Pleiades Interconnect

Category: Pleiades

Topology

InfiniBand (IB) is used for inter-node communication among all of the Pleiades nodes. A key feature of InfiniBand permits remote direct memory access (RDMA) between processing nodes, allowing direct access to other nodes' memory. This allows developers and application owners to bypass the TCP/IP stack, accelerating the application performance. Two devices are involved in the interconnect: the Mellanox ConnectX host channel adapter chip (HCA) on the motherboard of each node and the Mellanox IB switches. There are two IB switches per half- IRU (which includes 8 nodes). One of the switches is involved in the ib0 fabric, which is used mainly for MPI communication. The other is involved in the ib1 fabric which is used mainly for I/O. InfiniBand uses subnet manager (SM) software to manage the InfiniBand fabric and to monitor interconnect performance and health at the fabric level.

The network topology of each IB fabric of Pleiades is a partial 11-D hypercube. In a 11-D hypercube, each switch has 11 direct connections with 11 other specific switches in the network.

The ib1 hypercube fabric is extended by a set of nine switches connected to the Lustre servers (one for the MDSes, and eight for the OSSes). The plan is for each rack to connect directly to one of the OSS switches and each group of eight racks to connect directly to the MDS switch. Currently, most of the racks are connected this way, but some remain to be connected.

Another set of nine switches on the ib1 fabric provides direct access between hyperwall visualization nodes and Pleiades nodes and Lustre servers.

Latency

The shortest communication path in a Pleiades IB fabric will be for any two nodes located in the same half-IRU of the same rack such that the communication only needs to go through 1 switch. For a fully populated 11-D hypercube, the optimum communication path between any two nodes which are not in the same half-IRU varies from going through 2 to 12 switches, depending on which racks and half-IRUs the two nodes reside. Since the Pleiades IB fabric is not a full 11-D hypercube, some connections are missing that would facilitate the optimum path between some nodes, therefore, it is possible that some communications may go over more than 12 switches.

MPI half Ping-Pong latency starts around 1000 to 1500 ns for communication going through two switches. Each additional switch adds ~100 ns (QDR) to ~150 ns (DDR) to the latency.

Bandwidth

The HCA on each node uses either 4x DDR (double data rate) links or 4x QDR (quad data rate) links. Each link is bi-directional and contains 1 send channel and 1 receive channel. For each direction, the raw data transfer rates for 4x DDR and 4x QDR are 20 Gb/s and 40 Gb/s, respectively. These links use 8b/10b encoding such that every 10 bits sent carry 8 bits of useful data. Thus, for each direction, the effective maximum bandwidth for each node is 16 Gb/s (i.e, 2 GB/s) if 4x DDR HCA is used or 32 Gb/s (i.e. 4 GB/sec) if 4x QDR HCA is used.

The IB switch which every 8 nodes in each half-IRU share through a single path also has similar effective data transfer limits per port: 16 Gb/s for 4x DDR IB switches and 32 Gb/s for 4x QDR IB switches.

The Harpertown and Nehalem-EP nodes use 4x-DDR HCAs while the Westmere nodes use 4x-QDR HCAs. The Harpertown racks use 4x DDR IB switches while the Nehalem-EP and Westmere racks use 4x QDR switches.

These limits also apply to each OSS in the Lustre filesystem. For /nobackupp20 and /nobackupp50, QDR switches are used to connect to the IB fabric and DDR switches are used to connect to the DDNs of the hard disks. For /nobackupp[10,30,40,60], DDR switches are used to connect to both IB fabric and to the DDNs. With DDR switches, the theoretical bandwidth of each OSS is 2 GB/s for each direction. With 8 OSSes per Lustre filesystem, the theoretical peak aggregate bandwidth for each filesystem for each direction would be 16 GB/s. This bandwidth however is reduced to 10 GB/s due to bandwidth that the DDNs can provide. The best benchmark performance obtained for each Pleiades Lustre filesystem is 8 - 10 GB/sec (all read or all write).

The actual I/O bandwidth a user's application experiences is far less than the theoretical peak or even the benchmark data due to factors such as the I/O pattern the application is doing (for example, serial or parallel; for parallel, if the I/O requests are from nodes of different half-IRUs), the number of stripe count used (this affects the maximum aggregate bandwidth provided by the OSSs), how busy the Lustre is handling requests from many users, if there are bad links in the network, etc.

Follow the tips listed in Lustre Best Practices if you are not getting good performances out of the Lustre filesystem.

Article ID: 237
Last updated: 19 Dec, 2011
Computing at NAS -> Computing Hardware -> Pleiades -> Pleiades Interconnect
http://www.nas.nasa.gov/hecc/support/kb/entry/237/?ajax=1