NVIDIA InfiniBand Adaptive Routing Technology

Accelerating HPC and AI Applications

Whitepaper

Table of Contents

WP-10326-001_v01

InfiniBand Adaptive Routing Technology

Introduction

The exponential growth of data, the increasingly ubiquitous nature of AI applications, and the rapid expansion of data center infrastructure size are among several trends profoundly impacting today's data center networks. As enterprise and research institutions strive to maximize data center efficiency from all aspects, improving the use of the network resources is perhaps the most impactful technique.

InfiniBand is the preferred choice for world-leading supercomputers, displacing lower performance and proprietary interconnect options. The end-to-end NVIDIA® InfiniBand-based network enables extremely low latencies and high data throughput and message rates. Its high-value features, such as smart In-Network Computing acceleration engines, combined with advanced self-healing network capabilities, congestion control, quality of service, and adaptive routing, enable leading performance and scalability for high-performance computing, artificial intelligence, and other compute and data-intensive applications. The performance advantages of InfiniBand are second to none, while its open industry-standards backed guarantee of backward and forward compatibility across generations ensure users protect their data center investments.

In this white paper, we'll look at how adaptive routing from NVIDIA plays such an important role, eliminating congestion and increasing data center performance.

Avoiding Congestion with Adaptive Routing

The underlying cause of a congested network is quite similar to that of a highway during rush hour—everybody going to work at the same time, resulting in congestion and moving at a snail's pace. InfiniBand is inherently a lossless fabric and switches won't drop packets for flow control. Methods that reduce network congestion include congestion control and quality of service (QoS), both of which are supported by the InfiniBand network and can be very effective at reducing the effects of network congestion; however, there are other techniques included in the arsenal of InfiniBand features for robust performance: adaptive routing and NVIDIA Self-Healing Networking technology.

One of the best solutions for reducing congestion is to spread the traffic across routes—that is what adaptive routing is all about. Adaptive routing determines the optimal path a data packet should follow through a network to arrive at a specific destination. By allowing packets to avoid congested areas, adaptive routing improves network resource utilization, increasing efficiency, and performance.

NVIDIA InfiniBand is a full Software Defined Network (SDN) managed by a software management utility called the Subnet Manager (SM). This centralized entity configures the switches to pick and choose routes based on the network conditions. The switch ASIC selects the least loaded output port (from a set of outgoing ports); in effect, the route that achieves the best performance across the network. "Best performance" is determined by the lowest latency and maximum bandwidth for achieving the highest possible network efficiency.

Adaptive routing maximizes overall cluster performance by spreading the traffic across all network links and increasing links' utilization and balance, thus optimizing link bandwidth. The selection between different outgoing switch ports is based on a grading mechanism that considers egress port queue depth and path priority, where the shortest path has higher priority. Adaptive routing is supported on all types of InfiniBand topologies (for example Fat trees, DragonFly+, and Torus). Adaptive routing can also be configured for only a part of the topology (sub-topology).

Handling Out-of-Order Packets

The role of AR is to redirect traffic toward a less occupied outgoing port chosen from a set of potential ports. This can cause the network packets to arrive at their destination out-of-order. NVIDIA® ConnectX® InfiniBand network adapters (starting with ConnectX-5) include the in-hardware capability to manage out-of-order packet arrivals.

Traffic Classification

InfiniBand Quality of Service is a mechanism designed to allocate bandwidth within the system per service, per virtual lane, or per port. It uses service levels (SLs), virtual lanes (VLs), and per-port arbiters to schedule traffic departing from a given port. An SL categorizes end-to-end data flow, while a VL both categorizes a flow over a given link and is also associated with isolated network resources. By default, the Subnet Manager enables adaptive routing on all SLs. Users may opt to identify specific cases or applications where adaptive routing should not be used by configuring which SLs adaptive routing may or may not be applied. If adaptive routing is disabled on a specific SL, the packets on that SL will go through static routing only.

NVIDIA Self-Healing Networking

NVIDIA Self-Healing Networking technology enables the network to overcome link failures and achieve network recovery 1000X faster than any software-based solution. This technology enables switches to exchange information on link status. If a specific network link is suddenly detected as inactive, the switch connected to this link will broadcast this info to relevant switches in the network, so they can modify their adaptive routing mechanism to avoid selecting a path that may lead to this nonactive link. This mode allows the fastest traffic recovery in case of switch-to-switch port failures due to link flaps, or neighboring switch reboots, without intervention by the Subnet Manager or application downtime.

Configuration and Monitoring

Adaptive routing is enabled by default in networks based on NVIDIA Switch-IB® 2 (InfiniBand EDR), NVIDIA Quantum™ (InfiniBand HDR), and onwards, when deployed with ConnectX-5 or newer versions of host channel adapters (HCAs). This means that all application traffic receives the benefits of adaptive routing.

If a customer prefers using static routing for specific applications, the adaptive routing mechanism should be disabled and enabled per SL. For configuration procedures, monitoring options, and examples, please refer to the community post: How To Configure Adaptive Routing.

Performance Analysis

Micro-Benchmarks Performance

Micro-benchmarks are mini-applications that heavily test a specific function while trying to reach the performance limitation of the cluster.

