NEURAL INFORMATION PROCESSING SYSTEMS

Block-Diagonal LoRA for Eliminating Communication Overhead in Tensor Parallel LoRA Serving

aws

WARWICK
THE UNIVERSITY OF WARWICK

Xinyu Wang

Jonas M. Kübler

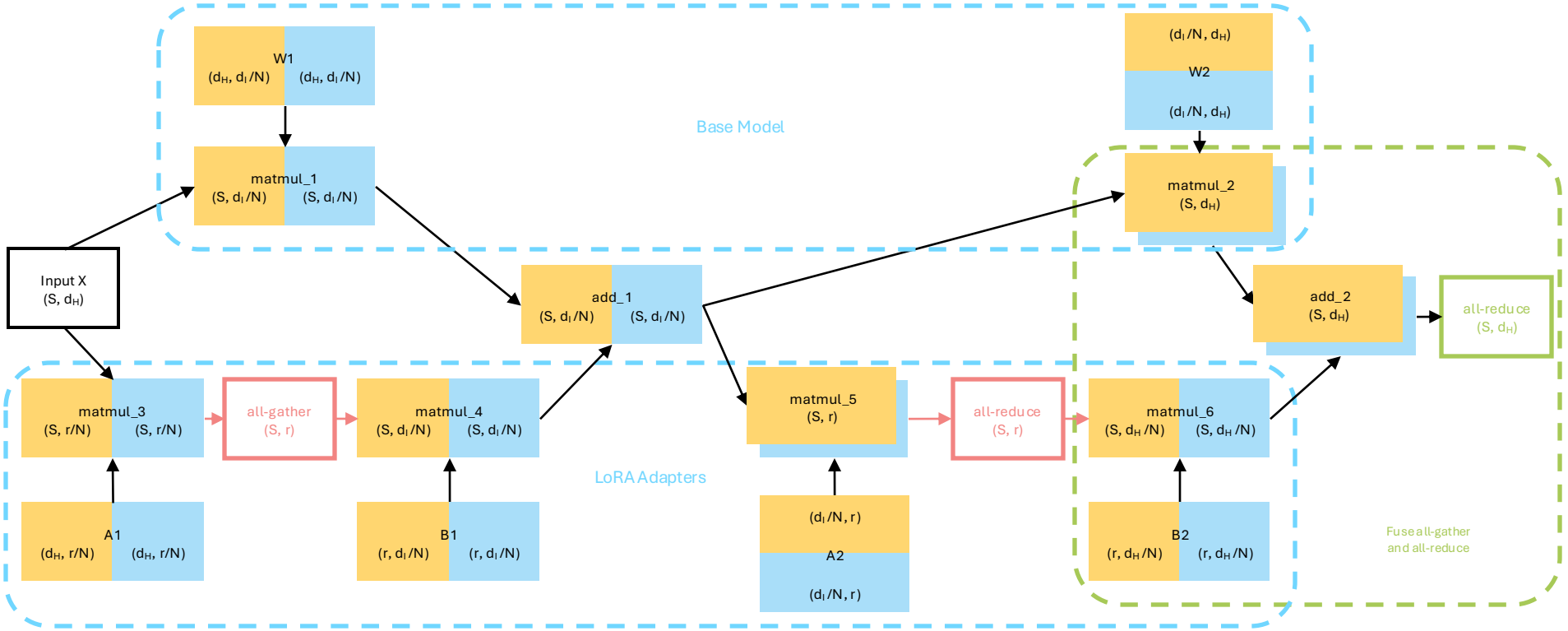
Kailash Budhathoki

Yida Wang

Matthäus Kleindessner

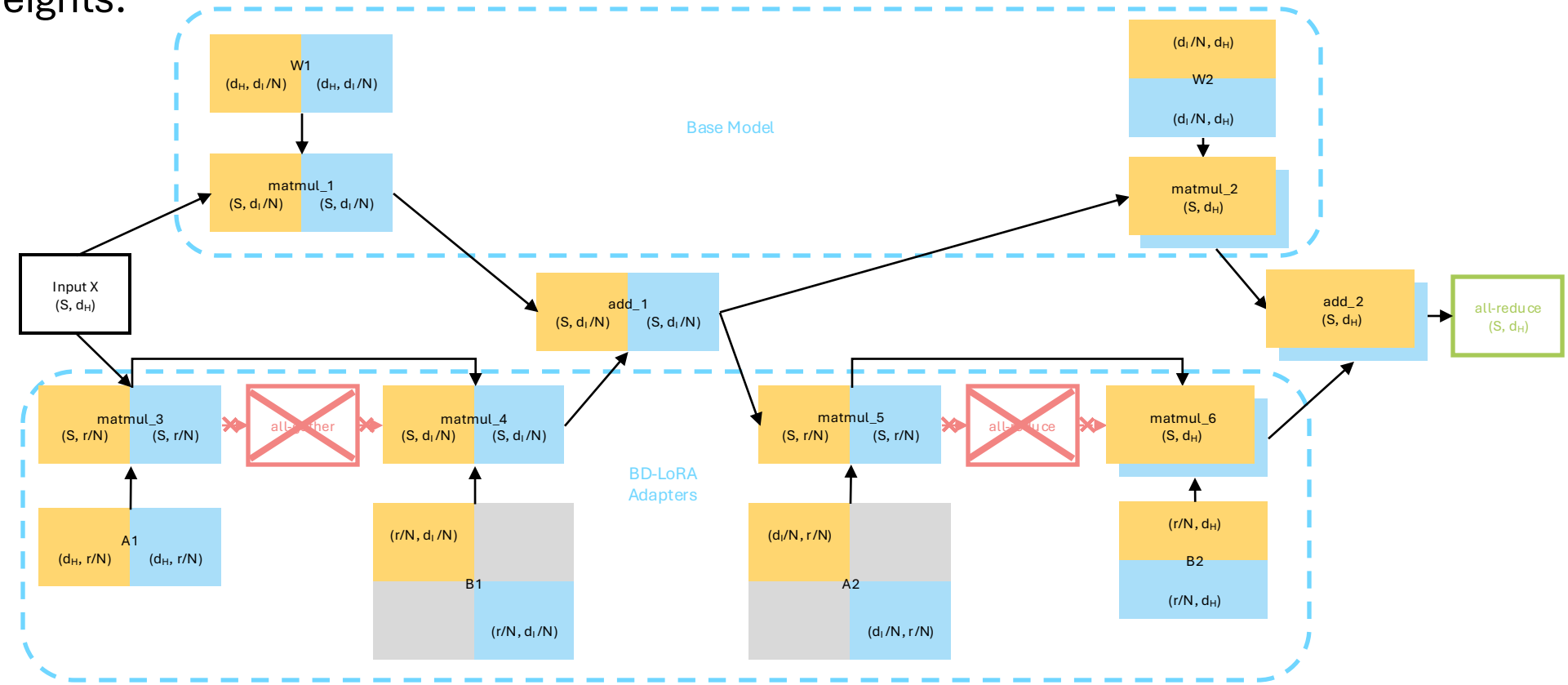
Problem and Motivation

- For multi-LoRA serving with a shared base LLM, merging adapters prevents batching and incurs swap overhead.
- S-LoRA TP strategy adds extra all-gather/all-reduce beyond the base model.
- The additional cost is significant in practice and exacerbated at larger ranks.



