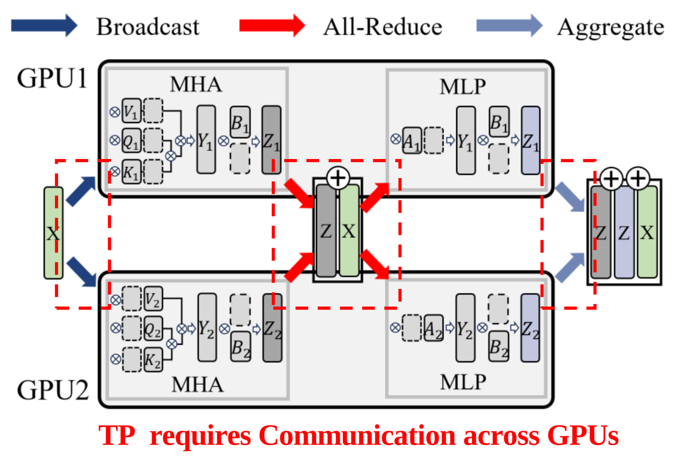
First Attentions Last: Better Exploiting First Attentions for Efficient Transformer Training

Gyudong Kim[1], Hyukju Na[1], Jin Hyeon Kim[1], Hyunsung Jang[2], Jaemin Park[2], Jaegi Hwang[2], Namkoo Ha[2], Seungryong Kim[3]†, Young Geun Kim[1]†

[1]Korea University. [2]LIG Nex1 Co., Ltd. [3]KAIST AI.

NEURAL INFORMATION PROCESSING SYSTEMS

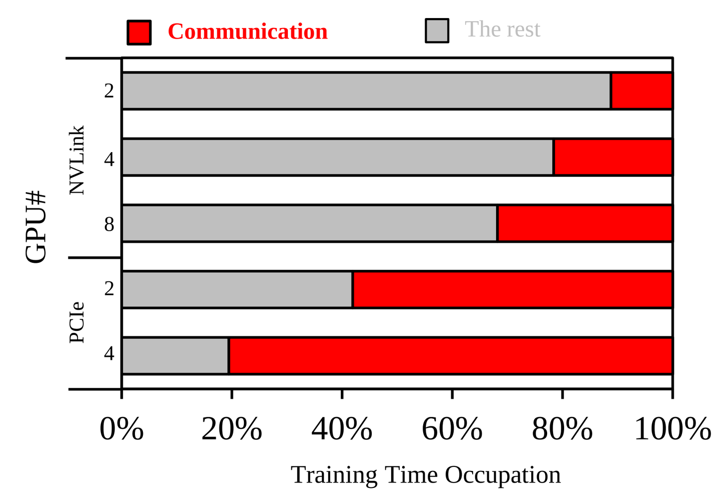
Target: Redundant Communication in Distributed Training



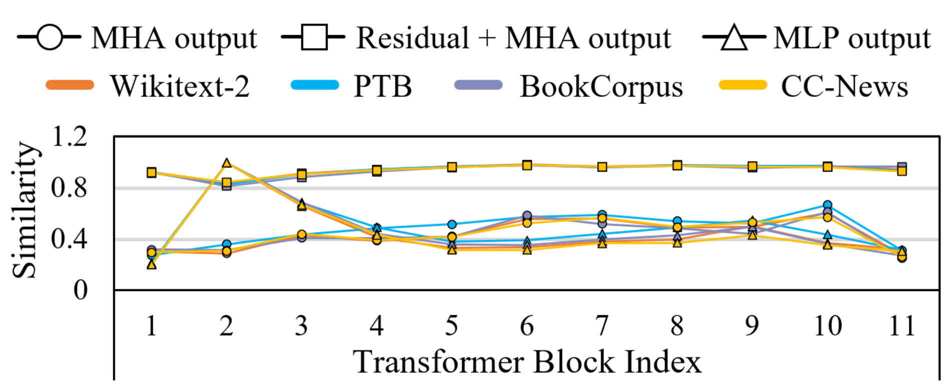
Sequential MHA→MLP linkage forces an All-Reduce in the block
(50% of the communication)

