DESIGN, AUTOMATION & TEST IN EUROPE
25 – 27 March 2024 · Valencia, Spain
The European Event for Electronic System Design & Test

Communication-Efficient Model Parallelism for Distributed In-Situ Transformer Inference

Yuanxin Wei, Shengyuan Ye, Jiazhi Jiang,
Xu Chen, Dan Huang*, Jiangsu Du*, Yutong Lu

Sun Yat-sen University, Guangzhou, China

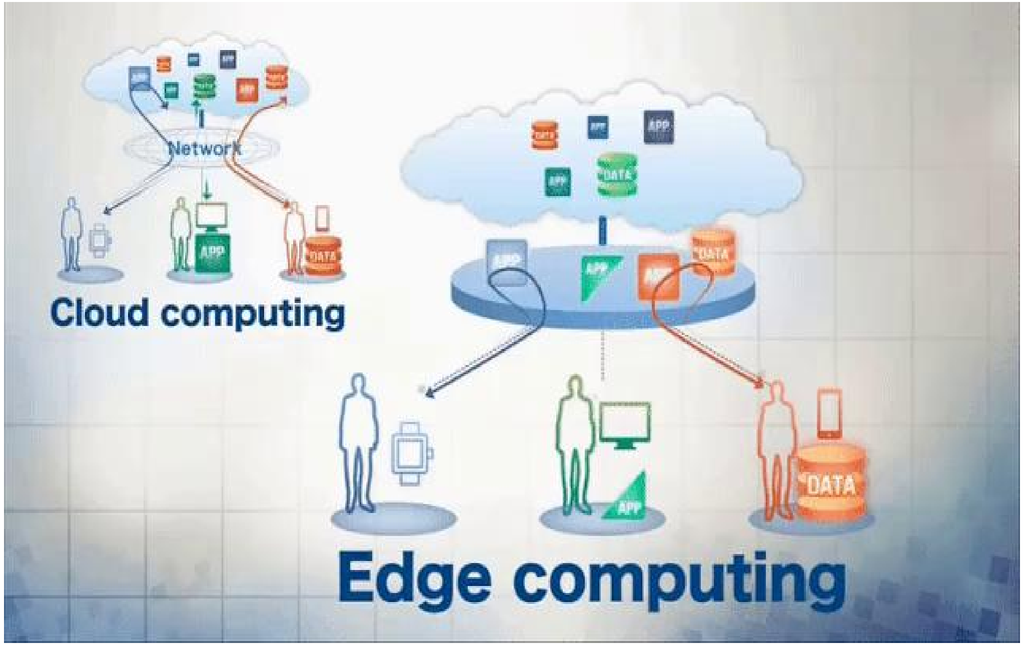
中山大學 SUN YAT-SEN UNIVERSITY

国家超级计算广州中心 NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

Background and Motivation

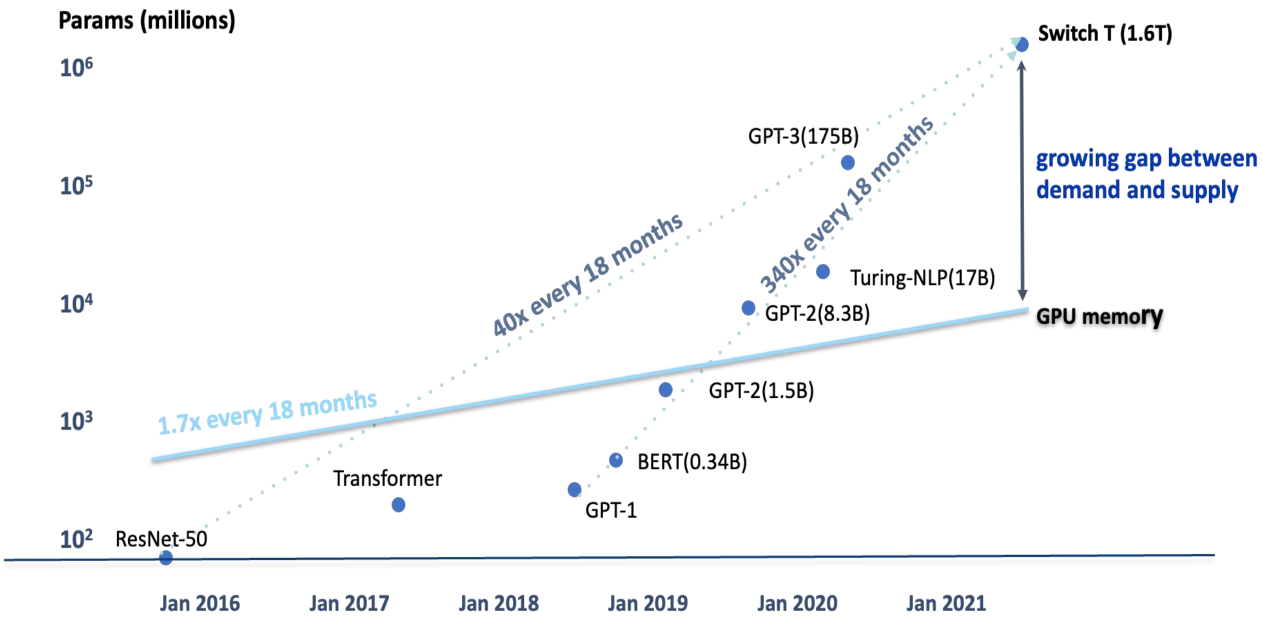
- **In-Situ Transformer Inference**

  - **Enhanced Privacy**
  - **Improved Efficiency**
  - **Better Robustness**



Background and Motivation

- **Memory Pressure Caused By Model Size**



Opportunity ->

