HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

Random Ring Effective Bandwidth with Adaptive Routing Analysis

May 2020

Note

- **The following research was performed under the HPC-AI Advisory Council activities**
  - Compute resource – HPC-AI Advisory Council Cluster Center

- **The following was done to provide best practices**
  - Effective Bandwidth benchmark on medium-large node count
  - InfiniBand Adaptive Routing effect on this benchmark, focusing on large message size.

- **Source Code of Effective Bandwidth**
  - https://fs.hlrs.de/projects/par/mpi/b_eff/

Natural Ring and Random Ring Tests

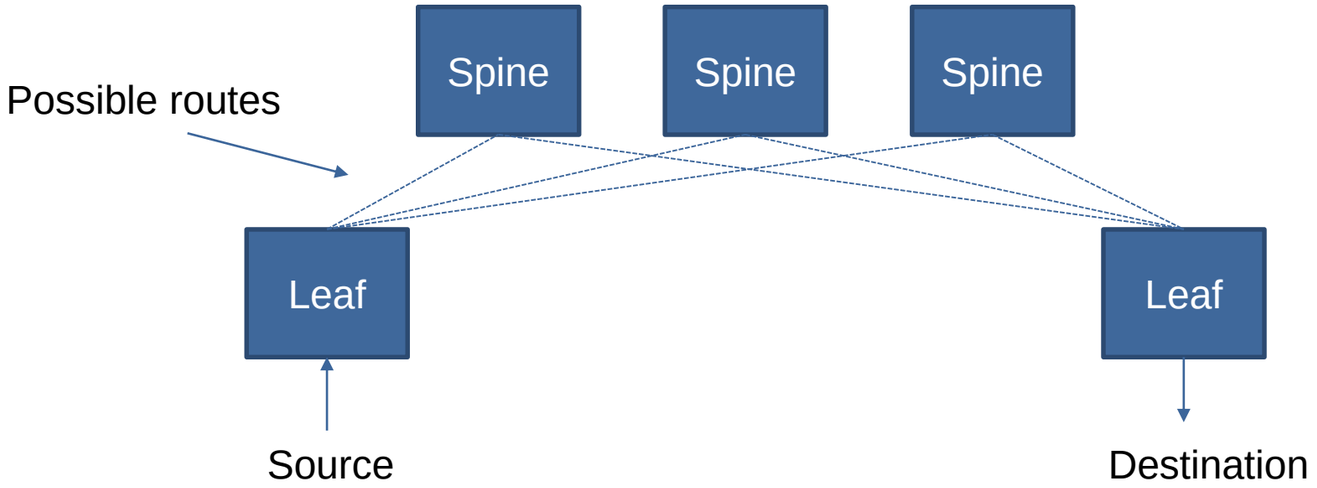
- **The effective bandwidth measures the bandwidth of the communication network of parallel and/or distributed computing systems**
- **Several message sizes, communication patterns and methods are used**
- **The algorithm uses an average to take into account that short and long messages are transferred with different bandwidth values in real applications**
- **The test generate several output tables**
- **The presentation will cover random ring bandwidth table**

  Random Ring Bandwidth – Creates random communication ring
  - Rank i communicates with Rank j and Rank k (randomly selected Rank j,k)
  - …
  - More overlaps between the routes

- **For more details of the test, refer to https://fs.hlrs.de/projects/par/mpi/b_eff/**

Adaptive Routing

- Adaptive routing enables network status to be taken into consideration when choosing the route for a network packet, providing an opportunity for improved fabric utilization
- Adaptive routing also provides enhancements to RAS features of the overall system and used to route around failed links and switches
- When enabled, the leaf switch on the network will select the egress port among the best possible routes available, based on the load on that route



Cluster Configuration 1

- **Setup 1**
  - Dual Socket Intel Xeon Platinum 8260L CPU @ 2.40GHz
  - Mellanox ConnectX-6 HDR InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 192GB DDR4 2677MHz RDIMMs per node

