Bandwidth Optimization Tree Leveraging eBPF for In-Kernel Gradient Aggregation

Duc Trung Vu†, Xuan Tung Hoang†, Duc Hai Bui†, Kim Khoa Nguyen‡
†VNU University of Engineering and Technology, Hanoi, Vietnam
‡École de Technologie Supérieure, Montreal, Canada
Email: {vdtrung, tunghx, 21020191}@vnu.edu.vn, kimkhoa.nguyen@etsmtl.ca

***Abstract*—Today's distributed machine training scenarios that rely on a single parameter server face an issue of bandwidth bottleneck on the link to this server, resulting in resource inefficiency and low learning rate. To achieve optimal performance, a more efficient communication structure is required. Unfortunately, prior work in the literature focuses only on a two-level tree topology and homogeneous links among the nodes, which is not scalable for AI-centric data centers. In this paper, we propose a framework to build a multi-level communication tree for training aggregation, named eBOT. Supported by the extended Berkeley Packet Filter (eBPF) technology, eBOT accelerates packet processing at each node by bypassing TCP/IP network stack in the kernel of Linux operating system. Experimental results demonstrate that our proposed solution outperforms state-of-the-art distributed training schemes in both homogeneous and heterogeneous bandwidth scenarios.**

***Index Terms*—Distributed Training, eBPF, Data Center, In-Kernel Hierarchical Aggregation**

References

[1] J. M. Steele, *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Appendix

A. Proof of Theorem 1

Proposition 1. *The delay caused when gradient fragments are transmitted through multiple layers during hierarchical gradient aggregation is negligible, allowing transmission without delay.*

Proof.
Let:

- F be the number of gradient fragments.
- L be the number of tree layers.
- T^{frag} be the transmission time of a single gradient fragment.

The link propagation time during aggregation is:

$$T^{links} = F \cdot T^{frag}.$$

The hierarchical delay time across layers is:

$$T^{delay} = (L-2) \cdot T^{frag}.$$

The total transmission time during aggregation is:

$$T^{trans} = T^{links} + T^{delay}.$$

$$T^{trans} = T^{links} + \frac{(L-2)}{F} \cdot T^{links}$$

Given that the number of fragments F is substantially larger than the number of layers L (i.e., $F \gg L$), the delay time T^{delay} converges to 0. Consequently, the total transmission time T^{trans} converges to the transmission time of the links.

$$T^{trans} \approx T^{links}$$

□

Proof of Theorem 1. *The aggregation time T is lower bound by:*

$$T \geq \frac{G}{\min_{i=1}^{N} b_i}$$

where G is the gradient size, and b_i is the normalized bandwidth defined as:

$$b_i = \frac{B_i}{\deg(i)} = \frac{B_i}{\sum_{j=1}^{N} x_{ij}}$$

Proof. First, we consider constraint (a),

$$\textbf{s.t.} \quad r_{ij} \leq \min(B_i, B_j), \quad \forall (i,j) \tag{a}$$

We have:

$$T = \max_{i,j=1}^{N} \frac{G x_{ij}}{r_{ij}}$$

$$T \geq \max_{i,j=1}^{N} \frac{G x_{ij}}{\min(B_i, B_j)}$$

$$T \geq \frac{\max_{i,j=1}^{N} G x_{ij}}{\min_{i,j=1}^{N} \min(B_i, B_j)}$$

$$T \geq \frac{G}{\min_{i=1}^{N} B_i} \tag{1}$$

Next we consider constraint (b),

$$\textbf{s.t.} \quad \sum_{j=1}^{N} r_{ij} \leq B_i, \quad \forall i \in \{1, \ldots, N\} \tag{b}$$

We denote t_{ij} as the transmission time from node i to node j. We have:

$$t_{ij} = \begin{cases} \frac{G x_{ij}}{r_{ij}}, & \text{if } r_{ij} > 0, \\ 0, & \text{if } r_{ij} = 0. \end{cases}$$

$$\sum_{j=1}^{N} t_{ij} = G \cdot \sum_{j=1}^{N} \frac{x_{ij}}{r_{ij}}, \quad \forall i \in \{1, \ldots, N\}$$

$$\sum_{j=1}^{N} t_{ij} \cdot x_{ij} = G \cdot \sum_{j=1}^{N} \frac{(x_{ij})^2}{r_{ij}}, \quad \forall i \in \{1, \dots, N\}$$

We denote T_i^{trans} as the transmission time from node i to the root node. We have $T_i^{trans} \geq \max_{j=1}^{N} t_{ij}$, so it follows that

$$T_i^{trans} \cdot \sum_{j=1}^{N} x_{ij} \geq G \cdot \sum_{j=1}^{N} \frac{(x_{ij})^2}{r_{ij}}, \quad \forall i \in \{1, \dots, N\}$$