HW FLOPS: 60,000x / 20 yrs
DRAM BW: 100x / 20 yrs
Interconnect BW: 30x / 20yrs

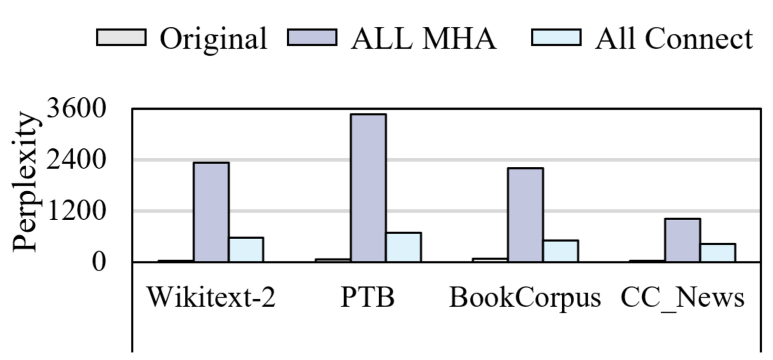
Gholami et al. "AI and Memory Wall"



Observation-①: MHA-MLP Connections Can Be Reconfigured



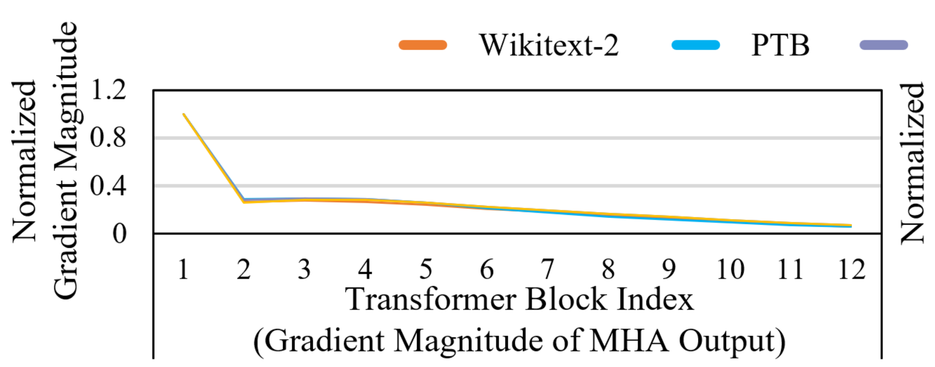
(a) Similarity analysis



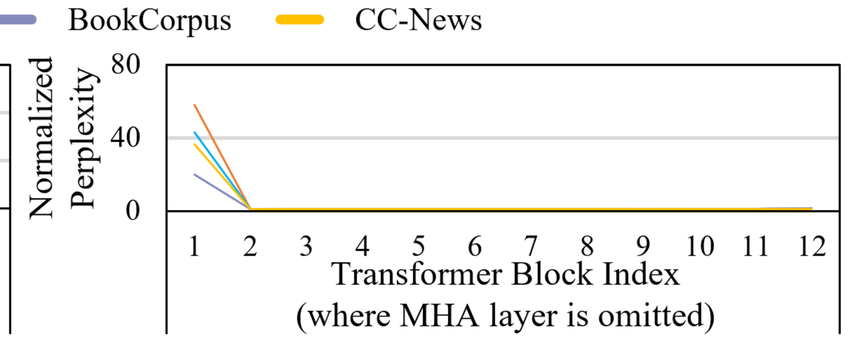
(b) Connection ablation

- Similarity analysis (a) shows that once MHA outputs pass through the residual path, MLP inputs across layers become almost identical.
- Connection ablation (b) indicates this dependency can be safely redesigned.