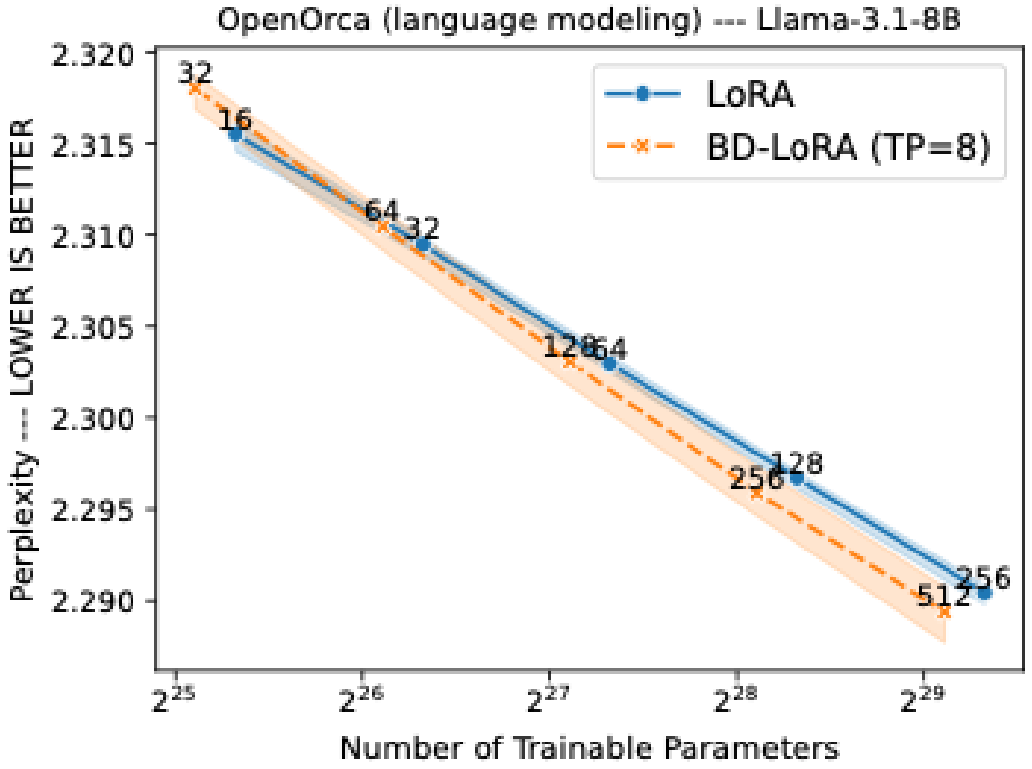
BD-LoRA

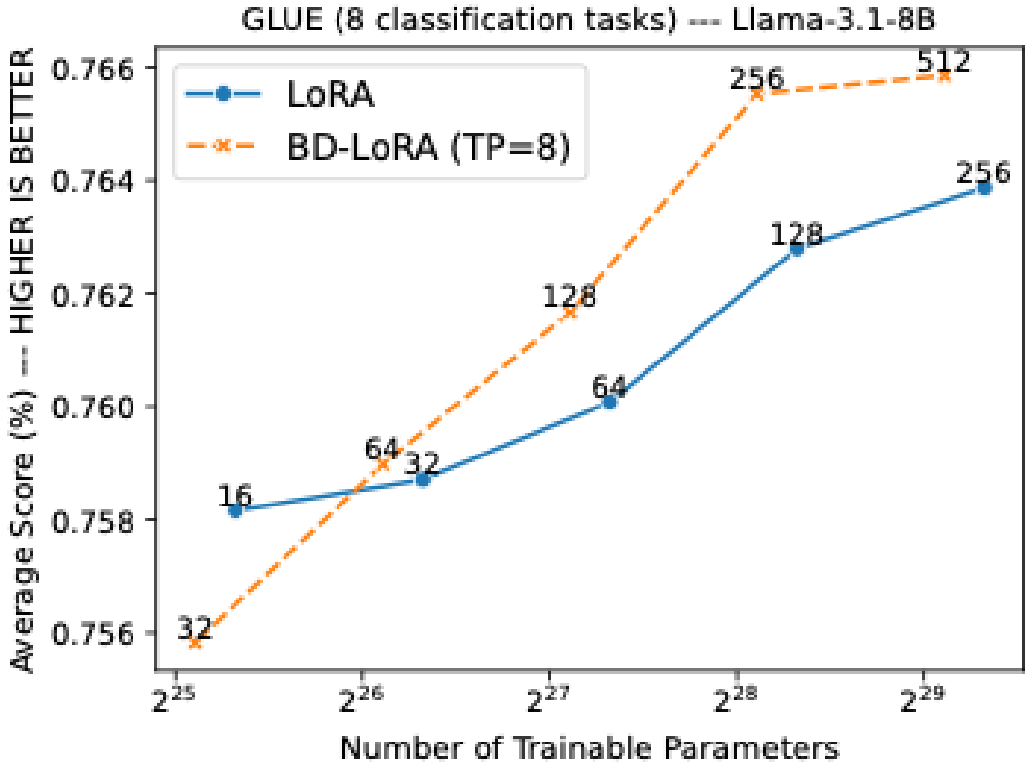
- Constrain specific LoRA factors (B_1 and A_2) to be block-diagonal so they align with base TP sharding -> no extra all-gathers/all-reduces for LoRA.
- BD-LoRA can be interpreted as attaching independent LoRA adapters to each shard of the base weights.



