CS 168 Spring 2025

Multicast and Collectives

Discussion 12

1 True or False

1.1 IP multicast is widely deployed across different domains for global Internet communication.

1.2 In the DVMRP protocol, the multicast forwarding table requires one entry per source, per multicast group.

1.3 Core-Based Trees (CBT) guarantee that packets are forwarded along the least-cost paths to all group members.

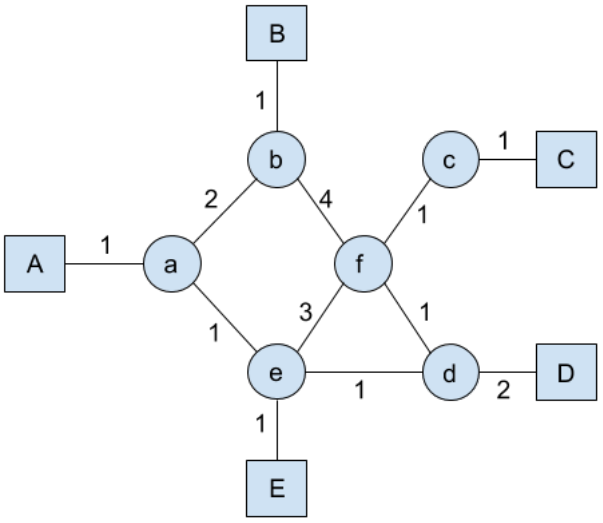
1.4 Overlay multicast requires routers to implement specialized multicast protocols, unlike IP multicast.

1.5 The AllReduce collective operation can be implemented efficiently using a ring topology, where each node sends and receives data to/from its neighbors.

1.6 The stretch factor in overlay multicast measures the ratio of the overlay path cost to the underlay path cost, with lower stretch indicating better performance.

1.7 Collective operations in AI training, such as Broadcast and Reduce, are unrelated and cannot be combined to form other operations like AllReduce.

2 Multicast



Consider the network topology above for both parts of this question. Hosts are squares with capital letters, routers are circles with lower case letters. Link cost is latency in seconds.

DVMRP

All parts of this section are cumulative. At the start, A and E are members of group 1.

2.1 (a) A sends a multicast packet to Group 1. After all Non-Membership Reports (NMRs) have propagated, which links remain active (i.e., used for forwarding packets) in the source-specific multicast tree for (Source A, Group 1)?

(b) Suppose C adds itself to group 1. Immediately after this, who knows C joined group 1?

(c) E sends a message to group 1. Which links remain active for (Source E, Group 1)?

(d) B sends a message to group 1, is this possible?

(e) What is the route that the packet takes from B to E?

(f) How long does it take for B's message to reach all members of group 1?

(g) Suppose that the pruning information expires. A, B, and D form a new group 2. A then sends messages to groups 1 and 2, B sends a message to group 2, and D sends a message to group 1. What state does router f have? (Give a list of (Source x, Group y)).

(h) What message(s) took the longest to send?

(i) What message(s) took the least time to send?

CBT

All parts of this section are cumulative, but independent from the previous section. a is the core for group 1.

2.2 (a) C decides to join group 1. What routers does its request go through?

(b) D decides to join group 1. What routers does its request go through?

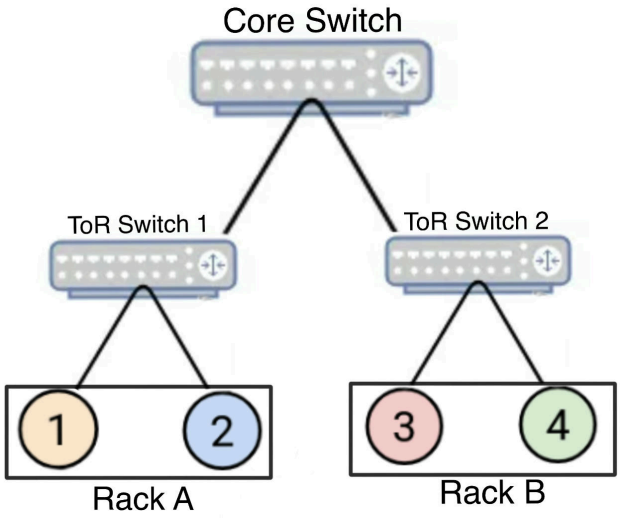
(c) B sends a message to group 1. How long does it take to reach all members?

(d) C sends a message to group 1. How long does it take to reach all members?

(e) The link between A and a goes down, but no other changes in control plane state occur. Who can still send/receive messages to/from group 1, if any?

3 Collective

You are designing a distributed AI training system in a datacenter with 4 GPUs (nodes 1, 2, 3, 4) performing collective operations. Each node holds a 4-element vector of data (total size D bytes). The datacenter uses a Clos topology, with nodes 1 and 2 in Rack A, and nodes 3 and 4 in Rack B, connected via a core switch. Intra-rack links (e.g., 1↔2) have 1 hop (stretch = 1); inter-rack links (e.g., 1↔3) have 3 hops (stretch = 3). Assume intra-rack links have unlimited bandwidth, inter-rack links have a bandwidth capacity of B bytes per time step, and nodes can send/receive on all links simultaneously unless otherwise specified.



