

Multicast for Computing and Storage

Yisong Liu (China Mobile)(Presenter)

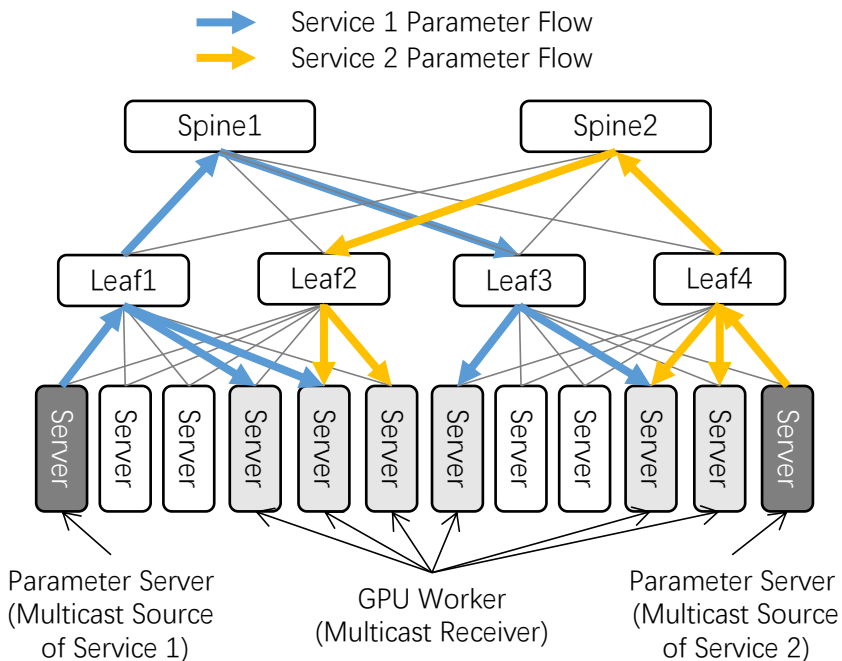
Xuesong Geng (Huawei)

Use Case: Multicast Request in large-scale DC Network

- Large-Scale DC Network could have a significant amount of potential multicast services running, including **AI training**, **HPC (High Performance Computing)** and **SAN (Storage Area Network)** scenarios in DCN.