Edge Collaborative Transformer Inference

Background and Motivation

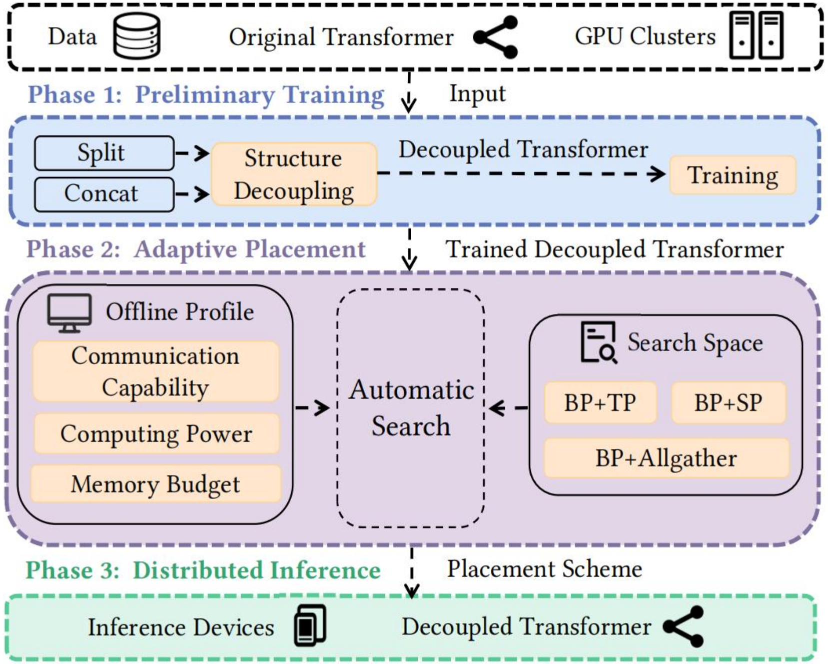
- **Parallelism Approaches in Distributed Inference**

Metrics	Data Parallelism (DP)	Pipeline Parallelism (PP)	Model Parallelism		
			Tensor Parallelism (TP)	Sequence Parallelism (SP)	<u>Block Parallelism (BP,Ours)</u>
Latency Reduction	×	×	√	√	√
Throughput Increase	√	√	√	√	√
Memory Reduction	×	√	√	×	√
Communication Friendly	×	√	×	×	√

Communication Bottleneck！

System and Methodologies

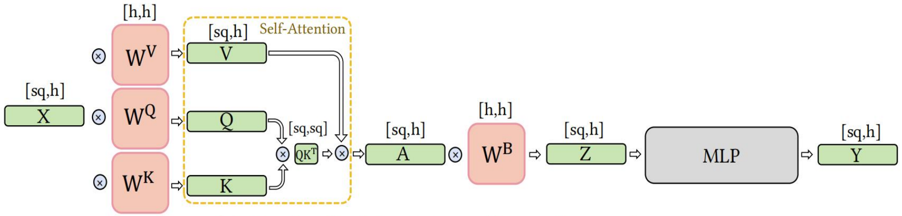
- **DeTransformer Overview**



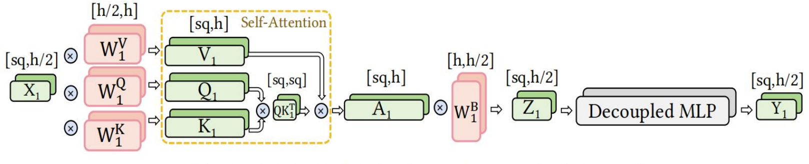
System and Methodologies

- **Block Parallelism through Structure Decoupling**

(1) Decouple an original layer into one decoupled layer:

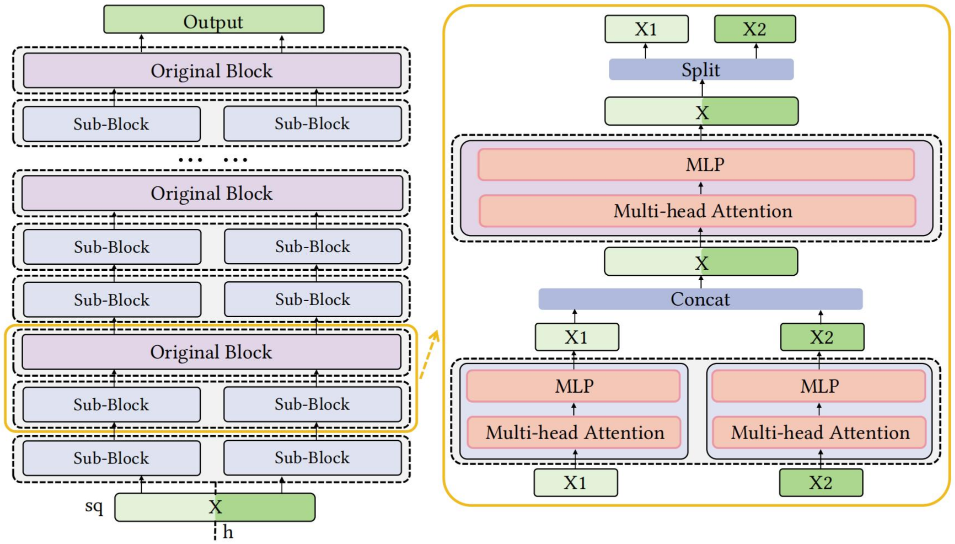


(a) Single Block in Original Transformer Layer



(b) Multiple Sub-Blocks in Decoupled Transformer Layer

(2) Build the decoupled Transformer model by stacking both the original layer and the decoupled layer:



System and Methodologies

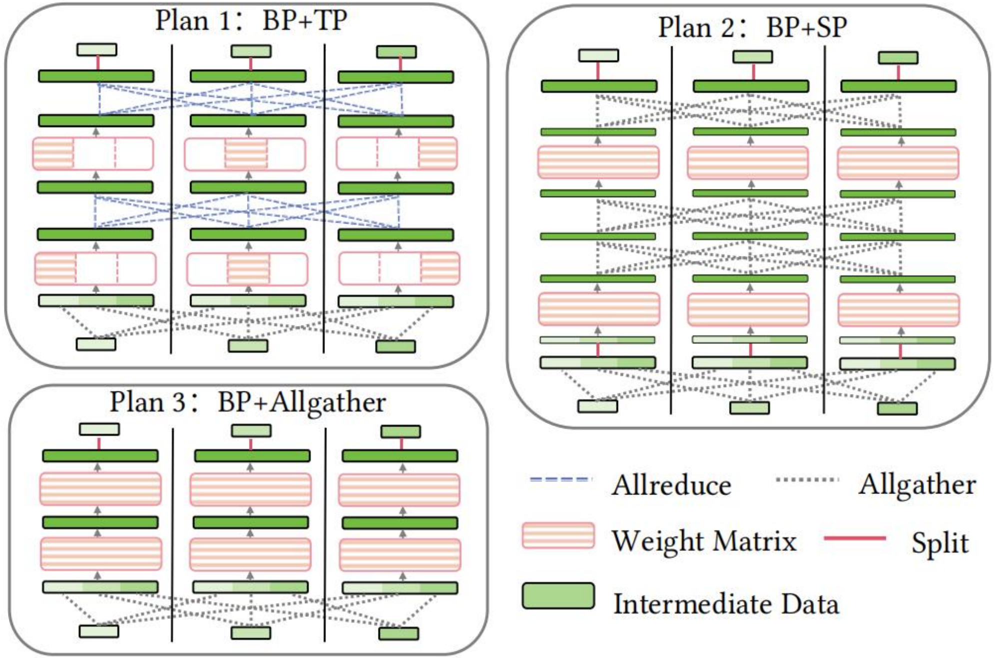
- **Adaptive Placement Approach**

Strike a balance among:

