NUS National University of Singapore | School of Computing

StarTrail: Concentric Ring Parallelism for Efficient Near-Infinite-Context Transformer Model Training (Neurips ‘25)

NUS National University of Singapore | School of Computing

Berkeley UNIVERSITY OF CALIFORNIA

Why do we need Sequence Parallelism

- Code base process
- Long document understanding and summary
- Agent contexts
- …

- GPT-4o: 128K
- Claude 3.5 Sonnet: 200K
- Gemini 2.5 Pro: 2M
- …

Problems of current solution?

- Mainstream solution: RingAttention.
- Large communication volume.
- Bottleneck of inter-node communication.

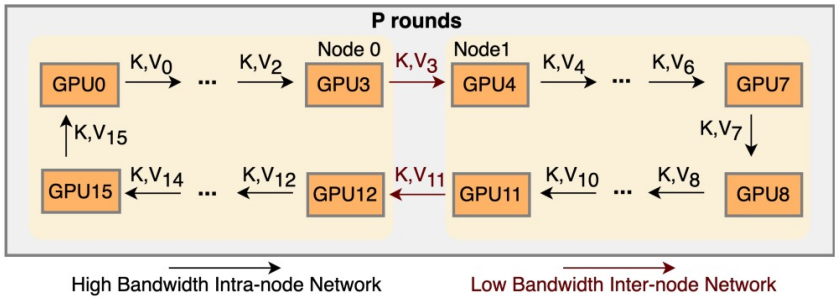
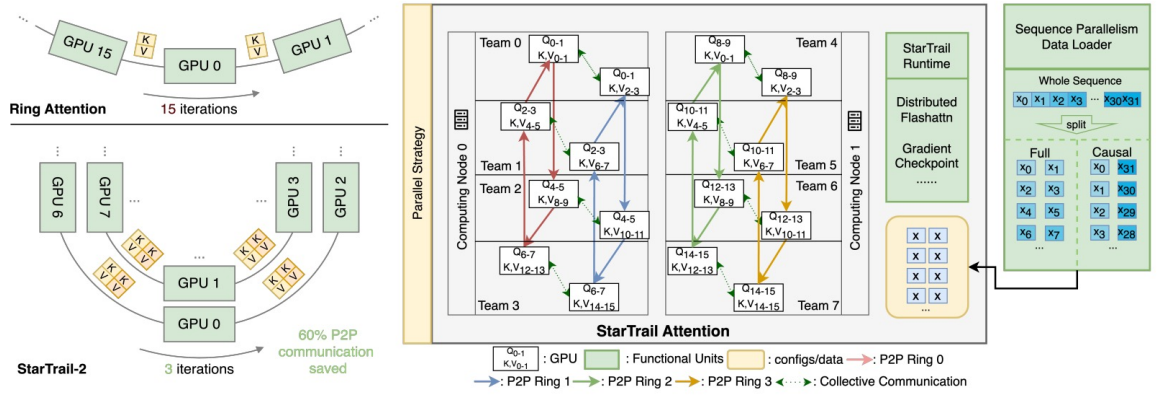


Figure 2: An example of Ring Attention Computation on 16 GPUs in two nodes. The Communication is largely limited by the inter-node bottleneck.

Our solution:

- Devide the computation and communication.
- C GPUs in a group, each contains the gathers sub-sequence in the group.
- Each GPU only communicate with $1/C^2$ GPUs from other groups.

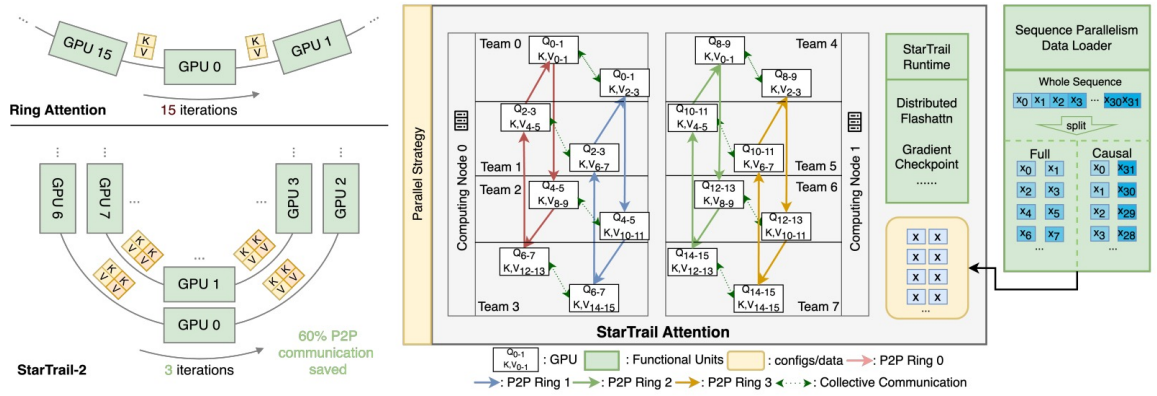


(a) StarTrail-2 reduces 60% P2P communication volume on 16 GPUs compared with ring attention.

(b) StarTrail divides one ring into four concentric sub-rings, with every two connected with collective communication.

Benefits

- P2P communication volume reduced by approximately C times.
- Inter-node communication can be avoided in certain scenarios.



(a) StarTrail-2 reduces 60% P2P communication volume on 16 GPUs compared with ring attention.

(b) StarTrail divides one ring into four concentric sub-rings, with every two connected with collective communication.

Experiments

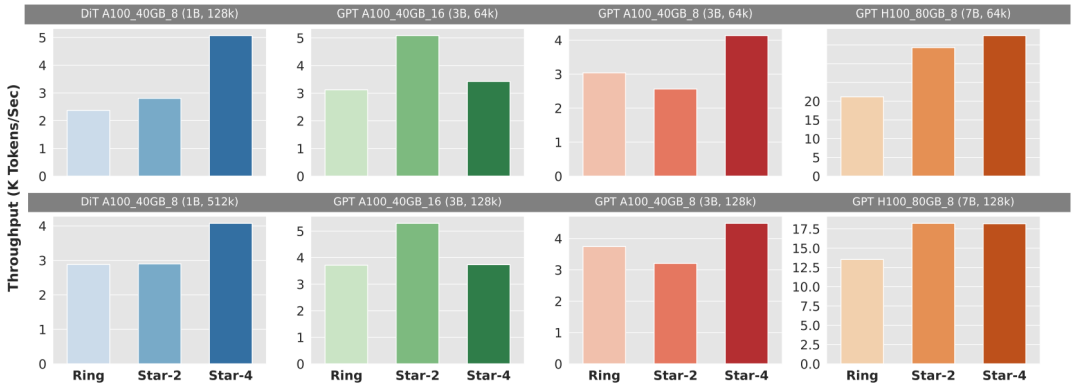


Figure 7: Throughput evaluation of Ring Attention and StarTrail on 32 GPUs from three different clusters. We place the performance of StarTrail with both C=2 and C=4 in the figure. The configurations are marked in the titles of the sub-figures. For instance, A100_40GB_8(1B, 512K) represents that the experiment is on machines with 8 Nvidia A100 40GB GPUs in each node, the model used has one billion parameters, and the sequence length is 512k.

Feel free to contact me!