The 29th Annual International Conference
On Mobile Computing And Networking

Efficient Federated Learning for Modern NLP

Dongqi Cai[1], Yaozong Wu[1], Shangguang Wang[1], Felix Xiaozhu Lin[2], Mengwei Xu[1]

1 Beiyou Shenzhen Institute
2 University of Virginia

How to understand the meaning of a word?

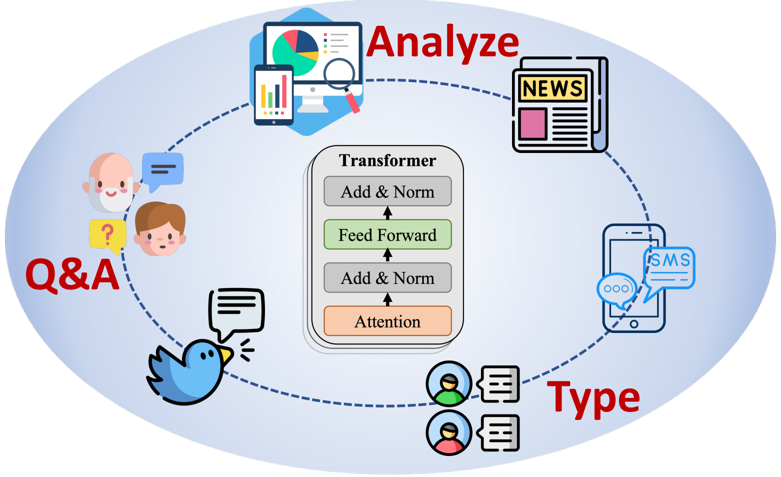
Natural Language Processing (NLP)

How to understand the meaning of a word?

Natural Language Processing (NLP)

What sparks modern NLP?

Attention-based Transformer

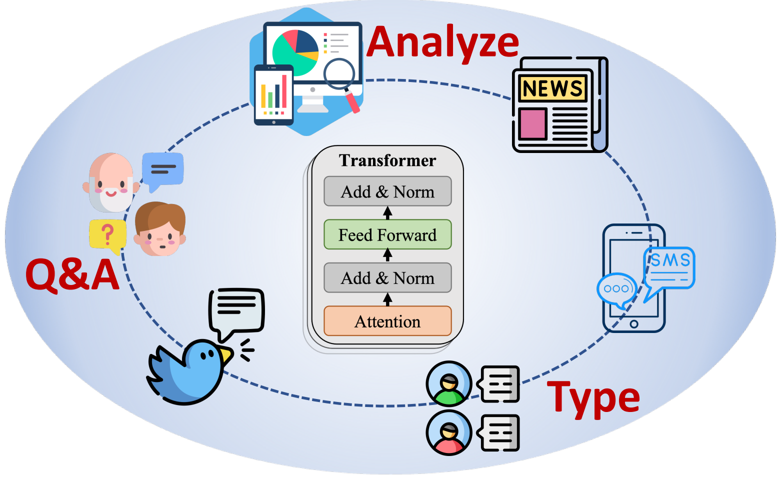


How to understand the meaning of a word?

Natural Language Processing (NLP)

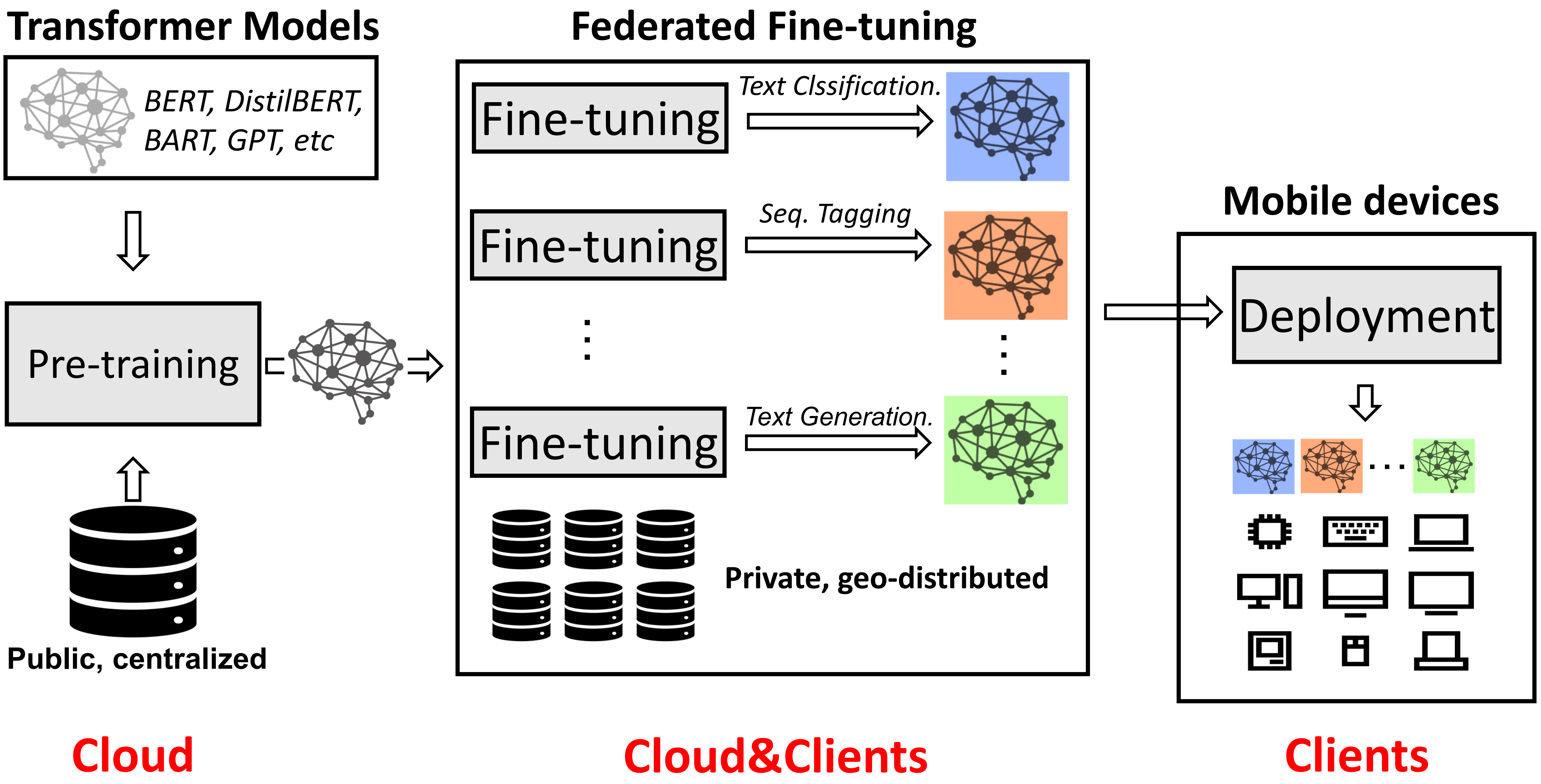
What sparks modern NLP?

Attention-based Transformer



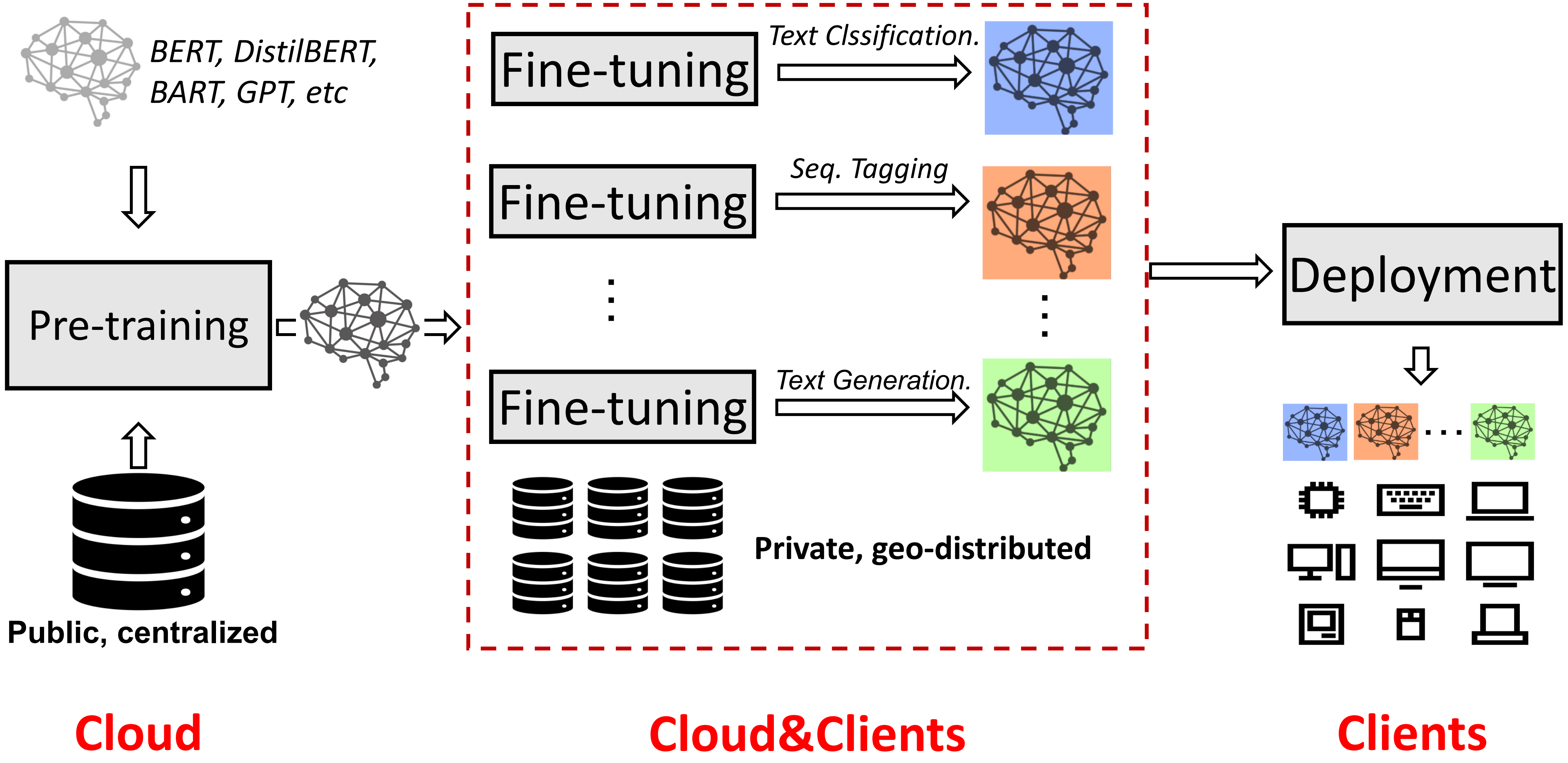
How to preserve the privacy of training data?