By Cauchy-Schwarz inequality [1], we obtain

$$\sum_{j=1}^{N} \frac{(x_{ij})^2}{r_{ij}} \geq \frac{(\sum_{j=1}^{N} x_{ij})^2}{\sum_{j=1}^{N} r_{ij}}, \quad \forall i \in \{1, \dots, N\}$$

Therefore,

$$T_i^{trans} \cdot \sum_{j=1}^{N} x_{ij} \geq G \cdot \frac{(\sum_{j=1}^{N} x_{ij})^2}{\sum_{j=1}^{N} r_{ij}}, \quad \forall i \in \{1, \dots, N\}$$

$$T_i^{trans} \geq G \cdot \frac{\sum_{j=1}^{N} x_{ij}}{\sum_{j=1}^{N} r_{ij}}, \quad \forall i \in \{1, \dots, N\}$$

$$T_i^{trans} \geq G \cdot \frac{\sum_{j=1}^{N} x_{ij}}{B_i}, \quad \forall i \in \{1, \dots, N\}$$

$$T_i^{trans} \geq \frac{G}{\frac{B_i}{\sum_{j=1}^{N} x_{ij}}}, \quad \forall i \in \{1, \dots, N\}$$

Since the aggregation time T is determined by the slowest transmission, we have $T = \max_{i=1}^{N} T_i^{trans}$. Therefore,

$$T \geq \max_{i=1}^{N} \frac{G}{\frac{B_i}{\sum_{j=1}^{N} x_{ij}}}$$

$$T \geq \frac{G}{\min_{i=1}^{N} \frac{B_i}{\sum_{j=1}^{N} x_{ij}}} \tag{2}$$

From (1) and (2), we derive

$$T \geq \frac{G}{\min_{i=1}^{N} \min(\frac{B_i}{\sum_{j=1}^{N} x_{ij}}, B_i)}$$

$$T \geq \frac{G}{\min_{i=1}^{N} \frac{B_i}{\sum_{j=1}^{N} x_{ij}}}$$

$$T \geq \frac{G}{\min_{i=1}^{N} b_i}$$

Equality holds when:

$$r^* = r_{ij}, \quad \forall (i, j) \in \{1, \dots, N\}$$

where r^* represents the optimal sending rate for the network. We derive $r^* = \min_{i=1}^{N} b_i$ □

B. Proof of Algorithm

Theorem 1. *Given a set $B = \{B_i \mid i = 1, \dots, N\}$ sorted in descending order ($B_1 > B_2 > \cdots > B_N$), the Max-Min Normalized Bandwidth Tree algorithm outputs (T, E) which is a max-min normalized bandwidth tree.*

Proof. We prove the theorem by induction.

Inductive Base ($|T| = 2$): $T = \{B_1, B_2\}$ and $E = \{(1, 2)\}$. The minimum normalized bandwidth is given by:

$$b_{\text{min}} = B_2 \geq B_i, \quad \forall i \geq 2,$$

Since b_{min} is the largest normalized bandwidth of all possible tree structures, the base case holds.

Inductive Hypothesis: Assume that for $|T| = k$, the minimum normalized bandwidth of k nodes holds:

$$b_{\text{min}}^k = \min\left(\frac{B_1}{\deg^k(1)}, \frac{B_2}{\deg^k(2)}, \dots, \frac{B_k}{\deg^k(k)}\right)$$

Inductive Step ($|T| = k + 1$): We add a new node $(k + 1)$ to the tree T and form a new edge $(m, k + 1)$ to E, where m is the node has the maximum next-step normalized bandwidth at step k. The new minimum normalized bandwidth b_{min}^{k+1} is calculated as:

$$b_{\text{min}}^{k+1} = \min\left(b_m^{k+1}, b_{k+1}^{k+1}, b_{\text{min}}^k\right),$$

$$b_{\text{min}}^{k+1} = \min\left(\frac{B_m}{\deg^{k+1}(m)}, B_{k+1}, b_{\text{min}}^k\right),$$

where $\frac{B_m}{\deg^{k+1}(m)} = \frac{B_m}{\deg^k(m)+1}$ is the maximum normalized bandwidth at step $(k + 1)$ and also the maximum next-step normalized bandwidth calculated from step k.

Case Analysis:

1) **If $b_{\text{min}}^{k+1} = \frac{B_m}{\deg^k(m)}$:** This is correct because $\frac{B_m}{\deg^{k+1}(m)}$ is the maximum normalized bandwidth at step $(k + 1)$.
2) **If $b_{\text{min}}^{k+1} = B_{k+1}$:** This is correct because $B_{k+1} \geq \frac{B_{k+1}}{\deg(k+1)}$ is the largest normalized bandwidth of all possible tree structures.
3) **If $b_{\text{min}}^{k+1} = b_{\text{min}}^k$:** Since b_{min}^k was correct by the inductive hypothesis, it remains valid when the node $(k + 1)$ is added.

Thus, in all cases, b_{min}^{k+1} is correct. By the principle of induction, the theorem holds for all $|T|$. □