Observation-②: First Attention is Key



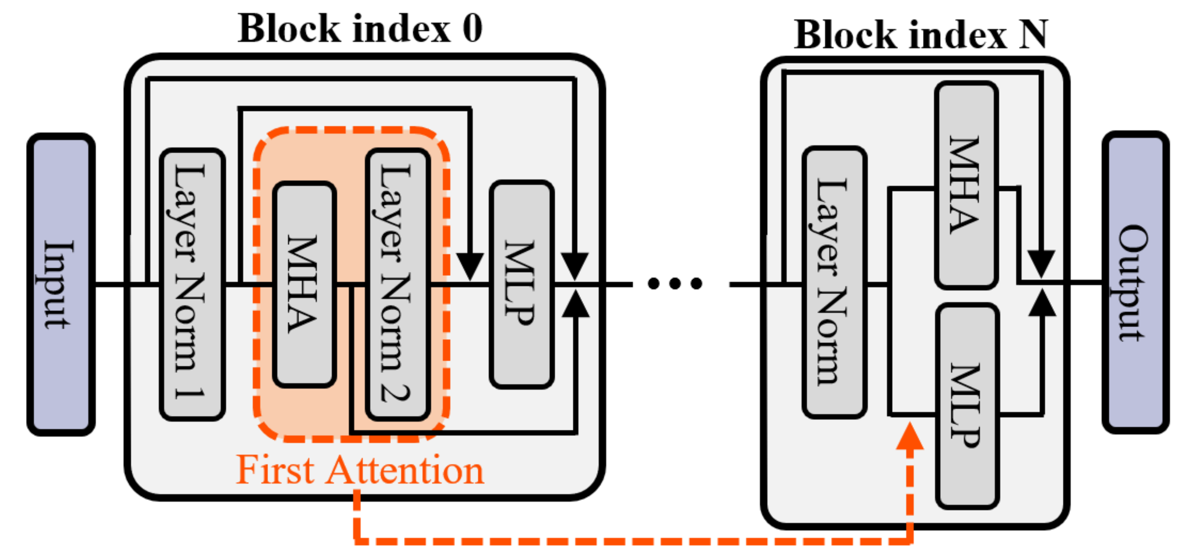
(a) Gradient magnitude analysis



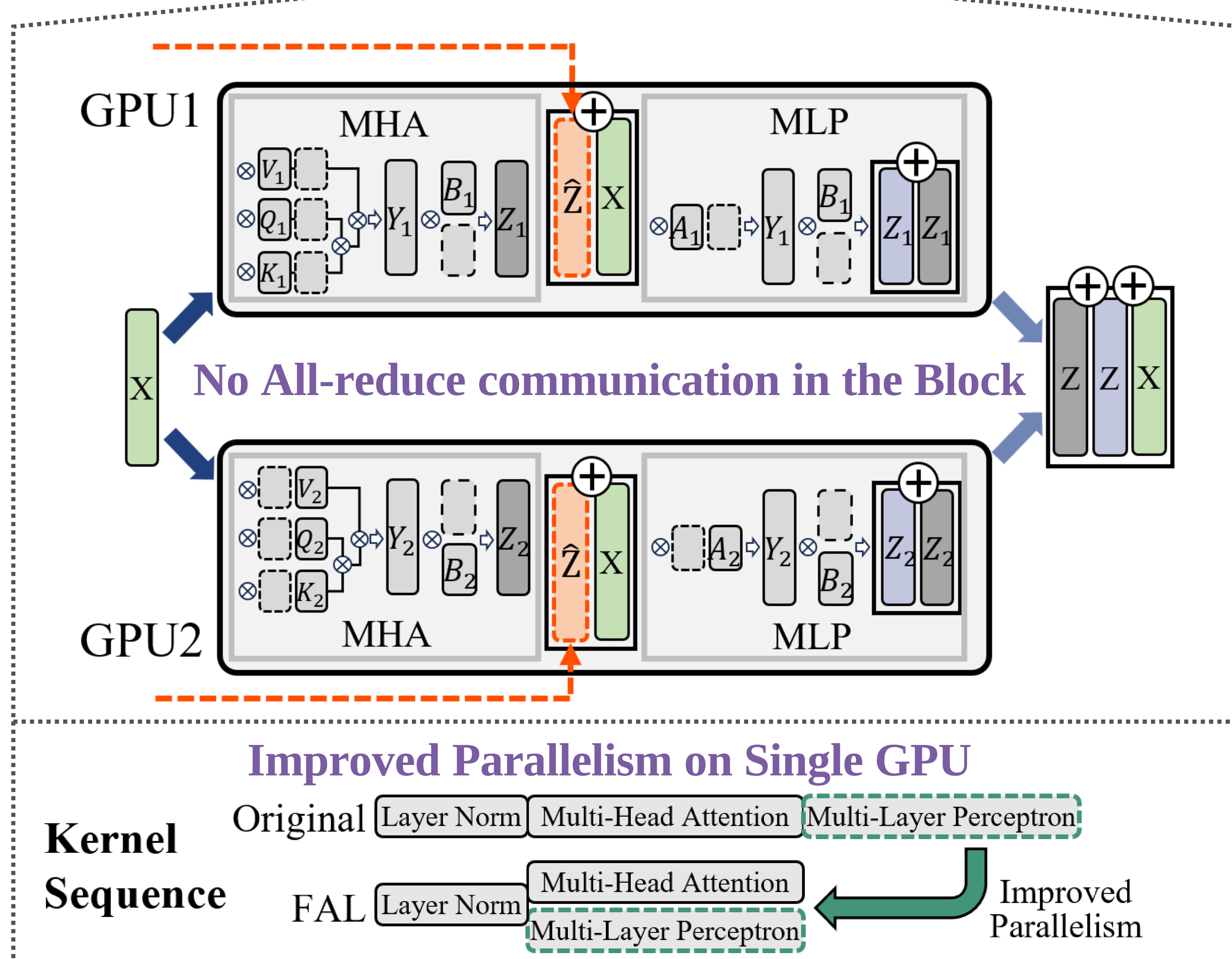
(b) Layer ablation

- Gradient analysis (a) and ablation analysis (b) show that the first attention layer has a much larger influence on the loss than later ones.
