Acceleration of All-to-all Communication on Multi-Layer Full Mesh, Low-cost Connectable Network Topology

Toshihiro Shimizu, Masahiro Miwa, Kohta Nakashima, FUJITSU LABORATORIES LTD.

Abstract

- Propose a cost-effective topology, Multi-layer full mesh topology, keeping high performance of communication

1. Introduction

- Fat Tree (FT)
  - Conventionally used topology for Super Computer cluster
  - (PROS) high communication throughput
    - All-to-all communication w/o congestion

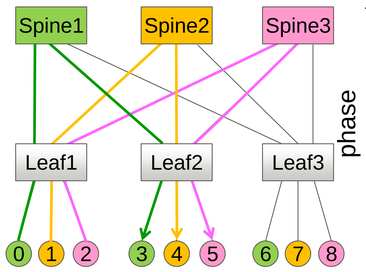


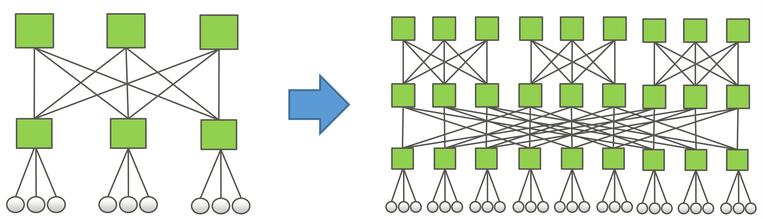
Fig.1 An example of Fat tree and its communication

source server

phase	0	1	2	3	4	5	6	7	8
0	0	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8	0
2	2	3	4	5	6	7	8	0	1
3	3	4	5	6	7	8	0	1	2
4	4	5	6	7	8	0	1	2	3
5	5	6	7	8	0	1	2	3	4
6	6	7	8	0	1	2	3	4	5
7	7	8	0	1	2	3	4	5	6
8	8	0	1	2	3	4	5	6	7

Fig.2 Shift communication pattern (effective All-to-all algorithm for Fat-tree)

  - (CONS) Needs too many switches
    - 3-level fat tree for larger scales lowers cost efficiency(cost of switch is high)



# of servers : 9
# of switches: 6
rate: 1.5 (per switch)

# of servers: 27
# of switches: 27
rate: 1.0 (per switch)

Fig.3 extending Fat-Tree for large scale

2. Proposal

- Multi-Layer full mesh (MLFM)
  - Low-cost connectable
    - only half number of switches compared to FT.
  - Keeps throughput of all-to-all Communication
    - All and partial servers
- Selecting partial servers
  - $n\ell m$ servers (n-layers, ℓ-leafs, m-servers)

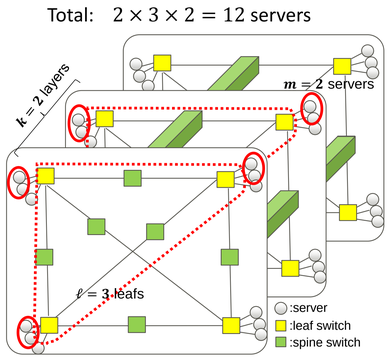


Fig.4 MLFM topology

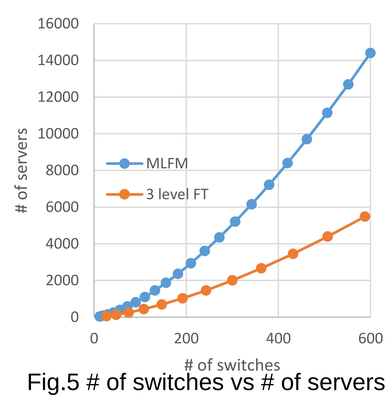


Fig.5 # of switches vs # of servers

- All-to-all communication
  - sending layer-to-layer(shift pattern)
  - execute all possible patterns;
    - Sending to the server of different column(Fig. 6a)
    - Sending to the server of same column(Fig. 6b)

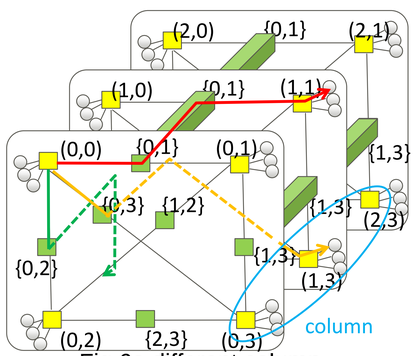


Fig.6a different column

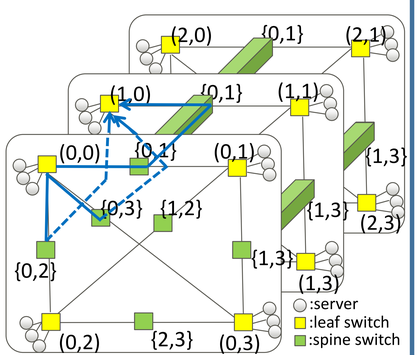


Fig.6b same column

3. Evaluation

- Measured performance of our communication algorithm
- Result:
  - No performance degradation
  - 2.2 times of throughput compared to shift communication pattern

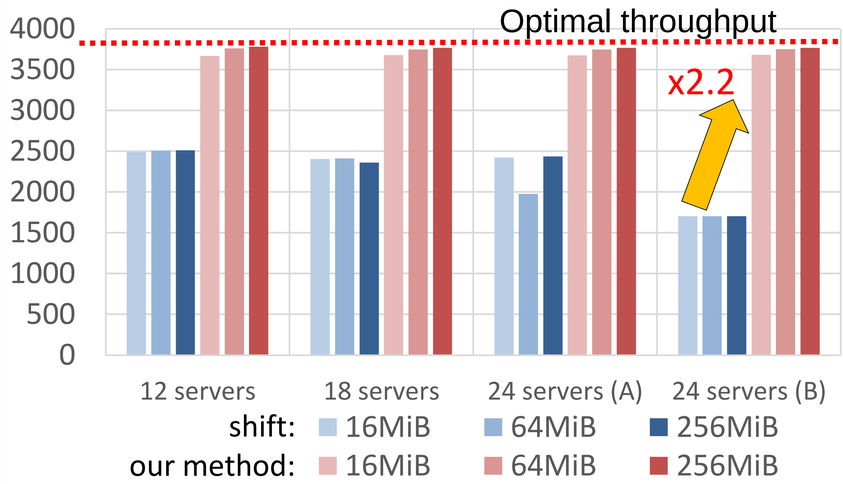


Fig.7 result of the experiment

Setup:
Topology is same as Fig. 4 (32 servers) except that the each leaf switch of the last layer has 2 servers
# of servers and (n, ℓ, m) : 12(2,3,2), 18(3,3,2), 24 pattern A(3,4,2), 24 pattern B(2,4,3)
Experiments are conducted by Infiniband(QDR, FDR). Software setup includes Cent OS 6.4(host servers), 7.2(VMs) and Intel MPI Benchmark(IMB) with OpenMPI 1.10.0. The message sizes are 16MiB, 64MiB, 256MiB

4. Conclusion & Future Work

- MLFM provides cost-efficiency and high communication performance
  - 2.2 times higher throughput
- Future work
  - fault tolerance
    - MLFM has few paths between servers