- Communication capability

- Computing power

- Memory budget



Evaluation

- **Accuracy Experiments**

BERT-Base (l = 12, h = 768, N_h = 12, 110M params)

BERT-Large (l = 24, h = 1024, N_h = 16, 340M params)

4 * NVIDIA A800 (80GB)

English Wikipedia data corpus (2.5B words)

5 downstream tasks: CoLA, SST-2, MRPC and MNLI from the popular GLUE benchmark and the SQUAD v1.1 benchmark

Model	N_b	N_{div}	GLUE Mcc/Acc(%)				SQUAD
			CoLA	SST-2	MRPC	MNLI	Acc(%)
Original Bert-Base	\	\	40.43	91.28	84.56	81.59	77.54
Decoupled Bert-Base (Ours)	1	4	39.85	89.45	80.64	78.80	74.77
	2	4	**41.26**	89.68	**87.75***	80.54	75.82
	3	4	**41.20**	90.37	83.82	80.77	76.66
	4	4	**42.06**	**92.20***	**86.52**	81.21	76.91
	4	8	36.75	**91.28**	83.58	80.40	74.82
	4	2	**47.51***	89.91	**84.56**	**81.90***	**78.37***
Original Bert-Large	\	\	51.00*	91.39*	80.39	81.73	79.29
Decoupled Bert-Large (Ours)	4	4	47.20	90.82	**83.82**	**81.96***	78.85
	6	4	47.40	91.28	**85.04**	**81.81**	**79.86**
	8	4	44.25	90.37	**86.01***	**81.85**	**79.88***

Comparable accuracy results

Evaluation

• **Performance Experiments**

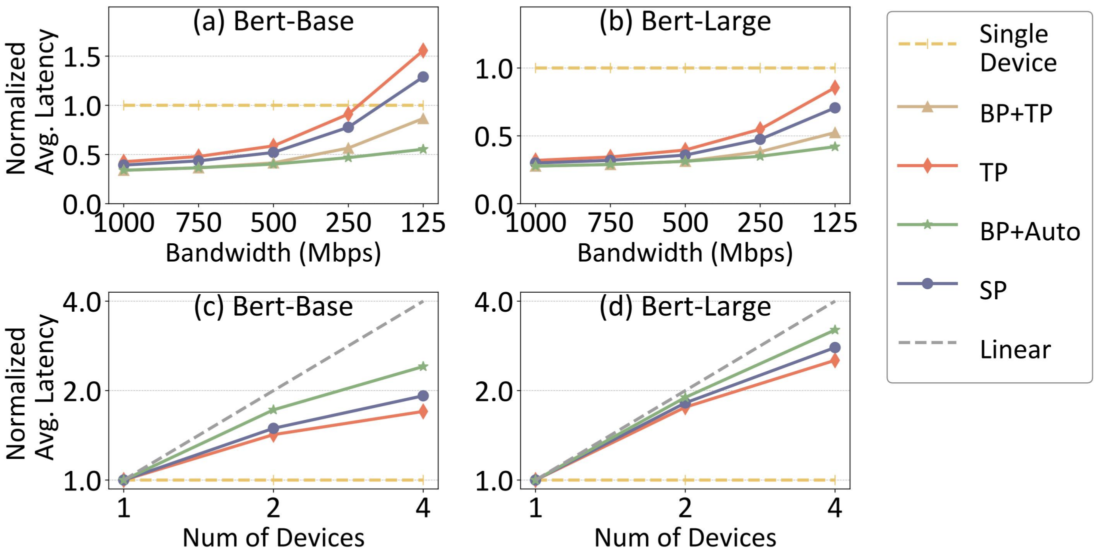
Edge devices: 4 * Raspberry Pi 4B

(a)(b)
Latency under various network conditions

Lower Latency ✓

(c)(d)
Throughput across different num of devices

Superior strong scaling ability ✓



Thanks for listening！