- This suggests the **first attention as the key signal for connection redesign.**

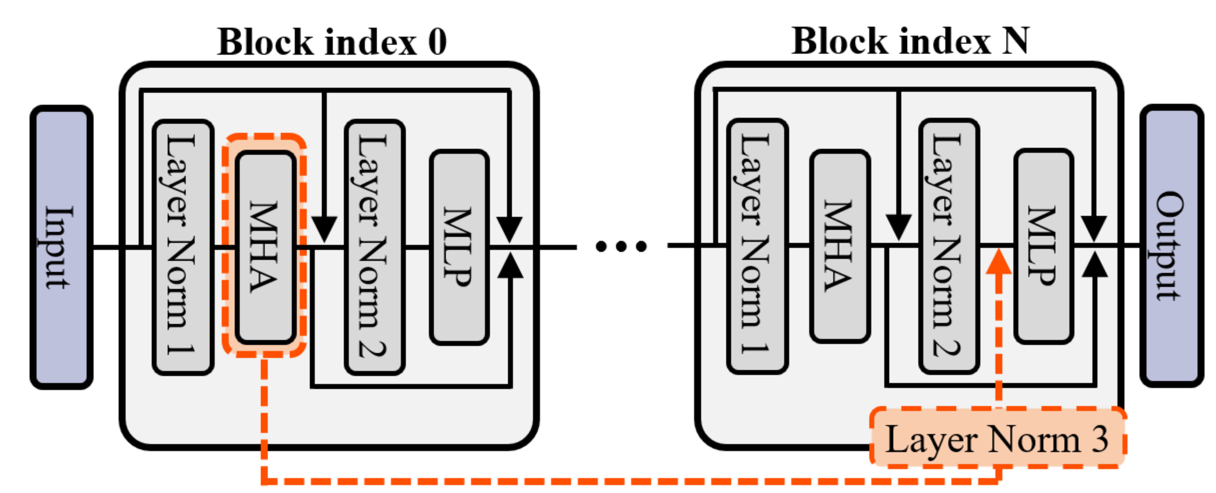
Proposed Design-①: FAL



FAL redirects the MLP input from the current MHA to the first attention, replacing the redundant per-block MHA→MLP dependency