Federated Learning

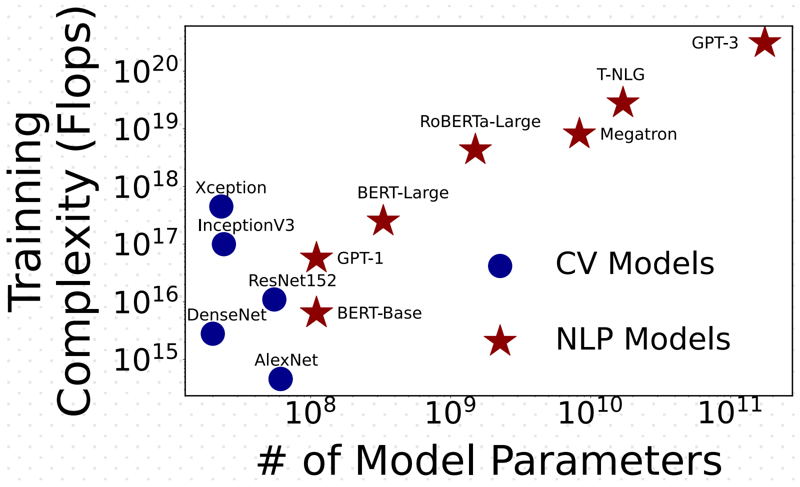


Transformer Models
BERT, DistilBERT, BART, GPT, etc
Pre-training
Public, centralized
Federated Fine-tuning
Fine-tuning
Text Clssification.
Fine-tuning
Seq. Tagging
Fine-tuning
Text Generation.
Private, geo-distributed
Mobile devices
Deployment
Cloud
Cloud&Clients
Clients

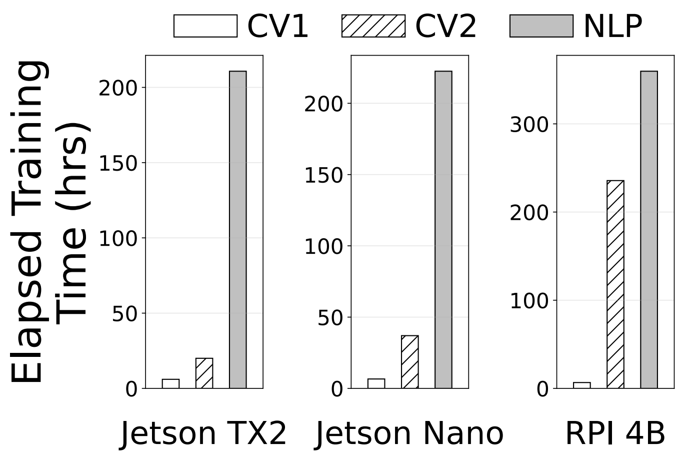
FedNLP: focus of this work



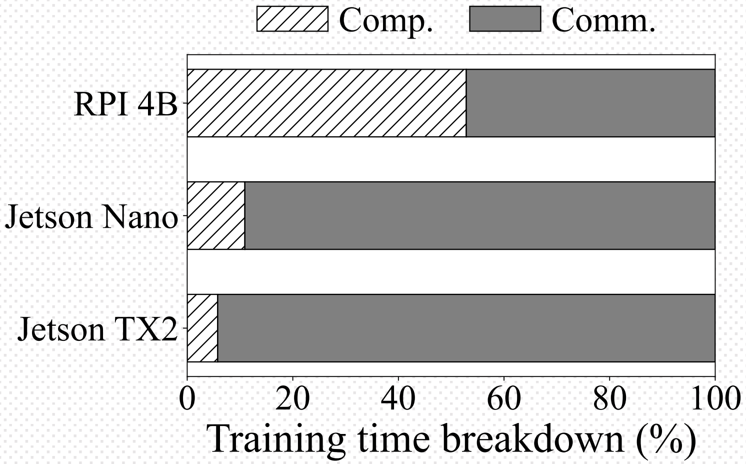
Is FedNLP practical on todays' mobile platforms?



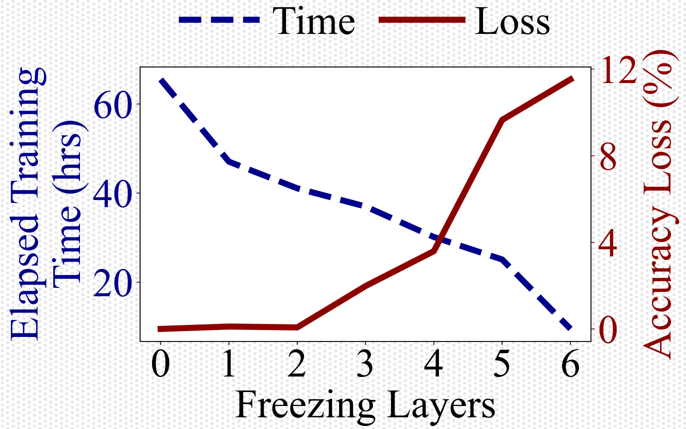
Observation 1: Transformer-based NLP models are highly costly.



Observation 2: FedNLP task is extremely slow.

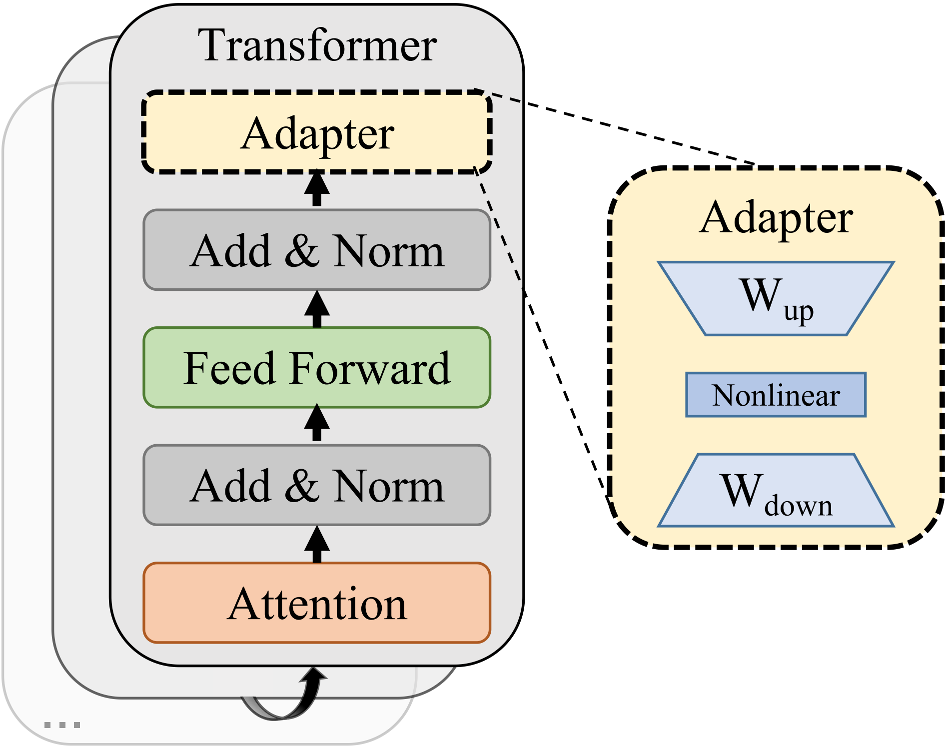


Observation 3: Network transmission dominates the training delay on high-end devices.



Observation 4: Existing techniques are inadequate for FedNLP.