MPI-GRAPH

In a paper entitled "The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems," published in SC18 by researchers from Oak Ridge National Lab (ORNL) and Lawrence Livermore National Lab (LLNL), adaptive routing was found to achieve 96% network efficiency. The measurements were taken using CORAL's (the Collaboration of Oak Ridge, Argonne, and Livermore) bisection bandwidth benchmark, based on an MPI-Graph that explores the bandwidth between possible MPI process pairs. The performance achieved using adaptive routing reached 11.8TB/s, which is 96% of the maximum bandwidth measured. In contrast, the single path static routing results achieve an average bandwidth of 10.2 TB/s, or only 80% of the maximum measured bandwidth, with much higher diversity, indicating many pairs reach 50% and below bandwidth.

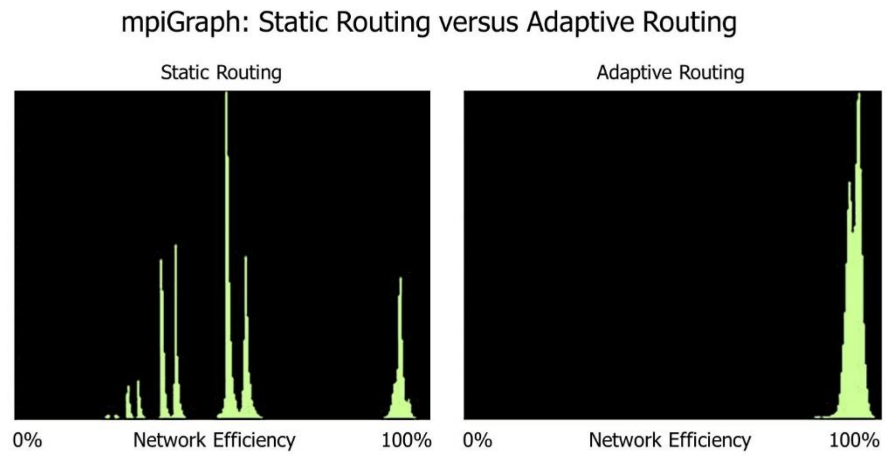

Figure 1. mpiGraph

Effective Bandwidth (b_eff)

The b_eff test created a ring of nodes; each of which, sends traffic of different message sizes to its neighbors.

The test presented in the following figure was measured on a Texas Advanced Computing Center (TACC) Frontera cluster using HDR100 InfiniBand Interconnect. In this example, we can see that once the number of nodes grows in the network, a bottleneck starts in one of the links, which causes a drop in the overall performance. Adaptive routing helps to load-balance the traffic between the network spines and thus, minimizes the bottleneck and improves the performance.

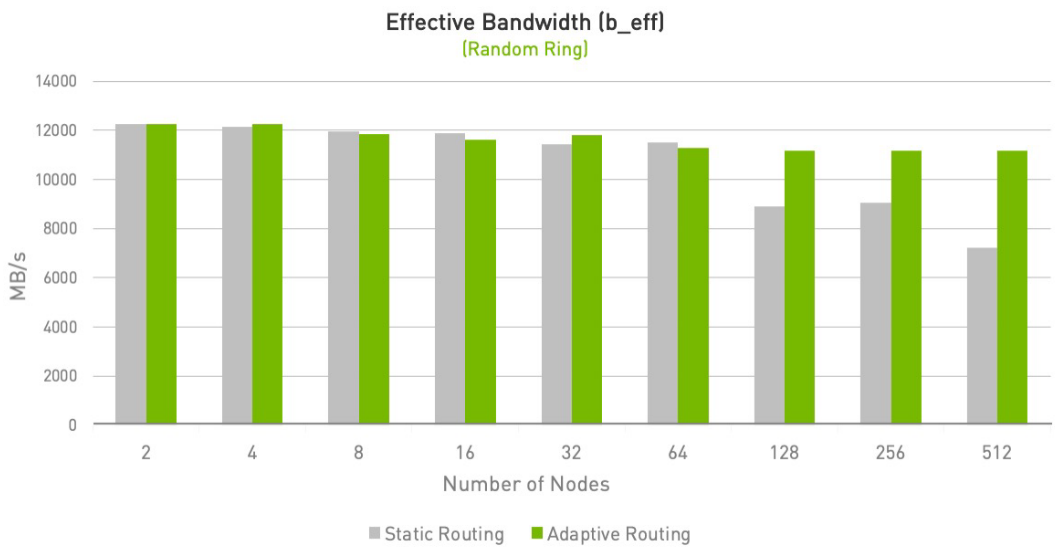

Figure 2. Effective bandwidth: Adaptive routing enabled and disabled

Application Performance

In this section, we review several applications and the effect adaptive routing has on them. We have run these applications on the Texas Advanced Computing Center "Frontera" supercomputer, which include the following elements: Intel(R) Xeon(R) Platinum 8280 CPUs at 2.70GHz, InfiniBand HDR100 network in a Fat-tree 28/26 blocking topology (minor blocking).