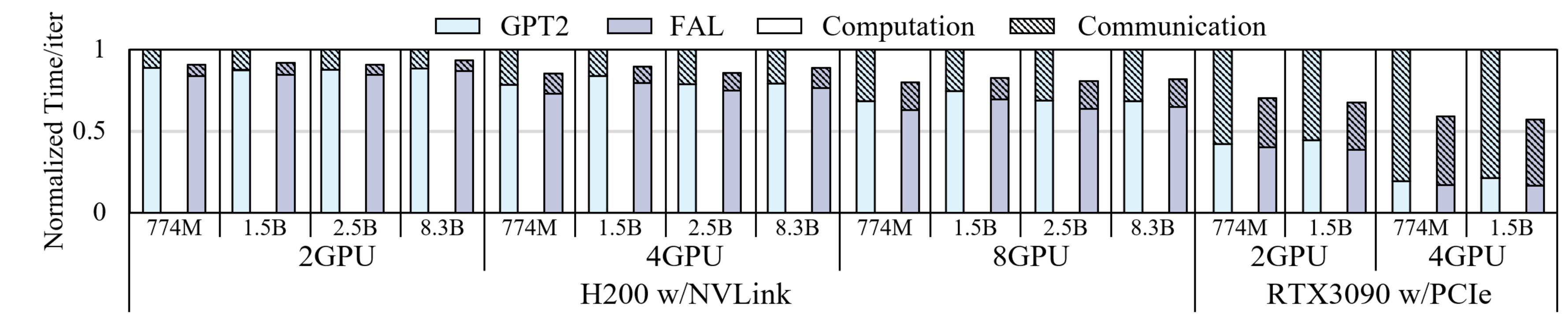


Proposed Design-②: FAL+

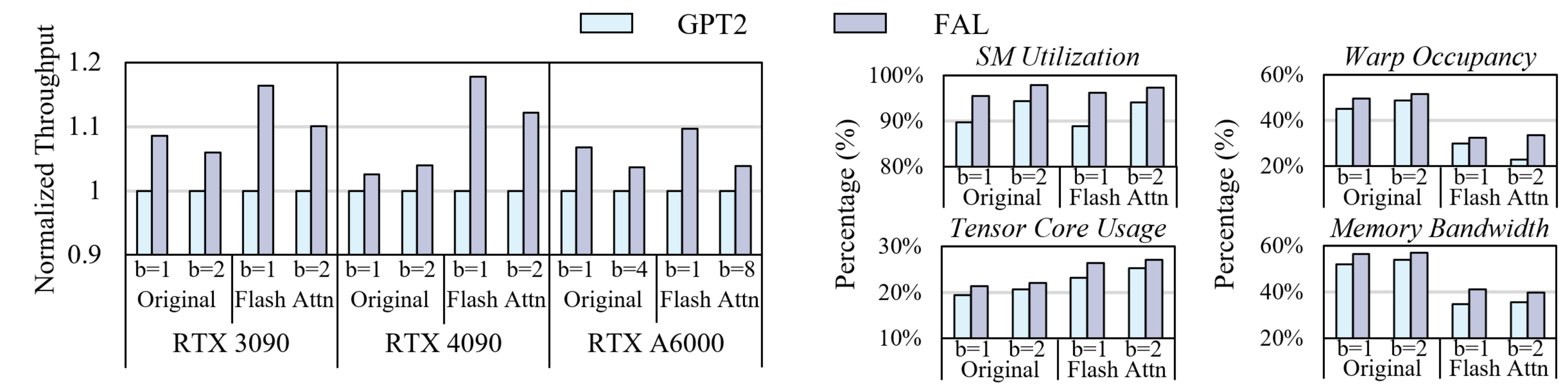


FAL+ augments the original MHA→MLP path with the first attention signal to enhance model quality.

Evaluation: Faster Speed, Higher Quality



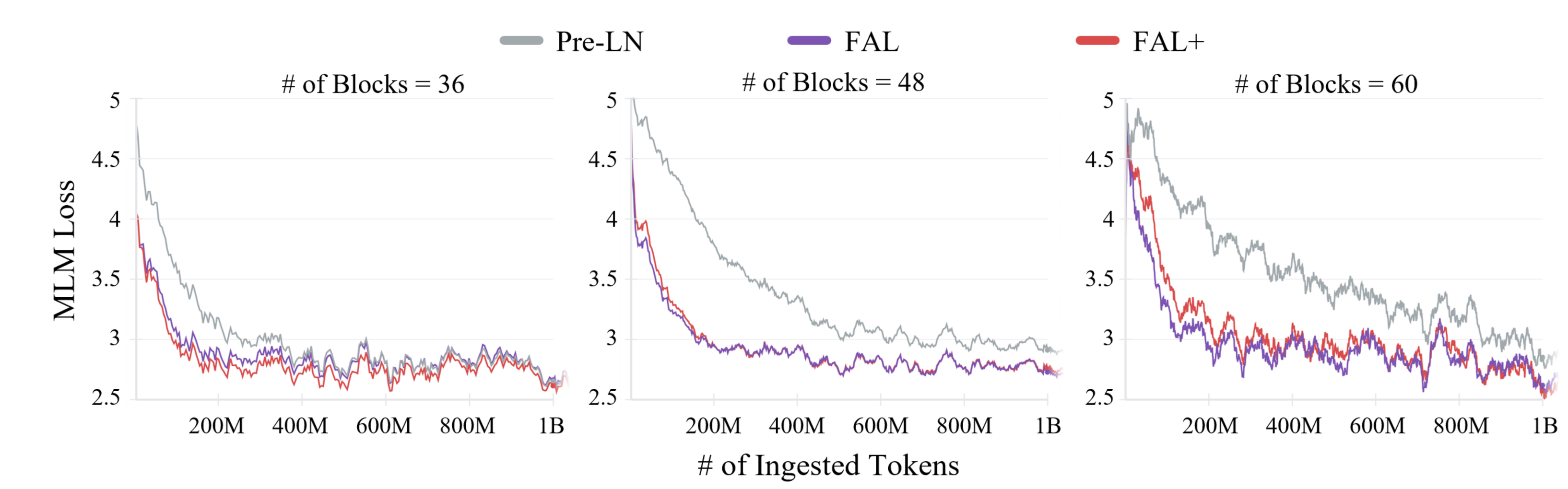
Shortens training time by up to 20.1% (NVLink) and 43.1% (PCIe)