Downstream performance

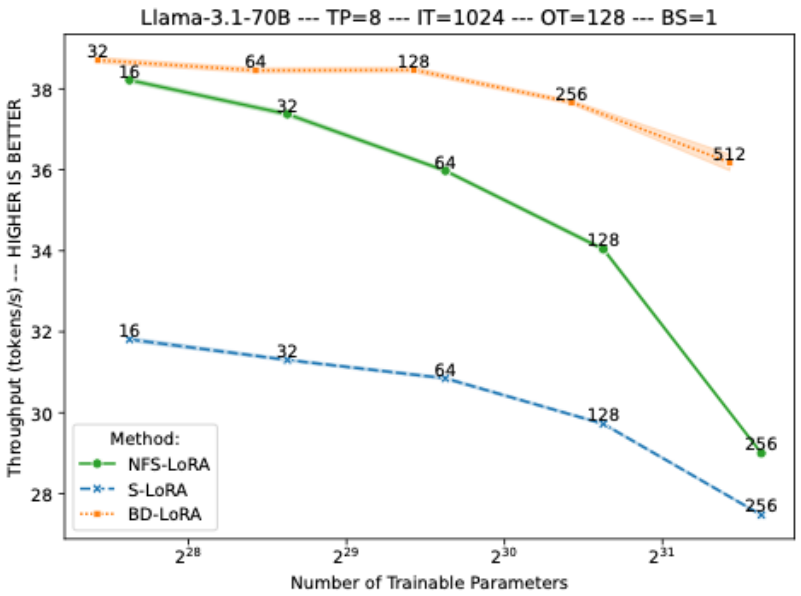
- For a similar number of trainable parameters, BD-LoRA shows comparable performance to LoRA.

