EchelonFlow Case Studies and Proofs

This report provides supplementary proofs for the EchelonFlow workshop paper [1]. Please refer to the paper for definitions and notations.

1. ECHELONFLOW AND ITS TARDINESS

The optimization goal of EchelonFlow scheduling is tardiness minimization. For an individual EchelonFlow H, the goal is to minimize its tardiness t_H:

$$\textbf{Minimize:} \quad z = t_H \tag{S1}$$

For multiple EchelonFlows, the goal is to minimize the sum of their tardiness. Let H_i be the i_{th} EchelonFlow in a set of EchelonFlows $\mathcal{H}$ and t_{H_i} be the tardiness of H_i, the goal is:

$$\textbf{Minimize:} \quad \hat{z} = \sum_{0 \le i < |\mathcal{H}|} t_{H_i} \tag{S2}$$

2. CASE STUDIES

Case I: Coflow

Coflow is a special EchelonFlow whose flows share the same ideal finish time. For a Coflow in the form of an EchelonFlow H, the ideal finish time d_j of the j_{th} flow f_j (from $|H|$ flows in total) should follow the arrangement function below, where r is the reference time, i.e., the start time of the head flow.

$$d_j = r, \quad 0 \le j < |H| \tag{S3}$$

We prove that minimizing the tardiness of EchelonFlow H equals to minimizing the Coflow completion time.

Proof. The real finish time of the j_{th} flow f_j is e_j. z is H's tardiness.

$$\begin{aligned} Minimize \quad z &= max(e_j - d_j) \\ &= max(e_j - r) \\ &= max(e_j) - r \\ &=> z = max(e_j) \end{aligned} \tag{S4}$$

The Coflow completion time is decided by the last-finish flow. $\hat{z}$ is H's completion time.

$$Minimize \quad \hat{z} = max(e_j) \tag{S5}$$

Eqn. S4 and Eqn. S5 are the same.

Case II: PP - GPipe

In Fig. 1 we have three workers and two EchelonFlows, H_1 and H_2. T_i represents the computation time of a mini batch on the i_{th} worker.

r_1 is the head flow start time in H_1. For H_1, we have:

$$d_j = \begin{cases} r_1, & j = 0 \\ d_{j-1} + T_2, & 1 \leqslant j < |H_1| \end{cases} \tag{S6}$$

r_2 is the head flow start time in H_2. For H_2, we have:

$$d_j = \begin{cases} r_2, & j = 0 \\ d_{j-1} + T_3, & 1 \leqslant j < |H_2| \end{cases} \tag{S7}$$

We prove that minimizing the tardiness of EchelonFlow H_1 equals to minimizing the task completion time.

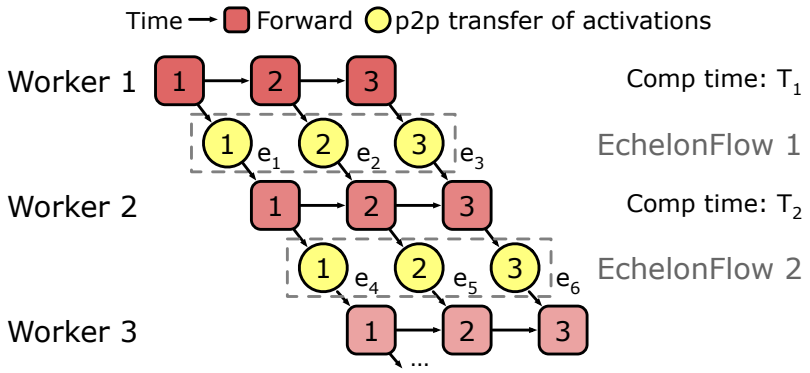

Fig. 1. EchelonFlow model for Pipeline

Proof. Take Fig. 1 as an example. For simplicity, we only prove the case with three mini batches. But the proof holds for any number of mini batches.

Let $\triangle t_j = e_j - d_j, 0 \leqslant j < |H_1|$. Minimizing H_1's tardiness z_1:

$$\textbf{Minimize:} \quad z_1 = max(\triangle t_1, \triangle t_2, \triangle t_3) \tag{S8}$$

The completion time $\hat{z}$ could be expressed as:

$$\begin{aligned} \textbf{Minimize:} \quad \hat{z} &= max(max(e_1 + T_2, e_2) + T_2, e_3) + T_2 \\ &= max(max(\triangle t_1 + d_1 + T_2, \triangle t_2 + d_2) + T_2, e_3) + T_2 \\ &= max(max(\triangle t_1 + d_2, \triangle t_2 + d_2) + T_2, e_3) + T_2 \\ &= max(max(\triangle t_1, \triangle t_2) + d_2 + T_2, e_3) + T_2 \\ &= max(max(\triangle t_1, \triangle t_2) + d_3, \triangle t_3 + d_3) + T_2 \\ &= max(max(\triangle t_1, \triangle t_2), \triangle t_3) + d_3 + T_2 \\ &= max(\triangle t_1, \triangle t_2, \triangle t_3) + d_3 + T_2 \\ &=> \hat{z} = max(\triangle t_1, \triangle t_2, \triangle t_3) \end{aligned} \tag{S9}$$

Eqn. S8 and Eqn. S9 are the same.

Case III: FSDP - ZeRO

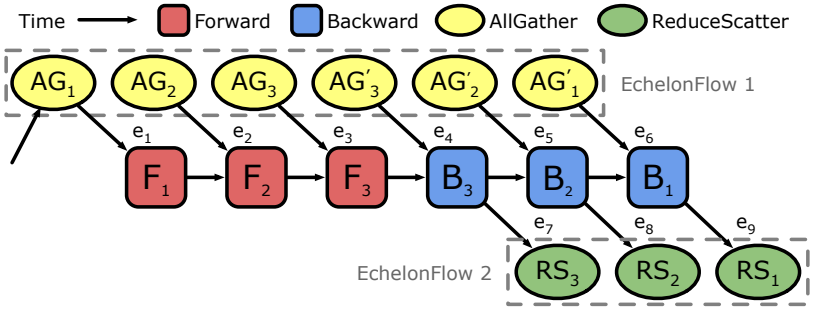

Fig. 2. EchelonFlow model for ZeRO

Fig. 2 shows the EchelonFlow model of ZeRO/FSDP. For ZeRO/FSDP, we need a two-level EchelonFlow model to represent it. Every AG or RS is a Coflow, which is also a special EchelonFlow. Furthermore, all AGs form a bigger EchelonFlow H_1, and all RSes form a bigger EchelonFlow H_2. We can use the same method in Case II to prove that minimizing the tardiness of EchelonFlow H1 and H2 equals to minimizing the task completion time.

REFERENCES

1. R. Pan, Y. Lei, J. Li, Z. Xie, B. Yuan, and Y. Xia, "Efficient flow scheduling in distributed deep learning training with echelon formation," in *The 21st ACM Workshop on Hot Topics in Networks (HotNets '22), November 14– 15, 2022, Austin, TX, USA,* (2022).