Key Building Block: Pluggable Adapters

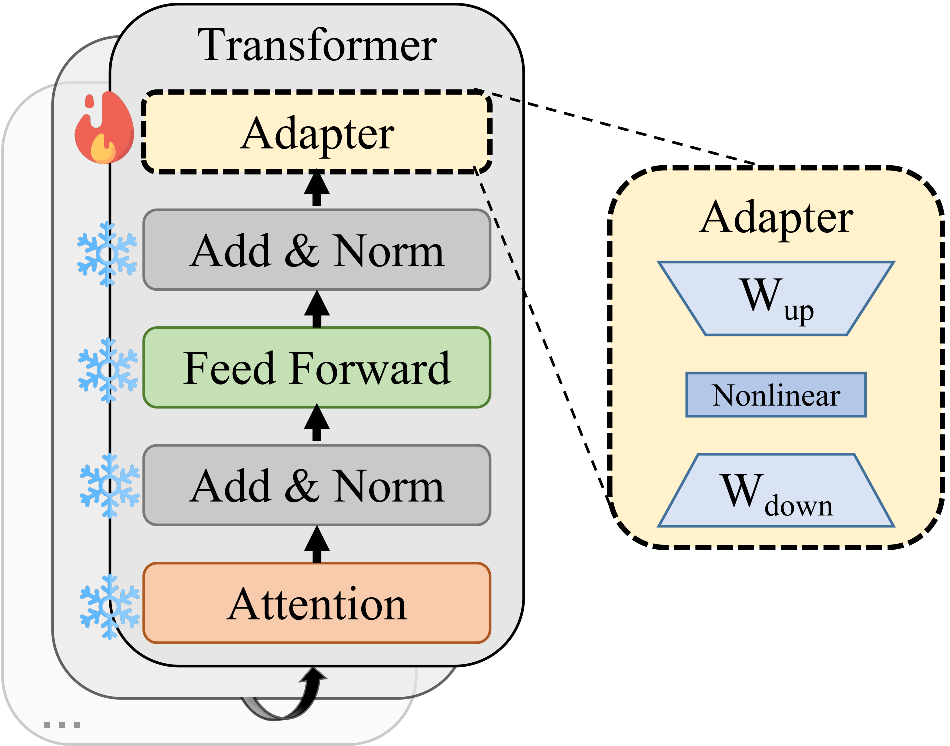


Model	Method	Training Time	Updated Paras.
BERT	Full Fine-tuning	1.86 sec	110.01×10^6
	Adapter	1.14 sec	0.61×10^6
DistilBERT	Full Fine-tuning	0.91 sec	67×10^6
	Adapter	0.56 sec	0.32×10^6

Table 1: **Computation** and **communication** cost of inserting adapters into each transformer block (width=32) and full model tuning. Batch size: 4. Device: Jetson TX2.

- Tiny adapters (**less than 1M** for each) are inserted to pre-trained Transformers.
- **Only adapters are updated** during training, most of Transformer parameters are freezing.

Key Building Block: Pluggable Adapters

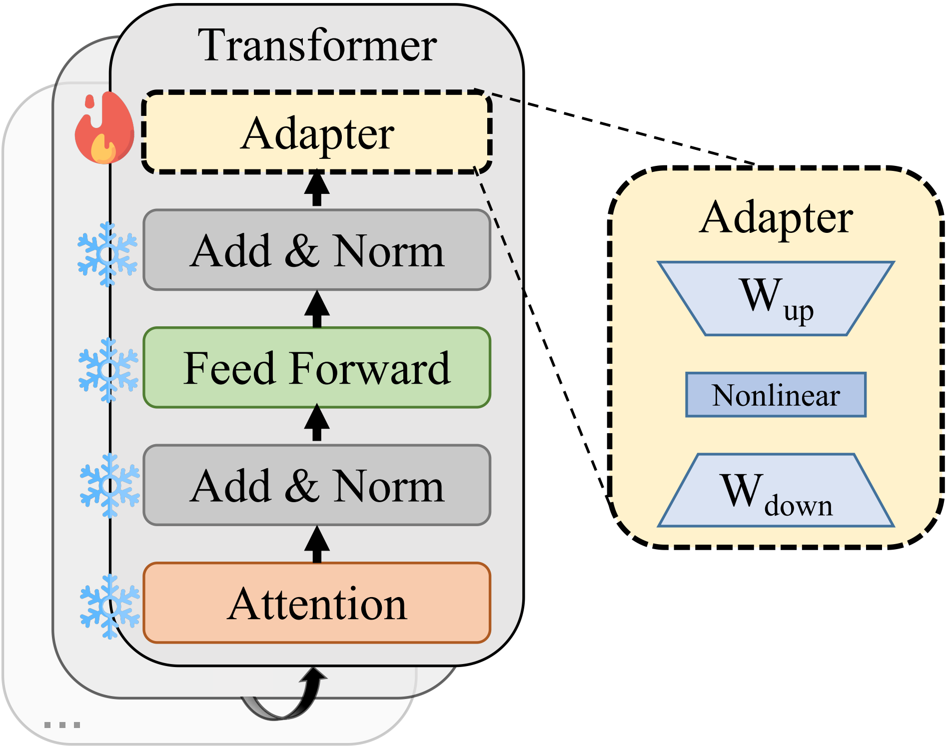


Model	Method	Training Time	Updated Paras.
BERT	Full Fine-tuning	1.86 sec	110.01 x 10^6
	Adapter	1.14 sec	0.61 x 10^6
DistilBERT	Full Fine-tuning	0.91 sec	67 x 10^6
	Adapter	0.56 sec	0.32 x 10^6

Table 1: **Computation** and **communication** cost of inserting adapters into each transformer block (width=32) and full model tuning. Batch size: 4. Device: Jetson TX2.

- Tiny adapters (**less than 1M** for each) are inserted to pre-trained Transformers.
- **Only adapters are updated** during training, most of Transformer parameters are freezing.

Key Building Block: Pluggable Adapters

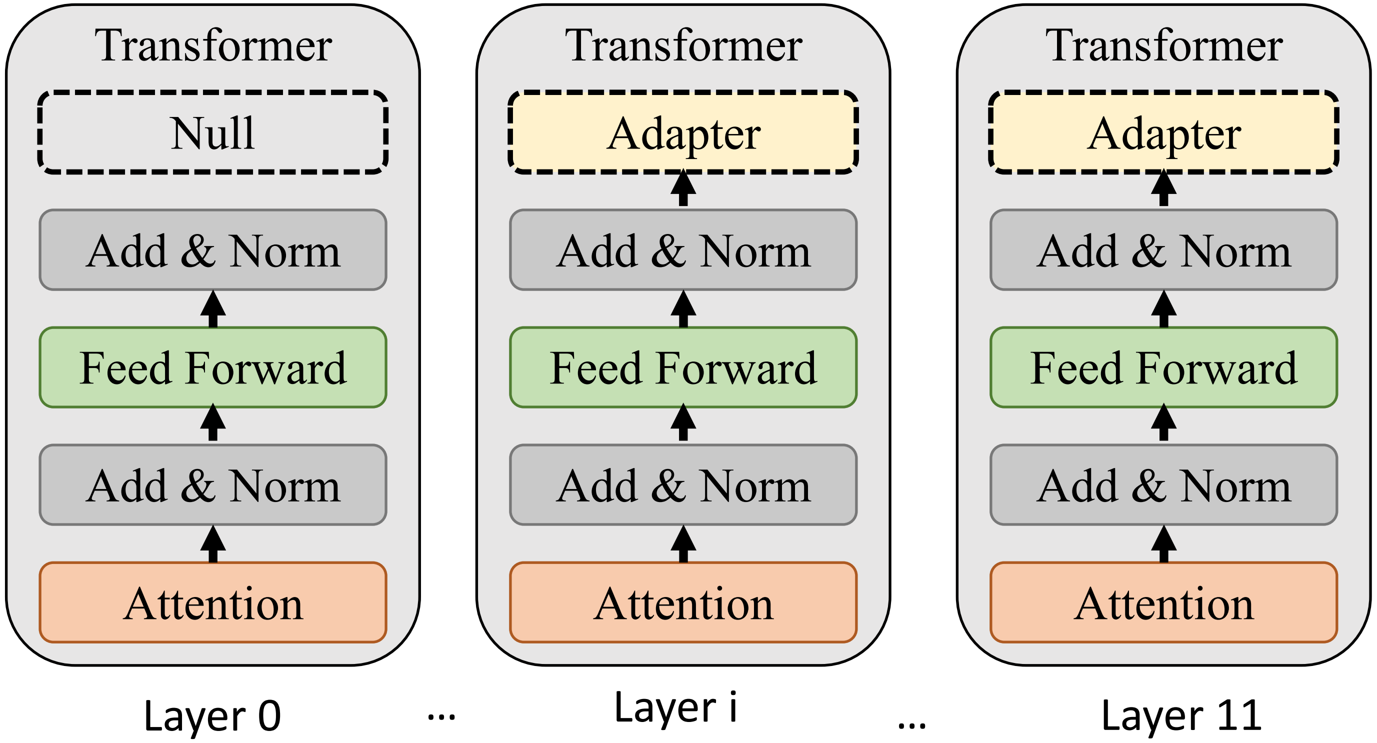


Model	Method	Training Time	Updated Paras.
BERT	Full Fine-tuning	1.86 sec	110.01×10^6
	Adapter	1.14 sec	0.61×10^6
DistilBERT	Full Fine-tuning	0.91 sec	67×10^6
	Adapter	0.56 sec	0.32×10^6

Table 1: **Computation** and **communication** cost of inserting adapters into each transformer block (width=32) and full model tuning. Batch size: 4. Device: Jetson TX2.

