT-Huge	2.04 ± 0.010	> 40	**1.51 ± 0.005**	1.34	> 26.32
	T5-Large [1)]	2.64 ± 0.110	MEM× [2)]	**0.91 ± 0.005**	2.90	2.90
	ViT-Huge	2.37 ± 0.180	> 40	**1.11 ± 0.011**	2.14	> 36.01
	Swin-Huge	4.29 ± 0.320	N/A [4)]	**2.29 ± 0.010**	1.87	1.87
EnvC	Llama-7B	6.84 ± 0.055	N/A [4)]	**0.58 ± 0.006**	11.83	11.83

Strategy opt. time
max 107x ⬇

Training throughput and strategy optimization time on NVIDIA GPUs.

3.2 Results on Hygon DCU

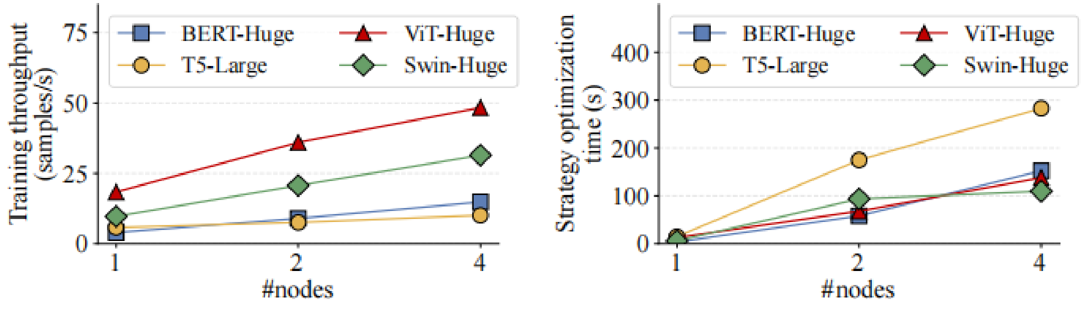
南京大學 NANJING UNIVERSITY

UniAP can find fastest parallel strategy while spending less time

Model	Training throughput (samples/s)			Strategy optimization time (min.)		
	Megatron	DeepSpeed	UniAP	Megatron	DeepSpeed	UniAP
Llama-7B	**2.01 ± 0.005**	SOL× [1)]	**2.01 ± 0.005**	> 8.0 hours	SOL× [1)]	**3.07 ± 0.121**
Llama-13B	**0.82 ± 0.001**	SOL× [1)]	**0.82 ± 0.001**	> 2.5 hours	SOL× [1)]	**1.95 ± 0.076**

Training throughput and strategy optimization time on 32 DCUs

3.3 Other experiments

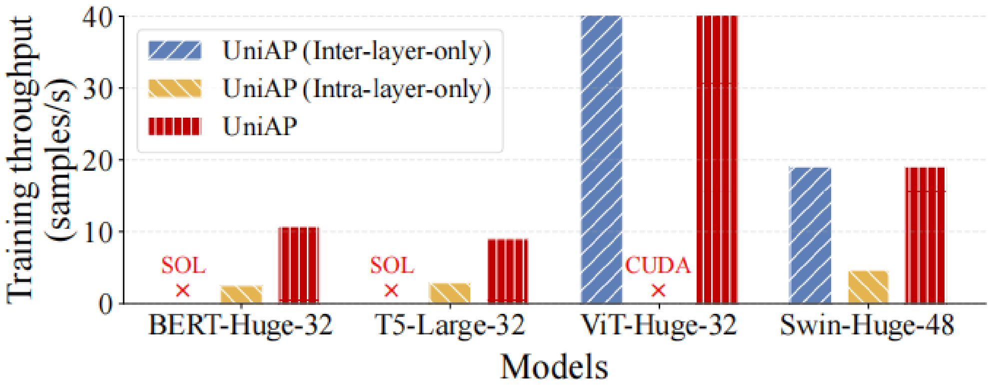


(a) Training throughput. (b) Strategy optimization time.

Scalability on 16 NVIDIA GPUs

Number of DCUs	Training throughput (samples/s)	Strategy optimization time(min.)
16	0.1726 ± 0.0005	0.75 ± 0.0100
32	0.3393 ± 0.0004	1.64 ± 0.0321
64	0.6656 ± 0.0023	1.55 ± 0.0058

Scalability on 64 Hygon DCUs



Ablation study

Contents

4.1 Conclusion

- We propose a novel AP method called UniAP to unify inter- and intra-layer AP by MIQP.

- To the best of our knowledge, UniAP is the FIRST parallel method that can optimize the two categories of parallel strategies jointly rather than separately to find an optimal solution.

- UniAP outperforms state-of-the-art methods to achieve the best training throughput.

4.2 Advertisement

- Please find our personal website at
  - Hao Lin: https://haolin-nju.github.io/
  - Ke Wu: https://bigai-nju.github.io/members/Ke Wu
  - Jie Li: https://dibaotian-xing.github.io/lijie/

THANKS

南京大學
NANJING UNIVERSITY