Improved Single GPU throughput by up to 1.18×

Openwebtext (↓)		SuperGLUE (↑) (CB, Record: F1 score, Others: Accuracy)								
Model	PPL / Time	BoolQ [42]	CB [43]	COPA [44]	MultiRC [45]	ReCoRD [46]	RTE [47]	WIC [48]	WSC [49]	Avg.
GPT-2 774M	17.75 / 13.2d	**55.7**	19.4	54.0	52.3	**57.4**	**54.2**	49.8	45.2	48.5
Parallel	17.80 / **8.6d**	50.0	19.4	58.0	53.8	48.6	51.6	49.1	36.5	45.9
FAL	17.55 / **8.6d**	50.2	**21.4**	**62.0**	54.5	52.6	51.6	46.6	**49.0**	48.5
FAL+	**17.24** / 13.2d	51.8	21.1	58.0	**55.7**	56.2	51.3	**51.3**	48.1	**49.2**
GPT-2 1.5B	14.72 / 24.1d	58.0	24.1	65.0	**57.2**	78.4	53.1	**50.0**	40.4	53.3
FAL	14.23 / **16.1d**	58.1	21.6	**72.0**	**57.2**	78.7	54.2	49.2	**64.4**	**56.9**
FAL+	**14.12** / 24.2d	**58.8**	**26.2**	65.0	**57.2**	**79.0**	**56.0**	49.8	51.0	55.4

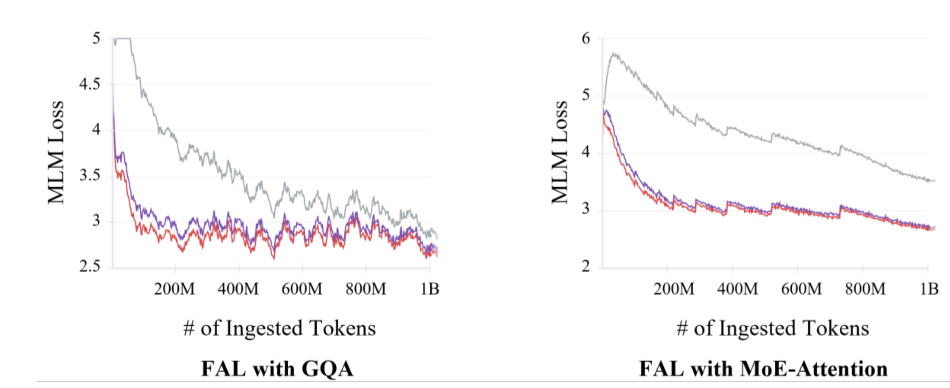
Achieved lower perplexity, higher SuperGLUE score



More effective as model depth increases

Model	LR	Δ Val PPL	Trained PPL
GPT2 1.5B	1e-5	0.01	30.93
	1e-4	0.00	30.92
	1e-3	0.01	28.10
	1e-2	1.39	**4.83**
FAL+ 1.5B	1e-5	**0.00**	**27.97**
	1e-4	**0.00**	**27.98**
	1e-3	**-0.01**	**27.22**
	1e-2	**0.61**	5.76

Robust adaptability-plasticity trade-off



Generalize to Transformer variants