Evaluating TCP Protocol Performance on High-Speed Networks

LA-UR

LA-UR-22-27685

Abstract

High performance computing (HPC) clusters rely on specialized low-latency, high-bandwidth communication networks to enable fast data transfer between compute nodes, with minimal processor overhead. This is implemented using a co-processor in the network card with direct access to system memory, which allows for modification of RAM contents without involving the CPU. These cards are connected via high-speed cabling and switches to form a cohesive network fabric with an architecture fundamentally incompatible with that of traditional Ethernet. Software interoperability layers exist, though they reintroduce the processor overhead that these networks were designed to avoid, increasing latency and reducing bandwidth. Since many HPC services, particularly high-performance filesystems and data transfer mechanisms, communicate via internet protocol (IP) networks, this reduced speed can be detrimental to system performance and utility. Here, we use a cluster of ten compute nodes and a single master node, connected by a Mellanox InfiniBand fabric, to evaluate the base performance of IP over InfiniBand (IPoIB) connections and fine-tune system parameters to maximize IPoIB link throughput. We also evaluate the effects of IPv6 addressing, and of relaying data to an Ethernet network through an intermediate input/output node. After system tuning, we consistently match vendor performance estimates (achieving roughly 3.2 times the out-of-box bandwidth), and our results suggest that IPoIB bandwidth on a capable host system can begin to approach that of a raw InfiniBand connection. The performance gains demonstrated here may enable future HPC systems to more efficiently communicate with IP networks using existing high-speed fabrics, reducing costs and maintenance requirements associated with dedicated Ethernet hardware.

Authors

Noah Jones
Jerrod Parten
Lucas Ritzdorf
Jesse Martinez
Thomas Areba
Chase Harrison