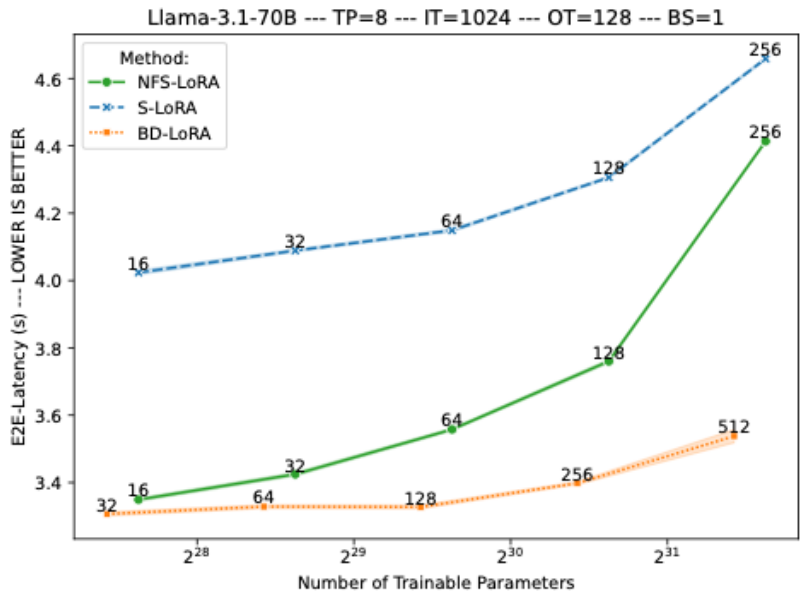


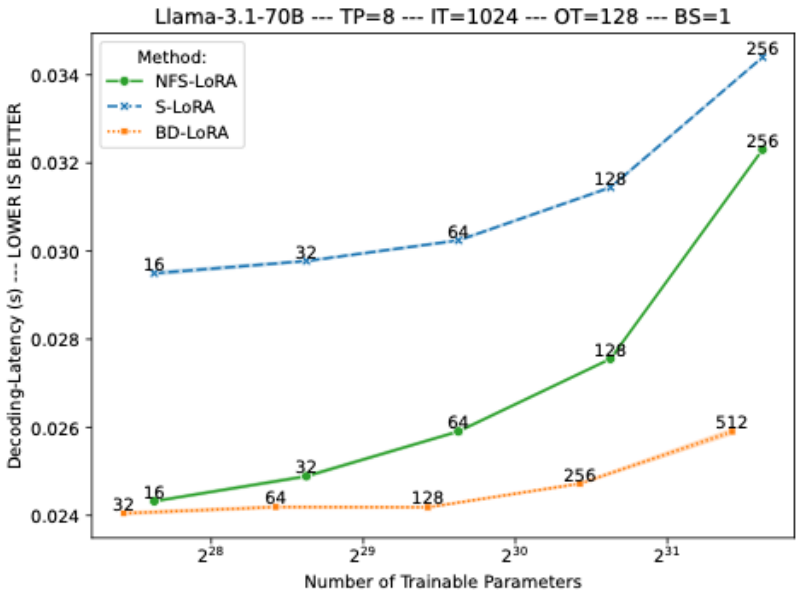


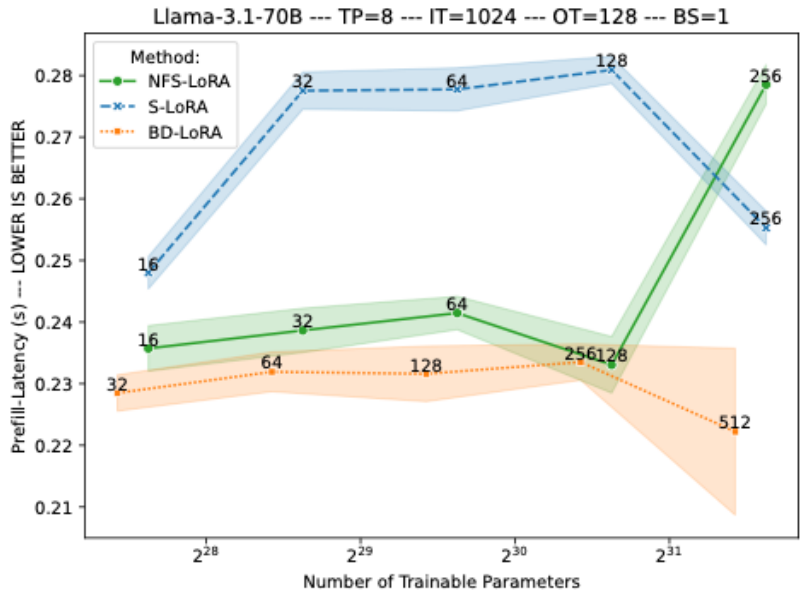
Throughput and Latency

- BD-LoRA yields 1.79x (1.23x) end-to-end speed-up with 0.87x (1.74x) number of adapter parameters on 70B model and 1.63x (1.3x) with 0.86x (1.73x) parameters on 8B model.