VASP

The Vienna Ab initio Simulation Package (VASP) is a computer program for atomic scale materials modeling, such as electronic structure calculations and quantum-mechanical molecular dynamics.

VASP gains a performance improvement of approximately 10% on Frontera when AR is enabled.

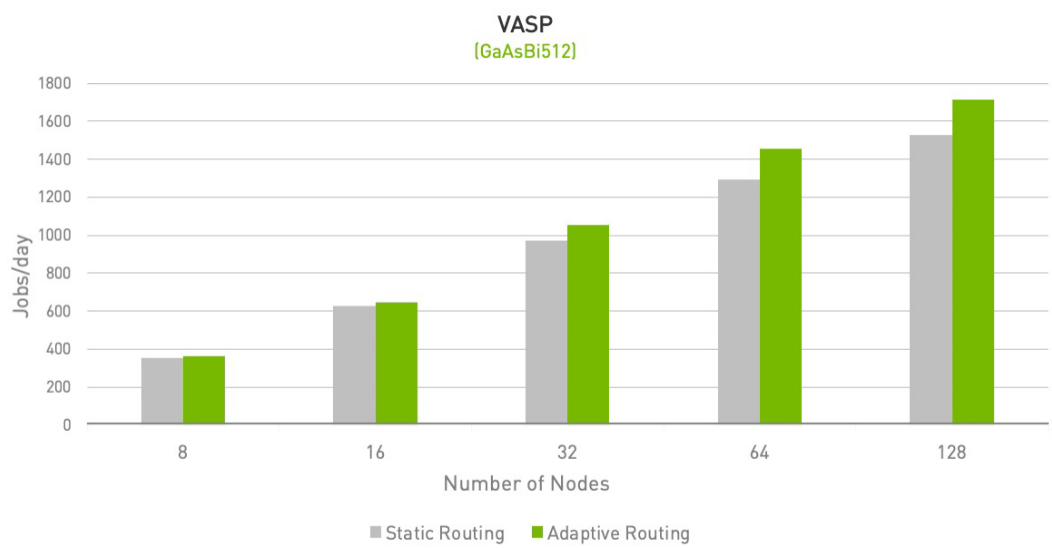

Figure 3. VASP

BSMBench

BSMBench is a flexible and scalable supercomputer benchmark from computational particle physics. As an open source benchmarking tool, it includes the ability to tune the ratio of communication to computation.

BSMBench is used to simulate workloads such as Lattice Quantum ChromoDynamics, and by extension, its parent field Lattice Gauge Theory; these make up a significant fraction of global supercomputing cycles.

The results demonstrated a 28% improvement with adaptive routing.

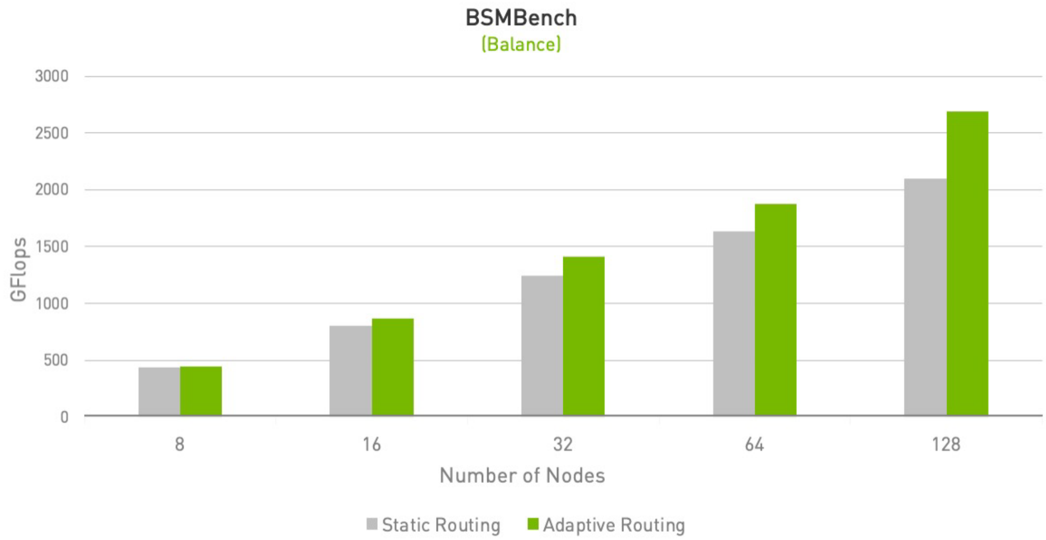

Figure 4. BSMBench

Summary and Conclusions

High-performance computing (HPC) and AI are the most essential tools fueling the advancement of science. To handle the ever-growing demands for higher computation performance and the increase in the complexity of research problems, the network needs to maximize its efficiency.

InfiniBand adaptive routing technology reroutes data to eliminate congestion, and therefore, increases data center performance. As presented, both HPC applications and AI applications utilizing adaptive routing achieve higher performance. Adaptive routing is an important network element that drives your HPC systems toward new levels of utilization that increase return on investment.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, Switch-IB, and ConnectX are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

DOCUMENT

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com