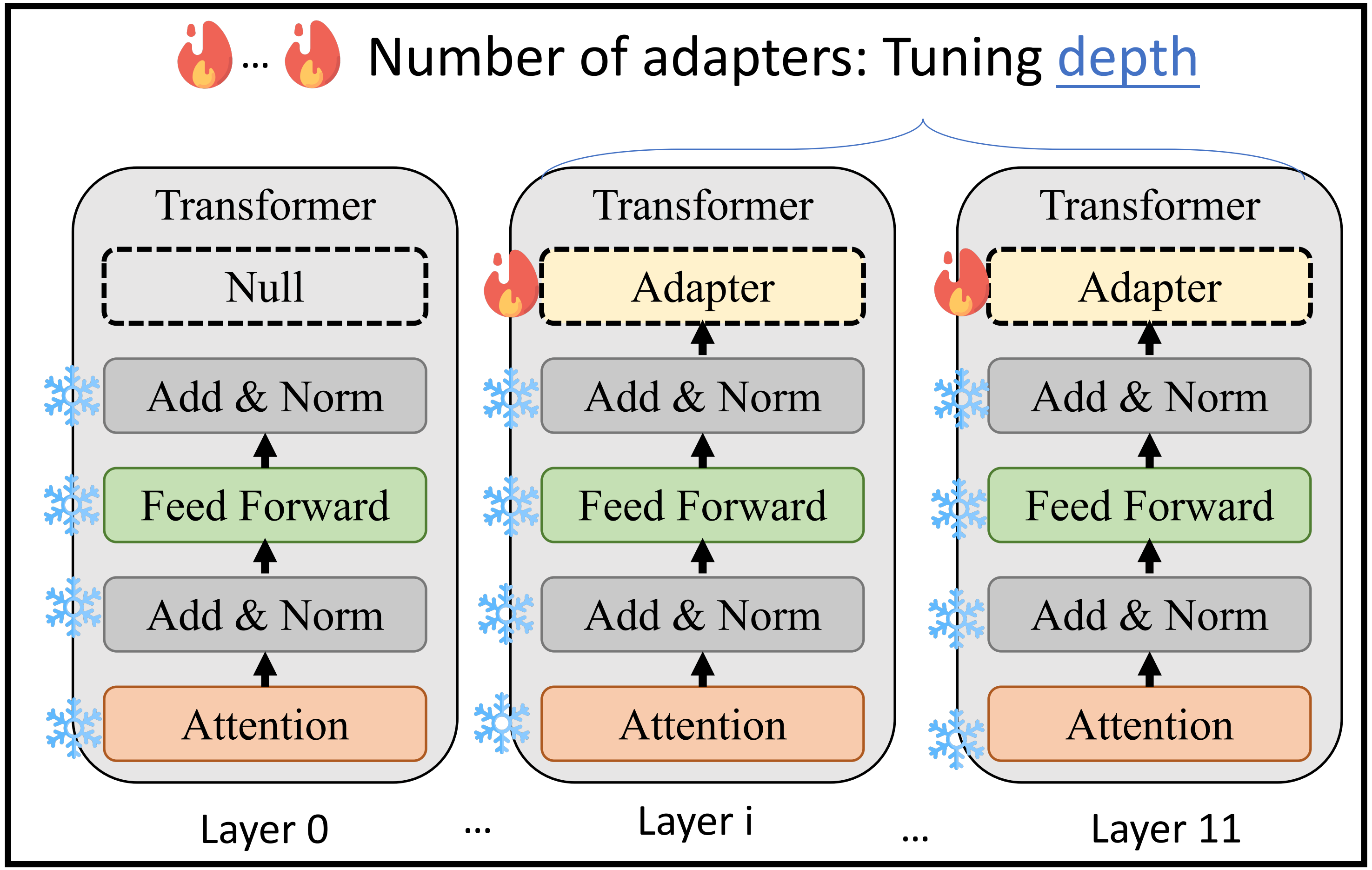
- Tiny adapters (**less than 1M** for each) are inserted to pre-trained Transformers.
- **Only adapters are updated** during training, most of Transformer parameters are freezing.

Challenge: Large Adapter Configuration Space



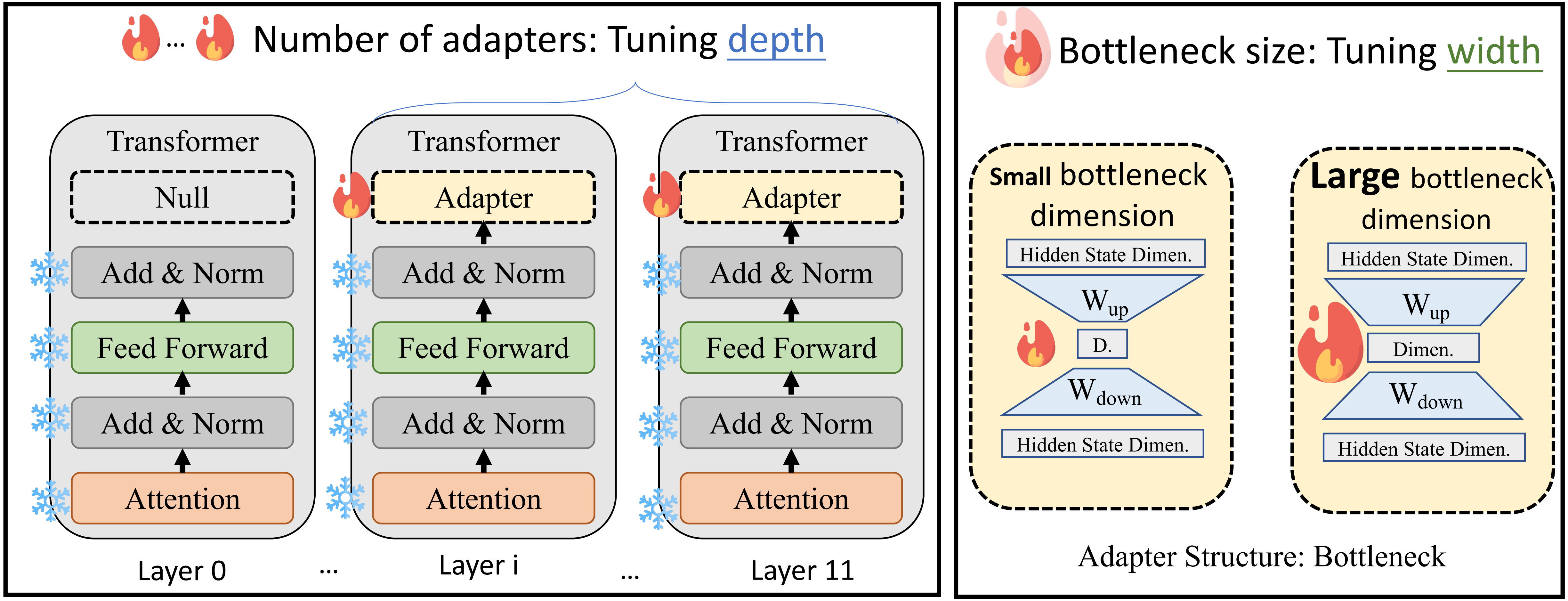
Different adapter configurations (depth, width) result in a variety of convergence delays, up to 4.7× gap.

Challenge: Large Adapter Configuration Space



Different adapter configurations (depth, width) result in a variety of convergence delays, up to 4.7× gap.

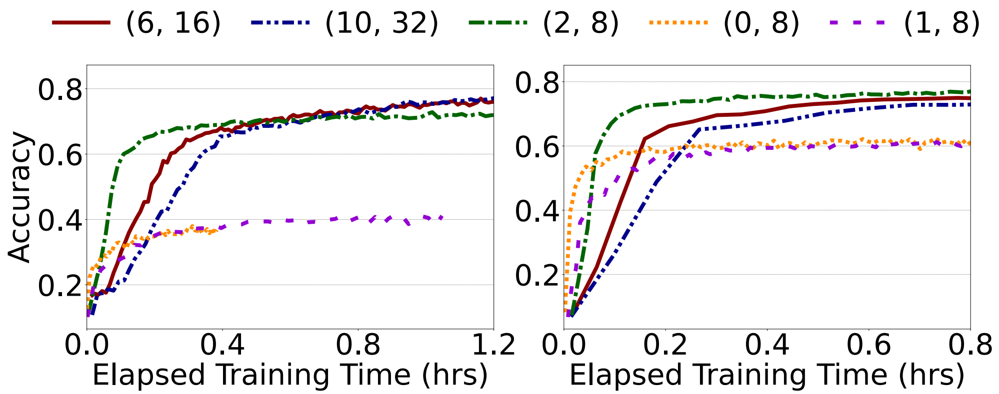
Challenge: Large Adapter Configuration Space



Different adapter configurations (depth, width) result in a variety of convergence delays, up to 4.7× gap.

Challenge: No Silver Bullet Configuration

- The optimal configuration can be **switched** across FL rounds.

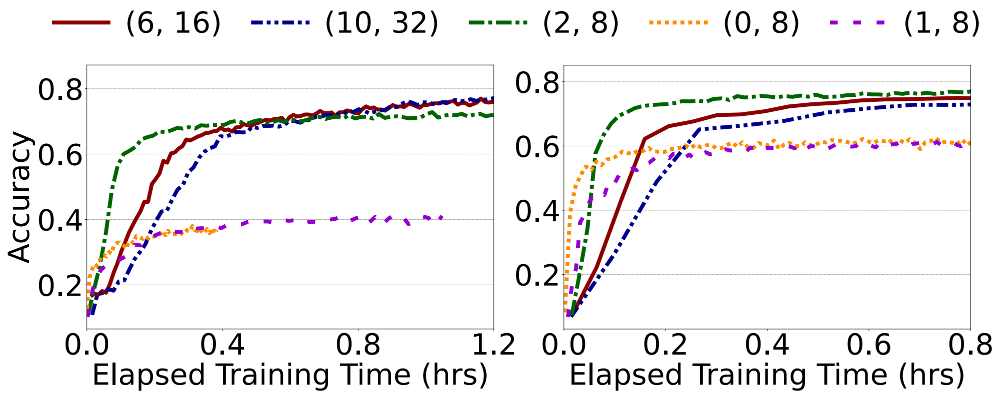


(a) SEMEVAL (b) 20NEWS

Across different target accuracy and target FedNLP tasks, the optimal adapter configuration (depth, width) varies. Model: BERT; device: Jetson TX2.

Challenge: No Silver Bullet Configuration

- The optimal configuration can be **switched** across FL rounds.
- Configuration **varies** across many factors:
  targeted accuracy, targeted NLP tasks and client resources.