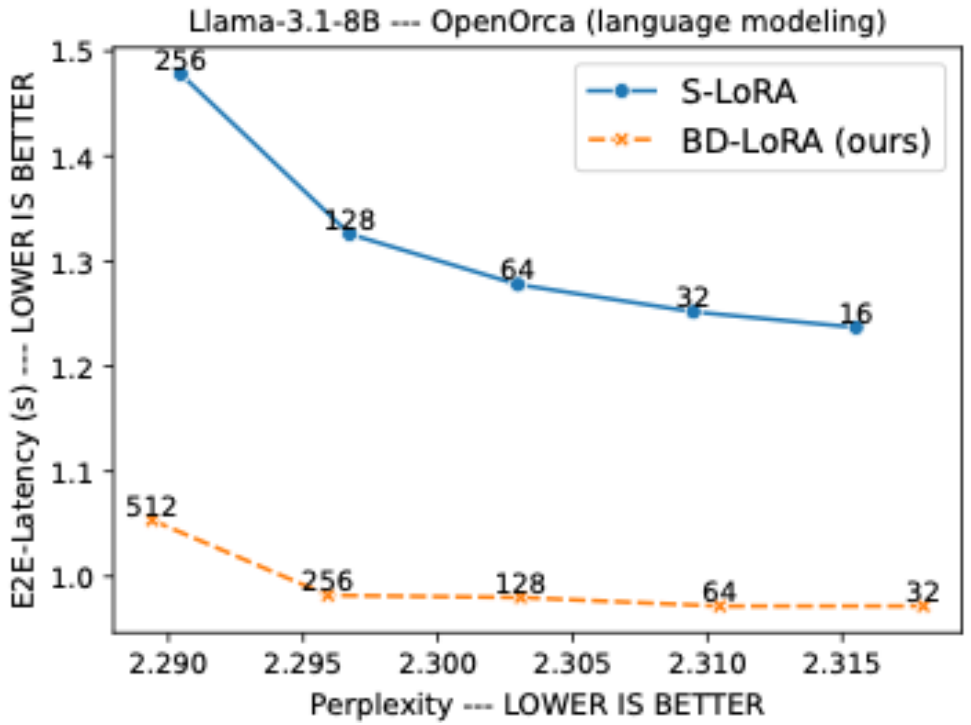


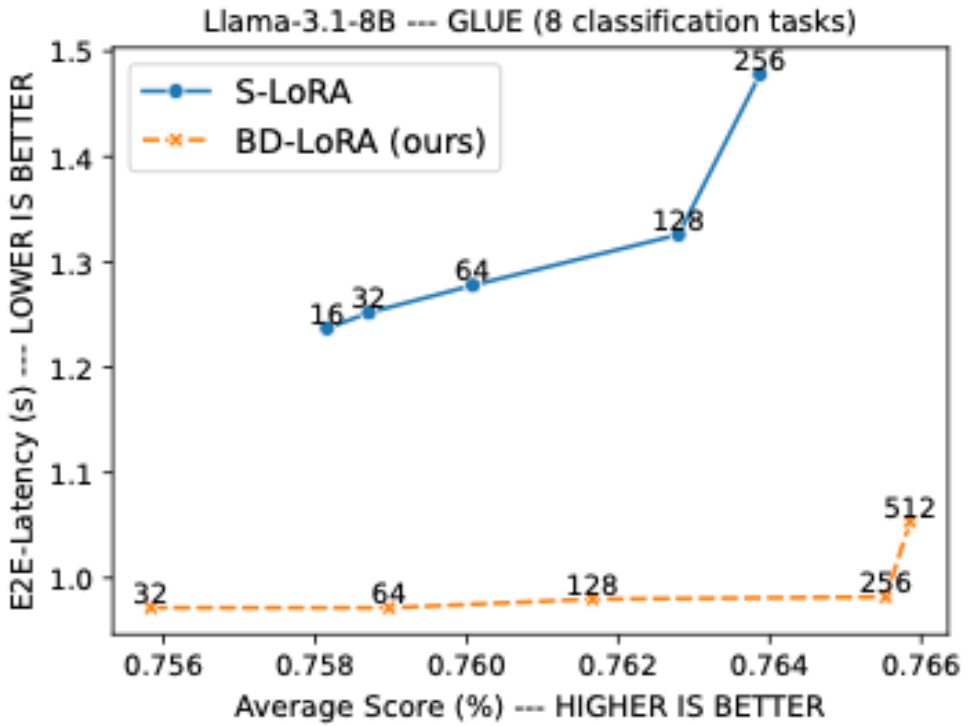




Downstream Performance vs. Run-time Trade-off

- At the same downstream performance, BD-LoRA achieves 1.27x – 1.51x end-to-end speedup.





Takeaway

- BD-LoRA eliminates the additional communication overhead in S-LoRA.
- With the same number of trainable parameters, BD-LoRA and LoRA achieve similar downstream performance.
- With the same number of trainable parameters, BD-LoRA is faster than S-LoRA.
- With the same downstream performance, BD-LoRA is faster than S-LoRA.

Thank You!