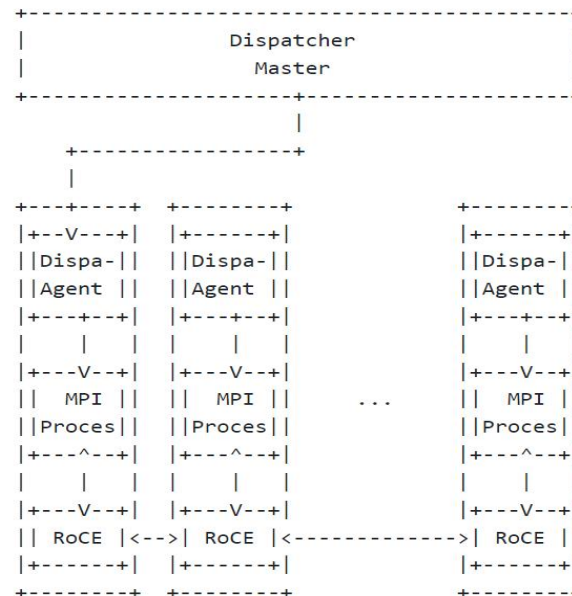
AI Training

- Training model distribution and parameter synchronization



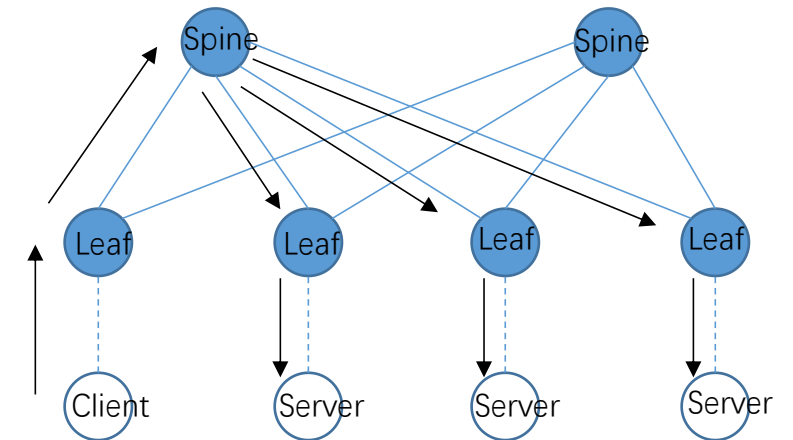
HPC

- Dispatcher master starts millions of rank MPIs and broadcasts messages to a scalable number of dispatcher agents



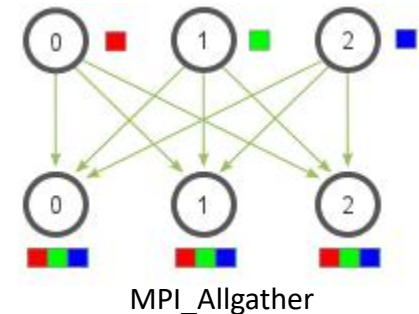
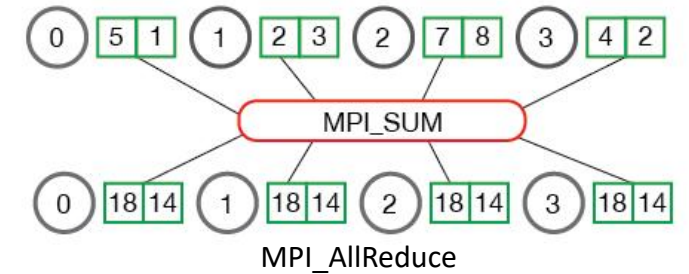
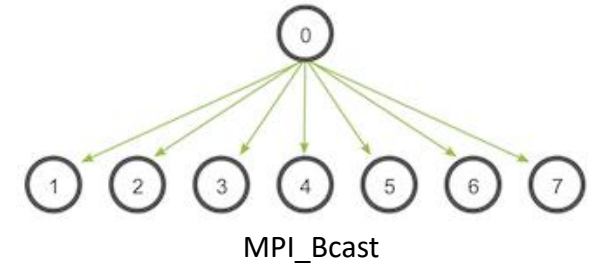
Storage

- “Multiple Copies” for reliability



What is MPI and how it Connect with Multicast?

- **The Message Passing Interface (MPI)** is an Application Program Interface that defines a model of parallel computing where each parallel process has its own local memory, and data must be explicitly shared by passing messages between processes. Using MPI allows programs to scale beyond the processors and shared memory of a single compute server, to the *distributed memory* and processors of multiple compute servers combined together.
- An MPI parallel code requires some changes from serial code, as **MPI function calls to communicate data are added**, and the data must somehow be divided across processes.
- There are three categories of functions that fall under collective communication operations. Some of them are related to P2MP or MP2MP , for example :
 - MPI_Bcast: Broadcast **sends a message from the process with rank “root” to all other processes in the group.**
 - MPI_AllReduce: Combines values from all processes and distributes the result back to all processes;
 - MPI_Allgather: An MPI_Allgather call gather the same data from all ranks and provides it to all ranks. It is logically identical to MPI_Gather to a root followed by an MPI_Bcast from that root, but is implemented more efficiently.
- MPI(or similarly interface implementation) is widely used in parallel computing for **HPC** and **AI Training**



Open Questions for Discussion

- Does network layer has a role in HPC/AI training scenario?
- What is the benefit of doing multicast in network layer rather than in application layer?
- Does it request new multicast protocol, for example new tree setting up signalling?
 - The traditional multicast tree set up process is suitable for multicast service like IPTV, which is triggered by the receiver; while a source-triggered signalling is more suitable for computing and storage;
- Does it request new multicast forwarding plane?
 - Flexible and scalable will be more
- Does it request new multicast transport plane?
 - There were reliable multicast transport work in IETF, like NORM. Does it enough?

Thanks