(a) SEMEVAL (b) 20NEWS

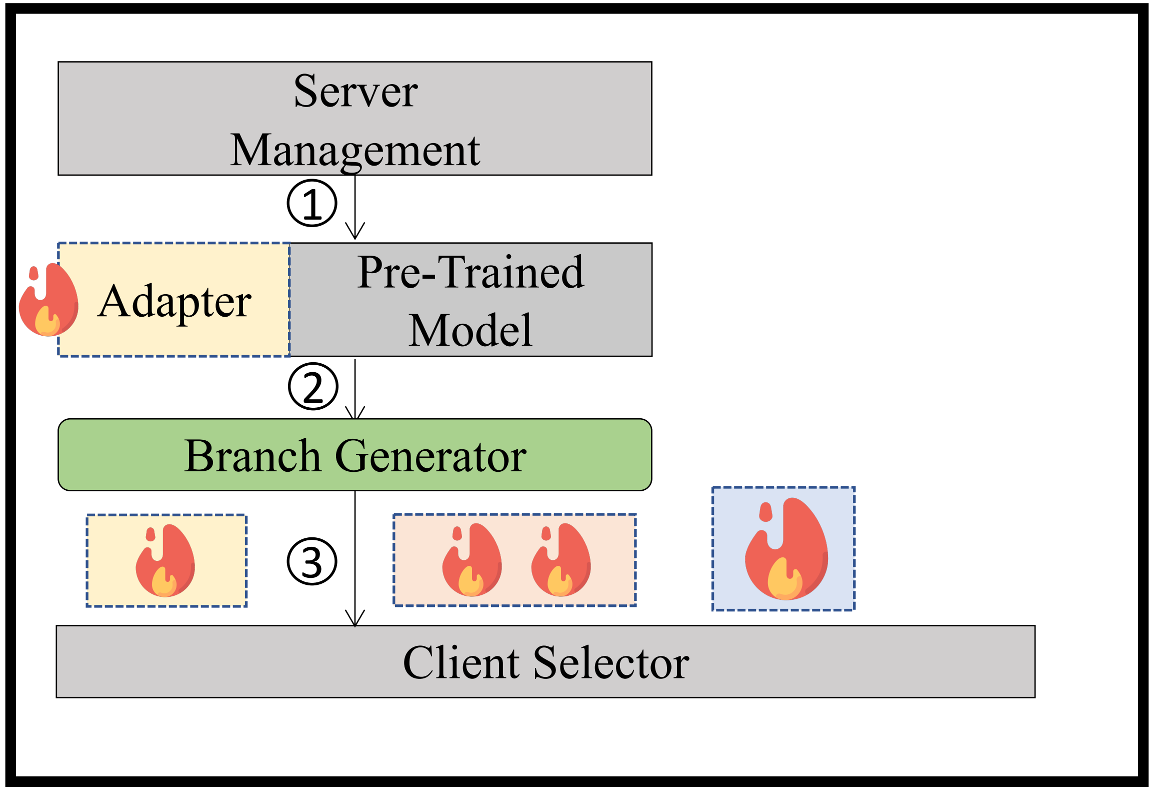
Across different target accuracy and target FedNLP tasks, the optimal adapter configuration (depth, width) varies. Model: BERT; device: Jetson TX2.

Model	Datasets	Optimal adapter configuration **(depth, width)** towards different target accuracy				
		99%	**95%**	**90%**	**80%**	**70%**
BERT	**20news**	(2,64)	(2,32)	(2,8)	(2,8)	(2,8)
	agnews	(3,16)	(2,16)	(2,8)	(0,8)	(0,8)
	semeval	(10,8)	(6,8)	(6,8)	(2,8)	(2,8)
	ontonotes	(12, 32)	(12, 32)	(10, 32)	(0, 16)	(0, 16)

The optimal adapter configuration (i.e., best time to-accuracy) for different target accuracy (ratio to the full convergence accuracy) and different datasets.

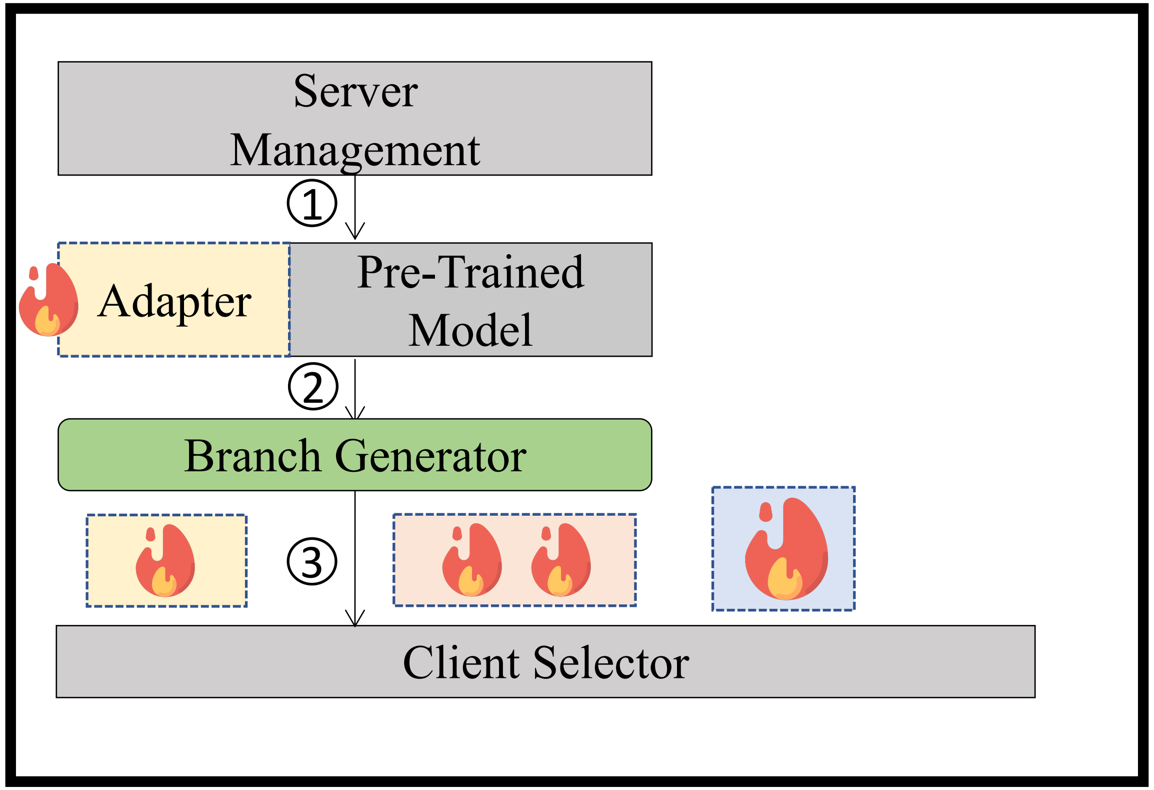
Design: Online Configurator

- **Progressive training:** curriculum upgrading adapter configuration.



Design: Online Configurator

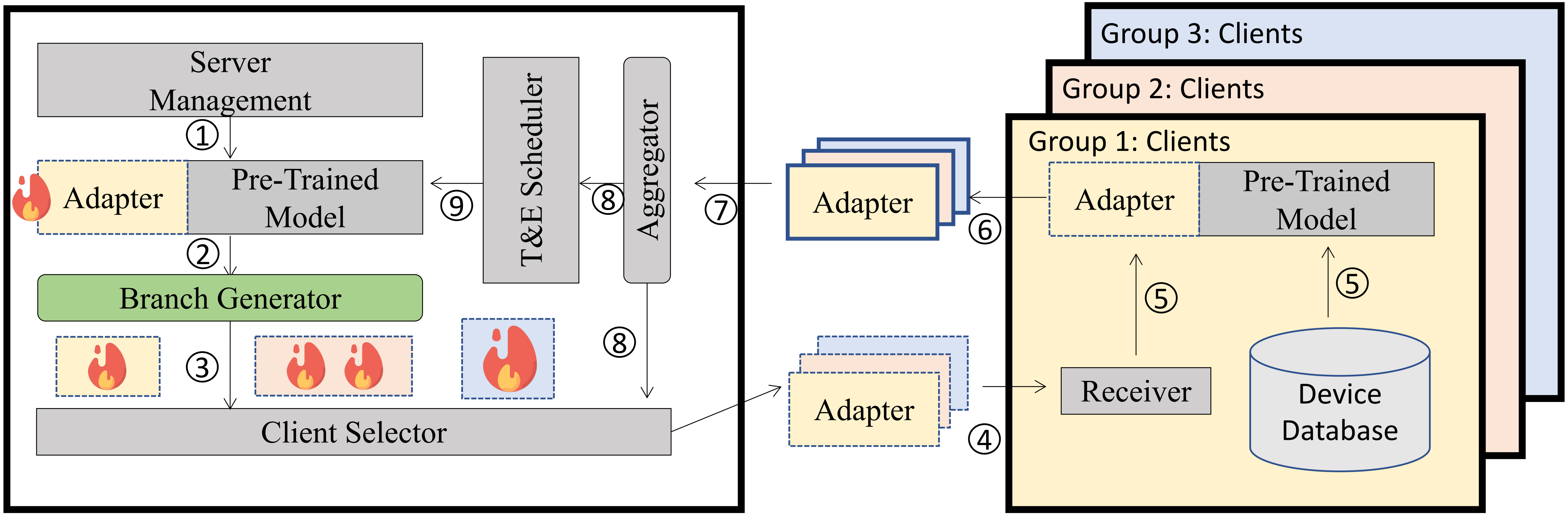
- **Progressive training:** curriculum upgrading adapter configuration.



When and how to upgrade the configuration?

