How Much Should We Invest for Network Facility:

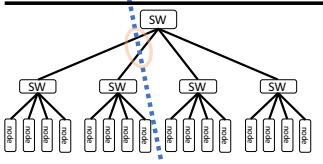
Quantitative Analysis on Network '*Fatness*' and Machine Learning Performance

Duo ZHANG, Mingxi LI (Univ. Tsukuba), Yusuke Tainimura, Hidemoto Nakada (AI Research Center, AIST)

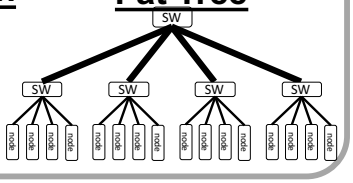
Overview

- In HPC area, **bi-section bandwidth** is considered to be important, since some application is quite sensitive to it.
- **QUESTION:**
  Do we also need to pay extra money for bi-section bandwidth for distributed machine learning?
- **Contribution:**
  Quantitative simulation using SimGrid [1]

Bi-Section Bandwidth



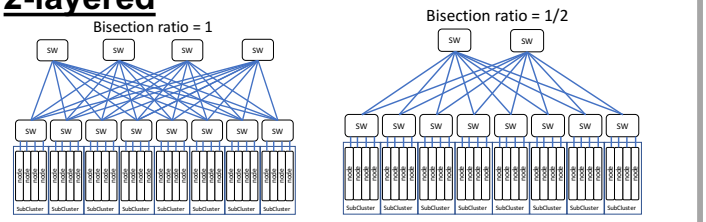
Fat Tree



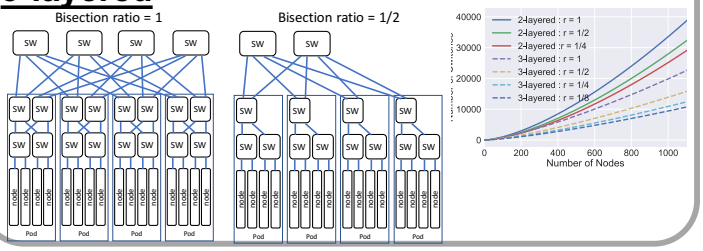
Network Structure

Clos Network[2]: implementations of Fat Tree

2-layered

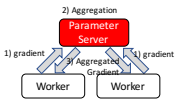


3-layered

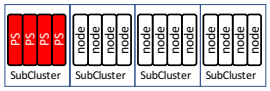


Gradient Exchange Methods

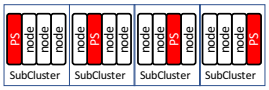
Via Parameter Server



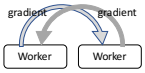
Packed Placement



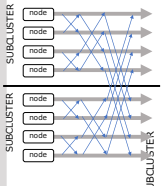
Distributed Placement



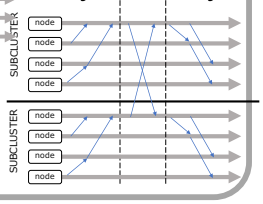
Direct Exchange [3]



Flat Butterfly



Layered Butterfly

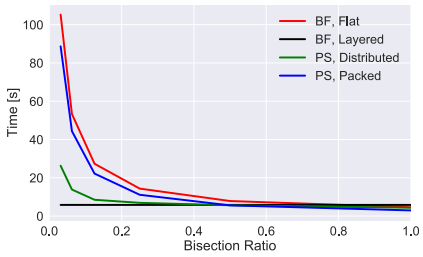


Experiments and Results

Setup

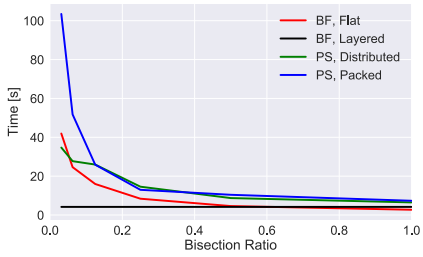
Measure time to exchange gradients once.

- Size of gradients: 1GBytes = 32bit x 256M
- 2 Layered: 128, 512, 2048 nodes
- 3 Layered: 256, 1204 nodes
- Bisection Ratio: 1/16, 1/8, 1/4, ½, 1
- Network Bandwidth: 1GBytes/s , 4GBytes/s
- Switch latency 0.2μs, 1μs
- Parameter Server: 1/8 of nodes are used for PS



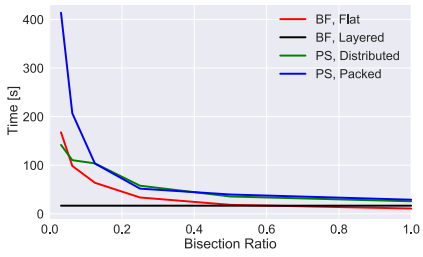
2-Layered: 2048 nodes
4GBytes/s, 0.2μs

- Bisection Ratio affects gradient exchange time
- BF, Layered is not affected



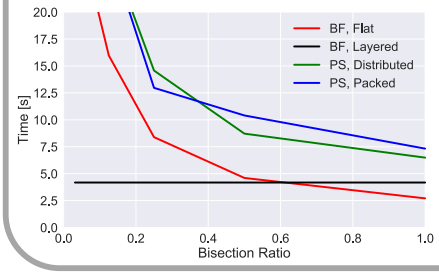
3-Layered: 1024nodes
4GBytes/s, 0.2μs

- Network configuration (2 /3 Layered) does not affect much for the performance.



3-Layered: 1024nodes
1GBytes/s, 0.2μs

- Link speed linearly affects the exchange speed.



3-Layered: 1024nodes
4GBytes/s, 0.2μs
Enlarged

- Flat Butterfly is faster than Layered Butterfly when the bi-section ratio is one.

Conclusion

- If you can afford 'full-bisection' (ratio = 1.0), Flat Butterfly is the best.
- If you cannot, Layered Butterfly provides flat performance. ½ and 1/16 are the same.
- Avoid using Parameter Server.
- Link speed is crucial. it make sense to pay more for faster link, such as 100GB infiniband.

[1] Henri Casanova, Arnaud Giersch, Arnaud Legrand, Martin Quinson, and Frédéric Suter. Versatile, scalable, and accurate simulation of distributed applications and platforms. *Journal of Parallel and Distributed Computing*, 74(10):2899–2917, June 2014.
[2] Charles Clos. A study of non-blocking switching networks. *Bell System Technical Journal*, 32(2):406—424, 1953.
[3] Rajeev Thakur and W. D. Gropp. Improving the performance of mpi collective communi- cation on switched networks. Technical Report ANL/MCS-P1007-1102, Argonne National Laboratory, 11/2002 2002.

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by JSPS KAKENHI Grant Number JP16K00116.