Collective Communication Optimization (CCO): Use cases, Problems, and Requirements

Personal I-Ds:

Kehan Yao, China Mobile
Shiping Xu, China Mobile
Yizhou Li, Huawei
Hongyi Huang, Huawei
Dirk Kutscher, HKUST(GZ)

IETF 118 hotRFC
**Concept:**

Collective communication is a communication model which plays a key role in high performance computing and modern distributed AI model training workloads such as recommender systems and natural language processing. It involves a group or groups of processes participating in collective operations like AllReduce or AllGather. The communication model can be one-to-all, all-to-one or all-to-all and is usually realized by a sequence of unicast messages.

**Use cases:**

- **Distributed AI Model Training**
- **Spark Shuffle in Big Data Analysis**
- **Distributed Storage**
Major Problems & Observation:

• P2P implementation of Collective Communication incurs much overhead, reflected in:
  • large bandwidth occupancy(duplications & redundancy)
  • much data movement(end-to-end transmission)
  • large number of data copies at endpoints(sending one pkt needs to copy at least one time).

Communication bottleneck & performance degradation

• It should
  • save bandwidth(This is extremely important for BW-sensitive Apps like distributed AI model training workloads, since BW is the new oil).

    "The metaphor is not from me, but I think it is quite impressive. 😊"

  • reduce data movement.
  • decrease data copies.

• Offloading collective operations to the network is important for achieving benefits above and very necessary, especially for these performance-driven Apps.
Transport Issues:
- Reliability: underlying network lacks collective communication reliability
- Semantic Gap: message passing vs packet delivering
- Blocking & Non-blocking: different optimizations for different communication modes

One-to-Group Transmission:
- IP Multicast for Message Bcast/AlltoAll/...
IP multicast is the most direct way, perhaps there is a better way

Data & Control & Management:
- In-network Primitives: collective operations based on unified in-network primitives
- Topology Awareness: to improve existing topology aware algorithms to support in-network computing

Design Issues[1],[2]:
Related Side Meeting in IETF118:

- https://wiki.ietf.org/meeting/118/sidemeetings

**Title**: Collective Communication Optimization (CCO),

**Time Schedule**: 9th, Nov, Thursday, 14:30 -- 16:00, Palmovka ½

**Agenda**: https://github.com/CCO-IETF/ietf118-side-meeting

Looking for collaborators to seek for potential standardization opportunity of the work in IETF, and welcome for more discussions and contributions.

Thanks!