Design: Online Configurator

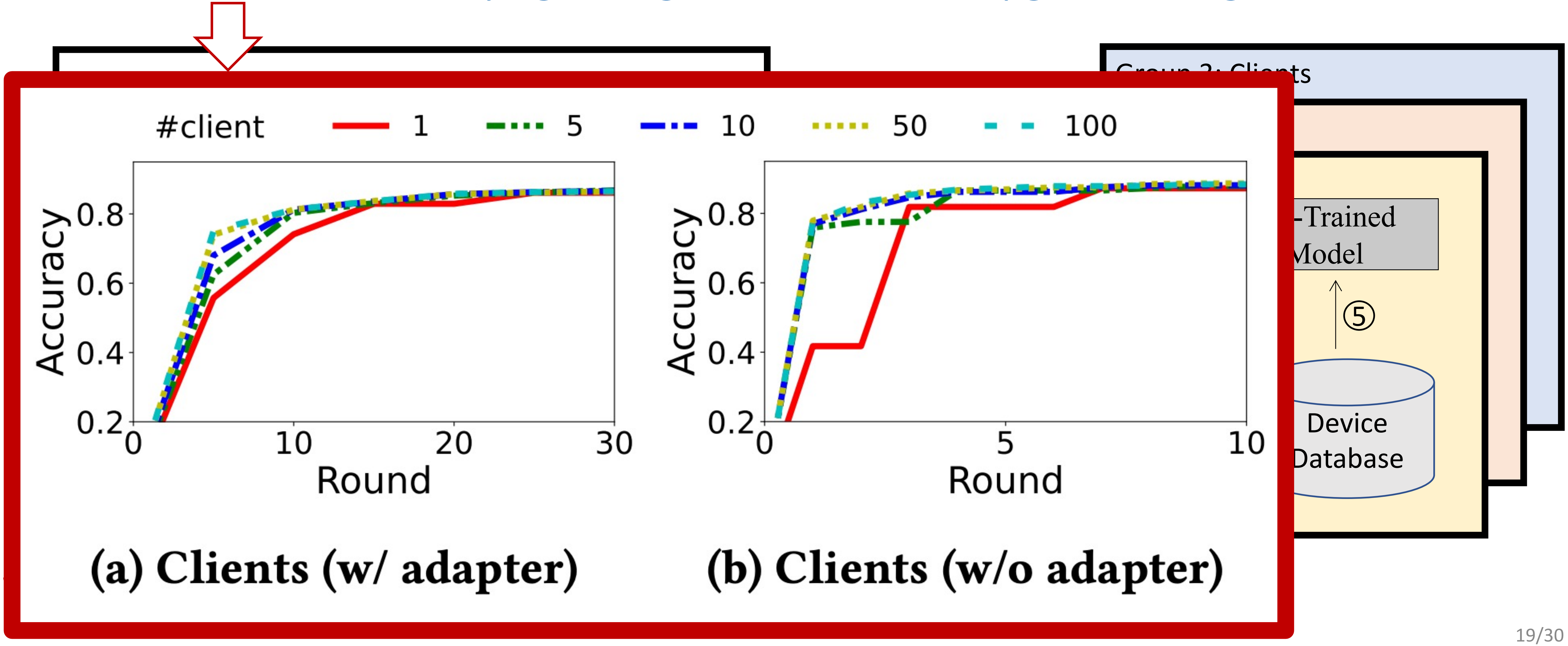
- **Progressive training:** curriculum upgrading adapter configuration.
- **Sideline trails:** identifying timing and direction to upgrade configuration.



When and how to upgrade the configuration?

Design: Online Configurator

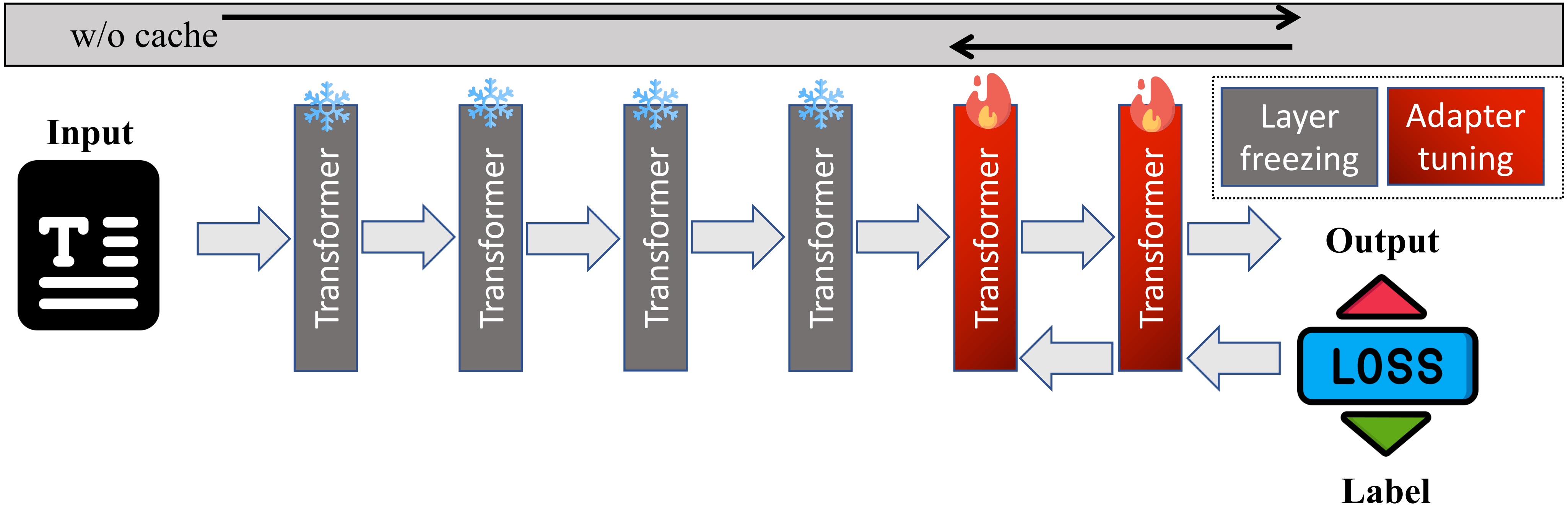
- **Progressive training:** curriculum upgrading adapter configuration.
- **Sideline trails:** identifying timing and direction to upgrade configuration.



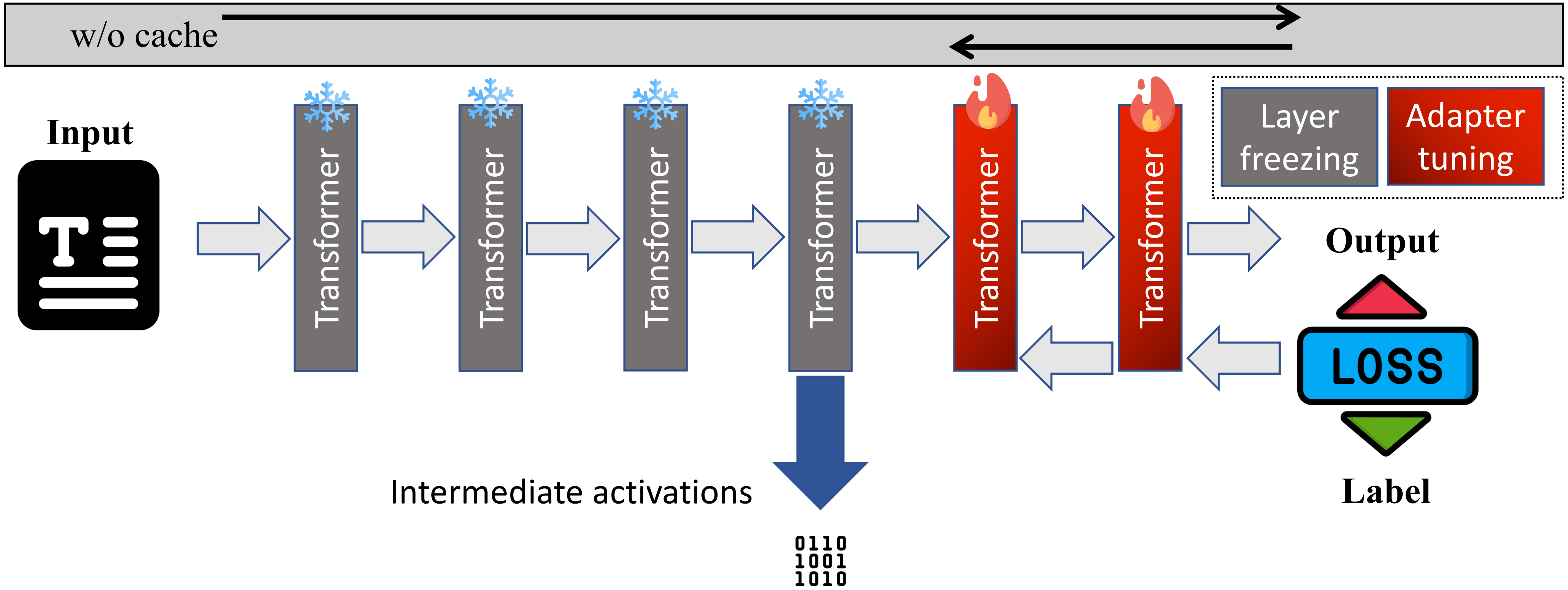
(a) Clients (w/ adapter)

(b) Clients (w/o adapter)