- **Software**
  - OS: CentOS 7.7, MLNX_OFED 4.7-3
  - MPI: HPC-X 2.6.0, UCX 1.8

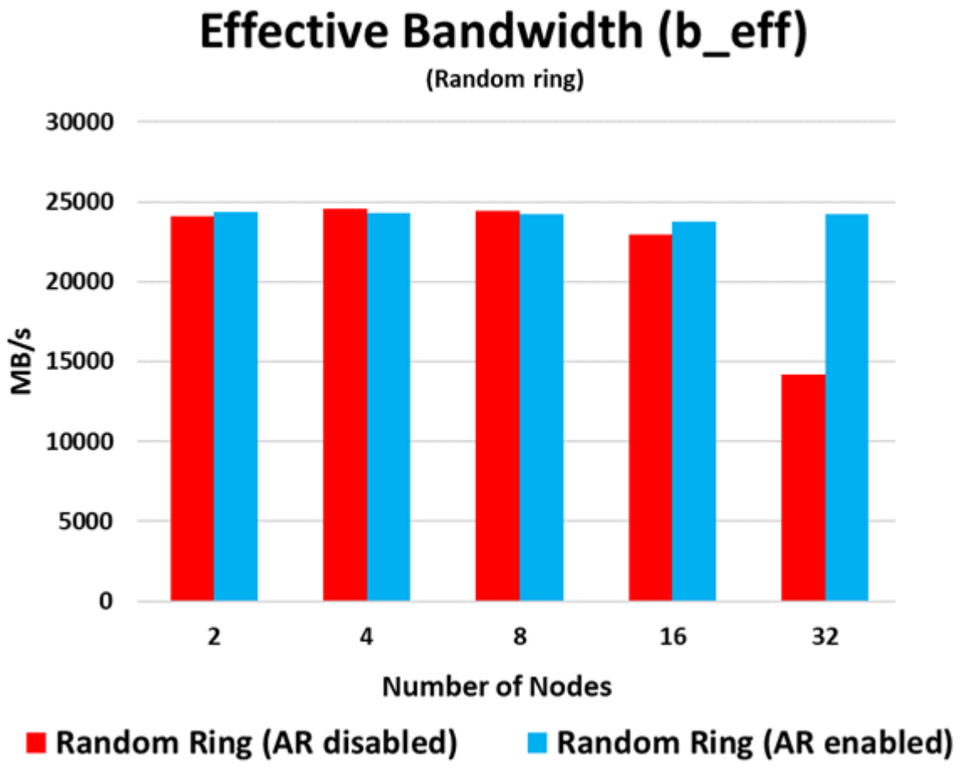
Cluster Configuration 2

- **Setup 2**
  - Dual Socket Intel Xeon Platinum 8280 CPU @ 2.70GHz
  - Mellanox ConnectX-6 HDR100 InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 192GB DDR4 2677MHz RDIMMs per node

- **Software**
  - OS: CentOS 7.6, MLNX_OFED 4.6.1
  - MPI: HPC-X 2.6.0, UCX 1.8

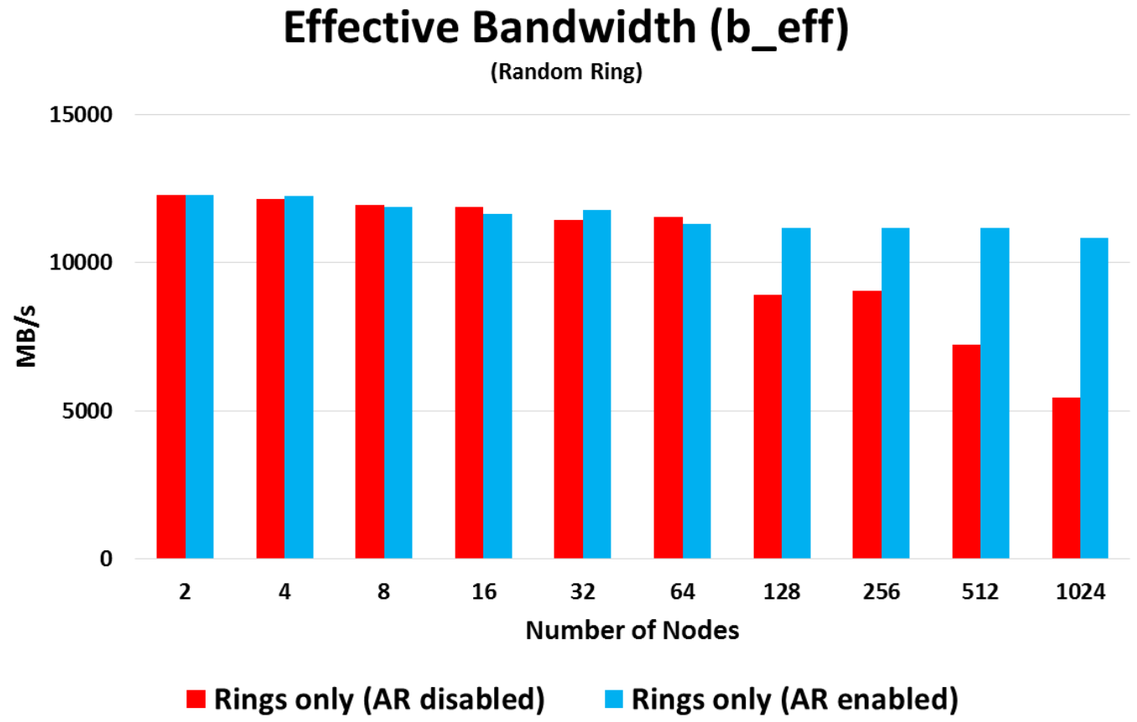
Performance Analysis – Random Ring – Setup 1

- With adaptive routing we observed 99% of the effective bandwidth
- 32 nodes compered to 2 nodes



Performance Analysis – Random Ring – Setup 2

- With adaptive routing we observed 90% of the effective bandwidth
- 1024 nodes compered to 2 nodes



References

- Effective Bandwidth Source Code
- How To Configure Adaptive Routing and SHIELD

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You