Understanding Congestion on Omni-Path Fabrics

Dan Cassidy
University of New Mexico
dcassidy@lanl.gov

Lauren Gillespie
Southwestern University
gillespl@southwestern.edu

Christopher Leap
University of New Mexico
cleap@unm.edu

LA-UR-17-26948

Overview

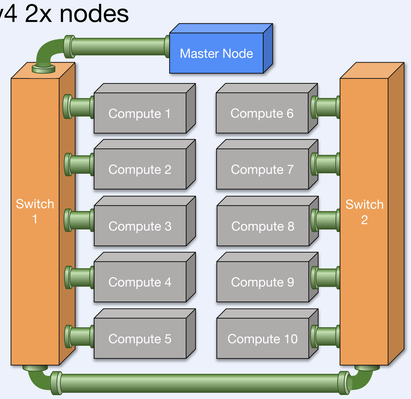
Recently, Intel has released a new high-speed interconnect similar to InfiniBand (IB) called Omni-Path Architecture (OPA). Both interconnects have counters to monitor fabric performance but OPA has a new error counter, Congestion Discards (CongDiscards), that specifically calls out congestion-related packet discards. The goal of this research is to better understand the effect congestion discards have on cluster performance by adjusting three different congestion parameters and comparing system performance to congestion discards. During this study, combined queue pair and psm2 traffic was observed to significantly lower performance without producing discards, while IPoIB and psm2 traffic showed discards but maintained performance. The interaction between verbs and psm2 traffic has implications for OpenMP work.

Methodology

- Compare bandwidth & latency performance to CongDiscards
- Modified HoqLife & SwitchLife parameters separately
- Also compare performance between IPoIB and verbs traffic
- IPoIB background noise: dd 29 GB file between each node pair
- Verbs background noise: run perftest between each node pair
- Bandwidth benchmark: OSU Bi-Directional Bandwidth
- Latency benchmark: OSU Alltoall Latency Test
- Ran each test for 50 runs

System Setup

- 11 Intel® Xeon® CPU E5-2620 v4 2x nodes
- OPA firmware: 10.3.0.0.66
- OPA software: 10.4.2.0.7
- OSU Benchmark: 5.3.2
- HFI driver: 0.9-294
- OpenMPI: 1.10.4
- OS: CentOS 7.3
- PerfTest: 3.0-7

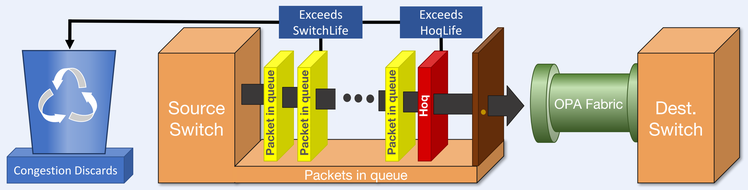


Fabric Topology. The fabric consists of 2 switches with a single interlink between and 5 compute nodes per switch. The compute nodes are paired with a counterpart on the other switch and communicate pairwise across the interlink to force congestion.

Congestion Discards

- CongDiscards are due to three settings:

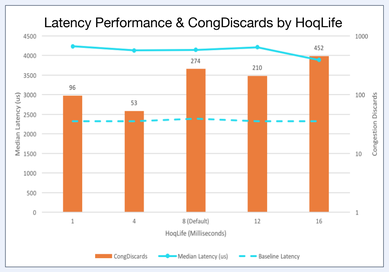
1. Head-of-Queue Lifetime (HoqLife): time for hoq packet to enter fabric
2. Switch Lifetime (SwitchLife): time for queued packet to enter fabric
3. VLStallCount: number of consecutive discards before queue flushed

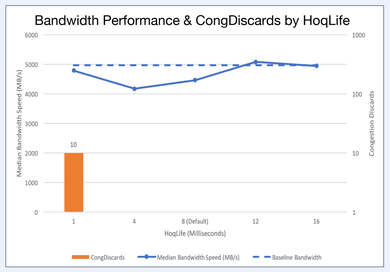


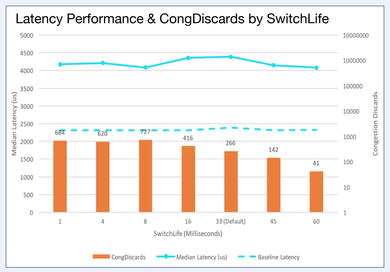
Discarding Packets. On the source switch, if the leading packet, also known as Head of queue (Hoq) does not enter the fabric in HoqLife milliseconds because the fabric is full, the packet is dropped and must be resent producing a congestion discard. If packets are queued for longer than SwitchLife, the queued packets are likewise dropped and resent.

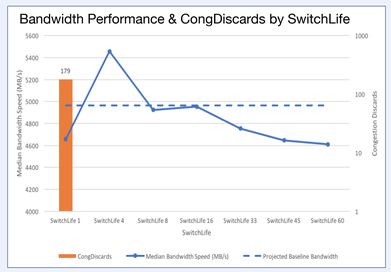
Congestion Results

- CongDiscards shows little correlation to network performance
- Saw bandwidth spikes and drops with no discards
- Also saw large latency spikes with many congestion discards
- HoqLife & SwitchLife 0 congestion broke the benchmarks
- Heavy network traffic led to performance degradation with or without discards



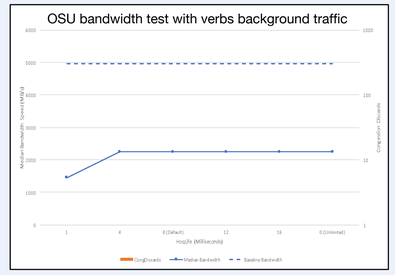


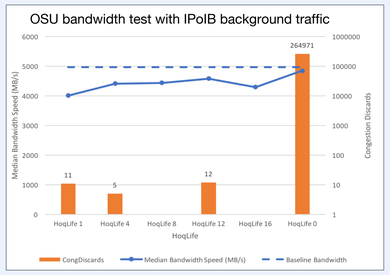




Verbs and PSM2 Results

- Verb queue pairs and psm2 traffic led large drop in bandwidth
- However, did not produce any congestion discards
- IPoIB and psm2 traffic had discards but bandwidth did not drop
- This interaction may be due to a bottleneck on NIC, not fabric





Conclusion

Performance had little correlation to congestion discards except when HoqLife or SwitchLife were 0 (disabled). SwitchLife & HoqLife 0 had such poor performance that the OSU benchmarks could not run effectively and never completed. However, our data shows that performance can still be negatively affected by heavy network traffic even without the presence of congestion discards. In conclusion, congestion discards alone are not indicative of network performance degradation. However, verbs and psm2 traffic led to significantly worse network performance but produced no discards. The bottleneck is no longer on the fabric but the NIC.

Future Work

- Experiment on larger cluster with different congestion patterns
- Study effect of VLStallCount parameter on congestion discards
- Further investigate impact of verbs interaction with psm2 library
- Understand effect of different psm2 protocols on performance

Acknowledgements

We would like to thank our mentors Susan Coulter, Jesse Martinez, and Howard Pritchard. We would also like to thank Brian McGinnis from Intel for assisting us with tuning our system. Finally, we would like to thank our instructor Joan Lucas, our teaching assistant Hunter Easterday and our program coordinators Amanda Bonnie and Alfred Torrez.

ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE

Southwestern

Los Alamos NATIONAL LABORATORY EST.1943