Further optimization: Activation Cache

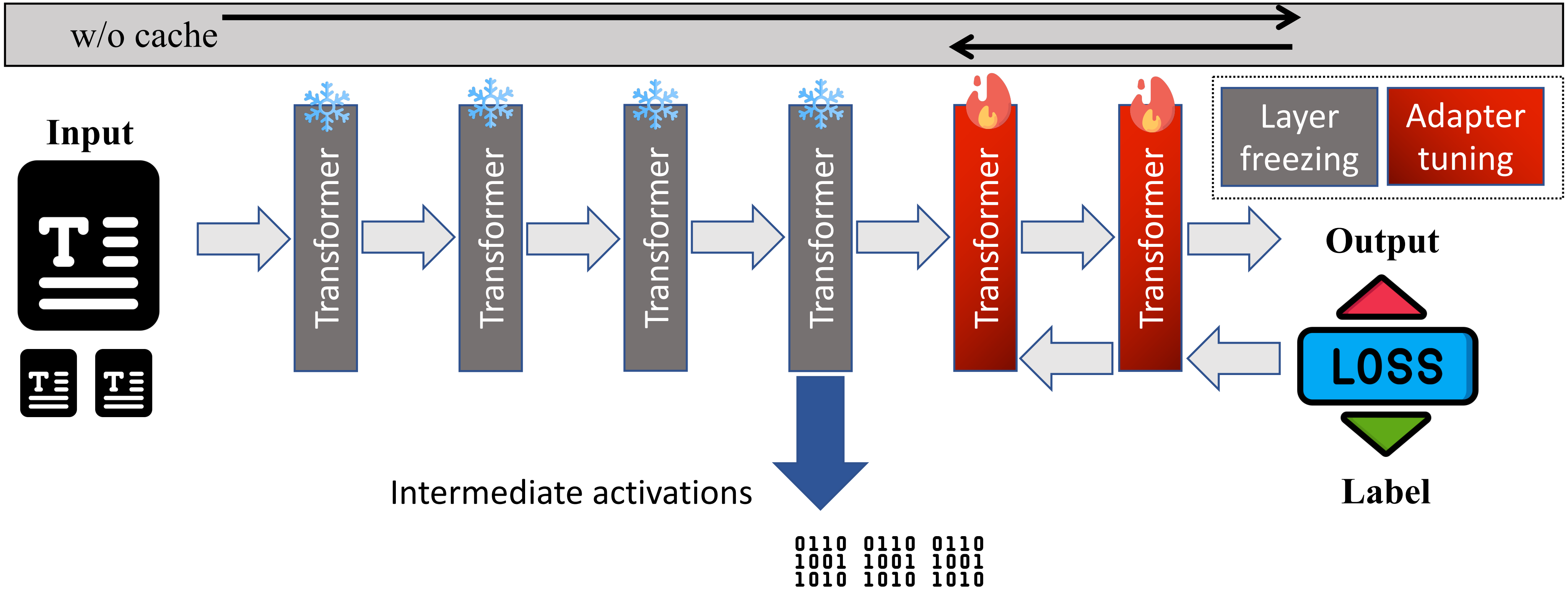


Further optimization: Activation Cache



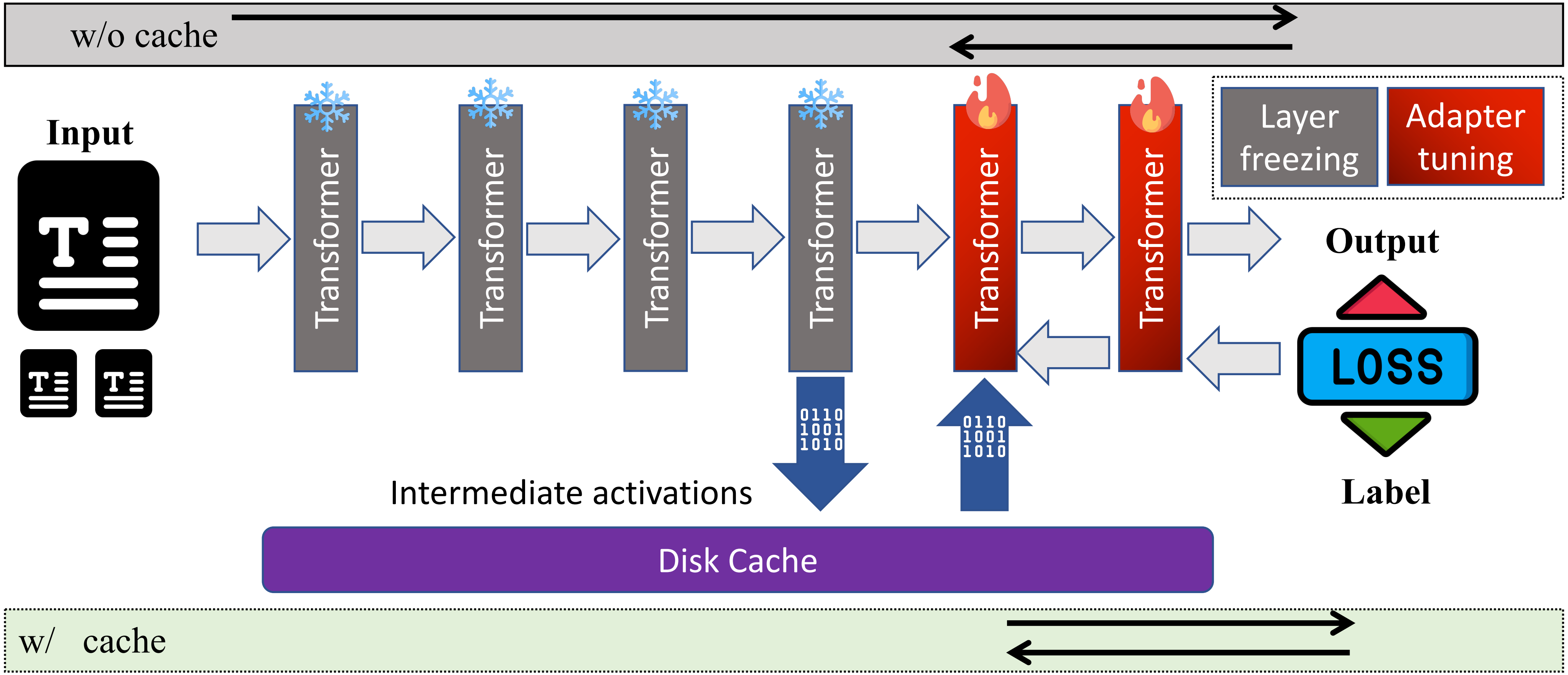
Further optimization: Activation Cache

An unique opportunity: Most of the Transformer parameters are freezing.



Further optimization: Activation Cache

An unique opportunity: Most of the Transformer parameters are freezing.



Evaluation: Setup

- **Implementation**
  - FedNLP[1]
  - AdapterHub[2]
- **Setups**
  - 3 devices
  - 2 models (BERT & DistilBERT)
  - 4 datasets
- **Baselines**
  1. Vanilla Fine-Tuning (FT)
  2. FineTuning-Quantized (FTQ)
  3. LayerFreeze-Oracle (LF$_{oracle}$)
  4. LayerFreeze-Quantized-Oracle (LFQ$_{oracle}$)

Device	**Processor**	**Per-batch Latency (s)**
Jetson TX2 [1]	256-core NVIDIA Pascal™ GPU.	0.88
Jetson Nano [2]	128-core NVIDIA CUDA® GPU.	1.89
RPI 4B [3]	Broadcom BCM2711B0 quad-core A72 64-bit @ 1.5GHz CPU.	18.27

Task	**Dataset**	**# of Clients**	**Labels**	**Non-IID**	**Samples**
TC	20NEWS [44]	100	20	/	18.8k
TC	AGNEWS [92]	1,000	4	a=10	127.6k
TC	SEMEVAL [31]	100	19	a=100	10.7k
ST	ONTONOTES [60]	600	37	a=10	5.5k

[1] Yuchen Lin B, He C, Zeng Z, et al. FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks[J]. Findings of NAACL, 2022.
[2] Pfeiffer J, Rücklé A, Poth C, et al. AdapterHub: A Framework for Adapting Transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020: 46-54

Evaluation: End-to-end Performance

- Our system reduces model convergence delays significantly.

Datasets	20NEWS			AGNEWS			SEMEVAL			ONTONOTES		
Relative Accuracy	99%	95%	90%	99%	95%	90%	99%	95%	90%	99%	95%	90%
FT	44.0	23.4	13.1	31.1	10.1	5.2	124.3	89.9	61.7	76.1	55.9	35.6
FTQ	12.7	6.8	3.8	9.1	2.6	1.7	32.0	23.1	15.9	21.2	15.5	9.9
LF_{oracle}	18.5	8.1	4.3	9.6	1.4	1.1	74.0	46.8	33.2	82.5	43.8	24.5
LFQ_{oracle}	5.2	2.5	1.1	1.6	0.3	0.2	16.8	11.0	7.7	23.9	12.9	7.2
AdaFL	**1.3**	**0.4**	**0.1**	**0.2**	**0.03**	**0.02**	**2.3**	**1.1**	**0.6**	**4.5**	**2.4**	**1.3**

260×

Table 1: Elapsed training time taken to reach different relative target accuracy. NLP model: BERT-base. Unit: Hour.

Evaluation: System Scalability

- Our system outperforms baselines Our systemrious **network** environments
- It outperforms baselines on various client **hardware**.

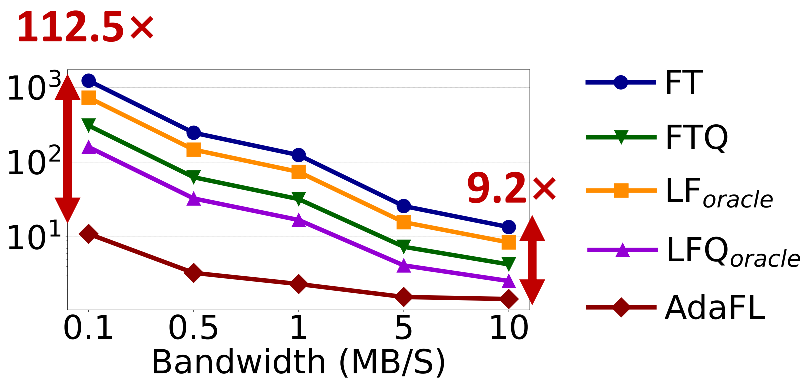


Fig. 1: Model convergence delays under different network bandwidths. Training targets 99% relative target accuracy.