The Clos topology has nodes 1 and 2 in Rack A, nodes 3 and 4 in Rack B, connected via a core switch. Intra-rack links have 1 hop (stretch = 1); inter-rack links have 3 hops (stretch = 3).

Part A: Full-Mesh AllReduce Analysis

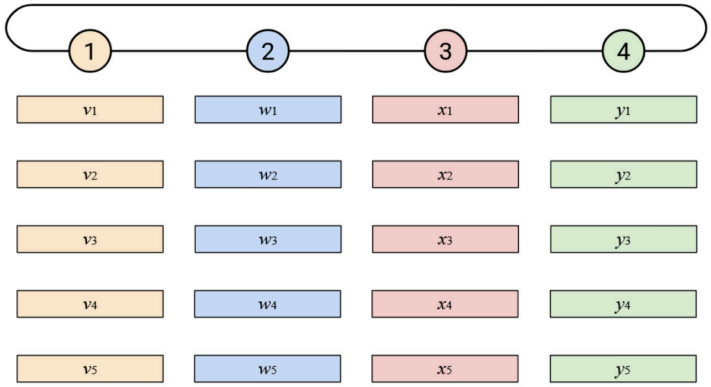
Consider an AllReduce operation where each node computes the element-wise sum of all vectors and receives the sum vector.

3.1 (a) Explain why a full-mesh topology is suitable for AllReduce in AI training, and discuss its scalability limitations for large p.

(b) Calculate the total bandwidth and number of steps for a full-mesh AllReduce with $p = 4$. If inter-rack link bandwidth is limited to $B = D$ bytes per time step and intra-rack links have unlimited bandwidth, how many time steps are required?

Part B: Naive Ring AllReduce

Assume a naive ring-based AllReduce (e.g., Node 1 ⟶ Node 2 ⟶ Node 3 ⟶ Node 4 ⟶ Node 1).



The ring topology connects nodes 1⟶2⟶3⟶4⟶1 in a closed loop.

3.2 (a) How many steps are required for a naive ring AllReduce with $p = 4$? Describe the data exchange process.

(b) Calculate the total bandwidth and bandwidth per node per step. If inter-rack links are limited to $B = \frac{D}{2}$ bytes per step, how does this affect the number of steps?

Part C: Optimized Ring vs. Tree-Based AllReduce

Compare naive ring, optimized ring, and tree-based AllReduce implementations for $p = 16$.

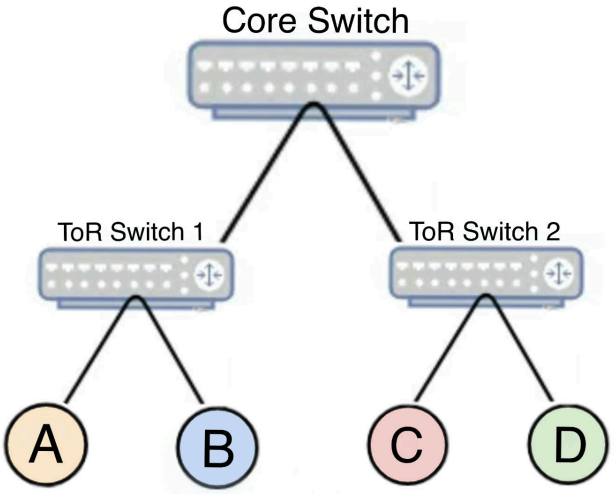
3.3 (a) For an optimized ring AllReduce, describe how it differs from the naive ring and calculate its total bandwidth and steps for $p = 16$.

(b) For a tree-based AllReduce, calculate the total bandwidth and steps for $p = 16$, assuming a balanced binary tree.

(c) Compare the three implementations (naive ring, optimized ring, tree-based) for $p = 16$, discussing trade-offs for AI training.

Part D: Overlay Topology Optimization

Optimize an overlay ring for AllReduce over the Clos network, with nodes A, B in Rack A, and C, D in Rack B.



Nodes A and B are in Rack A, nodes C and D are in Rack B, with intra-rack links (stretch = 1) and inter-rack links (stretch = 3).

3.4 (a) Propose an overlay ring topology by numbering nodes (A, B, C, D) as Nodes 1, 2, 3, 4 to minimize the average stretch factor. Calculate the stretch for each link and the average.

(b) Consider two alternative topologies: (i) A→C→D→B→A, and (ii) A→C→B→D→A. Calculate their average stretches and compare to your proposed topology. If inter-rack links have bandwidth $B = \frac{D}{4}$ bytes per step, which topology minimizes total latency for an optimized ring AllReduce, assuming each hop adds 1 time unit of latency?

Part E: ReduceScatter Implementation

Consider a ReduceScatter operation, where each node receives one summed element of the AllReduce output.

3.5 (a) Explain how ReduceScatter differs from AllReduce and how it can be implemented using a naive ring for $p = 4$. Calculate the total bandwidth and steps.

(b) If you implement ReduceScatter using the optimized ring topology from Part D.1 (A→B→C→D→A) with inter-rack bandwidth $B = \frac{D}{4}$, calculate the total latency (in steps) and compare to the naive ring.