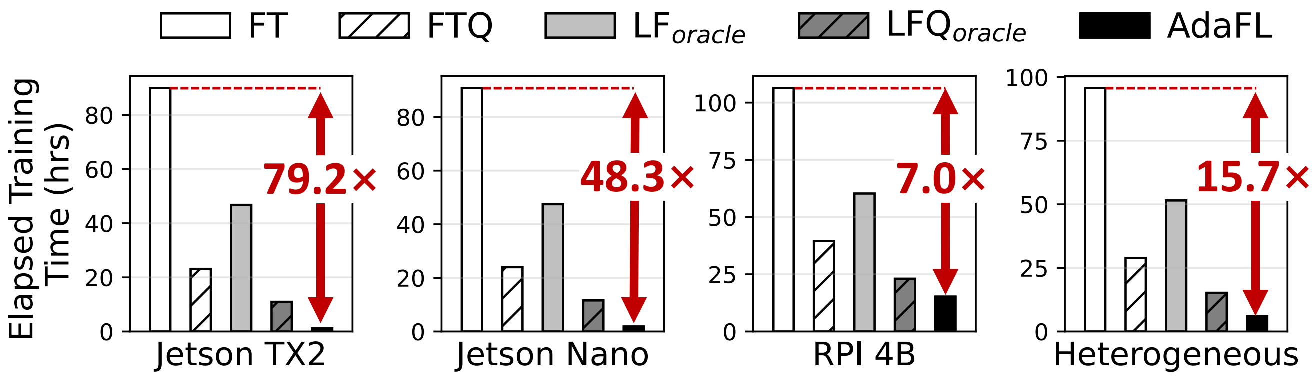


Fig. 2: Model convergence delays with a variety of client hardware. 'Heterogenous' means a mixture of heterogeneous hardware capacity.

Evaluation: Key deign

- Our key designs contribute to the results significantly.

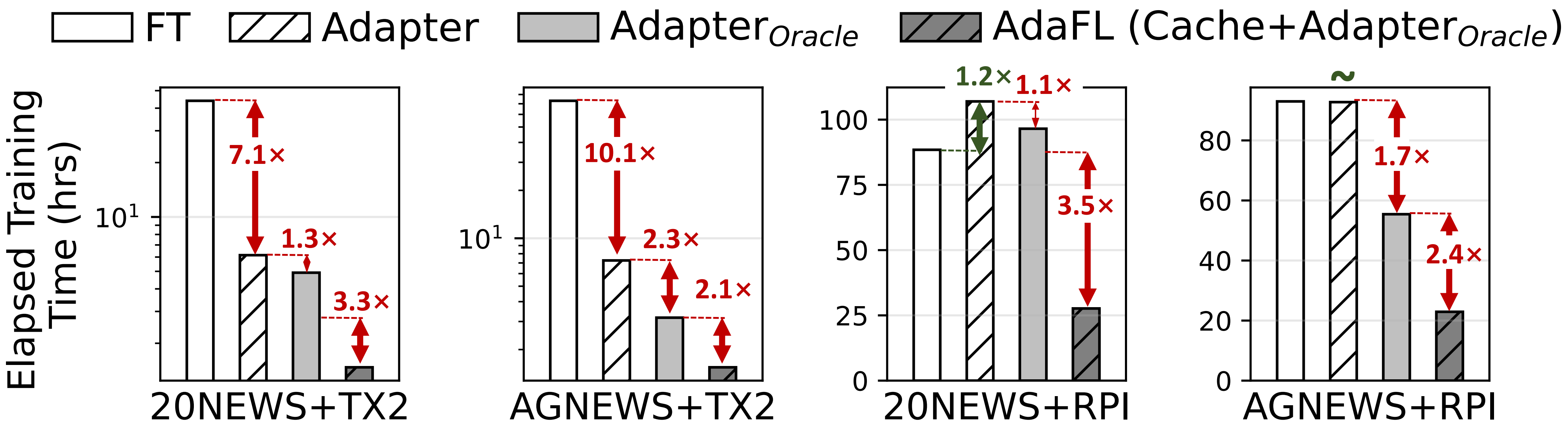


Fig. 1: Model convergence delays with and without our system's key designs, showing their significance.

Evaluation: System Cost

<u>Our system is resource-efficient.</u>

- It saves up to 220.7× **network traffic**. (Fig. 1)
- It reduces up to 32.2× **energy consumption**. (Fig. 2)
- It nontrivially reduces the **memory usage**. (Fig. 3)

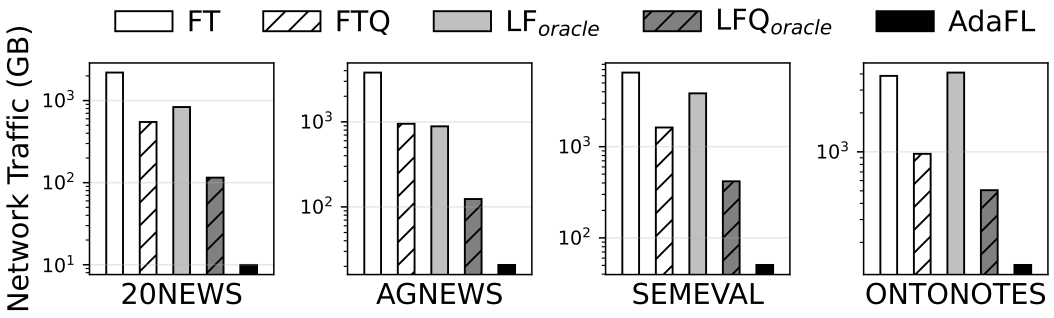


Fig. 1: Network traffic (downlink and uplink) of all 15 client devices.

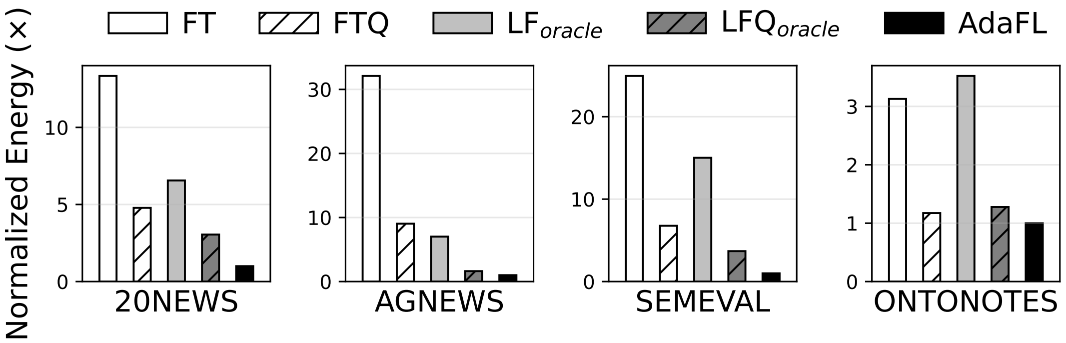
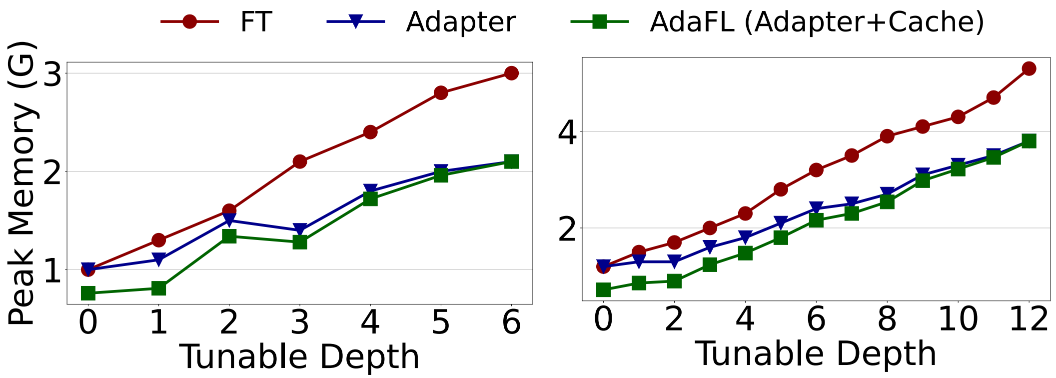


Fig. 2: Per-client average energy consumption, normalized to that of ours.



(a) DistilBERT **(b) BERT**

Fig. 3: Peak memory usage of a client device.

Conclusion

- Our system is a federated learning framework for fast NLP model fine-tuning.
- It uses adapter as the only trainable module in NLP model to reduce the training cost.
- To identify the optimal adapter configuration on the fly, it integrates a progressive training paradigm and trail-and-error profiling technique.
- It can reduce FedNLP's model convergence delay to no more than several hours, which is up to 155× faster compared to vanilla FedNLP and 48× faster compared to strong baselines.

Conclusion

- Our system is a federated learning framework for fast NLP model fine-tuning.
- It uses adapter as the only trainable module in NLP model to reduce the training cost.
- To identify the optimal adapter configuration on the fly, it integrates a progressive training paradigm and trail-and-error profiling technique.
- It can reduce FedNLP's model convergence delay to no more than several hours, which is up to 155× faster compared to vanilla FedNLP and 48× faster compared to strong baselines.

Thanks for listening!

<Efficient Federated Learning for Modern NLP>

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, Mengwei Xu

Scan for our code!